Investigation of supervised and unsupervised discovery–based chemometric tools to expand the scope of multidimensional gas chromatography

Paige E. Sudol

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Robert E. Synovec, Chair

Matthew Bush

Dan Fu

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Investigation of supervised and unsupervised discovery–based chemometric tools to expand the scope of multidimensional gas chromatography

Paige E. Sudol

Chair of the Supervisory Committee:

Robert E. Synovec

Chemistry

Comprehensive two-dimensional gas chromatography (GC×GC) has emerged as a powerful multidisciplinary tool for the separation of complex mixtures since its inception in 1991. When coupled to a time-of-flight mass spectrometer (TOFMS), GC×GC analysis can confidently identify up to thousands of analytes. The wealth of information provided within this three-dimensional (3D) data cube is too great for manual scrutinization, hence necessitating non-targeted chemometric analysis. Within non-targeted analysis, both supervised and unsupervised algorithms exist, which differ by the presence and absence of sample class labels, respectively. This dissertation describes several studies aimed at critically investigating non-targeted chemometric tools such as tile-based Fisher ratio (F-ratio) analysis and principal component analysis (PCA) for numerous multidimensional GC platforms, including GC×GC-FID, GC×GC-

TOFMS, and comprehensive 3D GC (GC^3)-TOFMS. First, the utility of data binning prior to PCA is demonstrated for a GC×GC-FID dataset of five diesel fuels, wherein the optimum bin size maintains chemical selectivity and improves the signal-to-noise ratio (*S/N*). Next, the advantage of ranking F-ratio hitlists using the top F-ratio *m/z* is demonstrated for low concentration comparisons of spiked and un-spiked JP8 jet fuel and the "limit of discovery" is identified as the limit of quantification (*LOQ*). Tile-based F-ratio analysis is then coupled with an offline one-way analysis of variance (*ANOVA*) to characterize the geographical differences between five Sicilian wines. An unsupervised algorithm known as tile-based variance rank initiated-unsupervised sample indexing (VRI-USI) is developed to identify sample-to-sample relationships in GC×GC-TOFMS data. Upon application of VRI-USI to a complex multi-fuel dataset, patterns in the resulting *k*-means clustering index assignments for each hit in the hitlist correctly revealed the presence of classes and sub-classes, which holds promise for future studies requiring chemical characterization of unknown samples. Finally, the first application of PCA to GC^3 -TOFMS data is reported herein, wherein the 3D loadings are used to distinguish four jet fuels.

TABLE OF CONTENTS

List of Figures vi

List of Tables ix

Chapter 1. Introduction to gas chromatography, multidimensional gas chromatography, and chemometric data analysis 1

1.1 Fundamentals of gas chromatography 1

1.1.1 Principles 1

1.1.2 Figures of merit 4

1.1.3 Introduction to comprehensive two-dimensional gas chromatography 7

1.2 Chemometric data analysis 13

1.2.1 Data analysis challenges 13

1.2.2 Pre-processing methods 17

1.2.2.1 Baseline correction 17

1.2.2.2 Smoothing 19

1.2.2.3 Normalization 20

1.2.2.4 Retention time alignment 21

1.2.3 Deconvolution methods 23

1.2.3.1 Parallel factor analysis (PARAFAC) 23

1.2.3.2 Multivariate curve resolution-alternating least squares (MCR-ALS) 24

1.2.4 Pattern recognition methods 26

1.2.4.1 Principal component analysis (PCA) 26

1.2.4.2 Partial least squares (PLS) regression & discriminant analysis (PLS-DA) 27
1.2.4.3 Hierarchical clustering analysis (HCA) 28
1.2.4.4 *k*-means clustering 29
1.2.4.5 Fisher-ratio (F-ratio) analysis 30
1.3 Overview of following chapters.....32
1.3.1 Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis 32
1.3.2 Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry 33
1.3.3 Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis 34
1.3.4 Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry 35
1.3.5 Principal component analysis with comprehensive three-dimensional gas chromatography time-of-flight mass spectrometry data 36
1.4 References.....37
Chapter 2. Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis 46
2.1 Introduction.....46
2.2 Experimental.....50
2.2.1 Diesel Fuel Samples, and GC×GC-FID Instrument and Separation Conditions.....50
2.2.2 Data analysis.....52

2.3 Results and discussion 56
2.4 Conclusions 66
2.5 References 67
Chapter 3. Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry 70
3.1 Introduction 70
3.2 Experimental 74
3.3 Results and discussion 78
3.4 Conclusions 97
3.5 References 98
Chapter 4. Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis 103
4.1 Introduction 103
4.2 Materials and methods 105
4.2.1 Chemicals, samples and sample preparation 105
4.2.2 Instrumentation 107
4.2.3 Data analysis 108
4.3 Results and discussion 110
4.4 Conclusions 128
4.5 Acknowledgements 129
4.6 References 130

Chapter 5. Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry 136
5.1 Introduction 136
5.2 Experimental 140
5.3 Results and discussion 144
5.4 Conclusions 162
5.5 Acknowledgements 163
5.6 References 163
Chapter 6. Principal component analysis with comprehensive three-dimensional gas chromatography time-of-flight mass spectrometry data 169
6.1 Introduction 169
6.2 Experimental 172
6.3 Results and discussion 175
6.4 Conclusions 191
6.5 References 192
Chapter 7. Conclusions and future directions 196
7.1 Chapter 2 summary and future directions 196
7.2 Chapter 3 summary and future directions 197
7.3 Chapter 4 summary and future directions 198
7.4 Chapter 5 summary and future directions 199
7.5 Chapter 6 summary and future directions 201
Bibliography 202
Appendix A 223

Appendix B 231

Appendix C 240

Appendix D 250

Appendix E 257

LIST OF FIGURES

Fig. 1.1. Illustration of peak width metrics on a representative Gaussian peak 4

Fig. 1.2. Demonstration of R_s values from 0.5 – 1.5.....5

Figure 1.3. Illustration of modulation process of GC×GC using an 18 s window of a jet fuel sample chromatogram.....10

Figure 1.4. Illustration of peak capacity advantage of GC×GC relative to 1D-GC using a jet fuel sample chromatogram.....11

Figure 1.5. Possible dimensionalities of 1D-GC and GC×GC data.....16

Figure 2.1. Schematic of GC×GC-FID instrument using high temperature diaphragm valve-based modulation 51

Figure 2.2. Illustration of binning experimental design.....55

Figure 2.3. GC×GC chromatogram examples of Fuel 1 binned to different 2D bin sizes 57

Figure 2.4. GC×GC chromatograms of the other four diesel fuels at the raw data pixel-level 58

Figure 2.5. GC×GC chromatograms of the remaining four diesel fuels at the SOP bin size 59

Figure 2.6. PCA results for pixel-level data.....60

Figure 2.7. PCA results for SOP-binned data 61

Figure 2.8. Heat map of DCS as a function of 2D bin size for each of the five fuel pairs studied 63

Figure 2.9. Summary of the optimal bin sizes obtained for each fuel pair 65

Figure 3.1. Total ion current (TIC) GC×GC-TOFMS chromatogram of 15-ppm spiked JP8 jet fuel 79

Figure 3.2. Illustration of the discovery challenge addressed by tile-based F-ratio analysis that leverages selective *m/z*.....80

Figure 3.3. Comparison of F-ratio selectivity obtained at 15 ppm, 3 ppm, and 1.5 ppm for 2,5-dimethylthiophene.....82

Figure 3.4. Plots of the logarithm of F-ratio calculated at each *m/z* for the three F-ratio comparisons 83

Figure 3.5. Summed ^{2}D peaks for the replicates of 2,5-dimethylthiophene 89

Figure 3.6. Summed ^{2}D peaks for the replicates of 1,4-oxathiane.....91

Figure 3.7. True positive hit numbers obtained using standard F-ratio methodology and the top F-ratio method relative to the analyte *LOD* and *LOQ* 96

Figure 3.8. True positive hit numbers obtained while varying the ^{2}D tile dimension and holding the ^{1}D tile dimension constant, relative to the analyte *LOD* and *LOQ* 96

Figure 4.1. Map displaying the regions of Sicily in which wines 1-5 were produced.....106

Figure 4.2. Total ion current (TIC) GC×GC chromatograms for wines 1-5, from 2 to 10 min (A-E) compared to a selected region from 5.5 to 6.5 min (F-J) 112

Figure 4.3. PCA scores plot of unfolded chromatograms.....115

Figure 4.4. Background-corrected summed ^{1}D and ^{2}D peaks for linalool ethyl ether at *m/z* 94 with tile dimensions indicated (A-B), and F-ratio distribution with location of linalool ethyl ether labeled (C) 115

Figure 4.5. Assessment of %RSD in peak areas of ChromaTOF Tile hits 119

Figure 4.6. One-way *ANOVA* and ROC curve results 123

Figure 4.7. PCA results using true positive hits.....125

Figure 4.8. Bar graphs displaying the averaged peak areas for three highly loaded hits and one lowly loaded hit.....127

Figure 5.1. Total ion current (TIC) GC×GC-TOFMS chromatogram for 30 ppm-spiked JP8 jet fuel and PCA scores plot of 30-ppm, 15-ppm, and neat chromatograms 144

Figure 5.2. Illustration of advantage of tiling method for non-targeted analysis of the 30-ppm spiked versus 15-ppm spiked versus neat JP8 jet fuel chromatograms 146

Figure 5.3. Results for tile-based VRI-USI analysis for the 30-ppm versus 15-ppm versus neat samples in Comparison (1) 148

Figure 5.4. Examination of the *k*-means clustering portion of VRI-USI analysis using 2-chloroethyl phenyl sulfide at its top RSD^2 *m/z* 123 150

Figure 5.5. Total ion current (TIC) GC×GC-TOFMS chromatograms for J1800A, JP4, and JP8 jet fuels used in the multi-fuel VRI-USI analysis 155

Figure 5.6. PCA results using the unfolded 30-ppm spiked and neat J1800A, JP4, and JP8 jet fuel chromatograms 155

Figure 5.7. S_{max} distributions for "ideal" fuel type clustering hits, "nearly ideal" fuel type clustering hits, and contradictory clustering hits 162

Figure 6.1. Schematic of GC^3-TOFMS instrumental platform 173

Figure 6.2. Illustration of 3D peak capacity ($n_{c,3D}$) achieved using a typical analyte in the "overall fuel sample", 1,1,3-trimethylcyclopentane 177

Figure 6.3. Overview of novel ^{2}D re-registration technique 178

Figure 6.4. Illustration of ^{3}D re-registration 183

Figure 6.5. Results of ^{2}D and ^{3}D re-registration procedures 184

Figure 6.6. Final 3D TIC chromatograms of jet fuels 186

Figure 6.7. Results of multi-fuel PCA 187

Figure 6.8. Results of pair-wise PCA using J1800A and JP7 190

LIST OF TABLES

Table 2.1. 110 bin sizes were studied, with every combination of the eleven ^{1}D bin sizes by each of the ten ^{2}D bin sizes 53

Table 2.2. The five fuel pairs for which DCS was calculated.....54

Table 3.1. Identity of sulfur-containing compounds in the equal-mass mixture.....76

Table 3.2. Hitlists generated from standard F-ratio methodology 86

Table 3.3. Hitlists generated when only the top F-ratio *m/z* is used to rank the hitlist 87

Table 3.4. Calculation of *LOD* and *LOQ* to 2 significant digits for the 14 sulfur-containing analyte compounds.....93

Table 4.1. Identities of the top 30 true positive hits identified following tile-based F-ratio.....121

Table 5.1. List of top 20 hits obtained from the 30 ppm vs. 15 ppm vs. neat VRI-USI comparison 153

Table 5.2. List of all index assignments at S_{max} which were indicative of fuel type clustering for Comparison (3) 157

Table 5.3. List of all index assignments at S_{max} which were not indicative of fuel type clustering for Comparison (3) 158

Table 6.1. Identities and retention times of straight chain/branched alkyl compounds identified in overall fuel sample.....179

Table 6.2. Identities and retention times of cycloalkyl compounds identified in overall fuel sample 180

Table 6.3. Identities and retention times of aromatic compounds identified in overall fuel sample 180

Table 6.4. DCS values for nearest neighbor fuel pairs in the multi-fuel PCA scores plot.....188

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Robert Synovec, for agreeing to let me join his group back in 2017. I never imagined I would grow to love my research so much, and I am just so grateful that I've had the opportunity to work on so many great projects. I'd also like to thank the senior members of my lab who were so helpful training me when I first joined the group – Dr. Sarah Prebihalo, Dr. Kelsey Berrier, and Dr. Derrick Gough. I would not have gotten to where I am today without all your help.

I never would have made it through my PhD program without the long-distance support of my family – my parents Ted and Debbie and my brother Brendan. Thank you to my parents for raising me with a strong work ethic. Mom, thanks for checking in every night to make sure I was home safe. Dad, thank you for always giving me level-headed advice, even when I didn't want to hear it. Thank you, Brendan, for always providing me with some much-needed comic relief with your SnapChats. I'm glad there's another scientist in the family for me to commiserate with. Thank you to my whole family for visiting me in Seattle when I got lonely.

My best friends Katie and Elena have always been an inspiration to me. They both got their Master's degrees when I was in graduate school, so they've always been such supportive sounding boards when graduate school got tough. Katie, thank you for visiting me in Seattle (almost) every Thanksgiving and sharing my love of cheesy wedding shows. Elena, thank you for introducing me to the one and only James Blunt and always keeping me laughing. You guys are the best friends I could ever ask for. Thanks for also loving true crime so I feel less weird.

Thank you to my friends from Towson – Madelyn, Bhavisha, Shaun, Mignon, Lea, Abbey, Jack, and Adam. Our Jackbox game nights during the pandemic were such a welcome

breath of socialization. Since my first year in Seattle, I like to think I've gotten much better about responding to the group chat! You are all honestly the funniest humans I have ever met, and I'm so glad we met in Douglass House back in 2013. I cannot wait until we are all back in the same state again.

And now to my friends in Seattle. First and foremost, I must thank Tammi for being such an awesome roommate our first two years in Seattle. I was honestly terrified of the city when we first moved here, but exploring Ballard and Downtown with you was way less scary. Thank you for being such a supportive homework buddy, fellow rom com fanatic, dog lover, and friend. You've accomplished so much and I'm so proud of you! To my lab mates, or as I like to call them, the "chromies" – Sonia, Caitlin, Tim, and Grant, and to the honorary chromies, Joe and Kristy. There is so much I could say about how much you all mean to me, but this dissertation would never end. Thank you for always pushing me to do better, and making me laugh, especially at myself. Thanks for coming to Pure Barre with me. Thank you for never saying no to goof off time at the office.

To my partner in life, Scott. I can't thank you enough for encouraging me to go to Seattle, even though we had just started dating. Thank you for saving up all your leave at work to come visit me. Thank you for staying up late to talk to me every night because of the time difference. Thank you for being such a great person to be with all the time. I love you and I can't wait until the next chapter of our lives together.

Finally, thank you to my beautiful dog, Angel. Thank you for being my companion for almost my entire tenure at UW. Seeing your smiling face and hearing your howls was always the highlight of my day. I'm so sorry that you lost your fight to cancer, but I'm proud of how hard you fought. You will forever be the goodest girl in my heart.

DEDICATION

To my beautiful pup Angel, for always putting a smile on my face

June 7, 2010 – January 31, 2022

RIP, sweet girl

Chapter 1. Introduction to Gas Chromatography, Multidimensional Gas Chromatography, and Chemometric Data Analysis

Some parts of this chapter have been reproduced from Paige E. Sudol, Karisa M. Pierce, Sarah E. Prebihalo, Kristen J. Skogerboe, Bob W. Wright, Robert E. Synovec, "Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review" *Analytica Chimica Acta* 1132 (2020) 157-186.

1.1 FUNDAMENTALS OF GAS CHROMATOGRAPHY

1.1.1 Principles

Chromatography can be broadly defined as a method for separating the components of a sample based on the distribution of chemical analytes between a mobile phase and a stationary phase [1]. The thermodynamics of this partitioning can be represented by the distribution coefficient, K_D,

$$K_D = \frac{[analyte]_{sp}}{[analyte]_{mp}} \quad (1.1)$$

wherein $[analyte]_{sp}$ and $[analyte]_{mp}$ represent the equilibrium concentrations of a given analyte in the stationary and mobile phases, respectively. Analytes with larger K_D values have a higher affinity for the stationary phase, while analytes with smaller K_D have a higher affinity for the mobile phase. Indeed K_D is a multi-faceted constant, as it depends on the analyte structure, the stationary and mobile phase compositions, and the temperature of the system [1,2]. Given that analyte structure is immutable, the goal of any chromatographic separation is thus to produce differences in K_D among the components of a sample, which is accomplished through optimization of the separation conditions. The specific nature of such optimization will depend on the chromatographic technique at hand (i.e., ion chromatography, thin-layer chromatography, liquid chromatography). This dissertation will focus on optimization of gas chromatographic separations, as said technique was used for all research presented herein, but it is important to

reiterate that the general tenants of optimization can be extended to other forms of chromatography.

Gas chromatography (GC) has evolved into a widely used multidisciplinary technique for the separation of volatile and semi-volatile analytes since its inception in 1952 [3]. In GC, mixtures of chemicals can be physically separated as the chemicals naturally partition between a gaseous mobile phase (*aka* carrier gas) and a thin layer of stationary phase that coats the inside of a silica capillary column, all of which is housed in a thermostatted oven. Hydrogen is typically the carrier gas of choice with a flame ionization detector (FID) due to its fast optimum linear flow velocity, whereas helium is used with mass spectrometry (MS) detection to allow efficient vacuum conditions to be achieved and for its inertness to minimize formation of ionization artifacts. Most GC stationary phases have a cross-linked dimethylsilicone (PDMS) backbone substituted with phenyl, cyanopropyl, or trifluropropyl groups, with typical column lengths (L) on the order of 15-60 m, stationary phase inner diameters (d_c) of 100-320 μm, and film thicknesses (d_f) of 0.1-1.5 μm. Polyethylene glycol (PEG) stationary phases are preferable for separating more polar functionalities such as aldehydes, esters, and alcohols, although they do have lower temperature limits than PDMS columns which must be considered [4,5].

For a homologous series of compounds, chemicals will elute in order of boiling point. However, for a non-homologous mixture of compounds (i.e., range of boiling points and varying chemical functional groups), the elution order will depend on the complex interrelationship between analyte boiling point and polarity. Unlike other chromatographic mobile phases, GC carrier gases are inert, hence K_D in GC separations is representative of a given analyte's affinity for *just* the stationary phase (Eq. (1.1)). Thus, for a given stationary phase, each analyte has a

unique retention factor, *k'*, that allows the analyst to relate retention time data to the thermodynamic relationships of the separation process. The *k'* is defined as,

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

where t_R is the retention time of an analyte, and t_0 is the dead time which is the time it takes a completely unretained analyte in the carrier gas stream to travel the length of the column. Analytes with small *k'* will have little affinity for the stationary phase and shorter t_R. Analytes with large *k'* spend more time retained by the stationary phase and have longer t_R [4].

Following separation, analytes are detected exhibiting an approximately Gaussian concentration distribution, defined as,

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}} \quad (1.3)$$

where μ is the mean of the distribution, σ is the standard deviation, and *x* is a given position along the Gaussian curve, with the distribution centered around $x = \mu$ (i.e., the peak height) [6]. A detector produces an electrical signal proportional to concentration, with the two most common detectors coupled to GC systems being the FID and MS. An FID is a highly sensitive univariate detector which is selective for hydrocarbons, with signal intensity increasing proportionally with the moles of carbon and hydrogen atoms and decreasing in the presence of heteroatoms (oxygen, sulfur, nitrogen) [7]. Conversely, MS is a universal multi-channel detector which measures signal abundance at multiple mass-to-charge ratios (*m/z*) following ionization and fragmentation of analyte molecules. The resulting mass spectrum per analyte effectively serves as a "fingerprint" for a given chemical compound; for more complex analytical scenarios, high resolution MS instruments enable the elucidation of structurally similar compounds with a mass accuracy < 5 ppm [8,9]. GC-MS chromatograms are thus two-dimensional in nature, with the resulting chromatographic profile and mass spectrum enabling the simultaneous

quantification and identification of unknown analytes, respectively. Using GC-FID, an analyst can inject standards and make identifications based on matching t_R. Given that the only requirement for GC-amenable samples is sufficient volatility and thermal stability, GC-FID and GC-MS have found widespread utility in several fields, including forensics [10–12], food chemistry [13–15], metabolomics [16–18], and petroleum-based fuels [19–21].

1.1.2 Figures of merit

The goal of any chromatographic separation is to maximize the chemical information which can be extracted. An important step to that end is the production of narrow peaks. According to the Gaussian peak model first presented in Eq. (1.3), the width-at-base (W_b) of a peak represents 95% of the total peak area, or 13.5% of H, and can be calculated as ±2σ from t_R, or $W_b = 4\sigma$ (t_R equal to μ in Eq. (1.3)). Note that the baseline is defined as t_R ±4σ. Another common measurement is the width-at-half-height ($W_{1/2}$) metric, which is equal to 2.35σ and can be converted to W_b by multiplying $W_{1/2}$ by a factor of 1.7. These relevant peak width metrics are illustrated on a representative Gaussian peak in Fig. 1.1.

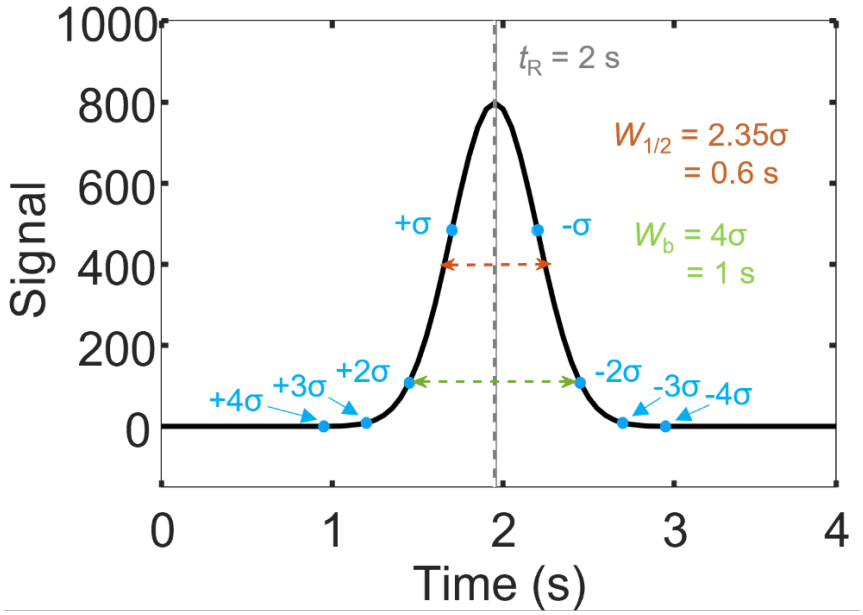

Fig. 1.1. Illustration of peak width metrics on a representative Gaussian peak, including W_b, $W_{1/2}$, and various multiples of σ from the t_R.

The degree of overlap between two adjacent Gaussian peaks in a chromatographic separation is assessed by calculating the resolution (R_S),

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

where $t_{R,1}$ and $t_{R,2}$ represent retention times and $W_{b,av}$ represents their average peak widths. A R_S = 1.5 is considered baseline-resolved, which is ideal for accurate identification and quantification. Unit resolution (R_s = 1) refers to when the time separation between adjacent peaks is equal to their average W_b; this level of resolution is much more easily obtainable in practice than R_s = 1.5, although it only enables accurate peak height determinations, not areas. At R_S <1, chemometric deconvolution software is needed to mathematically ascertain the presence of more than one chromatographic peak and accurately quantify said peaks, with two peaks becoming visually indistinguishable at a R_s ~ 0.5. Four values of R_s are illustrated in Fig. 1.2, with red and blue traces in Fig. 1.2(C-D) representing the peaks requiring deconvolution at R_s < 1.

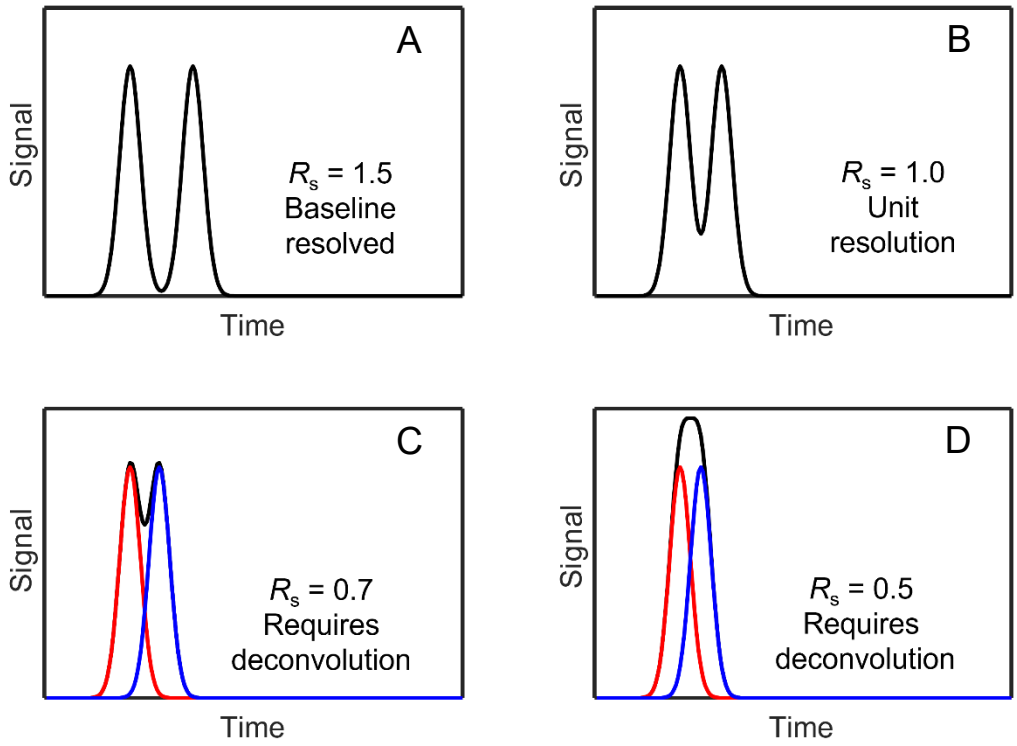

Fig. 1.2. Demonstration of R_s values equal to (A) 1.5, (B), 1, (C), 0.7, and (D) 0.5, calculated using Eq. (1.4).

Maximal resolution in minimal time saves resources and produces the greatest amount of information per unit time, a major goal in chromatographic applications. There are many experimental steps an analysts can take to maximize R_s between two peaks of interest, including decreasing the carrier gas flow rate or applying a temperature program (i.e., increasing the temperature by a constant rate over time, °C/min). Temperature programming mitigates the general elution problem (GEP) for GC separations, in which analytes with low *k'* have poor R_s and large *k'* produces excessive R_s, by ensuring that all analytes exit the column with a $k' \sim 0$ and approximately the same W_b. Thus, R_s is an ideal metric to guide targeted quantification exercises, which will be discussed in more depth later in this dissertation.

To assess the overall "performance" of a separation, one can calculate the peak capacity, n_c,

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

where n_c represents the maximum number of analyte peaks that can be ideally resolved, at unit resolution, during a given total separation run time, t_{sep}. Narrow peak widths produce a larger n_c as well as higher detection sensitivity, since peak height is inversely proportional to W_b. Overly wide peak widths will effectively waste separation time and space, producing smaller n_c and lower sensitivity, the latter of which will hamper detection of low concentration analytes. However, it is challenging to produce narrow peaks because the mass transfer kinetics during the separation process is hampered by the laminar flow of the carrier gas through the capillary GC column, stationary film thickness, and the overall kinetics of the analyte interacting with the stationary and mobile phases, all of which contribute to band broadening. To minimize peak band broadening, analysts can adjust various instrument parameters that include the carrier gas flow rate, temperature program rate, column dimensions, stationary phase composition and film thickness.

With regards to varying the column dimensions and stationary phase composition, shorter columns with narrower d_c and thinner d_f produce the fastest separations but are more likely to suffer from decreased chromatographic resolution, hence an analyst must find an optimum balance between separation time and resolution for a given application. To that end, the phase volume ratio (β) can be tuned to control the temperature at which analytes elute from a GC capillary column and minimize band broadening [22–25],

$$\beta = \frac{d_c}{4d_f} \tag{1.6}$$

where d_c is the column inner diameter and d_f is the stationary phase film thickness. At relatively low β (narrow d_c and thick d_f), analytes are retained on the column longer and elute at relatively hotter temperatures, which lends itself to short separation run times with sufficient chromatographic resolution and minimal band broadening relative to use of a high β (wide d_c and thin d_f) [22]. Other sources of band broadening are often referred to as "extra-column" band broadening since they are due to processes not associated with the on-column separation process. These include the injection technique and dead volumes in the inlet and detector [4]. Good instrumentation implementation practices can often mitigate extra-column band broadening.

1.1.3 Introduction to comprehensive two-dimensional gas chromatography

One-dimensional gas chromatography (1D-GC) is an invaluable interdisciplinary technique for elucidating the composition of sufficiently volatile samples, especially when coupled to a multi-channel detector like MS. Although various avenues can be explored with regards to maximizing chromatographic resolution in 1D-GC chromatograms, as was briefly discussed in the previous section, these efforts often become futile with increasingly complex samples. A notable example is petroleum-based fuels, which contain upwards of hundreds to thousands of analytes with similar boiling points *and* chemical functional groups. As a result,

1D-GC-MS chromatograms of gasoline and diesel fuels usually exhibit a large number of unresolved peaks observed as a non-flat baseline, referred to as an unresolved complex mixture (UCM) in the literature. Typically, only a small number of analyte peaks in said fuel sample chromatograms can be resolved, which leaves limited opportunity for analyte identification and quantification.

Numerous instrumental advancements have been made over the past few decades to overcome the limited resolving power (and hence limited n_c) of 1D-GC, which can be collectively referred to as multidimensional gas chromatography (MDGC). The initial form of MDGC was heart-cutting, or GC-GC, in which a selected fraction of the first-dimension (^{1}D) column effluent is transferred to a second-dimension (^{2}D) column for additional chromatographic separation of the unresolved components [4,26,27]. Column effluent transfer in GC-GC is generally accomplished using a simple flow switching device [27,28]. But GC-GC is inherently limited in scope, give that it is only designed to provide improved resolving power for a targeted number of analytes, not the entire sample mixture. The latter goal is accomplished using comprehensive two-dimensional gas chromatography (GC×GC), which was developed in 1991 by Liu and Phillips and is undoubtedly the most notable and highly influential form of MDGC to date [29].

In GC×GC, a capillary column of typical length (usually ~20 to 30 m), the ^{1}D column, is connected in series to a shorter ^{2}D capillary column (~1-5 m) of differing polarity by a modulator, which essentially acts as a secondary injector. Briefly, modulators can be broadly classified as thermal or flow-based, whereby thermal modulators utilize alternating cold and hot temperatures to trap and reinject ^{1}D effluent onto ^{2}D and flow-based modulators use directed carrier gas flow to transfer portions of the analyte stream to ^{2}D. The choice of modulator depends

on the specific experimental design, so we direct the reader to several excellent resources regarding GC×GC modulator development [30,31]. Following an initial ^{1}D separation and reinjection onto the ^{2}D column, the analyte mixture will undergo a series of very rapid secondary separations on ^{2}D and, ideally, analytes which were overlapped on ^{1}D will be resolved by the ^{2}D stationary phase. To fully take advantage of the enhanced resolving power of GC×GC, the orthogonality (i.e., opposing polarities) of the ^{1}D and ^{2}D column stationary phases should be maximized [32]. Typically, the ^{1}D column is non-polar (i.e., RTX-5 with 5% surface phenyl groups) and the ^{2}D column is more polar (i.e., polyethylene glycol or ionic liquid), but success has also been achieved using a reverse column configuration (polar ^{1}D column and non-polar ^{2}D column), particularly for separating alkanes and cycloalkanes in petrochemical samples [33,34]. More information on GC×GC method optimization will be provided later in this section.

An illustration of the modulation process is provided on a portion of a typical jet fuel sample chromatogram in Fig. 1.3. In Fig. 1.3(A), the ^{2}D separation dimension has been mathematically summed away to demonstrate what the total ion current (TIC) sample chromatogram would look like collected with 1D-GC-MS. In this view, only one peak is visible. The unfolded, modulated GC×GC-MS TIC chromatogram of the same time interval is provided in Fig. 1.3(B), with the dashed red lines indicating the modulation period of 3 s. Note that the overall Gaussian profile of the one peak observed in Fig. 1.3(A) is preserved in Fig. 1.3(B) across all six modulations. However, looking at just the first modulation (132 s to 135 s), two modulated peaks with unique ^{2}D retention times ($^{2}t_{R}$) are visible, indicating there are in fact *two analytes* which were originally overlapped on ^{1}D (dark green and orange stars). Note that there is an interfering peak from a previous modulation labeled accordingly. Overall, five analytes in this time interval are resolved, with the modulated peaks corresponding to unique analytes labeled

(dark green, orange, light green, pink, and blue stars), hence underscoring the impressive resolving power of GC×GC.

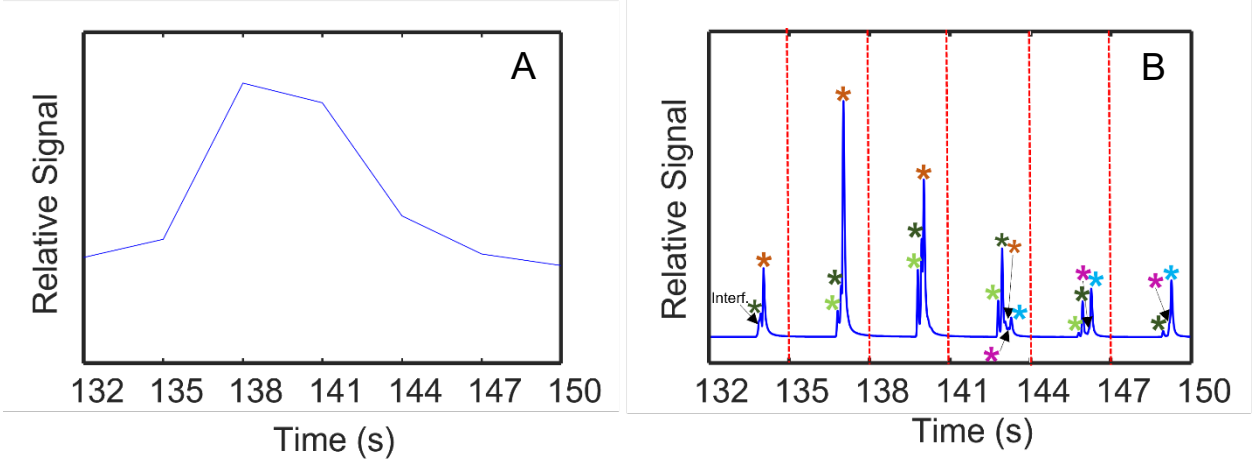

Figure 1.3. Illustration of modulation process of GC×GC using an 18 s window of a jet fuel sample chromatogram. (A) Unfolded TIC chromatogram with ^{2}D mathematically summed away. (B) Unfolded GC×GC-MS TIC chromatogram, with the modulations indicated with red dashed lines and the modulated peaks corresponding to unique analytes color-coded.

According to Klee et al., given equal separation times, GC×GC can theoretically provide a 10-fold to 20-fold peak capacity enhancement relative to 1D-GC, thus improving identification and quantification results for complex samples [35]. An example of this impressive peak capacity enhancement is illustrated using an entire jet fuel sample chromatogram in Fig. 1.4. In Fig. 1.4(A), the ^{2}D of a GC×GC-MS total ion current (TIC) separation has been mathematically summed away to illustrate what a 1D-GC separation theoretically would like. The extensive peak overlap observed in Fig. 1.4(A), visualized as a characteristic “petroleum hump”, would almost certainly complicate identification and quantification efforts. Conversely, the folded GC×GC TIC

contour plot is provided in Fig. 1.4(B), whereby the unfolded chromatogram has been cut, folded, and stacked along the P_M (i.e., the red dashed lines in Fig. 1.3(B)). Each analyte is represented by an individual "contour" with signal intensity indicated by a traditional red-to-blue color scale. Compared to Fig. 1.4(A), hundreds of analytes are visible in Fig 1.4(B), with impressive usage of the 2D separation space. This jet fuel chromatogram was collected using a reverse column configuration, producing distinct compound class bands for the alkane (^2t_R from ~2 to 3 s), cycloalkane (^2t_R from ~1 to 2 s), and aromatic compounds (^2t_R from ~0 to 0.5 s) on ^{2}D.

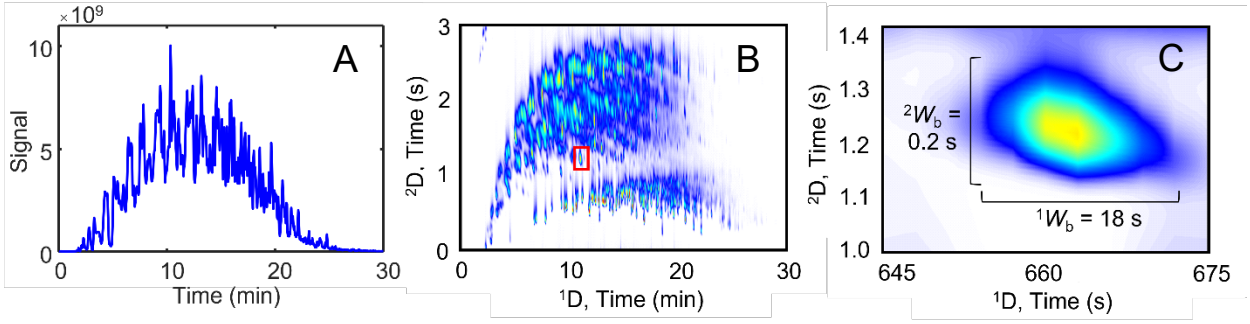

Figure 1.4. Illustration of peak capacity advantage of GC×GC relative to 1D-GC using a jet fuel sample chromatogram. (A) Unfolded TIC chromatogram with ^{2}D mathematically summed away. (B) Folded GC×GC-MS TIC contour plot, with a typical analyte enclosed in a red box. (C) Zoom-in of typical analyte labeled in (B), with the approximate ^1W_b and ^2W_b indicated accordingly.

More specifically, the two-dimensional (2D) peak capacity ($n_{c,2D}$) of a GC×GC separation can be calculated as follows,

$$n_{c,2D} = {}^1n_c \times {}^2n_c = \frac{{}^1t\,P_M}{{}^1W_b\,{}^2W_b} \tag{1.7}$$

where ^1n_c and ^2n_c represent the peak capacities of the individual ^{1}D and ^{2}D separations, respectively, calculated using Eq. (1.5). For the jet fuel separation shown in Fig. 1.4(B), a typical analyte in terms of peak widths along both dimensions (^1W_b and ^2W_b) is enclosed in a red box,

with a zoom-in of the analyte peak provided in Fig. 1.4(C). The ${}^1W_b$ and ${}^2W_b$ are calculated to be 18 s and 0.2 s, respectively, and are labeled accordingly in Fig. 1.4(C). Using a ^{1}D separation time (^1t) = 22.5 min (i.e., from 2.5 min to 25 min) and P_M = 3 s, the ${}^1n_c$ and ${}^2n_c$ are calculated to be 75 and 15 (Eq. (1.5)), respectively, resulting in $n_{c,2D}$ = 1,125 via Eq. (1.7). Thus, the 2D peak capacity produced in practice herein is in accordance with theory, with a particularly impressive ${}^2n_c$.

Given the multiplicative impact of ${}^1W_b$ and ${}^2W_b$ on 2D peak capacity (Eq. (1.7)), the importance of producing narrow peaks in GC×GC separations cannot be overstated. In optimizing a given GC×GC separation, analysts can vary the run time on ^{1}D column (^1t) and the modulation period (P_M) which controls the run time on the ^{2}D column to produce narrow W_b along both dimensions. However, when choosing a P_M, the analyst must consider the sampling density ρ_s (*aka* the modulation ratio, M_R),

$$\rho_s = \frac{{}^1W_b}{P_M} \tag{1.8}$$

which is equivalent to the number of times a given ^{1}D peak is modulated. In order for a GC×GC separation to be comprehensive, a ρ_s of 2 to 4 is considered sufficient [36–38]. A $\rho_s < 2$ is often considered under-sampled, as the ^{1}D peak is nearly unmodulated, and effectively broadened, and can negate much of the peak capacity advantage of GC×GC. Conversely, a $\rho_s > 4$ is considered oversampling, whereby the P_M could have been longer. Oversampling can potentially cause more retained compounds that will not elute during a shorter P_M to "wraparound" on ^{2}D, which may result in unnecessary peak overlap. Since the optimal P_M is related to the average ${}^1W_b$ and achieving a ρ_s of 2 to 4, the analyst will have to concurrently optimize both separation method conditions to minimize ${}^1W_b$ and ${}^2W_b$, including ^{1}D and ^{2}D flow rates, column lengths and internal

dimensions, and temperature programming rate, as was previously discussed for 1D-GC separations.

Just as the phase volume ratio β is highly useful for 1D-GC method optimization, the beta ratio metric, B_R, can be tuned to control the ^{2}D retention factor (^2k) range of analytes and improve $n_{c,2D}$,

$$\beta_R = \frac{^1\beta}{^2\beta} = \frac{^1d_c\ ^2d_f}{^2d_c\ ^1d_f} \tag{1.9}$$

where B_R is the ratio of ^{1}D β (^{1}β) to 2D β (2β), both calculated using Eq. (1.6). As ^1B decreases, analytes are retained longer on ^{1}D and elute at hotter temperatures (due to thicker ^1d_f relative to ^1d_c), so the resulting lower B_R produces a hotter, faster separation on ^{2}D and thus a lower starting ^2k. Conversely, when a larger ^1B is used, analytes are weakly retained on ^{1}D and elute at colder temperatures, which produces a colder, slower ^{2}D separation and a wider available ^2k range (larger B_R) [23–25]. Previous results have shown that by increasing ^2k, a larger B_R also improved the overall $n_{c,2D}$, although this was primarily accomplished by lowering ^1n_c and increasing ^2n_c [24]. Thus, depending on the analysis goals, B_R can be optimized to produce the desired ^1W_b and ^2W_b.

1.2 CHEMOMETRIC DATA ANALYSIS

1.2.1 Data analysis challenges

The traditional workflow for identification and quantification of analytes of interest in 1D-GC and GC×GC analyses involves manual mathematical determination of peak areas and/or peak volumes using the internal standard method, standard addition method, or external standard method [39–42]. Selective mass channels (*m/z*) can also be used to quantify individual analytes [33,42–44]. In fact, several commercial software platforms facilitate these calculations in a user-friendly interface [45–47]. Yet these routine data analysis methods are more ideal for application

to a limited number of known analytes, as application to *every single peak* in a GC×GC chromatogram (~hundreds to thousands of peaks) would be time consuming given the fact that one would have to identify the retention time boundaries along *both* dimensions. Additionally, although GC×GC chromatograms for a given sample exhibit considerably less peak overlap than 1D-GC chromatograms, especially complex samples such as petroleum-based fuels still exhibit a considerable number of co-elutions in GC×GC, even when using optimal stationary phases, flow rates, and temperature programs. Traditional instrument software platforms cannot produce accurate quantification results in the face of such extensive peak overlap. Off-line manual analysis of GC×GC chromatograms using platforms like MATLAB or Python is further complicated by the impressive data collection frequencies (up to *kHz* levels) of modern GC detector. Typical GC×GC-FID and GC×GC-TOFMS instruments produce chromatograms with upwards of millions of data points per sample run, which increases computation time, and lower signal-to-noise (*S/N*), which may complicate trace analyte identification and quantification. These effects are amplified when collecting numerous samples in a data collection campaign. Thus, powerful, automated data analysis methods (chemometrics) play an important role in computationally, rather than manually, elucidating the chemical information provided by large GC×GC datasets.

Herein, we broadly define chemometrics as the development and application of computational tools that apply linear algebra and statistical algorithms to optimally glean useful information from complex datasets obtained from chemical instrumentation measurements. Chemometric methods can be broadly categorized as signal pre-processing methods, peak deconvolution methods, and classification and pattern recognition methods. Although specific methods will be discussed in-depth in the remainder of this chapter, a brief overview will be

given herein. Signal-preprocessing methods are used to remove baseline fluctuations, retention time drift, and other analytical anomalies from chromatographic data prior to further chemometric analysis, so that ideally only true chemical differences between samples remain. Peak deconvolution methods and pattern recognition methods can be more generally classified as targeted and non-targeted, respectively, where targeted analysis focuses on identifying known analytes and non-targeted, or "discovery-based", analysis aims to uncover as much chemical information as possible. Discovery-based analyses can be further broken down into supervised and unsupervised methods, whereby sample class membership is known *a priori* in supervised studies but unknown in unsupervised studies. More details about supervised versus unsupervised analysis will be provided in section 1.2.4. More specifically, peak deconvolution methods are used to mathematically resolve physically overlapped analyte peaks, thus providing separate peak areas/volumes/spectra for each analyte. Pattern recognition methods can be used for classification and regression purposes; the goal of the former is to classify unknown samples by chemical composition, while the goal of the latter is to correlate chemical information with physical property information, or other known measurements.

Within each of these categories, the chemometric method-of-choice will depend on the dimensionality of the data. Although it is especially complex relative to its 1D counterpart, GC×GC data allows for more flexibility in terms of dimensionality, as the data can be unfolded (1D) or cut-and-folded along the modulation periods (2D). Coupling to MS detection adds an additional dimension to both 1D-GC-MS (2D data) and GC×GC-MS (3D data), and so does collection of multiple samples. Given that certain techniques require 2D or 3D data, GC×GC is amenable to a wider range of chemometric tools than 1D-GC. A summary of the different possible dimensionalities of 1D-GC and GC×GC data is provided in Fig. 1.5 [48,49].

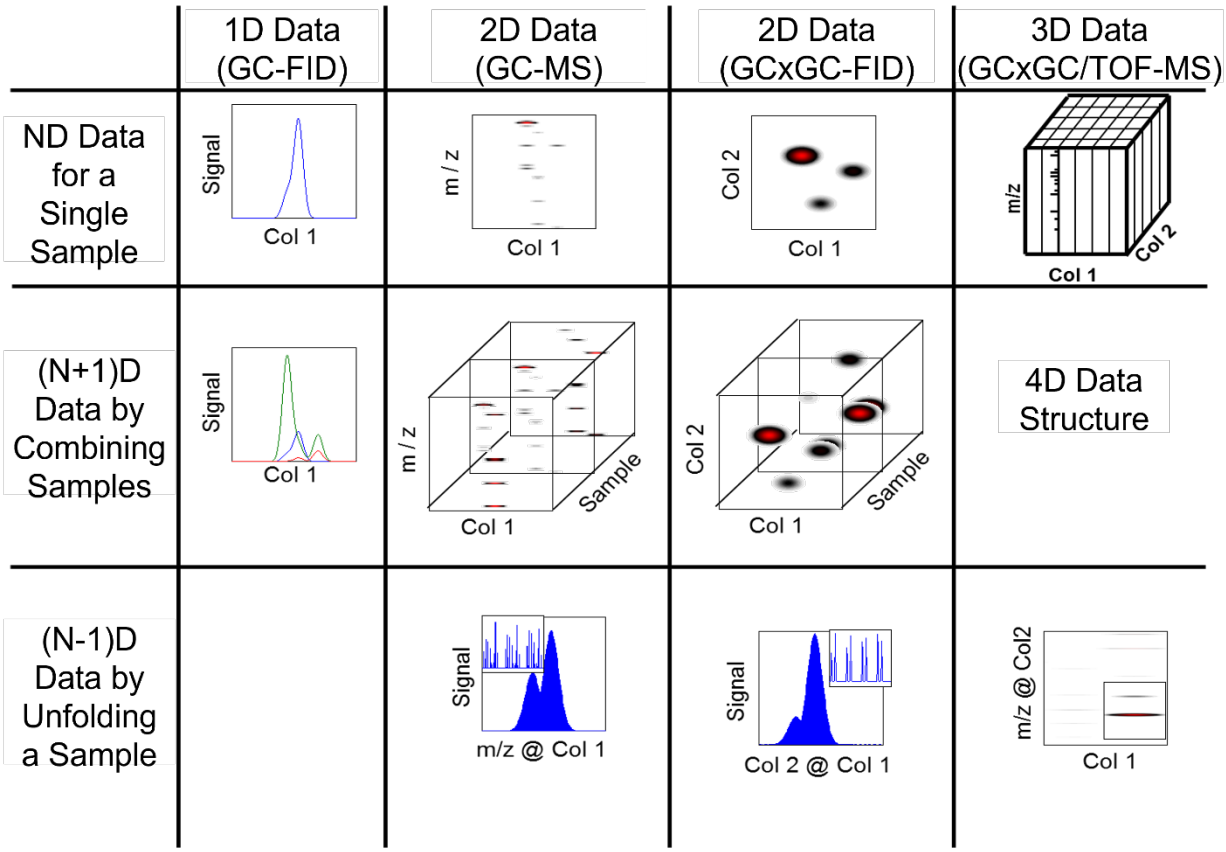

Figure 1.5. Possible dimensionalities of 1D-GC and GC×GC data [48,49].

1.2.2 Preprocessing methods

Principally, data preprocessing methods seek to "prepare" the dataset for a successful data analysis outcome. To some extent, data preprocessing methods are also used to minimize and/or remove some of the signal contributions from sources other than the analyte(s). There are four main preprocessing tools used in GC-based data analysis: baseline correction, smoothing, normalization, and retention time alignment.

1.2.2.1 Baseline correction

Baseline correction methods are designed to remove unwanted background contributions to signal due to low frequency detector noise, stationary phase degradation, and oven temperature fluctuations. Consideration must be given to the dimensionality of the data prior to baseline correction. GC-FID data is univariate and hence one can simply baseline correct the raw data vector. For GC-MS data, the analyst may choose to baseline correct the total ion current (TIC) chromatogram, i.e. the sum of all *m/z* signals, or baseline correct each individual *m/z* chromatogram in an iterative fashion. GC×GC data must be unfolded along the P_M intervals prior to baseline correction procedures. In GC×GC, consistent background signals (eg., column bleed along the ^{1}D separation) will be modulated like true analyte peaks, thus different contributions to baseline drift may appear to be separated chromatographically. Modulations of background signal contributions often are easily identifiable, and the analyst can remove these signals of known ^2t_R accordingly. Similar to GC-MS data, GC×GC-MS data can also be baseline corrected on a per *m/z* basis, which is useful for chemometric methods that rely upon all the *m/z* information, such as tile-based Fisher ratio and PARAFAC (more on these methods later).

The most basic baseline correction method is to acquire the signal of a reagent blank and subtract that blank pixel-level chromatogram from the sample chromatogram(s). Another

commonly employed strategy involves local linear fitting, in which a line is fit to each peak at its base and that line is subtracted from the peak signal [40,50]. However, local linear fitting requires knowledge of the precise retention time boundaries of each peak. Conversely, global baseline correction methods take advantage of the characteristic structures of GC chromatograms, in that the minimal signals are often taken to be noise regions between the more prominent eluting peaks, peaks are unimodal and Gaussian, noise is centered upon zero, peaks are positive, and negative peaks are absent. More specifically, a vector (or matrix) either (1) spanning approximately several peak widths or the entire length of the chromatogram, or (2) covering a user-defined percentage of the lowest signals, can be fit and subtracted [51,52]. Similarly, a baseline correction method for three-dimensional (3D) GC×GC-TOFMS data known as the "rolling-ball minimum method" assumes each second column separation will have a region of minimum signal in which no peaks elute, and that minimum value is subtracted from all signals in that vector representing the second column chromatogram [53]. This is algorithmically repeated across the entire GC×GC space for every *m/z* of the 3D chromatogram.

Although these baseline correction methods are highly useful for correcting minor baseline fluctuations, more severe baseline disturbances caused by numerous issues (i.e. a degrading stationary phase and detector noise working in concert) require advanced fitting procedures. Eilers published the penalized asymmetric least squares (PAsLS) baseline estimation method which avoids explicitly identifying regions of noise and instead uses a penalized-least-squares smoothing calculation with a user-controlled heavy smoothing parameter to fit a high-order polynomial to each data point [40,54,55]. The signal of each data point is replaced with the value provided by the polynomial fit, converging upon a best fit baseline. However, the parameters for PAsLS and other fitting procedures must be carefully optimized for a given

dataset to avoid overfitting the baseline, in which real chemical information will be included in the baseline and hence removed upon baseline subtraction [48].

1.2.2.2 Smoothing

Baseline correction procedures remove low frequency detector noise, but high frequency random noise can still interfere with successful chemometric analysis. In chromatographic data, high frequency random noise is observed due to uncontrollable fluctuations in the detector response [48]. Algorithmically smoothing chromatograms reduces such noise variations and simultaneously increases the signal-to-noise ratio (*S/N*), which is an obvious advantage for chemometrics. The classic Savitzky-Golay method is a smoothing function that fits a low order polynomial through the chromatographic signal of a small region and replaces the original value in the middle of the interval with the fitted value [56]. The interval of the small region is then shifted over one data point and the procedure is repeated with the same interval size until all original signals are replaced with new fitted values. The PAsLS algorithm described above was originally developed by Eilers as a smoothing algorithm to reduce high frequency noise and improve upon the Savitzky-Golay method [57,58]. Another running smoother that is much simpler is the moving average approach wherein the signal of each data point is replaced with the average signal of a constant number of data points surrounding the center data point. This running average smoother is simple and fast, but it risks distorting narrow peak signals if all values have the same weight when calculating the average. Therefore, smoothing functions that use polynomial fitting instead of a simple running average are preferred for chromatographic data. Regardless of the smoothing method chosen, excessive smoothing can distort peak shapes and remove small semi-overlapped peaks entirely, so rigorous optimization of the smoothing parameters is critical.

1.2.2.3 Normalization

When analyte concentration in principle should be equal either (1) between replicates of the same sample or (2) between different samples with similar compositions, normalization methods aim to remove erroneous analyte signal variations, which can be attributed to injection volume or sample preparation discrepancies. The most widely utilized normalization method in chromatography is the internal standard method, in which a non-native analyte is added at a known, constant concentration to all samples and standards, generally added prior to sample preparation steps [59–62]. The internal standard method assumes all analytes can be corrected by the same internal standard. If that assumption is not sufficiently valid, multiple internal standards may be useful. The internal standard method also requires that the non-native chemical must be inert and physically or mathematically resolvable from all other sample components. These prerequisites can make it difficult to choose a good internal standard for truly unknown samples. As an alternative, isotopically labeled internal standards may be useful, taking advantage of the isotope ratio method. However, the use of isotopic standards requires MS detection.

The sum-normalization method, also known as total area-normalization, is a viable alternative when it is not practical to use an internal standard. This method uses the total sum (area) of all baseline corrected signal in a chromatogram as the normalization factor, the value by which all given signals are divided [63]. Sum-normalization methods do not strictly preserve quantitative information as well as the internal standard method does, but they do preserve relative signal information for qualitative comparisons [64,65]. For all sum-normalization approaches, the analyst is relying upon the assumption that all samples in the dataset contain sufficiently similar components with similar response factors, such that equal volumes of the samples should have produced approximately equal total signal. While this assumption is reasonable for homogeneous datasets, for datasets that do not truly meet this criteria, sum-

normalization can introduce additional problematic signal variations [66]. The probabilistic quotient normalization method (PQN) was developed to address the problem of large concentration variations between samples outweighing "size effect" variations [66]. PQN calculates a most probable normalization factor by assessing the distribution of quotients of analyte signals in a test chromatogram relative to a reference chromatogram. In summary, careful consideration of the nature of the dataset and analysis goals is particularly important when selecting a normalization method.

1.2.2.4 Retention Time Alignment

Baseline-corrected and normalized chromatograms are often still hindered by run-to-run retention time shifting, in which the same analyte appears at a different location in subsequent chromatograms. Such shifting can be caused by variations in column flow, fluctuations in pressure and temperature, clipping the column during routine maintenance, or a degraded stationary phase, and is an obvious complication to chemometric data analysis. Retention time variations can be manually corrected at the peak-level by manually matching peaks in peak tables, or by application of retention indexing (RI) with standards as facilitated by instrument software implementation. However, the analyst must match peaks by making hundreds of subjective decisions, which is a very labor-intensive process. Alternatively, entire chromatograms or chromatographic regions can sometimes be automatically aligned at the raw data pixel-level using retention time correction algorithms. One of the most widely utilized algorithms for aligning entire chromatograms is correlation optimized warping (COW), a local linear alignment algorithm which linearly stretches and compresses local regions of a sample chromatogram to maximize the correlation with a target chromatogram [67]. The analyst selects the size of the local regions and the maximum amount of data points that each region is allowed

to be warped. Eilers similarly reported the parametric time warping algorithm (PTW) which explicitly models a warping function for a chromatographic data vector to match a target vector based on a parametric model [58]. PTW re-interpolates the entire data vector, yielding an aligned sample profile at the pixel-level. Another approach is rank-based alignment algorithms, which determine optimal corrections for selected regions of chromatograms rather than aligning entire chromatograms. The analyst estimates the rank of the subsection and the algorithm performs singular value decomposition (SVD) at each attempted retention time correction of a chromatographic region, which is concatenated with the same region from a target chromatogram. The correction that minimizes the sum of the noise eigenvectors and also produces the lowest number of relevant eigenvalues is the correction that is chosen as optimal [68,69].

Although automatic alignment algorithms should in principle preserve peak signal areas, there is always a risk of introducing distortions in peak shapes and areas. To avoid this potential shortcoming, a novel "tiling" approach was developed by our group to replace alignment within discovery-based sample class comparisons of GC×GC-TOFMS chromatograms [53,70]. In this approach, the sample class chromatograms are tiled (carefully binned and temporarily reduced in resolution), and correlations along with other metrics are calculated between all tiles of the sample class chromatograms to meet user-input criteria. The tiles are carefully allowed to be large enough to span ranges that encompass possible retention time shifting but small enough to avoid over-sampling neighboring peaks. The original raw data does not need to be re-interpolated to create an aligned chromatogram so peak shapes are not distorted by warping. However, a "hit list" of analyte specific information for peaks that are matched between the two or more sample classes is created at low resolution using a Fisher ratio statistical approach. The

hit list is checked for redundancy (split peaks) at the original high resolution to remove redundant listings. The chromatograms are not explicitly aligned and re-interpolated, but the information necessary for comparative analysis of samples with accurately matched peaks is produced.

1.2.3 Deconvolution Methods

1.2.3.1 Parallel Factor Analysis (PARAFAC)

PARAFAC is a peak deconvolution method that is based on alternating least-squares decomposition. It requires higher-order data (third-order or greater) and can mathematically resolve a sufficiently trilinear structure by modeling each component as the outer product of its three fixed vectors. PARAFAC decomposes the selected chromatographic region into *n* total components, given as the sum of the number of analyte, baseline and noise component contributions [30]. Each of the *n* total components is resolved into individual vectors describing the ^{1}D separation, ^{2}D separation, and mass spectral profiles, respectively. Not only does PARAFAC enable individual analyte identification and quantification, but it also allows an analyst to identify baseline contributions (as a separate component).

PARAFAC performs well for GC×GC-TOFMS data under the following requirements: the data array must be sufficiently trilinear (i.e. minimal retention time ($^{2}t_{R}$) shifting on ^{2}D so that individual analyte modulations are well-aligned), and it must contain chemically selective information (i.e. chromatographic peaks or *m/z* fragments unique to an analyte) for each component on two of the three dimensions [23,71–73]. Typically, only a small region containing the analyte peak or peaks of interest is selected for PARAFAC. In the absence of *a priori* knowledge about the number of components present, objective determination of such can be achieved by building successive models with increasing factors and automatically detecting the

number of factors at which overfitting, or splitting of analytes between multiple components, occurs [72]. Optional constraints such as nonnegative signals and unimodality in peak shape are often reasonably acceptable assumptions for GC×GC data that can be input into a PARAFAC algorithm to improve the efficiency of PARAFAC model convergence [30,74].

Following the determination of a successful PARAFAC model, the loadings can be further analyzed for analyte identification and quantification. For instance, the mass spectrum obtained via PARAFAC can be submitted to a reference library search for increased confidence in analyte identification. More recently, PARAFAC2 has been introduced as an advanced version of PARAFAC that is more equipped to handle retention time misalignment as it loosens the trilinearity requirement [30,75,76].

1.2.3.2 Multivariate curve resolution with alternating least squares (MCR-ALS)

MCR-ALS is a peak deconvolution method that decomposes chromatographic two-way arrays into the product of two matrices containing purified component information for each dimension [30,71,76]. For GC-MS, one dimension of information is the chromatographic contribution, while the second dimension of information is the mass spectral contribution. For GC×GC-MS, one dimension of information is the unfolded GC×GC chromatographic data contribution, while the second dimension of information is the mass spectral contribution. MCR-ALS operates on several assumptions, two of which are that the observed signal profile of a sample mixture is a linear combination of the signal profiles of the individual components in the mixture, and that chemically irrelevant sources of variation (noise) in observed signals are minimized. MCR-ALS only requires bilinear data in comparison with the trilinear requirement of PARAFAC. While MCR-ALS can be applied to data regardless of deviations in trilinearity such as retention time shifting on the ^{2}D time dimension, if there are multiple samples and retention

time shifting on ^{2}D as well as on ^{1}D between chromatograms, then the misalignment must be addressed prior to MCR-ALS. However, if multichannel detection is used (e.g., MS), MCR-ALS can handle both types of retention time shifting as the spectral mode provides reproducible responses that maintain data bilinearity [71,77].

The ways in which MCR-ALS can be applied are dependent on the type of data used. Prior to MCR-ALS decomposition, three-way arrays with univariate detection are often transformed back into two-way arrays based on the common information shared by the datasets. This is commonly achieved through column-wise augmentation of multiple samples [30]. In the case of multichannel detection, such as a GC×GC-TOFMS chromatogram, the three-way arrays must be first unfolded into a two-way array before MCR-ALS can be applied. This is most often accomplished by unfolding along the time dimension or analyzing modulations of the ^{1}D separation. If MCR-ALS is to be applied to multiple GC×GC-TOFMS chromatograms, a common pre-processing workflow is to unfold the ^{1}D time dimension (the modulations are augmented) and then the samples are concatenated column-wise, with *m/z* in the columns and with time in the rows [78].

Briefly, MCR-ALS is applied in the following way. First, constraints specified are applied (e.g. unimodality, nonnegativity, or local rank), and the algorithm makes an initial estimate of the individual components (chromatographic or mass spectral) in the model. After testing for convergence, an iterative process is applied until convergence criteria (i.e., minimum value of residuals, lack-of-fit improvement, etc.) are met. As in the case of PARAFAC, once a successful model has been obtained, the mass spectral components can be used for mass spectral library searching and/or accurate quantification.

1.2.4 Pattern recognition methods

1.2.4.1 Principal component analysis (PCA)

PCA is an unsupervised pattern recognition technique which finds linear combinations of latent variables referred to as principal components (PCs) in a bilinear 2D data matrix that succinctly model the reproducible variations and correlations (e.g. covariance) in that data matrix [79,80]. Given the requirement for 2D data, chromatograms for multiple samples must be concatenated row-wise and GC×GC chromatograms must be unfolded into vectors prior to PCA, so each column of the resulting matrix represents the signal intensity at a pixel-level chromatographic location. Multiway PCA (MPCA), also known as Tucker modeling, is also available for analysis of three-way data arrays [81]. To be bilinear, the 2D data matrix must be composed of unique, consistent, concentration-dependent signals from each independent and unique source present, and those signals must be additive. PCA produces a decomposition model that contains the scores and loadings, with the goal to define a high degree of the chemical variance (i.e., chemical information) in the data [79]. The scores contain information about how the sample classes relate to each other, whereas the loadings contain information about how the variables (peaks and corresponding t_R) relate to each other across individual samples. Conceptually, if two samples have perfectly matching chromatograms, they will have equal scores. If sample classes are very different from each other, then they will form distinct clusters on a scores plot, which is most commonly a plot of PC2 scores on the y-axis versus PC1 scores on the x-axis. The dissimilarity between two sample classes can be assessed using the degree-of-class separation (DCS) metric, which is based on the Euclidean distances between and within sample classes in a scores plot [82–84].

PC1 and PC2 capture the most variance within a given dataset, and this variance can be either sample-related (i.e., desired chemical information) or non-sample-related (i.e., instrument noise, injection variation, etc.). Assessing whether or not this variance is sample related requires examination of the loadings. For different sample classes, the variables (i.e., retention time pixel-level data locations corresponding to analyte peaks) that vary the most between them will have the highest loadings values. For a given PC axis, positive loadings correspond to variables characteristic of samples with positive scores, whereas negative loadings correspond to variables characteristic of samples with negative scores.

1.2.4.2. Partial least squares (PLS) regression & discriminant analysis (PLS-DA)

Partial least squares analysis (PLS) is a supervised multivariate calibration method that allows the analyst to model and predict quantitative information about samples using their chromatographic chemical fingerprints. PLS models the relationship of maximum covariance between chromatograms and the known quantitative information for each chromatogram, such as physical property data. PLS constructs a linear regression model by selecting latent variables for which variation in the chromatograms (chemical composition differences) is most predictive of physical property differences, hence the covariance between the two is maximized. Once the relationship between the independent variables and the dependent variables is modeled, the dependent variable value for unknown samples can be predicted through regression [85–87].

Like PCA, PLS requires input of a 2D data matrix, with individual chromatograms concatenated row-wise so that the columns represent pixel-level locations corresponding to analyte peaks. The quantitative information must be input as a separate vector, whereby the information in each row correlates to the rows (i.e. samples) of the 2D data matrix. The chromatograms in the 2D data matrix must be composed of unique, consistent, concentration-

dependent signals at the pixel-level, and those signals must be independently additive. PLS provides a calibration plot of the predicted physical property (y-axis) as a function of the known physical property (x-axis), ideally with a slope of unity. Leave-one-out cross validation is almost always applied with PLS, so for *n* samples, *n-1* of the samples are used to calibrate, and then the remaining n^{th} sample is treated as the "unknown" so is predicated in a round robin fashion until all *n* samples have been predicted. Along with the calibration (and prediction) plot, PLS also provides the linear regression vectors (LVRs) that provide insight into the chromatographic information that was leveraged in the analysis [198]. Plots of the LVRs reveal highly loaded chromatographic variables that are positively or negatively correlated to the predicted physical property values.

