Advanced chemometric techniques for the analysis of complex samples using one- and two-dimensional gas chromatography coupled with time-of-flight mass spectrometry

Brooke C. Reaser

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Robert E. Synovec, Chair

Matthew F. Bush

Bo Zhang

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Advanced chemometric techniques for the analysis of complex samples using one- and two-dimensional gas chromatography coupled with time-of-flight mass spectrometry

Brooke C. Reaser

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

Gas chromatography is a powerful separation technique that alone, and when coupled with mass spectrometric detection, can provide detailed information regarding the chemical composition of complex mixtures. Advanced chemometric algorithms are often applied to the data generated from these gas chromatographic separations in order to glean additional meaningful information from large and complex data sets. This dissertation presents several research investigations conducted on the development, optimization, application and study of several chemometric algorithms applied to one- and two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (TOFMS). The two-dimensional mass cluster method and principal component analysis (PCA) were applied to a non-targeted investigation of the stable-isotope incorporation of metabolites present in the metabolome of the methylotrophic

bacteria *Methylobacterium extorquens* AM1 using gas chromatography time-of-flight mass spectrometry (GC-TOFMS). The area under the curve (AUC) of receiver operating characteristic (ROC) curves were used as quantitative metrics for the optimization of the tile-based Fisher ratio method using diesel fuel spiked with native and non-native analytes using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC × GC – TOFMS). This optimized algorithm was then applied to a process analytical chemistry (PAC) investigation into the source of catalyst yield reduction in an industrial polymerization plant. Finally, a GC-TOFMS simulation-based study determined the chemometric limit of resolution for deconvoluting analytes using multivariate curve resolution alternating least squares (MCR-ALS) and compared the results to expected theory surrounding the probability of peak overlap.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my graduate research advisor, Rob Synovec. Thank you for the opportunity to do research under your tutelage and for your support and guidance throughout the scientific process. Secondly, thank you to Doug Beussman of St. Olaf College, who was the first professor to encourage me to pursue a graduate degree in Chemistry.

Thank you to the many colleagues I had the privilege of working with over the last four years. To Brian and Brendon, thank you for passing on your wealth of knowledge and for collaborating with me on my first few research projects. To Dave, for walking through all of grad school with me from co-TAing during our first Autumn to our final project, thank you for your friendship and your numerous pep-talks. To Nate, thank you for sharing an office, a plethora of life advice, and several research projects with me over the last two years. To Sarah and Kelsey, for the coffee breaks and for inheriting my old projects and data, thank you for supporting me, uplifting me, and giving me fellow female scientists to look up to. Thank you to Chris, Dan and Nick for being great group members and scientific collaborators.

Finally, thank you to my friends and family outside of the University of Washington, who over the years offered support and encouragement. Thank you Schinria and Bethany for always being there, listening without judgment and telling me you're proud of me. Mom and Dad, thank you for giving me every opportunity to succeed, and for being unwavering sources of love and support. To Blair, thank you for being my sister, my friend and my strong female role model. And to my husband, Christopher, thank you for your unconditional love, for uprooting your whole life to come support me in Seattle, for listening to me rant about science, for taking care of things at home so I could focus on work, and for being my biggest cheerleader. Thank you.

DEDICATION

To my original nuclear family: Mom, Dad and Blair

for instilling in me from a very young age the belief that I could do whatever I set my mind to;

and

To my new nuclear family: Christopher and Jack

for your love and support.

Reasers, Rorabaughs and Goolsbees—Thank you.

Chapter 1. Introduction to Gas Chromatography and Chemometrics

1.1 INTRODUCTION TO CHROMATOGRAPHY

1.1.1 *History*

Chromatography was first developed by Russian botanist Mikhail S. Tswett in or around the year 1900. Using a chalk-filled tube and organic solvent, he separated the colored pigments present in a green leaf and coined the new separation technique "chromatography" [1,2]. While Tswett may have named the technique after his initial separation, in Greek *chroma* means "color" and *graphein* means "to write," Tswett's last name also means "color" in Russian, and his choice of name may have resulted from the scientist taking 'an opportunity to indulge his sense of humour' [3]. Since its discovery, chromatography has grown into a popular analytical technique used for a variety of applications and with many different definitions. The International Union of Pure and Applied Chemistry (IUPAC) defines chromatography as "a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (stationary phase) while the other (the mobile phase) moves in a definite direction" [4].

This IUPAC definition of chromatography is specific, and yet still excludes some types of chromatography utilized in the field. Due to the plethora of different separation techniques and mechanisms, chromatographers aim to classify different types of chromatography into categories in order to easily communicate amongst themselves and the greater scientific community. There are many different ways to classify the numerous types of chromatography that exist: by mobile phase (e.g., liquid chromatography has a liquid mobile phase), by separation mechanism (e.g., size exclusion chromatography separates analytes based on size), by analyte type (e.g., ion chromatography separates ions), and more. Needless to say, chromatography and the separation

sciences as a whole, include a wide range of techniques that can be applied to a variety of different sample and analyte types. For more information, the reader is directed to several books and articles that address the various types of chromatography [3,5–11].

Gas chromatography (GC) refers to all separation techniques that employ a gaseous mobile phase consisting of an inert gas, usually hydrogen, helium or nitrogen. The gaseous mobile phase flows through a column with a liquid or polymer stationary phase and separates volatile and semi-volatile components based on their affinity for the stationary phase. A plethora of stationary phases exist, all with slightly different separation mechanisms, the most common of which separate based on boiling point or polarity. Samples separated via GC can be liquid, solid or gas, but must be easily vaporized upon injection into the inlet of the GC and must be thermally stable to avoid decomposition. A GC separation can be monitored by any number of detectors, both universal and specific to the sample type.

Mass spectrometery (MS) is a popular analytical technique in its own right, but is also one such universal detector utilized extensively for GC analyses. When coupled together, the technique is referred to as gas chromatography-mass spectrometry or GC-MS. Many different types of mass spectrometers exist, but they all share a similar analysis mechanism: sample entering the MS is ionized, fragmented and separated based on mass-to-charge ratio (*m/z*). Molecules detected by MS tend to exhibit unique fragmentation patterns, or mass spectra, which can be used to identify unknown molecules through analysis of the mass spectrum by hand or matching to a library of known spectra. When used together, the GC-MS can separate and then identify pure components in a complex mixture. Due to its robust, reproducible and informative nature, GC-MS has been used in a variety of scientific fields, including but not limited to astrophysics [12], medicine [13], environmental science [14,15], forensics [16,17], biochemistry [18], and food science [19,20].

1.1.2 *Fundamentals*

The fundamental aspects of chromatography, including gas chromatography, includes both qualitative and quantitative aspects. Below are descriptions of a very small subset of the overall fundamentals, those important for the understanding of the research reported in this dissertation. For further understanding of these and other principles of chromatographic methods the reader is directed to some excellent resources [3,5].

As compounds elute from the column and are observed by the detector, they generate peaks that can be approximated by the Gaussian equation:

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

Here, σ is the standard deviation, μ is the mean, and the fraction prior to the exponent is equal to the peak area and can be replaced by a constant, A. The analyte retention time, t_R, is equal to μ, and is the time at which the maximum of an analyte peak is detected eluting from the column. The width of the peak is generally measured either at half height, W_h, or at baseline, W_b, where the peak width at baseline is defined as 4σ or ± 2σ from t_R. Figure 1.1 shows the relationship between these parameters as well as the fraction of the total peak height or signal that occurs at various distances from t_R. These peak heights are important for the accurate measurement of mass spectra. Most often the mass spectrum of an analyte is measured at the apex of the peak, t_R. However, if different analytes elute too closely together, it may be necessary to extract the mass spectrum from some distance to the left or right of t_R in order to ensure that the extracted mass spectrum is pure. In these cases, the fraction of the total peak height shown in Figure 1.1 becomes important to avoid the inclusion of too much noise. This concept is described further in Chapter 2 when it is utilized to extract pure spectra from slightly overlapped peaks in a process referred to as "re-indexing."

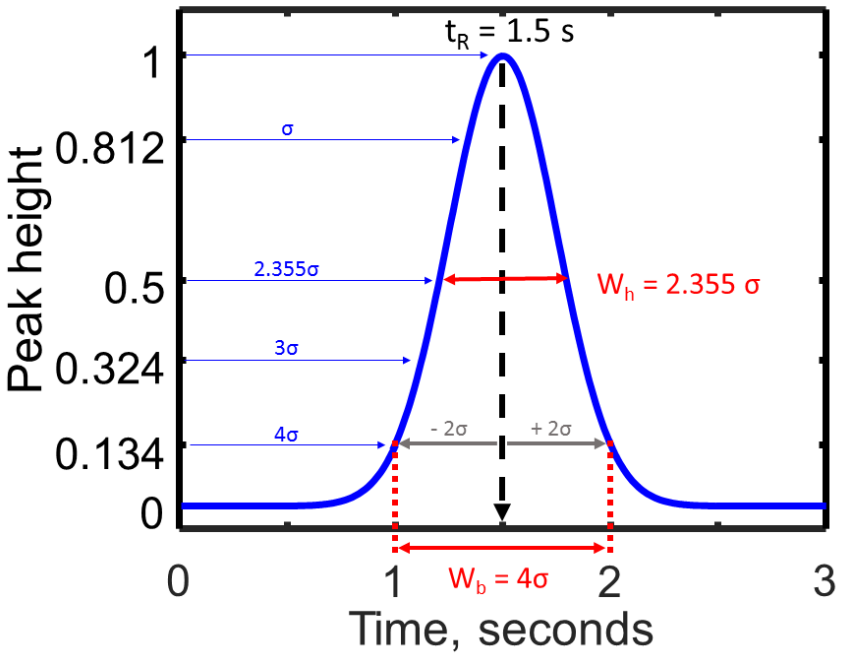

Figure 1.1 Gaussian peak with width at base and fractional peak heights identified.

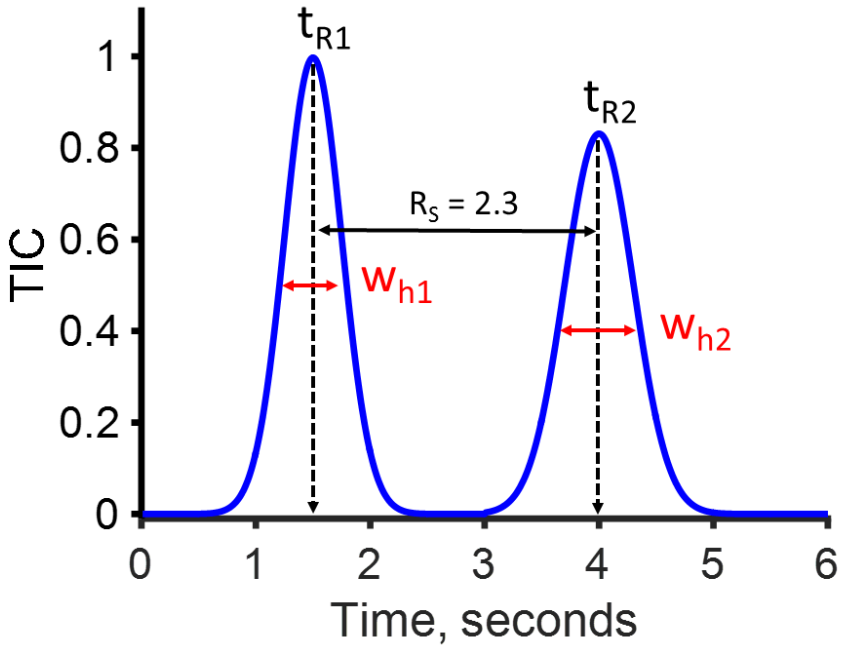

Figure 1.2. Visual representation of the resolution between two peaks.

Despite the separation power of GC, analyte peaks sometimes elute simultaneously or close enough together that part of the peaks overlap. Chromatographers quantify the extent of this peak overlap by reporting the resolution between peaks. Resolution is defined as:

$$R_S = \frac{2(t_{R2} - t_{R1})}{w_{b1} + w_{b2}} \quad (1.2)$$

Or, the difference in retention time between two peaks divide by the average width at base as seen in Figure 1.2. Peak widths are often measured at half height; the width at base can be approximated as 1.7 times the width at half height; $W_b = 1.7W_h$. Baseline resolution is defined as $R_S = 1.5$, but most chromatographers consider $R_S = 1.0$ adequate for quantitative analyses due to the minimal (~2%) peak overlap (Figure 1.4). $R_S = 0.5$ is the resolution at which two analytes will first appear as a single peak, and $R_S = 0.3$ is the resolution at which most deconvolution algorithms fail. Deconvolution will be discussed more in Section 1.2.2. These four resolution values are shown in Figure 1.3, where the red and blue curves each represent a single analyte, the thick black curve is what would be viewed by the detector response and the purple area is the peak overlap due to the coelution. Much time and energy is spent on improving the resolution of a separation, either by tweaking the various instrument parameters prior to separation, or through the use of chemometric techniques for the mathematical separation of analytes during the data anlysis stage.

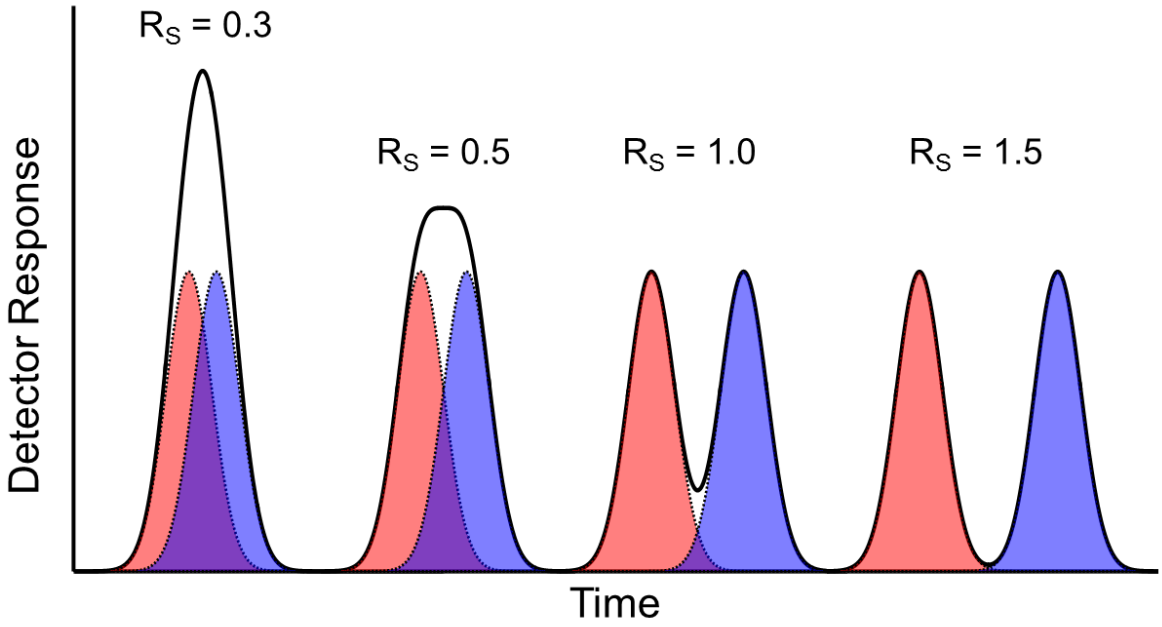

Figure 1.3. Representative peaks at various resolution values.

Just as pure analyte intensity varies at different locations in the Gaussian peak profile (Figure 1.1), so does the peak area. Peak areas at various σ values are shown in Figure 1.4, with ± σ in blue, ± 2σ in green and ± 3σ in red. As shown here, if the peak width at base (W_b) is taken to be 4σ or ± 2σ, then 95.44% of the total peak area is accounted for when quantifying. Peak overlap at higher R_S values will contain less peak area overlap than at lower R_S values. As mentioned above, R_S = 1.0 has about 2% peak overlap, this comes from the red region labeled 2.15% in Figure 1.4 being overlapped with another analyte eluting one second later. As R_S decreases and the peak overlap increases, the ability to deconvolute coeluting analytes becomes more important, especially when calculation of the peak area for quantification purposes is desired. Any peak overlap from interferent peaks could confound the quantification process of the target peak by adding excessive peak area where none would exist if the target and interferent(s) were resolved.

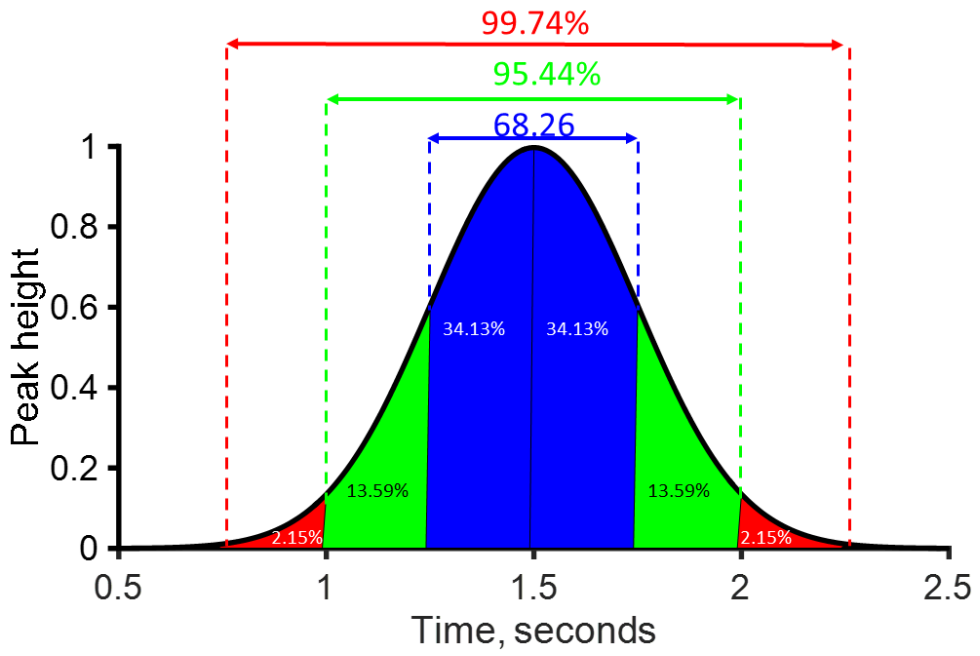

Figure 1.4. Visual representation of the areas under a Gaussian peak at various σ values.

1.1.3 *Comprehensive two-dimensional (2D) gas chromatography*

The 1980's brought the advent of comprehensive multidimensional separations [21]. Comprehensive two-dimensional (2D) separations include any technique where one separation dimension is serially coupled and complementary to a second separation dimension. The two separation dimensions need not be the same technique and ideally do not utilize the same separation mechanism. In comprehensive two-dimensional gas chromatography (GC × GC), a long first dimension column (20-30 m) and a short second dimension column (1-5 m) of complementary stationary phases are serially coupled with a modulator. The modulator may consist of a valve or a thermal modulator. Either type of modulator aims to collect small volumes of eluate from the first column, focus it into a small sample plug, and reinject it onto the second column. The first dimension separation generally follows a similar method to that used in one-dimensional GC, but the second dimension separation is usually only a few seconds long. The power of GC × GC lies

in the increased separation space given by the second dimension separation, which allows components that would otherwise elute simultaneously after the first column to be further separated prior to being detected.

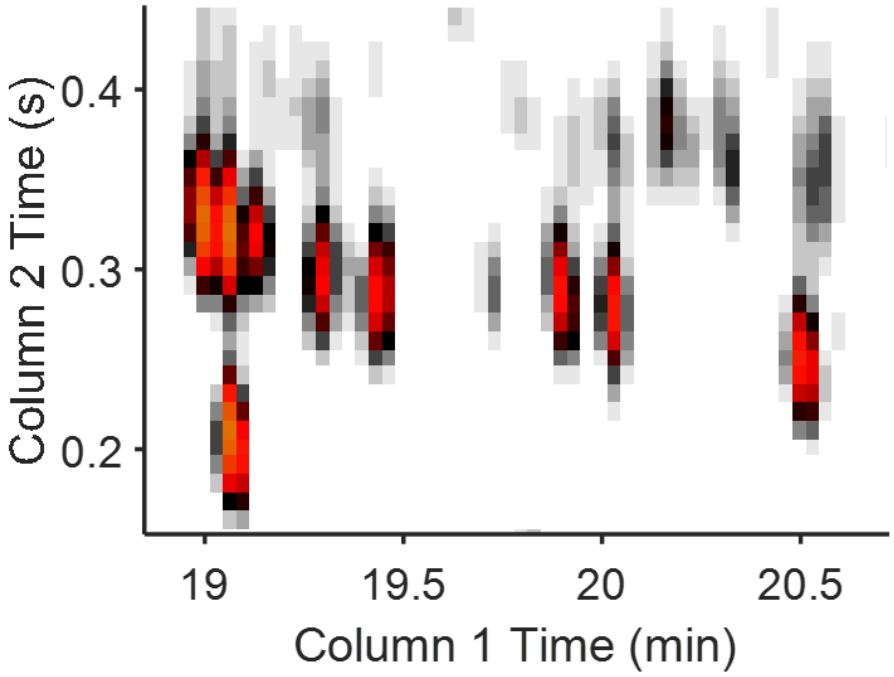

Figure 1.5. Section of a GC x GC chromatogram with each peak represented by a red contour.

While the data generated from GC × GC provides increased separation, it also has increased complexity. GC × GC data is generally viewed as a contour plot, as shown in Figure 1.5, where each red oval represents a single analyte peak, folded and projected in two dimensions. This can be seen in Figure 1.6, where Figure 1.6(A) is the unfolded 2D peak as seen by the detector, and Figure 1.6. Figure 1.6(B) is the folded contour of the same peak. The data shown in Figure 1.6 has a 2 second modulation or modulation period, P_M. The dotted lines in Figure 1.6(A) each represent a modulation, or the focusing and reinjecting of the eluate onto column 2. The P_M is the empty

space between these dotted lines where the two dimensional peaks elute. This is shown in Figure 1.6(B) as the length of the y-axis, also called the column 2 time.

As with GC, a variety of different detectors can be utilized with GC × GC, including mass spectrometry. However, due to the small peak widths on column 2, detectors with high data acquisition rates are generally required to avoid distortions of the peak shapes.

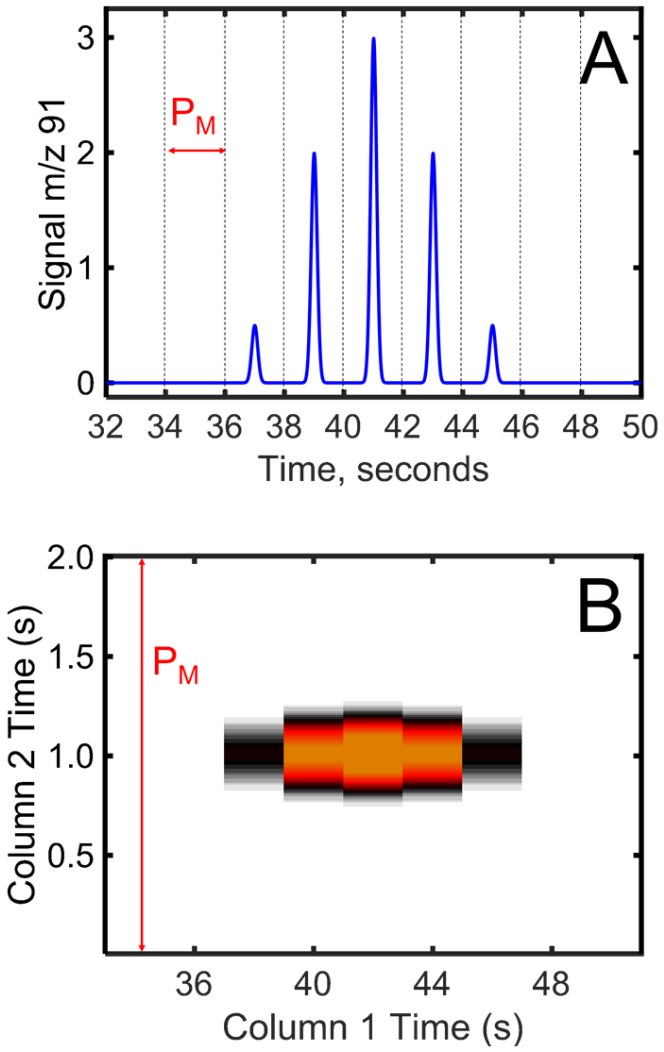

Figure 1.6. (A) Representation of GC × GC data as viewed by the detector; and (B), folded into 2D space.

1.2 Data Analysis and Advanced Chemometrics

1.2.1 *Introduction to chemometrics and data structure*

The data generated by one- and two-dimensional chromatography coupled with mass spectrometry data is inherently complex. But as technology improves, mass spectrometers continue to be able to detect at higher mass resolutions with faster scan rates, generating more data per chromatogram than ever before. The complexity of this data, whether it be second-order data from a one dimensional separation coupled with mass spectrometry or third-order data from a comprehensive two-dimensional separation with mass spectrometric detection, requires advanced chemometric techniques that can provide rapid and robust information about the system in question.

The IUPAC defines chemometrics as "the application of statistics to the analysis of chemical data (from organic, analytical or medicinal chemistry) and design of chemical experiments and simulations [4]." A field pioneered largely by Bruce Kowalski and Svante Wold, chemometric methods aim to glean useful information from complex data sets through the use of mathematics. Since its inception, chemometrics has been widely used throughout analytical chemistry [22], process analytical technology (PAT) [23], medicine [24], forensic science [23], metabolomics [25], and more [26].

Modern chemometrics includes a number of focus areas, including multivariate calibration; classification, pattern recognition and clustering; multivariate curve resolution; experimental design; structure modelling; and more [27,28]. Calibration methods include partial least squares (PLS) and principal component analysis (PCA); pattern recognition methods include hierarchical clustering analysis (HCA) and partial least squares discriminant analysis (PLS-DA); and curve resolution methods include multivariate curve resolution alternating least squares (MCR-ALS) and

parallel factor analysis (PARAFAC). These are just a few of examples of the many different chemometric methods available for analyzing data.

A discussion of chemometrics is not complete without first outlining data structures. For the purpose of this chapter, data structure will be place in terms of GC and GC × GC data, and discussed in more detail where appropriate. GC-MS data is generally taken to be bilinear, that is the MS dimension is linear, homogenous and independent of the separation dimension, and the data can be structured as a matrix of time × *m/z*. The data generated from GC × GC data is also bilinear, structured as a matrix of time on column 2 × time on column 1. However, data generated by GC × GC – MS data contains an extra dimension and can be viewed as a data cube of time on column 2 × modulations on column 1 × *m/z*. This data is trilinear, with each dimension being linear, homogenous and independent of the other two dimensions. The dimensionality, especially the bilinear and trilinear nature of the data is important for the efficacy of the chemometric methods employed. Several chemometric algorithms that were utilized for the various investigations presented in later chapters are described in the following section. These include both deconvolution techniques (Section 1.2.2) and comparative analyses (Section 1.2.3).

1.2.2 *Deconvolution Techniques*

Deconvolution is the process of mathematically resolving two or more analytes that are otherwise chromatographically overlapped. Deconvolution is generally performed during most analyses in order to obtain pure elution profiles from the chromatographic dimension(s) and pure mass spectra from the mass spectrometric dimension of coeluting analytes. This allows the analyst to better identify analytes based on their mass spectra and more accurately and precisely obtain pure component peak attributes such as peak height, area, and retention time for quantification. Generally speaking, peak integration methods suffice for quantification, but only when the peaks

are well resolved ($R_S \geq 1.0$) and have sufficient signal-to-noise ratios ($S/N \geq 10$). This is generally not the case for complex samples such as those found regularly analyzed by one- and two-dimensional chromatography coupled with mass spectrometric detection, so deconvolution methods are usually required. Many methods, with varying levels of automation, have been developed to perform deconvolution for specific instrument platforms and/or specific applications [29,30]. In addition to proprietary methods employed by commercial software packages, there are also several popular chemometric techniques for the deconvolution of chromatographically overlapped peaks. Below is a discussion of three of the most widely-utilized chemometrics techniques: classical least squares (CLS), multivariate curve resolution-alternating least squares (MCR-ALS), and parallel factor analysis (PARAFAC). Please note that for equations that deal with mathematically manipulating chromatograms in this chapter, lower case letters represent scalars, superscript T means the former variable is transposed, bold lower case letters represent vectors, bold uppercase letters represent matrices, and underlined bold uppercase letters represent 3D arrays.

Classical Least Squares (CLS)

Classical least squares (CLS) does not require a standard matrix for its deconvolution method as do some other deconvolution methods. However, it does require the user input the pure spectrum of each component in the mixture (**S**) as a matrix. The traditional CLS approach models the response-matrix **R**, containing the convoluted chromatographic profile coupled with mass spectrometric data as:

$$\mathbf{R} = \mathbf{CS} \tag{1.3}$$

where **C** is a matrix of pure component concentrations and **S** is a matrix of pure component mass spectra. CLS requires as many samples and at least as many spectra as there are components to be deconvoluted. Furthermore, *a priori* knowledge of either the identity of the components given by the pure mass spectra (**S**), or the pure chromatographic profile of each component (**C**). For example, if **S** is known, the concentrations of each component, **C**, can be determined from **R** by the following:

$$\mathbf{C} = \mathbf{R}\,\mathbf{S}^{\mathbf{T}}(\mathbf{S}\mathbf{S}^{\mathbf{T}})^{-1} \tag{1.4}$$

CLS can successfully deconvolute two-way data of a bilinear nature. We direct the reader to thorough reviews and discussion of CLS [31,32]. CLS is utilized as a deconvolution technique within the two-dimensional *m/z* cluster method in Chapter 2.

Multivariate Curve Resolution Alternating Least Squares (MCR-ALS)

MCR-ALS is another method for deconvoluting two-way, bilinear data by decomposing the observed data matrix **D**, that is the chromatographic data with mass spectrometric detection, into the two matrices **C**, containing the pure concentration profiles of the analytes and $\mathbf{S}^{\mathbf{T}}$, the pure mass spectra of *k* pure component species in the matrix, given by:

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

where **E** is the minimized error of residuals between the predicted and observed two-way chromatogram. We refer the reader to some excellent references [33–35].

Briefly, MCR requires an initial estimate of the number of components, *k*, and makes an initial estimate of the pure chromatograms and the pure spectra. The model tests for convergence, iterates **C** and **S**, tests for convergence again and continues until the convergence criterion is finally met. MCR-ALS outputs a pure elution profile and mass spectral vector for each pure component present in **D**. It is not necessary for the user to have *a priori* knowledge of either the mass spectrum or pure component profile of any of the components in the mixture, as is required of CLS. The user can simply define how many components are expected, choose appropriate constraints (such as unimodality, nonnegativity, or local rank) for each dimension, and define appropriate convergent criteria (such as threshold number of iterations, minimal value of **E**, or an improvement threshold for lack of fit between the predicted model and **D**). If there is some knowledge of one or more of the components present, the analyst can input an initial guess for **C** and/or **S** in addition to the previously mentioned inputs. While MCR-ALS is widely used for single dimension chromatography coupled with mass spectrometry, it is also widely employed for comprehensive two-dimensional gas chromatography-mass spectrometry, especially when trilinearity conditions are not met due to long modulation periods, with each second dimension modulated peak concatenated [11,12]. MCR-ALS is studied extensively in a theory-based simulation investigation described in Chapter 5.

Parallel Factor Analysis (PARAFAC)

Deconvolution techniques like CLS and MCR-ALS function well for bilinear data. However, trilinear data generated from GC × GC – TOFMS instrumentation must either be unfolded and concatenated to become bilinear, or must be deconvoluted using a different technique. Parallel factor Analysis (PARAFAC) is one such method that can leverage the

trilinearity of such data. PARAFAC can mathematically decompose the trilinear data into the individual first dimension and second dimension chromatographic peaks as well as extract the pure mass spectral vectors which recreate the original data matrix when multiplied. Mathematically, the PARAFAC model is described as:

$$\underline{\mathbf{R}} = \Sigma_{(\text{for } i = 1 \text{ to } n)} \mathbf{x}_i \otimes \mathbf{y}_i \otimes \mathbf{z}_i + \underline{\mathbf{E}} \quad (1.6)$$

where $\underline{\mathbf{R}}$ represents the detected data cube generated from the GC × GC – TOFMS separation. Each of the n total components is resolved into **x**, **y**, and **z** vectors which represent the ^{1}D, ^{2}D and mass spectral profiles respectively. Any remaining signal is retained in a residuals matrix, **E**. Numerous studies have been published on this technique and the reader is directed to these excellent resources for more details: PARAFAC [36–39] and an enhanced version called PARAFAC2 [40,40,41]. Employing PARAFAC is generally advantageous in cases where the trilinearity requirement is strictly followed [42,43]. As the signal response diverges from this criterion, usually due to long modulation periods or sub-optimal separation conditions, the PARAFAC result diverges from reality. It is also considered best practice for the analyst to create multiple models of varying numbers of factors and select the model which best represents the data. PARAFAC was utilized to extract pure spectra from low signal-to-noise analytes for identification purposes in the process analytical chemistry study discussed in Chapter 4.

1.2.3 *Comparative Analyses*

Comparative analysis encompasses all the techniques an analyst may employ to classify or sort data and/or results into useful groupings. In many cases data is collected in a manner where the classes are predefined as part of the experimental design. For example, a biologist subjects a

culture of cells to some perturbation while growing an equivalent culture which is not subjected to this perturbation. This kind of experiment naturally sorts into "control" and "experiment" classes. In other cases, these kinds of groups may not be known. In both circumstances a variety of classification and data reduction techniques are available. Generally, all these techniques center on using sample variance to determine classes or using predefined classes to identify sources of variance between those classes.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a classification method that performs an orthogonal transformation of possibly correlated variables into a new set of linearly uncorrelated variables known as principal components. The goal of PCA is to reduce the dimensionality of data sets with many variables such that the first few principal components describe the most variation of the original variables.

PCA is done by performing eigenvalue decomposition on the covariance matrix of the mean-centered data using singular value decomposition (SVD). This can be written:

$$\mathbf{X} = \mathbf{TDP}^{T} + \mathbf{E} \qquad (1.7)$$

Where **X** is the data matrix with m rows and n columns, with m equal to the number of samples and n equal to the number of variables. **T** is the matrix whose columns consist of the scores vectors, $\mathbf{t}_i$, and $\mathbf{P}^T$ is the matrix whose rows are loadings vectors, $\mathbf{p}^T_i$. **D** is the diagonal matrix with the diagonal elements equal to the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$, and **E** is the matrix of residuals. Each eigenvalue is equal to the variance in the data set associated with its corresponding eigenvector, or loadings vector p^T_i. The eigenvalues are ranked from greatest to least, with the

greatest eigenvalue corresponding to the first principal component of the model. The scores vectors in **T** describe the relationship between the samples in the original data and the loadings vectors in $\mathbf{P}^T$ describe the relationship between the variables in the original data set. It is worth noting that when viewing the plots formed by the scores and loadings matrices, it is meaningless to view the scores plots without the loadings plots and vice versa as they both describe important aspects of the original data matrix, **X**.

PCA is a popular chemometric method utilized not just in chemistry, but also in geology, statistics, electrical engineering, and medicine, to name a few [44–46]. We direct the readers interested in more about PCA to some very thorough reviews [22,44,47] and some interesting applications [48–53]. In the context of capillary chromatography, most often, PCA is performed on the pure, deconvoluted mass spectra of various samples; one-dimensional chromatograms at single, selective or total ion current mass channels; or on chromatographic signals generated from PARAFAC. PCA was performed on the purely extracted mass spectra to determine the extent of stable isotope uptake by metabolites in the investigation described in Chapter 2.

Fisher Ratio (F-ratio) Analysis

The Fisher Ratio (F-ratio) is a simple and useful statistical measurement of the variance within and between classes or populations. Mathematically, the F-ratio is described as

$$F - ratio = \frac{\sigma_1}{\sigma_2} \tag{1.8}$$

where variance 1 (σ_1) is the variance present between the classes and variance 2 (σ_2) is the sum of the variance within the classes. F-ratio values can range from zero, where there is no variance between two sample classes, and infinity, where the magnitude of the F-ratio scales with the

magnitude of the between class variance relative to the within class variance. Initially described by Fisher [54], the method is used in supervised, non-targeted analyses, where the experimental design allows for the arrangement of two populations or classes of samples but the identity of those components in the samples that distinguish between the samples is unknown.

The tile-based F-ratio method was developed in house as previously described [55,56], and has been utilized for various class comparisons on complex samples such as acid-altered diesel fuel [57] and a yeast metabolome [58]. The tile-based F-ratio method aims to find class-distinguishing chemical features in GC × GC – TOFMS using a novel tiling scheme in order to avoid the need for 2D alignment. This fast, robust F-ratio method has been able to leverage the density and complexity of GC × GC – TOFMS data in order to highlight chemical features with signal ratios as small as 1.06 [55]. The tile-based F-ratio method was employed for the investigations described in Chapters 3 and 4.

1.3 Overview of Following Chapters

The following chapters describe the bulk of the investigative research performed over the last several years. As indicated by the above introduction, they all employ gas chromatography coupled with mass spectrometry as the instrumental platform for the analysis of complex samples using advanced chemometrics. A brief abstract of each chapter is provided below.

1.3.1 *Chapter 2: Non-targeted determination of 13C-labeling in the Methylobacterium extorquens AM1 metabolome using the two-dimensional mass cluster method and principal component analysis*

A novel analytical workflow is presented for the analysis of time-dependent ^{13}C-labeling of the metabolites in the methylotrophic bacterium Methylobacterium extorquens AM1 using gas

chromatography time-of-flight mass spectrometry (GC-TOFMS). Using ^{13}C-methanol as the substrate in a time course experiment, the method aims to provide an accurate determination of the number of carbons converted to the stable isotope. The method also aims to extract a quantitative isotopic dilution time course profile for ^{13}C uptake of each metabolite labeled. This workflow combines both novel and traditional chemometric techniques, including the recently reported two-dimensional mass cluster plot method (2D *m/z* cluster plot method) as well as principal component analysis (PCA). It is hypothesized that the 2D *m/z* cluster plot method will effectively index all metabolites present in the sample and deconvolute metabolites at ultra-low chromatographic resolution (RS ≈ 0.04) as seen previously. Using the pure mass spectra extracted, PCA will be applied in two ways. Firstly, a PCA model will be created on the first and last time points of the time course experiment to determine and quantify the extent of ^{13}C uptake. Secondly, a PCA model will be performed on the full time course in order to quantitatively extract the time course profile for each metabolite. It is hypothesized that PCA will provide a novel, objective, time-efficient and quantitative method for the elucidation of ^{13}C incorporation by the metabolites.

1.3.2 *Chapter 3: Using ROC Curves to Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry*

A quantitative approach to optimize implementation of discovery-based software for comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) is described. The software performs a tile-based Fisher ratio (F-ratio) analysis, and facilitates a supervised non-targeted analysis based upon the experimental design to aid in the discovery of analytes with statistically different variances between sample classes. The quantitative approach for software optimization uses receiver operating characteristic

(ROC) curves. It is hypothesized that utilizing the area under the curve (AUC) for each ROC curve will provide a quantitative metric to optimize two key algorithm parameters: the signal-to-noise ratio (S/N) threshold of the data prior to calculating F-ratios at each *m/z* mass channel, and the number of these F-ratios per m/z used to calculate the average F-ratio of a tile. A total of 25 combinations of S/N threshold by number of *m/z* will be evaluated and compared using this AUC metric. Fifty analytes were spiked into a diesel fuel at two concentration levels to produce two sample classes that should in principle produce 50 positive instances in the ROC curves. It is hypothesized that through the evaluation of these 25 different parameterizations a "sweet spot" will be determined with a *S/N* threshold greater than the previously used 3, and a maximum number of the most chemically selective *m/z* that is some number fewer than utilizing all *m/z* above the *S/N* threshold in order to avoid using *m/z* with F-ratio signal due to spurious covariance between classes. It is the intent of this investigation to find a set of parameters that corresponds to a sizeable improvement in the discrimination of true positives relative to prior studies. Furthermore, optimization of these software parameters using this method should not depend upon *a priori* determination of the statistically correct number of positive instances in the sample classes. The AUC metric should be to be suitable for the evaluation of all data analysis methods that utilize the proper experimental design.

1.3.3 *Chapter 4: Application of the optimized tile-based Fisher ratio method to process analytical chemistry*

An application of the non-targeted tile-based Fisher ratio (F-ratio) method to a process analytical chemistry (PAC) investigation is presented. An industrial polymerization plant experiencing catalyst yield reduction due to one more unknown molecular poisons provided samples for analysis via comprehensive two-dimensional gas chromatography coupled with

time-of-flight mass spectrometry (GC × GC – TOFMS). Solvent samples were taken from two sampling points: Point I, after the feed tank and before the purification step, and Point II, after the purification step and before the polymerization process, on various date and times during "Excellent" polymerization processes. An additional sample was taken from Point II, after purification, during a "Bad" polymerization process. The solvent samples were analyzed with two goals in mind: first, to determine the difference between the composition of the solvent samples taken from Point I and Point II in the "Excellent" campaigns; and secondly, to determine what in the "Bad" sample might elucidate the cause of the catalyst yield reduction. It is hypothesized that the first question can be addressed through the application of the previously published tile-based F-ratio method to elucidate chemical features that differ between the two sample classes, that is, Point I and Point II. Using the information gleaned from the tile-based F-ratio method, the "Bad" sample will then be analyzed to determine what components correlated with the catalyst poison. From these results, it will be inferred that any molecules that appear in the "Bad" sample campaign at Point II that were otherwise successfully removed via the purification step prior to Point II in the "Excellent" campaigns were those correlating with the presence of the catalyst poison. This investigation will likely emphasize the importance of process analytical chemistry in the industry, and will provide a good argument for the use of on-line analysis during industrial processes for the rapid elucidation of issues on-site and in real time.

1.3.4 *Chapter 5: Chemometric Resolution Limit…*

An extensive theory and simulation-based investigation into the minimum chromatographic resolution at which a chemometric algorithm can successfully deconvolute coeluting GC-MS peaks is presented. Assuming that analyte peaks are randomly distributed across the separation space, a probabilistic description of peak overlap in GC-MS separations is presented. This theory will outline the probability of chemometric success for a deconvolution algorithm based on the saturation of the separation. The results of a simulation based study to investigate how the practical application of a deconvolution algorithm fits with the expected theory will then be presented. Simulations include the generation of chromatograms including one target analyte and one interferent analyte at various resolutions and two signal-to-noise levels. Applying, for this study alone, multivariate curve resolution-alternating least squares (MCR-ALS) as our chemometric algorithm, it is hypothesized that the minimum resolution at which the algorithm will successfully deconvolute peaks will be somewhat less than a resolution of 0.3, but greater than a resolution of 0.1. Once the minimum chemometric resolution is found, the results will be compared to the probabilistic theory primarily presented and the probability of overlap discussed based on saturation of the chromatogram.

1.4 REFERENCES

[1] K. Sakodynsky, M.S. Tswett—his life, J. Chromatogr. A. 49 (1970) 2–17. doi:10.1016/S0021-9673(00)93603-3.

[2] Deutsche Botanische Gesellschaft, Berichte der Deutschen Botanischen Gesellschaft, (1883) v.

[3] K. Robards, P.R. Haddad, P.E. Jackson, Principles and Practice of Modern Chromatographic Methods, Elsevier, Ltd., 2004.

[4] IUPAC, Compendium of Chemical Terminology (the "Gold Book")., 2nd ed., Oxford, 1997. http://goldbook.iupac.org/C01075.html.

[5] J.C. Giddings, Unified Separation Science, John Wiley & Sons, Inc., 1991.

[6] D.A. Skoog, F.J. Holler, S.R. Crouch, Principles of Instrumental Analysis, 6th ed., David Harris, 2007.

[7] J.W. Jorgenson, Capillary electrophoresis: An introduction, Methods. 4 (1992) 179–190. doi:10.1016/1046-2023(92)90033-5.

[8] F.L. Dorman, J.J. Whiting, J.W. Cochran, J. Gardea-Torresdey, Gas Chromatography, Anal. Chem. 82 (2010) 4775–4785. doi:10.1021/ac101156h.

[9] S. Dal Nogare, Gas Chromatography, Anal. Chem. 32 (1960) 19–25. doi:10.1021/ac60161a602.

[10] J.G. Dorsey, W.T. Cooper, B.A. Siles, J.P. Foley, H.G. Barth, Liquid Chromatography: Theory and Methodology, Anal. Chem. 70 (1998) 591–644. doi:10.1021/a1980022h.

[11] M.C. Henry, C.R. Yonker, Supercritical Fluid Chromatography, Pressurized Liquid Extraction, and Supercritical Fluid Extraction, Anal. Chem. 78 (2006) 3909–3916. doi:10.1021/ac0605703.

[12] C. Rodier, O. Vandenabeele-Trambouze, R. Sternberg, D. Coscia, P. Coll, C. Szopa, F. Raulin, C. Vidal-Madjar, M. Cabane, G. Israel, M.F. Grenier-Loustalot, M. Dobrijevic, D. Despois, Detection of martian amino acids by chemical derivatization coupled to gas chromatography: In situ and laboratory analysis, Adv. Space Res. 27 (2001) 195–199. doi:10.1016/S0273-1177(01)00047-3.

[13] J. Guo, Y. Shi, C. Xu, R. Zhong, F. Zhang, T. Zhang, B. Niu, J. Wang, Quantification of plasma myo-inositol using gas chromatography–mass spectrometry, Clin. Chim. Acta. 460 (2016) 88–92. doi:10.1016/j.cca.2016.06.022.

[14] M. del Olmo, A. González-Casado, N.A. Navas, J.L. Vilchez, Determination of bisphenol A (BPA) in water by gas chromatography-mass spectrometry, Anal. Chim. Acta. 346 (1997) 87–92. doi:10.1016/S0003-2670(97)00182-7.

[15] R. B. Gaines, G. S. Frysinger, C. M. Reddy, R. K. Nelson, 5 - Oil spill source identification by comprehensive two-dimensional gas chromatography (GC × GC) A2 - Wang, Zhendi, in: S.A. Stout (Ed.), Oil Spill Environ. Forensics, Academic Press, Burlington, 2007: p. 169–XI. http://www.sciencedirect.com/science/article/pii/B9780123695239500094 (accessed October 4, 2016).

[16] C. Brasseur, J. Dekeirsschieter, E.M.J. Schotsmans, S. de Koning, A.S. Wilson, E. Haubruge, J.-F. Focant, Comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry for the forensic study of cadaveric volatile organic compounds released in soil by buried decaying pig carcasses, J. Chromatogr. A. 1255 (2012) 163–170. doi:10.1016/j.chroma.2012.03.048.

[17] J.C. Hoggard, J.H. Wahl, R.E. Synovec, G.M. Mong, C.G. Fraga, Impurity Profiling of a Chemical Weapon Precursor for Possible Forensic Signatures by Comprehensive Two-Dimensional Gas Chromatography/Mass Spectrometry and Chemometrics, Anal. Chem. 82 (2010) 689–698. doi:10.1021/ac902247x.
[18] J. Börner, S. Buchinger, D. Schomburg, A high-throughput method for microbial metabolome analysis using gas chromatography/mass spectrometry, Anal. Biochem. 367 (2007) 143–151. doi:10.1016/j.ab.2007.04.036.
[19] T. Sasaki, E. Koshi, H. Take, T. Michihata, M. Maruya, T. Enomoto, Characterisation of odorants in roasted stem tea using gas chromatography–mass spectrometry and gas chromatography-olfactometry analysis, Food Chem. 220 (2017) 177–183. doi:10.1016/j.foodchem.2016.09.208.
[20] F. Magagna, L. Valverde-Som, C. Ruíz-Samblás, L. Cuadros-Rodríguez, S.E. Reichenbach, C. Bicchi, C. Cordero, Combined untargeted and targeted fingerprinting with comprehensive two-dimensional chromatography for volatiles and ripening indicators in olive oil, Anal. Chim. Acta. 936 (2016) 245–258. doi:10.1016/j.aca.2016.07.005.
[21] J.C. Giddings, TWO-DIMENSIONAL SEPARATIONS: CONCEPT AND PROMISE, Anal. Chem. 56 (1984) 1258A–1270A. doi:10.1021/ac00276a717.
[22] N. Kumar, A. Bansal, G.S. Sarma, R.K. Rawal, Chemometrics tools used in analytical chemistry: An overview, Talanta. 123 (2014) 186–199. doi:10.1016/j.talanta.2014.02.003.
[23] L.L. Simon, H. Pataki, G. Marosi, F. Meemken, K. Hungerbühler, A. Baiker, S. Tummala, B. Glennon, M. Kuentz, G. Steele, H.J.M. Kramer, J.W. Rydzak, Z. Chen, J. Morris, F. Kjell, R. Singh, R. Gani, K.V. Gernaey, M. Louhi-Kultanen, J. O'Reilly, N. Sandler, O. Antikainen, J. Yliruusi, P. Frohberg, J. Ulrich, R.D. Braatz, T. Leyssens, M. von Stosch, R. Oliveira, R.B.H. Tan, H. Wu, M. Khan, D. O'Grady, A. Pandey, R. Westra, E. Delle-Case, D. Pape, D. Angelosante, Y. Maret, O. Steiger, M. Lenner, K. Abbou-Oucherif, Z.K. Nagy, J.D. Litster, V.K. Kamaraju, M.-S. Chiu, Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review, Org. Process Res. Dev. 19 (2015) 3–62. doi:10.1021/op500261y.
[24] A.M. Yehia, H.M. Mohamed, Chemometrics resolution and quantification power evaluation: Application on pharmaceutical quaternary mixture of Paracetamol, Guaifenesin, Phenylephrine and p-aminophenol, Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 152 (2016) 491–500. doi:10.1016/j.saa.2015.07.101.
[25] R.E. Mohler, B.P. Tu, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Identification and evaluation of cycling yeast metabolites in two-dimensional comprehensive gas chromatography–time-of-flight-mass spectrometry data, J. Chromatogr. A. 1186 (2008) 401–411. doi:10.1016/j.chroma.2007.10.063.
[26] B.K. Lavine, J. Workman, Chemometrics, Anal. Chem. 85 (2013) 705–714. doi:10.1021/ac303193j.
[27] S. Wold, M. Sjöström, Chemometrics, present and future success, Chemom. Intell. Lab. Syst. 44 (1998) 3–14. doi:10.1016/S0169-7439(98)00075-6.
[28] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics, Chem. Rev. 110 (2010) 4582–4605. doi:10.1021/cr900394n.
[29] K. Hiller, J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, D. Schomburg, MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis, Anal. Chem. 81 (2009) 3429–3439. doi:10.1021/ac802689c.

[30] W. Niu, E. Knight, Q. Xia, B.D. McGarvey, Comparative evaluation of eight software programs for alignment of gas chromatography–mass spectrometry chromatograms in metabolomics experiments, J. Chromatogr. A. 1374 (2014) 199–206. doi:10.1016/j.chroma.2014.11.005.
[31] R. Kramer, Chemometric Techniques for Quantitative Analysis, Marcel Dekker, Inc., 1998.
[32] 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. doi:10.1021/ac5004344.
[33] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: New features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12. doi:10.1016/j.chemolab.2014.10.003.
[34] H. Parastar, N. Akvan, Multivariate curve resolution based chromatographic peak alignment combined with parallel factor analysis to exploit second-order advantage in complex chromatographic measurements, Anal. Chim. Acta. 816 (2014) 18–27. doi:10.1016/j.aca.2014.01.051.
[35] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, J. Chromatogr. A. 1218 (2011) 8411–8422. doi:10.1016/j.chroma.2011.09.057.
[36] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. doi:10.1016/S0169-7439(97)00032-4.
[37] R.A. Harshman, M.E. Lundy, PARAFAC: Parallel factor analysis, Comput. Stat. Data Anal. 18 (1994) 39–72. doi:10.1016/0167-9473(94)90132-5.
[38] J.C. Hoggard, R.E. Synovec, Parallel Factor Analysis (PARAFAC) of Target Analytes in GC × GC−TOFMS Data: Automated Selection of a Model with an Appropriate Number of Factors, Anal. Chem. 79 (2007) 1611–1619. doi:10.1021/ac061710b.
[39] J.C. Hoggard, W.C. Siegler, R.E. Synovec, Toward automated peak resolution in complete GC × GC–TOFMS chromatograms by PARAFAC, J. Chemom. 23 (2009) 421–431. doi:10.1002/cem.1239.
[40] 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. doi:10.1016/j.trac.2008.05.011.
[41] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling, J. Chromatogr. A. 1216 (2009) 4020–4029. doi:10.1016/j.chroma.2009.02.049.
[42] 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. doi:10.1016/j.aca.2015.02.040.
[43] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, Talanta. 160 (2016) 624–635. doi:10.1016/j.talanta.2016.08.005.
[44] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. doi:10.1016/0169-7439(87)80084-9.
[45] B.C. Reaser, S. Yang, B.D. Fitz, B.A. Parsons, M.E. Lidstrom, R.E. Synovec, Non-targeted determination of 13C-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. doi:10.1016/j.chroma.2015.12.088.
[46] V.G. Uarrota, R. Moresco, B. Coelho, E. da C. Nunes, L.A.M. Peruch, E. de O. Neubert, M. Rocha, M. Maraschin, Metabolomics combined with chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (Manihot esculenta Crantz) roots during postharvest physiological deterioration, Food Chem. 161 (2014) 67–78. doi:10.1016/j.foodchem.2014.03.110.
[47] B.M. Wise, N.B. Gallagher, The process chemometrics approach to process monitoring and fault detection, J. Process Control. 6 (1996) 329–348. doi:10.1016/0959-1524(96)00009-1.
[48] E.M. Humston, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Time-Dependent Profiling of Metabolites from Snf1 Mutant and Wild Type Yeast Cells, Anal. Chem. 80 (2008) 8002–8011. doi:10.1021/ac800998j.
[49] C.G. Fraga, G.A. Pérez Acosta, M.D. Crenshaw, K. Wallace, G.M. Mong, H.A. Colburn, Impurity Profiling to Match a Nerve Agent to Its Precursor Source for Chemical Forensics Applications, Anal. Chem. 83 (2011) 9564–9572. doi:10.1021/ac202340u.
[50] J.S. Nadeau, R.B. Wilson, J.C. Hoggard, B.W. Wright, R.E. Synovec, Study of the interdependency of the data sampling ratio with retention time alignment and principal component analysis for gas chromatography, J. Chromatogr. A. 1218 (2011) 9091–9101. doi:10.1016/j.chroma.2011.10.031.
[51] P.J. Dunlop, C.M. Bignell, J.F. Jackson, D.B. Hibbert, Chemometric analysis of gas chromatographic data of oils from Eucalyptus species, Chemom. Intell. Lab. Syst. 30 (1995) 59–67. doi:10.1016/0169-7439(95)00036-4.
[52] N.A. Sinkov, J.J. Harynuk, Three-dimensional cluster resolution for guiding automatic chemometric model optimization, Talanta. 103 (2013) 252–259. doi:10.1016/j.talanta.2012.10.040.
[53] 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. doi:10.1016/j.chroma.2006.06.087.
[54] R.A. Fisher, Statistical Methods for Research Workers, 14th ed., Oliver and Boyd, 1970.
[55] 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. doi:10.1016/j.talanta.2013.06.038.
[56] 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. doi:10.1021/ac504472s.
[57] 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. doi:10.1016/j.chroma.2016.02.067.
[58] 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. doi:10.1016/j.chroma.2016.06.067.

