Advances in Feature Selection in One- and Two-Dimensional Gas Chromatography with Mass Spectrometry

Kelsey L. Berrier

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Robert E. Synovec, Chair

Matthew F. Bush

František Tureček

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Advances in Feature Selection in One- and Two-Dimensional Gas Chromatography with Mass Spectrometry

Kelsey L. Berrier

Chair of the Supervisory Committee:
Professor Robert E. Synovec
Department of Chemistry

One- and two-dimensional gas chromatography coupled with mass spectrometry provides an enormous amount of quantitative data describing the chemical composition of complex samples. Besides quantification and identification of analytes, common analysis goals include classifying samples or predicting sample properties based upon the chemical information contained in the chromatographic data. The chemometric modeling techniques used to accomplish these goals often benefit from the removal of redundant or irrelevant chromatographic variables, which is achieved by feature selection. This dissertation presents several research studies detailing advances in and applications of feature selection applied to one- and two-dimensional gas chromatography with mass spectrometric detection. The two-dimensional mass cluster method was evaluated as a peak detection algorithm using simulations

of gas chromatography coupled with time-of-flight mass spectrometry (GC-TOFMS) data under varying sample and separation complexity. An unsupervised feature selection method based on variance thresholding was applied to simulated GC-MS chromatograms and a previously studied yeast metabolome dataset. A successful application of partial least squares (PLS) regression analysis to comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC-TOFMS) for the prediction of bulk physical properties of kerosene-based fuels is included to demonstrate a case where feature selection was not required. Finally, supervised feature selection was implemented on GC×GC-TOFMS data of rocket fuels to aid in the prediction of fuel thermal integrity by PLS.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my graduate research advisor, Rob Synovec. I will forever be grateful for the opportunity to do research under your supervision and for your guidance and support along the way. Secondly, thank you to my undergraduate research advisor at JMU, Dr. Downey. You facilitated my interest in analytical chemistry and helped me through an Honors Thesis. I will always cherish my time in the lab at JMU and in the field throughout the Shenandoah Valley.

Thank you to the many colleagues I have had the pleasure of working with over the past five years. To Brooke, thank you for being an awesome role model and gently pushing me into my first simulation project, because it was exactly what I needed. You were someone who I could always come talk to, whether it was about research or life, and you truly shaped my first years in graduate school. To Dave, thank you for answering my multitude of questions thoughtfully and thoroughly, and for generally being a sounding board throughout my time in graduate school, even after you had graduated. You have always been a fierce advocator and supporter, even when, and maybe especially when, I was most doubting of myself. To Nate, thank you for bringing your energy into the group and for being unapologetically candid. To Chris, thank you for all of your advice and for the great music and conversation. To Sarah, thank you for being present on this journey with me. To Derrick, thank you for the dad jokes and your generosity. To Warren, thank you for your friendship and for being my office buddy. Thank you to Brendon, Dan, Nick, Paige, Sonia, Grant, Tim, and Caitlin for being great group members and scientific collaborators.

Thank you to everyone who has seen me through this time: family, friends, and pets. Mom and Dad, thank you for all of the support, inspiration, and love it took to get me here. Kyle and Kaleb, thank you for being my brothers and for all the laughs. To Katie, Steph, Meghan, Vicki, Ashley, and Danika, thank you for listening and for always being there. To all the amazing people I have met along the way, of which there are too many to list individually, thank you for bringing your light into my life and making my time here more enjoyable. Thank you to my therapist, who listened patiently to my ranting and provided me with validation and the tools to better myself. Thank you to all the pets in my life, especially to my chocolate lab in Virginia, Coda, and my cat here in Seattle, Alcatraz. You are so loving and accepting, and I am deeply grateful for your being in my life. Finally, to Eedann, thank you for coming into my life and filling it with joy, love, and passion. Thank you for all of your sacrifices, for listening to my science and engaging in discussions with me, for understanding my challenges and supporting my endeavor to face them, for commiserating the failures and frustrations, and for celebrating the victories with me. Thank you.

DEDICATION

For my good bois, Coda and Alcatraz

Chapter 1. Introduction to Gas Chromatography and Advanced Data Handling Methods[1]

1.1 GAS CHROMATOGRAPHY

1.1.1 *Fundamentals*

Gas chromatography (GC) is an analytical separation technique used to separate complex mixtures of volatile and semi-volatile chemical compounds in the fields of petrochemistry [1–5], metabolomics [6–9], flavor and food [10–12], environmental [13–16], pharmaceutical [17,18], and forensics [19–21], among many others. In GC analysis, small volumes of sample mixtures containing semi-volatile and volatile analytes are vaporized (transferred into the gas phase) and then carried through a separation column by an inert gaseous mobile phase (usually helium, hydrogen, or nitrogen), where the analytes are subject to separation mechanisms based on the properties of the column. Generally, GC separation columns are comprised of narrow fused silica tubes with a film (liquid or polymer stationary phase) coating the inner wall. These are known as capillary columns and come available in many varieties of lengths (1-100 m), inner diameters (50-530 μm i.d.), film thickness (0.1-5 μm), and stationary phase composition. Analytes partition in and out of the stationary phase based on their individual affinities; the more time that an analyte spends in the stationary phase relative to the mobile phase, the longer that analyte is retained on the column, leading to the separation of analytes in a mixture. The amount of time it takes an analyte to exit the column is referred to as its retention time (t_R). Common separation mechanisms include boiling point and polarity, where compounds with lower boiling points elute sooner and

[1] Portions of this Chapter have been adapted from K.L. Berrier, S.E. Prebihalo, and R.E. Synovec, "Advanced Data Handling in Comprehensive Two-Dimensional Gas Chromatography" from Basic Multidimensional Gas Chromatography, Edited by Nicholas Snow (2020).

compounds with similar polarity to the stationary phase elute later (e.g., a polar compound will be more retained on a polar column and less retained on a nonpolar column, given that all other separation parameters remain the same).

Following the GC separation column is the detector, which occurs on-line as the separation progresses. Common detectors include universal detectors such as the mass spectrometer (MS) and the flame ionization detector (FID), which is nearly universal in that it responds to compounds with C-H bonds. Mass spectrometry itself is an established analytical technique, offering additional chemical selectivity when paired with GC. Analysis of an analyte by mass spectrometry is accomplished by ionization (e.g., through bombardment with electrons, known as electron ionization) followed by fragmentation of the analyte molecule into characteristic charged fragments, which are then separated according to their mass-to-charge ratio (*m/z*) and detected via signal amplification of the ions (e.g., using an electron multiplier). The result is a mass spectrum containing the signal intensities of the detected fragments, which is unique and reproducible for a given analyte. Regardless of the detector, there exists a range over which the measured signal is linear with analyte concentration; the range itself depends on the detector. With the presence of a detector, GC can be transformed into a powerful analytical platform, capable of the separation, detection, quantification, and identification of analytes in a diverse range of samples.

Typically, the detection of an analyte results in a peak that can be approximated by a Gaussian distribution, which takes the form

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{1.1}$$

where μ is the mean and σ is the standard deviation of the distribution. The t_R of a peak is defined at μ, which corresponds to the peak maximum, and the width-at-base (w_b) is defined as 4σ, corresponding to ~95% of the peak area. The reason for this distribution is due to band broadening,

the mechanism by which an initial narrow injection plug becomes diffuse over time and space during the separation, turning into a band of analyte molecules with the majority of molecules concentrated in the center of the band. Longer separation times (e.g., resulting from a long column, low flow rate, high affinity for the stationary phase) can exacerbate these effects in gas chromatography. Therefore, selection of appropriate column properties for the analysis at hand, such as length, inner diameter, film thickness, and stationary phase composition, and using an optimal flow rate for the mobile phase can help to mitigate broadening, although it is a complicated process with many conflicting considerations. A temperature program is also commonly implemented to avoid the general elution problem, in which the peaks of later eluting analytes become significantly more retained, wider, and shorter when the separation is optimized for early eluting compounds. Typically, the temperature is held at an initial temperature for a short time and then increased at one or more rates to a final temperature. Advances in injection techniques [14,22–26], column specifications [27–33], and detectors that provide high sensitivity and fast scan rates [34–36] have improved separations by reducing peak widths and allowing for more chemical information to be packed into shorter analysis times. Despite these and many other advances, GC separations are consistently plagued by component overlap [37], where two or more analytes elute with similar or identical retention times. The addition of a secondary separation dimension can help to mitigate this component overlap.

Comprehensive two-dimensional (2D) gas chromatography (GC×GC) is a powerful instrumental platform that is utilized to address the challenges that arise from the analysis of complex samples [38,39]. In GC×GC, a shorter column (1-5 m) with a complementary stationary phase is placed after the longer primary column. After the primary separation, analytes are trapped, refocused, and reinjected (modulated) onto the secondary column via a modulator.

Common modulators include valves and cryogenic modulators that utilize alternating cold and hot jets of nitrogen to achieve modulation. The secondary separation is much shorter than the primary separation, usually on the order of a few seconds and defined by the modulation period (P_M). Typically, the analyst aims to sample the first dimension peak 2-4 times, although larger sampling rates are also used; this is known as the sampling density (ρ_S) or formerly the modulation ratio (M_R). The result is two distinct separations, ideally based upon orthogonal separation mechanisms, where each peak can be defined by its retention time on the first dimension (^{1}D) and second dimension (^{2}D). Generally, the ^{1}D separation is achieved mainly based on boiling point, with the ^{2}D separation more dependent on the chosen stationary phase. The fundamental basis of how GC×GC relies on a temperature-based separation in both dimensions often results in highly structured 2D chromatograms, with chemical compound classes beautifully separated from each other. Fossil fuel sample separations are a prime example of this effect [4], whereas for example, metabolomics separations generally produce a more random appearing distribution of peaks in the 2D space [40].

1.1.2 *Figures of Merit*

There are several quantities associated with chromatographic separations that are used to describe the quality of separation and allow for comparisons when different separation parameters or instruments are used. Resolution (R_s) is used to describe the degree of chromatographic separation between two analytes and is defined as

$$R_S = \frac{\Delta t_R}{w_{b,avg}} = \frac{t_{R,2} - t_{R,1}}{\frac{1}{2}(w_{b,1} + w_{b,2})} \quad (1.2)$$

An R_s of 1.5 is considered baseline resolved, whereas an R_s of 1 corresponds to ~2% of overlap for normally distributed peaks with the same height. Peak capacity (n_c) is a metric that describes

the theoretical maximum number of peaks that can be separated in a given amount of time with the caveat that they are all spaced at a given resolution, usually an R_s of 1, given by the equation

$$n_c = \frac{t_{sep}}{w_{b,avg} \times R_s} = \frac{t_{sep}}{w_{b,avg}} (R_s = 1) \quad (1.3)$$

For GC×GC, peak capacity can be determined on both separation dimensions, e.g., ${}^1n_c$ and ${}^2n_c$, to yield an overall 2D peak capacity ($n_{c,2D}$) equal to ${}^1n_c \times {}^2n_c$.

The saturation factor (α) of a separation can be thought of as how "full" the separation is relative to its peak capacity, and is defined by

$$\alpha = \frac{m}{n_c} \quad (1.4)$$

where m is the number of components (i.e., analytes) in the separation. For most applications, this is unknown, so the separation factor is a more useful metric for theoretical studies and fundamental investigations of separations.

1.1.3 *Challenges*

GC separations generate sizeable data files, considering that data points are collected every 10 to 100 ms (correlating to common data collection frequencies of 100 and 10 Hz, respectively) over a 30 minute time period, on average. The comprehensive coupling of a secondary separation dimension to the primary separation dimension significantly increases the number of data points contained in the data file for a single sample. When the GC instrument is coupled with a multichannel detector such as a time-of-flight mass spectrometer (TOFMS), the size of the data files increase even further, resulting in an enormous data set for analysis especially when multiple samples or replicates are required.

Besides the potential enormity of the data files, the order or dimensionality of the data also increases with additional separation dimensions and multichannel detection. Following data

collection, the goal of data analysis is to answer complex analytical questions from these information-rich data sets. This is exceedingly difficult without the help of advanced data handling methods known as chemometrics, which aim to discover meaningful information from chemical data sets using mathematical means. Chemometric methods utilize linear algebra and statistical concepts to reveal underlying chemical relationships in the data that are related to the experimental design [41–44].

Chemical analysis is subject to challenges related to the chromatographic conditions applied with the 1D or 2D GC instrumentation. Optimization of the experimental and instrumental designs is imperative to obtain reliable and informative data sets that ultimately address the analytical goal(s) and question(s) [45]. Briefly, the goal is to separate as many analytes as possible concurrent with utilizing a majority, if not all, of the theoretical peak capacity n_c, which is a suitable metric that relates to how much information (i.e., peaks) can theoretically fit into a separation. Full use of the n_c is advantageous for meeting the ultimate goals for chemical analysis through advanced data handling.

There are many considerations that can be taken to improve n_c. Some of these include decreasing peak widths or increasing separation time. There are often other competing factors, such as hardware limitations or time constraints that complicate these approaches. Regardless of separation optimization, peak overlap is an unavoidable challenge in separations of complex mixtures. Based on a statistical theory presented by Davis and Giddings, the severity of peak overlap is dependent on the number of compounds to be separated and the available peak capacity [37]. Theory predicts that separations with more compounds and/or a smaller peak capacity will have greater peak overlap, manifesting as a smaller number of apparent peaks. This means that an observable peak in a separation may in fact be the consequence of several overlapped analyte

signals. Deconvolution is required to resolve overlapped analytes and isolate the individual signal contributions from each analyte, of which one or several may prove to be important sample distinguishing features in downstream analysis. Selective detection, such as mass spectrometry, can in some cases also be used to quantify overlapping analytes if fully selective mass channels (*m/z*) are present.

Other inherent properties of the chromatographic data can also present challenges in applying advanced data handling. These include baseline effects, noise, retention time shifting within and between samples, and other artifacts that require additional preprocessing prior to data analysis. Preprocessing of the raw data is a ubiquitous and necessary step in every data analysis procedure. The goal of preprocessing is to retain real, relevant chemical variation (i.e., information) while removing chemically-unrelated variation and noise that may mask the important chromatographic features. Common data preprocessing includes baseline correction, smoothing, retention time alignment, and normalization [43]. The former two address low frequency and high frequency noise, respectfully, and should be applied properly to prevent the introduction of new artifacts or excessive reduction of data density. Retention time alignment and normalization are particularly important preprocessing steps prior to cross-sample analysis, where differences between samples must be reflected accurately. Various alignment algorithms are available to correct retention time shifting between (and within) chromatographic runs so that the chromatograms may be compared on a pixel-based level [46]. Alterative options for dealing with retention time shifting include binning and tiling schemes that act on pixel-level data, yet result in an overall reduction in the size of the data files [47]. Normalization is also used to correct for variation between samples (e.g. caused by injection or sample preparation) so that representative comparisons may be made.

1.2 Advanced Data Handling Methods

1.2.1 *Data Analysis Goals*

Within the greater goal of answering specific analytical questions are the common analysis goals of analyte identification and quantification. When the analyst knows the analytes of interest beforehand, this is known as targeted analysis. Targeted analysis often does not require advanced data handling unless the targeted analytes are difficult to find in the 1D or 2D separation, since the end goal is principally to quantify analytes identity is known *a priori* [48]. For example, there are still challenges that may present in the data and call for the use of chemometrics to achieve targeted analysis goals. Most commonly, chemometric techniques referred to as deconvolution methods are applied in these situations to aid in the mathematical resolution, identification, and quantification of target analytes that may be overlapped on one or both chromatographic dimensions. Deconvolution may also be applied in non-targeted analysis, where the analyst does not know *a priori* what chemical features are of interest or relevant to the greater analytical question. In these cases, deconvolution may only be the first step in a non-targeted approach, also aptly referred to as discovery-based analysis. Non-targeted analysis goals include classifying samples based on chemical composition, discovering key chemical features or biomarkers that differentiate samples, or predicting sample properties based upon chemical measurements [49]. Chemometric methods that are applied to achieve these goals are referred to as pattern recognition methods. Data analysis may be approached in several different ways: on a pixel-level, peak table, or peak region basis. The latter two approaches are mainly provided by commercial software, which takes advantage of structured chromatograms to deliver the peak region analysis approach. Pixel-level data analysis consists of working with the raw data on the data point level.

1.2.2 *Deconvolution*

Component overlap in 1D GC separations is ubiquitous, and despite the increased peak capacity and selectivity offered by the additional separation dimension, peak overlap is still expected to occur [37] and is prevalent in GC×GC separations. Fortunately, the use of deconvolution methods to mathematically resolve analytical signals in regions of overlap can provide additional information that would otherwise be obscured. These deconvolution methods are based in linear algebra and can computationally separate analyte peaks on the separation dimension(s), as well as the spectral dimension if spectral detection is utilized. Deconvolution is applied to improve peak detection, reduce noise, remove background/baseline contributions, provide confident analyte identification, and assist with quantification, including calibration. It can also be applied prior to additional advanced data handling to provide a peak table for pattern recognition methods. Deconvolution methods can be used in either a targeted or non-targeted fashion. In other words, deconvolution can be used to "extract" the signal of a given target analyte from a region of overlap for quantification purposes, or alternatively, deconvolution can be applied to a chromatographic region of overlap where the peak identities are unknown and the goal is to identify those analytes. Due to the heavy computational requirement, some deconvolution algorithms are not well suited to be applied to entire chromatograms. Instead, a region of the chromatogram is selected and the deconvolution algorithm is applied to that region. Deconvolution may be performed on a given region in a single chromatogram or across multiple chromatograms. The most common deconvolution methods are parallel factor analysis (PARAFAC/PARAFAC2) [50,51] and multivariate curve resolution with alternating least squares (MCR-ALS) [52,53]. These methods have different data structure linearity requirements (i.e., data bilinearity or trilinearity) and are therefore appropriate in different analysis situations depending on the data to

be analyzed. Because of this, instrumental design and/or parameters, such modulation period, temperature programming rate, and sampling density, can have a large impact not only on the resulting data, but also on the success of applying various deconvolution methods.

Multivariate Curve Resolution with Alternating Least Squares (MCR-ALS)

MCR-ALS is an iterative chemometric resolution method that decomposes a chromatographic two-way array into the product of two matrices containing pure component information for each dimension (generally chromatographic retention time is in the rows and spectral information is in the columns, but not always). The algorithm works by making an initial guess of each dimension, testing for convergence, and then alternatively iterating the values for each dimension until the convergence criterion is met. The user must provide the data to be deconvoluted and the rank of the solution matrices (i.e., the number of analyte components in the region to be deconvoluted and/or anticipated number of factors in the model), but does not need to know the identity of the components. If the number of components is unknown, multiple MCR-ALS models can be built by varying the number of components and the best model can be selected. Alternatively, singular value decomposition (SVD) can provide information about the chemical rank of the data, i.e., the number of components [54]. In some applications, such as targeted deconvolution, the user may not be concerned with appropriately modeling all components and may instead be focused on modeling only one or a few particular analytes. Therefore, criteria for the selection of the "best" model will depend on the analyst's specific goals. An initial guess for the unmixed solution of either dimension for each component may be input to improve the likelihood of obtaining a meaningful result, but is not required. Methods such as simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) can be used to obtain pure estimates of

the profiles of one dimension to initialize ALS optimization [55]. The user should also choose appropriate constraints, such as unimodality, nonnegativity, component correspondence, spectral normalization, and convergence [56].

MCR-ALS and PARAFAC operate in essentially the same way; however MCR-ALS requires bilinear data instead of trilinear data. Bilinear and trilinear decomposition methods require each of the data dimensions to be linearly independent from the others. Additionally, at the pixel-level, the data must take the form of a linear combination of analyte signals that are unique, reproducible, and concentration-dependent for each analyte present. Second-order (or two-way data) such as GC-TOFMS or GC×GC-FID are well-suited for bilinear methods, whereas third-order or three-way data such as GC×GC-TOFMS generally are well-suited for trilinear methods. The second-order and third-order advantages are such that the analyst need not provide as much information as when working with lower order data, e.g. decomposing first-order data such as GC-FID by classical least squares (CLS) requires user input of a pure reference spectrum for each component in the mixture. Specifically, the second-order advantage is the ability to perform quantitative analysis of analytes of interest in the presence of unknown interferences, and the third-order advantage provides this ability with a single sample. It is also possible to increase the order of the data so that higher order chemometrics can be applied. For example, analyzing multiple samples can change second-order data into a third-order data structure. Alternatively, the order of the data can be reduced to mitigate issues with data trilinearity. In other words, if linear dependence between data dimensions or issues with retention time alignment or peak shape reproducibility exist in third-order data, the data can be unfolded or treated in a way such that it is reduced to second-order data. An advantage of MCR-ALS over PARAFAC is that it can be applied to chromatograms regardless of the retention time shifting in the ^{2}D time dimension, successfully

handling situations where the quantitative accuracy provided by PARAFAC would be problematic due to deviations from trilinearity [57]. Again, excessive deviations from trilinearity can result especially when long modulation periods, P_M, and high temperature programming rates are used, which can be mitigated by GC×GC system design [58]. However, the ability to apply MCR-ALS to problematic cases allows analysts to use a wide variety of instrumental conditions and forgo retention time shifting correction. Nonetheless, even with MCR-ALS, if there are multiple samples and retention time shifting is observed both within-run (2D retention time shifting between modulations) and between-run (1D retention time shifting between chromatograms), then the misalignment must be addressed to restore bilinearity before MCR-ALS is applied, particularly when univariate detection is implemented [52]. If multichannel detection is used, MCR-ALS can handle within and between run retention time shifting since the spectral mode provides a dimension with reproducible, linear responses that maintains data bilinearity [53,57].

The MCR-ALS model takes the form

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{1.5}$$

where **D** is the chromatographic two-way array, **C** is generally the chromatographic contribution for each component in the model, **S** is generally the spectral contribution for each component in the model, and **E** is a matrix of residuals not captured in the model. However, these general statements are not always the case due to the many different data structures that GC×GC data can take and whether univariate or multichannel detection is implemented. Individual two-way arrays of a singular GC×GC-FID chromatogram or the total ion current (TIC) of a GC×GC-TOFMS chromatogram are often extended into three-way arrays by analyzing multiple chromatograms (samples) at once. However, these three-way arrays with univariate detection must first be arranged back into a two-way array before MCR-ALS is applied, and this is often accomplished

through column-wise augmentation of the data for the different samples (can also be accomplished in a row-wise fashion, depending on the common information shared by the data sets). The augmentation can be performed on the 2D folded data or unfolded data. For three-way arrays with multichannel detection such as a single GC×GC-TOFMS run, the data must be unfolded to create a two-way array; usually this is accomplished by unfolding along the time dimension, or analyzing slices (modulations) of the ^{1}D separation. It is also possible to analyze four-way arrays, such as multiple GC×GC-TOFMS samples through superaugmentation of the data in which the time dimension is unfolded (or the slices are augmented) and the samples are concatenated column-wise. For three-way or four-way GC×GC-TOFMS data sets, *m/z* are commonly in the columns with time in the rows. MCR-ALS also allows for the augmented data (different samples or modulations) to have different numbers of rows (i.e., retention times). Depending on the data structure input into MCR-ALS, different information is contained in the resulting **C** and **S** matrices. MCR-ALS has been applied to generate calibration curves for essential oils [59], predict concentrations of biodiesel in blends of biodiesel and diesel [60], and resolve metabolites of interest from derivatization agents [8].

1.2.3 *Classification and Prediction*

Along with deconvolution (mathematical resolution), identification, and quantification, analysts commonly seek to classify samples, discover relevant chemical features that differentiate (or are common between) samples, and correlate chemical/physical properties to chemical composition of samples. Methods that are available for these goals can be referred to by several names, including pattern recognition methods, classification methods, cross-sample analysis methods, etc. Similar chemometric techniques implemented for relating measured sample properties (not restricted to chemical properties) and the chromatographic (i.e., chemical)

information are known as prediction methods. These pattern recognition and prediction methods can be characterized by a few commonalities: often the analyst is performing non-targeted analysis whereby the analytes of interest are unknown, the analysis is discovery-based where the goal is to reveal underlying relationships between sample composition and sample membership/classification/properties, and generally whole chromatograms are analyzed. A strategy involving feature selection can improve the success of downstream classification methods by removing irrelevant chromatographic information and reducing the size of the data files that need to be analyzed. Additionally, in many instances only select regions of the chromatogram are used instead of the entire chromatogram. Furthermore, the analytical work flow may involve generating and applying peak tables, either from commercial software packages or deconvolution methods. These methods can also be applied in either a supervised or unsupervised fashion, which means that sample class membership is either known or unknown *a priori*. A common unsupervised classification method is principal component analysis (PCA), while a common supervised classification method is partial least squares – discriminant analysis (PLS-DA). Partial least squares (PLS) regression analysis can also be used as a prediction method. Critical data preprocessing issues are commonly encountered and must be dealt with to optimize the performance of these chemometric methods: baseline correction, alignment (on various data density levels involving tiling, binning, pixel-level analysis), and normalization.

Principal Component Analysis (PCA)

PCA is used extensively in investigations of complex samples such as petroleum [61] and metabolomics [62], providing information on which variables, i.e., peaks, are responsible for sample class separation (loadings), which is manifested as the degree of sample clustering (scores).

PCA is generally applied as an unsupervised method, which means there is little *a priori* knowledge sample classifications prior to data analysis [63]. In short, rotation of the space axes is performed to encompass the greatest chemical variation in measurements (i.e., chemical information), with the new axes orthogonal to each other and called principal components (PCs). It is important to note that the term "component" does not refer to a specific chemical species (as has been the case in this chapter), but rather a fundamental part of the matrix **X**.

The data matrix, $\mathbf{X}(I,J)$, where I represents the number of data points in the time dimension and J refers to the number of *m/z* in the mass spectral dimension, is decomposed into two matrices $\mathbf{S}(I,F)$ and $\mathbf{L}^{\mathbf{T}}(F,J)$, which correspond to the scores and loadings, respectively, and depends on the number of PCs selected (F). Any remaining signal is captured in the residuals, **E**. Finally, the size of **E** will be a non-zero matrix of size I x J if the number of components used in the model, F, is less than the rank of the data matrix [64]. Briefly, rank can be defined as the number of linearly independent columns or rows (whichever is smaller) of a matrix, which is equivalent to the number of vector pairs used to completely decompose the data matrix. While many other matrix decompositions exist, PCA requires that the outer product of the columns of **S** and rows of **L** are mutually orthogonal and ordered by magnitude. That is, that the outer product of the first column of **S** ($\mathbf{s_1}$) and first row of **L** ($\mathbf{l_1}$) describes the largest variance in **X**. Therefore, it is important that the largest source of error is not due to chromatographic artifacts. Some pre-processing techniques to prevent, correct, or at least minimize artifacts will be discussed below.

One benefit of PCA is that is can be performed on the whole chromatogram across multiple samples. However, this requires either retention time alignment on both the ^{1}D and ^{2}D dimensions, or sufficient mitigation of retention time shift effects, often by data binning. In the case of sample analysis campaigns over the course of weeks or months, it may be necessary to align

chromatograms or apply PCA to a subset of potentially important regions in the chromatograms as identified by other classification techniques. This is particularly important if PCA is to be applied to pixel-level data (i.e., the raw data density has not been reduced). Various alignment algorithms are available, utilizing commercial or independent software [65–68].

Another useful option to deal with misalignment is through binning. Binning (or tiling) refers to the reduction of data density by summing the signal of a user-specified 2D region. If the proper 2D bin dimensions have been selected (i.e., one that encompasses the width of a peak but does not allow for multiple peaks to be summed together) misalignment can be corrected as the whole bin is considered one feature. For example, if the 2D peaks are ~ 3 modulations wide on ^{1}D, and ~ 10 data points wide on ^{2}D, a reasonable staring point to evaluate if binning improves the performance of PCA would be to bin the data into regions of these dimensions, which provides a 30-fold data reduction, some signal-to-noise (*S/N*) enhancement, and hopefully some mitigation of retention time shifting on both separation axes. Going to a larger bin size in some cases may be advantageous. While binning often corrects retention time shifting, artificial features of variation can sometimes be created if the bin splits critical analyte peaks for a given study; that is, if two or more bins describe the same analyte peak.

Other preprocessing techniques used to improve PCA models are baseline correction and normalization [69]. Removal of daily injection variation can be achieved through normalization. One approach is where the analyst multiplies each chromatogram by a factor that scales total signal to an average and reduces the impact of large variation. The consequence of not normalizing the data prior to analysis (or normalizing by the incorrect value) include artificially increasing or decreasing variation between sample classes or emphasizing analyte peaks with large signal that may not be the source of major variation (i.e., information).

Application of PCA to GC data requires a row vector for each sample. When PCA is applied to GC×GC data, the data matrix **X** is typically built by augmentation of several unfolded GC×GC chromatograms into a matrix. Extension of PCA to higher order data sets is possible through the use of multiway PCA (MPCA). After the successful correction of peak misalignment, PCA is applied to the unfolded data in the form of a two-way array, where the number of columns is equal to the number of variables (data points or bins) and the number of rows is equal to the number of samples. Ideally, scores from the same sample class should be in a tight cluster around a location in the scores plot. Two sample classes with significant differences would be located further apart in the scores plot than two sample classes with similar chemical features. This separation can be objectively determined by calculating the degree-of-class separation (DCS) metric. Briefly, the DCS is calculated by measuring Euclidean distances between the center of each sample class relative to the spatial scatter of individual samples in the scores plot [70].

Chemical information can often be obtained from the relative positions of clusters in the PCA model scores plot. As discussed above, chemically similar sample classes are located closer to each other than classes further apart. Specific chemical differences between sample classes correlate to one PC or another. The model loadings provide the locations of compounds in the chromatograms that are responsible for sample class separations observed in the scores plot. The analyst can then investigate these locations to identify and quantify the specific chemical compound(s) that are sample class distinguishing. PCA has been applied to biodiesel blends from various sources to determine characteristic peaks correlating to each vegetable oil source [60].

Partial Least Squares (PLS) Regression Analysis

PLS regression analysis is the most common chemometric method utilized for property prediction. The operation of PLS involves building a regression model that correlates independent variables (e.g. X-block; chromatographic data) to measured dependent variable(s) (e.g. Y-block; property data collected by other means). Generally speaking, PLS analysis aims to model the covariance between these two matrices by discovering the multidimensional direction in the X-block that explains the greatest variance in the Y-block. The X-block matrix, denoted **X**, has dimensions I x J, where I is the number of observations and J is the number of predictors; the Y-block matrix, denoted **Y**, has dimensions I x K, where K is the number of dependent variables (responses) for each of the I observations. Essentially, PLS seeks to predict **Y** from **X** using a set of orthogonal components LVs, similar to components in PCA, that maximizes the covariance between **X** and **Y**. In chromatographic applications, **X** contains I samples each with J chromatographic predictors (data points, peaks, etc. depending on the data structure input into PLS) while **Y** contains I samples with K responses (properties measured). In this context, oftentimes the number of predictors (chromatographic data points) is much greater than the number of observations (samples) and the predictors are collinear. PLS is well-suited to handle these situations, more-so than other regression models such as multiple linear regression (MLR) [71].

PLS analysis provides a linear predictive model of the sample properties based upon the chromatographic data. Although the model itself is linear, the data may be scaled to reflect other non-linear relationships that may exist (e.g. logarithmically, quadratically, etc.), although this is commonly not practiced. Typically, the chromatographic data is unfolded so that each sample is represented by a vector. For GC×GC data with univariate detection, this means simply unfolding

along the time dimension. However, for GC×GC data with multichannel detection, the data must be unfolded on both the time and mass spectral dimensions. There are a few approaches to unfolding this three-way data: all *m/z* values can be concatenated for each time data point ($t_1 m/z_1$, $t_1 m/z_2$, $t_1 m/z_3 \ldots t_1 m/z_n \ldots t_m m/z_n$), or the extracted ion chromatograms for all *m/z* can be concatenated ($t_1 m/z_1$, $t_2 m/z_1$, $t_3 m/z_1 \ldots t_m m/z_1 \ldots t_m m/z_n$), where *m* is the number of data points in the time dimension and *n* is the number of *m/z*. Another option involves summing along the mass spectral dimension to obtain the total ion current (TIC) chromatogram; in doing so, however, selective information contained in this dimension is initially lost in the PLS calibration step, but can be utilized in the interpretation of the LRVs. The application of PLS to unfolded three-way data is known as unfolded partial least squares (u-PLS), though u-PLS may be simply referred to as PLS in the literature. Extension of PLS to higher order data is possible using multiway partial least squares (NPLS) [72], which is essentially a three-way PLS algorithm. In PLS, the Y-block can contain either a single response variable (known as PLS1, where **Y** is a column vector) or multiple response variables (known as PLS2, where **Y** is a matrix and each response is contained in a different column) [72]. Since variation of chromatographic features from sample-to-sample is so important to building the PLS model, retention time shifting between samples can present a problem that can be solved by retention time alignment or binning of the data. As in the previously described chemometric methods, with binning approaches, the chromatographic signal within specified 2D bin dimensions is summed, reducing the size of the data concurrent with minimizing misalignment issues. Prior to PLS analysis, several more processing steps must be taken. Generally, the X-block and Y-block are mean-centered and auto-scaled, respectively. PLS is often performed using cross validation, which serves to evaluate the model. Several approaches exist for the cross validation step (e.g. leave one out, venetian blinds, etc.) that should be chosen on an

application-to-application basis. Leave one out cross validation (LOOCV) is generally applied when the number of samples is relatively small, while venetian blinds cross validation (VBCV) is applied when the number of samples is relatively large, and computation time becomes an issue if LOOCV were to be applied. During PLS modeling with cross validation, the analyst must choose the appropriate number of LVs to select the best model, whereby the analyst should aim to minimize the root mean square error of cross validation (RMSECV).

The main outcomes from PLS analysis are the regression plot showing the predicted property values (based on the chromatographic data) on the y-axis versus the measured property values on the x-axis and the linear regression vectors (LRVs) that contain information about what chromatographic features (i.e., time data points, *m/z*, analyte peaks) are positively or negatively correlated with the sample property being predicted. Chromatographic features that are positively correlated with the sample property will appear with positive values in the LRVs while features that are anti-correlated with the sample property will exhibit negative values. Features that are not correlated with the property will have values close to zero. Other results from PLS analysis, such as a plot of Q residuals vs. Hotelling's T^2 statistic, can inform the analyst about the samples and whether there are outliers or samples not being modeled well. With this information, the analyst can make decisions about whether to remove particular sample(s) from the analysis. For example, a sample that exhibits a high T^2 value is unlike all of the other samples, and accommodation of this sample could influence the model in an adverse way. It is possible that some samples that exhibit high Q residuals, meaning that they are not modelled well, do so because of the different sample throwing off the model. Removal of such samples prior to PLS analysis can also provide a better overall model.

PLS is often applied initially to a training or calibration set of samples that have known property values. Following the establishment of a satisfactory model, the property values of additional samples can be predicted from this model using the corresponding GC data of the new samples. This provides a fast and easy method of estimating a given property for a sample without having to directly measure the sample property. This is particularly useful for predicting properties that require time-consuming, expensive, or difficult analyses, such as may be the case in industrial applications. Additional value of PLS analysis comes from investigation of the LRVs and the greater understanding of how chemical composition of the samples is related to other sample properties. PLS analysis has been used in multiple GC×GC applications, particularly for fuel analysis and predicting composition [73,74], fouling [75], adulteration [76,77], boiling point [2], and other fuel properties [78,79].

Oftentimes, feature selection is performed prior to PLS analysis to reduce the number of predictors in **X** and improve ultimate prediction performance, especially when large data sets have been collected. The goal of feature selection strategies is to determine which variables hold the most importance with regard to predicting the property responses. This can be accomplished using feature selection techniques (e.g. Fisher ratio analysis) prior to PLS [75], omitting uninformative and/or low S/N *m/z* signals, removing chromatographic regions that do not exhibit peaks, or working with data summarized in peak tables acquired through deconvolution or means [80]. Interval multi-way partial least squares (iNPLS) has also been developed to build calibration models for target analytes, in which the 2D chromatogram is split into small sections and separate NPLS models are built for each section and then evaluated to select the best model [81].

1.2.4 *Feature Selection*

Feature selection is the name given to methods that are designed to reduce a dataset to its most important features. The benefits to these methods are reduction in data size (and analysis time) and removal of irrelevant data points. Feature selection is typically used prior to other data analysis methods, such as PCA or PLS, to improve the outcome of the model by reducing the chance of overfitting. In machine learning and data science, feature or variable selection is quite commonly implemented before modeling. Various approaches exist in these fields, such as filter methods and wrapper methods, which rank features according to a defined criterion and identify subsets based upon model performance, respectively [82–84]. Feature selection is often accomplished in a supervised fashion (class membership is known), such as in F-ratio analysis or ANOVA, more generally. In supervised feature selection, knowledge of the dataset is leveraged to isolate features that are related to a target variable (e.g., class membership). Unsupervised feature selection is based on inherent properties of the data (e.g., variance, mutual information) and can be used with unlabeled data.

Fisher Ratio (F-ratio) Analysis

F-ratio analysis is a popular supervised method (i.e., there is *a priori* knowledge about sample class membership) for non-targeted discovery of underlying differences in samples. This approach is particular useful if the analyst is interested in comparing two or more sample classes to ascertain the cause and effect impact of experimental design. The F-ratio approach is an analysis of variance (ANOVA) method that provides data reduction to elucidate sample class distinguishing features. The F-ratio quantitative metric is defined by

$$F - ratio = \frac{\sigma_{bc}^2}{\sum \sigma_{wc}^2} \tag{1.6}$$

where σ_{bc}^2 is the between-class variance, and $\sum\sigma_{wc}^2$ is the sum of the within-class variances. The output is the F-ratio value that scales in magnitude from zero to infinity as the variance between classes increases relative to the within-class variance. The F-ratio results can be summarized in what is referred to as a "hit-list" ordered from highest F-ratio value to lowest. The ordering of the hits by F-ratio enables the analyst to prioritize the next step of their work flow, which is often to start at the top of the hit list to identify and quantify the most class distinguishing analytes. It is important to note that since F-ratio calculation is based on variance, it prioritizes statistical significance over absolute signal (i.e., analyte concentration). If the F-ratio analysis and experimental conditions are not optimized, a high rate of false positives and false negatives may occur near the top of the hit list. False positives are the discovery of a feature that does not chemically distinguish sample classes [85]. On the other hand, false negatives occur when a chemically selective feature is not found by F-ratio analysis. F-ratio analysis can be applied to data in a multitude of ways, of which three will be discussed: F-ratios calculated from pixelated raw data, F-ratio calculation based on quantified analytes in peak tables, and F-ratios calculated from a "tile-based" method. Similar to PCA, with F-ratio analysis, data alignment is a critical aspect to gleaning the most important sample class characteristics without over-estimating others. If misalignment is significant, false positives or false negatives will be more frequent. Methods of mitigating this will be mentioned in the discussion of each F-ratio analysis approach.

Pixel-based F-ratio analysis applied to third-order chromatographic data (GC×GC-TOFMS) was introduced in 2006 by Pierce et al. [86]. Following basic preprocessing steps such as baseline correction and normalization, pixel-based F-ratio analysis compares every point in a GC×GC-TOFMS chromatogram to all of the other chromatograms. However, pixel-based F-ratio analysis faces a significant challenge if the data is misaligned, since a F-ratio value is calculated

at each data point. If retention time shifting between chromatograms occurs, F-ratio values artificially increase or decrease the significance (and F-ratio) of a given analyte.

One common way retention time misalignment has been addressed is through F-ratio calculations that rely upon using tabulated analyte peak areas, "peak tables", obtained from instrument software. Data processing is performed through commercial software that deconvolutes, identifies peaks, and provides a quantitative summary of each analyte in peak table format. The analyst can then correct for retention time shifting by aligning peaks according to their retention times and mass spectra and ultimately calculate F-ratio values [87]. Peak table based F-ratio analysis is performed on every analyte peak in the peak table from each chromatogram as produced by instrument data processing software. A critical distinction of utilizing peak tables output by instrument software is that preprocessing occurs prior to F-ratio analysis, whereas pixel-based F-ratio analysis (previously described) and tile-based F-ratio analysis (described next) both perform the F-ratio calculations on the data prior to analyte deconvolution, and so on. With peak table based F-ratio analysis, the preprocessing steps can include baseline correction, peak identification (and, if necessary, deconvolution) and integration.

Similar to peak table based F-ratio analysis, the utilization of a tiling scheme in F-ratio analysis allows for more retention time misalignment between samples concurrent with providing a high quality F-ratio analysis. However, in contrast to peak table approaches, tile-based F-ratio analysis is performed on raw data using a tiling scheme approach to mitigate the impact of retention time misalignment, effectively providing "smart binning." When F-ratio values are calculated using a tiling scheme approach, each tile is a sum of the GC×GC pixel-level data across a designated bin region, selected by the analyst to be approximately the 2D dimension of one GC×GC peak.

Hit list results can be simplified with the application of an F-ratio threshold. The basis of a threshold is so that the analyst can be more confident that each hit is meaningful to sample class separation for analyte hits at or above a suitable F-ratio value. The selection of an F-ratio threshold value has been a topic of discussion in the GC×GC community. Application of an F-ratio threshold allows an analyst to determine the point in a hit list where the false positives outweigh the true positive hits. Traditionally, this F-ratio value will be determined by manually inspecting the hit list and performing quantification (and statistical testing such as a *t*-test) of each hit until too many false positives are discovered [85]. This is not only a time-consuming method, but also introduces error due to subjectivity of the threshold selection. One way this has been addressed is through null distribution analysis [85]. Briefly, null distribution analysis can be automated and performed utilizing a pairwise rearrangement of samples within a class. Pairwise rearrangement works by switching two samples at a time from each class, while keeping balanced classes. Each of these null comparisons are then submitted for F-ratio analysis and combined into a histogram which can be used to objectively select an F-ratio threshold to achieve a desired confidence limit resulting in an acceptable false discovery rate. F-ratio analysis has been applied to clinical samples of bacterial infections in patients with cystic fibrosis [68] and to chemically fingerprint cocoa originating from different regions at various stages of processing [87] using a peak table-based approach, and in a tile-based approach to identify metabolites changing between fermenting and respiring yeast [88].

1.3 Overview of Chapters

1.3.1 *Chapter 2: Examination of the Two-Dimensional Mass Channel Cluster Plot Method for Gas Chromatography – Mass Spectrometry in the Context of the Statistical Model of Overlap*

Evaluation of a recently developed data reduction method for gas chromatography time-of-flight mass spectrometry (GC-TOFMS) is presented in the context of the statistical model of

overlap (SMO) using simulated chromatographic data. The two-dimensional mass cluster plot method (2D *m/z* cluster plot method) significantly improves separation visualization by measuring the retention time, t_R, and peak width-at-base, w_b, of each analyte peak on a per mass channel, *m/z*, basis and plotting w_b versus t_R as a single point for each peak. Additional selectivity is provided by the peak width dimension, allowing for the differentiation of "pure" or selective *m/z* and shared or overlapped *m/z*. Analyte clusters in the 2D mass cluster plot are defined based on clustering of individual points, representing the selective *m/z* for those analytes, and encompassed by a box of user-specified size. The method is applied to simulated chromatographic data with a random, independent distribution of analyte peaks and constant peak w_b. Two levels of chromatographic saturation factor, α, and two sets of analyte mass spectra with varying spectral similarity are studied to assess method performance. The percentage of analyte clusters found relative to the number of analytes simulated in the chromatogram increases as the box size (analogous to chromatographic resolution, R_s) is decreased, resulting in an R_s limit of 0.05 for the method. Additionally, the percentage of analyte clusters discovered also increases with lower α and greater dissimilarity between analyte mass spectra, demonstrating the immense benefit of improving the chromatographic separation and chemical selectivity in analyte discovery, identification, and quantification.

1.3.2 *Chapter 3: Unsupervised Feature Selection of Gas Chromatography with Mass Spectrometry by Variance Thresholding*

Feature selection is a commonly implemented step in a data analysis workflow, reducing the number of variables to only the most relevant features according to some criterion. Using only a relevant subset of the original variables can improve model performance and reduce the risk of overfitting. Feature selection can be accomplished in either a supervised (i.e., utilizing some *a*

priori knowledge of the data set such as class membership or an independently measured property) or unsupervised fashion. Supervised feature selection of gas chromatography coupled with mass spectrometry (GC-MS) data has been accomplished using Fisher ratio (F-ratio) analysis, which is an analysis of variance method. Herein, we demonstrate the application of a simple, unsupervised feature selection method known as variance thresholding to simulated and experimentally collected GC-MS datasets containing within-class variation approximating 30% relative standard deviation (*RSD*). A correlation coefficient of 0.71 relating the number of features discovered by F-ratio analysis and variance thresholding was determined for the 100 chromatographic simulations. Similarly, 27 out of 53 detected peaks comprising the metabolome of fermenting and respiring yeast were selected as features by both supervised and unsupervised methods, with a general positive correlation observed between the two quantities.

1.3.3 *Chapter 4: Predictive Modeling of Aerospace Fuel Properties Using Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry and Partial Least Squares Analysis*

Increasingly stringent requirements for aerospace propulsion system performance, reliability, and operability motivate quantitative connections between fuel composition, physical characteristics, and system performance. Chemically accurate assessment of aviation turbine fuels (Jet A, JP-8, etc.) and kerosene-based rocket propellants (RP-1 and RP-2) is requisite to mature these models. Comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC-TOFMS) is an excellent analytical tool for measuring detailed chemical information contained in complex fuels. Additionally, multivariate data analysis methods, referred to as chemometrics, are ideally suited to relate detailed chemical information contained within the GC×GC-TOFMS data to fuel properties and performance in a predictive manner. Herein, we apply

these techniques to a chemically diverse set of seventy-four distillate and multicomponent aerospace fuels, resulting in an improved understanding of the chemical compositional basis for physical and thermochemical behavior. Informed by GC×GC-TOFMS data, highly reliable partial least squares (PLS) models are developed and employed in the prediction of physical properties (measured separately using conventional test methods). Root mean square errors of cross validation (RMSECV) were relatively low: values of 0.0450 cSt, 41.3 Btu/lbm, 0.130 mass %, and 0.0064 g/mL were obtained for viscosity, heat of combustion, hydrogen content, and density, respectively. The corresponding normalized root mean square errors of cross validation (NRMSECV) were 6.01%, 10.3%, 8.71%, and 7.12%, respectively. Investigation of the linear regression vectors (LRV) provides valuable insight into the relationship between the chemical composition and physical properties, enabling in principle the model-informed selection of fuel chemical composition to achieve desired performance criteria.

1.3.4 *Chapter 5: Improvements to Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry Composition-Based Models for Kerosene-Based Fuel Thermal Integrity Using Supervised Feature Selection and Partial Least Squares Analysis*

Ensuring space launch system reliability, reusability, and operability places demands on rocket propulsion components and motivates quantitative connections between fuel composition, physical properties, and performance. In turn, the need for predictive models places greater emphasis on accurate fuel property measurements and detailed compositional information, especially for multicomponent kerosene-based fuels such as RP-1 and RP-2. To facilitate informed decisions regarding composition, specification, and fit-for-purpose behavior of complex fuels, we apply comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC-TOFMS) and multivariate data analysis techniques, referred to as chemometrics, in the

challenging context of fuel thermal stability in regenerative cooling systems. The recent availability of multivariate thermal integrity datasets (cooling channel pressure drop increase and carbonaceous deposit formation) acquired for an extensive set of rocket kerosene and advanced aerospace fuels offers an ideal opportunity to utilize powerful chemometric techniques to advance fundamental composition-performance relationships in the pursuit of predictive models. In this paper, we build upon previous progress by communicating results obtained using advancements in these approaches, namely: feature selection to isolate predominant chemical contributions to observed thermal stability performance metrics, thereby reducing computational time and improving model quality; the extension of partial least squares (PLS) model sets to mass-based carbonaceous deposit formation; and the discretization of test article deposition in an effort to differentiate the hydraulic effects of localized deposit from those of regionally summed deposit. 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 REFERENCES

[1] C. Vendeuvre, F. Bertoncini, L. Duval, J.-L. Duplan, D. Thiébaut, M.-C. Hennion, Comparison of conventional gas chromatography and comprehensive two-dimensional gas chromatography for the detailed analysis of petrochemical samples, J. Chromatogr. A. 1056 (2004) 155–162. https://doi.org/10.1016/j.chroma.2004.05.071.

[2] S.O. Fakayode, B.S. Mitchell, D.A. Pollard, Determination of boiling point of petrochemicals by gas chromatography–mass spectrometry and multivariate regression analysis of structural activity relationship, Talanta. 126 (2014) 151–156. https://doi.org/10.1016/j.talanta.2014.03.037.

[3] P.K. Kanaujia, Gas Chromatography | Petroleum and Petrochemical Applications, in: P. Worsfold, C. Poole, A. Townshend, M. Miró (Eds.), Encyclopedia of Analytical Science (Third Edition), Academic Press, Oxford, 2019: pp. 217–231. https://doi.org/10.1016/B978-0-12-409547-2.14104-6.

[4] 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.
[5] W. Fortunato de Carvalho Rocha, M.M. Schantz, D.A. Sheen, P.M. Chu, K.A. Lippa, Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data, Fuel. 197 (2017) 248–258. https://doi.org/10.1016/j.fuel.2017.02.025.
[6] E.A. Higgens Keppler, C.L. Jenkins, T.J. Davis, H.D. Bean, Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics, TrAC, Trends Anal. Chem. 109 (2018) 275–286. https://doi.org/10.1016/j.trac.2018.10.015.
[7] R. Fernández-Varela, G. Tomasi, J.H. Christensen, An untargeted gas chromatography mass spectrometry metabolomics platform for marine polychaetes, J. Chromatogr. A. 1384 (2015) 133–141. https://doi.org/10.1016/j.chroma.2015.01.025.
[8] 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.
[9] A. Garcia, C. Barbas, Gas chromatography-mass spectrometry (GC-MS)-based metabolomics, in: T.O. Metz (Ed.), Metabolic Profiling, Humana Press, 2011: pp. 191–204. https://doi.org/10.1007/978-1-61737-985-7_11.
[10] C. Cordero, H.-G. Schmarr, S.E. Reichenbach, C. Bicchi, Current Developments in Analyzing Food Volatiles by Multidimensional Gas Chromatographic Techniques, J. Agric. Food Chem. 66 (2018) 2226–2236. https://doi.org/10.1021/acs.jafc.6b04997.
[11] P.Q. Tranchida, G. Purcaro, M. Maimone, L. Mondello, Impact of comprehensive two-dimensional gas chromatography with mass spectrometry on food analysis, J. Sep. Science. 39 (2016) 149–161. https://doi.org/10.1002/jssc.201500379.
[12] 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.
[13] A.M. Muscalu, T. Górecki, Comprehensive two-dimensional gas chromatography in environmental analysis, TrAC, Trends Anal. Chem. 106 (2018) 225–245. https://doi.org/10.1016/j.trac.2018.07.001.
[14] R.B. Wilson, J.C. Hoggard, R.E. Synovec, High throughput analysis of atmospheric volatile organic compounds by thermal injection – isothermal gas chromatography – time-of-flight mass spectrometry, Talanta. 103 (2013) 95–102. https://doi.org/10.1016/j.talanta.2012.10.013.
[15] R.-Z. Zhou, J. Jiang, T. Mao, Y.-S. Zhao, Y. Lu, Multiresidue analysis of environmental pollutants in edible vegetable oils by gas chromatography–tandem mass spectrometry, Food Chem. 207 (2016) 43–50. https://doi.org/10.1016/j.foodchem.2016.03.071.
[16] S. Hashimoto, Y. Zushi, A. Fushimi, Y. Takazawa, K. Tanabe, Y. Shibata, Selective extraction of halogenated compounds from data measured by comprehensive multidimensional gas chromatography/high resolution time-of-flight mass spectrometry for non-target analysis of environmental and biological samples, J. Chromatogr. A. 1282 (2013) 183–189. https://doi.org/10.1016/j.chroma.2013.01.052.

[17] J. Aspromonte, K. Wolfs, E. Adams, Current application and potential use of GC × GC in the pharmaceutical and biomedical field, J. Pharm. Biomed. Anal. 176 (2019) 112817. https://doi.org/10.1016/j.jpba.2019.112817.
[18] G.C. Graffius, B.M. Jocher, D. Zewge, H.M. Halsey, G. Lee, F. Bernardoni, X. Bu, R. Hartman, E.L. Regalado, Generic gas chromatography-flame ionization detection method for quantitation of volatile amines in pharmaceutical drugs and synthetic intermediates, J. Chromatogr. A. 1518 (2017) 70–77. https://doi.org/10.1016/j.chroma.2017.08.048.
[19] 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.
[20] Y. Lu, P.B. Harrington, Forensic Application of Gas Chromatography–Differential Mobility Spectrometry with Two-Way Classification of Ignitable Liquids from Fire Debris, Anal. Chem. 79 (2007) 6752–6759. https://doi.org/10.1021/ac0707028.
[21] A.A.S. Sampat, M. Lopatka, G. Vivó-Truyols, P.J. Schoenmakers, A.C. van Asten, Towards chemical profiling of ignitable liquids with comprehensive two-dimensional gas chromatography: Exploring forensic application to neat white spirits, Forensic Sci. Int. 267 (2016) 183–195. https://doi.org/10.1016/j.forsciint.2016.08.006.
[22] 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.
[23] 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.
[24] R.B. Wilson, B.D. Fitz, B.C. Mannion, T. Lai, R.K. Olund, J.C. Hoggard, R.E. Synovec, High-speed cryo-focusing injection for gas chromatography: Reduction of injection band broadening with concentration enrichment, Talanta. 97 (2012) 9–15. https://doi.org/10.1016/j.talanta.2012.03.054.
[25] Á. Aragón, R.M. Toledano, S. Gea, J.M. Cortés, A.M. Vázquez, J. Villén, Large volume injection in gas chromatography using the through oven transfer adsorption desorption interface operating under vacuum, Talanta. 123 (2014) 39–44. https://doi.org/10.1016/j.talanta.2014.01.064.
[26] E. Hoh, K. Mastovska, Large volume injection techniques in capillary gas chromatography, J. Chromatogr. A. 1186 (2008) 2–15. https://doi.org/10.1016/j.chroma.2007.12.001.
[27] Z. Zajickova, I. Špánik, Applications of monolithic columns in gas chromatography and supercritical fluid chromatography, J. Sep. Science. 42 (2019) 999–1011. https://doi.org/10.1002/jssc.201801071.
[28] B. Gruber, F. David, P. Sandra, Capillary gas chromatography-mass spectrometry: Current trends and perspectives, TrAC, Trends Anal. Chem. 124 (2020) 115475. https://doi.org/10.1016/j.trac.2019.04.007.
[29] J. de Zeeuw, J. Luong, Developments in stationary phase technology for gas chromatography, TrAC, Trends Anal. Chem. 21 (2002) 594–607. https://doi.org/10.1016/S0165-9936(02)00809-9.

[30] C.F. Poole, N. Lenca, Gas chromatography on wall-coated open-tubular columns with ionic liquid stationary phases, J. Chromatogr. A. 1357 (2014) 87–109. https://doi.org/10.1016/j.chroma.2014.03.029.
[31] D. Gaddes, J. Westland, F.L. Dorman, S. Tadigadapa, Improved micromachined column design and fluidic interconnects for programmed high-temperature gas chromatography separations, J. Chromatogr. A. 1349 (2014) 96–104. https://doi.org/10.1016/j.chroma.2014.04.087.
[32] A. Kurganov, Monolithic column in gas chromatography, Anal. Chim. Acta. 775 (2013) 25–40. https://doi.org/10.1016/j.aca.2013.02.039.
[33] M.R. Jacobs, E.F. Hilder, R.A. Shellie, Applications of resistive heating in gas chromatography: A review, Anal. Chim. Acta. 803 (2013) 2–14. https://doi.org/10.1016/j.aca.2013.04.063.
[34] T.M. Gröger, U. Käfer, R. Zimmermann, Gas chromatography in combination with fast high-resolution time-of-flight mass spectrometry: Technical overview and perspectives for data visualization, TrAC, Trends Anal. Chem. 122 (2020) 115677. https://doi.org/10.1016/j.trac.2019.115677.
[35] M.M. van Deursen, J. Beens, H.-G. Janssen, P.A. Leclercq, C.A. Cramers, Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography, J. Chromatogr. A. 878 (2000) 205–213. https://doi.org/10.1016/S0021-9673(00)00300-9.
[36] M. Zoccali, P.Q. Tranchida, L. Mondello, Fast gas chromatography-mass spectrometry: A review of the last decade, TrAC, Trends Anal. Chem. 118 (2019) 444–452. https://doi.org/10.1016/j.trac.2019.06.006.
[37] 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.
[38] J.V. Seeley, S.K. Seeley, Multidimensional Gas Chromatography: Fundamental Advances and New Applications, Anal. Chem. 85 (2013) 557–578. https://doi.org/10.1021/ac303195u.
[39] 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.
[40] 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.
[41] A. de Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemom. 15 (2001) 749–771. https://doi.org/10.1002/cem.662.
[42] A. de Juan, R. Tauler, Factor analysis of hyphenated chromatographic data: Exploration, resolution and quantification of multicomponent systems, J. Chromatogr. A. 1158 (2007) 184–195. https://doi.org/10.1016/j.chroma.2007.05.045.
[43] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A. 1255 (2012) 3–11. https://doi.org/10.1016/j.chroma.2012.05.050.
[44] Z. Zeng, J. Li, H.M. Hugel, G. Xu, P.J. Marriott, Interpretation of comprehensive two-dimensional gas chromatography data using advanced chemometrics, TrAC, Trends Anal. Chem. 53 (2014) 150–166. https://doi.org/10.1016/j.trac.2013.08.009.

[45] B.A. Parsons, D.K. Pinkerton, R.E. Synovec, Implications of phase ratio for maximizing peak capacity in comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A. 1536 (2018) 16–26. https://doi.org/10.1016/j.chroma.2017.07.018.
[46] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, A Comprehensive Two-Dimensional Retention Time Alignment Algorithm To Enhance Chemometric Analysis of Comprehensive Two-Dimensional Separation Data, Anal. Chem. 77 (2005) 7735–7743. https://doi.org/10.1021/ac0511142.
[47] P.McA. Harvey, R.A. Shellie, Data Reduction in Comprehensive Two-Dimensional Gas Chromatography for Rapid and Repeatable Automated Data Analysis, Anal. Chem. 84 (2012) 6501–6507. https://doi.org/10.1021/ac300664h.
[48] J.A. Murray, Qualitative and quantitative approaches in comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1261 (2012) 58–68. https://doi.org/10.1016/j.chroma.2012.05.012.
[49] S.E. Reichenbach, X. Tian, C. Cordero, Q. Tao, Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography, J. Chromatogr. A. 1226 (2012) 140–148. https://doi.org/10.1016/j.chroma.2011.07.046.
[50] 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.
[51] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.
[52] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemom. Intell. Lab. Syst. 117 (2012) 80–91. https://doi.org/10.1016/j.chemolab.2012.02.003.
[53] H. Parastar, J.R. Radović, J.M. Bayona, R. Tauler, Solving chromatographic challenges in comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry using multivariate curve resolution–alternating least squares, Anal. Bioanal. Chem. 405 (2013) 6235–6249. https://doi.org/10.1007/s00216-013-7067-y.
[54] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numerische Mathematik. 14 (1970) 403–420.
[55] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432.
[56] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101–110. https://doi.org/10.1016/j.chemolab.2004.12.007.
[57] 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.
[58] 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.

[59] L.A.F. de Godoy, L.W. Hantao, M.P. Pedroso, R.J. Poppi, F. Augusto, Quantitative analysis of essential oils in perfume using multivariate curve resolution combined with comprehensive two-dimensional gas chromatography, Anal. Chim. Acta. 699 (2011) 120–125. https://doi.org/10.1016/j.aca.2011.05.003.
[60] N.G.S. Mogollon, F.A. de L. Ribeiro, M.M. Lopez, L.W. Hantao, R.J. Poppi, F. Augusto, Quantitative analysis of biodiesel in blends of biodiesel and conventional diesel by comprehensive two-dimensional gas chromatography and multivariate curve resolution, Anal. Chim. Acta. 796 (2013) 130–136. https://doi.org/10.1016/j.aca.2013.07.071.
[61] W. Zhang, S. Zhu, S. He, Y. Wang, Screening of oil sources by using comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry and multivariate statistical analysis, J. Chromatogr. A. 1380 (2015) 162–170. https://doi.org/10.1016/j.chroma.2014.12.068.
[62] H.D. Bean, C.A. Rees, J.E. Hill, Comparative analysis of the volatile metabolomes of Pseudomonas aeruginosa clinical isolates, J. Breath Res. 10 (2016) 047102. https://doi.org/10.1088/1752-7155/10/4/047102.
[63] J.E. Welke, V. Manfroi, M. Zanus, M. Lazzarotto, C. Alcaraz Zini, Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data, Food Chem. 141 (2013) 3897–3905. https://doi.org/10.1016/j.foodchem.2013.06.100.
[64] J.M. Amigo, M.J. Popielarz, R.M. Callejón, M.L. Morales, A.M. Troncoso, M.A. Petersen, T.B. Toldam-Andersen, Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis, J. Chromatogr. A. 1217 (2010) 4422–4429. https://doi.org/10.1016/j.chroma.2010.04.042.
[65] C. Couprie, L. Duval, M. Moreaud, S. Hénon, M. Tebib, V. Souchon, BARCHAN: Blob Alignment for Robust CHromatographic ANalysis, J. Chromatogr. A. 1484 (2017) 65–72. https://doi.org/10.1016/j.chroma.2017.01.003.
[66] T.-F. Tian, S.-Y. Wang, T.-C. Kuo, C.-E. Tan, G.-Y. Chen, C.-H. Kuo, C.-H.S. Chen, C.-C. Chan, O.A. Lin, Y.J. Tseng, Web Server for Peak Detection, Baseline Correction, and Alignment in Two-Dimensional Gas Chromatography Mass Spectrometry-Based Metabolomics Data, Anal. Chem. 88 (2016) 10395–10403. https://doi.org/10.1021/acs.analchem.6b00755.
[67] 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.
[68] 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.
[69] 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.
[70] 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.

[71] R.D. Tobias, An introduction to partial least squares regression, in: SUGI Proceedings, Orlando, FL, 1995: pp. 1250–1257.
[72] 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.
[73] L.A.F. de Godoy, E.C. Ferreira, M.P. Pedroso, C.H. de V. Fidélis, F. Augusto, R.J. Poppi, Quantification of Kerosene in Gasoline by Comprehensive Two-Dimensional Gas Chromatography and N-Way Multivariate Analysis, Anal. Lett. 41 (2008) 1603–1614. https://doi.org/10.1080/00032710802122222.
[74] 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.
[75] 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.
[76] H. Parastar, S. Mostafapour, G. Azimi, Quality assessment of gasoline using comprehensive two-dimensional gas chromatography combined with unfolded partial least squares: A reliable approach for the detection of gasoline adulteration, J. Sep. Science. 39 (2016) 367–374. https://doi.org/10.1002/jssc.201500720.
[77] M.P. Pedroso, L.A.F. de Godoy, E.C. Ferreira, R.J. Poppi, F. Augusto, Identification of gasoline adulteration using comprehensive two-dimensional gas chromatography combined to multivariate data processing, J. Chromatogr. A. 1201 (2008) 176–182. https://doi.org/10.1016/j.chroma.2008.05.092.
[78] 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.
[79] 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.
[80] C. Hurtado, H. Parastar, V. Matamoros, B. Piña, R. Tauler, J.M. Bayona, Linking the morphological and metabolomic response of Lactuca sativa L exposed to emerging contaminants using GC × GC-MS and chemometric tools, Sci. Rep. 7 (2017) 6546. https://doi.org/10.1038/s41598-017-06773-0.
[81] L.A.F. de Godoy, M.P. Pedroso, L.W. Hantao, R.J. Poppi, F. Augusto, Quantitative analysis by comprehensive two-dimensional gas chromatography using interval Multi-way Partial Least Squares calibration, Talanta. 83 (2011) 1302–1307. https://doi.org/10.1016/j.talanta.2010.08.015.
[82] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing. 300 (2018) 70–79. https://doi.org/10.1016/j.neucom.2017.11.077.
[83] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering. 40 (2014) 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

[84] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[85] 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.
[86] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Fisher Ratio Method Applied to Third-Order Separation Data To Identify Significant Chemical Components of Metabolite Extracts, Anal. Chem. 78 (2006) 5068–5075. https://doi.org/10.1021/ac0602625.
[87] 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.
[88] N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset, J. Chromatogr. A. 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.

Chapter 2. Examination of the Two-Dimensional Mass Channel Cluster Plot Method for Gas Chromatography – Mass Spectrometry in the Context of the Statistical Model of Overlap[2]

2.1 INTRODUCTION

Gas chromatography (GC) is a widely used analytical tool for a broad variety of complex samples. Detection with mass spectrometry (MS) provides additional chemical selectivity and a means to identify individual analyte species based upon unique molecular fragmentation patterns. Thus, GC-MS is a powerful platform that has gained routine implementation in the fields, including metabolomics, forensics, petrochemical, environmental, and food analysis [1–11]. Much effort has gone into improving separation efficiency and resolving power by gaining a deeper theoretical understanding, and then implementing the knowledge by improving instrumental design and methods [12–15]. Time-of-flight mass spectrometry (TOFMS) is commonly paired with GC (yielding GC-TOFMS) due to its sensitivity, extensive linear dynamic range, and high mass spectral scan rate [16].

Despite the recent advances in optimizing GC separations and detection, the sheer complexity of most samples makes the occurrence of co-eluting and/or overlapped sample components unavoidable. Component overlap leads to a loss of information due to problems with accurate identification and quantification of analytes of interest. Fortunately, the use of chemometric algorithms to obtain chemical information from complex data sets can greatly mitigate the burden of component overlap. Some chemometric methods can deconvolute

[2] This chapter has been reproduced from K.L. Berrier, B.C. Reaser, D.K. Pinkerton, and R.E. Synovec, Examination of the two-dimensional mass channel cluster plot method for gas chromatography – mass spectrometry in the context of the statistical model of overlap, *Journal of Chromatography A*, 1601 (2019), 319-326.

overlapped components, mathematically extracting the pure component peak profiles and mass spectra of overlapping analytes for confident analyte identification and quantification. The degree of overlap is represented by the chromatographic resolution, R_s, defined as the difference in the retention time between two analytes divided by their average peak width-at-base, w_b. For symmetric Gaussian peaks of equivalent size, at $R_s = 1.5$, two analytes are considered baseline resolved, whereas at $R_s \leq 0.5$, the two analytes would appear as one peak in the total ion current (TIC) chromatogram [17]. Chemometric methods such as multivariate curve resolution-alternating least squares (MCR-ALS) [18], generalized rank annihilation method (GRAM) [19,20], and parallel factor analysis (PARAFAC) [21] can deconvolute overlapped analytes at relatively low R_s, but start to fail at $R_s \leq 0.3$ [22].