PLS coupled to discriminant analysis (PLS-DA) is another variant of PLS regression that is commonly employed [88–91]. Like PLS regression, the goal of PLS-DA is to maximize the covariance between chromatographic data and the properties to be modeled, and analysis of a calibration dataset yields a predictive model, the accuracy and precision of which is evaluated with a test dataset. Additionally, sample classification is achieved via setting a threshold for the predicted property values [89]. Similar to how sample classes are visualized in PCA scores plots, sample classes can be visualized in PLS-DA by plotting the predicted values for each sample class versus one another. Unlike PCA, PLS-DA is a supervised technique and thus often shows improved clustering of samples which were difficult to distinguish with PCA. This added visualization represents the primary advantage of PLS-DA as opposed to simple PLS regression.

1.2.4.3 Hierarchical clustering analysis (HCA)

Hierarchical cluster analysis (HCA) is a well-known unsupervised pattern recognition method, being a simpler version of PCA. Briefly, HCA plots 1D data vectors (unfolded

chromatograms) in Euclidean space creating a dendrogram (without providing loadings like PCA) wherein similar chromatograms are closer together in that space in "nested clusters" and dissimilar chromatograms are more distant from each other [92,93]. The HCA algorithm can be performed in either an agglomerative or divisive manner; in the former, each data point is considered a single cluster and then combined with similar data points, whereas in the latter, all data points are originally in a single cluster and then split up based on their calculated similarity [94]. A common approach for unknown samples is to regress said samples onto an HCA model built using training set samples with known class memberships. The unknown samples can then be considered most similar to the nearest nested cluster. Although HCA is a great visual tool for small datasets, dendrograms become increasingly difficult to interpret as the number of data points (i.e. samples) increases, which makes it difficult to use HCA for strict classification exercises [94,95].

1.2.4.4 k-means clustering

k-means clustering is one of the most widely utilized partitioning clustering methods. Unlike hierarchical methods, partitioning methods assign all data points to their corresponding clusters simultaneously using a defined similarity metric and the number of expected clusters, *k*. More specifically, *k*-means clustering divides a defined number of *x* measurements into *k* clusters by minimizing the sum of squared errors between the cluster centroids and corresponding cluster members [94–97]. The output of *k*-means clustering is an index assignment for each data point indicating its cluster membership. Using a user-selected distance metric, the algorithm computes *k* number of initial cluster centroids, assigns data points to their closest clusters, and then recomputes the cluster centroids until the sum of squared errors is minimized [98]. The number of expected clusters *k* must be specified in *k*-means, which does add a slight amount of

supervision to a mostly “unsupervised” method. However, to utilize the algorithm when the expected *k* is unknown, one can test a range of *k* values and determine the optimum value of *k* using well-defined metrics. The most applied *k*-means metric in the literature is the silhouette value, which is a measure of within-cluster similarity relative to between-cluster dissimilarity for a given data point [97,99,100]. The silhouette value will be discussed in more depth in Chapter 5 of this dissertation.

Like other unsupervised methods such as PCA and PLS, *k*-means clustering requires input of a 2D matrix with the samples concatenated row-wise. For unfolded chromatograms, the columns would represent signal at pixel-level chromatographic locations. Although *k*-means clustering can be applied to raw data of various types [101–103], clustering performance will suffer in the presence of extraneous, non-chemical related variation. Thus, a common approach in the literature is to first input data to PCA and then perform *k*-means clustering on the resulting scores plot. Given that PCA is a data reduction step, this is a highly robust approach, with PCA coupled with *k*-means clustering being utilized in wide-ranging applications such as early diagnosis of diabetes [104], genomic ancestry determination [105], and prediction of dissolved oxygen content in water [106].

1.2.4.5 Fisher ratio (F-ratio) analysis

Fisher ratio (F-ratio) analysis is a supervised method for discovery of class distinguishing characteristics in samples for experimental designs in which there is *a priori* knowledge about class membership. The F-ratio approach is an analysis of variance (ANOVA) method that prioritizes statistical significance that is essentially normalized to the absolute signal [30,53,70,107–109]. Mathematically, the F-ratio is defined as the class-to-class variance divided by the sum of the within-class variance. As the class-to-class variance increases relative to the

within-class variance, the F-ratio increases. A high F-ratio is indicative of a peak location in the chromatographic data in which the analyte in one sample class is likely to have a statistically significant difference in concentration relative to the other sample class. Thus, the F-ratio analysis (an ANOVA method) is complementary to the student's t-test metric, which compares the means of concentration for a given analyte hit between two groups to confirm the statistical difference of the two means (true positive), or not confirm (false positive). The F-ratio results can be summarized in a table, or "hit list", with the most class-distinguishing features at the top.

As in other pattern recognition methods, F-ratio analysis can be sensitive to retention time shifting. While retention time alignment has been demonstrated to be a beneficial pre-processing tool for 1D-GC, it is often not desirable in F-ratio analysis due to the artifacts that can be introduced during the alignment process or ineffectiveness of alignment for GC×GC-TOFMS data. Therefore, the impact of retention time misalignment on F-ratio analysis has been minimized most successfully by a tile-based preprocessing approach. Tile-based F-ratio analysis allows for more retention time misalignment between samples by utilizing a novel tiling (smart binning) scheme. In contrast to peak-table based F-ratio, tile-based F-ratio analysis is performed on raw data prior to peak decomposition, analyte identification and quantification [53,70,107,108]. Briefly, the tile-based approach blurs retention time shifting with each tile consisting of the GC×GC raw pixel-level data summed to a user-defined tile size (typically the size of ^{1}D and ^{2}D peak widths, plus a couple extra data point pixels to encompass retention time shifting). Four tile schemes are utilized, and the tile which includes the most class-distinguishing information is saved. Once F-ratios have been calculated for the four tiling schemes, a pinning and clustering algorithm relates the calculated F-ratios back to the original raw pixel-level data resolution.

1.3 OVERVIEW OF FOLLOWING CHAPTERS

1.3.1 Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis

Principal component analysis (PCA) is a widely applied chemometric tool for classifying samples using comprehensive two-dimensional (2D) gas chromatography (GC×GC) separation data. Classification via PCA can be improved by 2D binning of the data. A "standard operating procedure (SOP) bin size" is often applied to improve the *S/N* and to mitigate potential retention time misalignment issues. The SOP bin size is generally selected to be slightly larger than the typical 2D peak dimensions. In this study we examine to what extent a single SOP bin size is optimal for all of the class comparisons that can be made in a single PCA scores plot. For this purpose, a GC×GC-FID dataset comprised of 5 different diesel fuels (i.e., 5 sample classes), each run with 4 replicates using a reverse column configuration (polar ^{1}D column and non-polar ^{2}D column) was utilized. The dataset was collected within about one day, which minimized retention time misalignment in order to allow the study to focus on *S/N* enhancement concurrent with maintaining the chemical selectivity provided by the GC×GC separations. A total of 110 bin sizes were evaluated. Degree-of-class separation (DCS) was utilized as a quantitative metric to assess the impact of binning in improving separation in the scores plot. The DCS was calculated pair-wise between nearest neighbor sample classes for each of the 5 sample classes in the scores plot (5 sample class pairs). Results indicated the SOP bin size did not provide the highest DCS for any of the 5 fuel pairs. Each fuel pair is found to have its own optimal bin size, suggesting the binning finds the balance between *S/N* optimization concurrent with leveraging the chemical selectivity information differences in the samples as manifested in their GC×GC

separation “patterns”. Robustness of the findings in this study were supported by leaving out one fuel at a time and re-running the PCA models.

1.3.2 Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry

Comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC×GC-TOFMS) is followed by tile-based Fisher ratio (F-ratio) analysis to investigate the “limit of discovery” for low concentration levels of sulfur-containing compounds in JP8 jet fuel. A mixture of 14 sulfur-containing compounds was spiked at 30 ppm, 15 ppm, 3 ppm and 1.5 ppm into the neat fuel prior to GC×GC-TOFMS analysis with a “reversed” column format (*aka* polar first dimension (^{1}D) and non-polar second dimension (^{2}D) column). Prior standard implementation of tile-based F-ratio analysis utilized an average F-ratio requiring a minimum of 3 mass channels (*m/z*) with the highest F-ratios. Herein, we explore the notion that use of the top F-ratio *m/z* for hitlist ranking is superior to the standard implementation for analytes near their limit-of-quantitation (*LOQ*), defined as an analyte concentration that produces a signal equal to ten times the standard deviation of the baseline noise ($10\sigma_n$). Hitlist ranking comparisons revealed that using only the top F-ratio *m/z* resulted in impressive improvements in discoverability for the low concentration comparisons. Specifically, for the 3 ppm versus neat hitlist, 1,4-oxathiane (*LOQ* = 2.5 ppm) improved from hit 114 via standard F-ratio analysis, to hit 25. For the 1.5 ppm versus neat hitlist, 2-propylthiophene (*LOQ* = 0.64 ppm) improved from hit 59 to 17, benzo[b]thiophene (*LOQ* = 1.1 ppm) from hit 98 to 28, and 2,5-dimethylthiophene (*LOQ* = 1.3 ppm) from hit 262 to 39. Additional hitlist ranking comparisons revealed the importance of proper tile size selection, as analyte discoverability deteriorated upon using either an inappropriately too small or too large of a tile.

1.3.3 Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis

The volatile fraction of food, also called the food volatilome, is increasingly used to develop new fingerprinting approaches. The characterization of the food volatilome is important to achieve desired flavor profiles in food production processes, or to differentiate different products, with winemaking being one popular area of interest. In the present research, headspace solid-phase microextraction (HS SPME) coupled to flow-modulated comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (FM GC×GC-TOFMS) was used to characterize geographical-based differences in the volatilome of five white "Grillo" wines (of Sicilian origin), comprising the five sample classes. All wines were produced with the same vinification method in 2019. To minimize the influence of minor bottle-to-bottle differences, three bottles of the same wine were randomly selected, and three samples were collected per bottle, resulting in nine sample replicates per wine. Particular emphasis was devoted to the operational conditions of a novel low duty cycle flow modulator. A fast FM GC×GC-TOFMS method with a modulation period of 700 ms and a re-injection period of 80 ms was developed. Following, the instrumental software was exploited to identify class-distinguishing analytes in the dataset *via* tile-based Fisher ratio analysis (i.e., ChromaTOF Tile). A tile size of 10 modulations (7 s) on the first dimension and 45 spectra (300 ms) on the second dimension was used to encompass average peak widths and to account for minor retention time shifting. Off-line software was used to apply an *ANOVA* test. A *p*-value of 0.01 was applied in order to select the most important class-distinguishing analytes, which were input to principal component analysis (PCA). The PCA scores plot showed distinct clustering of the wines according to geographical origin, although the loadings

revealed that only a few analytes were necessary to differentiate the wines. However, a comprehensive flavor profile assessment underscored the importance of all the information output by the ChromaTOF Tile software.

1.3.4 Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry

Tile-based variance rank initiated-unsupervised sample indexing (VRI-USI) analysis is introduced for comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOFMS). VRI-USI analysis addresses the challenge that irrelevant variables can often obscure true chemical variation when using other unsupervised chemometric tools. Implementation of VRI-USI analysis with GC×GC-TOFMS data incorporates the tile-based Fisher ratio (F-ratio) analysis software platform that mitigates the effects of retention shifting in both separation dimensions with an unsupervised variance metric (instead of the F-ratio metric) as the initial step of ranking the hitlist. Next, implementation of *k*-means clustering, *k*, per hit using the silhouette metric, S_{max}, is used to reveal to what extent recurring indexed sample clusters are uncovered. Finally, based upon a probability-based evaluation of how the individual samples cluster throughout the hitlist an unsupervised class membership is revealed. For a JP8 jet fuel dataset spiked with a sulfur-containing analyte mix at 30-ppm, 15-ppm, and neat, clustering by spike level at $k = 3$ was the most commonly re-occurring set of index assignments, occurring for 11 out of 14 spiked analytes. Upon application of these *k*-means index assignments to the entire hitlist, all 14 spiked hits had one way *ANOVA* *p*-values < 0.05, validating the presumption of classes. Next, application of VRI-USI to a 3-ppm spiked versus neat JP8 jet fuel comparison exhibited similar performance to F-ratio analysis for analyte discovery. In the last study, for a dataset of J1800A, JP4, and JP8 jet fuel, each spiked with the

sulfur-containing analyte mix at 30-ppm and neat, 453 out of 520 hits in the hitlist exhibited index assignments indicative of fuel type clustering. Scrutinization of the remaining 67 hits revealed nine hits with sample index assignments contradictory to the fuel type clustering, eight of which were spiked sulfur-containing analytes. Interestingly, these hits also had a large S_{max} indicative of a true sub-cluster. Thus, tile-based VRI-USI analysis appears to be a promising tool for unsupervised multi-class classification studies using GC×GC-TOFMS data.

1.3.5 Principal component analysis with comprehensive three-dimensional gas chromatography time-of-flight mass spectrometry data

Comprehensive three-dimensional (3D) gas chromatography (GC^3) coupled to time-of-flight-mass spectrometry (GC^3-TOFMS) is beginning to emerge as an intriguing augmentation of the well-established comprehensive two-dimensional gas chromatography (GC×GC) technique. Although impressive gains have been made in the instrumentation realm, the utility of non-targeted chemometric analysis of GC^3-TOFMS data has yet to be explored. Herein, we present the first application of principal component analysis (PCA) to elucidate a complex GC^3-TOFMS dataset of jet fuel samples. Five replicates each of four jet fuels (JP8, J1800A, JP4, and JP7) were collected by GC^3-TOFMS with commercial thermal modulation from the first-dimension column (^{1}D) to the second-dimension column (^{2}D), and dynamic pressure gradient modulation (DPGM) from the ^{2}D column to the third-dimension column (^{3}D), thus providing full transfer (100% duty cycle both modulation stages). A novel re-registration technique is introduced in which a linear series of vacant ^{3}D modulations are removed to correct ^{2}D shifting and effectively "center" the data. This shifting has consistently been observed in the ^{2}D versus ^{1}D view of GC^3-TOFMS chromatograms, likely due to the slowed flow on ^{2}D in DPGM and/or temperature programming effects. The 3D PCA loadings of the re-registered data revealed fine chemical

differences between the fuels which would not be as easily elucidated using PCA of GC×GC data. Finally, PCA of the two most chemically similar fuels (J1800A and JP7) revealed additional chemical differences which were drowned out in the initial multi-fuel PCA model, highlighting the potential advantage of "pairwise" PCA for multi-class GC^3-TOFMS datasets.

1.4 REFERENCES

[1] K. Robards, P.R. Haddad, P.E. Jackson, Princ. Pract. Mod. Chromatogr. Methods, Academic Press, San Diego, 1994.
[2] J.C. Giddings, Unified Separation Science, John Wiley & Sons, Inc., 1991.
[3] A.T. JAMES, A.J. MARTIN, Gas-liquid partition chromatography; the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid., Biochem. J. 50 (1952) 679–690. https://doi.org/10.1042/bj0500679.
[4] K. Robards, P.R. Haddad, P.E. Jackson, Gas Chromatography, in: Princ. Pract. Mod. Chromatogr. Methods, Academic Press, San Diego, 1994: pp. 117–120.
[5] C.F. Poole, S.K. Poole, Separation characteristics of wall-coated open-tubular columns for gas chromatography, J. Chromatogr. A 1184 (2008) 254–280. https://doi.org/10.1016/j.chroma.2007.07.028.
[6] D.C. Harris, Chapter 4: Statistics, in: Quant. Chem. Anal., Ninth, W. H. Freeman and Company, New York, NY, 2016: pp. 64–89.
[7] C.F. Poole, Ionization-based detectors for gas chromatography, J. Chromatogr. A 1421 (2015) 137–153. https://doi.org/10.1016/j.chroma.2015.02.061.
[8] R. Li, Y. Liu, Z. Wang, Q. Zhang, H. Bai, Q. Lv, High resolution GC–Orbitrap MS for nitrosamines analysis: Method performance, exploration of solid phase extraction regularity, and screening of children's products, Microchem. J. 162 (2021) 105878. https://doi.org/10.1016/j.microc.2020.105878.
[9] F. Rey-Stolle, D. Dudzik, C. Gonzalez-Riano, M. Fernández-García, V. Alonso-Herranz, D. Rojo, C. Barbas, A. García, Low and high resolution gas chromatography-mass spectrometry for untargeted metabolomics: A tutorial, Anal. Chim. Acta (2021) 339043. https://doi.org/10.1016/j.aca.2021.339043.
[10] A. Girod, C. Weyermann, Lipid composition of fingermark residue and donor classification using GC/MS, Forensic Sci. Int. 238 (2014) 68–82. https://doi.org/10.1016/j.forsciint.2014.02.020.
[11] N. Wiebelhaus, D. Hamblin, N.M. Kreitals, J.R. Almirall, Differentiation of marijuana headspace volatiles from other plants and hemp products using capillary microextraction of volatiles (CMV) coupled to gas-chromatography–mass spectrometry (GC–MS), Forensic Chem. 2 (2016) 1–8. https://doi.org/10.1016/j.forc.2016.08.004.
[12] A.L. Castro, S. Tarelho, P. Melo, J.M. Franco, A fast and reliable method for quantitation of THC and its 2 main metabolites in whole blood by GC–MS/MS (TQD), Forensic Sci. Int. 289 (2018) 344–351. https://doi.org/10.1016/j.forsciint.2018.06.003.
[13] M. Del Carlo, A. Pepe, G. Sacchetti, D. Compagnone, D. Mastrocola, A. Cichelli, Determination of phthalate esters in wine using solid-phase extraction and gas chromatography–

mass spectrometry, Food Chem. 111 (2008) 771–777. https://doi.org/10.1016/j.foodchem.2008.04.065.
[14] D. Giuffrida, M. Zoccali, L. Mondello, Recent developments in the carotenoid and carotenoid derivatives chromatography-mass spectrometry analysis in food matrices, TrAC Trends Anal. Chem. 132 (2020) 116047. https://doi.org/10.1016/j.trac.2020.116047.
[15] C.N. Cain, N.J. Haughn, H.J. Purcell, L.C. Marney, R.E. Synovec, C.T. Thoumsin, S.C. Jackels, K.J. Skogerboe, Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee, J. Agric. Food Chem. (2021). https://doi.org/10.1021/acs.jafc.1c00605.
[16] J.C.R. Demyttenaere, R.M. Moriña, P. Sandra, Monitoring and fast detection of mycotoxin-producing fungi based on headspace solid-phase microextraction and headspace sorptive extraction of the volatile metabolites, J. Chromatogr. A 985 (2003) 127–135. https://doi.org/10.1016/S0021-9673(02)01417-6.
[17] Q. Yang, L. Xu, L.-J. Tang, J.-T. Yang, B.-Q. Wu, N. Chen, J.-H. Jiang, R.-Q. Yu, Simultaneous detection of multiple inherited metabolic diseases using GC-MS urinary metabolomics by chemometrics multi-class classification strategies, Talanta 186 (2018) 489–496. https://doi.org/10.1016/j.talanta.2018.04.081.
[18] K. Berrou, C. Dunyach-Remy, J.-P. Lavigne, B. Roig, A. Cadiere, Multiple stir bar sorptive extraction combined with gas chromatography-mass spectrometry analysis for a tentative identification of bacterial volatile and/or semi-volatile metabolites, Talanta 195 (2019) 245–250. https://doi.org/10.1016/j.talanta.2018.11.042.
[19] C. Martín-Alberca, C. García-Ruiz, O. Delémont, Study of chemical modifications in acidified ignitable liquids analysed by GC–MS, Sci. Justice 55 (2015) 446–455. https://doi.org/10.1016/j.scijus.2015.06.006.
[20] M.E. Flood, M.P. Connolly, M.C. Comiskey, A.M. Hupp, Evaluation of single and multi-feedstock biodiesel – diesel blends using GCMS and chemometric methods, Fuel 186 (2016) 58–67. https://doi.org/10.1016/j.fuel.2016.08.069.
[21] L. Bai, J. Smuts, J. Schenk, J. Cochran, K.A. Schug, Comparison of GC-VUV, GC-FID, and comprehensive two-dimensional GC–MS for the characterization of weathered and unweathered diesel fuels, Fuel 214 (2018) 521–527. https://doi.org/10.1016/j.fuel.2017.11.053.
[22] Z.R. Roberson, J.V. Goodpaster, Preparation and characterization of micro-bore wall-coated open-tubular capillaries with low phase ratios for fast-gas chromatography–mass spectrometry: Application to ignitable liquids and fire debris, Sci. Justice (2019). https://doi.org/10.1016/j.scijus.2019.06.009.
[23] S.E. Prebihalo, D.K. Pinkerton, R.E. Synovec, Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on data trilinearity and parallel factor analysis deconvolution, J. Chromatogr. A 1605 (2019) 460368. https://doi.org/10.1016/j.chroma.2019.460368.
[24] B.A. Parsons, D.K. Pinkerton, R.E. Synovec, Implications of phase ratio for maximizing peak capacity in comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1536 (2018) 16–26. https://doi.org/JChrom.
[25] D.V. Gough, H.D. Bahaghighat, R.E. Synovec, Column selection approach to achieve a high peak capacity in comprehensive three-dimensional gas chromatography, Talanta 195 (2019) 822–829. https://doi.org/10.1016/j.talanta.2018.12.007.
[26] G. Schomburg, F. Weeke, F. Müller, M. Oreans, Multidimensional gas chromatography (MDC) in capillary columns using double oven instruments and a newly designed coupling piece

for monitoring detection after pre-separation, Chromatographia 16 (1982) 87–91. https://doi.org/10.1007/BF02258875.
[27] K.M. Sharif, S.-T. Chin, C. Kulsing, P.J. Marriott, The microfluidic Deans switch: 50 years of progress, innovation and application, TrAC Trends Anal. Chem. 82 (2016) 35–54. https://doi.org/10.1016/j.trac.2016.05.005.
[28] D.R. Deans, A new technique for heart cutting in gas chromatography, Chromatographia 1 (1968) 18–22. https://doi.org/10.1007/BF02259005.
[29] Z. Liu, J.B. Phillips, Comprehensive Two-Dimensional Gas Chromatography using an On-Column Thermal Modulator Interface, J. Chromatogr. Sci. 29 (1991) 227–231. https://doi.org/10.1093/chromsci/29.6.227.
[30] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications, Anal. Chem. 90 (2018) 505–532. https://doi.org/10.1021/acs.analchem.7b04226.
[31] H.D. Bahaghighat, C.E. Freye, R.E. Synovec, Recent advances in modulator technology for comprehensive two dimensional gas chromatography, TrAC Trends Anal. Chem. (2018). https://doi.org/10.1016/j.trac.2018.04.016.
[32] Zaiyou. Liu, D.G. Patterson, M.L. Lee, Geometric Approach to Factor Analysis for the Estimation of Orthogonality and Practical Peak Capacity in Comprehensive Two-Dimensional Separations, Anal. Chem. 67 (1995) 3840–3845. https://doi.org/10.1021/ac00117a004.
[33] M. Jennerwein, M. Eschner, T. Wilharm, T. Gröger, R. Zimmermann, Evaluation of reversed phase versus normal phase column combination for the quantitative analysis of common commercial available middle distillates using GC × GC-TOFMS and Visual Basic Script, Fuel 235 (2019) 336–338. https://doi.org/10.1016/j.fuel.2018.07.081.
[34] P. Vozka, G. Kilaz, How to obtain a detailed chemical composition for middle distillates via GC × GC-FID without the need of GC × GC-TOF/MS, Fuel 247 (2019) 368–377. https://doi.org/10.1016/j.fuel.2019.03.009.
[35] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain, J. Chromatogr. A 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.
[36] L.M. Blumberg, Accumulating resampling (modulation) in comprehensive two-dimensional capillary GC (GC×GC), J. Sep. Sci. 31 (2008) 3358–3365. https://doi.org/10.1002/jssc.200800424.
[37] W. Khummueng, J. Harynuk, P.J. Marriott, Modulation Ratio in Comprehensive Two-dimensional Gas Chromatography, Anal. Chem. 78 (2006) 4578–4587. https://doi.org/10.1021/ac052270b.
[38] W.C. Siegler, B.D. Fitz, J.C. Hoggard, R.E. Synovec, Experimental Study of the Quantitative Precision for Valve-Based Comprehensive Two-Dimensional Gas Chromatography, Anal. Chem. 83 (2011) 5190–5196. https://doi.org/10.1021/ac200302b.
[39] 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. https://doi.org/10.1021/ef501247h.
[40] S. Samanipour, P. Dimitriou-Christidis, J. Gros, A. Grange, 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, J. Chromatogr. A 1375 (2015) 123–139. https://doi.org/10.1016/j.chroma.2014.11.049.
[41] M.N. Dunkle, P. Pijcke, B. Winniford, G. Bellos, Quantification of the composition of liquid hydrocarbon streams: Comparing the GC-VUV to DHA and GCxGC, J. Chromatogr. A 1587 (2019) 239–246. https://doi.org/10.1016/j.chroma.2018.12.026.
[42] D. França, V.B. Pereira, D.M. Coutinho, L.M. Ainstein, D.A. Azevedo, Speciation and quantification of high molecular weight paraffins in Brazilian whole crude oils using high-temperature comprehensive two-dimensional gas chromatography, Fuel 234 (2018) 1154–1164. https://doi.org/10.1016/j.fuel.2018.07.145.
[43] T.N. Loegel, R.E. Morris, I. Leska, Detection and Quantification of Metal Deactivator Additive in Jet and Diesel Fuel by Liquid Chromatography, Energy Fuels 31 (2017) 3629–3634. https://doi.org/10.1021/acs.energyfuels.6b03128.
[44] R.L. Webster, P.M. Rawson, D.J. Evans, P.J. Marriott, Quantification of trace fatty acid methyl esters in diesel fuel by using multidimensional gas chromatography with electron and chemical ionization mass spectrometry, J. Sep. Sci. 39 (2016) 2537–2543. https://doi.org/10.1002/jssc.201600307.
[45] ChromaTOF® Software, LECO Corp. (2022). https://www.leco.com/product/chromatof-software (accessed January 17, 2022).
[46] Chromatography Method Development - OpenLab ChemStation | Agilent, (2022). https://www.agilent.com/en/product/software-informatics/analytical-software-suite/chromatography-data-systems/openlab-chemstation (accessed January 17, 2022).
[47] ChromCompare+, (2022). https://www.sepsolve.com/chromcompare/ (accessed November 16, 2021).
[48] K.M. Pierce, J.S. Nadeau, R.E. Synovec, Chapter 17 - Data Analysis Methods, in: C.F. Poole (Ed.), Gas Chromatogr., Elsevier, Amsterdam, 2012: pp. 415–434. https://doi.org/10.1016/B978-0-12-385540-4.00017-1.
[49] P.E. Sudol, K.M. Pierce, S.E. Prebihalo, K.J. Skogerboe, B.W. Wright, R.E. Synovec, Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review, Anal. Chim. Acta 1132 (2020) 157–186. https://doi.org/10.1016/j.aca.2020.07.027.
[50] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics, Chem. Rev. 110 (2010) 4582–4605. https://doi.org/10.1021/cr900394n.
[51] G.L. Alexandrino, J. Malmborg, F. Augusto, J.H. Christensen, Investigating weathering in light diesel oils using comprehensive two-dimensional gas chromatography–High resolution mass spectrometry and pixel-based analysis: Possibilities and limitations, J. Chromatogr. A 1591 (2019) 155–161. https://doi.org/10.1016/j.chroma.2019.01.042.
[52] K.M. Pierce, S.P. Schale, Predicting percent composition of blends of biodiesel and conventional diesel using gas chromatography–mass spectrometry, comprehensive two-dimensional gas chromatography–mass spectrometry, and partial least squares analysis, Talanta 83 (2011) 1254–1259. https://doi.org/10.1016/j.talanta.2010.07.084.
[53] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[54] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis, J. Chromatogr. A 996 (2003) 141–155. https://doi.org/10.1016/S0021-9673(03)00616-2.
[55] K.M. Pierce, B.W. Wright, R.E. Synovec, Unsupervised parameter optimization for automated retention time alignment of severely shifted gas chromatographic data using the piecewise alignment algorithm, J. Chromatogr. A 1141 (2007) 106–116. https://doi.org/10.1016/j.chroma.2006.11.101.
[56] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639.
[57] P.H.C. Eilers, A Perfect Smoother, Anal. Chem. 75 (2003) 3631–3636. https://doi.org/10.1021/ac034173t.
[58] P.H.C. Eilers, Parametric Time Warping, Anal. Chem. 76 (2004) 404–411. https://doi.org/10.1021/ac034800e.
[59] J.M. Hogan, R.A. Engel, H.F. Stevenson, Versatile internal standard technique for the gas chromatographic determination of water in liquids, Anal. Chem. 42 (1970) 249–252. https://doi.org/10.1021/ac60284a033.
[60] R.J. Strife, J.R. Simms, M.P. Lacey, Combined capillary gas chromatography/ion trap mass spectrometry quantitative methods using labeled or unlabeled internal standards, J. Am. Soc. Mass Spectrom. 1 (1990) 265–271. https://doi.org/10.1016/1044-0305(90)85044-M.
[61] R.H. Liu, G. Foster, E.J. Cone, S.D. Kumar, Selecting an appropriate isotopic internal standard for gas chromatography/mass spectrometry analysis of drugs of abuse--pentobarbital example, J. Forensic Sci. 40 (1995) 983–989.
[62] M. Fasciotti, T.V.C. Monteiro, A.A. Ferreira, M.N. Eberlin, L.A. Neves, Two-point normalization using internal and external standards for a traceable determination of δ13C values of fatty acid methyl esters by gas chromatography/combustion/isotope ratio mass spectrometry, Int. J. Mass Spectrom. 418 (2017) 41–50. https://doi.org/10.1016/j.ijms.2016.12.002.
[63] J.W. McIlroy, R.W. Smith, V.L. McGuffin, Assessing the effect of data pretreatment procedures for principal components analysis of chromatographic data, Forensic Sci. Int. 257 (2015) 1–12. https://doi.org/10.1016/j.forsciint.2015.07.038.
[64] J. Orzel, B. Krakowska, I. Stanimirova, M. Daszykowski, Detecting chemical markers to uncover counterfeit rebated excise duty diesel oil, Talanta 204 (2019) 229–237. https://doi.org/10.1016/j.talanta.2019.05.113.
[65] J.M. Baerncopf, V.L. McGuffin, R.W. Smith, Association of Ignitable Liquid Residues to Neat Ignitable Liquids in the Presence of Matrix Interferences Using Chemometric Procedures*,†, J. Forensic Sci. 56 (2011) 70–81. https://doi.org/10.1111/j.1556-4029.2010.01563.x.
[66] P. Filzmoser, B. Walczak, What can go wrong at the data normalization step for identification of biomarkers, J. Chromatogr. A 1362 (2014) 194–205. https://doi.org/10.1016/j.chroma.2014.08.050.
[67] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, 805 (1998) 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1.
[68] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Standardization of Second-Order Chromatographic/Spectroscopic Data for Optimum Chemical Analysis, Anal. Chem. 70 (1998) 218–225. https://doi.org/10.1021/ac9706335.

[69] C.G. Fraga, Chemometric approach for the resolution and quantification of unresolved peaks in gas chromatography–selected-ion mass spectrometry data, J. Chromatogr. A 1019 (2003) 31–42. https://doi.org/10.1016/S0021-9673(03)01329-3.
[70] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.
[71] H. Parastar, J.R. Radović, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and Quantification of Complex Mixtures of Polycyclic Aromatic Hydrocarbons in Heavy Fuel Oil Sample by Means of GC × GC-TOFMS Combined to Multivariate Curve Resolution, Anal. Chem. 83 (2011) 9289–9297. https://doi.org/10.1021/ac201799r.
[72] J.C. Hoggard, R.E. Synovec, Parallel Factor Analysis (PARAFAC) of Target Analytes in GC × GC−TOFMS Data: Automated Selection of a Model with an Appropriate Number of Factors, Anal. Chem. 79 (2007) 1611–1619. https://doi.org/10.1021/ac061710b.
[73] D.K. Pinkerton, B.A. Parsons, T.J. Anderson, R.E. Synovec, Trilinearity deviation ratio: A new metric for chemometric analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data, Anal. Chim. Acta 871 (2015) 66–76. https://doi.org/10.1016/j.aca.2015.02.040.
[74] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.
[75] L.A.F. de Godoy, E.C. Ferreira, M.P. Pedroso, C.H. de V. Fidélis, F. Augusto, R.J. Poppi, Quantification of Kerosene in Gasoline by Comprehensive Two-Dimensional Gas Chromatography and N-Way Multivariate Analysis, Anal. Lett. 41 (2008) 1603–1614. https://doi.org/10.1080/00032710802122222.
[76] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemom. Intell. Lab. Syst. 117 (2012) 80–91. https://doi.org/10.1016/j.chemolab.2012.02.003.
[77] H. Parastar, J.R. Radović, J.M. Bayona, R. Tauler, Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares, Anal. Bioanal. Chem. 405 (2013) 6235–6249. https://doi.org/10.1007/s00216-013-7067-y.
[78] L.W. Hantao, H.G. Aleme, M.P. Pedroso, G.P. Sabin, R.J. Poppi, F. Augusto, Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: A review, Anal. Chim. Acta 731 (2012) 11–23. https://doi.org/10.1016/j.aca.2012.04.003.
[79] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.
[80] D. Ballabio, A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure, Chemom. Intell. Lab. Syst. 149 (2015) 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003.
[81] N.G.S. Mogollón, F.A.L. Ribeiro, R.J. Poppi, A.L. Quintana, J.A.G. Chávez, D.A.P. Agualongo, H.G. Aleme, F. Augusto, N.G.S. Mogollón, F.A.L. Ribeiro, R.J. Poppi, A.L. Quintana, J.A.G. Chávez, D.A.P. Agualongo, H.G. Aleme, F. Augusto, Exploratory Analysis of Biodiesel by Combining Comprehensive Two-Dimensional Gas Chromatography and Multiway

Principal Component Analysis, J. Braz. Chem. Soc. 28 (2017) 740–746. https://doi.org/10.21577/0103-5053.20160222.
[82] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, J. Chromatogr. A 1096 (2005) 101–110. https://doi.org/10.1016/j.chroma.2005.04.078.
[83] 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. https://doi.org/10.1016/j.chroma.2011.10.031.
[84] P.E. Sudol, D.V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis, Talanta 206 (2020) 120239. https://doi.org/10.1016/j.talanta.2019.120239.
[85] K.M. Pierce, B.A. Parsons, R.E. Synovec, Chapter 10 - Pixel-Level Data Analysis Methods for Comprehensive Two-Dimensional Chromatography, in: A.M. de la Peña, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Eds.), Data Handl. Sci. Technol., Elsevier, 2015: pp. 427–463. https://doi.org/10.1016/B978-0-444-63527-3.00010-2.
[86] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202. https://doi.org/10.1021/ac00162a020.
[87] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.
[88] I. Barra, M.A. Mansouri, Y. Cherrah, M. Kharbach, A. Bouklouze, FTIR fingerprints associated to a PLS-DA model for rapid detection of smuggled non-compliant diesel marketed in Morocco, Vib. Spectrosc. 101 (2019) 40–45. https://doi.org/10.1016/j.vibspec.2019.02.001.
[89] L. Ranzan, C. Ranzan, L.F. Trierweiler, J.O. Trierweiler, Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy, Energy Fuels 31 (2017) 8942–8950. https://doi.org/10.1021/acs.energyfuels.7b00954.
[90] B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, Anal. Bioanal. Chem. 407 (2015) 1159–1170. https://doi.org/10.1007/s00216-014-8332-4.
[91] F. Mabood, S.A. Gilani, M. Albroumi, S. Alameri, M.M.O. Al Nabhani, F. Jabeen, J. Hussain, A. Al-Harrasi, R. Boqué, S. Farooq, A.M. Hamaed, Z. Naureen, A. Khan, Z. Hussain, Detection and estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with multivariate methods, Fuel 197 (2017) 388–396. https://doi.org/10.1016/j.fuel.2017.02.041.
[92] W.N.S. Mat-Desa, D. Ismail, N. NicDaeid, Classification and Source Determination of Medium Petroleum Distillates by Chemometric and Artificial Neural Networks: A Self Organizing Feature Approach, Anal. Chem. 83 (2011) 7745–7754. https://doi.org/10.1021/ac202315y.
[93] M. Novák, D. Palya, Z. Bodai, Z. Nyiri, N. Magyar, J. Kovács, Z. Eke, Combined cluster and discriminant analysis: An efficient chemometric approach in diesel fuel characterization, Forensic Sci. Int. 270 (2017) 61–69. https://doi.org/10.1016/j.forsciint.2016.11.025.
[94] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

[95] P. Govender, V. Sivakumar, Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), Atmospheric Pollut. Res. 11 (2020) 40–56. https://doi.org/10.1016/j.apr.2019.09.009.
[96] J.A. Hartigan, M.A. Wong, Algorithm AS 136: A K-Means Clustering Algorithm, J. R. Stat. Soc. Ser. C Appl. Stat. 28 (1979) 100–108. https://doi.org/10.2307/2346830.
[97] R. Lletí, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes, Anal. Chim. Acta 515 (2004) 87–100. https://doi.org/10.1016/j.aca.2003.12.020.
[98] H.-J. Chu, C.-J. Liau, C.-H. Lin, B.-S. Su, Integration of fuzzy cluster analysis and kernel density estimation for tracking typhoon trajectories in the Taiwan region, Expert Syst. Appl. 39 (2012) 9451–9457. https://doi.org/10.1016/j.eswa.2012.02.114.
[99] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 2009.
[100] A. Naghizadeh, D.N. Metaxas, Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means, Procedia Comput. Sci. 176 (2020) 205–214. https://doi.org/10.1016/j.procs.2020.08.022.
[101] H. Ji, H. Lu, Z. Zhang, Pure ion chromatogram extraction via optimal k -means clustering, RSC Adv. 6 (2016) 56977–56985. https://doi.org/10.1039/C6RA08409E.
[102] T.R. Noviandy, A. Maulana, N.R. Sasmita, R. Suhendra, Muslem, G.M. Idroes, M. Paristiowati, Z. Helwani, E. Yandri, S. Rahimah, Muhammad, Irvanizam, R. Idroes, The implementation of K-Means clustering in kovats retention index on gas chromatography, IOP Conf. Ser. Mater. Sci. Eng. 1087 (2021) 012051. https://doi.org/10.1088/1757-899X/1087/1/012051.
[103] A. Sancho, J.C. Ribeiro, M.S. Reis, F.G. Martins, Cluster analysis of crude oils with k-means based on their physicochemical properties, Comput. Chem. Eng. 157 (2022) 107633. https://doi.org/10.1016/j.compchemeng.2021.107633.
[104] C. Zhu, C.U. Idemudia, W. Feng, Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, Inform. Med. Unlocked. 17 (2019) 100179. https://doi.org/10.1016/j.imu.2019.100179.
[105] A. Budiarto, B. Mahesworo, J. Baurley, T. Suparyanto, B. Pardamean, Fast and Effective Clustering Method for Ancestry Estimation, Procedia Comput. Sci. 157 (2019) 306–312. https://doi.org/10.1016/j.procs.2019.08.171.
[106] X. Cao, Y. Liu, J. Wang, C. Liu, Q. Duan, Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network, Aquac. Eng. 91 (2020) 102122. https://doi.org/10.1016/j.aquaeng.2020.102122.
[107] B.C. Reaser, B.W. Wright, R.E. Synovec, Using Receiver Operating Characteristic Curves To Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry, Anal. Chem. 89 (2017) 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991.
[108] B.A. Parsons, D.K. Pinkerton, B.W. Wright, R.E. Synovec, Chemical characterization of the acid alteration of diesel fuel: Non-targeted analysis by two-dimensional gas chromatography coupled with time-of-flight mass spectrometry with tile-based Fisher ratio and combinatorial threshold determination, J. Chromatogr. A 1440 (2016) 179–190. https://doi.org/10.1016/j.chroma.2016.02.067.

[109] R.A. Fisher, A mathematical Examination of the Methods of determining the Accuracy of Observation by the Mean Error, and by the Mean Square Error, Mon. Not. R. Astron. Soc. 80 (1920) 758–770. https://doi.org/10.1093/mnras/80.8.758.

Chapter 2. Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis

This chapter was reproduced from Paige E. Sudol, Derrick V. Gough, Sarah E. Prebihalo, Robert E. Synovec, "Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis" *Talanta* 206 (2020) 120239.

2.1 INTRODUCTION

Comprehensive two-dimensional (2D) gas chromatography (GC×GC) has evolved into a widely applied separation technique for volatile and semi-volatile analytes since its inception in 1991 [1]. GC×GC utilizes two capillary columns connected serially by a modulator that transfers eluate from the first dimension (^{1}D) column to the second dimension (^{2}D) column according to the user-defined modulation period. The ^{1}D and ^{2}D columns should provide complementary separation selectivity to optimize the 2D separation [2,3], and analytes are separated in the 2D space according to their boiling point and stationary-phase interactions. GC×GC provides about a 10-fold increase in peak capacity versus one-dimensional GC, and significantly more peaks are produced in a given separation [4]. Hence, GC×GC is well suited for separating complex samples containing hundreds of analytes, and has been adapted for applications in diverse fields such as forensics [5,6], food chemistry [7], metabolomics [8,9], wastewater testing [10], and analysis of diesel fuels and biofuels [11–13]. However, for complex sample separations, it becomes difficult to visually ascertain differences between different samples, such as in the comparative analysis of sample classes. Indeed, extracting useful chemical information from complex GC×GC chromatograms is often a formidable challenge that requires implementation of chemometric techniques [14–19].

More specifically, discovery-based chemometric techniques are often utilized to leverage chemical differences between sample classes, in which the experimental design is either supervised or unsupervised, depending on whether the class information is known *a priori* [20,21]. A common unsupervised method for comparing GC×GC chromatograms is principal component analysis (PCA), a classification technique that aims to maximize the variation in a given dataset [22–24]. In PCA, the original data is projected onto a new set of axes called "loadings," and the projected coordinates on the new axes are called "scores." Hence PCA can be considered a data reduction step, since the PC axes better describe the overall variance in the dataset. Samples with little variation in chemical composition will have similar scores along these axes. Scores plots of the first few PC axes can thus be utilized to distinguish sample classes. More specifically, if the scores on PC2 are plotted against the scores on PC1, replicates of a given sample class should have sufficiently similar PC1 and PC2 scores that the replicates form distinct clusters in a PC scores plot. Sample classes may be distinguished from each other by differences in PC1 scores, PC2 scores, or both thereof.

However, it is critical to acknowledge that PCA is not very forgiving of non-sample related variations in a dataset. Retention time misalignment of peaks along both the ^{1}D and ^{2}D separation axes has the potential to be a major source of non-sample variation that may hamper the performance of PCA. Additionally, a low signal-to-noise ratio (*S/N*) may overpower important analyte signals and hamper the class distinguishing capability of PCA. A frequently employed remedy to provide *S/N* improvement is signal averaging, with the added benefit of reducing the data density to speed up subsequent preprocessing and chemometric steps [25]. However, an effective signal averaging approach for GC×GC data must maintain the advantages provided by using two separation dimensions, by preserving the 2D chemical selectivity

information. Thus, one can envisage the need to simultaneously average the data along the ^{1}D and ^{2}D axes by binning suitable 2D portions of the GC×GC data, whereby the "suitable" bin size optimizes the sample class distinguishing capability of PCA. Herein, binning is defined as averaging the signal along each separation dimension and can be equated to filling the entire GC×GC chromatogram with a grid of adjacent bins. Thus, the main purpose of binning GC×GC data prior to performing PCA is to improve the *S/N* and to correct retention time misalignment across a set of chromatograms [26–28].

Generally, the bin size applied should be sufficiently larger than the 2D dimension of a peak. So, based upon the typical width-at-base on each dimension, $^{1}W_{b}$ and $^{2}W_{b}$, for the ^{1}D and ^{2}D separations, respectively, the typical bin size would by fractionally larger than the product $^{1}W_{b} \times {}^{2}W_{b}$, depending upon the sample-to-sample retention time misalignment that needs to be overcome. For example, if the ^{1}D peaks are ~3 modulations and the ^{2}D peaks are ~100 ms, if the retention time shifting is ~50% to 100% of these peak width dimensions, a suitable bin size could be 7 modulations × 200 ms. Then, if signal is collected every 1 ms, each bin of 1400 raw data pixels is reduced to 1 binned pixel. For the purpose of this study, this typical bin size is referred to as the "standard operating procedure (SOP) bin size." Conventional wisdom is that application of a SOP bin size will optimize the *S/N* concurrent with minimizing artifacts due to any retention time misalignment, while preserving the inherent chemical selectivity information in the dataset. It is assumed that chemical information has not been compromised by binning. Herein we explore to what extent this assumption is valid. In this study we purposely study a dataset not hampered by alignment issues, so we can examine the interplay between optimizing *S/N* through binning, while not reducing the 2D chemical selectivity through over-binning.

Specifically, we study how bin size impacts distinguishing diesel fuel samples using PCA. Diesel fuel is an extremely complex sample, that produces an excellent GC×GC separation using a reverse column configuration (^{1}D polar and ^{2}D non-polar) [29]. The polar ^{1}D column provides elution of the compound classes in diesel from least polar to most polar, with analytes in each class eluting by boiling point, while the non-polar ^{2}D column subsequently produces distinct bands for each chemical compound class [11,29]. Hence, visually in a GC×GC chromatogram, analyte compounds are separated by boiling point along the ^{1}D axis and by functional groups (polarity) along the ^{2}D axis. The impact of binning on distinguishing sample classes is assessed by calculating the degree-of-class separation (DCS) between pairs of fuels in the scores plot [30].

A variety of methods exist for quantifying class separation in PCA, including calculation of Mahalanobis distances [31,32] and confidence ellipses [33–35] in scores plots. These metrics all utilize Euclidean distance in some fashion but require additional advanced statistical calculations (covariance matrices and singular value decomposition, for example). Thus, DCS was utilized as a suitable quantitative metric defined as follows,

$$DCS = \frac{D_{A,B}}{\sqrt{s_A^2 + s_B^2}} \tag{2.1}$$

where $D_{A,B}$ is the Euclidean distance between the centroids of two sample classes, A and B, and s_A^2 and s_B^2 are the variances in Euclidean distances between the centroids of classes A and B and the members of the respective classes.

Data for five diesel fuel samples was collected using GC×GC coupled to a flame ionization detector (FID). The FID functions at a high data collection frequency (data points/s), so the resulting chromatograms can benefit greatly from signal averaging, making it an excellent detector choice for this proof-of-principle study. The dataset was collected within about one day,

so relatively little misalignment along either separation dimension was observed. Thus, as a function of the degree of binning, a maximum DCS is interpreted as maintaining the inherent chemical selectivity provided by the separation, while concurrently optimizing the *S/N* ratio. One hundred and ten bin sizes were applied, ranging from no binning (1 modulation/bin on ^{1}D × 1 ms/bin on ^{2}D), to a very coarse level of binning (150 mod/bin on ^{1}D × 400 ms/bin on ^{2}D), which produced only 50 bins for each GC×GC separation (10 bins along ^{1}D × 5 bins along ^{2}D). A SOP bin size is defined, based on the peak widths obtained in the un-binned data, dictated by the 2D dimension of a typical peak given by $^{1}W_{b} \times {}^{2}W_{b}$. PCA was performed on all 110 binned datasets and the DCS was calculated between five pairs of diesel fuels in the scores plots. For each fuel pair, the "optimal" bin size, defined as producing the maximum DCS, was identified. Based upon the results obtained, we ascertain to what extent the SOP bin size will provide the maximum DCS between each of the five diesel fuel pairs, or conversely, if the optimal bin size varies from one fuel pair to another. Thus, the main goal herein is to determine to what extent a single bin size is capable of simultaneously maximizing DCS for all five diesel fuel pairs.

2.2 EXPERIMENTAL

2.2.1 Diesel Fuel Samples, and GC×GC-FID Instrument and Separation Conditions

Five diesel fuel samples were collected from fuel stations in the Seattle area, referred to as Fuel 1, Fuel 2, Fuel 3, Fuel 4, and Fuel 5. Four replicates of each fuel were analyzed by GC×GC-FID. The GC×GC-FID instrument illustrated in Fig. 2.1 was based upon a modified Agilent 6890 GC platform (Agilent Technologies, Palo Alto, CA, USA). The ^{1}D column (29.5 m, 250 µm inner diameter, and 0.25 µm film thickness) contained a Rxi-17Sil MS stationary phase (Restek, Bellefonte, PA, USA), and the ^{2}D column (2 m, 180 µm inner diameter, and 0.18 µm film thickness) contained a Rxi-1MS stationary phase (Restek, Bellefonte, PA, USA). A

high-speed, six-port diaphragm valve fitted with a 20 µL sample loop (VICI model DV-12-1116 T, Valco Instruments Company Inc., Houston, TX, USA) served as the flow-based modulator to transfer eluate from the ^{1}D column to the ^{2}D column. Sample injections were performed using a 7683B auto-injector (Agilent Technologies, Palo Alto, CA, USA). The inlet temperature was set to 275°C. A 1 µL sample volume was injected with a split ratio of 10:1. HPLC Grade acetone and hexane were used as solvent rinses prior to injection of a different diesel fuel type. All replicates of a given fuel were run within about one day to mitigate the effects of daily fluctuations in temperature, pressure, and other laboratory conditions.

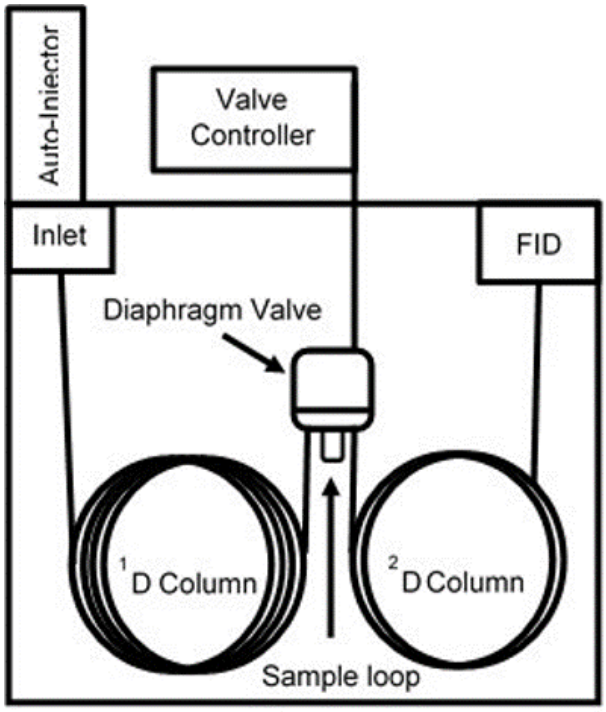

Figure 2.1. Schematic of GC×GC-FID instrument using high temperature diaphragm valve-based modulation. A 20 µL sample loop on the valve stored primary (^{1}D) column eluate before injection onto the secondary (^{2}D) column.

The GC×GC-FID operation conditions were as follows. The oven was held at 40 °C for 1 min, then increased 5 °C/min to 300 °C, and then held at 300 °C for 2 min. The total run time was 55 min. Ultra-high purity hydrogen (Grade 5, 99.999%) was utilized as the carrier gas (Praxair, Seattle, WA, USA) at a ^{1}D column flow of 1.5 mL/min and 20 mL/min on the ^{2}D

column. The pressure of the diaphragm valve (carrier gas flow) was ramped to ensure a constant flow rate on the ^{2}D column, in accordance with the oven temperature program. The pressure was held at 21.11 psig for 1 min, ramped at rate of 0.39 psig/min to 41.54 psig, and held at this final pressure for 2 min. Actuation of the diaphragm valve was controlled using LabVIEW (National Instruments, Austin, TX, USA). The modulation period was 2.0 s and the injection pulse width 20 ms. A high-speed electrometer built in-house was used for the FID so the data was collected at 100 kHz. The FID temperature was set to 250 °C and a makeup gas flow of 45 mL/min of nitrogen was utilized to minimize band broadening at the detector. The LabVIEW program performed boxcar averaging so the final raw data collection frequency was 1 kHz.

2.2.2 Data analysis

The raw data was imported into MATLAB R2018a (MathWorks, Inc., Natick, MA) as individual column vectors. Baseline correction was performed by applying an in-house written rolling minimum function. The baseline-corrected data vector was reshaped into a two-dimensional array, consisting of the number of ^{1}D modulations and the number of ^{2}D data points as the number of columns and rows, respectively. Each GC×GC separation was modulated 1650 times in 55 min. The last 150 modulations (5 min) were removed, since no peaks were observed after 50 min. The resulting size of the raw pixel-level (un-binned) data was 1500 modulations along ^{1}D and 2000 data points along ^{2}D. Binning was performed by dividing each axis by specified divisible integer values. The 110 bin sizes studied are listed in Table 2.1. The raw pixel-level data was 1 mod/bin on ^{1}D × 1 ms/bin on ^{2}D, and the SOP bin size was defined as 10 mod/bin on ^{1}D × 125 ms/bin on ^{2}D. The SOP bin size was chosen to be slightly larger than the experimentally obtained average $^{1}W_{b}$ and $^{2}W_{b}$ (14 s and 117 ms, respectively). A wide range of bin sizes were studied to include what was considered either “under-binning” to “over-binning,”

so as not to mistakenly exclude the optimal bin size for maximizing *S/N* while preserving chemical selectivity. The binned data was then unfolded back into a vector format for PCA.

Table 2.1. 110 bin sizes were studied, with every combination of the eleven ^{1}D bin sizes by each of the ten ^{2}D bin sizes.

^{1}D Bin Size (Number of modulations per bin)	^{2}D Bin Size (ms per bin)
1	1
2	2
3	4
5	10
6	20
10	40
15	80
20	125
30	200
60	400
150	

PCA was performed using PLS Toolbox 8.6.2 (Eigenvector Research, Inc., Wenatchee, WA, USA). The data was mean centered in PLS Toolbox prior to analysis. For each bin size a scores plot of PC2 versus PC1 was generated for the binned data. The DCS was calculated using Eq. (2.1) for the five directly adjacent pairs of fuels in the scores plot to quantify their relative separation (Table 2.2). Heat maps of DCS were generated using the imagesc function in MATLAB to examine the trend in DCS with binning for each fuel pair. Identical color scales were used to facilitate direct visual comparison. Using the DCS heat maps, the optimum bin size which produced the highest DCS for a given fuel pair was determined and compared to the DCS

at the raw data pixel-level (1 mod/bin on ^{1}D × 1 ms/bin on ^{2}D) and at the SOP bin size (10 mod/bin on ^{1}D × 125 ms/bin on ^{2}D). This comparison is conceptually illustrated in Fig. 2.2.

Table 2.2. The five fuel pairs for which DCS was calculated.

DCS Identifier	Pair of Fuels
A	Fuel 1 & Fuel 2
B	Fuel 2 & Fuel 3
C	Fuel 3 & Fuel 4
D	Fuel 4 & Fuel 5
E	Fuel 1 & Fuel 5

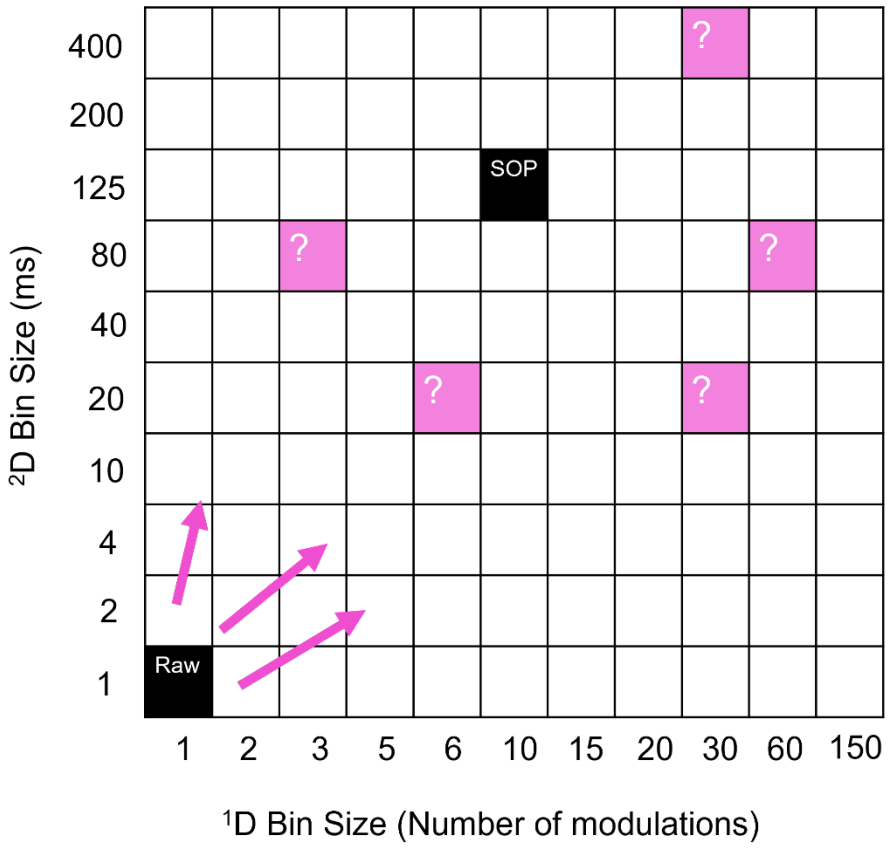

Figure 2.2. Illustration of the experimental design. The optimum bin size (i.e., highest DCS) is determined and compared to the DCS for the other 109 bin sizes (see Table 1), including the DCS at the raw data pixel-level (1 mod/bin on ^{1}D × 1 ms/bin on ^{2}D) and at the SOP bin size (10 mod/bin on ^{1}D × 125 ms/bin on ^{2}D), both indicated by the black boxes. The optimal bin sizes that may result for the five fuel pairs are indicated by the pink boxes.

2.3 RESULTS AND DISCUSSION

Examples of Fuel 1 binned to different bin sizes are shown in Fig. 2.3. When a reverse column configuration is used for GC×GC analysis of diesel fuel, three distinct roughly horizontal bands for each of the chemical compound classes (alkanes, cycloalkanes, and aromatics) are produced. While these bands are visible at the pixel-level in Fig. 2.3(A), one can imagine that some level of binning will affect how well class-distinguishing analytes manifest themselves in the 2D separation space. The *S/N* at the un-binned pixel-level is relatively lower, so it is more difficult to identify regions of importance. The signal intensity in each of the three chemical compound class bands appears to be relatively similar, and important, class-distinguishing compounds are not as visually obvious. Binning ^{1}D to 10 mod/bin while leaving ^{2}D un-binned (1 ms/bin) visually enhances the *S/N* while maintaining the 2D peak shape of individual analytes (Fig. 2.3(B)). At this level of binning, the aromatic region now appears to have analytes of the highest signal intensity. However, with the ^{2}D dimension left un-binned, the chemical selectivity of this dimension is understated. Binning ^{1}D to 10 mod/bin and ^{2}D to 125 ms/bin, which is the SOP bin size for this dataset, provides an even coarser image (Fig. 2.3(C)). The *S/N* has been improved to an even greater extent for all of the compound classes. Indeed, binning both dimensions appears to be necessary to maximize the *S/N* while preserving the chemical selectivity of each compound class. However, when ^{1}D and ^{2}D are binned to 150 mod/bin and 125 ms/bin respectively, distinct chemical compound class bands are no longer visible (Fig. 2.3(D)). It is now difficult to identify potential features of Fuel 1 since the *S/N* has increased greatly across all the bins, and one might suspect the data has been over-binned. Assessment of when over-binning has occurred is not possible by visual comparison of the chromatograms. This assessment can be rendered objective by maximizing the DCS in a comparative analysis using

PCA. We shall see that binning to such overly coarse levels, while counterintuitive, may provide added benefits to optimizing PCA.

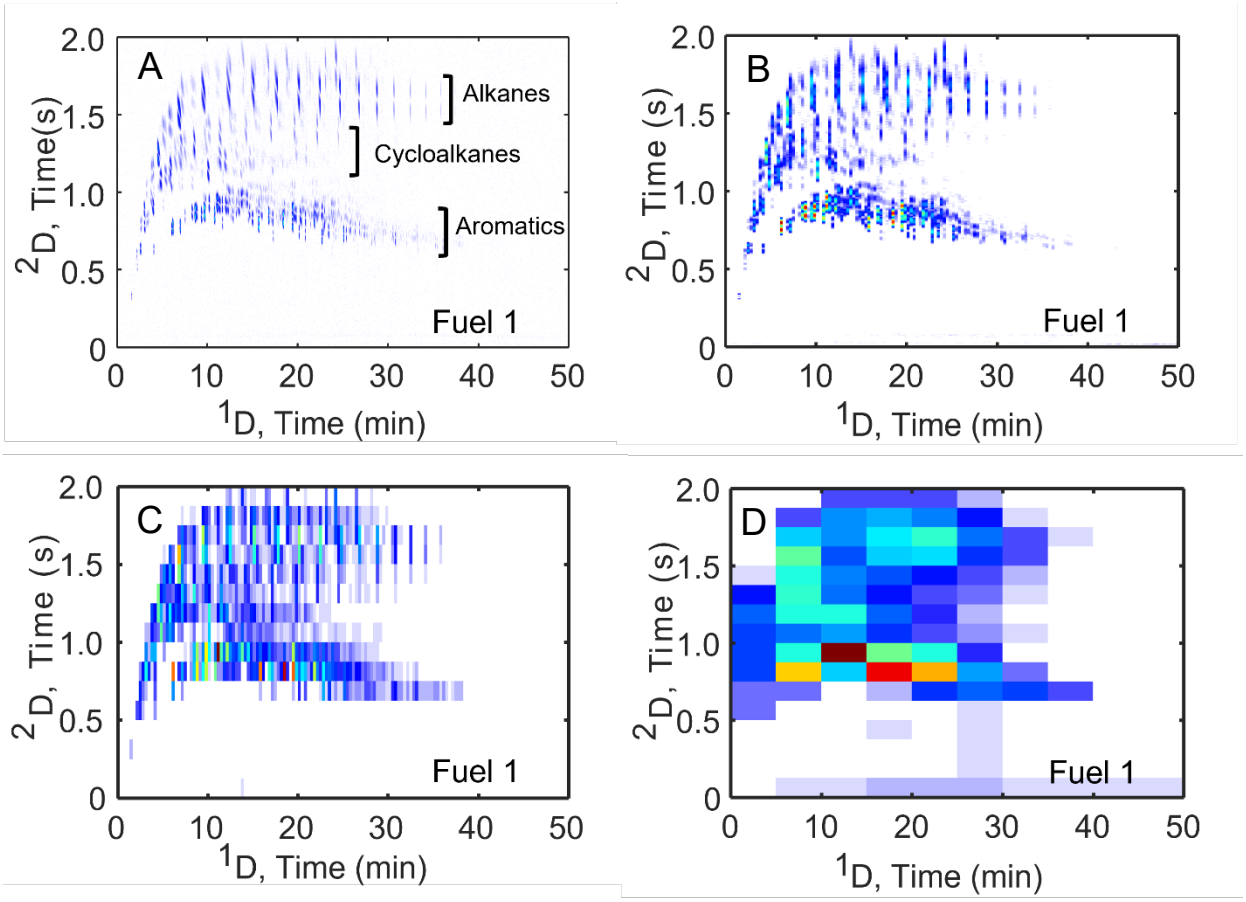

Figure 2.3. GC×GC chromatogram examples of Fuel 1 binned to different 2D bin sizes. (A) Raw data pixel level, 1 mod/bin on 1D × 1 ms/bin on 2D, corresponding to 1500 bins along 1D and 2000 bins along 2D. (B) 10 mod/bin on 1D × 1 ms/bin on 2D, corresponding to 150 bins along 1D and 2000 bins along 2D. (C) The SOP bin size, 10 mod/bin on 1D × 125 ms/bin on 2D, corresponding to 150 bins along 1D and 16 bins along 2D. (D) 150 mod/bin on 1D × 125 ms/bin on 2D, corresponding to 10 bins along 1D and 16 bins along 2D.

Raw data pixel-level chromatograms for the remaining four of the five diesel fuels are shown in Fig. 2.4 (Fuels 2 – 5). Due to the relatively lower *S/N*, only minute differences in signal intensity and breadth of the chemical compound class bands can be deduced. For example, Fuel 1 in Fig. 2.3(A) appears to have a relatively higher concentration of early-eluting cycloalkanes while Fuel 2 in Fig. 2.4(A) appears to have the most dense (highest concentration) aromatic region. Fuels 2 and 5 appear to have the longest aromatic bands (extending past ~40 min), so

they seem to contain the highest concentration of high boiling point aromatics. Conversely, Fuel 1 has a very short aromatic band (extending to ~30 to 35 min), which suggests it contains lower boiling point aromatic compounds. As in Fig. 2.3(C) for Fuel 1, when these chromatograms for Fuels 2 – 5 are binned to the SOP bin size, these chemical differences are more pronounced (Fig. 2.5). The next step was to investigate these apparent chemical fingerprint differences in the five fuels using PCA, using the DCS quantitative metric.

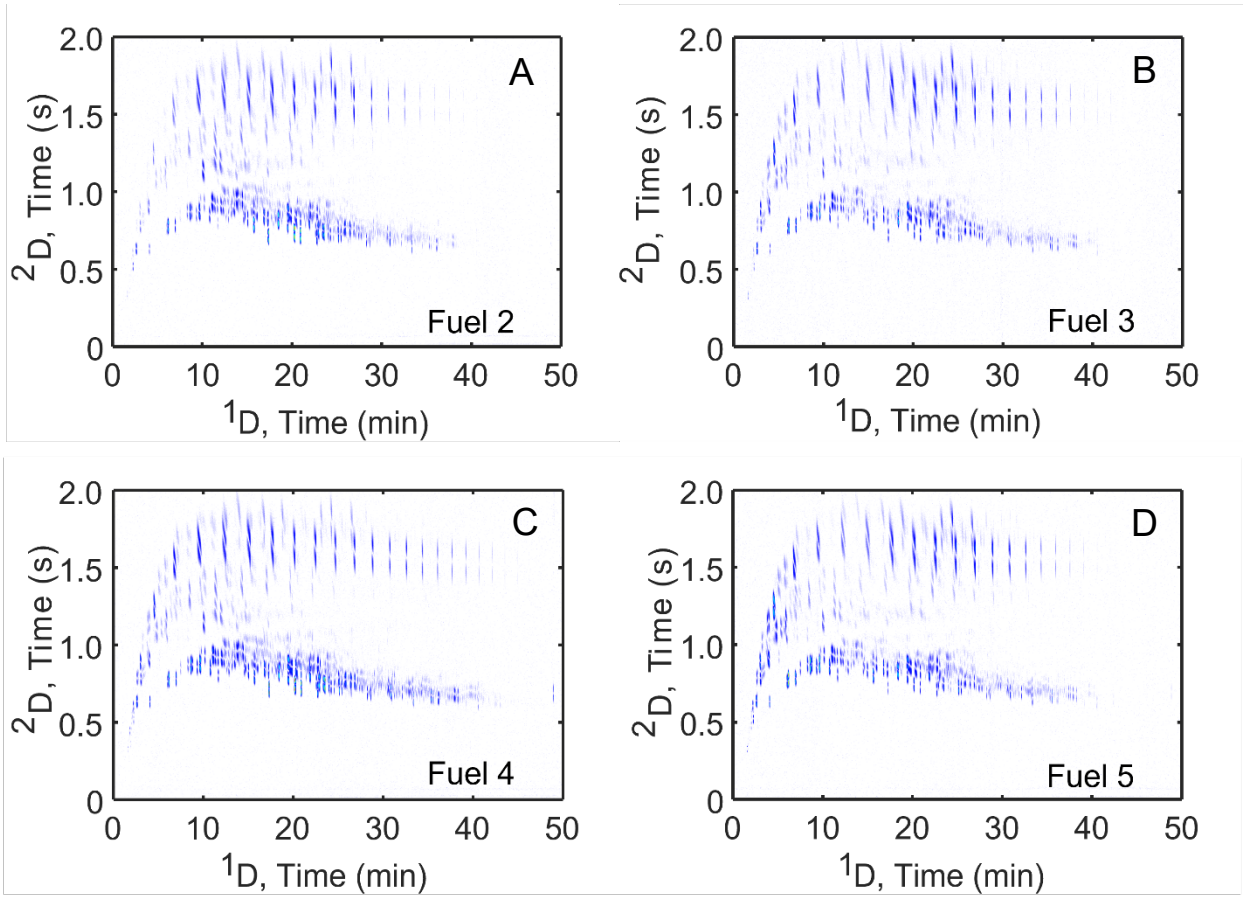

Figure 2.4. GC×GC chromatograms of the other four diesel fuels at the raw data pixel-level (1 mod/bin on ^{1}D × 1 ms/bin on ^{2}D). The color scale for each chromatogram is identical. (A) Fuel 2. (B) Fuel 3. (C) Fuel 4. (D) Fuel 5.