Chapter 2. Non-targeted determination of ^{13}C-labeling in the *Methylobacterium extorquens* AM1 metabolome using the two-dimensional mas cluster method and principal component analysis[1]

2.1 INTRODUCTION

Metabolomics is one of the many "-omics" fields, including proteomics, genomics, and transcriptomics that are of great interest to researchers that employ analytical chemistry methods [1]. The term metabolomics encompasses all studies of metabolites, the small molecules that make up an organism's metabolome and the cellular processes within. The metabolome consists of a wide variety of molecules, organic and inorganic, of varying sizes, polarities and volatilities [2,3]. Metabolomics provides a snapshot of the molecular content of a cell at a given moment. To provide further insight, time course or flux experiments aim to explore the dynamic and ever-changing nature of cellular metabolites as they move through the metabolic cycles within the organism. Time course studies provide additional insight into metabolic pathways and enzymatic activities. These experiments often utilize stable isotopes for labeling (^{13}C, ^{15}N, ^{2}H, etc.) to elucidate how concentrations [4], reaction rates [5], or isotopomer distributions [6] change over time [7].

Methyloacterium extorquens AM1 is a methylotrophic bacteria of great interest for its possible uses in the production of biofuels, valuable chemicals and catalysts [3]. Like all methylotrophic bacteria, *M. extorquens* AM1 is capable of using single carbon sources such as methane or methanol as its sole carbon source for all of its carbon and energy needs. Although it

[1] This chapter has been reproduced from B.C. Reaser, S. Yang, B.D. Fitz, B.A. Parsons, M.E. Lidstrom, and R.E. Synovec *Journal of Chromatography A*, 1432 (2016) 111-121.

is one of the most extensively studied methylotrophic bacteria, there exist metabolic cycles in *M. extorquens* AM1 that are still not completely understood, making it an ideal system for isotopic labeling-based time course studies. The pathway-specific rearrangement of molecules in the metabolome is highlighted through the use of an isotopic tracer, specifically ^{13}C in this study, facilitating better understanding of the metabolic processes of pathways or enzymes in a system [7–10]. For *M. extorquens* AM1, ^{13}C-labeled methanol is incorporated by metabolites in the 0 to 60 minute time frame, constituting the time course of our investigation.

The complex nature of the samples in time-dependent studies of the metabolome requires powerful analytical tools that can differentiate between stable isotopes, such as nuclear magnetic resonance (NMR) and mass spectrometry (MS) [9]. MS is often coupled with a separation technique, usually liquid chromatography (LC) [11] or gas chromatography (GC) [12,13]. GC-MS is one of the more popular analytical tools for complex samples such as those seen in metabolomics studies [10,14,15]. GC offers the ability to separate volatile and semi-volatile analytes in complex samples for better structural elucidation via the unique fragmentation pattern of MS with electron impact ionization [16–18]. Chemical derivatization is often required to make non-volatile metabolites ready for GC analysis [19]. Time-of-flight mass spectrometry (TOFMS) is an especially effective detector for GC due to the high acquisition rate (100-500 spectra/s), which allows for better deconvolution and analysis of chromatographically overlapped peaks [20].

Due to the significant data density that results from the use of GC-TOFMS as an instrumental platform, many analytical investigations employ chemometrics, which utilizes mathematics to extract useful information from complex data sets [21]. Commercial software packages are available to implement chemometric techniques for interpretation of complex data

sets, including software specifically for the analysis of metabolomics data [12,22,23]. Most of these software packages focus on non-targeted alignment and deconvolution or identification of metabolites in the data set, but many exist as "black boxes" allowing little room for input or interpretation by the researcher. Despite the prevalence of available software, the use of advanced chemometrics is limited in metabolomics. Use of GC-TOFMS for metabolomics studies, especially the use of stable isotope labeling for time course investigations, should provide several challenges for the analysis of the complex data sets. These challenges include large amounts of data, excessive chromatographic peak overlap, and difficulty in accurate extraction of pure mass spectra for identification of metabolites of interest. In order to successfully extract pure mass spectra, accurate deconvolution of coeluting metabolites is required. Many software packages provide deconvolution steps, but few explain how the deconvolution is mathematically performed or at what chromatographic resolution the algorithm fails. Deconvolution of coeluting/overlapped analytes down to a chromatographic resolution, R_S, of 0.1 has previously been obtained using advanced chemometrics in [10], however analytes elute at even lower chromatographic resolution with the extreme case of complete overlap. The challenge of coelution as it exists in time-dependent ^{13}C-labeling experiments requires chemometric deconvolution at a lower resolution for a multiplicity of samples (various time points and injection replicates) using less cumbersome but equally robust chemometric techniques, even in especially challenging examples, when more than two metabolites coelute. Successful identification of stable isotope incorporation (or lack thereof) adds an additional challenge. Few mass spectral databases include libraries for isotopically labeled metabolites for mass spectral matching, compounding the already difficult task of studying the hundreds of metabolites that have yet to be identified. Without a library, other methods must be proposed for the accurate identification of metabolites that have incorporated

the stable isotope over metabolites that have not. A successful time-dependent stable isotope labeling experiment would ideally address all of these challenges by incorporating the following into the overall method: (1) metabolite retention time indexing, (2) deconvolution of coeluting peaks, (3) extraction and normalization of pure mass spectra, (4) identification of metabolites that fully incorporated the stable isotope over the time course, (5) identification of the number of carbon atoms in these metabolites that became isotopically labeled, and (6) quantitative determination and visualization of the time course profile of the ^{13}C-labeled uptake of the metabolite over the time course. Metabolomics investigations use common practices that can become tedious when the researcher is faced with large data sets. Typically, mass spectra from labeled and unlabeled metabolites are plotted head-to-tail so an analyst may visually determine whether the metabolite was labeled or remained unlabeled. The analyst often must visually determine which mass channels (*m/z*) demonstrate the metabolite's isotopic uptake, and plot the intensity of these *m/z* individually in order to visualize the kinetic profile of the time course. This process can be arduous and subjective when metabolites are fragmented during MS detection and can show various different isotopomer distributions throughout the mass spectrum. The information from these isotopomer distributions can be readily misinterpreted, leading to an inaccurate determination and visualization of the time course (or isotopic dilution) profile and errant conclusions regarding the number of atoms labeled in the metabolite.

Many metabolomics studies, whether or not they involve stable isotope labeling, consider one or more of the analytical challenges outlined above [2,6,10,14,22,24]. Non-targeted identification of isotopically labeled metabolites [6,14] and quantification of isotopic incorporation [14], signal [10], and flux [6] have been performed on various data sets to study complex metabolomes. The current study reported herein aims to address the challenges noted

above while integrating and improving upon important aspects of previous studies, including improved deconvolution of coeluting metabolites [10], and the use of principal component analysis PCA to provide key aspects to the analytical method workflow.

Principal component analysis (PCA) is a popular chemometric tool [25] used extensively in many investigations of complex samples, including classification of gasoline [26], chemical forensics of nerve agents [16], species differentiation of tree oils [27], and metabolomics [2]. PCA is a valuable tool for mathematically separating classes of data, visualized in the scores plot, and the loadings plots are often used to provide complementary information to chemically interpret the basis of any observed class separations in the scores plot. PCA is often used in metabolomics, less often for the quantitative information provided by linear algebra, and more often as a visualization tool to separate classes of samples in studies of disease state [28], food decay [28,29], species or strain identification [2,30], and growth condition [2]. However, few studies utilize PCA to elucidate time course information [2,24] when chemical systems in flux would provide a compelling data set for the use of such a chemometric tool. For example, PCA loadings have been used to elucidate changes in metabolite concentrations in cacao beans in order to quantify the time course decay profile over several days [24].

We hypothesize that the quantitative benefits of PCA have yet to be fully exploited in studies using time-dependent ^{13}C-labeling. The scores plots from PCA can be used to quantitatively indicate whether or not a metabolite has been labeled over time course, eliminating the subjectivity of an analyst deciding by individually plotting each metabolite head-to-tail. The PCA loadings can also elucidate how many carbon atoms per metabolite have become ^{13}C-labeled. The PCA scores can also be readily used to quantitatively generate the profile of the time course for ^{13}C uptake, a calculation that was previously obtained from manually (i.e, generally

subjectively) determining the correct *m/z* to plot the mass spectral information of a metabolite's time course profile [7,14,31].

We report an analytical methodology, outlined in Figure 2.1, that provides a novel, non-targeted, time-dependent ^{13}C-labeling of the methylotrophic bacterium *M. extorquens* AM1, and, in general, provides a high throughput analytical platform that may be more attractive than previously reported methods for ^{13}C-labeling time course investigations. A complex GC-TOFMS data set, with many metabolites at a low signal-to-noise ratio (*S/N*) was studied in order to challenge the proposed method for proof-of-principle demonstration of overcoming difficulties in detecting and identifying metabolites. Our method starts with the non-targeted indexing of the retention time of every detectable metabolite in the chromatogram of *M. extorquens* AM1. Overlapped metabolites are deconvoluted, and the pure mass spectra are extracted and normalized for input into PCA. We previously reported the development of a novel data reduction and representation method for GC-TOFMS that significantly facilitates visualization and analyte peak deconvolution, referred to as the two-dimensional mass channel (*m/z*) cluster plot method (2D *m/z* cluster plot method) [32]. The 2D *m/z* cluster plot method provides an accurate and validated determination of the number of analytes (i.e., metabolites) in overlapped peak situations followed by quantitative deconvolution down to a chromatographic resolution of $R_s \sim 0.03$ [32], which is significantly better than typical deconvolution software [10,33]. After mass spectra are extracted via application of the 2D *m/z* cluster method, PCA is subsequently employed on the first (fully unlabeled) and last (fully ^{13}C-labeled) time points of the time course as a diagnostic tool for the determination of the extent, or lack thereof, of ^{13}C incorporation by each metabolite over the time course. A quantitative metric is implemented from the information gleaned from this primary PCA model to provide statistical insight into stable isotope incorporation. A second PCA model is then

performed including all time points in the time course of the GC-TOFMS data set to quantitatively determine ^{13}C labeling profiles of all metabolites, whether or not they incorporated the ^{13}C label.

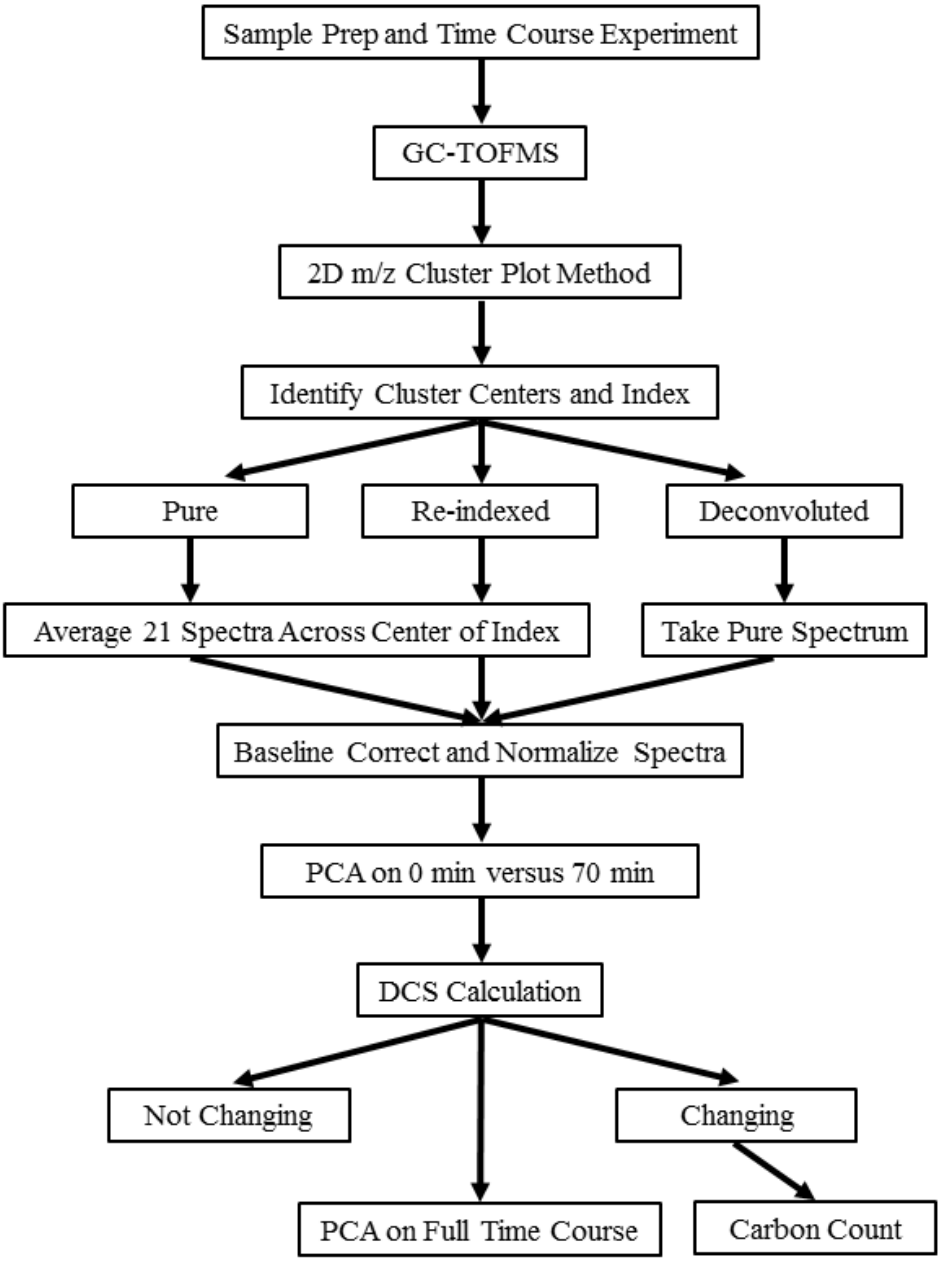

Figure 2.1. A flowchart of the novel method workflow used for the analysis of the time-dependent ^{13}C-labeling of *Methylobacterium extorquens* AM1.

2.2 EXPERIMENTAL

2.2.1 *Standards*

A standards mix of metabolites of biochemical interest was created using ^{12}C (naturally abundant) metabolites derivatized for GC-TOFMS analysis, listed in Table 1. Analytical grade metabolite standards for the mix were obtained from Sigma (St. Louis, MO, USA).

2.2.2 *Batch growth conditions of* M. extorquens *AM1*

The sample preparation for the time course experiment proceeded as follows (Fig. 1). *M. extorquens* AM1 (rifamycin-resistant strain) was grown in liquid batch cultures in a minimal medium, as previously described [13]. 120 mM of either ^{12}C (natural abundance) methanol or ^{13}C-labeled methanol was used as the sole carbon source, both at 99% purity, purchased from Sigma (St. Louis, MO, USA). In the middle of the exponential phase (OD_{600} = 0.60 to 0.70), 10 ml of the culture was extracted and rapidly passed through a membrane filter (S-Pak™ , Millipore, Billerica, MA, USA) using a pipette. The filter was immediately removed and placed on an agar plate of the same medium with ^{12}C-methanol for 20 min, and then transferred to another agar plate with the same concentration of ^{13}C-methanol. At six time points (0, 3, 6, 15 35, 70 min) for the time course experiment, replicates of the filter were immediately transferred to a petri dish located on the surface of a Cool Beans Chill Bucket™ (ISCBioexpress, Kaysville, UT, USA) at 5°C. To collect cells, the following three sequential rinse solutions were applied: (i) 0.5 mL of 25 mM ice cold HEPES buffer (pH 5.2), (ii) 0.5 ml of 20 °C ethanol solution (75/25,v/v, ethanol/aqueous 25 mM HEPES buffer, (pH 5.2), and (iii) 1.5 ml of 20 °C ethanol. The resulting solution was transferred to a pre-cooled tube and stored in a -80°C freezer until it was ready for subsequent extractions.

2.2.3 *Preparation of metabolites in* M. extorquens *AM1*

Extraction of metabolites from *M. extorquens* AM1 samples was carried out as previously published [11]. Briefly, samples were incubated in a 100 °C water bath for 3 min. The extracted cell suspension was cooled on ice for 5 min, and then cell debris was removed by centrifugation at 5,000 RPM for 5 min. The cell-free metabolite extract was centrifuged at14,000 RPM for 8 min. The supernatant was dried in a vacuum centrifuge (CentriVap®Concentrator System, Labconco, MO, USA) and stored at -80 °C. For GC-TOFMS analysis, each sample was further derivatized in two steps. First, keto groups were methoximated by adding 50 µl of methoxyamine solution (25 mg/ml methoxyamine hydrochloride in pyridine) and incubated at 60 °C for 30 min. Second, trimethylsilylation was performed by adding 50 µl of a TMS reagent (BSTFA/TMCS, 99:1) and incubated at 30 °C for 90 min.

2.2.4 *GC-TOFMS analysis of metabolites*

All GC-TOFMS data were collected using an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA) with a LECO Pegasus III TOFMS as the detector (LECO, St. Joseph, MI, USA). The GC column was a nonpolar BPX-5 (29.9 m x 0.25 mm x 1.0 µm film, SGE Inc., Austin, TX, USA). Ultra-high purity helium (Grade 5, 99.999%) was used as the carrier gas (Praxair, Seattle, WA, USA) at a constant flow of 1.5 ml/min, and 1 µl of a given sample was injected via an Agilent 7683 autosampler. All metabolite standards mix samples were injected with a 9:1 split to avoid chromatographic overloading, and all bacteria samples were injected in split-less mode. The GC oven method began at 60 °C with a hold time of 1.25 min after which it was ramped at a rate of 30 °C/min to 280 °C with a hold of 5 min. The ion source was set to 250 °C while the inlet temperature and transfer line were set to 280 °C. After an acquisition delay of 4 min, mass spectra were collected from *m/z* 50 to 600 at a rate of 100 spectra/s.

2.2.5 *Application of the 2D* m/z *cluster plot method*

GC-TOFMS data were imported from the instrument software (ChromaTOF, version 3.32, LECO, St. Joseph, MI, USA) into MATLAB R2012b (Mathworks, Natick, MA, USA) using an in-house software (peg2mat3p8) [34]. Replicates of the standards mix were baseline corrected and smoothed using a Savitzky-Golay filter with a span of 39 spectra (0.39 s) and a third degree polynomial. The mass spectra of the metabolite standards were extracted and averaged. These were imported into MS Search 2.0 (NIST, Gaithersburg, MD, USA) and matched to library entries. A custom library was created from the standards for mass spectral matching to corresponding metabolite peaks detected in the time course experiment.

Three replicate chromatograms per time point of the time course were aligned using total ion current (TIC) shift function retention time alignment [35]. After alignment, the data were analyzed using the method workflow outlined in Figure 2.1. The 2D *m/z* cluster plot method was used to determine the locations of the metabolites, both pure and overlapped, and to deconvolute the latter as previously described by Fitz, et al [32]. The 2D *m/z* cluster plot method takes a previously baseline corrected and smoothed GC-TOFMS chromatogram and, for each *m/z* of each metabolite peak, measures the peak location, t_R, and the peak width, *W*, with 10 ms precision in each dimension (retention time and peak width). The data are then plotted, *W* versus t_R, in a 2D scatter plot (one point per each *m/z*), with each data point occupying a 10 ms × 10 ms square as shown in Figure 2.2(A). Here, a single square can represent one or more *m/z* depending on its color: 1 *m/z*, gray; 2 *m/z*, black; 3 *m/z*, green; 4+ *m/z*, red, with the increased frequency given by the colors indicating a location of selective *m/z* corresponding to a pure analyte. Thus, each analyte peak is viewed as a "mass cluster" or collection of selective mass channels plotted on the *W* versus t_R axis as in Figure 2.2 (A) and (B). A signal threshold of 50 was applied at each *m/z*, corresponding

to a *S/N* of 3. A 70 ms × 50 ms "cluster box" was used to determine the selective *m/z* for each analyte present in the cluster plot. Pure analyte mass clusters are those which correspond to a pure metabolite peak whose spectrum is taken from the center of the cluster, like the first peak in Figure 2.2(A), where one cluster appears below the chromatographic peak. Re-indexed mass clusters are partially overlapped with a neighboring metabolite or interferent peak and therefore require their spectrum be taken from a region corresponding to a sufficiently pure region of the chromatographic peak. An example of two metabolites requiring re-indexing is the second peak in Figure 2.2(A) where two mass clusters appear below a single chromatographic peak. Convoluted mass clusters are those in which two or more metabolites (or interferents) are sufficiently overlapped and require deconvolution via Classical Least Squares (CLS) for extraction of the mass spectra. A more thorough description of the classification and indexing or re-indexing of the mass clusters is described in Section 2.2.6.

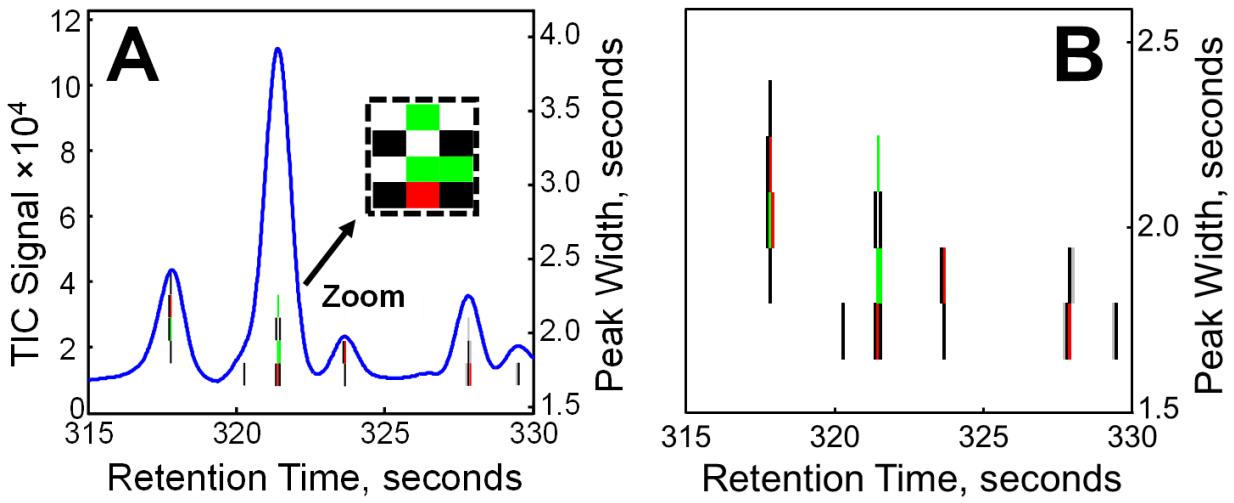

Figure 2.2. Demonstration of the 2D *m/z* cluster plot method

2.2.6 *Classification and re-indexing of mass clusters*

Every identified mass cluster was classed as a "pure" mass cluster, a "re-indexed" mass cluster, or a "deconvoluted" mass cluster. The classification of each mass cluster was dictated by the proximity of an adjacent cluster as defined by chromatographic resolution:

$$R_S = \frac{t_{R2} - t_{R1}}{\bar{w}_b} \tag{2.1}$$

Where R_S is the chromatographic resolution between two adjacent analytes, t_{R2} is the retention time of the later eluting analyte, t_{R1} is the retention time of the earlier eluting analyte and $\bar{w}_b$ is the average width at base of the two analyte peaks. The average peak width is defined as 4σ or ±2σ from the center of the cluster. The average peak width for our data set was 2.0 s, yielding an σ of 0.5 s. Pure mass clusters were defined as those mass clusters that were at least 1.5 s away from both adjacent mass clusters. A difference of 1.5 s corresponds to R_S = 0.75, where the interfering mass cluster is at least 3σ away from the analyte mass cluster. This means for the analyte mass cluster, the interferent's mass spectrum is 1.11% of its maximum or less at the analyte mass cluster's maximum intensity, providing a minimal interference to the mass spectrum of interest. For these pure mass clusters, the mass spectrum was extracted from averaging 21 spectra across the middle of the mass cluster, which is equivalent to 21 spectra or 0.21 s across the top of the pure chromatographic peak. Mass clusters that were less than 1.5 s away from an adjacent cluster as seen in Figure 2.3 (A) and (B) were evaluated for re-indexing, whereby a pure mass spectrum was taken from a pure portion of the analyte elution profile. Re-indexing was effective if the mass spectrum could be extracted from a new "center" that was 1.5 s away from the interferent mass cluster center but not more than 1.0 s away from the original analyte cluster center. This ensures that the interferent mass spectral intensity was no more than 1.11% of its maximum,

but the analyte mass spectrum was no more than 2σ away from the original mass cluster center, thus having no less than 13.5% of its peak maximum intensity, as seen in Figure 2.3(B). Here, 5-oxoproline (blue) is only 0.66 s from its interferent (red). The interferent is thus contributing almost 60% of its maximum mass spectral intensity to the mass spectrum of 5-oxoproline at the original cluster location, causing considerable convolution. However, if the mass cluster centers are "re-indexed" 0.84 s in opposing directions, a purer spectrum can be obtained from both analyte and interferent without the need for full deconvolution. This is demonstrated in Figure 2.3(C), which shows the re-indexed locations (marked by black dashed lines) from where the spectra are extracted, at least 1.5 s from the original interferent cluster center (marked by gray dashed lines) and now 2.34 s from each other. This ensures that the mass spectral intensity of the interferent is 1.11% or less of its peak maximum intensity at the new index location, while the intensity of the analyte of interest still remains high (in this case about 24% of its peak maximum intensity, but in all cases at least 13.7% of its peak maximum intensity). Re-indexing increases the resolution between the two new analyte indexes and minimizes the contribution of the mass spectrum from adjacent analytes or interferents. As with pure clusters, the mass spectra of re-indexed clusters are taken as the average of 21 spectra or 0.21s across the center of the re-indexed cluster. Finally, for mass clusters that were less than 1.5 s away from an adjacent mass cluster and could not be re-indexed, the mass cluster was designated as requiring deconvolution (i.e., deconvoluted mass cluster) and classical least squares (CLS) deconvolution was utilized via the 2D *m/z* cluster method to extract the pure mass spectrum for analysis. The location at which the mass spectrum was extracted, whether pure, re-indexed or deconvoluted, is referred to as the index of the metabolite and is included in Table 2-2.

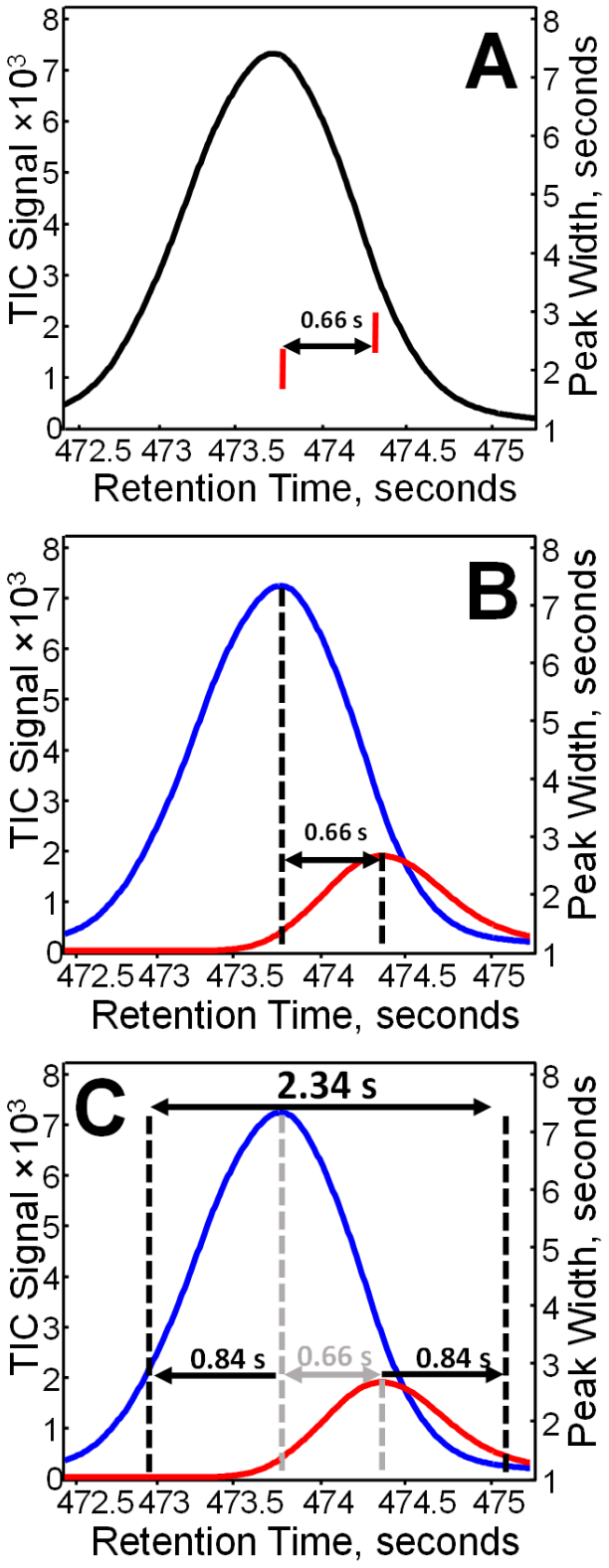

Figure 2.3. Visualization of the re-indexing process

Once extracted, the mass spectra (*m/z* 100-300, excluding *m/z* 73, and 146-148 so as to exclude peaks coming from the derivatization reagent) for all metabolites were baseline subtracted and normalized prior to analysis using PCA using PLS Toolbox 7.3.1. (Eigenvector Research, Inc., Wenatchee, WA, USA). As seen in Figure 2.4(A), the raw mass spectrum of 5-oxoproline is highly noisy. This makes finding time-dependent ^{13}C-labeling trends difficult, especially at higher *m/z* where the parent ion peak would be found. The signal intensity of a small noise region of the mass spectrum (*m/z*=350-400) was averaged and subsequently subtracted from the entire mass spectrum. The *m/z* desired for analysis (*m/z*=100-145, 149-300) were then normalized to the total signal of all summed mass channels. This was performed for every mass spectrum extracted prior to PCA analysis. Figure 2.4(B) shows the baseline corrected and normalized mass spectrum of 5-oxoproline with the noise greatly reduced and important mass channels clearly represented. The comparison between raw mass spectrum and baseline corrected and normalized mass spectrum is shown also for the unidentified interferent to 5-oxoproline in Figure 2.5 (A) and (B), respectively.

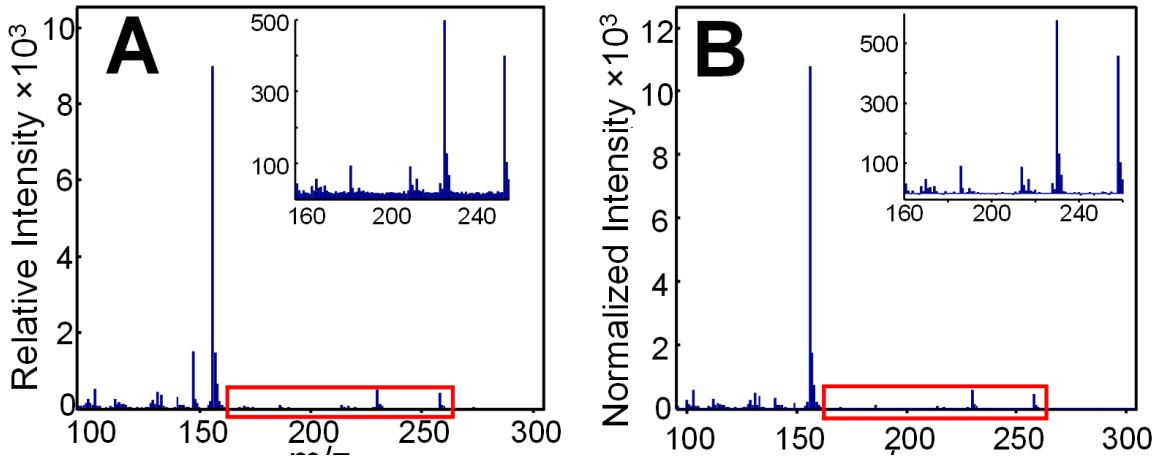

Figure 2.4. Mass spectrum of 5-oxoproline (A) before and (B) after baseline correction and normalization.

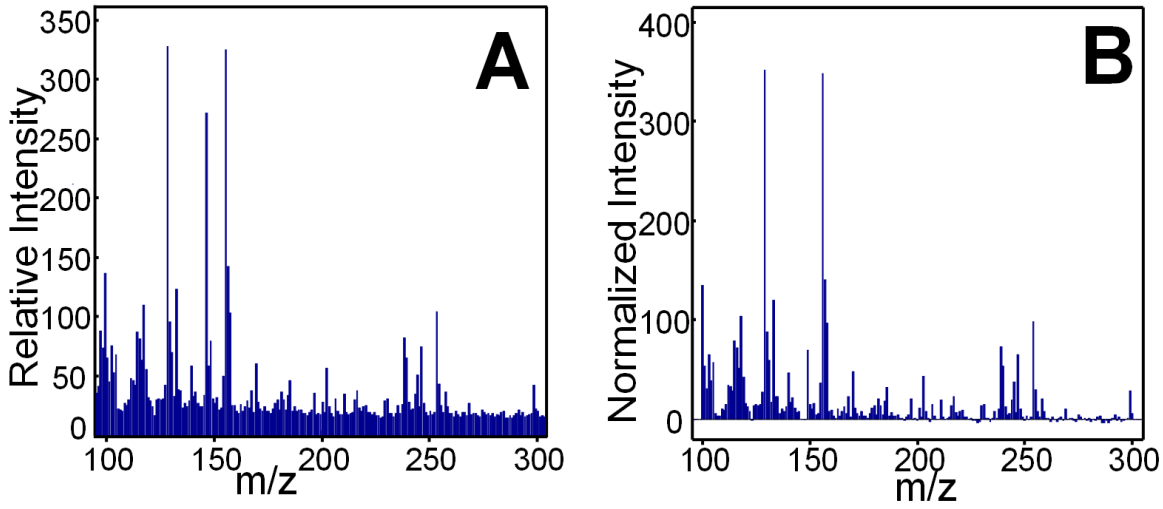

Figure 2.5. Mass spectrum of unknown interferent of 5-oxoproline (A) before and (B) after baseline correction and normalization.

Finally, after baseline subtraction, two PCA models were constructed, serving as vital steps in the method workflow outlined in Figure 2.1, denoted as "PCA on 0 min versus 70 min" and "PCA on Time Course," respectively. The first PCA step provided an accurate determination of the presence or absence of ^{13}C incorporation over the time course and, in the former, a straightforward method for determining the number of isotopically labeled carbon atoms. The second PCA step provided time course profiles of the ^{13}C labeled metabolites. For PCA on 0 min versus 70 min, baseline corrected and normalized mass spectra from the initial (0 min) and final (70 min) time points of the injection replicates of the time course were input to PCA for modeling. From the scores of the PCA model, a quantitative metric, degree-of-class separation (DCS) [27,36], was calculated for each metabolite (as indexed by its mass cluster). Based upon the DCS value, each metabolite was then classed as "changing," indicating ^{13}C uptake by the metabolite, or "not changing," indicating the lack of detectable ^{13}C uptake. Finally, in the PCA on Time Course

model step of the analysis, the mass spectra from the full time course (0, 3, 6, 15, 35, and 70 min) of each metabolite were input into PCA to obtain time course profiles for each metabolite.

2.3 Results and Discussion

2.3.1 *Metabolite standards and* M. extorquens *AM1 chromatograms*

A full chromatogram of the metabolite standards is provided Figure 2.6(A), with each standard peak numbered in the zoom view in Figure 2.6(B). The peak number, retention time, and mass spectral match value (NIST library, unless otherwise indicated) for each metabolite standard are summarized in Table 2-1. Additionally, the analyte number (far right column in Table 2-1) corresponds to the peak number location of each standard in the target chromatogram of the time course on *M. extorquens* AM1 (details of peak number labeling are provided below in the indexing section). The target chromatogram, to which all other chromatograms were aligned, is shown in Figure 2.7, which shows the complexity of the bacterial sample, low *S/N* and obvious overlap of many chromatographic peaks. It is this complexity of the bacteria separations that

Table 2-1: Table of metabolite standards.

Peak #	Metabolite Standard	t_R (s)	MV	Analyte #
1	Pyruvate	320.0	952*	43
2	Lactate	321.0	661$^+$	44
3	Alanine	338.0	885	52
4	Oxalate	347.0	918	59
5	3-Hydroxybutyrate	357.0	903	63
6	Methylmalonate	376.0	892	
7	Valine	377.0	869	70
8	Leucine	393.0	827	65
9	Ethylmalonate	396.5	884	
10	Isoleucine	401.0	844	
11	Threonine	403.5	874	80
12	Proline	408.0	722	
13	Glycine	408.0	746	82
14	Succinate	408.0	538*	83
15	Glycerate	408.0		
16	Methylsuccinate	411.0	898	85
17	Serine	417.0	844	
18	Fumarate	418.0	861	88
19	Methylmaleate	432.0	908	
20	Malate	452.5	861	102
21	Methionine	470.0	848	
22	5-Oxoproline	474.0	846	112
23	Phenylalanine	500.5	852	
24	Isocitrate	532.5	777	
25	Citrate	534.0	802	135

inspired the development and use of the proposed method workflow (provided in Figure 2.1) for the non-targeted discovery of metabolites that incorporate ^{13}C over the time course. The overall procedure includes the 2D *m/z* cluster plot method, a novel visualization and deconvolution software developed in-house. The following section describes the application of the 2D *m/z* cluster plot method to accurately extract all of the metabolite mass spectra with sufficient signal from the GC-TOFMS data down to a chromatographic resolution of $R_s \sim 0.04$.

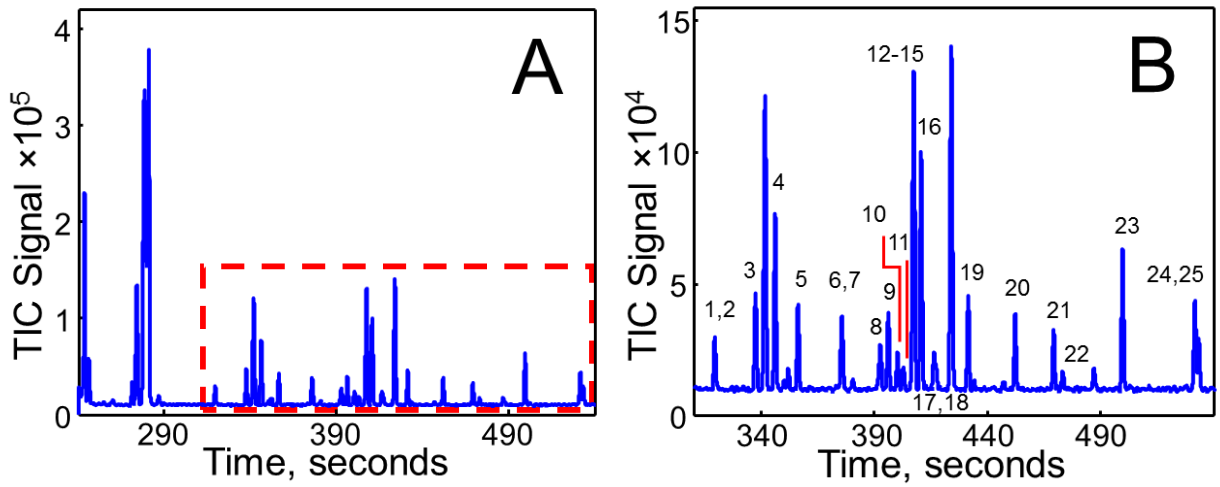

Figure 2.6. (A) Total ion current (TIC) chromatogram of metabolite standards; (B) with each standard peak numbered as in Table 1.

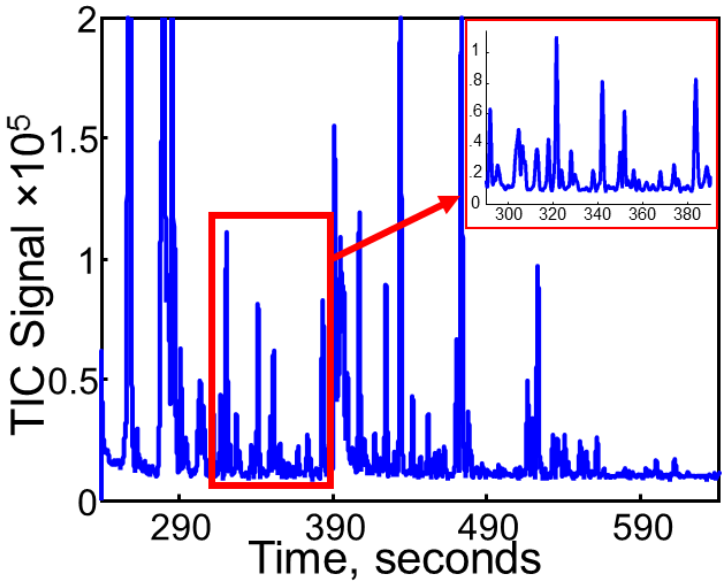

Figure 2.7. TIC chromatogram of *M. extorquens* AM1.

2.3.2 *Non-targeted metabolite indexing and spectra extraction for* M. extorquens *AM1*

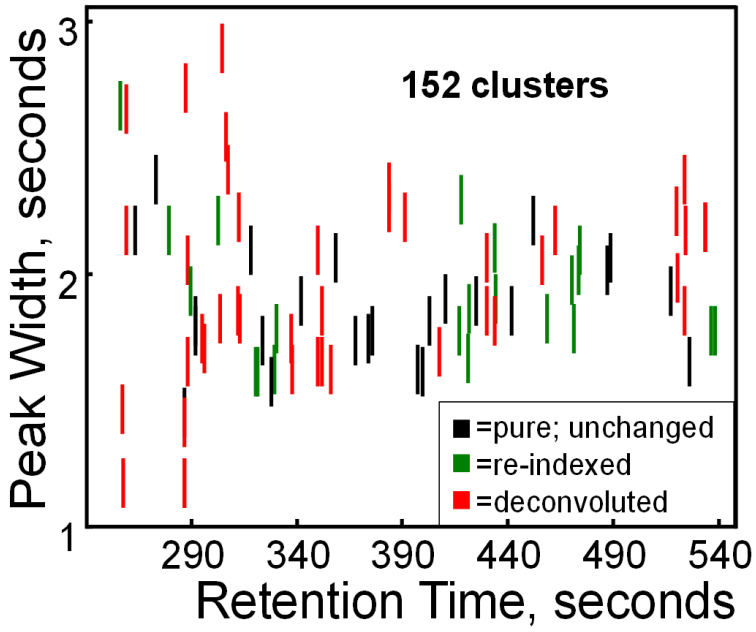

Figure 2.8. Mass cluster plot with pure (black), re-indexed (green) and deconvoluted (red) clusters.

Mass clusters (i.e., analyte peak center locations) in the GC-TOFMS separations were identified using the 2D *m/z* cluster method, in which peak width *W* is plotted versus the retention time, t_R, and further indexed according to the retention time center of the pure *m/z*. They were then classified into one of three groups: pure mass clusters, re-indexed mass clusters, or deconvoluted mass clusters, as discussed in Section 2.2.6. A single rectangle in Figure 2.8 represents one pure mass cluster as defined by the cluster box, including multiple selective *m/z*. Due to the dimensions of the axes in Figure 2.8, the singular rectangles representing individual analytes appear as vertical lines along the peak width *W* dimension. A pure mass cluster, represented by a black rectangular line in Figure 2.8, corresponds to a pure metabolite peak whose spectrum is taken from the center of the cluster. A re-indexed mass cluster, represented by a green rectangular line in Figure 2.8 corresponds to a metabolite peak partially overlapped with a neighboring metabolite or interferent peak and therefore required its spectrum be taken from a region corresponding to a sufficiently

pure region of the chromatographic peak, with the process described in Figure 2.3. A convoluted mass cluster, represented by a red rectangular line in Figure 2.8, indicates two or more metabolites (or interferents) are sufficiently overlapped and required deconvolution via classical least squares (CLS) for extraction of the mass spectra.

The 2D *m/z* cluster plot method provides visualization of the 1D GC-TOFMS data in two dimensions, and ultimately facilitates extraction of the pure mass spectra. In Figure 2.8, the separation between mass clusters is not only present in the retention time t_R dimension, but also in the peak width W dimension. An advantage of the method is that it optimizes the number of metabolites found and successfully analyzed, which is a direct consequence of the ability to identify and resolve analytes to an extremely low $R_s \sim 0.04$ (i.e., overlapped peaks that differ in retention time of only 4% of their peak width at base). The mass cluster provides a highly precise and accurate metabolite retention time index, as many metabolites, and therefore many mass clusters, can be clearly identified within the retention time span of one chromatographic peak width. Therefore, the number of mass clusters is a much better estimate of the number of metabolites since the chromatographic resolution in the mass cluster space is ~ 25 times better than in the original chromatogram at unit resolution, $R_s = 1.0$. Additionally, most peak finders fail to resolve peaks at a chromatographic resolution, $R_s < 0.5$. Indeed, while only 101 chromatographic peaks were counted in the TIC chromatogram in Figure 2.7, 152 mass clusters were counted in the 2D *m/z* cluster plot in Figure 2.8 (with more details provided in Table 2-2). Several of the metabolites coelute at very low R_s, and appear as a single chromatographic peak in the TIC chromatogram. Differences in peak width due to peak overlap or chromatographic-dependent band broadening differences cause some mass clusters to appear higher in the 2D scatter plot than others due to their larger peak widths, as seen in Figure 2.8.

Table 2-2 summarizes all of the mass clusters identified in the chromatogram of *M. extorquens* AM1. This table includes the analyte number, retention time, and index. The analyte number is sequential based on retention time, and the index is the location in time of the center of the mass cluster as determined by the 2D *m/z* cluster plot method. The white boxes are pure clusters (i.e., chromatographic resolved analytes), the green are re-indexed analytes (and therefore have an index referring to the re-indexed location of the center from which the mass spectrum was taken) and the red boxes are those analyte clusters that were deconvoluted using CLS. The analyte number in Table 2-2 also corresponds to the analyte number in the last column of Table 2-1, the table of standards, identifying where in the chromatogram the standards elute.

Table 2-2: Summary table of all 152 mass clusters discovered via the 2D *m/z* cluster method with their retention time index highlighted according to whether they were "pure" (no highlight), "re-indexed" (green), or "deconvoluted" (red) as in Figure 2.8. Mass cluster plot with pure (black), re-indexed (green) and deconvoluted (red) clusters.

Number	t_R (s)	Index (s)
1	256.4	255.79
2	257.8	257.29
3	257.8	257.49
4	257.8	257.74
5	259.0	258.21
6	259.0	258.94
7	259.0	258.98
8	259.0	260.48
9	263.3	263.23
10	265.3	265.09
11	265.3	266.90
12	273.2	273.20
13	279.5	279.09
14	281.1	280.59
15	281.1	282.09
16	286.7	286.64
17	286.7	286.80
18	286.7	286.86
19	286.7	286.93
20	286.7	287.35
21	288.1	288.05
22	288.1	288.26
23	288.1	289.76
24	291.8	291.77
25	293.8	293.56
26	295.1	295.07
27	295.1	295.87
28	297.4	297.36
29	299.8	299.87
30	304.5	300.20
31	304.5	302.14
32	304.5	303.64
33	304.5	304.57
34	304.5	304.93
35	304.5	305.34
36	306.2	306.28
37	307.3	306.91
38	307.3	307.30
39	312.6	312.03
40	312.6	312.35
41	312.6	313.10
42	317.8	317.82
43	321.4	319.93
44	321.4	321.80
45	323.7	323.65
46	323.7	326.30
47	327.8	327.81
48	329.5	328.87
49	330.5	330.98
50	330.5	334.53
51	337.6	337.40
52	337.6	337.68
53	337.6	340.28
54	341.8	341.79
55	341.8	344.30
56	341.8	346.09
57	350.0	350.00
58	350.0	350.11
59	351.9	351.93
60	351.9	352.00
61	353.7	354.42
62	356.1	355.91
63	356.1	356.04
64	358.3	358.24
65	361.7	361.54
66	361.7	361.64
67	364.5	364.42
68	367.8	367.80
69	374.0	373.99
70	375.7	375.70
71	380.0	380.16
72	383.8	383.75
73	383.8	383.94
74	388.7	386.70
75	388.6	388.60
76	391.3	391.16
77	395.3	395.64
78	397.1	397.30
79	399.6	399.74
80	403.1	403.10
81	407.5	405.91
82	407.5	407.41
83	407.5	407.50
84	409.3	409.03
85	410.6	410.53
86	412.0	412.03
87	413.7	413.63
88	417.3	416.27
89	417.3	418.65
90	421.4	420.14
91	421.4	422.67
92	424.7	424.67
93	428.3	428.41
94	430.3	429.95
95	430.3	430.43
96	430.3	431.20
97	434.1	432.56
98	434.1	434.06
99	434.1	435.56
100	441.9	441.92
101	447.1	446.96
102	452.1	452.05
103	456.6	456.23
104	456.6	456.52
105	456.6	456.76

Number	t_R (s)	Index (s)
106	458.9	458.88
107	460.4	460.40
108	462.5	462.21
109	462.5	462.80
110	471.0	469.61
111	471.0	471.99
112	473.8	472.88
113	473.8	475.22
114	478.2	477.85
115	478.2	478.21
116	478.2	478.27
117	479.9	479.83
118	483.3	483.33
119	487.2	487.26
120	488.9	488.79
121	490.9	490.55
122	495.8	495.72
123	498.2	498.12
124	502.2	502.25
125	509.3	509.34
126	517.1	517.11
127	520.4	520.19
128	520.4	520.53
129	520.4	520.65
130	523.8	523.73
131	523.8	524.16
132	523.8	524.27
133	528.3	528.45
134	533.6	533.65
135	533.6	533.76
136	536.6	535.60
137	536.6	537.10
138	538.5	538.60
139	541.0	541.06
140	543.2	543.24

Number	t_R (s)	Index (s)
141	551.1	551.03
142	551.1	551.13
143	554.9	553.53
144	554.9	555.97
145	561.9	561.81
146	593.8	593.81
147	600.3	600.32
148	612.2	612.00
149	664.3	663.93
150	737.3	731.12
151	743.7	743.48
152	784.1	784.20

2.3.3 *Principal component analysis to assess ^{13}C incorporation for* M. extorquens *AM1*

In order to assess whether or not a metabolite incorporated ^{13}C during the time course, PCA on the 0 min versus 70 min mass spectra was performed. The two time point extremes, 0 min versus 70 min, have been determined to be sufficient to provide the "fully unlabeled" and "fully labeled" states, respectively, for all of the metabolites for *M. extorquens* AM1 [10,14]. The PC2 versus PC1 scores plot of 5-oxoproline, provided in Figure 2.9, is a representative metabolite to demonstrate this stage of the method workflow in Figure 2.1. Although 5-oxoproline is not expected as a natural metabolite in the methylotrophic bacterium investigated, it can originate as a breakdown product of glutamine and can also be generated enzymatically as a bona fide metabolite [37]. The PCA scores plot for 5-oxoproline demonstrates how the two classes, 0 min (red triangles) and 70 min (green asterisks), separate from each other on PC1, which contains 99.97% of the variance between the data sets. As the only difference between the sample classes is the time exposed to the ^{13}C-methanol, the variance is attributable to changes in the mass spectra that occur as the metabolite incorporates ^{13}C over time.