A novel data analysis method known as the two-dimensional (2D) mass channel cluster plot method, or simply the mass cluster method (MCM), was developed and demonstrated to assist deconvolution of overlapped analytes in experimental GC-TOFMS chromatograms down to $R_s \sim 0.03$, surpassing the R_s limit of common deconvolution methods by a factor of ~10 [23]. The MCM has been utilized for data reduction, improved visualization, and assisted deconvolution of complex GC-TOFMS chromatograms [23,24], in which the MCM performed extremely well with real data. The MCM transforms raw GC-TOFMS data from one dimension (signal versus time vectors for each measured mass channel, *m/z*) into a 2D space that utilizes peak width as an added dimension of selectivity, i.e., a mass cluster plot. Broadly, each analyte peak per *m/z* is reduced into a single data point in the new 2D space with the mass cluster plot defined by retention time per *m/z* (x-axis) and accompanying width-at-base (y-axis). Visualization of the distribution of points in the mass cluster plot allows for the differentiation of pure (or selective) *m/z* and overlapped (or shared) *m/z*. Pure *m/z* for an analyte tend to cluster around specific retention time

and peak width coordinates, herein referred to as an "analyte cluster." Component overlap is readily visualized in the mass cluster plot, providing an estimate of the number of analyte components in a given overlap region. Therefore, the MCM can aid deconvolution of a given region of overlap by supplying selective information about the co-eluting analytes to use as inputs for various chemometric deconvolution and classification methods, going from unconstrained to MCM-assisted (constrained) models [2,23,24].

While the previously reported R_s limit for the MCM of $R_s \sim 0.03$ is impressive [23], it was principally a limited demonstration of the minimum R_s, and not a detailed, validated study of the capability of the MCM software in its current form. A more rigorous evaluation of the MCM is warranted to provide insight into determining a definitive minimum R_s, but also to put the MCM performance in the context of applying it to a wide range of chromatographic complexity. For this purpose, herein we study the MCM in the context of the statistical model of overlap (SMO) [25–27], referred to more broadly as a statistical overlap theory (SOT) [28]. By generating simulated GC-MS chromatograms based upon the principles of the SMO (randomly generated retention times of independent analyte components), using two sets of mass spectra (to challenge the MCM at two levels of sample complexity) and two levels of chromatographic saturation factor (functionally equivalent to two levels of separation efficiency, N), a relatively wide range of chromatographic sample complexity and spectral selectivity is explored, so that a minimum R_s for the MCM can be rigorously determined. To facilitate this study, the MCM is automated since 4000 SMO-simulated GC-MS chromatograms are analyzed, which is in contrast to previous studies that required substantially more user supervision [2,23,24].

2.2 THEORY

The SMO, pioneered by Davis and Giddings, is utilized herein [25]. Chromatograms are simulated by randomly and independently distributing components throughout the separation space. Resolution can be used as a quantitative metric to describe the ability of a method to successfully differentiate adjacent component peaks by some combination of chromatographic and mathematical separation,

$$R_s = \frac{x_0}{4\sigma} \tag{2.1}$$

where x_0 is equal to a minimum distinguishable distance of approach in the retention time dimension, Δt_R, which still allows two components to be identified as separate analyte peaks (analyte clusters in this study) at the corresponding R_s, and 4σ is equal to the average peak width-at-base, $w_{b,avg}$. Peak capacity is defined as

$$n_c = \frac{x_{sep}}{4\sigma R_s} = \frac{x_{sep}}{4\sigma}(R_s = 1) \tag{2.2}$$

where x_{sep} is a user defined separation distance along the retention time dimension, which could be either the entire chromatographic run time or some portion of it. The peak capacity, n_c, defines the number of peaks that can uniformly fit into a given separation window, and is often expressed at unit resolution peak spacing, $R_s = 1$. For the SMO, the saturation factor is defined as

$$\alpha = \frac{m}{n_c} \tag{2.3}$$

where m is the number of components occupying the user defined separation distance. The saturation factor is used to describe the degree of saturation of a given chromatogram or region of a chromatogram, and impacts the probability of successfully isolating a single component peak in the chromatogram at a specified x_0 (or allowed R_s). Davis and Giddings [25] define a probability that two consecutive local peak maxima with distance x fall within the distance x_0 as

$$P(x < x_0) = 1 - e^{-\alpha} \tag{2.4}$$

and the probability that the distance x between the two consecutive local peak maxima is equal to or greater than x_0 as

$$P(x \geq x_0) = e^{-\alpha} \tag{2.5}$$

which can be used to estimate the number of apparent peaks (singlets, doublets, etc.) expected under the given chromatographic conditions and some specified x_0 (or allowed R_s). The number of apparent peaks, p, and singlets, s, expected are given by the following,

$$p = me^{-\alpha} \tag{2.6}$$

and

$$s = me^{-2\alpha} \tag{2.7}$$

To examine the MCM in the context of the SMO, a user-defined cluster box size is applied to encompass the greatest number of selective *m/z*, deemed "pure" by exhibiting a width below a user-selected width threshold in the mass cluster plot, forming analyte clusters. The count of analyte clusters identified will increase as the cluster box size is decreased, and analogous to the SMO, as the number of components increases. The goal of this study is to determine the minimum cluster box size on the retention time axis that equates with x_0, and ultimately, R_s per Eq. 2.1. The impact of chromatographic complexity per the saturation factor α (related to separation efficiency, N) and analyte mass spectral selectivity are explored.

2.3 EXPERIMENTAL

2.3.1 *Chromatographic simulations*

All simulations and data manipulation/computations/analysis were performed in Matlab R2016a (The Mathworks, Inc., Natick, MA, U.S.A.) with the simulation parameters summarized in Table 2.1.

Table 2.1. Simulation and mass cluster method (MCM) parameters. Resolution values corresponding to the box sizes are equal to the box size divided by 100.

Parameter	**Conditions Studied**
Total separation time	20 s
Peak capacity, n_c	20
Number of components, m	10, 20
Saturation factor, α	0.5, 1
Peak width-at-base, w_b	1 s
Peak area	200,000 (TIC)
Signal-to-noise ratio, *S/N*	100 (TIC)
Data collection rate	100 Hz (spectra/s)
Analyte set	"Lower MV", "Higher MV"
Signal threshold (MCM)	20 (raw *S/N* threshold of ~10)
Width threshold (MCM)	1.1 s
Box sizes, R_s (MCM)	3, 5, 7, 9, 15, 21, 31, 41, 51, 61, 71, 81, 91, 101

Twenty second long chromatograms were simulated with either 10 components (α = 0.5) or 20 components (α = 1) randomly and independently distributed throughout the separation space. Each component was simulated as a Gaussian peak with a constant peak area and width-at-base (w_b, $\pm 2\sigma$) of 1 s at a mass spectral scan rate of 100 Hz (1 data point = 10 ms). Once the peaks were modeled, a randomly selected analyte mass spectrum was multiplied element-wise (i.e., outer product) across each peak to form a series of Gaussian peaks representative of the mass channels, *m/z*, having signal for that analyte. Analyte selection was performed in a way such that no chromatogram would have the same analyte simulated more than once, and each chromatogram had an independent, random selection from the analyte set applied. Analyte mass spectra at unit mass resolution were obtained from the NIST MS Search 2.0 database (NJ, U.S.A.). Table A.1 in

Appendix A lists the 200 chemical species and delineates the two analyte sets used in this study: Lower Match Value and Higher Match Value (hereafter, "Lower MV" and "Higher MV," respectively). For these analyte sets, the MV was determined between each analyte and every other analyte in the set in order to assess spectral similarity. Based on the equation presented by Stein [29], the MV is defined as the normalized dot product of two analyte mass spectra (generally a user-obtained spectrum and library spectrum) weighted by the *m/z* intensities. The Lower MV set is analogous to having a sample in which the analytes have very dissimilar mass spectra as shown in Figure 2.1A, while the Higher MV set is what one would expect for a more realistic sample, having a wide range of mass spectral similarity as shown in Figure 2.1B.

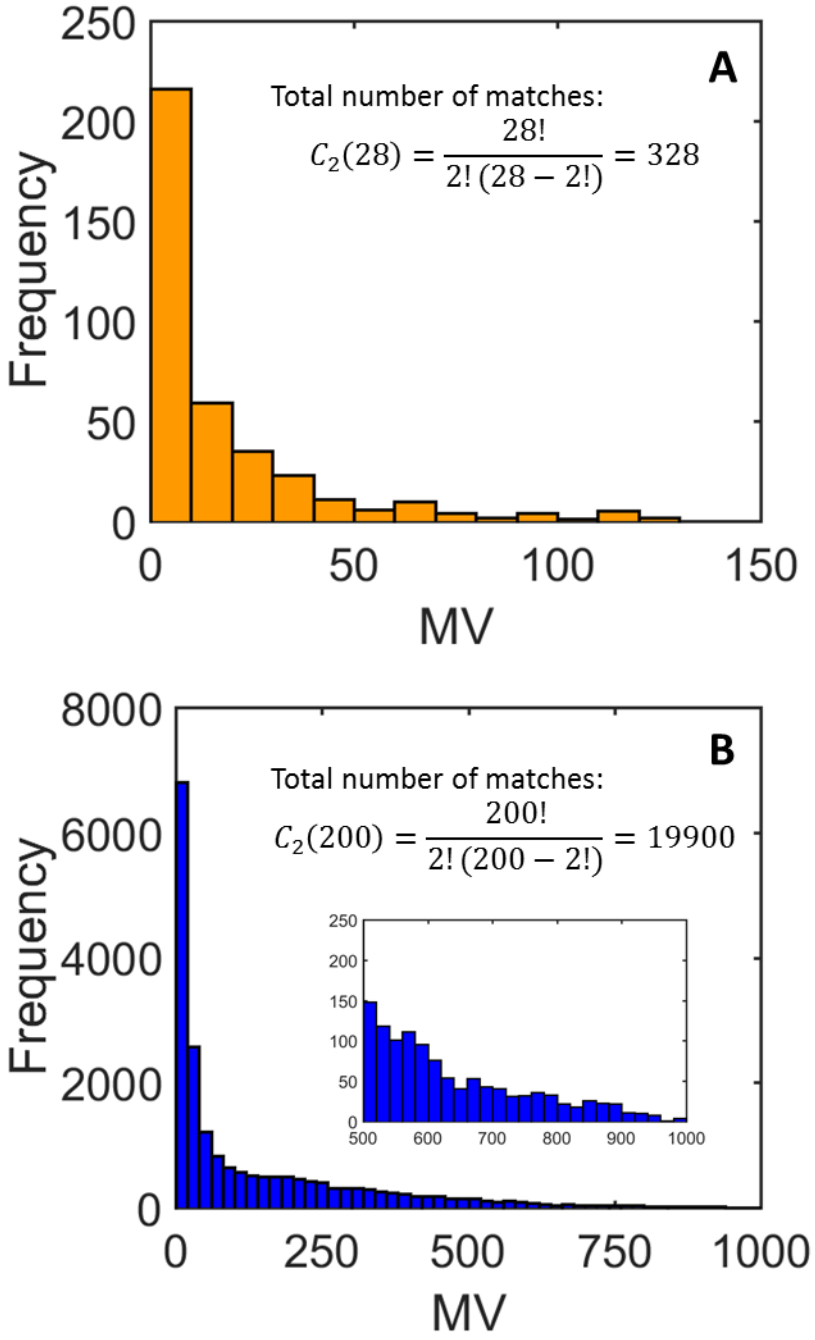


Figure 2.1. Histogram of match values (absolute frequency) for all analytes in the (A) Lower MV analyte set and (B) Higher MV analyte set. A zoom-in of the region of the histogram between MV of 500-1000 is provided inset. The total number of matches calculated was the number of possible combinations that can be obtained by choosing a sample of two elements from the set of 28 and 200 analytes for the Lower MV and Higher MV, respectively.

For the generation of the simulated GC-MS chromatograms, the mass spectrum of each analyte was normalized in a way such that the sum of the intensities of all the *m/z* would be equal to 1000 for all of the analytes to allow for a constant signal-to-noise ratio, *S/N*, of 100 in the TIC. However, the *S/N* per individual *m/z* for a particular analyte had a range dependent on the individual *m/z* intensities. Random Gaussian-distributed noise was generated independently for each *m/z* with a standard deviation that would provide an *S/N* of 100 in the TIC. This was accomplished using the formula

$$\sigma_{\mathrm{N}} = \frac{\left(\left(\frac{A}{\sigma\times\sqrt{2\pi}}\right)\times 1000\right)}{3\times S/N\times\sqrt{n}} \tag{2.8}$$

where A is equal to the area of a simulated Gaussian peak before a mass spectrum is multiplied element-wise across (constant at 200), σ is equal to the standard deviation of the Gaussian peak (equivalent to one-fourth of the peak w_{b}), the numerator is equal to the peak height in the TIC (constant at ~3200 for all simulated peaks before preprocessing and additive signal due to overlap), n is equal to the number of *m/z* (in this study, n = 360), and *S/N* is associated with the TIC. One thousand random chromatograms of each α (0.5 and 1) were simulated using both analyte sets with matched retention times to yield 4000 unique chromatograms total.

2.3.2 *Mass cluster method*

The MCM, consisting of a data reduction step and cluster location step, was implemented in an automated, unsupervised fashion utilizing the in-house algorithm (Figure A.1 in Appendix A) [23]. Briefly, the raw GC-TOFMS chromatograms are smoothed on a per *m/z* basis using a seven point moving boxcar to reduce variation in the apexes of each peak from the simulated noise without significantly affecting the peak width or height (< 5% change). Next, for each *m/z* peak above a user selected signal threshold, the retention time and peak width-at-base are measured,

binned according to the mass spectral scan rate, and plotted as a single point as measured w_b versus t_R to form a 2D mass cluster plot, as previously reported [23]. Each of these points in the mass cluster plot is referred to as a "cluster point", which are color coded according to the number of *m/z* present at a given location. Application of a user-selected width threshold trims the mass cluster plot to only include cluster points below a threshold of 1.1 s, selected to limit the number of false positives (FP) and concurrently maximize the number of true positives (TP) (Supporting Information, 2.5.1). We elected to measure the w_b of the peaks at 13% of the peak maxima, causing a 1% increase in the measured peak widths compared to the 4σ definition (i.e., 4.04σ). Proximity of cluster points in the trimmed mass cluster plot is then assessed by a density based spatial clustering algorithm to determine the location of analyte clusters.

A minimum of three *m/z* all falling within two data points of each other was required to be deemed an analyte cluster, whether or not it ultimately comprised a true positive (TP) or false positive (FP). Other definitions for the existence of a cluster are possible [23], and it should be noted that the analyst's confidence in the purity of a cluster should improve with greater numbers of *m/z* within the cluster box (i.e., more *m/z* below the width threshold for a given analyte). For each identified analyte cluster, a cluster box with a user-specified size is applied to encompass the greatest number of *m/z*. Initially centered on the location with the greatest number of *m/z*, the box is moved around this point and centered on the location that encompasses the most surrounding *m/z*. Location of cluster boxes was initially performed using a box size of 3 × 3 (30 ms × 30 ms at a 100 Hz scan rate), previously deemed to encompass ~95% of the selective *m/z* in both retention time and peak width dimensions for a pure, resolved analyte peak [23]. The box size was then systematically increased to a maximum of 101 × 101 (~ 1 s × 1 s) to ascertain how the density based spatial clustering algorithm groups neighboring analyte clusters falling within a distance less

than the box size for each box size tested. Thus, cluster box centers with a Euclidean distance less than the box size would be grouped as one cluster. Box size was kept square to facilitate this step. The box size dimension in the retention time dimension is analogous to some minimum distance of approach xo in the retention time dimension required to separate two components. Since the peak width-at-base (4σ) is 100 data points (1 s wide at 100 Hz) and R_s is equal to the minimum distance x_0 required to separate two components divided by 4σ (per Eq. 2.1), then a box size of 3 × 3 (corresponding to x_0 of 3 data points) is equivalent to an R_s of 0.03. The number of independent clusters at each box size for each chromatogram was compiled and averaged for each saturation factor/analyte set. A signal threshold of 20 (*S/N* threshold of ~10 prior to smoothing) was deemed appropriate, and thus was applied to the smoothed data to exclude the low intensity *m/z* from the algorithm. In general, the signal threshold should be selected with consideration of the analytical needs, e.g., the limit of detection (LOD) and the concentration range of analytes in the sample.

2.4 Results and Discussion

2.4.1 *Application of the MCM to the Lower MV analyte set*

A representative chromatogram was selected to demonstrate the workflow using the MCM as presented in Figure 2.2A. The total ion current (TIC) chromatogram contains 20 randomly selected components ($\alpha = 1$). The Lower MV set was selected to provide a simple data set that, along with constant simulated peak widths and areas, was amenable to systematic method evaluation, optimization, and validation. Selection of these simulation parameters allowed for interrogation of the MCM to determine the complicated mechanisms and variables that give rise to various outcomes. The same chromatogram is shown as the superimposed ion chromatogram with all *m/z* overlaid in Figure 2.2B, representing the data to which the MCM is applied. One would suspect that there are shared *m/z* present, based on the severity of overlap in a few particular

regions of the chromatogram (retention time regions of 8-12 s and 18-20 s) and the presence of *m/z* traces that appear to have elevated widths relative to other *m/z*.

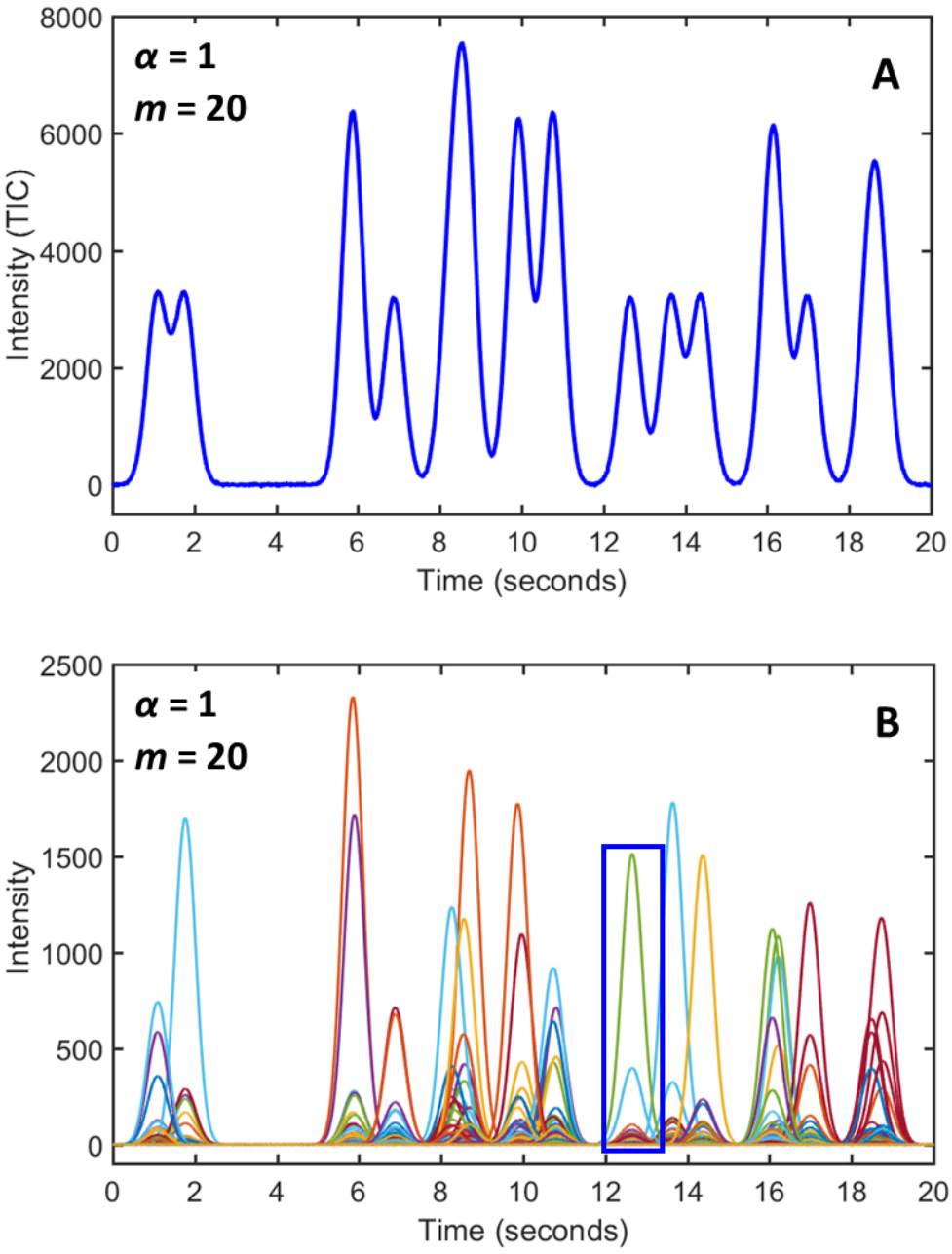


Figure 2.2. A representative simulated chromatogram in this study, as it is processed using the mass cluster method (MCM). (A) The total ion current (TIC) chromatogram, with 20 components ($\alpha = 1$). (B) The chromatogram with the signal traces for all *m/z* using the Lower MV analyte set provided in Table A.1 and Figure 2.1(A). The peak selected in the blue box is at $R_s \sim 1$ with respect to the subsequent peak.

The full mass cluster plot following the data reduction step is provided in Figure 2.3A, prior to applying the width threshold. Now it is apparent that there are shared *m/z* present as indicated by the cluster points at the elevated widths, with the majority of these cluster points located in the region of 8-12 s where a significant number of components appear in the TIC and superimposed ion chromatograms presented in Figure 2.2A and B. Since all analytes were simulated with the same peak width and area, component overlap is readily visualized in the TIC, appearing as peaks with increased peak width and signal intensity. A user-selected width threshold of 1.1 s (purple line) is applied to determine the analyte clusters present as illustrated in Figure 2.3B, using the 3 × 3 cluster box size (equivalent to an R_s of 0.03). Recall that all *m/z* above the width threshold are excluded in the cluster box location step by the density based spatial clustering algorithm, but are shown to put the analyte clusters in context in Figure 2.3B.

The distribution of the cluster points in a mass cluster plot reveals information about the R_s between analytes in addition to the degree of mass spectral similarity. An example of a well resolved analyte cluster is shown at a retention time of 12.66 s in Figure 2.3A and B. However, the presence of many *m/z* below the width threshold may also be due to the overlap of two or more analytes at a very low resolution (R_s ~ 0-0.1). An example of this is located at ~6 s, in which two analytes co-elute with R_s = 0.04. A "horseshoe" shaped distribution pattern of *m/z* is indicative of overlap and shared *m/z*, specifically at a relatively low resolution (R_s ~ 0.1-0.6); for example, the region between 8 and 9 s in Figure 2.3A demonstrates this pattern, where the R_s between the first two analytes is ~0.3 and the R_s between the second and third analyte is ~0.1. A third pattern occurs when analytes with shared *m/z* are overlapped at higher resolution (R_s ~ 0.6-1.5). In these cases, a vertical column of *m/z* is observed instead of a horseshoe pattern because the analytes are resolved enough not to give rise to a distribution of *m/z* in the retention time dimension; alternatively, some

m/z peaks are not baseline resolved, resulting in elevated measured widths. Examples of this are visible at ~ 1 s, 7 s, and 17 s in Figure 2.3A), though the pattern is much more obvious when the Higher MV set is applied to this same chromatogram.

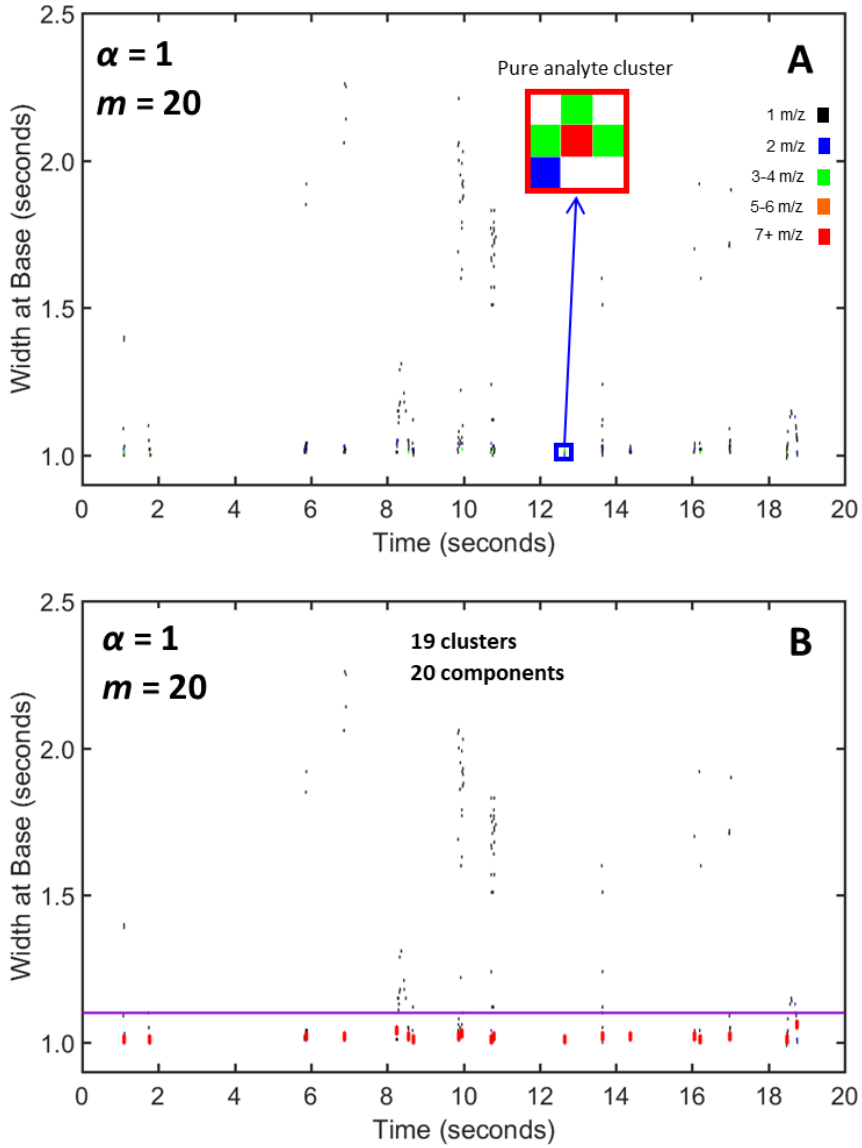


Figure 2.3. Application of MCM to the representative chromatogram with the Lower MV analyte set. (A) The cluster plot of the chromatogram following the data reduction step of the MCM. Each point represents at least one *m/z*, with locations having more than one *m/z* color-coded according to the scale. The cluster corresponding to the selected peak in Figure 2.2(B) is shown to illustrate a pure analyte cluster. (B) Cluster boxes determined from the cluster location step are represented by the red rectangles. The width threshold is represented by the purple line. All *m/z* with widths greater than or equal to this threshold were excluded from the cluster location step. These steps of the MCM are summarized in Figure A.1.

For this representative chromatogram, 19 analyte clusters of the 20 components were found (or 95%) for the 3 × 3 cluster box size (R_s = 0.03), due to the single co-elution (R_s = 0.04) at ~ 6 s. A discussion of *m/z* distribution patterns and the variables that impact them (R_s and mass spectral similarity) is included (Supporting Information, 2.5.2, Table 2.S1, Table 2.S2, Figure 2.S1, and Figure 2.S2).

A summary of the results for all 1000 chromatograms simulated under the described conditions (Figure 2.2, α = 1, Lower MV set, Table A.1) is provided in Figure 2.4. The average percentage of analyte clusters found by the MCM of the expected number of components is shown as a function of increasing R_s, which is related to the box size. The percentage of predicted apparent peaks (Eq. 2.6) and singlets (Eq. 2.7), relative to the expected number of components at the allowed R_s corresponding to each of the box sizes, was calculated from the formulas for the expected number of peaks and singlets according to the SMO [25]. It is noteworthy that the MCM results for the Lower MV set at α = 1 follow the curve for the predicted percentage of apparent peaks (Eq. 2.6), with some deviation as the R_s approaches zero. As the cluster box dimension is reduced, i.e., as R_s approaches zero, the percentage of theoretical apparent peaks approaches 100%, with the MCM experiencing a leveling off at a cluster box size equivalent to an R_s ~ 0.03-0.05 due to an inability to distinguish between neighboring analyte clusters from nearly complete or complete overlap. One can conclude that the R_s limit for the MCM is ~0.05 for the Lower MV set. Furthermore, although the use of the MCM to improve deconvolution of overlapped regions has been demonstrated previously [2,23,24], it has not yet been demonstrated quantitatively. A quantitative demonstration of the remarkable benefit of using information from the MCM as a constraint in the deconvolution of regions of overlap by a constrained version of MCR-ALS (i.e.,

"MCM assisted MCR-ALS") relative to implementing unconstrained MCR-ALS is provided (Supporting Information, 2.5.3, Table 2.S3, Figure 2.S3, Figure 2.S4, and Figure 2.S5).

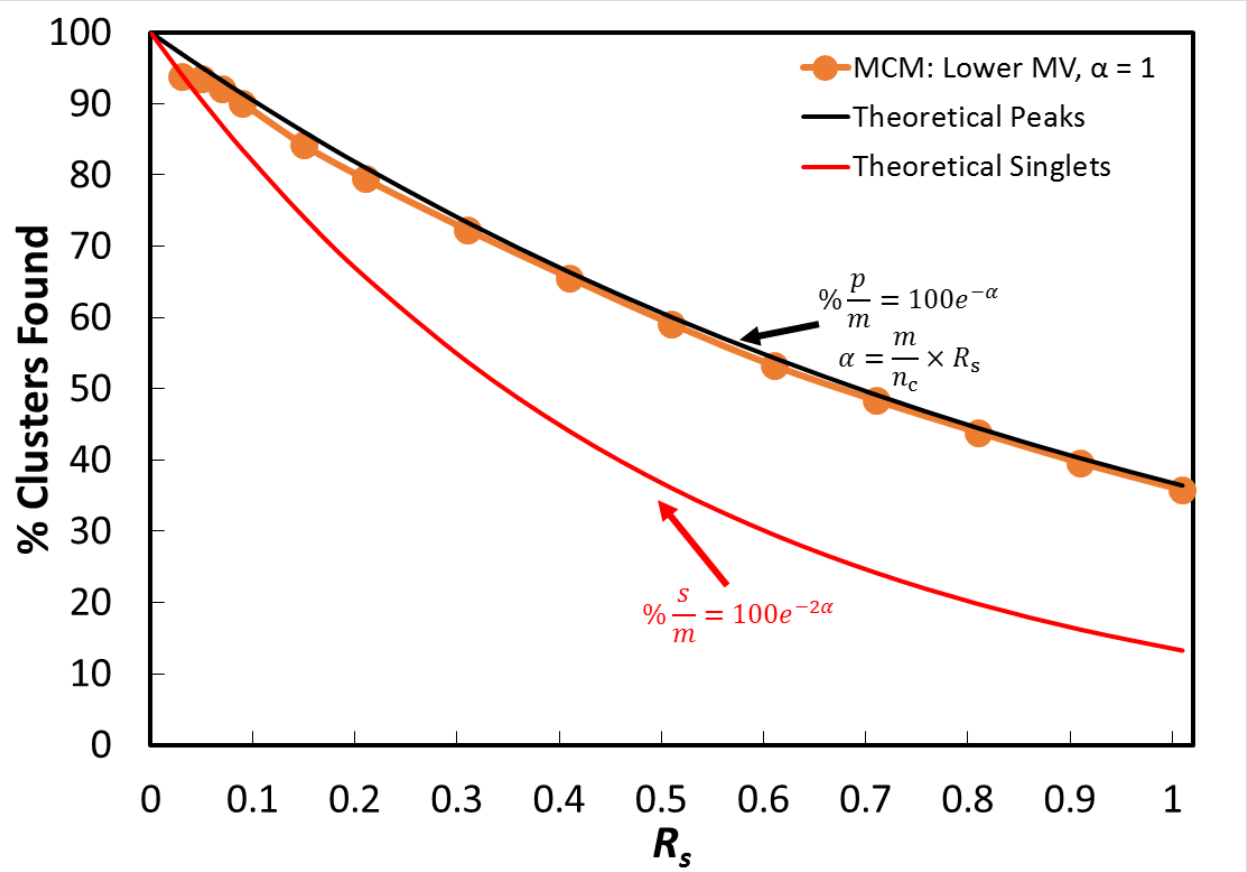


Figure 2.4. Validating the mass cluster method in the context of the statistical model of overlap (SMO). Percentage of clusters found (with respect to the number of components in each chromatogram) for the Lower MV analyte set and $\alpha = 1$ as a function of the resolution, R_s (related to cluster box size). Results shown are the average of 1000 chromatograms at each box size. The number of clusters found for each chromatogram consists of the number of true positives (TP) and the number of false positives (FP) discovered in each cluster plot. Based on the SMO, the theoretical percentage of apparent peaks (Eq. 2.6) and singlets (Eq. 2.7) expected to appear in a chromatogram with an $\alpha = 1$ are shown as a function of the allowed R_s.

2.4.2 *Application of the MCM to the Higher MV analyte set*

Following application of the MCM to the Lower MV set, the sample complexity of the chromatograms was increased by applying the Higher MV set (Table A.1) to the same chromatograms. In other words, the chromatograms were identical to the previous set of Lower

MV chromatograms with the exception of the analytes populating the chromatograms and the noise profiles simulated for each *m/z*. The MCM was then applied to illustrate the effect of mass spectral similarity of the analytes on the performance of the method. Hence, the same representative chromatogram as in Figure 2.2A was used for illustration purposes (same retention times, peak areas, and peak widths) for the Higher MV set. Analogous to previously shown in Figure 2.2B, the representative chromatogram with a random selection of analytes from the Higher MV set is provided in Figure 2.5.

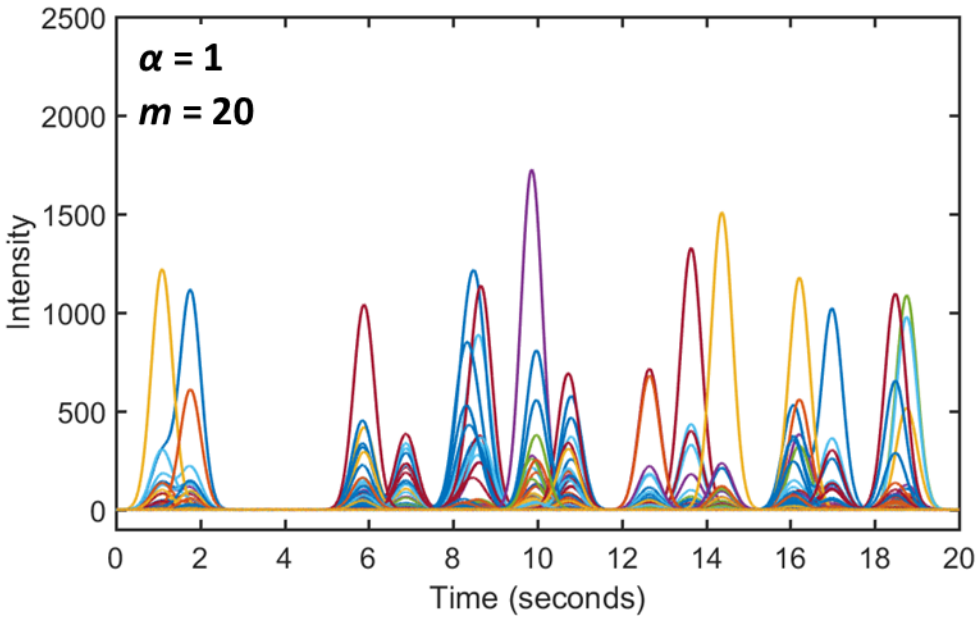


Figure 2.5. The representative chromatogram with the Higher MV analyte set, Table A.1 and Figure 2.1B, shown with the signal traces for all *m/z*.

Compared to the representative chromatogram in Figure 2.2B, we predict that more shared *m/z* are present. Indeed some shared *m/z* are visible, particularly the *m/z* traces with shoulders or humps between 0-2 s. The full mass cluster plot in Figure 2.6A contains an increased number of *m/z* located at elevated widths relative to when the Lower MV set is used in Figure 2.3A. This is consistent with the increased probability to have neighboring analytes with similar mass spectra, and therefore, greater numbers of shared *m/z*. Application of the width threshold in Figure 2.6B

located 16 analyte clusters out of the expected 20 components (corresponding to 80% found) as opposed to the 19 analyte clusters found (95%) in Figure 2.3B. This is due to overlap of analytes with similar mass spectra, causing many of the *m/z* to appear at elevated widths, which are removed by the width threshold. If there are not enough *m/z* below the width threshold to define an analyte cluster, then the analytes may not be found. An example of this is at ~17 s in Figure 2.6B, where only one selective *m/z* below the width threshold for that particular analyte is observed. Application of a higher width threshold would allow this analyte cluster to be found, but at the expense of also finding FP clusters.

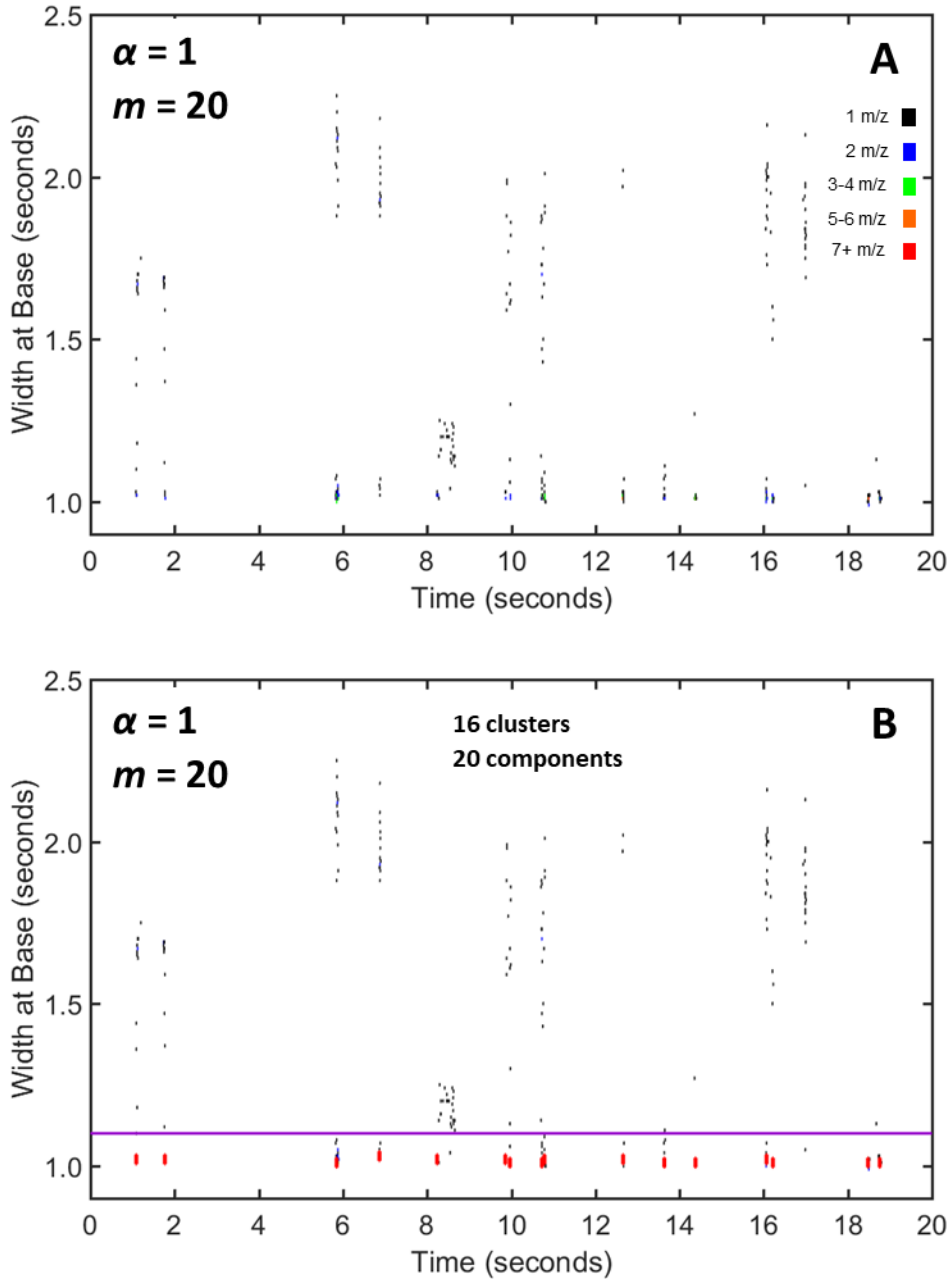


Figure 2.6. Application of the MCM to the representative chromatogram with the Higher MV analyte set. (A) The cluster plot of the chromatogram following the data reduction step of the MCM, as in Figure 2.3A. (B) Cluster box locations, as in Figure 2.3B, but now showing more *m/z* at elevated widths due to more mass spectral similarity across analytes.

2.4.3 *MCM performance in the context of sample and separation complexity*

A summary of the effect of sample complexity due to mass spectral similarity on the performance of the MCM at $\alpha = 1$ is provided in Figure 2.7, where the results for the Lower MV versus Higher MV sets are compared. Again, the average percentage of analyte clusters found for each analyte set is plotted as a function of R_s. A significant improvement in method performance using an optimum box size ($R_s \sim 0.03$-0.05) can be seen when the sample complexity is decreased by increasing chemical selectivity via decreasing the mass spectral similarity of the analytes in a given sample. The average percentage of analyte clusters found relative to the expected number of components increases from 80% to 94% when the analyte set applied to the chromatograms was changed from the Higher MV analyte set to the Lower MV set. Increased chemical selectivity reduces the R_s at which two analyte peaks can be distinguished [30], causing more analytes to be successfully found when the Lower MV set was applied relative to the Higher MV analyte set, even though the simulated R_s was the same for both sets. One could envision using high-resolution mass spectrometry (HRMS) to achieve significantly improved chemical selectivity and analyte detection compared to unit resolution mass spectrometry, somewhat analogous to the improvement observed when moving from the Higher MV to Lower MV set.

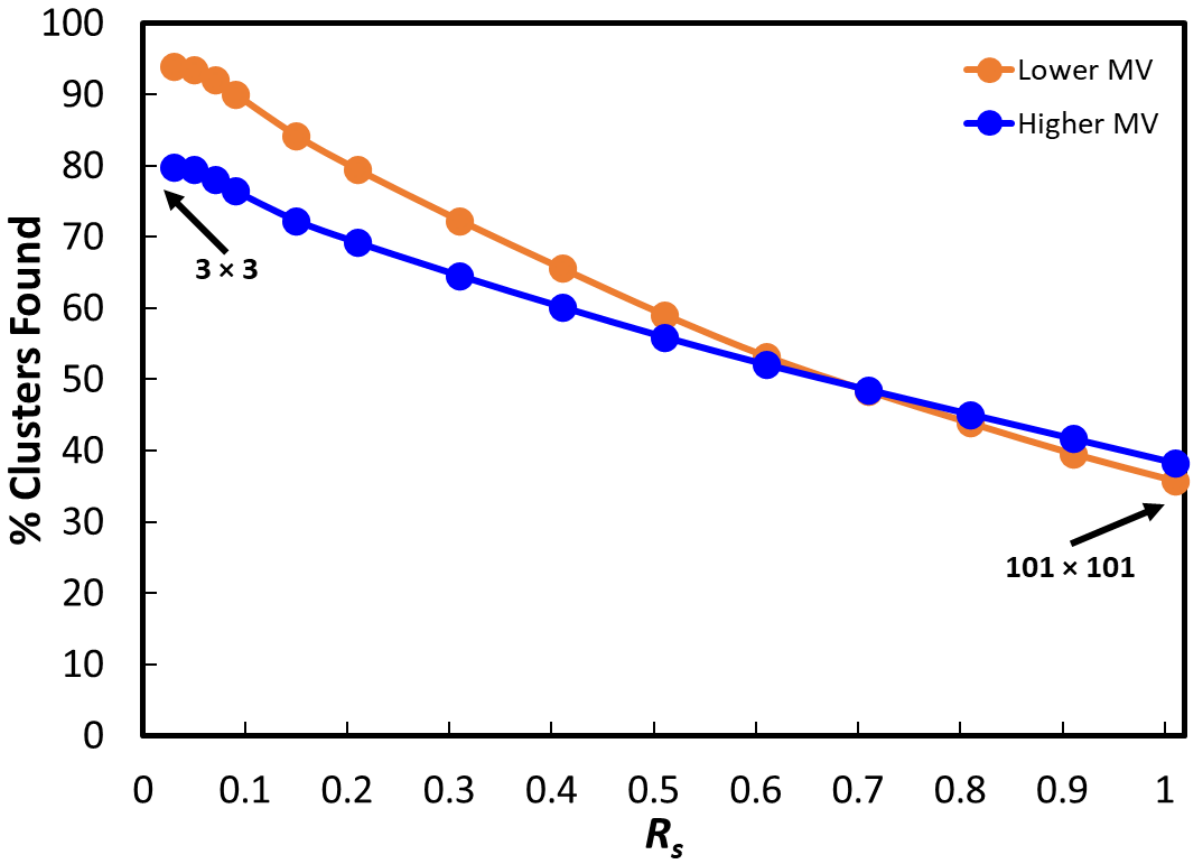


Figure 2.7. Investigating the effect of sample and separation complexity on the performance of the MCM. Percentage of clusters found (with respect to the number of components in each chromatogram) for an α = 1 as a function of R_s, emphasizing the Lower MV and Higher MV analyte set differences for clarity. Results are shown as the averages of 1000 chromatograms at each box size.

The effect of separation complexity (i.e., saturation factor, α) on the performance of the MCM is demonstrated in Figure 2.8. For brevity, representative chromatograms and cluster plots for α = 0.5 are not shown or discussed herein. The curves representing both analyte sets at α = 0.5 have been added to Figure 2.8 to illustrate the importance of improving the chromatography (e.g. through increasing efficiency, N) in achieving desirable results in terms of peak finding, identification, and quantification of analytes in a given sample. Improving the chromatography through increasing N can effectively lower the saturation of a separation and decrease the probability or severity of component overlap. Similar to Figure 2.4, the MCM results for the Lower

MV set at $\alpha = 0.5$ follow Eq. 2.6, not shown for brevity and clarity in Figure 2.8. We hypothesize that results obtained from the application of the MCM to the Lower MV set using a range of α would follow Eq. 2.6 due to the chemical selectivity provided by the orthogonality of these analyte mass spectra, as seen in Figure 2.1A. Without the effect of high mass spectral similarity, the chromatographic simulations in this study follow the conditions that govern Eq. 2.6 (i.e., random, independent component distribution). It is also notable that as R_s increases, the curves for $\alpha = 0.5$ converge at $R_s \sim 0.8$ and the curves for $\alpha = 1$ cross at $R_s \sim 0.7$. We would expect this trend of convergence for each saturation factor to continue with increasing R_s until some R_s value is reached where only one cluster is found (i.e., all analytes are encompassed in one cluster box), past which the percentage of clusters found would level off at 10% and 5% for $\alpha = 0.5$ and $\alpha = 1$, respectively. This is due to the method performance becoming less dependent on mass spectral similarity and more representative of the simulated R_s of the randomly distributed components in the chromatograms.

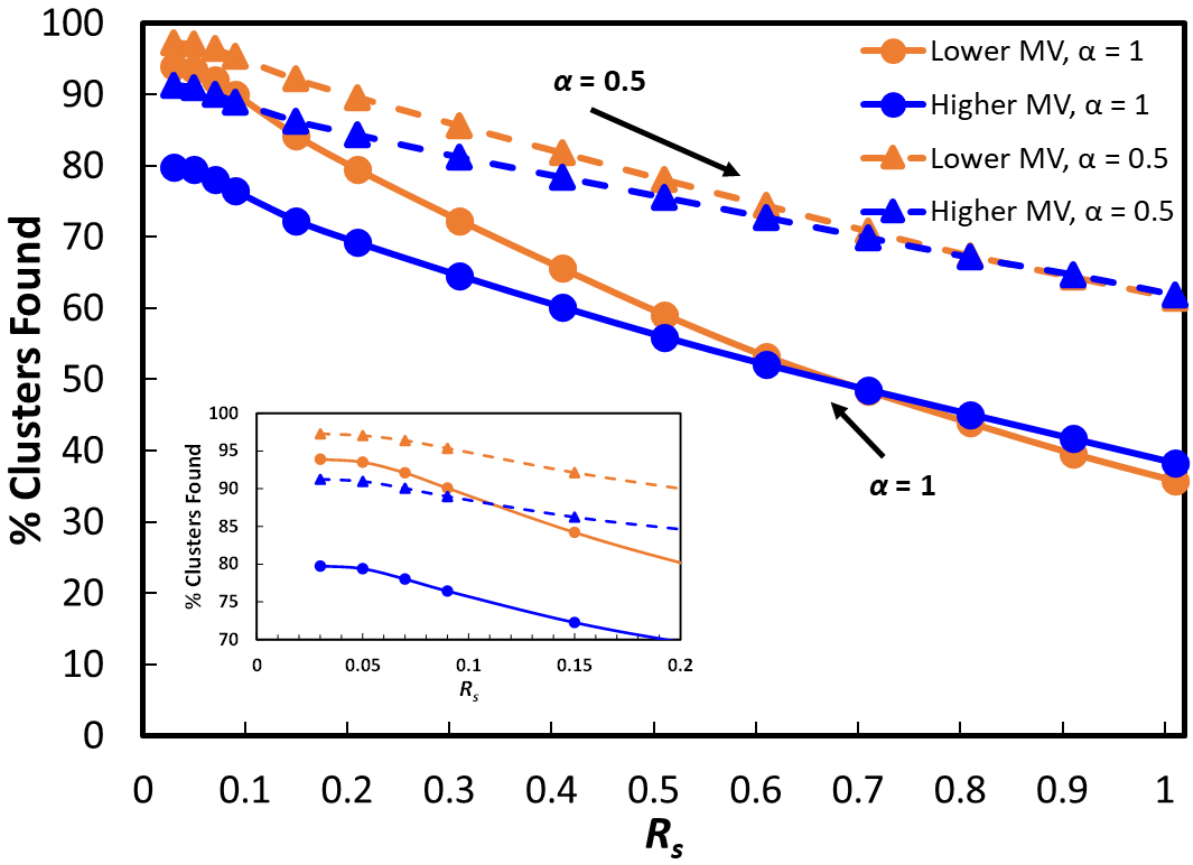


Figure 2.8. Percentage of clusters found (with respect to the number of components in each chromatogram) for the Lower MV and Higher MV analyte sets and an $\alpha = 0.5$ and 1 as a function of R_s. The decrease in α from 1 to 0.5 can be achieved through a 4-fold increase in separation efficiency, N. Inset is a zoomed-in region at low R_s.

2.4.4 *Factors that impact MCM performance*

In the analysis by the MCM, both FP and FN were observed. A FP occurs when the method locates an extra analyte cluster, while a FN results when the method fails to find a true analyte. False positives can arise for a few reasons, but principally because there is a sufficient number of pure *m/z* for a given analyte to comprise separate analyte clusters due to slight variation in retention time (i.e., splitting of a single analyte), or the overlap of two or more analytes at low R_s causes shared *m/z* to be detected as an extra analyte cluster. In the former case, an optimal signal threshold or increasing the *S/N* in the data could reduce the number of extra analyte clusters found, as noise

adversely impacts the retention time and peak width measurements. In the latter case, R_s needs to be low (but not completely overlapped, R_s ~ 0.1-0.6) with an adequate number of shared *m/z* that fall below the width threshold to result in a FP. Reducing the width threshold helps to eliminate the detection of FP that arise by these means, as the shared *m/z* would have elevated w_b values. False negatives can occur in cases of poor R_s and/or similar mass spectra, causing two or more analytes to be detected as one. Indeed, this effect is observed for the Higher MV set in Figure 2.7, which produces more FN than the Lower MV set due to the increased likelihood of increased mass spectral similarity of overlapping analyte peaks. As the saturation of the chromatogram increases, the number of FN increases due to increasing overlap. Nonetheless, operated in an automated fashion as in this study, the R_s limit of the MCM is ~0.05. Below this R_s limit, the automated MCM is frequently unable to distinguish between two or more analytes, and they would be detected as one analyte cluster. As determined previously, analyst supervision can bring down the R_s limit of the method to ~0.03 [23].

Other variables not studied herein, such as variation in peak widths, peak heights (concentration), and noise, would also affect MCM performance. Peak width variability throughout the chromatogram may actually improve the ability to find analytes at lower R_s due to separation on the peak width dimension. However, it may be more difficult to determine an appropriate width threshold to apply, and the benefit relies heavily on neighboring peaks having differences in peak widths. For the selected signal threshold, variation in peak heights (i.e., having a range of analyte concentrations) may lead to more FN for low concentration analytes. This may also make it more likely to observe *m/z* distribution patterns that manifest from extreme intensity ratios between shared *m/z* when R_s is sufficient to provide two local maxima (Supporting Information, 2.5.2). Alternatively, extreme peak height ratios between neighboring analytes may

result in more FN because the minimum R_s required to observe two local maxima increases with more extreme ratios [31]. Noise can also affect the precision of retention time and peak width measurements. High levels of noise may necessitate more smoothing and/or a higher signal threshold, which could reduce the number of *m/z* included.

The randomness of the chromatograms (distribution of the analytes) and analyte selection (mass spectral diversity) greatly impact the observed performance of the MCM. The observed standard deviations of approximately ±7% and ±9% (in terms of the percentage of analyte clusters found at the applied width threshold, reported as an average across all cluster box sizes) for the Lower MV and Higher MV sets at $\alpha = 1$, respectively, are mainly detailing the variation due to the chromatographic randomness. To give some insight into the potential variation between chromatograms, there are $\sim 4\times10^{45}$ possible permutations of random analyte selection from the Higher MV set at $\alpha = 1$, and $\sim 8\times10^{24}$ possible permutations from the Lower MV set. The user preferences of the method, such as signal threshold and width threshold selection, can also introduce some variation into the results; however, this variation is still inherent to the data as these variables are simply defining what data is to be analyzed by the method. Finally the findings in this study are tempered by the current performance capability of the MCM algorithm, which is an ongoing development.

2.5 SUPPORTING INFORMATION

The Supporting Information provides the determination of the width threshold applied in the MCM. Then, a demonstration of *m/z* distribution patterns in the cluster plot as a function of R_s and match value (MV) is included. An example of the benefit of utilizing MCM information in deconvolution of two analytes co-eluting at low R_s via MCR-ALS is illustrated. Finally, a

demonstration of a revised version of the MCM, the MCM 2.0, with different requirements for identifying clusters, is included.

2.5.1 *Width Threshold Determination*

The width threshold was determined using a separate set of chromatograms (n_c = 20, α = 1, Higher MV analyte set) with the prior knowledge that all peak widths were simulated to be 1 s, and therefore, sufficient deviation from this value was indicative of some degree of overlap. The width threshold was incrementally increased from 1.01 to 2.5 s. The number of false positives (FP) and false negatives (FN) were determined for selected potential width threshold values by comparing the simulated retention times with the analyte cluster center locations as discovered by the MCM.

A peak width threshold of 1.1 s (4.4σ) was determined, as this threshold simultaneously limited the number of FP (FDR < 5%) while maximizing the number of true positives (TP) relative to higher thresholds. Lower thresholds exhibited a severe loss in the number of TP found (or an increase in FN). If the peak widths were not constant, determination of a suitable threshold would be more challenging; however, one proposed method would be to apply a user-specified threshold, e.g., 5σ of the average peak width, taking into account that the peak width of a single analyte eluting at a constant retention time can typically differ by as much as 10-15% [23].

2.5.2 *Distribution Patterns of m/z as a Function of R_s and MV*

A demonstration was carried out to visually investigate *m/z* distribution patterns in the cluster plot format as a function of R_s and analyte MV, in which several pairs of analytes were simulated at decreasing R_s and the resulting cluster plots were observed. The four pairs of analytes and their corresponding MV are provided in Table 2.S1. Head-to-tail plots of the mass spectra for

these analyte pairs are shown in Figure 2.S1. Simulations were performed using the parameters in Table 2.S2, with constant peak heights, peak areas, and peak widths for all analytes and simulations. A single Gaussian-distributed random noise profile was generated and applied to every 10-s simulation for the purpose of representing the effect of noise (*S/N* = 100 in the TIC) consistently throughout the series of simulations. A 21-point boxcar was applied for smoothing prior to data reduction via the MCM, with MCM parameters summarized in Table 2.S2.

Table 2.S1. Analyte pairs and corresponding match values for the *m/z* distribution pattern demonstration.

Analyte Pair Set	**Analytes**	**Match Value**
Low MV	cyclohexane and benzene	35
Medium MV	tetradecane and 2-undecanone	501
High MV, selective *m/z*	undecane and dodecane	908
High MV, isomers	1,2,3-trichlorobenzene and 1,3,5-trichlorobenzene	981

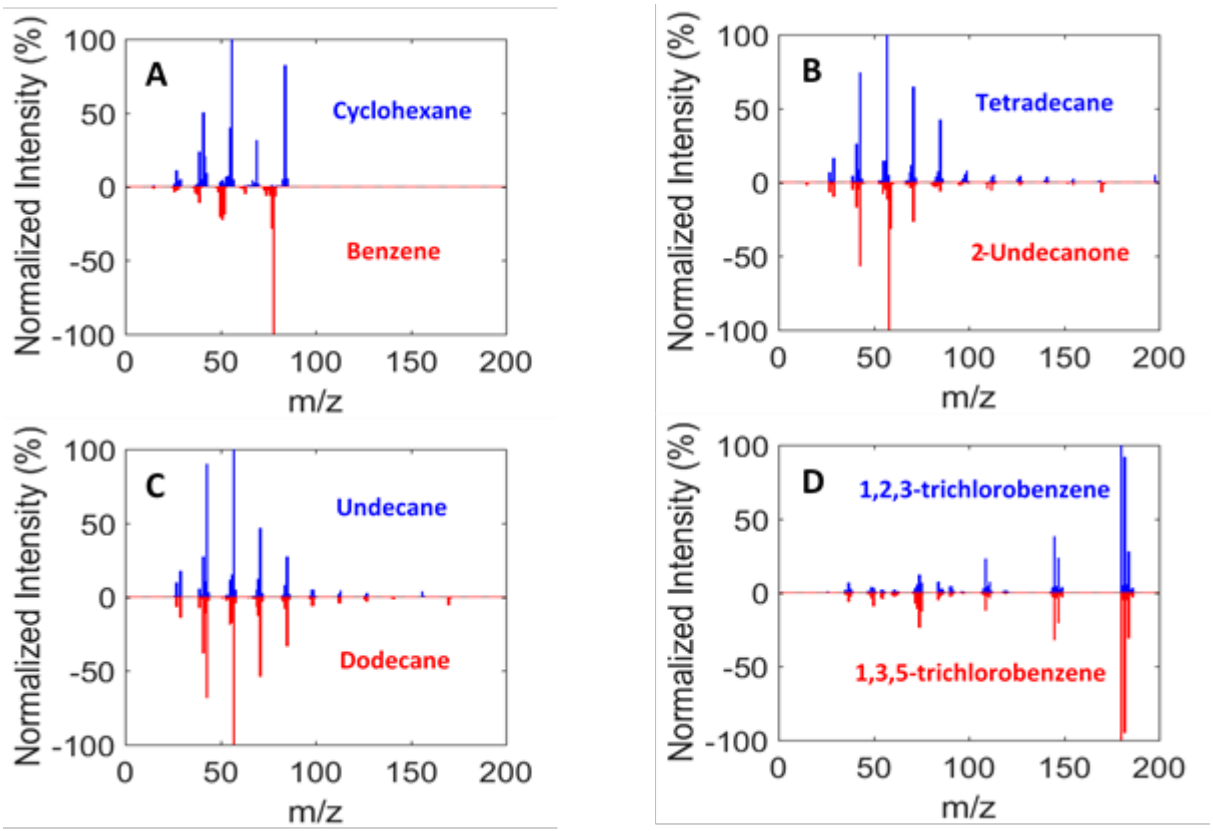


Figure 2.S1. Head-to-tail mass spectra (normalized to the base peak) of all four analyte pairs used in the *m/z* distribution pattern demonstration: (A) Low MV; (B) Medium MV; (C) High MV, selective *m/z*; and (D) High MV, isomers.

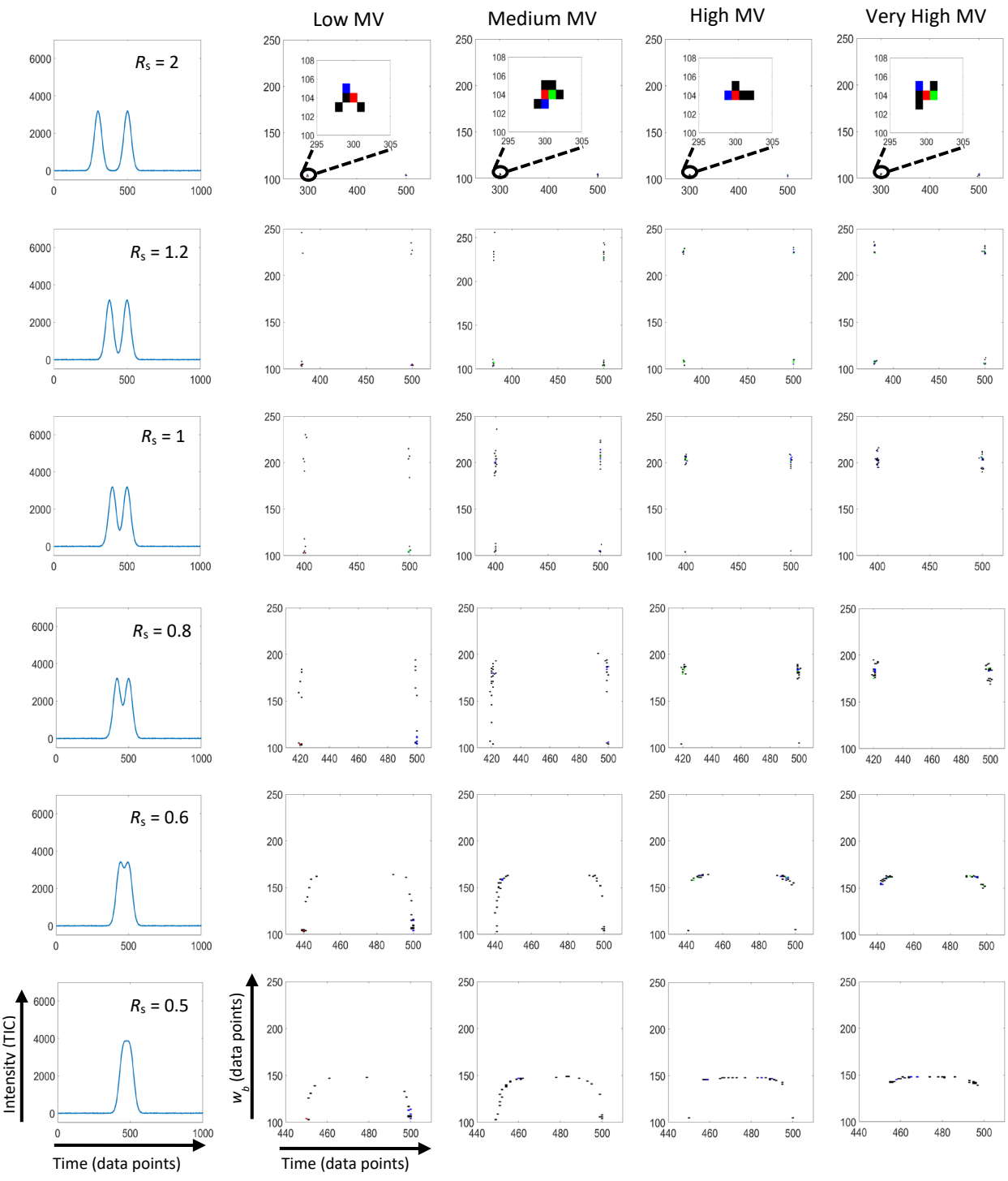
Table 2.S2. Simulation and MCM parameters for the *m/z* distribution pattern demonstration.

Parameter	**Conditions Studied**
Total separation time, t_{sep}	10 s
Number of components, m	2
Peak width-at-base, w_b	1 s
Peak area	200,000 (TIC)
Signal-to-noise ratio, *S/N*	100 (TIC)
Data collection rate	100 Hz
Resolution, R_s	2, 1.2, 1, 0.8, 0.6, 0.5, 0.3, 0.2, 0.1, 0.05
Signal threshold (MCM)	20
Width threshold (MCM)	5 s

The distribution of the cluster points in a cluster plot is related to the R_s as well as the degree of mass spectral similarity (represented by analyte MV) among other variables. Figure 2.S2 depicts how these distribution patterns change with R_s and analyte MV. When baseline resolved,

both analytes in the pair comprise a distinct, pure cluster. However, once overlap begins ($R_s < 1.5$), some shared *m/z* begin to appear at elevated widths in a vertical column. The vertical column persists until $R_s \approx 0.6$, where a horseshoe pattern begins to emerge. This pattern is observed due to changes in the retention time of the local maxima, where the maxima for a single *m/z* move to more intermediate retention times compared to more selective *m/z*. This continues until the two maxima become one ($R_s < 0.5$), assuming equal peak heights. For peak height ratios other than unity, the minimum R_s at which two maxima are observed increases with more extreme ratios [31]. At $R_s < 0.1$, the horseshoe pattern gives way to a dense cluster of *m/z* with a few locations of increased density.