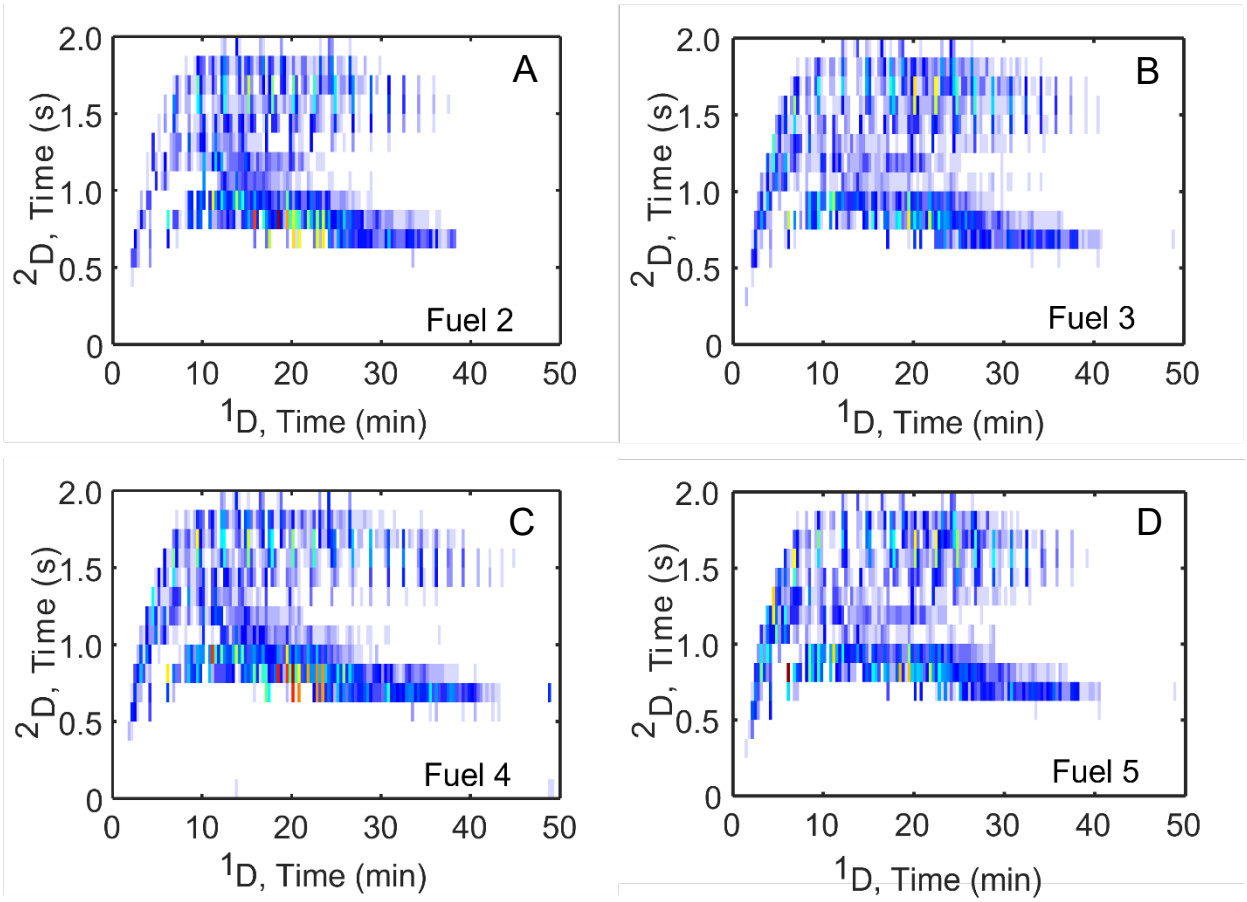

Figure 2.5. GC×GC chromatograms of the remaining four diesel fuels at the SOP bin size (10 mod/bin on ^{1}D × 125 ms/bin on ^{2}D). The color scale for each chromatogram is identical. (A) Fuel 2. (B) Fuel 3. (C) Fuel 4. (D) Fuel 5.

The scores plot for the pixel-level (un-binned) data is shown in Fig. 2.6(A). The adjacent fuel pairs for which DCS was calculated are listed in Table 2, since the adjacent fuels would likely benefit from increased distinction by binning prior to PCA. Assessing DCS between two non-adjacent sample classes (eg., Fuels 1 and 4) could also be performed, but was deemed not necessary for the purpose of this study. Except for fuel pair B, the fuels can be visually distinguished as unique classes in the scores plot. Note that only 49.03% of the variance is accounted for in the first two PCs, which suggests some improvement is possible [11,36–38]. The PC1 and PC2 loadings plots for the pixel-level data are shown in Fig. 2.6(B) and (C), respectively. The loadings are projected onto the GC×GC dimensions and the color scale

indicates their sign (positive or negative), with blue designating positive loadings and red designating negative loadings. Blue regions represent analyte peaks more concentrated in samples with positive scores and less concentrated in samples with negative scores. Conversely, red regions represent analyte peaks more concentrated in samples with negative scores and less concentrated in samples with positive scores.

Since there is no apparent pattern of blue and red in Fig. 2.6(B), there appears to be no broad correlation between PC1 and the separation dependencies along either the ^{1}D or ^{2}D dimension of the GC×GC separation, and one cannot conclude any cause/effect relationship between analyte boiling point or chemical compound class with PC1. Conversely, in the PC2 loadings plot, there appear to be primarily more blue peaks at the beginning of the temperature-programmed ^{1}D separation and primarily more red peaks at the end of the ^{1}D separation, suggesting a broad correlation between analyte boiling point and PC2.

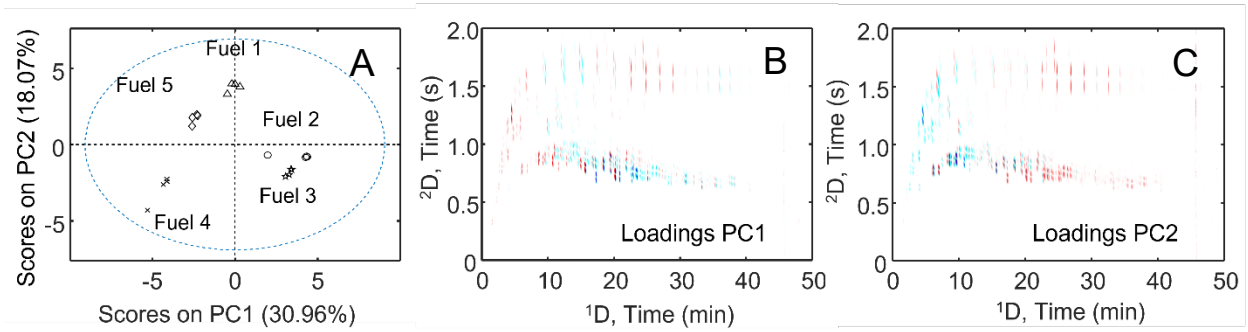

Figure 2.6. PCA results for pixel-level data. (A) Scores plot of the pixel-level diesel data (1 mod/bin on ^{1}D × 1 ms/bin on ^{2}D): Fuel 1 = triangles; Fuel 2 = circles; Fuel 3 = stars; Fuel 4 = x; Fuel 5 = diamonds. The DCS values for fuel pairs A-E as defined in Table 2 are as follows: A=9.7; B=1.9; C=13.6; D=8.5; and E=13.4. (B) PC1 loadings plot. Blue regions represent analyte peaks more concentrated in samples with positive scores and less concentrated in samples with negative scores. Conversely, red regions represent analyte peaks more concentrated in samples with negative scores and less concentrated in samples with positive scores. (C) PC2 loadings plot.

As in Fig. 2.6(A), the scores plot for the data binned to the SOP bin size is shown in Fig. 2.7(A). The individual diesel fuel samples are more tightly clustered together than in Fig. 2.6(A). Fuel pair B is still largely overlapped, which prompts investigation of additional bin sizes. However, binning has still resulted in a marked improvement in class distinction in the scores plot. The first two PCs now capture 78.77% of the variance in the dataset, a notable improvement compared to Fig. 2.6(A). The PC1 and PC2 loadings plots for the SOP-binned data are shown in Figs. 2.7(B) and (C), respectively. Binning makes patterns of blue and red, or a lack thereof, more identifiable in these plots compared to Figs. 2.6(B) and (C). The presence of alternating red and blue bins along both ^{1}D and ^{2}D in Fig. 2.7(B) once again suggests that there is no apparent broad correlation between PC1 and the separation dependencies along either the ^{1}D or ^{2}D dimension of the GC×GC separation. However, in Fig. 2.7(C), the benefits of binning are more pronounced than in Fig. 2.6(C). The PC2 loadings plot in Fig. 2.7(C) more clearly indicates primarily more blue peaks at the beginning of the ^{1}D separation and primarily more red peaks at the end of the ^{1}D separation, more confidently suggesting a broad correlation between analyte boiling point and PC2. The loadings plot interpretations are supported by comparison to subtraction plots, which are presented as Figs. A.1 and A.2 in Appendix A.

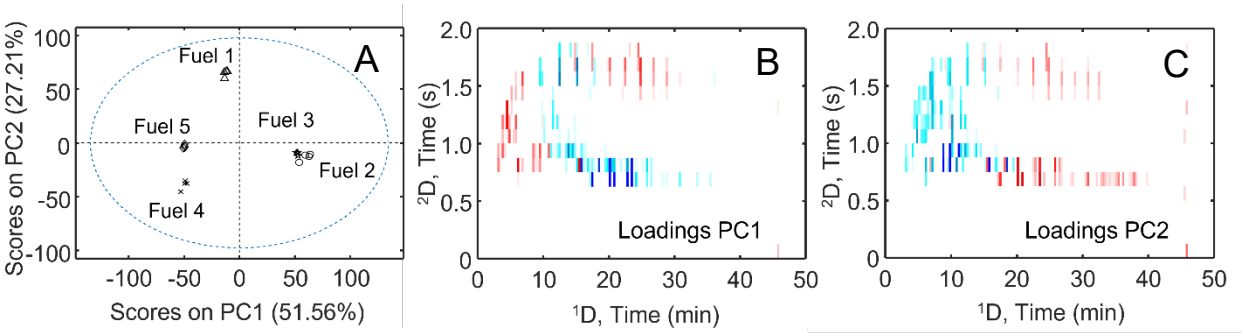

Figure 2.7. PCA results for SOP-binned data. (A) Scores plot of the SOP-binned diesel data (10 mod/bin on ^{1}D × 125 ms/bin on ^{2}D). The DCS values for fuel pairs A-E as defined in Table 2 are as follows: A=36.4; B=3.2; C=41.1; D=13.7; and E=46.7. (B) PC1 loadings plot. (C) PC2 loadings plot.

Following the heat map format illustrated in Fig. 2.2 (based upon Table 2.1), the DCS values for each of the five fuel pairs listed in Table 2.2 were prepared (Fig. 2.8). The magnitude of DCS at a given bin size is reflected in the bin color using a constant color scale. The pixel-level bin size is outlined with a red box, the SOP bin size is outlined with a black box, and the optimal bin size which produced the highest DCS for a given fuel pair is outlined with a purple box. As anticipated, the pixel-level data was oversampled, producing relatively small DCS values in Figs. 2.8(A-E). It is important to note that the SOP bin size does not produce the highest DCS for any of the fuel pairs in Figs. 2.8(A-E). Each fuel pair produced a different experimental DCS range (see color scale) and the SOP bin size produced a relatively intermediate DCS for all five fuel pairs. Each fuel pair, except for pairs C and D, has a different optimal bin size. Based on how a reverse GC×GC column configuration separates the compounds in diesel fuel, it was hypothesized that binning along ^{1}D relates primarily to boiling point selectivity, while binning along ^{2}D relates primarily to chemical compound class selectivity. If a given pair of fuels have minimal differences across a given dimension (^{1}D or ^{2}D), then they will be able to tolerate less binning along that dimension before their differences are summed away. For example, fuel pair A in Fig. 2.8(A) has an optimal bin size of 30 mod/bin on ^{1}D × 4 ms/bin on ^{2}D. This pair of fuels (Fuel 1 and Fuel 2) prefers minimal binning along ^{2}D, but very coarse binning along ^{1}D, suggesting boiling-point based chemical differences are the most important for their classification. These chemical differences are supported by examining the loadings plots for the SOP bin size in Figs. 2.7(B) and (C). Regardless of the bin size applied, interpretation of Fig. 2.7(A) indicates Fuel 2 always has more positive PC1 scores relative to Fuel 1, and Fuel 1 has more positive PC2 scores relative to Fuel 2. As indicated by the blue peaks in the PC1 and PC2 loadings plots in Figs. 2.7(B) and (C), Fuel 2 contains more mid-

boiling point to high-boiling point compounds and Fuel 1 contains predominantly low-boiling point compounds. It appears that Fuel 2 also contains more high-boiling point aromatic compounds compared to Fuel 1, but this appears to be the only substantial difference in the chemical compound classes present. It is noteworthy that these chemical features, namely the low boiling point compounds for Fuel 1 and the aromatic compounds for Fuel 2, were also identified in the subtraction plots provided in Appendix A (Figs. A.1 and A.2).

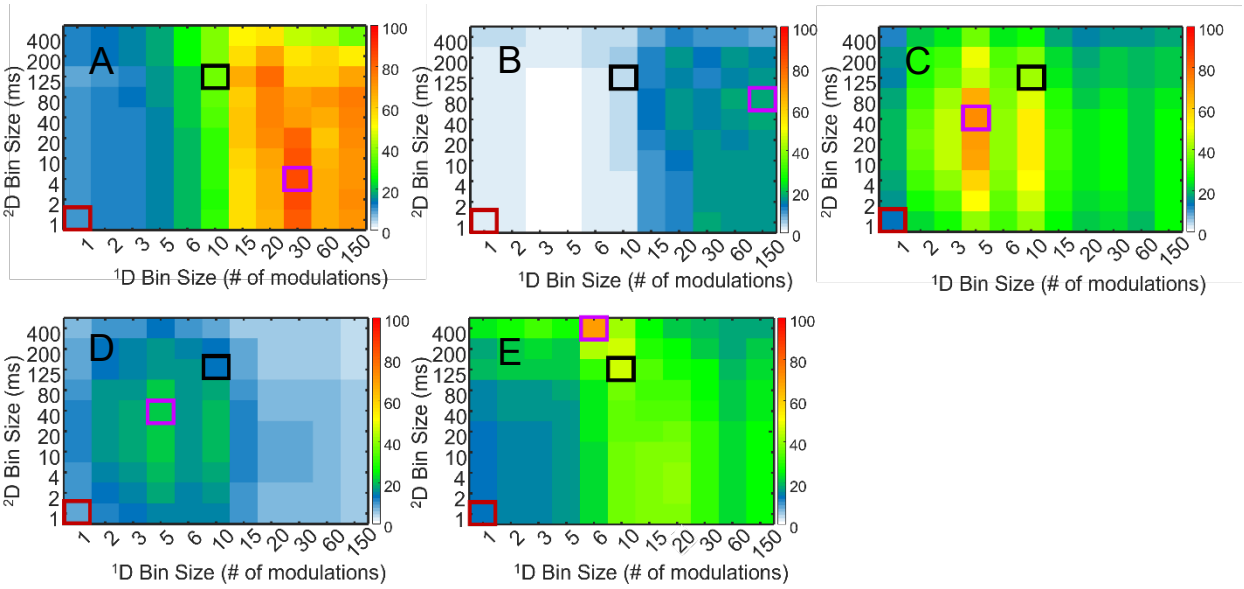

Figure 2.8. Heat map of DCS as a function of 2D bin size for each of the five fuel pairs studied (per Tables 2.1 and 2.2). The pixel-level bin size is outlined with a red box, the SOP bin size is outlined with a black box, and the optimal bin size which produced the highest DCS for a given fuel pair is outlined with a purple box. (A) Fuel pair A has an optimal bin size of 30 mod/bin on ^{1}D × 4 ms/bin on ^{2}D. (B) Fuel pair B has an optimal bin size of 150 mod/bin on ^{1}D × 80 ms/bin on ^{2}D. (C) Fuel pair C has an optimal bin size of 5 mod/bin on ^{1}D × 40 ms/bin on ^{2}D. (D) Fuel pair D has an optimal bin size of 5 mod/bin on ^{1}D × 40 ms/bin on ^{2}D. (E) Fuel pair E has an optimal bin size of 6 mod/bin on ^{1}D × 400 ms/bin on ^{2}D.

Unlike fuel pair A, fuel pair B benefits from coarse binning along both dimensions with an optimal bin size of 150 mod/bin on ^{1}D × 80 ms/bin on ^{2}D (Fig. 2.8(B)). Because the binning along ^{1}D is more extensive, this suggests the differences in boiling point between fuel pair B are very influential for their classification. Fuels 2 and 3 always have very similar scores along PC1

and PC2 (Figs. 2.6(A) and 2.7(A)), even at their optimal bin size (scores plot omitted for brevity), so the loadings could not be used to assess their chemical differences. The subtraction plot for this fuel pair was used to interpret their chemical differences. Fuel 2 contains more aromatics, whereas Fuel 3 contains more cycloalkanes (Fig. A.2(B) in Appendix A). Notably, there are alternating red and blue bins interspersed across the length of the alkane, cycloalkane, and aromatic bands, which suggest boiling point-based differences, or the presence of different compounds within the chemical compound class bands. Overall, Fuel 2 is characterized by low- and high-boiling point compounds and Fuel 3 is dominated more so by mid-boiling point compounds (Fig. A.2(B) in Appendix A). These results support the presented interpretation of the observed optimal bin size for this fuel pair. Note that the fuels in fuel pair B are categorically more similar in composition than fuel pair A, since they have a much smaller range in DCS values, as indicated in Figs. 2.8(A) and (B). Thus, for fuel pair B there is less room for improvement in the DCS by binning because the two fuels do not have as many chemical differences. Examination of the observed optimal bin size for fuel pairs C, D and E in the context of the loadings and subtraction plots yielded similar findings in Figs. 2.8(C)-(E).

Using the format in Fig. 2.2, a summary of the optimal bin sizes for each fuel pair (from Fig. 2.8) is provided in Fig. 2.9. Except for fuel pairs C and D, each fuel pair has a unique optimal bin size, and all markedly different than the SOP bin size. A major conclusion is that binning cannot be applied as a one-size-fits-all for a dataset with more than two sample classes. The optimal level of binning for improving DCS between any two sample classes depends on their chemical differences, and the chemical differences between two sample classes will naturally be different from one pair of classes to the next with more than two sample classes present.

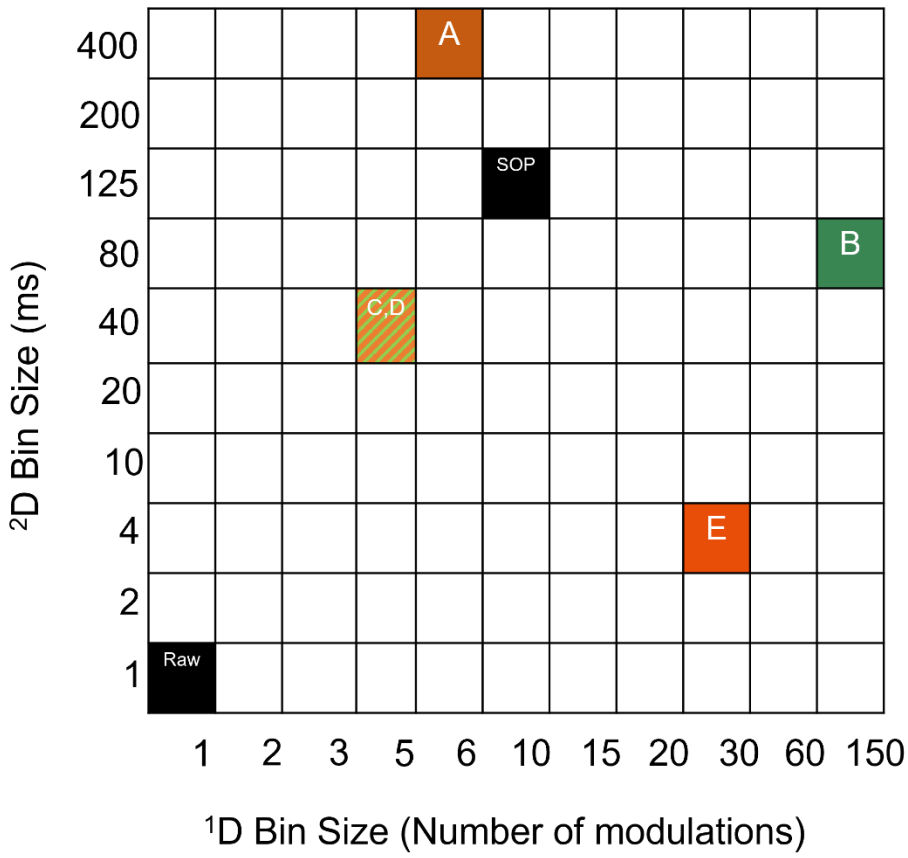

Figure 2.9. Summary of the optimal bin sizes obtained for each fuel pair, in the same format as the experimental design in Fig. 2.2. The results indicate that the optimal level of binning is fuel pair dependent and is not at a single SOP bin size.

In order to provide evidence of the robustness of this conclusion, a leave-one-sample class-out study was conducted. For all 110 bin sizes studied, Fuels 1-5 were individually excluded from the dataset prior to performing PCA, and the remaining DCS values were calculated using the same fuel pair scheme as in Table 2.2. The results of this study are provided in Figs. A.3-A.7 in Appendix A, and a summary is provided in Fig. A.8. Albeit with a few exceptions, the optimal bin sizes for the leave-one-sample class-out study are remarkably similar to the original optimal bin sizes, supporting the robustness of the conclusion that optimal bin size is sample class dependent in this PCA study of GC×GC data with more than two sample classes.

2.4 CONCLUSIONS

A proof-of-principle PCA study was performed on the effect of bin size with a GC×GC-FID diesel dataset consisting of five diesel samples. The dataset had minimal misalignment along either dimension, which facilitated studying the interplay between improving *S/N* by maximally binning versus suppressing chemical information by over-binning. At each bin size (110 total), the quantitative metric degree-of-class separation (DCS) was calculated for five adjacent fuel pair clusters in the scores plot. Each fuel pair demonstrated unique trends in DCS with bin size, with only fuel pairs C and D coincidentally having the same optimal bin size. None of the optimal bin sizes were equivalent to a single SOP bin size. Therefore, it was concluded that the optimal level of binning is dependent on maximizing the *S/N* concurrent with maintaining the chemical selectivity between adjacent fuel samples in the scores plot. Binning along ^{1}D primarily impacted boiling point selectivity, whereas binning along ^{2}D primarily impacted chemical compound class selectivity. For example, a pair of diesel samples with major differences in the chemical compound classes present was able to tolerate larger bin sizes along ^{2}D, whereas samples with major differences in boiling point distribution were able to tolerate larger bin sizes along ^{1}D. Binning compositionally similar samples to coarse levels summed away minute chemical differences, and the improved *S/N* on its own was less consequential for maximizing DCS. For validation, a leave-one-out study was conducted by excluding one fuel at a time from the PCA models, and recalculating DCS values for each binned dataset, and the optimal bin sizes remained relatively the same and/or the overall trend in DCS with bin size was preserved.

2.5 REFERENCES

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

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

[3] J. Jáčová, A. Gardlo, J.-M.D. Dimandja, T. Adam, D. Friedecký, Impact of sample dimensionality on orthogonality metrics in comprehensive two-dimensional separations, Anal. Chim. Acta 1064 (2019) 138–149. doi:10.1016/j.aca.2019.03.018.

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

[5] M. Marchini, C. Charvoz, L. Dujourdy, N. Baldovini, J.-J. Filippi, Multidimensional analysis of cannabis volatile constituents: identification of 5,5-dimethyl-1-vinylbicyclo[2.1.1]hexane as a volatile marker of hashish, the resin of Cannabis sativa L., J. Chromatogr. A 1370 (2014) 200–215. doi:10.1016/j.chroma.2014.10.045.

[6] B. Mitrevski, B. Veleska, E. Engel, P. Wynne, S.M. Song, P.J. Marriott, Chemical signature of ecstasy volatiles by comprehensive two-dimensional gas chromatography, Forensic Sci. Int. 209 (2011) 11–20. doi:10.1016/j.forsciint.2010.11.008.

[7] M.Z. Ozel, D.K. Yanık, F. Gogus, J.F. Hamilton, A.C. Lewis, Effect of roasting method and oil reduction on volatiles of roasted Pistacia terebinthus using direct thermal desorption-GC×GC-TOFMS, LWT - Food Sci. Technol. 59 (2014) 283–288. doi:10.1016/j.lwt.2014.05.004.

[8] J.C. Schoeman, I. du Preez, D.T. Loots, A comparison of four sputum pre-extraction preparation methods for identifying and characterising Mycobacterium tuberculosis using GC×GC-TOFMS metabolomics, J. Microbiol. Methods 91 (2012) 301–311. doi:10.1016/j.mimet.2012.09.002.

[9] H.K. Kandikattu, P. Rachitha, G.V. Jayashree, K. Krupashree, M. Sukhith, A. Majid, N. Amruta, F. Khanum, Anti-inflammatory and anti-oxidant effects of Cardamom (Elettaria repens (Sonn.) Baill) and its phytochemical analysis by 4D GCXGC TOF-MS, Biomed. Pharmacother. 91 (2017) 191–201. doi:10.1016/j.biopha.2017.04.049.

[10] P.C.F. Lima Gomes, B.B. Barnes, Á.J. Santos-Neto, F.M. Lancas, N.H. Snow, Determination of steroids, caffeine and methylparaben in water using solid phase microextraction-comprehensive two dimensional gas chromatography–time of flight mass spectrometry, J. Chromatogr. A 1299 (2013) 126–130. doi:10.1016/j.chroma.2013.05.023.

[11] G.L. Alexandrino, J. Malmborg, F. Augusto, J.H. Christensen, Investigating weathering in light diesel oils using comprehensive two-dimensional gas chromatography–high resolution mass spectrometry and pixel-based analysis: possibilities and limitations, J. Chromatogr. A 1591 (2019) 155-161. doi:10.1016/j.chroma.2019.01.042.

[12] N.G.S. Mogollón, F.A. de L. Ribeiro, M.M. Lopez, L.W. Hantao, R.J. Poppi, F. Augusto, Quantitative analysis of biodiesel in blends of biodiesel and conventional diesel by comprehensive two-dimensional gas chromatography and multivariate curve resolution, Anal. Chim. Acta 796 (2013) 130–136. doi:10.1016/j.aca.2013.07.071.

[13] K.M. Pierce, S.P. Schale, Predicting percent composition of blends of biodiesel and conventional diesel using gas chromatography–mass spectrometry, comprehensive two-dimensional gas chromatography–mass spectrometry, and partial least squares analysis, Talanta 83 (2011) 1254–1259. doi:10.1016/j.talanta.2010.07.084.
[14] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional gas chromatography: advances in instrumentation, chemometrics, and applications, Anal. Chem. 90 (2018) 505-532. doi:10.1021/acs.analchem.7b04226.
[15] K.J. Siebert, Using chemometrics to classify samples and detect misrepresentation, in: Prog. Authentication Food Wine, American Chemical Society, 2011, pp. 39–65. doi:10.1021/bk-2011-1081.ch004.
[16] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, J. Chromatogr. A 1488 (2017) 113–125. doi:10.1016/j.chroma.2017.01.052.
[17] H.D. Bean, J.E. Hill, J.-M.D. Dimandja, Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography–mass spectrometry data, J. Chromatogr. A 1394 (2015) 111–117. doi:10.1016/j.chroma.2015.03.001.
[18] F.N. Arslan, A. Kolk, H.-G. Janssen, Methods for one– and two–dimensional gas chromatography with flame ionization detection for identification of Mycobacterium tuberculosis in sputum, J. Chromatogr. B 1124 (2019) 204-217. doi:10.1016/j.jchromb.2019.06.012.
[19] 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.
[20] 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.
[21] K.M. Pierce, B.A. Parsons, R.E. Synovec, Chapter 10 - Pixel-level data analysis methods for comprehensive two-dimensional chromatography, in: A.M. de la Peña, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Eds.), Data Handl. Sci. Technol., Elsevier, 2015: pp. 427–463. doi:10.1016/B978-0-444-63527-3.00010-2.
[22] 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.
[23] J. Camacho, Visualizing big data with compressed score plots: approach and research challenges, Chemom. Intell. Lab. Syst. 135 (2014) 110–125. doi:10.1016/j.chemolab.2014.04.011.
[24] M.S. Park, J.Y. Choi, Theoretical analysis on feature extraction capability of class-augmented PCA, Pattern Recognit. 42 (2009) 2353–2362. doi:10.1016/j.patcog.2009.04.011.
[25] H.D. Bahaghighat, C.E. Freye, D.V. Gough, P.E. Sudol, R.E. Synovec, Ultrafast separations via pulse flow valve modulation to enable high peak capacity multidimensional gas chromatography, J. Chromatogr. A 1573 (2018) 115–124. doi:10.1016/j.chroma.2018.08.001.

[26] C.E. Freye, B.D. Fitz, M.C. Billingsley, R.E. Synovec, Partial least squares analysis of rocket propulsion fuel data using diaphragm valve-based comprehensive two-dimensional gas chromatography coupled with flame ionization detection, Talanta 153 (2016) 203–210. doi:10.1016/j.talanta.2016.03.016.
[27] 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.
[28] P.McA. Harvey, R.A. Shellie, Data reduction in comprehensive two-dimensional gas chromatography for rapid and repeatable automated data analysis, Anal. Chem. 84 (2012) 6501–6507. doi:10.1021/ac300664h.
[29] M. Jennerwein, M. Eschner, T. Wilharm, T. Gröger, R. Zimmermann, Evaluation of reversed phase versus normal phase column combination for the quantitative analysis of common commercial available middle distillates using GC×GC-TOFMS and Visual Basic Script, Fuel 235 (2019) 336–338. doi:10.1016/j.fuel.2018.07.081.
[30] 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.
[31] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemom. Intell. Lab. Syst. 50 (2000) 1–18. doi:10.1016/S0169-7439(99)00047-7.
[32] S.J. Dixon, N. Heinrich, M. Holmboe, M.L. Schaefer, R.R. Reed, J. Trevejo, R.G. Brereton, Use of cluster separation indices and the influence of outliers: application of two new separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles, J. Chemom. 23 (2009) 19–31. doi:10.1002/cem.1189.
[33] N.A. Sinkov, J.J. Harynuk, Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling, Talanta 83 (2011) 1079–1087. doi:10.1016/j.talanta.2010.10.025.
[34] B. Worley, S. Halouska, R. Powers, Utilities for quantifying separation in PCA/PLS-DA scores plots, Anal. Biochem. 433 (2013) 102–104. doi:10.1016/j.ab.2012.10.011.
[35] M. Cadoret, F. Husson, Construction and evaluation of confidence ellipses applied at sensory data, Food Qual. Prefer. 28 (2013) 106–115. doi:10.1016/j.foodqual.2012.09.005.
[36] J.W. McIlroy, R.W. Smith, V.L. McGuffin, Assessing the effect of data pretreatment procedures for principal components analysis of chromatographic data, Forensic Sci. Int. 257 (2015) 1–12. doi:10.1016/j.forsciint.2015.07.038.
[37] A.M. Hupp, L.J. Marshall, D.I. Campbell, R.W. Smith, V.L. McGuffin, Chemometric analysis of diesel fuel for forensic and environmental applications, Anal. Chim. Acta 606 (2008) 159–171. doi:10.1016/j.aca.2007.11.007.
[38] W. Fortunato de Carvalho Rocha, M.M. Schantz, D.A. Sheen, P.M. Chu, K.A. Lippa, Unsupervised classification of petroleum certified reference materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data, Fuel 197 (2017) 248–258. doi:10.1016/j.fuel.2017.02.025.

Chapter 3. Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry

This chapter was reproduced from Paige E. Sudol, Grant S. Ochoa, Robert E. Synovec, "Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry" *Journal of Chromatography A* 1644 (2021) 462092.

3.1. INTRODUCTION

Comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC×GC-TOFMS) is a powerful tool for the analysis of complex samples, with one of its most utilized applications being petroleum-based fuels [1–6]. Petroleum-based fuels, including gasoline, diesel fuel, and jet fuel, are composed of upwards of thousands of hydrocarbon components that often co-elute with one-dimensional gas chromatography (1D-GC) [7,8]. GC×GC harnesses the complementary separating power of two columns with "orthogonal" stationary phases to produce separations with up to ten times greater peak capacity [9,10]. Thus, because of the increased data dimensionality (additional separation dimension coupled with mass spectral detection), GC×GC-TOFMS analysis is much more amenable to complex mathematical and statistical methods of data interpretation, referred to as chemometrics [11–13].

Chemometric analyses can often be classified as either targeted or nontargeted [13–18]. Targeted analyses focus on known analytes, whereas nontargeted, discovery-based analyses, aim to uncover as much chemical information as possible, which often involves initially unknown analytes. Discovery-based analyses can be either supervised or unsupervised, whereby sample class membership is known *a priori* in supervised studies but unknown in unsupervised studies. Supervised and unsupervised discovery-based analyses of fuels are both popular approaches,

however, if sample class membership is known based upon the experimental design, then application of a supervised discovery-based analysis is often preferred [19–23]. While GC×GC-TOFMS is a powerful tool for reliable analyte identification and quantification in a supervised discovery-based analysis, such as Fisher ratio (F-ratio) analysis [24,25], the first step is to provide confident *discovery* of analytes that exhibit a sufficiently high F-ratio. Thus, just as we speak of there being a limit-of-detection (*LOD*) and limit-of-quantification (*LOQ*), there implicitly exists a limit-of-discovery referred to herein as the "discovery limit" (*DL*) for chemometric analyses. Since analytes of interest must be discovered first, this implies that the *DL* is the governing aspect of such investigations, and not necessarily the *LOD* and/or *LOQ*. The primary goal of this study is to investigate the interrelationship between the *DL*, *LOD* and *LOQ* as these analytical figures-of-merit relate to (supervised) discovery-based F-ratio analysis.

For this purpose, we study JP8 jet fuel spiked with a mixture of 14 sulfur-containing compounds at nominal concentrations of 30 ppm, 15 ppm, 3 ppm and 1.5 ppm into the neat fuel. These spiking levels produce five sample classes that are examined pairwise by F-ratio analysis: 30 ppm versus neat, 15 ppm versus neat, 3 ppm versus neat, and 1.5 ppm versus neat. We note that the quantitative analysis of low ppm levels of sulfur-containing compounds in fuels is an area of fuel quality assessment that remains largely understudied using GC×GC-TOFMS when coupled with discovery-based chemometrics [26]. Sulfur-containing compounds produce harmful emissions of sulfur oxides and cause engine corrosion, hence legal limits on allowable sulfur content in fuels have been lowered worldwide to between 10 and 15 ppm [27–31]. Thus, sulfur-contaminated jet fuel serves as a relevant application that would benefit from the findings of the current study.

F-ratio analysis is widely applied for discovery-based analyses using GC×GC-TOFMS [19,24,25,32–38]. The F-ratio is mathematically defined as the between-class variance divided by the sum of the within-class variance [39]. An analyte with a high F-ratio is statistically more likely to be class-distinguishing, so identification of class-distinguishing analytes, or "hits" is performed by constructing a "hitlist" in which analytes are ranked by their F-ratio. The analyte with the highest F-ratio is found at the top of the hitlist (i.e., hit number 1), followed by the analyte with the second highest F-ratio as hit 2, and so on. F-ratio analysis was originally performed on GC×GC-TOFMS data at the pixel-level, whereby the data is unfolded along the modulation period (P_M) and the F-ratio is calculated at each data point per mass channel (*m/z*) [40–46]. However, pixel-level F-ratio results are ultimately hindered by retention time variation. Tile-based F-ratio analysis was developed to improve the discoverability of true positives against false positive hits. Here, true positives are class-distinguishing hits the analyst seeks to discover (generally at or near the top of the hitlist), while false positives are non-class-distinguishing hits (generally further down from the top of the hitlist), which can be attributed to spurious detector noise, retention time variation, or a combination of both [24,25]. The tile-based F-ratio algorithm sums the chromatographic signal within defined tiles across the 2D space prior to calculating F-ratios, which when followed by the redundant hit removal step, simultaneously negates the effects of retention time shifting along both dimensions and produces a considerable signal-to-noise (*S/N*) improvement [24,25]. We direct the reader to previous reports describing the tile-based F-ratio analysis algorithmic details [19,24,25,32].

Although tile-based F-ratio analysis has been successfully used in many applications, maximizing the ranking of analyte hits as concentrations approach the *DL* (and thus the *LOD* and/or *LOQ*) remains a pressing issue [19,32]. The *DL* is inextricably tied to the true positive

discovery rate, which will decrease as analyte concentration decreases. Previous work indicates that tile-based F-ratio analysis discovered analytes at concentrations as low as ~1 ppm [25]. However, many false positives were interspersed among the true positive hits at low concentration, now understood to be an addressable issue by optimizing key software parameters: (1) tile size on the first and second chromatographic dimensions (^{1}D and ^{2}D, respectively), (2) cluster window size for redundant hit removal on ^{1}D and ^{2}D, (3) *S/N* threshold, and (4) number of F-ratio *m/z* to average for hitlist ranking. Recently, we demonstrated that optimization of the *S/N* threshold per tile and the number of F-ratio *m/z* to average for hitlist ranking improved the ranking of true positives by pushing false positives down the hitlist [34]. The conclusion of this prior study was that the optimum *S/N* threshold was 10, and the number of F-ratio *m/z* to average was 10 *m/z* with a minimum of 3 *m/z*. However, this prior study examined spiked analytes in diesel in which the lowest spike level was 10 ppm, so all of the spiked analytes had many *m/z* that cleared the *S/N* threshold and/or were not hampered by the background instrumental and chemical noise. Furthermore, using fewer than 3 *m/z* was not examined, as it was initially thought that use of only 1 or 2 *m/z* would introduce false positives, as was the case for pixel-based F-ratio analysis [42–45]. The study herein explores the notion that applying these previously reported parameters to lower concentration comparisons might actually diminish the rank of true positives [24,25]. Indeed, recent work has highlighted the importance of assessing *m/z* purity (i.e. how selective a *m/z* is for an analyte of interest versus an interferent) during tile-based F-ratio analysis to facilitate accurate quantification [33]. Three concentration comparisons were assessed, and the number of *m/z* that passed defined mass spectral purity metrics decreased as analyte concentrations decreased [33]. Essentially, this observation translates to the notion that at low concentration, very few F-ratio *m/z* per analyte

will likely be pure and use of impure *m/z* would likely reduce the effectiveness of ranking true positives. The logical conclusion was that of all the *m/z* passing defined F-ratio and *S/N* thresholds for a given analyte, the top F-ratio *m/z* is most likely to be "pure."

Herein we examine to what extent only utilizing the top F-ratio *m/z* (and not an average F-ratio) will be advantageous for F-ratio comparisons in which the "discovered" analytes exhibit a class-to-class concentration change approaching the *LOD* and/or *LOQ*. We also examine the impact of tile size selection on analyte discoverability. The F-ratio tile size chosen for a given analysis should encompass the full analyte peak width-at-base along both dimensions (^1W_b and ^2W_b) and any retention time shifting [24,25]; in this study the optimal tile size was selected using this approach. In principle, use of too large of a tile will allow more interferent signal to drown out the signal from low concentration analytes and push these hits further down the hitlist, while use of too small of a tile will increase the number of redundant hits in the hitlist and thus push hits for subsequent analytes further down the hitlist. Following optimization of key software parameters, eg., *S/N* threshold and number of F-ratio *m/z* to use, we demonstrate the resulting hit number rankings when the tile size is suboptimal. The cluster window dimensions for redundant hit removal must be smaller than the tile dimensions and thus cannot be optimized if the tile size is suboptimal, so the cluster window dimensions are held constant relative to the tile dimensions.

3.2 EXPERIMENTAL

JP8 jet fuel was spiked with a mixture of 14 sulfur-containing compounds, referred to as the "sulfur mix", at nominal concentrations of 30 ppm, 15 ppm, 3 ppm and 1.5 ppm per compound. The prepared concentrations of each sulfur-containing compound in the spiked fuel samples in ppm (mg/kg) are provided in Table B.1 in Appendix B. Table 3.1 lists the 14 compounds included in the sulfur mix, along with their ^{1}D and ^{2}D retention times (^1t_R and ^2t_R,

respectively), and top ten most intense *m/z* observed in the reference spectra, obtained with the sulfur mix spiked into toluene at a nominal concentration of 200 ppm/analyte. Four replicates of the spiked fuel samples (including the 200 ppm sulfur mix) were analyzed with the Pegasus BT 4D GC×GC-TOFMS instrument (LECO Corporation, St. Joseph, MI) using a quad-jet thermal modulator, an Agilent 7890 gas chromatograph (Agilent Technologies, Palo Alto, CA) and an L-PAL3 GC autosampler. Eight replicates of the neat fuel were collected, four with the 30 ppm and 15 ppm replicates and the other four with the 3 ppm and 1.5 ppm replicates. A 0.5 µl aliquot of each replicate was injected at a split ratio of 200:1 in the GC inlet, which was held constant at 275°C. The ^{1}D column (26.4 m × 250 µm i.d. × 0.25 µm film thickness) had a polar Rxi-17Sil MS stationary phase (Restek, Bellefonte, PA) and the ^{2}D column (1.9 m × 180 µm i.d. × 0.18 µm film thickness) had a non-polar Rxi-1 MS stationary phase. Ultrahigh purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) was the carrier gas at a constant flow of 2.0 ml/min. The ^{1}D column was held at 40°C for 1.5 min, increased to 200°C at 5°C/min, and held at 200°C for 1 min. The ^{2}D column and modulator block utilized the same temperature program with offsets of 12°C and 30°C, respectively. The modulation period was P_M of 3 s. After an acquisition delay of 10 s, *m/z* 45-200 were collected at a collection frequency of 100 spectra/s with a detector voltage of 2005 V. The ion source and transfer line were held constant at 225°C and 285°C, respectively.

Table 3.1. Identity of sulfur-containing compounds in the equal-mass mixture that was spiked into JP8 jet fuel at varying concentration levels, with locations in the separation indicated in Fig. 3.1.

	Analyte	^1t_R (min)	^2t_R (s)	Top 10 most intense *m/z* (highest to lowest)
1	thiophene	3.25	2.82	84, 58, 45, 57,50, 69, 83, 85, 51, 81
2	2-methylthiophene	5.00	0.34	97, 98, 45, 69, 50, 53, 63, 58, 99, 57
3	3-methylthiophene	5.20	0.37	97, 98, 45, 91, 65, 63, 92, 69, 51, 50
4	tetrahydrothiophene	6.30	0.53	60, 88, 45, 46, 59, 47, 87, 54, 58, 55
5	2,5-dimethylthiophene	7.25	0.69	111, 112, 97, 45, 59, 51, 77, 50, 69, 53
6	1,4-oxathiane	9.40	0.35	46, 104, 61, 45, 74, 47, 59, 60, 48, 58
7	2-propylthiophene	9.75	0.53	97, 126, 45, 53, 98, 99, 69, 58, 51, 71
8	2-butyl-5-ethylthiophene	17.80	0.76	125, 168, 97, 126, 91, 123, 110, 153, 77, 45
9	2-hexylthiophene	18.45	0.80	97, 98, 168, 45, 53, 99, 111, 84, 85, 110
10	benzo[b]thiophene	19.30	0.27	134, 89, 135, 63, 90, 69, 45, 67, 108, 50
11	3-acetyl-2,5-dimethylthiophene	20.45	0.55	139, 154, 111, 59, 67, 140, 141, 77, 45, 51
12	2-methylbenzothiophene	21.8	0.35	147, 148, 149, 69, 45, 115, 63, 74, 103, 77
13	3-methylbenzothiophene	22.45	0.26	147, 148, 149, 69, 45, 74, 115, 103, 77, 63
14	2-chloroethyl phenyl sulfide	23.85	0.35	123, 45, 172, 109, 65, 51, 110, 174, 69, 50

Data analysis was performed in MATLAB R2020a (The Mathworks Inc., Natick, MA, USA) after the data were imported from the LECO ChromaTOF for BT software. Four F-ratio comparisons were performed: 30 ppm versus neat, 15 ppm versus neat, 3 ppm versus neat, and 1.5 ppm versus neat. Note that the four neat chromatograms collected with the 30 ppm and 15 ppm replicates were used in those two comparisons, while the four neat chromatograms collected with the 3 ppm and 1.5 ppm replicates were used with those two comparisons. Each comparison

was initially performed with two different sets of parameters regarding the number of F-ratio *m/z* used to rank hits in the hitlist. For the first parameter set the top 10, minimum of 3 F-ratio *m/z* were averaged to get average F-ratios and rank the hitlist. For the second parameter set only the top F-ratio *m/z* was used to rank the hitlist. An F critical value (F_{crit}) of 5.99, corresponding to a four versus four comparison at a *p* value of 0.05, was utilized as an F-ratio threshold for all comparisons. The optimal tile size used for both sets of parameters was 12 s (4 modulations) on ^{1}D, and 300 ms on ^{2}D, based on the typical $^{1}W_b$ and $^{2}W_b$ of native fuel peaks observed in the chromatograms, as described in the Results and Discussion section. Previous work demonstrated that cluster window dimensions which are approximately half the tile dimensions perform well [19,24,25,32–35], so the cluster window dimensions used were 6 s (2 modulations) on ^{1}D and 150 ms on ^{2}D. To eliminate tiles with insufficient signal, *S/N* thresholds ranging from 1 to 50 times the standard deviation of the tiled baseline noise ($\sigma_{,BN}$) were tested, and a *S/N* threshold of $10\sigma_{,BN}$ produced the least false positive hits while maximizing the discovery and ranking of true positive hits. Hence, a *S/N* threshold of $10\sigma_{,BN}$ was utilized for all F-ratio comparisons. During the tile size and *S/N* threshold selection process, the 30 ppm versus neat and 15 ppm versus neat comparisons were found to produce essentially the same hitlist rankings, therefore we excluded the 30 ppm versus neat comparison from further examination. After this study with the optimal tile and cluster window sizes, F-ratio comparisons were repeated with two suboptimal sets of tile and cluster window sizes. For the first parameter set, the tile dimensions were 12 s on ^{1}D and 600 ms on ^{2}D (too large) with cluster window dimensions of 6 s on ^{1}D and 300 ms on ^{2}D (also too large). For the second parameter set, the tile dimensions were 12 s on ^{1}D and 100 ms on ^{2}D (too small) with cluster window dimensions of 6 s on ^{1}D and 50 ms on ^{2}D (also too small).

3.3 RESULTS AND DISCUSSION

The total ion current (TIC) GC×GC chromatogram obtained for the 15 ppm spiked JP8 jet fuel is presented in Fig. 3.1, with the locations of the 14 analytes in the sulfur mix circled. An impressive compound class-based separation was produced by the reverse GC×GC column configuration, with the alkanes (^2t_R from ~ 2 to 3 s), cycloalkanes (^2t_R from ~1 to 2 s), and aromatics (^2t_R from ~0.2 to 0.75 s) spread out in the 2D space. Temperature programming the ^{1}D separation produced approximately constant ^1W_b at 12 s (4 modulations) with negligible retention time shifting; this time interval was selected as the optimal tile dimension on ^{1}D. However, the ^{2}D separation in GC×GC is pseudo-isothermal, resulting in minor increases in ^2W_b as ^2k' increases [47,48], which complicates tile dimension selection on ^{2}D. To determine an optimal tile dimension on ^{2}D, the ^2W_b of ten representative native fuel peaks were measured. Their locations are labeled in Fig. B.1 in Appendix B, and their identities and ^2W_b are provided in Table B.2. These ten native peaks exhibited a ^2W_b range of 165-303 ms. Thus, an optimal tile dimension of 300 ms on ^{2}D was selected to encompass the widest peaks. Therefore, the cluster window dimensions of 6 s (2 modulations) on ^{1}D and 150 ms on ^{2}D were selected, being half the tile dimensions. We return to the issue of what happens with a sub-optimal tile size later, whether it be too large or too small.

The spiked analyte, 2,5-dimethylthiophene, is examined in detail in Fig. 3.2 to illustrate why performing tile-based F-ratio analysis on a per *m/z* basis is likely to be beneficial to elucidate class-to-class concentration differences at low concentration. This analyte was arbitrarily chosen for illustrative purposes. An overlay of the four 15 ppm replicates (red) and four neat replicates (blue) at the TIC signal in Fig. 3.2(A) indicates the presence of two ^{2}D peaks in the 15 ppm spiked fuel relative to the neat fuel. However, when the average TIC

chromatograms for the 15 ppm (Fig. 3.2(B)) and neat (Fig. 3.2(C)) replicates are folded into 2D contour plots, there is no discernable visual difference. In contrast, an overlay of the 15 ppm and neat replicates at *m/z* 112 in Fig. 3.2(D) more clearly indicates the ^{2}D peaks for 2,5-dimethylthiophene in the 15 ppm spiked fuel relative to the neat fuel. Although *m/z* 111 is the most intense in the reference spectrum of this analyte (Table 3.1), *m/z* 112 was chosen because we shall see that it produces the top F-ratio *m/z* with the 15 ppm versus neat comparison. The presence of 2,5-dimethylthiophene in the 15 ppm spiked fuel (Fig. 3.2(E)) relative to the neat fuel (Fig. 3.2(F)) is confirmed via examination of their 2D contour plots. The tile dimensions of 12 s (4 modulations) on ^{1}D and 300 ms on ^{2}D are indicated with a dashed black box around 2,5-dimethylthiophene in the 2D contour plots provided in Fig. 3.2. Although this optimal tile size was selected from ten representative native fuel peaks, without *a priori* information about the spiked sulfur-containing analytes, it is clearly suitable for this analyte, as the full $^{1}W_{b}$ and $^{2}W_{b}$ are captured. While illustrated with the analyte at the center of the tile for clarity, having the analyte centered in a tile is not a requirement of tile-based F-ratio analysis.

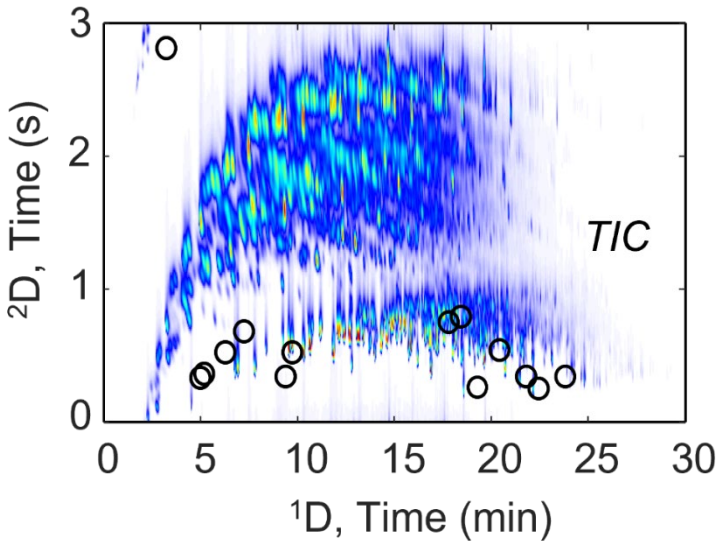

Figure 3.1. Total ion current (TIC) chromatogram from the GC×GC-TOFMS separation of JP8 jet fuel spiked with 15 ppm of sulfur-containing compound mix (Table 3.1). Locations of the 14 sulfur-containing compounds are circled.

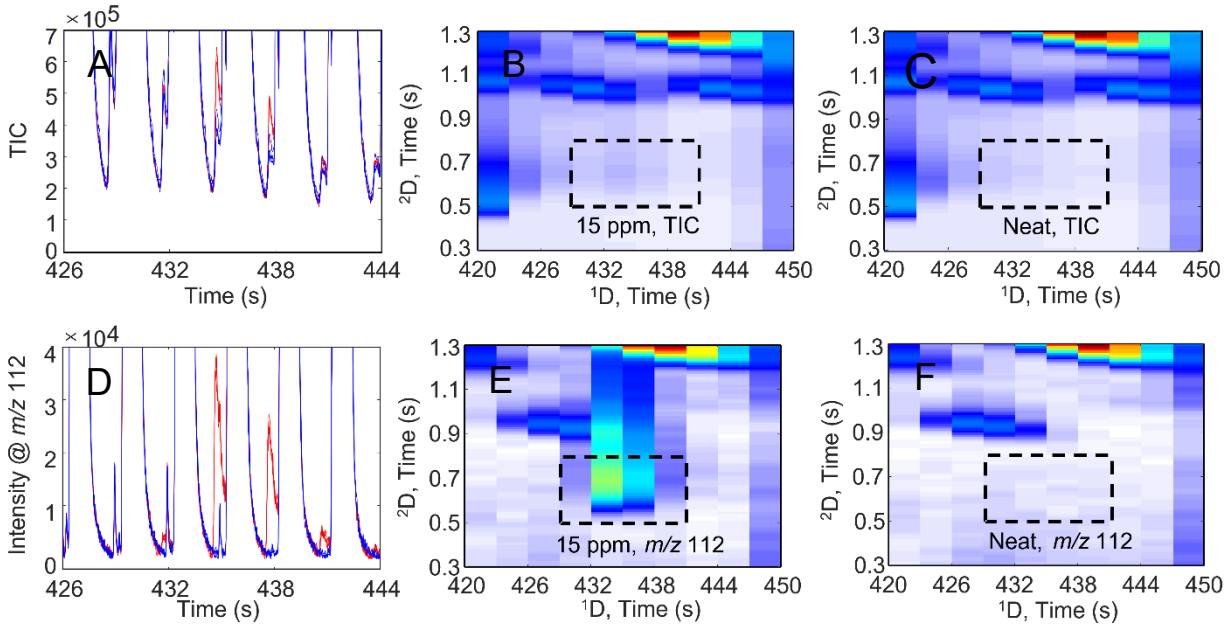

Figure 3.2. Illustration of the discovery challenge addressed by tile-based F-ratio analysis that leverages selective *m/z*, for 2,5-dimethylthiophene with the 15 ppm versus neat comparison. (A) Overlays of unfolded TIC data for the 15 ppm (red) and neat (blue) spiked JP8 fuel. (B) Average TIC GC×GC chromatogram for the 15 ppm spiked samples. The chosen tile size of 4 modulations on ^{1}D and 300 ms on ^{2}D is shown with a dashed black box, centered on the analyte for clarity. (C) Average TIC GC×GC chromatogram for the neat samples. (D) Overlays of unfolded data at *m/z* 112 for the 15 ppm (red) and neat (blue) spiked JP8 fuel. (E) Average GC×GC chromatogram at *m/z* 112 for the 15 ppm samples. (F) Average GC×GC chromatogram at *m/z* 112 for the neat samples.

Using 2,5-dimethylthiophene for a more in-depth discussion, the F-ratios obtained per *m/z* (i.e., F-ratio spectrum, blue) for the 15 ppm versus neat comparison are provided in a reflection plot in Fig. 3.3(A) relative to the corresponding *m/z* abundances from this analyte's reference mass spectrum (orange). Numerous highly abundant *m/z* from the reference spectrum have correspondingly high F-ratios ranging from ~ 200 to 2000 in Fig. 3.3(A), which implies these *m/z* are also highly selective for the purpose of applying F-ratio analysis. Therefore, even though the average F-ratio of 1091 for 2,5-dimethylthiophene is about 3-fold lower than its top F-ratio of 3738, since the average F-ratio is calculated using these highly selective *m/z* suggests

using average F-ratios to rank the 15 ppm versus neat hitlist may still provide a suitable ranking of true positives (i.e., spiked sulfur-containing analytes) relative to false positives (i.e., not spiked sulfur-containing analytes). When examining the analogous F-ratio spectrum for the 3 ppm versus neat comparison in Fig. 3.3(B), many of the highly selective and sensitive *m/z* from Fig. 3.3(A) exhibit diminished F-ratios, but *m/z* 112 and 111 still have relatively high F-ratios, 299 and 442, respectively. Clearly 2,5-dimethylthiophene has fewer "competitive" F-ratio *m/z* at 3 ppm versus neat compared to 15 ppm versus neat, but its average F-ratio of 111 is still only about 4-fold lower than its top F-ratio of 442. Therefore, the average F-ratio may not be a problematic metric for ranking for 2,5-dimethylthiophene, but it could hinder the rankings of other spiked analytes with less sensitivity with the 3 ppm versus neat comparison. When examining the 1.5 ppm versus neat comparison in Fig. 3.3(C), only the singular *m/z* 112 has a relatively high F-ratio of 186, while *m/z* 111 and 114 both exhibit large decreases in F-ratios from 442 to 25, and from 189 to 21, respectively, relative to the 3 ppm versus neat comparison (Fig. 3.3(B)). The resulting average F-ratio of 33 is about 6-fold lower than its top F-ratio of 186, which is a much greater loss of "F-ratio selectivity" than was observed at the other two concentration comparisons. Thus, these small F-ratio *m/z* are excessively diminishing the average F-ratio of 2,5-dimethylthiophene, suggesting that use of the average F-ratio will hinder this analyte's resulting hit ranking in the 1.5 ppm versus neat comparison.

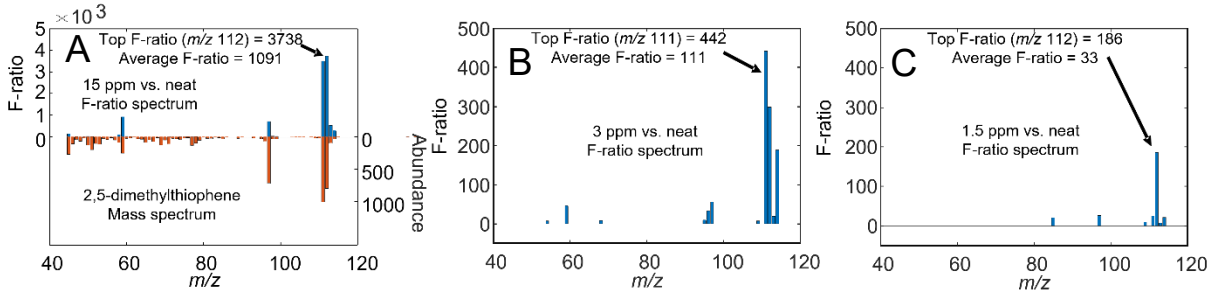

Figure 3.3. Comparison of F-ratio selectivity obtained at 15 ppm, 3 ppm, and 1.5 ppm for 2,5-dimethylthiophene. (A) F-ratios per *m/z* (blue), referred to as a F-ratio spectrum, for 2,5-dimethylthiophene from the 15 ppm versus neat comparison, presented in a reflection plot relative to in-house collected reference mass spectrum (orange). (B) F-ratio spectrum for 2,5-dimethylthiophene from the 3 ppm versus neat comparison. (C) F-ratio spectrum for 2,5-dimethylthiophene from the 1.5 ppm versus neat comparison.

To definitively assess the suitability, or lack thereof, of the average F-ratio for ranking the F-ratio concentration comparisons, the exercise conducted for a single analyte in Fig. 3.3 was broadened in Fig. 3.4 for all hit locations identified by the tile-based F-ratio algorithm, with the F-ratios per *m/z* of all tiles passing the *S/N* threshold of $10\sigma_{,BN}$ plotted against their corresponding *m/z* values. The y-axis (F-ratio) is shown on a log scale for ease of visualization. The red dots correspond to the top true positive F-ratio *m/z*, one for each sulfur-containing compound that is "discovered," while the yellow dots correspond to the secondary true positive F-ratio *m/z* (i.e. remaining *m/z* for true positives), and the blue dots correspond to the false positive F-ratio *m/z* due to the JP8 fuel background. When examining the 15 ppm versus neat comparison in Fig. 3.4(A), the top *m/z* and many of the secondary *m/z* for the true positives have noticeably higher F-ratios than the false positive *m/z* and all 14 spiked analytes are accounted for as 14 distinguishable red dots. Conversely, for the 3 ppm versus neat comparison in Fig. 3.4(B), only 10 of the 14 red dots for the top F-ratio *m/z* are observed, and substantially fewer secondary *m/z* have a higher F-ratio than many of the false positive *m/z*. For the 1.5 ppm versus neat

comparison in Fig. 3.4(C), an additional top F-ratio *m/z* has disappeared and only 9 out of 14 are observed, and the secondary *m/z* have even lower F-ratios than many of the false positive *m/z*. Using 2,5-dimethylthiophene as an example (Fig. 3.3(C)), its secondary F-ratio *m/z* have log(F-ratio) values of ~1.4 and less, which fall well below the large cluster of false positive F-ratio *m/z* with log(F-ratio) values from 0.5 to 2 (Fig. 3.4(C)). Overall, the evidence in Fig. 3.4 indicates that the spiked sulfur-containing analytes have considerably fewer highly sensitive *m/z* as the spike concentration decreases from 15 ppm to 3 ppm to 1.5 ppm. The inferior contribution of these small F-ratio *m/z* to the average F-ratio calculation could lower the average F-ratios of true positives below those of false positives and impact their hitlist rankings. The utility of using only the top F-ratio *m/z* to rank the hitlists is evident, since the top F-ratio *m/z* (i.e., red dots) have F-ratios distinguishable from the false positive F-ratio *m/z*. For the remainder of this report, we will discuss the success of said approach to rank the hitlists using the top F-ratio *m/z* compared to "standard" F-ratio methodology (i.e., averaging the top 10, minimum of 3, F-ratio *m/z*) in terms of hit numbers, rather than F-ratios, for the sake of clarity.

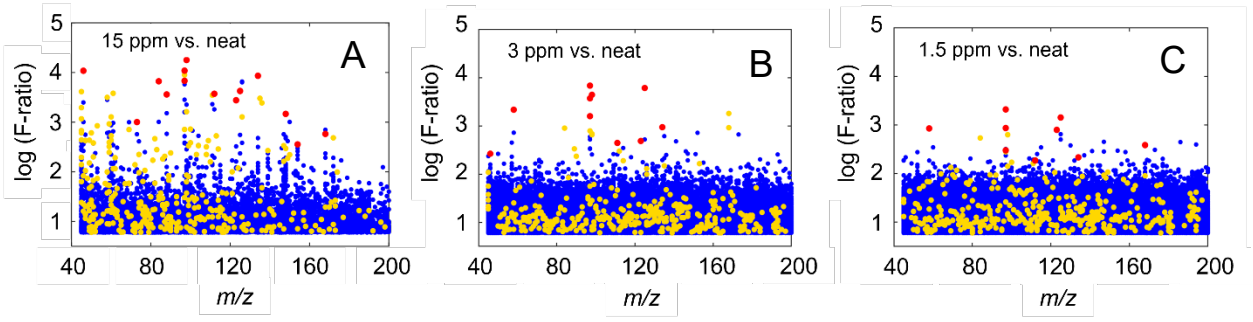

Figure 3.4. Plots of the logarithm of F-ratio calculated at each *m/z* from each of the following F-ratio comparisons: (A) 15 ppm versus neat comparison. (B) 3 ppm versus neat comparison. (C) 1.5 ppm versus neat comparison. The F-ratios must clear the *S/N* threshold and $F_{crit} = 5.99$. The red dots represent the top F-ratio for each sulfur-containing compound "discovered," while the yellow dots represent their secondary F-ratios. The blue dots represent F-ratios from the JP8 fuel "background."

The hitlists generated using the standard F-ratio methodology and only the top F-ratio *m/z* are provided in Tables 3.2 and 3.3, respectively. The F-ratios are indicated in parentheses beside the hit numbers. Note that the analytes are ordered in Tables 3.2 and 3.3 according to the 15 ppm versus neat hitlist rankings using the average F-ratios (first column of Table 3.2). Although further discussion of the hitlists will focus on hit number rankings, rather than F-ratios, we direct the reader to Fig. B.2 in Appendix B for a comparison of the average F-ratio distributions to the top F-ratio distributions for each of the concentration comparisons. For the 15 ppm versus neat comparison, all 14 sulfur-containing compounds were discovered in the top ~30 hits (Tables 3.2-3.3). The rankings of individual analytes are largely unchanged in Table 3.3 relative to Table 3.2, with minor variations resulting from simple re-orderings (i.e., 1,4-oxathiane goes from hit 2 to hit 3, while 2-propylthiophene goes from hit 5 to hit 2). However, for the 3 ppm versus neat comparison, only 10 of the 14 sulfur-containing compounds were discovered, with the other 4 compounds becoming undiscoverable (labeled "NF" in Tables 3.2-3.3) due to their insufficient mass spectral selectivity and/or sensitivity resulting in an F-ratio < F_{crit}. When the average F-ratios are used for ranking, 1,4-oxathiane experiences the most drastic change in hit number as concentration decreases: hit 2 in the 15 ppm versus neat comparison to hit 114 in the 3 ppm versus neat comparison (Table 3.2). However, when the top F-ratios are used to rank the hitlist instead, 1,4-oxathiane improves from hit 114 to hit 25 at the 3 ppm versus neat comparison (Table 3.3). At the lower concentration comparison of 1.5 ppm versus neat, 1,4-oxathiane joins the other 4 undiscoverable analytes from the 3 ppm versus neat comparison (Tables 3.2-3.3). Using the average F-ratios to rank the 1.5 ppm versus neat hitlist, 3 different analytes (2-propylthiophene, benzo[b]thiophene, and 2,5-dimethylthiophene) exhibit noticeable decreases in hit number from 3 to 59, 8 to 98, and 10 to 262, respectively, relative to the 3 ppm versus neat

comparison (Table 3.2). Yet, when only the top F-ratios are used for ranking, 2-propytlhiophene improves from hit 59 to 17; benzo[b]thiophene from hit 98 to 28; and 2,5-dimethylthiophene from hit 262 to 39 (Table 3.3). Thus, the results presented in Table 3.3 confirm that utilizing only the top F-ratio *m/z* for hitlist ranking is useful for maximizing discoverability as analyte concentrations approach *DL*. However, further examination is needed to uncover the relationship between this optimal hitlist ranking approach using the top F-ratio *m/z* and the analyte *LOD* and *LOQ*. We explore this issue next by putting the analysis into the context of the *S/N* of the analytes using their top F-ratio *m/z*.