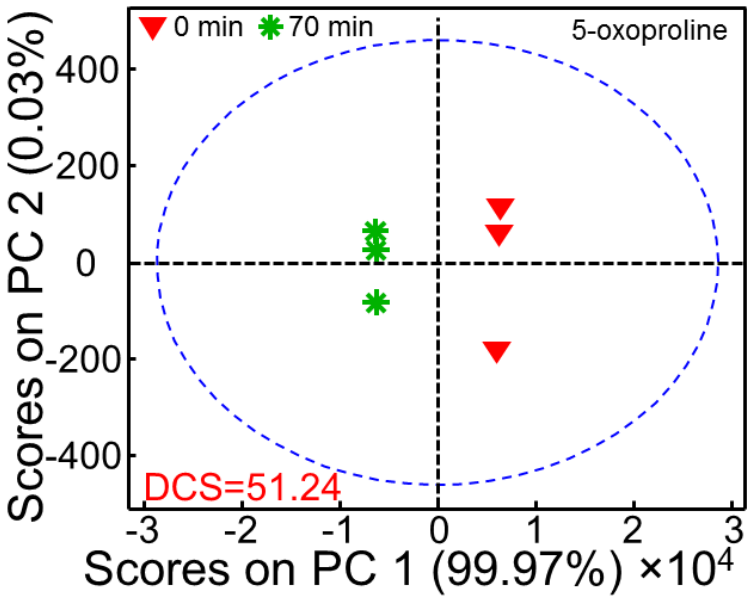

Figure 2.9. PCA Scores plot of 5-oxoproline.

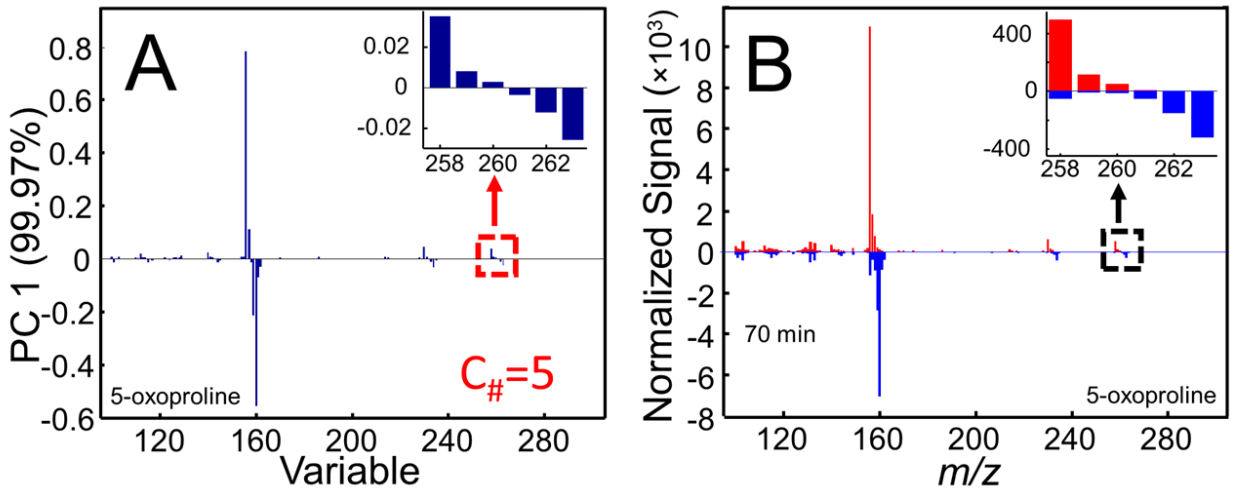

Figure 2.10. (A) PCA loadings plot; and (B) head-to-tail plot of 0 min (positive, red) and 70 minute (negative, blue) of 5-oxoproline.

Examination of the loadings plot for PC1 of the PCA model of 5-oxoproline is provided in Figure 2.10(A). Here, there is an obvious shift in the PC1 loadings intensity from positive to negative as ^{13}C is incorporated at various variables (i.e., the *m/z*). The TMS-derivatized molecular weight of 5-oxoproline is 273 amu. The intensity of higher *m/z* fragments is often too low for analysis due to the use of electron impact ionization (EI) with the TOFMS. Thus, for most metabolites in this study, the focus is on the M-15 fragment, which corresponds to a loss of a methyl group from one of the TMS groups [38]. For TMS derivatized 5-oxoproline, the M-15 peak is *m/z* = 258. A shift in intensity and sign from *m/z* = 258 to *m/z* = 263 in Figure 2.10(A) indicates an incorporation of five ^{13}C molecules, and, in fact, 5-oxoproline has that many carbons in its backbone. This loadings plot can be compared to the traditionally applied head-to-tail plot as provided in Figure 2.10(B), which shows the 0 min and 70 min mass spectra before PCA. The same mass channel range, *m/z* = 258 to *m/z* = 263, shows a shift in intensity from ^{12}C (red, positive) to ^{13}C (blue, negative), as in the loadings plot. Additionally, the comparison of the loadings and the head-to-tail plots highlights how PCA maximizes differences and minimizes similarities, as mass channels that are unchanged between the two samples are nearly invisible in the loadings plot. Even though 5-oxoproline was a representative metabolite, all metabolites classified as changing were interpolated in the same way, and it was found that the PCA method presented provided a direct means of determining the number ^{13}C atoms incorporated. Other metabolite examples are presented on the next page. To avoid analyst intervention to identify the number of ^{13}C atoms, a method to survey the metabolites quickly and reduce the number of loadings plots investigated was incorporated into the method, which is presented next.

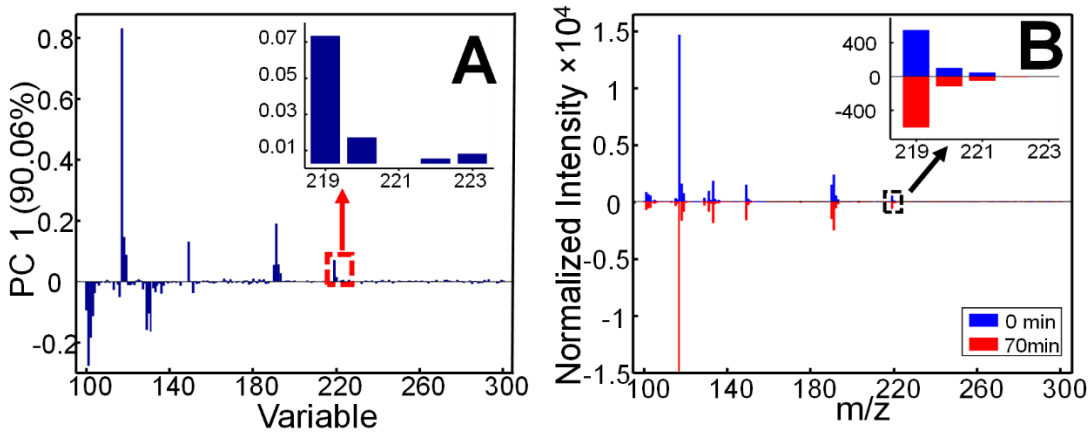

Figure 2.11. (A) PCA loadings plot; and (B) head-to-tail plot of lactate (DCS = 0.7) showing no incorporation of the ^{13}C during the time course.

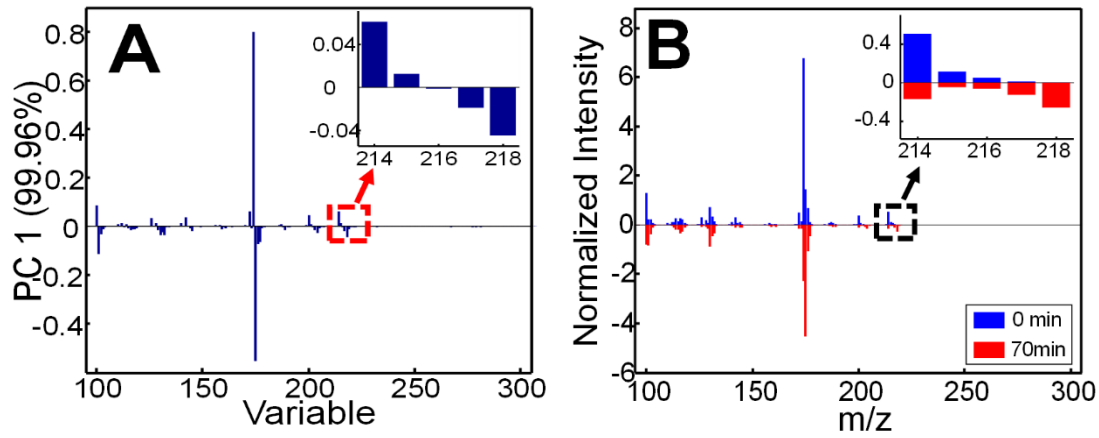

Figure 2.12. (A) PCA loadings plot; and (B) head-to-tail plot of putrescine (DCS = 44.1) showing successful incorporation of ^{13}C in all 4 carbons in the backbone of the molecule.

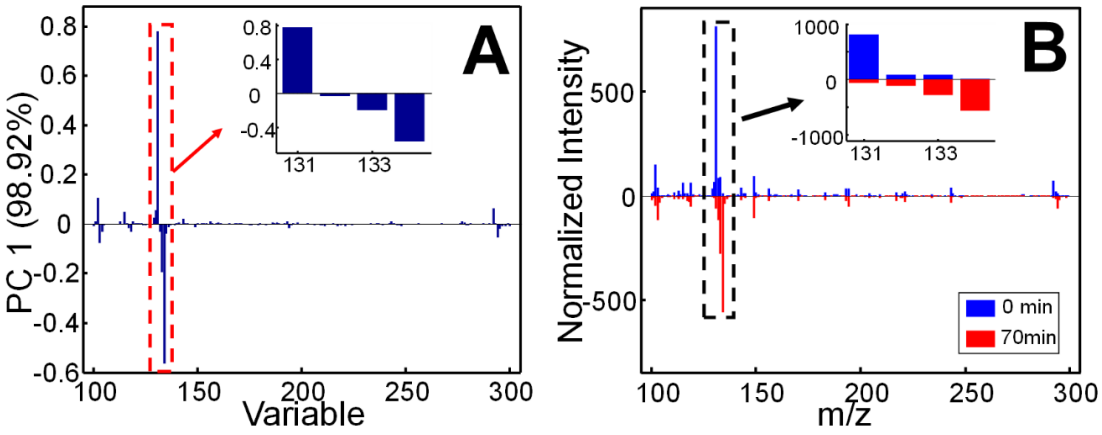

Figure 2.13. (A) PCA loadings plot; and (B) head-to-tail plot of unknown metabolite with the index 428.41 s (DCS = 7.98) showing successful incorporation of ^{13}C in at least three carbons.

In order to readily survey the 152 mass clusters (i.e., metabolites and reagent peaks) to assess ^{13}C incorporation, a quantitative metric of the two class separation between the 0 min replicates and 70 min replicates was utilized. Degree-of-class-separation (DCS) is a quantitative metric of the separation between two classes in the PC2 versus PC1 scores plot [27,36]:

$$DCS = \frac{D_{0,70}}{\sqrt{s_0^2 + s_{70}^2}} \tag{2.2}$$

The DCS takes into account both the Euclidean distance between the center of the two classes ($D_{0,70}$) as a measurement of the difference between classes, and the standard deviation of the Euclidean distances from the center of each group to its replicates (s_0 and s_{70}) as the spread within each class. A large DCS value indicates a large difference between the 0 min and 70 min replicates and a strong likelihood of incorporation of ^{13}C. DCS values were calculated for every PCA model of all 152 mass clusters, and those with a DCS value greater than 1.7 were deemed as changing (i.e., incorporating ^{13}C). A DCS value of 1.7 corresponds to a t-test t value of 2.57, with larger DCS values corresponding to larger t-values. With five degrees of freedom, a two sampled t-value less than 2.57 indicates no statistically significant difference between the two sample classes at the 95% confidence level, meaning there is no difference between the 0 min and 70 min mass spectra. For example, 5-oxoproline has a DCS value of 51.24, a value much larger than the DCS value threshold of 1.7, so 5-oxoproline was labeled as changing (see Table 2-3). DCS values were calculated for all metabolites in the time course based on their PCA models. A histogram of these DCS values is provided in Figure 2.14, showing the distribution of DCS values among the metabolites detected. For those metabolites with DCS greater than 1.7, the loadings plots were then investigated for the determination of the number of ^{13}C atoms incorporated. The incorporation of DCS as a quantitative metric helped eliminate analyst intervention for all 152 clusters,

emphasizing only those metabolites that were statistically significant in their scores values in the PCA modeling. This, together with the natural way in which PCA loadings plots emphasize changes, greatly reduces analyst intervention and makes unnecessary the close investigation of the entire mass spectrum as is typically performed. Note that it was imperative to use only the mass spectra from the 0 min (first) and 70 min (last) time points for the determination of the number of ^{13}C atoms incorporated, rather than the full time course, as these time points represent the fully unlabeled and fully labeled metabolite distributions, respectively. Although it requires an extra PCA model, it is necessary in order to avoid interferences due to partially labeled isotopomers of the metabolite at earlier time points, which could cause a misidentification of the number of carbons converted to ^{13}C.

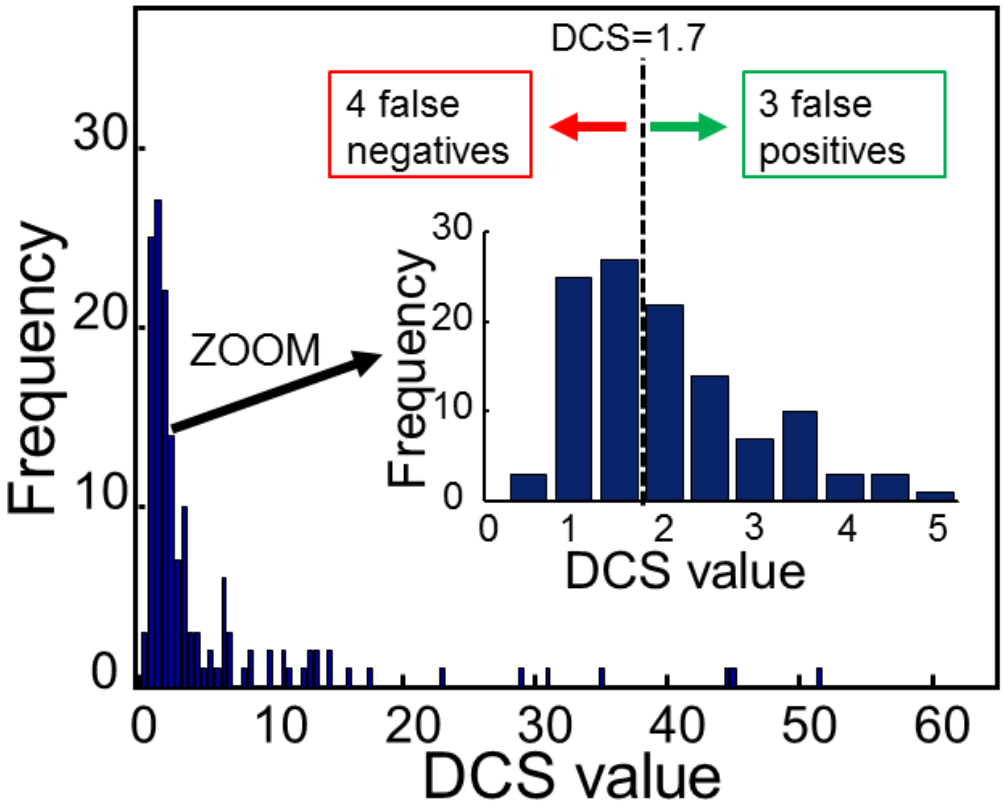

Figure 2.14. Histogram of DCS values on the scores plots of the first PCA model of the 152 mass clusters discovered.

2.3.4 *Principal component analysis to determine the time course*

To quantitatively visualize and assess the time-dependence of the ^{13}C-labeling, PCA on the entire time course data set was performed. The mass spectra of all injection replicates of the full time course study (0 min, 3 min, 6 min, 15 min, 35 min, and 70 min) were analyzed using PCA, using a method analogous to the previously described step using just the 0 and 70 min samples. The scores on PC1 were plotted versus time as shown in Figure 2.15(A), again for the representative metabolite 5-oxoproline. While a time course effect is visually confirmed by this use of PCA, we shall also demonstrate that an accurate quantitative time course is also readily provided by PCA, even elucidating how steep or gradual the change over 0 to 70 minutes for each metabolite as seen in Figure 2.18 through Figure 2.20. While beyond the scope of the study reported herein, the quantitative information provided by PCA could be used to calculate the metabolic fluxes of each metabolite. Furthermore, when PCA of the full time course was performed on the mass spectra of a metabolite that was classified as not changing based on the DCS value for the 0 min versus 70 min PCA, as in the case of lactate, it was confirmed to have no discernible time course effect when all time points and replicates were analyzed (see Table 2-3). Using this PCA-based method, the time course effect was quantitatively extracted for all 152 metabolites. Of these, 83 metabolites showed time course effects that indicated incorporation of ^{13}C, and 69 demonstrated no time course effect. Of the 83 changing metabolites, 77 had a DCS of 1.7 or greater as predicted by the PCA model of 0 min versus 70 min, indicating a time course effect was present prior to construction of the second model. As with any distribution of numbers, as in the case of the DCS histogram provided in Figure 2.14, there occur false positives and false negatives, but these account for less than 5% of the metabolites. Seven metabolites displayed time course effects contrary to their designation from the DCS value. Four metabolites were false

negatives, where the DCS value appeared below 1.7, but the PCA model of the time course indicated a slight change over time. Three metabolites were false positives having DCS greater than 1.7 but demonstrating no readily observable time course effect. The loadings plot on PC1 of the 0 min versus 70 min model was consulted for the shift in intensity characteristic of a shift from ^{12}C to ^{13}C before being re-classed from changing to not changing or vice versa.

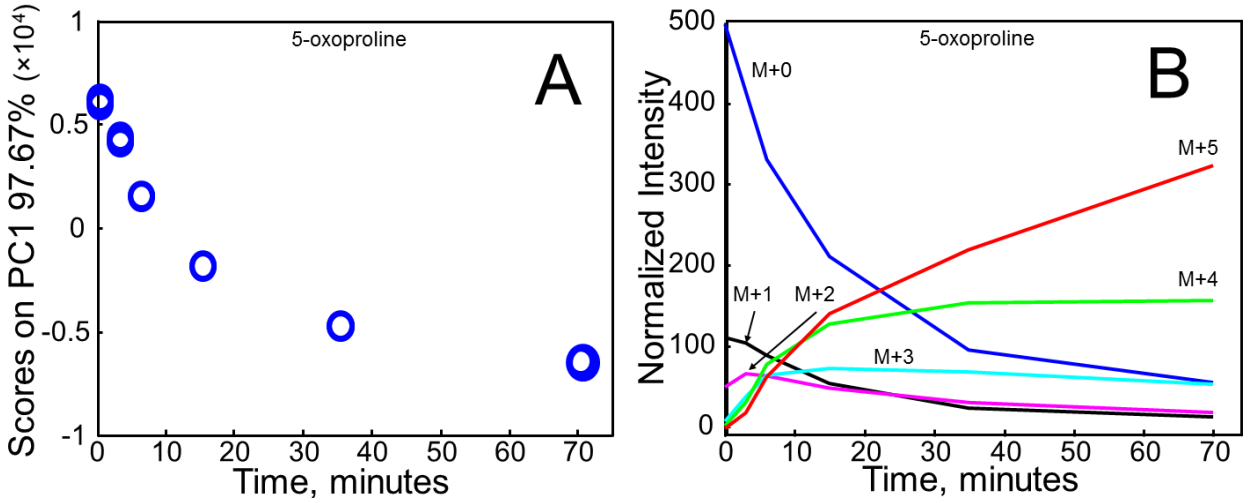

Figure 2.15. (A) PCA Scores vs. Time of the second PCA model of 5-oxoproline; and (B) the corresponding M+n plot with the intensity of *m/z* 258 vs. time shown as the blue M+0 line.

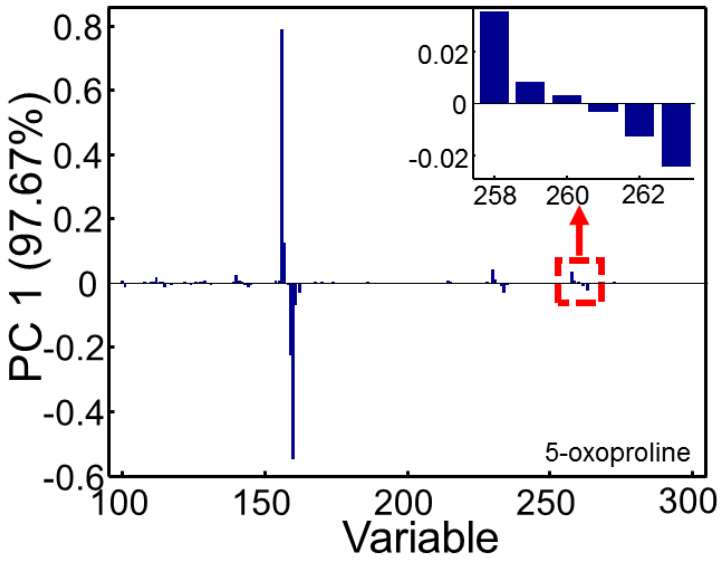

Figure 2.16. PCA Loadings plot of the second PCA model of 5-oxoproline.

Validation of the workflow method presented for elucidating the time course effect was performed by plotting the commonly applied mass intensity versus time plot, or M+n plot, as provided in Figure 2.15(B), where the M-15 peak for 5-oxoproline is plotted as M+0 (no increased abundance of ^{13}C). The intensity of M+0 decreases with time as the other *m/z* intensity values increase, indicating isotopic dilution of the ^{12}C-labeled molecules by the ^{13}C molecules. Specifically, the M+5 peak increases, showing that 5-oxoproline incorporates five ^{13}C, consistent with what was elucidated in the loadings plot discussed above. This M+n plot, traditionally used to elucidate a time course effect of a metabolite, is both cumbersome and subjective to obtain. It requires the analyst to correctly identify which *m/z* to plot, a tedious task made difficult for unidentified metabolites, before plotting the intensity of the *m/z* over time for hopeful elucidation of a time course that can be easily extracted via PCA as demonstrated herein. For metabolites having different time course shapes, the M+n plots demonstrated similar shapes to those seen in the PCA plots, with additional examples provided in the Figure 2.18, Figure 2.19, and Figure 2.20.

Using the information gleaned from both PCA models, the 0 versus 70 min model and the entire time course model, the time course effect was quantitatively reconstructed from the PCA model using Eq. 2.3:

$$X = 1\bar{x} + TP' \qquad (2.3)$$

In this expression X is the original baseline corrected and normalized mass spectra included in the model, $\bar{x}$ is the mean of the data (since the data is mean centered), T is the scores on PC1 (from Figure 2.15(B)) and P' is the loadings on PC1 (from Figure 2.16). Eq. 2.3 excludes the error or residuals matrix in order to elucidate how well the PCA model encompasses the variance and models the time course. For this reconstruction, the PCA model on 0 versus 70 min identified the

m/z that incorporated ^{13}C, and the PCA model on the full time course provided the values for T and P' for the aforementioned *m/z* to input into Eq. 2.3. Figure 2.17 shows the reconstructed model projected on top of the M+n plots from Figure 2.15(B) for the M+0 and the M+5 isotopomers. The models for the other isotopomers, (M+1, M+2, M+3, M+4), were calculated but not included for clarity. The coefficient of determination, R^2, was calculated as a metric for how well the PCA model correlated with the actual data, and the calculation includes all mass isotopomers. As seen in Figure 2.17 of 5-oxoproline, R^2 =0.98, indicating that the reconstruction from PCA is a very good model for the data. This further demonstrates that PCA is as effective as (if not more effective) and less tedious than traditional methods for modeling and quantitatively elucidating time course information in metabolomics. Because the analyst must investigate the loadings plot of the PCA model for 0 versus 70 minutes in order to determine which *m/z* to include for the determination of labeling, the reconstruction also serves as a validation step. Should the model be reconstructed using *m/z* that appear to shift, but do not actually contribute to the time course, the reconstruction will not correlate with the time course effect elucidated by the PCA on the time course model, and will have a lower R^2 value. This concept is shown in Figure 2.20(C) and (D), where a reconstruction of *m/z* =131 (M+0) and *m/z* =134 (M+3) follows the steep trajectory of the time course in Figure 2.20(A) and (B) with an R^2 =0.98, where as a reconstruction of *m/z* =243 (M+0) and *m/z* =248 (M+5), does not appear to follow the time course effect and has an R^2 =0.87, a much lower value, indicating a poor correlation.

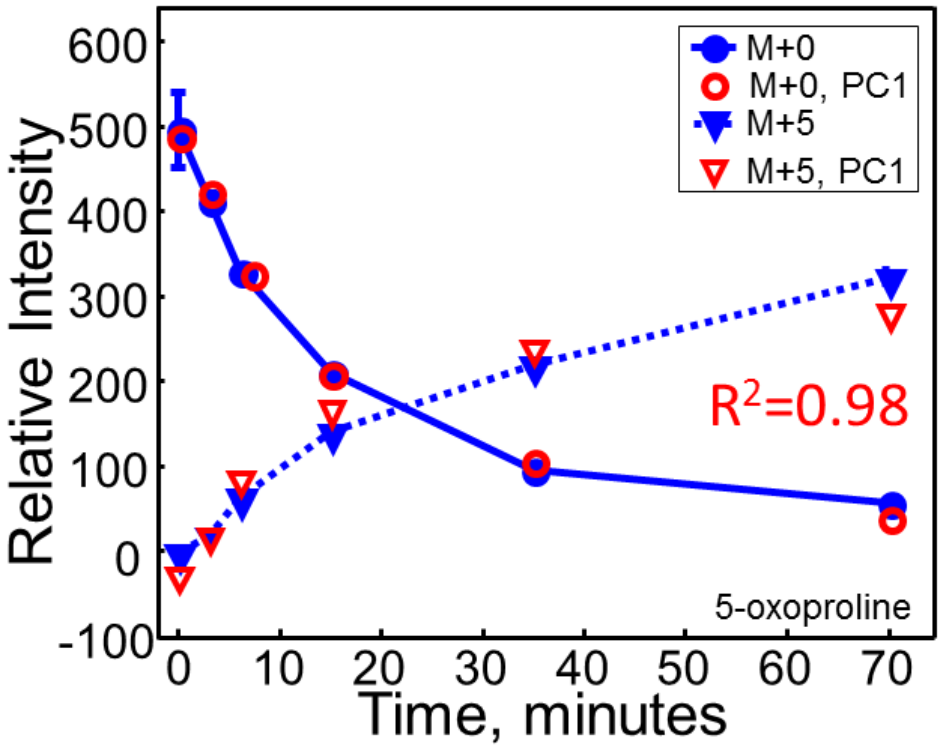

Figure 2.17. Reconstruction of the original data (in red) from the second PCA model of 5-oxoproline overlaid with the original data (blue).

For the PCA of time course models, Figure 2.18(A) shows the scores versus time plot of lactate, demonstrating no kinetic profile. This is confirmed in Figure 2.18(B), the traditional M+n line plot. Similarly, this is validated via the reconstruction of the data from the PCA model in Figure 2.18(C) with an $R^2 = 0.99$. Figure 2.19(A) shows the scores versus time plot for putrescine. This shows the time course profile that was expected, but of slight difference to that of 5-oxoproline as seen before. This time course profile as a gradual change at the start, where from 0 to 6 min, there is no change on PC1, after which the change become more apparent. This quantitative determination of the time course profile is confirmed in the traditional M+n plot in Figure 2.19(B) and is validated in the reconstruction of the data in Figure 2.19(C) with an $R^2 = 0.99$. Lastly, the unidentified metabolite with an index of 428.41 s in Figure 2.20(A) demonstrates a steep time course profile, where there is a drastic change in just the first 3 min and then the profile levels out. This is confirmed in Figure 2.20(B) and Figure 2.20(C) as with the previous metabolites. Figure 2.20(C), with an $R^2 = 0.93$, shows a well correlated reconstruction of the model, while Figure 2.20(D), with an $R^2 = 0.87$, shows how a model reconstructed from *m/z* that are not a part of the time course will not be well correlated with the original data. These examples demonstrate how PCA can elucidate the time course profile through the modeling of the entire mass spectrum, without subjective determination of which mass channels to plot as is traditionally done in the line plots. PCA can be used to visualize how gradual, steep or non-existent a time course profile a metabolite has, even if the metabolite's identity is not confirmed, as in the case of unidentified metabolite with index 428.41 s.

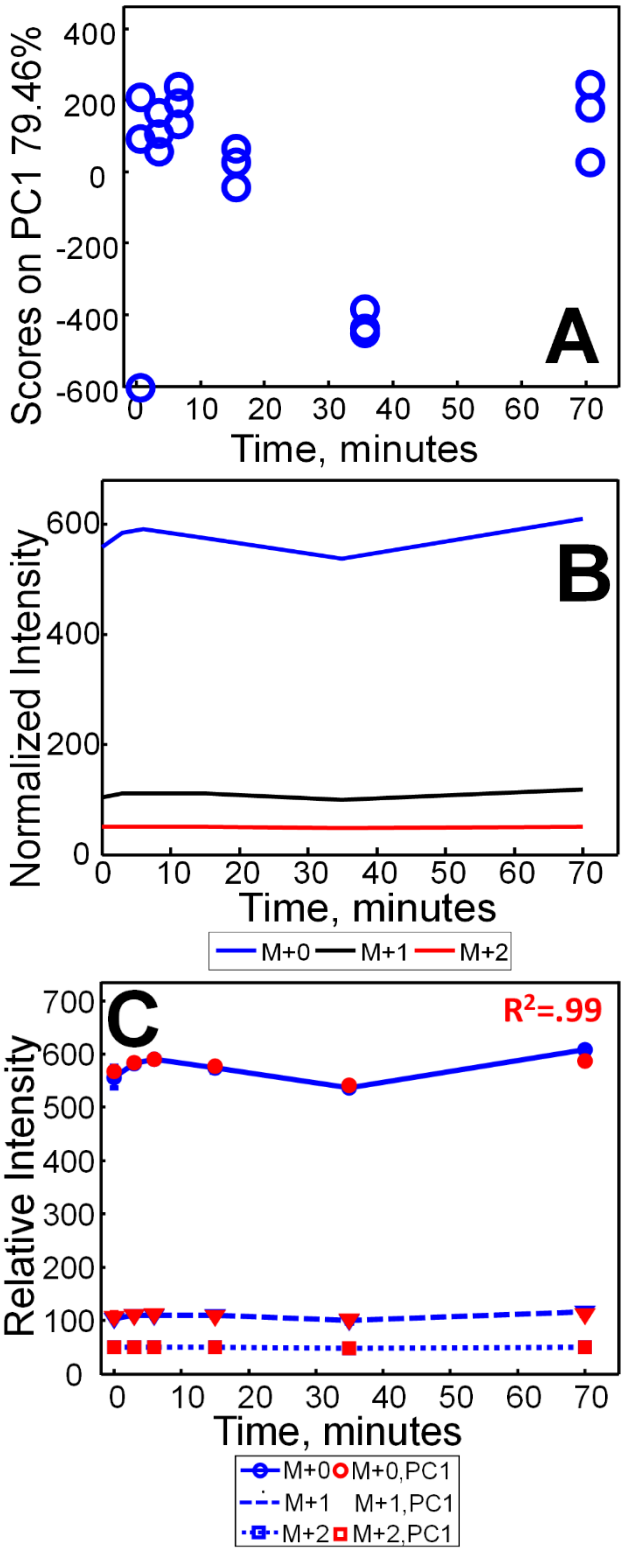

Figure 2.18. PCA results for lactate, including (A) PCA scores vs. time plot of the second PCA model; (B) corresponding M+n plot; and (C) comparison of reconstructed and original data.

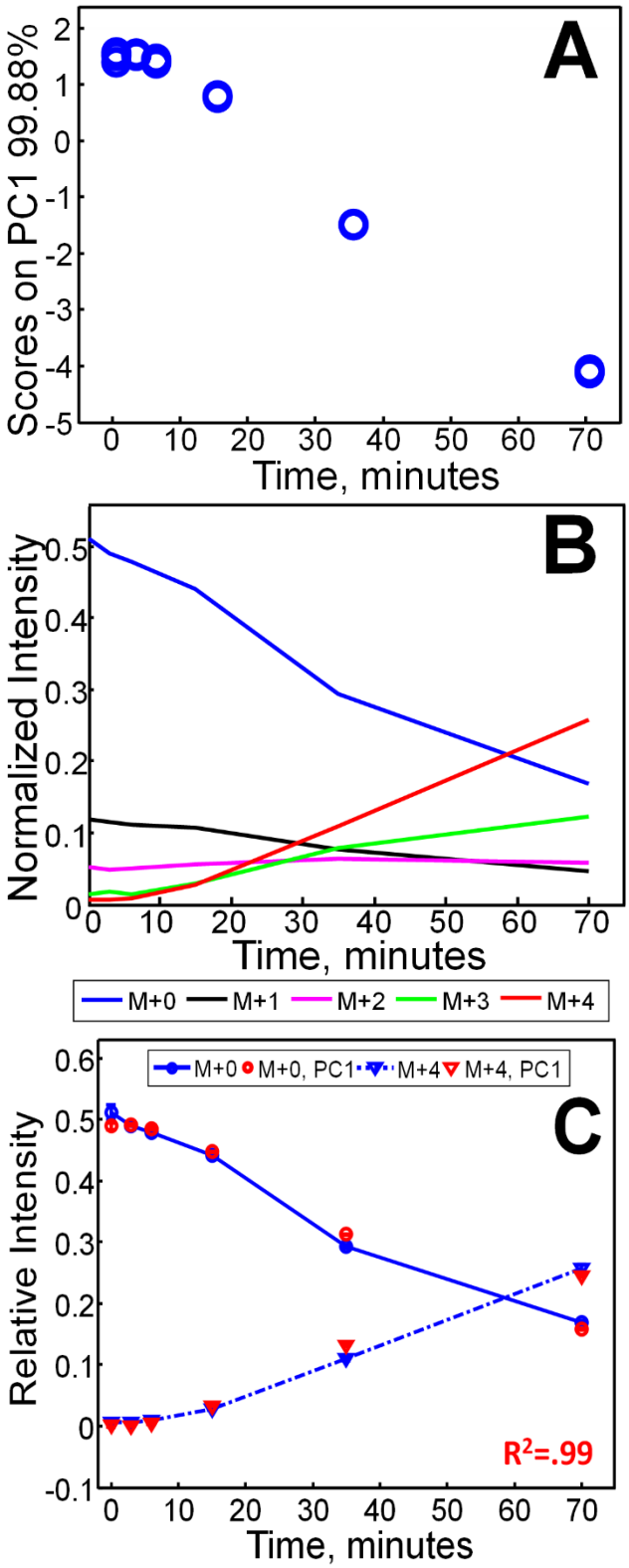

Figure 2.19. PCA results for putrescine, including (A) PCA scores vs. time plot of the second PCA model; (B) corresponding M+n plot; and (C) comparison of reconstructed and original data.

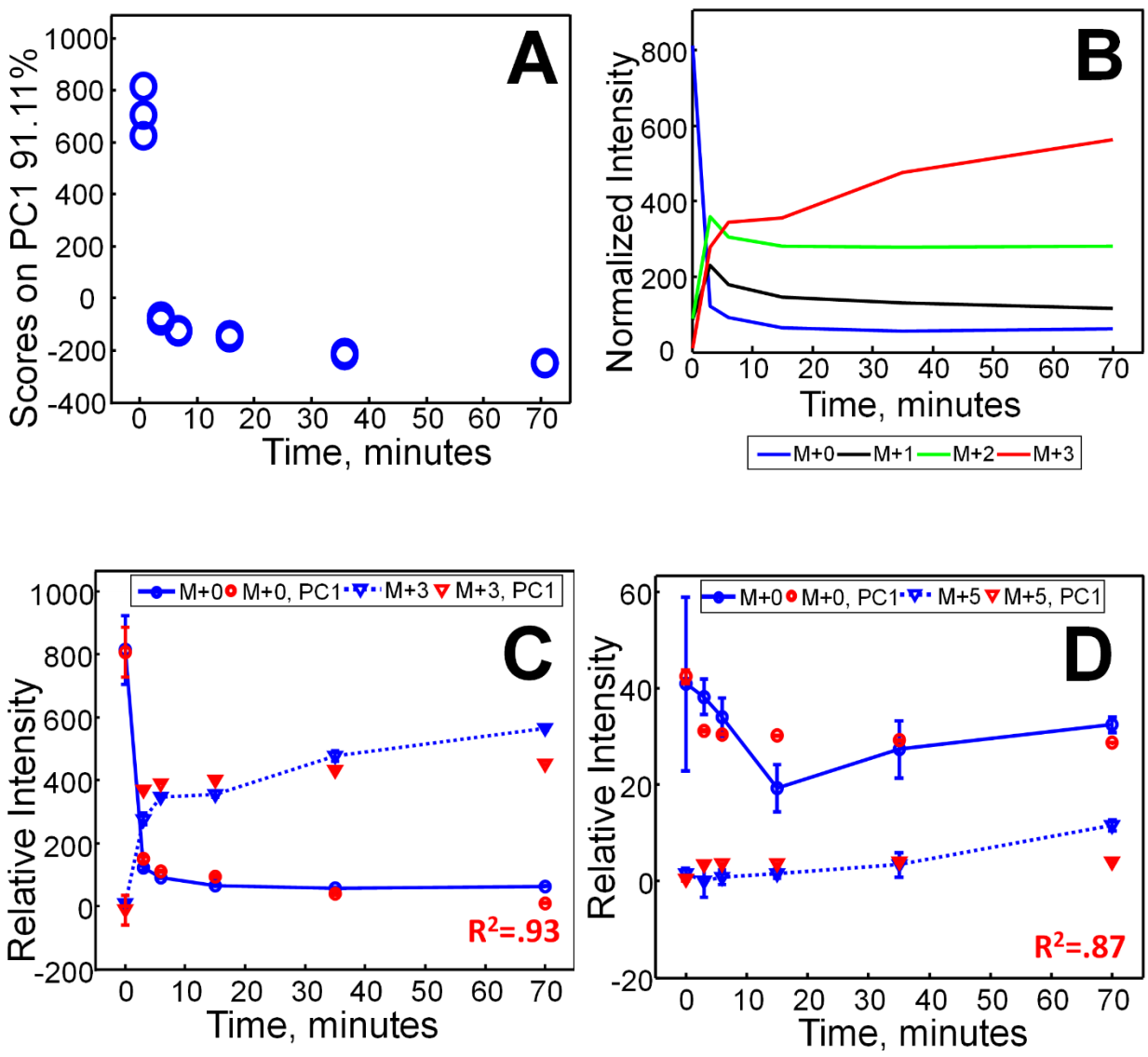

Figure 2.20. PCA results for unidentified metabolite with index 428.41 s, including (A) PCA scores vs. time plot of second PCA model; (B) corresponding M+n plot; (C) comparison of reconstructed and original data for *m/z* that are part of the time course; and (D) comparison of reconstructed and original data for *m/z* that are *not* influencing the time course.

A table summarizing all of the results follows, with the clusters ranked by DCS value such that the metabolite with the highest DCS value and therefore the greatest separation between the 0 mina and 70 minute classes in the first PCA model scores plot is first in the table. The PC% column indicates how much of the variance in the data is explained in PC1 for the PCA 0 min versus 70 min model. The analyte number corresponds to where in the chromatogram the metabolite eluted, as in Table 2-2. For those metabolites that could be identified, their name and match value is provided. A column indicates whether or not the metabolite incorporated ^{13}C successfully (C for Changing) or not (NC for Not Changing). The number of carbons that incorporated ^{13}C is given for those metabolites designated C as elucidated by the loadings plot of the PCA 0 min versus 70 min model. Finally, the % PC1 for the PCA on time course model indicates what percentage of the total variance was encompassed in the first PC. Note the number of metabolites that remain unidentified. It was not the goal of this investigation to successfully identify all metabolites indexed. Identifications were provided where possible, and further exploration into the identity of the metabolites, if desired, would be possible through thorough study of their mass spectra.

Table 2-3: Summary table of quantitative information for 152 mass clusters

Rank	DCS	PC1 %	Analyte #	Name	MV	Flux	#C	Fluxomic PC1 %
1	51.2	99.97	112	5-oxoproline	914	C	5	97.67
2	44.5	99.9	52	alanine	848	C	3	95.88
3	44.1	99.96	130	putrescine	871	C	4	99.88
4	34.6	99.91	100	l-aspartic acid	807	C	4	95.94
5	30.4	99.78	72	unidentified metabolite		C	5	98.67
6	28.3	99.63	139	c00029d: UDP glucose fragment shared with G1P	909	C	4	92.86
7	22.5	99.78	129	D-glucose 1-phosphate main TMS derivative	896	C	5	85.61
8	17.6	99.57	80	L-threonine	760	C	4	91.81
9	15.3	99.68	70	valine	860	C	5	94.11
10	14	99.26	79	unidentified metabolite		C	5	61.26
11	13.8	99.02	111	unidentified metabolite		C	5	95.51
12	13.2	99.41	85	methylsuccinate	771	C	5	89.18
13	13.1	98.78	101	unidentified metabolite		C	6	73.2
14	12.5	98.93	117	unidentified metabolite		C	7	83.95
15	12.4	98.51	62	unidentified metabolite		C	4	74.78
16	11.8	98.28	67	unidentified metabolite		C	5	91.78
17	11.2	99.5	35	unidentified metabolite		C	2	99.6
18	10.5	98.15	55	unidentified metabolite		C	3	87.91
19	10.3	97.61	147	unidentified metabolite		C	4	73.21
20	9.6	67.65	51	unidentified metabolite		C	2	80.57
21	9.5	99.21	140	unidentified metabolite		C	6	88.81
22	8	98.92	93	unidentified metabolite		C	3	90.11
23	7.9	96.63	134	unidentified metabolite		C	5	78.01
24	7.3	98.04	43	pyruvate	851	C	3	92.43
25	6.6	97.21	116	unidentified metabolite		C	5	79.82
26	6.5	95.91	64	unidentified metabolite		C	3	85.65
27	6.5	98.45	76	acetyl-CoA TMS phosphoric acid fragment	937	C	4	90.16
28	6.2	94.28	133	unidentified metabolite		C	5	60.94
29	5.9	96.34	82	glycine	852	C	2	62.87
30	5.9	99.11	34	unidentified metabolite		C	2	99.33
31	5.9	96.44	135	citrate	849	C	6	81.44
32	5.8	95.2	113	unidentified metabolite		C	6	79.43
33	5.8	95.81	95	unidentified metabolite		C	3	94.56
34	5.4	98.12	119	unidentified metabolite		C	6	90.74
35	5.2	96.98	131	unidentified metabolite		C	5	91.94
36	4.9	92.87	36	unidentified metabolite		C	3	99.96
37	4.4	93.41	102	malate	928	C	4	86.91
38	4.1	93.04	132	unidentified metabolite		C	4	86.14
39	4	93.08	105	unidentified metabolite		C	4	69.17
40	3.7	91.1	86	unidentified metabolite		C	7	81.38
41	3.6	86.7	65	leucine	743	C	6	71.4
42	3.4	93.7	87	unidentified metabolite		C	5	91.35
43	3.3	89.74	24	unidentified metabolite		C	2	85.92
44	3.2	86.7	63	3-hydroxybutyrate	806	C	4	44.73
45	3.2	85.9	88	fumarate	588	C	4	60.95
46	3.1	91.24	50	unidentified metabolite		C	3	84.28
47	3.1	86.5	103	unidentified metabolite		C	4	82.52
48	2.9	86.76	110	unidentified metabolite		C		54.73
49	2.9	96.65	42	benzene, 1,2,3-trimethyl-	899	C	2	97.15
50	2.9	81.08	91	unidentified metabolite		C	4	72.69

51	2.8	92.17	46	unidentified metabolite		C	3	70.2
52	2.8	81.06	152	unidentified metabolite		C	3	59.16
53	2.7	95.13	83	succinate	814	C	3	84.52
54	2.6	81.22	144	unidentified metabolite		C	6	46.86
55	2.6	83.92	122	unidentified metabolite		C	3	58.59
56	2.5	94.05	108	unidentified metabolite		C	3	92.92
57	2.5	98.32	60	Standard 381	577	C	7	93.11
58	2.5	77.27	121	unidentified metabolite		C	3	62.52
59	2.3	93.53	40	benzene, 1-ethyl-2-methyl-	878	C	3	91.95
60	2.2	66.24	142	unidentified metabolite		C	4	68.66
61	2.2	72.14	137	unidentified metabolite		C	4	46.07
62	2.1	81.74	99	unidentified metabolite		C	4	67
63	2	96.12	77	unidentified metabolite		C	2	95.21
64	2	79.27	94	unidentified metabolite		C	3	65.97
65	2	64.67	84	unidentified metabolite		C	2	66.38
66	1.9	69.79	141	galactose	856	C	3	88.63
67	1.9	79.91	25	unidentified metabolite		C	5	72.9
68	1.9	88.05	59	oxalate	710	C	2	83.63
69	1.9	88.92	13	trifluoromethyl-bis-(trimethylsilyl)methyl ketone	896	C	4	81.48
70	1.8	90.67	109	unidentified metabolite		C	3	79.63
71	1.8	58.01	104	unidentified metabolite		C	3	80.56
72	1.8	78.87	12	unidentified metabolite		C	2	64.51
73	1.8	55.05	29	unidentified metabolite		NC		50.01
74	1.8	87.1	126	unidentified metabolite		C	5	83
75	1.7	96.33	92	unidentified metabolite		C	5	75.46
76	1.7	89.02	10	unidentified metabolite		C	3	65.85
77	1.7	82.62	26	disilathiane, hexamethyl-	813	C	2	80.95
78	1.7	77.39	115	unidentified metabolite		C	4	89.64
79	1.6	74.56	45	pyridine, 3-trimethylsiloxy-	767	NC		74.79
80	1.6	68.32	39	unidentified metabolite		C	3	57.63
81	1.6	88.64	89	unidentified metabolite		C	3	65.08
82	1.6	56.38	124	unidentified metabolite		NC		28.63
83	1.6	96.89	98	trimethylsilyl 4-methylbenzoate	905	NC		95.84
84	1.6	77.52	8	unidentified metabolite		NC		85.87
85	1.5	98.83	54	unidentified metabolite		NC	4	97.89
86	1.5	61.25	96	unidentified metabolite		NC		48.11
87	1.5	62.57	71	unidentified metabolite		NC		49.07
88	1.5	65.97	61	unidentified metabolite		NC	4	49.01
89	1.4	94.2	148	unidentified metabolite		NC		88.33
90	1.4	55.07	143	unidentified metabolite		NC		45.99
91	1.4	59.01	146	unidentified metabolite		C	5	48.2
92	1.4	74.37	75	unidentified metabolite		NC		89.36
93	1.3	76.56	15	unidentified metabolite		C	3	77.44
94	1.3	41.35	120	unidentified metabolite		NC		25.93
95	1.3	92.88	32	unidentified metabolite		NC		86.46
96	1.2	61.7	57	pentasiloxane, dodecamethyl-	822	NC		93.29
97	1.2	84.91	66	unidentified metabolite		NC		49.28
98	1.2	53.26	107	3,6,9,12-Tetraoxa-2,13-disilatetradecane, 2,2,13,13-tetramethyl-	874	NC		56.71
99	1.2	87.5	123	unidentified metabolite		NC		75.17
100	1.2	90.74	4	unidentified metabolite		NC		65.49

101	1.2	41.62	145	D-sorbitol 6TMS	898	NC		57.63
102	1.2	80.13	97	unidentified metabolite		NC		64.6
103	1.1	80.74	58	unidentified metabolite		NC		84.67
104	1.1	50.17	81	unidentified metabolite		NC		50.79
105	1.1	56.33	27	1,2-bis(trimethylsiloxy)ethane	830	NC		48.85
106	1.1	92.82	69	butanoic acid, 3-methyl-3-[(trimethylsilyl)oxy]-, trimethylsilyl ester	873	C	4	77.32
107	1.1	62.04	41	silane, (cyclohexyloxy)trimethyl-	869	NC		61.69
108	1.1	56.81	106	unidentified metabolite		NC		41.02
109	1.1	44.47	138	unidentified metabolite		NC		41.87
110	1.1	45.64	125	unidentified metabolite		NC		55.52
111	1	53.03	11	unidentified metabolite		NC		31.71
112	1	76.35	14	unidentified metabolite		NC		66.06
113	1	72.43	56	unidentified metabolite		NC		69.92
114	0.9	54.02	128	unidentified metabolite		NC		82.8
115	0.9	80.36	48	unidentified metabolite		NC		79.84
116	0.9	76.4	47	glycolic acid	917	NC		56.06
117	0.9	77.51	73	unidentified metabolite		NC		76.77
118	0.9	60.07	2	unidentified metabolite		NC		52.95
119	0.8	87.23	136	unidentified metabolite		NC		71.49
120	0.8	76.33	78	unidentified metabolite		NC		88.26
121	0.8	63.7	114	unidentified metabolite		NC		96.59
122	0.8	93.6	3	unidentified metabolite		NC		64.94
123	0.7	63.58	74	unidentified metabolite		NC		89.13
124	0.7	90.06	44	lactate	932	NC		79.46
125	0.7	89.07	23	unidentified metabolite		NC		81.13
126	0.7	66.83	118	unidentified metabolite		NC		75.05
127	0.7	54.02	90	unidentified metabolite		NC		33.67
128	0.7	61.87	127	unidentified metabolite		NC		87.18
129	0.7	99.74	18	unidentified metabolite		NC		91.51
130	0.7	61.41	7	unidentified metabolite		NC		62.15
131	0.6	95.81	37	unidentified metabolite		NC		67.33
132	0.6	67.77	68	valeric	661	NC		60.05
133	0.6	87.88	9	unidentified metabolite		NC		73.43
134	0.6	73.13	5	unidentified metabolite		NC		94.91
135	0.6	74.89	150	unidentified metabolite		NC		64.47
136	0.6	93.52	149	unidentified metabolite		NC		71.08
137	0.6	67.49	31	unidentified metabolite		NC		82.95
138	0.5	53.41	19	unidentified metabolite		NC		78.9
139	0.5	73.08	53	unidentified metabolite		NC		80.37
140	0.5	86.33	28	unidentified metabolite		NC		75.21
141	0.5	73.59	38	unidentified metabolite		NC		84.4
142	0.5	97.98	6	unidentified metabolite		NC		62.63
143	0.4	56.65	1	unidentified metabolite		NC		79.74
144	0.4	96.34	17	unidentified metabolite		NC		81.99
145	0.3	54.99	49	unidentified metabolite		NC		43.57
146	0.3	86.68	20	unidentified metabolite		NC		74.16
147	0.3	88.69	151	unidentified metabolite		NC		81.09
148	0.2	95.37	30	unidentified metabolite		NC		88.86
149	0.2	95.2	33	unidentified metabolite		NC		91.95
150	0.1	98.1	16	unidentified metabolite		NC		86.14
151	NaN	99.87	21	unidentified metabolite		C	1	99.12
152	NaN	97.28	22	unidentified metabolite		NC		84.57

2.4 CONCLUSION

We have described a novel analytical method for analysis of a stable isotope time course using GC-TOFMS, based upon the 2D *m/z* cluster plot method and PCA. The 2D *m/z* cluster plot method allows for the discovery and deconvolution of coeluting metabolites at low chromatographic resolution. Metabolites were indexed according to their mass cluster locations and were classified as pure or convoluted. Convoluted mass clusters were either re-indexed or deconvoluted according to their proximity to an adjacent mass cluster. Of the 152 mass clusters identified, 54 mass clusters were pure and 98 mass clusters were convoluted. Of these convoluted mass clusters, 33 were re-indexed and 65 were deconvoluted using CLS. Performing PCA on the pure extracted spectra from the mass clusters differentiates those metabolites which are changing from those that are not changing, and the use of DCS as a quantitative metric allows for objective identification of metabolites that incorporated ^{13}C and visualization of the number of carbons converted to ^{13}C with minimal false positives and false negatives. Additionally, PCA performed on the full time course of a changing metabolite elucidates the time course profile for the ^{13}C-uptake of that metabolite. Although it was beyond the scope of this investigation to calculate the metabolic rate of fluxomic uptake, the quantitative information provided by PCA could be used to calculate the rate at which the cellular metabolism incorporates ^{13}C in a metabolic flux analysis investigation. Of the 152 metabolites surveyed, 83 were identified as changing with time and 69 unchanging with time.

Using a single DCS value as a cutoff for the identification of metabolites that incorporated ^{13}C may result in false positive or false negatives. The use of the 95% confidence interval was used for demonstration of the efficacy of the method, but each investigator should decide what confidence interval is relevant for the data set in question. A range of values near the cutoff value

can also be flagged for further investigation, and the researcher can manually decide by examination of the loadings plot whether the metabolite incorporated ^{13}C or not. Further studies into the time-dependent ^{13}C-labelingof mutant and wild type phenotypes as well as applications to comprehensive metabolic flux analysis studies can be investigated using the method described herein.

2.5 ACKNOWLEDGEMENTS

This work was supported by a grant from the DOE (DESC0006871). This study was previously published in the *Journal of Chromatography A*, 1432 (2016) 111-121.

2.6 REFERENCES

[1] J.K. Nicholson, J.C. Lindon, E. Holmes, "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, Xenobiotica. 29 (1999) 1181–1189. doi:10.1080/004982599238047.

[2] E.M. Humston, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Time-Dependent Profiling of Metabolites from Snf1 Mutant and Wild Type Yeast Cells, Anal. Chem. 80 (2008) 8002–8011. doi:10.1021/ac800998j.

[3] B.E. Alber, Biotechnological potential of the ethylmalonyl-CoA pathway, Appl. Microbiol. Biotechnol. 89 (2011) 17–25. doi:10.1007/s00253-010-2873-z.