The maximum width of the shared *m/z* is directly related to the R_s between analytes until baseline resolution is achieved, and second, the number of elevated *m/z* is related to the R_s and MV of the analyte pairs, where lower R_s ($0.2 < R_s < 1$) and higher MV give rise to more *m/z* at elevated widths. Visually, the distribution pattern is heavily dependent on the similarities (and differences) between analyte mass spectra, particularly the intensity ratios of shared *m/z* between the two analytes. In general, the higher intensity peak will present at a lower width than its shared lower intensity peak. Additionally, more extreme intensity ratios result in the lower intensity peak of the pair more readily appearing at an elevated width.


Low MV
Medium MV
High MV
Very High MV
R_s = 2
R_s = 1.2
R_s = 1
R_s = 0.8
R_s = 0.6
R_s = 0.5
Intensity (TIC)
Time (data points)
w_b (data points)
Time (data points)

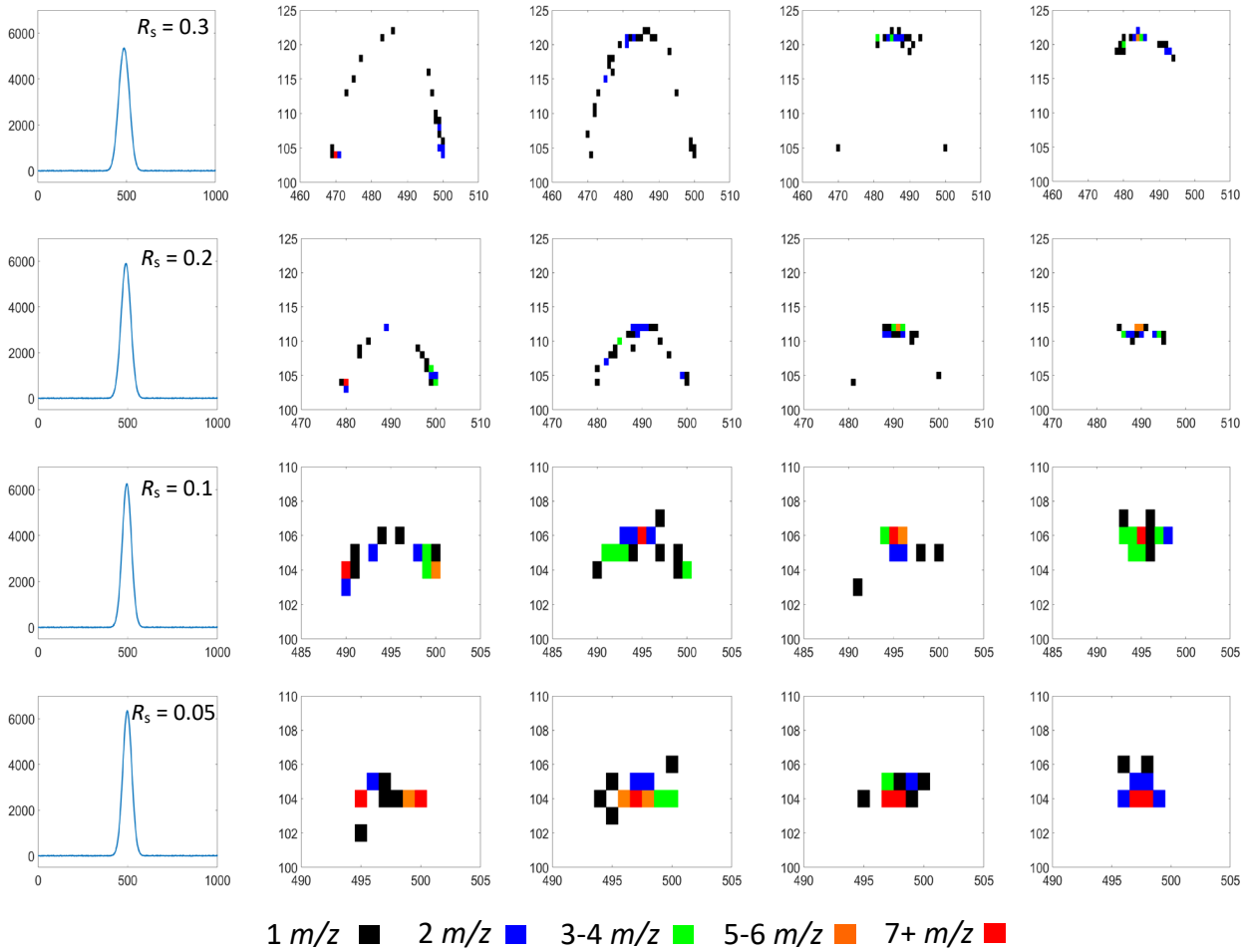


Figure 2.S2. Visualizing *m/z* distribution patterns as a function of R_s and MV. Each row is a different simulated R_s, with the TIC of the two analytes shown in the first column (plotted as TIC intensity versus time in data points). The remaining four columns are the cluster plots of the four analyte pairs (Table 2.S1): Low MV in column 2, Medium MV in column 3, High MV with selective *m/z* in column 4, and High MV isomers in column 5. Time (x-axis) and width-at-base (y-axis in columns 2-5) are reported in data points (1 data point = 10 ms). Axes labels are demonstrated in row 6. The x-axis and y-axis scales are adjusted to give the best view of the distribution patterns depending on the resolution and maximum width-at-base observed. A legend for the color-coded *m/z* is provided after the final row.

2.5.3 *MCM Assisted MCR-ALS versus Unconstrained MCR-ALS*

To illustrate the benefit of the MCM in the deconvolution of co-eluting compounds at a low R_s, cyclohexane and benzene were simulated at R_s = 0.04 and *S/N* of 10. A 10 s section of chromatogram with this co-elution was simulated four times, each with a unique noise profile.

Each of these chromatographic regions were subjected to the MCM to determine the selective *m/z* for each component. A width threshold and suitable signal threshold were applied after smoothing the data by a seven-point boxcar. The mass cluster plot of each replicate was observed and the selective *m/z* were chosen with supervision and summed together over the chromatographic region to create a "pure" peak profile for each component that was used as a hard constraint in MCR-ALS. MCR-ALS models were generated using PLS Toolbox version 8.1.1 (Eigenvector Research, Inc., Wenatchee, WA, U.S.A.) in Matlab. Two factor constrained (using the MCM peak profile as information) and unconstrained MCR-ALS models were applied to decompose the chromatographic region of overlap using the alternating least squares algorithm with a fast non-negativity constraint, based on an algorithm by Bro and de Jong [32]. Although technically all MCR-ALS models performed herein are constrained, the terminology "constrained" versus "unconstrained" here refers to the usage of equality constraints; i.e., constrained MCR-ALS models utilized "pure" concentration profiles obtained from the MCM as a contribution equality constraint, while unconstrained MCR-ALS models did not employ an equality constraint. Analyte identification and quantification were both evaluated using match value (MV) [29] and %error in peak area, respectively. The loadings were extracted to determine the deconvoluted mass spectrum and peak area of each isolated component from each model. The modeled peak areas were compared to the simulated peak areas using a percent error calculation with the formula

$$abs\,\%error = \left|\frac{A_{\mathrm{sim}} - A_{\mathrm{mod}}}{A_{\mathrm{sim}}}\right| \times 100 \qquad (2.S1)$$

where A_{sim} is the expected peak area defined during the simulation and A_{mod} is the modeled peak area resulting from the deconvolution. Match values between the deconvoluted mass spectra and the library mass spectra from NIST were calculated using a formula based on the one defined by Stein [29].

$$MV = \frac{1000 \times \left(\sum n \cdot [A_S \cdot A_L]^{1/2}\right)^2}{\sum n \cdot A_S \times \sum n \cdot A_L} \quad (2.S2)$$

In Eq. 2.S2, n is a vector of *m/z* values, A_S is a vector of *m/z* abundances in the user spectra and A_L is a vector of *m/z* abundances in the library/reference spectra.

Cyclohexane and benzene were chosen as analytes due to their use in the original development of the MCM [23] and their prevalence in real separations. The two analytes were simulated at an R_s of 0.3 with *S/N* 100 in the TIC, and no width or area distributions, presented in Figure 2.S3A. The corresponding mass cluster plot shows that there is an adequate number of selective *m/z* for each analyte, with a few shared *m/z* located at elevated peak widths and intermediate retention times shown in Figure 2.S3B. In order to demonstrate the utility of the MCM, it was intentionally applied to a difficult case of deconvolution (R_s = 0.04 and *S/N* = 10 in the TIC) to which the application of unconstrained MCR-ALS would expectedly result in significant errors in peak area and low MV as demonstrated in a previous study [22]. A representative simulated chromatogram of this overlap is shown in Figure 2.S4A. The analytes were simulated at low concentrations where the *S/N* in the TIC is 10 (having the same noise magnitude as the *S/N* = 100 simulations), and the *S/N* of individual *m/z* could potentially be much lower than 10. This caused fewer *m/z* to be included in the analysis compared to the *S/N* = 100 case, and with already few *m/z* for each analyte, there were only approximately five *m/z* included in the analysis for each analyte. For this case, often only one analyte cluster was found by the automated method due to noise-caused variation in the retention time measurements and the limited number of *m/z*. This case required supervised cluster box location to determine the selective *m/z* for each analyte to use in MCM assisted MCR-ALS, presented in Figure 2.S4B. For all replicates, both components were appropriately modeled by a two-factor model, with the noise being captured in the residuals. A comparison of the deconvolution results from MCM assisted

and unconstrained MCR-ALS is presented in Table 2.S3. In general, the %error in peak area and %RSD were significantly lower using MCM assisted MCR-ALS compared to unconstrained MCR-ALS. The deconvoluted peak profiles for MCM assisted and unconstrained MCR-ALS are shown overlaid in Figure 2.S5A and B, respectively. The deconvoluted peak profiles obtained by MCM assisted MCR-ALS were generally consistent for each analyte, giving low %RSD in Figure 2.S5A. On the other hand, the deconvoluted peak profiles obtained from unconstrained MCR-ALS were inconsistent for a single analyte and did not match the simulated R_s, presented in Figure 2.S5B. For unconstrained MCR-ALS, the deconvoluted loadings gave an overestimated peak area for one analyte and underestimated peak area for the other, and vice versa, resulting in high %RSD (Table 2.S3). Other than the deconvoluted peak area, these peak profiles also appeared more variable than those of MCM assisted MCR-ALS in terms of the retention time and peak widths of the deconvoluted peaks for each analyte. The match values (MV) between deconvoluted and reference library spectra improved upon deconvolution by MCM assisted MCR-ALS relative to unconstrained MCR-ALS. The deconvoluted mass spectra for cyclohexane and benzene obtained from MCM assisted MCR-ALS are shown in Figure 2.S5C and D, respectively. In general, an analyst may achieve more confident compound identification if MCM assisted MCR-ALS is utilized in such cases of overlap where both R_s and *S/N* are low.

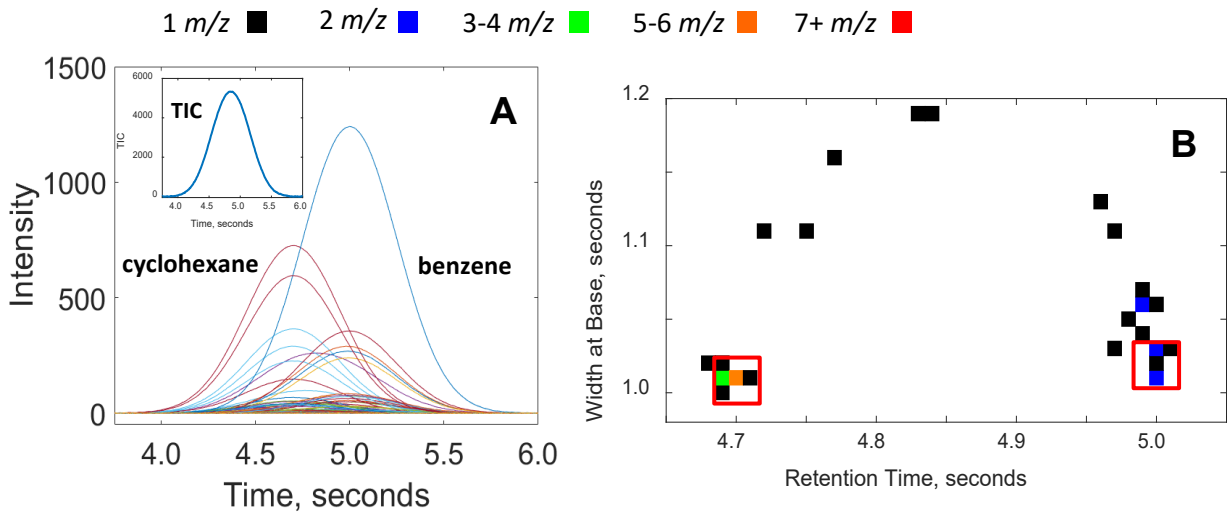


Figure 2.S3. Demonstration of component overlap of cyclohexane and benzene. (A) Region of component overlap in which cyclohexane and benzene are simulated at $R_s = 0.3$ with $S/N = 100$. The TIC for the region of overlap is shown inset. (B) Corresponding cluster plot with cluster boxes included. The cluster plot demonstrates that there are both selective and shared *m/z*, as seen by the horseshoe shape.

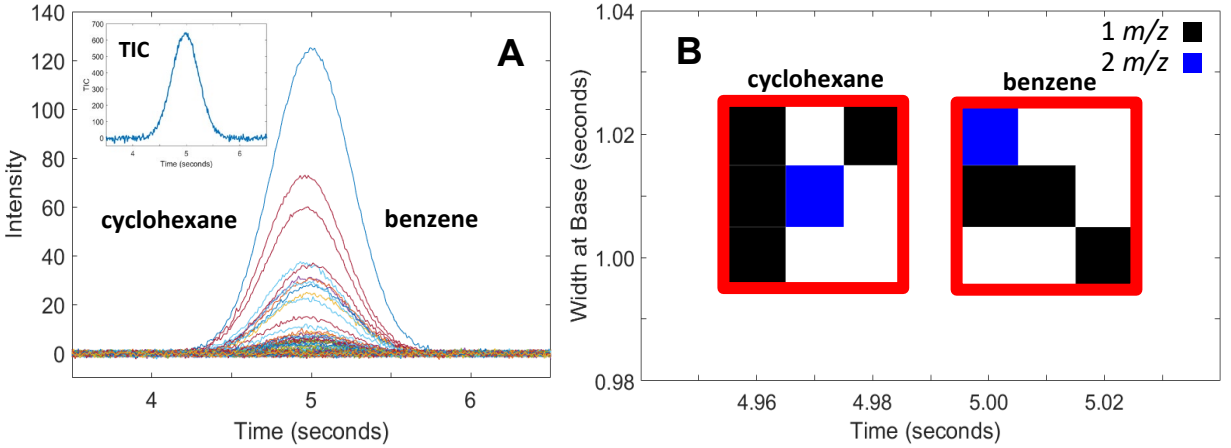


Figure 2.S4. Challenging case of overlap of cyclohexane and benzene for MCM and MCR-ALS. (A) Region of component overlap from one representative replicate in which cyclohexane and benzene are simulated at $R_s = 0.04$ with $S/N = 10$, which is generally considered a difficult case for deconvolution. The TIC for the region of overlap is shown inset. (B) Corresponding cluster plot with cluster boxes included. Cluster box location was performed in a supervised fashion, and the selective *m/z* were identified and summed to provide a pure peak profile for each analyte to use as constraints in MCM assisted MCR-ALS.

Table 2.S3. Results of unconstrained and MCM assisted MCR-ALS of the cyclohexane and benzene pair at $R_s = 0.04$ and $S/N = 10$. The absolute percent error (abs. %error) and %RSD of the peak areas of each component determined from the models with respect to the expected peak area of 20,000 are reported.

	Peak area (per replicate)					
Component	**1**	**2**	**3**	**4**	**Average abs. %error**	**%RSD**
cyclohexane (unconstrained)	31035	31287	6530	6390	61.8	75.8
benzene (unconstrained)	7916	7501	32557	32689	62.3	71.3
cyclohexane (MCM assisted)	18742	18848	19306	19127	5.0	1.4
benzene (MCM assisted)	20190	19817	19717	19825	1.0	1.0

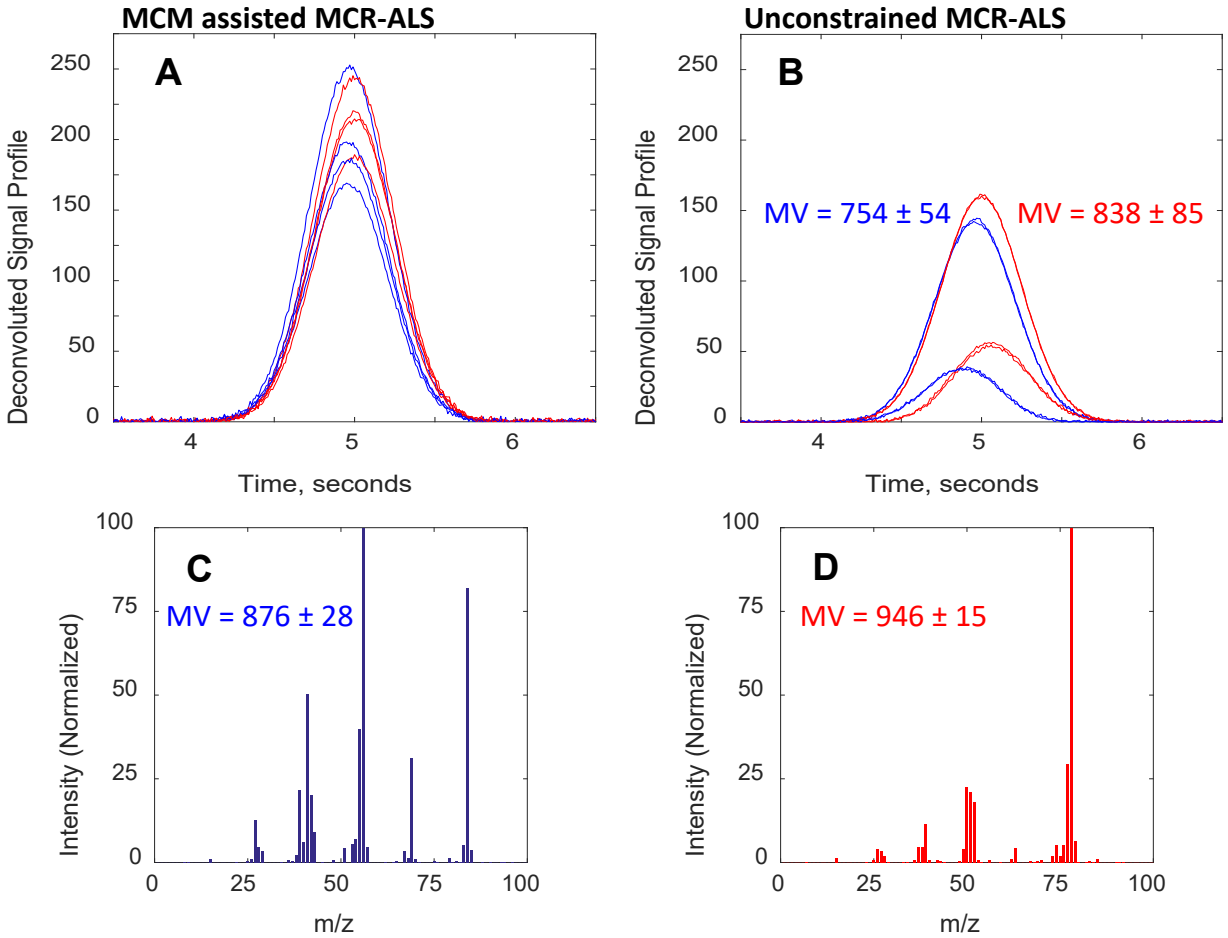


Figure 2.S5. Model loadings from MCM assisted MCR-ALS and unconstrained MCR-ALS. Deconvoluted peak profiles for each component with all replicates overlaid (cyclohexane: blue, benzene: red) from (A) MCM assisted MCR-ALS and (B) unconstrained MCR-ALS. The average match value and standard deviation of the deconvoluted mass spectrum of each component with respect to a reference library spectrum is shown. Deconvoluted mass spectra (normalized to the base peak) obtained from MCM assisted MCR-ALS for (C) cyclohexane and (D) benzene, with the average match value and standard deviation shown.

2.5.4 *Demonstration of Revised MCM 2.0*

The current limitations of the MCM, with regard to false negative occurrences due to severe overlap and co-elution of analytes with many shared *m/z*, derive from the "strict" requirements in defining a cluster. Improvements to the current MCM are necessary to develop "smarter" requirements for defining clusters, addressing the shortcomings by moving away from a *m/z* requirement (three *m/z* all within two data points) and current cluster classification method, which hinders analyte discovery at ultra-low resolutions. The new algorithm, referred to as the mass cluster method 2.0 (MCM 2.0), locates clusters based on minimum width because of the purity associated with low width *m/z*. However, the algorithm does not penalize higher width *m/z*, acting as a probe of the cluster plot to find the minimum width at each data point in the chromatogram. Additionally, the algorithm does not have a *m/z* requirement and can define a cluster using a single *m/z*, which addresses the challenge of analyte co-elution with high match values. Briefly, the new algorithm is similar to the original mass cluster method with an identical data reduction step. Following data reduction, the new algorithm finds the minimum width of each retention time data point in the cluster plot that has at least one *m/z*. Based upon a user defined window size (similar to box size), the algorithm starts at the beginning of the cluster plot and centers the window on the retention time of the first minimum width *m/z*, with the lower bound of the window located at the minimum width. The algorithm then tests the next few retention time data points for a minimum width less than the current minimum width, until it reaches the location of the next resolved center (a distance of one window length from the original center). If the minimum width does not decrease over that interval, the window location is optimized by shifting the window in the time dimension and choosing the location that encompasses the greatest number of *m/z* while keeping the original center/minimum width *m/z* in the window. The process then repeats, starting at the location of the

next resolved center. However, if the minimum width does decrease within the interval, the window is centered at the new minimum and the process of probing for a lower minimum width is repeated with a newly defined next resolved center. Other considerations in the cluster location step are included in the algorithm, however, descriptions are omitted for brevity.

A preliminary simulation to compare MCM and MCM 2.0 performance is shown in Figure 2.S6A [22]. The chromatogram was simulated with parameters listed in Table 2.S4. The analyte set consisted of a 100 analyte subset of the Higher MV analyte set (defined in Table A.1 in Appendix A), with MV distribution as seen in Figure 2.S6B. Both the MCM and MCM 2.0 were applied to the simulated chromatogram following smoothing accomplished by a 21-point boxcar, with method parameters located in Table 2.S4. The performance results are located in Table 2.S5, where it can be seen that the MCM 2.0 discovered more total clusters and had fewer false positives and false negatives than the original MCM. A proposed study is an extension of this preliminary simulation in which the new algorithm is tuned (window size, signal threshold, width threshold) and then used as a tool to study the effect of peak capacity, number of components, separation efficiency, and mass spectral similarity on the number of detectable analytes via simulated data. It is hypothesized that the MCM 2.0 will be extremely sensitive to changes in width threshold, making the selection of an appropriate width threshold very important.

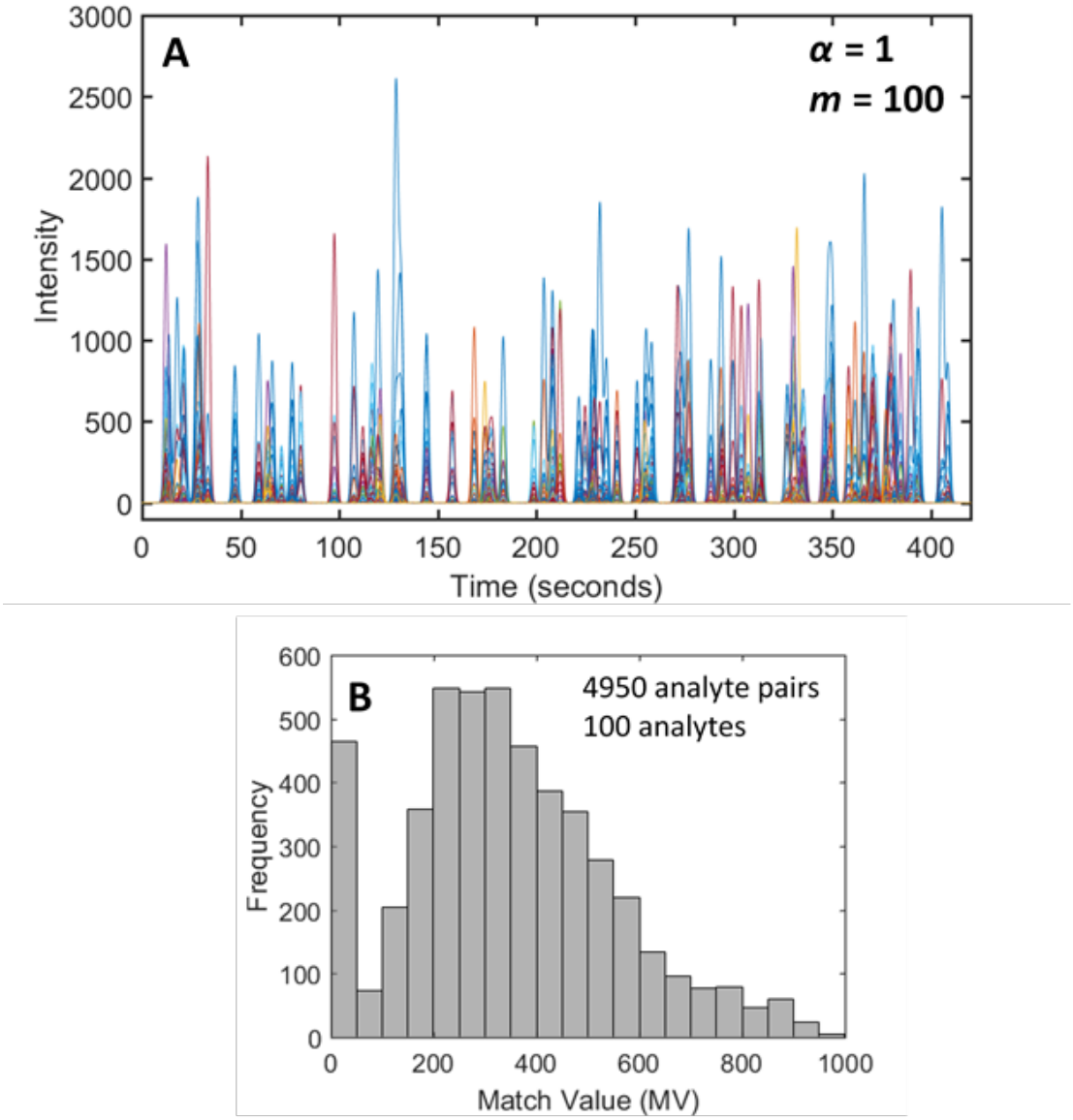


Figure 2.S6. Simulation for MCM versus MCM 2.0 comparison. (A) Chromatogram with all *m/z* overlaid, replicated from [22]. (B) Analyte MV histogram, where the calculated match value of each of the 4950 possible analyte pair combinations is shown.

Table 2.S4. Preliminary simulation and method parameters for MCM versus MCM 2.0 comparison.

Parameter	Conditions Studied
Separation window	400 s (centered on a 420 s separation)
Peak capacity, n_c	100
Number of components, m	100
Saturation factor, α	1
Peak width at base, w_b	4 s
Peak area	1,000,000 (TIC)
Signal-to-noise ratio, *S/N*	100 (TIC)
Data collection rate	100 Hz (spectra/s)
Analyte set	See Table A.1
Signal threshold (MCM/MCM 2.0)	20 (raw *S/N* threshold of ~10)
Width threshold (MCM/MCM 2.0)	4.4 s
Box size (MCM)	3 x 3
Window size (MCM 2.0)	19 x 19

Table 2.S5. Results for MCM versus MCM 2.0 comparison. Eleven analytes were undiscoverable due to no *m/z* present below the applied width threshold. Clusters were considered shifted if the center of the cluster was shifted by 8 data points or more from the simulated retention time.

Comparison Metric	MCM	MCM 2.0
Total clusters found	78	89
FP	17	7
FN	39	18
Undiscoverable	11	11
Shifted	2	8

2.6 CONCLUSION

This study provided an objective framework to determine to what extent all of the analyte components in a given sample can be confidently revealed by GC-MS using the current version of the MCM, with a prospectus on MCM-assisted chemometric deconvolution using MCR-ALS. While the MCM specifically was evaluated in the context of the SMO, the results in general may have implications for other similar software methods and tools. Specifically, the SMO-based approach utilized in this study provides the analytical community with a framework to more rigorously evaluate and validate deconvolution software tools of others. Additionally, the MCM

may be appropriate for applications of liquid chromatography coupled with TOFMS or diode array detection (DAD), though the focus on GC-TOFMS herein was implemented to assess the MCM under more ideal conditions for application (i.e., fast separations with small peak widths and utilization of high mass spectral scan rates of the TOFMS). Looking forward, a more detailed study of the performance of the MCM as a function of width threshold would be intriguing but was not the focus of this study. Additional improvements to the MCM algorithm would benefit such a study.

2.7 REFERENCES

[1] Y. Wang, L. Xu, H. Shen, J. Wang, W. Liu, X. Zhu, R. Wang, X. Sun, L. Liu, Metabolomic analysis with GC-MS to reveal potential metabolites and biological pathways involved in Pb & Cd stress response of radish roots, Sci. Rep. 5 (2015) 18296. https://doi.org/10.1038/srep18296.

[2] B.C. Reaser, S. Yang, B.D. Fitz, B.A. Parsons, M.E. Lidstrom, R.E. Synovec, Non-targeted determination of 13 C-labeling in the Methylobacterium extorquens AM1 metabolome using the two-dimensional mass cluster method and principal component analysis, J. Chromatogr. A. 1432 (2016) 111–121. https://doi.org/10.1016/j.chroma.2015.12.088.

[3] A. Garcia, C. Barbas, Gas chromatography-mass spectrometry (GC-MS)-based metabolomics, in: T.O. Metz (Ed.), Metabolic Profiling, Humana Press, 2011: pp. 191–204. https://doi.org/10.1007/978-1-61737-985-7_11.

[4] A.M. Leffler, P.B. Smith, A. de Armas, F.L. Dorman, The analytical investigation of synthetic street drugs containing cathinone analogs, Forensic Sci. Int. 234 (2014) 50–56. https://doi.org/10.1016/j.forsciint.2013.08.021.

[5] M. Carson, S. Kerrigan, Quantification of suvorexant in urine using gas chromatography/mass spectrometry, J. Chromatogr. B. 1040 (2017) 289–294. https://doi.org/10.1016/j.jchromb.2016.10.042.

[6] S.O. Fakayode, B.S. Mitchell, D.A. Pollard, Determination of boiling point of petrochemicals by gas chromatography–mass spectrometry and multivariate regression analysis of structural activity relationship, Talanta. 126 (2014) 151–156. https://doi.org/10.1016/j.talanta.2014.03.037.

[7] R.-Z. Zhou, J. Jiang, T. Mao, Y.-S. Zhao, Y. Lu, Multiresidue analysis of environmental pollutants in edible vegetable oils by gas chromatography–tandem mass spectrometry, Food Chem. 207 (2016) 43–50. https://doi.org/10.1016/j.foodchem.2016.03.071.

[8] B. Gruber, J. Schneider, M. Föhlinger, J. Buters, R. Zimmermann, G. Matuschek, A minimal-invasive method for systemic bio-monitoring of the environmental pollutant phenanthrene in humans: Thermal extraction and gas chromatography − mass spectrometry

from 1 mL capillary blood, J. Chromatogr. A. 1487 (2017) 254–257. https://doi.org/10.1016/j.chroma.2017.01.045.
[9] 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.
[10] L. Barp, G. Purcaro, F.A. Franchina, M. Zoccali, D. Sciarrone, P.Q. Tranchida, L. Mondello, Determination of phthalate esters in vegetable oils using direct immersion solid-phase microextraction and fast gas chromatography coupled with triple quadrupole mass spectrometry, Anal. Chim. Acta. 887 (2015) 237–244. https://doi.org/10.1016/j.aca.2015.06.039.
[11] D. Yan, Y.F. Wong, R.A. Shellie, P.J. Marriott, S.P. Whittock, A. Koutoulis, Assessment of the phytochemical profiles of novel hop (Humulus lupulus L.) cultivars: A potential route to beer crafting, Food Chem. 275 (2019) 15–23. https://doi.org/10.1016/j.foodchem.2018.09.082.
[12] 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.
[13] R.B. Wilson, J.C. Hoggard, R.E. Synovec, Fast, high peak capacity separations in gas chromatography-time-of-flight mass spectrometry, Anal. Chem. 84 (2012) 4167–4173. https://doi.org/10.1021/ac300481k.
[14] R.B. Wilson, B.D. Fitz, B.C. Mannion, T. Lai, R.K. Olund, J.C. Hoggard, R.E. Synovec, High-speed cryo-focusing injection for gas chromatography: Reduction of injection band broadening with concentration enrichment, Talanta. 97 (2012) 9–15. https://doi.org/10.1016/j.talanta.2012.03.054.
[15] 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.
[16] M.M. van Deursen, J. Beens, H.-G. Janssen, P.A. Leclercq, C.A. Cramers, Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography, J. Chromatogr. A. 878 (2000) 205–213. https://doi.org/10.1016/S0021-9673(00)00300-9.
[17] J.C. Giddings, Unified Separation Science, John Wiley & Sons, Inc., New York, 1991.
[18] 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.
[19] C.G. Fraga, C.A. Bruckner, R.E. Synovec, Increasing the number of analyzable peaks in comprehensive two-dimensional separations through chemometrics, Anal. Chem. 73 (2001) 675–683. https://doi.org/10.1021/ac0010025.
[20] 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.
[21] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.

[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] J.M. Davis, J.C. Giddings, Statistical method for estimation of number of components from single complex chromatograms: application to experimental chromatograms, Anal. Chem. 57 (1985) 2178–2182. https://doi.org/10.1021/ac00289a003.
[28] J.M. Davis, Theory of the probability of total resolution in chromatograms with systematic variation of average peak spacing and peak width, J. Chromatogr. A. 1588 (2019) 150–158. https://doi.org/10.1016/j.chroma.2018.12.031.
[29] S.E. Stein, An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, J. Am. Soc. Mass Spectrom. 10 (1999) 770–781. https://doi.org/10.1016/S1044-0305(99)00047-1.
[30] J.M. Davis, S.C. Rutan, P.W. Carr, Relationship between selectivity and average resolution in comprehensive two-dimensional separations with spectroscopic detection, J. Chromatogr. A. 1218 (2011) 5819–5828. https://doi.org/10.1016/j.chroma.2011.06.086.
[31] 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.
[32] R. Bro, S.D. Jong, A fast non-negativity-constrained least squares algorithm, J. Chemom. 11 (1997) 393–401. https://doi.org/10.1002/(SICI)1099-128X(199709/10)11:5<393::AID-CEM483>3.0.CO;2-L.

Chapter 3. Unsupervised Feature Selection of Gas Chromatography with Mass Spectrometry by Variance Thresholding[3]

3.1 INTRODUCTION

Gas chromatography with 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 [1–9]. Separation of a mixture by gas chromatography followed by the detection of individual analytes via mass spectrometry provides an ideal dataset for comparing chemical composition between multiple samples, as GC-MS quantitatively describes the volatile chemical profile of a sample. Common cross-sample analysis goals include classifying samples based on similarities and differences in their chemical composition and predicting sample properties based on chemical information [10], which can be carried out by chemometric methods such as principle component analysis (PCA) and partial least squares (PLS), respectively. Chemometric methods seek to extract relevant information, which is often obscured, from chemical data using mathematical means. However, the vast number of variables that result from typical GC-MS separations (30-minute separation with 200 mass channels collected at 10 Hz = 3,600,000 data points for a single sample) can cause issues for mathematical modeling. Many chemometric methods utilize the variation of these features to relate samples to each other or chemically-dependent sample properties. A vast majority of variables can be described as noise and there are often redundant variables. Model performance can suffer when forced to incorporate extraneous and irrelevant variables, and there is the potential for overfitting

[3] This chapter has been adapted from K.L. Berrier, C.N. Cain, and R.E. Synovec, Unsupervised feature selection of gas chromatography with mass spectrometry data by variance thresholding analysis, *in preparation*.

when the number of predictors significantly outweigh the number of samples [11]. In addition, attempting to model a large number of variables can increase computation time and expense.

Feature selection, also known as variable selection, is the process of selecting a subset of the original variables, consisting of only the most relevant features. Feature selection methods remove "junk" variables due to noise or those uncorrelated with a target variable. A target variable is the objective to which the data is to be correlated; depending on the particular analysis goals and experimental design, this could be class membership, an independently measured sample property, etc. When this target variable is known, supervised analysis is being performed. A common supervised feature selection method used in the analysis of GC-MS data is Fisher ratio (F-ratio) analysis, which is an analysis of variance (ANOVA) method [5–7,12,13]. In these methods, the target variable is class membership, which is utilized to find chromatographic variables that exhibit high between-class variance relative to within-class variance. The selected features are statistically likely to be class-distinguishing, i.e., statistically different between classes. While PCA and partial least squares – discriminant analysis (PLS-DA) are unsupervised and supervised counterparts that are often thought of as being feature selection methods, they are nominally feature extraction methods that achieve a reduction in dimensionality through the creation of new variables that are linear combinations of the original variables [8,14–16]. These new variables (i.e., principal components in PCA and latent variables in PLS-DA) are not easily interpretable with respect to the original data, and even interpretation of the loadings (contain information about the contribution of the original variables to the new variables) can have subtleties.

Many other variable selection methods are routinely applied in data science and other statistics- and mathematics-based disciplines. These methods are grouped into several categories, namely filtering and wrapping methods, both of which can be either supervised or unsupervised.

Filtering methods involve ranking the variables based upon a feature selection criterion (e.g., correlation, mutual information) and filtering out the variables below some threshold [17–19]. Often, filtering methods are used as preprocessing methods prior to implementation of a classification or prediction model. Alternatively, wrapper methods use predictor (i.e., model) performance as the criteria for evaluating subsets of features. At their core, wrapper methods are search problems that can be classified as exhaustive, heuristic, or random search [17–19]. Variance thresholding is an unsupervised filtering method used in data science to remove variables with low variance, which have low predictive power [20–24]. In variance thresholding, the analyst must choose a variance threshold, below which features are removed. To our knowledge, variance thresholding on chromatographic data has not been described in the literature.

Herein, we demonstrate the application of variance thresholding to simulated GC-MS data and a previously collected and analyzed yeast metabolome dataset [25–27]. Both datasets are comprised of two chemically distinct classes, although the analysis is performed in an unsupervised fashion where knowledge of the class membership is not considered. With both datasets, the features discovered by the unsupervised variance thresholding approach are compared with those elucidated by supervised F-ratio analysis. Feature selection is accomplished in a pixel-based approach using the simulated data and via a peak table based approach with the yeast samples due to minor retention time shifting. Both datasets contain within-class variation at a level consistent with biological variation (~30% *RSD*), where natural variation is typically the dominating variance term in chromatographic data. The presence of this variation challenges the application of variance thresholding and necessitates the selection of a threshold that differentiates naturally occurring and artificially introduced (e.g., through sample preparation steps, injection, and instrumentation) variation from chemically-significant variation. In this study, the variance of

each pixel/peak was normalized by the mean signal (the square of the mean signal, to give RSD^2) to eliminate the signal dependence on variance. A signal threshold, based upon the baseline noise of the data, is also applied to reduce the presence of artificially high relative variances due to low mean signal values. Application of this unsupervised feature selection method to one-dimensional (1D) GC-MS data is a natural starting point as currently many routine analyses are accomplished by GC coupled with a quadrupole mass spectrometer due to the availability and accessibility of this analytical platform.

3.2 THEORY

The basic premise is that for each analyte peak feature examined above a user-selected signal threshold, the "background" relative signal variance, $RSD^2_B = V_{R,B}$, is sufficiently independent of signal, S, and independent of concentration, and hence, $V_{R,B}$ can be treated as being relatively constant. 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. Additionally, another premise is that $V_{R,B}$ is relatively the same from one analyte to the next. Here, all sources of background variation are the common sources of variation that are present in GC-based experimental design and subsequent chemical measurements. Therefore, for any analyte peak feature in which the total relative signal variance, $V_{R,T} \geq$ user-selected $V_{R,T}$ threshold $\geq V_{R,B}$, this suggests the analyte peak feature has additional source(s) of signal variance, $V_{R,Info}$, that may be useful for contributing to the exploration of potential sample class distinguishing information in the dataset. Mathematical formulation relies upon all sources of measurement uncertainty being statistically independent and so the relative variances are additive. The total relative signal variance, $V_{R,T}$, is simply the sum of the additional source of signal variance and the background relative signal variance,

$$V_{R,T} = V_{R,Info} + V_{R,B} \quad (3.1)$$

The background relative signal variance, $V_{R,B}$, can be expressed as follows,

$$V_{R,B} = V_{R,NatVar} + V_{R,SP} + V_{R,Inj} + V_{R,Det} \quad (3.2)$$

with subscripts corresponding to the four major sources of uncertainty: natural variation (NatVar), such as biological variation, sample preparation (SP), injection (Inj), and detection (Det). In terms of *RSD*, $V_{R,B}$ is given by,

$$V_{R,B} = RSD^2_{NatVar} + RSD^2_{SP} + RSD^2_{Inj} + RSD^2_{Det} \quad (3.3)$$

The first variance contribution, RSD^2_{NatVar}, is often the dominating term, especially for bioanalytical studies such as metabolomics. However, in order to observe this term, the experimental design must incorporate true sample variation, e.g., having independent yeast samples from several different growth media. The RSD_{NatVar} can range from 30 – 50%. The second contribution, RSD_{SP}, is generally the second largest term, typically in the range of 10 – 20%, for solvent extraction-based sample preparation, but can be larger for SPME based sample preparation. The third contribution, RSD^2_{Inj}, is due to the variation in sample volume injection with GC-based instrumentation. The RSD_{Inj} can range from 5 – 10% [28], but can be readily minimized by use of an internal standard or other approaches to normalize the data. The underlying notion, though, is that all three of these contributions are sufficiently independent of concentration, and thus independent of signal. The final contribution, RSD^2_{Det}, is defined in terms of the limit-of-detection (LOD), with the signal at the LOD equal to $3\sigma_N$, where σ_N is the standard deviation of the baseline noise. Thus, RSD^2_{Det} will not be independent of concentration, but for many application purposes, the magnitude of the RSD^2_{Det} contribution in the determination of $V_{R,B}$ will be inconsequential for analyte peak features, except for analytes with signal approaching the LOD. Indeed, at the LOD, based upon the definition, RSD_{Det} = 33%. For an analyte signal 10-fold the

LOD, then RSD_{Det} = 3.3%, and so on. In this proof-of-principle study we accept the possibility that some analyte peaks with low signal, that in all actuality do not contain any additional source of interesting signal variance, may inadvertently be selected due to the inflation of their $V_{R,B}$ by the RSD^2_{Det} contribution. However this issue can also be readily addressed by applying a suitable signal threshold.

For the yeast data examined in this study, the following calculation provides a guide as to magnitude of $V_{R,B}$ to expect, and in turn this knowledge is used to set the $V_{R,T}$ threshold. Using $RSD_{NatVar} \approx 30\%$, $RSD_{SP} \approx 10\%$, $RSD_{Inj} \approx 5\%$, and for an analyte exhibiting a signal at 10-fold the LOD, so $RSD_{Det} \approx 3\%$, the $V_{R,B}$ is determined to be 0.103, or $RSD_B \approx 32\%$, i.e., dominated by the natural variation contribution.

3.3 EXPERIMENTAL

3.3.1 *Chromatographic Simulations*

All simulations were performed in Matlab R2020a (The Mathworks, Inc., Natick, MA, U.S.A.). Chromatograms were simulated to be 520 seconds long centered on a 500 second separation containing fifty analytes randomly and independently distributed throughout the separation space. Analogous to the peak height distributions observed in experimental GC data, an exponential distribution in peak heights was implemented herein with a mean of 100. For each simulation, a random, exponentially-distributed variable representing the peak heights for all 50 analytes was generated. Each simulation consisted of 10 replicates of two chemically distinct classes, Class A and Class B, to give 20 sample replicates total. For both sets of replicates, within class variation of 30% *RSD* was randomly generated for each analyte to mimic $V_{R,B}$ dominated by natural variation as in biological samples. In these simulations, 4 randomly selected analytes were designated to be class-distinguishing, reflected by a change in concentration (i.e., peak area)

between classes. Analyte concentrations in Class B remained nominally the same; concentrations of the selected analytes in Class A were changed by factors of 0.5, 0.67, 1.5, and 2 with respect to the nominal concentrations in Class B, leading to changes in the peak area of these analytes by the same factors. Each analyte was simulated as a Gaussian peak with a randomly generated peak area following an exponential distribution and 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). Once modeled, a randomly selected analyte mass spectrum was multiplied element-wise across each peak to form a series of Gaussian peaks representing the mass channels, *m/z*, having signal for that analyte. Analyte mass spectra at unit mass resolution were obtained from the NIST MS Search 2.0 database (NJ, U.S.A.). The mass spectrum of each analyte was normalized in a way such that the sum of the intensities of all *m/z* would be equal to 1000 for all of the analytes to allow for an average signal-to-noise ratio, *S/N*, of 20 in the TIC. The actual *S/N* for each analyte depended on the exponentially distributed peak height (related to area), and the *S/N* for each *m/z* depended on the intensity of the *m/z*. 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 of the exponentially distributed peak height variable in the TIC of each chromatogram (~100,000). Each of the 100 simulations were randomly and independently generated following the parameters summarized above and in Table 3.1.

Principal component analysis (PCA) was applied to the mean-centered, vectorized data using PLS Toolbox version 8.8.1 (Eigenvector Research Inc., Wenatchee, WA, USA). Creation of two component PCA models was automated for all simulations before and after feature selection. Degree of class separation (DCS) was used as a metric to quantify the classification observed in PCA and compare between other simulations. DCS is defined by

$$\mathrm{DCS} = \frac{D_{\mathrm{A,B}}}{\sqrt{s_{\mathrm{A}}^2 + s_B^2}} \tag{3.4}$$

where $D_{A,B}$ is the Euclidean distance between the centroids of two classes in the PC2 versus PC1 scores plot and s is the standard deviation in the distance of each score in a class from the centroid of that class [5,29].

Table 3.1. 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

3.3.2 *Benchmark Yeast Metabolome Dataset*

The benchmark experimental dataset utilized herein was collected in 2005 [25] and studied previously [25–27]. Briefly, the dataset 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. Data collection parameters can be found elsewhere [25], but briefly the yeast extracts were analyzed using an Agilent 6890N gas chromatograph equipped with an Agilent 7693 auto-injector (Agilent Technologies, Santa Clara, CA, USA) coupled to a LECO Pegasus III TOFMS with a 4D thermal modulator upgrade (LECO, St. Joseph, MI, USA). A sample volume of 1 μL was injected onto the ^{1}D column (RTX-5MS, 20 m × 250 μm i.d. × 0.5 μm, Restek, Bellefonte, PA, 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 ^{2}D column (RTX-200MS, 2 m × 180 μm i.d. × 0.2 μm, Restek, Bellefonte, PA, 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. This dataset has been studied extensively in applications of supervised analysis of variance, and most recently a tile-based F-ratio approach [27]. It is ideal for the present study because there exists natural variation from the three yeast cultures, sample preparation variation from the three extractions, and injection variation from the four injection replicates. For the present study, 6 samples for each class were selected so that two samples from each culture were present for each class. Many possible sample data files (70 injections) with collection dates spanning several months were available for analysis; samples collected within a few days of each other (April 28-May 2, 2005) were chosen to minimize retention time misalignment. The sample names included in this study can be found in Table 3.2.

Table 3.2. 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

All data preprocessing and analysis was done in Matlab 2020a. Sample data files were imported into Matlab using an in-house algorithm. The unfolded 2D data were baseline corrected, summed across the second dimension, centered with the mean of the baseline around 0, and normalized to the average total 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 amount of background noise stemming from the presence of CO_2.

3.3.3 *Feature Selection*

The unsupervised feature selection method applied herein was accomplished by calculating the pixel-based RSD^2 in peak height across all samples for each *m/z* in which a specified number of samples passed a signal threshold. The signal threshold was located at the maximum of signals corresponding to 10-fold the LOD across all *m/z*, determined by finding the greatest σ_N across all *m/z* in a user-defined region containing baseline noise. Inspection of the distribution of calculated RSD^2 values allows for the selection of an appropriate RSD^2 threshold above which corresponding features are more likely to be class-distinguishing or otherwise containing additional chemical information.

A pixel-based approach is appropriate when there is no retention time misalignment or minor retention time shifting that is easily correctible by retention time alignment. When alignment is undesired or cumbersome, a binning approach or peak table-based calculation is more suitable. Since the simulations in this study had no retention time shifting, a pixel-based approach was taken. For the yeast metabolome dataset, minor retention time shifting on the first dimension was observed (± 1 modulation); therefore, a peak table approach was deemed more acceptable.

In the simulations, an RSD^2 value was calculated at each data point per *m/z* in which at least one sample passed the signal threshold. A more conservative approach was taken for the yeast metabolome data, in which at least five samples were required to pass the signal threshold (i.e., one less than half of the samples) to comprise a peak. First, a peak finder was employed on the 1D data for each *m/z* and sample to locate the time associated with peak maxima. For each *m/z*, the peaks detected were aligned across all samples using an in-house automated peak table alignment algorithm followed by manual inspection. Peaks present in at least five samples were then aligned across *m/z* using the same methodology. In both alignment steps, deviation of ± 1 modulation with

a maximum range in retention time of 2 modulations was allowed. A minimum of three *m/z* per peak was required.

F-ratio analysis was applied identically to the simulated data in a pixel-based fashion and to the yeast metabolome dataset in a peak table approach. The same signal thresholds were applied prior to F-ratio analysis for the chromatographic simulations and the yeast metabolome dataset. The F-ratio calculation is defined as

$$\mathrm{F-ratio} = \frac{\sigma^2_{\mathrm{between}}}{\sum \sigma^2_{\mathrm{within}}} \tag{3.5}$$

where $\sigma^2_{\mathrm{between}}$ is the between-class variance and $\sigma^2_{\mathrm{within}}$ is the within-class variance.

3.4 Results and Discussion

3.4.1 *Unsupervised Feature Selection of Chromatographic Simulations*

Chromatographic replicates of two sample classes were simulated to demonstrate the proposed unsupervised variance thresholding method in a controlled setting where it was known which analytes were class-distinguishing. Reasonable natural variation for biological samples (30% *RSD*) was generated to occlude the chemically-significant variation, akin to real samples. One hundred unique simulations were generated to describe the random variation due to the utilization of an exponential peak height/area distribution, random assignment of retention times and analyte mass spectra throughout the separation space, and random selection of the four analytes to be class-distinguishing. A representative simulation was chosen to demonstrate the work flow, shown in Figure 3.1.

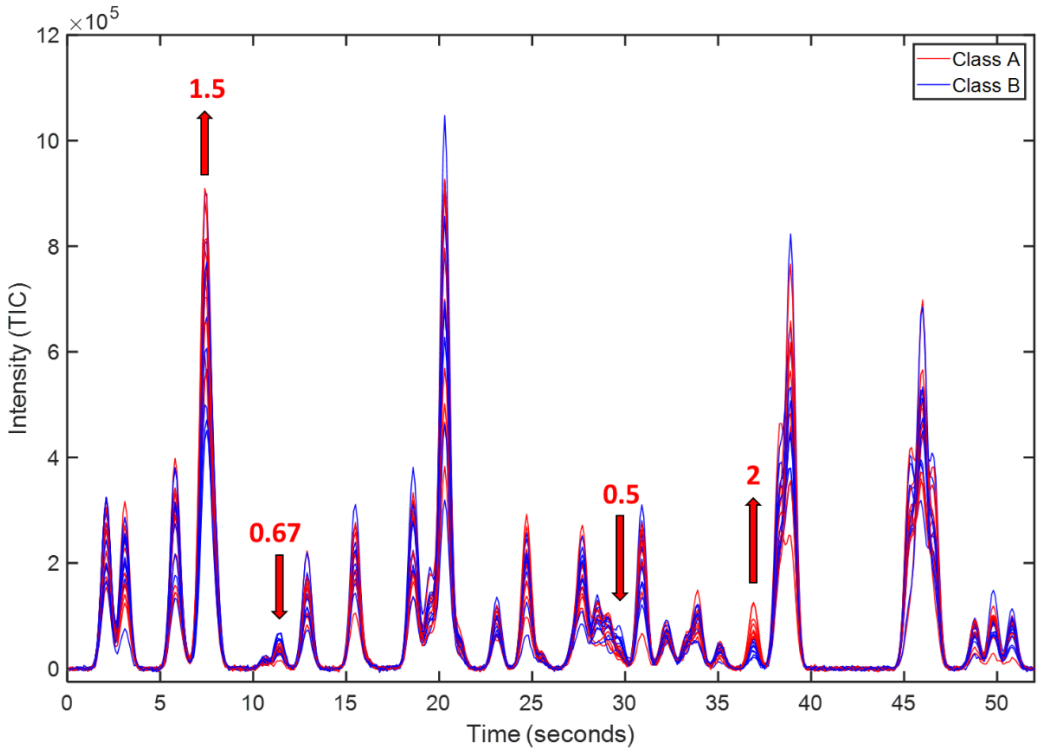


Figure 3.1. Overlay of TIC chromatograms of Class A (red) and Class B (blue) replicates from simulation 15. The four class-distinguishing analytes are indicated by the red arrows: upregulated by 1.5x at 7.3 s, downregulated by 0.67x at 11.4 s, downregulated by 0.5x at 29.7 s, and upregulated by 2x at 36.9 s.

PCA performed on the raw data (a row vector for each sample class replicate) for this simulation resulted in poor class separation, as visible in the scores plot and the associated DCS value of 0.18 (Figure 3.2). Clearly, the chemically insignificant within-class variation simulated to represent natural variation masks the variation corresponding to the chemical information contained in the four class-distinguishing analytes. Looking at the scores plot, an analyst would be unable to successfully group these samples into two classes. The loadings plot would be uninformative as well, with chemically insignificant peaks being highly loaded. A supervised feature selection technique, such as F-ratio analysis, would be well-equipped to deal with the high within-class variation that is mishandled by PCA; however, the application of supervised methods requires a priori knowledge of the sample class membership. These simulations were generated to

enact the scenario in which class membership was unknown prior to analysis, which precludes the use of supervised methods.

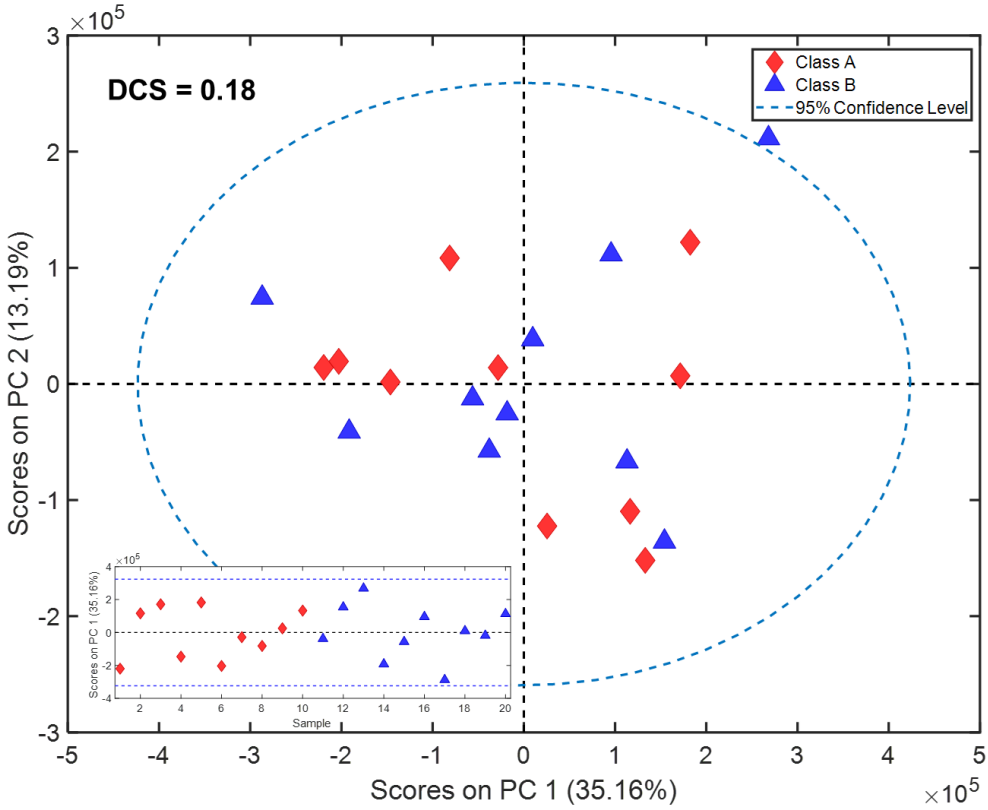


Figure 3.2. PCA scores plot of all data from simulation 15 (out of 100 simulations).

The unsupervised feature selection method was applied to each simulation, calculating the RSD^2 at each time point per *m/z* that had at least one sample passing the applied signal threshold. All other time points were assigned a value of 0 to distinguish between chromatographic peaks (which would pass the signal threshold unless an analyte was at a low concentration approaching or below the LOD) and noise. A plot of the RSD^2 metric calculated for each *m/z* throughout the chromatographic space is shown in Figure 3.3A. It appears that there is a RSD^2 "peak" corresponding to each peak in the chromatogram in Figure 3.1. Additionally, most of the "peaks" hover around an RSD^2 value of 0.9, which corresponds to *RSD* and %*RSD* quantities of 0.3 and

30%, respectively. There are several "peaks" with *m/z* that rise above the apparent baseline RSD^2 for this data, corresponding to the four analytes that were simulated to be class-distinguishing. The maxima of these "peaks" correspond to RSD^2 values of 0.13 and 0.23, equating to *RSD* values of 0.36 and 0.48, respectively. These numbers correspond to the simulated changes in concentration: 33% *RSD* for factors of 0.67 and 1.5, and 50% *RSD* for factors of 0.5 and 2.

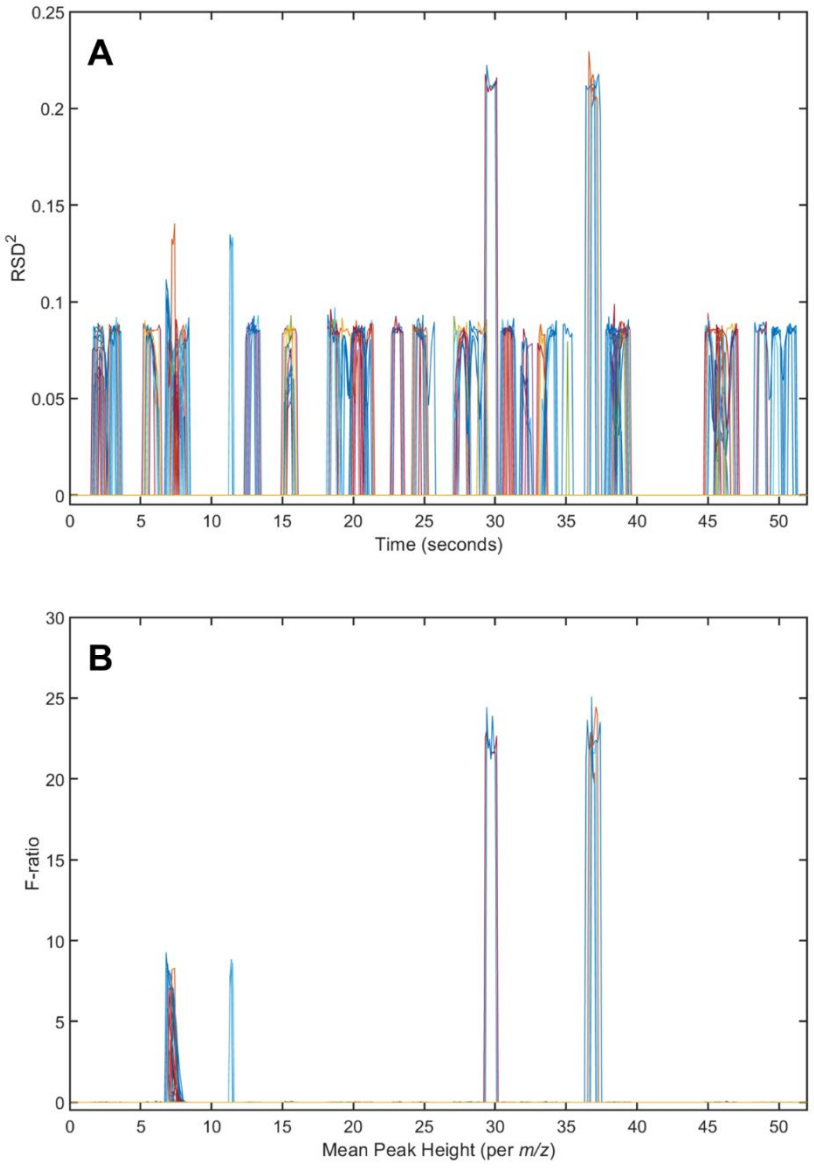


Figure 3.3. Unsupervised and supervised feature selection of simulation 15. (A) RSD^2 and (B) F-ratio calculated over the chromatographic space, with the requirement that at least one sample overcome the signal threshold.

To assess how this unsupervised feature selection method compares with a supervised method, pixel-based F-ratio analysis was applied to the data using the same signal threshold requirement. The F-ratio at each *m/z* over time is shown in Figure 3.3B. Notably, the same peaks discovered by the unsupervised method are present. Using an F_{crit} value of 4.41 ($\alpha = 0.05$, $df_{num} = 1$, $df_{denom} = 18$), these features are statistically significantly different between classes in terms of their peak heights, and thus, concentrations.

The variance thresholding results can be viewed in a different format in Figure 3.4A. Clearly, a large number of *m/z* fall around an RSD^2 of 0.9 with a large range in average peak height. This confirms that although peak height was varied, the relative standard deviation was simulated to be constant ($RSD_{R,B} = 30\%$) , implying that standard deviation in peak height between samples must increase linearly with increasing peak height. We will see in the next section whether this is the case with real data. The data above an RSD^2 of 0.9 is correlated with the class-distinguishing analytes. Two distinct clusters can be seen around an RSD^2 of 0.13 and 0.21, since two analytes correspond to each cluster (were changed by the same relative factor). Again we see that the RSD^2 remains relatively constant for each cluster as the average peak height increases. Some broadening can be observed at the base of each cluster, corresponding to lower signals. There are also random data points that do not seem to be associated with any cluster; these arise due to overlap of shared *m/z* between the class-distinguishing analytes and those at a comparable signal magnitude that have nominally the same concentration between classes. To demonstrate the relevance of the discovered features to differentiating the classes, PCA was repeated only using those *m/z* and timepoints that overcame an RSD^2 threshold of 0.1, which should eliminate the irrelevant variation in the data and retain some portion of $V_{R,Info}$. The scores plot resulting from PCA classification of

the reduced features is shown in Figure 3.4B. The classification of the samples is significantly improved compared to the random sample distribution observed in Figure 3.2, with a DCS of 3.2.

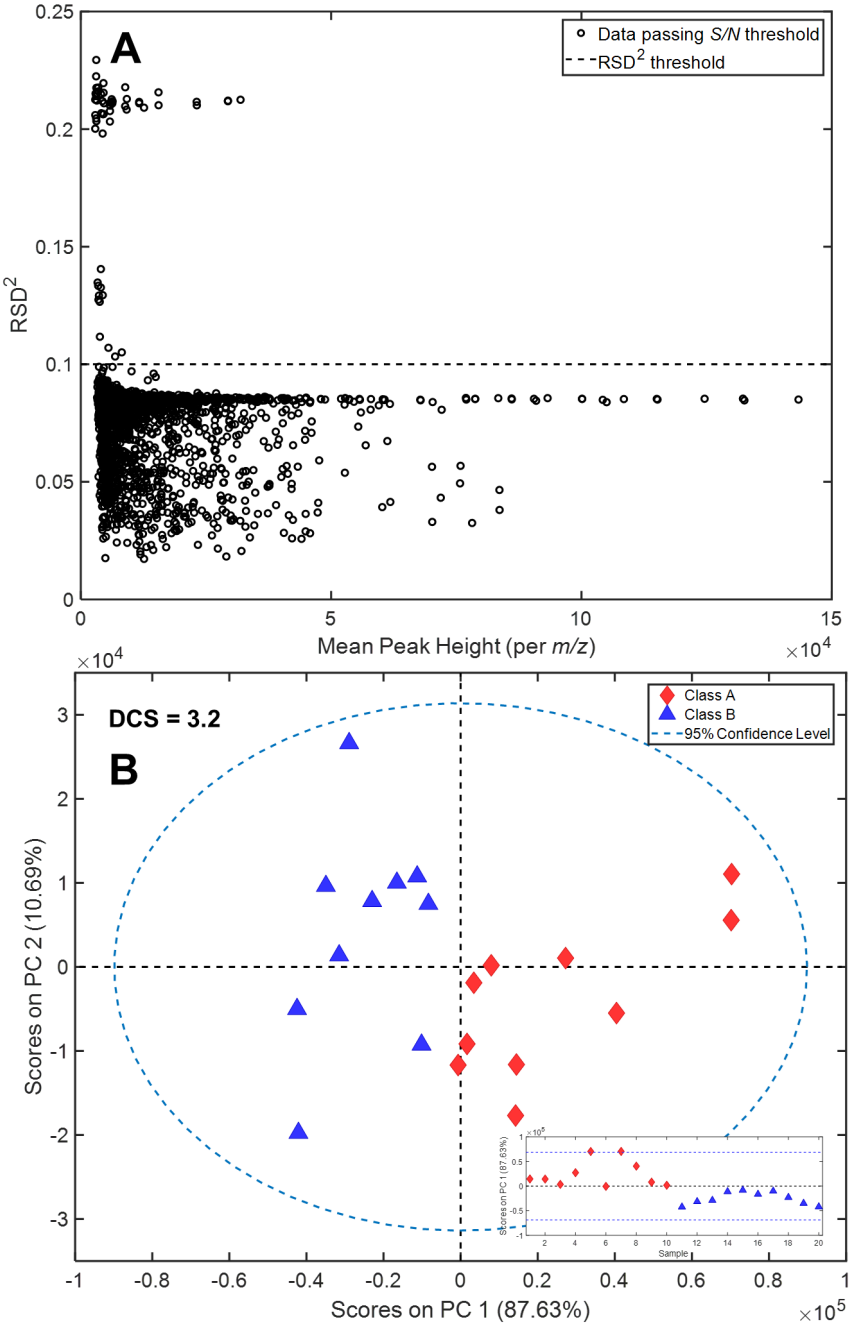


Figure 3.4. Unsupervised feature selection of simulation 15 by variance thresholding. (A) Application of an RSD^2 threshold (dotted line) at 0.1 to preserve the *m/z* and time points with the greatest relative variance. (B) PCA scores plot following unsupervised feature selection, using only the features in (A).