Table 3.2. Hitlists generated from standard F-ratio methodology for the following comparisons: 15 ppm versus neat, 3 ppm versus neat, and 1.5 ppm versus neat. A minimum of 3 *m/z* and a maximum of 10 *m/z* that cleared the *S/N* threshold of $10\sigma_{,BN}$ were used in the average F-ratio calculation. The optimum tile size of 12 s on ^{1}D and 300 ms on ^{2}D was utilized. Some compounds were not found (*NF*) on the given hitlist, due to F-ratio < F_{crit}.

Analyte	Hit Number (Avg F-ratio), 15 ppm vs. neat	Hit Number (Avg F-ratio), 3 ppm vs. neat	Hit Number (Avg F-ratio), 1.5 ppm vs. neat
2-methylthiophene	1 (2953)	2 (748)	4 (199)
1,4-oxathiane	2 (2448)	114 (37)	*NF (N/A)*
benzo[b]thiophene	3 (2208)	8 (194)	98 (49)
thiophene	4 (1520)	5 (631)	1 (284)
2-propylthiophene	5 (1369)	3 (703)	59 (60)
2,5-dimethylthiophene	6 (1091)	10 (111)	262 (33)
3-methylthiophene	9 (1040)	6 (249)	3 (223)
tetrahydrothiophene	13 (625)	*NF (N/A)*	*NF (N/A)*
3-methylbenzothiophene	15 (603)	*NF (N/A)*	*NF (N/A)*
2-chloroethyl phenyl sulfide	17 (483)	19 (78)	14 (100)
2-butyl-5-ethylthiophene	18 (454)	1 (750)	2 (250)
2-methylbenzothiophene	21 (311)	*NF(N/A)*	*NF (N/A)*
2-hexylthiophene	22 (275)	4 (649)	11 (102)
3-acetyl-2,5-dimethylthiophene	29 (129)	*NF (N/A)*	*NF (N/A)*

Table 3.3. Hitlists generated when only the top F-ratio *m/z* is used to rank the hitlist for the following comparisons: 15 ppm versus neat, 3 ppm versus neat, and 1.5 ppm versus neat. The optimum tile size of 12 s on ^{1}D and 300 ms on ^{2}D was utilized. Some compounds were not found (*NF*) on the given hitlist, due to F-ratio < F_{crit}. The analytes are listed top to bottom based on the ordering shown in Table 3.2 for the 15 ppm versus neat comparison.

Analyte	Hit Number (Top F-ratio), 15 ppm vs. neat	Hit Number (Top F-ratio), 3 ppm vs. neat	Hit Number (Top F-ratio), 1.5 ppm vs. neat
2-methylthiophene	1 (17880)	3 (4403)	3 (1386)
1,4-oxathiane	3 (10861)	25 (264)	*NF (N/A)*
benzo[b]thiophene	4 (8601)	7 (937)	28 (215)
thiophene	6 (6669)	5 (2158)	4 (848)
2-propylthiophene	2 (10905)	1 (6810)	17 (301)
2,5-dimethylthiophene	12 (3738)	19 (442)	39 (186)
3-methylthiophene	5 (6817)	6 (1856)	1 (2085)
tetrahydrothiophene	13 (3621)	*NF (N/A)*	*NF (N/A)*
3-methylbenzothiophene	19 (1470)	*NF (N/A)*	*NF (N/A)*
2-chloroethyl phenyl sulfide	15 (2769)	16 (485)	5 (792)
2-butyl-5-ethylthiophene	11 (4252)	2 (6082)	2 (1423)
2-methylbenzothiophene	23 (1006)	*NF (N/A)*	*NF (N/A)*
2-hexylthiophene	27 (579)	4 (3715)	11 (385)
3-acetyl-2,5-dimethylthiophene	34 (357)	*NF (N/A)*	*NF (N/A)*

Preparation of summed ^{2}D peaks to assess analyte *S/N* is demonstrated in Fig. 3.5 for 2,5-dimethylthiophene. For each concentration comparison, summed ^{2}D peaks were generated by defining a 2D chromatographic region of 12 s on ^{1}D (i.e., F-ratio tile dimension) and a sufficient time window to capture the analyte signal on ^{2}D, centered on the F-ratio pin location (^1t_R and ^2t_R) of each analyte. As an example, a ^{2}D window of 400 ms was used for 2,5-dimethylthiophene, but varying ^{2}D windows were selected for other analytes depending on their ^{2}D peak width. All the ^{2}D peaks for each analyte were summed by adding the ^{1}D modulations using the top F-ratio *m/z* of the lowest concentration comparison in which the given analyte was observed on the hitlist. Background correction of the JP8 fuel data at the given *m/z* was performed, as illustrated in Figs. 3.5(A,B) with the summed ^{2}D peaks for 2,5-dimethylthiophene at *m/z* 112 from the 15 ppm versus neat comparison before, and after, background correction, respectively. The corresponding summed ^{2}D peaks before, and after, background correction for the 3 ppm versus neat replicates are provided in Figs. 3.5(C,D), while the analogous summed ^{2}D peaks for 1.5 ppm versus neat are provided in Figs. 3.5(E,F), respectively.

The summed ^{2}D peak heights of the spiked fuel samples at known injected concentration (eg., the red traces in Figs. 3.5(B,D,F)) were used to calculate the *LOD* and *LOQ* for each analyte. The *LOD* is defined as an analyte concentration that yields a signal equal to three times the standard deviation of the baseline noise (σ_n), i.e., signal at $LOD = 3\sigma_n$, and the *LOQ* is defined as an analyte concentration that yields a signal equal to ten times the σ_n, signal at $LOQ = 10\sigma_n$. The standard deviations, σ_n, of the background-corrected neat peaks which are nominally baseline (the blue traces in Figs. 3.5(B,D,F)) were computed and averaged prior to the *LOD* and *LOQ* calculations. The *LOD* and *LOQ* determinations were performed using the lowest concentration comparison in which the given analyte was observed on the hitlist. This

corresponds to 1.5 ppm for 2,5-dimethylthiophene, at which concentration the average summed ^{2}D peak height exceeds the calculated *LOD* of 0.40 ppm by 4-fold but only barely exceeds the calculated *LOQ* of 1.3 ppm by a factor of 1.1 (Fig. 3.5(F)). Thus, for this example, the *DL* appears to be closely related to the *LOQ*, as 2,5-dimethylthiophene suffers a drastic decrease in hitlist ranking going from the 3 ppm versus neat comparison to the 1.5 ppm versus neat comparison (hit 19 to hit 39 in Table 3.3).

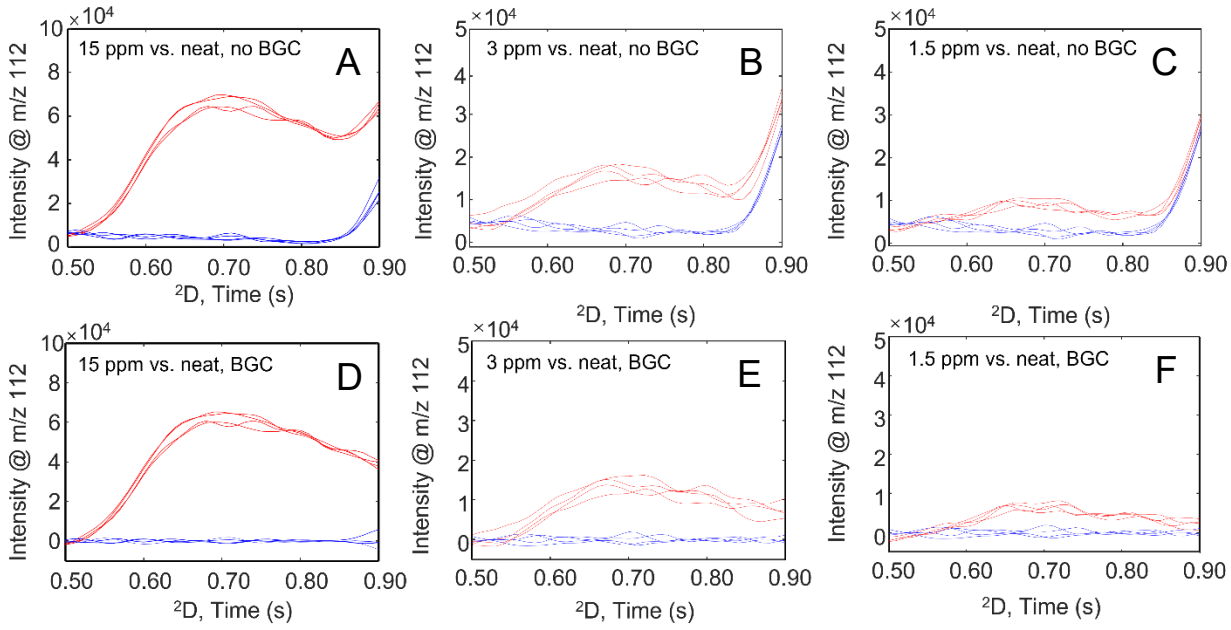

Figure 3.5. Summed ^{2}D peaks for the replicates of 2,5-dimethylthiophene: spiked (red) and neat (blue). (A) ^{2}D peaks for 15 ppm (red) and neat (blue), prepared by summing all the ^{2}D peaks along ^{1}D separation axis. (B) Background corrected (BGC) summed ^{2}D peaks generated by subtracting an average neat vector (blue) from all 15 ppm and neat replicates. (C) Original summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (D) BGC summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (E) Original summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). (F) BGC summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue).

An analogous set of figures for the analyte 1,4-oxathiane at *m/z* 46 are provided in Fig. 3.6. Due to 1,4-oxathiane having a wider ^{2}D peak width (^2W_b) than 2,5-dimethylthiophene at 15 ppm, a ^{2}D window of 830 ms was used for summed ^{2}D peak preparation (Fig. 3.6(A)) and background correction (Fig. 3.6(B)). 1,4-oxathiane also exhibits a notable decrease in ^2W_b as concentration decreases to 3 ppm, so a narrower ^{2}D window of 730 ms (versus 830 ms) was used for summed ^{2}D peak preparation (Fig. 3.6(C)) and background correction of the 3 ppm versus neat replicates (Fig. 3.6(D)) relative to the 15 ppm versus neat replicates. The need to background correct the summed ^{2}D peaks is especially evident for 1,4-oxathiane at 3 ppm, as a peak maximum cannot be readily measured due to the interferent peak (Fig. 3.6(C)), unlike 2,5-dimethylthiophene at the same concentration (Fig. 3.5(C)). The analogous before, and after, background correction figures for 1,4-oxathiane from the 1.5 ppm versus neat comparison are provided in Figs. 3.6(E,F), respectively. The lowest concentration comparison in which 1,4-oxathiane was observed on the hitlist was 3 ppm, whereby it exhibits a considerable decrease in hitlist ranking compared to the 15 ppm versus neat comparison (hit 3 to hit 25 in Table 3.3). In agreeance with the trend observed for 2,5-dimethylthiophene, the average ^{2}D peak height of 1,4-oxathiane at 3 ppm exceeds the calculated *LOD* of 0.76 ppm by 4-fold but only surpasses the calculated *LOQ* of 2.5 ppm by a factor of 1.2 (Fig. 3.6(D)). The average ^{2}D peak height at 1.5 ppm (Fig. 3.6(F)) falls well below the calculated *LOQ* and thus further highlights a probable link between *DL* and *LOQ*, since 1,4-oxathiane is not discovered for the 1.5 ppm versus neat comparison.

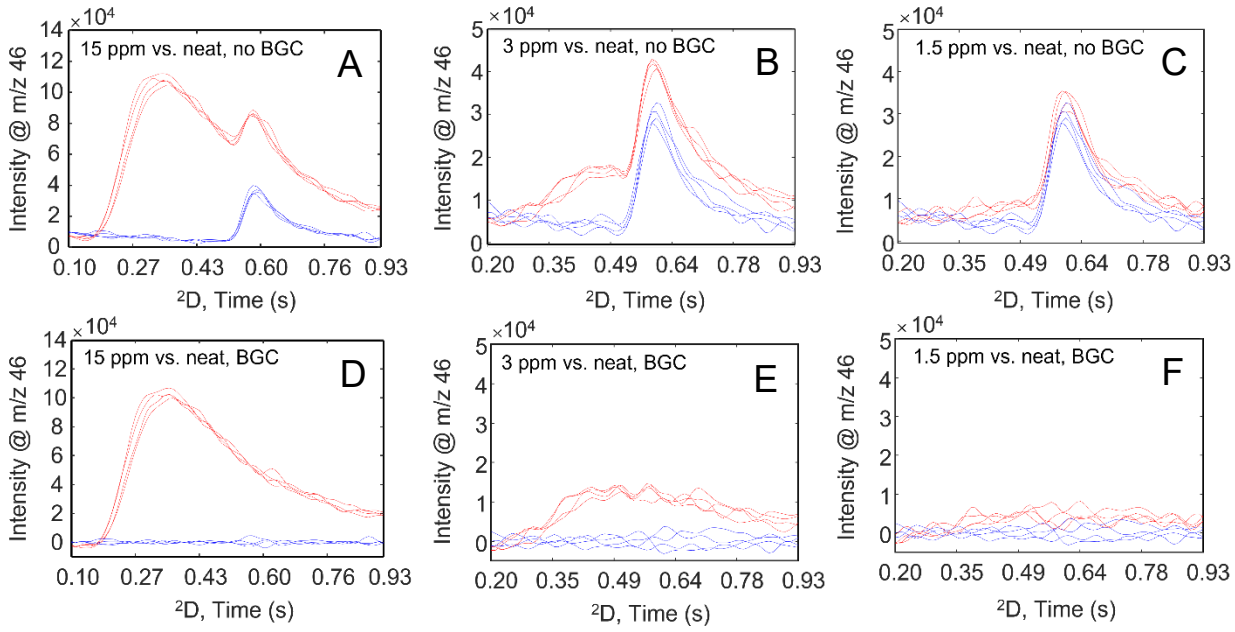

Figure 3.6. Summed ^{2}D peaks for the replicates of 1,4-oxathiane: spiked (red) and neat (blue). (A) ^{2}D peaks for 15 ppm (red) and neat (blue), prepared by summing all the ^{2}D peaks along ^{1}D separation axis. (B) Background corrected (BGC) summed ^{2}D peaks generated by subtracting an average neat vector (blue) from all 15 ppm and neat replicates. (C) Original summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (D) BGC summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (E) Original summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). (F) BGC summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue).

Additional evidence connecting the *DL*, *LOD* and *LOQ*, is provided in Table 3.4, which summarizes the *LOD* and *LOQ* calculations for all 14 sulfur-containing analyte compounds, reported to 2 significant figures. Indeed, for every analyte listed in Table 3.4, if its *LOQ* is greater than the spike concentration, then that analyte does not appear on the hitlist for that spike level. For example, the four analytes which were only discoverable in the 15 ppm versus neat comparison (3-methylbenzothiophene, tetrahydrothiophene, 2-methylbenzothiophene, and 3-acetyl-2,5-dimethylthiophene) had a *LOQ* ranging from 4.1-7.0 ppm, which all fall above the next lowest concentration level studied (3 ppm versus neat). Conversely their *LOD* range from

1.2-2.1 ppm, so if the *DL* is dictated by the *LOD*, then all four analytes should have been discovered with the 3 ppm versus neat F-ratio comparison, and only 2-methylbenzothiophene and 3-acetyl-2,5-dimethylthiophene would have not been discovered with the 1.5 ppm versus neat F-ratio comparison (Table 3.4). An additional F-ratio comparison of the 3 ppm versus 1.5 ppm classes identified the 10 hits which were found in the 3 ppm versus neat hitlist, which lends further credence to a relationship between *LOQ* and *DL* (results not shown for brevity). The summed ^{2}D peaks, along with discussion of how the calculated *LOD* and *LOQ* relate to the ^{2}D peak heights, for 2-propylthiophene, benzo[b]thiophene, and 3-methylbenzothiophene is provided in Appendix B as Figs. B.3-B.5, respectively.

Table 3.4. Calculation of *LOD* and *LOQ* to 2 significant digits for the 14 sulfur-containing analyte compounds based on peak heights in summed ^{2}D peaks.

Analyte	Quantification *m/z*	Lowest concentration F-ratio comparison found	*LOD* (ppm)	*LOQ* (ppm)
2-methylthiophene	97	1.5 ppm versus neat	0.15	0.50
1,4-oxathiane	46	3 ppm versus neat	0.76	2.5
benzo[b]thiophene	134	1.5 ppm versus neat	0.34	1.1
thiophene	58	1.5 ppm versus neat	0.12	0.40
2-propylthiophene	97	1.5 ppm versus neat	0.19	0.64
2,5-dimethylthiophene	112	1.5 ppm versus neat	0.40	1.3
3-methylthiophene	97	1.5 ppm versus neat	0.20	0.68
tetrahydrothiophene	88	15 ppm versus neat	1.2	4.1
3-methylbenzothiophene	148	15 ppm versus neat	1.3	4.2
2-chloroethyl phenyl sulfide	123	1.5 ppm versus neat	0.36	1.2
2-butyl-5-ethylthiophene	125	1.5 ppm versus neat	0.087	0.29
2-methylbenzothiophene	73	15 ppm versus neat	1.6	5.2
2-hexylthiophene	168	1.5 ppm versus neat	0.094	0.31
3-acetyl-2,5-dimethylthiophene	154	15 ppm versus neat	2.1	7.0

To summarize the analyte discoverability via F-ratio analysis, the F-ratio hitlist rank numbers obtained using the standard average F-ratio methodology (orange) and the top F-ratio methodology (blue) are plotted in Fig. 3.7 as a function of the *LOD* and *LOQ* for the 10 analytes that are discovered in the 3 ppm versus neat comparison (Fig. 3.7(A)), as well as the 9 analytes that are discovered in the 1.5 ppm versus neat comparison (Fig. 3.7(B)). The following observations can be made. First, at spike level concentrations approaching the analyte *LOQ*, ranking the F-ratio hitlist using only the top F-ratio *m/z* is superior to the standard implementation using the average F-ratio. This is illustrated by the orange dots falling well above the blue dots (i.e., being further down the hitlist) for 1,4-oxathiane at 3 ppm versus neat (Fig. 3.7(A)) and for 2-propylthiophene, benzo[b]thiophene, and 2,5-dimethylthiophene at 1.5 ppm versus neat (Fig. 3.7(B)). Furthermore, a clear relationship between the *LOD* and *LOQ* and hit number is observed; as *LOD* and *LOQ* increase, analytes fall further down the F-ratio hitlist, hence discoverability *decreases*. Given that the tile-based F-ratio algorithm requires a sufficient signal difference for analytes to have sizeable F-ratios, this is not necessarily surprising. However, the trend in Fig. 3.7 suggests that the discovery limit *DL* appears to be roughly equivalent to the *LOQ*, well above the *LOD*.

We now demonstrate how the tile size, and hence the cluster window size, impacts F-ratio discoverability. For this exercise, F-ratio hitlist rank numbers using the top F-ratio *m/z* were obtained with a constant ^{1}D tile dimension of 12 s (4 modulations), while the ^{2}D tile dimension was varied: optimal at 300 ms, too large at 600 ms, and too small at 100 ms. Results are plotted in Fig. 3.8 as a function of the *LOD* and *LOQ* for the 10 analytes discovered in the 3 ppm versus neat comparison (Fig. 3.8(A)), as well as the 9 analytes discovered in the 1.5 ppm versus neat comparison (Fig. 3.8(B)). Evidence that the 300 ms tile dimension on ^{2}D is superior for ranking

the F-ratio hitlist is provided by the ranking of 1,4-oxathiane and 2,5-dimethylthiophene in the 3 ppm versus neat comparison (Fig. 3.8(A)) and for 2,5-dimethylthiophene and 2-chloroethyl phenyl sulfide in the 1.5 ppm versus neat comparison (Fig. 3.8(B)), as the blue dots fall below the purple and green dots, thus are further up the hitlist. Notably, 2-chloroethyl phenyl sulfide is not discovered in the 1.5 ppm versus neat hitlist when a 600 ms ^{2}D tile dimension is used because it is drowned out by interferents in the excessively large F-ratio tile (Fig. 3.8(B)). Therefore, the results presented in Fig. 3.8 underscore the importance of proper tile selection for improving analyte discoverability. The detailed hitlists obtained using the 600 ms and 100 ms ^{2}D tile dimensions are provided in Appendix B as Table B.3-B.4.

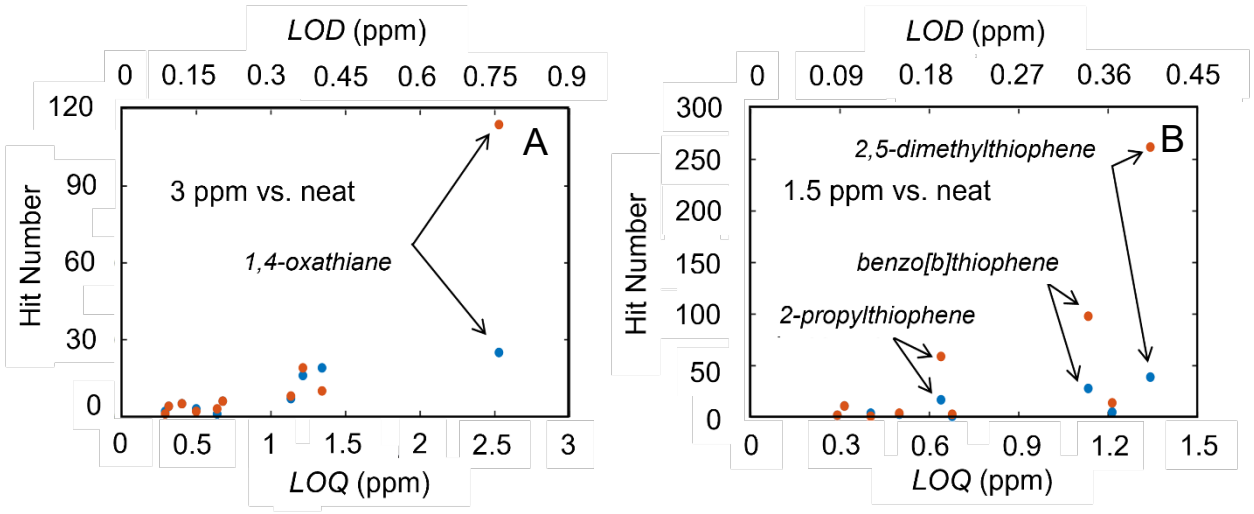

Figure 3.7. True positive hit numbers obtained using standard F-ratio methodology (orange) and the top F-ratio (blue) to rank the respective F-ratio hitlists, relative to the analyte *LOD* and *LOQ* from Table 3.4. The ^{1}D tile dimension was 12 s (4 modulations), with a ^{2}D tile dimension of 300 ms. (A) 3 ppm versus neat comparison. (B) 1.5 ppm versus neat comparison.

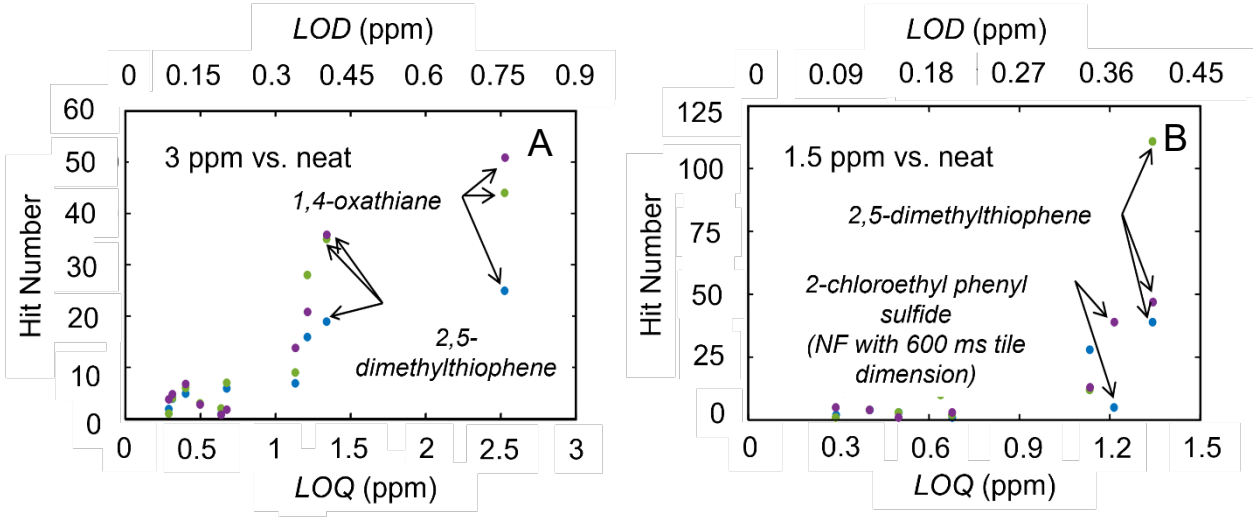

Figure 3.8. True positive hit numbers obtained using a ^{2}D tile dimension of 300 ms (blue), 600 ms (green), and 100 ms (purple) while holding the ^{1}D tile dimension constant at 12 s (4 modulations), relative to the analyte *LOD* and *LOQ* from Table 3.4. (A) 3 ppm versus neat comparison. (B) 1.5 ppm versus neat comparison.

3.4 CONCLUSIONS

Performance of supervised discovery-based analysis was examined for a JP8 jet fuel spiked with low levels of sulfur-containing compounds. Optimal parameters for tile-based F-ratio analysis improve the ranking of analytes approaching the discovery limit (*DL*). Specifically, using only the top F-ratio *m/z* to rank hits resulted in a marked improvement in discoverability relative to the previously applied "standard" F-ratio approach using an average F-ratio. Additionally, for every analyte, when its *LOQ* was greater than the spike concentration, then that analyte did not appear on the hitlist for that spike level. The advantage of using the top F-ratio *m/z* for hitlist ranking ultimately depends on a balance of *m/z* selectivity and sensitivity, but this is beyond the scope of the current study. Additional hitlist comparisons performed with varying the ^{2}D tile dimensions indicated the 300 ms ^{2}D tile dimension was superior for improving discoverability, given the maximum ^2W_b observed for the native fuel components. The results of this study serve to validate the performance of the tile-based F-ratio software, while providing guidance for optimizing the discoverability of analytes approaching their *LOQ* via key software parameters. F-ratio analysis has been demonstrated to be broadly applicable to numerous complex sample matrices, even in the presence of large within-class variance [13,19,35–38,40]. Thus, the demonstrated ability to perform the simple software optimization steps holds great promise for improving discoverability in F-ratio studies involving numerous sources of background variability.

3.5 REFERENCES

[1] W. Genuit, H. Chaabani, Comprehensive two-dimensional gas chromatography-field ionization time-of-flight mass spectrometry (GCxGC-FI-TOFMS) for detailed hydrocarbon middle distillate analysis, Int. J. Mass Spectrom. 413 (2017) 27–32. https://doi.org/10.1016/j.ijms.2016.12.001.

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

[3] 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. https://doi.org/10.1021/ef501247h.

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

[5] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications, Anal. Chem. (2017). https://doi.org/10.1021/acs.analchem.7b04226.

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

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

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

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

[10] Zaiyou. Liu, D.G. Patterson, M.L. Lee, Geometric Approach to Factor Analysis for the Estimation of Orthogonality and Practical Peak Capacity in Comprehensive Two-Dimensional Separations, Anal. Chem. 67 (1995) 3840–3845. https://doi.org/10.1021/ac00117a004.

[11] K.L. Berrier, S.E. Prebihalo, R.E. Synovec, Chapter 7 - Advanced data handling in comprehensive two-dimensional gas chromatography, in: N.H. Snow (Ed.), Sep. Sci. Technol., Academic Press, 2020: pp. 229–268. https://doi.org/10.1016/B978-0-12-813745-1.00007-6.

[12] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A 1255 (2012) 3–11. https://doi.org/10.1016/j.chroma.2012.05.050.

[13] B.J. Pollo, C.A. Teixeira, J.R. Belinato, M.F. Furlan, I. Cristina de Matos Cunha, C.R. Vaz, G.V. Volpato, F. Augusto, Chemometrics, Comprehensive Two-Dimensional Gas Chromatography And "Omics" Sciences: Basic Tools And Recent Applications, TrAC Trends Anal. Chem. (2020) 116111. https://doi.org/10.1016/j.trac.2020.116111.
[14] S. Li, Y. Hu, W. Liu, Y. Chen, F. Wang, X. Lu, W. Zheng, Untargeted volatile metabolomics using comprehensive two-dimensional gas chromatography-mass spectrometry – A solution for orange juice authentication, Talanta 217 (2020) 121038. https://doi.org/10.1016/j.talanta.2020.121038.
[15] Z. Ye, Z. Shang, M. Li, Y. Qu, H. Long, J. Yi, Evaluation of the physiochemical and aromatic qualities of pickled Chinese pepper (Paojiao) and their influence on consumer acceptability by using targeted and untargeted multivariate approaches, Food Res. Int. 137 (2020) 109535. https://doi.org/10.1016/j.foodres.2020.109535.
[16] Y.-Y. Zhang, Q. Zhang, Y.-M. Zhang, W.-W. Wang, L. Zhang, Y.-J. Yu, C.-C. Bai, J.-Z. Guo, H.-Y. Fu, Y. She, A comprehensive automatic data analysis strategy for gas chromatography-mass spectrometry based untargeted metabolomics, J. Chromatogr. A 1616 (2020) 460787. https://doi.org/10.1016/j.chroma.2019.460787.
[17] X.-D. Sun, H.-L. Wu, Z. Liu, Y. Chen, J.-C. Chen, L. Cheng, Y.-J. Ding, R.-Q. Yu, Target-based metabolomics for fast and sensitive quantification of eight small molecules in human urine using HPLC-DAD and chemometrics tools resolving of highly overlapping peaks, Talanta 201 (2019) 174–184. https://doi.org/10.1016/j.talanta.2019.03.090.
[18] T. Vrzal, J. Olšovská, Pyrolytic profiling nitrosamine specific chemiluminescence detection combined with multivariate chemometric discrimination for non-targeted detection and classification of nitroso compounds in complex samples, Anal. Chim. Acta 1059 (2019) 136–145. https://doi.org/10.1016/j.aca.2019.01.033.
[19] B.A. Parsons, D.K. Pinkerton, B.W. Wright, R.E. Synovec, Chemical characterization of the acid alteration of diesel fuel: Non-targeted analysis by two-dimensional gas chromatography coupled with time-of-flight mass spectrometry with tile-based Fisher ratio and combinatorial threshold determination, J. Chromatogr. A 1440 (2016) 179–190. https://doi.org/10.1016/j.chroma.2016.02.067.
[20] C.E. Freye, N.R. Moore, R.E. Synovec, Enhancing the chemical selectivity in discovery-based analysis with tandem ionization time-of-flight mass spectrometry detection for comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1537 (2018) 99–108. https://doi.org/10.1016/j.chroma.2018.01.008.
[21] K.J. Johnson, R.E. Synovec, Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis, Chemom. Intell. Lab. Syst. 60 (2002) 225–237. https://doi.org/10.1016/S0169-7439(01)00198-8.
[22] L. Ugena, S. Moncayo, S. Manzoor, D. Rosales, J.O. Cáceres, Identification and Discrimination of Brands of Fuels by Gas Chromatography and Neural Networks Algorithm in Forensic Research, J. Anal. Methods Chem. 2016 (2016). https://doi.org/10.1155/2016/6758281.
[23] P. Rearden, P.B. Harrington, J.J. Karnes, C.E. Bunker, Fuzzy Rule-Building Expert System Classification of Fuel Using Solid-Phase Microextraction Two-Way Gas Chromatography Differential Mobility Spectrometric Data, Anal. Chem. 79 (2007) 1485–1491. https://doi.org/10.1021/ac060527f.
[24] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of

comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.
[25] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.
[26] A.H. Hegazi, J.T. Andersson, 6 - Polycyclic aromatic sulfur heterocycles as source diagnostics of petroleum pollutants in the marine environment, in: S.A. Stout, Z. Wang (Eds.), Stand. Handb. Oil Spill Environ. Forensics Second Ed., Academic Press, Boston, 2016: pp. 313–342. https://doi.org/10.1016/B978-0-12-803832-1.00006-4.
[27] A. Stanislaus, A. Marafi, M.S. Rana, Recent advances in the science and technology of ultra low sulfur diesel (ULSD) production, Catal. Today 153 (2010) 1–68. https://doi.org/10.1016/j.cattod.2010.05.011.
[28] O. US EPA, Diesel Fuel Standards and Rulemakings, US EPA (2015). https://www.epa.gov/diesel-fuel-standards/diesel-fuel-standards-and-rulemakings (accessed October 1, 2020).
[29] Y. Han, Y. Zhang, C. Xu, C.S. Hsu, Molecular characterization of sulfur-containing compounds in petroleum, Fuel 221 (2018) 144–158. https://doi.org/10.1016/j.fuel.2018.02.110.
[30] M.T. Timko, E. Schmois, P. Patwardhan, Y. Kida, C.A. Class, W.H. Green, R.K. Nelson, C.M. Reddy, Response of Different Types of Sulfur Compounds to Oxidative Desulfurization of Jet Fuel, Energy Fuels 28 (2014) 2977–2983. https://doi.org/10.1021/ef500216p.
[31] V.V. Lobodin, W.K. Robbins, J. Lu, R.P. Rodgers, Separation and Characterization of Reactive and Non-Reactive Sulfur in Petroleum and Its Fractions, Energy Fuels 29 (2015) 6177–6186. https://doi.org/10.1021/acs.energyfuels.5b00780.
[32] N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset, J. Chromatogr. A 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.
[33] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.
[34] B.C. Reaser, B.W. Wright, R.E. Synovec, Using Receiver Operating Characteristic Curves To Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry, Anal. Chem. 89 (2017) 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991.
[35] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.0c03456.
[36] A.P. de la Mata, R.H. McQueen, S.L. Nam, J.J. Harynuk, Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics, Anal. Bioanal. Chem. 409 (2017) 1905–1913. https://doi.org/10.1007/s00216-016-0137-1.
[37] S. Carlin, U. Vrhovsek, P. Franceschi, C. Lotti, L. Bontempo, F. Camin, D. Toubiana, F. Zottele, G. Toller, A. Fait, F. Mattivi, Regional features of northern Italian sparkling wines,

identified using solid-phase micro extraction and comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry, Food Chem. 208 (2016) 68–80. https://doi.org/10.1016/j.foodchem.2016.03.112.
[38] I. Lukić, S. Carlin, I. Horvat, U. Vrhovsek, Combined targeted and untargeted profiling of volatile aroma compounds with comprehensive two-dimensional gas chromatography for differentiation of virgin olive oils according to variety and geographical origin, Food Chem. 270 (2019) 403–414. https://doi.org/10.1016/j.foodchem.2018.07.133.
[39] R.A. Fisher, A mathematical Examination of the Methods of determining the Accuracy of Observation by the Mean Error, and by the Mean Square Error, Mon. Not. R. Astron. Soc. 80 (1920) 758–770. https://doi.org/10.1093/mnras/80.8.758.
[40] E.M. Humston, J.D. Knowles, A. McShea, R.E. Synovec, Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics, J. Chromatogr. A 1217 (2010) 1963–1970. https://doi.org/10.1016/j.chroma.2010.01.069.
[41] 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. https://doi.org/10.1016/j.chroma.2009.02.030.
[42] A.C. Beckstrom, E.M. Humston, L.R. Snyder, R.E. Synovec, S.E. Juul, Application of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method to identify potential biomarkers of perinatal asphyxia in a non-human primate model, J. Chromatogr. A 1218 (2011) 1899–1906. https://doi.org/10.1016/j.chroma.2011.01.086.
[43] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Fisher Ratio Method Applied to Third-Order Separation Data To Identify Significant Chemical Components of Metabolite Extracts, Anal. Chem. 78 (2006) 5068–5075. https://doi.org/10.1021/ac0602625.
[44] P.-H. Stefanuto, K.A. Perrault, L.M. Dubois, B. L'Homme, C. Allen, C. Loughnane, N. Ochiai, J.-F. Focant, Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1507 (2017) 45–52. https://doi.org/10.1016/j.chroma.2017.05.064.
[45] S.E. Reichenbach, C.A. Zini, K.P. Nicolli, J.E. Welke, C. Cordero, Q. Tao, Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition, J. Chromatogr. A 1595 (2019) 158–167. https://doi.org/10.1016/j.chroma.2019.02.027.
[46] S.E. Reichenbach, X. Tian, Q. Tao, E.B. Ledford, Z. Wu, O. Fiehn, Informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography and high-resolution mass spectrometry (GCxGC–HRMS), Talanta 83 (2011) 1279–1288. https://doi.org/10.1016/j.talanta.2010.09.057.
[47] S.E. Prebihalo, D.K. Pinkerton, R.E. Synovec, Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on data trilinearity and parallel factor analysis deconvolution, J. Chromatogr. A 1605 (2019) 460368. https://doi.org/10.1016/j.chroma.2019.460368.
[48] D.K. Pinkerton, B.A. Parsons, T.J. Anderson, R.E. Synovec, Trilinearity deviation ratio: A new metric for chemometric analysis of comprehensive two-dimensional gas chromatography

time-of-flight mass spectrometry data, Anal. Chim. Acta 871 (2015) 66–76. https://doi.org/10.1016/j.aca.2015.02.040.

Chapter 4. Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis

This chapter was reproduced from Paige E. Sudol[†], Micaela Galletta[†], Peter Q. Tranchida, Mariosimone Zoccali, Luigi Mondello, Robert E. Synovec, "Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis" *Journal of Chromatography A* 1662 (2022) 462735.

[†]These authors contributed equally to this work.

4.1. INTRODUCTION

The introduction of comprehensive two-dimensional gas chromatography (GC×GC), an event which occurred 30 years ago, can be considered as one of the most important evolutions in the field of gas chromatography [1]. Nowadays, GC×GC can be considered as a well-established technique, widely employed also for routine analysis [2]. In brief, GC×GC is performed on two analytical columns connected in series by means of a modulator; the latter entraps sequential effluent fractions from the first-dimension (^{1}D) column, and then reinjects them onto a second-dimension (^{2}D) column, where a different chemical composition of the stationary phase for the two columns is exploited [3]. Several modulators are available on the market, mainly divided into cryogen thermal modulators and valve-based modulators [4], with the latter gaining popularity for their structural robustness and reduced economic costs [5].

Typically, GC×GC combined with time-of-flight mass spectrometry (TOFMS) is used for the separation and detection of complex samples, generating a high quantity of data. The amount of data becomes extremely large when many samples, and replicates of each sample, are analyzed. For such a reason, GC×GC-TOFMS data handling has become very demanding, with related

evolution occurring in recent years with the context of dedicated software [6–8]. The acquired data should be transformed into useful information after data processing; for such a purpose, chemometric methods can be exploited. In such a manner, different aims can be addressed, such as, sample characterization/differentiation, identification of key components, or to correlate chemical measurements to other properties of the samples [6]. By implementing an untargeted approach, several type of features can be investigated, such as tiles, datapoints, regions, and peak-regions [9].

Using either a supervised or unsupervised experimental design, GC×GC-TOFMS data collection followed by untargeted analysis is increasingly used for the analysis of a wide variety of chemical systems such as the volatile fraction of foods and beverages [10–18]. The distinction is that for a supervised experimental design sample class membership is known *a priori*, and a common data analysis method is to apply Fisher ratio analysis (F-ratio analysis), while for an unsupervised experimental design sample classes are not known, and the data can be examined by principal component analysis (PCA) to look for the presence of sample classes. We note that it is also common to apply F-ratio analysis, to find the class-distinguishing analyte features, then follow up with the use of PCA as a tool to visualize the success of the F-ratio analysis.

The study of wine aroma, which is the objective of this study, can be readily pursued using F-ratio analysis, since the aroma is extremely complex due to the presence of several classes of compounds. In fact, more than 1000 aroma constituents have been identified, covering a wide range of both polarities and volatilities [19]. The characteristic components of the wine aroma are commonly divided into three classes and relate to the geographical: grape (or varietal) aroma, fermentation aroma and aging aroma. However, these classes are not so clear-divided, most of them originate from grapes and are modified by the fermentation process or aging [20]. In the

present research, headspace solid-phase microextraction (HS SPME) coupled to flow-modulated comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (FM GC×GC-TOFMS) was used to determine geographical-based differences in the volatilome of five "Grillo" wines (of Sicilian origin). The flow modulator used was a low duty cycle one, based on a concept introduced by Seeley *et al.* [21]. The new tile-based Fisher-ratio software released by the LECO corporation known as ChromaTOF Tile was applied to compare the acquired raw data. Briefly, the algorithm works by creating four grids composed of tiles, that are offset from each other, where each tile should be wide enough to embed the average analyte peak widths and retention time shift along both dimensions, and sums all the signals at one or more *m/z* value. The four grids initially produce redundant hits, however, use of the tiling approach avoids the challenges of GC×GC chromatogram alignment and also provides an impressive signal-to-noise (*S/N*) enhancement for detection of low concentration chemical differences [22–27]. The redundant hits are readily removed via the pinning and clustering step of the software, leaving only the best hit for each analyte feature that is discovered. The hitlist of all analyte features obtained from ChromaTOF tile was subsequently subjected to an off-line one-way analysis of variance (*ANOVA*) to distinguish true and false positives, and the validity of this approach was assessed with a receiver operating characteristic (ROC) curve. Finally, PCA was performed to distinguish the wines using only the statistically significant chemical features identified via this *ANOVA* approach.

4.2 MATERIALS AND METHODS

4.2.1 Chemicals, samples and sample preparation

Five commercial Grillo wines, from different geographical zones (in Sicily) and wineries, were analyzed. Specifically, Colosi wine was produced in "Petrosino" and "Segesta", Barone wine

was produced in “Marsala”, Capovero was produced in “Sambuca di Sicilia”, Settesoli wine was produced in “Menfi”, and FeudoArancio wine was produced in “Acate”. All wines were obtained through the same winemaking process (Fig. 4.1). Three bottles of each of the five Grillo wines were utilized to study the reproducibility of the vinification procedure (15 total bottles of wine). Sodium chloride, ethanol, *n*-hexane and 3-octanol, used as internal standard (IS), were purchased from Merck Life Science (Merck KGaA, Darmstad, Germany). The IS was solubilized in ethanol and was added to each sample at a concentration of 170 mg/l .

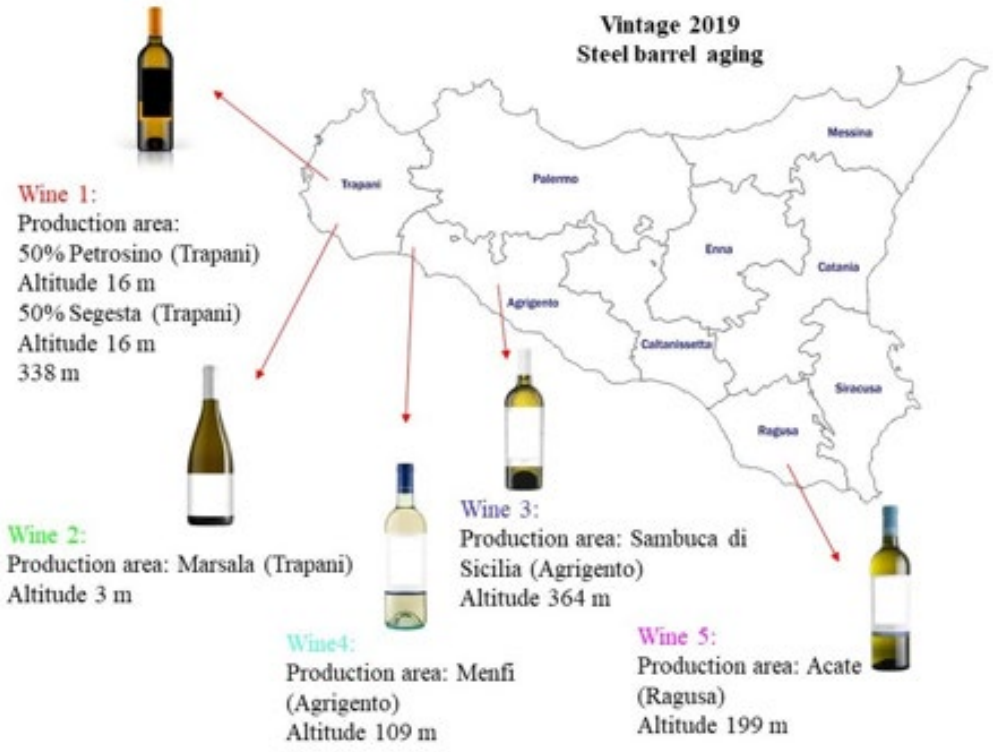

Figure 4.1. Map displaying the regions of Sicily in which wines 1-5 were produced.

4.2.2 Instrumentation

The extraction procedure was based on a previous report [13]. Headspace analysis was performed using a 50/30 μm divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) SPME fiber (Merck Life Science). Before the first use, the fiber was conditioned at 270 °C according to the conditioning guidelines. Five ml of sample was introduced into a 20 ml headspace vial, along with 2 g of sodium chloride and 10 μl of IS. The samples were incubated for 5 min at 45 °C, followed by 30 min of extraction at the same temperature. The agitation speed was 250 rpm. Three HS SPME replicates were obtained for each bottle per Grillo wine, resulting in 45 total samples for GC×GC-TOFMS analysis (five wines, three bottles each, three replicates per bottle). Desorption was performed for 2 min at 260 °C in the split mode with a 10:1 split ratio. The HS SPME procedure was performed automatically by using an L-PAL3 GC Autosampler (LECO, Mönchengladbach, Germany). All FM GC×GC-TOFMS analyses were performed using a Pegasus GC-BT 4D system, equipped with a diverting flow modulator, called the FLUX modulator (LECO). A schematic of this modulator is provided as Fig. C.1 in Appendix C. The ^{1}D column was a Supelcowax 10 [polyethylene glycol] with dimensions 10 m × 0.25 mm ID × 0.25 μm d_f, while the ^{2}D column was an SLB-35ms [silphenylene polymer with similar polarity to poly (35% diphenyl/65% dimethyl siloxane)] with dimensions 2 m × 0.10 mm ID × 0.10 μm d_f and with 0.3 m located inside the MS transfer line (280 °C). Both columns were supplied by Merck Life Science.

Helium was used as the carrier gas and was delivered at a constant flow of 1.9 ml/min. The main GC oven was held at 40 °C for 2 min, then ramped up to 280 °C at 20 °C/min, for a total analysis time of 14 min. The secondary oven offset was +20°C. The modulation period (P_M) was set at 700 ms, with a re-injection period of 80 ms. The auxiliary pressure unit (EPC) provided a

constant flow of 3.5 ml/min to the modulator. The MS parameters were as follows: acquisition delay 120 s, acquisition rate 150 spectra/s, electron ionization was performed at 70 eV, while mass spectra were acquired in the mass channel (*m/z)* range 35–360 amu.

4.2.3 Data Analysis

Following data acquisition, the 45 raw .SMP files were transferred from the LECO ChromaTOF for BT software to LECO ChromaTOF Tile v.1.01 (LECO Corporation, St. Joseph, MI) for tile-based Fisher ratio analysis (ChromaTOF Tile). The sample files were labeled according to Grillo wine type, resulting in five classes for F-ratio analysis with nine replicates per class. One-point normalization was performed prior to data analysis using the 3-octanol IS peak signal (^1t_R = 375.783 s, ^2t_R = 0.644 s) at *m/z* 59. For the remainder of this report, with the term peak area, we will refer to the normalized peak area. No fluctuation in the mass spectral response was observed between replicate chromatograms, so a mass spectrum drift correction was deemed unnecessary.

A tile size of 10 modulations (7 s) on ^{1}D and 45 spectra (300 ms) on ^{2}D was selected to encompass the average peak widths along both dimensions (see Results and Discussion) as well as modest retention time shifting of up to ~2 modulations on ^{1}D. Insignificant retention time shifting was observed on ^{2}D. A *S/N* threshold of 10 times the noise calculated on a per-tile basis for every *m/z* was applied to exclude low signal hits from the hitlist [22,28]. A minimum of 9 samples were required to exceed this *S/N* threshold, as this was the number of samples per F-ratio class. The hits were ranked in the hitlist according to their top F-ratio *m/z* to improve discoverability of true positive hits versus false positive hits [28]. No F-ratio threshold was set herein. The entire *m/z* range was included for tile-based F-ratio analysis.

Following hitlist generation, analyte identification was performed using the mainlib (NIST) and the Flavors and Fragrances of Natural and Synthetic Compounds [FFNSC version 4.0 (John Wiley & Sons, New Jersey, USA] library. Using the above mentioned parameters, ChromaTOF Tile produced an initial hitlist. Hits were excluded based on the following criteria: (1) artifact hits which streaked across multiple tile lengths, (2) hits which had noisy spectra, and (3) redundant analyte hits. Following this manual scrutinization, a final hitlist was generated and the ChromaTOF Tile-computed peak areas for these hits at their top F-ratio *m/z* were exported to MATLAB R2020a (The Mathworks Inc., Natick, MA, USA) for further data analysis. In order to ascertain the various contributions of chemical and non-chemical variation (i.e., injection, sample preparation, etc.) to the overall variation in this dataset [29], the percent relative standard deviation (%RSD) in peak areas for each hit was calculated in four ways. First, the overall RSD% was calculated for each hit collectively for all of the wines, defined by Eq. (4.1),

$$\%RSD_{all\ wines} = \frac{[std.dev.(areas_{wine\ 1}, areas_{wine\ 2}, areas_{wine\ 3}, areas_{wine\ 4}, areas_{wine\ 5})]*100}{mean\ (areas_{wine\ 1}, areas_{wine\ 2}, areas_{wine\ 3}, areas_{wine\ 4}, areas_{wine\ 5})} \quad (4.1)$$

Next, the %RSD was also calculated in Eq. (4.2) using the peak areas of the individual wines, one at a time, denoted as wine *n*,

$$\%RSD_{wine\ n} = \frac{[std.dev.(areas_{wine\ n})]*100}{mean\ (areas_{wine\ n})} \quad (4.2)$$

Hence, Eq. (4.2) was repeated for each wine, resulting in five values of %RSD for each analyte hit. Futhermore, Eqs. (4.1) and (4.2) were applied both without averaging the SPME replicate areas (subscript 'all areas'), and then following averaging the SPME replicate areas for each bottle of wine (subscript 'ave areas'). These four %RSD calulation methods will be referred to as $\%RSD_{\text{all wines, all areas}}$, $\%RSD_{\text{all wines, ave areas}}$, $\%RSD_{\text{wine } n\text{, all areas}}$, and $\%RSD_{\text{wine } n\text{, ave areas}}$. Examination of the %RSD distributions produced by collectively considering the %RSDs for all

analyte hits informs us as to the major sources of variation in the experimental design and chemical measurements.

Following examination of the %RSD distributions, a one-way analysis of variation (*ANOVA*) was performed using the average peak areas of the F-ratio hits (3 summed areas per wine, 15 areas total per hit). The one-way *ANOVA* enables calculation of a *p*-value across more than two sample classes, whereby the null hypothesis tested herein is that there is no statistically significant difference in the mean peak areas of the different types of wine [30]. A one-way *ANOVA* p-value < 0.01 was used to distinguish true positives from false positives. These true positive and false positive labels were used to construct a receiver operating characteristic (ROC) curve, from which the area under the curve (AUC) was calculated. More specifically, a ROC curve is a plot of the true positive probability (TPP, i.e. sensitivity), versus the false positive probability (FPP i.e., 1-specificity), which are calculated by calculating a running sum of the true and false positive instances divided by the total number of true and false positive instances, respectively [31]. This quantitative method has been widely used in the literature to evaluate the classifying capability of identified analytes, especially in biomarker research [22,31–33]. Finally, to highlight the class-distinguishing capability of the true positive F-ratio hits, the averaged peak areas were input to PCA. Mean centering was performed prior to PCA to ensure that analytes with the largest magnitude in peak areas did not unduly contribute to model performance [34,35]. The scores were used to classify respective "clusters" of the different wine samples, whereas the loadings were scrutinized to identify which analytes contributed most signficantly in distinguishing the wines.

4.3 RESULTS AND DISCUSSION

A typical total ion current (TIC) chromatogram for each wine (wines 1-5 per Fig. 4.1) is provided in Fig. 4.2(A-E). Although the total run time was 14 min, minimal chemical information

was present after a ^{1}D retention time ($^{1}t_{R}$) of 10 min, hence the ^{1}D axis of Fig. 4.2(A-E) was adjusted accordingly. At this level, all of the wines appear to be dominated by just a few highly concentrated analytes, such as octanoic acid, ethyl ester ($^{1}t_{R}$ = 6.6 min and $^{2}t_{R}$ = 300 ms), which is labeled with a star in Fig. 4.2(A-E). Thus, all of the wines at first glance appear chemically similar to each other, as the color scale is biased towards these highly concentrated analytes. However, closer examination of the chromatographic region centered around a $^{1}t_{R}$ of 6 min in Fig. 4.2(F-J) for each wine, reveals a high level of chemical complexity within the wine samples. Six analytes (linalool ethyl ether, heptanoic acid ethyl ester, hex-(3E)-enyl-acetate, octanoic acid methyl ester, n-hexanol, and 3-octanol) within this separation window have been identified and labeled accordingly as analytes 1-6 in Fig. 4.2(F-J), as the first five of these analytes exhibit noticeable concentration difference between these wines. For example, linalool ethyl ether (analyte 1) appears highly concentrated in wine 1 (Fig. 4.2(F)), whereas it is much less concentrated in wines 2 and 3 (Fig. 4.2(G-H)) and visually indistinguishable in wines 4 and 5 (Fig. 4.2(I-J)). 3-octanol (analyte 6) is the internal standard, so it serves as a useful visual control to confirm that it is present at the same concentration in Fig. 4.2(F-J). This preliminary visual examination of the chromatograms serves to underscore the chemical complexity of the wines and highlights the likely benefit of applying chemometrics to elucidate their chemical differences.

Although chemometric software tools such as ChromaTOF Tile are invaluable in extracting information from large GC×GC-TOFMS datasets, successful implementation requires a sound experimental design, particularly regarding the GC×GC separation conditions. For example, GC×GC chromatograms with minimal orthogonality, low sensitivity, wide ^{1}D and ^{2}D peak widths-at-base ($^{1}w_{b}$ and $^{2}w_{b}$), and/ or peak overlap and “wraparound” on ^{2}D can hamper extracting chemical information [7,8]. Herein, a relatively short 10 m ^{1}D column was utilized to produce

narrow $^{1}w_b$ to maximize the ^{1}D peak capacity ($^{1}n_c$) and the resulting 2D peak capacity ($n_{c,2D}$) [36]. Additionally, the FLUX modulator (Fig. C.1) was operated using a fast P_M of 700 ms and a long re-injection period of 80 ms (i.e., the longest allowed by the instrument given the P_M) to modulate the narrow ^{1}D peaks with an appropriate sampling density (p_s) of 2-4 while simultaneously maximizing the modulator duty cycle, the latter of which is critical for improving detection sensitivity and *S/N*. The resulting chromatograms shown in Fig. 4.2 exhibit minimal wraparound on ^{2}D that minimizes analyte co-elutions, which suggests the appropriateness of the modulation period selected. The impressive sensitivity afforded by the FLUX modulator can be indicated by the abundance of compounds observed in the zoom-ins provided in Fig. 4.2(F-J). Additionally, one of the most highly concentrated analytes in the wine samples, namely octanoic acid, ethyl ester (as discussed with Fig. 4.2) has an approximate $^{1}w_b$ of 3.6 s and $^{2}w_b$ of 240 ms, which equates to $n_{c,2D}$ of ~ 680, or a peak capacity production of ~ 50 peaks/min [37,38]. Thus, the experimental design from the instrumentation perspective was optimized to maximize the chemical information available for chemometric analysis.

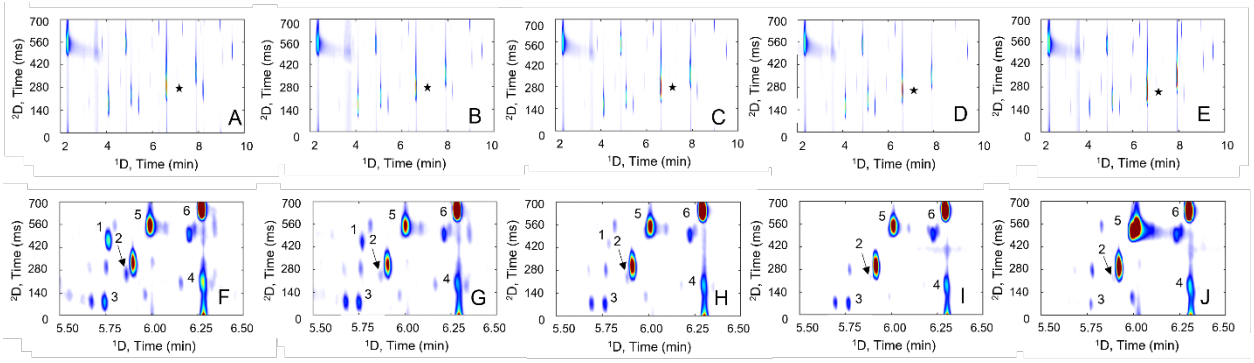

Figure 4.2. Total ion current (TIC) GC×GC chromatograms from 2 to 10 min for (A) wine 1, (B) wine 2, (C) wine 3, (D) wine 4, and (E) wine 5. Octanoic acid, ethyl ester is labeled with a star in A-E. A zoom-in region from 5.5 to 6.5 min is included to highlight the high peak capacity/chemical complexity of the separations for (F) wine 1, (G) wine 2, (H) wine 3, (I) wine 4, and (J) wine 5. Six analytes are labeled in F-J as 1 through 6, in the following order: linalool ethyl ether, heptanoic acid ethyl ester, hex-(3E)-enyl-acetate, octanoic acid methyl ester, n-hexanol, and 3-octanol.

It is instructive to initially apply PCA to all of the data to emphasize the need to apply ChromaTOF Tile for supervised feature selction to discover the analytes that best distinguish the five wines. The PCA scores plot of the 45 unfolded, normalized GC×GC-TOFMS chromatograms is provided in Fig. 4.3. It is important to note that only 58.43% of the variation within the dataset is captured in this PCA model (36.07% on PC1 and 22.36% on PC2), which suggests that potential class-distinguishing information is being buried by spurious chemical signal and noise. This is confirmed by the lack of clustering by wine type in the scores plot, as we would expect to see five distinct sample clusters if the wines were being properly classified by PCA at this stage (Fig. 4.3). Since PCA is an unsupervised chemometric tool, its ability to distinguish samples is hampered if spurious chemical signal and noise dominates the dataset input to PCA, which is the case prior to applying ChromaTOF Tile. Based on a preliminary examination of the TIC chromatograms in Fig. 4.2, these wines do have low concentration differences for a given analyte peak from one wine chromatogram to another, but rigorous examination of the entire chromatogram dataset would be laborious and highly prone to error. Thus, a supervised cheometric tool such as ChromaTOF Tile is necessary and ideally suited to elucidate the extent of chemical differences between these wines.

Arguably, the most critical input parameter to tile-based F-ratio analysis is the tile size. In this work, a ^{1}D tile dimension of 7 s (10 modulations) and a ^{2}D tile dimension of 300 ms were selected. The typical analyte linalool ethyl ether (analyte 1 in Fig. 4.2(F-J)), in terms of $^{1}w_{b}$ and $^{2}w_{b}$, is examined in greater detail in Fig. 4.4 to illustrate the justification behind this choice. The summed ^{1}D and ^{2}D peaks for linalool ethyl ether are provided in Fig. 4.4(A-B), respectively, with the same color coding by wine as was used in Fig. 4.3. The x-axes in Fig. 4.4(A-B) are equivalent in length to the respective tile dimensions. Using the replicates for wine 1 (red), linalool ethyl ether has an approximate $^{1}w_{b}$ = 2.6 s and $^{2}w_{b}$ = 130 ms. Retention time shifting of up to 1.4 s (2

modulations) on ^{1}D is observed in Fig. 4.4(A), whereas insignificant retention time shifting on ^{2}D is seen in Fig. 4.4(B). An especially large ^{1}D tile dimension of 7 s can be justified to correct for the ^{1}D run-to-run shifting, whereas the 300 ms ^{2}D tile dimension is appropriate, given that the maximum $^{2}w_{b}$ observed for other analytes is 300 ms with essentially no ^{2}D shifting (see discussion of octanoic acid, ethyl ester with Fig. 4.2). Thus, this tile size is suitable for capturing the full range of $^{1}w_{b}$ and $^{2}w_{b}$ observed.

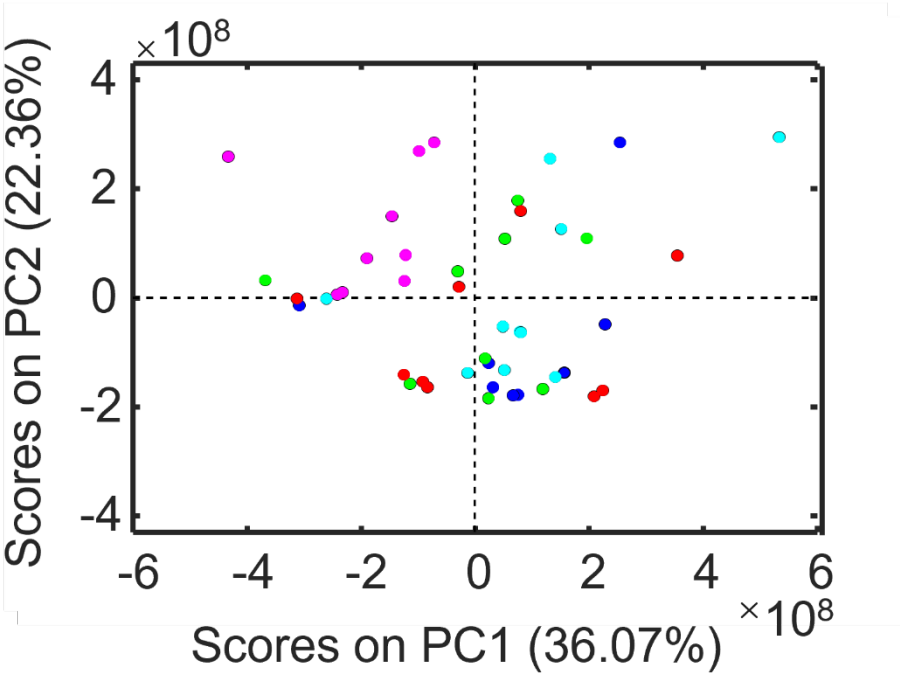

Figure 4.3. PCA scores plot of unfolded chromatograms using all the *m/z*, normalized to the IS. The abundance of chromatographic superfluous chemical signal and noise makes it impossible to distinguish the chromatograms in this scores plot, which necessitates application of a supervised discovery tool such as tile-based F-ratio analysis.

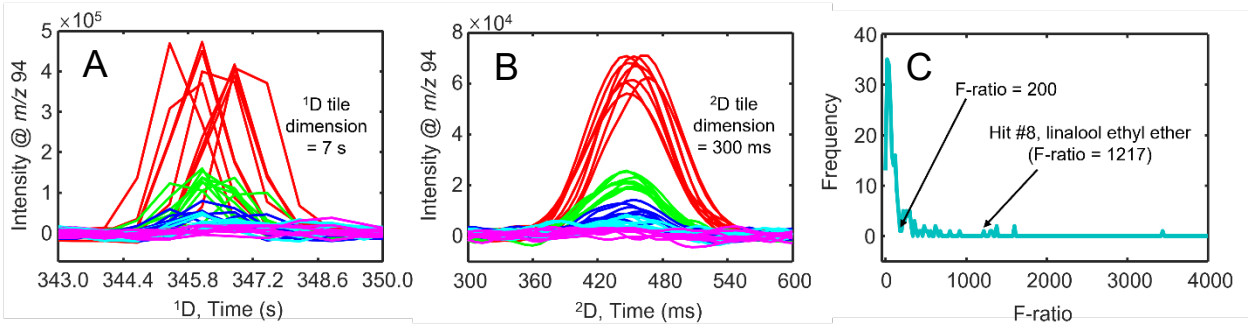

Figure 4.4. Background-corrected summed ^{1}D (A) and ^{2}D (B) peaks for linalool ethyl ether at *m/z* 94. The summed ^{1}D peaks were prepared by summing away the ^{2}D modulations, while the summed ^{2}D peaks were prepared by summing away the ^{1}D separation. The windows provided are the lengths of the F-ratio tile dimensions selected, with the sizes of the ^{1}D and ^{2}D tile dimensions labeled accordingly. The F-ratio distribution (bin size = 20) for the hitlist of 220 hits is provided in (C), with the F-ratio of linalool ethyl ether (1217) indicated accordingly.

Upon running the ChromaTOF Tile software using this tile size, an initial hitlist containing 899 hits was produced; 780/899 hits were assigned a p-value < 0.05 by the software, so these 780 hits were subjected to further examination. Following manual artifact and redundant hit removal (see Experimental), the initial hitlist was reduced to a final hitlist of 220 hits. The F-ratio distribution for these 220 final F-ratio hits is provided in Fig. 4.4(C), with the location of the analyte linalool ethyl ether indicated. Linalool ethyl ether is found at the top of the hitlist (hit 8), but is important to reiterate that the F-ratio magnitude of 1217 and hitlist ranking of linalool ethyl ether would be compromised had an inappropriate tile size been used, even though linalool ethyl ether is clearly class-distinguishing (Fig. 4.4(A-B)). Furthermore, approximately 77% of the F-ratio distribution shown in Fig. 4.4(C) falls below an F-ratio of 200 (170/220 hits), beyond which the frequency of hits levels off (frequency less than 10). Thus it appears that most of the hits in the hitlist are compounded by complex sources of background variation (sample preparation, injection variation, sensitivity, etc.) in addition to chemical variation, which requires further investigation to ascertain statistical significance in distinguishing the wines.