[4] G. Winter, J.O. Krömer, Fluxomics – connecting 'omics analysis and phenotypes, Environ. Microbiol. 15 (2013) 1901–1916. doi:10.1111/1462-2920.12064.

[5] U. Sauer, Metabolic networks in motion: 13C-based flux analysis, Mol. Syst. Biol. 2 (2006) 62. doi:10.1038/msb4100109.

[6] K. Hiller, C.M. Metallo, J.K. Kelleher, G. Stephanopoulos, Nontargeted Elucidation of Metabolic Pathways Using Stable-Isotope Tracers and Mass Spectrometry, Anal. Chem. 82 (2010) 6621–6628. doi:10.1021/ac1011574.

[7] W. Wiechert, 13C Metabolic Flux Analysis, Metab. Eng. 3 (2001) 195–206. doi:10.1006/mben.2001.0187.

[8] S.B. Crown, M.R. Antoniewicz, Publishing 13C metabolic flux analysis studies: A review and future perspectives, Metab. Eng. 20 (2013) 42–48. doi:10.1016/j.ymben.2013.08.005.

[9] S. Klein, E. Heinzle, Isotope labeling experiments in metabolomics and fluxomics, WIREs Syst. Biol. Med. 4 (2012) 261–272.

[10] S. Yang, J.S. Nadeau, E.M. Humston-Fulmer, J.C. Hoggard, M.E. Lidstrom, R.E. Synovec, Gas chromatography–mass spectrometry with chemometric analysis for determining 12C and 13C labeled contributions in metabolomics and 13C flux analysis, J. Chromatogr. A. 1240 (2012) 156–164. doi:10.1016/j.chroma.2012.03.072.

[11] S. Yang, M. Sadilek, M.E. Lidstrom, Streamlined pentafluorophenylpropyl column liquid chromatography–tandem quadrupole mass spectrometry and global 13C-labeled internal standards improve performance for quantitative metabolomics in bacteria, J. Chromatogr. A. 1217 (2010) 7401–7410. doi:10.1016/j.chroma.2010.09.055.
[12] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Metabolomics by numbers: acquiring and understanding global metabolite data, Trends Biotechnol. 22 (2004) 245–252. doi:10.1016/j.tibtech.2004.03.007.
[13] S. Yang, M. Sadilek, R.E. Synovec, M.E. Lidstrom, Liquid chromatography–tandem quadrupole mass spectrometry and comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry measurement of targeted metabolites of Methylobacterium extorquens AM1 grown on two different carbon sources, J. Chromatogr. A. 1216 (2009) 3280–3289. doi:10.1016/j.chroma.2009.02.030.
[14] S. Yang, J.C. Hoggard, M.E. Lidstrom, R.E. Synovec, Comprehensive discovery of 13C labeled metabolites in the bacterium Methylobacterium extorquens AM1 using gas chromatography–mass spectrometry, J. Chromatogr. A. 1317 (2013) 175–185. doi:10.1016/j.chroma.2013.08.059.
[15] W. Zeng, J. Hazebroek, M. Beatty, K. Hayes, C. Ponte, C. Maxwell, C.X. Zhong, Analytical Method Evaluation and Discovery of Variation within Maize Varieties in the Context of Food Safety: Transcript Profiling and Metabolomics, J. Agric. Food Chem. 62 (2014) 2997–3009. doi:10.1021/jf405652j.
[16] C.G. Fraga, G.A. Pérez Acosta, M.D. Crenshaw, K. Wallace, G.M. Mong, H.A. Colburn, Impurity Profiling to Match a Nerve Agent to Its Precursor Source for Chemical Forensics Applications, Anal. Chem. 83 (2011) 9564–9572. doi:10.1021/ac202340u.
[17] F.J. Santos, M.T. Galceran, Modern developments in gas chromatography–mass spectrometry-based environmental analysis, J. Chromatogr. A. 1000 (2003) 125–151. doi:10.1016/S0021-9673(03)00305-4.
[18] Z.S. Khan, R.K. Ghosh, R. Girame, S.C. Utture, M. Gadgil, K. Banerjee, D.D. Reddy, N. Johnson, Optimization of a sample preparation method for multiresidue analysis of pesticides in tobacco by single and multi-dimensional gas chromatography-mass spectrometry, J. Chromatogr. A. 1343 (2014) 200–206. doi:10.1016/j.chroma.2014.03.080.
[19] O. Fiehn, J. Kopka, R.N. Trethewey, L. Willmitzer, Identification of Uncommon Plant Metabolites Based on Calculation of Elemental Compositions Using Gas Chromatography and Quadrupole Mass Spectrometry, Anal. Chem. 72 (2000) 3573–3580. doi:10.1021/ac991142i.
[20] 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. doi:10.1016/S0021-9673(00)00300-9.
[21] K.R. Beebe, B.R. Kowalski, An Introduction to Multivariate Calibration and Analysis, Anal. Chem. 59 (1987) 1007A–1017A. doi:10.1021/ac00144a725.
[22] K. Hiller, C. Metallo, G. Stephanopoulos, Elucidation of Cellular Metabolism Via Metabolomics and Stable-Isotope Assisted Metabolomics, Curr. Pharm. Biotechnol. 12 (2011) 1075–1086.
[23] W. Niu, E. Knight, Q. Xia, B.D. McGarvey, Comparative evaluation of eight software programs for alignment of gas chromatography–mass spectrometry chromatograms in metabolomics experiments, J. Chromatogr. A. 1374 (2014) 199–206. doi:10.1016/j.chroma.2014.11.005.

[24] E.M. Humston, J.D. Knowles, A. McShea, R.E. Synovec, Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics, J. Chromatogr. A. 1217 (2010) 1963–1970. doi:10.1016/j.chroma.2010.01.069.
[25] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. doi:10.1016/0169-7439(87)80084-9.
[26] J.S. Nadeau, R.B. Wilson, J.C. Hoggard, B.W. Wright, R.E. Synovec, Study of the interdependency of the data sampling ratio with retention time alignment and principal component analysis for gas chromatography, J. Chromatogr. A. 1218 (2011) 9091–9101. doi:10.1016/j.chroma.2011.10.031.
[27] P.J. Dunlop, C.M. Bignell, J.F. Jackson, D.B. Hibbert, Chemometric analysis of gas chromatographic data of oils from Eucalyptus species, Chemom. Intell. Lab. Syst. 30 (1995) 59–67. doi:10.1016/0169-7439(95)00036-4.
[28] H. Gu, Z. Pan, B. Xi, V. Asiago, B. Musselman, D. Raftery, Principal component directed partial least squares analysis for combining nuclear magnetic resonance and mass spectrometry data in metabolomics: Application to the detection of breast cancer, Anal. Chim. Acta. 686 (2011) 57–63. doi:10.1016/j.aca.2010.11.040.
[29] V.G. Uarrota, R. Moresco, B. Coelho, E. da C. Nunes, L.A.M. Peruch, E. de O. Neubert, M. Rocha, M. Maraschin, Metabolomics combined with chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (Manihot esculenta Crantz) roots during postharvest physiological deterioration, Food Chem. 161 (2014) 67–78. doi:10.1016/j.foodchem.2014.03.110.
[30] I. Olivier, D.T. Loots, A metabolomics approach to characterise and identify various Mycobacterium species, J. Microbiol. Methods. 88 (2012) 419–426. doi:10.1016/j.mimet.2012.01.012.
[31] K. Nöh, K. Grönke, B. Luo, R. Takors, M. Oldiges, W. Wiechert, Metabolic flux analysis at ultra short time scale: Isotopically non-stationary 13C labeling experiments, J. Biotechnol. 129 (2007) 249–267. doi:10.1016/j.jbiotec.2006.11.015.
[32] 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. doi:10.1021/ac5004344.
[33] 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. doi:10.1016/S1044-0305(99)00047-1.
[34] K.M. Pierce, J.C. Hoggard, Chromatographic data analysis. Part 3.3.4: handling hyphenated data in chromatography, Anal. Methods. 6 (2014) 645–653. doi:10.1039/C3AY40965A.
[35] J.S. Nadeau, B.W. Wright, R.E. Synovec, Chemometric analysis of gas chromatography–mass spectrometry data using fast retention time alignment via a total ion current shift function, Talanta. 81 (2010) 120–128. doi:10.1016/j.talanta.2009.11.046.
[36] 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. doi:10.1016/j.chroma.2005.04.078.
[37] M. Mazelis, H.M. Pratt, In Vivo Conversion of 5-Oxoproline to Glutamate by Higher Plants, Plant Physiol. 57 (1976) 85–87. doi:10.1104/pp.57.1.85.

[38] T. Kind, G. Wohlgemuth, D.Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, FiehnLib – mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry, Anal. Chem. 81 (2009) 10038–10048. doi:10.1021/ac9019522.

Chapter 3. Using ROC Curves to Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry[2]

3.1 INTRODUCTION

Comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) is a powerful technique for the analysis of compounds in complex mixtures such as those found in food quality and control [1–3], waste water [4,5], metabolomics [6–9], and petroleum products [10–13]. The inherent complexity of GC × GC – TOFMS data often requires advanced chemometrics for an informative analysis. Analytical approaches and algorithms applied to GC × GC – TOFMS data sets generally fall into two categories: targeted and non-targeted, with the latter further divided into supervised and unsupervised. Targeted analyses focus on preselected, specific analytes of interest. In contrast, discovery-based, non-targeted approaches aim to discover any or all analytes that may be of interest without requiring *a priori* knowledge of their character or identity. Additionally, supervised non-targeted approaches include classification of samples with more emphasis on experimental design [14,15].

We previously reported the development of one such discovery-based software for the comprehensive analysis of GC × GC – TOFMS data. Tile-based Fisher ratio (F-ratio) analysis facilitates a supervised non-targeted method that uses information based upon the experimental design to aid in the discovery of analytes whose variances are statistically different between sample classes [15–18]. Since its initial development [15,16], the tile-based F-ratio software has been

[2] This chapter has been reproduced from B.C. Reaser, B.W. Wright, and R.E. Synovec *Analytical Chemistry*, 2017, 89 (6), 3606-3612.

applied to the analysis of yeast metabolites [18] and chemically-altered diesel fuel [17], successfully elucidating statistically significant class-distinguishing analytes in complex sample matrices.

Method validation is an important part of all analytical technology developments. Several agencies, including the International Union of Pure and Applied Chemistry (IUPAC) [19], Eurachem [20], and the Association of Analytical Communities (AOAC) [21] have published guidelines on the performance characteristics required to properly validate a new analytical method. While each agency has its own specific recommendations, they have a few key requirements in common: sensitivity, selectivity, accuracy, precision, working range, limit of detection (LOD) and limit of quantification (LOQ). While these guidelines are well established for new analytical methods, no such clearly published guidelines exist for the validation of new chemometric algorithms and software.

Introduction of a new chemometric algorithms and software should require rigorous validation and optimization; however, often new software algorithms are simply demonstrated and put into implementation, without rigorous validation. Validation may include applying the method on standard mixes of known composition [15,16] or employing previously analyzed data sets [18] where the "correct" answer has previously been determined. In the case of employing a standard mixture, the analyst will often assume that every analyte in the mixture is in fact a true positive when the sample matrix or analysis conditions may prevent that from actually being the case. The use of quantitative metrics, such as standard errors, correlation coefficients or statistical metrics like F-tests or Students' t-tests, are commonly utilized for validating a chemometric algorithm. The thorough study and evaluation of a software platform should also go beyond the validation stage. Key input parameters should be evaluated, to ensure algorithm optimization. The optimization of

software algorithms and their various parameters is rarely studied in detail, likely because it can be a tedious and subjective process if one lacks the proper quantitative metric. To address this issue, herein we report the optimization of the tile-based F-ratio software [15–18] using the area under the curve (AUC) of a receiver operating characteristic (ROC) curve as a quantitative metric [22,23]. While the AUC for ROC curves has been applied for other purposes, implementation of the AUC as a quantitative metric for software optimization as presented herein is a new approach.

ROC curves were developed during World War II in an attempt to quantify the ability of a radar operator to differentiate between the signals of ally and enemy aircraft [24]. Since then, ROC curves have been used extensively in medical diagnostics [25–27], evaluating analytical method performance [22,28,29], and machine learning [30], but ROC curves have remained under-utilized in analytical and chemometric analyses [23]. While machine learning and chemometric analyses are inherently interdisciplinary and often share similar algorithms, they have distinct differences. Machine learning is predominately a computer science methodology that involves computers learning without being explicitly programmed, while chemometrics is the application of linear algebra coupled with statistics to extract meaningful information from chemical systems.

A ROC curve is a plot of the true-positive probability (TPP) versus the false positive probability (FPP) at a particular threshold defined by the analyst, usually relating to a minimum signal or concentration. The terms true positive rate (TPR) and false positive rate (FPR) are also used throughout the literature and can be used interchangeably with TPP and FPP, respectively. For the purposes of the investigation reported herein, we will use TPP and FPP. The TPP, also known as the sensitivity, is equal to the sum of the true positives over the total number of positive instances. The FPP, also known as 1-specificity, is equal to the sum of the false positives over the total number of negative instances. The AUC is a popular metric for the quantitative, statistically-

based evaluation of ROC curves. It is equivalent to the probability of stochastic domination in non-parametric statistics; that is, the AUC defines the probability that a randomly selected value will be correctly ranked as either a true positive or false positive. Values for the AUC range from 0.5 for a "useless" test to 1.0 for a perfect test.

In the initial tile-based F-ratio software development, a signal-to-noise (*S/N*) threshold of 3 and all of the mass channels (*m/z*) were utilized [15]. The *S/N* threshold is used in the data reduction step prior to the calculation of F-ratios, and the number of *m/z* is used to calculate the average F-ratio value for each tile. Upon gaining more experience with the software, we have recognized that a rigorous evaluation of these two parameters is warranted. Accordingly, herein we employ the AUC metric for the ROC curve to evaluate and optimize these two key parameters. Each ROC curve quantitatively indicates how the *S/N* threshold and number of *m/z* both impact the sensitivity of the software to discover chemical features of interest. The hypothesis is that a higher *S/N* threshold (but not too high) coupled with use of less than all of the *m/z* would likely improve the capability of finding and more highly ranking true positives, while reducing the number of false positives.

Fifty analytes were spiked into a diesel fuel at two similar concentration spike levels (30 native and 20 non-native compounds), to serve as a pair of suitably complex samples to render challenging the software optimization using the AUC approach. Thus, the theoretical maximum number of positive instances is 50. The two spiked diesel samples defined two classes for F-ratio comparison. The *S/N* threshold used in the data reduction step prior to the calculation of F-ratios, and number of *m/z* used to calculate the average F-ratio values were varied, and F-ratio hit lists were evaluated. A total of 25 combinations of *S/N* threshold by number of *m/z* were studied. Varying the *S/N* threshold and number of *m/z* engenders optimization of the tile-based F-ratio

software, to find the "sweet spot" in terms of *S/N* threshold, coupled with use of only the most chemically selective *m/z* (i.e., those with the highest F-ratios). Additionally, it is demonstrated that the number of positive instances does not need to be known prior to analysis, as the AUC metric still provides adequate information regarding the sensitivity of the F-ratio analysis.

3.2 EXPERIMENTAL

Diesel fuel (~ 1 gallon) was collected from a fueling station in the Seattle area. A non-native internal standard, 1,2,3-trichlorobenzene (0.5170 g), was spiked into 1 L of diesel fuel to create a stock solution at 608 ppm trichlorobenzene. This stock solution was used for all serial dilutions. Two neat spike solutions were made of 30 native compounds and 20 non-native compounds by adding ~ 0.1 g of each component to a scintillation vial. A list of each compound and exact mass is provided in Table 3-1 (natives) and Table 3-2 (non-natives). The native and non-native compounds were then quantitatively transferred using stock solution to an amber bottle of known mass. Stock solution was then added to each of the two neat spike samples until the final mass reached ~ 100 g, creating ~ 1000 ppm native and ~ 1000 ppm non-native solutions. Approximately 80 g of the native spike solution and 8 g of the non-native spike solution were then transferred to a third amber bottle. Stock solution was added to bring the total mass of the contents to ~ 100 g, creating a spike solution containing ~ 800 ppm native compounds and ~ 80 ppm non-native compounds in diesel. This solution was diluted by half in stock solution successively until two spike solutions were obtained at the following nominal concentrations: 200/20 and 100/10 ppm native/non-native. The exact concentrations of the analytes in these spike solutions is provided in Table 3-1 and Table 3-2. A neat standards mix was also created in order to find accurate retention times for each analyte. This was created by adding 250 μL of each compound into a scintillation vial.

GC × GC–TOFMS data were collected using an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA) with a LECO Pegasus III TOFMS equipped with a 4D thermal modulator upgrade (LECO, St. Joseph, MI, USA). A reverse column configuration was employed with a polar 29.75 m × 250 μm × 0.25 μm Rxi-17Sil MS film primary column, ^{1}D, and a nonpolar 1.5 m × 180 μm × 0.18 μm Rxi-1ms film (Restek, Bellefonte, PA, USA) secondary column, ^{2}D. The GC inlet was set to 275 °C and the transfer line was set to 280 °C. Ultra-high purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) was used as the carrier gas at a constant flow of 2.0 mL/min. An Agilent 7683 autosampler was used to inject 1 μL for diesel samples or 0.1 μL of neat standards mix with a 300:1 split ratio. The primary column was maintained at 40 °C for 1 min, then ramped at 5 °C/min to 300 °C where it was held for 2 min. The secondary column and the modulator block followed the same temperature program with a +15 °C and +60 °C offset, respectively. The modulation period was 2 s with 0.5 s hot and cold pulses for each stage. The ion source was set to 225 °C, the electron impact energy was 70 eV, and the detector voltage was 1740 V. Mass channels, *m/z* 35-300, were collected at 100 spectra/s after a 10 s acquisition delay. Six injection replicates of each spike sample (200/20 and100/10 ppm native/non-native with internal standard) and the stock solution (diesel with internal standard), and two injection replicates of the neat standards were analyzed using the GC × GC–TOFMS method above. These latter two injection replicates were used to determine the retention times of the spiked standards. Due to the lack of variability between the retention times of the two injection replicates, it was determined that two injection replicates were sufficient to obtain an average retention time for each spiked analyte.

The GC × GC–TOFMS data were data processed using the instrument software (ChromaTOF, version 3.32, LECO, St. Joseph, MI, USA), including baseline correction, peak finding and peak area calculation. Peak areas were calculated at unique *m/z* for the 30 native

compounds in the stock solution and in each spike sample. Standard addition method (SAM) plots were prepared (see

Figure **3.5**) to determine the native concentration of these analytes in diesel as reported in Table 3-1. The GC × GC–TOFMS data were also imported from the instrument software into Matlab R2015b (Mathworks Inc., Natick, MA, USA) using an in-house data converting algorithm. The data were analyzed with the in-house developed tile-based F-ratio software, an algorithm that elucidates chemical features that distinguish between two classes of samples, such as two different spike concentration levels. We direct the reader to previous publications and their respective supplementary information for a thorough description of the tile-based F-ratio software [15–18]. For all F-ratio software executions, a ^{1}D tile size of 6 s, a ^{2}D tile size of 100 ms, a ^{1}D cluster size of 4 s, a ^{2}D cluster size of 60 ms were employed. All chromatograms were normalized to the sum of the total ion current (TIC), and retention time alignment was not required. The *S/N* threshold and number of *m/z* utilized for the calculation of the average F-ratio were varied to find the optimum values for software performance. The *S/N* threshold values examined were 3, 6, 10, 30 and 100, while the number of *m/z* examined were 3, 6, 10, 20, and all *m/z*, for a total of 25 *S/N* by *m/z* combinations. At least 3 *m/z* were required for a measureable F-ratio, and no F-ratio threshold minimum was employed.

The *S/N* threshold is applied by calculating the standard deviation of the noise on each tile region and multiplying that standard deviation by the factor given by the desired threshold (i.e., by a factor of 3, 6, 10, 30 or 100 per the *S/N* thresholds). Any *m/z* with a signal in the same tile falling below that *S/N* threshold is then excluded from the analysis. The number of *m/z* used for the analysis simply refers to the maximum number of *m/z* whose F-ratios are used to calculate the average F-ratio of the tile, with the minimum being 3 *m/z*. For example, if a tile has a feature

containing 52 *m/z* occurring above the *S/N* threshold and the number of *m/z* to be examined was set to 10 *m/z*, only the 10 *m/z* with the highest F-ratio values would be included in calculation of the average F-ratio. However, if a tile has a feature having only 6 *m/z*, all 6 *m/z* would be utilized for the calculation of the average F-ratio even though the software was set to examine 10 *m/z* because the tile meets the criteria of having at least 3 *m/z* above the *S/N* threshold.

3.3 Results and Discussion

The reverse column GC × GC configuration utilized a substantial portion of the 2D separation space, and was well suited for the separation presented in Figure 1.1(A) of the compound classes of interest in diesel fuel: alkanes, olefins, cyclic alkanes, and aromatics. The standards utilized for the spike study also encompass the 2D region of the chromatogram as shown in Figure 1(B), where each standard peak is numbered according to its respective analyte number in Table 3-1 and Table 3-2.

The F-ratio software was employed for the sample class comparison of the six injection replicates of each spiked diesel sample (200/20 versus 100/10 ppm native/non-native compounds) to generate a hit list of class distinguishing analytes. While the software was evaluated for 25 *S/N* by *m/z* combinations, we focus initially on the previously employed parameters, a *S/N* threshold of 3 and all *m/z* [15,17,18], followed by a detailed focus on the combination of parameters that turned out to be optimal, and finish with a summary for all 25 combinations. The distribution of the average F-ratio values in Figure 3.2(A) corresponds to the class comparison using a *S/N* threshold of 3 and all *m/z*. The maximum of the F-ratio distribution, and therefore the F-ratio with the greatest frequency, is about 0.5, and the sharp peak of the distribution drops nearly to baseline at an F-ratio of about 4. The tail of the distribution (shown in the inset figure) begins to be sparse from about 10 to 638.9 (hit #1, cyclooctane), and contains the class distinguishing chemical

features (at higher F-ratio) that are most likely to be true positives and thus most likely to be the spiked analytes. Using the known 2D retention times of the spiked standards, the hit list was evaluated. Hits that corresponded to spiked analytes were designated "true positives," while the rest were designated "false positives." The hit list was investigated until the 50th false positive was identified. Figure 3.2(B) shows the F-ratio value versus hit number in the F-ratio hit list. The blue circles correspond to true positives (spiked standards), while the red circles correspond to false positives. The F-ratios range from 638.9 (hit #1, cyclooctane) to 3.2 (hit #83, false positive).

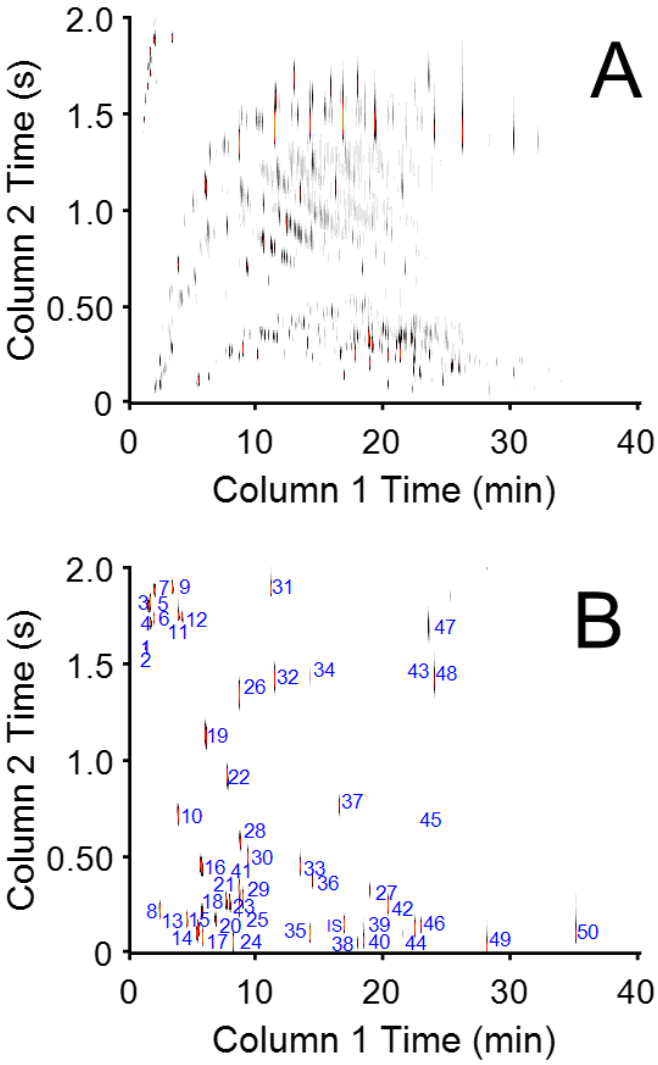

Figure 3.1. (A) GC × GC – TOFMS chromatogram of diesel fuel; and (B) chromatogram of native and non-native spiked analytes numbered according to Table 3-1 and Table 3-2. IS marks the elution location of the internal standard.

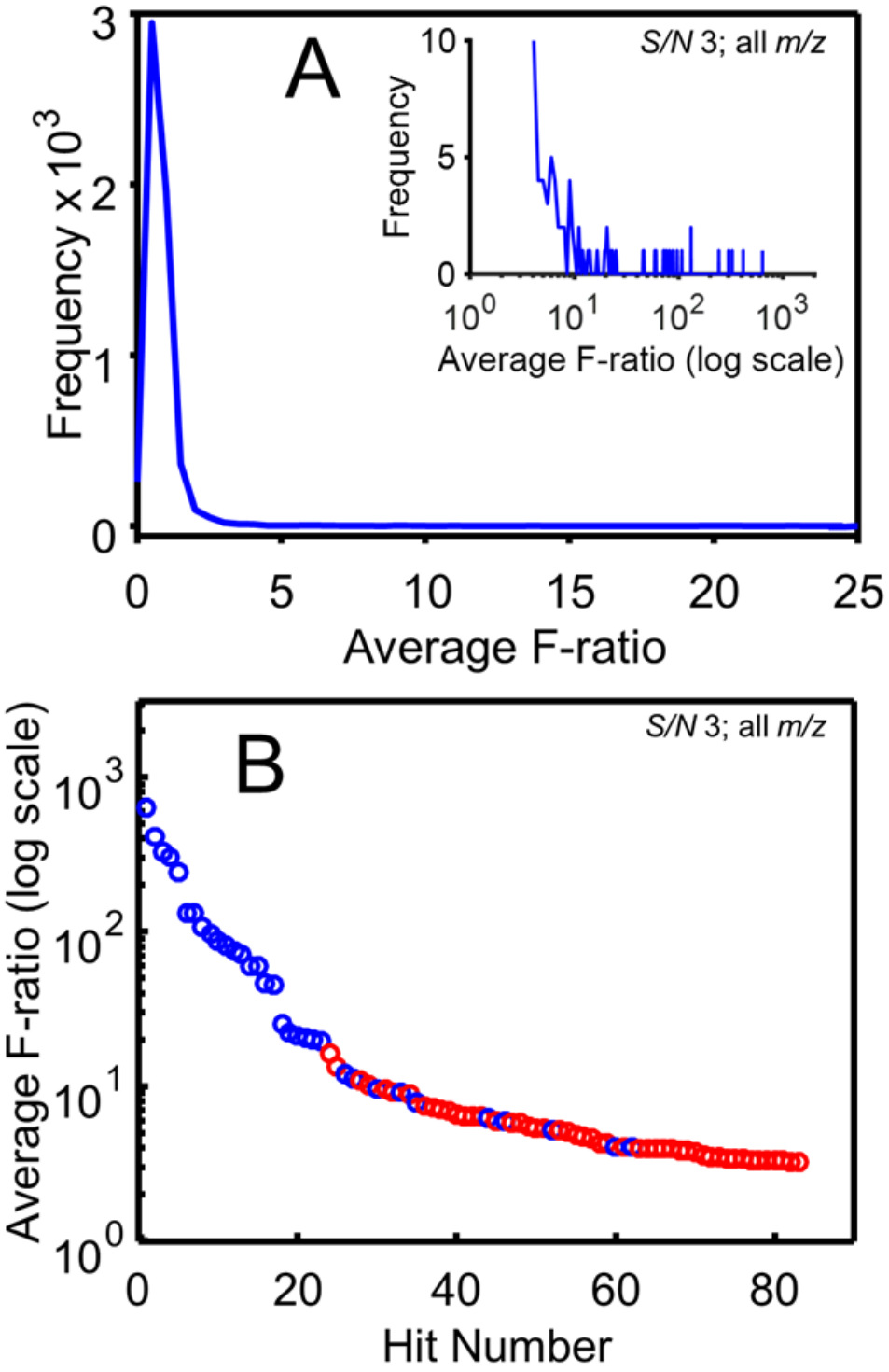

Figure 3.2. (A) The F-ratio distribution for the preliminary hit list of the 200/20 ppm versus 100/10 ppm comparison using a *S/N* threshold of 3 and all *m/z*; and (B) Log of average F-ratio versus hit number based upon the F-ratio distribution in (A), with true positives (spiked analytes) shown in blue and false positives shown in red.

Next, the results for the combination of parameters that turned out to be optimal, with a *S/N* threshold of 10 and 10 *m/z*, is provided in Figure 3.3(A). This class comparison distribution has a maximum F-ratio at about 1 and drops to baseline at an F-ratio of about 6.5. The area under the F-ratio distribution in Figure 3.3(A) is about 20% less than that of the curve in Figure 3.2(A). This indicates use of the stricter parameters reduced the number of total hits returned by the F-ratio software (4,854 for *S/N* threshold of 10 and 10 *m/z* versus 5780 for *S/N* threshold of 3 and all *m/z*). The shape of the curve is also shifted down and to the right, indicating a shift to higher F-ratio values as a consequence of using the top 10 *m/z* for calculating the average F-ratio, decreasing the ability of noisy *m/z* to lower the average. The tail of the distribution (shown in the inset figure) begins to be sparse from about 20 to 2014 (hit #1, cyclooctane), and contains the class distinguishing chemical features that are most likely to be true positives and thus spiked analytes. Figure 3.3(B) shows the F-ratio value versus hit number in the F-ratio hit list for this class comparison. Compared to the previous parameters utilized for Figure 3.2(B), that is *S/N* threshold 3; all *m/z*, the true positives (blue) are shifted up and to the left, while the false positives (red) fall down and to the right due to the better combination of parameters, that is *S/N* threshold 10; 10 *m/z*. The F-ratios in Figure 3.3(B) range from 2014 (hit #1, cyclooctane) to 5.0 (hit #87, false positive), a much greater range of average F-ratio values compared to that observed in Figure 3.2(B). In Figure 3.3, the exclusion of non-class distinguishing *m/z* increases the overall average F-ratio values, prioritizes true positives, and therefore improves the sensitivity of the discovery-based method.

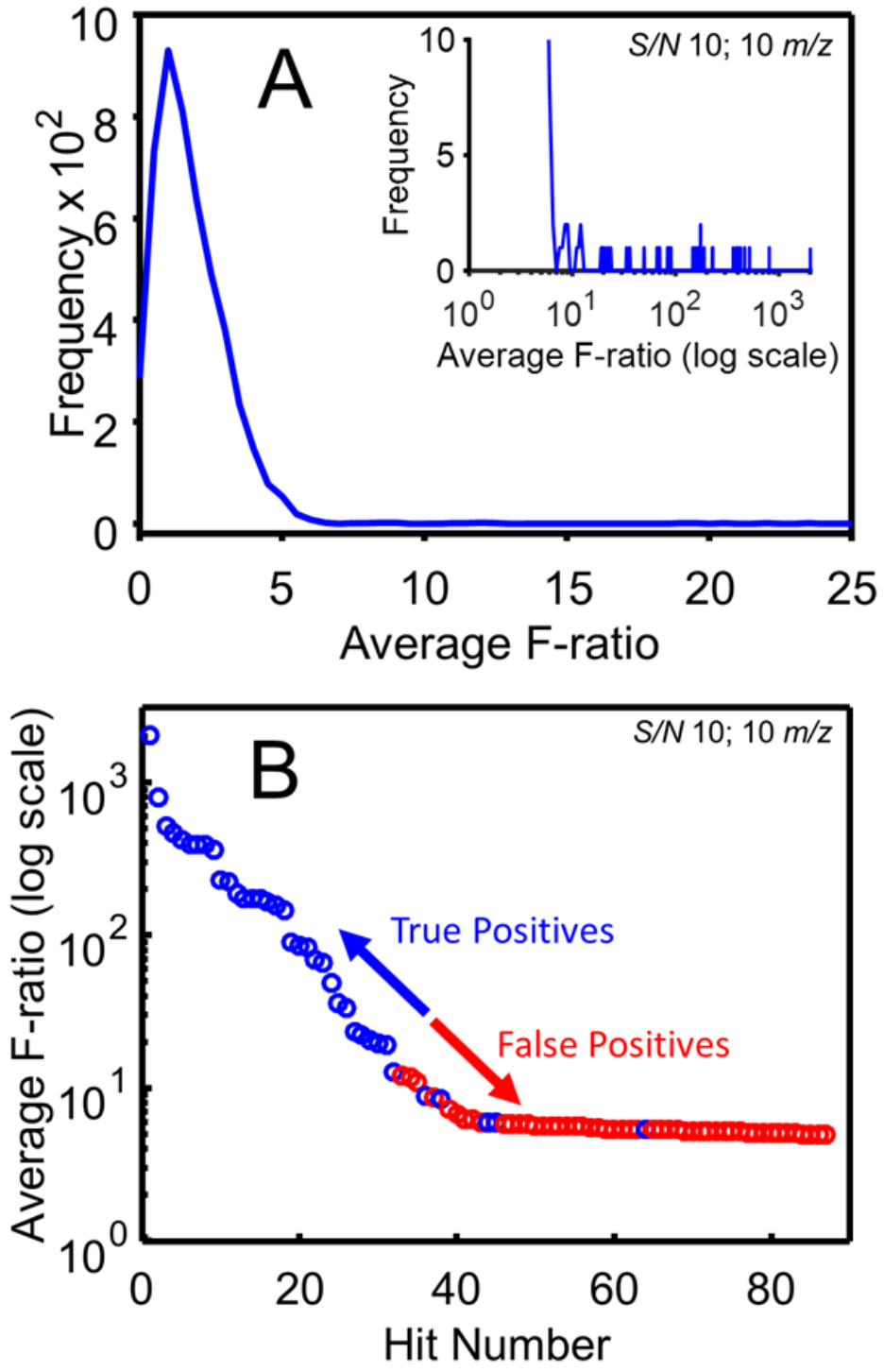

Figure 3.3. (A) The F-ratio distribution for the preliminary hits list of the 200/20 ppm versus 100/10 ppm comparison at optimized conditions of the tile-based F-ratio algorithm, *S/N* threshold of 10 and 10 *m/z*; and (B) Log of average F-ratio versus hit number based upon the F-ratio distribution in (A), with true positives (spiked analytes) shown in blue and false positives shown in red.

While the results in Figure 3.2(B) and Figure 3.3(B) were readily interpreted to gain insight into the F-ratio software performance, the interpretation is qualitative and not analytically rigorous. Therefore, using ROC curves, an AUC quantitative metric was investigated as a means to evaluate and optimize these two key parameters: *S/N* threshold and number of *m/z* required. The ROC curves for these two F-ratio analyses are presented in Figure 3.4 where the red curve corresponds to the parameters *S/N* threshold of 3 and all *m/z*, while the blue curve corresponds to the parameters *S/N* threshold of 10 and 10 *m/z*.

For the ROC curves shown in Figure 3.4, it was assumed that all 50 analytes spiked into the diesel fuel would be discovered as true positives such that the positive instances used for calculating the TPP was 50. For symmetry, 50 negative instances were evaluated for the FPP, effectively minimizing the length of the hit list evaluated while sufficiently investigating the ROC curves as they plateau.

Specifically, these ROC curves were constructed from the same hit list that created Figure 3.2(B) and Figure 3.3(B). Each hit was designated a true or false positive, and a running sum of true positives (TP) and a running sum of false positives (FP) were calculated until the 50^{th} false positive. These TP and FP running sums were divided by the number of positive instances (P) or number of negative instances (N), respectively, both of which were 50 as discussed in the previous paragraph. For example, for the blue curve in Figure 3.4(A), the first 32 hits were consecutive true positives, as shown by the blue circles in Figure 3.3(B). This gives the first point on the ROC curve with an TPP of 32/50 or 0.54 and an FPP of 0/50 or 0. The next three consecutive hits are false positives, given by the red circles in Figure 3(B). This yields the next point on the ROC curve with the TPP remaining equal to 32/50 or 0.64 and an FPP of 3/50 or 0.06. The plotting of the TPP

and FPP continues in this fashion until the 50th false positive, when FPP reaches 1.0, which for the blue ROC curve in Figure 3.4(A) corresponds to hit number 87, as seen in Figure 3.3(B)

As indicated by the shift upward and to the left of the true positives in Figure 3.3(B), the blue ROC curve in Figure 3.4 shows the increase in sensitivity, or TPP, for a given FPP, or 1-specificity, by using *S/N* 10 and 10 *m/z.* Increasing the *S/N* threshold from 3 to 10 decreased the influence of overly noisy *m/z* signals. Using 10 *m/z* rather than all *m/z* ensures that the 10 largest F-ratios for a given tile are used to determine the average F-ratio, thus decreasing the influence of smaller F-ratios at *m/z* with spurious covariance. A quantitative metric, the AUC can aid in the comparison of ROC curves generated by changing the algorithm parameters. For example, the AUC is 0.61 for the ROC curve using a *S/N* threshold of 3 and all *m/z*, while the AUC is 0.74 for the ROC curve using a *S/N* threshold of 10 and 10 *m/z*. This means, for the combination of a *S/N* threshold of 10 and 10 *m/z* is quantitatively superior to using a *S/N* threshold of 3 and all *m/z*, with a ~ 21% improvement in software performance.

One might presume that, upon software optimization, the AUC would approach the ideal value of 1.00. However, the experimental design provided very challenging datasets due to the number of native compounds naturally occurring at a high concentration in the diesel. Keeping in mind that native concentrations are not known *a priori*, some of these compounds when spiked in at the 200 ppm and 100 ppm level, were not apt to be *discovered* due to the lack of a statistically sufficient increase in the total concentration given by the spike. This is a major reason why the ROC curves level off at a sensitivity well below 1.00. For Figure 3.4, it was assumed that all 50 spiked analytes had a statistically sufficient concentration difference between the two spike levels. In a sense, the number of positive instances was not actually 50, but rather is likely much less. Despite this issue, the ROC curve and the AUC metric were able to differentiate the increase in

sensitivity using a *S/N* threshold of 10 and 10 *m/z* compared to using a *S/N* threshold of 3 and all *m/z*.

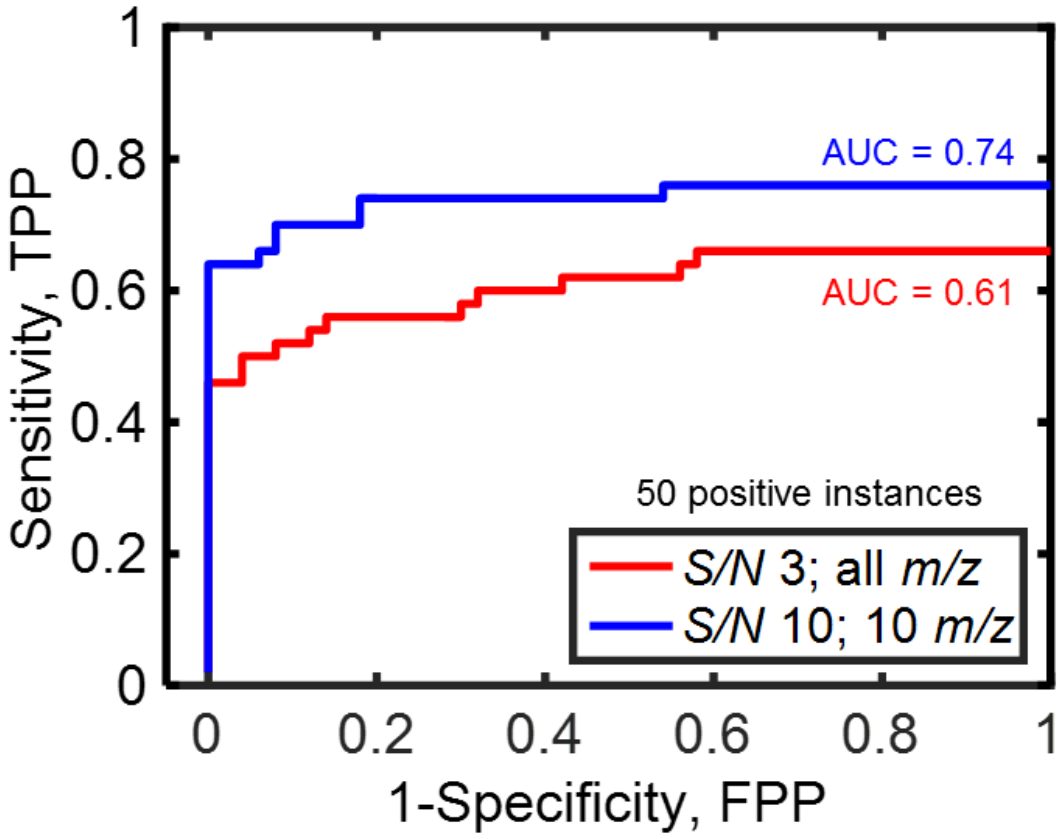

Figure 3.4. ROC curves for the two sets of parameters studied: (red) *S/N* threshold of 3 and all *m/z* (based upon the results in Figure 2), and (bue) *S/N* threshold of 10 and10 *m/z*.

In order to further investigate this issue, for the purpose of determining the number of spiked compounds that could (or should) serve as positive instances, the native spiked analyte concentrations were determined using the SAM (classical standard addition method) with the two spike levels, nominally 200 ppm and 100 ppm, and the stock solution, nominally 0 ppm. The concentrations of the native analytes calculated from the SAM plots, as demonstrated in

Figure **3.5**, are provided in Table 3-1. Worth noting are the three analytes with 0.0 ppm concentration in the diesel: 2,2,4-trimethylpentane, cyclohexene, and cyclooctane. These

compounds, though having negligible concentrations in this investigation so were not detected, can be found regularly in other blends of diesel. For each native compound, the native concentration was added to the actual spike concentration (at both spike levels), and the concentration ratio was determined as shown in Table 3-1. We have previously shown that tile-based F-ratio analysis is capable of discovering analytes at a concentration ratio between sample classes down to ~ 1.06 [16], Accordingly, the native compounds that have a concentration ratio less than 1.06 are shaded in grey in Table 3-1. In order to statistically determine which of the 30 spiked native compounds were not legitimate positive instances, a t-test was performed as described in Harris [31], with the results provided in the last column of Table 3-1. The t-test values were calculated using the peak areas of a selective *m/z* for each native analyte in all six injection replicates of the 200 ppm, 100 ppm and the stock solution (that is, nominally 0 ppm). The peak areas were normalized to the internal standard. At a 95% confidence interval, the t-table value for a two-tailed test with 5 degrees of freedom is 2.57, so all t-values less than that correspond to analytes that have no statistically significant difference in their concentration between the 200 ppm and 100 ppm spike levels, as designated by yellow highlighting in Table S1 in Supplemental. Nine analytes were determined not to be legitimate positive instances. There is good agreement between the concentration ratio and the t-test metric for the discovery of true positives. Other than hexane, that appears to be an outlier, the minimum concentration ratio corresponding to a true positive appears to be ~ 1.04, in good agreement with a previous study [16]. The cause of the outlying nature of hexane is unknown. Quantification was re-performed to confirm initial results, but inconsistencies may result from slight coelution in the second dimension with cyclopentane.

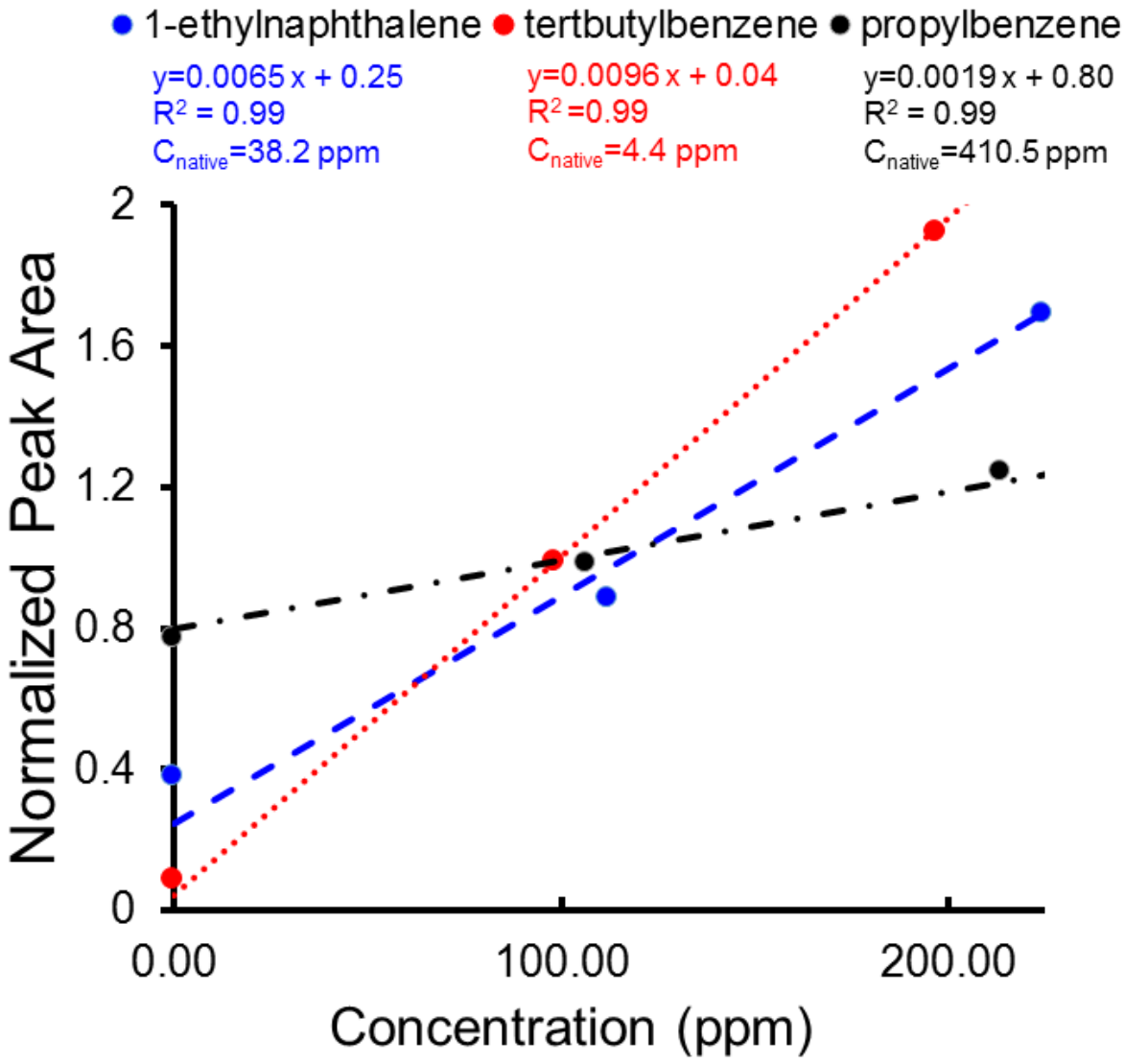

Figure 3.5. The standard addition method (SAM) plots for 1-ethylnapthalene (blue), tertbutylbenzene (red) and propylbenzene (black).

Table 3-1. Table of native components with the analyte number as shown in Figure 3.1(B), actual concentrations, concentration ratios and t-test values.

Natives	Name	Mass (g)	Actual Concentration (ppm)			Concentration Ratio	T-test
			200	100	Native		
1	hexane	0.1104	217.7	108.8	690.5	1.14	0.52
2	cyclopentane	0.1135	223.8	111.8	267.7	1.30	4.57
3	2,2,4-trimethylpentane	0.1090	214.9	107.4	0.0	2.00	12.51
4	cyclohexane	0.1037	204.5	102.2	907.0	1.10	3.45
5	heptane	0.1164	229.5	114.7	582.8	1.16	4.37
6	cyclohexene	0.1066	210.2	105.0	0.0	2.00	17.31
7	methylcyclohexane	0.1027	202.5	101.2	3978.2	1.02	2.10
8	octane	0.1052	207.4	103.7	1154.0	1.08	4.36
9	toluene	0.1180	232.6	116.3	1353.7	1.08	6.99
10	nonane	0.1129	222.6	111.2	2149.8	1.05	3.47
14	ethyl benzene	0.1118	220.4	110.2	475.3	1.19	7.30
15	p-xylene	0.1070	211.0	105.4	1882.4	1.05	5.24
16	cyclooctane	0.1131	223.0	111.4	0.0	2.00	32.92
19	decane	0.1186	233.8	116.9	4940.4	1.02	2.48
21	propylbenzene	0.1082	213.3	106.6	410.5	1.21	17.64
22	butylcylcohexane	0.1082	213.3	106.6	907.6	1.11	3.75
23	3-ethyltoluene	0.1106	218.1	109.0	1521.9	1.07	4.64
26	undecane	0.1015	200.1	100.0	9171.3	1.01	2.05
27	cyclohexylbenzene	0.1184	233.4	116.7	766.4	1.13	6.01
29	1,2,4-trimethylbenzene	0.1102	217.3	108.6	2797.8	1.04	4.19
32	dodecane	0.1065	210.0	104.9	16370.2	1.01	1.41
34	tridecane	0.1173	231.3	115.6	20322.6	1.01	1.88
37	bicyclohexyl	0.1200	236.6	118.2	510.3	1.19	6.19
41	tertbutylbenzene	0.0997	196.6	98.2	4.4	1.96	46.74
42	1,5-dimethyltetralin	0.1076	212.1	106.0	1957.7	1.05	3.05
43	hexadecane	0.1113	219.4	109.7	10483.0	1.01	1.62
44	1-ethylnaphthalene	0.1138	224.4	112.1	38.2	1.75	11.88
46	1,3-dimethylnaphthalene	0.1110	218.8	109.4	1756.5	1.06	4.90
47	pristane	0.1080	212.9	106.4	2648.8	1.04	1.13
48	heptadecane	0.1184	233.4	116.7	8272.5	1.01	0.65

Table 3-2. Table of non-native components and the actual concentrations present in the spike solution with the analyte number as shown in Figure 3.1(B). As designated by the asterisk, 2-mercaptoethanol is not found in the 200/20 vs 100/10 ppm class comparison.

Non-Natives	Name	Mass (g)	Actual Concentration (ppm)	
			20	10
11	pyridine	0.1026	21.7	10.9
12	2-mercaptoethanol*	0.1225	25.9	13.0
13	1-chlorohexane	0.1080	22.9	11.4
17	2,5-dimethylthiophene	0.1200	25.4	12.7
18	2-heptanol	0.1106	23.4	11.7
20	methyl caproate	0.1054	22.3	11.2
24	bromobenzene	0.1066	22.6	11.3
25	3-octanone	0.1166	24.7	12.3
28	limonene	0.1222	25.9	12.9
30	5-decyne	0.1039	22.0	11.0
31	aniline	0.1206	25.5	12.8
33	nonanol	0.1160	24.6	12.3
35	1,6-dichlorohexane	0.1112	23.5	11.8
36	2-decanone	0.1150	24.4	12.2
38	cyclohexyl isothiocyanate	0.1213	25.7	12.8
39	ethyl salicylate	0.1106	23.4	11.7
40	butyrophenone	0.1054	22.3	11.2
45	dodecanethiol	0.1120	23.7	11.9
49	diphenyl sulfide	0.1176	24.9	12.4
50	dibutylphthalate	0.1054	22.3	11.2

Using the list of native compounds corrected using the t-test metric, ROC curves were made in Figure 3.6 with a TPP out of 41 total positive instances instead of the previously assumed 50 for Figure 3.4. While the shape of each curve does not change, the AUC value obviously increases, from 0.61 to 0.74 for a *S/N* threshold of 3 and all *m/z*, and from 0.74 to 0.90 for a *S/N* threshold of 10 and 10 *m/z*. It is important to note that the analyst can draw the same conclusions from Figure 3.4 or (B), which has important ramifications for implementing the ROC curve-based AUC metric. Essentially, optimization of the software parameters does not require *a priori* determination of the statistically correct number of positive instances in the sample classes.