Degree of class separation was used to quantify the sample classification by PCA using all the data and only the discovered features. The DCS when all of the data was included is 0.66 ± 0.76 and it increased to 2.66 ± 0.60 when only the discovered features with RSD^2 above the threshold were included. In general, the DCS was very low when all the data was forwarded to PCA, indicating that the separation of the two classes was not successful due to few variables containing significant information ($V_{R,Info}$) being drowned out by the insignificant variation of the majority of the data ($V_{R,B}$). In addition, the poor distinction between classes can be a result of some analytes being simulated with concentrations below the LOD, causing them to be undetectable. This is particularly detrimental to the separation of classes when an analyte simulated below the LOD is one of the four class-distinguishing analytes. Overlap of the class-distinguishing analytes can also pose an issue to class separation before and after feature selection. If *m/z* are shared between these analytes and irrelevant analytes at considerably higher concentrations, the significant variation may be swamped out by the insignificant variation present at a similar or greater magnitude. It is also possible that two class-distinguishing analytes co-elute and the variation in shared *m/z* may be canceled out or reinforced depending on their relative concentration factors. It is important to note that for data with many class-distinguishing analytes and low within-class variation, PCA will be considerably more successful at differentiating the two classes when all data is used. Clearly, there is significant variation in the chromatographic simulations as encompassed by the 115% *RSD* in these DCS values. Using an exponential distribution in peak heights introduces a significant amount of variation between chromatograms, particularly when combined with random analyte distribution and mass spectral assignment. These factors cause immense variability in mass spectral similarity and resolution between analytes, which contribute to unique patterns of shared *m/z* along with the random exponentially distributed signals.

To quantify the correlation between variance thresholding and F-ratio feature selection, the number of features discovered by each method were compiled and related by the correlation coefficient. With a correlation coefficient of 0.71, the number of features discovered by supervised and unsupervised feature selection strategies have a fairly strong, positive correlation. When the retention times of the discovered features were compared with the simulated retention times of the class-distinguishing features, supervised F-ratio yielded an average of 3.5 features discovered out of 4 (88%) while unsupervised variance thresholding gave an average of 3.3 features discovered out of 4 (83%).

3.4.2 *Unsupervised Feature Selection of Benchmark Yeast Metabolome Dataset*

A previously studied metabolomics dataset was used to demonstrate the application of the variance thresholding method to real data. The dataset, consisting of two classes of metabolizing yeast, has been subject to supervised feature selection [26,27] and therefore many class-distinguishing metabolites have been discovered and identified. Furthermore, many of these metabolites have been statistically verified as being significantly different between classes (*t*-test), and a handful of discovered metabolites were found to be false positives; i.e., there was not a statistically significant difference between classes for these analytes.

For this study, the unsupervised feature selection method reported herein has only been applied to one-dimensional (1D) gas chromatography. However, application to comprehensive two-dimensional (2D) gas chromatography is a natural extension and could be applied in a similar vein as the tile-based F-ratio methodology to mitigate retention time misalignment on either chromatographic dimension [7,27,30–32]. For this reason, the GC × GC–TOFMS data was summed along the second dimension to artificially reduce the data to 1D GC. In doing so, the data collection rate was reduced to 0.67 Hz (1 data point = 1 modulation = 1.5 s). While severely

reduced, this sampling rate is closer to what is achieved by a quadrupole mass spectrometer compared to a TOFMS. In the process of reducing the dimensionality of the data, it can be expected that some analytes that were previously resolved in the second chromatographic dimension are now overlapped by varying degrees. Therefore it is likely that some of the peaks observed in the 1D chromatographic data are composed of two or more analytes. Compared to the discovery-based analysis of the 2D data, we would expect to see different results and perhaps fail to discover those metabolites that are obscured by co-eluting analytes present in higher concentrations. This is acceptable and within the scope of this study, with the main goal being to demonstrate that unsupervised feature selection applied to a real 1D dataset yields similar information to a supervised application to the same data. For that reason, we will not be comparing the results obtained herein with those of previous analyses, but instead with supervised discovery-based analysis (via F-ratio analysis) of the artificially reduced 1D dataset; however, we will use previous results to corroborate our findings.

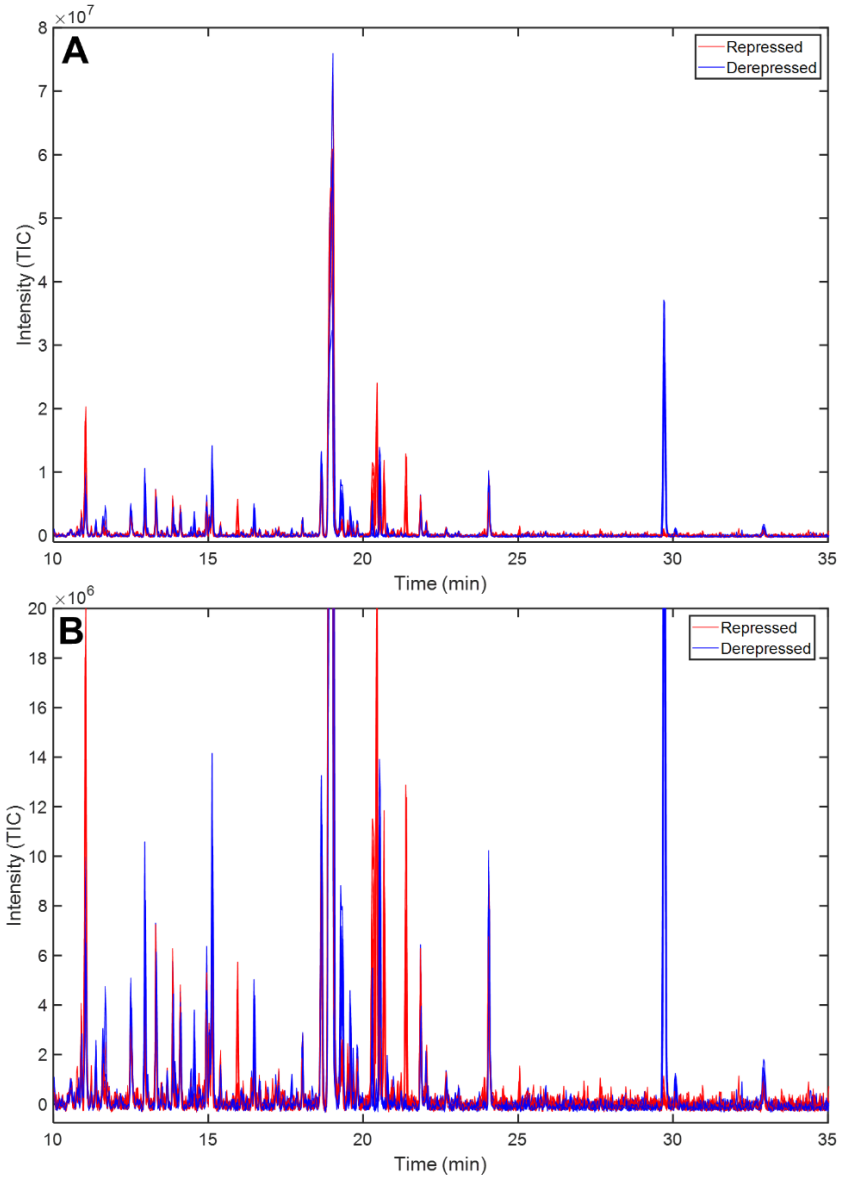


Figure 3.5. Overlay of 1D TIC chromatograms of Repressed (red) and Derepressed (DR) yeast metabolome samples. (A) Chromatographic separations from 10-35 min. (B) A zoom-in from the baseline to an intensity of 20×10^6 is shown to illustrate the many metabolites present at low intensities.

The TIC chromatograms of all 12 samples used in this demonstration are shown overlaid in Figure 3.5. Several class-distinguishing metabolites are visible to the eye. Some of these analytes were chosen to establish the baseline variation due to natural biological variation and justify the choice of an *RSD* threshold for feature selection. Based upon previous work, it was

determined that the average biological variation was approximately 30% [25]. Six metabolites were selected based on the following criteria: located easily in the 1D chromatograms, comprised top hits in the supervised discovery-based analysis [26,27], and varied in their DR/R ratio (i.e., the concentration ratio between classes, also indicative of in which class the analyte is present at a higher concentration). Inspection of the data revealed minor retention time shifting (±1 modulation) between samples. The analytes were found by locating the peak maxima for each sample within a three-modulation window centered on the previously reported retention times [26,27]. For each *m/z* in which at least five of the samples were above the signal threshold, the mean and standard deviation of the peak height at the peak maxima for each metabolite within each class was calculated. Figure 3.6A shows the scatter plot of these measurements. Regardless of the variation in standard deviation for different analytes (see Figure B.1 and Figure B.2 in Appendix B for plots with each class and analyte displayed separately) and/or *m/z*, a general linear trend can be observed. To confirm this observation, the data were transformed logarithmically and graphed in a log-log plot (Figure 3.6B). Once transformed, the data appear linear, except for a region in which the standard deviation levels off at lower average peak heights. Although these *m/z* likely did not pass the signal threshold for one of the two classes (as seen in the low mean peak height), they were included in this plot because the signal threshold was overcome in the other class, indicating that this *m/z* is important in differentiating between the classes. Additionally, the behavior of low signal *m/z* is important to demonstrate in the context of our theory to justify our reasoning for and the importance of instituting a signal threshold in such analyses. This region can be thought of as being "detection variation dominated", which is consistent with a significant contribution of RSD_{Det} relative to the other sources of variation at low signals approaching the LOD. In the context of the applied signal threshold, this region begins at peak heights less than the

threshold, which supports the selection of this threshold. When a line of best fit is fitted to logarithmically transformed data, the slope of this line is equal to the power to which the independent variable is raised to define the relationship between the dependent and independent variables of the untransformed data. A line with a slope of approximately 1 was fitted to the data above the signal threshold, indicating that the standard deviation increases linearly with the average signal. Therefore, the *relative* standard deviation (i.e., the slope of a line defining the standard deviation versus the mean peak height) should be somewhat constant for peak heights above the signal threshold. This is confirmed in Figure 3.7A, where a flat band of points above the signal threshold and a drastic increase in *RSD* below the threshold is observed. The average *RSD* above the signal threshold is 0.22, but the *RSD* ranges from ~0.1 to ~0.5. An *RSD* threshold of 0.4 was deemed acceptable to differentiate class-distinguishing features with the additional RSD_{info} term from analytes with background variation dominated by natural variation, as ~95% of the *m/z* above the signal threshold fell below an *RSD* of 0.4 (Figure 3.7B).

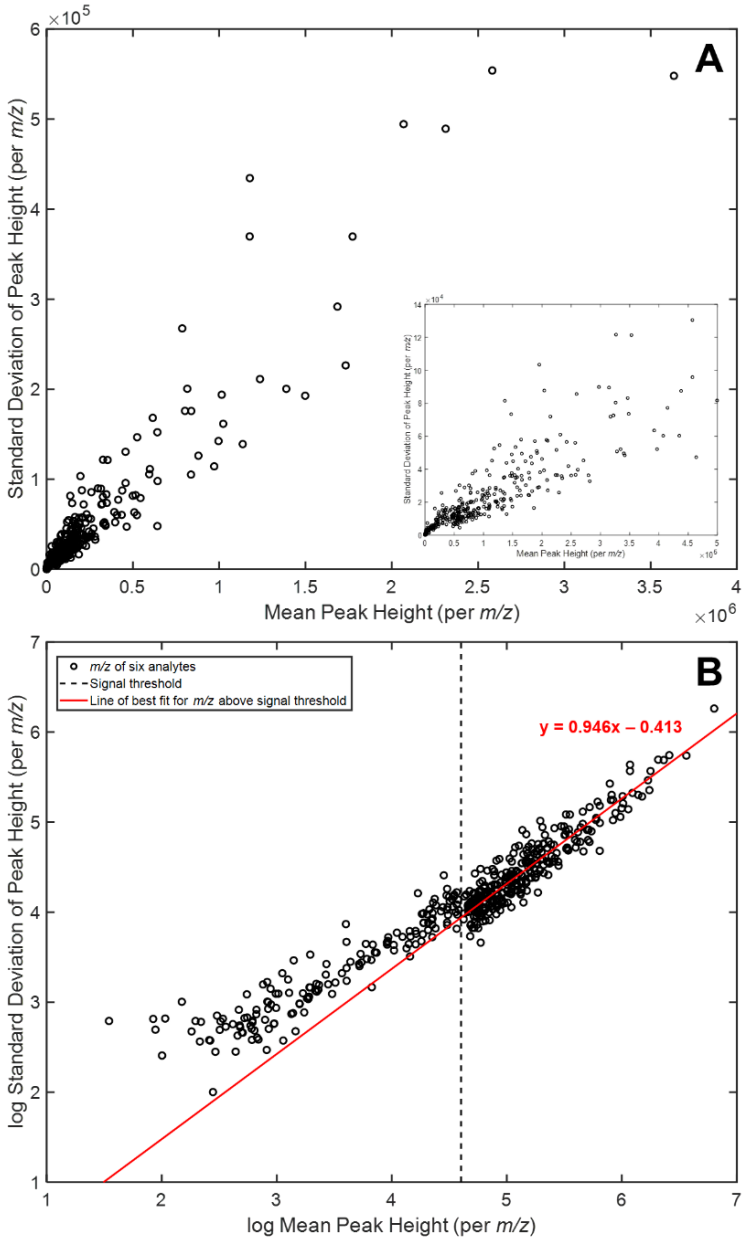


Figure 3.6. Investigation of within-class variation and determination of relationship between standard deviation and peak height (signal). The standard deviation and mean peak height per *m/z* were calculated for six analytes: trehalose, 1782.75 s; glucose, 1240.5 s, glycerol c00116, 663 s; threonine c00188, 777 s; malate c00149, 873 s; and 5-oxoproline c00025b, 907.5 s. (A) Scatter plot of the standard deviation versus mean of the peak height of *m/z* (with at least 5 samples that passed the signal threshold) for the 6 analytes. One data point with a mean peak height of ~6.3×10^7 and standard deviation of ~1.8×10^6 was left out. A zoom-in of the region between 0 and 5×10^5 in mean peak height is provided inset. (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) and defines a region to the left containing *m/z* with variation that is dominated by the RSD_{Det} term.

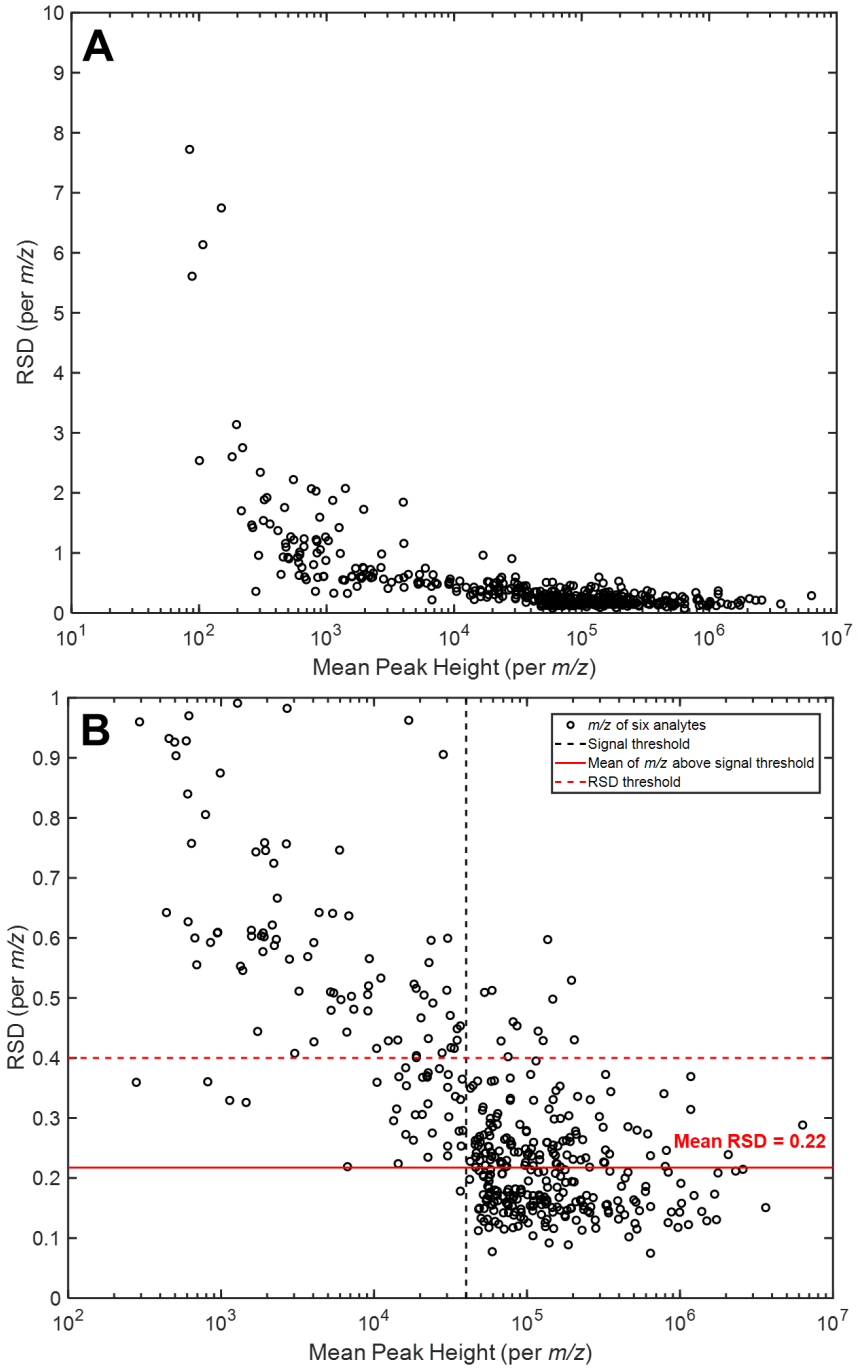


Figure 3.7. Estimation of RSD_B for the selection of an RSD^2 threshold. (A) Relative standard deviation versus mean of the peak height for the *m/z* of the six analytes described in Figure 3.6. One data point with a mean peak height of ~35 and *RSD* of ~18 was left out. (B) Zoom-in of (A) for *m/z* with *RSD* between 0 and 1. The signal threshold applied to the data on a per *m/z* basis (~4×10^4) is shown (black dotted line). The average *RSD* of the data above the signal threshold is 0.22 (red solid line). An *RSD* threshold of 0.4 (red dotted line) was chosen, as ~95% of the data above the signal threshold fell below this value.

Due to the retention time shifting in the data, the variance thresholding method was applied to peak tables rather than data pixels. A peak finder was applied to each sample and *m/z* to locate all peak maxima above the signal threshold. The retention time of each peak found at each *m/z* were then aligned between all samples, so that peaks within 1 modulation of each other were grouped together. Only the peaks that were present in at least 5 samples were retained and then aligned between all *m/z*. Finally, only the peaks that were defined by at least 3 *m/z* were retained for further analysis. This process resulted in 53 peaks, with 44 more present that had less than 3 *m/z*. For each of the 53 peaks, the mean and standard deviation in the peak height at the peak maxima across all samples was calculated for each *m/z* included. For samples in which the peak was not found, the signal was taken at the modulation closest to the median retention time of the samples in which the peak was present. A series of *RSD* and RSD^2 values were calculated for each peak with one value for each *m/z* included. F-ratio analysis was conducted on the aligned signal values for each *m/z* per peak. The average *RSD*, RSD^2, and F-ratio for all 53 peaks are reported in Table 3.3. Probable identification of the metabolites corresponding to the peaks was accomplished by reviewing the metabolite identities at similar retention times reported in previous work [25–27], and was supported by comparisons of *m/z* included for each analyte with reference mass spectra. Many of the peaks were unable to be identified with any level of certainty; this was also the case for many features in the previous work [25–27].

Table 3.3. Peak table of all peaks found sorted according to retention time (containing at least 3 *m/z* with at least 5 samples present above the signal threshold). The average *RSD*, RSD^2, and F-ratio reported were calculated by taking the mean of all *m/z* included for that peak. 44 additional peaks containing at least 1 *m/z* with at least 5 samples present were discovered and omitted from this table. Metabolites previously determined to be false positives are indicated by an asterisk (*).

Analyte	Retention time (minutes)	Retention time (seconds)	Average *RSD*	Average RSD^2	Average F-ratio	Number of *m/z* included
Unk1	10.55	633	0.207	0.0488	0.802	7
Unk2	10.80	648	0.516	0.268	1.33	4
leucine	10.90	654	0.265	0.0708	3.65	12
glycerol c00116	11.05	663	0.502	0.303	81.6	55
Unk3	11.23	673.5	0.208	0.0435	1.86	5
isoleucine*	11.38	682.5	0.330	0.111	7.21	11
Unk4	11.60	696	0.601	0.404	1.07	8
Unk5	11.66	699.75	0.879	0.836	9.07	4
serine*/o-toluic acid*	12.50	750	0.214	0.0468	0.556	22
threonine c00188	12.95	777	0.496	0.247	48.0	32
Unk6	13.03	781.5	0.334	0.120	17.8	4
Unk7	13.31	798.75	0.264	0.0722	1.78	23
Unk8	13.50	810	0.228	0.0520	1.26	4
Unk9	13.68	820.5	0.234	0.0548	1.11	6
Unk10	13.85	831	0.596	0.355	0.106	15
homoserine	13.93	835.5	0.780	0.612	107	6
Unk11	14.10	846	0.207	0.0448	1.97	23
malate c00149	14.55	873	0.899	0.809	72.4	10
methionine*	14.95	897	0.280	0.0792	0.251	18
Unk12	15.04	902.25	0.294	0.100	5.86	14
5-oxoproline c00025b	15.13	907.5	0.486	0.237	25.1	24
Unk13	15.40	924	0.815	0.687	0.617	5
Unk14	15.94	956.25	1.35	1.83	14.9	4
Unk15	16.40	984	0.366	0.136	2.55	4
glutamic acid	16.47	988.5	0.590	0.349	62.3	16
phenylalanine*	16.65	999	0.252	0.0642	0.202	5
Unk16	17.18	1030.5	0.330	0.112	1.74	3
asparginine c00152	17.28	1036.5	0.361	0.135	0.741	3
Unk17	18.04	1082.25	0.382	0.153	0.567	13
glutamine c00064/glucose-1-phosphate c00029b	18.65	1119	0.224	0.0542	12.2	39
Unk18	18.93	1135.5	0.233	0.0558	0.322	28
Unk19	18.94	1136.25	0.186	0.0351	0.333	5
Unk20	19.00	1140	0.226	0.0575	0.225	138
ornithine*	19.28	1156.5	0.700	0.495	59.5	17

citrate c00158	19.34	1160.25	0.732	0.562	125	20
Unk21	19.51	1170.75	0.610	0.406	22.4	6
Unk22	19.59	1175.25	0.568	0.324	19.2	15
Unk23	19.81	1188.75	0.241	0.0640	0.532	8
glucopyranose	20.28	1216.5	0.634	0.404	12.5	27
Unk24	20.34	1220.625	1.276	1.63	20.4	6
glucose	20.45	1227	1.083	1.17	135.9	53
lysine	20.54	1232.25	0.647	0.439	37.7	41
glucose	20.68	1240.5	1.084	1.18	78.0	24
tyrosine	20.79	1247.25	0.728	0.537	51.7	4
glucopyranose	21.38	1282.5	1.175	1.38	32.4	24
Unk25	21.85	1311	0.175	0.0312	0.0484	24
Unk26	22.05	1323	0.616	0.384	0.616	6
Unk27	22.68	1360.5	0.301	0.0918	1.46	4
stearic acid*	24.05	1443	0.172	0.0299	3.61	33
Unk28	25.05	1503	1.211	1.47	30.7	3
trehalose	29.71	1782.75	1.055	1.11	214	84
5’-S-methyl-5’-thioadenosine	30.08	1804.5	0.893	0.799	45.4	3
Unk29	32.91	1974.75	0.323	0.106	7.10	5

When the RSD^2 threshold of 0.16 is applied (corresponding to an *RSD* threshold of 0.4), 27 of the peaks are designated as hits. Using an F_{crit} value of 4.96 for this class comparison (α = 0.05, df_{num} = 1, df_{denom} = 10), 27 peaks are deemed statistically significantly different between classes. For both of these statistics, all but 5 hits are discovered by the other method. Notably, metabolites determined to be false positives (presenting with an F-ratio above F_{crit} but not differing statistically significantly between classes according to a *t*-test) in previous analyses [26,27] were not discovered to be hits in the present study. These metabolites (isoleucine, serine, o-toluic acid, methionine, phenylalanine, and stearic acid) had RSD^2 values less than the RSD^2 threshold. The only exception was ornithine, which was a borderline false positive and purported to be a true positive [27]. Likewise, many of the previously statistically verified true positives were discovered by the current method. While many other true positive hits went undiscovered in the current demonstration due to aforementioned reasons, we believe that the present study has shown that

similar information revealed by supervised feature selection can be found by unsupervised feature selection.

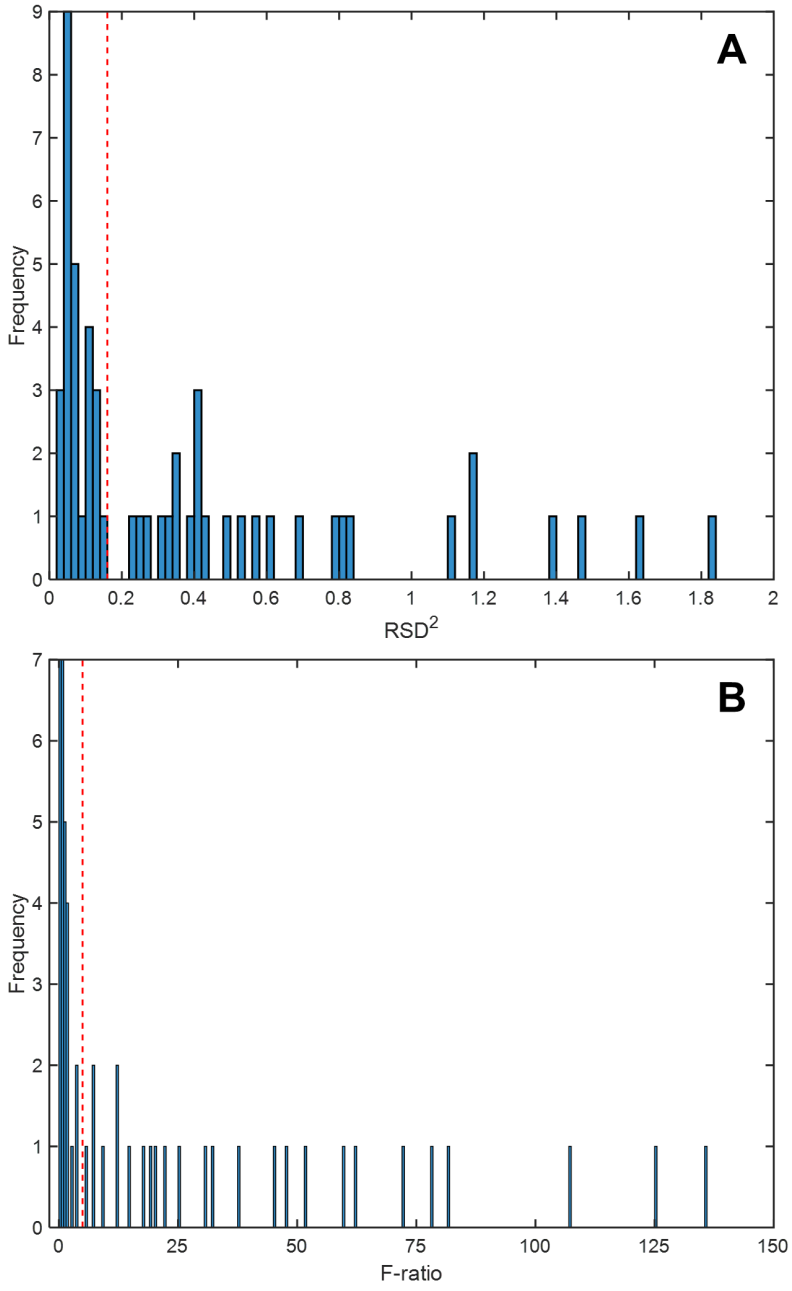


Figure 3.8. Histograms for unsupervised and supervised feature selection of yeast metabolome peak data. Distribution of (A) RSD^2 and (B) F-ratio values for the 53 peaks.

Of the 53 peaks, 27 had average RSD^2 above the RSD^2 threshold of 0.16 (corresponding to the *RSD* threshold of 0.4). The distribution of the average RSD^2 values is shown in Figure 3.8A. Similarly, the distribution of average F-ratio values is shown in Figure 3.8B. To examine the relationship between unsupervised and supervised feature selection methods, the average RSD^2 and F-ratio (i.e., variances) for each peak was plotted in Figure 3.9, where a general positive correlation can be seen. Out of the 27 features deemed hits, 22 were selected by both unsupervised variance thresholding and supervised F-ratio analysis. A perfect correlation is not expected due to F-ratio normalization to the within-class variance, which can result in smaller F-ratio values in the event that a feature has somewhat higher within-class variation, whereas the total RSD^2 would encompass this variation and may observe an increase in the calculated value and its relative significance (i.e., hit number) compared with an analyte with more typical within-class variation. A related, potential benefit of variance thresholding is that since it does not distinguish between within-class variation and between-class variation, an analyte exhibiting comparatively high within-class variation will be discovered by this method but would not be discovered by F-ratio analysis. The presence of high within-class variation may be worthy of investigation as it may be indicative of subclasses.

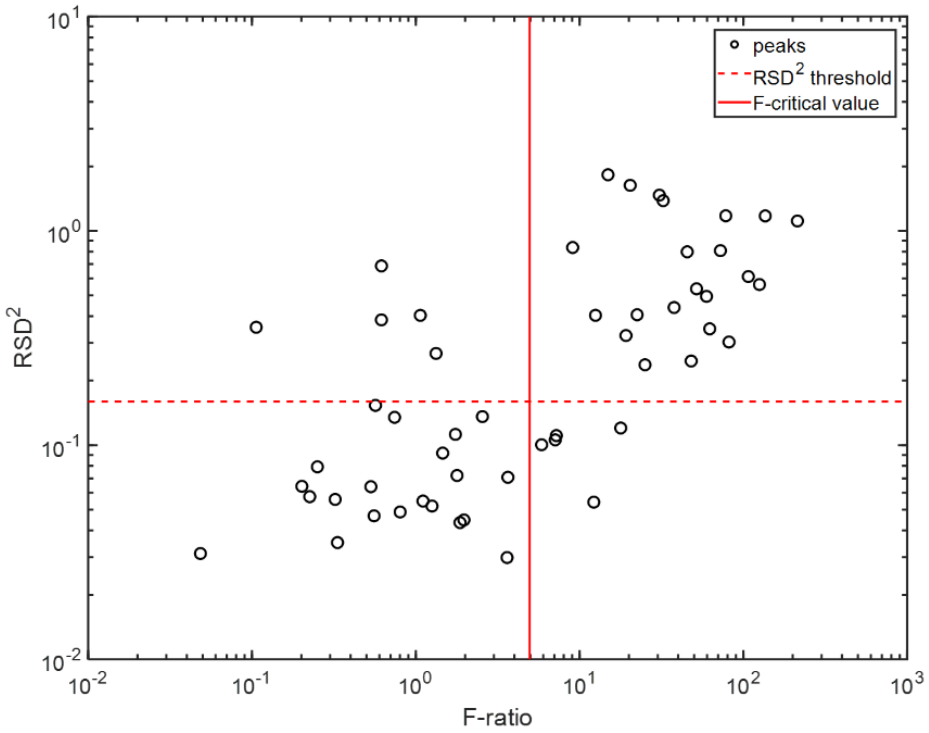


Figure 3.9. Scatter plot of the average RSD^2 versus average F-ratio for the 53 peaks.

Several analytical ion current (AIC) chromatograms obtained by summing the signal for the *m/z* included for the peak are shown in Figure 3.10. The metabolites shown are hits that were deemed to be class distinguishing by their RSD^2 values. A variety of concentration ratios between classes can be seen; some metabolites are only present in the R class, others are only present in the DR class, and still more are present in both classes but in different amounts. The concentration range that is spanned by these metabolites is also notable; some metabolites shown are the most prevalent peaks in the separation, while others are orders of magnitude smaller and approach the signal threshold applied in this study. Still more metabolites may have been discovered if the requirements were loosened (i.e., decreasing the signal threshold, requiring fewer samples, <5, to have the peak present, eliminating the minimum number of *m/z*, 3, required); however, the purpose of this study was to demonstrate the comparative ability of the unsupervised feature selection

method to supervised methods, not to perform an exhaustive discovery-based investigation of the dataset as previously accomplished [26,27].

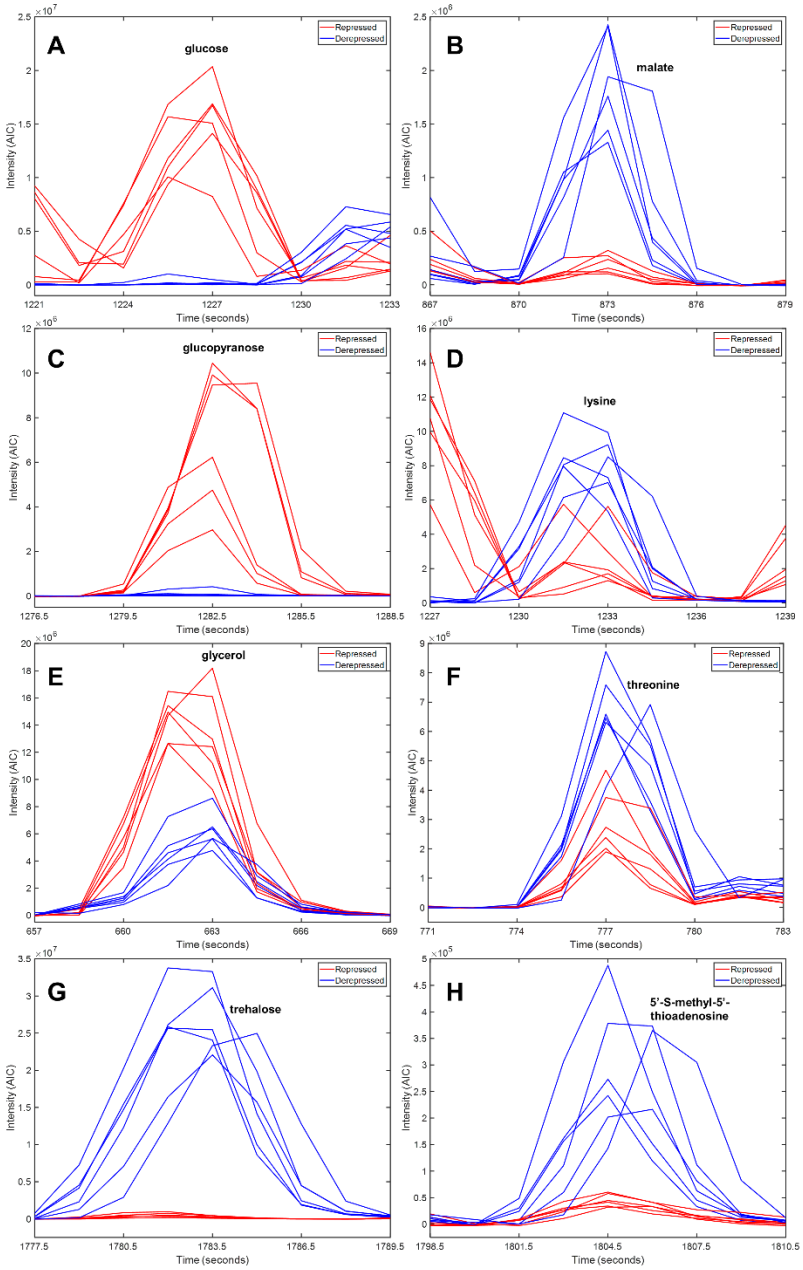


Figure 3.10. Analytical ion current (AIC) chromatograms of 8 identified metabolites with RSD^2 above the threshold. The *m/z* that had at least 5 samples present were summed for each peak. (A) Glucose, 53 *m/z* included; (B) malate c00149, 10 *m/z* included; (C) glucopyranose, 24 *m/z* included; (D) lysine, 41 *m/z* included; (E) glycerol, 55 *m/z* included; (F) threonine, 32 *m/z* included; (G) trehalose, 84 *m/z* included; and (H) 5'-S-methyl-5'-thioadenosine.

3.5 CONCLUSION

Variance thresholding provides an analogous unsupervised alternative to supervised feature selection techniques such as F-ratio analysis for GC-MS data. The present study demonstrates the feasibility of variance thresholding on simulated and previously analyzed GC-MS datasets with significant (~30% *RSD*) within-class variation. The selection of signal and variance thresholds removed low signal data points with deceptively high variance and data points exhibiting background variance as a result of natural and other experimental sources of variation, respectively. The simulations were analyzed by a pixel-based approach, which provided features that improved sample classification by PCA with an average ~ 50-fold increase in DCS with respect to PCA applied pre-feature selection. Due to minor retention time shifting, a peak table approach was utilized on the yeast metabolome dataset, wherein 27 features were discovered and an overall positive correlation was observed between the unsupervised variance thresholding method and supervised F-ratio analysis. Foreseen challenges of this method lie in the selection of an appropriate variance threshold and mitigation of retention time misalignment. Whereas in this proof-of-principle study class membership was known so the variance threshold was readily calibrated, in practice the analyst would need to approach this consideration by other means such as starting at a high variance threshold and working down to lower thresholds in a systematic exploration of the hitlist, and/or calibrate with similar samples to mimic the major source of anticipated background variation. Extension of the variance thresholding to 2D GC data, such as in a tile-based method, is expected to be straightforward.

3.6 REFERENCES

[1] B. Lehallier, J. Ratel, M. Hanafi, E. Engel, Systematic ratio normalization of gas chromatography signals for biological sample discrimination and biomarker discovery, Anal. Chim. Acta. 733 (2012) 16–22. https://doi.org/10.1016/j.aca.2012.04.019.

[2] 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.

[3] S. Abbasi, S. Gharaghani, A. Benvidi, A. Latif, Identifying the novel natural antioxidants by coupling different feature selection methods with nonlinear regressions and gas chromatography-mass spectroscopy, Microchem. J. 139 (2018) 372–379. https://doi.org/10.1016/j.microc.2018.03.012.

[4] 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.

[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] 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.

[8] 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.

[9] Z. Wang, P. de B. Harrington, Feature selection of gas chromatography/mass spectrometry chemical profiles of basil plants using a bootstrapped fuzzy rule-building expert system, Anal. Bioanal. Chem. 405 (2013) 9219–9234. https://doi.org/10.1007/s00216-013-7327-x.

[10] S.E. Reichenbach, X. Tian, C. Cordero, Q. Tao, Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography, J. Chromatogr. A. 1226 (2012) 140–148. https://doi.org/10.1016/j.chroma.2011.07.046.

[11] N.M. Faber, R. Rajkó, How to avoid over-fitting in multivariate calibration - The conventional validation approach and an alternative, Anal. Chim. Acta. 595 (2007) 98–106.

[12] 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.

[13] 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.
[14] W. Zhao, C.E. Davis, Autoregressive model based feature extraction method for time shifted chromatography data, Chemom. Intell. Lab. Syst. 96 (2009) 252–257. https://doi.org/10.1016/j.chemolab.2009.02.010.
[15] R. Setiono, H. Liu, Feature extraction via Neural networks, in: H. Liu, H. Motoda (Eds.), Feature Extraction, Construction and Selection: A Data Mining Perspective, Springer US, Boston, MA, 1998: pp. 191–204. https://doi.org/10.1007/978-1-4615-5725-8_12.
[16] P. Ciosek, Z. Brzózka, W. Wróblewski, E. Martinelli, C. Di Natale, A. D'Amico, Direct and two-stage data analysis procedures based on PCA, PLS-DA and ANN for ISE-based electronic tongue—Effect of supervised feature extraction, Talanta. 67 (2005) 590–596. https://doi.org/10.1016/j.talanta.2005.03.006.
[17] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing. 300 (2018) 70–79. https://doi.org/10.1016/j.neucom.2017.11.077.
[18] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering. 40 (2014) 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
[19] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[20] L. Haar, K. Anding, K. Trambitckii, G. Notni, Comparison between Supervised and Unsupervised Feature Selection Methods:, in: Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019), SCITEPRESS - Science and Technology Publications, Prague, Czech Republic, 2019: pp. 582–589. https://doi.org/10.5220/0007385305820589.
[21] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, Machine Learning Mastery, 2020.
[22] X. He, D. Cai, P. Niyogi, Laplacian Score for Feature Selection, in: Y. Weiss, B. Schölkopf, J.C. Platt (Eds.), Advances in Neural Information Processing Systems 18, MIT Press, 2006: pp. 507–514. http://papers.nips.cc/paper/2909-laplacian-score-for-feature-selection.pdf.
[23] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10, ACM Press, Washington, DC, USA, 2010: p. 333. https://doi.org/10.1145/1835804.1835848.
[24] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature Selection: A Data Perspective, ACM Comput. Surv. 50 (2018) 1–45. https://doi.org/10.1145/3136625.
[25] 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.
[26] 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.
[27] N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset, J. Chromatogr. A. 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.
[28] 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.
[29] B.C. Reaser, S. Yang, B.D. Fitz, B.A. Parsons, M.E. Lidstrom, R.E. Synovec, Non-targeted determination of 13 C-labeling in the Methylobacterium extorquens AM1 metabolome using the two-dimensional mass cluster method and principal component analysis, J. Chromatogr. A. 1432 (2016) 111–121. https://doi.org/10.1016/j.chroma.2015.12.088.
[30] 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.
[31] 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.
[32] 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.

Chapter 4. Predictive Modeling of Aerospace Fuel Properties Using Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry and Partial Least Squares Analysis[4]

4.1 INTRODUCTION

Chemical composition has a demonstrable effect on the properties of multicomponent hydrocarbon fuels and ultimately the performance and reliability of the systems that use them [1–8]. For this reason, specifications for aerospace kerosene fuels intentionally limit impurities and hydrocarbon classes (e.g., sulfur, olefins, oxygenates, and aromatic compounds) with detrimental system level impacts [9]. Given the diversity in chemical sources and production methods for commodity fuels and the increasingly stringent operational requirements in aerospace energy conversion devices, determining quantitative relationships between fuel chemical composition, specification properties, and performance is vitally important [10–20]. With regard to advancing a foundational connection between fuel composition and physical properties, evaluation of compositionally controlled laboratory blends and "field" fuel samples is instructive [10–14,16–20]. However, as the number of fuels analyzed increases, a reliable and straightforward analytical protocol is needed in order to glean significant information while keeping analysis time reasonably short.

Gas chromatography (GC) is a conventional analytical technique that is ideal for the separation and analysis of volatile and semi-volatile mixtures [21,22]. When GC is coupled with mass spectrometry (MS), spectral information is gathered that allows further selectivity and

[4] This chapter has been reproduced from K.L. Berrier, C.E. Freye, M.C. Billingsley, and 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.

increases confidence in chemical compound identification. GC-MS has been shown to be convenient and effective in the analysis of kerosene-based fuel [10–15]. Comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC×GC-TOFMS) significantly improves upon the separation power of one-dimensional (1D) GC and provides additional insight into complex mixtures of volatile compounds including kerosene-based fuels [17–19,23–26]. In typical GC×GC column configurations, the first separation dimension (^{1}D) uses a non-polar stationary phase while the second separation dimension (^{2}D) uses a polar stationary phase. However, a reverse-column configuration with a polar ^{1}D column and a non-polar ^{2}D column has been shown to provide better selectivity for petroleum-based samples [17–20,25,27], and is implemented herein. As with any technology that produces inordinate volumes of data, a critical issue in the practical application of GC×GC-TOFMS to the analysis of complex fuels is the conversion of that data into useful information. This challenge is overcome through application of powerful chemometric software methods that aid in the interpretation of such complex and voluminous data sets [28,29].

Partial least squares (PLS) analysis is a chemometric method that associates the differences in measurable information for two different data sets. Briefly, PLS analysis mathematically relates, via linear algebra, two data matrices (X- and Y-block) through calculation of new variables referred to as latent variables (LV), each of which is a linear combination of the original variables. Ultimately, PLS analysis utilizes the relationship between the X-block and Y-block to predict the responses contained in the Y-block based upon predictor variables in the X-block. In the study herein, we construct models using PLS to account for the variance (ideally, the relevant chemical differences) in both the GC×GC-TOFMS data for a fuel sample set (i.e., signal intensities, which constitute the X-block) and measured property values for the same sample set (which constitute

the Y-block). A key output of PLS modeling is a one-to-one correspondence between the predicted values relative to the measured values based upon the relationship to the chemical composition encoded in the GC×GC-TOFMS data (referred to as the PLS calibration). Another important output obtained from PLS modeling is the linear regression vectors (LRV), which indicate how the variables of the X-block (chromatographic and mass spectral information) relate to the Y-block (physical properties). PLS has been used in many applications with GC×GC instrumental systems [19,20,30–34] to relate the chemical information to physical properties of interest. Additional details on PLS methods are available in the literature [35–38].

A large sample set with predominantly natural compositional diversity (as opposed to chemical diversity introduced by blending a limited number of chemical streams in varying proportion) is crucial in the development of PLS models that provide meaningful, chemically accurate conclusions. In the absence of this compositional diversity, collinearity between compounds (or classes of compounds) can result in misleading information and errant conclusions, even when PLS calibrations are accurate [19,20]. To address this issue, this study encompasses seventy-four compositionally diverse, multicomponent fuels comprising both middle distillate commodity products and blended formulations of hydrocarbon solvents.

Early property-composition models were based on simple linear relationships that correlated bulk compound class content (e.g., n- and iso-alkanes, cycloalkanes, and aromatics) with fuel properties (e.g., smoke point, hydrogen content, heat of combustion, specific gravity) to ascertain the chemical makeup necessary to produce specification fuels [1–4]. Fuel composition-property relationships have also been investigated using artificial neural network approaches coupled with determination of chemical composition by GC-MS [5], PLS modeling applied to chemical data obtained from near infra-red spectroscopy and GC-MS [39–41], and a comparison

of the weighted average method, PLS, and support vector machines combined with chemical composition determined by GC×GC-MS and GC×GC-FID [33,42]. Though the application of GC×GC methods in determining chemical composition of fuels has significantly increased the ability to quantify individual chemical compounds, the work to date has classified the hydrocarbons into composition matrices based on hydrocarbon class and carbon numbers. A key goal of our study is to demonstrate the creation of chemically specific models that are based on relationships between the property and individual hydrocarbon species. Thus, investigation of the LRV from PLS modeling performed on the "raw" GC×GC-TOFMS data (i.e., not on peak tables or otherwise severely reduced data sets) provide a more comprehensive and accurate understanding of the relationship of the physical properties to the chemical composition of this diverse fuel set. Specifically, using PLS the chemical information obtained from GC×GC-TOFMS is correlated with measured viscosity, heat of combustion, hydrogen content, and temperature-dependent density, and the LRV are examined to provide insight into the chemical compounds that are directly responsible for the observed correlations. The current study constitutes a significant expansion of our previous studies to gain further insight and create robust PLS models with a large, diverse set of fuels [19,20].

4.2 EXPERIMENTAL

4.2.1 *Fuel Sample Set*

An expanded fuel set of seventy-four hydrocarbon fuels, produced in most cases to meet specifications for aerospace kerosene products (e.g., Jet A, JP-5, JP-8, RP-1, and RP-2), was acquired from the Air Force Research Laboratory (AFRL, Edwards AFB, CA and Wright-Patterson AFB, OH) and analyzed as received. Table 4.1 summarizes all the samples. Briefly, forty-nine fuels are rocket-grade kerosene formulations, seventeen are refined jet fuels, four are

jet fuels from alternative sources, two are specialty aerospace fuels, and two are commercial hydrocarbon products.

Table 4.1. List of the 74 kerosene-based fuels analyzed. Fuel specifications (i.e., type) and thermophysical properties are listed.

Sample Number	Sample Name	Type	Viscosity 40 °C (cSt)	Heat of Combustion (Btu/lbm)	Hydrogen Content (mass %)	Density 15 °C (g/mL)	Density 45 °C (g/mL)	Density 85 °C (g/mL)
1	YA2921HW10	RP-2	1.609	18626	14.202	0.81022	0.78841	0.75906
2	BG1121GP04	RP-1	1.696	18626	14.184	0.80879	0.78714	0.75797
3	GRC/0-100 HEP	RP-1	1.612	18594	14.187	0.81011	0.78801	0.75861
4	WC0721HW01	RP-2	1.760	18660	14.368	0.80318	0.78149	0.75241
5	LB073009-05	RP-1	1.710	18555	14.054	0.81362	0.79184	0.76259
6	ZI1521HW10	RP-1	1.675	18559	14.130	0.81447	0.79259	0.76330
7	CG0721HW10	RP-2	1.645	18583	14.215	0.80999	0.78814	0.75887
8	LB073009-08	RP-1	1.600	18580	14.183	0.80498	0.78308	0.75378
9	BB0821HW10	RP-2	1.705	18556	14.178	0.80890	0.78711	0.75786
10	LB080409-05	RP-1	1.663	18587	14.160	0.81318	0.79139	0.76210
11	ZI2621HW01	RP-2	1.852	18603	14.276	0.80822	0.78667	0.75770
12	ZJ1321GP01	RP-2	1.662	18608	14.218	0.80864	0.78690	0.75768
13	RG3021LS06	UL-RP-1	1.593	18604	14.242	0.80743	0.78562	0.75630
14	RG3021LS05	RP-TS-5	1.584	18602	14.216	0.80745	0.78559	0.75622
15	POSF 3327	JP-7	1.537	18690	14.536	0.79124	0.76944	0.74008
16	POSF 4765	JP-900	1.889	18289	13.073	0.86994	0.84793	0.81854
17	LB073009-02	RP-1	1.588	18568	14.055	0.81131	0.78939	0.76003
18	B0112868	RP-1	1.472	18590	14.125	0.80520	0.78324	0.75375
19	VI2621LS01	RP-1	1.593	18598	14.210	0.80962	0.78769	0.75832
20	DB131014	RP-1	2.324	18683	14.451	0.80694	0.78587	0.75771
21	DC310925	RP-1	1.600	18619	14.234	0.80672	0.78486	0.75549
22	DC310923	RP-1	1.623	18623	14.211	0.80762	0.78579	0.75647
23	DB131013	RP-1	1.647	18644	14.249	0.80738	0.78554	0.75634
24	DB131015	RP-1	2.117	18489	13.790	0.83545	0.81387	0.78472
25	CL031236	RP-1	1.031	18607	14.215	0.80329	0.78066	0.74973
26	CB1121HW10	RP-2	1.688	18546	14.122	0.81415	0.79237	0.76320
27	EA130720	RP-1	1.652	18514	13.966	0.81743	0.79545	0.76616
28	EB220705	RP-1	1.559	18538	14.005	0.81227	0.79032	0.76088
29	CHC JP-5	JP-5	1.347	18594	14.117	0.80160	0.77923	0.74923
30	LB080409-01	RP-1	1.654	18553	14.179	0.81103	0.78927	0.76001
31	LB073009-01	RP-1	1.720	18559	14.146	0.81583	0.79401	0.76494
32	A0072256	RP-1	1.433	18636	14.057	0.80272	0.78069	0.75090
33	CL11-3089	D80	1.635	18668	14.570	0.79370	0.77212	0.74289
34	LB100413-40	RP-1	1.644	18582	14.093	0.80887	0.78719	0.75782
35	LB073009-03	RP-1	1.652	18590	14.157	0.81352	0.79178	0.76240
36	LB073009-10	RP-1	1.593	18563	14.188	0.81222	0.79030	0.76096
37	SA1421LS03	UL-RP-1	1.603	18584	14.213	0.80990	0.78809	0.75878

38	ED060739	RP-1	1.667	18497	13.970	0.81794	0.79616	0.76695
39	CB1121HW10	RP-2	1.694	18526	14.157	0.81424	0.79244	0.76325
40	LB073009-09	RP-1	1.608	18552	14.160	0.81269	0.79075	0.76135
41	LB073009-06	RP-1	1.726	18592	14.266	0.80938	0.78767	0.75852
42	XC2521HW10	RP-1	1.588	18582	14.253	0.80924	0.78736	0.75796
43	ZK0821HW20	RP-2	1.661	18550	14.184	0.81349	0.79164	0.76243
44	ZK2121HW10	RP-2	1.526	18550	14.180	0.81250	0.79057	0.76083
45	CL11-2928	JP-8	1.344	18487	13.779	0.80371	0.78144	0.75137
46	CL11-2929	IPK	1.128	18815	15.192	0.75995	0.73768	0.70713
47	CL11-2654	HRJ	1.522	18819	15.097	0.76476	0.74310	0.71342
48	CL12-3493	Decalin	1.989	18188	12.944	0.88432	0.86154	0.83123
49	CA2021HW10	RP-2	1.645	18572	14.188	0.81090	0.78919	0.75989
50	POSF 9641	ATJ-8	1.481	18797	15.247	0.75843	0.73674	0.70738
51	41910	RP-1	1.512	18602	14.266	0.80469	0.78283	0.75327
52	B01001634-01	RP-1	1.620	18579	14.263	0.80732	0.78554	0.75633
53	POSF 4751	JP-8	1.340	18394	13.932	0.80370	0.78150	0.75150
54	POSF 10359	JP-8	1.330	18553	13.988	0.80240	0.78030	0.75020
55	POSF 10314	JP-8	1.510	18462	13.908	0.81240	0.79050	0.76080
56	POSF 10312	JP-8	1.370	18463	13.751	0.81140	0.78930	0.75940
57	POSF 10316	JP-8	1.290	18522	13.971	0.80480	0.78250	0.75220
58	POSF 9326	Jet-A	1.190	18483	13.761	0.80500	0.78240	0.75190
59	POSF 10358	Jet-A	1.430	18459	13.946	0.80550	0.78440	0.75470
60	POSF 10311	Jet-A	1.370	18450	13.499	0.82010	0.79790	0.76790
61	POSF 10315	Jet-A	1.140	18623	14.282	0.79180	0.76970	0.73960
62	POSF 10369	Jet-A	1.180	18533	13.985	0.79730	0.77480	0.74430
63	POSF 10313	Jet-A	1.480	18488	13.618	0.81020	0.78830	0.75900
64	POSF 10337	JP-5	1.370	18514	13.852	0.80800	0.78580	0.75590
65	POSF 10325	Jet-A	1.310	18506	13.920	0.80320	0.78100	0.75090
66	POSF 10264	JP-8	1.140	18584	14.410	0.77990	0.75740	0.72860
67	POSF 10289	JP-5	1.570	18429	13.580	0.82680	0.80480	0.77540
68	POSF 9698	JP-8	1.300	18530	14.078	0.79920	0.77690	0.74670
69	EA0721BE01	RP-1	1.680	18611	14.246	0.80962	0.78735	0.75821
70	CL1621HW10	RP-1	1.601	18610	14.289	0.80667	0.78404	0.75473
71	DA2321HW20	RP-1	1.602	18627	14.260	0.80684	0.78492	0.75556
72	DJ2621BE10	RP-1	2.148	18508	13.644	0.83641	0.81429	0.78558
73	Sample A	RP-1	1.563	18624	14.441	0.80416	0.78234	0.75298
74	Sample B	RP-1	1.586	18625	14.436	0.80608	0.78435	0.75520

4.2.2 *GC×GC-TOFMS Analysis*

Analysis of fuel chemical composition was performed using a GC×GC-TOFMS instrument consisting of an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA), a thermal modulator (4D upgrade, LECO, St. Joseph, MI, USA), and a Pegasus III TOFMS (LECO, St.

Joseph, MI, USA). Aliquots of the fuel samples were introduced to the GC×GC-TOFMS instrument via a 7683B auto-injector (Agilent Technologies, Palo Alto, CA, USA). The auto-injector was set to inject 1 μL at a 200:1 split at an inlet temperature of 275 °C. Prior to injection, HPLC grade acetone and hexane (Fisher Scientific) were used as solvent rinses. The ^{1}D column was a Rxi-17Sil MS: 29.5 m × 250 μm inner diameter (i.d.) × 0.25 μm film thickness, and the ^{2}D column was a Rxi-1MS 1.5 m × 180 μm i.d. × 0.18 μm film thickness. Ultrahigh purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) was used as the carrier gas at a constant flow rate of 2.0 mL/min. The ^{1}D oven was held at 40 °C for 1.5 min before being temperature programmed at 5 °C/min to 200 °C where it was held for 1 min. The ^{2}D oven was held at a +12 °C offset relative to the ^{1}D oven and the modulator block was held at a +30 °C offset relative to the ^{1}D oven. The modulation period, P_M, was 3 s (^{2}D separation run time) with 0.75 s hot and cold pulses for each stage. The transfer line was set to 285 °C and the TOFMS ion source was 225 °C. Mass channels, *m/z*, 35-334 at unit resolution were collected with an electron impact ionization voltage of 70 eV at 100 spectra/s after a 10 s acquisition delay. All samples (Table 4.1) were analyzed with two GC×GC-TOFMS replicates collected for each sample. Samples 1-68 and 69-74 analyses were performed in two distinct campaigns separated by six months.

4.2.3 *Measured Fuel Properties*

Fuel properties of relevance to fluid physical behavior and thermochemistry were acquired for the sample set and used in the current effort. Kinematic viscosity, net heat of combustion, and density for samples 1-52 and 69-74 were measured by AFRL/RQRP (Edwards AFB, CA); the same properties for samples 53-68 were provided by AFRL/RQTF (Wright-Patterson AFB, OH). Hydrogen content for all seventy-four fuels was measured by the Air Force Petroleum Office Fuels Laboratory (Vandenberg AFB, CA). Precision of reported property values is crucial in the

development of mathematical models that encompass variance in quantitative fuel data. Viscosity was measured at 40 °C in accordance with ASTM D445/446 using a capillary viscometer system (AVS 370, Xylem Analytics Germany). The method states repeatability of 0.56% at 40 °C for successive measurements (for kerosene and biodiesel fuels) [43]. Net heat of combustion was determined in duplicate or triplicate according to ASTM D4809 (Model 6200 calorimeter, Parr Instruments). Relative standard deviation (RSD) was <0.5% for all samples and <0.1% for all but four. According to D4809, the difference in successive results is expected to exceed 0.096 MJ/kg for no more than one in twenty samples [44]. For the fifty-eight fuels analyzed at Edwards AFB, CA, the largest discrepancy in successive results was 0.074 MJ/kg, and the average discrepancy was 0.022 MJ/kg. Hydrogen content was measured by time-domain nuclear magnetic resonance (TD-NMR) analysis in accordance with ASTM D7171 (minispec, Bruker) [45]. Duplicate measurements were performed on eight samples to evaluate precision. For these repeat determinations, the difference between successive results varied from 0.000-0.034 percent hydrogen by mass (hereafter mass %), with an average of 0.007 mass % (significantly better than the stated method repeatability of 0.12 and 0.13 mass % for hydrogen content values of 13.00 and 15.50 mass %, respectively). Density was measured at 15, 45, and 85 °C following ASTM D4052 (DDM2911 density meter, Rudolph Research Analytical) [46]. Method precision suggests that the expected difference in successive density measurements is <0.00016 g/mL with 95% confidence for the applicable density range (0.80-0.88 g/mL). Property data presented herein adopt the following units to maintain consistency with aerospace propulsion development: kinematic viscosity, mm^2/s (cSt); density, g/mL; heat of combustion, Btu/lbm (1 MJ/kg = 429.9 Btu/lbm); and hydrogen content, mass %. A summary of the property data is provided in Table 4.1. Ancillary compositional analyses of several fuels were performed by University of Dayton Research Institute

(UDRI) at AFRL/RQTF, Wright-Patterson AFB, OH using a GC×GC-MS / flame ionization detection (FID) method described elsewhere [26]. These quantitative data facilitated the interpretation of results obtained herein as they pertain to general hydrocarbon group-types.

4.2.4 *Data Analysis*

The GC×GC-TOFMS chromatograms were imported into MATLAB 2015b (MathWorks, Natick, MA) using an in-house converting algorithm. The data were baseline corrected and binned by 4 data points (i.e., 4 modulations) on the ^{1}D dimension and by 15 data points (150 ms) on the ^{2}D dimension, thereby reducing the sample matrix from (600 ^{1}D data points × 300 ^{2}D data points × 300 *m/z*) to (150 ^{1}D data points × 20 ^{2}D data points × 300 *m/z*). This binning procedure mitigated any minor run-to-run GC retention time misalignment while also reducing computation time [37]. Because Samples 69-74 were analyzed 6 months later than Samples 1-68, there was minor ^{1}D retention time shifting, which was effectively corrected by the binning. After the preprocessing steps, the chromatographic data was forwarded to the PLS Toolbox 8.6.1 analysis software (Eigenvector Research Inc., Wenatchee, WA, USA) for initial mean centering. The variables forwarded to PLS were the bins of chromatographic data per *m/z* obtained by summing the signal within each bin, which is analogous to peak area in that an increase in analyte concentration will result in a larger summed signal for a given bin.

When building predictive models such as in PLS, the sample set is divided into calibration and external validation sets, both of which are selected to be representative of the entire sample set [47]. When sample size is relatively small, achieving representative splits is difficult and leads to a reduction in model robustness due to a small calibration set size [48]. Furthermore, it is not always feasible to obtain an independent test set with similar chemical composition and other properties collected using the same instrument at a different time. In these cases, internal validation

(e.g., cross validation) is well suited for the approximation of error for similar samples obtained using the same instrumentation that are not included in the sample set [49]. Thus, leave-one-out cross validation (LOOCV) is often implemented with PLS [19,20,30,37]. However, with a large data set, LOOCV would result in excessively long computation time and potential overfitting of the data. Therefore, the PLS model construction implemented an alternative cross validation method known as Venetian blinds cross validation to test the overall predictive ability of the models and to determine the correct number of latent variables [50,51]. For Venetian blinds cross validation, the data for N total samples are split s times into groups of m samples, if N is evenly divisible by s. Therefore, s calibration models are prepared using each group of $N - m$ samples, and the m samples left out of the calibration are predicted. If N is not evenly divisible by s, then some of the groups have more than m samples to make up the difference. For example, if the PLS analysis was for sixty-six fuels (N evenly divisible by $s = 11$), the first sub-validation step would exclude six fuels in the calibration (Samples 1, 12, 23, 34, 45, and 56) that are subsequently predicted. The next sub-validation step would exclude Samples 2, 13, 24, 35, 46, and 57 in the calibration, which are then predicted, and so on. Since this study investigates seventy-four fuels (N is not divisible by $s = 11$), eight groups excluded from the sub-validation step had seven fuels instead of six fuels. The LRV were investigated to determine the relationship between the chemical composition and physical measurements. Analyte compounds that exhibit positive LRV values correlate with increased measured property values, while analyte compounds that exhibit negative LRV values correlate with decreased property values [17–20,30,35–38]. The GC×GC-TOFMS chromatogram replicates were analyzed in separate PLS models and the average predicted values were plotted. In order to determine the "goodness-of-fit" for the PLS models, the root mean square

error of cross validation (RMSECV) was calculated for each model and then averaged, as defined by:

$$RMSECV = \left[\frac{1}{N} \times \sum\left(y_{i,cv} - y_{i,meas}\right)^2\right]^{0.5} \quad (4.1)$$

where N is the number of fuel samples, $y_{i,cv}$ is the cross validation predicted value of sample i and $y_{i,meas}$ is the measured value for the same sample. RMSECV can be normalized (NRMESCV) by dividing the RMSECV by the range of the $y_{i,meas}$ values. In addition, principal component analysis (PCA) was performed on the three LRV from the density modeling for the purpose of interrelating three separate PLS models at 15, 45, and 85 °C.

4.3 Results and Discussion

4.3.1 *GC×GC-TOFMS Dataset*

The GC×GC-TOFMS instrument in the reversed column configuration provides an excellent 2D chromatographic separation of the fuel components, which enables a much more detailed examination of the physical properties in relation to the chemical composition than previously demonstrated [1–4]. Using the total ion current (TIC) chromatogram of Sample 67 (JP-5) for illustration purposes, Figure 4.1A shows the approximate location in the 2D separation of the primary compound classes (alkanes, cycloalkanes, and aromatics). The approximate location of these compound classes are delimited in the LRV to provide sound chemical interpretation of overall trends. The fuels used in this study, and multicomponent distillate fractions in general, possess individual chemical "fingerprints" comprised of a unique combination of compound classes and sub-classes. Note that data collected using a 1D-GC separation of the same complex fuel results in significant peak overlap (Figure 4.1B), making compound class visualization difficult and ambiguous.