Bar plots displaying the peak areas of linalool ethyl ether in all 45 chromatograms are provided in Fig. 4.5(A), with the same color coding previously applied to distinguish wines 1-5. The peak area differences observed in Fig. 4.5(A) coincide with the peak profiles provided in Fig. 4.4(A-B), with wine 1 and wine 2 having the highest and second highest concentrations of linalool ethyl ether, respectively, and wines 3, 4, and 5 having successively lower concentrations. Note that three bottles of each wine were analyzed, with three SPME replicates collected per bottle. For each of the wines in Fig. 4.5(A), the bars are ordered by injection replicate for each consecutive bottle (i.e., bottle 1 replicate 1, bottle 1 replicate 2, bottle 1 replicate 3, bottle 2 replicate 1, etc.). Visual inspection suggests that the injection replicates (recall a replicate refers to a separate wine aliquot

with a separate SPME injection from the same bottle) are only marginally contributing to the overall variation in peak areas for a given analyte hit. Thus, averaging the replicates was performed. The resulting bar plot for linalool ethyl ether is shown in Fig. 4.5(B), where now only three bars per wine are provided to represent the summed peak areas per bottle. Note that the averaged bars in Fig. 4.5(B) exhibit the same trends in peak areas between bottles of a given wine as were observed in Fig. 4.5(A), which suggests that summing the replicates does not remove and/or add a significant portion of the total variation associated with the linalool ethyl ether hit. Indeed, for wine 2, the %RSD between peak areas in Fig. 4.5(A) is approximately 13.9% and only slightly rises to 14.4% in Fig. 4.5(B) after averaging the injection replicates, which indicates that the variation due to injection only amounts to 0.5% RSD and is thus negligible relative to the ~14% RSD due to bottle-based differences. However, this conclusion that the injection replicates contribute a neglibile amount of variation can not necessarily be extrapoloated to the remaining 219 hits in the hitlist, which necessitates a comprehensive assessment of %RSD in the peak areas.

Using Eq. (4.1) and Eq. (4.2) for the four %RSD calculations described in the Experimental, four %RSD distributions in peak areas were generated, which are provided in Fig. 4.5(C): $\%RSD_{\text{all wines, all areas}}$ (gold solid line), $\%RSD_{\text{all wines, ave areas}}$ (gold dashed line), $\%RSD_{\text{wine n, all areas}}$ (purple solid line), and $\%RSD_{\text{wine n, ave areas}}$ (purple dashed line). The $\%RSD_{\text{wine n, all areas}}$ and $\%RSD_{\text{wine n, ave areas}}$ values were boxcar averaged from 1100 to 220 total %RSD values (boxcar size = 5) to facilite the comparison between all four RSD% distributions. Note that the $\%RSD_{\text{all wines, all areas}}$ and $\%RSD_{\text{all wines, ave areas}}$ distributions have a maximum frequency of occurence at %RSD of 48%, whereas the $\%RSD_{\text{wine n, all areas}}$ and $\%RSD_{\text{wine n, ave areas}}$ distributions are shifted to the left (i.e., lower %RSD) and have a maximum frequency of occurence at %RSD values of 12% and 8%, respectively. Thus, for most of the 220 total hits in the F-ratio hitlist, the %RSD in peak areas

between wines 1-5 (wine-to-wine variation) is much larger than the %RSD in peak areas for a given wine (bottle-to-bottle variation), so it can be concluded that the wine-to-wine differences are the most significant source of variation in the dataset, rather than the bottle-to-bottle differences for a given wine (Fig. 4.5(C)). Therefore, during the tile-based F-ratio analysis, the presence of only minor bottle-to-bottle variation should not signficantly hinder the discovery of chemically relavent differences between the wines. Furthermore, in comparing the $\%RSD_{\text{all wines, all areas}}$ and $\%RSD_{\text{all wines, ave areas}}$ distributions to each other, as well as the $\%RSD_{\text{wine n, all areas}}$ and $\%RSD_{\text{wine n, ave areas}}$ distributions to each other, they appear largely overlapped, with the %RSD distributions using the averaged areas slightly shifted to lower %RSD values relative to the distributions using all 45 original peak areas. What this reveals, is that injection/sample preparation-based variation is largely negligible in the context of the chemically-based variation, as it amounts to an RSD% of ~ 5% when comparing the respective distributions. Thus averaging the peak areas of the injection replicates is justified and necessary prior to further statistical analysis, as this will simultaneously discount the influence of uninformative variables while reducing data density.

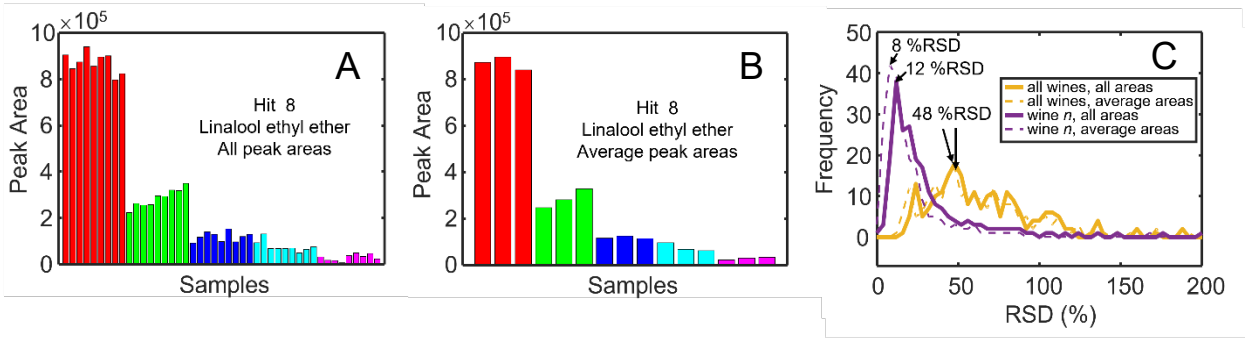

Figure 4.5. Assessment of %RSD in peak areas of ChromaTOF Tile hits. (A) All peak areas for linalool ethyl ether in wines 1-5, prior to averaging. (B) Averaged peak areas for linalool ethyl ether. The following color coding is used in (A) and (B): red for wine 1, green for wine 2, blue for wine 3, cyan for wine 4, and pink for wine 5. (C) Overlaid RSD distributions for assessing the contributions of wine-to-wine variation, bottle-to-bottle variation, and replicate-to-replicate variation to the overall chemical variation provided by each hit. The distributions are color-coded accordingly: $\%RSD_{\text{all wines, all areas}}$ (gold solid line), $\%RSD_{\text{all wines, ave areas}}$ (gold dashed line), $\%RSD_{\text{wine n, all areas}}$ (purple solid line), and $\%RSD_{\text{wine n, ave areas}}$ (purple dashed line). The %RSD of maximum frequency is labeled for each distribution accordingly: 48% for $\%RSD_{\text{all wines, all areas}}$ and $\%RSD_{\text{all wines, ave areas}}$; 12% for $\%RSD_{\text{wine n, all areas}}$; and 8% for $\%RSD_{\text{wine n, ave areas}}$.

The tentative identities of the top 30 hits in the F-ratio hitlist are provided in Table 4.1, along with their retention times on ^{1}D and ^{2}D (^1t_R and ^2t_R), F-ratios, top F-ratio *m/z*, *p*-values from a one-way *ANOVA* test, the corresponding wine with the highest concentration, and abbreviated sensory information [39]. The entire hitlist for all 220 hits is provided in Appendix C as Table C.1, with the corresponding concentration and detailed sensory information in Table C.2. Note that 137/220 hits could be identified with a match value (MV) of at least 800 (1000 perfect match to library), while only 83 hits could not be identified (Table C.1). For the top 30 hits in Table 4.1, 23 have been identified in wines using 1D-GC-MS and GC×GC-MS in prior studies [19,20,40–51]. Interestingly, the advantage of using GC×GC relative to 1D-GC is underscored by examining just the top 30 hits in Table 4.1, with (1) geranyl isobutyrate and nerol oxide, and (2) myrcene and α-phellandrene, respectively, being overlapped on ^{1}D but having distinct ^2t_R. Further down the hitlist

beyond hit 30, there are several other instances in which the benefits of using GC×GC relative to 1D-GC occur. Regarding the seven hits in the top 30 not previously identified in wine, five are structurally similar to other previously identified flavor volatiles in wine, so these compounds are shaded in blue in Table 4.1. The remaining two hits (shaded in orange in Table 4.1) are not flavor volatiles, but artifacts of the vinification process. For example, phthalate esters such as di-isobutyl phthalate (hit #16) are common contaminants from plastics introduced during the winemaking process [52]. This information is still highly useful in the differentiation of the wines, as the vinification procedure for wine 1 appears to result in the highest concentration of di-isobutyl phthalate (Table 4.1). Future work could involve collecting additional replicate chromatograms of wine 1 to quantify the level of di-isobutyl phthalate contamination more accurately. Another non-flavor associated compound found in the top 30 hits is allyl isothiocyanate (hit #26), which has known anti-fungal properties and has been used widely in food preservation [53,54]. Thus, it appears that allyl isothiocyanate is used more widely as an antimicrobial agent in the north end of Trapani, where wine 1 is produced, as it has the highest concentration in wine 1 (Fig. 4.1). Interestingly, most of the flavor volatiles in the top 30 hits are also at their highest concentration in wine 1, with *p*-values four to ten orders of magnitude smaller than the typical *p*-value thresholds of 0.05 or 0.01 typically applied (Table 4.1). Thus it appears that wine 1 may be the most "chemically unique" wine in this dataset. However, examination of Table C.2 reveals that hits further down the hitlist are present at higher concentrations in wines 2-5. Thus, application of one-way *ANOVA* to all 220 hits is necessary to fully characterize the chemical differences between the wines, as trace level volatiles could be missed if one were just to rely on a subset of the top hits.

Table 4.1. Identities of the top 30 true positive hits identified following tile-based F-ratio analysis. Retention time and sensory information are also provided. Volatile flavor compounds which were not identified in previous studies are highlighted in blue, while non-flavor components are highlighted in orange.

Hit Number	^1t_R (min)	^2t_R (s)	F-ratio	Top *m/z*	Identity	*p*-value	Wine, highest concentration	Sensory profile
1	8.17	0.27	3434.2	152	ethyl trans-4-decenoate	2.9E-12	1	fatty
2	6.69	0.51	1608.4	87	cis-ocimenol	1.6E-12	1	**unknown**
3	8.24	0.01	1594.0	136	α-terpineol	6.9E-09	1	citrus
4	7.29	0.64	1385.7	72	linalool	1.9E-12	1	orange
5	7.07	0.35	1370.9	41	geranyl vinyl ether	9.7E-12	1	**unknown**
6	8.09	0.63	1323.6	174	diethyl succinate	5.9E-12	5	fruity
7	6.92	0.18	1301.6	55	ethyl 7-octenoate	9.4E-06	1	**unknown**
8	5.75	0.45	1217.2	94	linalool ethyl ether	6.4E-12	1	floral
9	7.47	0.59	917.3	70	isoamyl lactate	2.1E-11	5	fruity
10	6.86	0.35	806.6	93	geranyl isobutyrate	3.6E-10	1	sweet
11	5.96	0.52	689.6	75	ethyl lactate	1.1E-10	5	sweet
12	4.45	0.37	631.6	93	myrcene	1.1E-08	1	woody
13	5.21	0.32	623.0	93	3-carene	4.1E-07	1	citrus
14	7.68	0.61	610.4	71	2,6-dimethyl-3,7-octadiene-2,6-diol	2.2E-09	1	**unknown**
15	5.07	0.33	551.6	93	ethyl hexanoate	7.5E-07	1	sweet
16	12.95	0.07	529.2	149	di-isobutyl phthalate	9.3E-11	1	**unknown**
17	8.58	0.61	503.9	123	citronellol	1.2E-09	1	floral
18	4.45	0.54	475.1	93	α-phellandrene	2.4E-09	1	terpenic
19	9.26	0.47	472.4	91	benzyl alcohol	1.2E-10	3	chemical
20	9.35	0.04	442.9	129	ethyl isopentyl succinate	5.7E-10	5	**unknown**
21	6.83	0.19	412.7	68	nerol oxide	2.7E-10	1	green
22	4.76	0.52	373.0	92	limonene	2.8E-09	1	citrus
23	8.79	0.60	361.6	69	nerol	1.2E-09	1	lemon
24	9.42	0.44	353.7	70	isopentyl undecenoate	1.6E-07	1	floral
25	10.95	0.28	353.5	221	2,3-dihydro-1,1,3-trimethyl-3-phenyl-1H-indene	5.3E-09	2	**unknown**
26	6.09	0.60	318.6	99	allyl isothiocyanate	5.6E-07	1	mustard
27	5.56	0.37	317.1	70	3-methylbutyl pentanoate	3.3E-07	1	strawberry
28	3.02	0.13	305.2	55	isobutyl acetate	8.4E-08	1	sweet
29	9.04	0.59	301.0	68	geraniol	3.6E-06	1	floral
30	9.31	0.34	300.2	88	ethyl 10-undecenoate	1.2E-07	1	fatty

The results of the one-way *ANOVA* for all 220 hits in the F-ratio hitlist are provided as a scatterplot in Fig. 4.6(A), with *p*-value plotted versus hit number. A *p*-value threshold of 0.01 is shown with a dashed red line, whereby hits with a *p*-value < 0.01 indicate that the mean peak areas are significantly different and hits with a *p*-value ≥ 0.01 that the mean peak areas are not significantly different at the 99% confidence level. Since 220 hits were identified, the 99% confidence level equates to ~2 analyte hits erroneously exhibiting wine-to-wine concentration differences, whereas the 95% confidence level equates to 11 hits. Thus, the 99% confidence level was applied to more robustly avoid inclusion of false positives, while still including a substantial number of true positives. Indeed, 187 hits fall below a *p*-value < 0.01 in Fig. 4.6(A) and can be deemed "true positive hits" (i.e., class distinguishing), whereas the 33 hits which fall above the *p*-value cutoff can be considered "false positive hits" (i.e., random noise and other background variation). Note that the first false positive (hit 133) occurs at an F-ratio of 55 and the last true positive (hit 213) occurs at an F-ratio of 6 (Fig. 4.6(A)), so 54 true positives are intermingled with false positives at lower F-ratios. It is important to consider that the F-ratios in Table 4.1 and Table C.1 were calculated using all 45 original peak areas in ChromaTOF Tile, whereby F-critical cutoffs at the 95% (*p*-value = 0.05) and 99% (*p*-value = 0.01) confidence levels equated to F-ratios of 2.6 and 3.8, respectively. Thus, simply applying either cut-off to the original ChromaTOF hitlist without off-line statistical analysis would have been problematic, as all 33 false positives would have been incorrectly deemed true positives. Using a large sample size of 45 samples, with 9 samples per class, led to an underestimation in *p*-values and a resulting overestimation in statistical significance, an effect that has been widely studied by statisticians and termed the "large sample size fallacy" in numerous research disciplines [55–58]. By applying an off-line *ANOVA* to the averaged peak areas, effectively reducing the total sample size from 45 to 15 peak areas for a given

analyte hit, the F-values were recalculated and the new F-critical value of 6 at the 99% confidence level identified numerous false positives in the dataset.

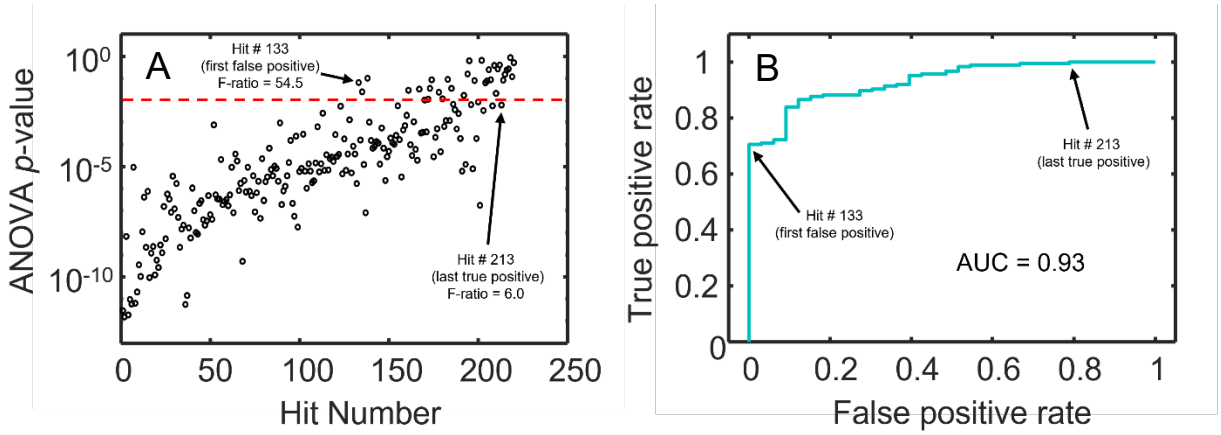

Figure 4.6. One-way *ANOVA* and ROC curve results. (A) Results of an off-line one-way ANOVA applied to the averaged peak areas of the 220 hits identified via ChromaTOF tile analysis. A stringent *p*-value threshold of 0.01 (dashed red line) was selected for true positive identification, with the first false positive (hit 133) and last true positive (hit 213) labeled accordingly. (B) Receiver operating characteristic (ROC) curve prepared using the true and false positive labels obtained via one-way ANOVA in (A), with the first false positive (hit 133) and last true positive (hit 213) labeled. The calculated area under the curve (AUC) of 0.93 is also provided, which indicates that a *p*-value threshold of 0.01 has a 93% probability of correctly labeling true and false positive hits with this particular dataset.

A receiver operating characteristic (ROC) curve using the true positive and false positive labels identified with off-line one-way *ANOVA* in Fig. 4.6(A) is provided in Fig. 4.6(B), with the true positive rate (i.e., sensitivity) plotted versus the false positive rate (i.e., 1-specificity). The apparent "steps" in the ROC curve are indicative of intermingling of true and false positives, so the first false positive (hit 133) and last true positive (hit 213) are labeled accordingly. The area under the curve (AUC) was calculated to be 0.93. Note that AUC values range from 0.5 to 1, with an AUC of 1 indicating maximum classifying power for a given variable and an AUC of 0.5 (i.e.,

a diagonal line with a slope of 1) equivalent to random chance decisions, meaning that a given variable has no classifying power. In other words, the AUC represents the probability of a given variable correctly distinguishing true and false positives [22,31–33]. In this work, the variable being evaluated is a *p*-value threshold of 0.01. The large AUC of 0.93 indicates that a *p*-value threshold of 0.01 is highly accurate for distinguishing the statistically significant chemical differences between the wines from superflous chemical signal, erroneous noise, and background variation. None the less, there is a 93% probability that the 33 hits with *p*-values ≥ 0.01 are false positives, so these hits should be excluded from further chemometric endeavors to distinguish these wines.

The resulting PCA scores plot using the average peak areas of the 187 true positive hits is provided in Fig. 4.7(A) to serve as a visualization tool to highlight the performance of ChromaTOF Tile. It is important to note that 85.76% variance is captured along both PC axes, which is considerably greater than the 58.43% variance captured in the original PCA model using all 45 unfolded chromatograms (Fig. 4.3). Furthermore, the averaged samples (3 per wine) now exhibit clustering by wine type in the PCA space, with minimal variation between the samples of a given wine cluster and no overlap between neighboring clusters (Fig. 4.7(A)). Essentially, the workflow presented herein provides variable reduction; even though PCA itself is a dimensionality data reduction tool, variable reduction tools are often necessary to reduce noise and thus improve the discriminatory power of PCA models [59–61]. It is interesting to note that PC1 (64.50% variance) captures the chemical differences between wines 1 and the remaining wines 2-5, whereas PC2 (21.26% variance) captures the finer detail in the chemical differences between wines 2-5 themselves. The fact that PC1 accounts for ~ 40% more variance than PC2 reveals that, as was

suggested by examining the top 30 hits in Table 4.1, wine 1 is highly "chemically unique" relative to the other wines.

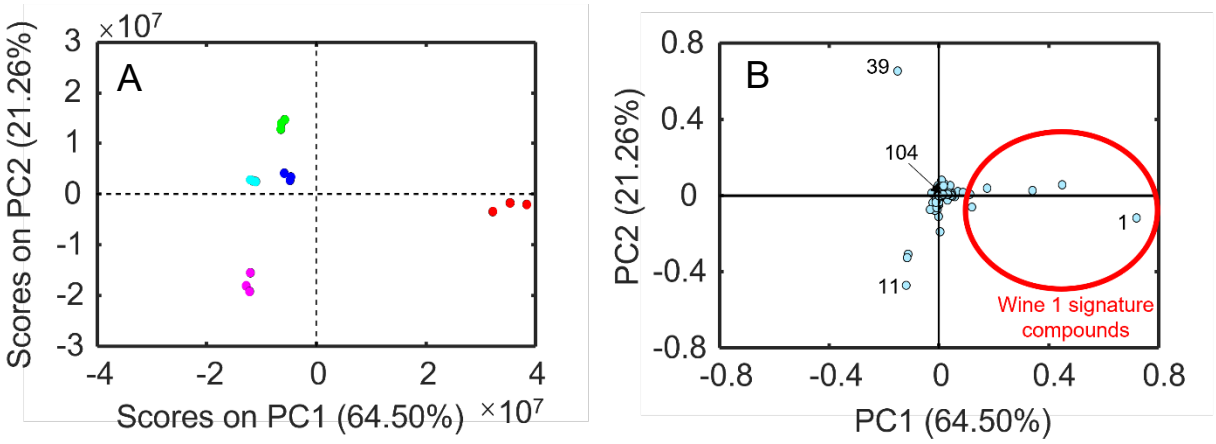

Figure 4.7. PCA results using true positive hits. (A) Scores plot obtained from PCA of the peak areas of the 187 true positive hits. (B) Two-dimensional loadings plot, wherein each blue dot represents one of the 187 true positives. Three highly loaded hits (1, 11, and 39) and one lowly loaded hit (104) are labeled for discussion purposes.

Examination of the two-dimensional loadings plot in Fig. 4.7(B) enables identification of which analyte peak areas contributed most significantly to the PCA model, wherein the blue circles represent the 187 true positive analyte hits. Since wine 1 is the only group with positive PC1 scores in Fig. 4.7(A), the five hits with highly positive PC1 loadings in Fig. 4.7(B) (enclosed by the red circle) must be more highly concentrated in wine 1 relative to the other wines, effectively making these five hits signature compounds of wine 1. Similarly, based on their PC1 and PC2 loadings, hits 11 and 39 are signature compounds of wine 5 and wine 2, respectively (Fig. 4.7). The fact that a large portion of the 187 hits, including hit 104, cluster around 0 in Fig. 4.7(B) indicates that these hits contribute minimal variance to the PCA clustering of the wines. Thus, information about most

of the analyte hits is not necessarily needed to quickly classify "unknown" samples of these wines, if such a need arose. However, the peak areas for all 187 hits have a p-value < 0.01, which indicates more subtle differences in concentration that are worth exploring for comprehensive fingerprinting purposes.

To explore this idea further, bar plots displaying the averaged peak areas of three highly loaded hits (hits 1, 11, and 39) and one lowly loaded hit (hit 104) are provided in Fig. 4.8, along with their chemical identities and one-way *ANOVA* p-values. Via examination of Fig. 4.8(A-C), it is obvious why ethyl trans-4-decenoate (hit 1), ethyl lactate (hit 11), and butyl alcohol (hit 39) have large loadings values in Fig. 4.7(B), as these compounds are highly concentrated in only one wine, namely wine 1, wine 5, and wine 2, respectively. Conversely, methyl 2-oxononanoate (Fig. 4.8(D)) has a similar concentration in all of the wines, which is why it is lowly loaded on PC1 and PC2 in Fig. 4.7(B). Thus, although methyl 2-oxononanoate is not one of the most notable signature volatiles in distinguishing the wines, its low p-value of 10^{-6} indicates that its higher concentration in wine 3 is indeed statistically significant (Fig. 4.8(D)). These trace concentration differences are critically important to assess the overall chemical differences between the wines, such as their sensory profiles.

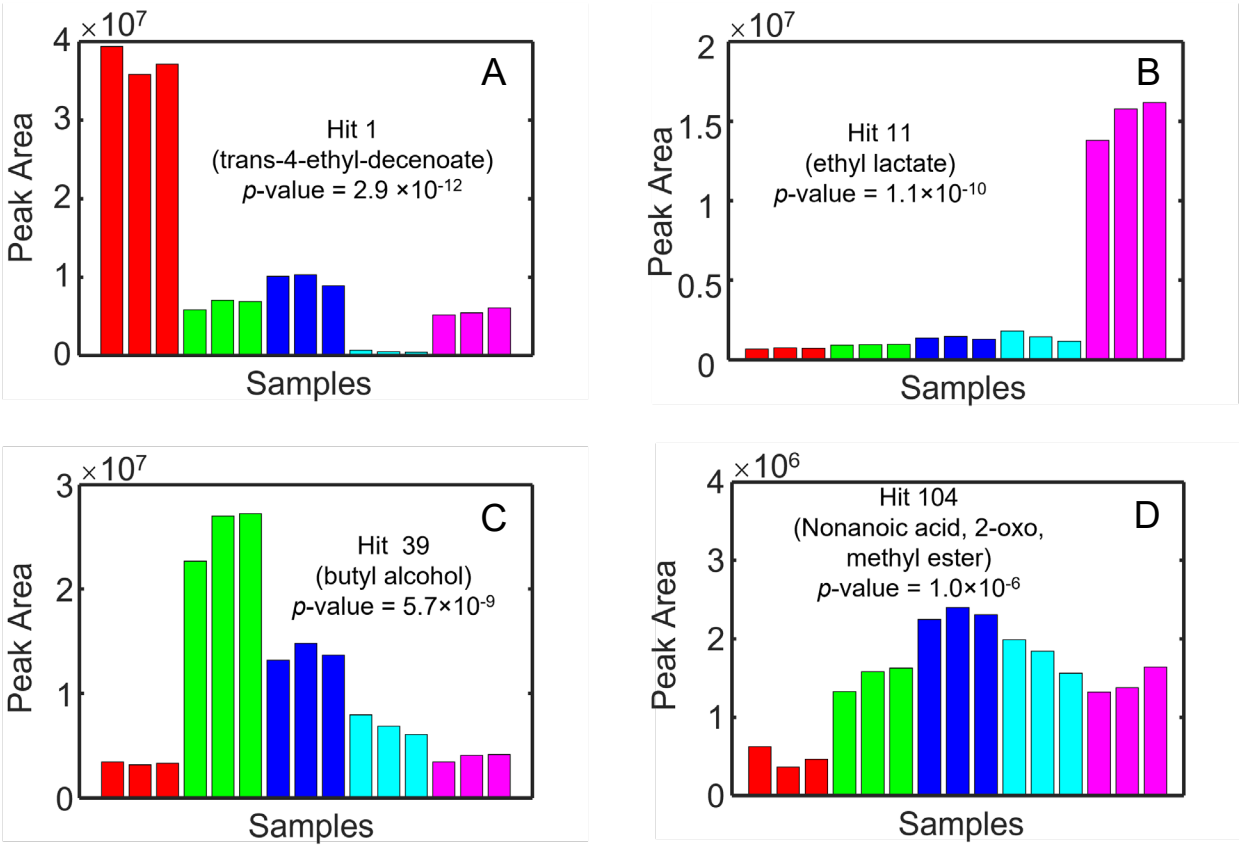

Figure 4.8. Bar graphs displaying the averaged peak areas for (A) hit 1, ethyl trans-4-decenoate (B) hit 11, ethyl lactate, (C) hit 39, butyl alcohol, and (D) hit 104, methyl 2-oxononanoate, which were indicated in the loadings plot provided in Fig. 4.7(B). The corresponding *p*-values are also provided to underscore the significant chemical information given by lowly loaded hits, such as (D) hit 104.

Of the 187 true positve hits identified herein, 90 analytes were the most highly concentrated (i.e., signature compounds) in wine 1; 24 analytes in wine 2; 21 analytes in wine 3; 7 analytes in wine 4; and 45 analytes in wine 5 (Table C.2). Pie charts displaying the distributions of sensory descriptors for these signature compounds of each wine are provided in Appendix C as Fig. C.2. It is important to note that Fig. C.2 does not reflect quantitative concentrations of these signature compounds, but rather the relative proportions of signature compounds with certain flavor attributes. For example, of the 90 signature volatile compounds for wine 1, 19 compounds have been described as fruity; 14 as floral; and 14 as green, which is why these flavors are the largest "slices" for wine 1 (Fig. C.2(A)). Similarly, sweet (7/24) , banana (5/24), and fruity (5/24) are the top flavor descriptors for wine 2; fruity (3/21) for wine 3; ethereal (3/7) and fruity (2/7) for wine 4; and fruity (12/45) and sweet (9/45) for wine 5 (Fig. C.2(B-E)). On a finer level, wine 3 is the only wine characterized by compounds with bacon flavor (4-vinylguaiacol, hit 54 (Table C.2)) and earthy/mushroom flavor (hit 74, 7-methylbenzofuran (Table C.2)), while wine 5 has the highest proportion of compounds with creamy and related descriptions (i.e., cheesy, dairy, egg nog, etc.). Such a holistic assessment of the sensory characteristics of each wine is not possible using only the few highly loaded analytes identified with PCA in Fig. 4.7(B), which underscores the utility of the off-line one-way *ANOVA* performed herein. The combination of ChromaTOF Tile, followed by one-way *ANOVA* and PCA is a powerful workflow for enabling comprehensive fingerprinting and classification by wine type.

4.4 CONCLUSIONS

Chromatographic fingerprinting through HS SPME coupled with FM GC×GC-TOFMS has demonstrated to be a valuable tool to characterize geographical-based differences in the volatilome of five white "Grillo" wines. A fast FM GC×GC approach was developed following, a supervised

chemometric approach was carried out, exploiting the ChromaTOF Tile software, to elucidate the chemical differences between the wines. Of the 220 hits, 187 hits were discovered to be true positive, “class-distinguishing” hits via an off-line one-way ANOVA p-value threshold of 0.01, the validity of which was verified via a ROC curve.

PCA using the average peak areas of the 187 true positive hits showed distinct clustering of the wines according to geographical origin in the scores plot, but only a handful of analytes were highly loaded and thus needed for classification purposes. The wines have distinct flavor profiles which would have been overlooked using only the most highly loaded analytes in PCA. Such sensory information may be critically important to winemakers looking to optimize their vinification process and/or experiment with new flavor profiles. Additional work to this end could include correlating geographical information (i.e., altitude, soil conditions, etc.) with GC×GC-TOFMS signatures, or even building complex neural networks to distinguish highly similar wines according to trace volatile concentrations. Thus, the work performed herein highlights the utility of the new ChromaTOF Tile software, both as a standalone supervised method and as a feature selection tool prior to additional chemometric/machine learning endeavors.

4.5 ACKNOWLEDGEMENTS

The Authors acknowledge LECO Corporation and Merck Life Science for their continuous support. The research was conducted within the project AIM1808474-1 – AREA SNSI Energia (CUP J44I18000140006).

"This article is based upon work from the Sample Preparation Task Force and Network, supported by the Division of Analytical Chemistry of the European Chemical Society."

The Authors thanks Cantine Madaudo and Cantine Colosi for the provided wines.

4.6 REFERENCES

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

[2] M.S.S. Amaral, Y. Nolvachai, P.J. Marriott, Comprehensive Two-Dimensional Gas Chromatography Advances in Technology and Applications: Biennial Update, Anal. Chem. 92 (2020) 85–104. https://doi.org/10.1021/acs.analchem.9b05412.

[3] Nicholas H. Snow, Basic Multidimensional Gas Chromatography, Elsevier, 2020.

[4] H.D. Bahaghighat, C.E. Freye, R.E. Synovec, Recent advances in modulator technology for comprehensive two dimensional gas chromatography, TrAC Trends Anal. Chem. (2018). https://doi.org/10.1016/j.trac.2018.04.016.

[5] A. Ferracane, M. Zoccali, F. Cacciola, T.M.G. Salerno, P.Q. Tranchida, L. Mondello, Determination of multi-pesticide residues in vegetable products using a “reduced-scale” Quechers method and flow-modulated comprehensive two-dimensional gas chromatography-triple quadrupole mass spectrometry, J. Chromatogr. A 1645 (2021) 462126. https://doi.org/10.1016/j.chroma.2021.462126.

[6] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications, Anal. Chem. (2017). https://doi.org/10.1021/acs.analchem.7b04226.

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

[8] F. Stilo, C. Bicchi, A.M. Jimenez-Carvelo, L. Cuadros-Rodriguez, S.E. Reichenbach, C. Cordero, Chromatographic fingerprinting by comprehensive two-dimensional chromatography: Fundamentals and tools, TrAC Trends Anal. Chem. 134 (2021) 116133. https://doi.org/10.1016/j.trac.2020.116133.

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

[10] L.T. Vaz-Freire, M.D.R.G. da Silva, A.M.C. Freitas, Comprehensive two-dimensional gas chromatography for fingerprint pattern recognition in olive oils produced by two different techniques in Portuguese olive varieties Galega Vulgar, Cobrançosa e Carrasquenha, Anal. Chim. Acta 633 (2009) 263–270. https://doi.org/10.1016/j.aca.2008.11.057.

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

[12] J.E. Welke, M. Zanus, M. Lazzarotto, F.H. Pulgati, C.A. Zini, Main differences between volatiles of sparkling and base wines accessed through comprehensive two dimensional gas chromatography with time-of-flight mass spectrometric detection and chemometric tools, Food Chem. 164 (2014) 427–437. https://doi.org/10.1016/j.foodchem.2014.05.025.

[13] R. Costa, C. Fanali, G. Pennazza, L. Tedone, L. Dugo, M. Santonico, D. Sciarrone, F. Cacciola, L. Cucchiarini, M. Dachà, L. Mondello, Screening of volatile compounds composition of white truffle during storage by GCxGC-(FID/MS) and gas sensor array analyses, LWT - Food Sci. Technol. 60 (2015) 905–913. https://doi.org/10.1016/j.lwt.2014.09.054.
[14] S. Carlin, U. Vrhovsek, P. Franceschi, C. Lotti, L. Bontempo, F. Camin, D. Toubiana, F. Zottele, G. Toller, A. Fait, F. Mattivi, Regional features of northern Italian sparkling wines, identified using solid-phase micro extraction and comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry, Food Chem. 208 (2016) 68–80. https://doi.org/10.1016/j.foodchem.2016.03.112.
[15] P.-H. Stefanuto, K.A. Perrault, L.M. Dubois, B. L'Homme, C. Allen, C. Loughnane, N. Ochiai, J.-F. Focant, Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1507 (2017) 45–52. https://doi.org/10.1016/j.chroma.2017.05.064.
[16] J.M. Muñoz-Redondo, M.J. Ruiz-Moreno, B. Puertas, E. Cantos-Villar, J.M. Moreno-Rojas, Multivariate optimization of headspace solid-phase microextraction coupled to gas chromatography-mass spectrometry for the analysis of terpenoids in sparkling wines, Talanta 208 (2020) 120483. https://doi.org/10.1016/j.talanta.2019.120483.
[17] F. Stilo, M. del P. Segura borrego, C. Bicchi, S. Battaglino, R.M. Callejón fernadez, M.L. Morales, S.E. Reichenbach, J. Mccurry, D. Peroni, C. Cordero, Delineating the extra-virgin olive oil aroma blueprint by multiple headspace solid phase microextraction and differential-flow modulated comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1650 (2021) 462232. https://doi.org/10.1016/j.chroma.2021.462232.
[18] F. Stilo, E. Liberto, N. Spigolon, G. Genova, G. Rosso, M. Fontana, S.E. Reichenbach, C. Bicchi, C. Cordero, An effective chromatographic fingerprinting workflow based on comprehensive two-dimensional gas chromatography – Mass spectrometry to establish volatiles patterns discriminative of spoiled hazelnuts (Corylus avellana L.), Food Chem. 340 (2021) 128135. https://doi.org/10.1016/j.foodchem.2020.128135.
[19] B. Mendes, J. Gonçalves, J.S. Câmara, Effectiveness of high-throughput miniaturized sorbent- and solid phase microextraction techniques combined with gas chromatography–mass spectrometry analysis for a rapid screening of volatile and semi-volatile composition of wines—A comparative study, Talanta 88 (2012) 79–94. https://doi.org/10.1016/j.talanta.2011.10.010.
[20] T. Ilc, D. Werck-Reichhart, N. Navrot, Meta-Analysis of the Core Aroma Components of Grape and Wine Aroma, Front. Plant Sci. 7 (2016) 1472. https://doi.org/10.3389/fpls.2016.01472.
[21] J.V. Seeley, N.E. Schimmel, S.K. Seeley, The multi-mode modulator: A versatile fluidic device for two-dimensional gas chromatography, J. Chromatogr. A (2017). https://doi.org/10.1016/j.chroma.2017.06.030.
[22] B.C. Reaser, B.W. Wright, R.E. Synovec, Using Receiver Operating Characteristic Curves To Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry, Anal. Chem. 89 (2017) 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991.
[23] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of

comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.
[24] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.
[25] N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset, J. Chromatogr. A 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.
[26] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.
[27] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury, Anal. Chem. 92 (2020) 15526–15533. https://doi.org/10.1021/acs.analchem.0c03456.
[28] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1644 (2021) 462092. https://doi.org/10.1016/j.chroma.2021.462092.
[29] C.N. Cain, P.E. Sudol, K.L. Berrier, R.E. Synovec, Development of variance rank initiated-unsupervised sample indexing for gas chromatography-mass spectrometry analysis, Talanta 233 (2021) 122495. https://doi.org/10.1016/j.talanta.2021.122495.
[30] R.M. Heiberger, E. Neuwirth, R Through Excel, Springer New York, New York, NY, 2009. https://doi.org/10.1007/978-1-4419-0052-4.
[31] C.D. Brown, H.T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemom. Intell. Lab. Syst. 80 (2006) 24–38. https://doi.org/10.1016/j.chemolab.2005.05.004.
[32] J. Yin, J. Xie, X. Guo, L. Ju, Y. Li, Y. Zhang, Plasma metabolic profiling analysis of cyclophosphamide-induced cardiotoxicity using metabolomics coupled with UPLC/Q-TOF-MS and ROC curve, J. Chromatogr. B 1033–1034 (2016) 428–435. https://doi.org/10.1016/j.jchromb.2016.08.042.
[33] I. Ruisánchez, A.M. Jiménez-Carvelo, M.P. Callao, ROC curves for the optimization of one-class model parameters. A case study: Authenticating extra virgin olive oil from a Catalan protected designation of origin, Talanta 222 (2021) 121564. https://doi.org/10.1016/j.talanta.2020.121564.
[34] G. Ivosev, L. Burton, R. Bonner, Dimensionality Reduction and Visualization in Principal Component Analysis, Anal. Chem. 80 (2008) 4933–4944. https://doi.org/10.1021/ac800110w.
[35] R.A. van den Berg, H.C. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7 (2006) 142. https://doi.org/10.1186/1471-2164-7-142.

[36] I. Aloisi, B. Giocastro, A. Ferracane, T.M.G. Salerno, M. Zoccali, P.Q. Tranchida, L. Mondello, Preliminary observations on the use of a novel low duty cycle flow modulator for comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1643 (2021) 462076. https://doi.org/10.1016/j.chroma.2021.462076.
[37] S. Schöneich, T.J. Trinklein, C.G. Warren, R.E. Synovec, A systematic investigation of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry with dynamic pressure gradient modulation for high peak capacity separations, Anal. Chim. Acta 1134 (2020) 115–124. https://doi.org/10.1016/j.aca.2020.08.023.
[38] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain, J. Chromatogr. A 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.
[39] The Good Scents Company - Flavor, Fragrance, Food and Cosmetics Ingredients information, Good Scents Co. (n.d.). http://www.thegoodscentscompany.com/ (accessed August 23, 2021).
[40] E. Paula Barros, N. Moreira, G. Elias Pereira, S.G.F. Leite, C. Moraes Rezende, P. Guedes de Pinho, Development and validation of automatic HS-SPME with a gas chromatography-ion trap/mass spectrometry method for analysis of volatiles in wines, Talanta 101 (2012) 177–186. https://doi.org/10.1016/j.talanta.2012.08.028.
[41] E. Campo, V. Ferreira, A. Escudero, J. Cacho, Prediction of the Wine Sensory Properties Related to Grape Variety from Dynamic-Headspace Gas Chromatography−Olfactometry Data, J. Agric. Food Chem. 53 (2005) 5682–5690. https://doi.org/10.1021/jf047870a.
[42] S.-T. Chin, G.T. Eyres, P.J. Marriott, Identification of potent odourants in wine and brewed coffee using gas chromatography-olfactometry and comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1218 (2011) 7487–7498. https://doi.org/10.1016/j.chroma.2011.06.039.
[43] J.E. Welke, M. Zanus, M. Lazzarotto, C. Alcaraz Zini, Quantitative analysis of headspace volatile compounds using comprehensive two-dimensional gas chromatography and their contribution to the aroma of Chardonnay wine, Food Res. Int. 59 (2014) 85–99. https://doi.org/10.1016/j.foodres.2014.02.002.
[44] A.L. Robinson, P.K. Boss, H. Heymann, P.S. Solomon, R.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, J. Chromatogr. A 1218 (2011) 504–517. https://doi.org/10.1016/j.chroma.2010.11.008.
[45] G. Dugo, F.A. Franchina, M.R. Scandinaro, I. Bonaccorsi, N. Cicero, P.Q. Tranchida, L. Mondello, Elucidation of the volatile composition of Marsala wines by using comprehensive two-dimensional gas chromatography, Food Chem. 142 (2014) 262–268. https://doi.org/10.1016/j.foodchem.2013.07.061.
[46] K.P. Nicolli, A.C.T. Biasoto, É.A. Souza-Silva, C.C. Guerra, H.P. dos Santos, J.E. Welke, C.A. Zini, Sensory, olfactometry and comprehensive two-dimensional gas chromatography analyses as appropriate tools to characterize the effects of vine management on wine aroma, Food Chem. 243 (2018) 103–117. https://doi.org/10.1016/j.foodchem.2017.09.078.
[47] B.T. Weldegergis, A. de Villiers, C. McNeish, S. Seethapathy, A. Mostafa, T. Górecki, A.M. Crouch, Characterisation of volatile components of Pinotage wines using comprehensive two-dimensional gas chromatography coupled to time-of-flight mass

spectrometry (GC×GC–TOFMS), Food Chem. 129 (2011) 188–199. https://doi.org/10.1016/j.foodchem.2010.11.157.
[48] K. Furdíková, L. Bajnociová, F. Malík, I. Špánik, Investigation of volatile profile of varietal Gewürztraminer wines using two-dimensional gas chromatography, J. Food Nutr. Res. 56 (2017) 73–85.
[49] M. Salinas, A. Zalacain, F. Pardo, G.L. Alonso, Stir Bar Sorptive Extraction Applied to Volatile Constituents Evolution during Vitis vinifera Ripening, J. Agric. Food Chem. 52 (2004) 4821–4827. https://doi.org/10.1021/jf040040c.
[50] R. López, M. Aznar, J. Cacho, V. Ferreira, Determination of minor and trace volatile compounds in wine by solid-phase extraction and gas chromatography with mass spectrometric detection, J. Chromatogr. A 966 (2002) 167–177. https://doi.org/10.1016/S0021-9673(02)00696-9.
[51] J.E. Welke, V. Manfroi, M. Zanus, M. Lazzarotto, C. Alcaraz Zini, Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data, Food Chem. 141 (2013) 3897–3905. https://doi.org/10.1016/j.foodchem.2013.06.100.
[52] M. Del Carlo, A. Pepe, G. Sacchetti, D. Compagnone, D. Mastrocola, A. Cichelli, Determination of phthalate esters in wine using solid-phase extraction and gas chromatography–mass spectrometry, Food Chem. 111 (2008) 771–777. https://doi.org/10.1016/j.foodchem.2008.04.065.
[53] H. Chen, H. Gao, X. Fang, L. Ye, Y. Zhou, H. Yang, Effects of allyl isothiocyanate treatment on postharvest quality and the activities of antioxidant enzymes of mulberry fruit, Postharvest Biol. Technol. 108 (2015) 61–67. https://doi.org/10.1016/j.postharvbio.2015.05.011.
[54] B. Yang, L. Li, H. Geng, C. Zhang, G. Wang, S. Yang, S. Gao, Y. Zhao, F. Xing, Inhibitory effect of allyl and benzyl isothiocyanates on ochratoxin a producing fungi in grape and maize, Food Microbiol. 100 (2021) 103865. https://doi.org/10.1016/j.fm.2021.103865.
[55] B. Lantz, The large sample size fallacy, Scand. J. Caring Sci. 27 (2013) 487–492. https://doi.org/10.1111/j.1471-6712.2012.01052.x.
[56] J.W. Tukey, The Philosophy of Multiple Comparisons, Stat. Sci. 6 (1991) 100–116. https://www.jstor.org/stable/2245714 (accessed August 6, 2021).
[57] L. Held, M. Ott, How the Maximal Evidence of P-Values Against Point Null Hypotheses Depends on Sample Size, Am. Stat. 70 (2016) 335–341. https://doi.org/10.1080/00031305.2016.1209128.
[58] R.G. Brereton, P values and multivariate distributions: Non-orthogonal terms in regression models, Chemom. Intell. Lab. Syst. 210 (2021) 104264. https://doi.org/10.1016/j.chemolab.2021.104264.
[59] B. Xiao, Y. Li, B. Sun, C. Yang, K. Huang, H. Zhu, Decentralized PCA modeling based on relevance and redundancy variable selection and its application to large-scale dynamic process monitoring, Process Saf. Environ. Prot. 151 (2021) 85–100. https://doi.org/10.1016/j.psep.2021.04.043.
[60] V.E. de Almeida, D.D. de Sousa Fernandes, P.H.G.D. Diniz, A. de Araújo Gomes, G. Véras, R.K.H. Galvão, M.C.U. Araujo, Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data, Food Chem. 363 (2021) 130296. https://doi.org/10.1016/j.foodchem.2021.130296.

[61] H. Yamamoto, H. Yamaji, Y. Abe, K. Harada, D. Waluyo, E. Fukusaki, A. Kondo, H. Ohno, H. Fukuda, Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables, Chemom. Intell. Lab. Syst. 98 (2009) 136–142. https://doi.org/10.1016/j.chemolab.2009.05.006.

Chapter 5. Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry

This chapter was reproduced from Paige E. Sudol, Grant S. Ochoa, Caitlin N. Cain, Robert E. Synovec, "Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry" Submitted to *Analytica Chimica Acta* (2021).

5.1. INTRODUCTION

Since its inception in 1991 by Liu and Phillips [1], comprehensive two-dimensional (2D) gas chromatography (GC×GC) has evolved into a powerful, interdisciplinary analytical tool for the separation of volatile and semi-volatile mixtures, including petroleum-based fuels [2,3], aerosols [4–6], food products [7–9], wastewater effluent [10–12], and biological fluids such as blood and urine [13–16]. GC×GC offers an approximately ten-fold increase in peak capacity relative to one-dimensional gas chromatography (1D-GC) [17,18], but especially complex mixtures still exhibit considerable peak overlap [19–21]. Furthermore, when GC×GC is coupled to time-of-flight mass spectrometry (TOFMS) detection, analysis of multiple samples produces an array of three-dimensional (3D) data which is too complex for manual interpretation. Thus, targeted and non-targeted chemometric analysis tools are invaluable for extracting relevant chemical information from complex GC×GC-TOFMS datasets [22–25].

Non-targeted, or "discovery-based" analyses can be classified as supervised or unsupervised, depending upon whether or not sample class information is known *a priori* [26–28]. Tile-based Fisher ratio (F-ratio) analysis is an example of a widely utilized supervised method for the discovery of class-distinguishing analytes in GC×GC-TOFMS chromatograms [29–35]. Briefly, prior to the F-ratio metric calculation, the raw GC×GC-TOFMS data is

summed within defined tiles along the first and second dimension (^{1}D and ^{2}D), which offers numerous advantages relative to a pixel-level approach, namely (1) mitigation of retention time misalignment, (2) *S/N* enhancement, and (3) reduction of redundant hits per unique analyte via a novel pinning and clustering algorithm [30]. The output is a hitlist with "hits" (i.e., 2D chromatographic locations) ranked in order of descending F-ratio, whereby the hits at the top of the hitlist are more likely to be class-distinguishing based upon passing a *t*-test. This software platform has evolved into a commercially available chemometric tool of choice [36,37]. While tile-based F-ratio analysis requires a supervised experimental design governed by knowing sample class membership, in numerous analytical scenarios lack of class membership information is a severe impediment for its use. For example, pharmaceutical drug evaluation [38], food quality and authenticity studies [39–41], fuel compliance investigations [42–44], and forensic document examinations [45,46], can often be poorly suited to implementation of a supervised data analysis. Hence, development of an unsupervised counterpart to a supervised F-ratio analysis, but *implemented within* the tile-based software platform, is of keen interest to address this analytical challenge.

Principal component analysis (PCA), hierarchical cluster analysis (HCA), and *k*-means clustering are examples of heavily utilized unsupervised methods for elucidating diverse GC×GC datasets, including white truffles stored under varying conditions [47], real-time pyrolytic interactions between PVC and other plant materials [48], and weathered versus un-weathered ignitable liquids [49]. Yet, the main hindrance with these unsupervised tools is irrelevant variables that can often obscure meaningful chemical variation. Therein lies the usefulness of unsupervised feature ranking tools embedded within an unsupervised classification workflow,

which include high-level eigen-problem optimizations [50], distance-based rankings such as Euclidean distance and Laplacian score [51,52], and statistical metrics such as variance [53].

Recently, Cain et al. introduced variance rank initiated-unsupervised sample indexing (VRI-USI) as an unsupervised algorithm for finding sample classes in GC-MS data [54]. This algorithm utilizes the relative variance (RSD^2) per chromatographic peak for feature ranking prior to *k*-means clustering. The total relative variance, RSD^2_T, can be expressed as

$$RSD^2_T = RSD^2_{Chem} + RSD^2_B \quad (5.1)$$

where RSD^2_{Chem} is the chemically significant relative signal variance and RSD^2_B is the "background" relative signal variance. The RSD^2_B can be further broken down as follows,

$$RSD^2_B = RSD^2_{NV} + RSD^2_{SP} + RSD^2_{Inj} + RSD^2_{Det} \quad (5.2)$$

with subscripts corresponding to the four major potential sources of uncertainty in GC×GC-TOFMS datasets: natural variation (NV), such as biological variation, sample preparation (SP), injection (Inj), and detection (Det). Unlike hierarchical clustering methods, *k*-means clustering is a partitioning technique which outputs an index assignment for each sample to one of a defined number of clusters, *k* [55–57]. Upon application of VRI-USI analysis to a complex human cancer dataset of malignant and control patients ($k = 2$), Cain et al. discovered 5 out of 48 peaks with matching sample index assignments. Although these 5 hits were not at the top of the variance ranked hitlist, based upon a probabilistic argument, the re-occurrence of these 5 sample index assignments indicated an underlying "true" sample-to-sample relationship in the data [54].

While this previous study demonstrated VRI-USI analysis as a powerful unsupervised tool for identifying up to two sample classes in GC-MS datasets, the limits of VRI-USI for finding sample-to-sample relationships in more complex analytical settings such as GC×GC-TOFMS datasets requires development. For example, in the study of the human cancer dataset,

VRI-USI was applied to a previously curated peak table rather than to the raw data. In a truly unsupervised workflow, an analyst would not want to initially exclude any data, and the number of matching index assignments for true chemical differences versus noise would impact accurate assessment of the sample classes present. Furthermore, with the immense advancements in multidimensional GC-based instrumentation in recent years and the resulting ability of analysts to collect increasingly large, complex datasets, multi-class chemometric classification techniques have become highly desirable relative to their binary counterparts [58–61].

Herein, we introduce tile-based VRI-USI analysis for finding sample-to-sample relationships in unsupervised GC×GC-TOFMS datasets. Using the original tile-based F-ratio analysis software platform, we use a different, unsupervised metric to rank the hitlist, namely the relative signal variance (RSD^2). Each hit in the variance ranked hitlist is then interrogated to determine its cluster index assignments. Relative to our previous work, we broaden the scope of VRI-USI to finding binary *and* multi-class sample index assignments by performing *k*-means clustering on numerous numbers of clusters, *k*, and selecting the best clustering solution with the well-known silhouette metric [62–64]. We then examine the re-occurrence of index assignments indicative of initially unknown sample classes via a probabilistic argument. First, the principles of tile-based VRI-USI analysis are explored using a GC×GC-TOFMS dataset of a JP8 jet fuel spiked with a sulfur-containing analyte mix at three concentrations: 30-ppm, 15-ppm, and 0-ppm (i.e., neat) [35,65]. Next, to study the applicability of VRI-USI analysis at concentrations approaching the analyte limit of quantification (*LOQ*), VRI-USI is performed on 3-ppm versus neat spiked JP8 jet fuel. Finally, VRI-USI is applied to a more complex multi-class dataset consisting of three jet fuels (J1800A, JP4, and JP8), each spiked with the sulfur-containing analyte mix at 30-ppm and neat. Ultimately, the utility of the silhouette metric for multi-class

VRI-USI is examined by comparing the silhouette metrics across the identified classes and/or sub-classes.

5.2 EXPERIMENTAL

JP8 jet fuel was spiked with a mixture of 14 sulfur-containing analytes at nominal concentrations of 30-ppm, 15-ppm, and 3-ppm to go along with the un-spiked neat fuel [35,65]. J1800A and JP4 jet fuel were spiked with the sulfur-containing analyte mix at a nominal concentration of 30-ppm to go along with the neat fuels. Four replicates each of the spiked and four replicates of the neat fuels (except JP8, eight neat replicates collected) were analyzed with the Pegasus BT 4D GC×GC-TOFMS instrument (LECO Corporation, St. Joseph, MI) using a quad-jet thermal modulator, an Agilent 7890 gas chromatograph (Agilent Technologies, Palo Alto, CA) and an L-PAL3 GC autosampler. A 0.5 µl aliquot of each replicate was injected at a split ratio of 200:1. The ^{1}D column (26.4 m × 250 µm i.d. × 0.25 µm film thickness) contained a polar Rxi-17Sil MS stationary phase (Restek, Bellefonte, PA) and the ^{2}D column (1.9 m × 180 µm i.d. × 0.18 µm film thickness) contained a non-polar Rxi-1 MS stationary phase. Ultrahigh purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) was the carrier gas at a constant flow of 2.0 ml/min. The ^{1}D column was initially held at 40°C for 1.5 min, increased to 200°C with a temperature program rate of 5°C/min, and finally held at 200°C for 1 min. The ^{2}D column and modulator block utilized the same temperature program with offsets of 12°C and 30°C, respectively. A modulation period of P_M = 3 s was used. After an acquisition delay of 10 s, mass channels (*m/z*) 45-200 were collected at a collection frequency of 100 spectra/s.

After the data were transferred from the instrument PC using the LECO ChromaTOF for BT software, the data analysis was performed in MATLAB R2021a (The Mathworks Inc., Natick, MA, USA) on a lab top PC. The chromatograms were baseline corrected with an in-

house written rolling minimum function and normalized to the total ion current (TIC) signal. VRI-USI analysis was performed on three increasingly complex and/or challenging comparisons: Comparison (1) 30-ppm versus 15-ppm versus neat JP8 jet fuel; Comparison (2) 3-ppm versus neat JP8 jet fuel, and Comparison (3) 30-ppm versus neat for the three jet fuels J1800A, JP4, and JP8. For all three VRI-USI comparisons, the ^{1}D window before 2.5 min was excluded due to solvent peaks. For Comparisons (1) and (2), the ^{1}D window from 2.5 min to 25 min was utilized, as minimal chemical information was present past 25 min. For Comparison (3), the ^{1}D window from 2.5 min to 30 min was selected for study due to the higher boiling point composition of J1800A. A tile size of 4 modulations (12 s) on ^{1}D and 600 ms on ^{2}D was selected for tile-based VRI-USI, along with a cluster window size of 12 s on ^{1}D and 350 ms on ^{2}D for redundant hit removal. A *S/N* threshold of ten times the standard deviation of a tiled noise region (σ_{BN}, defined as the first 7000 pixels) was applied to remove hits with low signal. In light of previous work demonstrating the advantage of ranking F-ratio hitlists with the top F-ratio *m/z*, rather than an average F-ratio, the VRI-USI hitlist was ranked by the top RSD^2 *m/z* per hit [35].

Using the variance ranked hitlist, summed ^{2}D peaks were prepared by extracting the hit tiles (12 s on ^{1}D and 600 ms on ^{2}D) at the top RSD^2 *m/z* per hit, centered around $^{1}t_R$ and $^{2}t_R$, and summing away the ^{1}D separation. Although the tile-based software performs an initial "global" baseline correction (as previously mentioned), for each hit a "secondary" baseline correction was performed on the summed ^{2}D peaks by fitting a line-of-best-fit to the first three and last three data points and subtracting this line from the data (more information in Appendix D). The secondary baseline correction was deemed necessary to optimize the performance of the subsequent *k*-means clustering step by refining the hitlist, i.e., by moving true hits that exhibited

a peak up the hitlist and moving hits down the hitlist that only exhibited enough signal to exceed the user-selected *S/N* threshold.

These secondary baseline-corrected summed ^{2}D peaks per hit (12 chromatograms for Comparison (1); 8 chromatograms for Comparison (2); and 24 chromatograms for Comparison (3)) were submitted to PCA using PLS Toolbox 8.6.2 (Eigenvector Research, Inc., Wenatchee, WA, USA). For Comparisons (1) and (2), the resulting PC1 and PC2 scores were analyzed with *k*-means clustering at $k = 2$, 3, and 4. For Comparison (3), the PC1 and PC2 scores were analyzed with k-means clustering at $k = 2$, 3, 4, 5, and 6. For each value of *k* tested per hit, 100 sets of random initial cluster centroids were selected and a maximum of 1000 iterations per cluster centroid initializations were run to reach the global minimum solution, the clustering solution with the lowest sum of distances between the centroids and respective cluster members. More specifically, the Manhattan distance (i.e., the city-block distance) was used as the distance metric within *k*-means. The *k*-means clustering solutions were run in parallel using the Parallel Computing Toolbox in MATLAB R2021a (The Mathworks Inc., Natick, MA, USA).

For the clustering solutions obtained at the various values of *k* tested, the silhouette value S_i per individual sample *i* was calculated as follows,

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \tag{5.3}$$

where a_i is the average distance from *i* to members of the same cluster and b_i is the minimum of the average distances from *i* to members of different clusters [62–64]. S_i ranges from -1 to 1, with S_i approaching 1 indicating optimum clustering. The typical approach in the literature is to average the individual S_i values to get *S*; the largest *S* across clustering solutions at different values of *k* is indicative of the optimum number of clusters and will be termed S_{max} herein.

Following the determination of S_{max} for each hit, all the unique index assignments at S_{max} were identified and their number of occurrences tallied. As was defined in our previous work, the number of possible combinations (N) of samples in individual k-means clusters at a given value of k can be defined by the multinomial coefficient formula [54,66],

$$N = \binom{s}{r_1, r_2, \ldots, r_k} = \frac{s!}{r_1! r_2! \ldots r_k!} \tag{5.4}$$

where s is the total number of samples for $s>2$ and $\{r_1, r_2, \ldots r_k\}$ is the number of samples assigned to each cluster, k. Simpler combinatorial scenarios involving only one or two subsets, r, of s (i.e., $k = 2$ in k-means clustering) enable use of the binomial coefficient formula, defined as,

$$N = \binom{s}{r_k} = \frac{s!}{r_k!(s-r_k)!} \tag{5.5}$$

Although in our previous work Eq. (5.4) or (5.5) were used to calculate the number of possible combinations of *specific* index assignments at a given k, herein we extend application of Eq. (5.5) to determining the number of possible *patterns* in index assignments for the more complex Comparison (3). Given a probability of success, $p_k = 1/N$, for N calculated using Eq. (5.4) or (5.5), binomial probabilities P_x were calculated as follows,

$$P_x = \binom{n}{x} {p_k}^x (1 - p_k)^{n-x} \tag{5.6}$$

where n is the number of hits in the hitlist, and x is the number of hits with matching sample index assignments [54,66]. Index assignments with a smaller P_x have a lower chance of occurring due to random chance and are thus more likely to be due to underlying sample class differences. For Comparison (3), given the complex nature of the dataset, a MATLAB script was written to identify patterns in index assignments which contradicted the most frequently re-occurring index assignments, specifically using the “ismember” and “intersect” functions.

5.3 RESULTS AND DISCUSSION

A GC×GC total ion current (TIC) chromatogram for the 30-ppm spiked JP8 jet fuel is provided in Fig. 5.1(A). The locations of the 14 spiked sulfur-containing analytes (Table D.1 in Appendix D) are indicated with black circles. It is important to note that the spiked analytes are not visible in the TIC signal relative to the JP8 jet fuel background. This fact is manifested in the PCA scores plot of the 30-ppm (green), 15-ppm (blue), and neat (red) JP8 chromatograms for Comparison (1) in Fig. 5.1(B), wherein the samples do not cluster by spike level, indicating that the small concentration spike level differences are overwhelmed by the run-to-run variation in the background JP8 jet fuel peaks that are present even after chromatogram baseline correction and normalization. To address this challenge, we introduce tile-based VRI-USI analysis which begins with an unsupervised RSD^2 ranking as an extension of the tile-based F-ratio software platform. This first step in VRI-USI analysis is to simply replace the F-ratio metric calculation in the tile-based software with the RSD^2 metric calculation.

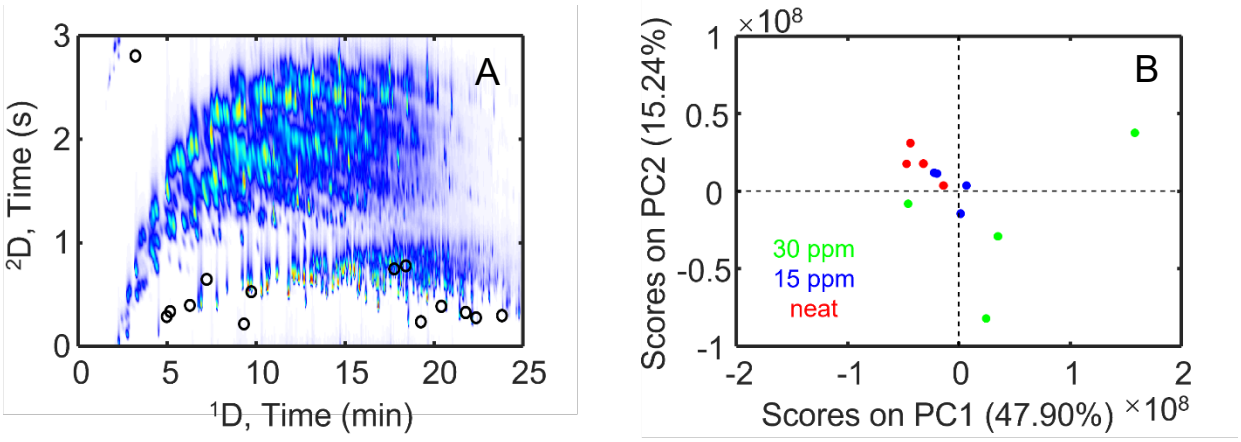

Figure 5.1. (A) Total ion current (TIC) GC×GC-TOFMS chromatogram for 30 ppm-spiked JP8 jet fuel, with locations of 14 spiked sulfur-containing analytes labeled with black circles. (B) PCA scores plot of unfolded 30-ppm (green), 15-ppm (blue), and neat (red) chromatograms.

The tile dimension is illustrated in Fig. 5.2 in terms of peak widths along ^{1}D and ^{2}D ($^{1}W_{b}$ and $^{2}W_{b}$) for 2-butyl-5-ethylthiophene (Table D.1) in Comparison (1), where the tile dimensions of 12 s on ^{1}D and 600 ms on ^{2}D are shown with dashed black boxes centered on the ^{1}D and ^{2}D retention times ($^{1}t_{R}$ and $^{2}t_{R}$, respectively). For the TIC signal data, the neat fuel chromatogram (Fig. 5.2(A)) is indistinguishable from the 15-ppm (Fig. 5.2(B)) and 30-ppm (Fig. 5.2(C)) chromatograms, as there is a native fuel peak which elutes at approximately the same $^{1}t_{R}$ and $^{2}t_{R}$ as 2-butyl-5-ethylthiophene. However, for the selective *m/z* 125, there is a clear difference between the neat (Fig. 5.2(D)), 15-ppm (Fig. 5.2(E)), and 30-ppm (Fig. 5.2(F)) tiles, as 2-butyl-5-ethylthiophene is not present natively in JP8 but observed with an increased signal intensity at 30-ppm relative to the neat and 15-ppm spike level.

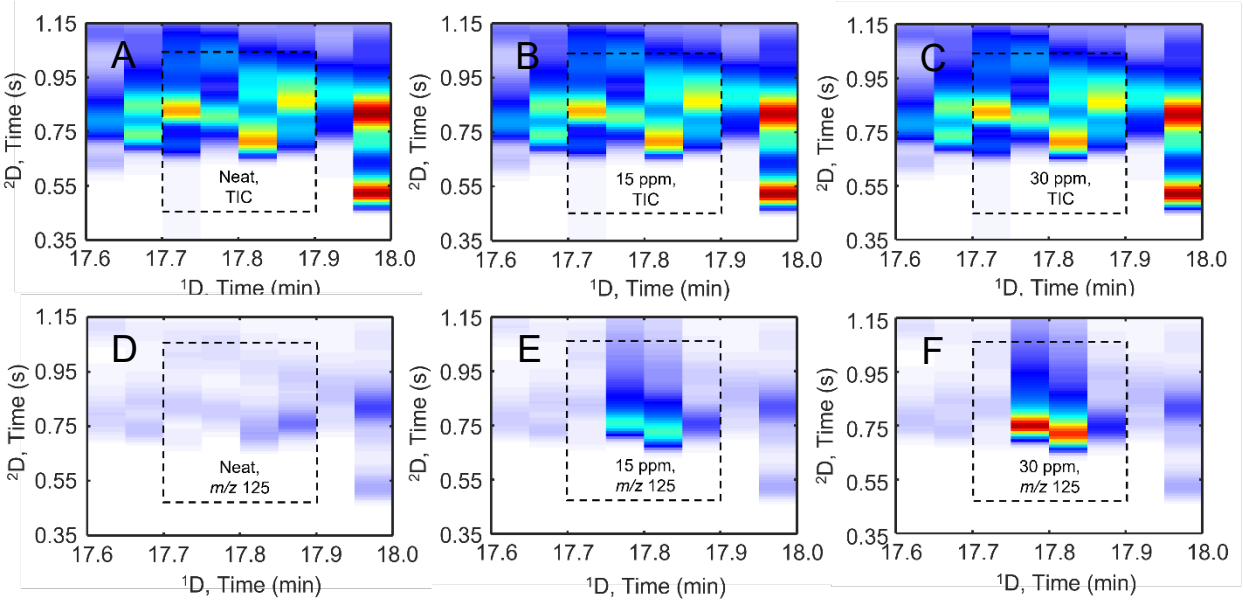

Figure 5.2. Illustration of advantage of tiling method for non-targeted analysis of the 30-ppm spiked versus 15-ppm spiked versus neat JP8 jet fuel chromatograms, using 2-butyl-5-ethylthiophene at its selective *m/z* 125. (A) Average TIC GC×GC chromatogram 2D section for the neat sample. The chosen tile size of 4 modulations on ^{1}D and 600 ms on ^{2}D is shown with a dashed black box, centered on the location of 2-butyl-5-ethylthiophene for illustration purposes. (B) Average TIC GC×GC chromatogram 2D section for the 15-ppm spiked sample. (C) Average TIC GC×GC chromatogram 2D section for the 30-ppm spiked sample. (D) Average GC×GC chromatogram at *m/z* 125 for the neat sample. (E) Average GC×GC chromatogram at *m/z* 125 for the 15-ppm spiked sample. (F) Average GC×GC chromatogram at *m/z* 125 for the 30-ppm spiked sample.