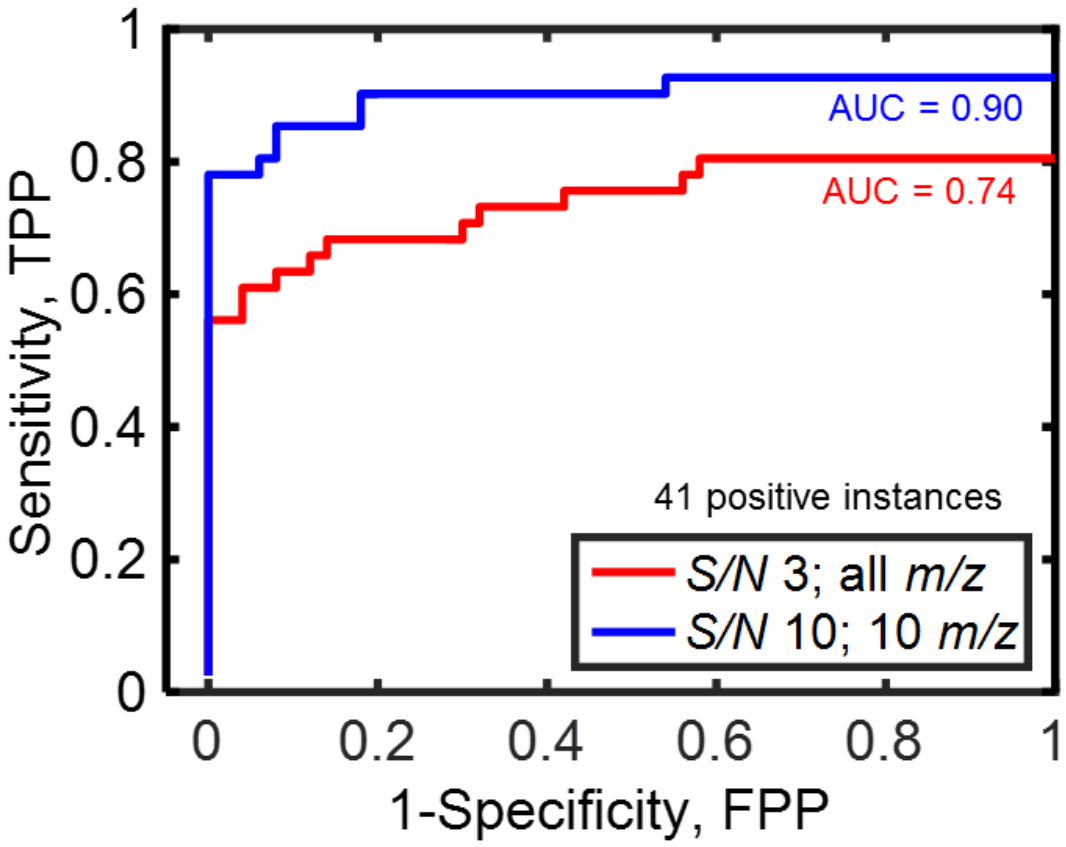

Figure 3.6. ROC curves for the two sets of parameters studied: (red) *S/N* threshold of 3 and all *m/z* (based upon the results in Figure 2), and (blue) *S/N* threshold of 10 and10 *m/z* with the TPP calculated using only the 41 statistically significant positive instances as calculated by the t-test in Table 3-1.

To more fully explore the effects of *S/N* threshold and number of *m/z* on F-ratio software performance, ROC curves were generated for the 25 combinations of *S/N* threshold and number of *m/z* defined in the Experimental Section. The area under the ROC curve (AUC) was calculated for each combination of *S/N* threshold and *m/z*, whereby the AUC provided a quantitative metric to assess the effect of these parameters on the discrimination of the F-ratio software. The AUC values for all 25 parameter combinations are shown as a heat map in Figure 3.7, where Figure 3.7(A) corresponds to AUCs determined assuming all 50 spiked analytes correspond to positive instances and Figure 3.7(B) corresponds to AUCs determined using only the 41 statistically significant positive instances. Most notably, in Figure 3.7(A) the optimum parameters remain an *S/N* 10 coupled with 10 *m/z*, producing an AUC of 0.74 assuming 50 positive instances, and in Figure 3.7(B) an AUC of 0.90 using only 41 positive instances. The previously employed parameters (*S/N* 3 coupled with all *m/z*) produced an AUC of 0.61 or 0.74, respectively. The difference between these two AUC values (0.13) relative to the prior value (0.61) is 0.21, or a 21% improvement in the F-ratio software performance. Hence, optimization of the *S/N* threshold and number of *m/z* used for this particular data set provides a 21% increase in the discrimination of the true positives from the false positives.

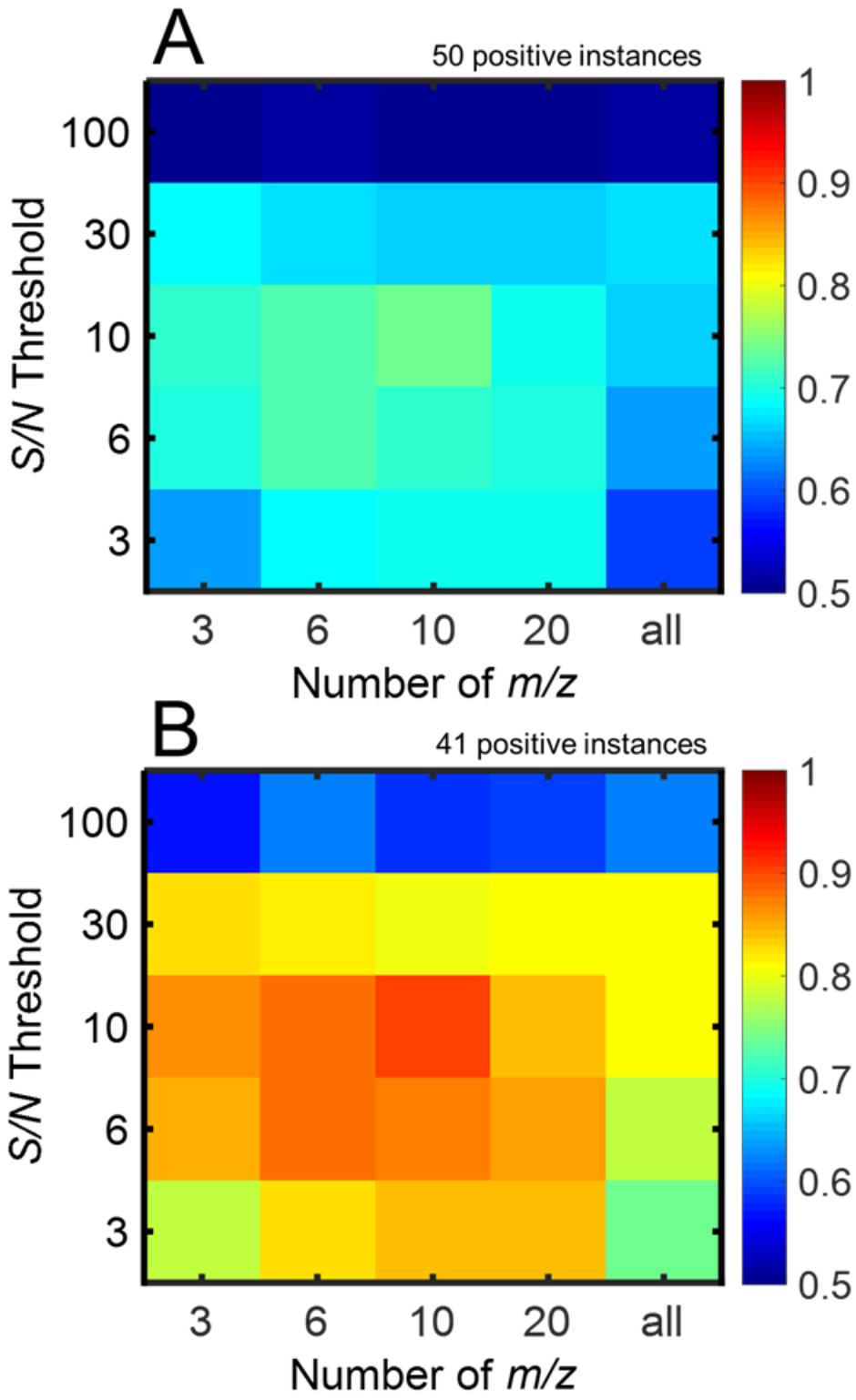

Figure 3.7. Heat maps showing the area under the curve (AUC) at the 25 *S/N* threshold and number of *m/z* combinations studied (A) assuming all 50 spiked standards were positive instances; and (B) using only 41 statistically significant positive instances.

Beyond interpretation of the AUC, perhaps most elucidative of the increase in software performance is the comparison of the number of true positives discovered at different FPP thresholds. For example, the previously utilized parameters (*S/N* 3; all *m/z*) found a total of 33 true positives, while the optimized parameters (*S/N* 10; 10 *m/z*) found a total number of 38 true positives of the 41 possible true instances. Not only did the optimized parameters find 5 additional hits total (i.e., at an FPP of 1.0), but using the optimized parameters allowed the F-ratio algorithm to discover the vast majority of those true positives at much lower FPP thresholds than those required by the previously utilized parameters. For example, the greatest distance between the two ROC curves shown in Figure 4 occurs at an FPP of about 0.2. An FPP of 0.2 corresponds to when the F-ratio hit list has reached 10 false positives. At this FPP threshold, the optimized parameters (*S/N* 10; 10 *m/z*) have discovered 37 of the 38 true positives that were discovered overall, which corresponds to 9 more true positives than the 28 that the software discovered using the previous parameters (*S/N* 3; all *m/z*). The optimization of the parameters (using *S/N* 10 and 10 *m/z)* allowed the method to discover nearly all of the true positives at the cost of only 10 false positives.

It is worth mentioning that the AUC is not the only metric of ROC performance, and other metrics may be utilized that are more efficacious for other investigations. For example, as previously mentioned, one may count the number of true positives at a lower FPP threshold, such as the 0.2 or 10 false positives. The results of this metric are shown in

Figure **3.8**, where the same conclusion, that the parameters of *S/N* 10 and 10 *m/z* are most optimal, remains the same. However, for most analysts, evaluating a hit list to the 10^{th} false positive may be a sufficient investigation of the hit list and more time-efficient. Therefore, this

may serve as a better ROC metric than the AUC, including in cases where the AUC may not indicate clearly that one set of parameters is obviously superior to another.

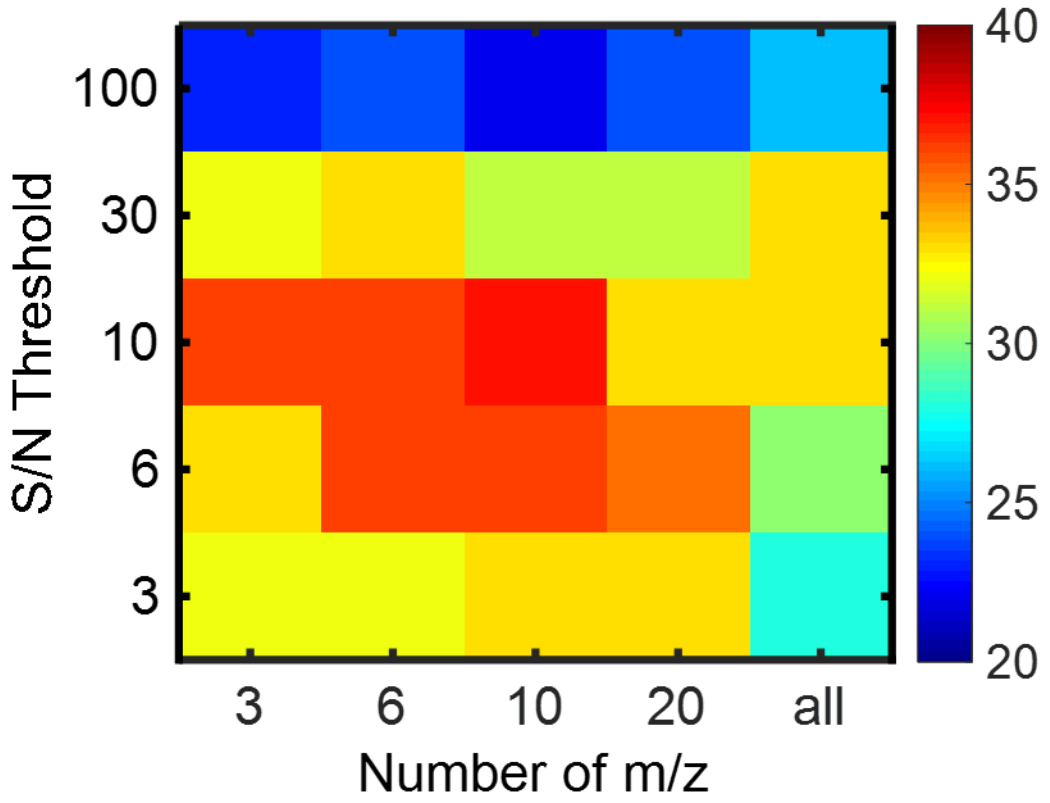

Figure 3.8. Heat map of the number of true positives (TP) at a false positive probability (FPP) of 0.2, that is, 10 allowed false positives, for each of the parameter combinations.

3.4 CONCLUSION

The AUC is a quantitative metric to compare ROC curves to study and optimize chemometric software performance and to demonstrate how various input parameters employed by the software can be tuned for better analysis. We report the implementation of such an approach on the tile-based F-ratio analysis software of GC × GC – TOFMS data using diesel fuel spiked with 30 native and 20 non-native compounds at two nominal concentrations, 200/20 ppm and 100/10 ppm, native/non-native respectively. By varying the *S/N* threshold and the number of F-ratios per *m/z* used to calculate the average F-ratio of each tile, the AUC quantitatively determined the best parameters, found to be a *S/N* threshold of 10 and 10 *m/z*. Compared to the previously employed parameters of *S/N* threshold of 3 and all *m/z*, the use of these parameters provided a ~21% improvement in the ability of the software to discriminate between true positives and false positives in the hit list. It also provided the method with the ability to find 9 additional true positives at an FPP of 0.2 (10 allowed false positives) and 5 additional true positives overall. Furthermore, we demonstrate that the AUC metric can be used effectively without prior knowledge of whether or not the spiked analytes (i.e., the intended true positives) were in fact statistically significant positive instances.

For practical application of this method to other sample matrices and chemometric software algorithms, a few comments of merit remain. Firstly, the optimization process explained herein used two sample classes fabricated and spiked into a diesel matrix to serve as the classes for optimization of the tile-based F-ratio method specifically. For other software algorithms, it may be beneficial to pick a matrix of interest and fabricate the sample classes or spike test mixture for comparison based on the research goals or software type. For example, a software algorithm designed for isotopic labeling investigations of metabolomes should utilize isotopically labeled

metabolites in a biological matrix. The spiked standards chosen for this study were chosen based on sample classes of interest in diesel and in no way were optimized for the software algorithm or GC × GC separation. Evaluation of whatever test mixture one chooses to employ may provide increased optimization of the software algorithm where the identities of the analytes in the test mixture are of increased importance and affect the outcome of the AUC metric for determining the optimized software parameters. Secondly, due to the time-consuming and repetitive nature of the software optimization method presented herein, it can be assumed that once the software algorithm is optimized, it can then be applied to real samples. This is worth stating explicitly for two reasons. Primarily that the calculations and evaluations need not be repeated once a software algorithm is optimized. Secondarily, the application of these optimized parameters to the real samples will aid in the discovery of true chemical markers in the most efficient way, that is, elucidating the highest number of true positives with the fewest number of false positives possible. Lastly, this optimization method, namely the use of the AUC of a ROC curve as a quantitative metric, can be applied to many other software methods and investigative questions. For example, peak detection is a common algorithmic step in many software packages applied to chromatographic data. This method could be applied to any peak table generating software algorithm in order to find the optimal parameters. Results of such a study are not presented here with the results of the tile-based F-ratio software due to the latter's proven ability to perform better than peak tables methods at finding class distinguishing chemical features, especially at low concentration signal differences [15].

3.5 ACKNOWLEDGEMENTS

This work was supported by the Internal Revenue Service (IRS) under an Interagency Agreement with the U.S. Department of Energy (DOE) under Contract DE-AC-5-76RLO with the Pacific Northwest National Laboratory. The authors thank Dr. Brendon Parsons for his assistance throughout this investigation. This published in *Analytical Chemistry* online February 16, 2017.

3.6 REFERENCES

[1] L. Zhang, Z. Zeng, C. Zhao, H. Kong, X. Lu, G. Xu, A comparative study of volatile components in green, oolong and black teas by using comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry and multivariate data analysis, J. Chromatogr. A. 1313 (2013) 245–252. doi:10.1016/j.chroma.2013.06.022.

[2] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, P. Schieberle, S.E. Reichenbach, Q. Tao, Profiling food volatiles by comprehensive two-dimensional ga schromatography coupled with mass spectrometry: Advanced fingerprinting approaches for comparative analysis of the volatile fraction of roasted hazelnuts (Corylus avellana L.) from different origins, J. Chromatogr. A. 1217 (2010) 5848–5858. doi:10.1016/j.chroma.2010.07.006.

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

[4] S. Prebihalo, A. Brockman, J. Cochran, F.L. Dorman, Determination of emerging contaminants in wastewater utilizing comprehensive two-dimensional gas-chromatography coupled with time-of-flight mass spectrometry, J. Chromatogr. A. 1419 (2015) 109–115. doi:10.1016/j.chroma.2015.09.080.

[5] N. Ochiai, T. Ieda, K. Sasamoto, Y. Takazawa, S. Hashimoto, A. Fushimi, K. Tanabe, Stir bar sorptive extraction and comprehensive two-dimensional gas chromatography coupled to high-resolution time-of-flight mass spectrometry for ultra-trace analysis of organochlorine pesticides in river water, J. Chromatogr. A. 1218 (2011) 6851–6860. doi:10.1016/j.chroma.2011.08.027.

[6] S. Castillo, I. Mattila, J. Miettinen, M. Orešič, T. Hyötyläinen, Data Analysis Tool for Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry, Anal. Chem. 83 (2011) 3058–3067. doi:10.1021/ac103308x.

[7] W. Welthagen, R.A. Shellie, J. Spranger, M. Ristow, R. Zimmermann, O. Fiehn, Comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC-TOF) for high resolution metabolomics: biomarker discovery on spleen tissue extracts of obese NZO compared to lean C57BL/6 mice, Metabolomics. 1 (n.d.) 65–73. doi:10.1007/s11306-005-1108-2.

[8] 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. doi:10.1021/ac052106o.
[9] R.E. Mohler, B.P. Tu, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Identification and evaluation of cycling yeast metabolites in two-dimensional comprehensive gas chromatography–time-of-flight-mass spectrometry data, J. Chromatogr. A. 1186 (2008) 401–411. doi:10.1016/j.chroma.2007.10.063.
[10] M.K. Jennerwein, M. Eschner, T. Gröger, T. Wilharm, R. Zimmermann, Complete Group-Type Quantification of Petroleum Middle Distillates Based on Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC×GC-TOFMS) and Visual Basic Scripting, Energy Fuels. 28 (2014) 5670–5681. doi:10.1021/ef501247h.
[11] K.D. Nizio, T.M. MGinitie, J.J. Harynuk, Comprehensive multidimensional separations for the analysis of petroleum, J. Chromatogr. A. 1255 (2012) 12–23. doi:10.1016/j.chroma.2012.01.078.
[12] C. von Mühlen, C.A. Zini, E.B. Caramão, P.J. Marriott, Applications of comprehensive two-dimensional gas chromatography to the characterization of petrochemical and related samples, J. Chromatogr. A. 1105 (2006) 39–50. doi:10.1016/j.chroma.2005.09.036.
[13] 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. doi:10.1016/j.chroma.2013.12.060.
[14] 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. doi:10.1016/j.chroma.2012.05.050.
[15] 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. doi:10.1021/ac504472s.
[16] 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. doi:10.1016/j.talanta.2013.06.038.
[17] 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. doi:10.1016/j.chroma.2016.02.067.
[18] 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. doi:10.1016/j.chroma.2016.06.067.
[19] M. Thompson, S.L.R. Ellison, R. Wood, Harmonized Guidelines for Single-Laboratory Validation of Methods of Analysis, Pure Appl. Chem. 74 (2002) 835–855.
[20] B. Magnusson, U. Ornemark, Eurachem Guide: The Fitness for Purpose of Analytical Methods -- A Laboratory Guide of Method Validation and Related Topics, (2014).
[21] Official methods of Analysis of AOAC International, 6th ed., AOAC International, 1995.

[22] C.G. Fraga, A.M. Melville, B.W. Wright, ROC-curve approach for determining the detection limit of a field chemical sensor, Analyst. 132 (2007) 230–236. doi:10.1039/B607843E.
[23] C.D. Brown, H.T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemom. Intell. Lab. Syst. 80 (2006) 24–38. doi:10.1016/j.chemolab.2005.05.004.
[24] D.M. Green, J.A. Swets, Signal Detection Theory and Psychophysics, John Wiley & Sons, Inc., 1966.
[25] C.E. Metz, Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems, J. Am. Coll. Radiol. 3 (2006) 413–422. doi:10.1016/j.jacr.2006.02.021.
[26] S.G. Baker, The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer, J. Natl. Cancer Inst. 95 (2003) 511–515. doi:10.1093/jnci/95.7.511.
[27] J.V. Carter, J. Pan, S.N. Rai, S. Galandiuk, ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves, Surgery. 159 (2016) 1638–1645. doi:10.1016/j.surg.2015.12.029.
[28] B.J. Blaise, Data-Driven Sample Size Determination for Metabolic Phenotyping Studies, Anal. Chem. 85 (2013) 8943–8950. doi:10.1021/ac4022314.
[29] S. Duraipandian, W. Zheng, J. Ng, J.J.H. Low, A. Ilancheran, Z. Huang, Simultaneous Fingerprint and High-Wavenumber Confocal Raman Spectroscopy Enhances Early Detection of Cervical Precancer In Vivo, Anal. Chem. 84 (2012) 5913–5919. doi:10.1021/ac300394f.
[30] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159. doi:10.1016/S0031-3203(96)00142-2.
[31] D.C. Harris, *Quantitative Chemical Analysis*, 6th ed.; W. H. Freeman and Company: New York, NY, 2003.

Chapter 4. Application of the Optimized Tile-based Fisher Ratio Method to a Process Analytical Chemistry Investigation

4.1 INTRODUCTION

Process analytical chemistry (PAC) is a specialization of analytical chemistry that aims to optimize a chemical process through the analysis of the physical and chemical composition of the starting material, products and/or bi-products of a manufacturing process. PAC may employ various instrumental analysis platforms, chemometric algorithms, and sampling methods to elucidate both qualitative and quantitative information about the chemical process of interest [1,2]. PAC is generally related to and often employed in process analytical technology (PAT) used in the pharmaceutical industry [3].

Gas chromatography coupled with mass spectrometry (GC-MS) is one popular instrumental platforms employed in PAC investigations due to the ability to simultaneously separate and identify components in complex mixtures. GC-MS is a robust technique for evaluating both qualitative and quantitative information regarding a sample. Comprehensive two dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC - TOFMS) is an analytical technique employing two complementary GC separations in series, allowing for the separation of components that otherwise may be unresolved on a single column. GC × GC – TOFMS has been used extensively for the analysis of complex samples such as essential oils [4], food products [5], forensics [6], petrochemicals [7], metabolomics [8], and more [9].

The data generated from modern analytical platforms such as GC × GC – TOFMS requires advanced chemometric algorithms in order to elucidate meaningful information regarding the samples. One such chemometric technique is the Fisher ratio (F-ratio) method, a statistical analysis of variance between sample classes as dictated by the experimental design. An F-ratio is calculated

as the variance between classes divided by the sum of the variances within each class, such that the greater the magnitude of the F-ratio, the more variance there is between two sample classes. We have previously reported on the development of a tile-based F-ratio software specifically designed to elucidate class-distinguishing features in GC × GC – TOFMS in supervised, non-targeted investigations [10,11]. Since then, the software has been successfully applied to the characterization of chemically altered diesel fuel [12], evaluated on a yeast metabolome [13], and optimized using spiked diesel samples, as discussed in Chapter 3.

Herein, we report the application of the newly optimized tile-based F-ratio method based on the results from the aforementioned study to a process analytical chemistry application. We apply both the optimized chromatographic and tile-based F-ratio software parameters to evaluate solvent samples from a polymerization plant experiencing catalyst yield reduction due to one or more catalyst poisons present in the samples. The objective of the study was two-fold: first, to evaluate the differences between the composition of solvent samples from the two sampling points in the polymerization process; and secondly, to determine which, if any, components present in the solvent samples might be contributing to the catalyst yield reduction. The first objective was obtained by performing the tile-based F-ratio method with the two sampling points sampled during excellent polymerization processes serving as the two classes for comparison. Using the information from these results, a sample from a poor polymerization process was evaluated, and compounds of interest were quantified. Parallel factor analysis (PARAFAC) was utilized in an attempt to remove the noise and obtain clean mass spectra from these compounds of interest for improved identification.

4.2 EXPERIMENTAL

4.2.1 *Samples*

Approximately 2 mL of solvent were taken from two sampling points in the polymerization process on various days with specific date and time stamps. A flowchart of the polymerization process is shown in Fig 1, with the two sampling points of interest shown as roman numerals in red. Each sampling date and time point will be heretofore called a "campaign," as shown in Table 4-1. For example, campaign E1 refers to all "Excellent" samples collected on 11/12/2014 at 00:30 hours. Samples from Points I and II were taken at four different campaigns during "Excellent" polymerization processes (E1, E2, E3 and E4), while a sample from Point II for a single campaign was taken during a "Bad" polymerization process (B1). Samples were placed into 2 mL GC vials for analysis. GC vial caps and septa were changed after each injection to avoid solvent evaporation.

Table 4-1. Table of campaigns and sampling points analyzed.

Campaign	Date	Hour	Point	Purification
B1	12/09/2012	20:50	II	Bad
E1	11/12/2014	00:30	I	Excellent
E1	11/12/2014	00:30	II	Excellent
E2	17/12/2014	22:00	I	Excellent
E2	17/12/2014	22:00	II	Excellent
E3	19/12/2014	21:15	I	Excellent
E3	19/12/2014	21:15	II	Excellent
E4	23/12/2014	04:10	I	Excellent
E4	23/12/2014	04:10	II	Excellent

4.2.2 *Analysis of samples via GC × GC – TOFMS*

GC × GC–TOFMS data were collected using an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA) with a LECO Pegasus III TOFMS equipped with a 4D thermal modulator upgrade (LECO, St. Joseph, MI, USA). The same columns and instrumental parameters were utilized as in Chapter 3. Briefly, a polar 29.75 m × 250 μm × 0.25 μm Rxi-17Sil MS film primary column (column 1) and a nonpolar 1.3 m × 180 μm × 0.18 μm Rxi-1ms film (Restek, Bellefonte, PA, USA) secondary column (column 2) were used with ultra-high purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) as the carrier gas at a constant flow of 2.0 mL/min. An Agilent 7683 autosampler was used to inject 1 μL of the solvent samples with a 300:1 split ratio. Column 1 was maintained at 40 °C for one minute, then ramped at 5 °C/min to 300 °C where it was held for 2 minutes. Column 2 and the modulator block followed the same temperature program with a +15 °C and +60 °C offset, respectively. The modulation period was 2 seconds with 0.5 second hot and cold pulses for each stage. Mass channels, *m/z* 35-400, were collected at 100 spectra/s after a 10 second acquisition delay. Two injection replicates of each solvent sample outlined in Table 4-1 were obtained and analyzed in a mathematically randomized order.

Two standards, 2-hexanone and 2-ethyl-1-hexanol were obtained (Sigma-Aldrich, St. Louis MO, USA). Approximately 60 mg of each standard was measured using an analytical balance into a scintillation vial and diluted with approximately 60 g of hexane standard (Sigma-Aldrich, St. Louis MO, USA), to create a solution with ~1000 ppm of each standard. Using hexane, the solution was serially diluted, by mass, to the following nominal concentrations: 100 ppm, 10 ppm, 1 ppm, and 0.1 ppm. These were analyzed using the aforementioned GC × GC – TOFMS method with triplicate injection replicates.

4.2.3 *Data Analysis*

As with the instrumental parameters, the data analysis was performed as in Chapter 3. The GC × GC–TOFMS data were data processed using the instrument software (ChromaTOF, version 3.32, LECO, St. Joseph, MI, USA), including baseline correction, peak finding and peak area calculation. Peak areas were calculated using the total ion current (TIC) for the two analytical standards as well as the 6 oxygenates of interest. Standard addition method (SAM) plots of 2-hexanone and 2-ethyl-1-hexanol were utilized to determine the approximate concentrations of the unknown oxygenates in the sample solvent solutions as seen in Table 4-2.

The GC × GC–TOFMS data were also imported from the instrument software into Matlab R2015b (Mathworks Inc., Natick, MA, USA) using an in-house data converting algorithm. The data were analyzed with the in-house developed tile-based F-ratio software, an algorithm that elucidates chemical features that distinguish between two classes of samples, such as two samples coming from different steps in an industrial process. We direct the reader to previous publications and their respective supplementary information for a thorough understanding of the tile-based F-ratio software [10–13]. For all executions of the method a ^{1}D tile size of 6 seconds, a ^{2}D tile size of 100 milliseconds, a ^{1}D cluster size of 4 seconds, a ^{2}D cluster size of 60 milliseconds were employed; all chromatograms were normalized to the sum of the total ion current (TIC); and no retention time alignment was performed. The signal-to-noise (*S/N*) threshold and number of mass channels (*m/z*) utilized for the calculation of the average F-ratio were both set to 10, as indicated to be the optimum parameters in Chapter 3. At least 3 *m/z* were required for a measureable F-ratio and no F-ratio threshold minimum was employed.

Finally, PLS Toolbox (Version 8.1; Eigenvector Research, Inc., Manson, WA, USA) was used to perform parallel factor analysis (PARAFAC) on the six oxygenates listed in Table 4-2.

PARAFAC is a widely-used, quantitative multivariate chemometric technique that can be applied to trilinear or three-way data such as those generated by GC × GC – TOFMS. PARAFAC can decompose GC × GC – TOFMS data into the pure component profile on column 1, the pure component profile on column 2 and the pure component mass spectrum, allowing for more accurate and precise quantification and identification of analytes that may be convoluted due to noise or overlapping interferents. PARAFAC was performed on equivalent tiles encompassing the six oxygenates used for the F-ratio analysis, using *m/z* 35-150 of the E1-I sample, which had the most signal of all six oxygenates of interest. A two factor model was employed with a unimodality constraint on the column 1 dimension and nonnegativity constraints on column 2 and the mass spectral dimensions in order to obtain cleaner mass spectra. These mass spectra were matched to the National Institute of Standards and Technology (NIST) library using NIST MS Search 2.0 (NIST, Gaithersburg, MD, USA).

4.3 Results and Discussion

A flowchart of the industrial polymerization process from which samples were taken for this investigation is outlined in Figure 4.1. The feedstock tank consists of new feed, mostly isomers of hexane. The feed tank includes both new feedstock and waste from the centrifuge. Sampling Point I occurs after the feed tank and prior to the purification step. The purification step includes both a distillation tower and a molecular sieve column. Sampling Point II occurs after the purification steps. Following purification, the polymerization process occurs, which includes a

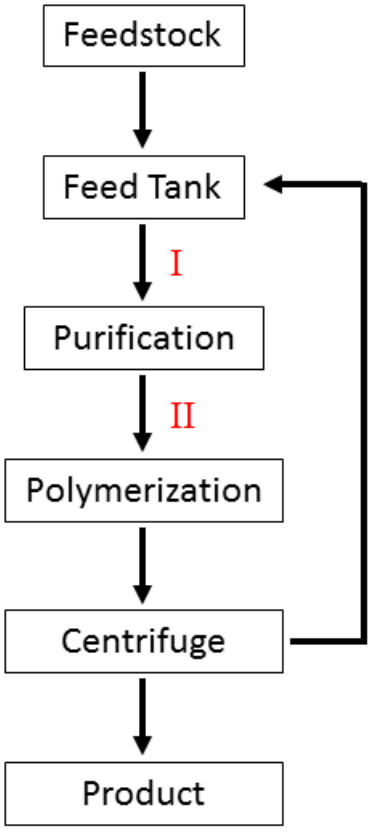

Figure 4.1. Flowchart of industrial process

proprietary process aid we will refer to as "Substance A." After polymerization, a centrifuge separates the product and the waste, the latter of which is returned to the feed tank for reuse.

Figure 4.2 shows the selective ion chromatogram (SIC) at *m/z* 57 of E1-I; that is the sample taken from Point I for campaign E1. The vast majority of the signal in the chromatogram comes from the solvent peaks, consisting mainly of the isomers of hexane, eluting prior to 4 minutes on column 1. This can be seen clearly in Figure 4.2(A), with the inset showing the overloaded and convoluted nature of the sample consisting mostly of solvent. Figure 4.2(B) shows the same chromatogram, excluding the solvent peaks. Here, numerous other components can be seen at lower concentrations that were otherwise drowned out by the signal due to the solvent. Eluting early in the chromatogram on column 1 are the branched alkanes, eluting early on column 2; and the oxygenates, eluting between 1.0 and 1.5 seconds on column 2. The process aid, "Substance A" can also be seen eluting at about 25 minutes on column 1 and at about 1.6 seconds on column 2. A zoom-in of the oxygenate region shown in the red box can be seen in Figure 4.2(C), with at least six obvious oxygenate peaks designated by their numbers in red. It is obvious that these analytes are at a rather low signal-to-noise, especially relative to the solvent peaks, and are likely contaminants of the polymerization process due to the centrifuge waste being recycled (Figure 4.1).

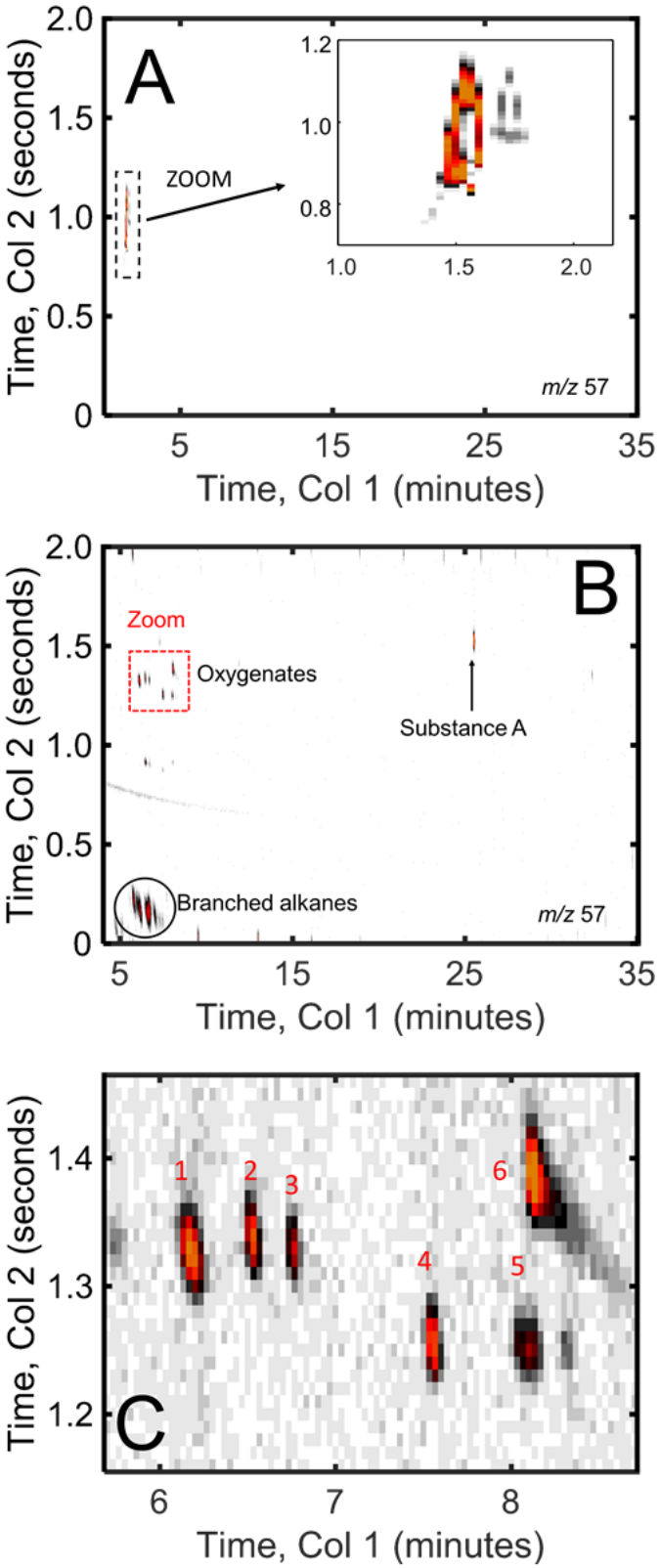

Figure 4.2. (A) Chromatogram of E1-I; (B) E1-I excluding solvent peaks; and (C) Zoom in of oxygenate region of E1-I from (B).

Figure 4.3 shows the F-ratio results of the 8 vs. 8 comparison of the injection replicates from the excellent campaigns sampled at Point I vs. those sampled at Point II. Each hit in the F-ratio hit list is plotted in the chromatographic space with the size of the blue circle scaling with the magnitude of the F-ratio value: the larger the circle, the greater the F-ratio, and therefore the greater the class-to-class (Point I vs. II) difference. As in Figure 4.2(B), the oxygenates and the branched alkanes are clearly visible as F-ratio hits in their corresponding chromatographic region. These analytes are present in the Point I samples and absent in the Point II samples, indicating these classes of compounds are removed via the purification process.

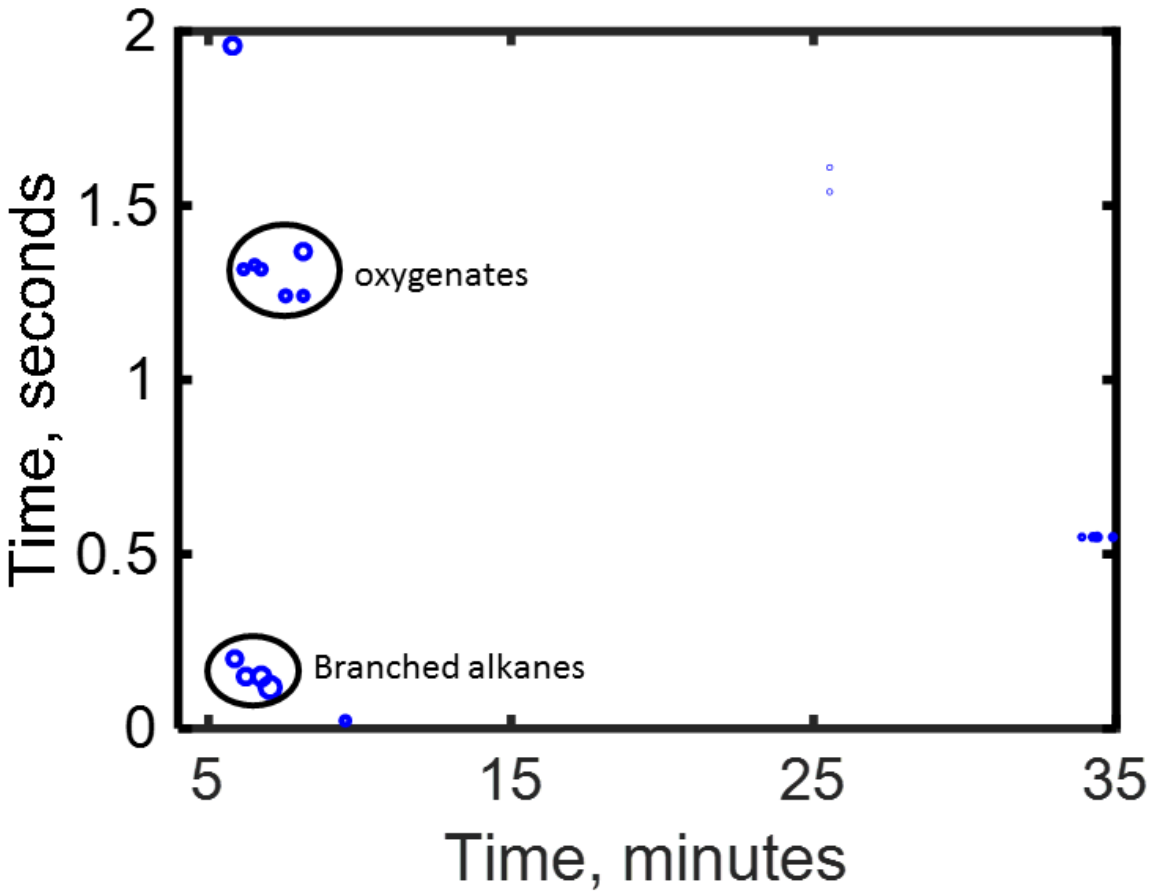

Figure 4.3. Scatter of F-ratio hits at their ^{1}D and ^{2}D retention times, the size of the circles scales with the magnitude of the F-ratio.

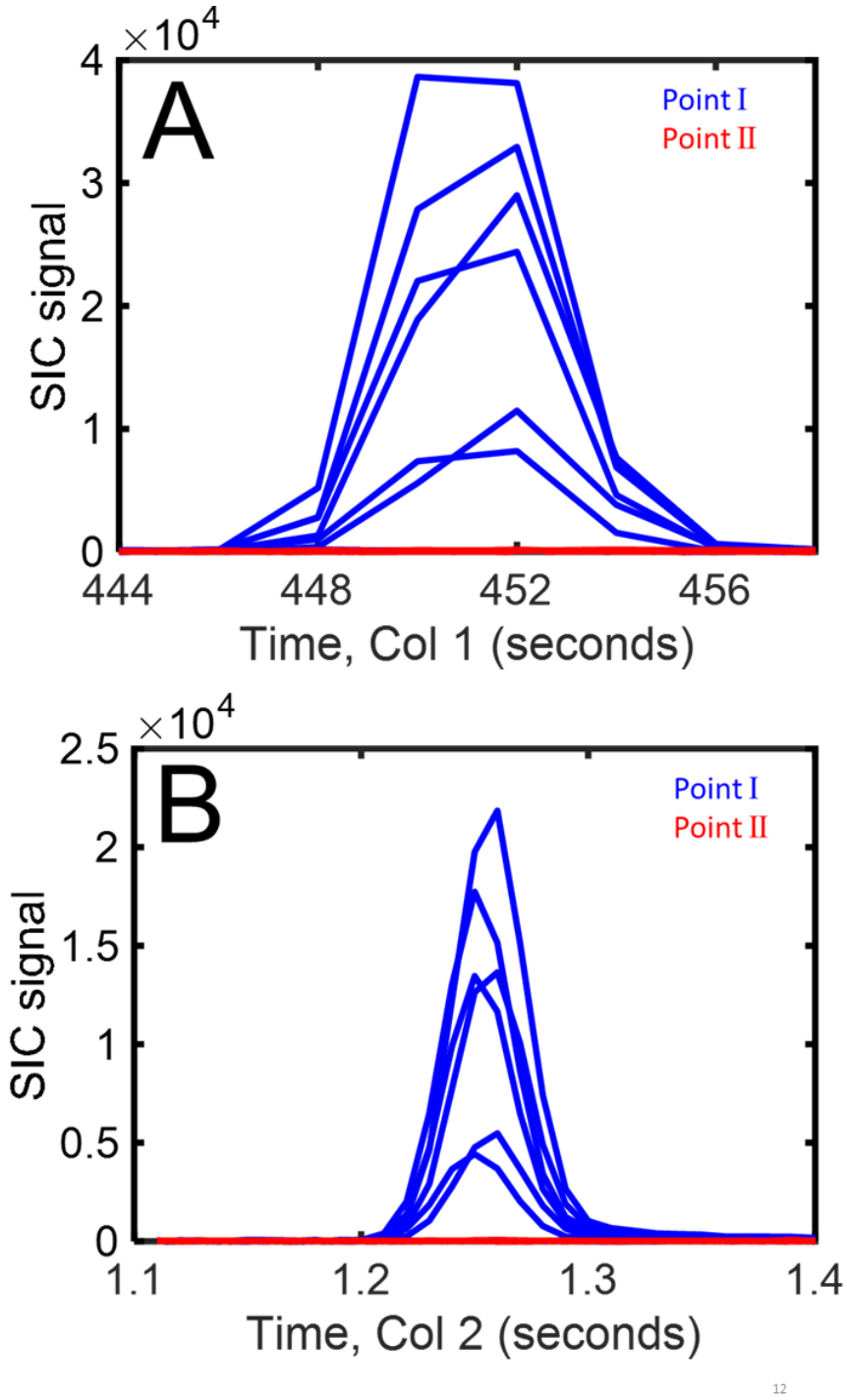

Figure 4.4. Ruler plots showing the elution profile of oxygenate #4 (A) on column 1 and (B) on column 2 in the Excellent campaigns at Point I (blue) and Point II (red) used for the F-ratio analysis.

Figure 4.4 (A) and (B) are the ruler plots of oxygenate 4 from Figure 4.2(C) of all sixteen samples used for the 8 vs. 8 F-ratio comparison with results shown in Figure 4.3. Figure 4.4(A) is the chromatographic profile of oxygenate 4 on column 1, while Figure 4.4(B) is that on column 2. These figures show the obvious signal presence of the oxygenate at sampling Point I in at least 6 of the 8 samples (that is, 3 of the 4 campaigns with duplicate injections) and complete absence of the analyte at sampling Point II. The variance between the classes shown in Figure 4.4**Error! R eference source not found.** is what generates a large F-ratio hit (Hit #7; F-ratio = 11.6) as seen in Figure 4.3.

With the knowledge gained from the tile-based F-ratio method, mainly that the oxygenates and branched alkanes were successfully removed via the purification process during the Excellent polymerization campaigns, the next objective was to discover any possible compounds present in the Bad samples that could be contributing to or correlating with the reduction in polymerization efficiency. Figure 4.5(A) shows a chromatogram (*m/z* 43) of an "Excellent" campaign sampled at Point II, and Figure 4.5(B) shows one of a "Bad" campaign at Point II. While Figure 4.5(A) shows mostly noise with only an obvious peak where the process aid, Substance A, elutes, Figure 4.5(B) very obviously contains the oxygenates that were removed via the purification process in the "Excellent" campaigns as indicated by the F-ratio results. This indicates that during poor polymerization processes the oxygenates were not successfully removed via the purification process as with the excellent processes, and the presence of these analytes correlate with the presence of the catalyst poison. This is confirmed in Fig 6 (A) and (B), which show the elution profiles of oxygenate 4 on column 1 and column 2, respectively, for the "Excellent" (red) and "Bad" (blue) processes at Point II.

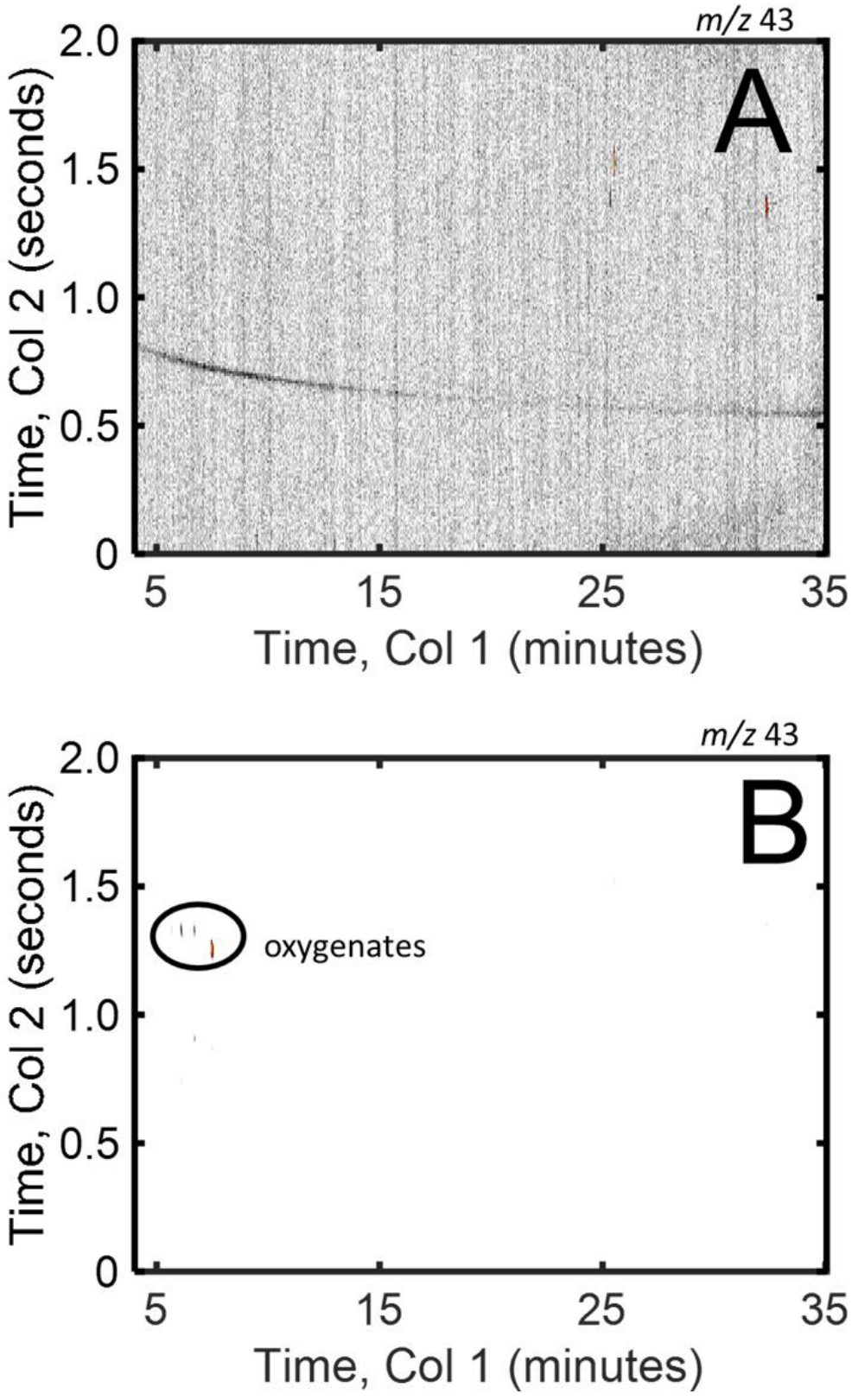

Figure 4.5. (A) Chromatogram of E1-II and (B) Chromatogram of B1-II

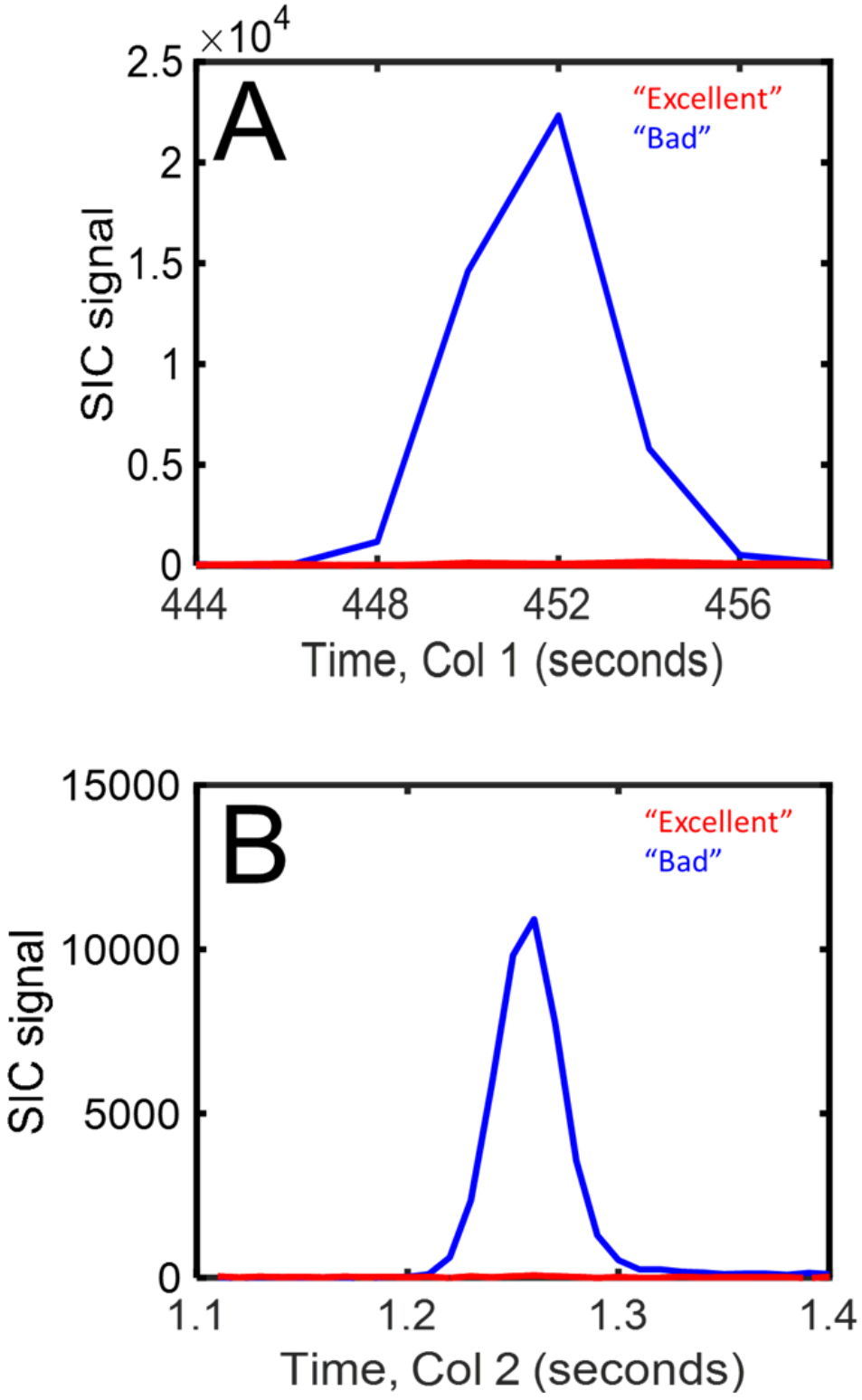

Figure 4.6. Ruler Ruler plots showing the elution profile of oxygenate #4 (A) on column 1 and (B) on column 2 in the E1-II and B1-II chromatograms.