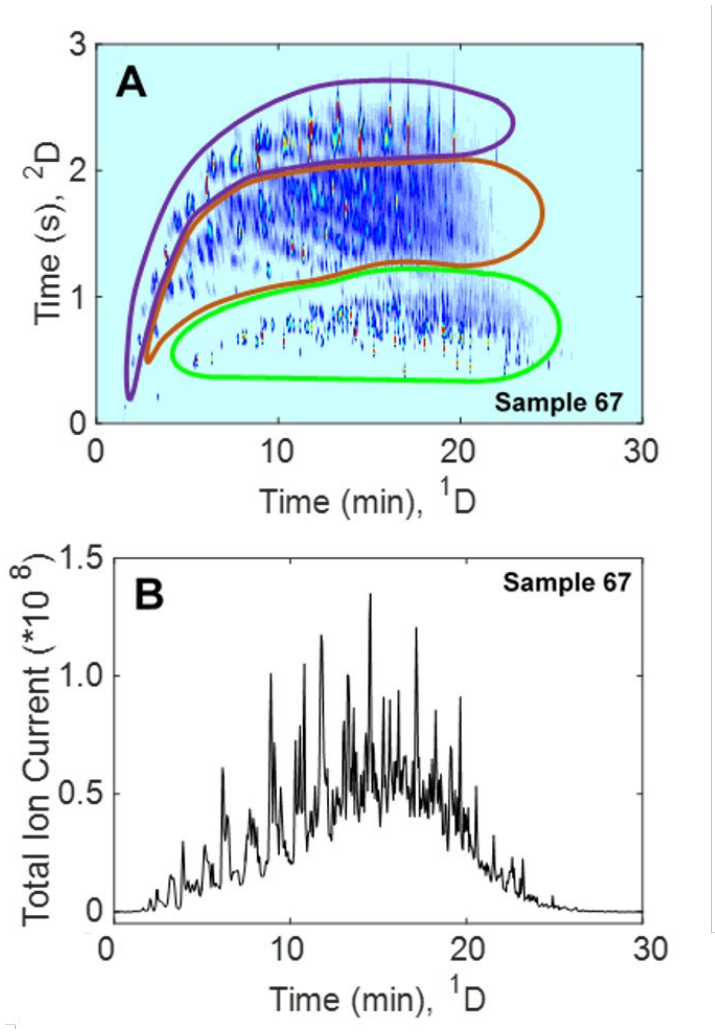


Figure 4.1. Demonstration of GC×GC separation capabilities. (A) Representative GC×GC-TOFMS TIC chromatogram for the fuel sample set (Sample 67: JP-5, POSF 10289) with the three major chemical classes highlighted: alkanes (paraffins) in purple (top), cycloalkanes (cycloparaffins) in orange (middle), and aromatics in green (bottom). (B) One-dimensional GC separation of Sample 67, obtained by summing the second dimension separations of the GC×GC TIC chromatogram in (A) onto the first dimension axis.

Additionally, because a multivariate detector (TOFMS) was employed, informative utilization of *m/z* that are representative of chemical compound classes present in the fuels is possible. Figure 4.2 illustrates this capability by showing several analytical ion chromatograms (AIC) for Sample 67, thereby demonstrating the power of using selective *m/z* coupled with advantageous 2D separations to isolate and, in principle, quantify hydrocarbon classes and compounds in multicomponent fuels. Additionally, for each compound class (alkane, cycloalkane, and aromatic)

it is possible to discern sub-classes (e.g., n-alkanes, iso-alkanes, monocycloalkanes). An alternative to rigorous identification and quantification of all compounds within the major hydrocarbon classes, the inclusion of all chromatographic data per *m/z* enables all potential peaks, including those attributed to co-elutions, heteroatom-containing analytes, and unknown compounds, to contribute to the PLS model. Including all *m/z* also increases the probability of resolving co-eluting compounds through potential selective *m/z* that describe the pure compounds.

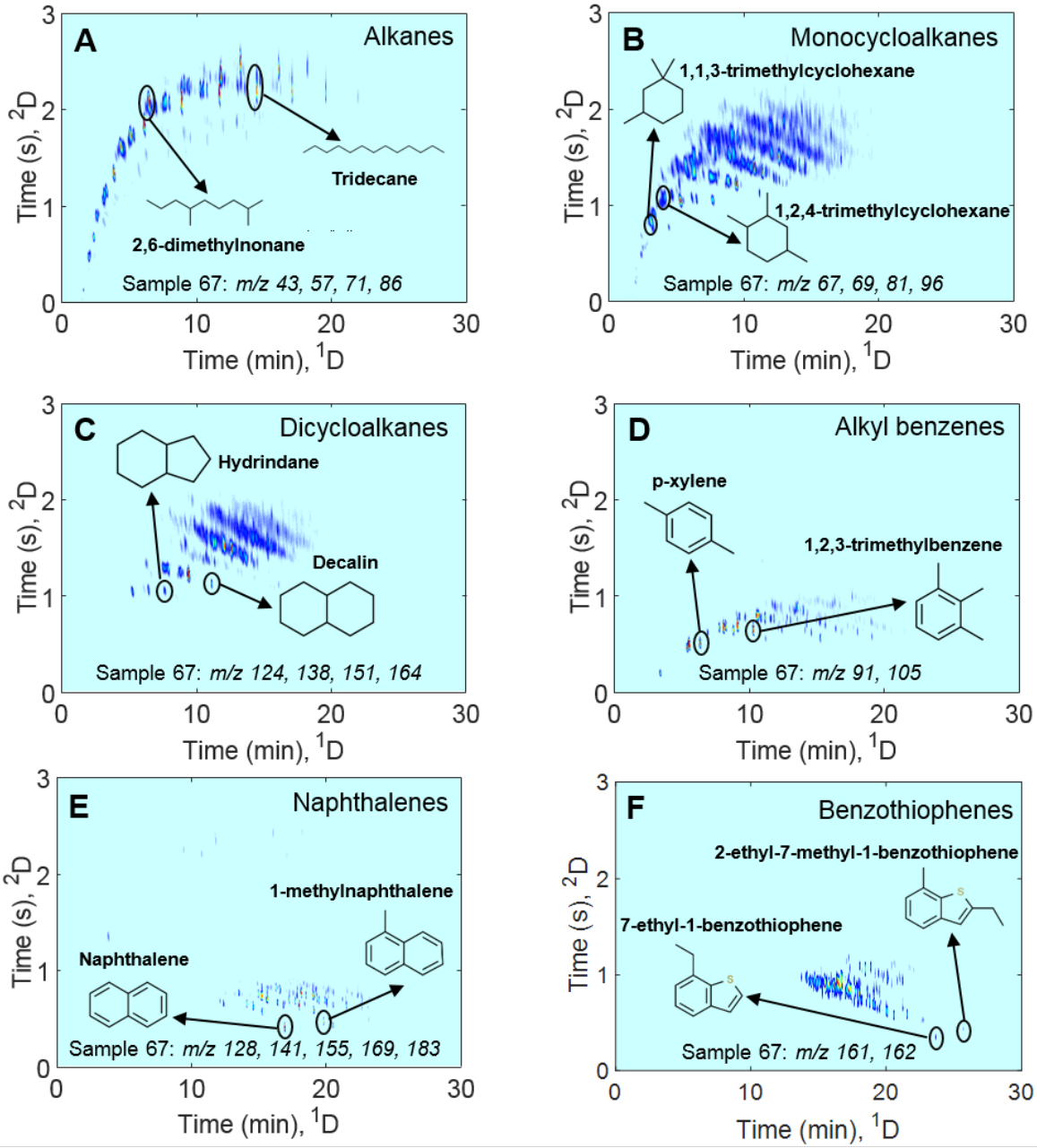


Figure 4.2. Analytical ion chromatograms (AIC) for Sample 67 showing the ability to use selective *m/z* to identify the chemical compound classes and subclasses (e.g., monocycloalkanes, dicycloalkanes, alkyl benzenes). For each figure, the AIC is the sum of the signal for only the *m/z* indicated. Within each figure are names and structure diagrams of representative compounds that have been circled. (A) AIC for alkanes. (B) AIC for monocycloalkanes. (C) AIC for dicycloalkanes. (D) AIC for alkyl benzenes. (E) AIC for naphthalenes. (F) AIC for sulfur-containing compounds (benzothiophenes).

Figure 4.3 shows the total ion current (TIC) chromatograms of four representative fuels exhibiting "typical" chemical composition for a distillate or blended multicomponent fuel and two fuels that exhibit "atypical" composition. Herein, we refer to a fuel as having an atypical composition if intense, overloaded chromatographic peaks are present in the TIC chromatogram, whereas a typical composition does not exhibit intense, overloaded peaks. Since chemical composition was unknown prior to GC×GC-TOFMS analysis, all fuel samples were analyzed under the same conditions (e.g., injection volume, inlet split ratio). For atypical fuels, this could have resulted in overloading of the detector, manifesting as peak broadening and co-elution of compounds in some areas of the chromatogram. Figure 4.3A-B show representative chromatograms of typical RP-2 fuels, and Figure 4.3E-F illustrate representative chromatograms of typical jet fuels. Chromatograms of atypical fuels are displayed in Figure 4.3C-D, as indicated by the disproportionately intense chromatographic peaks. Reflecting upon Figure 4.1, Figure 4.2, and Figure 4.3, evidently these six fuels are compositionally very diverse, with the differences in hydrocarbon classes and specific compound concentrations effectively captured by the GC×GC separations.

GC×GC chromatograms (TIC) of all seventy-four fuels are provided in Appendix C (Figure C.1), serving as a visual resource. The compositional variations among fuels can be chemometrically modeled in such a way that was previously not possible [1–4]. As mentioned, the fuel sample set demonstrates diversity in terms of chemical composition and more specifically volatility (relative location of peaks in the 2D separation space), an ideal feature in composition-based modeling of thermophysical properties. However, eight fuels (Samples 20, 24, 25, 46, 47, 48, 50, and 72) exhibit a significantly atypical chemical fingerprint compared with the remaining

sixty-six samples, a characteristic largely attributable to differences in fuel production. The consequences of extensive compositional disparity are discussed subsequently.

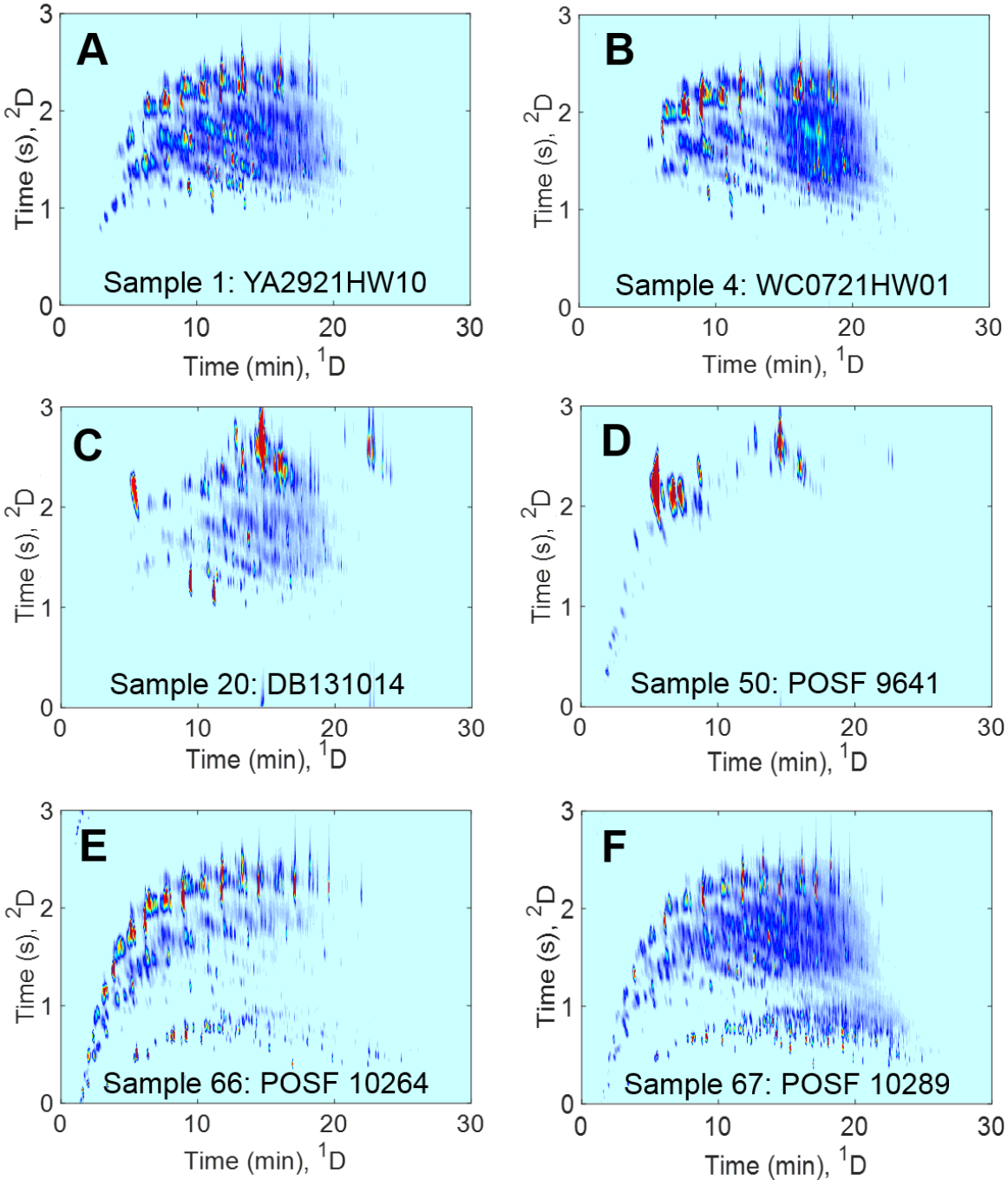


Figure 4.3. Six total ion current (TIC) GC×GC chromatograms depicting the compositional diversity of fuels analyzed. (A) Sample 1: RP-2. (B) Sample 4: RP-2. (C) Sample 20: RP-1 produced by unconventional blend stocks. (D) Sample 50: ATJ-8. (E) Sample 66: JP-8, POSF 10264 (low aromatics). (F) Sample 67: JP-5, POSF 10289 (high aromatics).

4.3.2 *PLS Modeling*

We now turn our attention to the PLS modeling of physical properties. Most of the properties modeled herein were previously modeled using simple linear property-composition relationships involving bulk measurements of n-alkanes (n-paraffins), iso-alkanes (iso-paraffins) and cycloalkanes (naphthenes), and aromatic compounds [1–4]. Quantitative analysis of the hydrocarbon compound classes in these studies required several analytical procedures (i.e., HPLC and GC or ^{13}C NMR), as a single method was shown to be insufficient to accurately determine the fuel composition in terms of the above compound classes. For example, when ^{13}C NMR was applied alone, the measurement of n-alkyl carbon content <15% proved very difficult due to resonance overlap attributed to other aliphatic compounds [3,4]. When several properties measured from a subset of the fuels in the present study (inverse density, heat of combustion, and hydrogen content) were modeled using three parameter linear equations based on measured weight fractions of n-alkanes, combined iso-alkanes and cycloalkanes, and aromatics, and experimentally determined coefficients assumed to be applicable to all kerosene fuels [2], normalized root mean square error (NRMSE) values observed were 54, 17, and 35%, respectively. Also notable is that nearly all of the RP-1 and RP-2 fuels in the present work have <15% alkyl carbon content, which would produce inaccurate results if ^{13}C NMR analysis were to be used to evaluate hydrocarbon content. While property-composition models constructed using hydrocarbon type are effective, they disregard the relationship between the number of carbon atoms (related to boiling point) within a single compound class and the modeled property. Consideration of this relationship is particularly critical when fuels of variable boiling point range are analyzed; this point was addressed with the addition of several temperature terms accounting for the temperatures at which certain percentages of the fuel has vaporized [4]. However, creation of sensitive models that are

responsive to minute differences in sample composition at the level of individual chemical species would better address boiling point differences within compound classes and, more generally, between fuels. For the work herein, a single, comprehensive analytical platform is provided in which all relevant hydrocarbon compound classes are determined simultaneously, with sensitivity and specificity that allows quantification of individual chemical species over a large concentration range. With the vast increase in chemical variables provided by such a platform, an extension of multiple linear regression that can handle highly correlated variables is required. GC×GC-TOFMS analysis provides a means to investigate the chemical composition of fuels at a deeper level, while PLS modeling leverages these chemical composition differences in the samples at the specific-compound level instead of the bulk-compound level.

Based on GC×GC-TOFMS chromatograms for each fuel, viscosity of each sample was predicted using PLS models with five latent variables. Figure 4.4A shows the calibration results from the modeled viscosity for the full fuel sample set with the average predicted result of both replicates of the GC×GC-TOFMS analyses shown. The NRMSECV for the modeling is less than 10%. However, many of the previously mentioned compositionally atypical samples do not model correctly. The LRV in Figure 4.4B indicate pronounced chemical features affecting the model, namely, compounds corresponding to a large red feature at a ^{1}D retention time of $^1t_R \approx 8$ min and large blue features at $^1t_R \approx 10$ and 12 min. All sample chromatograms were inspected and those containing large, overloaded peaks were excluded from subsequent analyses. Furthermore, examination of the Hotelling's T^2 statistic (Figure 4.S1) reinforced this decision by revealing outliers that adversely influence the PLS modeling due to previously noted chromatographic differences relative to typical fuels.

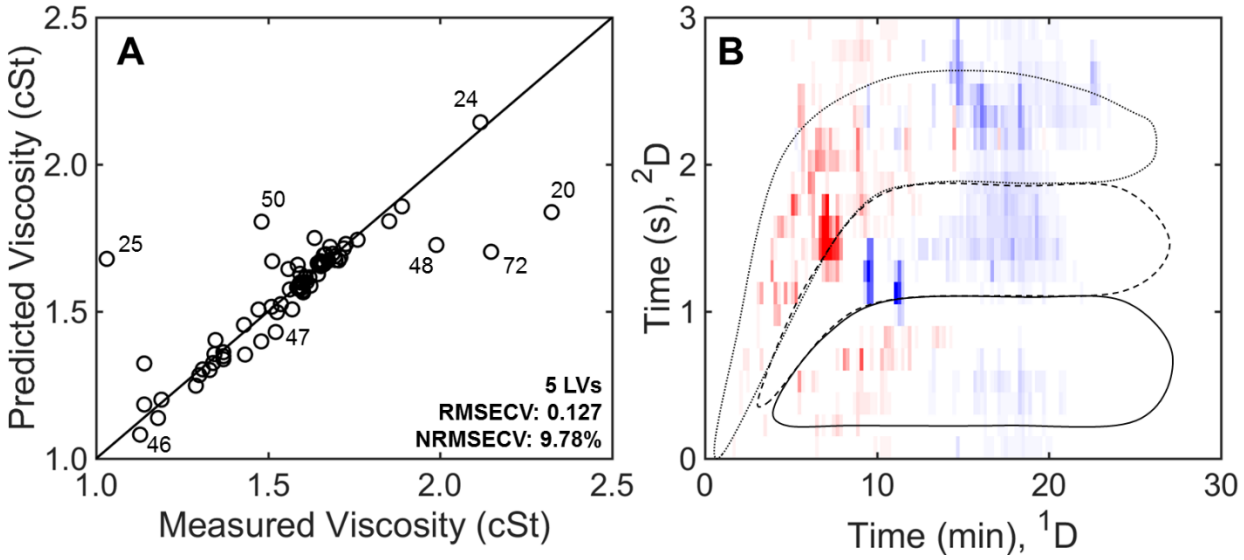


Figure 4.4. PLS prediction of viscosity for 74 fuels. (A) PLS calibration of viscosity using GC×GC-TOFMS chromatograms of all 74 fuel samples (Table 4.1). The average of the two predicted results for the two GC×GC-TOFMS replicates is shown. The black line represents the ideal agreement between the predicted and measured values, with slope = 1. There are several fuels that are incorrectly predicted. (B) Linear regression vectors from the PLS prediction of the viscosity. Blue indicates positive values in the LRV while red indicates negative values in the LRV for the chromatographic variables. Alkanes are located within the dotted line, cycloalkanes are located within the dashed line, and aromatics are located within the solid black line.

For example, Samples 20 and 50, the atypical fuels featured in Figure 4.3C-D, contain overloaded peaks amongst compounds that may be expected to comprise kerosene-based fuels. These overly concentrated compositional features render these fuels and others like them difficult from a modeling perspective. The chromatograms of all eight excluded outlier fuels (Samples 20, 24, 25, 46, 47, 48, 50, and 72) can be inspected in Figure C.1 in Appendix C. Other fuels also exhibited atypical chemical composition to a lesser extent; as they did not exhibit overloaded peaks, these fuels were included in subsequent modeling. Sample 16, in particular, was an atypical fuel that nearly qualified for exclusion, so its inclusion serves to test the limits of the PLS modeling predictive capabilities. Investigation of the physical properties in Table 4.1 indicate that the differences in chemical composition of the excluded fuels correspond to significant deviation of some physical property values from the observed range of the typical fuels. For example, Sample

50 falls on the high end for heat of combustion (18797 Btu/lbm) and hydrogen content (15.247 mass %), and on the low end for density (0.75843 g/mL at 15°C, 0.73674 g/mL at 45°C, and 0.70738 g/mL at 85°C). These observations are consistent with the chemical composition of Sample 50, which exclusively contains alkanes.

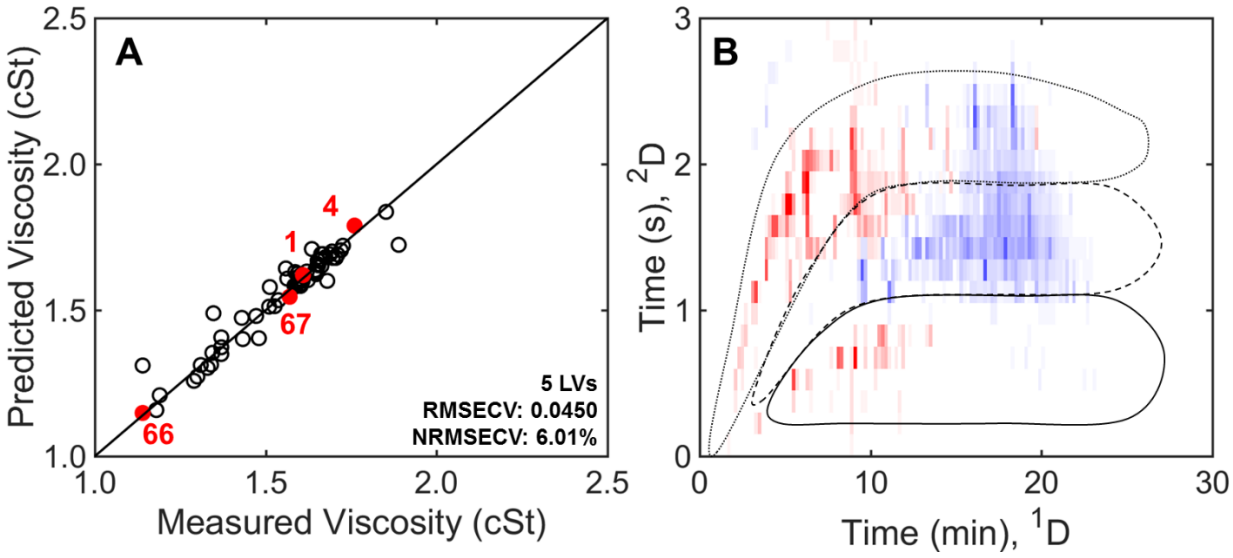


Figure 4.5. PLS prediction of viscosity for 66 fuels (without outliers). (A) PLS calibration of the viscosity using GC×GC-TOFMS chromatograms using 66 of the initial 74 fuels from Table 4.1, with excluded outliers: Samples 20, 24, 25, 46, 47, 48, 50, and 72. The average of the two PLS modeling results for the two GC×GC-TOFMS replicates is shown. The black line represents the ideal agreement between the predicted and measured values, with slope = 1. Samples 1, 4, 66, and 67 (from Figure 4.3) have been highlighted red in order to discuss the relationship between their chemical composition, location on the calibration plot, and LRV. (B) LRV from the PLS prediction of viscosity using 66 fuels. Blue indicates positive values in the LRV while red indicates negative values in the LRV for the chromatographic variables. Alkanes are located within the dotted line, cycloalkanes are located within the dashed line, and aromatics are located within the solid black line.

A more meaningful PLS model for viscosity was created using the remaining sixty-six fuels. Following the removal of the eight outlier fuels, all models were constructed using five latent variables. Scores plots for the various latent variables are available in Figure C.2 in Appendix C for inspection. Figure 4.5A shows the calibration plot based on the average of the chromatographic analyses, with the typical fuels from Figure 4.3 indicated in red (Samples 1, 4, 66, 67); the resulting

model error is significantly lower (NRMSECV ≈ 6%). The range of viscosity values over which the model accurately represents experimentally measured data is noteworthy. The LRV (Figure 4.5B) is absent of excessively overloaded compounds and provides a sound chemical interpretation. Compounds with a higher boiling point range (longer retention times on the ^{1}D separation, $^{1}t_{R}$) correlate with increasing viscosity and are indicated with blue, while lower boiling compounds (shorter $^{1}t_{R}$ on the ^{1}D separation) correlate with lower viscosity and are indicated with red; this result is consistent with expectations regarding the influence of molecular size and shape on viscosity [52]. Based on these observations, comparison of the chromatograms of the representative fuels presented in Figure 4.3 with the LRV in Figure 4.5B suggests the following relationship for relative magnitude of viscosity: Sample 4 > Sample 1 ≈ Sample 67 > Sample 66. This is appropriately reflected in the PLS viscosity calibration plot in Figure 4.5A.

The PLS modeling for heat of combustion is provided in Figure 4.6A, performed with eight outliers (Samples 20, 24, 25, 46, 47, 48, 50, and 72) excluded. Heat of combustion modeling results in slightly higher error than viscosity, with NRMSECV of ~10%. However, the method precision (repeatability) for ASTM D4809 is 0.096 MJ/kg (~ 40 Btu/lbm), thus the model quality (RMSECV = 41.3 Btu/lbm) is reasonable. Chemical interpretation of the LRV in Figure 4.6B indicates that as expected, alkanes correlate positively with heat of combustion while aromatics correlate negatively. The cycloalkanes contribute to a lesser extent based on the PLS modeling. Based on these relationships, visual inspection of the representative fuel chromatograms in Figure 4.3 in light of the LRV leads to the expectation that heat of combustion for Samples 1 and 4 is relatively high given their minimal aromatic content; Sample 66 heat of combustion is moderate since it is abundant in n-alkanes but also contains aromatics; and a relatively low heating value is expected for Sample 67 owing to its high aromatic content.

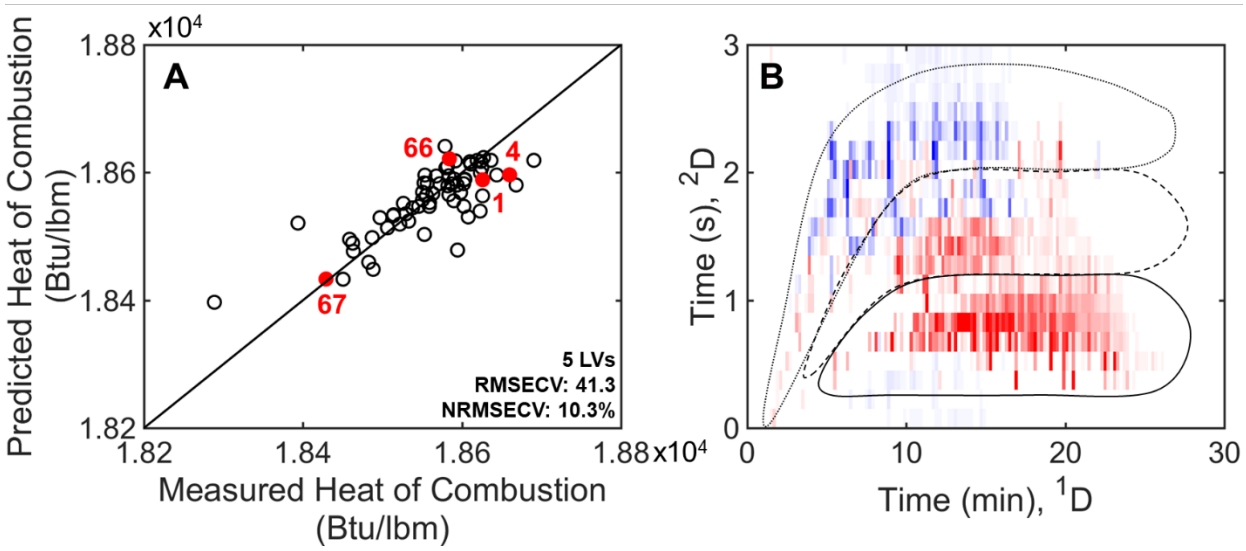


Figure 4.6. PLS prediction of heat of combustion for 66 fuels (without outliers). (A) PLS calibration of the heat of combustion using GC×GC-TOFMS chromatograms using 66 of the initial 74 fuels from Table 4.1, with Samples 20, 24, 25, 46, 47, 48, 50, and 72 excluded. The average of the two PLS modeling results for the two GC×GC-TOFMS replicates is shown. The black line represents the ideal agreement between the predicted and measured values. Samples 1, 4, 66, and 67 (from Figure 4.3) have been highlighted red in order to discuss the relationship between their chemical composition, location on the calibration plots, and LRV. (B) LRV from the PLS prediction of heat of combustion. Blue indicates positive values in the LRV while red indicates negative values in the LRV for the chromatographic variables. Alkanes are located within the dotted line, cycloalkanes are located within the dashed line, and aromatics are located within the solid black line.

Location of these four fuels in the PLS heat of combustion calibration plot shown in Figure 4.6A confirms this visual assessment. One sample with high residuals is observed with a measured heat of combustion of approximately 18300 Btu/lbm. This is Sample 16, one of the fuels that exhibited an atypical chemical composition but was included in the modeling. Sample 16 (Figure C.1) appears to be composed primarily of cycloalkanes that span a wide boiling point range. The near complete lack of alkanes results in a low heating value for this fuel; however the cycloalkane region is not weighted as significantly in the LRV as the alkanes (positive) and aromatics (negative), which causes Sample 16 to be predicted with a higher heat of combustion of approximately 18400 Btu/lbm.

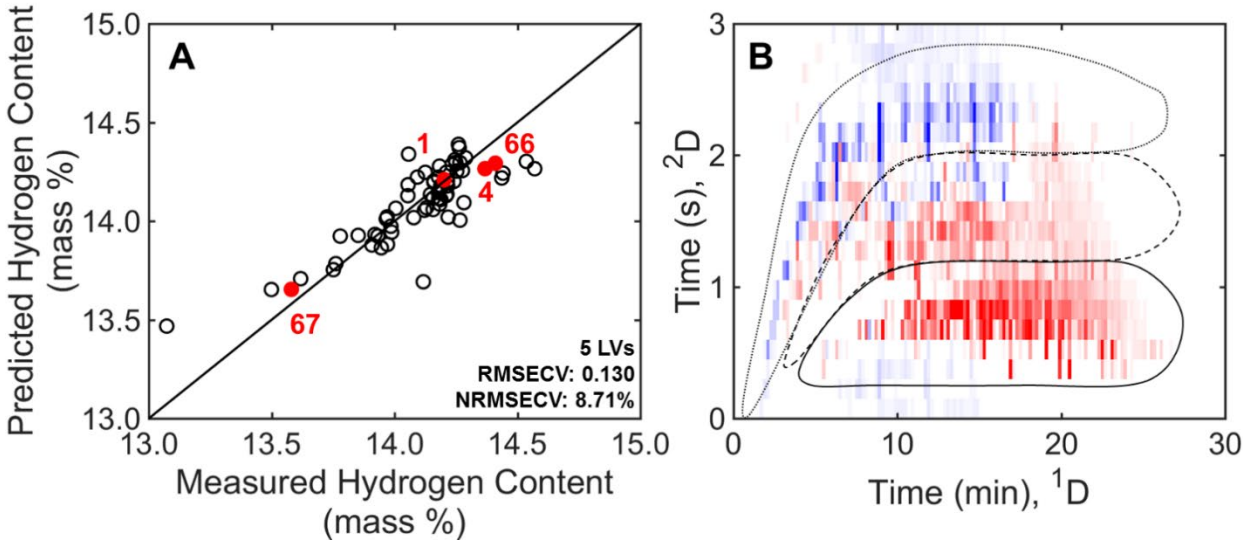


Figure 4.7. PLS prediction of hydrogen content for 66 fuels (without outliers). (A) PLS calibration of the hydrogen content using GC×GC-TOFMS chromatograms using 66 of the initial 74 fuels from Table 4.1, with Samples 20, 24, 25, 46, 47, 48, 50, and 72 excluded. The average of the two PLS modeling results for the two GC×GC-TOFMS replicates is shown. The black line represents the ideal agreement between the predicted and measured values. Samples 1, 4, 66, and 67 (Figure 4.3) have been highlight red in order to discuss the relationship between their chemical composition, location on the calibration plots, and LRV. (B) LRV from the PLS prediction of hydrogen content. Blue indicates positive values in the LRV while red indicates negative values in the LRV for the chromatographic variables. Alkanes are located within the dotted line, cycloalkanes are located within the dashed line, and aromatics are located within the solid black line.

PLS prediction of hydrogen content was performed in a manner similar to viscosity and heat of combustion, with the same samples excluded due to atypical chromatograms marked by excessively concentrated compounds. The calibration plot is shown in Figure 4.7A; the NRMSECV is less than 9%. Chemical interpretation of the LRV (Figure 4.7B) indicates that alkanes are positively correlated while aromatics are negatively correlated, an expected result given the relative unsaturation (higher carbon-to-hydrogen ratio) of aromatic rings. Similarly, monocycloalkanes (C_nH_{2n}) and dicycloalkanes (C_nH_{2n-2}) exhibit a negative correlation compared with n- and iso-alkanes (C_nH_{2n+2}), but to a lesser extent than aromatics. Again, samples from Figure 4.3 are indicated in red in Figure 4.7A to illustrate compositional dependence. As expected, the

low hydrogen content of Sample 67 (one of the lowest observed, at 13.58 mass %) is consistent with the high aromatic content of this aviation turbine fuel (20.4 mass % compared with an average value of 3.9 mass % for twenty-three fuels in this study). As noted, Sample 66 is a jet fuel with low aromatic content (13.6 mass %) compared with fuels encountered in practice, resulting in a relatively high hydrogen content (14.41 mass % compared with the study average of 13.90 mass % for refined jet fuels). In fact, Sample 66 is slightly higher in measured hydrogen content than Samples 1 and 4 (14.20 and 14.37 mass %, respectively), despite the very low aromatic content of the latter fuels. This arises from the abundance of saturated cycloalkanes in Samples 1 and 4 (61 and 46 mass % total cycloalkanes, respectively) compared with Sample 66 (23 mass %). Thus, the acyclic aliphatic contribution of Sample 66 (64 mass % combined n- and iso-alkanes) counteracts its aromatic content, leading to measured hydrogen content surpassing that of very low aromatic fuels (Samples 1 and 4). Again, Sample 16 appears as an outlier in Figure 4.7A with a significantly higher predicted hydrogen content than its true value (13.07 mass %). Similar to the heat of combustion model, the cycloalkane region of the LRV appears to contribute less to hydrogen content than the alkane and aromatic regions. This work is notable considering the limitation of previous modeling strategies that grouped iso-alkanes and cycloalkanes in a single class (both of which are in high abundance in kerosene fuels) and did not distinguish between aromatics with cyclic side chains and simple alkyl side chains, both of which groupings combine chemical species with significantly different hydrogen content for modeling purposes [2]. Although more recent studies have classified hydrocarbon species into more representative hydrocarbon classes in addition to separation by carbon number, the consideration of impurities such as olefins and heteroatom-containing compounds, such as sulfur and oxygen (oxygenates), has not been included [33,53]. The analytical procedure we employed utilizes the resolution between compound classes

and every chemical species to allow all of the chromatographic variables to contribute to the modeling.

Finally, an overlay of the three PLS density models at 15, 45, and 85 °C is provided in Figure 4.8A with the average of GC×GC-TOFMS replicates shown. The PLS model has an NRMSECV of ~7%. Samples 66, 4, 1, and 67 (left to right in Figure 4.8A) have been highlighted for all three temperatures in order to provide context for the relationship between the LRV and chemical composition of the fuels. The LRV for the PLS modeling of the density at 15, 45, and 85 °C are shown in Figure 4.8B-D, respectively. Inspection of the LRV indicates that alkanes are negatively correlated with density, while cycloalkanes and aromatics are positively correlated. As expected, the relative abundance of aromatics in Sample 67 leads to a higher density. Samples 1 and 4 both contain few aromatics (effecting a lower density with respect to Sample 67), but are differentiated from one another by cycloalkane and n-alkane content as discussed previously, thereby resulting in slightly higher density for Sample 1. Sample 66 has the lowest density given its abundance in straight and branched alkanes and relatively low aromatic content. For each temperature modeled, Sample 16 was severely under predicted with the highest measured density across all temperatures. Although cycloalkanes are shown to correlate positively with density, many of the chemical features present in Sample 16 fall outside of the region with highly loaded variables (i.e., dark blue). As was the case for heat of combustion and hydrogen content, density LRV indicate compositional relationships that are consistent with expected contributions from various hydrocarbon classes due to their geometry and relative packing efficiency. Upon examination, the three LRV in Figure 4.8B-D appear identical, however their relative magnitudes are not identical.

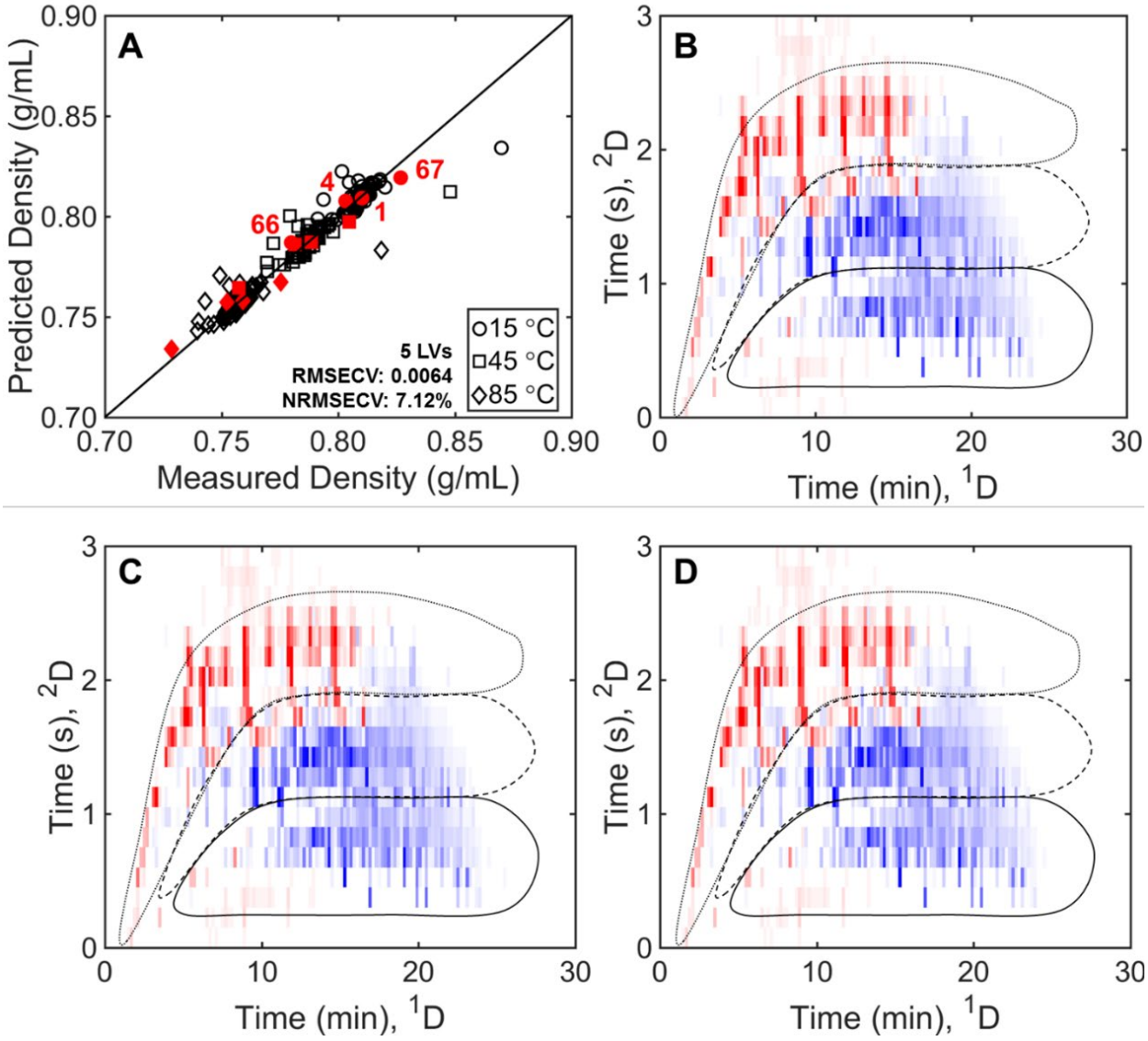


Figure 4.8. PLS prediction of temperature-dependent density for 66 fuels (without outliers). (A) Overlay of the three PLS calibration models of density at temperatures of 15 °C, 45°C, and 85 °C using GC×GC-TOFMS chromatograms with Samples 20, 24, 25, 46, 47, 48, 50, and 72 excluded. The average of the two PLS modeling results for the two GC×GC-TOFMS replicates is shown. The back line represents the ideal agreement between the predicted and measured values. Samples 66, 4, 1, and 67 from Figure 4.3 (left to right) have been highlighted for each temperature so they can be discussed in the context of their chemical composition, location on the calibration plot, and LRV. The samples have been labeled only for the density at 15 °C model. (B) LRV for PLS prediction of density at 15 °C. (C) LRV for PLS prediction of density at 45 °C. (D) LRV for PLS prediction of density at 85 °C. Blue indicates positive values in the LRV while red indicates negative values in the LRV for the chromatographic variables. Alkanes are located within the dotted line, cycloalkanes are located within the dashed line, and aromatics are located within the solid black line. Although the LRV for the three temperatures appear to be equivalent, PCA in Figure 4.9 reveals that they are different, and change with temperature.

To investigate this aspect further and to gain insight into the compositional dependence of density as a function of temperature, the LRV from each PLS density model (2 replicates × 3 different temperatures) were analyzed using PCA; results are presented in Figure 4.9. In the scores plot (Figure 4.9A), the LRV separate on PC2 (PC1 captures the variance between the two chromatographic replicates), forming a linear trend with respect to temperature. Over a sufficiently short temperature range, density of kerosene-based fuel is essentially linearly dependent on temperature. Inspection of the loadings plot for the PCA in Figure 4.9B reflects what was originally observed in Figure 4.8. Namely, alkanes negatively influence density, cycloalkanes have a moderately positive correlation, and aromatics are positively correlated. However, the loadings plot provides additional insight into the relationship between density and temperature. For each compound class, density increases at a different rate, which is caused by intermolecular forces in the fuel matrix [54].

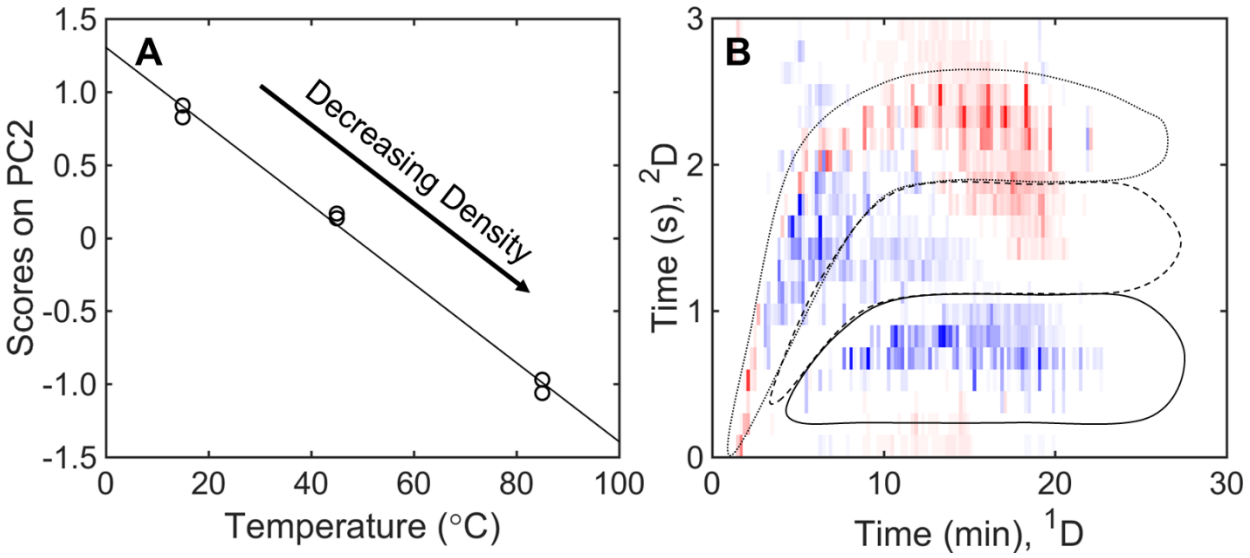


Figure 4.9. PCA modeling on the LRV from the PLS prediction of density at 15, 45 and 85 °C. (A) Scores on PC2 form a linear function with respect to temperature. (B) Loadings plots from the PCA on the LRV of the PLS prediction of density. Blue indicates positive values in the loadings while red indicates negative values in the loadings. Alkanes are located within the dotted line, cycloalkanes are located within the dashed line, and aromatics are located within the solid black line.

To further support our findings determined by PLS with Venetian blinds cross validation (i.e., internal validation), a follow-up pseudo-external validation study was performed in which the remaining sixty-six samples in the data set were split into a calibration set of fifty-five samples (~85%) and a validation set of eleven samples (~15%). The details and results of this study are presented in the Supporting Information (4.4.2, Figure 4.S2 and Table 4.S1). Figure 4.S2 provides a comparison for the modeling of viscosity presented in Figure 4.5. The calibration plot and LRV in Figure 4.5 (internal validation) are visually similar to those in Figure 4.S2 (pseudo-external validation) and yield comparable predicted values and RMSECV, indicating good agreement between the two validation methods. For all models, the root mean square error of prediction (RMSEP) was consistent with the RMSECV for the pseudo-externally calibrated model. Additionally, the RMSECV for the pseudo-externally validated models were slightly higher than the RMSECV for the internally validated models. This is indicative of slightly more robust models when all sixty-six samples were included in the calibration set. Since a goal of this study is to demonstrate proof-of-concept for using raw GC×GC-TOFMS data to predict physical properties of compositionally diverse kerosene-based fuels, an estimate of model performance for samples measured under identical conditions is satisfactory. In lieu of a representative, independent test set and other limitations in experimental design, internal calibration was determined to be sufficient for this purpose.

4.4 SUPPORTING INFORMATION

The Supporting Information contains supplementary information regarding the PLS modeling and identification of outliers. Additionally, a pseudo-external validation of the viscosity model is demonstrated.

4.4.1 *Outliers in modeling viscosity of 74 fuels*

Following PLS modeling, several summary statistics are informative for identifying outliers. The Q residuals statistic indicates how well each sample is modeled, providing a measure of the variation that is not explained by the model. On the other hand, Hotelling's T^2 statistic is a measure of the variation of each sample within the model. Examining a plot of the Q residuals versus Hotelling's T^2 can guide the identification of outliers (samples with high Q residuals and/or Hotelling's T^2 values). Samples that exhibit high Hotelling's T^2 values may skew the model and result in high Q residuals for other samples in the model, which may constitute grounds for removal. Figure 4.S1 shows the Q residuals and Hotelling's T^2 for the PLS modeling of viscosity for all 74 fuels and both GC×GC-TOFMS replicate sets. It can be seen that there is a handful of numbered samples with high Hotelling's T^2 values and/or Q residuals. Through investigation of these plots, we confirmed that all samples identified as having overloaded peaks by visual inspection of chromatograms in Figure C.1 had Hotelling's T^2 reduced values greater than 0.5. This supported our decision to remove the following eight outliers from subsequent modeling: Samples 20, 24, 25, 46, 47, 48, 50, and 72. The remaining samples with relatively high Hotelling's T^2 values were measured chromatographically at a separate point in time (Samples 69, 70, 71, 73, and 74) or displayed markedly different compositional features (Sample 16).

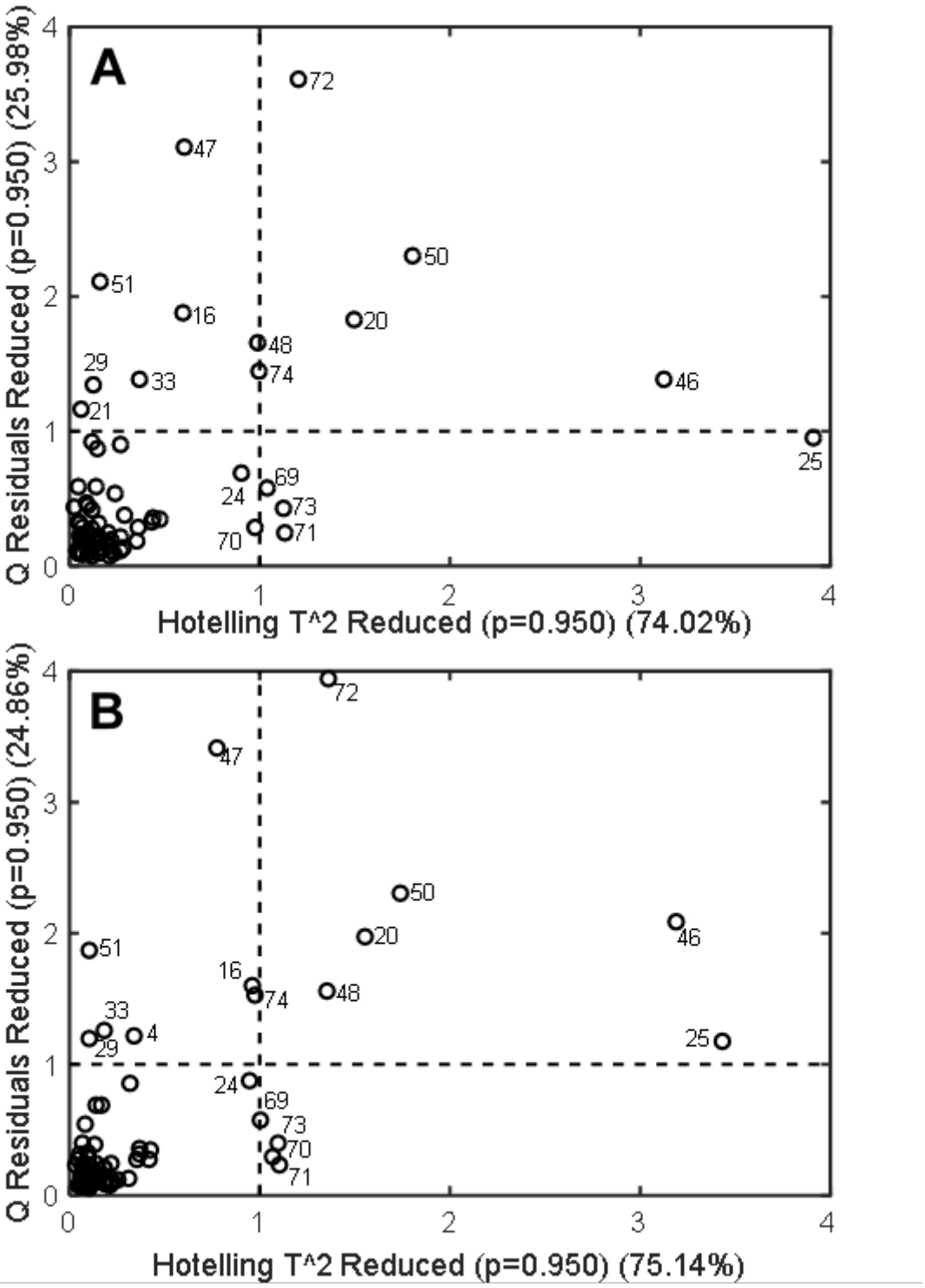


Figure 4.S1. Q residuals vs. Hotelling's T^2 statistic for PLS model of viscosity with all 74 fuels. (A) Model for GC×GC-TOFMS replicate set one. (B) Model for GC×GC-TOFMS replicate set two. Sample labels are shown for samples with borderline to high Q residuals and/or Hotelling's T^2 value. The dashed lines indicate the 95% confidence limits for each statistic.

4.4.2 *Pseudo-external validation*

Following the removal of the eight outlier samples, a pseudo-external validation study was performed in which the sixty-six remaining samples were automatically split into a calibration set and external validation set using the Kennard-Stone algorithm. In the automated calibration/validation splitting, 83% of the samples (fifty-five samples) comprising a uniform distribution over the original sample space were selected by the Kennard-Stone algorithm and kept in the calibration set. The remaining eleven samples were removed from the calibration set and assigned to the validation set (Samples 1, 2, 14, 22, 26, 30, 37, 39, 43, 54, and 65). Venetian blinds cross validation with 11 splits was performed on the calibration data. Five latent variables were used in the modeling to provide a comparison with the internal validation modeling presented in the main paper. The same eleven samples were used as the validation set for all properties modeled. Figure 4.S2 captures the calibration and prediction of the calibration and validation sets, respectively, for viscosity modeled using the pseudo-external validation procedure. The average results from the models corresponding to the two chromatographic replicates are shown. Figures for the remaining properties modeled using pseudo-external validation are omitted for brevity. Table 4.S1 summarizes the results of all models and provides a comparison with models for which internal validation alone (i.e., cross validation) was used. For the pseudo-external validation, RMSEP was consistently lower than the RMSECV, indicating that the model is generalizable. The RMSECV values for internal validation were slightly lower than those for the pseudo-external validation, which is a result of having more samples in the training set for each cross-validation step. Overall, the results show that both approaches perform similarly in terms of estimating the error for similar samples outside of the calibration set that are analyzed using the same instrumentation.

This is a pseudo-external validation because the samples included in the validation set do not comprise an independent test and the size of the set is small. A potential independent test set could be comprised of Samples 69-74, as the chromatographic data was collected six months after the rest of the fuels. However, the chemical composition does not appear to come from the same statistical population as the calibration set (Figure C.1) and five samples (after the removal of Sample 72) is an extremely small size for an external validation set.

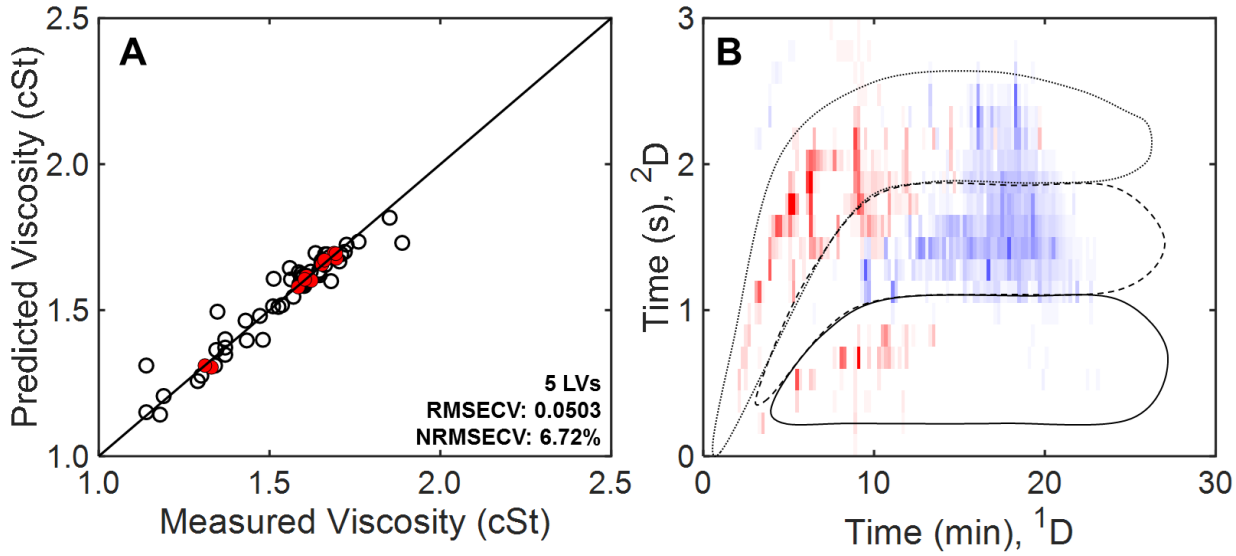


Figure 4.S2. PLS prediction of viscosity for 66 fuels (without outliers) using pseudo-external validation. (A) PLS calibration of viscosity using GC×GC-TOFMS chromatograms of 66 fuel samples using pseudo-external validation. Samples in the external validation set (Samples 1, 2, 14, 22, 26, 30, 37, 39, 43, 54, and 65) have been highlighted red. The average of the two predicted results for the two GC×GC-TOFMS replicates is shown. The black line represents the ideal agreement between the predicted and measured values, with slope = 1. (B) LRV from the PLS prediction of viscosity using 66 fuels. Blue indicates positive values in the LRV while red indicates negative values in the LRV for the chromatographic variables. Alkanes are located within the dotted line, cycloalkanes are located within the dashed line, and aromatics are located within the solid black line.

Table 4.S1. Summary of RMSEC, RMSECV, and RMSEP values for pseudo-external validation and RMSECV values for internal validation.

	Pseudo-external validation			Internal validation
Property	**RMSEC**	**RMSECV**	**RMSEP**	**RMSECV**
Viscosity (cSt)	0.0247	0.0503	0.0132	0.0450
Heat of Combustion (Btu/lbm)	22.8	42.3	32.4	41.3
Hydrogen Content (mass %)	0.0603	0.136	0.0682	0.130
Density 15°C (g/mL)	0.0018	0.0063	0.0018	0.0064
Density 45°C (g/mL)	0.0017	0.0064	0.0018	0.0064
Density 85°C (g/mL)	0.0017	0.0064	0.0017	0.0064

4.5 CONCLUSION

The underlying goal of this research is to gain a fundamental comprehension of the chemical basis for fuel performance through chemometric modeling by relating measured fuel property data to chemical composition data acquired by GC×GC-TOFMS. The predictive PLS modeling approach presented is intended to instruct fuel selection and application-specific fuel engineering. Proof-of-principle demonstration of PLS modeling was performed in the context of four accurately measured properties of relevance to propulsion system operation and design: viscosity, heat of combustion, hydrogen content, and density. With chromatographic and mass spectral data acquired for seventy-four compositionally diverse rocket and gas turbine aviation fuels, reliable PLS models were constructed that mathematically relate the chemical information provided by GC×GC-TOFMS analyses to physical properties. PLS model accuracy was encouraging: respective normalized root mean square errors of cross validation (NRMSECV) of 6.01, 10.3, 8.71, and 7.12% were obtained for viscosity, heat of combustion, hydrogen content, and temperature-dependent density, respectively. For these fuel sample and analytical data sets, model quality is improved compared with previous three-parameter composition-property

relationships based on quantitative hydrocarbon type analysis. Additionally, LRV were shown to relate fuel properties with chemical composition; the qualitative influence of specific hydrocarbon classes on properties was consistent with expected behavior. Eight samples with atypical chemical composition were excluded from the model training sets based on visual inspection supported by statistical analysis. The extension of this methodology to broader compositional diversity (gasoline, diesel fuel, etc.) is expected to be straightforward.

4.6 REFERENCES

[1] D.J. Cookson, J.L. Latten, I.M. Shaw, B.E. Smith, Property-composition relationships for diesel and kerosene fuels, Fuel. 64 (1985) 509–519. https://doi.org/10.1016/0016-2361(85)90086-9.

[2] D.J. Cookson, C.P. Lloyd, B.E. Smith, Investigation of the chemical basis of kerosene (jet fuel) specification properties, Energy Fuels. 1 (1987) 438–447. https://doi.org/10.1021/ef00005a011.

[3] D.J. Cookson, B.E. Smith, Calculation of jet and diesel fuel properties using carbon-13 NMR spectroscopy, Energy Fuels. 4 (1990) 152–156. https://doi.org/10.1021/ef00020a004.

[4] D.J. Cookson, P. Iliopoulos, B.E. Smith, Composition-property relations for jet and diesel fuels of variable boiling range, Fuel. 74 (1995) 70–78. https://doi.org/10.1016/0016-2361(94)P4333-W.

[5] G. Liu, L. Wang, H. Qu, H. Shen, X. Zhang, S. Zhang, Z. Mi, Artificial neural network approaches on composition–property relationships of jet fuels based on GC–MS, Fuel. 86 (2007) 2551–2559. https://doi.org/10.1016/j.fuel.2007.02.023.

[6] M.L. Huber, E.W. Lemmon, T.J. Bruno, Effect of RP-1 Compositional Variability on Thermophysical Properties, Energy Fuels. 23 (2009) 5550–5555. https://doi.org/10.1021/ef900597q.

[7] 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.

[8] R.R. Mallepally, B.A. Bamgbade, M.A. McHugh, H.O. Baled, R.M. Enick, M.C. Billingsley, Measurements and modeling of the density of rocket propellant RP-2 at temperatures to 573 K and pressures to 100 MPa, Fuel. 253 (2019) 1193–1203. https://doi.org/10.1016/j.fuel.2019.05.089.

[9] I.C. Lee, H.C. Ubanyionwu, Determination of sulfur contaminants in military jet fuels, Fuel. 87 (2008) 312–318. https://doi.org/10.1016/j.fuel.2007.05.010.

[10] M.C. Billingsley, J.T. Edwards, L.M. Shafer, T.J. Bruno, Extent and Impacts of Hydrocarbon Fuel Compositional Variability for Aerospace Propulsion Systems, in: Proceedings of the 46th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit,

Nashville, TN, 2010; American Institute of Aeronautics and Astronautics: Reston, VA, 2010. https://doi.org/10.2514/6.2010-6824.
[11] 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.
[12] 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.
[13] 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.
[14] J.L. Burger, T.J. Bruno, Application of the Advanced Distillation Curve Method to the Variability of Jet Fuels, Energy Fuels. 26 (2012) 3661–3671. https://doi.org/10.1021/ef3006178.
[15] 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.
[16] T.J. Bruno, L.S. Ott, T.M. Lovestead, M.L. Huber, The composition-explicit distillation curve technique: Relating chemical analysis and physical properties of complex fluids, J. Chromatogr. A. 1217 (2010) 2703–2715. https://doi.org/10.1016/j.chroma.2009.11.030.
[17] M. Billingsley, N. Keim, R. Synovec, B. Hill-Lam, C. Wilhelm, Progress Toward Quality Assurance Standards for Advanced Hydrocarbon Fuels Based on Thermal Performance Testing and Chemometric Modeling, in: Proceedings of the IASH 14th International Symposium on Stability, Handling, and Use of Liquid Fuels, Charleston, SC, 2015; International Association for Stability, Handling, and Use of Liquid Fuels: Atlanta, GA, 2015.
[18] M. Billingsley, N. Keim, B. Hill-Lam, R. Synovec, Hydrocarbon Fuel Thermal Performance Modeling based on Systematic Measurement and Comprehensive Chromatographic Analysis, in: Proceedings of the 52nd AIAA/SAE/ASEE Joint Propulsion Conference, Salt Lake City, UT, 2016; American Institute of Aeronautics and Astronautics: Reston, VA, 2016. https://doi.org/10.2514/6.2016-4903.
[19] 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.
[20] 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.
[21] 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.
[22] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Comprehensive Two-Dimensional High-Speed Gas Chromatography with Chemometric Analysis, Anal. Chem. 70 (1998) 2796–2804. https://doi.org/10.1021/ac980164m.