The tile-based VRI-USI analysis results (Fig. 5.3(A-C)) for 30-ppm versus 15-ppm versus neat samples in Comparison (1) is provided in Fig. 5.3. In Fig. 5.3(A), the logarithm of RSD^2 for all *m/z* is plotted versus the logarithm of summed ^{2}D peak area for all 467 hits in the hitlist, with the top RSD^2 *m/z* of the 14 spiked analytes (i.e., true positives) indicated by red dots, the secondary RSD^2 *m/z* of these true positives by gold dots, and the false positive RSD^2 *m/z* by blue dots. An analogous plot utilizing just the top RSD^2 *m/z* for each hit is shown in Fig. 5.3(B), while the distribution of the log (RSD^2) from Fig. 5.3(B) is provided in Fig. 5.3(C). Although VRI-USI is unsupervised and which "dots" in Fig. 5.3(A-B) are true versus false positives would not be available in practice at this stage of the analysis, these plots are informative as we are demonstrating method development; in Fig. 5.3(A-B), the red dots of the highest ranked true positives are clearly distinguishable from the large cloud of false positive blue dots, which indicates that the spiked sulfur-containing analytes are discovered at the top of the RSD^2 ranked hitlist. Also, since the red dots are largely distinguishable from the gold dots in Fig. 5.3(A), this validates the advantage of ranking the hits by their top RSD^2 *m/z* [35]. It is interesting to note in Fig. 5.3(A), the overall cloud of *m/z* dots exhibit a slight trend of decreasing RSD^2 with increasing peak signal, which is due to the bias introduced by the RSD^2_{Det} contribution in Eq. (5.2) to the RSD^2 metric in Eq. (5.1). An analogous set of plots for the tile-based F-ratio results are provided in Appendix D as Fig. D.1.

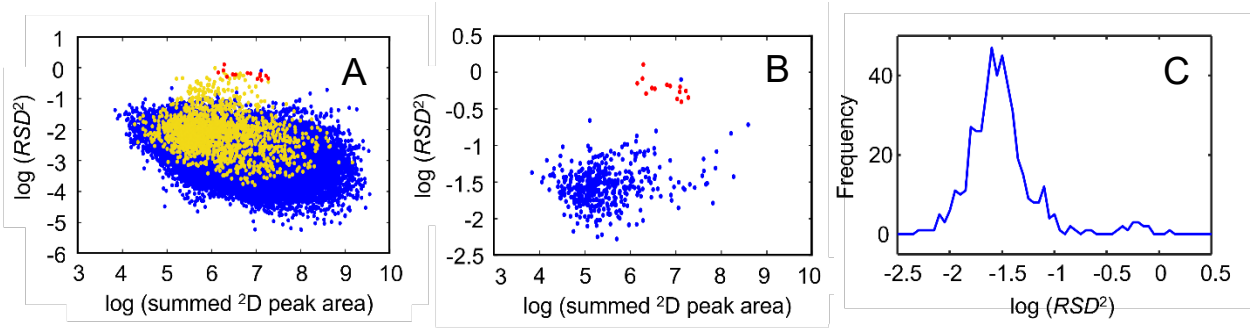

Figure 5.3. Results for tile-based VRI-USI analysis for the 30-ppm versus 15-ppm versus neat samples in Comparison (1). (A) Scatter plot of log (RSD^2) versus log (summed ^{2}D peak area per *m/z*) for all 467 hits in the VRI-USI hitlist: (red dots) top RSD^2 *m/z* for the 14 spiked sulfur-containing analytes, (gold dots) secondary RSD^2 *m/z* for the 14 spiked analytes, (blue dots) background RSD^2 *m/z*. (B) Scatter plot of log (RSD^2) versus log (summed ^{2}D peak area per *m/z*) for just the top RSD^2 *m/z*: (red dots) spiked sulfur-containing analyte *m/z* are red, (blue dots) top background RSD^2 *m/z*. (C) Distribution of log (RSD^2) for the top RSD^2 *m/z* using a bin size = 0.05.

Following ranking the hitlist by RSD^2, VRI-USI analysis is performed outside of the tile-based software platform. Next, *k*-means clustering is implemented with an example provided in Fig. 5.4 using the spiked analyte, 2-chloroethyl phenyl sulfide (hit 4) from Comparison (1). The original summed ^{2}D peaks in the neat (red), 15-ppm (blue), and 30-ppm (green) chromatograms at *m/z* 123 are provided in Fig. 5.4(A). Notably, the approximate ^2W_b of 2-chloroethyl phenyl sulfide falls well within the ^{2}D hit tile dimension, but other spiked analytes tail into neighboring hit tiles. These tailing peaks produce artifact and/or redundant hits with matching index assignments at $k = 3$ in the VRI-USI hitlist, so the summed ^{2}D peaks are baseline corrected prior to *k*-means clustering. This "secondary" baseline correction procedure is demonstrated on a representative redundant hit in Appendix D (Fig. D.2). The final baseline-corrected summed ^{2}D peaks for 2-chloroethyl phenyl sulfide are provided in Fig. 5.4(B), and subsequently used to construct the PCA scores plot in Fig. 5.4(C), wherein the 30-ppm, 15-ppm, and neat samples appear to cluster into distinct groups, hence explaining the large S_{max} of 0.933 obtained at $k = 3$. The individual S_i values (Eq. (5.3)) at $k = 2$, 3, and 4 can be visualized using the silhouette plots in Figs. 5.4(D-F), respectively. In Figs. 5.4(D-F), S_i for $i = 12$ samples is plotted on the x-axis from 0 to 1 and the number of clusters is plotted on the y axis. The individual samples are represented by blue bars, wherein the samples are grouped by index assignments along the y-axis. For 2-chloroethyl phenyl sulfide at $k = 2$, the neat and 15-ppm samples are clustered together, but these samples have low S_i from ~0.60 to 0.75, resulting in a lower $S = 0.780$ (Fig. 5.4(D)) relative to S_{max} of 0.933 at $k = 3$ (Fig. 5.4(E)). The index assignments at $k = 4$ are even less favorable, with samples 5, 6, and 9 in cluster 2 having extremely low S_i from ~ 0.2 to 0.5, resulting in $S = 0.731$ (Fig. 5.4(F)).

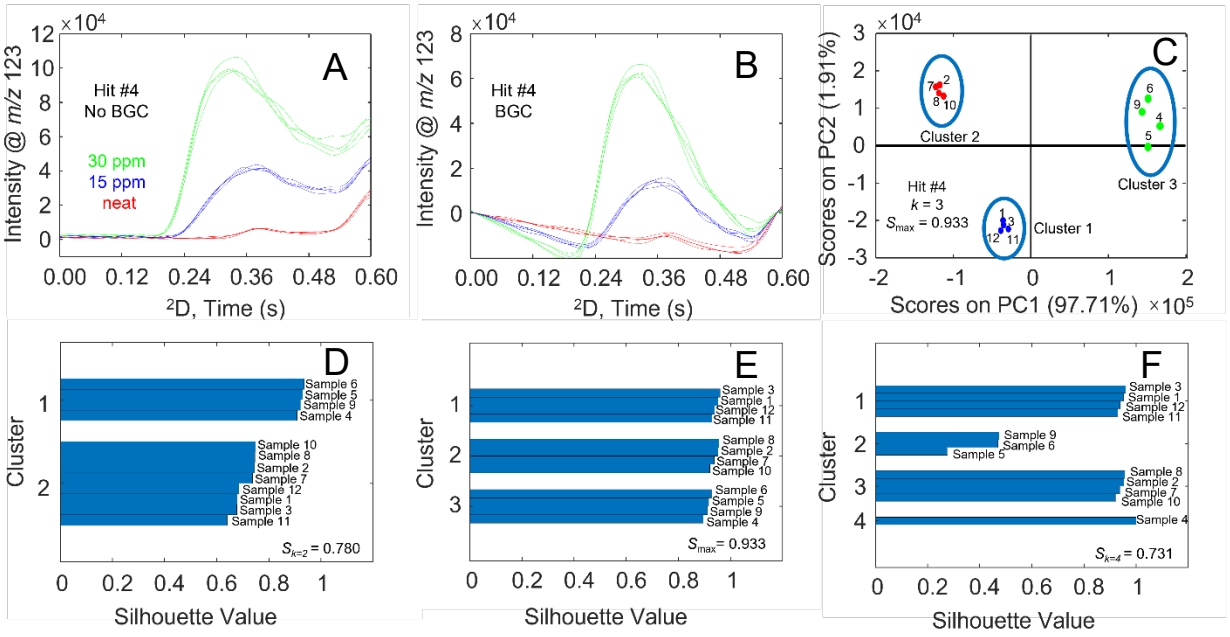

Figure 5.4. Examination of the *k*-means clustering portion of VRI-USI analysis using 2-chloroethyl phenyl sulfide at its top RSD[2] *m/z* 123. (A). Original summed ^{2}D peaks for the 30-ppm (green), 15-ppm (blue), and neat (red) samples. (B) Summed ^{2}D peaks for 30ppm, 15-ppm, and neat samples following the baseline subtraction procedure on the hit tile. (C) PCA scores plot of the baseline corrected summed ^{2}D peaks, with S_{max} and the corresponding *k* value labeled, and the samples circled according to sample index assignments at S_{max}. (D) Silhouette plot at k = 2, with samples 4-6, 9 in cluster 1 and samples 1-3,7,8,10-12 in cluster 2. (E) Silhouette plot at k = 3, with samples 1,3,11,12 in cluster 1; samples 2,7,8,10 in cluster 2; and samples 4,5,6,9 in cluster 3. S_{max} is at k = 3 for this analyte, so S has been labeled as S_{max} here. (F) Silhouette plot at k = 4, with samples 1,3,11,12 in cluster 1; samples 5,6,9 in cluster 2; samples 2,7,8,10 in cluster 3; and sample 4 in cluster 4.

The top 20 hits in the 30-ppm vs. 15-ppm vs. neat VRI-USI hitlist for Comparison (1) are provided in Table 5.1, along with the top RSD^2, top RSD^2 *m/z*, *S* at $k = 2$, 3, and 4, and index assignments at S_{max}. Eleven out of 14 of the spiked sulfur-containing analytes have an S_{max} at $k = 3$ and cluster into samples 1,3,11,12; samples 2,7,8,10; and samples 4,5,6,9, which corresponds to clustering by spike level (15-ppm, 0-ppm, and 30-ppm classes, respectively). The remaining three spiked sulfur-containing analytes exhibit different index assignments at S_{max}, with hit 1 clustering the neat and 15 ppm samples into one class (S_{max} at $k = 2$) and hits 10 and 14 splitting up the 30 ppm samples into two classes (S_{max} at $k = 4$). None the less, clustering by spike level at $k = 3$ is the most frequently re-occurring set of index assignments in the hitlist, so application of a probabilistic argument enables us to assign these index assignments to the entire hitlist. More specifically, clustering by spike level at $k = 3$ occurs for 14 hits, 13 of which are in the top 20 hits and highlighted in green in Table 5.1. The total number of combinations *N* of three clusters with four samples each is calculated to be 34,650 using Eq. (5.4); $N = 34{,}650$ is then input to Eq. (5.6) for 14 matching index assignments out of 467 total hits, returning a probability $P_x = 6.07\times10^{-38}$ of the spike level index assignments occurring by random chance. Furthermore, all of the spiked analytes had a statistically significant difference in concentration with a one-way *ANOVA p*-value < 0.05 (Table 5.1). The second most commonly re-occurring set of index assignments for Comparison (1) is samples 1-3,7,8,10-12 and samples 4-6,9 at $k = 2$, occurring for 10 hits. In this case, because there are only two *k*-means clusters, $N = 495$ using Eq. (5.5); input of this *N* to Eq. (5.6) yields a probability $P_x = 5.54\times10^{-8}$ of random chance occurrence. Interestingly, this set of index assignments is complementary to clustering by spike level at $k = 3$, in that samples 4-6,9 are still in their own cluster. In fact, aside from tetrahydrothiophene (hit 1), 9 out of 10 of hits with this set of index assignments at S_{max} are due to spurious variations in the

jet fuel background and/or interferences of the spiked sulfur-containing analytes, which highlights the weight which can be given to the most frequently re-occurring index assignments as being "class-indicating." Given previous work exploring the ability of tile-based F-ratio analysis to discover hits approaching the *LOQ*, in Comparison (2) we similarly explored the utility of VRI-USI for identifying the sample differences between 3-ppm spiked JP8 jet fuel and neat JP8 jet fuel [35]. The results of this comparison are provided in Appendix D as Fig. D.3. With VRI-USI, 8 out of 10 of the discovered sulfur-containing analytes had matching sample index assignments at $k = 2$ indicative of the two spike levels (Table D.2), which underscores the widespread applicability of VRI-USI to increasingly challenging analytical scenarios.

Table 5.1. List of top 20 hits obtained from the 30 ppm vs. 15 ppm vs. neat VRI-USI comparison, including retention time information and *S* values at $k = 2$, 3, and 4 and the corresponding index assignments at S_{max}. The common index assignments are color coded, with green shading for clustering by spike level (30 ppm = samples 4-6, 9; 15 ppm = samples 1,3,11,12; neat = samples 2,7,8,10) at $k = 3$, orange shading for clustering sample 5 on its own at $k = 2$, and yellow shading for splitting up the 30 ppm samples at $k = 4$. The *p*-values from a one-way *ANOVA* using the spike level index assignments are provided, with *p*-values < 0.05 highlighted in green.

Hit Number	[1] t_R (min)	[2] t_R (s)	RSD[2] (top *m/z*)	Analyte	$S_{k=2}$	$S_{k=3}$	$S_{k=4}$	k at S_{max}	Index assignments at S_{max}	*p*-value
1	6.30	0.40	1.28 (60)	Tetrahydrothiophene	0.888	0.764	0.666	2	samples 1-3,7,8,10-12; samples 4-6,9	8.4E-79
2	9.35	0.22	0.822 (104)	1,4-oxathiane	0.818	0.848	0.675	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	1.3E-47
3	11.05	2.51	0.797 (167)	unknown	0.854	0.430	0.392	2	samples 1-4,6-12; sample 5	9.9E-01
4	23.85	0.30	0.705 (123)	2-chloroethyl phenyl sulfide	0.780	0.933	0.731	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	1.8E-37
5	5.00	0.29	0.673 (97)	2-methylthiopheme	0.738	0.957	0.762	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	3.3E-50
6	5.20	0.34	0.648 (97)	3-methylthiophene	0.743	0.952	0.774	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	2.6E-52
7	19.30	0.24	0.630 (134)	benzo[b]thiophene	0.568	0.852	0.826	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	1.3E-37
8	7.25	0.65	0.609 (112)	2,5-dimethylthiophene	0.774	0.955	0.704	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	2.6E-66
9	3.25	2.81	0.595 (58)	thiophene	0.709	0.944	0.775	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	2.1E-33
10	18.45	0.78	0.555 (97)	2-hexylthiophene	0.662	0.857	0.898	4	samples 1,3,11,12; samples 2,7,8,10; samples 4,9; samples 5,6	8.9E-29
11	9.75	0.53	0.539 (97)	2-propylthiophene	0.729	0.919	0.821	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	6.8E-15
12	20.45	0.39	0.511 (139)	3-acetyl-2,5-dimethylthiophene	0.823	0.885	0.691	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	3.4E-33
13	21.80	0.33	0.451 (147)	2-methylbenzothiophene	0.729	0.884	0.812	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	4.1E-17
14	17.80	0.75	0.432 (125)	2-butyl-5-ethylthiophene	0.634	0.885	0.894	4	samples 1,3,11,12; samples 2,7,8,10; samples 4,9; samples 5,6	6.9E-21
15	22.40	0.28	0.395 (148)	3-methylbenzothiophene	0.722	0.901	0.767	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	6.3E-17
16	2.70	2.45	0.219 (76)	unknown	0.727	0.721	0.491	2	samples 1-3,7-12; samples 4-6	2.1E-22
17	11.05	1.33	0.192 (138)	unknown	0.865	0.493	0.387	2	samples 1-4,6-12; sample 5	1.0E+00
18	9.35	0.71	0.158 (46)	unknown	0.716	0.797	0.624	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	2.0E-30
19	11.05	1.92	0.146 (154)	unknown	0.854	0.556	0.482	2	samples 1-4,6-12; sample 5	1.0E+00
20	6.30	0.71	0.135 (88)	unknown	0.684	0.868	0.790	3	samples 1,3,11,12; samples 2,7,8,10; samples 4-6,9	8.7E-161

Next, in Comparison (3), VRI-USI analysis was rigorously evaluated with a dataset to discover "layers" of classes. This complex dataset was based upon three jet fuels (J1800A, JP4, and JP8) spiked at 30-ppm with the sulfur-containing analyte mix versus the neat fuels. Representative TIC GC×GC chromatograms of these 30-ppm fuels are provided in Fig. 5.5(A-C), respectively. J1800A is characterized by higher boiling point compounds in the alkane (^2t_R ~ 2 to 3 s), cycloalkane (^2t_R ~ 1 to 2 s), and aromatic (^2t_R ~ 0 to 1 s) regions (Fig. 5.5(A)), whereas JP4 contains several highly concentrated compounds from ~2.5 to 5 min on ^{1}D (Fig. 5.5(B)) and JP8 appears to have the most highly concentrated aromatic region (Fig. 5.5(C)). The stark chemical differences between J1800A, JP4, and JP8 are underscored in the PCA scores plot in Fig. 5.6(A), such that the 30- ppm spike versus neat differences within each fuel are not observed. Additionally, the distinguishing chemical features that produced the clustering in Fig. 5.6(A) are provided in the PC1 and PC2 loadings in Fig. 5.6(B) and Fig. 5.6(C), respectively [67]. On PC1, the highly concentrated, low boiling point compounds from ~2.5 to 5 min on ^{1}D found only in JP4 are positively loaded, along with a few select alkanes and aromatics which were not noticeably characteristic of JP4 in comparing the TIC chromatograms (Fig. 5.5). Conversely, the negatively loaded analytes on PC1 compose most of the chromatogram and thus do not help in distinguishing J1800A and JP8 (Fig. 5.6(B)). Similarly, the PC2 loadings in Fig. 5.6(C) display primarily positively loaded analytes which are essentially representative of the J1800A chromatogram, but not the specific differences between J1800A and the other fuels. Even though PCA "classifies" these fuels as chemically different, the loadings do not reveal all of the underlying sample differences such as the spiked sulfur-containing analytes, underscoring the need for VRI-USI analysis to untangle the layers of classes in this dataset.

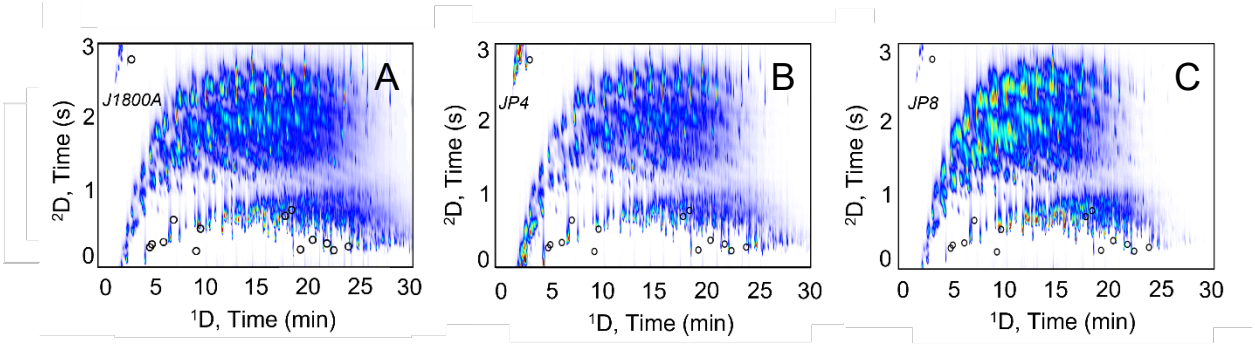

Figure 5.5. Total ion current (TIC) GC×GC-TOFMS chromatograms for (A) J1800A, (B) JP4, and (C) JP8 jet fuel used in the multi-fuel VRI-USI analysis. The locations of the 14 spiked sulfur-containing analytes are labeled accordingly.

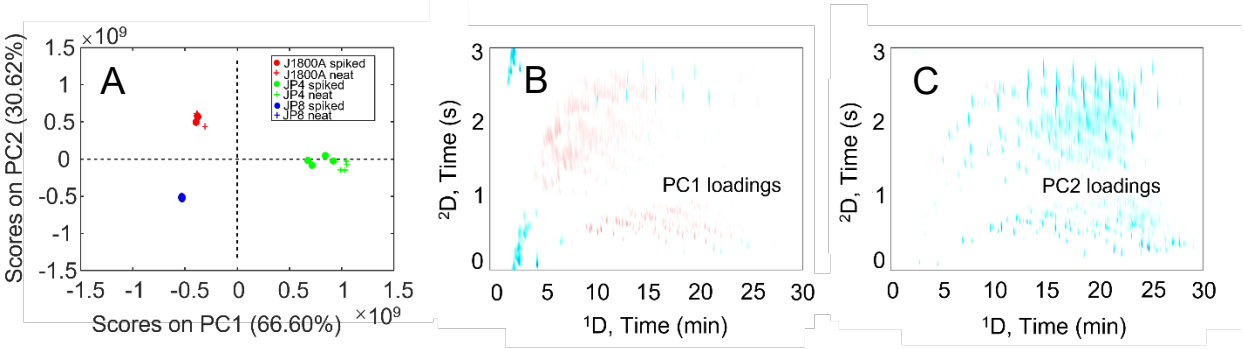

Figure 5.6. (A) PCA scores plot of unfolded 30-ppm spiked samples (circles) and neat samples (+) of J1800A (red), JP4 (green), and JP8 (blue). (B) Total ion current (TIC) GC×GC-TOFMS chromatogram of PC1 loadings, with positively loaded variable in blue and negatively loaded variables in red. (C) TIC GC×GC-TOFMS chromatogram of PC2 loadings.

Implementing VRI-USI analysis, the hitlist for Comparison (3) resulted in 520 total hits, with the results summarized in Tables 5.2 and 5.3. In Table 5.2, based upon the sample index assignments, 306 of the 520 hits grouped the samples with the following eight samples always together in some combination in the same *k*-means cluster, referred to as "ideal" clustering: samples 1-8 (the spiked and neat J1800A), samples 9-16 (the spiked and neat JP4), and samples 17-24 (the spiked and neat JP8). Another 147 hits in Table 5.2 possessed at least one of the groups of eight samples (samples 1-8; samples 9-16; samples 17-24) together in the same *k*-means cluster so the clustering by fuel type was clear, but they exhibited a slight deviation from the ideal index assignment combinations. Finally, investigation of the remaining 67 hits summarized in Table 5.3 confirmed that none of these groups of eight samples due to fuel type were together in a *k*-means cluster, with 58 of the hits exhibiting clustering contradictory to the clustering by fuel type, with seemingly random sample index assignments. However, for the other 9 of the 67 hits in Table 5.3, at least one of these "split" combinations occurred: samples 1-4 and samples 5-8 in separate clusters; samples 9-12 and samples 13-16 in separate clusters; and samples 17-20 and samples 21-24 in separate clusters. Through an unsupervised lens, this pattern in the sample index assignments for these 9 hits would be suggestive of a sub-class in the data. In fact, this pattern corresponds to the 30-ppm spike versus neat differences.

Table 5.2. List of all index assignments at S_{max} which were indicative of fuel type clustering for Comparison (3), along with the corresponding number of occurrences (i.e., hits in the hitlist) and relevant *k*. The "ideal" clustering assignments are shaded blue and the index assignments which deviated from "ideal" are shaded green. The single-occurring index assignments are grouped together for brevity.[a]

Number of occurrences (hits)	Index assignments at S_{max}	*k* at S_{max}
125	samples 1-8; samples 9-16; samples 17-24	3
116	samples 1-8; samples 9-24	2
38	samples 1-8, 17-24; samples 9-16	2
27	samples 1-16; samples 17-24	2
6	samples 1-7; samples 8-24	2
5	samples 1-8,16; samples 9-15,17-244	2
5	samples 1-7; sample 8; samples 9-16; samples 17-24	4
4	samples 1-7; samples 8-16; samples 17-24	3
4	samples 1-8,17-24; sample 9; sample 10; samples 11-16	4
3	samples 1,2,3,5-8; samples 4,9-24	2
2	samples 1,2,3,9-24; samples 4,5,6,7,8	2
2	samples 10,11,17-24; samples 1-9,12-16	2
2	samples 1-9; samples 10-24	2
2	samples 1,4-8; samples 2,3,9-24	2
2	samples 1-7; samples 9-16; samples 8,17-24	3
2	samples 1-8; samples 9-12,14,15,16; samples 13,17-24	3
2	samples 1-8; samples 9,10,11; samples 12-16; samples 17-24	4
2	samples 1-8; samples 9-12; samples 13-16; samples 17-24	4
2	samples 1-8; samples 9,12,14,15; samples 10,11; samples 13,16; samples 17-24	5
2	samples 1-8; samples 9,12; samples 10,11; samples 13,16; samples 14,15; samples 17-24	6
[a]100	[a]single-occurring assignments	[a]2-6

[a] Of the 147 hits which deviated from ideal fuel type clustering (green), 100 hits had single-occurring *k*-means clustering index assignments at various values of *k*. These index assignments are grouped together in Table 2, as they all contained *at least* one of the groups of eight samples (samples 1-8, samples 9-16, samples 17-24) in a *k*-means cluster.

Table 5.3. List of all index assignments at S_{max} which were not indicative of fuel type clustering for Comparison (3), including the corresponding number of occurrences, hit numbers, and relevant *k*. The contradictory index assignments are shaded yellow[a] and the index assignments representing sub-clustering by spike level are shaded pink. The eight discovered sulfur-containing compounds are identified accordingly, and the one false positive hit is labeled.[b]

Number of occurrences (hits)	Index assignments at S_{max}	Hit number (s)	Identity	*k* at S_{max}
2	samples 1-4,17-20; samples 9-12; samples 5-8,13-16,21-24	16 161	tetrahydrothiophene 2-hexylthiophene	3
2	samples 1-4; samples 5-8; samples 9-12; samples 13-16; samples 17-20; samples 21-24	153 289	benzo[b]thiophene 2-methylbenzothiophene	6
1	samples 1-4,9-12,17-20; samples 5-8,13-16,21-24	182	2,5-dimethylthiophene	2
1	samples 1-4,12,17-20; samples 9,10,11; samples 5-8,13-16,21-24	162	1,4-oxathiane	3
1	samples 1-4; samples 5-8; samples 9-12; samples 13-16,21-24; samples 17-20	269	3-acetyl-2,5-dimethylthiophene	5
1	samples 1-4; samples 5-8; samples 9-13; samples 14,15,16; samples 17-20; samples 21-24	210	2-chloroethyl phenyl sulfide	6
1	samples 1-4,9,10,12,13,16,19,20; samples 5-8,11,14,15,17,18,21-24	263	[b]false positive	2
[a]58	[a]single-occurring assignments with no evident patterns	[a]n/a	[a]n/a	[a]2-6

[a]The 58 contradictory index assignments are seemingly random, with no evidence of fuel type clustering or "splitting" these fuel type groups of eight samples (samples 1-4 and 5-8, samples 9-12 and 13-16, and samples 17-20 and 21-24). Since these assignments are single-occurring and have no evident patterns, they are grouped together in Table 3 for brevity.

[b]Hit 263 is labeled a false positive because it is a native fuel peak which split up the J1800A spiked and neat samples into separate *k*-means clusters due to low peak *S/N*.

The fuel type clustering most commonly observed in Table 5.2 corresponded to the set of index assignments at $k = 3$ with samples 1-8 (the spiked and neat J1800A), samples 9-16 (the spiked and neat JP4), and samples 17-24 (the spiked and neat JP8), with 125 total occurrences giving a probability $P_x = 1.40\times10^{-1124}$ of occurrence by random chance via Eq. (5.6) with $N = 9.47\times10^{9}$ using Eq. (5.4). Three other combinations of index assignments were observed at $k = 2$ with either samples 1-8, samples 9-16, or samples 17-24 in their own cluster and the remaining 16 samples in a separate cluster with $N = 73{,}5471$ combinations using Eq. (5.5). These index assignments occur for 116, 38, and 27 hits, respectively, corresponding to probabilities of occurring by random chance of 9.29×10^{-563}, 9.08×10^{-166}, and 3.97×10^{-114}. Notably, the next most commonly occurring set of index assignments (samples 1-7; samples 8-24 at $k = 2$) occurs for 6 hits, corresponding to a much greater random chance probability of 1.55×10^{-20} ($N = 346{,}104$ using Eq. (5.5)). This set of index assignments is a result of a single chromatogram (sample 8) deviating from the ideal clustering arrangement. There were 146 other hits with a similarly small deviation from the ideal sample index assignment combination for fuel type clustering.

In contrast to the hits exhibiting clustering by fuel type, in Table 5.3, of the nine hits which exhibited evidence of sub-class clustering, eight hits were due to the spiked sulfur-containing analytes, while one hit (hit 263) was a native fuel peak. Unsupervised manual examination would reveal that this is a false positive, as the J1800A spiked and neat samples coincidently split due to the low *S/N*. It is interesting to note that these spiked sulfur-containing analyte hits display S_{max} at varying values of k, depending on the native composition of each fuel. For example, 2,5-dimethylthiophene is natively present at insignificant levels in J1800A, JP4, and JP8, hence its S_{max} occurs at $k = 2$ with all the spiked and neat samples in separate clusters. Conversely, benzo[b]thiophene is present at varying concentrations in the three fuels, so its S_{max}

occurs at $k = 6$ with the spiked and neat samples per fuel forming separate clusters. While eight out of the 14 spiked sulfur-containing analytes were discovered, the remaining six spiked analytes were overshadowed by a fuel type clustering hit with a higher RSD^2 *m/z* at the given pin location. However, given the reoccurring patterns in index assignments observed, this was more than sufficient to identify the spike level sub-class clustering via a probabilistic argument, which will now be outlined in greater depth.

Since the eight spiked sulfur hits and one false positive hit have varying index assignments at S_{max} (Table 5.3), calculating the binomial probability of a specific index assignment would not be reflective of the significance of the underlying pattern. Using Eq. (5.5), the number of combinations N of four samples from a total of eight samples (samples 1-8, samples 9-16, or samples 17-24) is 70, the probability of which is 0.0143 (1/70). The probability of the eight samples "splitting" into two specific groups of four samples is thus $(1/70)^2 = 0.000204$. The random chance probability of 9/67 contradictory clustering hits in the hitlist having at least one of these splitting patterns in their index assignments is then 2.60×10^{-23} (Eq. (5.6)). Interestingly, this was the only observable pattern for the 67 contradictory clustering hits. Thus, through an unsupervised view, this pattern in index assignments is highly probable of representing a true underlying sample difference, and indeed we know that this pattern represents the spike level sub-class. This example illustrates the power of the probabilistic aspect of VRI-USI, as the spiked sulfur analytes were discovered to be class-indicating even though they were found further down the RSD^2 ranked hitlist (Table 5.3). Although the RSD^2 ranking is an important organizational step, the application of combinatorics to the *k*-means index assignments enables an analyst to dissect "layers" of classes in complex datasets.

Next, we verify the overall confidence in which the major classes and sub-classes were identified with *k*-means clustering based upon distributions of S_{max} obtained for all 520 hits, where each S_{max} was obtained as in Fig. 5.4(E) for 2-chloroethyl phenyl sulfide. For this purpose, a larger S_{max} correlates with a more confident clustering assignment. An overlay of the S_{max} distributions for the ideal fuel type clustering hits (306 hits), nearly ideal fuel type clustering hits (147 hits), and contradictory clustering hits (67 hits) is provided in Fig. 5.7(A). The ideal fuel type clustering distribution has a maximum frequency at $S_{max} \sim 0.88$, whereas the nearly ideal fuel type clustering distribution is shifted to the left to lower S_{max} with its maximum frequency at $S_{max} \sim 0.65$. These lower S_{max} values reinforce the notion that the nearly ideal fuel type clustering hits have lower *S/N* and/or smaller concentration differences, resulting in slight deviations in their sample index assignments which in turn render them less effective in indicating the three fuel classes. The contradictory clustering hits distribution has two distinct regions, with the first portion having a maximum frequency at $S_{max} \sim 0.45$ and the second portion having a maximum frequency at $S_{max} \sim 0.80$ (Fig. 5.7(B)). In fact, this second portion corresponds to the eight discovered sulfur-containing analytes. This distinction in S_{max} lends further credence to our labeling of hit 263 as a false positive hit, as hit 263 has an S_{max} value in line with the first portion of the distribution ($S_{max} = 0.431$). Since these 59 contradictory clustering hits have such low S_{max} strongly supports their spurious nature, whereas the eight spiked sulfur-containing analytes each have a large S_{max} in the same S_{max} range as the ideal fuel type clustering hits (Fig. 5.7(A)).

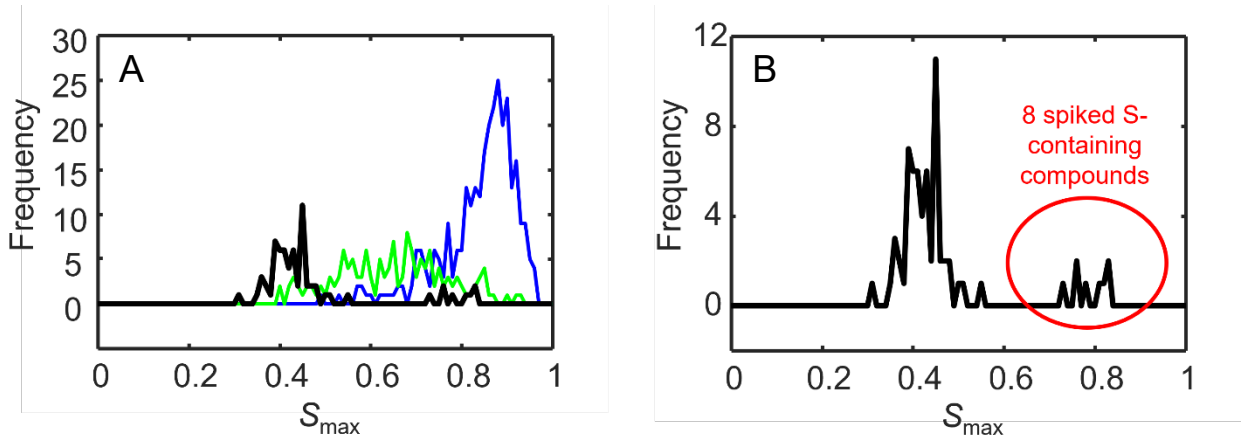

Figure 5.7. (A) Overlay of S_{max} distributions for "ideal" fuel type clustering hits (blue), "nearly ideal" fuel type clustering hits (green), and contradictory clustering hits (black). (B) Zoom plot of the S_{max} distribution for the 67 hits with contradictory index assignments. The S_{max} values of the eight discovered sulfur-containing analyte spikes are enclosed by the red circle.

5.4 CONCLUSIONS

Tile-based VRI-USI analysis has been demonstrated to be a valuable tool for unsupervised multi-class classification of complex GC×GC-TOFMS datasets. Incorporation of the S_{max} metric permitted evaluation of multiple possible numbers of clusters, k, per hit, thereby extending this technique from binary to multi-class classification scenarios. In our initial demonstration of tile-based VRI-USI analysis on a 30-ppm versus 15-ppm versus neat JP8 jet fuel in Comparison (1), clustering by spike level at $k = 3$ occurred for 14/467 hits in the hitlist, making it the most commonly re-occurring set of index assignments. Additionally, using 3-ppm versus neat JP8 jet fuel in Comparison (2), VRI-USI analysis was demonstrated to be successful for low concentration analyte discovery. Finally, in Comparison (3), VRI-USI analysis applied on a dataset of three jet fuels, spiked with the 30-ppm sulfur-containing analyte mix versus neat

samples, 453/520 hits had index assignments indicating clustering by fuel type. The remaining 67 hits had contradictory index assignments, and examination of the patterns in index assignments revealed nine hits (eight sulfur-containing analytes and one false positive) which consistently split up the fuel type groupings. Given that six of the 14 sulfur-containing analytes were "undiscoverable" amidst the three-native fuel backgrounds, future work may involve performing *k*-means clustering on all the RSD^2 *m/z* per hit and identifying patterns amongst all the index assignments per hit. Yet, the overall performance of the multi-fuel study of Comparison (3) still underscores the utility of VRI-USI analysis to untangle "layers" of classes in complex datasets.

5.5 ACKNOWLEDGEMENTS

C. N. Cain acknowledges the U.S. National Science Foundation Graduate Research Fellowship (DGE-1762114).

5.6 REFERENCES

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

[2] N. Boegelsack, C. Sandau, D.W. McMartin, J.M. Withey, G. O'Sullivan, Development of retention time indices for comprehensive multidimensional gas chromatography and application to ignitable liquid residue mapping in wildfire investigations, J. Chromatogr. A 1635 (2021) 461717. https://doi.org/10.1016/j.chroma.2020.461717.

[3] G.L. Alexandrino, J. Malmborg, F. Augusto, J.H. Christensen, Investigating weathering in light diesel oils using comprehensive two-dimensional gas chromatography–High resolution mass spectrometry and pixel-based analysis: Possibilities and limitations, J. Chromatogr. A 1591 (2019) 155–161. https://doi.org/10.1016/j.chroma.2019.01.042.

[4] Z. An, H. Ren, M. Xue, X. Guan, J. Jiang, Comprehensive two-dimensional gas chromatography mass spectrometry with a solid-state thermal modulator for in-situ speciated measurement of organic aerosols, J. Chromatogr. A 1625 (2020) 461336. https://doi.org/10.1016/j.chroma.2020.461336.

[5] Z. An, X. Li, Z. Shi, B.J. Williams, R.M. Harrison, J. Jiang, Frontier review on comprehensive two-dimensional gas chromatography for measuring organic aerosol, J. Hazard. Mater. Lett. 2 (2021) 100013. https://doi.org/10.1016/j.hazl.2021.100013.

[6] B. Savareear, J. Escobar-Arnanz, M. Brokl, M.J. Saxton, C. Wright, C. Liu, J.-F. Focant, Comprehensive comparative compositional study of the vapour phase of cigarette mainstream

tobacco smoke and tobacco heating product aerosol, J. Chromatogr. A 1581–1582 (2018) 105–115. https://doi.org/10.1016/j.chroma.2018.10.035.
[7] F. Stilo, C. Bicchi, A. Robbat, S.E. Reichenbach, C. Cordero, Untargeted approaches in food-omics: The potential of comprehensive two-dimensional gas chromatography/mass spectrometry, TrAC Trends Anal. Chem. 135 (2021) 116162. https://doi.org/10.1016/j.trac.2020.116162.
[8] F. Stilo, E. Liberto, N. Spigolon, G. Genova, G. Rosso, M. Fontana, S.E. Reichenbach, C. Bicchi, C. Cordero, An effective chromatographic fingerprinting workflow based on comprehensive two-dimensional gas chromatography – Mass spectrometry to establish volatiles patterns discriminative of spoiled hazelnuts (Corylus avellana L.), Food Chem. 340 (2021) 128135. https://doi.org/10.1016/j.foodchem.2020.128135.
[9] A.C. Paiva, L.W. Hantao, Exploring a public database to evaluate consumer preference and aroma profile of lager beers by comprehensive two-dimensional gas chromatography and partial least squares regression discriminant analysis, J. Chromatogr. A 1630 (2020) 461529. https://doi.org/10.1016/j.chroma.2020.461529.
[10] K.A. Murrell, F.L. Dorman, A comparison of liquid-liquid extraction and stir bar sorptive extraction for multiclass organic contaminants in wastewater by comprehensive two-dimensional gas chromatography time of flight mass spectrometry, Talanta 221 (2021) 121481. https://doi.org/10.1016/j.talanta.2020.121481.
[11] M. Kopperi, J. Ruiz-Jiménez, J.I. Hukkinen, M.-L. Riekkola, New way to quantify multiple steroidal compounds in wastewater by comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry, Anal. Chim. Acta 761 (2013) 217–226. https://doi.org/10.1016/j.aca.2012.11.059.
[12] A.M. Muscalu, T. Górecki, Comprehensive two-dimensional gas chromatography in environmental analysis, TrAC Trends Anal. Chem. 106 (2018) 225–245. https://doi.org/10.1016/j.trac.2018.07.001.
[13] L.M. Dubois, K.A. Perrault, P.-H. Stefanuto, S. Koschinski, M. Edwards, L. McGregor, J.-F. Focant, Thermal desorption comprehensive two-dimensional gas chromatography coupled to variable-energy electron ionization time-of-flight mass spectrometry for monitoring subtle changes in volatile organic compound profiles of human blood, J. Chromatogr. A 1501 (2017) 117–127. https://doi.org/10.1016/j.chroma.2017.04.026.
[14] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury, Anal. Chem. 92 (2020) 15526–15533. https://doi.org/10.1021/acs.analchem.0c03456.
[15] S.M. Rocha, M. Caldeira, J. Carrola, M. Santos, N. Cruz, I.F. Duarte, Exploring the human urine metabolomic potentialities by comprehensive two-dimensional gas chromatography coupled to time of flight mass spectrometry, J. Chromatogr. A 1252 (2012) 155–163. https://doi.org/10.1016/j.chroma.2012.06.067.
[16] C.H. Weinert, B. Egert, S.E. Kulling, On the applicability of comprehensive two-dimensional gas chromatography combined with a fast-scanning quadrupole mass spectrometer for untargeted large-scale metabolomics, J. Chromatogr. A 1405 (2015) 156–167. https://doi.org/10.1016/j.chroma.2015.04.011.
[17] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in

peak capacity gain, J. Chromatogr. A 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.
[18] Zaiyou. Liu, D.G. Patterson, M.L. Lee, Geometric Approach to Factor Analysis for the Estimation of Orthogonality and Practical Peak Capacity in Comprehensive Two-Dimensional Separations, Anal. Chem. 67 (1995) 3840–3845. https://doi.org/10.1021/ac00117a004.
[19] Z.-D. Zeng, S.-T. Chin, H.M. Hugel, P.J. Marriott, Simultaneous deconvolution and re-construction of primary and secondary overlapping peak clusters in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1218 (2011) 2301–2310. https://doi.org/10.1016/j.chroma.2011.02.028.
[20] J.M. Davis, Statistical theory of spot overlap in two-dimensional separations, Anal. Chem. 63 (1991) 2141–2152. https://doi.org/10.1021/ac00019a014.
[21] F.J. Oros, J.M. Davis, Comparison of statistical theories of spot overlap in two-dimensional separations and verification of means for estimating the number of zones, J. Chromatogr. A 591 (1992) 1–18. https://doi.org/10.1016/0021-9673(92)80218-J.
[22] P.-H. Stefanuto, A. Smolinska, J.-F. Focant, Advanced chemometric and data handling tools for GC×GC-TOF-MS: Application of chemometrics and related advanced data handling in chemical separations, TrAC Trends Anal. Chem. 139 (2021) 116251. https://doi.org/10.1016/j.trac.2021.116251.
[23] K.M. Pierce, T.J. Trinklein, J.S. Nadeau, R.E. Synovec, Chapter 20 - Data analysis methods for gas chromatography, in: C.F. Poole (Ed.), Gas Chromatogr. Second Ed., Elsevier, Amsterdam, 2021: pp. 525–546. https://doi.org/10.1016/B978-0-12-820675-1.00007-1.
[24] S.E. Reichenbach, C.A. Zini, K.P. Nicolli, J.E. Welke, C. Cordero, Q. Tao, Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition, J. Chromatogr. A 1595 (2019) 158–167. https://doi.org/10.1016/j.chroma.2019.02.027.
[25] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications, Anal. Chem. (2017). https://doi.org/10.1021/acs.analchem.7b04226.
[26] K.M. Pierce, J.S. Nadeau, R.E. Synovec, Chapter 17 - Data Analysis Methods, in: C.F. Poole (Ed.), Gas Chromatogr., Elsevier, Amsterdam, 2012: pp. 415–434. https://doi.org/10.1016/B978-0-12-385540-4.00017-1.
[27] K.L. Berrier, S.E. Prebihalo, R.E. Synovec, Chapter 7 - Advanced data handling in comprehensive two-dimensional gas chromatography, in: N.H. Snow (Ed.), Sep. Sci. Technol., Academic Press, 2020: pp. 229–268. https://doi.org/10.1016/B978-0-12-813745-1.00007-6.
[28] B.J. Pollo, C.A. Teixeira, J.R. Belinato, M.F. Furlan, I. Cristina de Matos Cunha, C.R. Vaz, G.V. Volpato, F. Augusto, Chemometrics, Comprehensive Two-Dimensional Gas Chromatography And “Omics” Sciences: Basic Tools And Recent Applications, TrAC Trends Anal. Chem. (2020) 116111. https://doi.org/10.1016/j.trac.2020.116111.
[29] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.
[30] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[31] N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset, J. Chromatogr. A 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.
[32] B.A. Parsons, D.K. Pinkerton, B.W. Wright, R.E. Synovec, Chemical characterization of the acid alteration of diesel fuel: Non-targeted analysis by two-dimensional gas chromatography coupled with time-of-flight mass spectrometry with tile-based Fisher ratio and combinatorial threshold determination, J. Chromatogr. A 1440 (2016) 179–190. https://doi.org/10.1016/j.chroma.2016.02.067.
[33] B.C. Reaser, B.W. Wright, R.E. Synovec, Using Receiver Operating Characteristic Curves To Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry, Anal. Chem. 89 (2017) 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991.
[34] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.
[35] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1644 (2021) 462092. https://doi.org/10.1016/j.chroma.2021.462092.
[36] ChromaTOF® Tile Analytical Software, LECO Corp. (2021). https://www.leco.com/product/chromatof-tile (accessed November 16, 2021).
[37] ChromCompare+, (2021). https://www.sepsolve.com/chromcompare/ (accessed November 16, 2021).
[38] H. Rebiere, Y. Grange, E. Deconinck, P. Courselle, J. Acevska, K. Brezovska, J. Maurin, T. Rundlöf, M.J. Portela, L.S. Olsen, C. Offerlé, M. Bertrand, European fingerprint study on omeprazole drug substances using a multi analytical approach and chemometrics as a tool for the discrimination of manufacturing sources, J. Pharm. Biomed. Anal. (2021) 114444. https://doi.org/10.1016/j.jpba.2021.114444.
[39] N.P. Kalogiouri, R. Aalizadeh, M.E. Dasenaki, N.S. Thomaidis, Application of High Resolution Mass Spectrometric methods coupled with chemometric techniques in olive oil authenticity studies - A review, Anal. Chim. Acta 1134 (2020) 150–173. https://doi.org/10.1016/j.aca.2020.07.029.
[40] E. Parente, T. Zotta, Chemometric Approaches for Identity and Authenticity Testing, Quality Assurance and Process Control☆, in: P.L.H. McSweeney, J.P. McNamara (Eds.), Encycl. Dairy Sci. Third Ed., Academic Press, Oxford, 2022: pp. 327–347. https://doi.org/10.1016/B978-0-12-818766-1.00117-3.
[41] N. Nikzad, H. Parastar, Evaluation of the effect of organic pollutants exposure on the antioxidant activity, total phenolic and total flavonoid content of lettuce (Lactuca sativa L.) using UV–Vis spectrophotometry and chemometrics, Microchem. J. 170 (2021) 106632. https://doi.org/10.1016/j.microc.2021.106632.
[42] P.E. Sudol, K.M. Pierce, S.E. Prebihalo, K.J. Skogerboe, B.W. Wright, R.E. Synovec, Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review, Anal. Chim. Acta 1132 (2020) 157–186. https://doi.org/10.1016/j.aca.2020.07.027.

[43] V.S. Pinto, Use of 1H nuclear magnetic resonance and chemometrics to detect the percentage of ethanol anhydrous in Brazilian type C premium gasoline, Fuel 276 (2020) 118015. https://doi.org/10.1016/j.fuel.2020.118015.
[44] D. Vrtiška, P. Vozka, V. Váchová, P. Šimáček, G. Kilaz, Prediction of HEFA content in jet fuel using FTIR and chemometric methods, Fuel 236 (2019) 1458–1464. https://doi.org/10.1016/j.fuel.2018.09.102.
[45] M.N.M. Asri, N.F. Nestrgan, N.A.M. Nor, R. Verma, On the discrimination of inkjet, laser and photocopier printed documents using Raman spectroscopy and chemometrics: Application in forensic science, Microchem. J. 165 (2021) 106136. https://doi.org/10.1016/j.microc.2021.106136.
[46] S. Materazzi, R. Risoluti, S. Pinci, F. Saverio Romolo, New insights in forensic chemistry: NIR/Chemometrics analysis of toners for questioned documents examination, Talanta 174 (2017) 673–678. https://doi.org/10.1016/j.talanta.2017.06.044.
[47] R. Costa, C. Fanali, G. Pennazza, L. Tedone, L. Dugo, M. Santonico, D. Sciarrone, F. Cacciola, L. Cucchiarini, M. Dachà, L. Mondello, Screening of volatile compounds composition of white truffle during storage by GCxGC-(FID/MS) and gas sensor array analyses, LWT - Food Sci. Technol. 60 (2015) 905–913. https://doi.org/10.1016/j.lwt.2014.09.054.
[48] S. Kumagai, A. Matsukami, F. Kabashima, M. Sakurai, M. Kanai, T. Kameda, Y. Saito, T. Yoshioka, Combining pyrolysis–two-dimensional gas chromatography–time-of-flight mass spectrometry with hierarchical cluster analysis for rapid identification of pyrolytic interactions: Case study of co-pyrolysis of PVC and biomass components, Process Saf. Environ. Prot. 143 (2020) 91–100. https://doi.org/10.1016/j.psep.2020.06.036.
[49] J. Pandohee, J.G. Hughes, J.R. Pearson, O. A.H Jones, Chemical fingerprinting of petrochemicals for arson investigations using two-dimensional gas chromatography - flame ionisation detection and multivariate analysis, Sci. Justice 60 (2020) 381–387. https://doi.org/10.1016/j.scijus.2020.04.004.
[50] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Association for Computing Machinery, New York, NY, USA, 2010: pp. 333–342. https://doi.org/10.1145/1835804.1835848.
[51] S.P. Patel, S.H. Upadhyay, Euclidean distance based feature ranking and subset selection for bearing fault diagnosis, Expert Syst. Appl. 154 (2020) 113400. https://doi.org/10.1016/j.eswa.2020.113400.
[52] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Proceedings of the 18th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, 2005: pp. 507–514.
[53] L. Haar, K. Anding, K. Trambitckii, G. Notni, Comparison between Supervised and Unsupervised Feature Selection Methods:, in: Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods, SCITEPRESS - Science and Technology Publications, Prague, Czech Republic, 2019: pp. 582–589. https://doi.org/10.5220/0007385305820589.
[54] C.N. Cain, P.E. Sudol, K.L. Berrier, R.E. Synovec, Development of variance rank initiated-unsupervised sample indexing for gas chromatography-mass spectrometry analysis, Talanta 233 (2021) 122495. https://doi.org/10.1016/j.talanta.2021.122495.
[55] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

[56] P. Govender, V. Sivakumar, Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), Atmospheric Pollut. Res. 11 (2020) 40–56. https://doi.org/10.1016/j.apr.2019.09.009.
[57] F. Nielsen, Hierarchical Clustering, in: Introduction to HPC with MPI for Data Science, Springer International Publishing, Cham, 2016: pp. 195–211. https://doi.org/10.1007/978-3-319-21903-5_8.
[58] X. Wang, X. Liu, J. Wang, G. Wang, Y. Zhang, L. Lan, G. Sun, Study on multiple fingerprint profiles control and quantitative analysis of multi-components by single marker method combined with chemometrics based on Yankening tablets, Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 253 (2021) 119554. https://doi.org/10.1016/j.saa.2021.119554.
[59] Q. Yang, L. Xu, L.-J. Tang, J.-T. Yang, B.-Q. Wu, N. Chen, J.-H. Jiang, R.-Q. Yu, Simultaneous detection of multiple inherited metabolic diseases using GC-MS urinary metabolomics by chemometrics multi-class classification strategies, Talanta 186 (2018) 489–496. https://doi.org/10.1016/j.talanta.2018.04.081.
[60] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, Dynamic classifier selection for One-vs-One strategy: Avoiding non-competent classifiers, Pattern Recognit. 46 (2013) 3412–3424. https://doi.org/10.1016/j.patcog.2013.04.018.
[61] J. Zhao, H.-L. Wu, J.-F. Niu, Y.-J. Yu, L.-L. Yu, C. Kang, Q. Li, X.-H. Zhang, R.-Q. Yu, Chemometric resolution of coeluting peaks of eleven antihypertensives from multiple classes in high performance liquid chromatography: A comprehensive research in human serum, health product and Chinese patent medicine samples, J. Chromatogr. B 902 (2012) 96–107. https://doi.org/10.1016/j.jchromb.2012.06.032.
[62] R. Lletí, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes, Anal. Chim. Acta 515 (2004) 87–100. https://doi.org/10.1016/j.aca.2003.12.020.
[63] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 2009.
[64] A. Naghizadeh, D.N. Metaxas, Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means, Procedia Comput. Sci. 176 (2020) 205–214. https://doi.org/10.1016/j.procs.2020.08.022.
[65] G.S. Ochoa, P.E. Sudol, T.J. Trinklein, R.E. Synovec, Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry, Talanta 236 (2022) 122844. https://doi.org/10.1016/j.talanta.2021.122844.
[66] V.K. Rohatgi, A.K.M.E. Saleh, An Introduction to Probability and Statistics, John Wiley & Sons, 2015.
[67] P.E. Sudol, D.V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis, Talanta 206 (2020) 120239. https://doi.org/10.1016/j.talanta.2019.120239.

Chapter 6. Principal component analysis with comprehensive three-dimensional gas chromatography time-of-flight mass spectrometry data

Paige E. Sudol, Sonia Schöneich, Robert E. Synovec

6.1 INTRODUCTION

Comprehensive three-dimensional gas chromatography (GC^3) was first introduced as an extension of comprehensive two-dimensional gas chromatography (GC×GC) for exceedingly complex samples in 2000 [1]. Since then, impressive gains in the GC^3 instrumentation realm have been made, including coupling GC^3 to time-of-flight mass spectrometry (TOFMS) detection [2]. The primary challenge in GC^3 method development is that the modulator connecting the second- (^{2}D) and third-dimensions (^{3}D) (^{2}M) must be able to operate at exceptionally fast P_M in order to comprehensively sample the narrow ^{2}D peaks produced by the first-dimension (^{1}D) to ^{2}D modulation process (^{1}M) [3–5]. The first GC^3-TOFMS instrument utilized a high-speed diaphragm valve as ^{1}M and a commercial quad-jet thermal modulator as ^{2}M [2,6]. Using P_M as fast as 200 ms, these studies only achieved modest 3D peak capacities ($n_{c,3D}$) of up to 9600 (~190 peaks/min) [2], which is relatively comparable to peak capacities achieved with GC×GC [7]. More recent GC^3-TOFMS studies achieved much greater $n_{c,3D}$ approaching 30,000 via (1) switching the order of the thermal modulator and flow modulator to be ^{1}M and ^{2}M, in that order, and (2) utilizing dynamic pressure gradient modulation (DPGM) with a pulse valve as ^{2}M instead of a diaphragm valve [8,9]. The use of flow modulation as ^{2}M enables the use of higher flow rates and thus produces narrower ^{3}D peaks [10]. Additionally, whereas the diaphragm valve only transfers about ~10-15% of the column effluent to the subsequent column, resulting in a loss of sensitivity, DPGM is a total-transfer (i.e., 100% duty cycle) method that provides a considerable *S/N* enhancement [11–13]. Although the peak capacity enhancement

provided by GC^3 is notable, the true hallmark of GC^3 is the increased selectivity, which can be enhanced by using unique stationary phase combinations [14,15]. In recent years, several groups have harnessed the selectivity offered by GC^3 to elucidate diverse samples, such as allergens in perfume [16], wampee essential oil [17], hop oil [18], and sulfur-spiked jet fuels [8].

The considerable advancements in GC^3-TOFMS technology in recent years are part of larger trend amongst analytical chemists towards replacing one-dimensional techniques with multidimensional and hyphenated counterparts, with multidimensional NMR spectroscopy [19], 2D fluorescence spectroscopy [20], and liquid chromatography-gas chromatography (LC-GC/LC×GC) [21] representing a handful of additional examples. This trend toward multidimensional instrumentation has necessitated the development of more powerful data analysis tools known as chemometrics, as the resulting data arrays are too complex for manual interpretation. Chemometrics can be broadly defined as the application of linear algebra and statistical tools to extract useful chemical information from a dataset [22,23]. Using GC×GC-TOFMS as an example, consider a typical GC ^{1}D run time of 50 min and a ^{2}D separation time of 3 s. With a typical TOFMS collection frequency of 100 *Hz* and a mass range of 45-300 *amu*, a single chromatogram would produce a 3D array (1000 × 300 × 256 *m/z*) which, when unfolded, contains 7.68×10^7 data points. An analyst could spend years manually identifying analytes and/or comparing samples. Chemometric algorithms are thus indispensable tools in the analytical chemist's toolbox, both for *known* analyte identification (*targeted* studies) and *unknown* analyte discovery (*untargeted* studies) in complex, multidimensional datasets [24–30].

To date, the scope of chemometric analysis of GC^3 data is currently limited to targeted deconvolution tools, with several groups having performed variations of parallel factor analysis (PARAFAC) to resolve overlapped peaks in GC^3 chromatograms [16,31,32]. For example, in an

interesting adaptation of three-way PARAFAC for four-dimensional (4D) GC^3-TOFMS datasets, Watson et al. considered ^1D to be a preparative "fractional distillation" and performed PARAFAC on the individual $^2D \times {}^3D \times m/z$ chromatograms [6]. Although targeted deconvolution tools are invaluable, they require a significant amount of *a priori* knowledge on the part of the analyst, such as precise analyte retention times, which is often unavailable for truly unknown samples. Additionally, targeted deconvolution tools are only applied to a select region of the chromatogram, even though valuable information unbeknownst to the analyst may be found in other regions. Therein lies the utility of non-targeted, i.e., "discovery-based" chemometric analysis tools, which find as much chemical information as possible via classification or regression algorithms. Likely one of the main hindrances in applying non-targeted chemometric tools to GC^3-TOFMS datasets is limited computing power. Given our previous example, a ^3D separation time of 250 ms would increase the size of the resulting unfolded data, which was already quite large, by 25-fold. Yet the potential benefits of non-targeted chemometric analysis are too intriguing to ignore, as there is a strong possibility that, given the increased peak capacity and selectivity offered in GC^3 relative to GC×GC, chemometric analysis of GC^3 data would elucidate more chemical information. Indeed, the technical aspects of multi-dimensional instrumentation are interesting, but amenability to non-targeted chemometric analysis elevates multidimensional instrumentation beyond proof-of-principle studies to large-scale data collection campaigns [33–35].

Herein we present the first application of non-targeted chemometric analysis, specifically principal component analysis (PCA), to a complex GC^3-TOFMS dataset of four jet fuels. PCA is an unsupervised (i.e., no class label information available) data reduction method which aims to minimize the variance in a given dataset, with the resulting scores describing relationships

between samples and the loadings describing relationships between variables (i.e., peaks) [36,37]. Given its simplicity and widespread usage in commercial software platforms, PCA is often the initial step in a data processing workflow, and so it seems appropriate that the amenability of GC^3-TOFMS data to PCA should be tested before more complex techniques. Five replicates per jet fuel are collected on a GC^3-TOFMS instrument utilizing commercial thermal modulation and DPGM as ^1M and ^2M, respectively. Prior to PCA, the data is aligned and re-registered using a novel ^2D re-registration technique in which vacant ^3D modulations are subtracted from the data. The latter step removes the ^2D shifting observed in the ^1D x ^2D view and effectively "centers" the data. This issue will be described in more depth later in this manuscript, but the purpose of this is to make the data visually interpretable, which will in turn make the resulting PCA model interpretable. We explore the potential of the 3D PCA loadings as a visual tool for identifying the distinguishing features between the jet fuels. We also examine to what extent GC^3 provides a "chemometric advantage" to GC×GC by comparing the 3D chromatograms to the corresponding $^1D \times ^2D$ view. Finally, we highlight the utility of performing "pair-wise" PCA on chemically similar fuels, whose differences will be drowned out in a multi-class PCA model.

6.2 EXPERIMENTAL

A schematic of the GC^3-TOFMS instrumental platform is provided in Fig. 6.1. This instrument was based on a modified Pegasus 4D GC×GC-TOFMS (LECO Corporation, St. Joseph, MI, USA) platform comprised of an Agilent 6890 gas chromatograph, Agilent 7683B auto-injector (Agilent Technologies, Santa Clara, CA, USA), and a 4D thermal modulator upgrade (LECO Corporation, St. Joseph, MI, USA). The ^1D column (26 m × 250 µm inner diameter × 0.25 µm film thickness) contained a polar Rxi-17Sil MS stationary phase (Restek,

Bellefonte, PA); the ^{2}D column (4.5 m × 100 µm inner diameter × 0.10 film thickness) contained a non-polar Rxi-1 MS stationary phase; and the ^{3}D column (2.5 m × 180 µm inner diameter × 0.14 µm film thickness) contained a highly polar ionic liquid (IL-60) stationary phase. The ^{1}D and ^{2}D columns were interfaced via the thermal modulator, and the ^{2}D and ^{3}D columns were interfaced via a microvolume T-union (Model MT.5CXS6, Valco Instruments Company Inc., Houston, TX, USA) directly connected to the solenoid pulse valve (i.e., the total transfer flow modulator performing DPGM; model 009–0347–900, Parker Hannifin, Hollis, NH). We direct the reader to additional publications providing detailed descriptions of the in-house fabrication process for the pulse valve flow modulator [8,13].

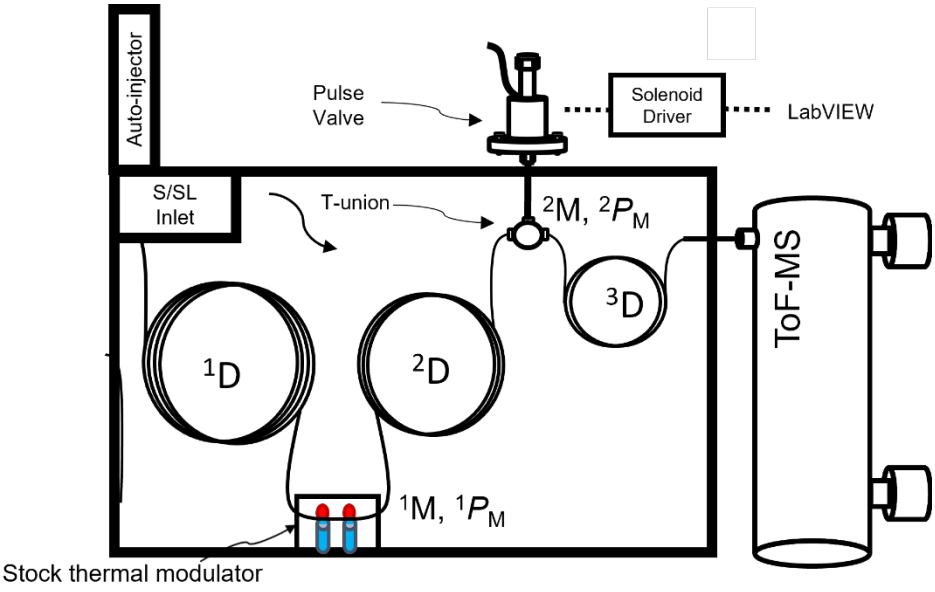

Figure 6.1. Schematic of GC^{3}-TOFMS instrumental platform.

Five replicate chromatograms per each of the four jet fuels (JP8, J1800, JP4, and JP7) were collected using the GC^3-TOFMS instrument. A 1 ul aliquot of each replicate was injected with a split ratio of 50:1 into the GC inlet, which was held constant at 250°C. Ultrahigh purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) was utilized as the carrier gas, with the inlet pressure program (P_{inlet}) set to constant flow mode at 0.5 ml/min on ^{1}D. The ^{1}D column was held at 40°C for 1 min, increased to 260°C at 5°C/min, and held at 260°C for 5 min. The modulator block utilized the same temperature program with a 20°C offset. Hot and cold pulse times of 0.40 s and 1.60 s, respectively, were utilized with a P_M = 4 s using the thermal modulator (^1P_M, from ^{1}D to ^{2}D). The following auxiliary pressure program (P_{aux}) provided full modulation from ^{2}D to ^{3}D using ^2P_M = 250 ms and pulse width (p_w) = 140 ms: held at 13.8 psig for 1 min, increased to 30.2 prig at 0.3727 psig/min, and held at 30.2 psig for 5 min. As discussed by Trinklein et al., this P_{aux} was chosen to provide ~ 6 ml/min flow on the ^{3}D column [8]. Pulse valve actuation was controlled by an in-house LabVIEW program operating concomitantly with the autoinjector and thermal modulator. The TOFMS detector collected *m/z* 45-300 at a collection frequency of 100 spectra/s with a detector voltage of 1657 V.