The standards, 2-hexanone and 2-ethyl-1-hexanol were analyzed, as discussed in the Section 2.3, and a traditional standard addition method (SAM) plot was made. The SAM plot of 2-hexanone (blue), 2-ethyl-1-hexanol (red), and the average of the two (black) is shown in Figure 4.7. These analytes were chosen due to their mass spectral similarity to the unknown oxygenates of interest shown in Figure 4.2(C). The linear fit equation of the average line (black; y=11689x) was used to calculate the relative concentrations of the unknown oxygenates in the "Excellent" campaigns at Points I and II and in the "Bad" campaign at Point II. The average result, including both the TIC peak areas and calculated concentrations, for the oxygenates are shown in Table 2-2. As expected, the E-II campaigns have no oxygenates present, while the B-II campaign do. It is worth noting the complete absence of oxygenate 6 from the B-II campaign. Whether this is due to its absence from the Feed Tank or its successful removal during the purification process is unknown due to the lack of an equivalent B-I campaign sample.

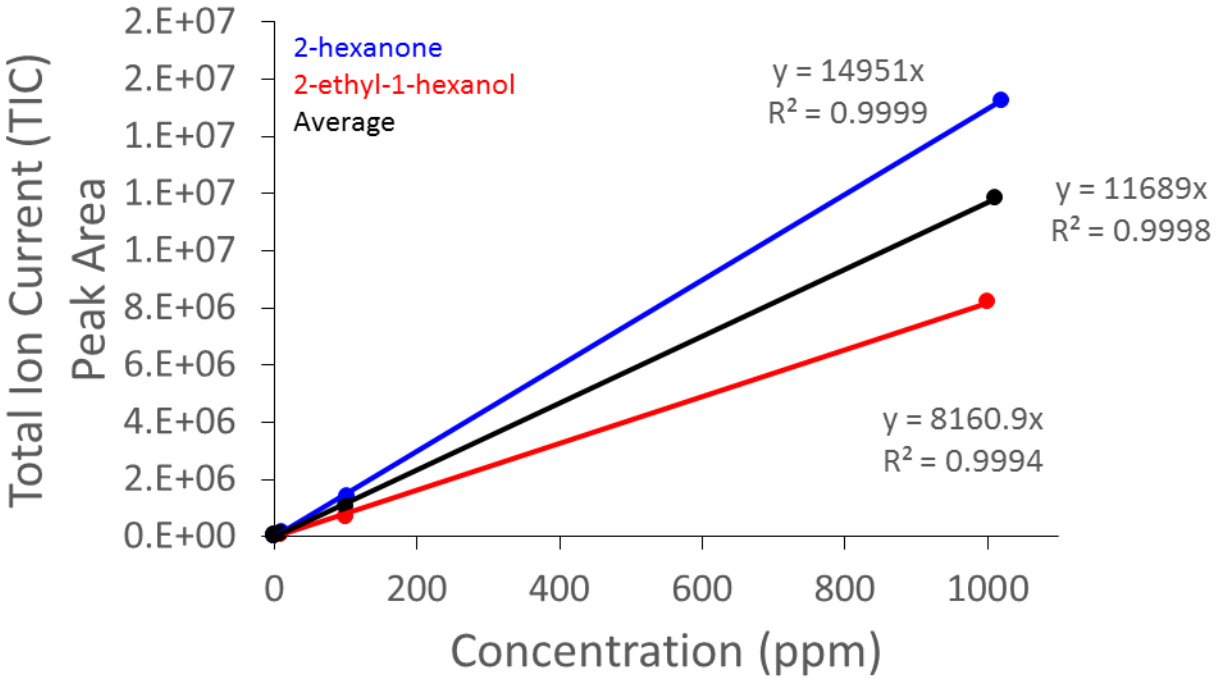

Figure 4.7. Standard addition method (SAM) plot of the standards 2-hexanone (blue), 2-ethyl-1-hexanol (red), and the average of the two (black).

Table 4-2. Table of peak areas and concentrations of oxygenates.

Oxygenate #	E-I		E-II		B-II	
	Peak Area	Concentration (ppm)	Peak Area	Concentration (ppm)	Peak Area	Concentration (ppm)
1	52761	4.5	0	0	82948	7.1
2	27189	2.3	0	0	45408	3.9
3	35433	3.0	0	0	52420	4.5
4	208665	17.9	0	0	271806	23.3
5	32968	2.8	0	0	43252	3.7
6	29904	2.6	0	0	0	0.0

Finally, while LECO's ChromaTOF software was able to effectively quantify the six oxygenates, with results shown in Table 2, the spectra were noisy enough that poor library matches resulted, making identification of the oxygenates difficult. In an attempt to obtain cleaner mass spectra, PARAFAC was performed as discussed in Section 2.3. Figure 4.8(A) shows the raw mass spectrum of oxygenate 4 as detected by the instrument, and Figure 4.8(B) shows the mass spectrum resulting from the PARAFAC model with the noise obviously removed. This pure mass spectrum in Figure 4.8(B) was matched to the NIST database as 3,5-hexadien-2-ol with a match value of 778 and a reverse match of 780. The NIST MS Search 2.0 identifications, match values and reverse match values of the PARAFAC modeled spectra of the six oxygenates are shown in Table 4-2.

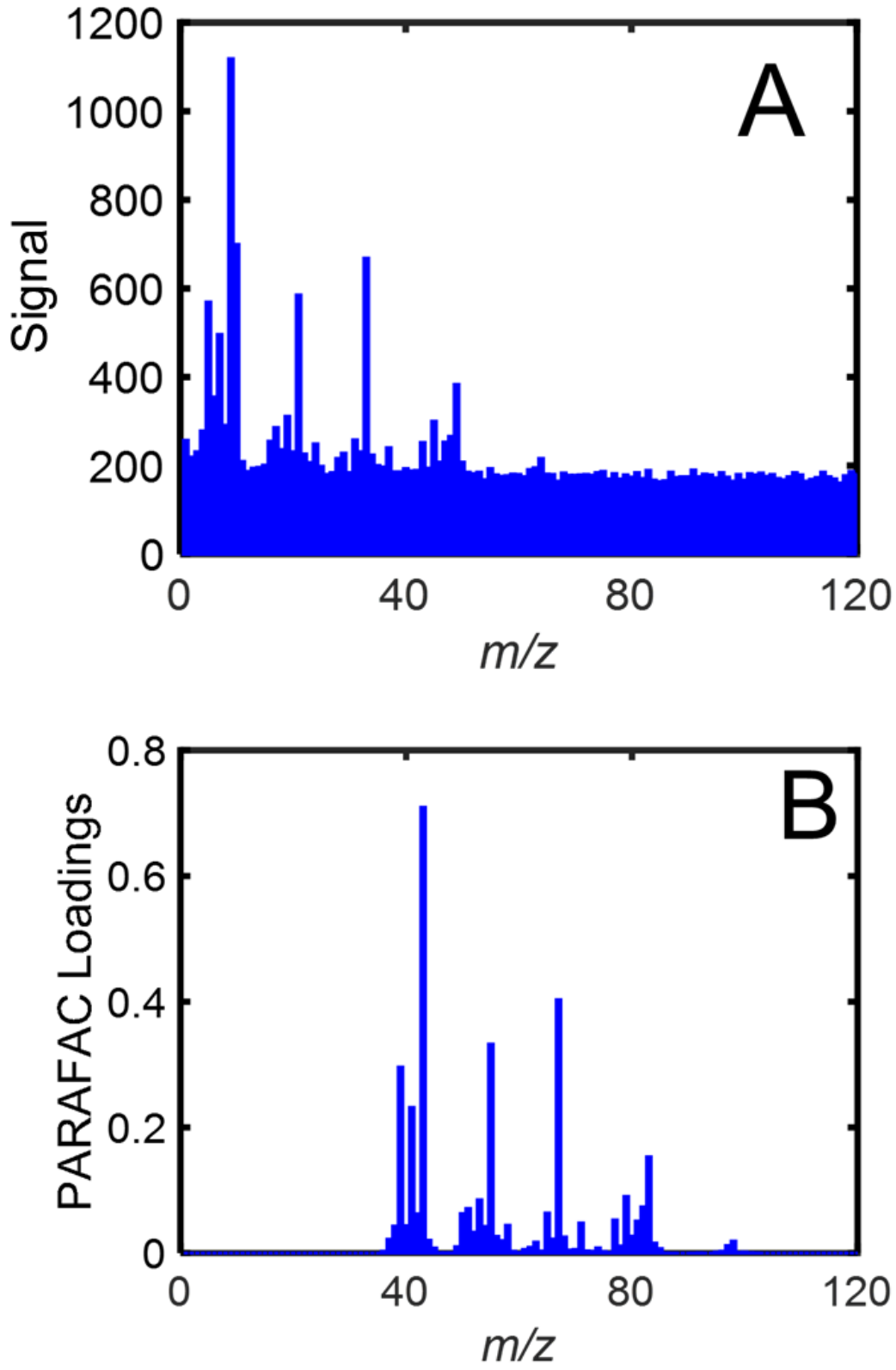

Figure 4.8. (A) Raw signal and (B) PARAFAC loadings of the *m/z* of oxygenate 4.

4.4 CONCLUSION

Solvent samples from an industrial polymerization plant were analyzed in order to determine the cause of a catalyst yield reduction. First, the optimized tile-based F-ratio method, a non-targeted software that analyzes variance between sample classes, was used to elucidate differences between the two sampling points occurring before (Point I) and after (Point II) the purification step in the polymerization plant during a "Excellent" polymerization process campaigns. This supervised, non-targeted software elucidated compounds that were removed via the purification process, most notably several branched alkanes and six oxygenate compounds. Using the information gleaned from the tile-based F-ratio method, a single sample from the post-purification sampling point (Point II) during a "Bad" polymerization process was analyzed. This sample clearly contained oxygenates that were otherwise removed via purification during the previously analyzed "Excellent" campaigns. These oxygenates were quantified and identified using the pure mass spectra generated from PARAFAC. The results of this investigation revealed that the presence of oxygenate compounds entering the polymerization reactor correlated with the catalyst yield reduction. This illustrates the importance of online process monitoring during industrial processes, which can provide answers regarding any changes in the industrial process in real-time to analysts on-site. For the process analyzed in this investigation, monitoring may be most illustrative at points just before and after the purification steps, in order to avoid further instances of catalyst yield reduction.

4.5 REFERENCES

[1] K.R. Beebe, W.W. Blaser, R.A. Bredeweg, J.P. Chauvel, R.S. Harner, M. LaPack, A. Leugers, D.P. Martin, L.G. Wright, E.D. Yalvac, Process analytical chemistry, Anal. Chem. 65 (1993) 199R–216R. doi:10.1021/ac00060a012.

[2] Workman Jerome, D.J. Veltkamp, S. Doherty, B.B. Anderson, K.E. Creasy, M. Koch, J.F. Tatera, A.L. Robinson, L. Bond, L.W. Burgess, G.N. Bokerman, A.H. Ullman, G.P. Darsey, F. Mozayeni, J.A. Bamberger, M.S. Greenwood, Process Analytical Chemistry, Anal. Chem. 71 (1999) 121–180. doi:10.1021/a1990007s.

[3] United States Food and Drug Administration, Guidance for Indsutry: PAT--A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance, (2004). http://www.fda.gov/downloads/drugs/guidances/ucm070305.pdf (accessed January 4, 2017).

[4] R. Shellie, P. Marriott, P. Morrison, Concepts and Preliminary Observations on the Triple-Dimensional Analysis of Complex Volatile Samples by Using GC×GC−TOFMS, Anal. Chem. 73 (2001) 1336–1344. doi:10.1021/ac000987n.

[5] C. Cordero, J. Kiefl, P. Schieberle, S.E. Reichenbach, C. Bicchi, Comprehensive two-dimensional gas chromatography and food sensory properties: potential and challenges, Anal. Bioanal. Chem. 407 (2014) 169–191. doi:10.1007/s00216-014-8248-z.

[6] A. Sampat, M. Lopatka, M. Sjerps, G. Vivo-Truyols, P. Schoenmakers, A. van Asten, Forensic potential of comprehensive two-dimensional gas chromatography, TrAC Trends Anal. Chem. 80 (2016) 345–363. doi:10.1016/j.trac.2015.10.011.

[7] C. von Mühlen, C.A. Zini, E.B. Caramão, P.J. Marriott, Applications of comprehensive two-dimensional gas chromatography to the characterization of petrochemical and related samples, J. Chromatogr. A. 1105 (2006) 39–50. doi:10.1016/j.chroma.2005.09.036.

[8] S. Risticevic, E.A. Souza-Silva, J.R. DeEll, J. Cochran, J. Pawliszyn, Capturing Plant Metabolome with Direct-Immersion in Vivo Solid Phase Microextraction of Plant Tissues, Anal. Chem. 88 (2016) 1266–1274. doi:10.1021/acs.analchem.5b03684.

[9] J.V. Seeley, S.K. Seeley, Multidimensional Gas Chromatography: Fundamental Advances and New Applications, Anal. Chem. 85 (2013) 557–578. doi:10.1021/ac303195u.

[10] 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. doi:10.1016/j.talanta.2013.06.038.

[11] 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. doi:10.1021/ac504472s.

[12] 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. doi:10.1016/j.chroma.2016.02.067.

[13] 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. doi:10.1016/j.chroma.2016.06.067.

Chapter 5. Determining the Probability of Chemometric Success for Gas Chromatography-Mass Spectrometry Based on Saturation Factor and the Chemometric Enhanced Peak Capacity

5.1 INTRODUCTION

Gas chromatography (GC) is a powerful separation technique used to separate volatile and semi-volatile compounds in complex mixtures. Often coupled with mass spectrometry (MS) as a hyphenated technique, GC-MS is a popular analytical platform for a variety of applications, such as metabolomics,[1,2] forensics,[3,4] food products,[5,6] and more.[7–9] Complex samples generate complex chromatograms, often with excessive peak overlap due to coeluting analytes. Such chromatograms often require advance chemometrics techniques for successful deconvolution and extraction of useful information about the sample components.

Deconvolution algorithms are used to mathematically resolve two or more analytes that are chromatographically overlapped. Compared to simple peak integration techniques, deconvolution can provide more precise and accurate peak attributes, including the pure analyte peak area and pure analyte mass spectrum, for better analyte quantification and identification. Many chemometric algorithms and software algorithms have been developed to perform deconvolution on chromatographic peaks. Each algorithm functions slightly differently in its requirements of the data structure and its efficacy for certain applications. However, all deconvolution algorithms are accuracy-and precision-limited in their performance by three main factors: (1) chromatographic resolution between the target analyte and nearby interferent(s); (2) signal-to-noise (*S/N*) of the analytes of interest; and (3) mass spectral similarity between the target and nearby interferent(s).

Many metrics exist to evaluate and compare the efficacy and power of various GC separations. One popular metric is peak capacity, which is representative of the amount of

information that can be contained in a given separation, defined as the maximum number of peaks that can be evenly resolved, at $R_s = 1.0$, in a certain separation time.[10] However, if one can decrease the resolution at which peaks can be resolved through chemometric means, such as the application of a deconvolution algorithm, the requirement for physical separation can be loosened and the peak capacity can be chemometrically enhanced and increased, allowing more peaks per separation window.

The relative "crowdedness" of a chromatogram, that is, how many analytes are present in a separation window, is referred to as the saturation factor, α.[11] The saturation factor is mathematically defined as the number of components present in a separation divided by the peak capacity.[11] Therefore, if the peak capacity can be chemometrically enhanced, it will nominally enhance the saturation factor. Consequently, the metrics chromatographers use to compare and evaluate separations, especially resolution, peak capacity and saturation factor, are interrelated by definition and can be chemometrically enhanced through the application of an effective deconvolution algorithm.

The relation between peak overlap and saturation factor has previously been reported by Davis and Giddings as part of their statistical overlap theory. [11,12] Their theory expressly states that with complex samples, the overlap of peaks can be statistically estimated, specifically that the number of observed peaks can be statistically approximated. However, their theory does not include a discussion of chemometrically enhanced resolution and was suggested to fail at saturation factors greater than one. We apply a similar theory here, which is nominally consistent with that presented by Davis and Giddings, but aims to describe the effects of chemometrically enhanced resolution on peak capacity and the probability of successful chemometric deconvolution.

We present an extensive simulation-based investigation into the minimum chromatographic resolution at which a chemometric algorithm can successfully deconvolute coeluting GC-MS peaks, hereafter referred to as the limit of chemometric resolution, R_s*. Operating under the assumption that analyte peaks are randomly distributed across the separation space, we first present a probabilistic description of peak overlap in GC-MS separations to determine the probability of chemometric success. In particular, chemometric success for a target analyte requires that any interfering peaks have $R_s \geq R_s$* to the target peak. Secondly, we present the results of a simulation based study to investigate how the practical application of a deconvolution algorithm fits with the expected theory. Simulations include the generation of chromatograms including one target analyte and one interferent analyte (referred to as a "target-interferent pair) at various resolutions and two signal-to-noise levels. The target-interferent pairs are varied to include pairs that with similar mass spectral character and with different mass spectral character. Any deconvolution algorithm could be applied to the simulated data to examine its limit of resolution. For the purpose of this study, multivariate curve resolution with alternating least squares (MCR-ALS) was chosen as the deconvolution algorithm. However, other deconvolution algorithms, such as generalized rank annihilation method (GRAM), classical least squares (CLS) or stacking replicates using parallel factor analysis (PARAFAC), could also be applied to determine their respective chromatographic limits of deconvolution. The results of more than 31,000 simulations and their deconvolution algorithm model results are presented, indicating that MCR-ALS has a R_s* of approximately 0.13 or better corresponding to an 87% probability of successful deconvolution of a target analyte at a saturation factor of 0.5.

5.2 THEORY

The equations presented here operate under the assumption that chromatographic peaks are randomly distributed across the separation space. While this is not necessarily the case for all separations, depending on sample composition and column stationary phase, the theory presented can also be applied as local peak capacities by replacing t_{sep} with a value t_{window} such that $t_{window} \leq t_{sep}$, assuming only local randomness and allowing for variations in peak distribution across the chromatogram. The assumption that peaks are randomly distributed across the entire separation is used for the purposes of demonstration, but all equations presented are readily applied with the assumption of local randomness.

In chromatography, the amount of information contained in a given separation is inferred from and is directly related to the ideal peak capacity of the separation, defined as

$$n_c = \frac{t_{sep}}{W_b R_s} = \frac{t_{sep}}{W_b} \tag{5.1}$$

where t_{sep} is the total separation time and W_b is taken as the average peak width at base. Herein, the peak capacity, n_c, represents the number of evenly resolved peaks at unit resolution, $R_s = 1$, that can fit into a given separation. The resolution between two chromatographic peaks is defined by

$$R_s = \frac{t_{R,2} - t_{R,1}}{\frac{1}{2}(W_{b,1} + W_{b,2})} = \frac{\Delta t_R}{W_b} \tag{5.2}$$

where Δt_R is the difference in retention time for peaks one and two, $t_{R,1}$ and $t_{R,2}$, and W_b is the average of the individual peak widths at base for peaks one and two, $W_{b,1}$ and $W_{b,2}$. For temperature

programmed separations peak widths tend vary only slightly, such that the average width at base terms, W_b, from Equations (1) and (2) are nominally equivalent.

The application of chemometrics to chromatographic data allows for mathematical resolution of peaks that may not have been resolved physically by chromatography. The ability to mathematically resolve peaks that are physically overlapped, $R_s < 1$, significantly relaxes the requirement for chromatographic resolution across the separation and increases the number of evenly resolvable peak, resulting in a new chemometric enhanced peak capacity,

$$n_c{}^* = \frac{t_{sep}}{W_b R_s{}^*} = \frac{n_c}{R_s{}^*} \tag{5.3}$$

where R_s* is defined to be the minimum chemometric resolution, which is the minimum resolution between two peaks for a particular chemometric method to deconvolute them, that is, to resolve them mathematically. It should be noted that the minimum chemometric resolution, R_s*, is distinct from chromatographic resolution defined by Equation (5.2); in particular, R_s* depends on the chemometric method and must be determined experimentally, while R_s is a measured chromatographic value between two peaks. In general, R_s* suggests that the chemometric method can be successfully applied for all values Rs ≥ R_s*. Figure 5.1(A) and (B) show simulated selective ion chromatograms of a single target analyte (solid line with t_R = 4.5) and the two nearest interferent peak profiles eluting with $R_s = R_s$* for R_s* = 1.0 and 0.3 respectively. The R_s* = 1.0 in Figure 5.1(A) is representative of analysis without chemometrics while Figure 5.1(B), shows a separation readily analyzed by a chemometric technique with R_s* = 0.3, illustrating the substantial gain in peak capacity when chemometrics are considered.

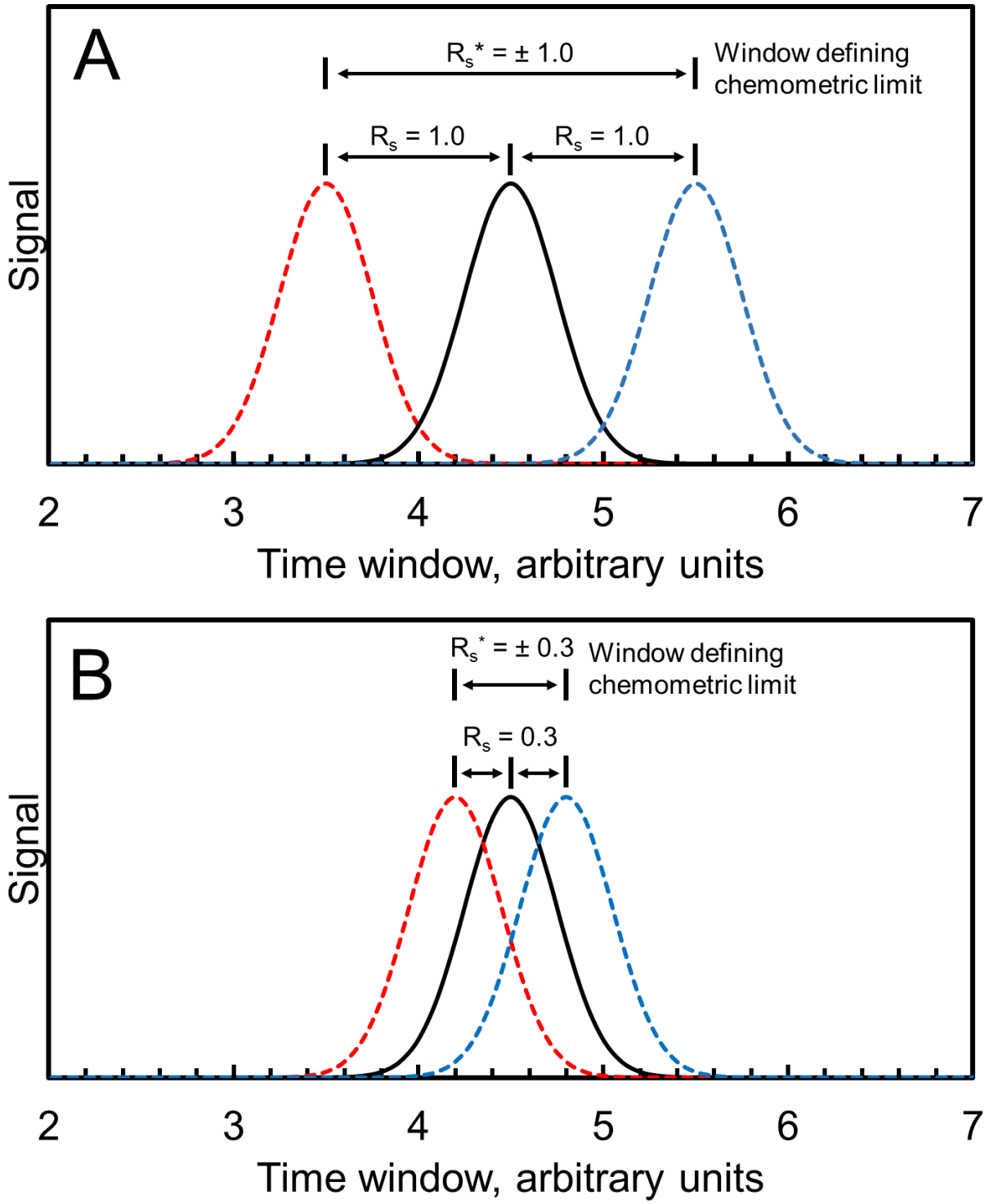

Figure 5.1. Simulated chromatograms of a target analyte (black solid line) with two neighboring interferents (red and blue dotted lines) at (A) $R_s = 1.0$; and (B) $R_s = 0.3$.

The assumption of randomly distributed peaks generally leads to the Poisson distribution, a generalized approximation of the binomial distribution. However, the assumption of randomly distributed peaks means that the probability of chemometric success can be determined rigorously with the binomial distribution with each chromatographic peak representing an independent Bernoulli random variable, since chemometric success requires that there are no interferent peaks at Rs < R_s* to the target analyte. For a single target target-interferent pair (m = 1) the probability of chemometric failure, Rs < R_s*, is defined by

$$p_{fail} = \frac{2}{n_c^*} \tag{5.4}$$

where n_c* is as in Equation (5.3). This leads to the determination that the probability of chemometric success for a single target target-interferent pair, Rs ≥ R_s*, is

$$p_{success} = 1 - p_{fail} = 1 - \frac{2}{n_c^*} \tag{5.5}$$

Equation (5.5) can be generalized to samples of any complexity to define the probability of chemometric success for a target analyte with $m \geq 0$ independent intereferent peaks

$$P(success) = (p_{success})^m = \left(1 - \frac{2}{n_c^*}\right)^m \qquad m \geq 0 \tag{5.6}$$

The saturation factor of a chromatographic separation is given by

$$\alpha^\circ = \frac{\#\,components}{n_c} = \frac{m+1}{n_c} \tag{5.7}$$

where m is the number of interferent peaks as in Equation (5.6) so the total number of components is m interferents + 1 target analyte. Figure 5.2 (A-D) are simulated total ion current (TIC)

chromatograms with n_c = 100 and various numbers of randomly distributed analytes, corresponding to α^o = 0.1, 0.5, 1.0, and 2.0 respectively. An alternative expression for *P*(success), as a function of α^o, can now be provided by combining Equation (5.3) and (5.7) with Equation (5.6)

$$P(success) = \left(1 - \frac{2R_s^*}{n_c}\right)^{(\alpha^\circ n_c - 1)} \quad (5.8)$$

Clearly, *P*(success) depends on the separation peak capacity and the sample complexity but it also requires some approximation of the limit of chemometric resolution, R_s*, for the chosen method. The most generalized benefit of Equation (5.8) is that allows for a quantitative comparison of chemometric method performance and chromatographic separation parameters selection for samples of varying complexity.

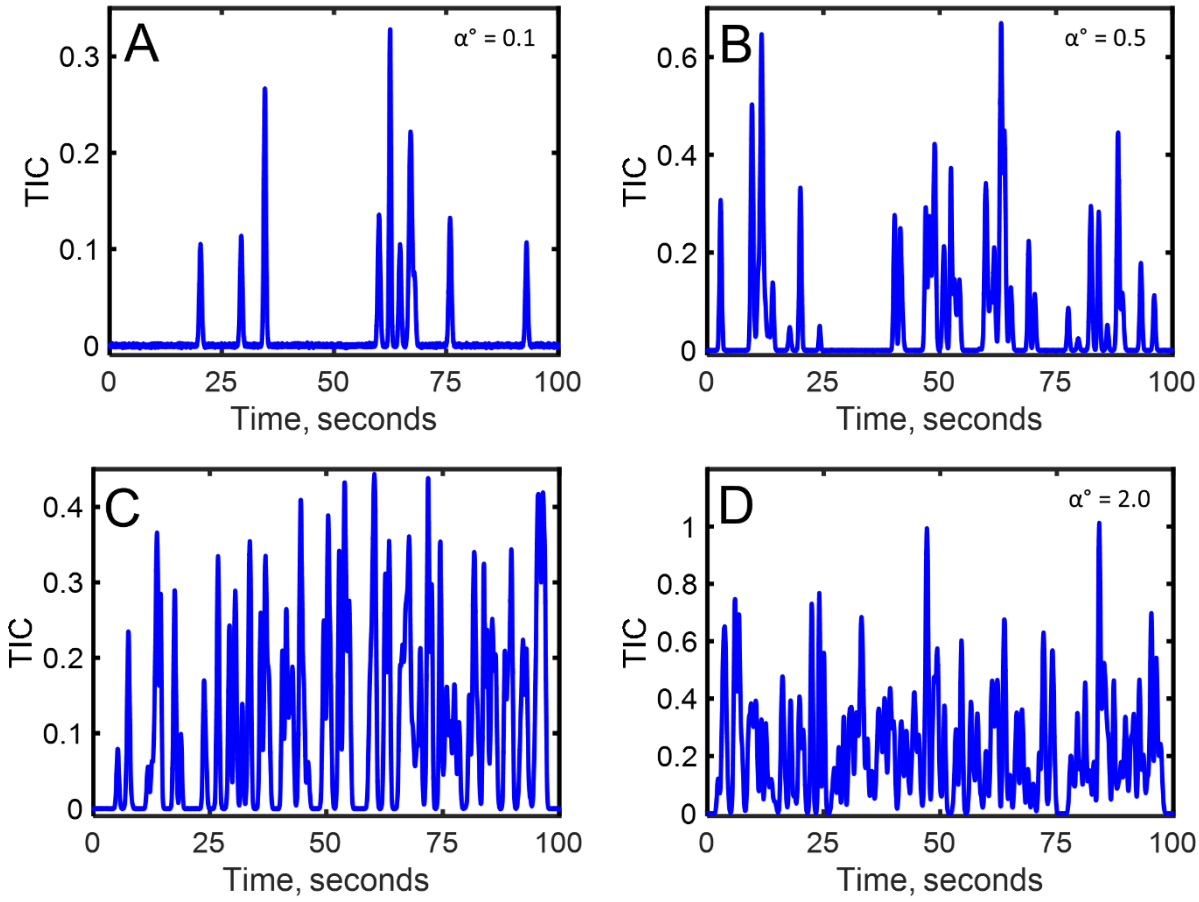

Figure 5.2. Simulated total ion current (TIC) chromatograms with n_c = 100 and various numbers of randomly distributed analytes, corresponding to (A) α^o = 0.1; (B) α^o = 0.5; (C) α^o = 1.0; and (D) α^o = 2.0.

5.3 EXPERIMENTAL

All data was simulated and analyzed using Matlab R2015b (Mathworks, Inc., Natwick, MA, USA), with the modeling conditions summarized in Table 5-1. Three-second long GC-TOFMS chromatograms were simulated consisting of one analyte and one interferent at a mass spectral collection rate of 100 Hz. Peaks were modeled as Gaussians with the same peak area and a width at base (± 2σ) of 1 second. Random Gaussian-distributed noise was generated independently for each *m/z*. The parameters for the noise were determined such that the mean and standard deviation of the noise would provide the desired signal-to-noise ratio (*S/N*) in the total ion current (TIC) of the peak. Two *S/N* values were studied, the high *S/N* was defined as *S/N* of 100 (hereafter, "*S/N* 100") and the low *S/N* was defined as *S/N* of 10 (hereafter, "*S/N* 10"), where *N* was defined as 3σ, or three times the standard deviation of the noise.

Table 5-1: Table of experimental simulation conditions

	Conditions studied
R_S	0.02; 0.04; 0,06; 0.08; 0.10; 0.12; 0.14; 0.16; 0.18; 0.2; 0.25; 0.30; 0.35; 0.40; 0.45; 0.50
S/N	10; 100
Number of analytes	45
Number of analyte-interferent combinations	990
Total number of MCR Models	31,680

In order to create a large enough sample of simulations for study, a total of 45 analytes were chosen to act as the targets and interferents with their mass spectra taken from the NIST database. Here, a "target" is taken to be an analyte of interest and "interferent" an analyte that is chromatographically overlapped with the desired target. A list of the 45 analytes can be found in Table 5-2. Since two of the 45 analytes were used for each simulation (one target, one interferent), a total of 990 simulations were created at each *S/N* ($_{45}C_2$, or 45 choose 2). These 45 analytes were chosen such that there was a good distribution of target-interferent pairs with very similar mass spectra and pairs with very dissimilar mass spectra. A match value (MV) was calculated between each target and interferent to determine whether there was a "High" MV (that is, the mass spectra were very similar) "Mid" MV, or "Low" MV (that is, the mass spectra are very dissimilar). MV was calculated based on the equation outlined by Stein[13]. "High" MV target-interferent pairs were those with $600 \leq MV < 1000$; "Mid" MV pairs were those with $300 \leq MV < 600$; and "Low" MV pairs were those with $MV < 300$. The calculated match values for all 990 target-interferent pairs are shown in the heat map in Figure 5.3 (A) and the histogram in Figure 5.3 (B).

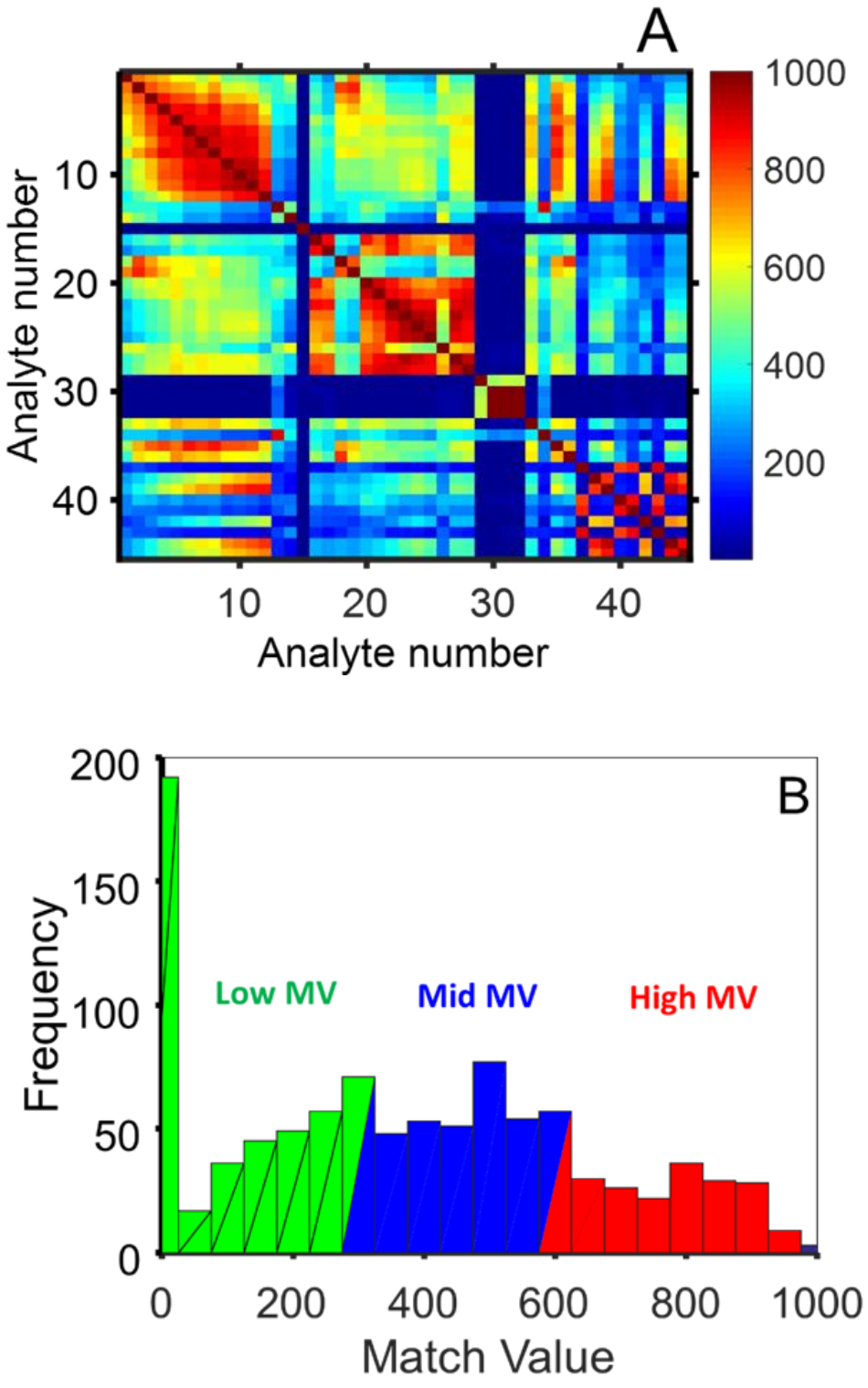

Figure 5.3. (A) Heat map; and (B) histogram of match values of target and interferent pairs.

Table 5-2: Table of compounds used as analytes and interferents for simulations.

Number	Compound
1	Hexane
2	Heptane
3	Octane
4	Nonane
5	Decane
6	Undecane
7	Dodecane
8	Tridecane
9	Tetradecane
10	Pentadecane
11	Hexadecane
12	Pristane
13	1-Chlorohexane
14	1-chlorobutane
15	carbon tetrachloride
16	cyclooctane
17	cis-1,2-dimethylcyclohexane
18	2,3,4-trimethylpentane
19	2-methylpentane
20	1-octanol
21	1-nonanol
22	1-decanol
23	1-dodecanol
24	1-tetradecanol
25	1-hexadecanol
26	1-heptene
27	dodecene
28	1-undecene
29	propylbenzene
30	p-xylene
31	o-xylene
32	m-xylene
33	1-bromo-2-ethylhexane
34	1-chloro-5-methylhexane
35	2,3,6,7-tetramethyloctane
36	3,4-diethylhexane
37	butanoic acid
38	docosane
39	heneicosane
40	heptanoic acid
41	hexanoic acid
42	pentacosane
43	pentanoic acid
44	tetracosane
45	tricosane

In addition to varying the target-interferent pair and the *S/N*, the resolution between the target and interferent peaks was varied in order to determine the minimum resolution for the deconvolution algorithm to successfully deconvolute the target from the interferent. Sixteen resolution values (R_s = 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, and 0.5) were chosen. For each target-interferent pair (990 options), a chromatogram of each resolution (16 options) at each *S/N* (2 options) was generated, resulting in a total of 31,680 total chromatograms generated. All parameters for the simulation study are outlined in Table 1.

Next, an MCR-ALS model was created for each simulation. One component, two component and three component models were generated, and it was determined that the two-component models effectively modeled both target and interferent while pulling out the noise in the residuals. All chemometric models shown and discussed hereafter will refer to the two-component models. Models were generated without any constraints. All chromatograms and models were generated over the course of 36 hours on a personal computer with an Intel® Core™ i7-4770 processor (3.4 GHz), 24 GB of random access memory (RAM), a 250 GB Samsung 840 solid state hard drive, and with Windows 7 SP1 as the operating system.

Two parameters were calculated to evaluate each chemometric deconvolution model. First, the peak area of the chromatographic component from the model was compared to the actual simulated peak area via a percent error calculation. The percent error on the peak area was defined as:

$$\%error = \frac{A_{Model} - A_{Sim}}{A_{Sim}} \times 100 \qquad (5.9)$$

Where the A_{Model} is the area of the peak resulting from the chemometric deconvolution model and A_{Sim} is the area of the peak generated during the simulation. Percent error was calculated for the

target in each of the 31,680 models. Secondly, the match value (MV, with possible values ranging from 0 to 1000) was calculated between the mass spectrum of the target generated from the loadings of the model and the original library mass spectrum for the target from the NIST library based on the equation outlined by Stein [13].

5.4 Results and Discussion

Figure 5.4 (A) is a plot of the probability of chemometric success, as a function of the saturation factor, α^o, Equation (5.8), for various chemometric resolutions ranging from 0.05 to 1.0. The previously reported mass cluster method[1,14,15] was found to be effective with an R_s* approximately equal to 0.05, if not slightly below, and R_s* = 1.0 is representative of analysis without chemometrics, relying solely on the physical chromatographic resolution of peaks. The dashed vertical lines are located at the four saturation factors shown in Figure 5.2, α^o = 0.1, 0.5, 1.0, and 2.0. The nonlinear relationships between *P*(success), α^o, and R_s* means that for a saturation α^o = 0.7, improving from R_s* = 1.0 to 0.5 results in a gain in *P*(success) from ~0.25 to ~0.5. Figure 5.4 (B) shows a zoomed in view of the plot in Figure 5.4 (A) to illustrate the significant gains realized when chemometrics with various R_s* values are applied. However, since R_s* depends on the particular chemometric method to be used and the conditions under which the method will be applied, R_s* must be determined experimentally.

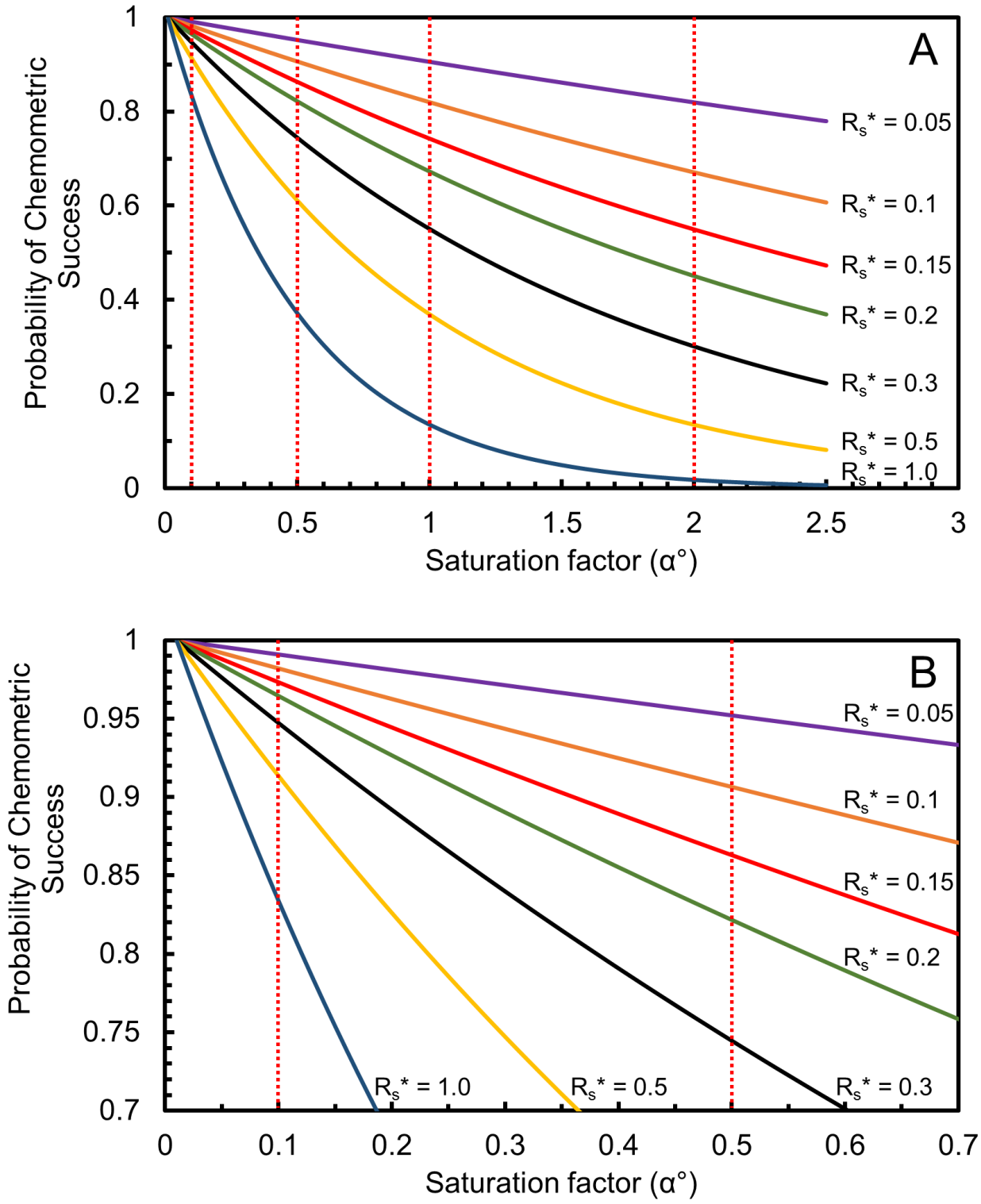

Figure 5.4. (A) Plot of probability of chemometric success as function of saturation factor, α^{o}; and (B) a zoomed in view of the plot in (A).

We now consider the simulation based study to determine an approximate R_s* for the chemometric deconvolution algorithm. An example of the results of two representative chemometric models at *S/N* 10 and R_s = 0.20 are shown in Figure 5.5. Figure 5.5(A) shows the component profiles of the target (1-chloro-5-methylhexane, blue) and interferent (decane, red). This target-interferent pair is an example of a "Low" MV pair, as the match value between 1-chloro-5-methylhexane and decane is 284, meaning they have dissimilar mass spectra. This, ideally, would make it easier for the model to distinguish between the two analytes. The percent error is low, at 13.22%, meaning the model attributed about 13.22% more peak area to the target than what was original modeled. The component mass spectra generated from the model are shown in Figure 5.5(B) with the mass spectra of the target, 1-chloro-5-methylhexane, in blue and the interferent, decane, in red. Their match values are shown as well, with matches to the library spectra well into the 900's, demonstrating that the model easily distinguished between the two analytes.

Figure 5.5 (C) and (D) show the same figures as in (A) and (B), but for a target-interferent pair from the "High" MV group. The target (2,3,4-trimethylpentane, blue) and the interferent (3,4-diethylhexane, red) have a match value to each other of 816, meaning their mass spectra are very similar. This should make a more difficult case for the algorithm to deconvolute, and in fact, the percent error is 47.36%, as shown in Figure 5.5 (C) with their component chromatographic profiles. Here, the target has significantly more peak area than what was originally modeled because the model falsely attributed some of the interferent signal to the target due to their mass spectral similarities. Figure 5.5 (D) shows the deconvoluted mass spectra of the two components generated by the model, with a MV of 928 between the deconvoluted 2,3,4-trimethylpentane and its library spectrum, while the MV of 3,4-diethylhexane to its library spectrum is only 851, likely

because of the model contributing some of its mass spectral signal to the 2,3,4-trimethypentane. These figures demonstrate how the "High" MV cases tend to have larger percent error, but still maintain high target-to-library MV due to the similarity in the mass spectra between the target and interferent.

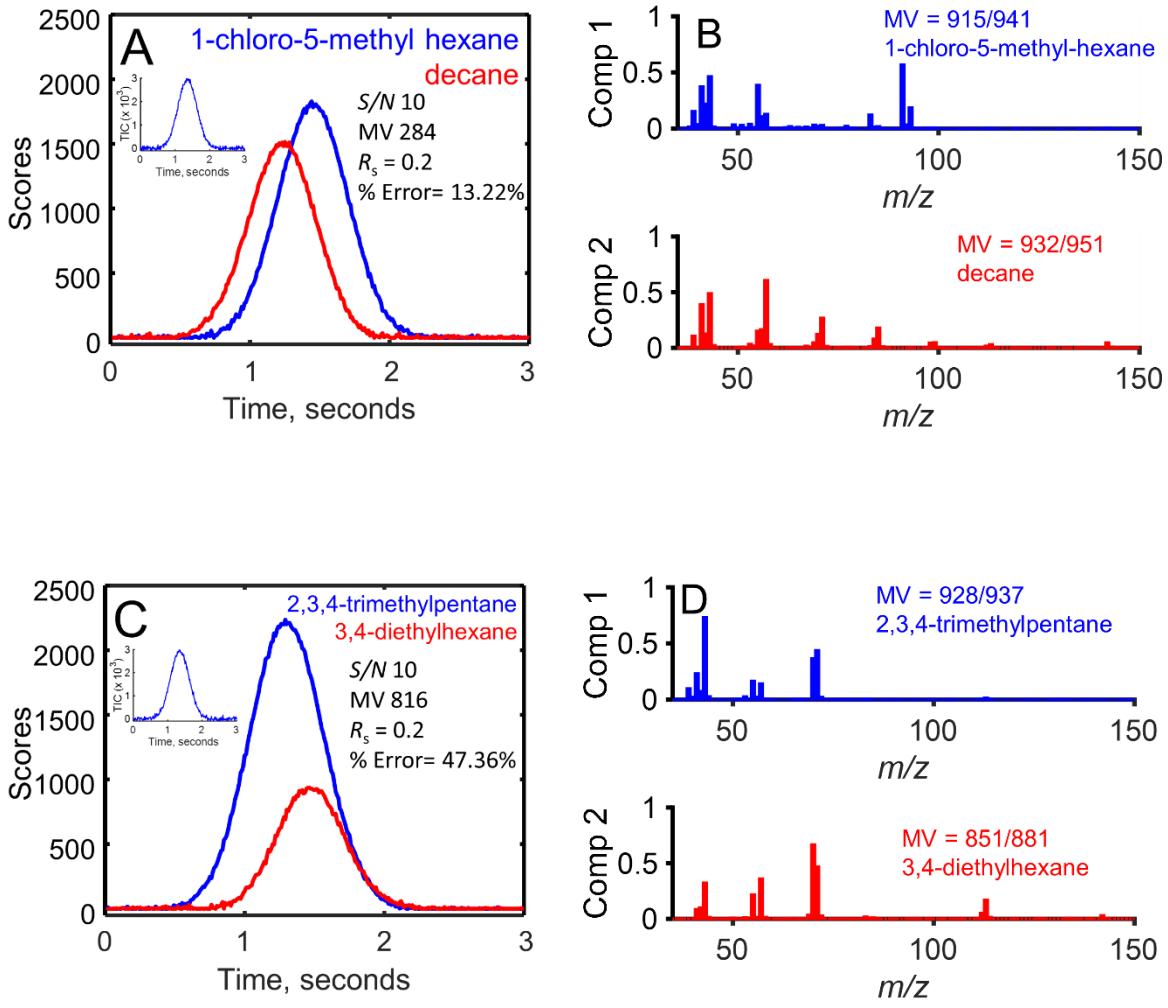

Figure 5.5. Representative chemometric models at *S/N* 10 and R_s = 0.20. (A) and (B) are results of a "Low" MV pair; (C) and (D) are results of a "High" MV pair.

Figure 5.6 summarizes the average MV and average percent error results for the 31,680 chemometric models. The "Low" MV results are shown in green, the "Mid" MV in blue and the "High" MV in red, with *S/N* 100 results as solid lines and *S/N* 10 results in dashed lines. Figure 5.6 (A) shows the average target-to-library MV results, where an increase in resolution results in an increase in the MV, as expected, due to a more facile and accurate extraction of the pure analyte mass spectra by the model. The *S/N* 100 results obviously have higher target-to-library MV at all resolutions simulated, starting at a MV of approximately 950 at the lowest resolutions and capping out at a maximum MV at 990 at the larger resolution values, while the *S/N* 10 simulations start at much lower MV (as low as 648 for the Low MV at R_s = 0.02) and converge on a maximum MV of about 915. This pattern is likely due to noisy *m/z* contributing spurious signals that are extracted with the true signal in the models. The extent of this inclusion of noisy mass channels increases as resolution decreases and the summed signal of these noisy mass channels increases due to excessive peak overlap between the target and interferent.

Contrary to what might be hypothesized is the order of the colored lines, with "Low" MV target-interferent pairs having chemometric models that result in lower MV to the library spectrum than the "High" MV. One might assume that target-interferent pairs that are very alike ("High" MV pairs) would be more difficult to correctly deconvolute and result in lower MV between the extracted and library spectrum. However, because the target and interferent share so many mass channels for those "High" MV pairs, any *m/z* signal attributed to the target that actually came from the interferent in the model will not appreciably affect the target-to-library MV because so many of the *m/z* are shared. On the other hand, those "Low" MV target-interferent pairs have many *m/z* that are not shared whatsoever, so any signal from an interferent mistakenly attributed to the target will automatically decrease the target-interferent MV.