[23] 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.
[24] C.G. Fraga, B.J. Prazen, R.E. Synovec, Comprehensive two-dimensional gas chromatography and chemometrics for the high-speed quantitative analysis of aromatic isomers in a jet fuel using the standard addition method and an objective retention time alignment algorithm, Anal. Chem. 72 (2000) 4154–4162. https://doi.org/10.1021/ac000303b.
[25] B. Omais, M. Courtiade, N. Charon, D. Thiébaut, A. Quignard, M.-C. Hennion, Investigating comprehensive two-dimensional gas chromatography conditions to optimize the separation of oxygenated compounds in a direct coal liquefaction middle distillate, J. Chromatogr. A. 1218 (2011) 3233–3240. https://doi.org/10.1016/j.chroma.2010.12.049.
[26] R.C. Striebich, L.M. Shafer, R.K. Adams, Z.J. West, M.J. DeWitt, S. Zabarnick, Hydrocarbon Group-Type Analysis of Petroleum-Derived and Synthetic Fuels Using Two-Dimensional Gas Chromatography, Energy Fuels. 28 (2014) 5696–5706. https://doi.org/10.1021/ef500813x.
[27] 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.
[28] F. Westad, N.K. Afseth, R. Bro, Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares regression, Anal. Chim. Acta. 595 (2007) 323–327. https://doi.org/10.1016/j.aca.2007.02.015.
[29] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A. 1255 (2012) 3–11. https://doi.org/10.1016/j.chroma.2012.05.050.
[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] L.A.F. de Godoy, M.P. Pedroso, L.W. Hantao, R.J. Poppi, F. Augusto, Quantitative analysis by comprehensive two-dimensional gas chromatography using interval Multi-way Partial Least Squares calibration, Talanta. 83 (2011) 1302–1307. https://doi.org/10.1016/j.talanta.2010.08.015.
[32] B.J. Prazen, K.J. Johnson, A. Weber, R.E. Synovec, Two-Dimensional Gas Chromatography and Trilinear Partial Least Squares for the Quantitative Analysis of Aromatic and Naphthene Content in Naphtha, Anal. Chem. 73 (2001) 5677–5682. https://doi.org/10.1021/ac010637g.
[33] 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.
[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] 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.
[36] T. Rajalahti, O.M. Kvalheim, Multivariate data analysis in pharmaceutics: A tutorial review, Int. J. Pharm. 417 (2011) 280–290. https://doi.org/10.1016/j.ijpharm.2011.02.019.
[37] A.A. Gowen, G. Downey, C. Esquerre, C.P. O'Donnell, Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients, J. Chemom. 25 (2011) 375–381. https://doi.org/10.1002/cem.1349.
[38] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.
[39] R.E. Morris, M.H. Hammond, J.A. Cramer, K.J. Johnson, B.C. Giordano, K.E. Kramer, S.L. Rose-Pehrsson, Rapid Fuel Quality Surveillance through Chemometric Modeling of Near-Infrared Spectra, Energy Fuels. 23 (2009) 1610–1618. https://doi.org/10.1021/ef800869t.
[40] J.A. Cramer, M.H. Hammond, K.M. Myers, I.A. Leska, R.E. Morris, Expanded Framework for the Prediction of Alternative Fuel Content and Alternative Fuel Blend Performance Properties Using Near-Infrared Spectroscopic Data, Energy Fuels. 29 (2015) 7026–7035. https://doi.org/10.1021/acs.energyfuels.5b01660.
[41] J.A. Cramer, M.H. Hammond, K.M. Myers, T.N. Loegel, R.E. Morris, Novel Data Abstraction Strategy Utilizing Gas Chromatography–Mass Spectrometry Data for Fuel Property Modeling, Energy Fuels. 28 (2014) 1781–1791. https://doi.org/10.1021/ef4021872.
[42] 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.
[43] ASTM D445-19, "Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity)," ASTM International, West Conshohocken, PA, 2019, www.astm.org.
[44] ASTM D4809-18, "Standard Test Method for Heat of Combustion of Liquid Hydrocarbon Fuels by Bomb Calorimeter (Precision Method)," ASTM International, West Conshohocken, PA, 2018, www.astm.org.
[45] ASTM D7171-16, "Standard Test Method for Hydrogen Content of Middle Distillate Petroleum Products by Low-Resolution Pulsed Nuclear Magnetic Resonance Spectroscopy," ASTM International, West Conshohocken, PA, 2016, www.astm.org.
[46] ASTM D4052-18a, "Standard Test Method for Density, Relative Density, and API Gravity of Liquids by Digital Density Meter," ASTM International, West Conshohocken, PA, 2018, www.astm.org.
[47] R.G. Brereton, Chemometrics for Pattern Recognition, John Wiley & Sons: West Sussex, U.K., 2009; 311-390.
[48] D. Pérez-Guaita, G. Quintás, J. Kuligowski, Discriminant analysis and feature selection in mass spectrometry imaging using constrained repeated random sampling - Cross validation (CORRS-CV), Anal. Chim. Acta. 1097 (2020) 30–36. https://doi.org/10.1016/j.aca.2019.10.039.
[49] Y. Xu, R. Goodacre, On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization

Performance of Supervised Learning, J. Anal. Test. 2 (2018) 249–262. https://doi.org/10.1007/s41664-018-0068-2.
[50] L. Ranzan, C. Ranzan, L.F. Trierweiler, J.O. Trierweiler, Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy, Energy Fuels. 31 (2017) 8942–8950. https://doi.org/10.1021/acs.energyfuels.7b00954.
[51] 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.
[52] S.L. Outcalt, A. Laesecke, K.J. Brumback, Thermophysical Properties Measurements of Rocket Propellants RP-1 and RP-2, J. Propul. Power. 25 (2009) 1032–1040. https://doi.org/10.2514/1.40543.
[53] 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.
[54] J.L. Trenzado, J.S. Matos, L. Segade, E. Carballo, Densities, Viscosities, and Related Properties of Some (Methyl Ester + Alkane) Binary Mixtures in the Temperature Range from 283.15 to 313.15 K, J. Chem. Eng. Data. 46 (2001) 974–983. https://doi.org/10.1021/je0100286.

Chapter 5. Improvements to Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry Composition-Based Models for Kerosene-Based Fuel Thermal Integrity Using Supervised Feature Selection and Partial Least Squares Analysis

5.1 INTRODUCTION

Decomposition of hydrocarbons into carbonaceous deposits ("coking") is a major problem found in the cracking of petroleum products [1–3], pyrolysis of biomass for biofuels [4–6], production of PVC [7,8], and regenerative cooling in rocket and jet engines [9–11]. Formation of carbonaceous deposits is a relatively well understood phenomenon, and it is generally accepted that there are three basic mechanisms governing the formation of carbon deposits [1,3,8,12–18]. These mechanisms include the production of condensation (amorphous) carbon via formation of polycyclic aromatic hydrocarbons (PAHs) in the hydrocarbon stream which condense on the tubing, catalyst, and/or reactor wall surfaces; free radical growth reactions between the hydrocarbon stream and the surface coke; and formation of filamentous carbon via catalytic reactions with surface metals found in the tubing, catalyst, and/or reactor wall surfaces. Several studies using single component model compounds have shown that coking propensity as a function of chemical structure is n-alkanes (n-paraffins) < iso-alkanes < alkenes ≈ cycloalkanes < aromatics < PAH [3,16,19]. Furthermore, the presence of heteroatoms such as those containing sulfur can increase coking rates [1,20].

The use of regenerative cooling in liquid-hydrocarbon-fueled rockets and hypersonic vehicles has been a staple for the aerospace field, wherein fuel flows through tubes or channels of a heat-exchanger in order to act as a heat-sink [12,21–24]. While regenerative cooling may be advantageous to reduce the temperature of the combustion chamber, nozzle of the engine, and/or

other sub-systems, the formation of undesirable carbonaceous deposits compromises system reliability and performance [9–12,25,26]. Traditionally, multicomponent distillate fuels were analyzed for known impurities (e.g. sulfur, olefins, oxygenates, and aromatic compounds) and effects on fuel-wetted surfaces (e.g. copper corrosivity) [27], with additional testing in representative engine components and environments providing full assurance of system reliability. However, this analysis approach can become expensive and time consuming as engine operating environments and the number of candidate fuel formulations increases, and skilled technicians may be required both for testing as well as analysis and interpretation of results. In order to evaluate fuel performance, an ideal approach is to implement a platform that exposes candidate fuels to perturbations in the most relevant environmental conditions occurring in rocket cooling system environments. Investigation into the thermal stability of fuels has taken a three-fold approach: (a) fuel composition [28,29], (b) cooling surface substrates (i.e. copper, stainless steel, etc.) [29–32], and (c) testing conditions (i.e. pressure and flow rates) [33].

It is important to note that the influence of certain types of compounds on fuel thermal performance is widely accepted. For example, in rocket grade kerosene, aromatic, olefinic, and sulfur-containing compounds are restricted to relatively low levels due to their detrimental influence on cooling performance and other engine behavior [34–38]. However, questions regarding existing specifications must be addressed before candidate fuels are adopted for use in systems imposing high levels of thermal stress and surface reactivity. These include assessment of (a) the validity of current test methods controlling deleterious compounds, (b) the validity of current maximum allowable levels for contaminants and unsaturated species, (c) specific types of compounds in a given class or family, including their relative detriment to cooling performance,

and (d) other compounds not governed by specification limits but suggested to influence fuel reactivity, decomposition, and deposit formation.

Gas chromatography (GC) is a traditional analytical technique that is amenable for the separation and analysis of volatile and semi-volatile mixtures [39,40]. When GC is coupled with mass spectrometry (MS), spectral information can be gathered allowing for further selectivity and the ability to identify chemical compounds. GC-MS has been shown to be a powerful tool for the analysis of kerosene-based fuel [41–46]. Comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC×GC–TOFMS) can further improve upon the separation power of one dimensional GC and provide additional insight into complex mixtures of volatile compounds such as those present in kerosene-based fuels [47–52]. Traditionally in GC×GC, the first separation dimension uses non-polar stationary phase while the second dimension uses a polar stationary phase. However, a reverse-column configuration (polar first dimension and non-polar second dimension) has been shown to provide better selectivity for petroleum based samples [47–49,53]. A potential drawback in the application of GC×GC–TOFMS is the sheer magnitude of information that is produced in a single analysis of fuels. This data analysis challenge is compounded by injection replicates for a large set of fuels. Yet, the challenge of obtaining meaningful information can be overcome through application of powerful chemometric software methods that can aid in the interpretation of such complex data sets.

Partial least squares (PLS) analysis is a chemometric method that associates the differences in measurable information for two different data sets. For example, PLS has been used to associate the chemical information obtained from GC×GC–TOFMS and GC×GC-FID to physical properties of kerosene-based fuels [47–49,53]. Details about PLS can be found elsewhere [54–57]. Briefly, PLS analysis mathematically relates, via linear algebra, two data matrices (X and Y block) through

extraction of factors referred to as latent variables (LVs). Using PLS, in the study herein, models are constructed to account for the variance (ideally, the relevant chemical differences) in both the GC×GC–TOFMS data for a fuel sample set, i.e., the signal intensities (which constitute the X-block) and the respective measured property values, for the same fuel sample set (which constitute the Y-block). PLS yields two important outcomes: (1) a linear correspondence of the chemical/physical properties to the GC×GC–TOFMS data, which can subsequently be used to predict chemical/physical properties without having to directly measure these properties in new samples, and (2) the underlying relationship between the chemical composition of the samples and the predicted chemical/physical measurements, provided by the linear regression vectors (LRVs). In order to achieve reliable, robust PLS modeling, as well as determine the correct number of latent variables, cross validation is performed [49,53,58–62].

Herein, we use a fuel set of 38 kerosene-based fuels (i.e., jet and rocket fuels, with a focus on RP-1 and RP-2 fuels) in order to gain further insight into the relationship between fuel properties and chemical composition. The goal of this study is to dig more deeply into the application-specific performance of these fuels as gauged by thermal integrity in the CRAFTI (Compact Rapid Assessment of Fuel Thermal Integrity) experimental platform [63–66]. CRAFTI is a laboratory scale experiment providing quantitative data under conditions relevant to rocket regenerative cooling systems. Briefly, fuel flows through a copper test article, a section of which is resistively heated (the heated zone) to temperatures that promote the onset of fuel chemical degradation. 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. Three generally accepted mechanisms give rise to the formation of amorphous carbon, chemisorbed carbon, and filamentous carbon deposits

[1,3,8,12–18]. The physical metrics obtained from CRAFTI analysis and the subsequent temperature-programmed oxidation of carbonaceous test article deposits were (a) the change in test article pressure drop as function of time, and (b) the carbon deposition in mass as a function of test article position. Pressure change in the test article has been identified as a critical physical property since it may be indicative of significant deposition leading to local wall overheating and, in the worst case, catastrophic failure of a rocket engine. A major goal of this research is to determine if there are informative connections between the chemical information provided by GC×GC–TOFMS and the measured physical characteristics occurring during the complex process of fuel thermal stressing in the CRAFTI apparatus.

5.2 EXPERIMENTAL

5.2.1 *GC×GC Analysis*

To investigate the chemical composition of the fuels, a GC×GC–TOFMS instrumental platform was used, consisting of an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA), a thermal modulator (4D upgrade, LECO, St. Joseph, MI, USA), and a Pegasus III TOFMS (LECO, St. Joseph, MI, USA). Aliquots of the fuel samples were introduced to the GC×GC–TOFMS instrument via a 7683B auto-injector (Agilent Technologies, Palo Alto, CA, USA). The auto-injector was set to inject 1 µL of sample at a 200:1 split at an inlet temperature of 275 °C. Prior to injection, HPLC grade acetone and hexane (Fisher Scientific) were used as solvent rinses. The primary GC×GC column (^{1}D) was a Rxi-17Sil MS: 29.5 m × 250 µm inner diameter (i.d.) × 0.25 µm film thickness, and the secondary GC×GC column (^{2}D) was a Rxi-1MS 1.5 m × 180 µm I.D. × 0.18 µm film thickness. Ultrahigh purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) was used as the carrier gas at a constant flow rate of 2.0 ml/min. The primary ^{1}D oven was held at 40 °C for 1.5 min before being ramped at 5 °C/min to 200 °C where it was held for 1 min.

The secondary oven was held at a +12 °C offset relative to the primary oven and the modulator block was held at a +30 °C offset to the primary oven. The modulation period was 3 s (separation run time of the ^{2}D column) with 0.75 s hot and cold pulses for each stage. The transfer line was set to 285 °C and the ion source was 225 °C. Mass channels, *m/z*, 35-334 at unit resolution were collected with an ionization voltage of 70 eV at 100 spectra/s after a 10 s acquisition delay. Two replicates were collected for each fuel. A summary of the fuels is provided in Table 5.1. Note that while 38 total fuels were analyzed by the CRAFTI platform, only 34 fuels were included in the modeling for this report. Reasons for excluding fuels were the presence of outliers and replicate samples. The included fuels are indicated in the second column of Table 5.1, and will be referred to by those sample numbers throughout this chapter.

5.2.2 *Fuel Thermal Performance in CRAFTI Apparatus*

The thermal stability of thirty-eight fuels was assessed using the CRAFTI experimental platform at standard run conditions, complemented by subsequent analysis of the types and amount of carbonaceous deposits via temperature-programmed oxidation (TPO) using an RC612 Carbon Determinator (LECO Corporation). During the course of the experiment, the pressure drop ΔP, the difference between the test article inlet and outlet pressure, was measured as a function of time. The pressure data for repeat experimental tests were averaged together to obtain a single ΔP versus time vector for each fuel. To account for the variation in starting (run time = 0) ΔP values for each fuel, we determined the *change* in ΔP over the time period of 300 to 900 s, or Δ(ΔP), defined as ΔP at 900 s minus the ΔP at 300 s. Note that the ΔP values used to calculate Δ(ΔP) are the average of 10 s intervals surrounding 300 s, and just prior to and including 900s. Following each 900 s duration thermal integrity test, the test article was cut into sections for TPO analysis. In previous reports, carbon deposition was reported in counts as obtained directly from the Carbon

Determinator. To provide a more meaningful measurement, the raw electronic signal corresponding to the measurement of CO_2 was converted to carbon mass via calibration. For each type of carbon (determined by several time ranges during which CO_2 was detected in the TPO analysis), the mass was normalized by the area of the corresponding test article section to yield mass/area for each section of the test article. For each run, the normalized mass/area was calculated for each zone (i.e., heated and exit) by calculating the total mass deposited in the zone and dividing by the total area of the sections that comprise the zone. Multiple runs were averaged to provide a single value for each sample. Table 5.1 provides ΔP values at 900 s, and $\Delta(\Delta P)$ values, along with the deposit mass of three types of carbon (chemisorbed, amorphous, and filamentous) measured in the heated and exit zones of the test article via TPO analysis, respectively. Table 5.2 provides the same information with the carbonaceous deposit reported in counts as in previous reports for comparison [63–66]. Additional details on experimental conditions, instrumentation, and test article analysis can be found in our previous reports [63–66].

Table 5.1. Summary of the fuel properties measured via CRAFTI and the LECO RC612 Carbon Determinator (mass). The change in pressure ΔP, at 900 s, and Δ(ΔP), the difference of ΔP at 900 s and 300 s, is measured in psi, and carbon measurements are provided in mass per unit area (μg/cm^2).

Sample Number	Sample Number (Report)	ΔP	Δ(ΔP)	Chemisorbed (Heated)	Chemisorbed (Exit)	Amorphous (Heated)	Amorphous (Exit)	Filamentous (Heated)	Filamentous (Exit)
1	1	61	2.7	5.9	13.6	11.7	45.9	2.2	3.4
2	2	139	54.3	397.2	16.7	33.2	106.1	6.7	4.3
3	3	68	4.0	4.7	9.7	14.4	58.0	4.0	3.7
4	4	62	3.1	3.7	7.1	6.7	24.2	1.5	2.0
5	5	125	40.7	20.3	7.2	15.9	107.1	1.9	3.7
6	6	65	5.0	3.7	8.2	7.0	27.5	2.0	3.6
7	7	60	3.4	-0.9	4.8	7.4	28.1	1.3	2.7
8	8	117	39.5	53.1	14.6	28.3	151.3	6.9	5.4
9	9	56	3.6	1.1	7.1	8.2	40.2	1.9	2.8
10	10	57	3.5	-0.1	9.1	10.7	57.2	1.9	4.0
11	11	62	4.4	5.5	14.0	14.6	41.7	2.3	4.7
12	12	60	3.5	6.0	13.6	12.7	32.2	2.4	3.7
13	13	55	1.4	6.1	13.0	83.4	96.9	241.0	3.9
14	14	68	6.8	8.6	13.1	16.6	83.3	4.2	4.8
15	15	80	14.2	20.5	12.8	11.8	65.0	2.5	3.7
16	16	76	8.0	7.0	6.6	9.7	41.3	2.5	2.4
17	17	142	65.9	9.3	15.4	26.5	208.0	0.2	1.9
18	-			1294.0	43.4	67.1	287.9	43.5	4.0
19	18	57	4.4	4.6	8.6	8.3	42.5	3.3	4.3
20	19	55	3.9	3.5	9.5	20.7	41.8	87.5	2.8
21	-	65	3.3	12.1	14.3	24.5	58.0	351.7	3.5
22	-	59	2.8	3.6	11.2	19.7	73.6	95.1	4.4
23	20	52	3.8	3.2	7.7	6.9	27.4	2.1	2.9
24	21	56	4.7	6.6	15.4	44.7	89.6	7.5	4.1
25	22	57	3.3	3.3	8.6	9.0	41.2	2.5	4.5
26	-			305.5	17.9	69.4	81.8	103.7	4.1
27	23	85	14.0	34.7	22.6	33.7	187.6	11.7	15.4
28	24	58	3.6	5.1	15.0	6.7	45.7	1.2	-1.9
29	25	65	4.7	3.3	7.5	4.4	43.2	1.5	4.2
30	26	60	3.7	5.6	11.7	11.7	69.8	3.1	3.7
31	27	65	4.7	6.7	11.5	13.0	72.8	3.2	4.2
32	28	68	4.5	8.6	12.2	13.0	76.0	2.9	3.4
33	29	64	3.2	4.2	7.5	9.4	35.0	3.3	3.4
34	30	61	2.1	4.1	11.2	172.4	53.4	47.7	3.9
35	31	65	4.4	6.3	12.9	12.5	97.0	2.5	4.1
36	32	63	3.9	5.3	9.7	6.7	61.1	1.4	2.2
37	33	58	4.9	2.7	7.6	8.4	31.7	2.8	2.1
38	34	86	14.3	10.3	24.2	42.1	225.7	6.9	32.2

Table 5.2. Summary of the fuel properties measured via CRAFTI and the LECO RC612 Carbon Determinator (counts). The change in pressure ΔP, at 900 s, and Δ(ΔP), the difference of ΔP at 900 s and 300 s, is measured in psi, and carbon measurements are provided in counts.

Sample Number	Sample Number (Report)	ΔP	Δ(ΔP)	Chemisorbed (Heated)	Chemisorbed (Exit)	Amorphous (Heated)	Amorphous (Exit)	Filamentous (Heated)	Filamentous (Exit)
1	1	61	2.7	11132	22324	18692	61411	3761	6234
2	2	139	54.3	221919	26964	37868	124977	8557	6593
3	3	68	4.0	10183	18488	20504	78314	5076	5038
4	4	62	3.1	9933	18169	15314	50304	3425	5155
5	5	125	40.7	32812	18182	27719	195672	3948	8074
6	6	65	5.0	7004	14275	10510	35407	2458	4080
7	7	60	3.4	6940	15767	11129	37951	2128	4190
8	8	117	39.5	56238	22605	31728	185758	7307	7683
9	9	56	3.6	7279	16072	11896	53411	2462	4042
10	10	57	3.5	7935	20909	13576	71105	2293	5159
11	11	62	4.4	8356	17601	15792	42855	2684	5327
12	12	60	3.5	9606	20187	16469	41663	3060	5122
13	13	55	1.4	9207	19371	80050	114791	217066	5321
14	14	68	6.8	10302	16170	18106	81679	4875	5021
15	15	80	14.2	20503	18060	13995	72899	2459	4297
16	16	76	8.0	10248	12408	13008	50925	3246	3910
17	17	142	65.9	12149	21712	33347	236008	3162	5742
18	-								
19	18	57	4.4	7835	14196	10587	51332	2140	3700
20	19	55	3.9	8583	19153	29843	66772	101423	5431
21	-	65	3.3	25828	33114	39461	99501	634837	6045
22	-	59	2.8	8466	21518	27106	116794	131354	5955
23	20	52	3.8	8689	17831	13912	49201	3074	4985
24	21	56	4.7	9979	22646	44914	107655	6705	4467
25	22	57	3.3	7386	13319	11944	39685	2478	4612
26	-								
27	23	85	14.0	29865	25369	27150	179473	3560	7695
28	24	58	3.6	10538	24095	16430	64247	6896	5030
29	25	65	4.7	11614	25561	31869	185521	4205	25103
30	26	60	3.7	8811	15271	9228	51076	961	3948
31	27	65	4.7	8777	16638	12523	70307	3032	4249
32	28	68	4.5	9153	16228	12894	74876	2108	3611
33	29	64	3.2	10232	16236	13512	77808	2307	3306
34	30	61	2.1	8945	14684	11890	42479	2270	2898
35	31	65	4.4	9487	17693	114567	49673	29524	3000
36	32	63	3.9	9064	16388	13693	83661	2123	3931
37	33	58	4.9	7200	15763	14381	49580	3998	3849
38	34	86	14.3	9770	16170	12238	67564	1613	2839

5.2.3 *Data Analysis*

The GC×GC–TOFMS chromatograms were imported into MATLAB 2015b (MathWorks, Natick, MA) using an in-house converting algorithm. The data were baseline correction and binned 3 modulations (9 s) on the ^{1}D separation dimension and 200 ms (20 mass spectral scans) on the ^{2}D separation dimension that facilitated the definition of tiled GC×GC–TOFMS data. The binning resulted in a total of 200 tile intervals on the ^{1}D dimension by 15 tile intervals on the ^{2}D dimension by 250 *m/z*, or a grand total of 750,000 tiles by *m/z* (200×15×250). Binning within each tile (summing the data per *m/z*) was performed to minimize retention time misalignment and reduce PLS modeling computation time [67]. After the preprocessing steps, the chromatographic data was forwarded to PLS Toolbox 8.61 (Eigenvector Research Inc., Wenatchee, WA, USA), where it was mean-centered. PLS analysis was performed on both replicates of the thirty-four samples in Table 5.1, using Venetian blinds cross validation with 10 splits and a blind thickness of 2 (i.e. grouping together the replicates for each fuel). In order to determine the "goodness of fit" of the PLS models, the root mean square error of cross validation (RMSECV) was calculated as defined by

$$RMSECV = \left[\frac{1}{N} \times \sum\left(y_{i,cv} - y_{i,meas}\right)^2\right]^{0.5} \quad (5.1)$$

A feature selection technique was employed to reduce modeling errors, in which binned signal per tile and *m/z* were regressed against the fuel properties using LOOCV and significant tiles were selected based on a requirement of having at least three *m/z* below a user-selected NRMSECV threshold in the tile [65].

5.3 RESULTS AND DISCUSSION

5.3.1 *Chemical Composition*

Using GC×GC–TOFMS in a reverse column configuration, an excellent two-dimensional (2D) chromatographic separation of the compound classes (alkanes, cycloalkanes, and aromatics) was achieved. Figure 5.1 shows the 2D separation of two representative fuels: Sample 3: GRC RP-1 and Sample 8: LB073009-08. The alkanes are located from 2.0-3.0 s on the ^{2}D dimension, the cycloalkanes are located from 1.0-2.0 s on the ^{2}D dimension, and the aromatics are located from 0-1.0 s on ^{2}D. In each chromatogram, one can visually discern subclasses for the cycloalkanes and aromatic classes (i.e. monocyclics, dicyclics, tricyclics, mono-aromatics, and di-aromatics). In principle, every hydrocarbon class (indeed every hydrocarbon) could be identified and quantified (utilizing retention indices and mass-to-charge ratios, *m/z*). However, in order to create compositional descriptions of the fuels, we have elected to approach the chemical analysis of the fuels using a PLS-based approach. The identification of the general elution times for the three classes (alkanes, cycloalkanes, aromatics) in the 2D chromatographic space serves a more instructive purpose of correlating general fuel chemical composition with differences in measured performance behavior. The 34 fuels exhibit a large composition range leading to their various physical properties and thermal stability.

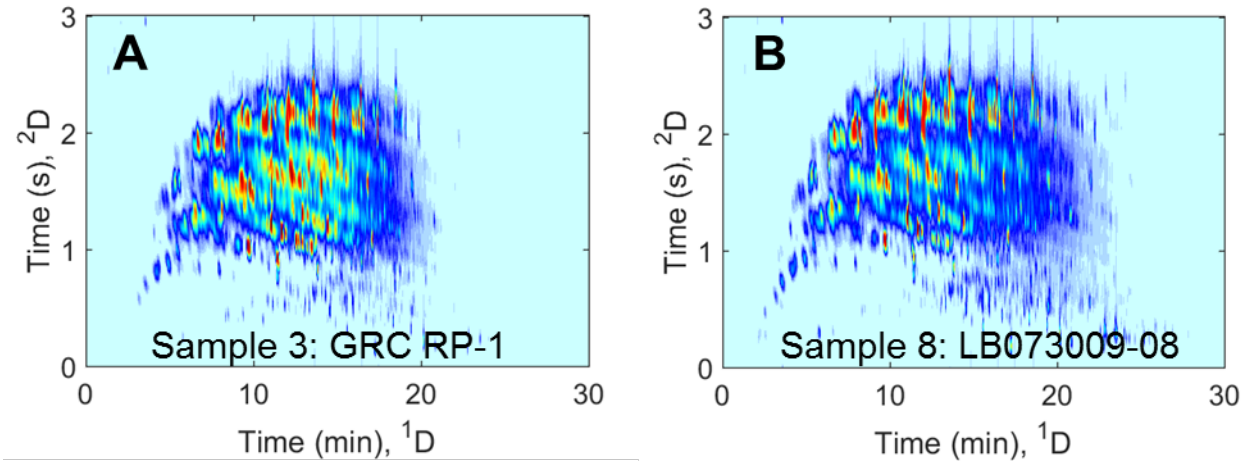


Figure 5.1. Two total ion current (TIC) GC×GC chromatograms representing examples of well-behaving fuels and poorly behaving fuels. (A) Sample 3: GRC RP-1. (B) Sample 8: LB073009-08.

5.3.2 *Fuel Thermal Performance*

This time region was chosen because the ΔP of the well-behaving fuels reaches pseudo-steady state at roughly 300 s, producing Δ(ΔP) values approaching 0. Alternatively, ΔP of the poorly-behaving fuels continues to increase, resulting in higher Δ(ΔP) values. Figure 5.2 shows CRAFTI test article pressure drop time history of the thirty-four fuels. The relationship between Δ(ΔP) and ΔP measured at 900 s as reported in the previous report [66] is provided in Figure 5.3. The data in Figure 5.3 indicates that four fuels have a relatively large Δ(ΔP), Samples 2, 5, 8, and 17, with an additional four fuels that have a moderate Δ(ΔP) relative to the rest of the fuels that have relatively small Δ(ΔP).

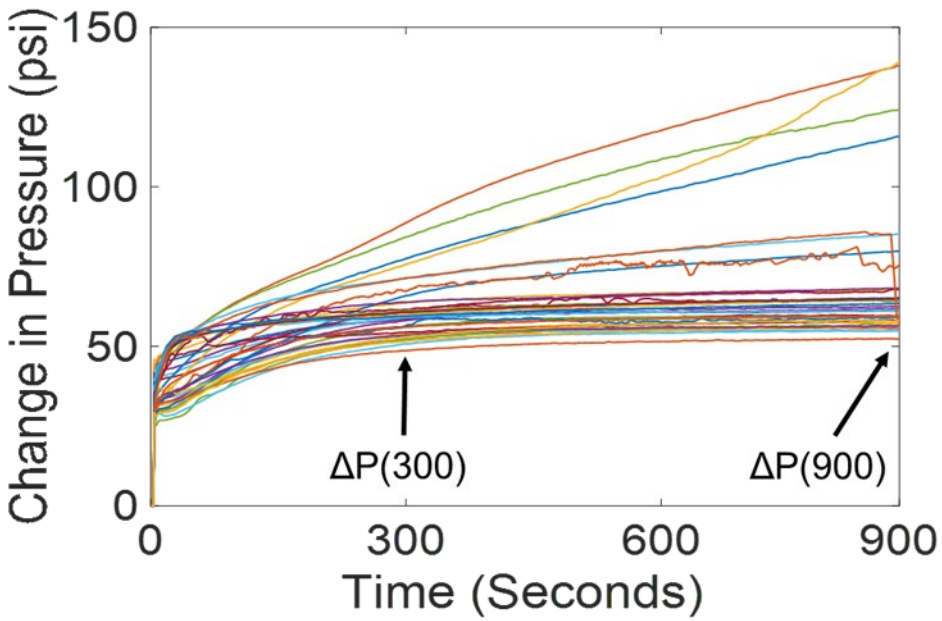


Figure 5.2. Pressure vectors for the fuels obtained from the CRAFTI analysis. Pressure drop as a function of time is shown, where ΔP is defined as the pressure of the inlet minus pressure of the outlet. A corrected change in pressure, Δ(ΔP), was calculated by taking the ΔP at 900 s minus ΔP at 300 s to account for variation in starting ΔP between fuels.

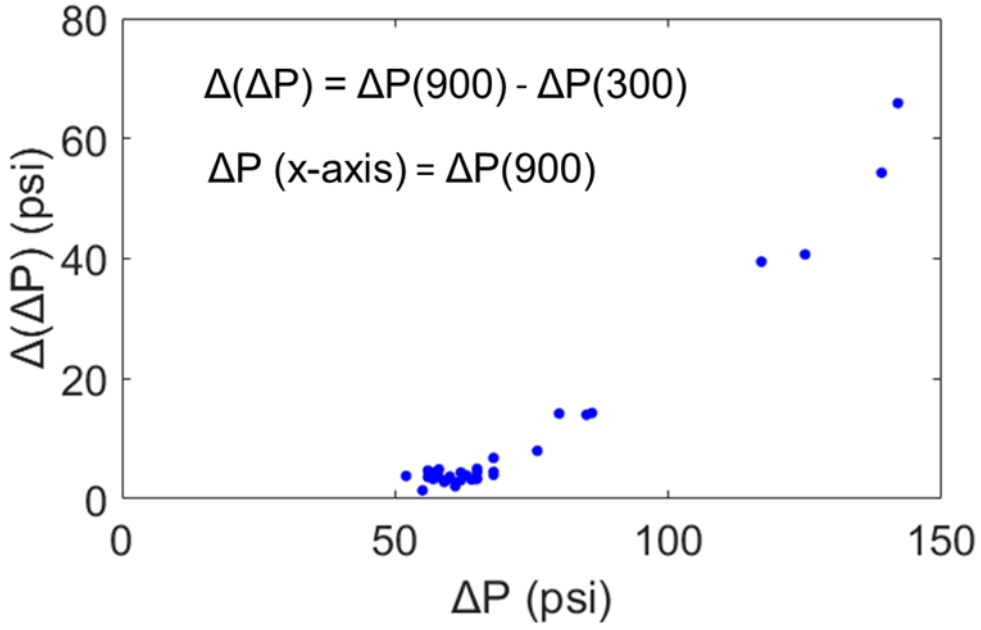


Figure 5.3. Relationship between Δ(ΔP) and ΔP at 900 s.

Figure 5.4 shows the 2D TPO data, detailing the deposition of the various types of carbon and location for two different fuels. Sample 3: GRC RP-1 deposits a relatively small amount of chemisorbed carbon in the heated zone and modest amount of amorphous carbon in the exit zone, while Sample 8: LB073007-08 deposits a relatively significant amount of chemisorbed carbon, which was concentrated in the heated zone, and more amorphous carbon in the exit zone. Sample 3 is representative of fuels that perform "well" in terms of yielding a low Δ(ΔP) and depositing a limited amount of carbon, while Sample 8 is representative of fuels that perform "poorly" yielding a large Δ(ΔP) and depositing relatively high amounts of carbon.

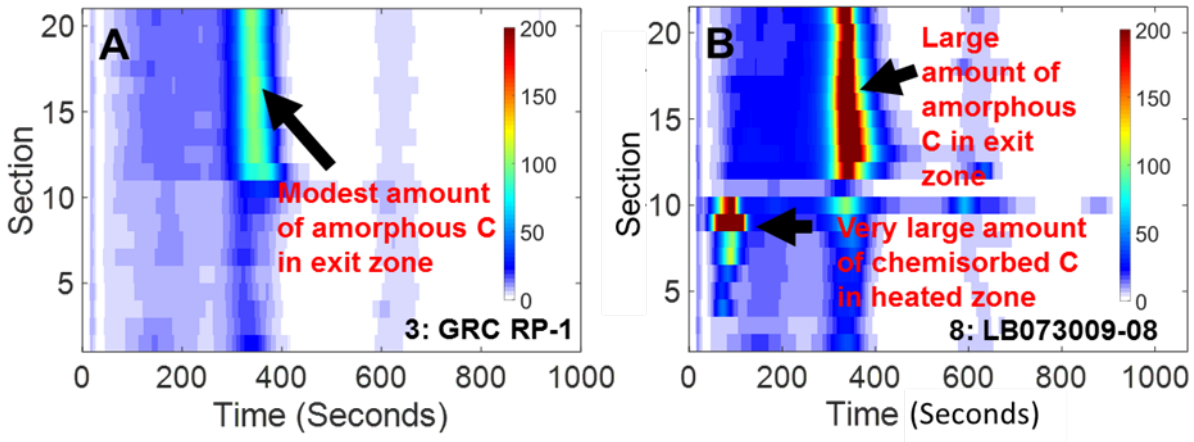


Figure 5.4. Instrument data from Temperature-Programmed Oxidation (TPO) is shown for two fuels that deposit different amounts of carbon. (A) Carbon deposition for fuel GRC RP-1 is representative of fuels that exhibit low pressure changes. These fuels deposit a relatively small amount of chemisorbed and amorphous carbon. (B) Carbon deposition for fuel LB073007-08 is representative of fuels that perform poorly. These fuels deposit a relatively large amount of chemisorbed carbon in the heated zone (sections 4-11) as well as more amorphous carbon in the exit zone (12-21).

Figure 5.5 shows the carbon deposition data of amorphous carbon and chemisorbed carbon for Sample 3 and Sample 8 in a slightly different form. Amorphous carbon deposits in each section of the test article for Sample 3 (Figure 5.5A) and Sample 8 (Figure 5.5B) corroborates the observation that Sample 8 has significantly more amorphous carbon in the exit zone. Similarly, depictions of

chemisorbed carbon deposits for Sample 3 (Figure 5.5C) and Sample 8 (Figure 5.5D) illustrate that Sample 8 deposits a large amount of chemisorbed carbon in the heated zone, specifically concentrated in segments 9 and 10.

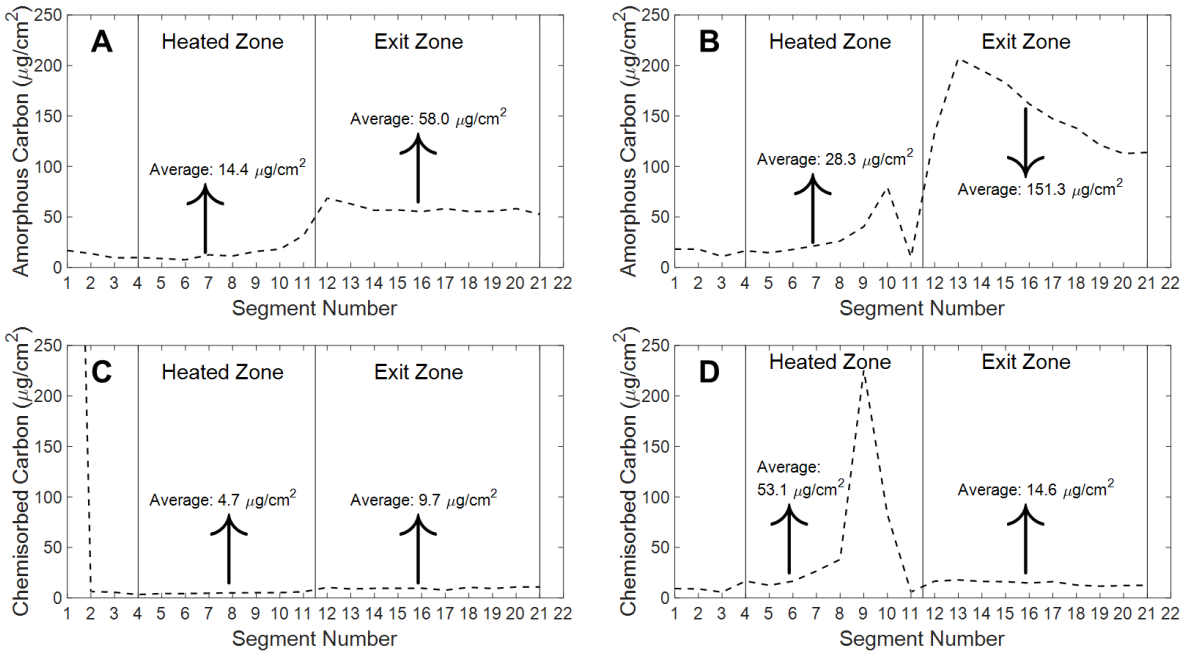


Figure 5.5. Amorphous carbon and chemisorbed carbon deposits per section in the test article shown for Sample 3: GRC RP-1 and Sample 8: LB073009-08. (A) Amorphous carbon deposit for Sample 3. (B) Amorphous carbon deposit for Sample 8. (C) Chemisorbed carbon deposit for Sample 3. (D) Chemisorbed carbon deposit for Sample 8.

Initially, PLS modeling was used to predict the Δ(ΔP) from the CRAFTI study based upon the entire GC×GC–TOFMS chromatograms of the thirty-four fuels as shown in Figure 5.6. The PLS model in Figure 5.6 had a NRMSECV of 19.4% but there were several fuels whose Δ(ΔP) was predicted to be rather high when their measured Δ(ΔP) was actually much lower. The PLS modeling using the entire GC×GC–TOFMS chromatograms for the various forms of carbon, e.g., logarithmically scaled chemisorbed carbon in the heated zone and amorphous carbon in the exit zone, had similar shortcomings [65], so a tile-based feature selection method was implemented to improve modeling of Δ(ΔP) and the various forms of carbon deposition by providing the statistically significant subset of the GC×GC–TOFMS chromatographic data for each PLS model.

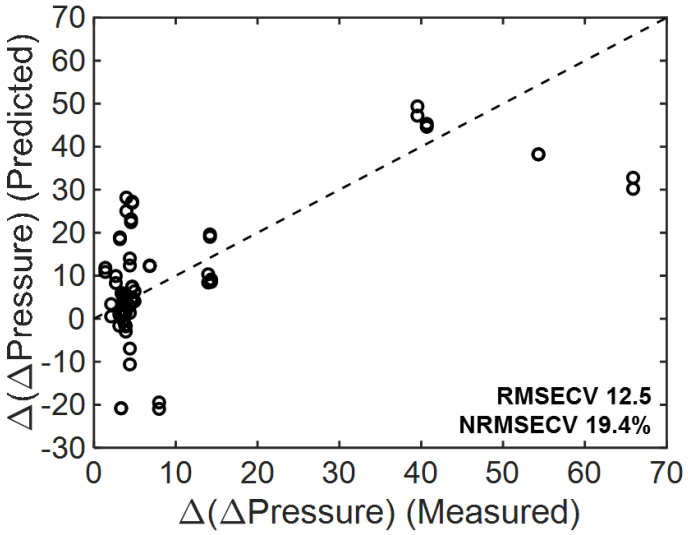


Figure 5.6. PLS prediction of Δ(ΔP) using entire GC×GC–TOFMS chromatograms of 34 samples (100% of data). The black dashed line represents the ideal agreement between the predicted and measured values. The model yields an error of ~20%, but more troublesome is that some fuels have relatively high predicted Δ(ΔP) when their measured Δ(ΔP) is low while other fuels exhibit negative predicted Δ(ΔP) when Δ(ΔP) is always positive.

Figure 5.7A shows the relevant subset of the feature selection results from the LOOCV regressions of the Δ(ΔP) data. As the slope approaches unity, the NRMSECV approaches a minimum value, which is to be expected. A threshold NRMSECV of 13% was chosen as this yielded an acceptable PLS model for Δ(ΔP) concurrent with having sufficient chemical features to lead to a robust PLS model. Figure 5.7B and C show the analytical ion chromatogram (AIC) of Sample 8 for the tiles identified by the LOOCV regressions for the Δ(ΔP) at a NRMSECV threshold of 13%. All of the tiles identified are located within the aromatic region of the chromatogram with a majority found within the di-aromatic region that includes naphthalene derivatives, and heteroatom containing (sulfur) compounds. The feature selection resulted in a large reduction of data with only 0.13% of the data identified across 83 tiles correlating reasonably well with Δ(ΔP). This result highlights the disproportionate importance of relatively low concentration compounds and classes on application-specific thermal integrity as measured by cooling channel pressure drop increase. Figure 5.7D shows the subsequent PLS model based on those selected tiles and *m/z*. The resulting

PLS NRMSECV error is 11.3%, which is significantly lower than the non-feature selection PLS model (cf. Figure 5.6); there are no significant sample outliers. The feature selection results in a large reduction in the data while focusing the modeling on chemical features that are important.

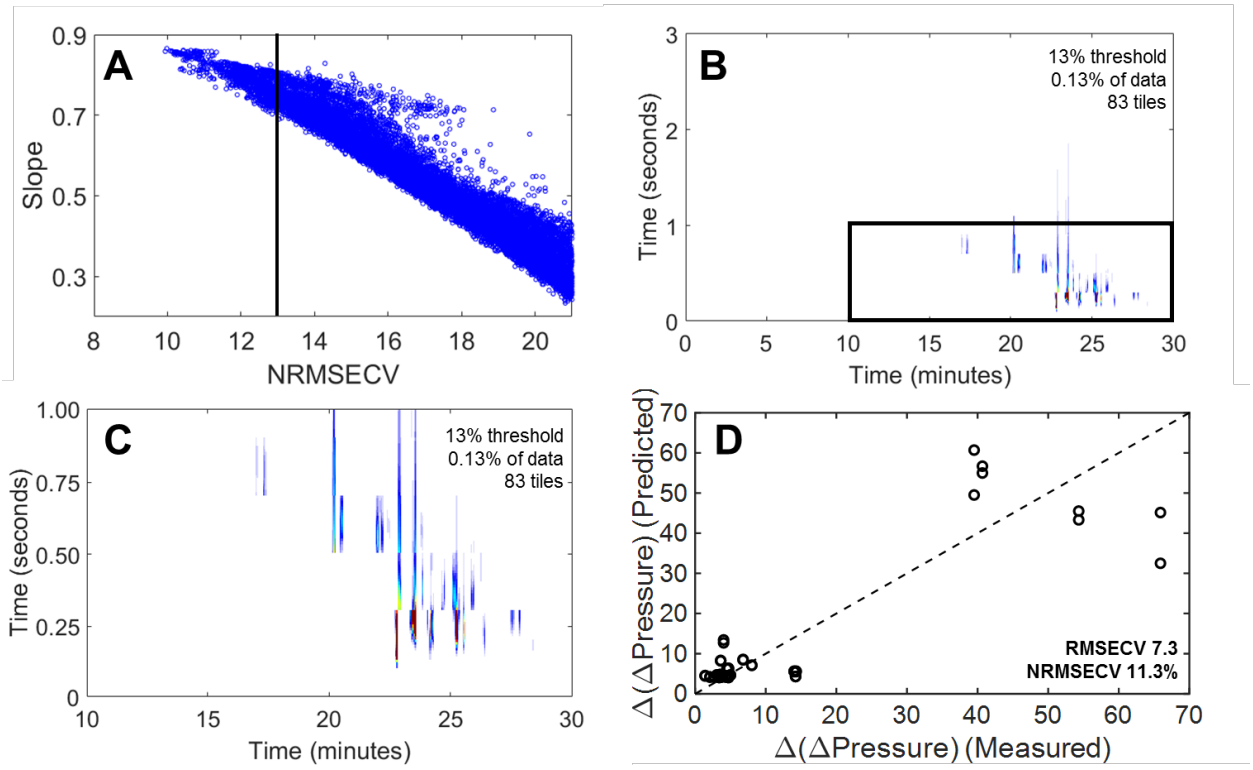


Figure 5.7. Supervised feature selection and subsequent PLS prediction of Δ(ΔP). (A) Shown are a subset of the LOOCV regressions for the Δ(ΔP) data. (B) The LOOCV regressions for the change in pressure from each tile per *m/z* which had a NRMSECV below 13% and contained at least 3 *m/z* were used to create an analytical ion current (AIC) chromatogram using Sample 8 to highlight the tiles identified. A total of 83 tiles were identified which were predominantly located within the aromatic compound and heteroatom containing compound portion of the chromatogram. (C) An enhanced zoom-in view of (B) is shown. (D) PLS prediction of the Δ(ΔP) was performed using the tile locations and their respective *m/z* which were below the NRMSECV threshold of 13%.

Figure 5.8 demonstrates the feature selection method by comparing the TIC and AIC of Samples 3 and 8. Recall that Sample 3: GRC RP-1 is representative of a well-behaving fuel while Sample 8: LB073009-08 is representative of a poorly behaving fuel. Although Samples 3 and 8 both contain aromatics, the AIC chromatograms indicate that Sample 8 is relatively abundant in

compounds responsible for the large Δ(ΔP) while Sample 3 possesses fewer of those compounds and thus has a low Δ(ΔP). Comparison of Sample 3 and Sample 8 total sulfur (6.2 and 33 mg/kg, respectively, measured in accordance with ASTM D5453) and aromatic content (1.3 and 3.7 mass%, respectively, measured using GC×GC-FID/MS) in light of the differences in Δ(ΔP) (4.0 and 39.5, respectively, cf. Table 5.1) is instructive: the approximate ten-fold increase in Δ(ΔP) for Sample 8 is likely attributable primarily to aromatic and sulfur-containing compounds present in the chromatographic region shown in Figure 5.8. Hence, a correlation is provided between Δ(ΔP) and chemical composition within the context of the 2D chromatogram.

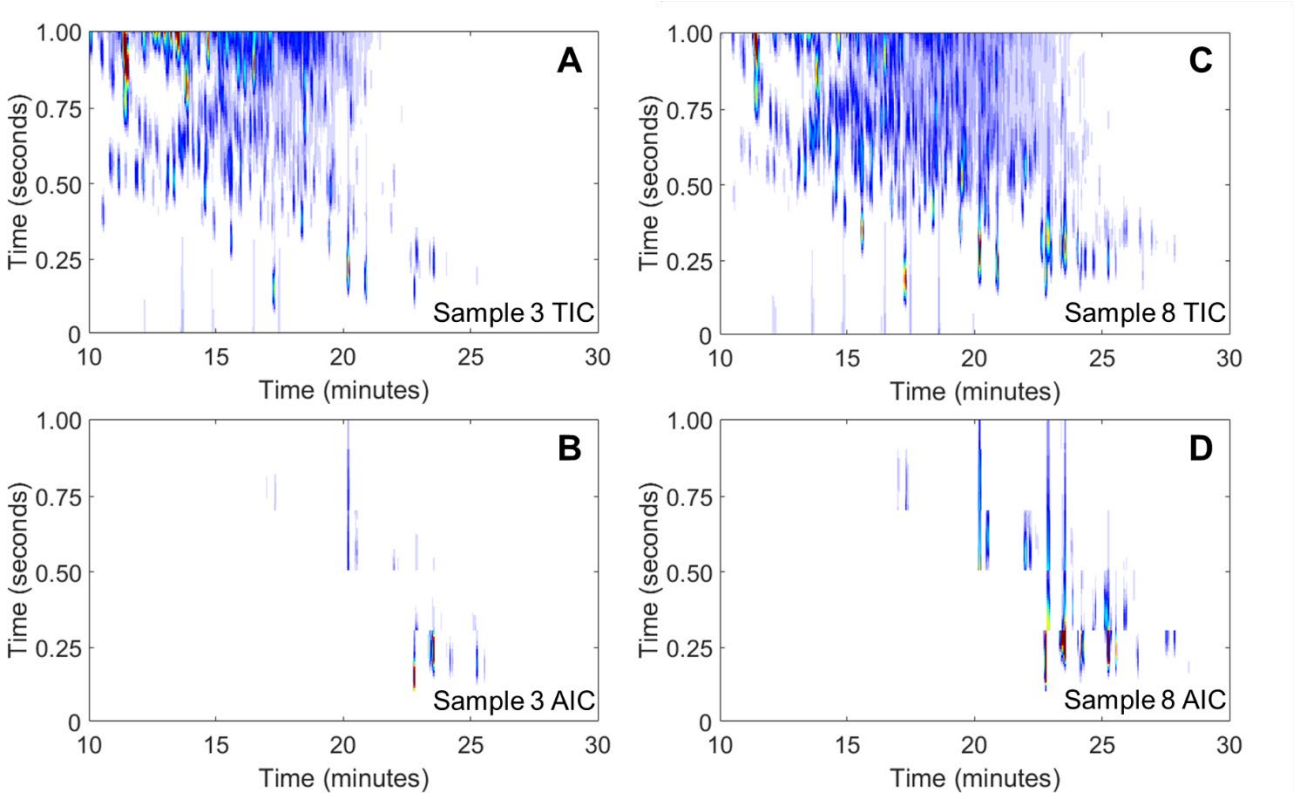


Figure 5.8. Chromatographic comparison of a well-behaving fuel and a poorly behaving fuel. (A) TIC (sum of all *m/z*) for Sample 3 showing the chromatographic region where aromatic compounds elute. (B) AIC for Sample 3 highlighting compounds that were identified by the LOOCV regressions of Δ(ΔP). (C) TIC (sum of all *m/z*) for Sample 8 showing the chromatographic region where aromatic compounds elute. (D) AIC for Sample 8 highlighting compounds that were identified by the LOOCV regressions of Δ(ΔP).

Table 5.3 provides a summary of the number of tiles identified at various NRMSECV thresholds for the Δ(ΔP) and various forms of carbon. The table also indicates the error in the subsequent PLS modeling (i.e. the greater the number of tiles found at lower NRMSECV will result in a better PLS model). The lowest NRMSECV that had at least 100 tiles (or close to 100) was selected for subsequent inquiry in order to determine the relationship between the different forms of carbon and most importantly their relationship to Δ(ΔP).

Table 5.3. Summary of the number of tiles that are identified by the LOOCV regressions for the eight properties: Δ(ΔP) data, ΔP data, logarithmically scaled chemisorbed carbon in the heated zone (CCH), amorphous carbon in the exit zone (ACE), amorphous carbon in the heated zone (ACH), chemisorbed carbon in the exit zone (CCE), filamentous carbon in the exit zone (FCE), and filamentous carbon in the heated zone (FCH). The greater the number of tiles identified at lower NRMSECV thresholds will result in better PLS modeling. The lowest NRMSECV that had at least 100 tiles (or close to 100) was selected (highlighted green) for subsequent inquiry in order to determine the relationship between the different forms of carbon and most importantly their relationship to Δ(ΔP).

	NRMSECV (%) Threshold										
	11	12	13	14	15	16	17	18	19	20	21
Δ(ΔP)	5	22	83	167	248	336	431	534	646	785	942
ΔP	3	8	48	113	187	272	354	430	520	619	748
CCH	-	-	-	3	6	11	52	258	726	2503	3105
ACE	-	-	-	3	7	21	51	81	115	227	355
ACH	2	4	12	15	15	18	24	57	3105	3105	3105
CCE	-	-	-	-	-	6	158	254	329	417	527
FCE	30	78	141	258	429	634	3105	3105	3105	3105	3105
FCH	-	-	-	-	-	-	1	514	3105	3105	3105

Based on our previous findings, we included the results for amorphous carbon in the exit zone and chemisorbed carbon in the heated zone and excluded the remaining types and locations of carbonaceous deposits due to their lack of correlation with increased Δ(ΔP). Using the tiled GC×GC–TOFMS data, the PLS prediction of the amorphous carbon in the exit zone was accomplished in a similar manner to the Δ(ΔP) modeling. The feature selection technique was

applied, and Figure 5.9A shows the AIC for Sample 8 for the tiles identified by the LOOCV regression of the amorphous carbon in the exit zone that were below a NRMESCV threshold of 18%. Figure 5.9B shows the PLS prediction of the amorphous carbon in the exit zone using the tiles identified. Relatively speaking, the PLS model performs reasonably well at predicting the amorphous carbon in the exit zone. Taking this approach one step further, the tiles identified by the amorphous carbon in the exit zone LOOCV regressions were then used to predict the Δ(ΔP) using PLS (Figure 5.9C). This model significantly over-predicts and under-predicts several samples. Notably, one fuel is predicted to have a negative Δ(ΔP) value although all changes in pressure are positive. Using the same approach, Figure 5.9D shows the tiles selected for the logarithmically scaled chemisorbed carbon in the heated zone for a NRMSECV threshold of 17%. A total of 52 tiles were identified and the corresponding chromatographic features have some similarities to the features in the tiles identified by the Δ(ΔP) at a threshold of 13% (Figure 5.7B and C). Figure 5.9E shows the PLS prediction of the chemisorbed carbon in the heated zone using the tiles identified from the feature selection. In addition, these tiles were then used in the PLS prediction of the Δ(ΔP) (Figure 5.9F). The PLS modeling error is low at 12% and is very similar to the PLS model of Δ(ΔP) in Figure 5.7D, which is a very critical observation. To summarize the PLS models presented in Figure 5.9C and F, the feature-selected tiles from the regression of amorphous carbon in the exit zone and logarithmically-scaled chemisorbed carbon in the heated zone are demonstrated to "predict" something very different, the increase in Δ(ΔP). This experimental and modeling platform is providing clear evidence for the connection between which form of carbon is primarily responsible for the increase in pressure during the CRAFTI experiment (chemisorbed carbon in the heated zone), and will be shown to provide insight into which chemical compounds are most influential in doing so.

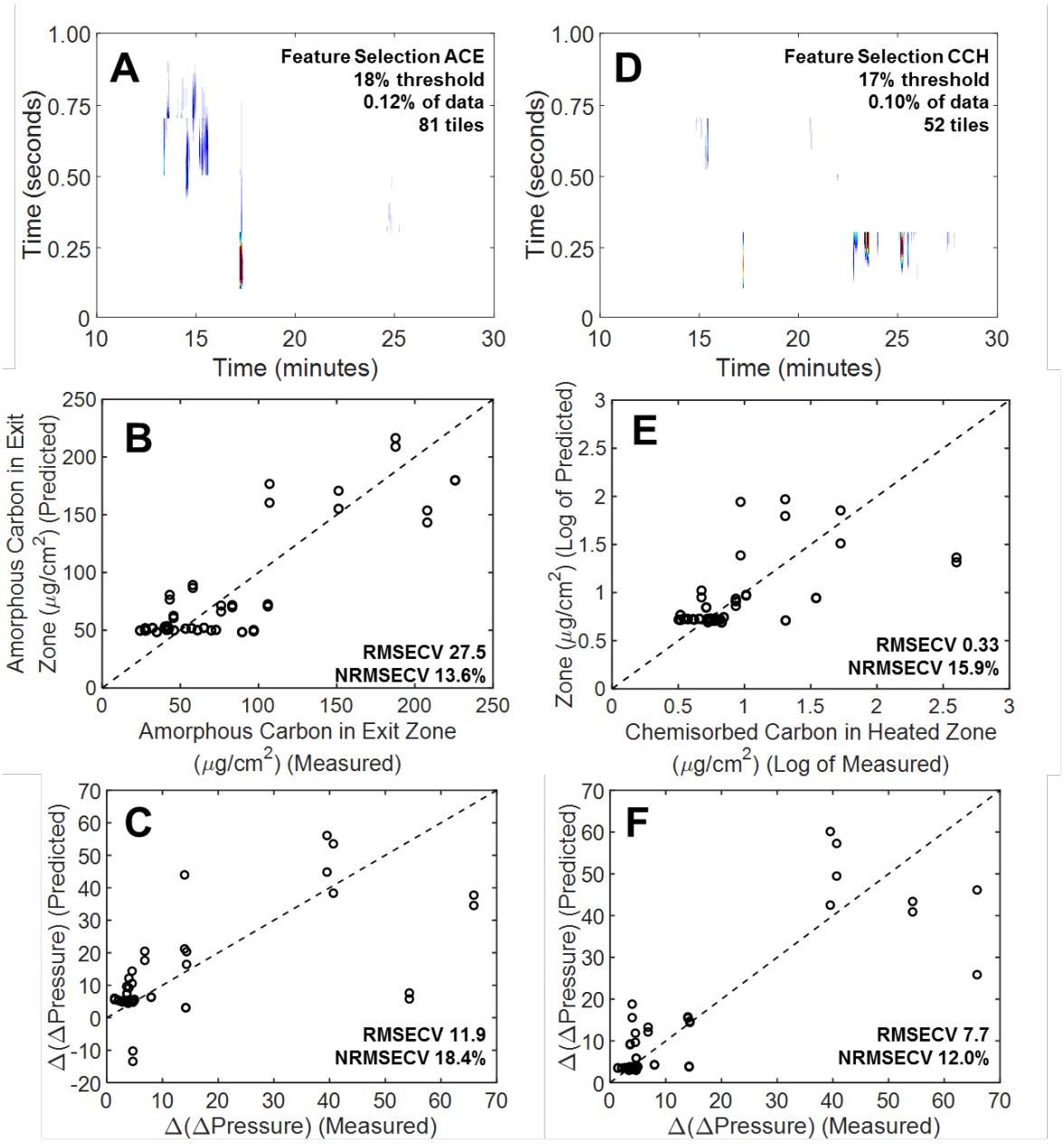


Figure 5.9. Supervised feature selection and subsequent prediction of amorphous carbon in the exit zone (ACE) and chemisorbed carbon in the heated zone (CCH). (A) The LOOCV regressions for the ACE from each tile per *m/z* which had a NRMSECV below 18% and contained at least 3 *m/z* were used to create an AIC chromatogram using Sample 8 to highlight the tiles identified. (B) PLS prediction of the ACE was performed using the tile locations and their respective *m/z*. (C) PLS prediction of Δ(ΔP) from tiles identified from the LOOCV regression of the ACE. (D) The LOOCV regressions for the logarithmically scaled CCH from each tile per *m/z* which had a NRMSECV below 17% and contained at least 3 *m/z* were used to create an AIC chromatogram using Sample 8 to highlight the tiles identified. (E) PLS prediction of the logarithmically CCH was performed using the tile locations and their respective *m/z*. (F) PLS prediction of Δ(ΔP) from tiles identified from the LOOCV regression of the logarithmically scaled CCH.

In order to provide a metric for the correlation between the tiles identified by the Δ(ΔP) and the various forms of carbon per test article locations (either heated zone or exit zone), the number of common tiles identified in the LOOCV regressions were compared. Table 5.4 provides a list of the top 20 tiles identified by the LOOCV regressions of Δ(ΔP) ordered in rank of number of *m/z*. For the *same 20 tile locations*, the number of *m/z* identified by LOOCV regressions of logarithmically-scaled chemisorbed carbon in the heated zone and amorphous carbon in the exit zone is included. The other carbonaceous deposit types and locations are excluded as they did not have many *m/z* in common within these tile locations. In addition, tentative identification of the compounds contained within the tiles identified by the LOOCV regressions of Δ(ΔP) are included. Some of the tiles contained more than one chemical species, with most of the compounds identified being naphthalene derivatives. Of the 83 tiles identified by the LOOCV regressions of Δ(ΔP), the logarithmically scaled chemisorbed carbon in the heated zone had 24 tiles in common with Δ(ΔP) and 33% total *m/z* shared, while the amorphous carbon in the exit zone had 7 tiles in common with Δ(ΔP) and 0% total *m/z* shared. Based on the number of shared tiles, it appears that the chemical species primarily associated with chemisorbed carbon deposition in the heated zone also accounts for the a significant number of the chemical species primarily associated with the resulting Δ(ΔP), which is consistent with previous reports [63–66]. Compared with our previous report, fewer tiles were included following feature selection of the logarithmically-scaled chemisorbed carbon in the heated zone (52 tiles compared with 120 tiles previously) [66]. This could explain why fewer features were observed in Figure 5.9D in general compared to our previous report, in addition to fewer features in common with the feature selection of Δ(ΔP), demonstrated in Figure 5.7C. Figure 5.10 illustrates the similarity between mass spectra of features selected by LOOCV regressions of Δ(ΔP) and chemisorbed carbon in the heated zone for the top four tiles in Table 5.4.

Table 5.4. List of compounds tentatively identified by the LOOCV regressions of Δ(ΔP) sorted by the highest number (#) of *m/z* identified in the tile. The top 20 tiles are shown. Also included are the number of *m/z* that were identified in the same tiles by the LOOCV regressions of chemisorbed carbon in the heated zone (CCH) and amorphous carbon in the exit zone (ACE).

Tile Hit Number	Name	Retention Time	# Δ(ΔP) *m/z*	# CCH *m/z*	# ACE *m/z*	# *m/z* in common Δ(ΔP) and CCH
1	4-methyl-biphenyl	^1t_R = 1522 s ^2t_R = 1.67 s	81	82	0	79
	2-(1-methylethyl)-naphthalene	^1t_R = 1528 s ^2t_R = 1.75 s				
2	1,3-dimethylnaphthalene	^1t_R = 1420 s ^2t_R = 1.73 s	75	70	0	68
3	4-methyl-biphenyl	^1t_R = 1540 s ^2t_R = 1.67 s	66	65	0	64
4	Biphenyl	^1t_R = 1375 s ^2t_R = 1.70 s	60	6	0	3
5	1,4,5-trimethylnaphthalene	^1t_R = 1519 s ^2t_R = 1.81 s	47	2	2	0
6	4,4'-dimethylbiphenyl	^1t_R = 1660 s ^2t_R = 1.73 s	44	47	0	40
7	1,5-dimethylnaphthalene	^1t_R = 1468 s ^2t_R = 1.69 s	43	1	0	0
8	4-methyl-biphenyl	^1t_R = 1522 s ^2t_R = 1.69 s	37	62	1	37
9	1,2,3,4-tetrahydro-3-isopropyl-5-methylnaphthalene	^1t_R = 1381 s ^2t_R = 2.09 s	36	0	0	0
10	1,4,5-trimethylnaphthalene	^1t_R = 1678 s ^2t_R = 1.72 s	32	20	2	10
11	1-methyl-3-(phenylmethyl)-benzene	^1t_R = 1591 s ^2t_R = 1.65 s	30	3	0	2
12	2,3,6-trimethylnaphthalene	^1t_R = 1564 s ^2t_R = 1.80 s	27	6	4	0
13	2-(1,1-dimethylethyl)-naphthalene	^1t_R = 1525 s ^2t_R = 1.88 s	23	5	12	0
14	7-butyl-tricyclo[4.2.2.0(2,5)]dec-7-ene	^1t_R = 1414 s ^2t_R = 2.22 s	20	0	0	0
15	(-)-3,7,7-trimethyl-11-methylene-spiro[5.5]undec-2-ene	^1t_R = 1384 s ^2t_R = 2.38 s	19	0	0	0
16	2-(1-methylethyl)-naphthalene	^1t_R = 1438 s ^2t_R = 1.82 s	18	0	1	0
17	4,5,9,10-dehydro-isolongifolene	^1t_R = 1543 s ^2t_R = 1.99 s	18	3	1	1
18	11-methylene-tricycle[5.3.1.1(2,6)]dodecane	^1t_R = 1219 s ^2t_R = 2.32 s	17	0	0	0
19	3-butyl-1-methyl-1H-indene	^1t_R = 1459 s ^2t_R = 1.92 s	16	5	0	0
20	1-cyclohexyl-3-methyl-benzene	^1t_R = 1327 s ^2t_R = 2.01 s	16	2	0	1

The mass spectra for the first three tiles align very well, consistent with the high number of *m/z* in common between the two fuel properties in Table 5.4. For tile 4, the number of *m/z* selected using

the chemisorbed carbon in the heated zone was considerably less than the number of *m/z* included using Δ(ΔP), which is also reflected in Figure 5.10D.

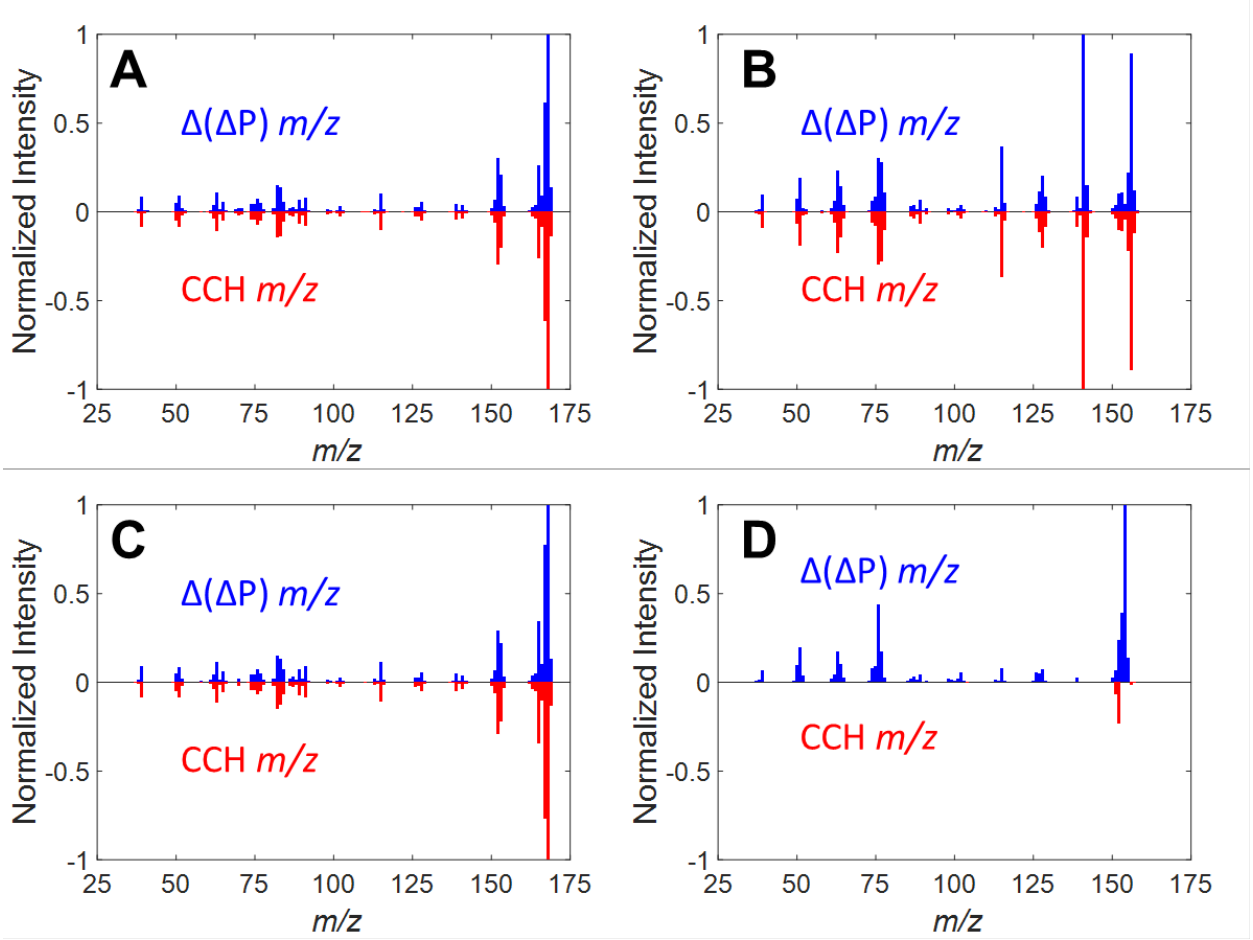


Figure 5.10. Mass spectral comparison of analytes in top 4 tiles for Δ(ΔP) with corresponding tiles for CCH. (A) Head-to-tail mass spectra of *m/z* in tile 1 containing 4-methyl-1,1'-diphenyl and 2-(1-methylethyl)-naphthalene identified by LOOCV regressions of Δ(ΔP) compared with the *m/z* in the tile identified by LOOCV regression of CCH. (B) Head-to-tail mass spectra of *m/z* in tile 2 containing 1,3-dimethylnaphthalene identified by LOOCV regressions of Δ(ΔP) compared with the *m/z* in the tile identified by LOOCV regression of CCH. (C) Head-to-tail mass spectra of *m/z* in tile 3 containing 4-methyl-1,1'-diphenyl identified by LOOCV regressions of Δ(ΔP) compared with the *m/z* in the tile identified by LOOCV regression of CCH. (D) Head-to-tail mass spectra of *m/z* in tile 4 containing diphenyl and identified by LOOCV regressions of Δ(ΔP) compared with the *m/z* in the tile identified by LOOCV regression of CCH.