Data analysis was performed in MATLAB R2021a (The Mathworks Inc., Natick, MA, USA) after the data were imported from the LECO ChromaTOF software. Prior to data analysis, the chromatograms were baseline corrected on a per *m/z* basis using an in-house written rolling minimum function. Since this function was nominally written for GC×GC-TOFMS data, the 4D GC^3-TOFMS chromatograms were re-folded into 3D arrays ((^{3}D × ^{2}D) × ^{1}D × *m/z*) prior to baseline correction [38]. To correct for retention time offsets due to minor fluctuations in pulse valve actuation time, the unfolded (i.e., vectorized) chromatograms per *m/z* were aligned using the circular shift (circshift) function in MATLAB. More specifically, the replicates of a given

fuel were aligned *to each other* prior to a so-called "cycle of alignment", in which all five replicates of one fuel were aligned to the replicates of *another* fuel, which were aligned to the next fuel, and so on. Following baseline correction and alignment, the overall 3D peak capacity ($n_{c,3D}$) was calculated as follows,

$$n_{c,3D} = {}^{1}n_c \times {}^{2}n_c \times {}^{3}n_c = \frac{{}^{1}t \; {}^{1}P_M \; {}^{2}P_M}{{}^{1}W_b \; {}^{2}W_b \; {}^{3}W_b} \tag{6.1}$$

wherein ${}^{1}n_c$, ${}^{2}n_c$, and ${}^{3}n_c$ represent the peak capacities on the individual chromatographic dimensions, each equal to t/W_b; ${}^{1}t$ is the total run time; ${}^{1}P_M$ and ${}^{2}P_M$ are the modulation periods between ^{1}D and ^{2}D, and ^{2}D and ^{3}D, respectively, and ${}^{1}W_b$, ${}^{2}W_b$, and ${}^{3}W_b$ are the calculated widths-at base. Rather than apply Eq. (6.1) to the individual fuel chromatograms, the fuel chromatograms were summed to produce a representative sample, which we will refer to as the "overall fuel sample," for peak capacity calculations. Prior to and following data re-registration on ^{2}D and ^{3}D (more regarding this in Results and Discussion), the fuel chromatograms were visualized in 2D with the imagesc function and in 3D with the vol3d function in MATLAB [39]. Analyte identifications were performed using the mainlib and replib NIST mass spectral libraries. Principal component analysis (PCA) was performed using PLS Toolbox 8.9.2 (Eigenvector Research, Inc., Wenatchee, WA, USA), with mean centering of the data prior to model generation. The 3D PCA loadings were also prepared using vol3d [39].

6.3 RESULTS AND DISCUSSION

An evaluation of the experimental 3D peak capacity achieved by the GC^3-TOFMS platform used herein is provided in Fig 6.2. For the sake of simplicity, a representative replicate chromatogram of each fuel was summed together to produce what we will call the "overall fuel sample." The total ion current (TIC) chromatogram of the overall fuel sample will serve as a

platform throughout this manuscript. Here, the ^{1}M (^{1}D to ^{2}D) modulations of a typical analyte in terms of peak widths, 1,1,3-trimethylcyclopentane, are shown in Fig. 6.2(A), with a zoom-in of the most intense ^{1}M modulation ($^{1}t_{R}$ ~ 857 s) provided in Fig. 6.2(B) to highlight the ^{2}M (^{2}D to ^{3}D) modulation process. Given careful optimization of the P_{aux} and p_{w}, DPGM provides total transfer (i.e., 100% duty cycle) and full modulation, which is clearly observed in Fig. 6.2(B). The following peak width measurement were obtained using this analyte in Fig. 6.2(A-B): $^{1}W_{b}$ = 12 s, $^{2}W_{b}$ = 290 ms, and $^{3}W_{b}$ = 60 ms for this analyte. Given an effective ^{1}D run time of ~29.9 min (1796 s) and the $^{2}P_{M}$ and $^{3}P_{M}$ utilized, this equates to $^{1}n_{c}$, $^{2}n_{c}$, and $^{3}n_{c}$ of 150, 14, and 4, respectively. Thus this instrumental configuration resulted in an overall ideal 3D peak capacity ($n_{c,3D}$) of 8,400, or ~281 peaks/min. Previous work by Trinklein et al. using DPGM as ^{2}M in GC^{3} achieved average $n_{c,3D}$ values of 8,000 [40], 11,470 [9], and ~28,000 [8], making the $n_{c,3D}$ achieved herein comparable, albeit at the lower end. However, we shall see that the *selectivity* provided by the third dimension plays the most important role in chemometric distinguishment, rather than the *peak capacity*. The calculation of peak widths and peak capacity values merely serves to validate that the data quality is suitable for chemometric analysis, specifically PCA.

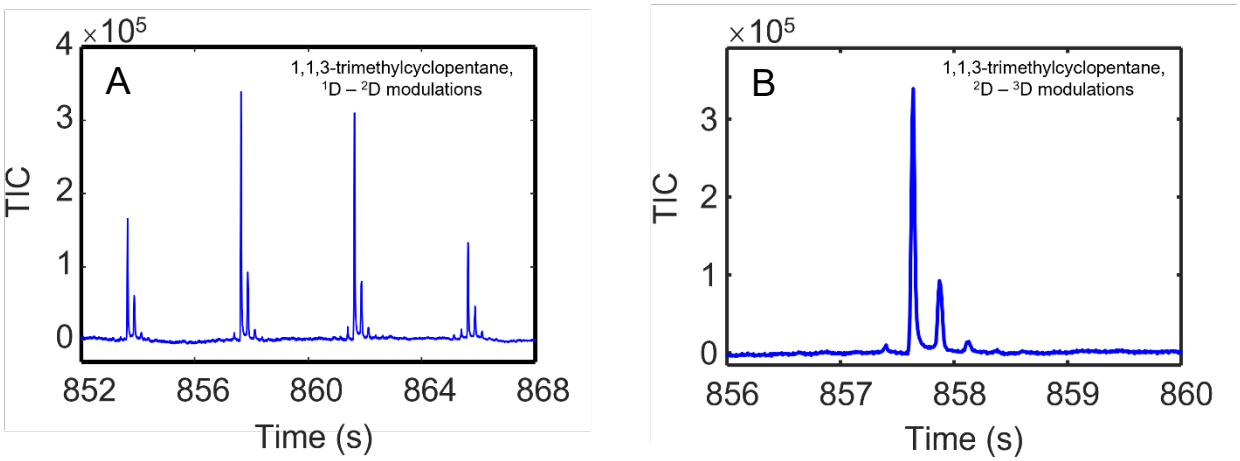

Figure 6.2. Illustration of 3D peak capacity ($n_{c,3D}$) achieved using a typical analyte in the "overall fuel sample", 1,1,3-trimethylcyclopentane. (A) ^{1}M (^{1}D to ^{2}D) modulations. (B) ^{2}M (^{2}D to ^{3}D) modulations.

An overview of the ^{2}D re-registration procedure employed herein is provided in Fig. 6.3. The ^{1}D × ^{2}D TIC chromatogram (i.e., produced by summing ^{3}D) of the overall fuel sample is shown in Fig. 6.3(A). For reference, the ^{1}D ×^{2}D chromatograms of the four individual fuels are provided in Appendix E as Fig. E.1(A-D). To ensure the accuracy of the ^{2}D re-registration procedure, 40 native compounds comprising the major chemical compound classes in jet fuel (straight chain and branched alkyl compounds; cycloalkyl compounds; and aromatic compounds) were identified in the overall fuel sample and are labeled accordingly in Fig. 6.3(A). These compounds are listed in Tables 6.1-6.3, with the straight chain and branched alkyl compounds in Table 6.1, the cycloalkyl compounds in Table 6.2, and the aromatic compounds in Table 6.3. Note that the final ^{1}D and ^{2}D retention times ($^{1}t_R$ and $^{2}t_R$) following ^{2}D re-registration are listed in Table 6.1-6.3 and the compounds are listed in order of $^{1}t_R$.

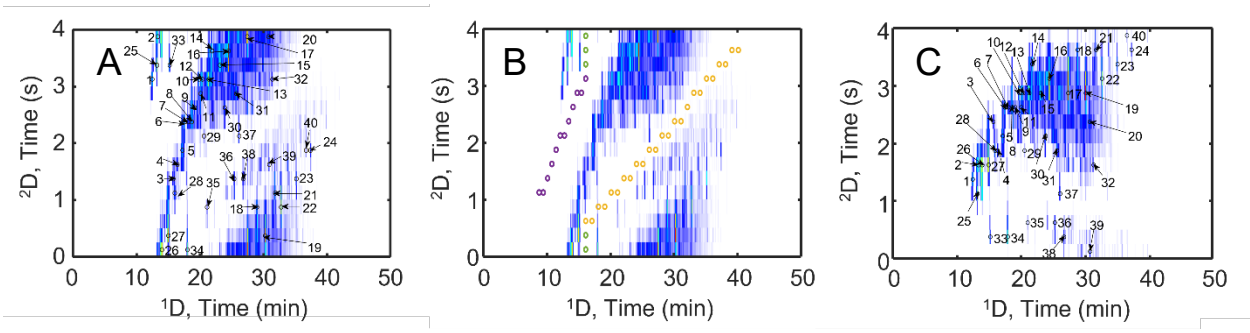

Figure 6.3. Overview of novel ^{2}D re-registration technique. (A) ^{1}D × ^{2}D chromatogram of overall fuel sample, with 40 identified analytes consisting of alkyl, cycloalkyl, and aromatic components (Table 6.1-6.3) labeled accordingly. (B) ^{1}D × ^{2}D chromatogram with ^{3}D modulations subtracted from the GC^3-TOFMS chromatogram indicated with colored circles (purple, gold, and green). The different colored circled regions have different slopes (see Results and Discussion). (C) ^{1}D × ^{2}D chromatogram after ^{2}D re-registration, with new locations of 40 compounds indicated accordingly.

Table 6.1. Identities and retention times of straight chain/branched alkyl compounds identified in overall fuel sample.

Number	Identity	$^{1}t_{R}$ (min)	$^{2}t_{R}$ (s)	$^{3}t_{R}$ (ms)
1	n-hexane	12.47	1.25	130
2	3-methylhexane	13.33	1.50	140
3	2,4-dimethylhexane	15.67	2.75	130
4	2,5-dimethylheptane	16.33	2.25	130
5	4-methyloctane	17.13	2.50	120
6	2,2-dimethyloctane	17.47	3.00	120
7	4,6-dimethylundecane	18.00	3.00	120
8	3,5-dimethyloctane	18.60	2.75	120
9	tetrahydrocitronellene	19.07	2.75	120
10	2,2,6-trimethyloctane	19.73	3.00	110
11	3-ethyl-2-methylheptane	20.00	2.75	110
12	2,2,7,7-Tetramethyloctane	20.33	3.00	110
13	decane	21.27	3.00	100
14	2,2,4,6,6-pentamethylheptane	21.73	3.50	100
15	2-methyldecane	23.07	3.00	100
16	undecane	24.40	3.25	90
17	dodecane	27.33	3.00	70
18	2,3,7-trimethyloctane	28.80	3.75	70
19	tridecane	30.07	3.00	60
20	(E)-5-tridecene	30.67	2.50	70
21	2,6,10-trimethyldodecane	31.67	3.75	50
22	*substituted decane	32.60	3.25	50
23	*substituted decane	35.00	3.50	40
24	hexadecane	37.20	3.75	30

Table 6.2. Identities and retention times of cycloalkyl compounds identified in overall fuel sample.

Number	Identity	$^{1}t_{R}$ (min)	$^{2}t_{R}$ (s)	$^{3}t_{R}$ (ms)
25	methyl cyclopentane	13.20	1.00	130
26	cis-1,3-dimethylcyclopentane	13.93	1.50	140
27	methyl cyclohexane	14.93	1.75	140
28	(cis/trans)-1,2-dimethylcyclohexane	15.93	2.25	130
29	propylcyclohexane	20.53	2.00	110
30	(1-methylpropyl)cyclohexane	23.80	2.25	100
31	(2-methylbutyl)cyclohexane	25.60	2.00	90
32	(E)-cyclododecene	31.27	1.75	60

Table 6.3. Identities and retention times of aromatic compounds identified in overall fuel sample.

Number	Identity	$^{1}t_{R}$ (min)	$^{2}t_{R}$ (s)	$^{3}t_{R}$ (ms)
33	benzene	15.20	0.75	190
34	toluene	17.93	0.75	190
35	p-xylene	21.00	0.75	170
36	1,2,4-trimethylbenzene	25.27	0.75	170
37	trans-decahydronaphthalene	26.07	1.25	80
38	1-methyl-3-propylbenzene	26.67	0.50	150
39	1,2,4,5-tetramethylbenzene	30.73	0.25	150
40	2-methylnaphthalene	36.47	4.00	80

The ^{2}D re-registration template is given in Fig. 6.3(B). The colored circles (purple, green, and gold) represent the individual ^{3}D modulations which were subtracted. Traditionally, data re-registration of GC×GC-TOFMS chromatograms simply "resets" the 0-time mark on the ^{2}D time axis to correct wrap-around, which occurs when the ideal P_M is not chosen. As discussed in the Introduction, herein we present a novel ^{2}D r-registration technique for GC^{3}-TOFMS data to correct ^{2}D shifting by subtracting a linear series of vacant ^{3}D modulations to remove the ^{1}D × ^{2}D slant and "center" the data. This will ideally produce a ^{1}D × ^{2}D chromatogram with the major compound classes in the correct location given the column set. Following manual determination of best fit lines through the vacant chromatographic regions, the purple and gold modulations were selected in Fig. 6.3(B) as fitting most closely along these lines. For the purple line, one ^{3}D modulation is subtracted every 10 ^{2}D modulations, and for the gold line, one ^{3}D modulation is subtracted every 15 ^{2}D modulations. Given the length of these two vacant regions, use of these slopes results in 12 purple modulations (3 s) and 25 gold modulations (6.25 s) being subtracted, which would effectively re-register the data by 9.25 s over the course of ~31 minutes (from left to right, first purple modulation to last gold modulation). Note that the purple line ends and the gold line starts at the same modulation ($^{1}t_R$ = 16.13 min); if they overlapped, peaks in overlapping regions would be re-registered multiple times and become inconsistent with the rest of the chromatogram. Lastly, note the green modulations also all occur at $^{1}t_R$ = 16.13 min but at different $^{2}t_R$. Although the purple and gold modulations are sufficient to remove the slant and "restructure" the ^{1}D × ^{2}D chromatogram (not shown for brevity), a final vertical re-registration (green modulations) is needed to ensure that the compound classes are in the correct order given the ^{1}D × ^{2}D column set. More specifically, since the ^{2}D column in more non-polar relative to ^{1}D, the saturated alkyl compounds should appear at the top of the chromatogram ($^{2}t_R$ ~ 3 to 4 s) and

the aromatics should appear at the bottom (^2t_R ~ 0 to 1 s). The final product of application of the re-registration template in Fig. 6.3(B) is provided in Fig. 6.3(C), with the new locations of the 40 compounds listed in Table 6.1-6.3 labeled. Except for compound #40 (2-methylnaphthalene, Table 6.3), which wraps around into the alkane band, the ^2t_R locations of the alkyl, cycloalkyl, and aromatic constituents are now consistent with a reverse GC×GC column configuration [41,42]. Most notably, the ^{2}D shifting has been corrected in a consistent manner *without* changing the inherent chemical information present in the GC^3-TOFMS data.

Although our novel ^{2}D re-registration method effectively corrects the ^{2}D shifting, it does not correct wrap-around on ^{3}D. To illustrate this phenomenon, the ^{1}D × ^{3}D chromatogram (^{2}D summed) of the overall fuel sample *after* ^{2}D re-registration is provided in Fig. 6.4(A). Labels for the 40 compounds have been omitted for brevity, as the wrap-around is clear with several peaks between 10 and 20 minutes being split across the ^{3}D time axis. An alternative view of the ^{3}D wrap-around is provided in Appendix E as Fig. E.2(A-D) via 3D chromatograms of the four jet fuels following ^{2}D re-registration. This wrap-around can be corrected with a traditional uniform re-registration of ^{3}D at 120 ms, at which no peaks elute to avoid peak splitting. The result is provided in Fig. 6.4(B), with the 40 compounds in Table 6.1-6.3 labeled accordingly. Notably, this re-registration produces a ^{1}D x ^{3}D chromatogram consistent with the "normal" column configuration utilized, in which the ^{1}D column (Rxi-17 MS) is more non-polar relative to the highly polar ionic liquid ^{3}D column. More specifically, the aromatics (Table 6.3) are more retained on ^{3}D and hence at the top of the chromatogram relative to the cycloalkyl (Table 6.2) and alkyl (Table 6.1) compounds, in that order.

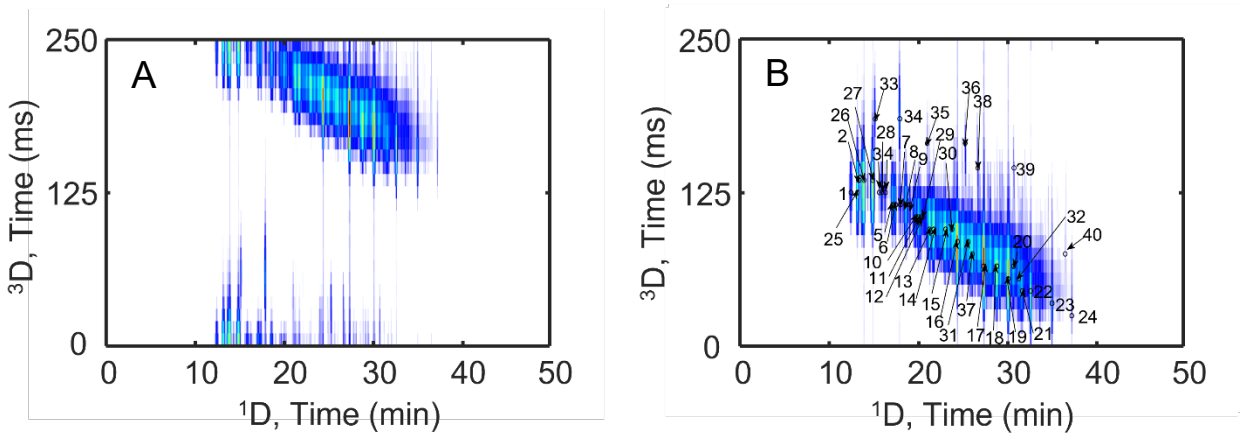

Figure 6.4. Illustration of ^{3}D re-registration. (A) ^{1}D × ^{3}D chromatogram before re-registration. (B) ^{1}D × ^{3}D chromatogram after re-registration by 120 ms on ^{3}D in (A), with locations of 40 compounds listed in Table 6.1-6.3 labeled accordingly.

The final 3D chromatogram of the overall fuel sample, following ^{2}D and ^{3}D re-registration, is provided in Fig. 6.5(A). Note that for all 3D chromatograms and loadings presented in the remainder of this chapter, the ^{1}D axis has been truncated to the 10 to 40 min range, as no significant chemical information is present in the fuels before and after these time points. For reference, the corresponding locations of the 40 compounds in the 3D plane are shown as colored dots in Fig. 6.5(B), with the compounds color-coded by compound class (red = straight-chain and branched alkyls; green = cycloalkyls; blue = aromatics). The GC^3 column set

employed herein achieved an exceptional compound-based separation, particularly along ^{2}D, as evidenced by Fig. 6.5(B). In comparing Fig. 6.5(B) to the structure of Fig. 6.5(A), an additional band of compounds can be observed in Fig. 6.5(A), with $^{1}t_{R}$ from 20 to 40 min, $^{2}t_{R}$ at ~ 4 s and $^{3}t_{R}$ ranging from ~80 to 180 ms. Identification via the NIST libraries reveals that these compounds are substituted benzenes. Notably, these compounds were difficult to isolate from the more highly concentrated alkane band in the ^{1}D x ^{2}D chromatogram in Fig. 6.3(C), hence highlighting the advantage of GC^{3}-TOFMS relative to GC×GC-TOFMS herein.

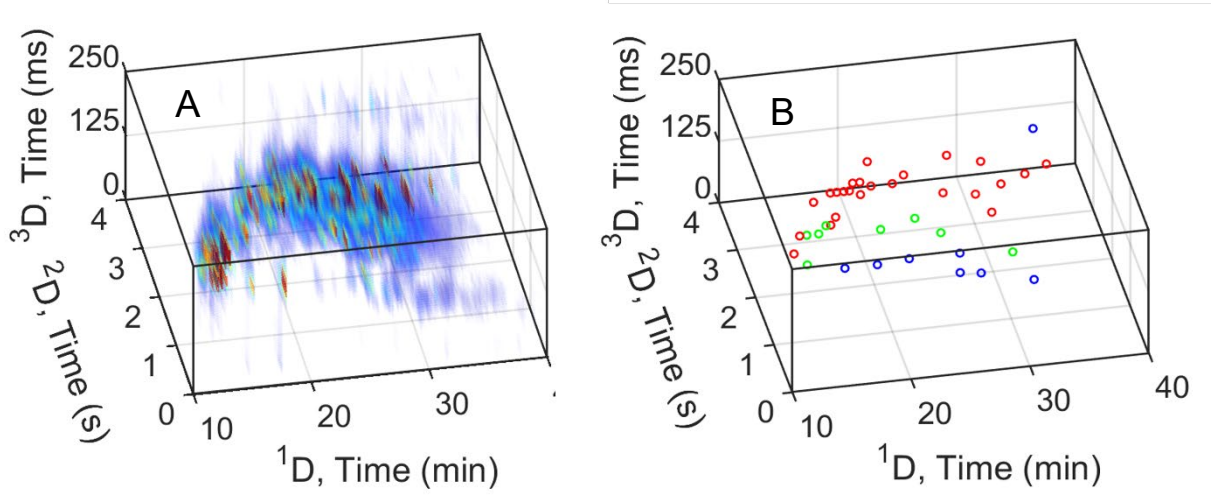

Figure 6.5. Results of ^{2}D and ^{3}D re-registration procedures. (A) 3D TIC chromatogram of overall fuel sample. (B) Locations of 40 identified compounds in 3D space, color-coded by compound class: straight chain/branched alkyl compounds in red (Table 6.1); cycloalkyl compounds in green (Table 6.2); aromatic compounds in blue (Table 6.3).

The 3D chromatograms of the four jet fuels following ^{2}D and ^{3}D re-registration are provided in Fig. 6.6(A-D). A thorough comparison of Fig. 6.6(A-D) to the retention time locations of the 40 identified analytes (Table 6.1-6.3), as well as the general compound class locations highlighted in Fig. 6.5(B), enables us to thoroughly characterize these fuels' chemical composition. JP8 appears to primarily contain mid-boiling point alkanes, such as 2,5-dimethylheptane, 3,5-dimethyloctane, and 4,6-dimethylundecane (Fig. 6.6(A)). There appear to be some compounds in the cycloalkane region, albeit none of the eight identified herein (Table 6.2), but notably there is no discernable aromatic band in JP8. Conversely, J1800A has a very highly concentrated aromatic band, including trans-decahydronaphthalene and 1,2,4,5-tetramethylbenzene, and contains several distinct high-boiling point alkanes such as hexadecane and other substituted decanes (Fig. 6.6(B)). Additionally, J1800A contains the additional band of substituted benzenes observed in the overall fuel chromatogram (Fig. 6.5(A)), and it is important to once again reiterate that these compounds are difficult to distinguish in the ^{1}D × ^{2}D view of J1800A alone, without the additional ^{3}D dimension. Relative to the other fuels, JP4 is characterized by several low-boiling point alkanes and cycloalkanes including hexane, 3-methylhexane, and methylcyclohexane, and appears to be the only fuel with highly concentrated benzene and toluene peaks (Fig. 6.6(C)). Like J1800A, JP4 also contains several high-boiling point alkanes and appears to have trace amounts of the substituted benzenes. Unlike the other fuels, JP7 has no compounds eluting between 10 and 20 min and is thus a relatively "heavy" fuel. It contains several high-boiling point alkanes, such as 2,6,10-trimethyldodecane, tridecane, and other substituted decanes, along with numerous cycloalkanes including 2-methylbutylcyclohexane, 1-methylpropylcyclohexane, and (E)-cyclododecene (Fig. 6.6(D)). Like JP8, JP7 also has no noticeable aromatic band.

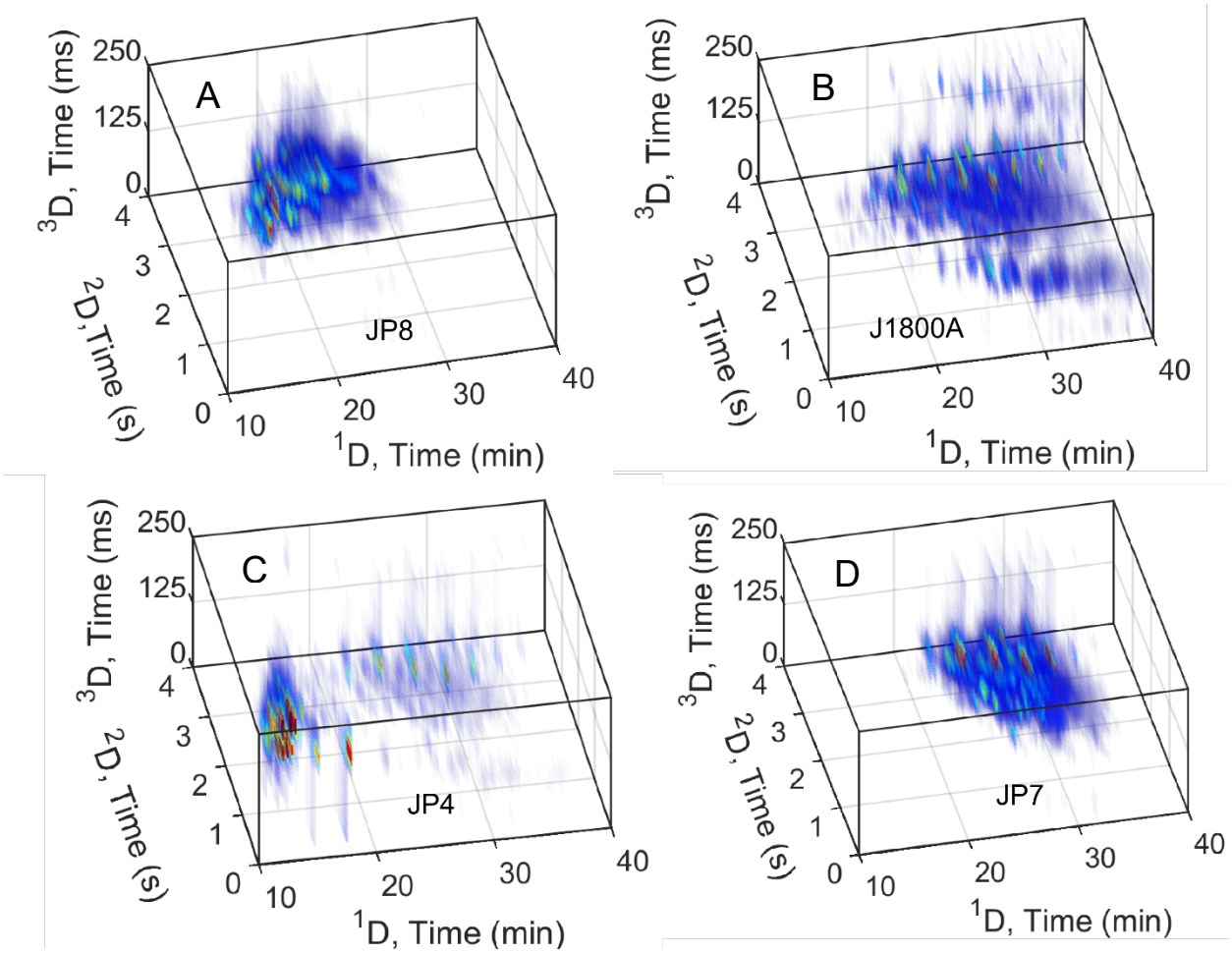

Figure 6.6. Final 3D TIC chromatograms of jet fuels. (A) JP8. (B) J1800A. (C) JP4. (D) JP7.

The results following PCA of the re-registered fuel samples are provided in Fig. 6.7. The scores plot is given in Fig. 6.7(A), with the fuels color-coded as follows: red= JP8; green = J1800A; blue = JP4; cyan = JP7. DCS values for the nearest neighbor pairs of fuels are indicated in Table 6.4 for reference. The replicates for a given fuel are well clustered in Fig. 6.7(A), which highlights the effectiveness of the alignment procedure described in the Experimental section. The DCS calculations allow us to conclude that JP8 and JP4 are the most "chemically different", which makes sense given that JP8 does not contain the low- and high-boiling point alkanes or aromatic band characteristic of JP4, and JP4 is missing the large cluster of mid-boiling point alkanes in JP8. J1800A and JP7 are the most "chemically similar", which is again consistent with both fuels being relatively heavy (i.e., most compounds eluting after 20 min on ^{1}D) and having highly similar alkyl/cycloalkyl compositions.

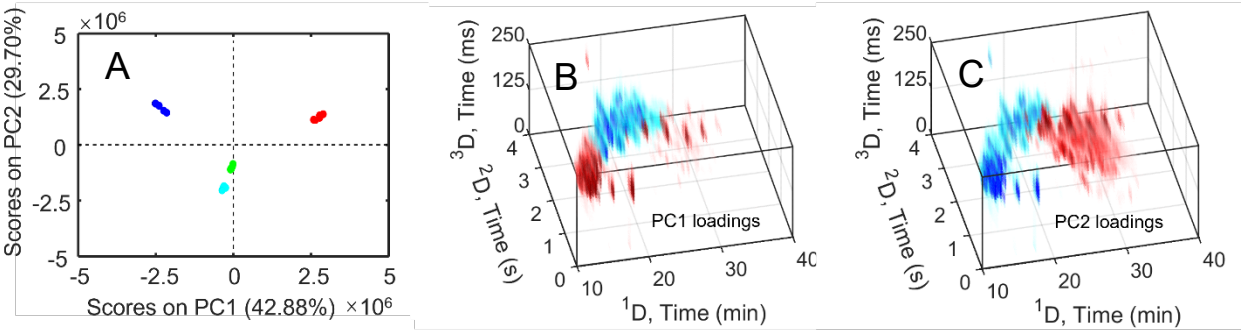

Figure 6.7. Results of multi-fuel PCA. (A) Scores plot, with jet fuels color-coded as follows: red=JP8; green=J1800A; blue=JP4; cyan=JP7. (B) 3D TIC PC1 loadings, with blue representing positively loaded analytes and red representing negatively loaded analytes. (C) 3D TIC PC2 loadings.

Table 6.4. DCS calculations for nearest neighbor fuel pairs in the scores plot in Fig. 6.7(A).

Fuel pair	DCS
JP8 & JP4	45.6
JP8 & J1800A	44.2
JP4 & JP7	41.6
J1800A & JP4	34.5
J1800A & JP7	15.0

The precise chemical differences/similarities between the fuels can be more comprehensively assessed using the 3D PC1 and PC2 loadings provided in Fig. 6.7(B-C). The individual peaks are color-coded by their loadings values in Fig. 6.7, with blue representing positively loaded analytes and red representing negatively loaded analytes. Positively loaded analytes are more concentrated in samples with positive scores on a given PC axis, and vice versa for negatively loaded analytes. In terms of the PCA model generated herein for the four fuels, given that the JP8 and JP4 samples have highly positive and negative scores in Fig. 6.7(A), respectively, with J1800A and JP7 approximately at zero, PC1 describes the chemical differences between JP8 and JP4. The individual chromatographic features in Fig. 6.6 are illuminated in the PC1 loadings, with the mid-boiling point alkane band in JP8 appearing as positive loadings, and the low-boiling point alkyl/cycloalkyl compounds and high-boiling point alkanes characteristic of JP4 appearing as negative loadings (Fig. 6.7(B)). Looking at the PC2 axis, JP8 and JP4 have highly positive scores, whereas J1800A and JP7 have highly negative

scores (Fig. 6.7(A)). Thus, PC2 must describe the differences *between* these two sets of fuels. Indeed, we see the low-boiling point compounds in JP4 appear as positive loadings, along with a portion of the mid-boiling point alkanes in JP8 (Fig. 6.7(C)). It is intriguing to note that the JP7 chromatogram almost exactly resembles the red, negatively loaded analytes in Fig. 6.7(C), whereas no distinguishing features of J1800A are observed in the PC2 loadings. Therefore, the 3D PCA loadings are a highly useful visual tool for distinguishing precise chemical differences between the fuels which would not have been as easily identified using PCA of GC×GC data.

Although the multi-fuel PCA model provided in Fig. 6.7 is highly useful for illuminating the major differences between the fuels which were clearly different, as observed by large DCS values in the scores plot, it is insufficient to fully distinguish the highly similar J1800A and JP7 fuels. This is because these minor chemical differences were drowned out by the much greater differences between JP8, JP4, and J1800A and JP7 *together,* namely the presence/absence of low- to mid-boiling point alkyl and cycloalkyl compounds. Thus, PCA was re-run using just these two fuels, the results of which are presented in Fig. 6.8. Interestingly, the DCS between J1800A and JP7 in Fig. 6.8(A) has decreased relative to Fig. 6.7(A) since minor replicate-to-replicate differences between the fuels have been amplified by removing the other chemical information. However, the PC1 loadings in Fig. 6.8(B) are highly informative, with the cycloalkane region in JP7 (positive PC1 scores in Fig. 6.8(A)) appearing as positively loaded analytes, and the distinctive aromatic band in J1800A (negative PC1 scores in Fig. 6.8(A)) appearing as negatively loaded analytes. The alternating red and blue high-boiling point alkane peaks in Fig. 6.8(B) reveal minor differences in the alkanes present in J1800A versus JP7, which was not clear via manual comparison of the chromatograms in Fig. 6.6. Equally notable is the substituted benzenes in J1800A appearing as negatively loaded analytes on PC1 (Fig. 6.8(B)), as

PCA of GC×GC data would *not* have revealed these compounds in the loadings. Thus, a chemometric advantage of GC^3 relative to GC×GC is revealed in this individual fuel pair PCA model. Given that the PC2 axis in Fig. 6.8(A) describes mainly replicate-to-replicate differences within the fuels and accounts for minimal variance in the PCA model (11.74% on PC2 versus 64.55% on PC1 in Fig. 6.8(A)), it is not surprising that the PC2 loadings in Fig. 6.8(C) are not particularly useful for distinguishing J1800A and JP7. However, given the abundance of chemical information revealed by the PC1 axis, this example highlights the potential of "pairwise" PCA for multi-class GC^3-TOFMS datasets.

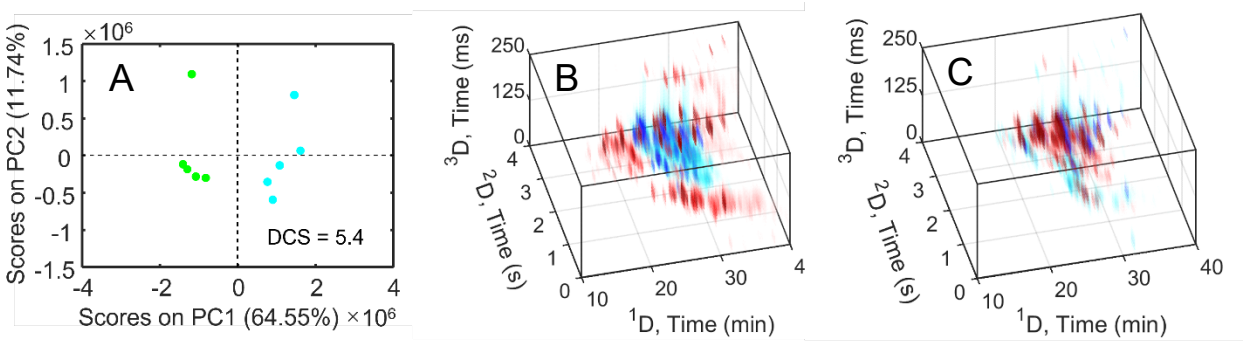

Figure 6.8. Results of pair-wise PCA using J1800A and JP7. (A) Scores plot, with jet fuels color-coded as follows: green=J1800A; cyan=JP7. (B) 3D TIC PC1 loadings, with blue representing positively loaded analytes and red representing negatively loaded analytes. (C) 3D TIC PC2 loadings.

6.4 CONCLUSIONS

Herein we have demonstrated the analytical utility of performing non-targeted chemometric analysis on GC^3-TOFMS data. The novel ^2D re-registration technique applied prior to PCA effectively centered the data without changing the inherent chemical information, thus making the data interpretable for PCA. An impressive compound class-based separation was achieved in the $^1D \times ^2D$ orientation. The 3D TIC chromatograms revealed a band of substituted benzenes in J1800A jet fuel, which were not as easily distinguishable in the $^1D \times ^2D$ view, thus highlighting the advantage of GC^3 relative to GC×GC. Although these benzenes were in fact wrapped around on ^2D, the ^2D re-registration template was a suitable compromise for ensuring the remaining compound classes were in the correct locations given the reverse column configuration. The 3D PCA loadings for the multi-fuel PCA model were highly useful for elucidating the chemical differences between the jet fuels. An individual PCA model using the two most similar jet fuels, J1800A and JP7, revealed finer chemical differences which were drowned out in the multi-fuel PCA model. It appears that highly complex, chemically similar samples such as the J1800A and JP7 jet fuels would benefit from GC^3-TOFMS followed by PCA, as opposed to GC×GC and subsequent PCA. Such samples may also benefit from supervised tools such as F-ratio analysis following GC^3-TOFMS, which necessitates additional studies.

6.5 REFERENCES

[1] E.B. Ledford, C.A. Billesbach, Q. Zhu, GC3: Comprehensive Three-Dimensional Gas Chromatography, J. High Resolut. Chromatogr. 23 (2000) 205–207. https://doi.org/10.1002/(SICI)1521-4168(20000301)23:3<205::AID-JHRC205>3.0.CO;2-U.

[2] N.E. Watson, H.D. Bahaghighat, K. Cui, R.E. Synovec, Comprehensive Three-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry, Anal. Chem. 89 (2017) 1793–1800. https://doi.org/10.1021/acs.analchem.6b04112.

[3] L.M. Blumberg, Accumulating resampling (modulation) in comprehensive two-dimensional capillary GC (GC×GC), J. Sep. Sci. 31 (2008) 3358–3365. https://doi.org/10.1002/jssc.200800424.

[4] W. Khummueng, J. Harynuk, P.J. Marriott, Modulation Ratio in Comprehensive Two-dimensional Gas Chromatography, Anal. Chem. 78 (2006) 4578–4587. https://doi.org/10.1021/ac052270b.

[5] W.C. Siegler, B.D. Fitz, J.C. Hoggard, R.E. Synovec, Experimental Study of the Quantitative Precision for Valve-Based Comprehensive Two-Dimensional Gas Chromatography, Anal. Chem. 83 (2011) 5190–5196. https://doi.org/10.1021/ac200302b.

[6] N.E. Watson, S.E. Prebihalo, R.E. Synovec, Targeted analyte deconvolution and identification by four-way parallel factor analysis using three-dimensional gas chromatography with mass spectrometry data, Anal. Chim. Acta 983 (2017) 67–75. https://doi.org/10.1016/j.aca.2017.06.017.

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

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

[9] T.J. Trinklein, C.G. Warren, R.E. Synovec, Determination of the Signal-To-Noise Ratio Enhancement in Comprehensive Three-Dimensional Gas Chromatography, Anal. Chem. 93 (2021) 8526–8535. https://doi.org/10.1021/acs.analchem.1c01190.

[10] D.V. Gough, D.H. Song, S. Schöneich, S.E. Prebihalo, R.E. Synovec, Development of Ultrafast Separations Using Negative Pulse Partial Modulation To Enable New Directions in Gas Chromatography, Anal. Chem. 91 (2019) 7328–7335. https://doi.org/10.1021/acs.analchem.9b01085.

[11] T.J. Trinklein, D.V. Gough, C.G. Warren, G.S. Ochoa, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography, J. Chromatogr. A (2019) 460488. https://doi.org/10.1016/j.chroma.2019.460488.

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

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

[14] W.C. Siegler, J.A. Crank, D.W. Armstrong, R.E. Synovec, Increasing selectivity in comprehensive three-dimensional gas chromatography via an ionic liquid stationary phase

column in one dimension, J. Chromatogr. A 1217 (2010) 3144–3149. https://doi.org/10.1016/j.chroma.2010.02.082.
[15] D.V. Gough, H.D. Bahaghighat, R.E. Synovec, Column selection approach to achieve a high peak capacity in comprehensive three-dimensional gas chromatography, Talanta 195 (2019) 822–829. https://doi.org/10.1016/j.talanta.2018.12.007.
[16] V.H.C. Ferreira, L.W. Hantao, R.J. Poppi, Consumable-free Comprehensive Three-Dimensional Gas Chromatography and PARAFAC for Determination of Allergens in Perfumes, Chromatographia 83 (2020) 581–592. https://doi.org/10.1007/s10337-020-03863-6.
[17] D. Sciarrone, S. Pantò, A. Rotondo, L. Tedone, P.Q. Tranchida, P. Dugo, L. Mondello, Rapid collection and identification of a novel component from Clausena lansium Skeels leaves by means of three-dimensional preparative gas chromatography and nuclear magnetic resonance/infrared/mass spectrometric analysis, Anal. Chim. Acta 785 (2013) 119–125. https://doi.org/10.1016/j.aca.2013.04.069.
[18] D. Yan, Y.F. Wong, S.P. Whittock, A. Koutoulis, R.A. Shellie, P.J. Marriott, Sequential Hybrid Three-Dimensional Gas Chromatography with Accurate Mass Spectrometry: A Novel Tool for High-Resolution Characterization of Multicomponent Samples, Anal. Chem. 90 (2018) 5264–5271. https://doi.org/10.1021/acs.analchem.8b00142.
[19] Q. Shi, J. Yan, B. Jiang, X. Chi, J. Wang, X. Liang, X. Ai, A general strategy for the structural determination of carbohydrates by multi-dimensional NMR spectroscopies, Carbohydr. Polym. 267 (2021) 118218. https://doi.org/10.1016/j.carbpol.2021.118218.
[20] L. Ranzan, C. Ranzan, L.F. Trierweiler, J.O. Trierweiler, Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy, Energy Fuels 31 (2017) 8942–8950. https://doi.org/10.1021/acs.energyfuels.7b00954.
[21] D.M. Rasheed, A. Serag, Z.T. Abdel Shakour, M. Farag, Novel trends and applications of multidimensional chromatography in the analysis of food, cosmetics and medicine bearing essential oils, Talanta 223 (2021) 121710. https://doi.org/10.1016/j.talanta.2020.121710.
[22] K.J. Siebert, Using Chemometrics To Classify Samples and Detect Misrepresentation, in: Prog. Authentication Food Wine, American Chemical Society, 2011: pp. 39–65. https://doi.org/10.1021/bk-2011-1081.ch004.
[23] R.G. Brereton, Pattern recognition in chemometrics, Chemom. Intell. Lab. Syst. 149 (2015) 90–96. https://doi.org/10.1016/j.chemolab.2015.06.012.
[24] R. Ríos-Reina, S. Elcoroaristizabal, J.A. Ocaña-González, D.L. García-González, J.M. Amigo, R.M. Callejón, Characterization and authentication of Spanish PDO wine vinegars using multidimensional fluorescence and chemometrics, Food Chem. 230 (2017) 108–116. https://doi.org/10.1016/j.foodchem.2017.02.118.
[25] V. Sharma, R. Kumar, Trends of chemometrics in bloodstain investigations, TrAC Trends Anal. Chem. 107 (2018) 181–195. https://doi.org/10.1016/j.trac.2018.08.006.
[26] R. Ríos-Reina, J.M. Camiña, R.M. Callejón, S.M. Azcarate, Spectralprint techniques for wine and vinegar characterization, authentication and quality control: Advances and projections, TrAC Trends Anal. Chem. 134 (2021) 116121. https://doi.org/10.1016/j.trac.2020.116121.
[27] Md.M. Rahman, M.V. Bui, M. Shibata, N. Nakazawa, Mst.N.A. Rithu, H. Yamashita, K. Sadayasu, K. Tsuchiyama, S. Nakauchi, T. Hagiwara, K. Osako, E. Okazaki, Rapid noninvasive monitoring of freshness variation in frozen shrimp using multidimensional fluorescence imaging coupled with chemometrics, Talanta 224 (2021) 121871. https://doi.org/10.1016/j.talanta.2020.121871.

[28] M.B. Anzardi, J.A. Arancibia, A.C. Olivieri, Processing multi-way chromatographic data for analytical calibration, classification and discrimination: A successful marriage between separation science and chemometrics, TrAC Trends Anal. Chem. 134 (2021) 116128. https://doi.org/10.1016/j.trac.2020.116128.

[29] M. Pérez-Cova, J. Jaumot, R. Tauler, Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches, TrAC Trends Anal. Chem. 137 (2021) 116207. https://doi.org/10.1016/j.trac.2021.116207.

[30] P.-H. Stefanuto, A. Smolinska, J.-F. Focant, Advanced chemometric and data handling tools for GC×GC-TOF-MS: Application of chemometrics and related advanced data handling in chemical separations, TrAC Trends Anal. Chem. 139 (2021) 116251. https://doi.org/10.1016/j.trac.2021.116251.

[31] N.E. Watson, W.C. Siegler, J.C. Hoggard, R.E. Synovec, Comprehensive Three-Dimensional Gas Chromatography with Parallel Factor Analysis, Anal. Chem. 79 (2007) 8270–8280. https://doi.org/10.1021/ac070829x.

[32] N. Abdulhussain, S. Nawada, P. Schoenmakers, Latest Trends on the Future of Three-Dimensional Separations in Chromatography, Chem. Rev. 121 (2021) 12016–12034. https://doi.org/10.1021/acs.chemrev.0c01244.

[33] R. Kumar, V. Sharma, Chemometrics in forensic science, TrAC Trends Anal. Chem. 105 (2018) 191–201. https://doi.org/10.1016/j.trac.2018.05.010.

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

[35] B.J. Pollo, C.A. Teixeira, J.R. Belinato, M.F. Furlan, I.C. de M. Cunha, C.R. Vaz, G.V. Volpato, F. Augusto, Chemometrics, Comprehensive Two-Dimensional gas chromatography and “omics” sciences: Basic tools and recent applications, TrAC Trends Anal. Chem. 134 (2021) 116111. https://doi.org/10.1016/j.trac.2020.116111.

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

[37] G. Ivosev, L. Burton, R. Bonner, Dimensionality Reduction and Visualization in Principal Component Analysis, Anal. Chem. 80 (2008) 4933–4944. https://doi.org/10.1021/ac800110w.

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

[39] vol3d v2, (2018). https://www.mathworks.com/matlabcentral/fileexchange/22940-vol3d-v2 (accessed January 7, 2022).

[40] T.J. Trinklein, S.E. Prebihalo, C.G. Warren, G.S. Ochoa, R.E. Synovec, Discovery-based analysis and quantification for comprehensive three-dimensional gas chromatography flame ionization detection data, J. Chromatogr. A 1623 (2020) 461190. https://doi.org/10.1016/j.chroma.2020.461190.

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

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

Chapter 7. Conclusion and future directions

Comprehensive two-dimensional gas chromatography (GC×GC) is an immensely powerful tool for the separation of complex samples, particularly when it is coupled with time-of-flight mass spectrometry (TOFMS) detection. However, the resulting 4D data cube of information is so complex that it necessitates the application of appropriate chemometric tools, depending on the analysis goals. The goal of the work presented herein was to rigorously evaluate numerous non-targeted, i.e. "discovery-based", chemometric tools in increasingly complex analytical scenarios. Ultimately, the work presented in this dissertation will ideally extend the applicability of non-targeted chemometric tools to lower concentration (low-ppm to ppb) samples and higher-dimensional instrumental platforms (GC^3-TOFMS), as well as improve analyte discovery in unsupervised scenarios.

7.1 CHAPTER 2 SUMMARY AND FUTURE DIRECTIONS

In Chapter 2, the effect of 2D binning on PCA of GC×GC-FID was explored. A total of 110 bin sizes were applied to a GC×GC-FID dataset of five diesel fuels prior to PCA. The quantitative metric degree-of-class separation (DCS) was utilized to assess the "success" of each bin size, whereby DCS was calculated between the nearest-neighbor pairs of fuels in the scores plot (5 pairs total). The results of this work showed that each fuel pair had a unique optimum bin size, depending on the precise chemical differences between the fuels. Fuels with greater boiling point (^{1}D)-based differences were able to tolerate larger ^{1}D bin sizes, whereas fuels with greater polarity (^{2}D)-based differences were able to tolerate larger ^{2}D bin sizes. Thus, this work revealed that binning cannot be applied in a "one size fits all" manner for a multi-class dataset. Given the challenge this presents to optimizing data processing, future work may envision devising a new

metric for multi-class PCA scores plots. For example, one could imagine calculating the area between multiple classes and then repeating the binning study to identify a possible optimum bin size.

Unlike our novel tile-based F-ratio approach, binning only uses one grid of tiles. Therefore, it is possible that analytes were split between bins in this study, which would complicate analyte identification with an MS detector. It would be interesting to repeat the binning study detailed in Chapter 2 with the tile-based code to confirm if the optimum bin sizes change or stay the same. Additionally, given that the fuels studied herein were highly different samples, it would be interesting to repeat this binning study on a spiked versus un-spiked comparison, whereby the samples will likely be completely overlapped in the un-binned scores plot.

7.2 CHAPTER 3 SUMMARY AND FUTURE DIRECTIONS

In Chapter 3, the limit of discovery of tile-based F-ratio analysis was quantitatively defined. Tile-based F-ratio analysis was performed on three increasingly complex comparisons of GC×GC-TOFMS jet fuel chromatograms; 15-ppm spiked versus neat, 3-ppm spiked versus neat, and 1.5-ppm spiked versus neat. For each F-ratio comparison, two key software parameters were varied: the number of F-ratio *m/z* to average for ranking, and the ^{2}D tile dimension. The results of this work showed that ranking with the top-F-ratio *m/z* was advantageous at the low concentration comparisons (3-ppm versus neat and 1.5-ppm versus neat), as the number of selective *m/z* decreases with concentration. It was concluded that analytes spiked at concentrations below their limit of quantification (*LOQ*) would not be discovered by tile-based F-ratio analysis, and the analytes spiked at concentrations approaching their *LOQ* were the ones which benefited from ranking with the top F-ratio *m/z*. Lastly, use of too large or too small of a

^{2}D tile dimension worsened the F-ratio hitlist rankings by drowning out analyte signal and producing excessive redundant hits, respectively. The ideal ^{2}D tile dimension depends on the average ^2W_B observed in the dataset.

A logical next step in this study would be to explore even lower spike concentrations at ppb-levels, given that several spiked sulfur analytes had *LOQ* in the ppb range. A limitation of the presented work was that the sulfur analytes were overly retained on the ^{2}D stationary phase, which produced tailing peaks and likely lowered the resulting sensitivity of these peaks. Future work would involve rigorous optimization of the column set to maximize sensitivity and improve the potential concentration range of subsequent tile-based F-ratio analysis. In the same vein, application of tile-based F-ratio analysis with these optimized parameters to high resolution GC×GC-MS data would be interesting to explore.

7.3 CHAPTER 4 SUMMARY AND FUTURE DIRECTIONS

In Chapter 4, the commercial tile-based F-ratio software ChromaTOF Tile was applied to differentiate five Grillo wines from different geographical regions of Sicily. Three bottles of each wine were obtained, and three SPME replicates per bottle were collected using GC×GC-TOFMS, resulting in nine replicates per wine for F-ratio analysis. A rigorous evaluation of the %*RSD* in peak areas per F-ratio hit revealed that the chemical variation (i.e., different wines) was far more significant than the replicate-based variation. Thus, to reduce data density and improve the accuracy of subsequent chemometric analysis, the injection replicates per bottle were averaged (3 replicates per wine now). An offline one-way *ANOVA* and ROC curve analysis identified 187 true positives out of 220 total hits in the hitlist. PCA using these true positive hits accurately differentiated the wines, although the loadings revealed that only a handful of highly loaded compounds were needed to classify wines 1-5. However, an in-depth sensory analysis revealed

the value in identifying all the true positive hits from F-ratio analysis, as the wines exhibited highly unique sensory profiles.

It is important to note that the sensory analysis performed herein was solely based on sensory databases and previous work; no sensory data was collected by the authors involved. Thus, future work may involve performing GC×GC coupled with olfactometric detection on wines 1-5. The significance of all 187 true positive hits in the F-ratio hitlist, in terms of sensory properties, could then be assessed by performing partial least squares (PLS) regression with the peak areas and sensory data. Future work may also involve repeating this study with more chemically similar wines, to really challenge the commercial ChromaTOF software.

7.4 CHAPTER 5 SUMMARY AND FUTURE DIRECTIONS

In Chapter 5, tile-based variance rank initiated-unsupervised sample indexing (VRI-USI) was introduced as an unsupervised tool for GC×GC-TOFMS data. Briefly, this method works by replacing the F-ratio metric in the tile-based platform with RSD^2. Each hit in the resulting RSD^2 ranked hitlist is then subjected to *k*-means clustering at various values of *k* to find the sample index assignments. Based on a probabilistic argument, repeating sample index assignments are indicative of classes in a dataset. An initial evaluation of this method on a 30-ppm spiked versus 15-ppm spiked versus neat JP8 jet fuel dataset identified all 14 spiked analytes at the top of the hitlist, 11 of which had matching index assignments. This was the most frequently re-occurring index assignment in the hitlist, correctly indicating the spike level class. Application of this tool to a 3-ppm spiked versus neat JP8 jet fuel dataset revealed that 8/10 of the discovered spiked hits had matching index assignments, thus highlighting the utility of VRI-USI for low concentration analyte discovery. Finally, tile-based VRI-USI was applied to a more complex dataset of three jet fuels (J1800A, JP4, and JP8), each spiked with 30-ppm sulfur mix and neat. 453/520 hits in the

hitlist had index assignments indicative of clustering by fuel type. However, of the remaining 67 hits, 9 hits exhibited specific patterns in their index assignments which indicated clustering by spike level. 8 of these hits were in fact the spiked compounds, with only one false positive. Evaluation of S_{max} distributions revealed that these 8 hits had high S_{max} in line with the S_{max} values of the ideal fuel type clustering hits, which validated the presence of a true sub-class (spike level) in the data.

Thus far, tile-based VRI-USI has only been applied to this contrived jet fuel dataset which exhibited minimal between-replicate variation. It is highly probable that application of VRI-USI to metabolomics samples would be more challenging, since metabolomics samples experience considerable variability due to culture conditions, sample preparation, and other factors. A necessary next step is then to apply VRI-USI to a metabolomics dataset with known classes and sub-classes and compare the resulting index assignments to these class labels. Although this would be a "supervised" application of VRI-USI like the work presented herein, it is necessary to validate the method.

The high S_{max} values associated with the fuel type clustering and spike level sub-clustering hits were an intriguing find that demands further analysis. Only a qualitative assessment of high versus low S_{max} was made herein, but it would be intriguing to define an S_{max} threshold for identifying classes and sub-classes. One approach could involve simulating a series of chromatograms at varying *S/N* and concentration differences, performing VRI-USI, and determining when S_{max} begins to decrease.

7.5 CHAPTER 6 SUMMARY AND FUTURE DIRECTIONS

In Chapter 6, the first application of PCA to GC^3-TOFMS data was presented. Five replicates each of four jet fuels were collected on a full-transfer GC^3-TOFMS instrument. Prior to PCA, the data was re-registered using a novel ^2D re-registration technique in which vacant ^3D modulations were removed. The resulting $^1D \times {}^2D$ chromatogram was properly centered, with the alkyl, cycloalkyl, and aromatic components in the correct locations given the reverse column format. A simple ^3D re-registration was also performed to remove wrap-around. The resulting 3D chromatograms exhibited differences which were not as easily elucidated in the $^1D \times {}^2D$ view. Most notably, the 3D PCA loadings served as a highly useful visual tool for distinguishing the fuels. An additional PCA model of the two most similar fuels proved highly useful, as the loadings revealed differences which were overshadowed in the multi-fuel PCA model.

This study was conducted using noticeably different fuels for proof-of-principle. However, it would be intriguing to collect GC^3-TOFMS data of highly similar samples and see if the 3D loadings reveal unknown differences. It is important to note that highly similar samples might not be distinguished in a $^1D \times {}^2D$ scores plot, so utilizing the third principal component (PC3) would be needed in this scenario. A logical next step in applying chemometrics to GC^3-TOFMS data is the application of supervised F-ratio analysis, which is an ongoing project.

BIBLIOGRAPHY

Abdulhussain, N., Nawada, S., Schoenmakers, P., 2021. Latest Trends on the Future of Three-Dimensional Separations in Chromatography. Chem. Rev. 121, 12016–12034. https://doi.org/10.1021/acs.chemrev.0c01244

Alexandrino, G.L., Malmborg, J., Augusto, F., Christensen, J.H., 2019. Investigating weathering in light diesel oils using comprehensive two-dimensional gas chromatography–High resolution mass spectrometry and pixel-based analysis: Possibilities and limitations. J. Chromatogr. A 1591, 155–161. https://doi.org/10.1016/j.chroma.2019.01.042

Aloisi, I., Giocastro, B., Ferracane, A., Salerno, T.M.G., Zoccali, M., Tranchida, P.Q., Mondello, L., 2021. Preliminary observations on the use of a novel low duty cycle flow modulator for comprehensive two-dimensional gas chromatography. J. Chromatogr. A 1643, 462076. https://doi.org/10.1016/j.chroma.2021.462076

Amaral, M.S.S., Nolvachai, Y., Marriott, P.J., 2020. Comprehensive Two-Dimensional Gas Chromatography Advances in Technology and Applications: Biennial Update. Anal. Chem. 92, 85–104. https://doi.org/10.1021/acs.analchem.9b05412

Amigo, J.M., Skov, T., Bro, R., 2010. ChroMATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics. Chem. Rev. 110, 4582–4605. https://doi.org/10.1021/cr900394n

An, Z., Li, X., Shi, Z., Williams, B.J., Harrison, R.M., Jiang, J., 2021. Frontier review on comprehensive two-dimensional gas chromatography for measuring organic aerosol. J. Hazard. Mater. Lett. 2, 100013. https://doi.org/10.1016/j.hazl.2021.100013

An, Z., Ren, H., Xue, M., Guan, X., Jiang, J., 2020. Comprehensive two-dimensional gas chromatography mass spectrometry with a solid-state thermal modulator for in-situ speciated measurement of organic aerosols. J. Chromatogr. A 1625, 461336. https://doi.org/10.1016/j.chroma.2020.461336

Anzardi, M.B., Arancibia, J.A., Olivieri, A.C., 2021. Processing multi-way chromatographic data for analytical calibration, classification and discrimination: A successful marriage between separation science and chemometrics. TrAC Trends Anal. Chem. 134, 116128. https://doi.org/10.1016/j.trac.2020.116128

Arslan, F.N., Kolk, A., Janssen, H.-G., 2019. Methods for one– and two–dimensional gas chromatography with flame ionization detection for identification of Mycobacterium tuberculosis in sputum. J. Chromatogr. B. https://doi.org/10.1016/j.jchromb.2019.06.012

Asri, M.N.M., Nestrgan, N.F., Nor, N.A.M., Verma, R., 2021. On the discrimination of inkjet, laser and photocopier printed documents using Raman spectroscopy and chemometrics: Application in forensic science. Microchem. J. 165, 106136. https://doi.org/10.1016/j.microc.2021.106136

Baerncopf, J.M., McGuffin, V.L., Smith, R.W., 2011. Association of Ignitable Liquid Residues to Neat Ignitable Liquids in the Presence of Matrix Interferences Using Chemometric Procedures*,†. J. Forensic Sci. 56, 70–81. https://doi.org/10.1111/j.1556-4029.2010.01563.x

Bahaghighat, H.D., Freye, C.E., Gough, D.V., Sudol, P.E., Synovec, R.E., 2018a. Ultrafast separations via pulse flow valve modulation to enable high peak capacity multidimensional gas chromatography. J. Chromatogr. A 1573, 115–124. https://doi.org/10.1016/j.chroma.2018.08.001

Bahaghighat, H.D., Freye, C.E., Synovec, R.E., 2018b. Recent advances in modulator technology for comprehensive two dimensional gas chromatography. TrAC Trends Anal. Chem. https://doi.org/10.1016/j.trac.2018.04.016

Bai, L., Smuts, J., Schenk, J., Cochran, J., Schug, K.A., 2018. Comparison of GC-VUV, GC-FID, and comprehensive two-dimensional GC–MS for the characterization of weathered and unweathered diesel fuels. Fuel 214, 521–527. https://doi.org/10.1016/j.fuel.2017.11.053

Ballabio, D., 2015. A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. Chemom. Intell. Lab. Syst. 149, 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003

Barra, I., Mansouri, M.A., Cherrah, Y., Kharbach, M., Bouklouze, A., 2019. FTIR fingerprints associated to a PLS-DA model for rapid detection of smuggled non-compliant diesel marketed in Morocco. Vib. Spectrosc. 101, 40–45. https://doi.org/10.1016/j.vibspec.2019.02.001

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

Beckstrom, A.C., Humston, E.M., Snyder, L.R., Synovec, R.E., Juul, S.E., 2011. Application of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method to identify potential biomarkers of perinatal asphyxia in a non-human primate model. J. Chromatogr. A 1218, 1899–1906. https://doi.org/10.1016/j.chroma.2011.01.086

Berrier, K.L., Prebihalo, S.E., Synovec, R.E., 2020. Chapter 7 - Advanced data handling in comprehensive two-dimensional gas chromatography, in: Snow, N.H. (Ed.), Separation Science and Technology, Basic Multidimensional Gas Chromatography. Academic Press, pp. 229–268. https://doi.org/10.1016/B978-0-12-813745-1.00007-6

Berrou, K., Dunyach-Remy, C., Lavigne, J.-P., Roig, B., Cadiere, A., 2019. Multiple stir bar sorptive extraction combined with gas chromatography-mass spectrometry analysis for a tentative identification of bacterial volatile and/or semi-volatile metabolites. Talanta 195, 245–250. https://doi.org/10.1016/j.talanta.2018.11.042

Blumberg, L.M., 2008. Accumulating resampling (modulation) in comprehensive two-dimensional capillary GC (GC×GC). J. Sep. Sci. 31, 3358–3365. https://doi.org/10.1002/jssc.200800424

Boegelsack, N., Sandau, C., McMartin, D.W., Withey, J.M., O'Sullivan, G., 2021. Development of retention time indices for comprehensive multidimensional gas chromatography and application to ignitable liquid residue mapping in wildfire investigations. J. Chromatogr. A 1635, 461717. https://doi.org/10.1016/j.chroma.2020.461717

Brereton, R.G., 2021. P values and multivariate distributions: Non-orthogonal terms in regression models. Chemom. Intell. Lab. Syst. 210, 104264. https://doi.org/10.1016/j.chemolab.2021.104264

Brereton, R.G., 2015. Pattern recognition in chemometrics. Chemom. Intell. Lab. Syst. 149, 90–96. https://doi.org/10.1016/j.chemolab.2015.06.012

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

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

Budiarto, A., Mahesworo, B., Baurley, J., Suparyanto, T., Pardamean, B., 2019. Fast and Effective Clustering Method for Ancestry Estimation. Procedia Comput. Sci., The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019) : Enabling Collaboration to Escalate Impact of Research Results for Society 157, 306–312. https://doi.org/10.1016/j.procs.2019.08.171

Cadoret, M., Husson, F., 2013. Construction and evaluation of confidence ellipses applied at sensory data. Food Qual. Prefer. 28, 106–115. https://doi.org/10.1016/j.foodqual.2012.09.005

Cai, D., Zhang, C., He, X., 2010. Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10. Association for Computing Machinery, New York, NY, USA, pp. 333–342. https://doi.org/10.1145/1835804.1835848

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

Cain, C.N., Sudol, P.E., Berrier, K.L., Synovec, R.E., 2021. Development of variance rank initiated-unsupervised sample indexing for gas chromatography-mass spectrometry analysis. Talanta 233, 122495. https://doi.org/10.1016/j.talanta.2021.122495

Camacho, J., 2014. Visualizing Big data with Compressed Score Plots: Approach and research challenges. Chemom. Intell. Lab. Syst. 135, 110–125. https://doi.org/10.1016/j.chemolab.2014.04.011

Campo, E., Ferreira, V., Escudero, A., Cacho, J., 2005. Prediction of the Wine Sensory Properties Related to Grape Variety from Dynamic-Headspace Gas Chromatography−Olfactometry Data. J. Agric. Food Chem. 53, 5682–5690. https://doi.org/10.1021/jf047870a

Cao, X., Liu, Y., Wang, J., Liu, C., Duan, Q., 2020. Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network. Aquac. Eng. 91, 102122. https://doi.org/10.1016/j.aquaeng.2020.102122

Carlin, S., Vrhovsek, U., Franceschi, P., Lotti, C., Bontempo, L., Camin, F., Toubiana, D., Zottele, F., Toller, G., Fait, A., Mattivi, F., 2016. Regional features of northern Italian sparkling wines, identified using solid-phase micro extraction and comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry. Food Chem. 208, 68–80. https://doi.org/10.1016/j.foodchem.2016.03.112

Castro, A.L., Tarelho, S., Melo, P., Franco, J.M., 2018. A fast and reliable method for quantitation of THC and its 2 main metabolites in whole blood by GC–MS/MS (TQD). Forensic Sci. Int. 289, 344–351. https://doi.org/10.1016/j.forsciint.2018.06.003

Chen, H., Gao, H., Fang, X., Ye, L., Zhou, Y., Yang, H., 2015. Effects of allyl isothiocyanate treatment on postharvest quality and the activities of antioxidant enzymes of mulberry fruit. Postharvest Biol. Technol. 108, 61–67. https://doi.org/10.1016/j.postharvbio.2015.05.011

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

ChromaTOF® Software [WWW Document], 2022. LECO Corp. URL https://www.leco.com/product/chromatof-software (accessed 1.17.22).

ChromaTOF® Tile Analytical Software [WWW Document], 2022. LECO Corp. URL https://www.leco.com/product/chromatof-tile (accessed 11.16.21).

Chromatography Method Development - OpenLab ChemStation | Agilent [WWW Document], 2022 URL https://www.agilent.com/en/product/software-informatics/analytical-software-suite/chromatography-data-systems/openlab-chemstation (accessed 1.17.22).

ChromCompare+ [WWW Document], 2022. URL https://www.sepsolve.com/chromcompare/ (accessed 11.16.21).