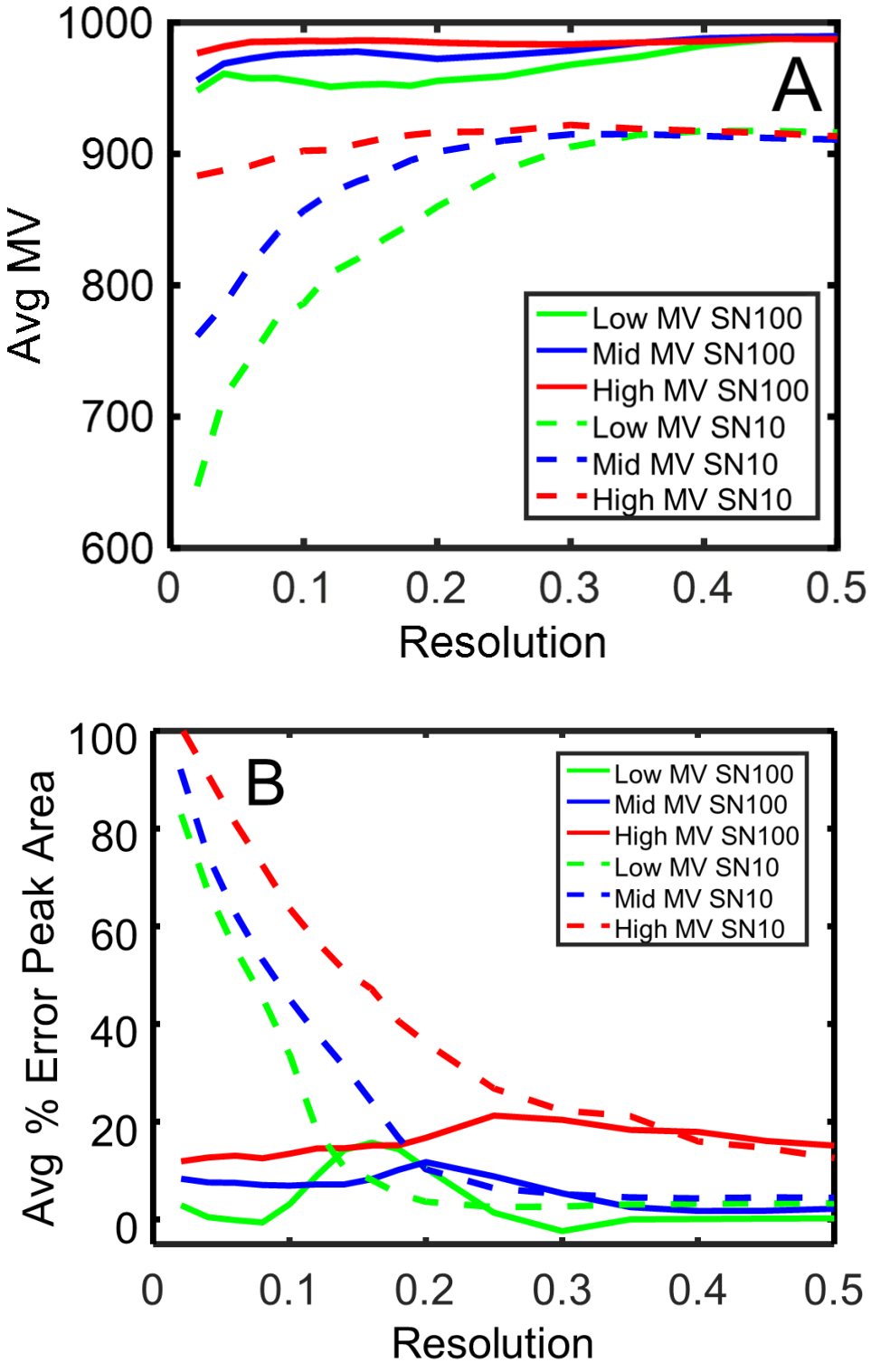

Figure 5.6. Summary of the results of the 31,680 chemometric models with (A) average MV vs. Resolution; and (B) average %Error vs. Resolution.

It is always important when looking at the results of a chemometric model that the mass spectral loadings are not the only model loadings regarded to determine the efficacy of the model, even though it might be the easiest to "match" to a library spectrum to determine the likely accuracy of a model used to deconvolute an unknown. As discussed in the previous paragraph, relying solely on the mass spectral loadings may prevent the analyst from discovering signal falsely attributed to the target of interest when the interferent has very similar mass spectral features. Figure 5.6 (B) aims to evaluate the chromatographic model loadings resulting from the chemometric models in conjunction with Figure 5(A). Looking at either Figure 5.6 (A) or (B) alone will not provide a full and accurate picture of the quality of model generated; but rather, both should be viewed simultaneously.

Figure 5.6 (B) contains the average peak area percent error results. The *S/N* 10 results show a sharp increase in the percent error between the resolution values of 0.3 and 0.2, especially for the "High" MV case, where mass spectral signals are easily misattributed to the target peak resulting in a large percent error. While the *S/N* 100 cases have much lower average percent error, there is an obvious increase in the percent error for the "Low" and "Mid" MV subsets after a resolution of 0.3. The cause of the slight peak in the "Low" MV, *S/N* 100 subset between $R_s = 0.3$ and $R_s = 0.1$ is unknown.

The smallest R_s with an average percent error less than 20% was deemed to be the limit of chemometric success. This avoided confusion due to strange artifacts in the data and provided the strictest and most objective definition. Because the large dataset was divided into subsets, this left us with a swath of R_s* values ranging from 0.02 to 0.4, depending on the *S/N* and MV designation. The average of all of these values was approximately R_s* = 0.13. Based on the results shown in Figure 5.4 (B), this would equate with *P*(success) = 0.87 for a target analyte at a saturation factor

of 0.5 (Figure 5.2 B). Figure 5.6(B) shows that for $R_s = 0.2$ only the "High" MV, *S/N* 10 curve has percent error greater than 20%. While the R_s* for MCR for this experiment is determined to be approximately 0.13, a more slightly more conservative R_s* = 0.2, with *P*(success) = 0.83 at α^o = 0.5, could be readily extended to other GC-MS experiments with only a minor decrease in the probability of chemometric success for all values α^o.

5.5 CONCLUSION

We report herein an investigation into the limit of chemometric resolution, R_s*, of a chemometric algorithm using simulations of analyte and interferent pairs with varying degrees of mass spectral similarity at various resolutions and signal-to-noise values. Theory outlining the probability of component overlap in a chromatographic separation is also. Although any deconvolution algorithm could have been used, the study here utilized MCR-ALS to deconvolute the target from the interferent. The efficacy of the chemometric deconvolution was evaluated by comparing the modeled peak area to the original simulated area via a percent error calculation and by comparing the deconvoluted mass spectrum to the original simulated mass spectrum via match value. The results of 31,680 chemometric models, their percent errors and match values were summarized and it was determined that the R_s* of MCR-ALS is about 0.13, corresponding to an 87% probability of deconvoluting a target analyte in a chromatogram with a saturation factor of 0.5.

Chapter 6. Conclusions and Future Directions

6.1 Chapter 2 Summary, Limitations and Future Directions

A new method for the analysis of stable isotope incorporation in the methylotrophic bacteria *M. extorquens* AM1 was presented. The new method workflow employed the 2D *m/z* cluster method to discover and extract the pure mass spectra of coeluting metabolites, discovering 152 *m/z* clusters where only 101 peaks were found using a traditional peak finder. Two principal component analysis (PCA) models were created using the pure mass spectra: the first model was to determine whether or not a metabolite had incorporated the stable isotope; the second model extracted the time course profile. Of the 152 metabolites surveyed, 83 were identified as changing with time and 69 unchanging with time.

The application of the 2D *m/z* cluster plot method to real, challenging samples was of great importance in this investigation. For this reason, the chromatographic method was shortened to ensure the presence of chromatographic overlap and to ensure all samples were analyzed in a timely manner. For future investigations, especially in which the biological ramifications are of utmost interest, it is suggested that the chromatographic method be optimized. Further investigations into *M. extorquens* AM1 may also include the analysis of various mutant bacterial strains and how they compare to the wild type. These analyses could also be performed using PCA.

While it was beyond the scope of the investigation presented herein, the second PCA model used for extracting the time course profile could have been used to calculate the metabolic flux rates of each metabolite. This would be an efficacious continuation of the previous study, especially with a thorough comparison to traditional metabolic flux calculation methods. Using a new metabolic system, it is hypothesized that a comparison of these flux calculations would

reveal that utilizing PCA is more efficient, objective and robust in the calculation of these fluxes in a comprehensive metabolic flux analysis.

Further study and validation of the 2D *m/z* cluster method would also be efficacious. Using simulated chromatograms of varying saturation, mass spectral variability and noise, the various parameters of the software algorithm could be studied and evaluated. The parameters of interest include the box size used to cluster *m/z* with similar peak width and retention time indices, the *S/N* threshold, and the maximum width threshold applied to consider a mass channel pure. Current research is being conducted by K.L. Berrier, with promising results that verify the previously published experimental findings.

6.2 Chapter 3 Summary, Limitations and Future Directions

Presented in Chapter 3 was the optimization of the tile-based F-ratio method using the area under the curve (AUC) of the receiver operating characteristic (ROC) curves generated from the results of the F-ratio comparison of the 200/20 ppm and 100/10 ppm native/non-native spike solutions in diesel. Two parameters, the *S/N* threshold applied before the F-ratio calculation and the number of *m/z* used to calculate the average F-ratio, were varied. It was determined that the optimal parameters were the *S/N* 10 and 10 *m/z*, which gave the software a ~21% improvement on its ability to discriminate between true positives and false positives. Furthermore, the use of the AUC was consistent in its conclusions without prior knowledge of whether the spiked analytes were true positive instances or not.

The sensitivity of the tile-based F-ratio software resulted in several limitations to the originally framed investigation. Originally four spike levels (800/80, 400/40, 200/20, 100/10 ppm native/non-native) were to be presented and analyzed, both as adjacent class comparisons

(800/80 vs. 400/40, 400/40 vs 200/20, and 200/20 vs 100/10) and absolute class comparisons (each spike level compared to the stock solution, that is, 0 ppm). However, due to the sheer volume of spike standards required to make enough solution for the various dilutions and injection replicates, the higher concentration spike solutions actually diluted the diesel solvent. The slight dilutions of the diesel used as the matrix were readily picked up in the F-ratio analysis, and calculations of the diesel signal had to be performed to confirm there was no nominal dilution in the 200/20 vs 100/10 ppm solutions that were eventually used for the investigation. While it seems simply a pesky limitation, it is mentioned in detail here as a "word of warning," so to speak, to future researchers hoping to fabricate class comparisons with large numbers of spiked analytes.

New research on the tile-based F-ratio software continues. Future investigations will build upon those previously published. Specifically, the optimization of other parameters of the tile-based F-ratio software, such as tile size or cluster window, including the efficacy of employing a tile size that changes throughout the separation as peak widths may increase due to the general elution problem in isothermal or pseudo-isothermal regions of the separation, may be studied. The alteration of diesel fuel with regards to different chemical alteration methods or perhaps physical alterations could also be investigated. Finally, adaptation of the F-ratio code for other data structures, such as 2D *m/z* cluster plots or high resolution mass spectrometric data may be an important step to encourage other investigators to utilize the F-ratio software. Additionally, the spiked diesel fuel separations are further being utilized by S.E. Prebihalo to demonstrate the efficacy of PARAFAC for deconvolution of coeluting GC × GC – TOFMS peaks for quantification as long as there is no deviation from trilinearity of the data.

6.3 Chapter 4 Summary, Limitations and Future Directions

The optimized parameters elucidated in Chapter 3 were applied to a process analytical chemistry (PAC) investigation of catalyst poisons in an industrial polymerization plant. Solvent samples from two sampling points were analyzed via GC × GC -TOFMS and the tile-based F-ratio software. Samples taken from "Excellent" polymerization processes at Point I, prior to purification, and Point II, post purification, were compared in a supervised, non-targeted F-ratio class comparison to elucidate those analytes that were removed via the purification process. It was discovered that oxygenate compounds and branched alkanes were removed via the purification process. A single sample from Point II during a "Bad" polymerization process was also analyzed and revealed the presence of oxygenates, even after the purification step. It is likely these oxygenates were indicative of the presence of the catalyst poison, even if they weren't the catalyst poison itself. The conclusions of this investigation led to the recommendation for on-site, online monitoring pre- and post-purification during the polymerization process in order to prevent the proliferation of catalyst poisons in the system in the future.

This investigation was elucidative of a real-world industrial problem. Obvious limitations, especially in terms of academic analytical chemistry, abound. Most notably, the single "Bad" sample from Point II lacked both a partner sample from Point I and samples from different days. Therefore, very few conclusions could be drawn regarding the overall system during polymerization yield reduction. Future investigations in regards to this, or other, industrial processes, would ideally have multiple samples from consistent sampling points during both high and low yield for statistical comparison. Additionally, the use of a solvent delay or programmable temperature vaporization (PVT) inlet, may help in the discovery of low-

concentration containments whose signal is drowned out due to the overwhelming signal of the solvent. Finally, while PARAFAC was performed in order to better determine the identity of the analytes from their pure spectra, using a high-resolution mass spectrometer (HRMS) might help in the identification of the unknown oxygenates, or at least provide a more accurate molecular weight.

6.4 Chapter 5 Summary, Limitations and Future Directions

Chapter 5 included a simulation-based investigation of the minimum chromatographic resolution for a chemometric algorithm to successfully deconvolute coeluting GC-MS peaks. This is referred to as the limit of chemometric resolution, R_s*. The probabilistic theory of component overlap in separations was presented defining chemometric success and calculating the probability of peak overlap in a GC-MS separation. Chromatograms consisting of a target analyte and interferent analyte coeluting at various *S/N* thresholds and chromatographic resolutions were generated. A chemometric algorithm, in this case MCR-ALS, was applied to deconvolute the target from the interferent. The results of the chemometric model, both the peak area and mass spectrum of the target, were compared to the originally simulated data to determine the efficacy of the deconvolution method. The results of a simulation based study indicated that the R_s* of MCR-ALS is about 0.13, corresponding to an 87% probability of deconvoluting a target analyte in a chromatogram with a saturation factor of 0.5.

The limitation of this study is twofold. Firstly, the data was simulated. Although it was the aim of the investigation to simulate as realistic data as possible by including isomers for overlap and both high and low *S/N*, it is likely that in real-world applications with complex matrices, deconvolution may be more difficult than implied by the results reported in Chapter 5.

Secondly, only one chemometric algorithm was investigated, MCR-ALS. Further investigations on GRAM, PARAFAC, and CLS may provide more input into how various chemometric models perform differently in various situations. As mentioned in the introduction, each deconvolution algorithm requires different data structure, input parameters and may be more efficacious in some situations over others. Therefore, the choice of algorithm should fit with the type of data and questions of interest, and each algorithm will likely have a different R_s*.

6.5 Final Thoughts

The research presented herein represents the vast majority of the work performed over the last four years. Many of the chemometric models, statistical calculations, and sample preparations were performed multiple times before they were done correctly or meaningfully. While every side-track, misstep and re-do of the research cannot and should not be discussed in detail, it is worth mentioning, for the sake of current and future scientists, that it is all a part of the scientific and learning processes, and contribute to the final results nonetheless.

Finally, it is my hope and aim that these research investigations elucidate the separation power of one- and two-dimensional gas chromatography coupled with mass spectrometry for the analysis of complex samples as well as the efficacy of chemometric methods for gleaning meaningful information from these complex data sets. And while the specific methods laid out in these aforementioned investigations may not prove useful to every investigator, GC as an instrumental platform and chemometric models for data analytics are important tools that belong in the wheelhouse of every analytical chemist.

BIBLIOGRAPHY

Adahchour, M., J. Beens, R. J. J. Vreuls, and U. A. Th. Brinkman. "Recent Developments in Comprehensive Two-Dimensional Gas Chromatography (GC × GC): II. Modulation and Detection." *TrAC Trends in Analytical Chemistry* 25, no. 6 (June 2006): 540–53. doi:10.1016/j.trac.2006.04.004.

Alber, Birgit E. "Biotechnological Potential of the Ethylmalonyl-CoA Pathway." *Applied Microbiology and Biotechnology* 89, no. 1 (January 1, 2011): 17–25. doi:10.1007/s00253-010-2873-z.

Amigo, José Manuel, Thomas Skov, and Rasmus Bro. "ChroMATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics." *Chemical Reviews* 110, no. 8 (August 11, 2010): 4582–4605. doi:10.1021/cr900394n.

Amigo, José Manuel, Thomas Skov, Rasmus Bro, Jordi Coello, and Santiago Maspoch. "Solving GC-MS Problems with PARAFAC2." *TrAC Trends in Analytical Chemistry* 27, no. 8 (September 2008): 714–25. doi:10.1016/j.trac.2008.05.011.

Andersson, Claus A, and Rasmus Bro. "The N-Way Toolbox for MATLAB." *Chemometrics and Intelligent Laboratory Systems* 52, no. 1 (August 14, 2000): 1–4. doi:10.1016/S0169-7439(00)00071-X.

"Applied Chemometrics." Accessed December 9, 2016. http://www.chemometrics.com/.

B. Gaines, Richard, Glenn S. Frysinger, Christopher M. Reddy, and Robert K. Nelson. "5 - Oil Spill Source Identification by Comprehensive Two-Dimensional Gas Chromatography (GC × GC) A2 - Wang, Zhendi." In *Oil Spill Environmental Forensics*, edited by Scott A. Stout, 169–XI. Burlington: Academic Press, 2007. http://www.sciencedirect.com/science/article/pii/B9780123695239500094.

Bailey, Hope P., Sarah C. Rutan, and Peter W. Carr. "Factors That Affect Quantification of Diode Array Data in Comprehensive Two-Dimensional Liquid Chromatography Using Chemometric Data Analysis." *Journal of Chromatography A* 1218, no. 46 (November 18, 2011): 8411–22. doi:10.1016/j.chroma.2011.09.057.

Baker, Stuart G. "The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer." *Journal of the National Cancer Institute* 95, no. 7 (April 2, 2003): 511–15. doi:10.1093/jnci/95.7.511.

Bean, Heather D., Jane E. Hill, and Jean-Marie 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." *Journal of Chromatography A* 1394 (May 15, 2015): 111–17. doi:10.1016/j.chroma.2015.03.001.

Beebe, Kenneth R., Wayne W. Blaser, Robert A. Bredeweg, Jean Paul Chauvel, Richard S. Harner, Mark LaPack, Anne Leugers, Daniel P. Martin, Larry G. Wright, and E. Deniz Yalvac. "Process Analytical Chemistry." *Analytical Chemistry* 65, no. 12 (June 1, 1993): 199R–216R. doi:10.1021/ac00060a012.

Beebe, Kenneth R., and Bruce R. Kowalski. "An Introduction to Multivariate Calibration and Analysis." *Analytical Chemistry* 59, no. 17 (September 1, 1987): 1007A–1017A. doi:10.1021/ac00144a725.

Beebe, Kenneth R., Randy J. Pell, and Mary Beth Seasholtz. *Chemometrics: A Practical Guide*. John Wiley & Sons, Inc., 1998.

Berg, Frans van den, Christian B. Lyndgaard, Klavs M. Sørensen, and Søren B. Engelsen. "Process Analytical Technology in the Food Industry." *Trends in Food Science & Technology* 31, no. 1 (May 2013): 27–35. doi:10.1016/j.tifs.2012.04.007.

Blaise, Benjamin J. "Data-Driven Sample Size Determination for Metabolic Phenotyping Studies." *Analytical Chemistry* 85, no. 19 (October 1, 2013): 8943–50. doi:10.1021/ac4022314.

Blumberg, L. M., and M. S. Klee. "Optimal Heating Rate in Gas Chromatography." *Journal of Microcolumn Separations* 12, no. 9 (January 1, 2000): 508–14. doi:10.1002/1520-667X(2000)12:9<508::AID-MCS5>3.0.CO;2-Y.

Blumberg, Leonid M., Frank David, Matthew S. Klee, and Pat Sandra. "Comparison of One-Dimensional and Comprehensive Two-Dimensional Separations by Gas Chromatography." *Journal of Chromatography A*, 30th International Symposium on Capillary Chromatography and Electrophoresis and 4th Comprehensive Two-Dimensional Gas Chromatography Symposium, 1188, no. 1 (April 18, 2008): 2–16. doi:10.1016/j.chroma.2008.02.044.

Blumberg, Leonid M., and Matthew S. Klee. "Elution Parameters in Constant-Pressure, Single-Ramp Temperature-Programmed Gas Chromatography." *Journal of Chromatography A* 918, no. 1 (May 18, 2001): 113–20. doi:10.1016/S0021-9673(01)00659-8.

Bordawekar, Shailendra, Arani Chanda, Adrian M. Daly, Aaron W. Garrett, John P. Higgins, Mark A. LaPack, Todd D. Maloney, et al. "Industry Perspectives on Process Analytical Technology: Tools and Applications in API Manufacturing." *Organic Process Research & Development* 19, no. 9 (September 18, 2015): 1174–85. doi:10.1021/acs.oprd.5b00088.

Börner, Jana, Sebastian Buchinger, and Dietmar Schomburg. "A High-Throughput Method for Microbial Metabolome Analysis Using Gas Chromatography/Mass Spectrometry." *Analytical Biochemistry* 367, no. 2 (August 15, 2007): 143–51. doi:10.1016/j.ab.2007.04.036.

Bradley, Andrew P. "ROC Curves and the X2 Test." *Pattern Recognition Letters* 17, no. 3 (March 6, 1996): 287–94. doi:10.1016/0167-8655(95)00121-2.

Brasseur, Catherine, Jessica Dekeirsschieter, Eline M. J. Schotsmans, Sjaak de Koning, Andrew S. Wilson, Eric Haubruge, and Jean-Francois Focant. "Comprehensive Two-Dimensional Gas Chromatography–time-of-Flight Mass Spectrometry for the Forensic Study of Cadaveric Volatile Organic Compounds Released in Soil by Buried Decaying Pig Carcasses." *Journal of Chromatography A*, Hyphenated and Multidimensional Chromatography Techniques, 1255 (September 14, 2012): 163–70. doi:10.1016/j.chroma.2012.03.048.

Brereton, Richard G. "Consequences of Sample Size, Variable Selection, and Model Validation and Optimisation, for Predicting Classification Ability from Analytical Data." *TrAC Trends in Analytical Chemistry*, Use and abuse of chemometrics, 25, no. 11 (December 2006): 1103–11. doi:10.1016/j.trac.2006.10.005.

Bro, Rasmus. "PARAFAC. Tutorial and Applications." *Chemometrics and Intelligent Laboratory Systems* 38, no. 2 (October 1997): 149–71. doi:10.1016/S0169-7439(97)00032-4.

Brokl, Michał, Louise Bishop, Christopher G. Wright, Chuan Liu, Kevin McAdam, and Jean-François Focant. "Multivariate Analysis of Mainstream Tobacco Smoke Particulate Phase by Headspace Solid-Phase Micro Extraction Coupled with Comprehensive Two-Dimensional Gas Chromatography–time-of-Flight Mass Spectrometry." *Journal of Chromatography A* 1370 (November 28, 2014): 216–29. doi:10.1016/j.chroma.2014.10.057.

Brown, Christopher D., and Herbert T. Davis. “Receiver Operating Characteristics Curves and Related Decision Measures: A Tutorial.” *Chemometrics and Intelligent Laboratory Systems* 80, no. 1 (January 20, 2006): 24–38. doi:10.1016/j.chemolab.2005.05.004.

Bruckner, Carsten A., Bryan J. Prazen, and Robert E. Synovec. “Comprehensive Two-Dimensional High-Speed Gas Chromatography with Chemometric Analysis.” *Analytical Chemistry* 70, no. 14 (July 1, 1998): 2796–2804. doi:10.1021/ac980164m.

Callis, James B., Deborah L. Illman, and Bruce R. Kowalski. “Process Analytical Chemistry.” *Analytical Chemistry* 59, no. 9 (May 1, 1987): 624A–637A. doi:10.1021/ac00136a001.

Carter, Jane V., Jianmin Pan, Shesh N. Rai, and Susan Galandiuk. “ROC-Ing along: Evaluation and Interpretation of Receiver Operating Characteristic Curves.” *Surgery* 159, no. 6 (June 2016): 1638–45. doi:10.1016/j.surg.2015.12.029.

Castillo, Sandra, Ismo Mattila, Jarkko Miettinen, Matej Orešič, and Tuulia Hyötyläinen. “Data Analysis Tool for Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry.” *Analytical Chemistry* 83, no. 8 (April 15, 2011): 3058–67. doi:10.1021/ac103308x.

Cochran, Jack. “Evaluation of Comprehensive Two-Dimensional Gas Chromatography – Time-of-Flight Mass Spectrometry for the Determination of Pesticides in Tobacco.” *Journal of Chromatography A*, Trends and Developments in Gas Chromatography, 1186, no. 1–2 (April 4, 2008): 202–10. doi:10.1016/j.chroma.2008.01.043.

Comas, Enric, R. Ana Gimeno, Joan Ferré, Rosa M. Marcé, Francesc Borrull, and F. Xavier Rius. “Quantification from Highly Drifted and Overlapped Chromatographic Peaks Using Second-Order Calibration Methods.” *Journal of Chromatography A* 1035, no. 2 (May 7, 2004): 195–202. doi:10.1016/j.chroma.2004.02.069.

Cordero, Chiara, Johannes Kiefl, Peter Schieberle, Stephen E. Reichenbach, and Carlo Bicchi. “Comprehensive Two-Dimensional Gas Chromatography and Food Sensory Properties: Potential and Challenges.” *Analytical and Bioanalytical Chemistry* 407, no. 1 (October 30, 2014): 169–91. doi:10.1007/s00216-014-8248-z.

Cordero, Chiara, Erica Liberto, Carlo Bicchi, Patrizia Rubiolo, Peter Schieberle, Stephen E. Reichenbach, and Qingping Tao. “Profiling Food Volatiles by Comprehensive Two-Dimensional Ga Schromatography Coupled with Mass Spectrometry: Advanced Fingerprinting Approaches for Comparative Analysis of the Volatile Fraction of Roasted Hazelnuts (Corylus Avellana L.) from Different Origins.” *Journal of Chromatography A* 1217, no. 37 (September 10, 2010): 5848–58. doi:10.1016/j.chroma.2010.07.006.

Crown, Scott B., and Maciek R. Antoniewicz. “Publishing 13C Metabolic Flux Analysis Studies: A Review and Future Perspectives.” *Metabolic Engineering* 20 (November 2013): 42–48. doi:10.1016/j.ymben.2013.08.005.

Dal Nogare, Stephen. “Gas Chromatography.” *Analytical Chemistry* 32, no. 5 (April 1, 1960): 19–25. doi:10.1021/ac60161a602.

Dauner, Michael. “From Fluxes and Isotope Labeling Patterns towards in Silico Cells.” *Current Opinion in Biotechnology*, Analytical Biotechnology, 21, no. 1 (February 2010): 55–62. doi:10.1016/j.copbio.2010.01.014.

Davis, Joe M., and Leonid M. Blumberg. “Probability Theory for Number of Mixture Components Resolved by N Independent Columns.” *Journal of Chromatography A*, Chemical Separations and Chemometrics, 1096, no. 1–2 (November 25, 2005): 28–39. doi:10.1016/j.chroma.2005.03.137.

Davis, Joe M., and J. Calvin Giddings. “Statistical Method for Estimation of Number of Components from Single Complex Chromatograms: Theory, Computer-Based Testing, and Analysis of Errors.” *Analytical Chemistry* 57, no. 12 (October 1, 1985): 2168–77. doi:10.1021/ac00289a002.

Davis, Joe M., Dwight R. Stoll, and Peter W. Carr. “Effect of First-Dimension Undersampling on Effective Peak Capacity in Comprehensive Two-Dimensional Separations.” *Analytical Chemistry* 80, no. 2 (January 1, 2008): 461–73. doi:10.1021/ac071504j.

Deursen, M. M. van, J. Beens, H. -G. Janssen, P. A. Leclercq, and C. A. Cramers. “Evaluation of Time-of-Flight Mass Spectrometric Detection for Fast Gas Chromatography.” *Journal of Chromatography A* 878, no. 2 (May 12, 2000): 205–13. doi:10.1016/S0021-9673(00)00300-9.

Deutsche Botanische Gesellschaft. “Berichte Der Deutschen Botanischen Gesellschaft,” 1883, v.

Dorman, Frank L., Joshua J. Whiting, Jack W. Cochran, and Jorge Gardea-Torresdey. “Gas Chromatography.” *Analytical Chemistry* 82, no. 12 (June 15, 2010): 4775–85. doi:10.1021/ac101156h.

Dorsey, John G., William T. Cooper, Barbara A. Siles, Joe P. Foley, and Howard G. Barth. “Liquid Chromatography: Theory and Methodology.” *Analytical Chemistry* 70, no. 12 (June 1, 1998): 591–644. doi:10.1021/a1980022h.

Dovichi, Norman J. “Bioinstrumental Analysis.” In *Teaching Bioanalytical Chemistry*, 1137:155–69. ACS Symposium Series 1137. American Chemical Society, 2013. http://dx.doi.org/10.1021/bk-2013-1137.ch008.

Dromey, R. G., Mark J. Stefik, Thomas C. Rindfleisch, and Alan M. Duffield. “Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry Data.” *Analytical Chemistry* 48, no. 9 (August 1, 1976): 1368–75. doi:10.1021/ac50003a027.

Dunlop, Peter J., C. M. Bignell, J. F. Jackson, and D. Brynn Hibbert. “Chemometric Analysis of Gas Chromatographic Data of Oils from Eucalyptus Species.” *Chemometrics and Intelligent Laboratory Systems*, InCINC ’94 Selected papers from the First International Chemometrics Internet Conference, 30, no. 1 (November 1995): 59–67. doi:10.1016/0169-7439(95)00036-4.

Duraipandian, Shiyamala, Wei Zheng, Joseph Ng, Jeffrey J.H. Low, A. Ilancheran, and Zhiwei Huang. “Simultaneous Fingerprint and High-Wavenumber Confocal Raman Spectroscopy Enhances Early Detection of Cervical Precancer In Vivo.” *Analytical Chemistry* 84, no. 14 (July 17, 2012): 5913–19. doi:10.1021/ac300394f.

“Eigenvector Research: Chemometrics Software, Consulting and Training.” Accessed December 9, 2016. http://eigenvector.com/.

Eilers, Paul H. C. “Unimodal Smoothing.” *Journal of Chemometrics* 19, no. 5–7 (May 1, 2005): 317–28. doi:10.1002/cem.935.

Erkel, Arian R van, and Peter M. Th Pattynama. “Receiver Operating Characteristic (ROC) Analysis: Basic Principles and Applications in Radiology.” *European Journal of Radiology* 27, no. 2 (May 1998): 88–94. doi:10.1016/S0720-048X(97)00157-5.

Ernst, Madeleine, Denise Brentan Silva, Ricardo Roberto Silva, Ricardo Z. N. Vêncio, and Norberto Peporine Lopes. “Mass Spectrometry in Plant Metabolomics Strategies: From Analytical Platforms to Data Acquisition and Processing.” *Natural Product Reports* 31, no. 6 (May 15, 2014): 784–806. doi:10.1039/C3NP70086K.

Fang, Mingliang, Julijana Ivanisevic, H. Paul Benton, Caroline H. Johnson, Gary J. Patti, Linh T. Hoang, Winnie Uritboonthai, Michael E. Kurczy, and Gary Siuzdak. “Thermal Degradation of Small Molecules: A Global Metabolomic Investigation.” *Analytical Chemistry*, October 4, 2015. doi:10.1021/acs.analchem.5b03003.

Fawcett, Tom. “An Introduction to ROC Analysis.” *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition, 27, no. 8 (June 2006): 861–74. doi:10.1016/j.patrec.2005.10.010.

Fiehn, Oliver, Joachim Kopka, Richard N. Trethewey, and Lothar Willmitzer. “Identification of Uncommon Plant Metabolites Based on Calculation of Elemental Compositions Using Gas Chromatography and Quadrupole Mass Spectrometry.” *Analytical Chemistry* 72, no. 15 (2000): 3573–80. doi:10.1021/ac991142i.

Fisher, Ronald A. *Statistical Methods for Research Workers*. 14th ed. Oliver and Boyd, 1970.

Fitz, Brian D., Brandyn C. Mannion, Khang To, Trinh Hoac, and Robert E. Synovec. “Evaluation of Injection Methods for Fast, High Peak Capacity Separations with Low Thermal Mass Gas Chromatography.” *Journal of Chromatography A* 1392 (May 1, 2015): 82–90. doi: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. “Enhancing Gas Chromatography–Time of Flight Mass Spectrometry Data Analysis Using Two-Dimensional Mass Channel Cluster Plots.” *Analytical Chemistry* 86, no. 8 (April 15, 2014): 3973–79. doi:10.1021/ac5004344.

Fitz, Brian D., and Robert 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.” *Analytica Chimica Acta* 913 (March 24, 2016): 160–70. doi:10.1016/j.aca.2016.01.045.

Fraga, Carlos G., Carsten A. Bruckner, and Robert E. Synovec. “Increasing the Number of Analyzable Peaks in Comprehensive Two-Dimensional Separations through Chemometrics.” *Analytical Chemistry* 73, no. 3 (February 1, 2001): 675–83. doi:10.1021/ac0010025.

Fraga, Carlos G., Angela M. Melville, and Bob W. Wright. “ROC-Curve Approach for Determining the Detection Limit of a Field Chemical Sensor.” *Analyst* 132, no. 3 (February 26, 2007): 230–36. doi:10.1039/B607843E.

Fraga, Carlos G., Gabriel A. Pérez Acosta, Michael D. Crenshaw, Krys Wallace, Gary M. Mong, and Heather A Colburn. “Impurity Profiling to Match a Nerve Agent to Its Precursor Source for Chemical Forensics Applications.” *Analytical Chemistry* 83, no. 24 (December 15, 2011): 9564–72. doi:10.1021/ac202340u.

Geladi, Paul, and Bruce R. Kowalski. “Partial Least-Squares Regression: A Tutorial.” *Analytica Chimica Acta* 185 (January 1, 1986): 1–17. doi:10.1016/0003-2670(86)80028-9.

Gerdes, K. R., and G. J. Suppes. “Miscibility of Ethanol in Diesel Fuels.” *Industrial & Engineering Chemistry Research* 40, no. 3 (February 1, 2001): 949–56. doi:10.1021/ie000566w.

Giddings, J. Calvin. “TWO-DIMENSIONAL SEPARATIONS: CONCEPT AND PROMISE.” *Analytical Chemistry* 56, no. 12 (October 1, 1984): 1258A–1270A. doi:10.1021/ac00276a717.

Giddings, J. Calvin. *Unified Separation Science*. John Wiley & Sons, Inc., 1991.

Gigliarano, Chiara, Silvia Figini, and Pietro Muliere. “Making Classifier Performance Comparisons When ROC Curves Intersect.” *Computational Statistics & Data Analysis* 77 (September 2014): 300–312. doi:10.1016/j.csda.2014.03.008.

Goodacre, Royston, Seetharaman Vaidyanathan, Warwick B. Dunn, George G. Harrigan, and Douglas B. Kell. “Metabolomics by Numbers: Acquiring and Understanding Global Metabolite Data.” *Trends in Biotechnology* 22, no. 5 (May 1, 2004): 245–52. doi:10.1016/j.tibtech.2004.03.007.

Goodwin, Cody R., Stacy D. Sherrod, Christina C. Marasco, Brian O. Bachmann, Nicole Schramm-Sapyta, John P. Wikswo, and John A. McLean. “Phenotypic Mapping of Metabolic Profiles Using Self-Organizing Maps of High-Dimensional Mass Spectrometry Data.” *Analytical Chemistry* 86, no. 13 (July 1, 2014): 6563–71. doi:10.1021/ac5010794.

Gorrochategui, Eva, Joaquim Jaumot, Sílvia Lacorte, and Romà Tauler. “Data Analysis Strategies for Targeted and Untargeted LC-MS Metabolomic Studies: Overview and Workflow.” *TrAC Trends in Analytical Chemistry* 82 (September 2016): 425–42. doi:10.1016/j.trac.2016.07.004.

Green, David Marvin, and John A. Swets. *Signal Detection Theory and Psychophysics*. John Wiley & Sons, Inc., 1966.

Grey, D. R, and B. J. T Morgan. “Some Aspects of ROC Curve-Fitting: Normal and Logistic Models.” *Journal of Mathematical Psychology* 9, no. 1 (February 1, 1972): 128–39. doi:10.1016/0022-2496(72)90009-0.

Gröger, Th., M. Schäffer, M. Pütz, B. Ahrens, K. Drew, M. Eschner, and R. Zimmermann. “Application of Two-Dimensional Gas Chromatography Combined with Pixel-Based Chemometric Processing for the Chemical Profiling of Illicit Drug Samples.” *Journal of Chromatography A*, 31st International Symposium on Capillary Chromatography31st International Symposium on Capillary Chromatography, 1200, no. 1 (July 18, 2008): 8–16. doi:10.1016/j.chroma.2008.05.028.

Gromski, Piotr S., Howbeer Muhamadali, David I. Ellis, Yun Xu, Elon Correa, Michael L. Turner, and Royston Goodacre. “A Tutorial Review: Metabolomics and Partial Least Squares-Discriminant Analysis – a Marriage of Convenience or a Shotgun Wedding.” *Analytica Chimica Acta* 879 (June 16, 2015): 10–23. doi:10.1016/j.aca.2015.02.012.

Gu, Haiwei, Zhengzheng Pan, Bowei Xi, Vincent Asiago, Brian Musselman, and Daniel Raftery. “Principal Component Directed Partial Least Squares Analysis for Combining Nuclear Magnetic Resonance and Mass Spectrometry Data in Metabolomics: Application to the Detection of Breast Cancer.” *Analytica Chimica Acta* 686, no. 1–2 (February 7, 2011): 57–63. doi:10.1016/j.aca.2010.11.040.

Guo, Jin, Yingfei Shi, Chengbao Xu, Rugang Zhong, Feng Zhang, Ting Zhang, Bo Niu, and Jianhua Wang. “Quantification of Plasma Myo-Inositol Using Gas Chromatography–mass Spectrometry.” *Clinica Chimica Acta* 460 (September 1, 2016): 88–92. doi:10.1016/j.cca.2016.06.022.

Halket, John M., Anna Przyborowska, Stephen E. Stein, W. Gary Mallard, Stephen Down, and Ronald A. Chalmers. “Deconvolution Gas Chromatography/Mass Spectrometry of Urinary Organic Acids – Potential for Pattern Recognition and Automated Identification of Metabolic Disorders.” *Rapid Communications in Mass Spectrometry* 13, no. 4 (February 28, 1999): 279–84. doi:10.1002/(SICI)1097-0231(19990228)13:4<279::AID-RCM478>3.0.CO;2-I.

Hantao, Leandro W., Ali Najafi, Cheng Zhang, Fabio Augusto, and Jared L. Anderson. “Tuning the Selectivity of Ionic Liquid Stationary Phases for Enhanced Separation of Nonpolar Analytes in Kerosene Using Multidimensional Gas Chromatography.” *Analytical Chemistry* 86, no. 8 (April 15, 2014): 3717–21. doi:10.1021/ac5004129.

Harris, Daniel C. *Quantitative Chemical Analysis*. 6th ed. New York, NY: W. H. Freeman and Company, 2003.
Harshman, Richard A., and Margaret E. Lundy. "PARAFAC: Parallel Factor Analysis." *Computational Statistics & Data Analysis* 18, no. 1 (August 1994): 39–72. doi:10.1016/0167-9473(94)90132-5.
Harvey, Paul McA., and Robert A. Shellie. "Data Reduction in Comprehensive Two-Dimensional Gas Chromatography for Rapid and Repeatable Automated Data Analysis." *Analytical Chemistry* 84, no. 15 (August 7, 2012): 6501–7. doi:10.1021/ac300664h.
Henry, Matthew C., and Clement R. Yonker. "Supercritical Fluid Chromatography, Pressurized Liquid Extraction, and Supercritical Fluid Extraction." *Analytical Chemistry* 78, no. 12 (June 1, 2006): 3909–16. doi:10.1021/ac0605703.
Hilden, Jørgen. "The Area under the ROC Curve and Its Competitors." *Medical Decision Making* 11, no. 2 (June 1, 1991): 95–101. doi:10.1177/0272989X9101100204.
Hiller, Karsten, Jasper Hangebrauk, Christian Jäger, Jana Spura, Kerstin Schreiber, and Dietmar Schomburg. "MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis." *Analytical Chemistry* 81, no. 9 (May 1, 2009): 3429–39. doi:10.1021/ac802689c.
Hiller, Karsten, Christian M. Metallo, Joanne K Kelleher, and Gregory Stephanopoulos. "Nontargeted Elucidation of Metabolic Pathways Using Stable-Isotope Tracers and Mass Spectrometry." *Analytical Chemistry* 82, no. 15 (August 1, 2010): 6621–28. doi:10.1021/ac1011574.
Hiller, Karsten, Christian Metallo, and Gregory Stephanopoulos. "Elucidation of Cellular Metabolism Via Metabolomics and Stable-Isotope Assisted Metabolomics." *Current Pharmaceutical Biotechnology* 12 (2011): 1075–86.
Hnatyshyn, S., P. Shipkova, and M. Sanders. "Expedient Data Mining for Nontargeted High-Resolution LC-MS Profiles of Biological Samples." *Bioanalysis* 5, no. 10 (2013): 1195–1210. doi:10.4155/bio.13.86.
Hobbs, Andria L., and José R. Almirall. "Trace Elemental Analysis of Automotive Paints by Laser Ablation–inductively Coupled Plasma–mass Spectrometry (LA–ICP–MS)." *Analytical & Bioanalytical Chemistry* 376, no. 8 (August 15, 2003): 1265–71. doi:10.1007/s00216-003-1918-x.
Hoggard, Jamin C. "peg2mat3p8," n.d. http://depts.washington.edu/synlab/software/.
Hoggard, Jamin C., W. Christopher Siegler, and Robert E. Synovec. "Toward Automated Peak Resolution in Complete GC × GC–TOFMS Chromatograms by PARAFAC." *Journal of Chemometrics* 23, no. 7–8 (July 1, 2009): 421–31. doi:10.1002/cem.1239.
Hoggard, Jamin C., and Robert E. Synovec. "Parallel Factor Analysis (PARAFAC) of Target Analytes in GC × GC−TOFMS Data: Automated Selection of a Model with an Appropriate Number of Factors." *Analytical Chemistry* 79, no. 4 (February 1, 2007): 1611–19. doi:10.1021/ac061710b.
Hoggard, Jamin C., Jon H. Wahl, Robert E. Synovec, Gary M. Mong, and Carlos G. Fraga. "Impurity Profiling of a Chemical Weapon Precursor for Possible Forensic Signatures by Comprehensive Two-Dimensional Gas Chromatography/Mass Spectrometry and Chemometrics." *Analytical Chemistry* 82, no. 2 (January 15, 2010): 689–98. doi:10.1021/ac902247x.
Hope, Janiece L., Amanda E. Sinha, Bryan J. Prazen, and Robert E. Synovec. "Evaluation of the DotMap Algorithm for Locating Analytes of Interest Based on Mass Spectral Similarity in

Data Collected Using Comprehensive Two-Dimensional Gas Chromatography Coupled with Time-of-Flight Mass Spectrometry." *Journal of Chromatography A*, 2nd International Symposium on Comprehensive Multidimensional Gas Chromatography 2nd International Symposium on Comprehensive Multidimensional Gas Chromatography, 1086, no. 1–2 (September 9, 2005): 185–92. doi:10.1016/j.chroma.2005.06.026.

Hua, Qiang, Chen Yang, and Bernhard Ø. Palsson. "Analysis of Experimental Evolution of Escherichia Coli Based on Flux Profiling." *Journal of Biotechnology*, Biotechnology for the Sustainability of Human Society IBS 2008 Abstracts 13th International Biotechnology Symposium and Exhibition, 136, Supplement (October 2008): S353–54. doi:10.1016/j.jbiotec.2008.07.811.

Hubaux, Andre, and Gilbert Vos. "Decision and Detection Limits for Calibration Curves." *Analytical Chemistry* 42, no. 8 (July 1, 1970): 849–55. doi:10.1021/ac60290a013.

Humston, Elizabeth M., Kenneth M. Dombek, Jamin C. Hoggard, Elton T. Young, and Robert E. Synovec. "Time-Dependent Profiling of Metabolites from Snf1 Mutant and Wild Type Yeast Cells." *Analytical Chemistry* 80, no. 21 (November 1, 2008): 8002–11. doi:10.1021/ac800998j.

Humston, Elizabeth M., Joshua D. Knowles, Andrew McShea, and Robert E. Synovec. "Quantitative Assessment of Moisture Damage for Cacao Bean Quality Using Two-Dimensional Gas Chromatography Combined with Time-of-Flight Mass Spectrometry and Chemometrics." *Journal of Chromatography A* 1217, no. 12 (March 19, 2010): 1963–70. doi:10.1016/j.chroma.2010.01.069.

Humston, Elizabeth M., Yan Zhang, Gregory F. Brabeck, Andrew McShea, and Robert E. Synovec. "Development of a GC×GC–TOFMS Method Using SPME to Determine Volatile Compounds in Cacao Beans." *Journal of Separation Science* 32, no. 13 (July 1, 2009): 2289–95. doi:10.1002/jssc.200900143.

Ichihara, Ken 'ichi, Chihiro Kohsaka, Naohiro Tomari, Tamami Kiyono, Jun Wada, Kiyoo Hirooka, and Yoshihiro Yamamoto. "Fatty Acid Analysis of Triacylglycerols: Preparation of Fatty Acid Methyl Esters for Gas Chromatography." *Analytical Biochemistry* 495 (February 15, 2016): 6–8. doi:10.1016/j.ab.2015.11.009.

Ipsen, Andreas. "Derivation from First Principles of the Statistical Distribution of the Mass Peak Intensities of MS Data." *Analytical Chemistry*, 2015. doi:10.1021/ac503554u.

IUPAC. *Compendium of Chemical Terminology (the "Gold Book").* 2nd ed. Oxford, 1997. http://goldbook.iupac.org/C01075.html.

Jalali-Heravi, Mehdi, Hadi Parastar, Mohsen Kamalzadeh, Roma Tauler, and Joaquim Jaumot. "MCRC Software: A Tool for Chemometric Analysis of Two-Way Chromatographic Data." *Chemometrics and Intelligent Laboratory Systems* 104, no. 2 (December 15, 2010): 155–71. doi:10.1016/j.chemolab.2010.08.002.

Jaumot, Joaquim, Anna de Juan, and Romà Tauler. "MCR-ALS GUI 2.0: New Features and Applications." *Chemometrics and Intelligent Laboratory Systems* 140 (January 15, 2015): 1–12. doi:10.1016/j.chemolab.2014.10.003.

Jazmin, Lara J., and Jamey D. Young. "Isotopically Nonstationary 13C Metabolic Flux Analysis." In *Systems Metabolic Engineering*, edited by Hal S. Alper, 367–90. Methods in Molecular Biology 985. Humana Press, 2013. doi:10.1007/978-1-62703-299-5_18.

Jennerwein, Maximilian K., Markus Eschner, Thomas Gröger, Thomas Wilharm, and Ralf Zimmermann. "Complete Group-Type Quantification of Petroleum Middle Distillates Based on Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass

Spectrometry (GC×GC-TOFMS) and Visual Basic Scripting." *Energy & Fuels* 28, no. 9 (September 18, 2014): 5670–81. doi:10.1021/ef501247h.

Jennerwein, Maximilian K., Aimée Celeste Sutherland, Markus Eschner, Thomas Gröger, Thomas Wilharm, and Ralf Zimmermann. "Quantitative Analysis of Modern Fuels Derived from Middle Distillates – The Impact of Diverse Compositions on Standard Methods Evaluated by an Offline Hyphenation of HPLC-Refractive Index Detection with GC×GC-TOFMS." *Fuel* 187 (January 1, 2017): 16–25. doi:10.1016/j.fuel.2016.09.033.

Johnson, Kevin J, and Robert E Synovec. "Pattern Recognition of Jet Fuels: Comprehensive GC×GC with ANOVA-Based Feature Selection and Principal Component Analysis." *Chemometrics and Intelligent Laboratory Systems*, Fourth International Conference on Environ metrics and Chemometrics held in Las Vegas, NV, USA, 18-20 September 2000, 60, no. 1–2 (January 28, 2002): 225–37. doi:10.1016/S0169-7439(01)00198-8.

Jones, Melissa A., Ashley Kramer, Matthew Humbert, Tyler Vanadurongvan, Jonathan Maurer, Michael T. Bowser, and Anthony J. Borgerding. "Analysis and Monitoring of Volatile Analytes from Aqueous Solutions by Extractions into the Gas Phase Using Microdialysis Membranes and Coupling to Fast GC." *Analytical Chemistry* 80, no. 1 (January 1, 2008): 123–28. doi:10.1021/ac071530h.

Jorgenson, James W. "Capillary Electrophoresis: An Introduction." *Methods* 4, no. 3 (December 1992): 179–90. doi:10.1016/1046-2023(92)90033-5.

Kallio, Minna, Maarit Kivilompolo, Sami Varjo, Matti Jussila, and Tuulia Hyötyläinen. "Data Analysis Programs for Comprehensive Two-Dimensional Chromatography." *Journal of Chromatography A*, 32nd International Symposium on Capillary Chromatography and 5th GCxGC Symposium, 1216, no. 14 (April 3, 2009): 2923–27. doi:10.1016/j.chroma.2008.11.037.

Kawada, T. "Receiver Operating Characteristic Curve Analysis, Sensitivity Comparison and Individual Difference." *Clinical Radiology* 67, no. 9 (September 2012): 940. doi:10.1016/j.crad.2012.04.002.

Kehimkar, Benjamin, Jamin C. Hoggard, Luke C. Marney, Matthew C. Billingsley, Carlos G. Fraga, Thomas J. Bruno, and Robert 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." *Journal of Chromatography A* 1327 (January 31, 2014): 132–40. doi:10.1016/j.chroma.2013.12.060.

Kehimkar, Benjamin, Jamin C. Hoggard, Jeremy S. Nadeau, and Robert E. Synovec. "Targeted Mass Spectral Ratio Analysis: A New Tool for Gas Chromatography—mass Spectrometry." *Talanta* 103 (January 15, 2013): 267–75. doi:10.1016/j.talanta.2012.10.043.

Khan, Zareen S., Rakesh Kumar Ghosh, Rushali Girame, Sagar C. Utture, Manasi Gadgil, Kaushik Banerjee, D. Damodar Reddy, and Nalli Johnson. "Optimization of a Sample Preparation Method for Multiresidue Analysis of Pesticides in Tobacco by Single and Multi-Dimensional Gas Chromatography-Mass Spectrometry." *Journal of Chromatography A* 1343 (May 23, 2014): 200–206. doi:10.1016/j.chroma.2014.03.080.

Khanipov, K., G. Golovko, M. Rojas, L. Albayrak, O. Dobretsberger, M. Pimenova, N. Olson, S. Chumakov, and Y. Fofanov. "CoCo: An Application to Store High-Throughput Sequencing Data in Compact Text and Binary File Formats." In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1117–22, 2015. doi:10.1109/BIBM.2015.7359838.

Khodayari, Ali, Ali R. Zomorrodi, James C. Liao, and Costas D. Maranas. “A Kinetic Model of Escherichia Coli Core Metabolism Satisfying Multiple Sets of Mutant Flux Data.” *Metabolic Engineering* 25 (September 2014): 50–62. doi:10.1016/j.ymben.2014.05.014.

Kim, Seongho, and Xiang Zhang. “Discovery of False Identification Using Similarity Difference in GC-MS-Based Metabolomics: Discovery of False Identification in Metabolomics.” *Journal of Chemometrics* 29, no. 2 (February 2015): 80–86. doi:10.1002/cem.2665.

Kind, Tobias, Gert Wohlgemuth, Do Yup Lee, Yun Lu, Mine Palazoglu, Sevini Shahbaz, and Oliver Fiehn. “FiehnLib – Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry.” *Analytical Chemistry* 81, no. 24 (December 15, 2009): 10038–48. doi:10.1021/ac9019522.

Klee, Matthew S., and Leonid M. Blumberg. “Measurement of Retention in Comprehensive Two-Dimensional Gas Chromatography Using Flow Modulation with Methane Dopant.” *Journal of Chromatography A* 1217, no. 11 (March 12, 2010): 1830–37. doi:10.1016/j.chroma.2010.01.027.

Klee, Matthew S., and Leonid M. Blumberg. “Theoretical and Practical Aspects of Fast Gas Chromatography and Method Translation.” *Journal of Chromatographic Science* 40, no. 5 (May 1, 2002): 234–47. doi:10.1093/chromsci/40.5.234.

Klee, Matthew S., Jack Cochran, Mark Merrick, and Leonid M. Blumberg. “Evaluation of Conditions of Comprehensive Two-Dimensional Gas Chromatography That Yield a near-Theoretical Maximum in Peak Capacity Gain.” *Journal of Chromatography A* 1383 (February 27, 2015): 151–59. doi:10.1016/j.chroma.2015.01.031.

Klein, Sebastian, and Elmar Heinzle. “Isotope Labeling Experiments in Metabolomics and Fluxomics.” *WIREs Systems Biology and Medicine* 4 (June 2012): 261–72.

Kohl, Anja, Jack Cochran, and Donald M. Cropek. “Characterization of Military Fog Oil by Comprehensive Two-Dimensional Gas Chromatography.” *Journal of Chromatography A* 1217, no. 4 (January 22, 2010): 550–57. doi:10.1016/j.chroma.2009.11.054.

Koo, Imhoi, Seongho Kim, and Xiang Zhang. “Comparative Analysis of Mass Spectral Matching-Based Compound Identification in Gas Chromatography–mass Spectrometry.” *Journal of Chromatography A* 1298 (July 12, 2013): 132–38. doi:10.1016/j.chroma.2013.05.021.

Koo, Imhoi, Xue Shi, Seongho Kim, and Xiang Zhang. “iMatch2: Compound Identification Using Retention Index for Analysis of Gas Chromatography–mass Spectrometry Data.” *Journal of Chromatography A* 1337 (April 11, 2014): 202–10. doi:10.1016/j.chroma.2014.02.049.

Krakowska, Barbara, Ivana Stanimirova, Joanna Orzel, Michal Daszykowski, Ireneusz Grabowski, Grzegorz Zaleszczyk, and Miroslaw Sznajder. “Detection of Discoloration in Diesel Fuel Based on Gas Chromatographic Fingerprints.” *Analytical and Bioanalytical Chemistry* 407, no. 4 (November 19, 2014): 1159–70. doi:10.1007/s00216-014-8332-4.

Kramer, Richard. *Chemometric Techniques for Quantitative Analysis*. Marcel Dekker, Inc., 1998.

Krupčík, Ján, Pavel Májek, Roman Gorovenko, Pat Sandra, and Daniel W. Armstrong. “On Retentivity Tuning by Flow in the Second Column of Different Comprehensive Two Dimensional Gas Chromatographic Configurations.” *Journal of Chromatography A*, Selected Papers from the 34th ISCC and the 7th GCxGC Symposium34th International Symposium on Capillary Chromatography and 7th GCxGC Symposium, 1218, no. 21 (May 27, 2011): 3186–89. doi:10.1016/j.chroma.2011.03.042.