5.4 CONCLUSION

The overarching goal of this research is to better understand and enhance fuel performance through chemometric modeling by relating fuel property data (ASTM measurements) and thermal performance (CRAFTI data) to chemical composition data (GC×GC–TOFMS). The predictive PLS modeling approach presented herein may ultimately be used to educate fuel selection. A feature selection method based on PLS analysis was demonstrated in the context of fuel thermal stability (test article pressure drop increase and carbon deposition data) in relation to chemical information present in GC×GC–TOFMS chromatograms. Using a subset of thirty-four fuels and selected features, PLS modeling was successful in predicting fuel properties based on chemical composition and determining chemical species responsible for large changes in test article pressure over time and carbonaceous deposition, particularly of chemisorbed carbon in the heated zone. This information can be used to further tailor chemical composition of kerosene-based rocket fuels to achieve optimal fuel performance. Future work should focus on investigating chemical compound identification that correlates to each physical property.

5.5 REFERENCES

[1] M.-F.S.G. Reyniers, G.F. Froment, Influence of Metal Surface and Sulfur Addition on Coke Deposition in the Thermal Cracking of Hydrocarbons, Ind. Eng. Chem. Res. 34 (1995) 773–785. https://doi.org/10.1021/ie00042a009.

[2] G.C. Reyniers, G.F. Froment, F.-D. Kopinke, G. Zimmermann, Coke Formation in the Thermal Cracking of Hydrocarbons. 4. Modeling of Coke Formation in Naphtha Cracking, Ind. Eng. Chem. Res. 33 (1994) 2584–2590. https://doi.org/10.1021/ie00035a009.

[3] F.D. Kopinke, G. Zimmermann, G.C. Reyniers, G.F. Froment, Relative rates of coke formation from hydrocarbons in steam cracking of naphtha. 3. Aromatic hydrocarbons, Ind. Eng. Chem. Res. 32 (1993) 2620–2625. https://doi.org/10.1021/ie00023a027.

[4] R. French, S. Czernik, Catalytic pyrolysis of biomass for biofuels production, Fuel Process. Technol. 91 (2010) 25–32. https://doi.org/10.1016/j.fuproc.2009.08.011.

[5] Y. Zhang, S. Kajitani, M. Ashizawa, Y. Oki, Tar destruction and coke formation during rapid pyrolysis and gasification of biomass in a drop-tube furnace, Fuel. 89 (2010) 302–309. https://doi.org/10.1016/j.fuel.2009.08.045.

[6] S. Du, J.A. Valla, G.M. Bollas, Characteristics and origin of char and coke from fast and slow, catalytic and thermal pyrolysis of biomass and relevant model compounds, Green Chem. 15 (2013) 3214–3229. https://doi.org/10.1039/C3GC41581C.
[7] K. Zhou, J. Jia, X. Li, X. Pang, C. Li, J. Zhou, G. Luo, F. Wei, Continuous vinyl chloride monomer production by acetylene hydrochlorination on Hg-free bismuth catalyst: From lab-scale catalyst characterization, catalytic evaluation to a pilot-scale trial by circulating regeneration in coupled fluidized beds, Fuel Process. Technol. 108 (2013) 12–18. https://doi.org/10.1016/j.fuproc.2012.03.018.
[8] A.G. Borsa, A.M. Herring, J.T. McKinnon, R.L. McCormick, G.H. Ko, Coke and Byproduct Formation during 1,2-Dichloroethane Pyrolysis in a Laboratory Tubular Reactor, Ind. Eng. Chem. Res. 40 (2001) 2428–2436. https://doi.org/10.1021/ie0006460.
[9] O. Altin, S. Eser, Analysis of Solid Deposits from Thermal Stressing of a JP-8 Fuel on Different Tube Surfaces in a Flow Reactor, Ind. Eng. Chem. Res. 40 (2001) 596–603. https://doi.org/10.1021/ie0004491.
[10] 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.
[11] G. Liu, Y. Han, L. Wang, X. Zhang, Z. Mi, 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.
[12] T. Edwards, Cracking and Deposition Behavior of Supercritical Hydrocarbon Aviation Fuels, Combust. Sci. Technol. 178 (2006) 307–334. https://doi.org/10.1080/00102200500294346.
[13] B.M. Fabuss, J.O. Smith, C.N. Satterfield, Thermal Cracking of Pure Saturated Hydrocarbons, in: Advances in Petroleum Chemistry and Refining, Interscience, 1964.
[14] O. Altin, S. Eser, Carbon deposit formation from thermal stressing of petroleum fuels, Am. Chem. Soc. Div. Fuel Chem. 49 (2004).
[15] R.T.K. Baker, D.J.C. Yates, J.A. Dumesic, Filamentous Carbon Formation over Iron Surfaces, in: Coke Formation on Metal Surfaces, American Chemical Society, 1983: pp. 1–21. https://doi.org/10.1021/bk-1983-0202.ch001.
[16] F.D. Kopinke, G. Zimmermann, S. Nowak, On the mechanism of coke formation in steam cracking—conclusions from results obtained by tracer experiments, Carbon. 26 (1988) 117–124. https://doi.org/10.1016/0008-6223(88)90027-9.
[17] S. Shao, H. Zhang, Y. Wang, R. Xiao, L. Heng, D. Shen, Catalytic Pyrolysis of Biomass-Derived Compounds: Coking Kinetics and Formation Network, Energy Fuels. 29 (2015) 1751–1757. https://doi.org/10.1021/ef5026505.
[18] H.M. Jeong, M.W. Seo, S.M. Jeong, B.K. Na, S.J. Yoon, J.G. Lee, W.J. Lee, Pyrolysis kinetics of coking coal mixed with biomass under non-isothermal and isothermal conditions, Bioresour. Technol. 155 (2014) 442–445. https://doi.org/10.1016/j.biortech.2014.01.005.
[19] F.D. Kopinke, G. Zimmermann, G.C. Reyniers, G.F. Froment, Relative rates of coke formation from hydrocarbons in steam cracking of naphtha. 2. Paraffins, naphthenes, mono-, di-, and cycloolefins, and acetylenes, Ind. Eng. Chem. Res. 32 (1993) 56–61. https://doi.org/10.1021/ie00013a009.

[20] J. Wang, M.-F. Reyniers, G.B. Marin, Influence of Dimethyl Disulfide on Coke Formation during Steam Cracking of Hydrocarbons, Ind. Eng. Chem. Res. 46 (2007) 4134–4148. https://doi.org/10.1021/ie061096u.
[21] N. Gascoin, G. Abraham, P. Gillard, Synthetic and jet fuels pyrolysis for cooling and combustion applications, J. Anal. Appl. Pyrolysis. 89 (2010) 294–306. https://doi.org/10.1016/j.jaap.2010.09.008.
[22] H. Huang, L.J. Spadaccini, D.R. Sobel, Fuel-Cooled Thermal Management for Advanced Aeroengines, J. Eng. Gas Turbines Power. 126 (2004) 284–293. https://doi.org/10.1115/1.1689361.
[23] H. Lander, A.C. Nixon, Endothermic fuels for hypersonic vehicles, J. Aircr. 8 (1971) 200–207. https://doi.org/10.2514/3.44255.
[24] H. Huang, D.R. Sobel, L.J. Spadaccini, Endothermic Heat-Sink of Jet Fuels for Scramjet Cooling, in: 38th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Indianapolis, IN, 2002; American Institute of Aeronautics and Astronautics: Reston, VA, 2002. https://doi.org/10.2514/6.2002-3871.
[25] R. Jiang, G. Liu, X. Zhang, Thermal Cracking of Hydrocarbon Aviation Fuels in Regenerative Cooling Microchannels, Energy Fuels. 27 (2013) 2563–2577. https://doi.org/10.1021/ef400367n.
[26] R. Bates, J. Edwards, M. Meyer, Heat Transfer and Deposition Behavior of Hydrocarbon Rocket Fuels, in: 41st Aerospace Sciences Meeting and Exhibit, Reno, NV, 2003; American Institute of Aeronautics and Astronautics: Reston, VA, 2003. https://doi.org/10.2514/6.2003-123.
[27] I.C. Lee, H.C. Ubanyionwu, Determination of sulfur contaminants in military jet fuels, Fuel. 87 (2008) 312–318. https://doi.org/10.1016/j.fuel.2007.05.010.
[28] 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.
[29] B. Stiegemeier, M. Meyer, R. Taghavi, A Thermal Stability and Heat Transfer Investigation of Five Hydrocarbon Fuels, in: 38th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Indianapolis, IN, 2002; American Institute of Aeronautics and Astronautics, Reston: VA, 2002. https://doi.org/10.2514/6.2002-3873.
[30] R. Venkataraman, S. Eser, Characterization of Solid Deposits Formed from Short Durations of Jet Fuel Degradation: Carbonaceous Solids, Ind. Eng. Chem. Res. 47 (2008) 9337–9350. https://doi.org/10.1021/ie8010066.
[31] 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.
[32] S. Tang, N. Shi, J. Wang, A. Tang, Comparison of the anti-coking performance of CVD TiN, TiO2 and TiC coatings for hydrocarbon fuel pyrolysis, Ceram. Int. 43 (2017) 3818–3823. https://doi.org/10.1016/j.ceramint.2016.12.036.
[33] 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.

[34] M.L. Huber, E.W. Lemmon, T.J. Bruno, Effect of RP-1 Compositional Variability on Thermophysical Properties, Energy Fuels. 23 (2009) 5550–5555. https://doi.org/10.1021/ef900597q.
[35] 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.
[36] MIL-DTL-25576E, Detail Specification: Propellant, Rocket Grade Kerosene, 2006.
[37] B.C. Windom, T.J. Bruno, Assessment of the Composition and Distillation Properties of Thermally Stressed RP-1 and RP-2: Application to Fuel Regenerative Cooling, Energy Fuels. 25 (2011) 5200–5214. https://doi.org/10.1021/ef201077a.
[38] 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.
[39] 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.
[40] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Comprehensive Two-Dimensional High-Speed Gas Chromatography with Chemometric Analysis, Anal. Chem. 70 (1998) 2796–2804. https://doi.org/10.1021/ac980164m.
[41] M.C. Billingsley, J.T. Edwards, L.M. Shafer, T.J. Bruno, Extent and Impacts of Hydrocarbon Fuel Compositional Variability for Aerospace Propulsion Systems, in: 46th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Nashville, TN, 2010; American Institute of Aeronautics and Astronautics: Reston, VA, 2010. https://doi.org/10.2514/6.2010-6824.
[42] 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.
[43] 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.
[44] 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.
[45] J.L. Burger, T.J. Bruno, Application of the Advanced Distillation Curve Method to the Variability of Jet Fuels, Energy Fuels. 26 (2012) 3661–3671. https://doi.org/10.1021/ef3006178.
[46] 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.
[47] M. Billingsley, N. Keim, R. Synovec, B. Hill-Lam, C. Wilhelm, Progress Toward Quality Assurance Standards for Advanced Hydrocarbon Fuels Based on Thermal Performance Testing and Chemometric Modeling, in: Proceedings of the IASH 14th International Symposium on Stability, Handling, and Use of Liquid Fuels, Charleston, SC, 2015; International Association for Stability, Handling, and Use of Liquid Fuels: Atlanta, GA, 2015.
[48] M. Billingsley, N. Keim, B. Hill-Lam, R. Synovec, Hydrocarbon Fuel Thermal Performance Modeling based on Systematic Measurement and Comprehensive

Chromatographic Analysis, in: 52nd AIAA/SAE/ASEE Joint Propulsion Conference, Salt Lake City, UT, 2016; American Institute of Aeronautics and Astronautics: Reston, VA, 2016. https://doi.org/10.2514/6.2016-4903.
[49] 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.
[50] 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.
[51] C.G. Fraga, B.J. Prazen, R.E. Synovec, Comprehensive two-dimensional gas chromatography and chemometrics for the high-speed quantitative analysis of aromatic isomers in a jet fuel using the standard addition method and an objective retention time alignment algorithm, Anal. Chem. 72 (2000) 4154–4162. https://doi.org/10.1021/ac000303b.
[52] B. Omais, M. Courtiade, N. Charon, D. Thiébaut, A. Quignard, M.-C. Hennion, Investigating comprehensive two-dimensional gas chromatography conditions to optimize the separation of oxygenated compounds in a direct coal liquefaction middle distillate, J. Chromatogr. A. 1218 (2011) 3233–3240. https://doi.org/10.1016/j.chroma.2010.12.049.
[53] 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.
[54] F. Westad, N.K. Afseth, R. Bro, Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares regression, Anal. Chim. Acta. 595 (2007) 323–327. https://doi.org/10.1016/j.aca.2007.02.015.
[55] T. Rajalahti, O.M. Kvalheim, Multivariate data analysis in pharmaceutics: a tutorial review, Int. J. Pharm. 417 (2011) 280–290. https://doi.org/10.1016/j.ijpharm.2011.02.019.
[56] A.A. Gowen, G. Downey, C. Esquerre, C.P. O'Donnell, Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients, J. Chemom. 25 (2011) 375–381. https://doi.org/10.1002/cem.1349.
[57] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.
[58] 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.
[59] 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.
[60] 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.

[61] J.R. Radović, K.V. Thomas, H. Parastar, S. Díez, R. Tauler, J.M. Bayona, Chemometrics-Assisted Effect-Directed Analysis of Crude and Refined Oil Using Comprehensive Two-Dimensional Gas Chromatography–Time-of-Flight Mass Spectrometry, Environ. Sci. Technol. 48 (2014) 3074–3083. https://doi.org/10.1021/es404859m.
[62] P. de Peinder, T. Visser, R. Wagemans, J. Blomberg, H. Chaabani, F. Soulimani, B.M. Weckhuysen, Sulfur Speciation of Crude Oils by Partial Least Squares Regression Modeling of Their Infrared Spectra, Energy Fuels. 24 (2010) 557–562. https://doi.org/10.1021/ef900908p.
[63] R.E. Synovec, C.E. Freye, B.A. Parsons, M.C. Billingsley, N. Keim, B. Hill-Lam, J.C. Wilhelm, 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, June 2015, Abstract Number: 2015-0001FU.
[64] 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 Liquid Propulsion Meeting, Phoenix, AZ, December 2016.
[65] R.E. Synovec, C.E. Freye, M.C. Billingsley, N. Keim, B. Hill-Lam, A. Bishop, 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, May 2018, Abstract Number: 2018-0001CV.
[66] R.E. Synovec, K.L. Berrier, C.E. Freye, S.E. Prebihalo, M.C. Billingsley, N. Keim, B. Hill-Lam, A. Bishop, Gaining a Fundamental Understanding of Fuel Performance through Advanced Chemical Composition Measurements, in: 66th JANNAF Propulsion Meeting, June 2019.
[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.

Chapter 6. Conclusions

6.1 Integration of Presented Work

One- and two-dimensional gas chromatography coupled with mass spectrometry is a powerful tool for the analysis of complex samples. When such chemically detailed and dense datasets are obtained, as in the implementation of GC-MS or GC×GC-MS, advanced data analysis methods are often used to extract meaningful information, such as when the data analysis goals include classification of samples or prediction of an independently measured sample property based on the chemical data. These advanced data analysis methods benefit from reduced datasets that contain only the most relevant chemical information, which can be achieved by feature selection. Feature selection methods, as well as chemometric methods, may be applied on raw data (pixel level), individual analyte features (peak tables), or binned data. The latter two data structure options additionally serve to reduce the dataset and remove the effects of noise and retention time misalignment.

Advances to and applications of feature selection with chemometric analysis in GC-MS and GC×GC-MS datasets were discussed herein. The first two chapters dealt with improvements to feature selection methods for GC-MS data, focusing on evaluating a method for analyte discovery and demonstrating the extension of an unsupervised feature selection method to chromatographic data. The last two chapters illustrated the application of feature selection and chemometric methods to GC×GC-MS datasets, providing a side-by-side comparison for circumstances in which feature selection is beneficial versus unnecessary.

6.2 Chapter 2 Summary, Limitations, and Future Directions

Chapter 2 described the evaluation of the 2D *m/z* cluster method (MCM) as an analyte discovery tool under varying conditions of chromatographic and sample complexity. Two levels of chromatographic complexity (saturation factor) and sample complexity (mass spectral similarity) were reflected in simulated chromatograms. Following the assumptions of the statistical model of overlap (SMO), the MCM behavior aligned with predictions of the theoretical percentage of peaks expected to be found as a function of resolution when the sample complexity was low (chromatograms contained analytes with mainly orthogonal spectra). Approximately 95% of analytes were discovered by the MCM at the optimum box size (analogous to resolution) when the analyte mass spectra set was of lower similarity (Low MV analyte set). The percentage of analytes discovered fell to 80% for the most challenging case studied: analyte mass spectra with high similarity (High MV analyte set) and chromatographically "full" of analytes with a saturation factor of 1. However, this result is much improved over traditional peak finders, which are expected to find only ~60% of the analytes present based on the SMO. Such improvements to analyte discovery can benefit feature selection by providing more chemically detailed and precise peak tables to which feature selection methods can be applied.

Many variables that affect component overlap and/or the MCM are worthy of additional study. For example, in the work presented herein, peak area was kept constant with a constant peak height and peak width for each analyte (prior to signal overlap and noise). However, peak heights in real chromatographic data are generally characterized by an exponentially distributed random variable. The minimum resolution at which two analytes can be distinguished is dependent upon the relative peak heights of the two analytes. Additionally, the MCM will only be more successful in analyte discovery at low resolutions when the peak width dimension is fully utilized as an

additional dimension of selectivity (i.e., variation in peak widths). Objective methodology for the selection of an appropriate width threshold deserves more study, especially when the aforementioned variables bear a greater resemblance to the values and distributions observed in real datasets.

Following the revisions made to the MCM, yielding the MCM 2.0, a study evaluating the performance of this revised method is warranted. A similar analytical platform to the one used in Chapter 2 would be appropriate, in which a rigorous and systematic evaluation of method performance in terms of window size, signal threshold, width threshold, peak capacity, number of components, separation efficiency, and mass spectral similarity can be achieved through application to realistic simulations. Future improvements to the MCM may be necessary. Further understanding of the effect of resolution, analyte pair match value, analyte peak height, analyte peak width, and intensity ratios of shared *m/z* on *m/z* distribution patterns could inform the development of a pattern recognition technique. Finally, application of the MCM to GC×GC-TOFMS data could be used to quantitatively demonstrate the 2D advantage, which is primarily observed through the theoretical increase in peak capacity of 2D separations relative to 1D separations. The MCM could be used as a tool to more realistically evaluate the increase in observable/discoverable analytes through application to both simulated and experimental data of 1D and 2D separations collected under similar conditions.

6.3 Chapter 3 Summary, Limitations, and Future Directions

Chapter 3 demonstrated the application of an unsupervised feature selection method based on variance thresholding to two GC-MS datasets. One dataset was comprised of simulated data with four analytes that were nominally designated to be different in terms of concentration between two classes. Within-class variation of 30% *RSD* was added to obscure the class-to-class

differences. The clustering and separation of the two classes in principal component analysis (PCA), quantified by the degree-of-class separation (DCS), improved from 0.66 ± 0.76 using all data to 2.66 ± 0.60 when only the selected features (data points with RSD^2 above the RSD^2 threshold of 0.1) were used. The correlation coefficient between the number of features discovered by the unsupervised variance thresholding method and supervised Fisher ratio (F-ratio) analysis for these simulations was 0.71. The second dataset consisted of GC×GC-TOFMS chromatograms of yeast metabolome samples from fermenting (repressed) and respiring (derepressed) cells, which was artificially reduced to GC-MS data by summing along the second separation dimension. Out of 53 discovered peaks, 27 had an RSD^2 above the RSD^2 threshold of 0.16, compared with 27 peaks with F-ratio values above the $F_{critical}$ value; 22 of these peaks were selected as features by both feature selection methods. The work in this study showed that unsupervised variance thresholding can provide similar information about important class-distinguishing features as supervised F-ratio analysis, expanding feature selection to cases in which class membership may not be known. The study also demonstrates feature selection applied to pixel level data and peak tables, a result of varying degrees of retention time misalignment.

A continuing challenge is the unbiased selection of an RSD^2 threshold to discriminate important features from irrelevant features in real-world unsupervised applications. A conservative approach is suggested in which the hits in the top 5% are investigated. Since these hits exhibit the greatest relative variance, they are statistically more likely to be class-indicating; therefore, the peak heights/areas of these hits can be input to PCA and class membership may be estimated based upon sample clustering (e.g., k-nearest neighbors classification). A more appropriate RSD^2 threshold can then be determined by estimating the within-class variation of the hypothesized classes. An alternative iterative approach consists of starting with a high RSD^2 threshold and

working down the hitlist until analysis becomes unfruitful. A less exhaustive approach involves analyzing a similar set of samples to describe the background variance expected in the dataset. Selection of samples must be done carefully to replicate the greatest source(s) of within-class variation without introducing chemical variation relevant to the investigation at hand.

Analogous to F-ratio analysis, variance thresholding can be applied to GC×GC data in a tile-based format. Additional studies are necessary to demonstrate the application of this method; the yeast metabolome dataset in its original GC×GC-TOFMS form is an ideal candidate with previous applications of tile-based F-ratio analysis providing a direct comparison for unsupervised variance thresholding.

6.4 Chapter 4 Summary, Limitations, and Future Directions

Chapter 4 represented a case in which feature selection was not required prior to the application of chemometrics. In the work described in this chapter, predictive modeling of the physical properties of kerosene-based fuels was accomplished based upon their chemical composition data, which was measured by GC×GC-TOFMS. Partial least squares (PLS) regression models were generated for viscosity, heat of combustion, hydrogen content, and temperature-dependent density measured for 74 fuels, of which 8 were designated to be outliers based upon their chromatographic data and statistical measures such as Q residuals and Hotelling's T^2. The remaining 66 fuels were modeled successfully with normalized root mean square error of cross validation (NRMSECV) less than 10% for all but one model, and RMSECV values within 10% of ASTM specification reproducibility. Linear regression vectors, which represent the significant chromatographic variables, for each property modeled were chemically sound. Overall, the research presented in this chapter demonstrated that GC×GC-TOFMS in combination with PLS can be used to model and subsequently predict various physical properties of rocket and jet fuels

using a single analysis method (once model calibration and validation is achieved) instead of multiple standard test methods. Feature selection was unnecessary in this endeavor due to the bulk properties measured herein; every analyte in a fuel contributed to its measured property, which can be thought of as a weighted average of all of the individual property values for each analyte present.

Generally in PLS, calibration of a model is followed by validation, in which an independent test set containing similar samples collected at a different time is used to estimate model performance (prediction error) when predicting new samples not contained in the training set. Unfortunately, in this study, an independent test set was not available for this purpose. Instead, a pseudo-external validation was performed in which the dataset was automatically split into calibration (55 samples) and validation (11 samples) sets using the Kennard-Stone algorithm. Root mean square error of prediction (RMSEP) values for the models were within the range spanned by the root mean square error of calibration (RMSEC) on the lower end and RMSECV on the upper end. Models generated by internal validation (cross validation) and the pseudo-external validation were very similar with slightly lower RMSECV values for the more robust, internally validated models. Since a truly independent test set could not be obtained, this pseudo-external validation exercise demonstrated that in this context, internal validation was sufficient for approximating the error in prediction for new samples of similar composition, measured by identical means.

Predictive capabilities of these models can be improved by including a greater number of samples that span a larger range of physical property values and contain more variety in chemical composition. Incorporating a truly independent test set would provide an accurate estimate of the prediction error expected for future samples. Similar models can be created for different types of samples (e.g., diesel, biodiesel) or additional physical properties of interest.

6.5 Chapter 5 Summary, Limitations, and Future Directions

Chapter 5 exemplified a situation in which feature selection prior to chemometric analysis was warranted. In this chapter, the thermal integrity of rocket fuels based upon their chemical composition (measured by GC×GC-TOFMS) was predictively modeled by PLS. A lab-scale platform created to simulate a regeneratively-cooled rocket engine was used to measure the pressure drop (ΔP) across the test article (representing a cooling channel) for a 900 s experiment, followed by an analysis of the type and amount of carbon deposited along the test article. These carbonaceous deposits included amorphous carbon, chemisorbed carbon, and filamentous carbon, which were characterized by location in the test article as being in either the heated zone or exit (unheated) zone. Based on previous research, increases in the pressure drop over time were indicative of a poorly performing fuel, while increases in the deposition of amorphous carbon in the exit zone (ACE) and chemisorbed carbon in the heated zone (CCH) were shown to be most correlated with these poorly behaving fuels. Due to variability in the starting pressure drop among different fuels, a correction was made by subtracting the pressure drop once most fuels had achieved steady state from the pressure drop at the end of the experiment. Additionally, carbon deposition data was provided in mass units instead of counts, as previously modeled. Initial modeling of the corrected change in pressure drop (Δ(ΔP)) resulted in negative predicted values for some fuels and NRMSECV close to 20%. Supervised feature selection was implemented to identify the chromatographic features most correlated with each property. Simple linear regression along with a cross validation method analogous to leave-one-out cross validation (LOOCV) for replicate samples was used to select features based upon the slope of the predicted property values relative to the measured values and error in the predicted values (NRMSECV). For the Δ(ΔP), a NRMSECV threshold of 13% was deemed acceptable, and gave close to 100 independent

chromatographic features (0.13% of the data). Modeling Δ(ΔP) using only these features reduced the NRMSECV close to 10% and eliminated the negative Δ(ΔP) predictions. Subsequently, the features selected for ACE and CCH (discovered using NRMSECV thresholds to yield approximately 100 individual features) were used to predict Δ(ΔP). While the ACE features did not provide a good model (error close to 20% and negative Δ(ΔP) predictions), CCH features provided a model of similar quality to the one achieved by the Δ(ΔP) features. Further investigation of the top 20 features related to Δ(ΔP) revealed that at the present NRMSECV thresholds, many of these features were selected for both Δ(ΔP) and CCH, and were characterized by a significant number of *m/z* in common between the Δ(ΔP) and corresponding CCH features. Tentative identification of these features illustrated that the chemical compounds most related to the increase in pressure drop and deposition of CCH were aromatics. Feature selection was appropriate in this situation due to the relatively small number of chromatographic features expected to contribute to poor fuel thermal integrity.

A significant limitation to this study was the lack of fuel samples with Δ(ΔP) in the middle and upper ranges. Most of the fuels analyzed were well-behaving with low Δ(ΔP), and the modeling would benefit from a more uniform distribution of samples across each of the properties of interest. It is recommended that more poorly behaving and mediocre fuels be analyzed and included in the modeling if possible. Additionally, the supervised feature selection methodology implemented herein is very computationally expensive, taking hours for a single property, even when parallel computing is utilized. With the addition of more fuels, this feature selection method will become more impractical. An alternative to LOOCV, venetian blinds cross validation could be used with a blind thickness of 2 to account for the replicate chromatographic data.

The current dataset may be amenable to F-ratio analysis, where the fuels are classified as either poorly behaving or well-behaving. F-ratio analysis can be used to determine which chemical features are statistically different in terms of concentration between classes, and ideally these features will be related to the increasingly poor performance of the fuels. F-ratio analysis could be utilized as a feature selection method prior to PLS modeling or as a stand-alone method to identify potential undesirable analytes that should be considered in future fuel formulations and specifications. Further study into the carbon deposition in specific sections of the test article is possible.

6.6 Final Thoughts

The research presented herein demonstrates the majority of work performed over the last five years. However, behind the final versions of the datasets, simulations, and models shown herein were countless attempts at modeling and revisions of code that contributed greatly to my learning in graduate school. In addition to the work presented herein are some additional datasets and projects, bits of code, and many collaborative efforts that are not represented in this dissertation. Nonetheless, I am grateful for the projects that never materialized into something more and for the stumbles and falls along the way, as they have helped me learn and grow as a scientist and person, possibly more so than the successes.

It has been an honor to contribute to the fields of gas chromatography, feature selection, and chemometrics, and I hope that these small, yet significant, contributions aid in a greater understanding of the fundamentals, applications, and advantages of gas chromatography coupled with feature selection and chemometrics in the analysis of complex samples.

BIBLIOGRAPHY

Abbasi, Saleheh, Sajjad Gharaghani, Ali Benvidi, and AliMohammad Latif. 2018. "Identifying the novel natural antioxidants by coupling different feature selection methods with nonlinear regressions and gas chromatography-mass spectroscopy." *Microchemical Journal* 139: 372-379. https://doi.org/10.1016/j.microc.2018.03.012.

Abrahamsson, Victor, Nenad Ristic, Kristina Franz, and Kevin Van Geem. 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.

Altin, Orhan, and Semih Eser. 2001. "Analysis of Solid Deposits from Thermal Stressing of a JP-8 Fuel on Different Tube Surfaces in a Flow Reactor." *Industrial & Engineering Chemistry Research* 40 (2): 596-603. https://doi.org/10.1021/ie0004491.

---. 2004. "Carbon deposit formation from thermal stressing of petroleum fuels." *Preprints of Papers - American Chemical Society, Division of Fuel Chemistry* 49 (2): 764-766.

American Society for Testing and Materials. 2016. ASTM D7171-16: Standard Test Method for Hydrogen Content of Middle Distillate Petroleum Products by Low-Resolution Pulsed Nuclear Magnetic Resonance Spectroscopy. ASTM International, West Conshohocken, PA.

---. 2018a. ASTM D4052-18a: Standard Test Method for Density, Relative Density, and API Gravity of Liquids by Digital Density Meter. ASTM International, West Conshohocken, PA.

---. 2018b. ASTM D4809-18: Standard Test Method for Heat of Combustion of Liquid Hydrocarbon Fuels by Bomb Calorimeter (Precision Method). ASTM International, West Conshohocken, PA.

---. 2019. ASTM D445-19: Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity). ASTM International, West Conshohocken, PA.

Amigo, José Manuel, Marta J. Popielarz, Raquel M. Callejón, Maria L. Morales, Ana M. Troncoso, Mikael A. Petersen, and Torben B. Toldam-Andersen. 2010. "Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis." *Journal of Chromatography A* 1217 (26): 4422-4429. https://doi.org/10.1016/j.chroma.2010.04.042.

Amigo, José Manuel, Thomas Skov, Rasmus Bro, Jordi Coello, and Santiago Maspoch. 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.

Aragón, Álvaro, Rosa M. Toledano, Sara Gea, José M. Cortés, Ana M. Vázquez, and Jesús Villén. 2014. "Large volume injection in gas chromatography using the through oven transfer adsorption desorption interface operating under vacuum." *Talanta* 123: 39-44. https://doi.org/10.1016/j.talanta.2014.01.064.

Aspromonte, Juan, Kris Wolfs, and Erwin Adams. 2019. "Current application and potential use of GC × GC in the pharmaceutical and biomedical field." *Journal of Pharmaceutical and Biomedical Analysis* 176: 112817. https://doi.org/10.1016/j.jpba.2019.112817.

Baker, R. T. K., D. J. C. Yates, and J. A. Dumesic. 1983. "Filamentous Carbon Formation over Iron Surfaces." In *Coke Formation on Metal Surfaces*, In ACS Symposium Series, 1-21. American Chemical Society.

Ballabio, Davide, and Viviana Consonni. 2013. "Classification tools in chemistry. Part 1: linear models. PLS-DA." *Analytical Methods* 5 (16): 3790-3798. https://doi.org/10.1039/c3ay40582f.

Barp, Laura, Giorgia Purcaro, Flavio A. Franchina, Mariosimone Zoccali, Danilo Sciarrone, Peter Q. Tranchida, and Luigi Mondello. 2015. "Determination of phthalate esters in vegetable oils using direct immersion solid-phase microextraction and fast gas chromatography coupled with triple quadrupole mass spectrometry." *Analytica Chimica Acta* 887: 237-244. https://doi.org/10.1016/j.aca.2015.06.039.

Barwick, Vicki 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.

Bates, Ronald, James Edwards, and Michael Meyer. 2003. "Heat Transfer and Deposition Behavior of Hydrocarbon Rocket Fuels." 41st Aerospace Sciences Meeting and Exhibit, Reno, NV.

Bean, Heather D., Jane E. Hill, and Jean-Marie D. Dimandja. 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.

Bean, Heather D., Christiaan A. Rees, and Jane E. Hill. 2016. "Comparative analysis of the volatile metabolomes of Pseudomonas aeruginosaclinical isolates." *Journal of Breath Research* 10 (4): 047102. https://doi.org/10.1088/1752-7155/10/4/047102.

Begue, Nathan J., Jeffery A. Cramer, Chris Von Bargen, Kristina M. Myers, Kevin J. Johnson, and Robert E. Morris. 2011. "Automated Method for Determining Hydrocarbon Distributions in Mobility Fuels." *Energy & Fuels* 25 (4): 1617-1623. https://doi.org/10.1021/ef101635a.

Billingsley, Matthew C., J. Tim Edwards, Linda M. Shafer, and Thomas J. Bruno. 2010. "Extent and Impacts of Hydrocarbon Fuel Compositional Variability for Aerospace Propulsion Systems." Proceedings of the 46th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Nashville, TN.

Billingsley, Matthew C., Nicholas Keim, Benjamin Hill-Lam, and Robert E. Synovec. 2016. "Hydrocarbon Fuel Thermal Performance Modeling based on Systematic Measurement and Comprehensive Chromatographic Analysis." Proceedings of the 52nd AIAA/SAE/ASEE Joint Propulsion Conference, Salt Lake City, UT.

Billingsley, Matthew C., Nicholas Keim, Robert E. Synovec, Benjamin Hill-Lam, and Claire Wilhelm. 2015. "Progress Toward Quality Assurance Standards for Advanced Hydrocarbon Fuels Based on Thermal Performance Testing and Chemometric Modeling." Proceedings of the IASH 14th International Symposium on Stability, Handling, and Use of Liquid Fuels, Charleston, SC.

Borsa, Alessandro G., Andrew M. Herring, J. Thomas McKinnon, Robert L. McCormick, and Glen H. Ko. 2001. "Coke and Byproduct Formation during 1,2-Dichloroethane Pyrolysis in a Laboratory Tubular Reactor." *Industrial & Engineering Chemistry Research* 40 (11): 2428-2436. https://doi.org/10.1021/ie0006460.

Brereton, Richard G. 2009. *Chemometrics for Pattern Recognition*. West Sussex, U.K.: John Wiley & Sons, Ltd.

Bro, Rasmus. 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.
---. 1997. "PARAFAC. Tutorial and applications." *Chemometrics and Intelligent Laboratory Systems* 38 (2): 149-171. https://doi.org/10.1016/S0169-7439(97)00032-4.
Bro, Rasmus, and Sijmen De Jong. 1997. "A fast non-negativity-constrained least squares algorithm." *Journal of Chemometrics* 11 (5): 393-401. https://doi.org/10.1002/(SICI)1099-128X(199709/10)11:5<393::AID-CEM483>3.0.CO;2-L.
Brownlee, Jason. 2020. *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery.
Bruckner, Carsten A., Bryan J. Prazen, and Robert E. Synovec. 1998. "Comprehensive Two-Dimensional High-Speed Gas Chromatography with Chemometric Analysis." *Analytical Chemistry* 70 (14): 2796-2804. https://doi.org/10.1021/ac980164m.
Bruno, Thomas J., Lisa S. Ott, Tara M. Lovestead, and Marcia L. Huber. 2010. "The composition-explicit distillation curve technique: Relating chemical analysis and physical properties of complex fluids." *Journal of Chromatography A* 1217 (16): 2703-2715. https://doi.org/10.1016/j.chroma.2009.11.030.
Burger, Jessica L., and Thomas J. Bruno. 2012. "Application of the Advanced Distillation Curve Method to the Variability of Jet Fuels." *Energy & Fuels* 26 (6): 3661-3671. https://doi.org/10.1021/ef3006178.
Cai, Deng, Chiyuan Zhang, and Xiaofei He. 2010. "Unsupervised feature selection for multi-cluster data." 16th ACM SIGKDD International Conference, Washington, DC, USA.
Cai, Jie, Jiawei Luo, Shulin Wang, and Sheng Yang. 2018. "Feature selection in machine learning: A new perspective." *Neurocomputing* 300: 70-79. https://doi.org/10.1016/j.neucom.2017.11.077.
Carson, Mariah, and Sarah Kerrigan. 2017. "Quantification of suvorexant in urine using gas chromatography/mass spectrometry." *Journal of Chromatography B* 1040: 289-294. https://doi.org/10.1016/j.jchromb.2016.10.042.
Chakravarthy, Ramachandra, Chhayakanta Acharya, Anilkumar Savalia, Ganesh N. Naik, Asit Kumar Das, Chandra Saravanan, Anurag Verma, and Kalagouda B. Gudasi. 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.
Chandrashekar, Girish, and Ferat Sahin. 2014. "A survey on feature selection methods." *Computers & Electrical Engineering* 40 (1): 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
Ciosek, P., Z. Brzózka, W. Wróblewski, E. Martinelli, C. Di Natale, and A. D'Amico. 2005. "Direct and two-stage data analysis procedures based on PCA, PLS-DA and ANN for ISE-based electronic tongue—Effect of supervised feature extraction." *Talanta* 67 (3): 590-596. https://doi.org/10.1016/j.talanta.2005.03.006.
Cookson, David J., Peter Iliopoulos, and Brian E. Smith. 1995. "Composition-property relations for jet and diesel fuels of variable boiling range." *Fuel* 74 (1): 70-78. https://doi.org/10.1016/0016-2361(94)P4333-W.

Cookson, David J., Jozef L. Latten, Ian M. Shaw, and Brian E. Smith. 1985. "Property-composition relationships for diesel and kerosene fuels." *Fuel* 64 (4): 509-519. https://doi.org/10.1016/0016-2361(85)90086-9.

Cookson, David J., C. Paul Lloyd, and Brian E. Smith. 1987. "Investigation of the chemical basis of kerosene (jet fuel) specification properties." *Energy & Fuels* 1 (5): 438-447. https://doi.org/10.1021/ef00005a011.

Cookson, David J., and Brian E. Smith. 1990. "Calculation of jet and diesel fuel properties using carbon-13 NMR spectroscopy." *Energy & Fuels* 4 (2): 152-156. https://doi.org/10.1021/ef00020a004.

Cordero, Chiara, Hans-Georg Schmarr, Stephen E. Reichenbach, and Carlo Bicchi. 2018. "Current Developments in Analyzing Food Volatiles by Multidimensional Gas Chromatographic Techniques." *Journal of Agricultural and Food Chemistry* 66: 2226-2236. https://doi.org/10.1021/acs.jafc.6b04997.

Couprie, Camille, Laurent Duval, Maxime Moreaud, Sophie Hénon, Mélinda Tebib, and Vincent Souchon. 2017. "BARCHAN: Blob Alignment for Robust CHromatographic ANalysis." *Journal of Chromatography A* 1484: 65-72. https://doi.org/10.1016/j.chroma.2017.01.003.

Cramer, Jeffrey A., Mark H. Hammond, Kristina M. Myers, Iwona A. Leska, and Robert E. Morris. 2015. "Expanded Framework for the Prediction of Alternative Fuel Content and Alternative Fuel Blend Performance Properties Using Near-Infrared Spectroscopic Data." *Energy & Fuels* 29 (11): 7026-7035. https://doi.org/10.1021/acs.energyfuels.5b01660.

Cramer, Jeffrey A., Mark H. Hammond, Kristina M. Myers, Thomas N. Loegel, and Robert E. Morris. 2014. "Novel Data Abstraction Strategy Utilizing Gas Chromatography–Mass Spectrometry Data for Fuel Property Modeling." *Energy & Fuels* 28 (3): 1781-1791. https://doi.org/10.1021/ef4021872.

Davis, Joe 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.

---. 2019. "Theory of the probability of total resolution in chromatograms with systematic variation of average peak spacing and peak width." *Journal of Chromatography A* 1588: 150-158. https://doi.org/10.1016/j.chroma.2018.12.031.

Davis, Joe M., and J. Calvin Giddings. 1983. "Statistical theory of component overlap in multicomponent chromatograms." *Analytical Chemistry* 55 (3): 418–424. https://doi.org/10.1021/ac00254a003.

---. 1985a. "Statistical method for estimation of number of components from single complex chromatograms: application to experimental chromatograms." *Analytical Chemistry* 57 (12): 2178-2182. https://doi.org/10.1021/ac00289a003.

---. 1985b. "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, Joe M., Sarah C. Rutan, and Peter W. Carr. 2011. "Relationship between selectivity and average resolution in comprehensive two-dimensional separations with spectroscopic detection." *Journal of Chromatography A* 1218 (34): 5819-5828. https://doi.org/10.1016/j.chroma.2011.06.086.

de Godoy, Luiz Antonio F., Ernesto Correa Ferreira, Marcio Pozzobon Pedroso, Carlos Henrique de V. Fidélis, Fabio Augusto, and Ronei Jesus Poppi. 2008. "Quantification of Kerosene

in Gasoline by Comprehensive Two-Dimensional Gas Chromatography and *N* -Way Multivariate Analysis." *Analytical Letters* 41 (9): 1603-1614. https://doi.org/10.1080/00032710802122222.

de Godoy, Luiz Antonio Fonseca, Leandro Wang Hantao, Marcio Pozzobon Pedroso, Ronei Jesus Poppi, and Fabio Augusto. 2011. "Quantitative analysis of essential oils in perfume using multivariate curve resolution combined with comprehensive two-dimensional gas chromatography." *Analytica Chimica Acta* 699 (1): 120-125. https://doi.org/10.1016/j.aca.2011.05.003.

de Godoy, Luiz Antonio Fonseca, Marcio Pozzobon Pedroso, Leandro Wang Hantao, Ronei Jesus Poppi, and Fabio Augusto. 2011. "Quantitative analysis by comprehensive two-dimensional gas chromatography using interval Multi-way Partial Least Squares calibration." *Talanta* 83 (4): 1302-1307. https://doi.org/10.1016/j.talanta.2010.08.015.

de Juan, Anna, Joaquim Jaumot, and Romà Tauler. 2014. "Multivariate curve resolution (MCR). Solving the mixture analysis problem." *Analytical Methods* 6 (14): 4964-4976. https://doi.org/10.1039/C4AY00571F.

de Juan, Anna, and Romà Tauler. 2001. "Comparison of three-way resolution methods for non-trilinear chemical data sets." *Journal of Chemometrics* 15 (10): 749-771. https://doi.org/10.1002/cem.662.

---. 2007. "Factor analysis of hyphenated chromatographic data exploration, resolution and quantification of multicomponent systems." *Journal of Chromatography A* 1158 (1-2): 184-195. https://doi.org/10.1016/j.chroma.2007.05.045.

de Peinder, Peter, Tom Visser, Rudy Wagemans, Jan Blomberg, Hassan Chaabani, Fouad Soulimani, and Bert M. Weckhuysen. 2010. "Sulfur Speciation of Crude Oils by Partial Least Squares Regression Modeling of Their Infrared Spectra." *Energy & Fuels* 24 (1): 557-562. https://doi.org/10.1021/ef900908p.

DeWitt, Matthew J., Tim Edwards, Linda Shafer, David Brooks, Richard Striebich, Sean P. Bagley, and Mary J. Wornat. 2011. "Effect of Aviation Fuel Type on Pyrolytic Reactivity and Deposition Propensity under Supercritical Conditions." *Industrial & Engineering Chemistry Research* 50 (18): 10434-10451. https://doi.org/10.1021/ie200257b.

Du, Shoucheng, Julia A. Valla, and George M. Bollas. 2013. "Characteristics and origin of char and coke from fast and slow, catalytic and thermal pyrolysis of biomass and relevant model compounds." *Green Chemistry* 15 (11): 3214-3229. https://doi.org/10.1039/c3gc41581c.

Edwards, Tim. 2006. "Cracking and Deposition Behavior of Supercritical Hydrocarbon Aviation Fuels." *Combustion Science and Technology* 178 (1-3): 307-334. https://doi.org/10.1080/00102200500294346.

Eser, Semih, Ramya Venkataraman, and Orhan Altin. 2006. "Deposition of Carbonaceous Solids on Different Substrates from Thermal Stressing of JP-8 and Jet A Fuels." *Industrial & Engineering Chemistry Research* 45 (26): 8946-8955. https://doi.org/10.1021/ie060968p.

Faber, N. M., and R. Rajkó. 2007. "How to avoid over-fitting in multivariate calibration - The conventional validation approach and an alternative." *Analytica Chimica Acta* 595: 98-106. https://doi.org/10.1016/j.aca.2007.05.030.

Fabuss, BM, JO Smith, and CN Satterfield. 1964. "Thermal cracking of pure saturated hydrocarbons." *Advances in petroleum chemistry and refining* 9: 157-201.

Fakayode, Sayo O., Breanna S. Mitchell, and David A. Pollard. 2014. "Determination of boiling point of petrochemicals by gas chromatography–mass spectrometry and multivariate

regression analysis of structural activity relationship." *Talanta* 126: 151-156. https://doi.org/10.1016/j.talanta.2014.03.037.

Fernández-Varela, R., G. Tomasi, and J. H. Christensen. 2015. "An untargeted gas chromatography mass spectrometry metabolomics platform for marine polychaetes." *Journal of Chromatography A* 1384: 133-141. https://doi.org/10.1016/j.chroma.2015.01.025.

Fitz, Brian D., Brandyn C. Mannion, Khang To, Trinh Hoac, and Robert E. Synovec. 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.

Fitz, Brian D., Brooke C. Reaser, David K. Pinkerton, Jamin C. Hoggard, Kristen J. Skogerboe, and Robert E. Synovec. 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, Brian D., and Robert E. Synovec. 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.

Fortin, Tara J., and Thomas J. Bruno. 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.

Fortunato de Carvalho Rocha, Werickson, Michele M. Schantz, David A. Sheen, Pamela M. Chu, and Katrice A. Lippa. 2017. "Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data." *Fuel* 197: 248-258. https://doi.org/10.1016/j.fuel.2017.02.025.

Fraga, Carlos 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 (1): 31-42. https://doi.org/10.1016/S0021-9673(03)01329-3.

Fraga, Carlos G., Carsten A. Bruckner, and Robert E. Synovec. 2001. "Increasing the number of analyzable peaks in comprehensive two-dimensional separations through chemometrics." *Analytical Chemistry* 73 (3): 675-683. https://doi.org/10.1021/ac0010025.

Fraga, Carlos G., Bryan J. Prazen, and Robert E. Synovec. 2000. "Comprehensive Two-Dimensional Gas Chromatography and Chemometrics for the High-Speed Quantitative Analysis of Aromatic Isomers in a Jet Fuel Using the Standard Addition Method and an Objective Retention Time Alignment Algorithm." *Analytical Chemistry* 72 (17): 4154-4162. https://doi.org/10.1021/ac000303b.

French, Richard, and Stefan Czernik. 2010. "Catalytic pyrolysis of biomass for biofuels production." *Fuel Processing Technology* 91 (1): 25-32. https://doi.org/10.1016/j.fuproc.2009.08.011.

Freye, Chris E., Patrick R. Bowden, Margo T. Greenfield, and Bryce C. Tappan. 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.

Freye, Chris E., Brian D. Fitz, Matthew C. Billingsley, and Robert E. Synovec. 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.
Furbo, Søren, Asger B. Hansen, Thomas Skov, and Jan H. Christensen. 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.
Gaddes, David, Jessica Westland, Frank L. Dorman, and Srinivas Tadigadapa. 2014. "Improved micromachined column design and fluidic interconnects for programmed high-temperature gas chromatography separations." *Journal of Chromatography A* 1349: 96-104. https://doi.org/10.1016/j.chroma.2014.04.087.
Garcia, Antonia, and Coral Barbas. 2011. "Gas chromatography-mass spectrometry (GC-MS)-based metabolomics." In *Metabolic Profiling*, edited by Thomas O. Metz, In Methods in Molecular Biology, 191-204. Humana Press.
Gascoin, N., G. Abraham, and P. Gillard. 2010. "Synthetic and jet fuels pyrolysis for cooling and combustion applications." *Journal of Analytical and Applied Pyrolysis* 89 (2): 294-306. https://doi.org/10.1016/j.jaap.2010.09.008.
Geladi, Paul, and Bruce R. Kowalski. 1986. "Partial least-squares regression: a tutorial." *Analytica Chimica Acta* 185: 1-17. https://doi.org/10.1016/0003-2670(86)80028-9.
Giddings, J. Calvin. 1991. *Unified Separation Science*. New York: John Wiley & Sons, Inc.
Golub, Gene H., and Christian Reinsch. 1970. "Singular value decomposition and least squares solutions." *Numerische Mathematik* 14 (5): 403–420. https://doi.org/10.1007/BF02163027.
Gough, R. V., and T. J. Bruno. 2012. "Composition-Explicit Distillation Curves of Alternative Turbine Fuels." *Energy & Fuels* 27 (1): 294-302. https://doi.org/10.1021/ef3016848.
Gowen, A. A., G. Downey, C. Esquerre, and C. P. O'Donnell. 2011. "Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients." *Journal of Chemometrics* 25 (7): 375-381. https://doi.org/10.1002/cem.1349.
Graffius, Gabriel C., Brandon M. Jocher, Daniel Zewge, Holst M. Halsey, Gary Lee, Frank Bernardoni, Xiaodong Bu, Robert Hartman, and Erik L. Regalado. 2017. "Generic gas chromatography-flame ionization detection method for quantitation of volatile amines in pharmaceutical drugs and synthetic intermediates." *Journal of Chromatography A* 1518: 70-77. https://doi.org/10.1016/j.chroma.2017.08.048.
Gröger, Thomas M., Uwe Käfer, and Ralf Zimmermann. 2020. "Gas chromatography in combination with fast high-resolution time-of-flight mass spectrometry: Technical overview and perspectives for data visualization." *TrAC, Trends in Analytical Chemistry* 122: 115677. https://doi.org/10.1016/j.trac.2019.115677.
Gruber, B., B. A. Weggler, R. Jaramillo, K. A. Murrell, P. K. Piotrowski, and F. L. Dorman. 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.
Gruber, Beate, Frank David, and Pat Sandra. 2020. "Capillary gas chromatography-mass spectrometry: Current trends and perspectives." *TrAC, Trends in Analytical Chemistry* 124: 115475. https://doi.org/10.1016/j.trac.2019.04.007.
Gruber, Beate, Julian Schneider, Michael Föhlinger, Jeroen Buters, Ralf Zimmermann, and Georg Matuschek. 2017. "A minimal-invasive method for systemic bio-monitoring of the

environmental pollutant phenanthrene in humans: Thermal extraction and gas chromatography − mass spectrometry from 1 mL capillary blood." *Journal of Chromatography A* 1487: 254-257. https://doi.org/10.1016/j.chroma.2017.01.045.

Guyon, Isabelle, and André Elisseeff. 2003. "An introduction to variable and feature selection." *Journal of Maching Learning Research* 3: 1157-1182.

Haar, Lilli, Katharina Anding, Konstantin Trambitckii, and Gunther Notni. 2019. "Comparison between Supervised and Unsupervised Feature Selection Methods." 8th International Conference on Pattern Recognition Applications and Methods, Prague, Czech Republic.

Harvey, Paul McA., and Robert A. Shellie. 2012. "Data Reduction in Comprehensive Two-Dimensional Gas Chromatography for Rapid and Repeatable Automated Data Analysis." *Analytical Chemistry* 84 (15): 6501-6507. https://doi.org/10.1021/ac300664h.

Hashimoto, Shunji, Yasuyuki Zushi, Akihiro Fushimi, Yoshikatsu Takazawa, Kiyoshi Tanabe, and Yasuyuki Shibata. 2013. "Selective extraction of halogenated compounds from data measured by comprehensive multidimensional gas chromatography/high resolution time-of-flight mass spectrometry for non-target analysis of environmental and biological samples." *Journal of Chromatography A* 1282: 183-189. https://doi.org/10.1016/j.chroma.2013.01.052.

He, Xiaofei, Deng Cai, and Partha Niyogi. 2006. "Laplacian Score for Feature Selection." In *Advances in Neural Information Processing Systems 18*, edited by Y. Weiss, B. Schölkopf and J. C. Platt, 507-514. MIT Press.

Higgens Keppler, Emily A., Carrie L. Jenkins, Trenton J. Davis, and Heather D. Bean. 2018. "Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics." *TrAC, Trends in Analytical Chemistry* 109: 275-286. https://doi.org/10.1016/j.trac.2018.10.015.

Hoh, Eunha, and Katerina Mastovska. 2008. "Large volume injection techniques in capillary gas chromatography." *Journal of Chromatography A* 1186 (1): 2-15. https://doi.org/10.1016/j.chroma.2007.12.001.

Hsieh, Peter Y., Kathryn R. Abel, and Thomas J. Bruno. 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, H., David R. Sobel, and Louis J. Spadaccini. 2002. "Endothermic Heat-Sink of Jet Fuels for Scramjet Cooling." 38th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Indianapolis, IN.

Huang , He, Louis J. Spadaccini , and David R. Sobel. 2004. "Fuel-Cooled Thermal Management for Advanced Aeroengines." *Journal of Engineering for Gas Turbines and Power* 126 (2): 284-293. https://doi.org/10.1115/1.1689361.

Huber, M. L., E. W. Lemmon, and T. J. Bruno. 2009. "Effect of RP-1 Compositional Variability on Thermophysical Properties." *Energy & Fuels* 23 (11): 5550-5555. https://doi.org/10.1021/ef900597q.

Hurtado, Carlos, Hadi Parastar, Víctor Matamoros, Benjamín Piña, Romà Tauler, and Josep M. Bayona. 2017. "Linking the morphological and metabolomic response of Lactuca sativa L exposed to emerging contaminants using GC × GC-MS and chemometric tools." *Scientific Reports* 7 (1): 6546. https://doi.org/10.1038/s41598-017-06773-0.

Izadmanesh, Yahya, Elba Garreta-Lara, Jahan B. Ghasemi, Silvia Lacorte, Victor Matamoros, and Roma Tauler. 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.

Jacobs, Matthew R., Emily F. Hilder, and Robert A. Shellie. 2013. "Applications of resistive heating in gas chromatography: A review." *Analytica Chimica Acta* 803: 2-14. https://doi.org/10.1016/j.aca.2013.04.063.

Jaumot, Joaquim, Raimundo Gargallo, Anna De Juan, and Romà Tauler. 2005. "A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB." *Chemometrics and Intelligent Laboratory Systems* 76 (1): 101-110. https://doi.org/10.1016/j.chemolab.2004.12.007.

Jennerwein, Maximilian, Markus Eschner, Thomas Wilharm, Thomas Gröger, and Ralf Zimmermann. 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.

Jeong, Ha Myung, Myung Won Seo, Sang Mun Jeong, Byung Ki Na, Sang Jun Yoon, Jae Goo Lee, and Woon Jae Lee. 2014. "Pyrolysis kinetics of coking coal mixed with biomass under non-isothermal and isothermal conditions." *Bioresource Technology* 155: 442-445. https://doi.org/10.1016/j.biortech.2014.01.005.

Jiang, Rongpei, Guozhu Liu, and Xiangwen Zhang. 2013. "Thermal Cracking of Hydrocarbon Aviation Fuels in Regenerative Cooling Microchannels." *Energy & Fuels* 27 (5): 2563-2577. https://doi.org/10.1021/ef400367n.

Jin, Baitang, Kai Jing, Jie Liu, Xiangwen Zhang, and Guozhu Liu. 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, Kevin J., and Robert E. Synovec. 2002. "Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis." *Chemometrics and Intelligent Laboratory Systems* 60 (1): 225-237. https://doi.org/10.1016/S0169-7439(01)00198-8.

Kanaujia, Pankaj K. 2019. "Gas Chromatography | Petroleum and Petrochemical Applications." In *Encyclopedia of Analytical Science (Third Edition)*, edited by Paul Worsfold, Colin Poole, Alan Townshend and Manuel Miró, 217-231. Oxford: Academic Press.

Kehimkar, Benjamin, Jamin C. Hoggard, Luke C. Marney, Matthew C. Billingsley, Carlos G. Fraga, Thomas J. Bruno, and Robert E. Synovec. 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, Benjamin, Brendon A. Parsons, Jamin C. Hoggard, Matthew C. Billingsley, Thomas J. Bruno, and Robert E. Synovec. 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): 321-30. https://doi.org/10.1007/s00216-014-8233-6.

Kondo, Elsuida, Philip John Marriott, Rhiannon M. Parker, Konstantinos A. Kouremenos, Paul Morrison, and Mike Adams. 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: 135-42. https://doi.org/10.1016/j.aca.2013.11.004.
Kopinke, Frank Dieter, Gerhard Zimmermann, and S. Nowak. 1988. "On the mechanism of coke formation in steam cracking—conclusions from results obtained by tracer experiments." *Carbon* 26 (2): 117-124. https://doi.org/10.1016/0008-6223(88)90027-9.
Kopinke, Frank Dieter, Gerhard Zimmermann, Geerd C. Reyniers, and Gilbert F. Froment. 1993a. "Relative rates of coke formation from hydrocarbons in steam cracking of naphtha. 2. Paraffins, naphthenes, mono-, di-, and cycloolefins, and acetylenes." *Industrial & Engineering Chemistry Research* 32 (1): 56-61. https://doi.org/10.1021/ie00013a009.
---. 1993b. "Relative rates of coke formation from hydrocarbons in steam cracking of naphtha. 3. Aromatic hydrocarbons." *Industrial & Engineering Chemistry Research* 32 (11): 2620-2625. https://doi.org/10.1021/ie00023a027.
Kurganov, A. 2013. "Monolithic column in gas chromatography." *Analytica Chimica Acta* 775: 25-40. https://doi.org/10.1016/j.aca.2013.02.039.
Lander, H., and A. C. Nixon. 1971. "Endothermic fuels for hypersonic vehicles." *Journal of Aircraft* 8 (4): 200-207. https://doi.org/10.2514/3.44255.
Lebanov, Leo, Laura Tedone, Alireza Ghiasvand, and Brett Paull. 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: 120471. https://doi.org/10.1016/j.talanta.2019.120471.
Lee, I., and H. Ubanyionwu. 2008. "Determination of sulfur contaminants in military jet fuels." *Fuel* 87 (3): 312-318. https://doi.org/10.1016/j.fuel.2007.05.010.
Leffler, Amanda M., Philip B. Smith, Adriana de Armas, and Frank L. Dorman. 2014. "The analytical investigation of synthetic street drugs containing cathinone analogs." *Forensic Science International* 234: 50-56. https://doi.org/10.1016/j.forsciint.2013.08.021.
Lehallier, Benoist, Jérémy Ratel, Mohamed Hanafi, and Erwan Engel. 2012. "Systematic ratio normalization of gas chromatography signals for biological sample discrimination and biomarker discovery." *Analytica Chimica Acta* 733: 16-22. https://doi.org/10.1016/j.aca.2012.04.019.
Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2018. "Feature Selection: A Data Perspective." *ACM Computing Surveys* 50 (6): 1-45. https://doi.org/10.1145/3136625.
Liu, Guozhu, Yongjin Han, Li Wang, Xiangwen Zhang, and Zhentao Mi. 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.
Liu, Guozhu, Li Wang, Haijie Qu, Huiming Shen, Xiangwen Zhang, Shuting Zhang, and Zhentao Mi. 2007. "Artificial neural network approaches on composition–property relationships of jet fuels based on GC–MS." *Fuel* 86 (16): 2551-2559. https://doi.org/10.1016/j.fuel.2007.02.023.
Liu, Zaiyou, and John B. Phillips. 1991. "Comprehensive Two-Dimensional Gas Chromatography using an On-Column Thermal Modulator Interface." *Journal of Chromatographic Science* 29 (6): 227-231. https://doi.org/10.1093/chromsci/29.6.227.
Lovestead, Tara M., Bret C. Windom, Jennifer R. Riggs, Christopher Nickell, and Thomas J. Bruno. 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.
Lu, Yao, and Peter B. Harrington. 2007. "Forensic Application of Gas Chromatography–Differential Mobility Spectrometry with Two-Way Classification of Ignitable Liquids from Fire Debris." *Analytical Chemistry* 79 (17): 6752-6759. https://doi.org/10.1021/ac0707028.
Magagna, Federico, Alessandro Guglielmetti, Erica Liberto, Stephen E. Reichenbach, Elena Allegrucci, Guido Gobino, Carlo Bicchi, and Chiara Cordero. 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.
Mallepally, R. R., B. A. Bamgbade, M. A. McHugh, H. O. Baled, R. M. Enick, and M. C. Billingsley. 2019. "Measurements and modeling of the density of rocket propellant RP-2 at temperatures to 573 K and pressures to 100 MPa." *Fuel* 253: 1193-1203. https://doi.org/10.1016/j.fuel.2019.05.089.
Marney, Luke C., W. Christopher Siegler, Brendon A. Parsons, Jamin C. Hoggard, Bob W. Wright, and Robert E. Synovec. 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.
MIL-DTL-25576E - Detail Specification: Propellant, Rocket Grade Kerosene. 2006.
Mogollon, Noroska Gabriela Salazar, Fabiana Alves de Lima Ribeiro, Monica Mamian Lopez, Leandro Wang Hantao, Ronei Jesus Poppi, and Fabio Augusto. 2013. "Quantitative analysis of biodiesel in blends of biodiesel and conventional diesel by comprehensive two-dimensional gas chromatography and multivariate curve resolution." *Analytica Chimica Acta* 796: 130-136. https://doi.org/10.1016/j.aca.2013.07.071.
Mohan, Arun Ram, and Semih Eser. 2010. "Analysis of Carbonaceous Solid Deposits from Thermal Oxidative Stressing of Jet-A Fuel on Iron- and Nickel-Based Alloy Surfaces." *Industrial & Engineering Chemistry Research* 49 (6): 2722-2730. https://doi.org/10.1021/ie901283r.
Mohler, Rachel E., Kenneth M. Dombek, Jamin C. Hoggard, Karisa M. Pierce, Elton T. Young, and Robert E. Synovec. 2007. "Comprehensive analysis of yeast metabolite GC x GC-TOFMS data: combining discovery-mode and deconvolution chemometric software." *Analyst* 132 (8): 756-67. https://doi.org/10.1039/b700061h.
Mohler, Rachel E., Kenneth M. Dombek, Jamin C. Hoggard, Elton T. Young, and Robert E. Synovec. 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.
Morris, Robert E., Mark H. Hammond, Jeffrey A. Cramer, Kevin J. Johnson, Braden C. Giordano, Kirsten E. Kramer, and Susan L. Rose-Pehrsson. 2009. "Rapid Fuel Quality Surveillance through Chemometric Modeling of Near-Infrared Spectra." *Energy & Fuels* 23 (3): 1610-1618. https://doi.org/10.1021/ef800869t.
Murray, Jacolin A. 2012. "Qualitative and quantitative approaches in comprehensive two-dimensional gas chromatography." *Journal of Chromatography A* 1261: 58-68. https://doi.org/10.1016/j.chroma.2012.05.012.

Muscalu, Alina M., and Tadeusz Górecki. 2018. "Comprehensive two-dimensional gas chromatography in environmental analysis." *TrAC, Trends in Analytical Chemistry* 106: 225-245. https://doi.org/10.1016/j.trac.2018.07.001.

Omais, Badaoui, Marion Courtiade, Nadège Charon, Didier Thiebaut, Alain Quignard, and Marie-Claire Hennion. 2011. "Investigating comprehensive two-dimensional gas chromatography conditions to optimize the separation of oxygenated compounds in a direct coal liquefaction middle distillate." *Journal of Chromatography A* 1218 (21): 3233-40. https://doi.org/10.1016/j.chroma.2010.12.049.

Ott, Lisa Starkey, Amelia B. Hadler, and Thomas J. Bruno. 2008. "Variability of The Rocket Propellants RP-1, RP-2, and TS-5: Application of a Composition- and Enthalpy-Explicit Distillation Curve Method." *Industrial & Engineering Chemistry Research* 47 (23): 9225-9233. https://doi.org/10.1021/ie800988u.

Outcalt, Stephanie L., Arno Laesecke, and Karin J. Brumback. 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.

Parastar, Hadi, Mehdi Jalali-Heravi, and Roma Tauler. 2012. "Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution." *Chemometrics and Intelligent Laboratory Systems* 117: 80-91. https://doi.org/10.1016/j.chemolab.2012.02.003.

Parastar, Hadi, Sara Mostafapour, and Gholamhasan Azimi. 2016. "Quality assessment of gasoline using comprehensive two-dimensional gas chromatography combined with unfolded partial least squares: A reliable approach for the detection of gasoline adulteration." *Journal of Separation Science* 39 (2): 367-374. https://doi.org/10.1002/jssc.201500720.