Kumar, Naveen, Ankit Bansal, G. S. Sarma, and Ravindra K. Rawal. “Chemometrics Tools Used in Analytical Chemistry: An Overview.” *Talanta* 123 (June 2014): 186–99. doi:10.1016/j.talanta.2014.02.003.

Latha, Indu, Stephen E. Reichenbach, and Qingping Tao. “Comparative Analysis of Peak-Detection Techniques for Comprehensive Two-Dimensional Chromatography.” *Journal of Chromatography A* 1218, no. 38 (September 23, 2011): 6792–98. doi:10.1016/j.chroma.2011.07.052.

Lavine, Barry K., and Jerome Workman. “Chemometrics.” *Analytical Chemistry* 85, no. 2 (January 15, 2013): 705–14. doi:10.1021/ac303193j.

Li, Shuifu, Jian Cao, and Shouzhi Hu. “Analyzing Hydrocarbon Fractions in Crude Oils by Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry under Reversed-Phase Column System.” *Fuel* 158 (October 15, 2015): 191–99. doi:10.1016/j.fuel.2015.05.026.

Lind, Pehr A, Lawrence B Marks, Donna Hollis, Ming Fan, Su-Min Zhou, Michael T Munley, Timothy D Shafman, Ronald J Jaszczak, and R. Edward Coleman. “Receiver Operating Characteristic Curves to Assess Predictors of Radiation-Induced Symptomatic Lung Injury.” *International Journal of Radiation Oncology*Biology*Physics* 54, no. 2 (October 1, 2002): 340–47. doi:10.1016/S0360-3016(02)02932-2.

Lisec, Jan, Friederike Hoffmann, Clemens Schmitt, and Carsten Jaeger. “Extending the Dynamic Range in Metabolomics Experiments by Automatic Correction of Peaks Exceeding the Detection Limit.” *Analytical Chemistry* 88, no. 15 (August 2, 2016): 7487–92. doi:10.1021/acs.analchem.6b02515.

Long, Gary L., and J. D. Winefordner. “Limit of Detection A Closer Look at the IUPAC Definition.” *Analytical Chemistry* 55, no. 7 (June 1, 1983): 712A–724A. doi:10.1021/ac00258a724.

Ludwig, Katelyn R., Liangliang Sun, Guijie Zhu, Norman J. Dovichi, and Amanda B. Hummon. “Over 2300 Phosphorylated Peptide Identifications with Single-Shot Capillary Zone Electrophoresis-Tandem Mass Spectrometry in a 100 Min Separation.” *Analytical Chemistry* 87, no. 19 (October 6, 2015): 9532–37. doi:10.1021/acs.analchem.5b02457.

Luong, Jim, Ronda Gras, Robert Mustacich, and Hernan Cortes. “Low Thermal Mass Gas Chromatography: Principles and Applications.” *Journal of Chromatographic Science* 44, no. 5 (May 1, 2006): 253–61. doi:10.1093/chromsci/44.5.253.

Magagna, Federico, Lucia Valverde-Som, Cristina Ruíz-Samblás, Luis Cuadros-Rodríguez, Stephen E. Reichenbach, Carlo Bicchi, and Chiara Cordero. “Combined Untargeted and Targeted Fingerprinting with Comprehensive Two-Dimensional Chromatography for Volatiles and Ripening Indicators in Olive Oil.” *Analytica Chimica Acta* 936 (September 14, 2016): 245–58. doi:10.1016/j.aca.2016.07.005.

Magnusson, B., and U. Ornemark. “Eurachem Guide: The Fitness for Purpose of Analytical Methods -- A Laboratory Guide of Method Validation and Related Topics,” 2014.

Marney, Luke C., W. Christopher Siegler, Brendon A. Parsons, Jamin C. Hoggard, Bob W. Wright, and Robert 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 (October 15, 2013): 887–95. doi:10.1016/j.talanta.2013.06.038.

Marney, Luke C., Stephen C. Kolwicz Jr., Rong Tian, and Robert E. Synovec. “Sample Preparation Methodology for Mouse Heart Metabolomics Using Comprehensive Two-

Dimensional Gas Chromatography Coupled with Time-of-Flight Mass Spectrometry." *Talanta* 108 (April 15, 2013): 123–30. doi:10.1016/j.talanta.2013.03.005.
Mazelis, Mendel, and Helen M. Pratt. "In Vivo Conversion of 5-Oxoproline to Glutamate by Higher Plants." *Plant Physiology* 57, no. 1 (January 1, 1976): 85–87. doi:10.1104/pp.57.1.85.
Megson, David, Eric J. Reiner, Karl J. Jobst, Frank L. Dorman, Mathew Robson, and Jean-François Focant. "A Review of the Determination of Persistent Organic Pollutants for Environmental Forensics Investigations." *Analytica Chimica Acta* 941 (October 19, 2016): 10–25. doi:10.1016/j.aca.2016.08.027.
Mercier, Sarah M., Bas Diepenbroek, Rene H. Wijffels, and Mathieu Streefland. "Multivariate PAT Solutions for Biopharmaceutical Cultivation: Current Progress and Limitations." *Trends in Biotechnology* 32, no. 6 (June 2014): 329–36. doi:10.1016/j.tibtech.2014.03.008.
Metz, Charles E. "Basic Principles of ROC Analysis." *Seminars in Nuclear Medicine* 8, no. 4 (October 1978): 283–98. doi:10.1016/S0001-2998(78)80014-2.
Metz, Charles E. "Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems." *Journal of the American College of Radiology*, Special Issue: Image PerceptionSpecial Issue: Image Perception, 3, no. 6 (June 2006): 413–22. doi:10.1016/j.jacr.2006.02.021.
Mi, Xiaoxia, Sicong Li, Yanhua Li, Kaiqiang Wang, Dan Zhu, and Gang Chen. "Quantitative Determination of 26 Steroids in Eggs from Various Species Using Liquid Chromatography–triple Quadrupole-Mass Spectrometry." *Journal of Chromatography A* 1356 (August 22, 2014): 54–63. doi:10.1016/j.chroma.2014.05.084.
Milman, Boris L., and Inna K. Zhurkovich. "Identification of Toxic Cyclopeptides Based on Mass Spectral Library Matching." *Analytical Chemistry Research* 1 (August 2014): 8–15. doi:10.1016/j.ancr.2014.06.002.
Mohler, Rachel E., Kenneth M. Dombek, Jamin C. Hoggard, Elton T. Young, and Robert E. Synovec. "Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Analysis of Metabolites in Fermenting and Respiring Yeast Cells." *Analytical Chemistry* 78, no. 8 (April 1, 2006): 2700–2709. doi:10.1021/ac052106o.
Mohler, Rachel E., Benjamin P. Tu, Kenneth M. Dombek, Jamin C. Hoggard, Elton T. Young, and Robert E. Synovec. "Identification and Evaluation of Cycling Yeast Metabolites in Two-Dimensional Comprehensive Gas Chromatography–time-of-Flight-Mass Spectrometry Data." *Journal of Chromatography A*, Trends and Developments in Gas Chromatography, 1186, no. 1–2 (April 4, 2008): 401–11. doi:10.1016/j.chroma.2007.10.063.
Moise, Alain, Bernard Clement, and Marios Raissis. "A Test for Crossing Receiver Operating Characteristic (Roc) Curves." *Communications in Statistics - Theory and Methods* 17, no. 6 (January 1, 1988): 1985–2003. doi:10.1080/03610928808829727.
Mondello, Luigi, Peter Quinto Tranchida, Paola Dugo, and Giovanni Dugo. "Comprehensive Two-Dimensional Gas Chromatography-Mass Spectrometry: A Review." *Mass Spectrometry Reviews* 27, no. 2 (March 1, 2008): 101–24. doi:10.1002/mas.20158.
Mostafa, Ahmed, Matthew Edwards, and Tadeusz Górecki. "Optimization Aspects of Comprehensive Two-Dimensional Gas Chromatography." *Journal of Chromatography A*, Hyphenated and Multidimensional Chromatography Techniques, 1255 (September 14, 2012): 38–55. doi:10.1016/j.chroma.2012.02.064.
Mou, Si, Liangliang Sun, and Norman J. Dovichi. "Accurate Determination of Peptide Phosphorylation Stoichiometry Via Automated Diagonal Capillary Electrophoresis Coupled

with Mass Spectrometry: Proof of Principle." *Analytical Chemistry* 85, no. 22 (November 19, 2013): 10692–96. doi:10.1021/ac402858a.

Mueller, Daniel, and Elmar Heinzle. "Stable Isotope-Assisted Metabolomics to Detect Metabolic Flux Changes in Mammalian Cell Cultures." *Current Opinion in Biotechnology*, Analytical biotechnology, 24, no. 1 (February 2013): 54–59. doi:10.1016/j.copbio.2012.10.015.

Mühlen, Carin von, Claudia Alcaraz Zini, Elina Bastos Caramão, and Philip J. Marriott. "Applications of Comprehensive Two-Dimensional Gas Chromatography to the Characterization of Petrochemical and Related Samples." *Journal of Chromatography A*, 28TH INTERNATIONAL SYMPOSIUM ON CAPILLARY CHROMATOGRAPHY AND ELECTROPHORESIS28TH INTERNATIONAL SYMPOSIUM ON CAPILLARY CHROMATOGRAPHY AND ELECTROPHORESIS, 1105, no. 1–2 (February 10, 2006): 39–50. doi:10.1016/j.chroma.2005.09.036.

Murphy, Kathleen R., Philip Wenig, Gavin Parcsi, Thomas Skov, and Richard M. Stuetz. "Characterizing Odorous Emissions Using New Software for Identifying Peaks in Chemometric Models of Gas Chromatography–mass Spectrometry Datasets." *Chemometrics and Intelligent Laboratory Systems* 118 (August 15, 2012): 41–50. doi:10.1016/j.chemolab.2012.07.006.

Nadeau, Jeremy S., Ryan B. Wilson, Brian D. Fitz, Jason T. Reed, and Robert E. Synovec. "Utilizing a Constant Peak Width Transform for Isothermal Gas Chromatography." *Journal of Chromatography A* 1218, no. 23 (June 10, 2011): 3718–24. doi:10.1016/j.chroma.2011.04.007.

Nadeau, Jeremy S., Ryan B. Wilson, Jamin C. Hoggard, Bob W. Wright, and Robert E. Synovec. "Study of the Interdependency of the Data Sampling Ratio with Retention Time Alignment and Principal Component Analysis for Gas Chromatography." *Journal of Chromatography A* 1218, no. 50 (December 16, 2011): 9091–9101. doi:10.1016/j.chroma.2011.10.031.

Nadeau, Jeremy S., Bob W. Wright, and Robert E. Synovec. "Chemometric Analysis of Gas Chromatography–mass Spectrometry Data Using Fast Retention Time Alignment via a Total Ion Current Shift Function." *Talanta* 81, no. 1–2 (April 15, 2010): 120–28. doi:10.1016/j.talanta.2009.11.046.

Navarro-Reig, Meritxell, Joaquim Jaumot, Teris A. van Beek, Gabriel Vivó-Truyols, and Romà Tauler. "Chemometric Analysis of Comprehensive LC×LC-MS Data: Resolution of Triacylglycerol Structural Isomers in Corn Oil." *Talanta* 160 (November 1, 2016): 624–35. doi:10.1016/j.talanta.2016.08.005.

Naz, Shama, Maria Vallejo, Antonia García, and Coral Barbas. "Method Validation Strategies Involved in Non-Targeted Metabolomics." *Journal of Chromatography A*, Method Validation, 1353 (August 1, 2014): 99–105. doi:10.1016/j.chroma.2014.04.071.

Nicholson, J. K., J. C. Lindon, and E. Holmes. "'Metabonomics': Understanding the Metabolic Responses of Living Systems to Pathophysiological Stimuli via Multivariate Statistical Analysis of Biological NMR Spectroscopic Data." *Xenobiotica* 29, no. 11 (January 1, 1999): 1181–89. doi:10.1080/004982599238047.

Niedenführ, Sebastian, Wolfgang Wiechert, and Katharina Nöh. "How to Measure Metabolic Fluxes: A Taxonomic Guide for 13C Fluxomics." *Current Opinion in Biotechnology*, Systems biology • Nanobiotechnology, 34 (August 2015): 82–90. doi:10.1016/j.copbio.2014.12.003.

Niu, Weihuan, Elisa Knight, Qingyou Xia, and Brian D. McGarvey. "Comparative Evaluation of Eight Software Programs for Alignment of Gas Chromatography–mass Spectrometry

Chromatograms in Metabolomics Experiments." *Journal of Chromatography A* 1374 (December 29, 2014): 199–206. doi:10.1016/j.chroma.2014.11.005.
Nizio, Katie D., and James J. Harynuk. "Analysis of Alkyl Phosphates in Petroleum Samples by Comprehensive Two-Dimensional Gas Chromatography with Nitrogen Phosphorus Detection and Post-Column Deans Switching." *Journal of Chromatography A* 1252 (August 24, 2012): 171–76. doi:10.1016/j.chroma.2012.06.070.
Nizio, Katie D., Teague M. MGinitie, and James J. Harynuk. "Comprehensive Multidimensional Separations for the Analysis of Petroleum." *Journal of Chromatography A*, Hyphenated and Multidimensional Chromatography Techniques, 1255 (September 14, 2012): 12–23. doi:10.1016/j.chroma.2012.01.078.
Nöh, Katharina, Karsten Grönke, Bing Luo, Ralf Takors, Marco Oldiges, and Wolfgang Wiechert. "Metabolic Flux Analysis at Ultra Short Time Scale: Isotopically Non-Stationary 13C Labeling Experiments." *Journal of Biotechnology*, Molecular Systems Biology, 129, no. 2 (April 30, 2007): 249–67. doi:10.1016/j.jbiotec.2006.11.015.
Nordström, Anders, Grace O'Maille, Chuan Qin, and Gary Siuzdak. "Nonlinear Data Alignment for UPLC−MS and HPLC−MS Based Metabolomics: Quantitative Analysis of Endogenous and Exogenous Metabolites in Human Serum." *Analytical Chemistry* 78, no. 10 (May 1, 2006): 3289–95. doi:10.1021/ac060245f.
Oberoi, Harinder Singh, Praveen Venkata Vadlani, Ronald L. Madl, Lavudi Saida, and Jithma P. Abeykoon. "Ethanol Production from Orange Peels: Two-Stage Hydrolysis and Fermentation Studies Using Optimized Parameters through Experimental Design." *Journal of Agricultural and Food Chemistry* 58, no. 6 (March 24, 2010): 3422–29. doi:10.1021/jf903163t.
O'Callaghan, Sean, David P. De Souza, Andrew Isaac, Qiao Wang, Luke Hodkinson, Moshe Olshansky, Tim Erwin, et al. "PyMS: A Python Toolkit for Processing of Gas Chromatography-Mass Spectrometry (GC-MS) Data. Application and Comparative Study of Selected Tools." *BMC Bioinformatics* 13, no. 1 (May 30, 2012): 115. doi:10.1186/1471-2105-13-115.
Ochiai, Nobuo, Teruyo Ieda, Kikuo Sasamoto, Yoshikatsu Takazawa, Shunji Hashimoto, Akihiro Fushimi, and Kiyoshi Tanabe. "Stir Bar Sorptive Extraction and Comprehensive Two-Dimensional Gas Chromatography Coupled to High-Resolution Time-of-Flight Mass Spectrometry for Ultra-Trace Analysis of Organochlorine Pesticides in River Water." *Journal of Chromatography A* 1218, no. 39 (September 28, 2011): 6851–60. doi:10.1016/j.chroma.2011.08.027.
Official Methods of Analysis of AOAC International. 6th ed. AOAC International, 1995.
Olivier, Ilse, and Du Toit Loots. "A Metabolomics Approach to Characterise and Identify Various Mycobacterium Species." *Journal of Microbiological Methods* 88, no. 3 (March 2012): 419–26. doi:10.1016/j.mimet.2012.01.012.
Olmo, M. del, A. González-Casado, N. A. Navas, and J. L. Vilchez. "Determination of Bisphenol A (BPA) in Water by Gas Chromatography-Mass Spectrometry." *Analytica Chimica Acta*, Papers presented at Euroanalysis IX, Session on "Emerging Techniques in Environmental Analysis," 346, no. 1 (June 30, 1997): 87–92. doi:10.1016/S0003-2670(97)00182-7.
O'Neill, Dennis T., Elizabeth A. Rochette, and Philip J. Ramsey. "Method Detection Limit Determination and Application of a Convenient Headspace Analysis Method for Methyl Tert-Butyl Ether in Water." *Analytical Chemistry* 74, no. 22 (November 1, 2002): 5907–11. doi:10.1021/ac0203239.

Ouyang, Xiyu, Jana M. Weiss, Jacob de Boer, Marja H. Lamoree, and Pim E. G. Leonards. "Non-Target Analysis of Household Dust and Laundry Dryer Lint Using Comprehensive Two-Dimensional Liquid Chromatography Coupled with Time-of-Flight Mass Spectrometry." *Chemosphere* 166 (January 2017): 431–37. doi:10.1016/j.chemosphere.2016.09.107.

Parastar, Hadi, and Nadia Akvan. "Multivariate Curve Resolution Based Chromatographic Peak Alignment Combined with Parallel Factor Analysis to Exploit Second-Order Advantage in Complex Chromatographic Measurements." *Analytica Chimica Acta* 816 (March 13, 2014): 18–27. doi:10.1016/j.aca.2014.01.051.

Parastar, Hadi, Mehdi Jalali-Heravi, and Roma 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." *Chemometrics and Intelligent Laboratory Systems*, Special Issue Section: Selected Papers from the 1st African-European Conference on Chemometrics, Rabat, Morocco, September 2010 Special Issue Section: Preprocessing methods Special Issue Section: Spectroscopic imaging, 117 (August 1, 2012): 80–91. doi:10.1016/j.chemolab.2012.02.003.

Parsons, Brendon A., Luke C. Marney, W. Christopher Siegler, Jamin C. Hoggard, Bob W. Wright, and Robert 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." *Analytical Chemistry* 87, no. 7 (April 7, 2015): 3812–19. doi:10.1021/ac504472s.

Parsons, Brendon A., David K. Pinkerton, Bob W. Wright, and Robert 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." *Journal of Chromatography A* 1440 (April 1, 2016): 179–90. doi:10.1016/j.chroma.2016.02.067.

Pell, Randy J., Mary Beth Seasholtz, Kenneth R. Beebe, and Mel V. Koch. "Process Analytical Chemistry and Chemometrics, Bruce Kowalski's Legacy at The Dow Chemical Company." *Journal of Chemometrics* 28, no. 5 (May 1, 2014): 321–31. doi:10.1002/cem.2535.

Pierce, Karisa M., and Jamin C. Hoggard. "Chromatographic Data Analysis. Part 3.3.4: Handling Hyphenated Data in Chromatography." *Analytical Methods* 6, no. 3 (January 17, 2014): 645–53. doi:10.1039/C3AY40965A.

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. "Fisher Ratio Method Applied to Third-Order Separation Data To Identify Significant Chemical Components of Metabolite Extracts." *Analytical Chemistry* 78, no. 14 (July 1, 2006): 5068–75. doi:10.1021/ac0602625.

Pierce, Karisa M., Jamin C. Hoggard, Rachel E. Mohler, and Robert E. Synovec. "Recent Advancements in Comprehensive Two-Dimensional Separations with Chemometrics." *Journal of Chromatography A*, 50 Years Journal of Chromatography, 1184, no. 1–2 (March 14, 2008): 341–52. doi:10.1016/j.chroma.2007.07.059.

Pierce, Karisa M., Janiece L. Hope, Kevin J. Johnson, Bob W. Wright, and Robert E. Synovec. "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*, Chemical Separations and Chemometrics, 1096, no. 1–2 (November 25, 2005): 101–10. doi:10.1016/j.chroma.2005.04.078.

Pierce, Karisa M., Benjamin Kehimkar, Luke C. Marney, Jamin C. Hoggard, and Robert E. Synovec. “Review of Chemometric Analysis Techniques for Comprehensive Two Dimensional Separations Data.” *Journal of Chromatography A*, Hyphenated and Multidimensional Chromatography Techniques, 1255 (September 14, 2012): 3–11. doi:10.1016/j.chroma.2012.05.050.

Pinkerton, David K., Brendon A. Parsons, Todd J. Anderson, and Robert E. Synovec. “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 (April 29, 2015): 66–76. doi:10.1016/j.aca.2015.02.040.

Płotka, Justyna M., Calum Morrison, David Adam, and Marek Biziuk. “Chiral Analysis of Chloro Intermediates of Methylamphetamine by One-Dimensional and Multidimentional NMR and GC/MS.” *Analytical Chemistry* 84, no. 13 (July 3, 2012): 5625–32. doi:10.1021/ac300503g.

Poe, Russel B., and Sarah C. Rutan. “Effects of Resolution, Peak Ratio and Sampling Frequency in Diode-Array Fluorescence Detection in Liquid Chromatography.” *Analytica Chimica Acta* 283, no. 2 (November 26, 1993): 845–53. doi:10.1016/0003-2670(93)85298-X.

Prazen, Bryan J., Carsten A. Bruckner, Robert E. Synovec, and Bruce R. Kowalski. “Second-Order Chemometric Standardization for High-Speed Hyphenated Gas Chromatography: Analysis of GC/MS and Comprehensive GC×GC Data.” *Journal of Microcolumn Separations* 11, no. 2 (January 1, 1999): 97–107. doi:10.1002/(SICI)1520-667X(1999)11:2<97::AID-MCS2>3.0.CO;2-Z.

Prazen, Bryan J., Robert E. Synovec, and Bruce R. Kowalski. “Standardization of Second-Order Chromatographic/Spectroscopic Data for Optimum Chemical Analysis.” *Analytical Chemistry* 70, no. 2 (January 1, 1998): 218–25. doi:10.1021/ac9706335.

Prebihalo, Sarah, Adrienne Brockman, Jack Cochran, and Frank L. Dorman. “Determination of Emerging Contaminants in Wastewater Utilizing Comprehensive Two-Dimensional Gas-Chromatography Coupled with Time-of-Flight Mass Spectrometry.” *Journal of Chromatography A* 1419 (November 6, 2015): 109–15. doi:10.1016/j.chroma.2015.09.080.

Quigley, Wes W. C., Carlos G. Fraga, and Robert E. Synovec. “Comprehensive LC×GC for Enhanced Headspace Analysis.” *Journal of Microcolumn Separations* 12, no. 3 (January 1, 2000): 160–66. doi:10.1002/(SICI)1520-667X(2000)12:3<160::AID-MCS5>3.0.CO;2-8.

Ramos, L. Scott, Eugenio Sanchez, and Bruce R. Kowalski. “Generalized Rank Annihilation Method.” *Journal of Chromatography A* 385 (January 9, 1987): 165–80. doi:10.1016/S0021-9673(01)94630-8.

Reaser, Brooke C., Song Yang, Brian D. Fitz, Brendon A. Parsons, Mary E. Lidstrom, and Robert E. Synovec. “Non-Targeted Determination of 13C-Labeling in the Methylobacterium Extorquens AM1 Metabolome Using the Two-Dimensional Mass Cluster Method and Principal Component Analysis.” *Journal of Chromatography A* 1432 (February 5, 2016): 111–21. doi:10.1016/j.chroma.2015.12.088.

Reichenbach, Stephen E. “Chapter 4 Data Acquisition, Visualization, and Analysis.” In *Comprehensive Analytical Chemistry*, edited by Lourdes Ramos, 55:77–106. Comprehensive Two Dimensional Gas Chromatography. Elsevier, 2009. http://www.sciencedirect.com/science/article/pii/S0166526X09055044.

Reichenbach, Stephen E., Visweswara Kottapalli, Mingtian Ni, and Arvind Visvanathan. “Computer Language for Identifying Chemicals with Comprehensive Two-Dimensional Gas Chromatography and Mass Spectrometry.” *Journal of Chromatography A*, 27th

International Symposium on Capillary ChromatographyRIVA 2004, 1071, no. 1–2 (April 15, 2005): 263–69. doi:10.1016/j.chroma.2004.08.125.
Reichenbach, Stephen E., Mingtian Ni, Visweswara Kottapalli, and Arvind Visvanathan. "Information Technologies for Comprehensive Two-Dimensional Gas Chromatography." *Chemometrics and Intelligent Laboratory Systems* 71, no. 2 (May 28, 2004): 107–20. doi:10.1016/j.chemolab.2003.12.009.
Reichenbach, Stephen E., Xue Tian, Chiara Cordero, and Qingping Tao. "Features for Non-Targeted Cross-Sample Analysis with Comprehensive Two-Dimensional Chromatography." *Journal of Chromatography A*, Selected Papers from the 35th International Symposium on Capillary Chromatography, the 26th International Symposium on MicroScale Bioseparations and the 8th GC×GC Symposium, San Diego, CA, USA, 1-5 May 201135th International Symposium on Capillary Chromatography, 26th International Symposium on MicroScale Bioseparations and 8th GC×GC Symposium, 1226 (February 24, 2012): 140–48. doi:10.1016/j.chroma.2011.07.046.
Reid, Vanessa R., Adam D. McBrady, and Robert E. Synovec. "Investigation of High-Speed Gas Chromatography Using Synchronized Dual-Valve Injection and Resistively Heated Temperature Programming." *Journal of Chromatography A* 1148, no. 2 (May 4, 2007): 236–43. doi:10.1016/j.chroma.2007.03.029.
Risticevic, Sanja, Heather Lord, Tadeusz Górecki, Catherine L. Arthur, and Janusz Pawliszyn. "Protocol for Solid-Phase Microextraction Method Development." *Nature Protocols* 5, no. 1 (January 2010): 122–39. doi:10.1038/nprot.2009.179.
Risticevic, Sanja, Erica A. Souza-Silva, Jennifer R. DeEll, Jack Cochran, and Janusz Pawliszyn. "Capturing Plant Metabolome with Direct-Immersion in Vivo Solid Phase Microextraction of Plant Tissues." *Analytical Chemistry* 88, no. 2 (January 19, 2016): 1266–74. doi:10.1021/acs.analchem.5b03684.
Robards, Kevin, Paul R. Haddad, and Peter E. Jackson. *Principles and Practice of Modern Chromatographic Methods*. Elsevier, Ltd., 2004.
Robinson, Anthony L., Paul K. Boss, Hildegarde Heymann, Peter S. Solomon, and Robert D. Trengove. "Development of a Sensitive Non-Targeted Method for Characterizing the Wine Volatile Profile Using Headspace Solid-Phase Microextraction Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry." *Journal of Chromatography A* 1218, no. 3 (January 21, 2011): 504–17. doi:10.1016/j.chroma.2010.11.008.
Rodier, C., O. Vandenabeele-Trambouze, R. Sternberg, D. Coscia, P. Coll, C. Szopa, F. Raulin, et al. "Detection of Martian Amino Acids by Chemical Derivatization Coupled to Gas Chromatography: In Situ and Laboratory Analysis." *Advances in Space Research* 27, no. 2 (January 1, 2001): 195–99. doi:10.1016/S0273-1177(01)00047-3.
Rontani, Jean-François, and Claude Aubert. "Hydrogen and Trimethylsilyl Transfers During EI Mass Spectral Fragmentation of Hydroxycarboxylic and Oxocarboxylic Acid Trimethylsilyl Derivatives." *Journal of the American Society for Mass Spectrometry* 19, no. 1 (January 2008): 66–75. doi:10.1016/j.jasms.2007.10.014.
Ruckebusch, C., and L. Blanchet. "Multivariate Curve Resolution: A Review of Advanced and Tailored Applications and Challenges." *Analytica Chimica Acta* 765 (February 26, 2013): 28–36. doi:10.1016/j.aca.2012.12.028.
Sadoughi, Navideh, Leigh M. Schmidtke, Guillaume Antalick, John W. Blackman, and Christopher C. Steel. "Gas Chromatography–Mass Spectrometry Method Optimized Using

Response Surface Modeling for the Quantitation of Fungal Off-Flavors in Grapes and Wine." *Journal of Agricultural and Food Chemistry*, February 21, 2015. doi:10.1021/jf505444r.

Sakodynsky, K. "M.S. Tswett—his Life." *Journal of Chromatography A* 49 (January 1, 1970): 2–17. doi:10.1016/S0021-9673(00)93603-3.

Samanipour, Saer, Petros Dimitriou-Christidis, Jonas Gros, Aureline Grange, and J. Samuel Arey. "Analyte Quantification with Comprehensive Two-Dimensional Gas Chromatography: Assessment of Methods for Baseline Correction, Peak Delineation, and Matrix Effect Elimination for Real Samples." *Journal of Chromatography A* 1375 (January 2, 2015): 123–39. doi:10.1016/j.chroma.2014.11.049.

Sampat, Andjoe, Martin Lopatka, Marjan Sjerps, Gabriel Vivo-Truyols, Peter Schoenmakers, and Arian van Asten. "Forensic Potential of Comprehensive Two-Dimensional Gas Chromatography." *TrAC Trends in Analytical Chemistry* 80 (June 2016): 345–63. doi:10.1016/j.trac.2015.10.011.

Sanchez, Eugenio, L. Scott Ramos, and Bruce R. Kowalski. "Generalized Rank Annihilation Method." *Journal of Chromatography A* 385 (January 9, 1987): 151–64. doi:10.1016/S0021-9673(01)94629-1.

Santhanam, Ganesh Ram. "Qualitative Optimization in Software Engineering: A Short Survey." *Journal of Systems and Software* 111 (January 2016): 149–56. doi:10.1016/j.jss.2015.09.001.

Santos, F. J., and M. T. Galceran. "Modern Developments in Gas Chromatography–mass Spectrometry-Based Environmental Analysis." *Journal of Chromatography A*, A Century of Chromatography 1903-2003, 1000, no. 1–2 (June 6, 2003): 125–51. doi:10.1016/S0021-9673(03)00305-4.

Sasaki, Tetsuya, Erina Koshi, Harumi Take, Toshihide Michihata, Masachika Maruya, and Toshiki Enomoto. "Characterisation of Odorants in Roasted Stem Tea Using Gas Chromatography–mass Spectrometry and Gas Chromatography-Olfactometry Analysis." *Food Chemistry* 220 (April 1, 2017): 177–83. doi:10.1016/j.foodchem.2016.09.208.

Sauer, Uwe. "Metabolic Networks in Motion: 13C-Based Flux Analysis." *Molecular Systems Biology* 2 (November 14, 2006): 62. doi:10.1038/msb4100109.

Schmarr, Hans-Georg, and Jörg Bernhardt. "Profiling Analysis of Volatile Compounds from Fruits Using Comprehensive Two-Dimensional Gas Chromatography and Image Processing Techniques." *Journal of Chromatography A* 1217, no. 4 (January 22, 2010): 565–74. doi:10.1016/j.chroma.2009.11.063.

Schoenherr, Regine M., Mingliang Ye, Michael Vannatta, and Norman J. Dovichi. "CE-Microreactor-CE-MS/MS for Protein Analysis." *Analytical Chemistry* 79, no. 6 (March 1, 2007): 2230–38. doi:10.1021/ac061638h.

Scoazec, Marie, Sylvere Durand, Alexis Chery, Lorenzo Galluzzi, and Guido Kroemer. "Chapter Eight - Metabolomic Profiling of Cultured Cancer Cells." In *Methods in Enzymology*, edited by Lorenzo Galluzzi and Guido Kroemer, Volume 543:165–78. Cell-Wide Metabolic Alterations Associated with Malignancy. Academic Press, 2014. http://www.sciencedirect.com/science/article/pii/B9780128013298000088.

Seeley, John V., and Stacy K. Seeley. "Multidimensional Gas Chromatography: Fundamental Advances and New Applications." *Analytical Chemistry* 85, no. 2 (January 15, 2013): 557–78. doi:10.1021/ac303195u.

Sgorbini, Barbara, Carlo Bicchi, Cecilia Cagliero, Chiara Cordero, Erica Liberto, and Patrizia Rubiolo. “Herbs and Spices: Characterization and Quantitation of Biologically-Active Markers for Routine Quality Control by Multiple Headspace Solid-Phase Microextraction Combined with Separative or Non-Separative Analysis.” *Journal of Chromatography A* 1376 (January 9, 2015): 9–17. doi:10.1016/j.chroma.2014.12.007.

Shellie, Robert, Philip Marriott, and Paul Morrison. “Concepts and Preliminary Observations on the Triple-Dimensional Analysis of Complex Volatile Samples by Using GC×GC−TOFMS.” *Analytical Chemistry* 73, no. 6 (March 1, 2001): 1336–44. doi:10.1021/ac000987n.

Simon, Levente L., Hajnalka Pataki, György Marosi, Fabian Meemken, Konrad Hungerbühler, Alfons Baiker, Srinivas Tummala, et al. “Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review.” *Organic Process Research & Development* 19, no. 1 (January 16, 2015): 3–62. doi:10.1021/op500261y.

Sinanian, Melanie M., Daniel W. Cook, Sarah C. Rutan, and Dayanjan S. Wijesinghe. “Multivariate Curve Resolution-Alternating Least Squares Analysis of High-Resolution Liquid Chromatography–Mass Spectrometry Data.” *Analytical Chemistry* 88, no. 22 (November 15, 2016): 11092–99. doi:10.1021/acs.analchem.6b03116.

Sinha, Amanda E, Carlos G Fraga, Bryan J Prazen, and Robert E Synovec. “Trilinear Chemometric Analysis of Two-Dimensional Comprehensive Gas Chromatography–time-of-Flight Mass Spectrometry Data.” *Journal of Chromatography A*, 26th International Symposium on Capillary Chromatography and Electrophoresis, 1027, no. 1–2 (February 20, 2004): 269–77. doi:10.1016/j.chroma.2003.08.081.

Sinha, Amanda E., Janiece L. Hope, Bryan J. Prazen, Carlos G. Fraga, Erik J. Nilsson, and Robert E. Synovec. “Multivariate Selectivity as a Metric for Evaluating Comprehensive Two-Dimensional Gas Chromatography–time-of-Flight Mass Spectrometry Subjected to Chemometric Peak Deconvolution.” *Journal of Chromatography A*, 8th International Symposium on Hyphenated Techniques in Chromatography and Hyphenated Chromatographic Analyzers, 1056, no. 1–2 (November 12, 2004): 145–54. doi:10.1016/j.chroma.2004.06.110.

Sinha, Amanda E., Janiece L. Hope, Bryan J. Prazen, Erik J. Nilsson, Rhona M. Jack, and Robert E. Synovec. “Algorithm for Locating Analytes of Interest Based on Mass Spectral Similarity in GC × GC–TOF-MS Data: Analysis of Metabolites in Human Infant Urine.” *Journal of Chromatography A*, Mass Spectrometry: Innovation and Application. Part III, 1058, no. 1–2 (November 26, 2004): 209–15. doi:10.1016/j.chroma.2004.08.064.

Sinkov, Nikolai A., and James J. Harynuk. “Three-Dimensional Cluster Resolution for Guiding Automatic Chemometric Model Optimization.” *Talanta* 103 (January 15, 2013): 252–59. doi:10.1016/j.talanta.2012.10.040.

Sinkov, Nikolai A., Brandon M. Johnston, P. Mark L. Sandercock, and James J. Harynuk. “Automated Optimization and Construction of Chemometric Models Based on Highly Variable Raw Chromatographic Data.” *Analytica Chimica Acta* 697, no. 1–2 (July 4, 2011): 8–15. doi:10.1016/j.aca.2011.04.029.

Skoog, Douglas A., F. James Holler, and Stanley R. Crouch. *Principles of Instrumental Analysis*. 6th ed. David Harris, 2007.

Skov, Thomas, Jamin C. Hoggard, Rasmus Bro, and Robert E. Synovec. “Handling within Run Retention Time Shifts in Two-Dimensional Chromatography Data Using Shift Correction

and Modeling." *Journal of Chromatography A* 1216, no. 18 (May 1, 2009): 4020–29. doi:10.1016/j.chroma.2009.02.049.

Skrobot, Vinicius L., Eustáquio V. R. Castro, Rita C. C. Pereira, Vânya M. D. Pasa, and Isabel C. P. Fortes. "Identification of Adulteration of Gasoline Applying Multivariate Data Analysis Techniques HCA and KNN in Chromatographic Data." *Energy & Fuels* 19, no. 6 (November 1, 2005): 2350–56. doi:10.1021/ef050031l.

"Speed Optimized Flow and Optimal Heating Rate in Gas Chromatography « ChromaBLOGraphy: Restek's Chromatography Blog." Accessed January 19, 2017. http://blog.restek.com/?p=1735.

Stadler, Sonja, Pierre-Hugues Stefanuto, Jonathan D. Byer, Michał Brokl, Shari Forbes, and Jean-François Focant. "Analysis of Synthetic Canine Training Aids by Comprehensive Two-Dimensional Gas Chromatography–time of Flight Mass Spectrometry." *Journal of Chromatography A*, Hyphenated and Multidimensional Chromatography Techniques, 1255 (September 14, 2012): 202–6. doi:10.1016/j.chroma.2012.04.001.

Stee, L. L. P. van, and U. A. Th. Brinkman. "Peak Detection Methods for GC × GC: An Overview." *TrAC Trends in Analytical Chemistry* 83, Part B (October 2016): 1–13. doi:10.1016/j.trac.2016.07.009.

Stein, S. E. "An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data." *Journal of the American Society for Mass Spectrometry* 10, no. 8 (August 1999): 770–81. doi:10.1016/S1044-0305(99)00047-1.

Subtil, Fabien, and Muriel Rabilloud. "An Enhancement of ROC Curves Made Them Clinically Relevant for Diagnostic-Test Comparison and Optimal-Threshold Determination." *Journal of Clinical Epidemiology* 68, no. 7 (July 2015): 752–59. doi:10.1016/j.jclinepi.2015.01.003.

Thompson, Michael, Stephen L. R. Ellison, and Roger Wood. "Harmonized Guidelines for Single-Laboratory Validation of Methods of Analysis." *Pure and Applied Chemistry* 74, no. 5 (2002): 835–55.

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. "Web Server for Peak Detection, Baseline Correction, and Alignment in Two-Dimensional Gas Chromatography Mass Spectrometry-Based Metabolomics Data." *Analytical Chemistry*, September 27, 2016. doi:10.1021/acs.analchem.6b00755.

Tomasi, Giorgio, Frans van den Berg, and Claus Andersson. "Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data." *Journal of Chemometrics* 18, no. 5 (May 1, 2004): 231–41. doi:10.1002/cem.859.

Uarrota, Virgílio Gavicho, Rodolfo Moresco, Bianca Coelho, Eduardo da Costa Nunes, Luiz Augusto Martins Peruch, Enilto de Oliveira Neubert, Miguel Rocha, and Marcelo Maraschin. "Metabolomics Combined with Chemometric Tools (PCA, HCA, PLS-DA and SVM) for Screening Cassava (Manihot Esculenta Crantz) Roots during Postharvest Physiological Deterioration." *Food Chemistry* 161 (October 15, 2014): 67–78. doi:10.1016/j.foodchem.2014.03.110.

Ubukata, Masaaki, Karl J. Jobst, Eric J. Reiner, Stephen E. Reichenbach, Qingping Tao, Jiliang Hang, Zhanpin Wu, A. John Dane, and Robert B. Cody. "Non-Targeted Analysis of Electronics Waste by Comprehensive Two-Dimensional Gas Chromatography Combined with High-Resolution Mass Spectrometry: Using Accurate Mass Information and Mass Defect Analysis to Explore the Data." *Journal of Chromatography A* 1395 (May 22, 2015): 152–59. doi:10.1016/j.chroma.2015.03.050.

United States Food and Drug Administration. “Guidance for Indsutry: PAT--A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance,” September 2004. http://www.fda.gov/downloads/drugs/guidances/ucm070305.pdf.

Veriotti, Tincuta, and Richard Sacks. “High-Speed GC and GC/Time-of-Flight MS of Lemon and Lime Oil Samples.” *Analytical Chemistry* 73, no. 18 (September 1, 2001): 4395–4402. doi:10.1021/ac010239d.

Verouden, Maikel P. H., Johan A. Westerhuis, Mariët J. van der Werf, and Age K. Smilde. “Exploring the Analysis of Structured Metabolomics Data.” *Chemometrics and Intelligent Laboratory Systems* 98, no. 1 (August 15, 2009): 88–96. doi:10.1016/j.chemolab.2009.05.004.

Wan, Katty X., Ilan Vidavsky, and Michael L. Gross. “Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle.” *Journal of the American Society for Mass Spectrometry* 13, no. 1 (January 2002): 85–88. doi:10.1016/S1044-0305(01)00327-0.

Wang, Yan-Feng, Juan Tian, Zhi-Hua Ji, Mao-Yong Song, and Hao Li. “Intracellular Metabolic Changes of Clostridium Acetobutylicum and Promotion to Butanol Tolerance during Biobutanol Fermentation.” *The International Journal of Biochemistry & Cell Biology* 78 (September 2016): 297–306. doi:10.1016/j.biocel.2016.07.031.

Wang, Zhengfang, Mengliang Zhang, and Peter de B. Harrington. “Comparison of Three Algorithms for the Baseline Correction of Hyphenated Data Objects.” *Analytical Chemistry* 86, no. 18 (September 16, 2014): 9050–57. doi:10.1021/ac501658k.

Ward, Joe. H. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58, no. 301 (March 1963): 236–44.

Watson, Nathanial E., Brendon A. Parsons, and Robert E. Synovec. “Performance Evaluation of Tile-Based Fisher Ratio Analysis Using a Benchmark Yeast Metabolome Dataset.” *Journal of Chromatography A* 1459 (August 12, 2016): 101–11. doi:10.1016/j.chroma.2016.06.067.

Watson, Nathanial E., Matthew M. VanWingerden, Karisa M. Pierce, Bob W. Wright, and Robert E. Synovec. “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, no. 1 (September 29, 2006): 111–18. doi:10.1016/j.chroma.2006.06.087.

Weggler, Benedikt A., Thomas Gröger, and Ralf Zimmermann. “Advanced Scripting for the Automated Profiling of Two-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry Data from Combustion Aerosol.” *Journal of Chromatography A* 1364 (October 17, 2014): 241–48. doi:10.1016/j.chroma.2014.08.091.

Wei, Xiaoli, Imhoi Koo, Seongho Kim, and Xiang Zhang. “Compound Identification in GC-MS by Simultaneously Evaluating the Mass Spectrum and Retention Index.” *Analyst* 139, no. 10 (April 15, 2014): 2507–14. doi:10.1039/C3AN02171H.

Weindl, Daniel, André Wegner, Christian Jäger, and Karsten Hiller. “Isotopologue Ratio Normalization for Non-Targeted Metabolomics.” *Journal of Chromatography A* 1389 (April 10, 2015): 112–19. doi:10.1016/j.chroma.2015.02.025.

Welthagen, Werner, Robert A. Shellie, Joachim Spranger, Michael Ristow, Ralf Zimmermann, and Oliver Fiehn. “Comprehensive Two-Dimensional Gas Chromatography–time-of-Flight Mass Spectrometry (GC × GC-TOF) for High Resolution Metabolomics: Biomarker Discovery on Spleen Tissue Extracts of Obese NZO Compared to Lean C57BL/6 Mice.” *Metabolomics* 1, no. 1 (n.d.): 65–73. doi:10.1007/s11306-005-1108-2.

Wiechert, Wolfgang. “13C Metabolic Flux Analysis.” *Metabolic Engineering* 3, no. 3 (July 2001): 195–206. doi:10.1006/mben.2001.0187.
Wiklund, Susanne, Erik Johansson, Lina Sjöström, Ewa J. Mellerowicz, Ulf Edlund, John P. Shockcor, Johan Gottfries, Thomas Moritz, and Johan Trygg. “Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models.” *Analytical Chemistry* 80, no. 1 (January 1, 2008): 115–22. doi:10.1021/ac0713510.
Wilson, Ryan B., Brian D. Fitz, Brandyn C. Mannion, Tina Lai, Roy K. Olund, Jamin C. Hoggard, and Robert E. Synovec. “High-Speed Cryo-Focusing Injection for Gas Chromatography: Reduction of Injection Band Broadening with Concentration Enrichment.” *Talanta* 97 (August 15, 2012): 9–15. doi:10.1016/j.talanta.2012.03.054.
Wilson, Ryan B., Jamin C. Hoggard, and Robert E. Synovec. “Fast, High Peak Capacity Separations in Gas Chromatography–Time-of-Flight Mass Spectrometry.” *Analytical Chemistry* 84, no. 9 (2012): 4167–73. doi:10.1021/ac300481k.
Wilson, Ryan B., Jamin C. Hoggard, and Robert E. Synovec. “High Throughput Analysis of Atmospheric Volatile Organic Compounds by Thermal Injection – Isothermal Gas Chromatography – Time-of-Flight Mass Spectrometry.” *Talanta* 103 (January 15, 2013): 95–102. doi:10.1016/j.talanta.2012.10.013.
Winter, Gal, and Jens O. Krömer. “Fluxomics – Connecting ‘omics Analysis and Phenotypes.” *Environmental Microbiology* 15, no. 7 (July 1, 2013): 1901–16. doi:10.1111/1462-2920.12064.
Wise, Barry M., and Neal B. Gallagher. “The Process Chemometrics Approach to Process Monitoring and Fault Detection.” *Journal of Process Control* 6, no. 6 (December 1, 1996): 329–48. doi:10.1016/0959-1524(96)00009-1.
Wold, Svante, Kim Esbensen, and Paul Geladi. “Principal Component Analysis.” *Chemometrics and Intelligent Laboratory Systems*, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, 2, no. 1–3 (August 1987): 37–52. doi:10.1016/0169-7439(87)80084-9.
Wold, Svante, and Michael Sjöström. “Chemometrics, Present and Future Success.” *Chemometrics and Intelligent Laboratory Systems* 44, no. 1–2 (December 14, 1998): 3–14. doi:10.1016/S0169-7439(98)00075-6.
Woldegebriel, Michael, John Gonsalves, Arian van Asten, and Gabriel Vivó-Truyols. “Robust Bayesian Algorithm for Targeted Compound Screening in Forensic Toxicology.” *Analytical Chemistry* 88, no. 4 (February 16, 2016): 2421–30. doi:10.1021/acs.analchem.5b04484.
Workman, Jerome, Barry Lavine, Ray Chrisman, and Mel Koch. “Process Analytical Chemistry.” *Analytical Chemistry* 83, no. 12 (June 15, 2011): 4557–78. doi:10.1021/ac200974w.
Workman, Jerome, David J. Veltkamp, Steve Doherty, Brian B. Anderson, Ken E. Creasy, Mel Koch, James F. Tatera, et al. “Process Analytical Chemistry.” *Analytical Chemistry* 71, no. 12 (June 1, 1999): 121–80. doi:10.1021/a1990007s.
Xu, Weichao, Jisheng Dai, Y. S. Hung, and Qinruo Wang. “Estimating the Area under a Receiver Operating Characteristic (ROC) Curve: Parametric and Nonparametric Ways.” *Signal Processing* 93, no. 11 (November 2013): 3111–23. doi:10.1016/j.sigpro.2013.05.010.
Yang, Song, Jamin C. Hoggard, Mary E. Lidstrom, and Robert E. Synovec. “Comprehensive Discovery of 13C Labeled Metabolites in the Bacterium Methylobacterium Extorquens AM1 Using Gas Chromatography–mass Spectrometry.” *Journal of Chromatography A*,

Advanced Analytical Separation Methods for Studying Man and his Environment InHonour of Marja-Liisa Riekkola's 60th Birthday, 1317 (November 22, 2013): 175–85. doi:10.1016/j.chroma.2013.08.059.

Yang, Song, Jamin C. Hoggard, Mary E. Lidstrom, and Robert E Synovec. "Gas Chromatography and Comprehensive Two-Dimensional Gas Chromatography Hyphenated with Mass Spectrometry for Targeted and Nontargeted Metabolomics." In *Metabolomics in Practice: Successful Strategies to Generate and Analyze Metabolic Data*, 69–92. Weinheim, Germany: Wiley-VHC, 2013.

Yang, Song, Jeremy S. Nadeau, Elizabeth M. Humston-Fulmer, Jamin C. Hoggard, Mary E. Lidstrom, and Robert E. Synovec. "Gas Chromatography–mass Spectrometry with Chemometric Analysis for Determining 12C and 13C Labeled Contributions in Metabolomics and 13C Flux Analysis." *Journal of Chromatography A* 1240 (June 1, 2012): 156–64. doi:10.1016/j.chroma.2012.03.072.

Yang, Song, Martin Sadilek, and Mary E. Lidstrom. "Streamlined Pentafluorophenylpropyl Column Liquid Chromatography–tandem Quadrupole Mass Spectrometry and Global 13C-Labeled Internal Standards Improve Performance for Quantitative Metabolomics in Bacteria." *Journal of Chromatography A* 1217, no. 47 (November 19, 2010): 7401–10. doi:10.1016/j.chroma.2010.09.055.

Yang, Song, Martin Sadilek, Robert E. Synovec, and Mary E. Lidstrom. "Liquid Chromatography–tandem Quadrupole Mass Spectrometry and Comprehensive Two-Dimensional Gas Chromatography–time-of-Flight Mass Spectrometry Measurement of Targeted Metabolites of Methylobacterium Extorquens AM1 Grown on Two Different Carbon Sources." *Journal of Chromatography A* 1216, no. 15 (April 10, 2009): 3280–89. doi:10.1016/j.chroma.2009.02.030.

Yehia, Ali M., and Heba M. Mohamed. "Chemometrics Resolution and Quantification Power Evaluation: Application on Pharmaceutical Quaternary Mixture of Paracetamol, Guaifenesin, Phenylephrine and P-Aminophenol." *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 152 (January 5, 2016): 491–500. doi:10.1016/j.saa.2015.07.101.

Zeng, Weiqing, Jan Hazebroek, Mary Beatty, Kevin Hayes, Christine Ponte, Carl Maxwell, and Cathy Xiaoyan Zhong. "Analytical Method Evaluation and Discovery of Variation within Maize Varieties in the Context of Food Safety: Transcript Profiling and Metabolomics." *Journal of Agricultural and Food Chemistry* 62, no. 13 (April 2, 2014): 2997–3009. doi:10.1021/jf405652j.

Zeng, Zhong-Da, Sung-Tong Chin, Helmut M. Hugel, and Philip J. Marriott. "Simultaneous Deconvolution and Re-Construction of Primary and Secondary Overlapping Peak Clusters in Comprehensive Two-Dimensional Gas Chromatography." *Journal of Chromatography A* 1218, no. 16 (April 22, 2011): 2301–10. doi:10.1016/j.chroma.2011.02.028.

Zeng, Zhong-Da, Helmut M. Hugel, and Philip J. Marriott. "Component Correlation between Related Samples by Using Comprehensive Two-Dimensional Gas Chromatography–time-of-Flight Mass Spectrometry with Chemometric Tools." *Journal of Chromatography A* 1254 (September 7, 2012): 98–106. doi:10.1016/j.chroma.2012.07.032.

Zeng, Zhongda, Jia Li, Helmut M. Hugel, Guowang Xu, and Philip J. Marriott. "Interpretation of Comprehensive Two-Dimensional Gas Chromatography Data Using Advanced Chemometrics." *TrAC Trends in Analytical Chemistry* 53 (January 2014): 150–66. doi:10.1016/j.trac.2013.08.009.

Zhang, Dabao, Xiaodong Huang, Fred E. Regnier, and Min Zhang. “Two-Dimensional Correlation Optimized Warping Algorithm for Aligning GC×GC−MS Data.” *Analytical Chemistry* 80, no. 8 (April 15, 2008): 2664–71. doi:10.1021/ac7024317.

Zhang, Lei, Zhongda Zeng, Chunxia Zhao, Hongwei Kong, Xin Lu, and Guowang Xu. “A Comparative Study of Volatile Components in Green, Oolong and Black Teas by Using Comprehensive Two-Dimensional Gas Chromatography–time-of-Flight Mass Spectrometry and Multivariate Data Analysis.” *Journal of Chromatography A*, Advances in Food Analysis, 1313 (October 25, 2013): 245–52. doi:10.1016/j.chroma.2013.06.022.

Zhao, Yimeng, Liangliang Sun, Michael D. Knierman, and Norman J. Dovichi. “Fast Separation and Analysis of Reduced Monoclonal Antibodies with Capillary Zone Electrophoresis Coupled to Mass Spectrometry.” *Talanta* 148 (February 1, 2016): 529–33. doi:10.1016/j.talanta.2015.11.020.

Zhou, Xiao H., and Constantine A. Gatsonis. “A Simple Method for Comparing Correlated Roc Curves Using Incomplete Data.” *Statistics in Medicine* 15, no. 15 (August 15, 1996): 1687–93. doi:10.1002/(SICI)1097-0258(19960815)15:15<1687::AID-SIM324>3.0.CO;2-S.

Zhu, Jiangjiang, Danijel Djukovic, Lingli Deng, Haiwei Gu, Farhan Himmati, E. Gabriela Chiorean, and Daniel Raftery. “Colorectal Cancer Detection Using Targeted Serum Metabolic Profiling.” *Journal of Proteome Research* 13, no. 9 (September 5, 2014): 4120–30. doi:10.1021/pr500494u. Accessed January 19, 2017. https://www.chem.agilent.com/cag/cabu/carriergas.htm.

VITA

Brooke C. Reaser was born in Las Vegas, Nevada in 1989. She grew up in Reno, where she was an avid downhill ski racer and flautist. After graduating from Reno High School in 2007, she attended St. Olaf College in Northfield, MN, where she earned degrees in Spanish and Chemistry in 2012. After completing the requirements of the doctoral degree, Brooke will relocate to New Jersey to begin a career as a Senior Analytical Chemist at L'Oréal USA.