Parastar, Hadi, Jagoš R. Radović, Josep M. Bayona, and Roma Tauler. 2013. "Solving chromatographic challenges in comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry using multivariate curve resolution–alternating least squares." *Analytical and Bioanalytical Chemistry* 405 (19): 6235-6249. https://doi.org/10.1007/s00216-013-7067-y.

Parastar, Hadi, Jagoš R. Radović, Mehdi Jalali-Heravi, Sergi Diez, Josep Maria Bayona, and Roma Tauler. 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, Brendon A., Luke C. Marney, W. Christopher Siegler, Jamin C. Hoggard, Bob W. Wright, and Robert E. Synovec. 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, Brendon A., David K. Pinkerton, and Robert E. Synovec. 2018. "Implications of phase ratio for maximizing peak capacity in comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry." *Journal of Chromatography A* 1536: 16-26. https://doi.org/10.1016/j.chroma.2017.07.018.

Parsons, Brendon A., David K. Pinkerton, Bob W. Wright, and Robert E. Synovec. 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.

Pedroso, Marcio Pozzobon, Luiz Antonio Fonseca de Godoy, Ernesto Correa Ferreira, Ronei Jesus Poppi, and Fabio Augusto. 2008. "Identification of gasoline adulteration using comprehensive two-dimensional gas chromatography combined to multivariate data processing." *Journal of Chromatography A* 1201 (2): 176-182. https://doi.org/10.1016/j.chroma.2008.05.092.

Pérez-Guaita, David, Guillermo Quintás, and Julia Kuligowski. 2020. "Discriminant analysis and feature selection in mass spectrometry imaging using constrained repeated random sampling - Cross validation (CORRS-CV)." *Analytica Chimica Acta* 1097: 30-36. https://doi.org/10.1016/j.aca.2019.10.039.

Pierce, Karisa M., Jamin C. Hoggard, Janiece L. Hope, Petrie M. Rainey, Andrew N. Hoofnagle, Rhona M. Jack, Bob W. Wright, and Robert E. Synovec. 2006. "Fisher Ratio Method Applied to Third-Order Separation Data To Identify Significant Chemical Components of Metabolite Extracts." *Analytical Chemistry* 78 (14): 5068-5075. https://doi.org/10.1021/ac0602625.

Pierce, Karisa M., Janiece L. Hope, Kevin J. Johnson, Bob W. Wright, and Robert E. Synovec. 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 (1-2): 101-10. https://doi.org/10.1016/j.chroma.2005.04.078.

Pierce, Karisa M., Benjamin Kehimkar, Luke C. Marney, Jamin C. Hoggard, and Robert E. Synovec. 2012. "Review of chemometric analysis techniques for comprehensive two dimensional separations data." *Journal of Chromatography A* 1255: 3-11. https://doi.org/10.1016/j.chroma.2012.05.050.

Pierce, Karisa M., and Stephen P. Schale. 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.

Pierce, Karisa M., Lianna F. Wood, Bob W. Wright, and Robert E. Synovec. 2005. "A Comprehensive Two-Dimensional Retention Time Alignment Algorithm To Enhance Chemometric Analysis of Comprehensive Two-Dimensional Separation Data." *Analytical Chemistry* 77 (23): 7735-7743. https://doi.org/10.1021/ac0511142.

Pinkerton, David K., Brendon A. Parsons, Todd J. Anderson, and Robert E. Synovec. 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, David K., Brooke C. Reaser, Kelsey L. Berrier, and Robert E. Synovec. 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., I. Esteban-Diez, C. Saenz-Gonzalez, and J. M. Gonzalez-Saiz. 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.

Pollo, Breno J., Guilherme L. Alexandrino, Fabio Augusto, and Leandro W. Hantao. 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.

Poole, Colin F., and Nicole Lenca. 2014. "Gas chromatography on wall-coated open-tubular columns with ionic liquid stationary phases." *Journal of Chromatography A* 1357: 87-109. https://doi.org/10.1016/j.chroma.2014.03.029.

Prazen, Bryan J., Kevin J. Johnson, Andrew Weber, and Robert E. Synovec. 2001. "Two-Dimensional Gas Chromatography and Trilinear Partial Least Squares for the Quantitative Analysis of Aromatic and Naphthene Content in Naphtha." *Analytical Chemistry* 73 (23): 5677-5682. https://doi.org/10.1021/ac010637g.

Prebihalo, Sarah E., Kelsey L. Berrier, Chris E. Freye, H. Daniel Bahaghighat, Nicholas R. Moore, David K. Pinkerton, and Robert E. Synovec. 2018. "Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications." *Analytical Chemistry* 90 (1): 505-532. https://doi.org/10.1021/acs.analchem.7b04226.

Radović, Jagoš R., Kevin V. Thomas, Hadi Parastar, Sergi Díez, Romà Tauler, and Josep M. Bayona. 2014. "Chemometrics-Assisted Effect-Directed Analysis of Crude and Refined Oil Using Comprehensive Two-Dimensional Gas Chromatography–Time-of-Flight Mass Spectrometry." *Environmental Science & Technology* 48 (5): 3074-3083. https://doi.org/10.1021/es404859m.

Rajalahti, T., and O. M. Kvalheim. 2011. "Multivariate data analysis in pharmaceutics: a tutorial review." *International Journal of Pharmaceutics* 417 (1-2): 280-90. https://doi.org/10.1016/j.ijpharm.2011.02.019.

Ranzan, Lucas, Cassiano Ranzan, Luciane F. Trierweiler, and Jorge O. Trierweiler. 2017. "Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy." *Energy & Fuels* 31 (9): 8942-8950. https://doi.org/10.1021/acs.energyfuels.7b00954.

Reaser, Brooke C., Song Yang, Brian D. Fitz, Brendon A. Parsons, Mary E. Lidstrom, and Robert E. Synovec. 2016. "Non-targeted determination of 13 C-labeling in the Methylobacterium extorquens AM1 metabolome using the two-dimensional mass cluster method and principal component analysis." *Journal of Chromatography A* 1432: 111-121. https://doi.org/10.1016/j.chroma.2015.12.088.

Reichenbach, Stephen E., Xue Tian, Chiara Cordero, and Qingping Tao. 2012. "Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography." *Journal of Chromatography A* 1226: 140-148. https://doi.org/10.1016/j.chroma.2011.07.046.

Reyniers, Geert C., Gilbert F. Froment, Frank-Dieter Kopinke, and Gerhard Zimmermann. 1994. "Coke Formation in the Thermal Cracking of Hydrocarbons. 4. Modeling of Coke Formation in Naphtha Cracking." *Industrial & Engineering Chemistry Research* 33 (11): 2584-2590. https://doi.org/10.1021/ie00035a009.

Reyniers, Marie-Francoise S. G., and Gilbert F. Froment. 1995. "Influence of Metal Surface and Sulfur Addition on Coke Deposition in the Thermal Cracking of Hydrocarbons." *Industrial & Engineering Chemistry Research* 34 (3): 773-785. https://doi.org/10.1021/ie00042a009.

Ribeiro, J. S., F. Augusto, T. J. G. Salva, R. A. Thomaziello, and M. M. C. Ferreira. 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.

Sampat, Andjoe A. S., Martin Lopatka, Gabriel Vivó-Truyols, Peter J. Schoenmakers, and Arian C. van Asten. 2016. "Towards chemical profiling of ignitable liquids with comprehensive two-dimensional gas chromatography: Exploring forensic application to neat white spirits." *Forensic Science International* 267: 183-195. https://doi.org/10.1016/j.forsciint.2016.08.006.

Seeley, John V., and Stacy K. Seeley. 2013. "Multidimensional Gas Chromatography: Fundamental Advances and New Applications." *Analytical Chemistry* 85 (2): 557-578. https://doi.org/10.1021/ac303195u.

Setiono, Rudy, and Huan Liu. 1998. "Feature extraction via Neural networks." In *Feature Extraction, Construction and Selection: A Data Mining Perspective*, edited by Huan Liu and Hiroshi Motoda, In The Springer International Series in Engineering and Computer Science, 191-204. Boston, MA: Springer US.

Shao, Shanshan, Huiyan Zhang, Yun Wang, Rui Xiao, Lijun Heng, and Dekui Shen. 2015. "Catalytic Pyrolysis of Biomass-Derived Compounds: Coking Kinetics and Formation Network." *Energy & Fuels* 29 (3): 1751-1757. https://doi.org/10.1021/ef5026505.

Shi, Xiangpeng, Haijing Li, Zhaoyu Song, Xiangwen Zhang, and Guozhu Liu. 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.

Stein, Stephen E. 1999. "An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data." *Journal of the American Society for Mass Spectrometry* 10 (8): 770-781. https://doi.org/10.1016/S1044-0305(99)00047-1.

Stiegemeier, Benjamin, Michael Meyer, and Ray Taghavi. 2002. "A Thermal Stability and Heat Transfer Investigation of Five Hydrocarbon Fuels." 38th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Indianapolis, IN. https://doi.org/10.2514/6.2002-3873.

Striebich, Richard C., Linda M. Shafer, Ryan K. Adams, Zachary J. West, Matthew J. DeWitt, and Steven Zabarnick. 2014. "Hydrocarbon Group-Type Analysis of Petroleum-Derived and Synthetic Fuels Using Two-Dimensional Gas Chromatography." *Energy & Fuels* 28 (9): 5696-5706. https://doi.org/10.1021/ef500813x.

Synovec, Robert E., Kelsey L. Berrier, Chris E. Freye, Sarah E. Prebihalo, Matthew C. Billingsley, Nicholas Keim, Benjamin Hill-Lam, and A. Bishop. 2019. "Gaining a Fundamental Understanding of Fuel Performance through Advanced Chemical Composition Measurements." 66th JANNAF Propulsion Meeting, Dayton, OH.

Synovec, Robert, Chris E. Freye, M. C. Billingsley, Nicholas Keim, and Benjamin Hill-Lam. 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." JANNAF 9th Liquid Propulsion Meeting, Phoenix, AZ.

Synovec, Robert, Chris E. Freye, Matthew C. Billingsley, Nicholas Keim, Benjamin Hill-Lam, and A. Bishop. 2018. "Recent Advances in Relating Chemical Compositional Variation in RP-1, RP-2, and Similar Fuels to Thermal Integrity Data." JANNAF 10th Liquid Propulsion Meeting, Long Beach, CA.

Synovec, Robert, Chris E. Freye, Brendon A. Parsons, Matthew C. Billingsley, Nicholas Keim, Benjamin Hill-Lam, and J. C. Wilhelm. 2015. "Development of Chemical Analysis Tools to Relate Compositional Variation to Thermal Integrity Data for RP-1, RP-2, and Related Fuels." JANNAF 8th Liquid Propulsion Meeting, Nashville, TN.

Tang, Shiyun, Ning Shi, Jianli Wang, and Anjiang Tang. 2017. "Comparison of the anti-coking performance of CVD TiN, TiO 2 and TiC coatings for hydrocarbon fuel pyrolysis." *Ceramics International* 43 (4): 3818-3823. https://doi.org/10.1016/j.ceramint.2016.12.036.

Tian, Tze-Feng, San-Yuan Wang, Tien-Chueh Kuo, Cheng-En Tan, Guan-Yuan Chen, Ching-Hua Kuo, Chi-Hsin Sally Chen, Chang-Chuan Chan, Olivia A. Lin, and Y. Jane Tseng. 2016. "Web Server for Peak Detection, Baseline Correction, and Alignment in Two-Dimensional Gas Chromatography Mass Spectrometry-Based Metabolomics Data." *Analytical Chemistry* 88 (21): 10395-10403. https://doi.org/10.1021/acs.analchem.6b00755.

Tobias, Randall D. 1995. "An introduction to partial least squares regression." SUGI Proceedings, Orlando, FL.

Tranchida, Peter Q., Giorgia Purcaro, Mariarosa Maimone, and Luigi Mondello. 2016. "Impact of comprehensive two-dimensional gas chromatography with mass spectrometry on food analysis." *Journal of Separation Science* 39 (1): 149-161. https://doi.org/10.1002/jssc.201500379.

Trenzado, José L., José S. Matos, Luisa Segade, and Enrique Carballo. 2001. "Densities, Viscosities, and Related Properties of Some (Methyl Ester + Alkane) Binary Mixtures in the Temperature Range from 283.15 to 313.15 K." *Journal of Chemical and Engineering Data* 46 (4): 974-983. https://doi.org/10.1021/je0100286.

van Deursen, M. M., J. Beens, H. -G. Janssen, P. A. Leclercq, and C. A. Cramers. 2000. "Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography." *Journal of Chromatography A* 878 (2): 205-213. https://doi.org/10.1016/S0021-9673(00)00300-9.

Vendeuvre, Colombe, Fabrice Bertoncini, Laurent Duval, Jean-Luc Duplan, Didier Thiébaut, and Marie-Claire Hennion. 2004. "Comparison of conventional gas chromatography and comprehensive two-dimensional gas chromatography for the detailed analysis of petrochemical samples." *Journal of Chromatography A* 1056 (1): 155-162. https://doi.org/10.1016/j.chroma.2004.05.071.

Venkataraman, Ramya, and Semih Eser. 2008. "Characterization of Solid Deposits Formed from Short Durations of Jet Fuel Degradation: Carbonaceous Solids." *Industrial & Engineering Chemistry Research* 47 (23): 9337-9350. https://doi.org/10.1021/ie8010066.

Vozka, Petr, Huaping Mo, Pavel Šimáček, and Gozdem Kilaz. 2018. "Middle distillates hydrogen content via GCxGC-FID." *Talanta* 186: 140-146. https://doi.org/10.1016/j.talanta.2018.04.059.

Vozka, Petr, Brent A. Modereger, Anthony C. Park, Wan Tang Jeff Zhang, Rodney W. Trice, Hilkka I. Kenttämaa, and Gozdem Kilaz. 2019. "Jet fuel density via GC × GC-FID." *Fuel* 235: 1052-1060. https://doi.org/10.1016/j.fuel.2018.08.110.

Wang, Jidong, Marie-Françoise Reyniers, and Guy B. Marin. 2007. "Influence of Dimethyl Disulfide on Coke Formation during Steam Cracking of Hydrocarbons." *Industrial & Engineering Chemistry Research* 46 (12): 4134-4148. https://doi.org/10.1021/ie061096u.

Wang, Yan, Liang Xu, Hong Shen, Juanjuan Wang, Wei Liu, Xianwen Zhu, Ronghua Wang, Xiaochuan Sun, and Liwang Liu. 2015. "Metabolomic analysis with GC-MS to reveal potential metabolites and biological pathways involved in Pb & Cd stress response of radish roots." *Scientific Reports* 5: 18296. https://doi.org/10.1038/srep18296.

Wang, Zhengfang, and Peter de B. Harrington. 2013. "Feature selection of gas chromatography/mass spectrometry chemical profiles of basil plants using a bootstrapped fuzzy rule-building expert system." *Analytical and Bioanalytical Chemistry* 405 (28): 9219-9234. https://doi.org/10.1007/s00216-013-7327-x.

Watson, Nathanial E., Brendon A. Parsons, and Robert E. Synovec. 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.

Watson, Nathanial E., Matthew M. Vanwingerden, Karisa M. Pierce, Bob W. Wright, and Robert E. Synovec. 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): 111-8. https://doi.org/10.1016/j.chroma.2006.06.087.

Welke, Juliane Elisa, Vitor Manfroi, Mauro Zanus, Marcelo Lazzarotto, and Cláudia Alcaraz Zini. 2013. "Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data." *Food Chemistry* 141 (4): 3897-3905. https://doi.org/10.1016/j.foodchem.2013.06.100.

Westad, Frank, Nils Kristian Afseth, and Rasmus Bro. 2007. "Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares regression." *Analytica Chimica Acta* 595 (1-2): 323-7. https://doi.org/10.1016/j.aca.2007.02.015.

Wilson, Ryan B., Brian D. Fitz, Brandyn C. Mannion, Tina Lai, Roy K. Olund, Jamin C. Hoggard, and Robert E. Synovec. 2012. "High-speed cryo-focusing injection for gas chromatography: Reduction of injection band broadening with concentration enrichment." *Talanta* 97: 9-15. https://doi.org/10.1016/j.talanta.2012.03.054.

Wilson, Ryan B., Jamin C. Hoggard, and Robert E. Synovec. 2012. "Fast, high peak capacity separations in gas chromatography-time-of-flight mass spectrometry." *Analytical Chemistry* 84 (9): 4167-4173. https://doi.org/10.1021/ac300481k.

---. 2013. "High throughput analysis of atmospheric volatile organic compounds by thermal injection – isothermal gas chromatography – time-of-flight mass spectrometry." *Talanta* 103: 95-102. https://doi.org/10.1016/j.talanta.2012.10.013.

Wilson, Ryan B., W. Christopher Siegler, Jamin C. Hoggard, Brian D. Fitz, Jeremy S. Nadeau, and Robert E. Synovec. 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, Willem, and Jean Guilment. 1991. "Interactive self-modeling mixture analysis." *Analytical Chemistry* 63 (14): 1425–1432. https://doi.org/10.1021/ac00014a016.

Windom, Bret C., and Thomas J. Bruno. 2011. "Assessment of the Composition and Distillation Properties of Thermally Stressed RP-1 and RP-2: Application to Fuel Regenerative Cooling." *Energy & Fuels* 25 (11): 5200-5214. https://doi.org/10.1021/ef201077a.

Xu, Y., and R. Goodacre. 2018. "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning." *Journal of Analysis and Testing* 2 (3): 249-262. https://doi.org/10.1007/s41664-018-0068-2.

Yan, DanDan, Yong Foo Wong, Robert A. Shellie, Philip J. Marriott, Simon P. Whittock, and Anthony Koutoulis. 2019. "Assessment of the phytochemical profiles of novel hop (Humulus lupulus L.) cultivars: A potential route to beer crafting." *Food Chemistry* 275: 15-23. https://doi.org/10.1016/j.foodchem.2018.09.082.

Zajickova, Zuzana, and Ivan Špánik. 2019. "Applications of monolithic columns in gas chromatography and supercritical fluid chromatography." *Journal of Separation Science* 42 (5): 999-1011. https://doi.org/10.1002/jssc.201801071.

Zeeuw, Jaap de, and Jim Luong. 2002. "Developments in stationary phase technology for gas chromatography." *TrAC, Trends in Analytical Chemistry* 21 (9): 594-607. https://doi.org/10.1016/S0165-9936(02)00809-9.

Zeng, Zhongda, Jia Li, Helmut M. Hugel, Guowang Xu, and Philip J. Marriott. 2014. "Interpretation of comprehensive two-dimensional gas chromatography data using advanced chemometrics." *TrAC, Trends in Analytical Chemistry* 53: 150-166. https://doi.org/10.1016/j.trac.2013.08.009.

Zhang, Wanfeng, Shukui Zhu, Sheng He, and Yanxin Wang. 2015. "Screening of oil sources by using comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry and multivariate statistical analysis." *Journal of Chromatography A* 1380: 162-170. https://doi.org/10.1016/j.chroma.2014.12.068.

Zhang, Yan, Shiro Kajitani, Masami Ashizawa, and Yuso Oki. 2010. "Tar destruction and coke formation during rapid pyrolysis and gasification of biomass in a drop-tube furnace." *Fuel* 89 (2): 302-309. https://doi.org/10.1016/j.fuel.2009.08.045.

Zhao, Weixiang, and Cristina E. Davis. 2009. "Autoregressive model based feature extraction method for time shifted chromatography data." *Chemometrics and Intelligent Laboratory Systems* 96 (2): 252-257. https://doi.org/10.1016/j.chemolab.2009.02.010.

Zhou, Kai, Jinchao Jia, Xiaogang Li, Xiaodong Pang, Chunhua Li, Jun Zhou, Guohua Luo, and Fei Wei. 2013. "Continuous vinyl chloride monomer production by acetylene hydrochlorination on Hg-free bismuth catalyst: From lab-scale catalyst characterization, catalytic evaluation to a pilot-scale trial by circulating regeneration in coupled fluidized beds." *Fuel Processing Technology* 108: 12-18. https://doi.org/10.1016/j.fuproc.2012.03.018.

Zhou, Rui-Ze, Jie Jiang, Ting Mao, Ya-Song Zhao, and Yong Lu. 2016. "Multiresidue analysis of environmental pollutants in edible vegetable oils by gas chromatography–tandem mass spectrometry." *Food Chemistry* 207: 43-50. https://doi.org/10.1016/j.foodchem.2016.03.071.

Zoccali, Mariosimone, Peter Q. Tranchida, and Luigi Mondello. 2019. "Fast gas chromatography-mass spectrometry: A review of the last decade." *TrAC, Trends in Analytical Chemistry* 118: 444-452. https://doi.org/10.1016/j.trac.2019.06.006.

Zushi, Yasuyuki, Jonas Gros, Qingping Tao, Stephen E. Reichenbach, Shunji Hashimoto, and J. Samuel Arey. 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

Table A.1. List of analytes selected for the MCM study. All 200 compromise the Higher MV set, where analyte par MV ranges from 0-1000. The asterisk (*) indicates the 28 analytes in the Lower MV set, where analyte pair MV ranges from 0-150. The superscript plus sign ($^+$) indicates the 100 analytes in the analyte set applied in the preliminary simulation for the comparison of the MCM and the MCM 2.0, where analyte pair MV ranges from 0-1000.

Pentane$^+$	1-hexadecanol$^+$	Cyclohexylbenzene	2-octene-4-ol$^+$
Hexane$^+$	2-pentanol	sec-butyl benzene	3,4,5-trimethylheptane$^+$
Heptane$^+$	1-butanol$^+$	Phenanthrene	3,4-diethylhexane$^+$
Octane$^+$	1-eicosanol$^+$	p-xylene$^+$	4,5-dipropyloctane$^+$
Nonane$^+$	1-octadecanol$^+$	o-xylene$^+$	4-ethyl-2,2,6,6-tetramethylheptane$^+$
Decane$^+$	cyclohexanol$^+$	m-xylene$^+$	4-methylpentanamide
Undecane$^+$	neopentyl alcohol$^+$	1,2,4,5-tetrachlorobenzene*	4-propylheptane$^+$
Dodecane$^+$	2-methyl-2-propanol	1,2,4-trimethylbenzene	5-phenyl-1-pentanol
Tridecane$^+$	tert-amyl alcohol	Anisole	Acetonitrile*
Tetradecane$^+$	isobutyl alcohol$^+$	Benzophenone*	acetylsalicylic acid
Pentadecane$^+$	isopropyl alcohol*	1-hexyne$^+$	Acrylophenone
Hexadecane$^+$	benzyl alcohol	1-heptyne$^+$	Aniline
Pristane$^+$	1,2-propanediol	1-nonyne$^+$	Anthracene*
Octadecane$^+$	ethyl formate	5-decyne	Benzamide
Eicosane$^+$	methyl decanoate	Phenylacetaldehyde	Benzamidine*
Heptadecane$^+$	methyl caprylate	2-Carboxybenzaldehyde	Benzenethiol
Nonadecane$^+$	methyl salicylate	9-phenanthrenol*	Butanamine*
Chloroform*	ethyl salicylate	benzoylformic acid	butanoic acid$^+$
1-Chlorohexane$^+$	methyl laurate	Nonanoic Acid$^+$	Cyclopentadiene*
1-Bromohexane$^+$	methyl caproate	4-hydroxybenzoic acid*	Dibenzothiophene*
1-Bromoheptane$^+$	diethyl phthalate	Methyl hexadecanoate	Dimethylphenylphosphonate
1-Bromooctane$^+$	1-hexene$^+$	Acetanilide	Docosane$^+$
1,6-dichlorohexane$^+$	1-heptene$^+$	4-hydroxypyridine*	Ethanethiol
1-Chlorobutane$^+$	Dodecene$^+$	benzene-1,2,4-triol*	ethyl butanoate
1,1,1-trichloroethane*	1-undecene$^+$	2,3-dihydroxypyridine*	Heneicosane$^+$
1,2-dichloroethane*	2-butanone	Decanoic acid$^+$	Heptanal$^+$
Carbon tetrachloride$^+$	2-pentanone$^+$	Resorcinol	heptanoic acid$^+$

Methylcyclopentane^{+}	3-hexanone^{+}	2,3-dihydroxybenzoic acid	Hexanal^{+}
Cyclohexane^{+}	3-heptanone^{+}	methyl heptadecanoate*	hexanedioic acid^{+}
Methylcyclohexane^{+}	3-Octanone^{+}	Glucose	hexanoic acid^{+}
Cyclooctane^{+}	2-Nonanone^{+}	Tryptophan	isopentyl alcohol^{+}
Butylcyclohexane^{+}	2-decanone^{+}	Phenylalanine	Methanethiol*
Adamantane	2-undecanone^{+}	Tyrosine	N,N-dimethylhexanamine*
Bicyclohexyl	2-dodecanone^{+}	Serine	N,N-dimethylpentanamide
Cyclopentane^{+}	2-pentadecanone^{+}	Benzoic acid	N-benzophenylhydroxylamine
Cis-1,2-dimethylcyclohexane^{+}	2-hexanone^{+}	1,2,3,4,5-pentamethylcyclopentadiene	Nitrobenzene
2,2,4-trimethylpentane^{+}	2-heptanone^{+}	1,3,5-triethylbenzene*	N-methylhexanamine
2,3,4-trimethylpentane^{+}	Benzene	1,3,5-triphenylbenzene*	Octanoic acid^{+}
2-methylpentane^{+}	Toluene	1,3,6-heptatriene	Pentacosane^{+}
1-propanol	Naphthalene*	1,4-butanediol^{+}	Pentanal^{+}
2-butanol	Mesitylene	1,4-pentadiene	Pentanoic acid^{+}
1-pentanol^{+}	Ethylbenzene	1-bromo-2-ethylhexane^{+}	Phenol
1-hexanol^{+}	Butylbenzene	1-bromoadamantane*	Phenyl benzoate
2-heptanol^{+}	Isobutylbenzene	1-chloro-5-methylhexane^{+}	Phenylphosphonic acid
1-octanol^{+}	Tert-butyl benzene	2,3,6,7-tetramethyloctane^{+}	p-methoxyanisole
1-nonanol^{+}	Propylbenzene^{+}	2,4,6-trichlorophenol*	Propanoic acid
1-decanol^{+}	Chlorobenzene*	2-amino-1-butanol	Pyrene*
1-geraniol	Bromobenzene*	2-cyclohexyloctane^{+}	Tetracosane^{+}
1-dodecanol^{+}	1,3,5-trichlorobenzene	2-mercaptoethanol	2,2,2-trichloroacetic acid
1-tetradecanol^{+}	1,2,3-trichlorobenzene	2-methyl-1,3-pentanediol^{+}	Tricosane^{+}

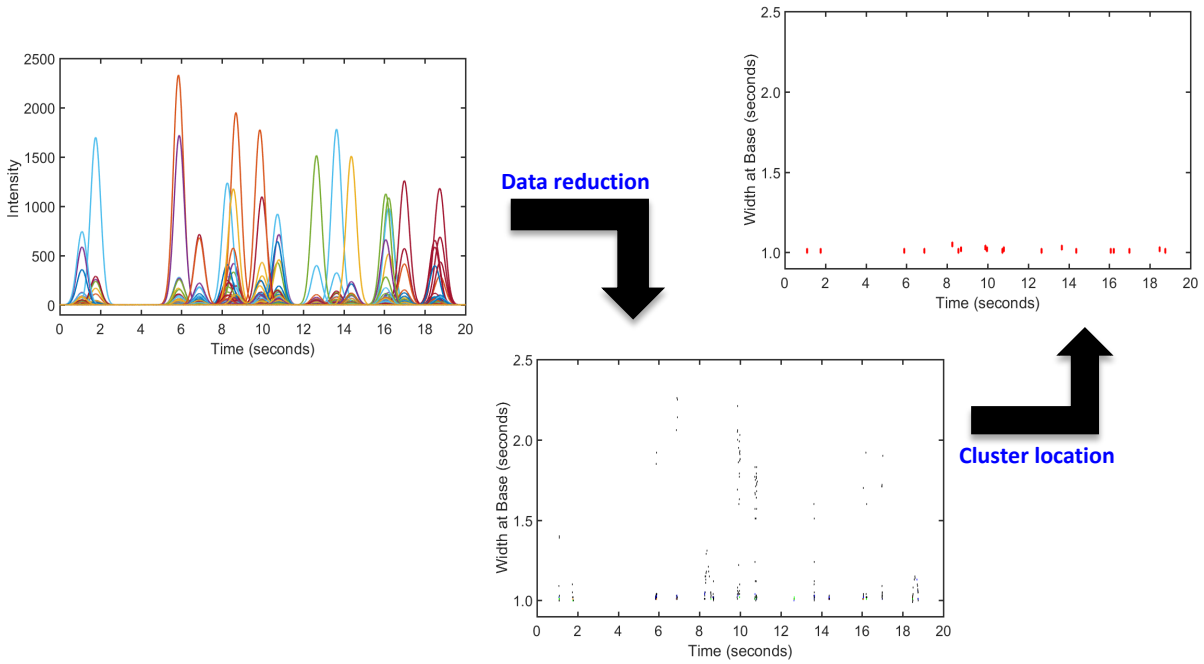


Figure A.1. Outline of the workflow of the MCM study. Chromatograms were simulated with the chosen number of components, *m*, corresponding to a saturation factor, α. A representative chromatogram with 20 components from the Lower MV analyte set and a saturation factor of 1 is depicted. The MCM consists of a data reduction step, which transforms the *m/z* peaks into cluster points in the 2D *m/z* cluster plot, followed by a cluster location step, which identifies the *m/z* deemed selective for each analyte cluster and encompasses them in a cluster box of user-specified dimensions (represented by the red rectangles).

Appendix B

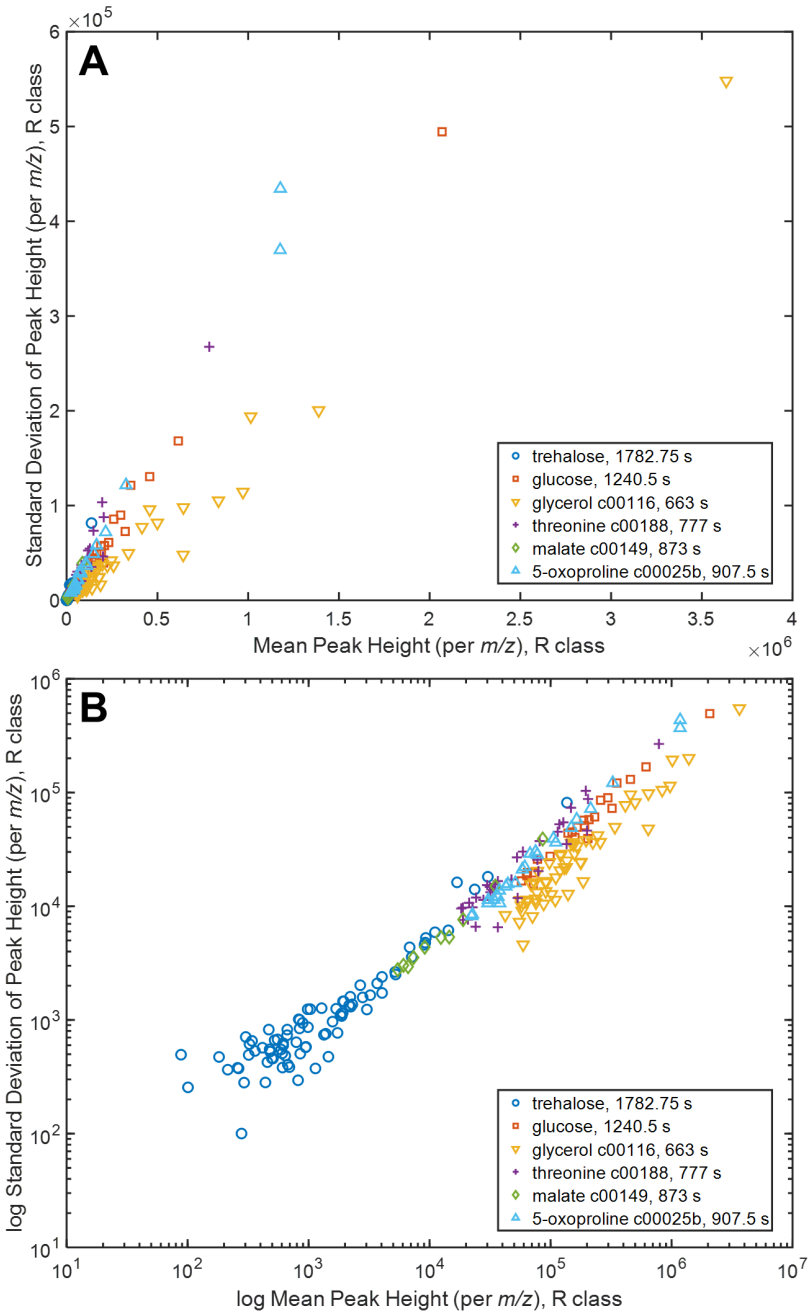


Figure B.1. Standard deviation versus mean of the peak height for six analytes in the repressed samples individually. (A) Scatter plot of the *m/z* (with at least 5 samples that passed the threshold) for: trehalose (blue circle), glucose (orange square), glycerol c00116 (yellow upside down triangle), threonine c00188 (purple plus sign), malate c00149 (green diamond), and 5-oxoproline c00025b (light blue triangle). (B) Logarithmically transformed standard deviation and peak height data from (A).

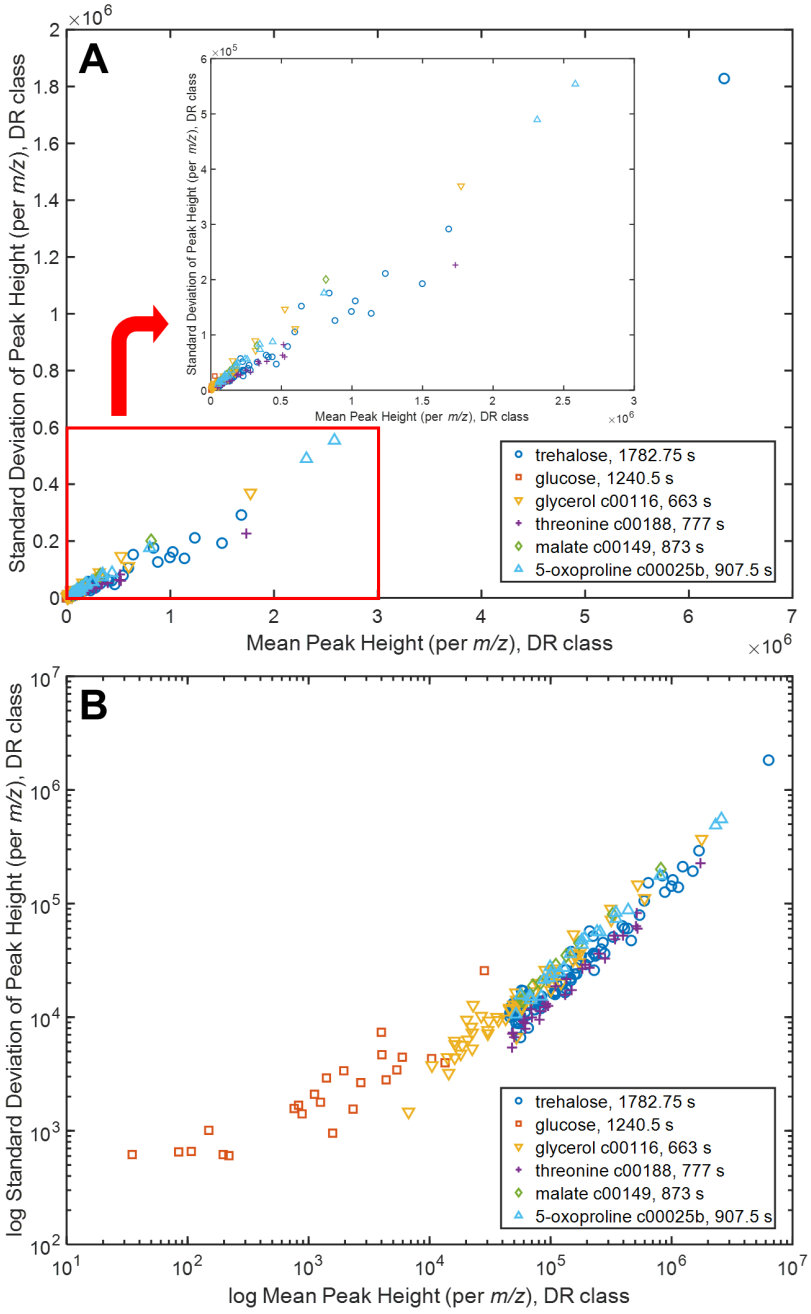
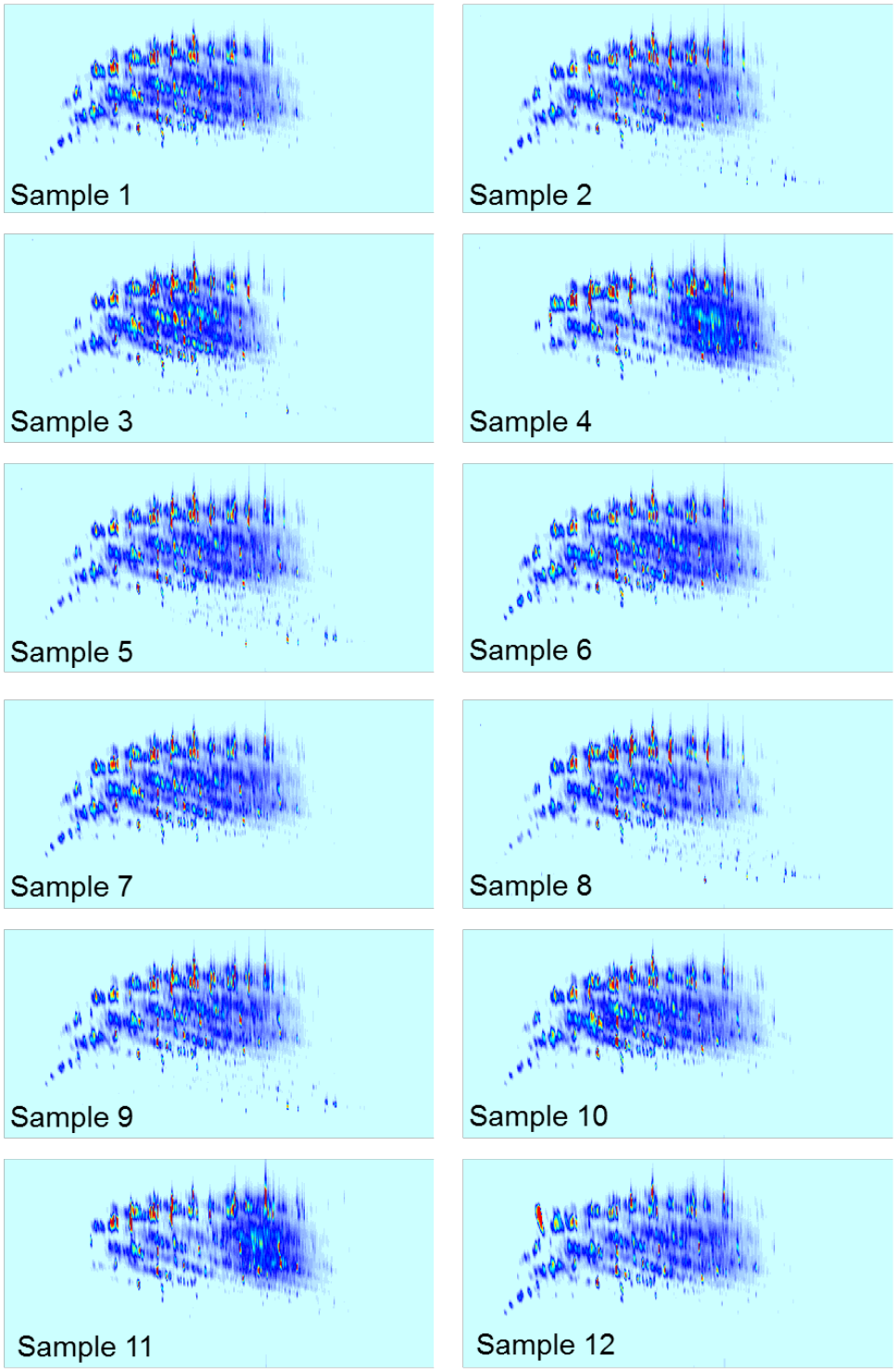
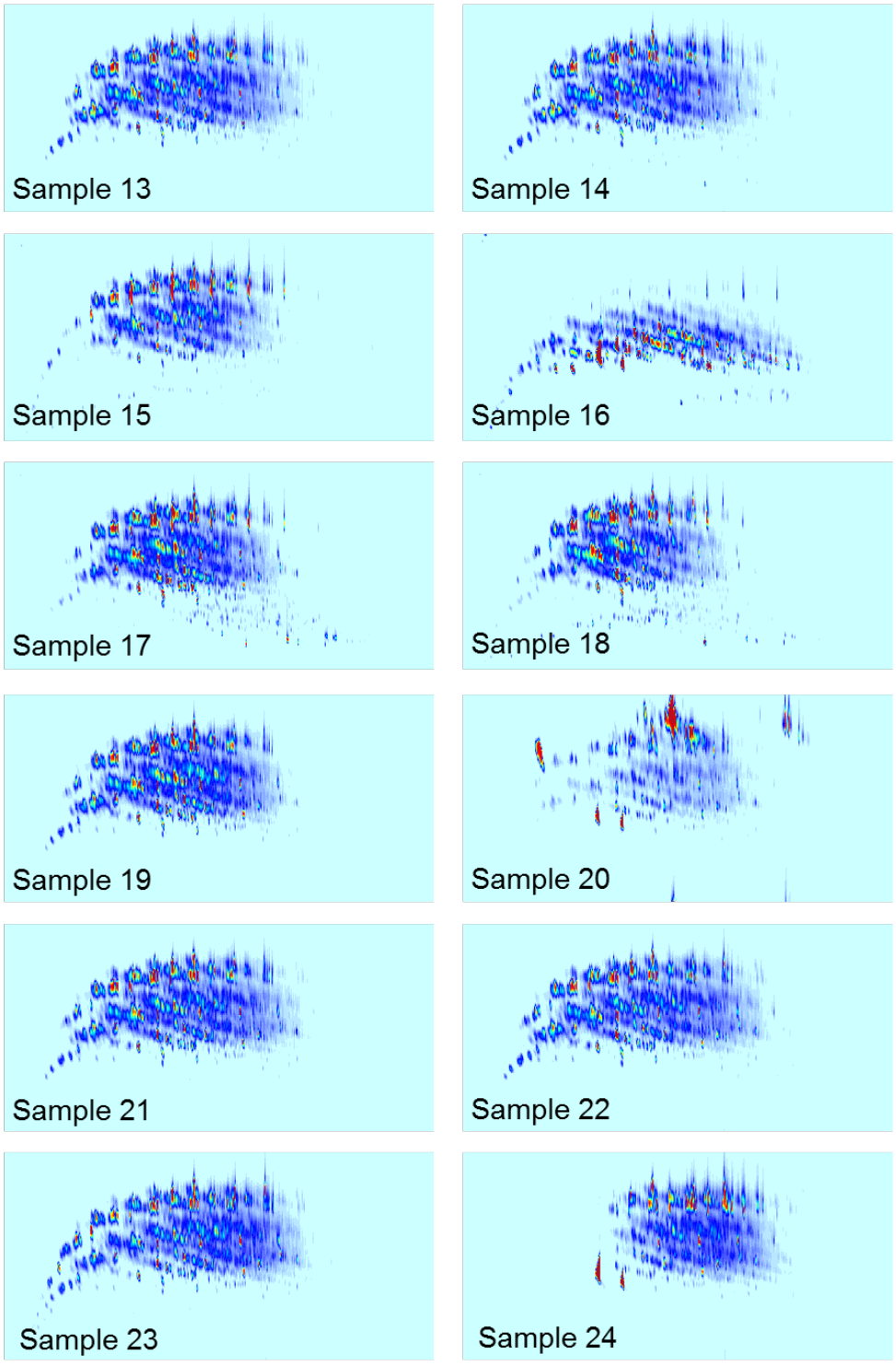
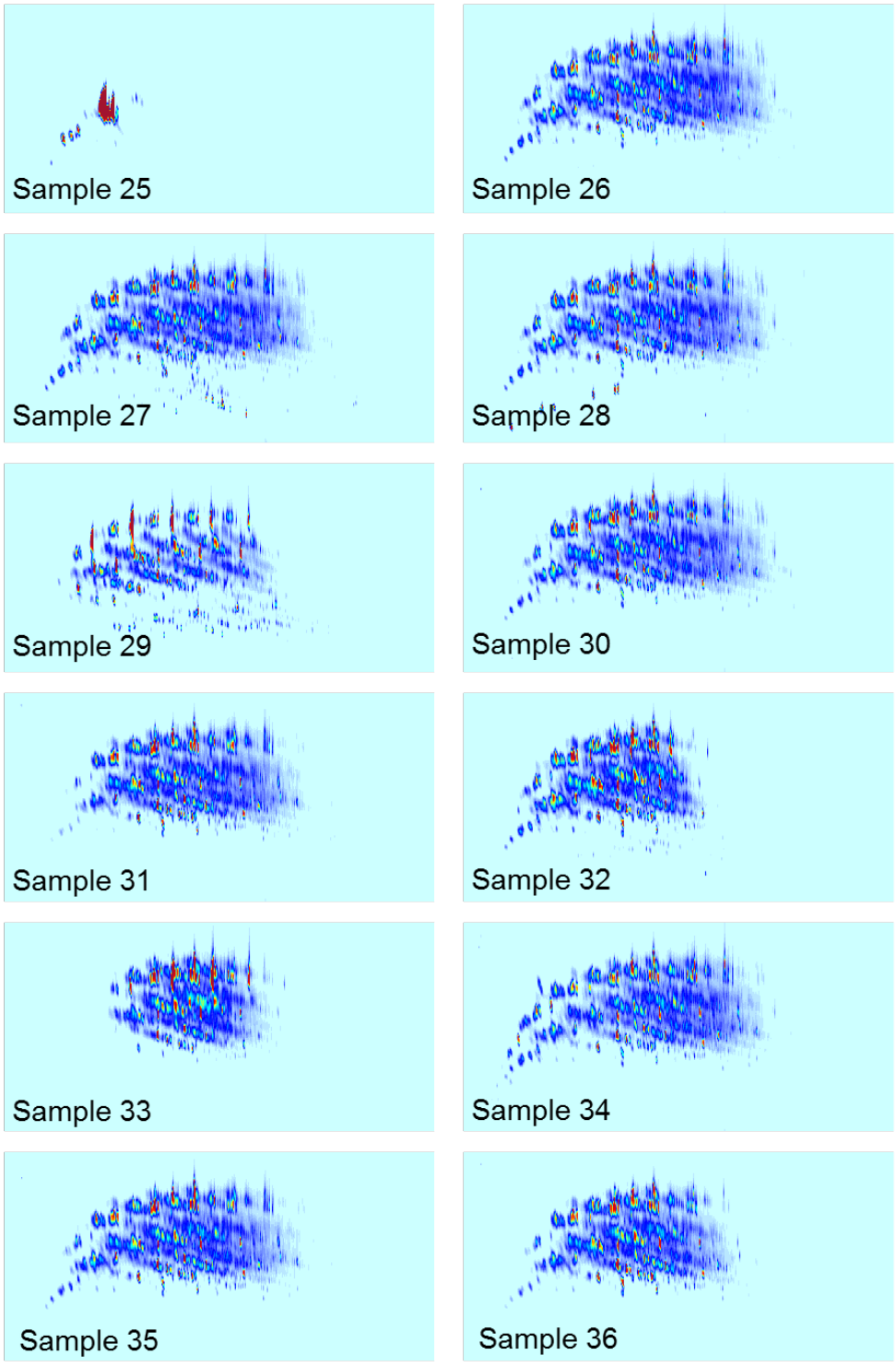


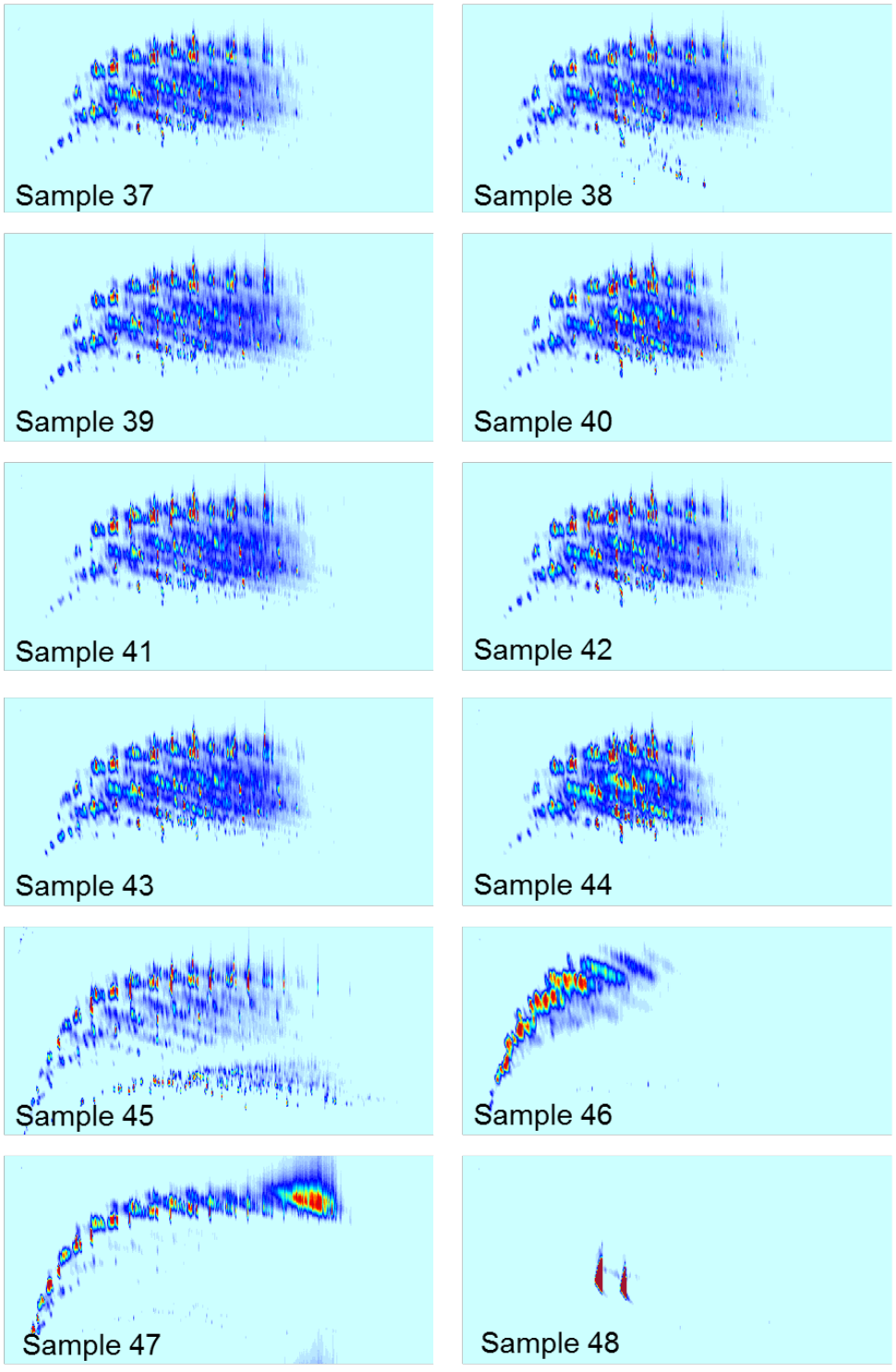
Figure B.2. Standard deviation versus mean of the peak height for six analytes in the derepressed samples individually. (A) Scatter plot of the *m/z* (with at least 5 samples that passed the threshold) for: trehalose (blue circle), glucose (orange square), glycerol c00116 (yellow upside down triangle), threonine c00188 (purple plus sign), malate c00149 (green diamond), and 5-oxoproline c00025b (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).

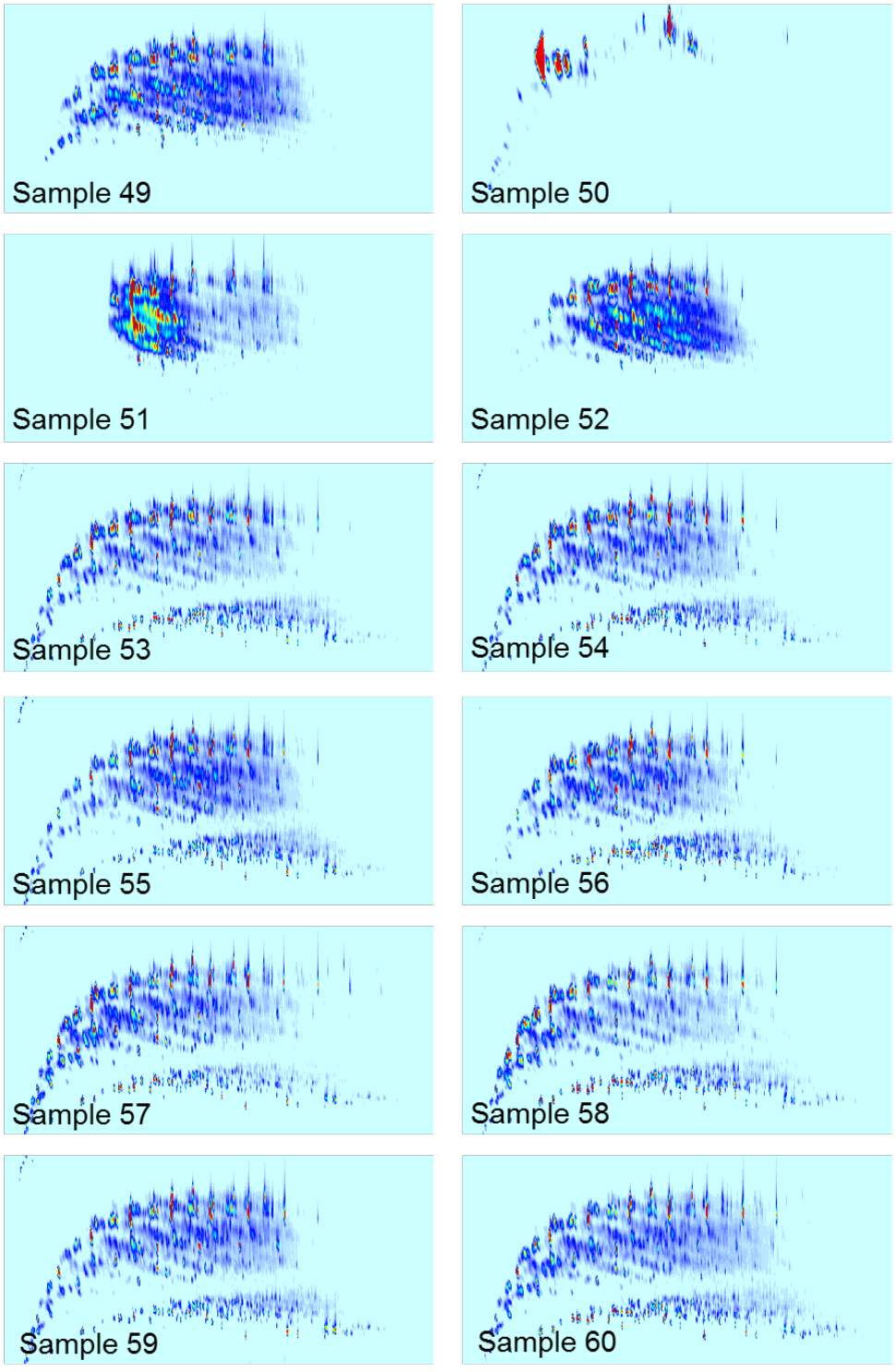
Appendix C

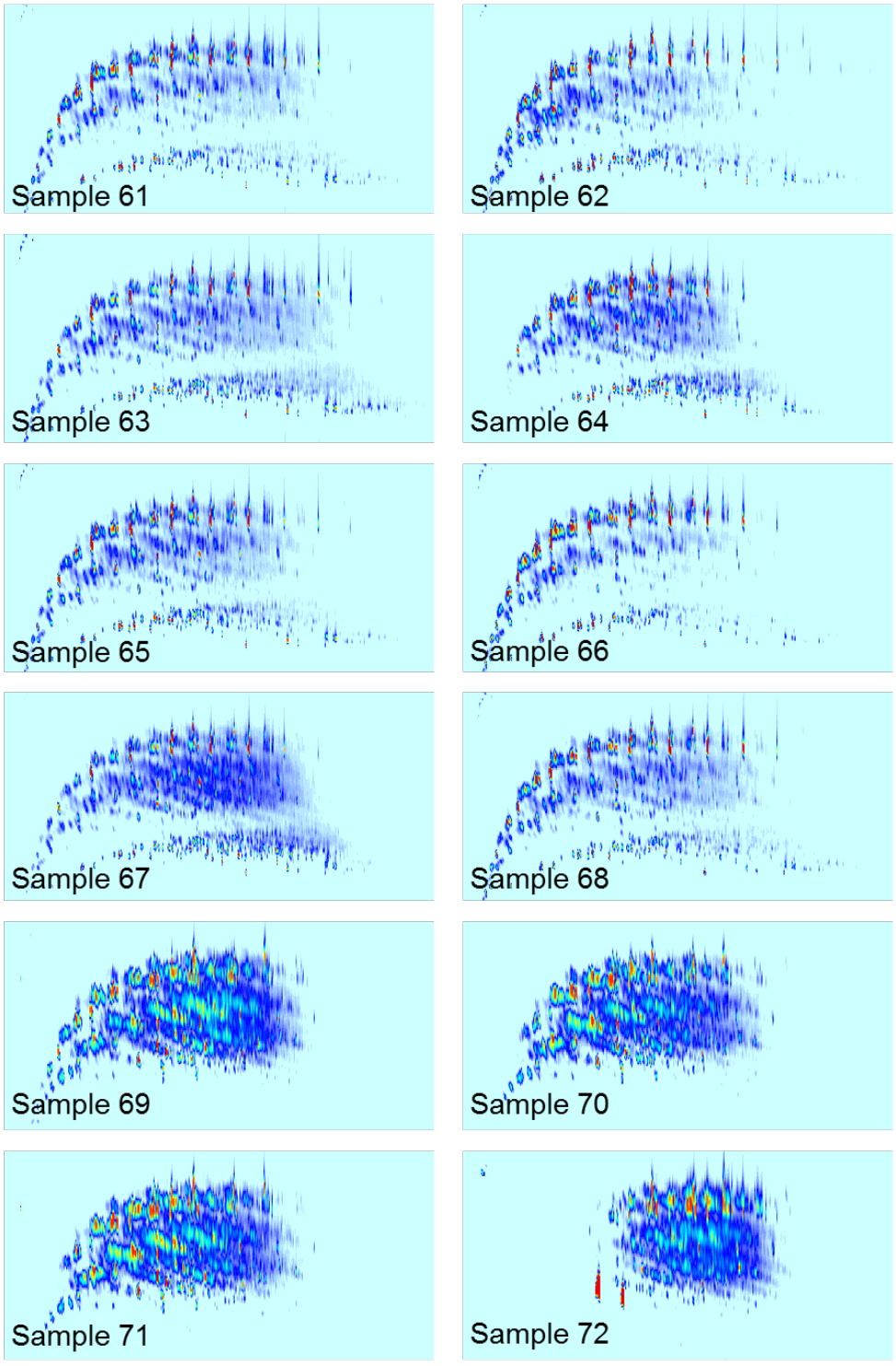

Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12


Sample 13
Sample 14
Sample 15
Sample 16
Sample 17
Sample 18
Sample 19
Sample 20
Sample 21
Sample 22
Sample 23
Sample 24


Sample 25
Sample 26
Sample 27
Sample 28
Sample 29
Sample 30
Sample 31
Sample 32
Sample 33
Sample 34
Sample 35
Sample 36


Sample 37
Sample 38
Sample 39
Sample 40
Sample 41
Sample 42
Sample 43
Sample 44
Sample 45
Sample 46
Sample 47
Sample 48


Sample 49
Sample 50
Sample 51
Sample 52
Sample 53
Sample 54
Sample 55
Sample 56
Sample 57
Sample 58
Sample 59
Sample 60


Sample 61
Sample 62
Sample 63
Sample 64
Sample 65
Sample 66
Sample 67
Sample 68
Sample 69
Sample 70
Sample 71
Sample 72

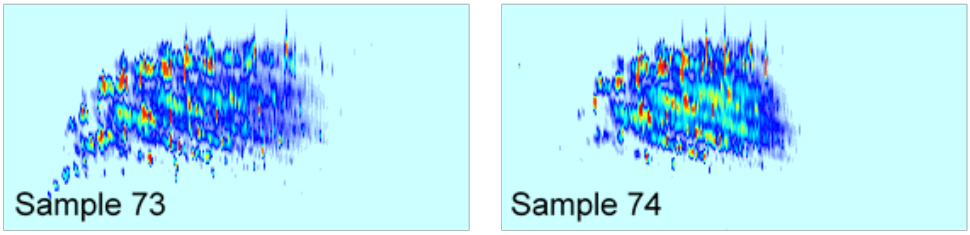


Figure C.1. GC×GC-TOFMS chromatograms of all 74 fuel samples. There are 8 samples (Samples 20, 24, 25, 46, 47, 48, 60, and 72) that are very dissimilar in their chemical fingerprint relative to the majority of the samples.

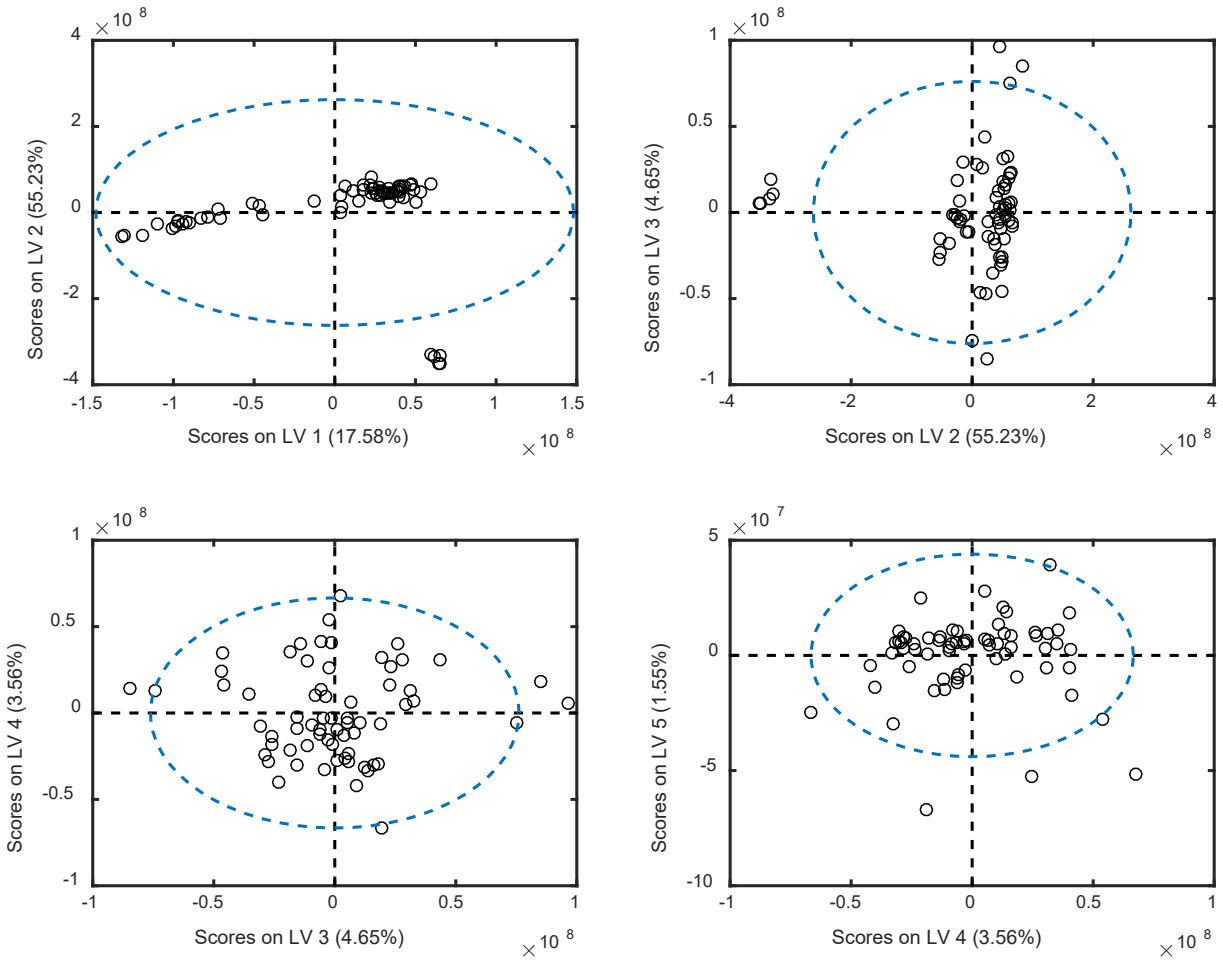


Figure C.2. Scores plots for the PLS model of viscosity with 66 fuels (replicate set one) and internal validation. (A) Scores on LV 2 versus scores on LV 1. (B) Scores on LV 3 versus scores on LV 2. (C) Scores on LV 4 versus scores on LV 3. (D) Scores on LV 5 versus scores on LV 4.

VITA

Kelsey L. Berrier was born in Baltimore, Maryland in 1993 and spent her childhood years in Maryland, South Carolina, and Virginia. After graduating from Briar Woods High School in Ashburn, VA in 2011, she attended James Madison University in Harrisonburg, VA. At JMU, she was a member of the Marching Royal Dukes, playing cymbals on the drumline for two years. She also engaged in undergraduate research in the environmental and analytical chemistry lab of Dr. Daniel Downey, where she was lucky enough to collect water samples as an excuse for hiking. She graduated *magna cum laude* in 2015 with a degree in Chemistry. Having completed her degree, Kelsey will begin working as a chemometrician with JP3 Measurement, LLC.