Chu, H.-J., Liau, C.-J., Lin, C.-H., Su, B.-S., 2012. Integration of fuzzy cluster analysis and kernel density estimation for tracking typhoon trajectories in the Taiwan region. Expert Syst. Appl. 39, 9451–9457. https://doi.org/10.1016/j.eswa.2012.02.114

Costa, R., Fanali, C., Pennazza, G., Tedone, L., Dugo, L., Santonico, M., Sciarrone, D., Cacciola, F., Cucchiarini, L., Dachà, M., Mondello, L., 2015. Screening of volatile compounds composition of white truffle during storage by GCxGC-(FID/MS) and gas sensor array analyses. LWT - Food Sci. Technol. 60, 905–913. https://doi.org/10.1016/j.lwt.2014.09.054

Davis, J.M., 1991. Statistical theory of spot overlap in two-dimensional separations. Anal. Chem. 63, 2141–2152. https://doi.org/10.1021/ac00019a014

de Almeida, V.E., de Sousa Fernandes, D.D., Diniz, P.H.G.D., de Araújo Gomes, A., Véras, G., Galvão, R.K.H., Araujo, M.C.U., 2021. Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data. Food Chem. 363, 130296. https://doi.org/10.1016/j.foodchem.2021.130296

de la Mata, A.P., McQueen, R.H., Nam, S.L., Harynuk, J.J., 2017. Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics. Anal. Bioanal. Chem. 409, 1905–1913. https://doi.org/10.1007/s00216-016-0137-1

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

Deans, D.R., 1968. A new technique for heart cutting in gas chromatography [1]Eine neue Technik des "heart cuttings" in der Gas-ChromatographieUne nouvelle technique de "Heart Cutting" dans la chromatographie en phase gazeuse. Chromatographia 1, 18–22. https://doi.org/10.1007/BF02259005

Del Carlo, M., Pepe, A., Sacchetti, G., Compagnone, D., Mastrocola, D., Cichelli, A., 2008. Determination of phthalate esters in wine using solid-phase extraction and gas chromatography–mass spectrometry. Food Chem. 111, 771–777. https://doi.org/10.1016/j.foodchem.2008.04.065

Demyttenaere, J.C.R., Moriña, R.M., Sandra, P., 2003. Monitoring and fast detection of mycotoxin-producing fungi based on headspace solid-phase microextraction and headspace sorptive extraction of the volatile metabolites. J. Chromatogr. A, 25th International Symposium on Capillary Chromatography 985, 127–135. https://doi.org/10.1016/S0021-9673(02)01417-6

Dixon, S.J., Heinrich, N., Holmboe, M., Schaefer, M.L., Reed, R.R., Trevejo, J., Brereton, R.G., 2009. Use of cluster separation indices and the influence of outliers: application of two new separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles. J. Chemom. 23, 19–31. https://doi.org/10.1002/cem.1189

Dubois, L.M., Perrault, K.A., Stefanuto, P.-H., Koschinski, S., Edwards, M., McGregor, L., Focant, J.-F., 2017. Thermal desorption comprehensive two-dimensional gas chromatography coupled to variable-energy electron ionization time-of-flight mass spectrometry for monitoring subtle changes in volatile organic compound profiles of human blood. J. Chromatogr. A 1501, 117–127. https://doi.org/10.1016/j.chroma.2017.04.026

Dugo, G., Franchina, F.A., Scandinaro, M.R., Bonaccorsi, I., Cicero, N., Tranchida, P.Q., Mondello, L., 2014. Elucidation of the volatile composition of Marsala wines by using comprehensive two-dimensional gas chromatography. Food Chem. 142, 262–268. https://doi.org/10.1016/j.foodchem.2013.07.061

Dunkle, M.N., Pijcke, P., Winniford, B., Bellos, G., 2019. Quantification of the composition of liquid hydrocarbon streams: Comparing the GC-VUV to DHA and GCxGC. J. Chromatogr. A 1587, 239–246. https://doi.org/10.1016/j.chroma.2018.12.026

Eilers, P.H.C., 2004. Parametric Time Warping. Anal. Chem. 76, 404–411. https://doi.org/10.1021/ac034800e

Eilers, P.H.C., 2003. A Perfect Smoother. Anal. Chem. 75, 3631–3636. https://doi.org/10.1021/ac034173t

Fasciotti, M., Monteiro, T.V.C., Ferreira, A.A., Eberlin, M.N., Neves, L.A., 2017. Two-point normalization using internal and external standards for a traceable determination of δ13C values of fatty acid methyl esters by gas chromatography/combustion/isotope ratio mass spectrometry. Int. J. Mass Spectrom., SI: Jose Riveros 418, 41–50. https://doi.org/10.1016/j.ijms.2016.12.002

Ferracane, A., Zoccali, M., Cacciola, F., Salerno, T.M.G., Tranchida, P.Q., Mondello, L., 2021. Determination of multi-pesticide residues in vegetable products using a "reduced-scale" Quechers method and flow-modulated comprehensive two-dimensional gas chromatography-triple quadrupole mass spectrometry. J. Chromatogr. A 1645, 462126. https://doi.org/10.1016/j.chroma.2021.462126

Ferreira, V.H.C., Hantao, L.W., Poppi, R.J., 2020. Consumable-free Comprehensive Three-Dimensional Gas Chromatography and PARAFAC for Determination of Allergens in Perfumes. Chromatographia 83, 581–592. https://doi.org/10.1007/s10337-020-03863-6

Filzmoser, P., Walczak, B., 2014. What can go wrong at the data normalization step for identification of biomarkers. J. Chromatogr. A 1362, 194–205. https://doi.org/10.1016/j.chroma.2014.08.050

Fisher, R.A., 1920. A mathematical Examination of the Methods of determining the Accuracy of Observation by the Mean Error, and by the Mean Square Error. Mon. Not. R. Astron. Soc. 80, 758–770. https://doi.org/10.1093/mnras/80.8.758

Flood, M.E., Connolly, M.P., Comiskey, M.C., Hupp, A.M., 2016. Evaluation of single and multi-feedstock biodiesel – diesel blends using GCMS and chemometric methods. Fuel 186, 58–67. https://doi.org/10.1016/j.fuel.2016.08.069

Fortunato de Carvalho Rocha, W., Schantz, M.M., Sheen, D.A., Chu, P.M., Lippa, K.A., 2017. Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data. Fuel 197, 248–258. https://doi.org/10.1016/j.fuel.2017.02.025

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

França, D., Pereira, V.B., Coutinho, D.M., Ainstein, L.M., Azevedo, D.A., 2018. Speciation and quantification of high molecular weight paraffins in Brazilian whole crude oils using high-temperature comprehensive two-dimensional gas chromatography. Fuel 234, 1154–1164. https://doi.org/10.1016/j.fuel.2018.07.145

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

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

Furdíková, K., Bajnociová, L., Malík, F., Špánik, I., 2017. Investigation of volatile profile of varietal Gewürztraminer wines using two-dimensional gas chromatography. J. Food Nutr. Res. 56, 73–85.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F., 2013. Dynamic classifier selection for One-vs-One strategy: Avoiding non-competent classifiers. Pattern Recognit. 46, 3412–3424. https://doi.org/10.1016/j.patcog.2013.04.018

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Genuit, W., Chaabani, H., 2017. Comprehensive two-dimensional gas chromatography-field ionization time-of-flight mass spectrometry (GCxGC-FI-TOFMS) for detailed hydrocarbon middle distillate analysis. Int. J. Mass Spectrom., SI: Nico Nibbering Issue 413, 27–32. https://doi.org/10.1016/j.ijms.2016.12.001

Girod, A., Weyermann, C., 2014. Lipid composition of fingermark residue and donor classification using GC/MS. Forensic Sci. Int. 238, 68–82. https://doi.org/10.1016/j.forsciint.2014.02.020

Giuffrida, D., Zoccali, M., Mondello, L., 2020. Recent developments in the carotenoid and carotenoid derivatives chromatography-mass spectrometry analysis in food matrices. TrAC Trends Anal. Chem. 132, 116047. https://doi.org/10.1016/j.trac.2020.116047

Godoy, L.A.F. de, Ferreira, E.C., Pedroso, M.P., Fidélis, C.H. de V., Augusto, F., Poppi, R.J., 2008. Quantification of Kerosene in Gasoline by Comprehensive Two-Dimensional Gas Chromatography and N-Way Multivariate Analysis. Anal. Lett. 41, 1603–1614. https://doi.org/10.1080/00032710802122222

Gough, D.V., Bahaghighat, H.D., Synovec, R.E., 2019a. Column selection approach to achieve a high peak capacity in comprehensive three-dimensional gas chromatography. Talanta 195, 822–829. https://doi.org/10.1016/j.talanta.2018.12.007

Gough, D.V., Song, D.H., Schöneich, S., Prebihalo, S.E., Synovec, R.E., 2019b. Development of Ultrafast Separations Using Negative Pulse Partial Modulation To Enable New Directions in Gas Chromatography. Anal. Chem. 91, 7328–7335. https://doi.org/10.1021/acs.analchem.9b01085

Govender, P., Sivakumar, V., 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollut. Res. 11, 40–56. https://doi.org/10.1016/j.apr.2019.09.009

Haaland, D.M., Thomas, E.V., 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60, 1193–1202. https://doi.org/10.1021/ac00162a020

Haar, L., Anding, K., Trambitckii, K., Notni, G., 2019. Comparison between Supervised and Unsupervised Feature Selection Methods:, in: Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods. Presented at the 8th International Conference on Pattern Recognition Applications and Methods, SCITEPRESS - Science and Technology Publications, Prague, Czech Republic, pp. 582–589. https://doi.org/10.5220/0007385305820589

Han, Y., Zhang, Y., Xu, C., Hsu, C.S., 2018. Molecular characterization of sulfur-containing compounds in petroleum. Fuel 221, 144–158. https://doi.org/10.1016/j.fuel.2018.02.110

Hantao, L.W., Aleme, H.G., Pedroso, M.P., Sabin, G.P., Poppi, R.J., Augusto, F., 2012. Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: A review. Anal. Chim. Acta 731, 11–23. https://doi.org/10.1016/j.aca.2012.04.003

Harris, D.C., 2016. Chapter 4: Statistics, in: Quantitative Chemical Analysis. W. H. Freeman and Company, New York, NY, pp. 64–89.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. J. R. Stat. Soc. Ser. C Appl. Stat. 28, 100–108. https://doi.org/10.2307/2346830

Harvey, P.McA., Shellie, R.A., 2012. Data Reduction in Comprehensive Two-Dimensional Gas Chromatography for Rapid and Repeatable Automated Data Analysis. Anal. Chem. 84, 6501–6507. https://doi.org/10.1021/ac300664h

He, X., Cai, D., Niyogi, P., n.d. Laplacian Score for Feature Selection 8.

Hegazi, A.H., Andersson, J.T., 2016. 6 - Polycyclic aromatic sulfur heterocycles as source diagnostics of petroleum pollutants in the marine environment, in: Stout, S.A., Wang, Z. (Eds.), Standard Handbook Oil Spill Environmental Forensics (Second Edition). Academic Press, Boston, pp. 313–342. https://doi.org/10.1016/B978-0-12-803832-1.00006-4

Heiberger, R.M., Neuwirth, E., 2009. R Through Excel. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4419-0052-4

Held, L., Ott, M., 2016. How the Maximal Evidence of P-Values Against Point Null Hypotheses Depends on Sample Size. Am. Stat. 70, 335–341. https://doi.org/10.1080/00031305.2016.1209128

Hierarchical Clustering | SpringerLink [WWW Document], n.d. URL https://link.springer.com/chapter/10.1007/978-3-319-21903-5_8 (accessed 11.25.20).

Hogan, J.M., Engel, R.A., Stevenson, H.F., 1970. Versatile internal standard technique for the gas chromatographic determination of water in liquids. Anal. Chem. 42, 249–252. https://doi.org/10.1021/ac60284a033

Hoggard, J.C., Synovec, R.E., 2007. Parallel Factor Analysis (PARAFAC) of Target Analytes in GC × GC−TOFMS Data: Automated Selection of a Model with an Appropriate Number of Factors. Anal. Chem. 79, 1611–1619. https://doi.org/10.1021/ac061710b

Hoggard, J.C., Wahl, J.H., Synovec, R.E., Mong, G.M., Fraga, C.G., 2010. Impurity Profiling of a Chemical Weapon Precursor for Possible Forensic Signatures by Comprehensive Two-Dimensional Gas Chromatography/Mass Spectrometry and Chemometrics. Anal. Chem. 82, 689–698. https://doi.org/10.1021/ac902247x

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

Hupp, A.M., Marshall, L.J., Campbell, D.I., Smith, R.W., McGuffin, V.L., 2008. Chemometric analysis of diesel fuel for forensic and environmental applications. Anal. Chim. Acta 606, 159–171. https://doi.org/10.1016/j.aca.2007.11.007

Ilc, T., Werck-Reichhart, D., Navrot, N., 2016. Meta-Analysis of the Core Aroma Components of Grape and Wine Aroma. Front. Plant Sci. 7, 1472. https://doi.org/10.3389/fpls.2016.01472

Ivosev, G., Burton, L., Bonner, R., 2008. Dimensionality Reduction and Visualization in Principal Component Analysis. Anal. Chem. 80, 4933–4944. https://doi.org/10.1021/ac800110w

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

Jáčová, J., Gardlo, A., Dimandja, J.-M.D., Adam, T., Friedecký, D., 2019. Impact of sample dimensionality on orthogonality metrics in comprehensive two-dimensional separations. Anal. Chim. Acta 1064, 138–149. https://doi.org/10.1016/j.aca.2019.03.018

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett., Award winning papers from the 19th International Conference on Pattern Recognition (ICPR) 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

JAMES, A.T., MARTIN, A.J., 1952. Gas-liquid partition chromatography; the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid. Biochem. J. 50, 679–690. https://doi.org/10.1042/bj0500679

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

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

Jennerwein, M.K., Eschner, M., Gröger, T., Wilharm, T., Zimmermann, R., 2014. Complete Group-Type Quantification of Petroleum Middle Distillates Based on Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC×GC-TOFMS) and Visual Basic Scripting. Energy Fuels 28, 5670–5681. https://doi.org/10.1021/ef501247h

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

Ji, H., Lu, H., Zhang, Z., 2016. Pure ion chromatogram extraction via optimal k -means clustering. RSC Adv. 6, 56977–56985. https://doi.org/10.1039/C6RA08409E

Johnson, K.J., Synovec, R.E., 2002. Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. Chemom. Intell. Lab. Syst., Fourth International Conference on Environ metrics and Chemometrics held in Las Vegas, NV, USA, 18-20 September 2000 60, 225–237. https://doi.org/10.1016/S0169-7439(01)00198-8

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

K. Robards, P.R. Haddad, P.E. Jackson, 2004. Principles and Practice of Modern Chromatographic Methods, 1st ed. Elsevier.

Kalogiouri, N.P., Aalizadeh, R., Dasenaki, M.E., Thomaidis, N.S., 2020. Application of High Resolution Mass Spectrometric methods coupled with chemometric techniques in olive oil authenticity studies - A review. Anal. Chim. Acta 1134, 150–173. https://doi.org/10.1016/j.aca.2020.07.029

Kandikattu, H.K., Rachitha, P., Jayashree, G.V., Krupashree, K., Sukhith, M., Majid, A., Amruta, N., Khanum, F., 2017. Anti-inflammatory and anti-oxidant effects of Cardamom (Elettaria repens (Sonn.) Baill) and its phytochemical analysis by 4D GCXGC TOF-MS. Biomed. Pharmacother. 91, 191–201. https://doi.org/10.1016/j.biopha.2017.04.049

Kaufman, L., Rousseeuw, P.J., 2009. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

Khummueng, W., Harynuk, J., Marriott, P.J., 2006. Modulation Ratio in Comprehensive Two-dimensional Gas Chromatography. Anal. Chem. 78, 4578–4587. https://doi.org/10.1021/ac052270b

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

Kopperi, M., Ruiz-Jiménez, J., Hukkinen, J.I., Riekkola, M.-L., 2013. New way to quantify multiple steroidal compounds in wastewater by comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry. Anal. Chim. Acta 761, 217–226. https://doi.org/10.1016/j.aca.2012.11.059

Krakowska, B., Stanimirova, I., Orzel, J., Daszykowski, M., Grabowski, I., Zaleszczyk, G., Sznajder, M., 2015. Detection of discoloration in diesel fuel based on gas chromatographic fingerprints. Anal. Bioanal. Chem. 407, 1159–1170. https://doi.org/10.1007/s00216-014-8332-4

Kumagai, S., Matsukami, A., Kabashima, F., Sakurai, M., Kanai, M., Kameda, T., Saito, Y., Yoshioka, T., 2020. Combining pyrolysis–two-dimensional gas chromatography–time-of-flight mass spectrometry with hierarchical cluster analysis for rapid identification of pyrolytic interactions: Case study of co-pyrolysis of PVC and biomass components. Process Saf. Environ. Prot. 143, 91–100. https://doi.org/10.1016/j.psep.2020.06.036

Kumar, R., Sharma, V., 2018. Chemometrics in forensic science. TrAC Trends Anal. Chem. 105, 191–201. https://doi.org/10.1016/j.trac.2018.05.010

Lantz, B., 2013. The large sample size fallacy. Scand. J. Caring Sci. 27, 487–492. https://doi.org/10.1111/j.1471-6712.2012.01052.x

Ledford, E.B., Billesbach, C.A., Zhu, Q., 2000. GC3: Comprehensive Three-Dimensional Gas Chromatography. J. High Resolut. Chromatogr. 23, 205–207. https://doi.org/10.1002/(SICI)1521-4168(20000301)23:3<205::AID-JHRC205>3.0.CO;2-U

Li, R., Liu, Y., Wang, Z., Zhang, Q., Bai, H., Lv, Q., 2021. High resolution GC–Orbitrap MS for nitrosamines analysis: Method performance, exploration of solid phase extraction regularity, and screening of children's products. Microchem. J. 162, 105878. https://doi.org/10.1016/j.microc.2020.105878

Li, S., Hu, Y., Liu, W., Chen, Y., Wang, F., Lu, X., Zheng, W., 2020. Untargeted volatile metabolomics using comprehensive two-dimensional gas chromatography-mass spectrometry – A solution for orange juice authentication. Talanta 217, 121038. https://doi.org/10.1016/j.talanta.2020.121038

Lima Gomes, P.C.F., Barnes, B.B., Santos-Neto, Á.J., Lancas, F.M., Snow, N.H., 2013. Determination of steroids, caffeine and methylparaben in water using solid phase microextraction-comprehensive two dimensional gas chromatography–time of flight mass spectrometry. J. Chromatogr. A 1299, 126–130. https://doi.org/10.1016/j.chroma.2013.05.023

Liu, R.H., Foster, G., Cone, E.J., Kumar, S.D., 1995. Selecting an appropriate isotopic internal standard for gas chromatography/mass spectrometry analysis of drugs of abuse--pentobarbital example. J. Forensic Sci. 40, 983–989.

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

Liu, Zaiyou., Patterson, D.G., Lee, M.L., 1995. Geometric Approach to Factor Analysis for the Estimation of Orthogonality and Practical Peak Capacity in Comprehensive Two-Dimensional Separations. Anal. Chem. 67, 3840–3845. https://doi.org/10.1021/ac00117a004

Lletí, R., Ortiz, M.C., Sarabia, L.A., Sánchez, M.S., 2004. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. Anal. Chim. Acta, Papers presented at the 5th COLLOQUIUM CHEMIOMETRICUM MEDITERRANEUM 515, 87–100. https://doi.org/10.1016/j.aca.2003.12.020

Lobodin, V.V., Robbins, W.K., Lu, J., Rodgers, R.P., 2015. Separation and Characterization of Reactive and Non-Reactive Sulfur in Petroleum and Its Fractions. Energy Fuels 29, 6177–6186. https://doi.org/10.1021/acs.energyfuels.5b00780

Loegel, T.N., Morris, R.E., Leska, I., 2017. Detection and Quantification of Metal Deactivator Additive in Jet and Diesel Fuel by Liquid Chromatography. Energy Fuels 31, 3629–3634. https://doi.org/10.1021/acs.energyfuels.6b03128

López, R., Aznar, M., Cacho, J., Ferreira, V., 2002. Determination of minor and trace volatile compounds in wine by solid-phase extraction and gas chromatography with mass spectrometric detection. J. Chromatogr. A 966, 167–177. https://doi.org/10.1016/S0021-9673(02)00696-9

Lukić, I., Carlin, S., Horvat, I., Vrhovsek, U., 2019. Combined targeted and untargeted profiling of volatile aroma compounds with comprehensive two-dimensional gas chromatography for differentiation of virgin olive oils according to variety and geographical origin. Food Chem. 270, 403–414. https://doi.org/10.1016/j.foodchem.2018.07.133

Mabood, F., Gilani, S.A., Albroumi, M., Alameri, S., Al Nabhani, M.M.O., Jabeen, F., Hussain, J., Al-Harrasi, A., Boqué, R., Farooq, S., Hamaed, A.M., Naureen, Z., Khan, A., Hussain, Z., 2017. Detection and estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with multivariate methods. Fuel 197, 388–396. https://doi.org/10.1016/j.fuel.2017.02.041

Marchini, M., Charvoz, C., Dujourdy, L., Baldovini, N., Filippi, J.-J., 2014. Multidimensional analysis of cannabis volatile constituents: Identification of 5,5-dimethyl-1-vinylbicyclo[2.1.1]hexane as a volatile marker of hashish, the resin of Cannabis sativa L. J. Chromatogr. A 1370, 200–215. https://doi.org/10.1016/j.chroma.2014.10.045

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

Martín-Alberca, C., García-Ruiz, C., Delémont, O., 2015. Study of chemical modifications in acidified ignitable liquids analysed by GC–MS. Sci. Justice 55, 446–455. https://doi.org/10.1016/j.scijus.2015.06.006

Mat-Desa, W.N.S., Ismail, D., NicDaeid, N., 2011. Classification and Source Determination of Medium Petroleum Distillates by Chemometric and Artificial Neural Networks: A Self Organizing Feature Approach. Anal. Chem. 83, 7745–7754. https://doi.org/10.1021/ac202315y

Materazzi, S., Risoluti, R., Pinci, S., Saverio Romolo, F., 2017. New insights in forensic chemistry: NIR/Chemometrics analysis of toners for questioned documents examination. Talanta 174, 673–678. https://doi.org/10.1016/j.talanta.2017.06.044

McIlroy, J.W., Smith, R.W., McGuffin, V.L., 2015. Assessing the effect of data pretreatment procedures for principal components analysis of chromatographic data. Forensic Sci. Int. 257, 1–12. https://doi.org/10.1016/j.forsciint.2015.07.038

Mendes, B., Gonçalves, J., Câmara, J.S., 2012. Effectiveness of high-throughput miniaturized sorbent- and solid phase microextraction techniques combined with gas chromatography–mass spectrometry analysis for a rapid screening of volatile and semi-volatile composition of wines—A comparative study. Talanta 88, 79–94. https://doi.org/10.1016/j.talanta.2011.10.010

Mitrevski, B., Veleska, B., Engel, E., Wynne, P., Song, S.M., Marriott, P.J., 2011. Chemical signature of ecstasy volatiles by comprehensive two-dimensional gas chromatography. Forensic Sci. Int. 209, 11–20. https://doi.org/10.1016/j.forsciint.2010.11.008

Mogollón, N.G.S., Ribeiro, F.A. de L., Lopez, M.M., Hantao, L.W., Poppi, R.J., Augusto, F., 2013. Quantitative analysis of biodiesel in blends of biodiesel and conventional diesel by comprehensive two-dimensional gas chromatography and multivariate curve resolution. Anal. Chim. Acta 796, 130–136. https://doi.org/10.1016/j.aca.2013.07.071

Mogollón, N.G.S., Ribeiro, F.A.L., Poppi, R.J., Quintana, A.L., Chávez, J.A.G., Agualongo, D.A.P., Aleme, H.G., Augusto, F., Mogollón, N.G.S., Ribeiro, F.A.L., Poppi, R.J., Quintana, A.L., Chávez, J.A.G., Agualongo, D.A.P., Aleme, H.G., Augusto, F., 2017. Exploratory Analysis of Biodiesel by Combining Comprehensive Two-Dimensional Gas Chromatography and Multiway Principal Component Analysis. J. Braz. Chem. Soc. 28, 740–746. https://doi.org/10.21577/0103-5053.20160222

Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., Puertas, B., Cantos-Villar, E., Moreno-Rojas, J.M., 2020. Multivariate optimization of headspace solid-phase microextraction coupled to gas chromatography-mass spectrometry for the analysis of terpenoids in sparkling wines. Talanta 208, 120483. https://doi.org/10.1016/j.talanta.2019.120483

Murrell, K.A., Dorman, F.L., 2021. A comparison of liquid-liquid extraction and stir bar sorptive extraction for multiclass organic contaminants in wastewater by comprehensive two-dimensional gas chromatography time of flight mass spectrometry. Talanta 221, 121481. https://doi.org/10.1016/j.talanta.2020.121481

Muscalu, A.M., Górecki, T., 2018. Comprehensive two-dimensional gas chromatography in environmental analysis. TrAC Trends Anal. Chem. 106, 225–245. https://doi.org/10.1016/j.trac.2018.07.001

Nadeau, J.S., Wilson, R.B., Hoggard, J.C., Wright, B.W., Synovec, R.E., 2011. Study of the interdependency of the data sampling ratio with retention time alignment and principal component analysis for gas chromatography. J. Chromatogr. A 1218, 9091–9101. https://doi.org/10.1016/j.chroma.2011.10.031

Naghizadeh, A., Metaxas, D.N., 2020. Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means. Procedia Comput. Sci., Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020 176, 205–214. https://doi.org/10.1016/j.procs.2020.08.022

Nicholas H. Snow, 2020. Basic Multidimensional Gas Chromatography, Separation Science and Technology. Elsevier.

Nicolli, K.P., Biasoto, A.C.T., Souza-Silva, É.A., Guerra, C.C., dos Santos, H.P., Welke, J.E., Zini, C.A., 2018. Sensory, olfactometry and comprehensive two-dimensional gas chromatography analyses as appropriate tools to characterize the effects of vine management on wine aroma. Food Chem. 243, 103–117. https://doi.org/10.1016/j.foodchem.2017.09.078

Nielsen, N.-P.V., Carstensen, J.M., Smedsgaard, J., 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping 805, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

Nikzad, N., Parastar, H., 2021. Evaluation of the effect of organic pollutants exposure on the antioxidant activity, total phenolic and total flavonoid content of lettuce (Lactuca sativa L.) using UV–Vis spectrophotometry and chemometrics. Microchem. J. 170, 106632. https://doi.org/10.1016/j.microc.2021.106632

Novák, M., Palya, D., Bodai, Z., Nyiri, Z., Magyar, N., Kovács, J., Eke, Z., 2017. Combined cluster and discriminant analysis: An efficient chemometric approach in diesel fuel characterization. Forensic Sci. Int. 270, 61–69. https://doi.org/10.1016/j.forsciint.2016.11.025

Noviandy, T.R., Maulana, A., Sasmita, N.R., Suhendra, R., Muslem, Idroes, G.M., Paristiowati, M., Helwani, Z., Yandri, E., Rahimah, S., Muhammad, Irvanizam, Idroes, R., 2021. The implementation of K-Means clustering in kovats retention index on gas chromatography. IOP Conf. Ser. Mater. Sci. Eng. 1087, 012051. https://doi.org/10.1088/1757-899X/1087/1/012051

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

Ochoa, G.S., Sudol, P.E., Trinklein, T.J., Synovec, R.E., 2022. Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. Talanta 236, 122844. https://doi.org/10.1016/j.talanta.2021.122844

Oros, F.J., Davis, J.M., 1992. Comparison of statistical theories of spot overlap in two-dimensional separations and verification of means for estimating the number of zones. J. Chromatogr. A 591, 1–18. https://doi.org/10.1016/0021-9673(92)80218-J

Orzel, J., Krakowska, B., Stanimirova, I., Daszykowski, M., 2019. Detecting chemical markers to uncover counterfeit rebated excise duty diesel oil. Talanta 204, 229–237. https://doi.org/10.1016/j.talanta.2019.05.113

Ozel, M.Z., Yanık, D.K., Gogus, F., Hamilton, J.F., Lewis, A.C., 2014. Effect of roasting method and oil reduction on volatiles of roasted Pistacia terebinthus using direct thermal desorption-GCxGC-TOF/MS. LWT - Food Sci. Technol. 59, 283–288. https://doi.org/10.1016/j.lwt.2014.05.004

Paiva, A.C., Hantao, L.W., 2020. Exploring a public database to evaluate consumer preference and aroma profile of lager beers by comprehensive two-dimensional gas chromatography and partial least squares regression discriminant analysis. J. Chromatogr. A 1630, 461529. https://doi.org/10.1016/j.chroma.2020.461529

Pandohee, J., Hughes, J.G., Pearson, J.R., A.H Jones, O., 2020. Chemical fingerprinting of petrochemicals for arson investigations using two-dimensional gas chromatography - flame ionisation detection and multivariate analysis. Sci. Justice 60, 381–387. https://doi.org/10.1016/j.scijus.2020.04.004

Parastar, H., Jalali-Heravi, M., Tauler, R., 2012. Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution. Chemom. Intell. Lab. Syst., 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, 80–91. https://doi.org/10.1016/j.chemolab.2012.02.003

Parastar, H., Radović, J.R., Bayona, J.M., Tauler, R., 2013. Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares. Anal. Bioanal. Chem. 405, 6235–6249. https://doi.org/10.1007/s00216-013-7067-y

Parastar, H., Radović, J.R., Jalali-Heravi, M., Diez, S., Bayona, J.M., Tauler, R., 2011. Resolution and Quantification of Complex Mixtures of Polycyclic Aromatic Hydrocarbons in Heavy Fuel Oil Sample by Means of GC × GC-TOFMS Combined to Multivariate Curve Resolution. Anal. Chem. 83, 9289–9297. https://doi.org/10.1021/ac201799r

Parente, E., Zotta, T., 2022. Chemometric Approaches for Identity and Authenticity Testing, Quality Assurance and Process Control☆, in: McSweeney, P.L.H., McNamara, J.P. (Eds.), Encyclopedia of Dairy Sciences (Third Edition). Academic Press, Oxford, pp. 327–347. https://doi.org/10.1016/B978-0-12-818766-1.00117-3

Park, M.S., Choi, J.Y., 2009. Theoretical analysis on feature extraction capability of class-augmented PCA. Pattern Recognit. 42, 2353–2362. https://doi.org/10.1016/j.patcog.2009.04.011

Parsons, B.A., Marney, L.C., Siegler, W.C., Hoggard, J.C., Wright, B.W., Synovec, R.E., 2015. Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-

of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach. Anal. Chem. 87, 3812–3819. https://doi.org/10.1021/ac504472s

Parsons, B.A., Pinkerton, D.K., Synovec, R.E., 2018. Implications of phase ratio for maximizing peak capacity in comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. J. Chromatogr. A, Comprehensive two-dimensional chromatography 1536, 16–26. https://doi.org/JChrom

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

Patel, S.P., Upadhyay, S.H., 2020. Euclidean distance based feature ranking and subset selection for bearing fault diagnosis. Expert Syst. Appl. 154, 113400. https://doi.org/10.1016/j.eswa.2020.113400

Paula Barros, E., Moreira, N., Elias Pereira, G., Leite, S.G.F., Moraes Rezende, C., Guedes de Pinho, P., 2012. Development and validation of automatic HS-SPME with a gas chromatography-ion trap/mass spectrometry method for analysis of volatiles in wines. Talanta 101, 177–186. https://doi.org/10.1016/j.talanta.2012.08.028

Pérez-Cova, M., Jaumot, J., Tauler, R., 2021. Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches. TrAC Trends Anal. Chem. 137, 116207. https://doi.org/10.1016/j.trac.2021.116207

Pierce, K.M., Hoggard, J.C., Hope, J.L., Rainey, P.M., Hoofnagle, A.N., Jack, R.M., Wright, B.W., Synovec, R.E., 2006. Fisher Ratio Method Applied to Third-Order Separation Data To Identify Significant Chemical Components of Metabolite Extracts. Anal. Chem. 78, 5068–5075. https://doi.org/10.1021/ac0602625

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

Pierce, K.M., Kehimkar, B., Marney, L.C., Hoggard, J.C., Synovec, R.E., 2012a. Review of chemometric analysis techniques for comprehensive two dimensional separations data. J. Chromatogr. A, Hyphenated and Multidimensional Chromatography Techniques 1255, 3–11. https://doi.org/10.1016/j.chroma.2012.05.050

Pierce, K.M., Nadeau, J.S., Synovec, R.E., 2012b. Chapter 17 - Data Analysis Methods, in: Poole, C.F. (Ed.), Gas Chromatography. Elsevier, Amsterdam, pp. 415–434. https://doi.org/10.1016/B978-0-12-385540-4.00017-1

Pierce, K.M., Parsons, B.A., Synovec, R.E., 2015. Chapter 10 - Pixel-Level Data Analysis Methods for Comprehensive Two-Dimensional Chromatography, in: de la Peña, A.M., Goicoechea, H.C., Escandar, G.M., Olivieri, A.C. (Eds.), Data Handling in Science and Technology, Fundamentals and Analytical Applications of Multiway Calibration. Elsevier, pp. 427–463. https://doi.org/10.1016/B978-0-444-63527-3.00010-2

Pierce, K.M., Schale, S.P., 2011. Predicting percent composition of blends of biodiesel and conventional diesel using gas chromatography–mass spectrometry, comprehensive two-dimensional gas chromatography–mass spectrometry, and partial least squares analysis. Talanta, Enhancing Chemical Separations with Chemometric Data Analysis 83, 1254–1259. https://doi.org/10.1016/j.talanta.2010.07.084

Pierce, K.M., Trinklein, T.J., Nadeau, J.S., Synovec, R.E., 2021. Chapter 20 - Data analysis methods for gas chromatography, in: Poole, C.F. (Ed.), Gas Chromatography (Second Edition), Handbooks in Separation Science. Elsevier, Amsterdam, pp. 525–546. https://doi.org/10.1016/B978-0-12-820675-1.00007-1

Pierce, K.M., Wright, B.W., Synovec, R.E., 2007. Unsupervised parameter optimization for automated retention time alignment of severely shifted gas chromatographic data using the piecewise alignment algorithm. J. Chromatogr. A 1141, 106–116. https://doi.org/10.1016/j.chroma.2006.11.101

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

Pinto, V.S., 2020. Use of 1H nuclear magnetic resonance and chemometrics to detect the percentage of ethanol anhydrous in Brazilian type C premium gasoline. Fuel 276, 118015. https://doi.org/10.1016/j.fuel.2020.118015

Pollo, B.J., Teixeira, C.A., Belinato, J.R., Furlan, M.F., Cunha, I.C. de M., Vaz, C.R., Volpato, G.V., Augusto, F., 2021. Chemometrics, Comprehensive Two-Dimensional gas chromatography and "omics" sciences: Basic tools and recent applications. TrAC Trends Anal. Chem. 134, 116111. https://doi.org/10.1016/j.trac.2020.116111

Poole, C.F., 2015. Ionization-based detectors for gas chromatography. J. Chromatogr. A, Instrumentation and Automation for the Separation Sciences 1421, 137–153. https://doi.org/10.1016/j.chroma.2015.02.061

Poole, C.F., Poole, S.K., 2008. Separation characteristics of wall-coated open-tubular columns for gas chromatography. J. Chromatogr. A, 50 Years Journal of Chromatography 1184, 254–280. https://doi.org/10.1016/j.chroma.2007.07.028

Prazen, B.J., Synovec, R.E., Kowalski, B.R., 1998. Standardization of Second-Order Chromatographic/Spectroscopic Data for Optimum Chemical Analysis. Anal. Chem. 70, 218–225. https://doi.org/10.1021/ac9706335

Prebihalo, S.E., Berrier, K.L., Freye, C.E., Bahaghighat, H.D., Moore, N.R., Pinkerton, D.K., Synovec, R.E., 2018. Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications. Anal. Chem. 90, 505–532. https://doi.org/10.1021/acs.analchem.7b04226

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

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

Rahman, Md.M., Bui, M.V., Shibata, M., Nakazawa, N., Rithu, Mst.N.A., Yamashita, H., Sadayasu, K., Tsuchiyama, K., Nakauchi, S., Hagiwara, T., Osako, K., Okazaki, E., 2021. Rapid noninvasive monitoring of freshness variation in frozen shrimp using multidimensional fluorescence imaging coupled with chemometrics. Talanta 224, 121871. https://doi.org/10.1016/j.talanta.2020.121871

Ranzan, L., Ranzan, C., Trierweiler, L.F., Trierweiler, J.O., 2017. Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy. Energy Fuels 31, 8942–8950. https://doi.org/10.1021/acs.energyfuels.7b00954

Rasheed, D.M., Serag, A., Abdel Shakour, Z.T., Farag, M., 2021. Novel trends and applications of multidimensional chromatography in the analysis of food, cosmetics and medicine bearing essential oils. Talanta 223, 121710. https://doi.org/10.1016/j.talanta.2020.121710

Rearden, P., Harrington, P.B., Karnes, J.J., Bunker, C.E., 2007. Fuzzy Rule-Building Expert System Classification of Fuel Using Solid-Phase Microextraction Two-Way Gas Chromatography Differential Mobility Spectrometric Data. Anal. Chem. 79, 1485–1491. https://doi.org/10.1021/ac060527f

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

Rebiere, H., Grange, Y., Deconinck, E., Courselle, P., Acevska, J., Brezovska, K., Maurin, J., Rundlöf, T., Portela, M.J., Olsen, L.S., Offerlé, C., Bertrand, M., 2021. European fingerprint study on omeprazole drug substances using a multi analytical approach and chemometrics as a tool for the discrimination of manufacturing sources. J. Pharm. Biomed. Anal. 114444. https://doi.org/10.1016/j.jpba.2021.114444

Reichenbach, S.E., Tian, X., Tao, Q., Ledford, E.B., Wu, Z., Fiehn, O., 2011. Informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography and high-resolution mass spectrometry (GCxGC–HRMS). Talanta, Enhancing Chemical Separations with Chemometric Data Analysis 83, 1279–1288. https://doi.org/10.1016/j.talanta.2010.09.057

Reichenbach, S.E., Zini, C.A., Nicolli, K.P., Welke, J.E., Cordero, C., Tao, Q., 2019. Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition. J. Chromatogr. A 1595, 158–167. https://doi.org/10.1016/j.chroma.2019.02.027

Rey-Stolle, F., Dudzik, D., Gonzalez-Riano, C., Fernández-García, M., Alonso-Herranz, V., Rojo, D., Barbas, C., García, A., 2021. Low and high resolution gas chromatography-mass spectrometry for untargeted metabolomics: A tutorial. Anal. Chim. Acta 339043. https://doi.org/10.1016/j.aca.2021.339043

Ríos-Reina, R., Camiña, J.M., Callejón, R.M., Azcarate, S.M., 2021. Spectralprint techniques for wine and vinegar characterization, authentication and quality control: Advances and projections. TrAC Trends Anal. Chem. 134, 116121. https://doi.org/10.1016/j.trac.2020.116121

Ríos-Reina, R., Elcoroaristizabal, S., Ocaña-González, J.A., García-González, D.L., Amigo, J.M., Callejón, R.M., 2017. Characterization and authentication of Spanish PDO wine vinegars using multidimensional fluorescence and chemometrics. Food Chem. 230, 108–116. https://doi.org/10.1016/j.foodchem.2017.02.118

Robards, K., Haddad, P.R., Jackson, P.E., 1994. Gas Chromatography, in: Principles and Practice of Modern Chromatographic Methods. Academic Press, San Diego, pp. 117–120.

Roberson, Z.R., Goodpaster, J.V., 2019. Preparation and characterization of micro-bore wall-coated open-tubular capillaries with low phase ratios for fast-gas chromatography–mass spectrometry: Application to ignitable liquids and fire debris. Sci. Justice. https://doi.org/10.1016/j.scijus.2019.06.009

Robinson, A.L., Boss, P.K., Heymann, H., Solomon, P.S., Trengove, R.D., 2011. 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. J. Chromatogr. A 1218, 504–517. https://doi.org/10.1016/j.chroma.2010.11.008
Rocha, S.M., Caldeira, M., Carrola, J., Santos, M., Cruz, N., Duarte, I.F., 2012. Exploring the human urine metabolomic potentialities by comprehensive two-dimensional gas chromatography coupled to time of flight mass spectrometry. J. Chromatogr. A 1252, 155–163. https://doi.org/10.1016/j.chroma.2012.06.067
Rohatgi, V.K., Saleh, A.K.M.E., 2015. An Introduction to Probability and Statistics. John Wiley & Sons.
Ruisánchez, I., Jiménez-Carvelo, A.M., Callao, M.P., 2021. ROC curves for the optimization of one-class model parameters. A case study: Authenticating extra virgin olive oil from a Catalan protected designation of origin. Talanta 222, 121564. https://doi.org/10.1016/j.talanta.2020.121564
Salinas, M., Zalacain, A., Pardo, F., Alonso, G.L., 2004. Stir Bar Sorptive Extraction Applied to Volatile Constituents Evolution during Vitis vinifera Ripening. J. Agric. Food Chem. 52, 4821–4827. https://doi.org/10.1021/jf040040c
Samanipour, S., Dimitriou-Christidis, P., Gros, J., Grange, A., Samuel Arey, J., 2015. Analyte quantification with comprehensive two-dimensional gas chromatography: Assessment of methods for baseline correction, peak delineation, and matrix effect elimination for real samples. J. Chromatogr. A 1375, 123–139. https://doi.org/10.1016/j.chroma.2014.11.049
Sancho, A., Ribeiro, J.C., Reis, M.S., Martins, F.G., 2022. Cluster analysis of crude oils with k-means based on their physicochemical properties. Comput. Chem. Eng. 157, 107633. https://doi.org/10.1016/j.compchemeng.2021.107633
Savareear, B., Escobar-Arnanz, J., Brokl, M., Saxton, M.J., Wright, C., Liu, C., Focant, J.-F., 2018. Comprehensive comparative compositional study of the vapour phase of cigarette mainstream tobacco smoke and tobacco heating product aerosol. J. Chromatogr. A 1581–1582, 105–115. https://doi.org/10.1016/j.chroma.2018.10.035
Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639.
Schoeman, J.C., du Preez, I., Loots, D.T., 2012. A comparison of four sputum pre-extraction preparation methods for identifying and characterising Mycobacterium tuberculosis using GCxGC-TOFMS metabolomics. J. Microbiol. Methods 91, 301–311. https://doi.org/10.1016/j.mimet.2012.09.002
Schomburg, G., Weeke, F., Müller, F., Oreans, M., 1982. Multidimensional gas chromatography (MDC) in capillary columns using double oven instruments and a newly designed coupling piece for monitoring detection after pre-separation. Chromatographia 16, 87–91. https://doi.org/10.1007/BF02258875
Schöneich, S., Gough, D.V., Trinklein, T.J., Synovec, R.E., 2020a. Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection. J. Chromatogr. A 1620, 460982. https://doi.org/10.1016/j.chroma.2020.460982
Schöneich, S., Trinklein, T.J., Warren, C.G., Synovec, R.E., 2020b. A systematic investigation of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry with dynamic pressure gradient modulation for high peak capacity separations. Anal. Chim. Acta 1134, 115–124. https://doi.org/10.1016/j.aca.2020.08.023
Sciarrone, D., Pantò, S., Rotondo, A., Tedone, L., Tranchida, P.Q., Dugo, P., Mondello, L., 2013. Rapid collection and identification of a novel component from Clausena lansium Skeels leaves by means of three-dimensional preparative gas chromatography and nuclear magnetic

resonance/infrared/mass spectrometric analysis. Anal. Chim. Acta 785, 119–125. https://doi.org/10.1016/j.aca.2013.04.069
Seeley, J.V., Schimmel, N.E., Seeley, S.K., 2017. The multi-mode modulator: A versatile fluidic device for two-dimensional gas chromatography. J. Chromatogr. A. https://doi.org/10.1016/j.chroma.2017.06.030
Seeley, J.V., Seeley, S.K., 2013. Multidimensional Gas Chromatography: Fundamental Advances and New Applications. Anal. Chem. 85, 557–578. https://doi.org/10.1021/ac303195u
Sharif, K.M., Chin, S.-T., Kulsing, C., Marriott, P.J., 2016. The microfluidic Deans switch: 50 years of progress, innovation and application. TrAC Trends Anal. Chem. 82, 35–54. https://doi.org/10.1016/j.trac.2016.05.005
Sharma, V., Kumar, R., 2018. Trends of chemometrics in bloodstain investigations. TrAC Trends Anal. Chem. 107, 181–195. https://doi.org/10.1016/j.trac.2018.08.006
Shi, Q., Yan, J., Jiang, B., Chi, X., Wang, J., Liang, X., Ai, X., 2021. A general strategy for the structural determination of carbohydrates by multi-dimensional NMR spectroscopies. Carbohydr. Polym. 267, 118218. https://doi.org/10.1016/j.carbpol.2021.118218
Siebert, K.J., 2011. Using Chemometrics To Classify Samples and Detect Misrepresentation, in: Progress in Authentication of Food and Wine, ACS Symposium Series. American Chemical Society, pp. 39–65. https://doi.org/10.1021/bk-2011-1081.ch004
Siegler, W.C., Crank, J.A., Armstrong, D.W., Synovec, R.E., 2010. Increasing selectivity in comprehensive three-dimensional gas chromatography via an ionic liquid stationary phase column in one dimension. J. Chromatogr. A 1217, 3144–3149. https://doi.org/10.1016/j.chroma.2010.02.082
Siegler, W.C., Fitz, B.D., Hoggard, J.C., Synovec, R.E., 2011. Experimental Study of the Quantitative Precision for Valve-Based Comprehensive Two-Dimensional Gas Chromatography. Anal. Chem. 83, 5190–5196. https://doi.org/10.1021/ac200302b
Sinkov, N.A., Harynuk, J.J., 2011. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. Talanta, Enhancing Chemical Separations with Chemometric Data Analysis 83, 1079–1087. https://doi.org/10.1016/j.talanta.2010.10.025
Stanislaus, A., Marafi, A., Rana, M.S., 2010. Recent advances in the science and technology of ultra low sulfur diesel (ULSD) production. Catal. Today, MONOGRAPH: Recent advances in the science and technology of ultra low sulfur diesel (ULSD) production 153, 1–68. https://doi.org/10.1016/j.cattod.2010.05.011
Stefanuto, P.-H., Perrault, K.A., Dubois, L.M., L'Homme, B., Allen, C., Loughnane, C., Ochiai, N., Focant, J.-F., 2017. Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry. J. Chromatogr. A 1507, 45–52. https://doi.org/10.1016/j.chroma.2017.05.064
Stefanuto, P.-H., Smolinska, A., Focant, J.-F., 2021. Advanced chemometric and data handling tools for GC×GC-TOF-MS: Application of chemometrics and related advanced data handling in chemical separations. TrAC Trends Anal. Chem. 139, 116251. https://doi.org/10.1016/j.trac.2021.116251
Stilo, F., Bicchi, C., Jimenez-Carvelo, A.M., Cuadros-Rodriguez, L., Reichenbach, S.E., Cordero, C., 2021a. Chromatographic fingerprinting by comprehensive two-dimensional chromatography: Fundamentals and tools. TrAC Trends Anal. Chem. 134, 116133. https://doi.org/10.1016/j.trac.2020.116133
Stilo, F., Bicchi, C., Robbat, A., Reichenbach, S.E., Cordero, C., 2021b. Untargeted approaches in food-omics: The potential of comprehensive two-dimensional gas chromatography/mass spectrometry. TrAC Trends Anal. Chem. 135, 116162. https://doi.org/10.1016/j.trac.2020.116162

Stilo, F., Liberto, E., Spigolon, N., Genova, G., Rosso, G., Fontana, M., Reichenbach, S.E., Bicchi, C., Cordero, C., 2021c. An effective chromatographic fingerprinting workflow based on comprehensive two-dimensional gas chromatography – Mass spectrometry to establish volatiles patterns discriminative of spoiled hazelnuts (Corylus avellana L.). Food Chem. 340, 128135. https://doi.org/10.1016/j.foodchem.2020.128135

Stilo, F., Segura borrego, M. del P., Bicchi, C., Battaglino, S., Callejón fernadez, R.M., Morales, M.L., Reichenbach, S.E., Mccurry, J., Peroni, D., Cordero, C., 2021d. Delineating the extra-virgin olive oil aroma blueprint by multiple headspace solid phase microextraction and differential-flow modulated comprehensive two-dimensional gas chromatography. J. Chromatogr. A 1650, 462232. https://doi.org/10.1016/j.chroma.2021.462232

Strife, R.J., Simms, J.R., Lacey, M.P., 1990. Combined capillary gas chromatography/ion trap mass spectrometry quantitative methods using labeled or unlabeled internal standards. J. Am. Soc. Mass Spectrom. 1, 265–271. https://doi.org/10.1016/1044-0305(90)85044-M

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

Sudol, P.E., Ochoa, G.S., Synovec, R.E., 2021. Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. J. Chromatogr. A 1644, 462092. https://doi.org/10.1016/j.chroma.2021.462092

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

Sun, X.-D., Wu, H.-L., Liu, Z., Chen, Y., Chen, J.-C., Cheng, L., Ding, Y.-J., Yu, R.-Q., 2019. Target-based metabolomics for fast and sensitive quantification of eight small molecules in human urine using HPLC-DAD and chemometrics tools resolving of highly overlapping peaks. Talanta 201, 174–184. https://doi.org/10.1016/j.talanta.2019.03.090

The Good Scents Company - Flavor, Fragrance, Food and Cosmetics Ingredients information [WWW Document], n.d. . Good Scents Co. URL http://www.thegoodscentscompany.com/ (accessed 8.23.21).

Timko, M.T., Schmois, E., Patwardhan, P., Kida, Y., Class, C.A., Green, W.H., Nelson, R.K., Reddy, C.M., 2014. Response of Different Types of Sulfur Compounds to Oxidative Desulfurization of Jet Fuel. Energy Fuels 28, 2977–2983. https://doi.org/10.1021/ef500216p

Trinklein, T.J., Gough, D.V., Warren, C.G., Ochoa, G.S., Synovec, R.E., 2019. Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography. J. Chromatogr. A 460488. https://doi.org/10.1016/j.chroma.2019.460488

Trinklein, T.J., Prebihalo, S.E., Warren, C.G., Ochoa, G.S., Synovec, R.E., 2020a. Discovery-based analysis and quantification for comprehensive three-dimensional gas chromatography flame ionization detection data. J. Chromatogr. A 1623, 461190. https://doi.org/10.1016/j.chroma.2020.461190

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

Trinklein, T.J., Warren, C.G., Synovec, R.E., 2021. Determination of the Signal-To-Noise Ratio Enhancement in Comprehensive Three-Dimensional Gas Chromatography. Anal. Chem. 93, 8526–8535. https://doi.org/10.1021/acs.analchem.1c01190

Tukey, J.W., 1991. The Philosophy of Multiple Comparisons. Stat. Sci. 6, 100–116.

Ugena, L., Moncayo, S., Manzoor, S., Rosales, D., Cáceres, J.O., 2016. Identification and Discrimination of Brands of Fuels by Gas Chromatography and Neural Networks Algorithm in Forensic Research. J. Anal. Methods Chem. 2016. https://doi.org/10.1155/2016/6758281

US EPA, O., 2015. Diesel Fuel Standards and Rulemakings [WWW Document]. US EPA. URL https://www.epa.gov/diesel-fuel-standards/diesel-fuel-standards-and-rulemakings (accessed 10.1.20).

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7, 142. https://doi.org/10.1186/1471-2164-7-142

Vaz-Freire, L.T., da Silva, M.D.R.G., Freitas, A.M.C., 2009. Comprehensive two-dimensional gas chromatography for fingerprint pattern recognition in olive oils produced by two different techniques in Portuguese olive varieties Galega Vulgar, Cobrançosa e Carrasquenha. Anal. Chim. Acta 633, 263–270. https://doi.org/10.1016/j.aca.2008.11.057

Vendeuvre, C., Bertoncini, F., Duval, L., Duplan, J.-L., Thiébaut, D., Hennion, M.-C., 2004. Comparison of conventional gas chromatography and comprehensive two-dimensional gas chromatography for the detailed analysis of petrochemical samples. J. Chromatogr. A, 8th International Symposium on Hyphenated Techniques in Chromatography and Hyphenated Chromatographic Analyzers 1056, 155–162. https://doi.org/10.1016/j.chroma.2004.05.071

vol3d v2 [WWW Document], n.d. URL https://www.mathworks.com/matlabcentral/fileexchange/22940-vol3d-v2 (accessed 1.7.22).

Vozka, P., Kilaz, G., 2019. How to obtain a detailed chemical composition for middle distillates via GC × GC-FID without the need of GC × GC-TOF/MS. Fuel 247, 368–377. https://doi.org/10.1016/j.fuel.2019.03.009

Vrtiška, D., Vozka, P., Váchová, V., Šimáček, P., Kilaz, G., 2019. Prediction of HEFA content in jet fuel using FTIR and chemometric methods. Fuel 236, 1458–1464. https://doi.org/10.1016/j.fuel.2018.09.102

Vrzal, T., Olšovská, J., 2019. Pyrolytic profiling nitrosamine specific chemiluminescence detection combined with multivariate chemometric discrimination for non-targeted detection and classification of nitroso compounds in complex samples. Anal. Chim. Acta 1059, 136–145. https://doi.org/10.1016/j.aca.2019.01.033

Wang, X., Liu, X., Wang, J., Wang, G., Zhang, Y., Lan, L., Sun, G., 2021. Study on multiple fingerprint profiles control and quantitative analysis of multi-components by single marker method combined with chemometrics based on Yankening tablets. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 253, 119554. https://doi.org/10.1016/j.saa.2021.119554

Watson, N.E., Bahaghighat, H.D., Cui, K., Synovec, R.E., 2017a. Comprehensive Three-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry. Anal. Chem. 89, 1793–1800. https://doi.org/10.1021/acs.analchem.6b04112

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

Watson, N.E., Prebihalo, S.E., Synovec, R.E., 2017b. Targeted analyte deconvolution and identification by four-way parallel factor analysis using three-dimensional gas chromatography with mass spectrometry data. Anal. Chim. Acta 983, 67–75. https://doi.org/10.1016/j.aca.2017.06.017

Watson, N.E., Siegler, W.C., Hoggard, J.C., Synovec, R.E., 2007. Comprehensive Three-Dimensional Gas Chromatography with Parallel Factor Analysis. Anal. Chem. 79, 8270–8280. https://doi.org/10.1021/ac070829x

Webster, R.L., Rawson, P.M., Evans, D.J., Marriott, P.J., 2016. Quantification of trace fatty acid methyl esters in diesel fuel by using multidimensional gas chromatography with electron and chemical ionization mass spectrometry. J. Sep. Sci. 39, 2537–2543. https://doi.org/10.1002/jssc.201600307

Weinert, C.H., Egert, B., Kulling, S.E., 2015. On the applicability of comprehensive two-dimensional gas chromatography combined with a fast-scanning quadrupole mass spectrometer for untargeted large-scale metabolomics. J. Chromatogr. A 1405, 156–167. https://doi.org/10.1016/j.chroma.2015.04.011

Weldegergis, B.T., Villiers, A. de, McNeish, C., Seethapathy, S., Mostafa, A., Górecki, T., Crouch, A.M., 2011. Characterisation of volatile components of Pinotage wines using comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC×GC–TOFMS). Food Chem. 129, 188–199. https://doi.org/10.1016/j.foodchem.2010.11.157

Welke, J.E., Manfroi, V., Zanus, M., Lazzarotto, M., Alcaraz Zini, C., 2013. Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data. Food Chem. 141, 3897–3905. https://doi.org/10.1016/j.foodchem.2013.06.100

Welke, J.E., Zanus, M., Lazzarotto, M., Alcaraz Zini, C., 2014a. Quantitative analysis of headspace volatile compounds using comprehensive two-dimensional gas chromatography and their contribution to the aroma of Chardonnay wine. Food Res. Int. 59, 85–99. https://doi.org/10.1016/j.foodres.2014.02.002

Welke, J.E., Zanus, M., Lazzarotto, M., Pulgati, F.H., Zini, C.A., 2014b. Main differences between volatiles of sparkling and base wines accessed through comprehensive two dimensional gas chromatography with time-of-flight mass spectrometric detection and chemometric tools. Food Chem. 164, 427–437. https://doi.org/10.1016/j.foodchem.2014.05.025

Wiebelhaus, N., Hamblin, D., Kreitals, N.M., Almirall, J.R., 2016. Differentiation of marijuana headspace volatiles from other plants and hemp products using capillary microextraction of volatiles (CMV) coupled to gas-chromatography–mass spectrometry (GC–MS). Forensic Chem. 2, 1–8. https://doi.org/10.1016/j.forc.2016.08.004

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemom. Intell. Lab. Syst., Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists 2, 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

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

Xiao, B., Li, Y., Sun, B., Yang, C., Huang, K., Zhu, H., 2021. Decentralized PCA modeling based on relevance and redundancy variable selection and its application to large-scale dynamic process monitoring. Process Saf. Environ. Prot. 151, 85–100. https://doi.org/10.1016/j.psep.2021.04.043

Yamamoto, H., Yamaji, H., Abe, Y., Harada, K., Waluyo, D., Fukusaki, E., Kondo, A., Ohno, H., Fukuda, H., 2009. Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. Chemom. Intell. Lab. Syst. 98, 136–142. https://doi.org/10.1016/j.chemolab.2009.05.006

Yan, D., Wong, Y.F., Whittock, S.P., Koutoulis, A., Shellie, R.A., Marriott, P.J., 2018. Sequential Hybrid Three-Dimensional Gas Chromatography with Accurate Mass Spectrometry: A Novel Tool for High-Resolution Characterization of Multicomponent Samples. Anal. Chem. 90, 5264–5271. https://doi.org/10.1021/acs.analchem.8b00142

Yang, B., Li, L., Geng, H., Zhang, C., Wang, G., Yang, S., Gao, S., Zhao, Y., Xing, F., 2021. Inhibitory effect of allyl and benzyl isothiocyanates on ochratoxin a producing fungi in grape and maize. Food Microbiol. 100, 103865. https://doi.org/10.1016/j.fm.2021.103865

Yang, Q., Xu, L., Tang, L.-J., Yang, J.-T., Wu, B.-Q., Chen, N., Jiang, J.-H., Yu, R.-Q., 2018. Simultaneous detection of multiple inherited metabolic diseases using GC-MS urinary metabolomics by chemometrics multi-class classification strategies. Talanta 186, 489–496. https://doi.org/10.1016/j.talanta.2018.04.081

Yang, S., Sadilek, M., Synovec, R.E., Lidstrom, M.E., 2009. 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, 3280–3289. https://doi.org/10.1016/j.chroma.2009.02.030

Ye, Z., Shang, Z., Li, M., Qu, Y., Long, H., Yi, J., 2020. Evaluation of the physiochemical and aromatic qualities of pickled Chinese pepper (Paojiao) and their influence on consumer acceptability by using targeted and untargeted multivariate approaches. Food Res. Int. 137, 109535. https://doi.org/10.1016/j.foodres.2020.109535

Yin, J., Xie, J., Guo, X., Ju, L., Li, Y., Zhang, Y., 2016. Plasma metabolic profiling analysis of cyclophosphamide-induced cardiotoxicity using metabolomics coupled with UPLC/Q□TOF□ MS and ROC curve. J. Chromatogr. B 1033–1034, 428–435. https://doi.org/10.1016/j.jchromb.2016.08.042

Zeng, Z.-D., Chin, S.-T., Hugel, H.M., Marriott, P.J., 2011. Simultaneous deconvolution and re-construction of primary and secondary overlapping peak clusters in comprehensive two-dimensional gas chromatography. J. Chromatogr. A 1218, 2301–2310. https://doi.org/10.1016/j.chroma.2011.02.028

Zhang, Y.-Y., Zhang, Q., Zhang, Y.-M., Wang, W.-W., Zhang, L., Yu, Y.-J., Bai, C.-C., Guo, J.-Z., Fu, H.-Y., She, Y., 2020. A comprehensive automatic data analysis strategy for gas chromatography-mass spectrometry based untargeted metabolomics. J. Chromatogr. A 1616, 460787. https://doi.org/10.1016/j.chroma.2019.460787

Zhao, J., Wu, H.-L., Niu, J.-F., Yu, Y.-J., Yu, L.-L., Kang, C., Li, Q., Zhang, X.-H., Yu, R.-Q., 2012. Chemometric resolution of coeluting peaks of eleven antihypertensives from multiple classes in high performance liquid chromatography: A comprehensive research in human serum, health product and Chinese patent medicine samples. J. Chromatogr. B 902, 96–107. https://doi.org/10.1016/j.jchromb.2012.06.032

Zhu, C., Idemudia, C.U., Feng, W., 2019. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Inform. Med. Unlocked 17, 100179. https://doi.org/10.1016/j.imu.2019.100179

APPENDIX A

This Appendix is reproduced from the Electronic Supplementary Content of Paige E. Sudol, Derrick V. Gough, Sarah E. Prebihalo, Robert E. Synovec, "Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis" *Talanta* 206 (2020) 120239.

Subtraction plots

Subtraction plots were prepared to highlight distinguishing chemical features of the five fuels. These plots were generated by subtracting the representative GC×GC chromatograms (number of modulations along ^{1}D × number of data points along ^{2}D) of two fuels from one another. The color scale was adjusted so that blue indicated positive signal whereas red indicated negative signal. Hence if the chromatogram for Fuel 2 was subtracted from that of Fuel 1, blue would represent analyte peaks more prevalent in Fuel 1 and red would represent analyte peaks more prevalent in Fuel 2. Thus, subtraction plots provide complementary information to loadings in Figs. 2.6(B-C) and 2.7(B-C).

To glean as much chemical information from the dataset as possible, two sets of subtraction plots were prepared, using Fuel 1 and Fuel 2, respectively, as positive (blue) reference points. Since Fuel 1 has the most positive PC2 scores, correlations between the PC2 axis and the GC×GC axes can be assessed when this fuel is used as a reference point. Similarly, Fuel 2 has the most positive PC1 scores, so correlations between the PC1 axis and the GC×GC axes can be determined when Fuel 2 is used as a reference point. Utilizing defined reference points simplifies interpretation of the subtractions plots as well.

For ease of visual comparison, all subtraction plots were prepared using the SOP-binned chromatograms (Fig. 2.5). When Fuel 1 is used as a positive (blue) reference point, the resulting 2D plots have predominantly blue bins at the start of the ^{1}D separation and red bins at the end of

the separation (Fig. A.1(A)-(D)). Hence Fuel 1 appears to be characterized by a high concentration of low boiling point compounds, and this is reflected in its highly positive PC2 scores. Therefore these results further support the presence of a correlation between analyte boiling point and PC2, which was originally deduced using the PC2 loadings (Figs. 2.6(C) and 2.7(C)).

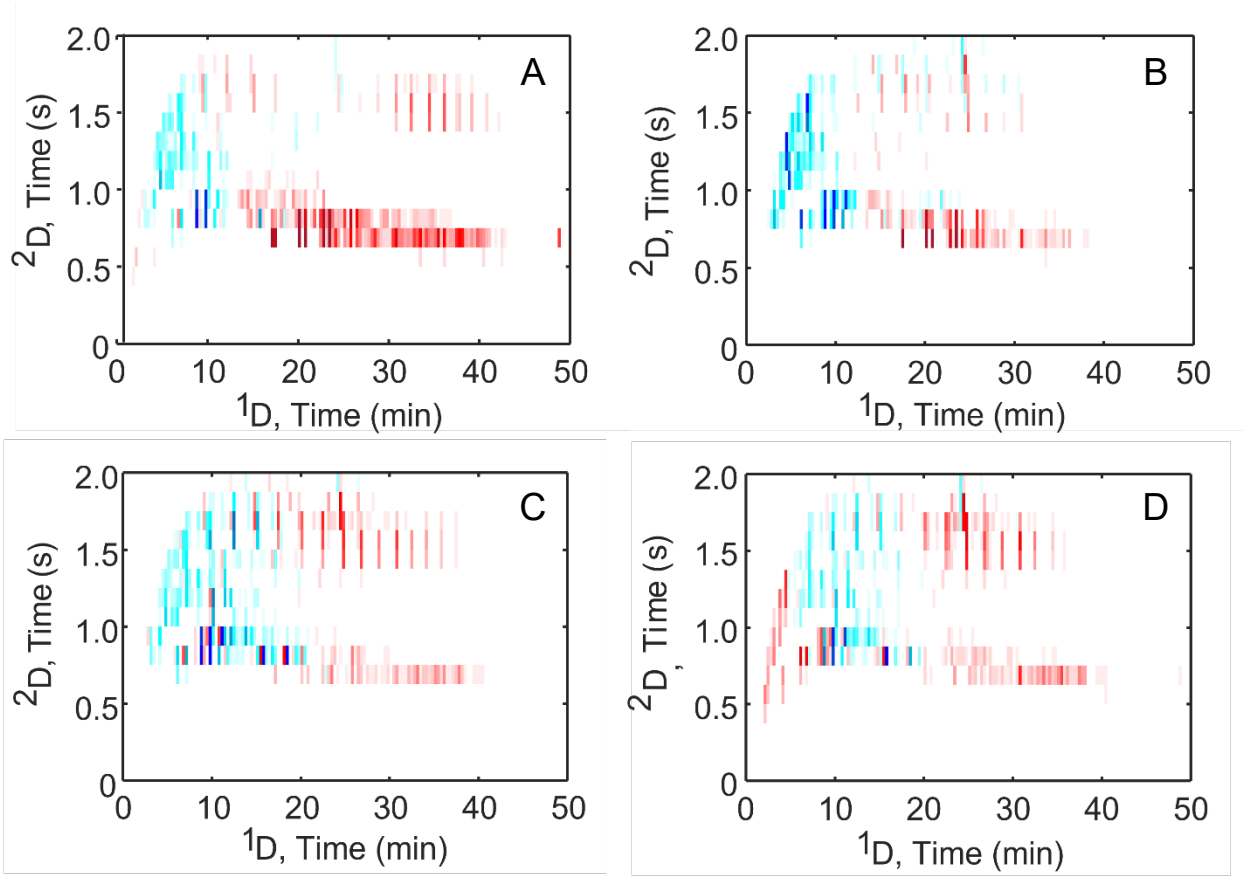

Figure A.1. Subtraction plots (binned to the SOP bin size) using Fuel 1 as a reference point (positive). This fuel has the most positive PC2 scores, so correlations between the PC2 axis and the GC×GC axes can be assessed. (A) Fuel 1 – Fuel 2. (B) Fuel 1 – Fuel 3. (C) Fuel 1 – Fuel 4. (D) Fuel 1 – Fuel 5.

When Fuel 2 is used as a positive reference point, the aromatic band is mostly blue after a ^{1}D retention time of 10 min in Fig. A.2(A-D), which suggests that Fuel 2 is characterized by a higher concentration of aromatic compounds. However, this is not reflected in its highly positive PC1 scores, since the aromatic band is blue regardless of which other fuel (positive PC1 scores or negative PC1 scores) is compared to Fuel 2 in Fig. A.2(A-D). Hence there is no clear correlation between PC1 and ^{2}D, supporting the conclusions drawn using the PC1 loadings plots (Figs. 2.6(B) and 2.7(B)).

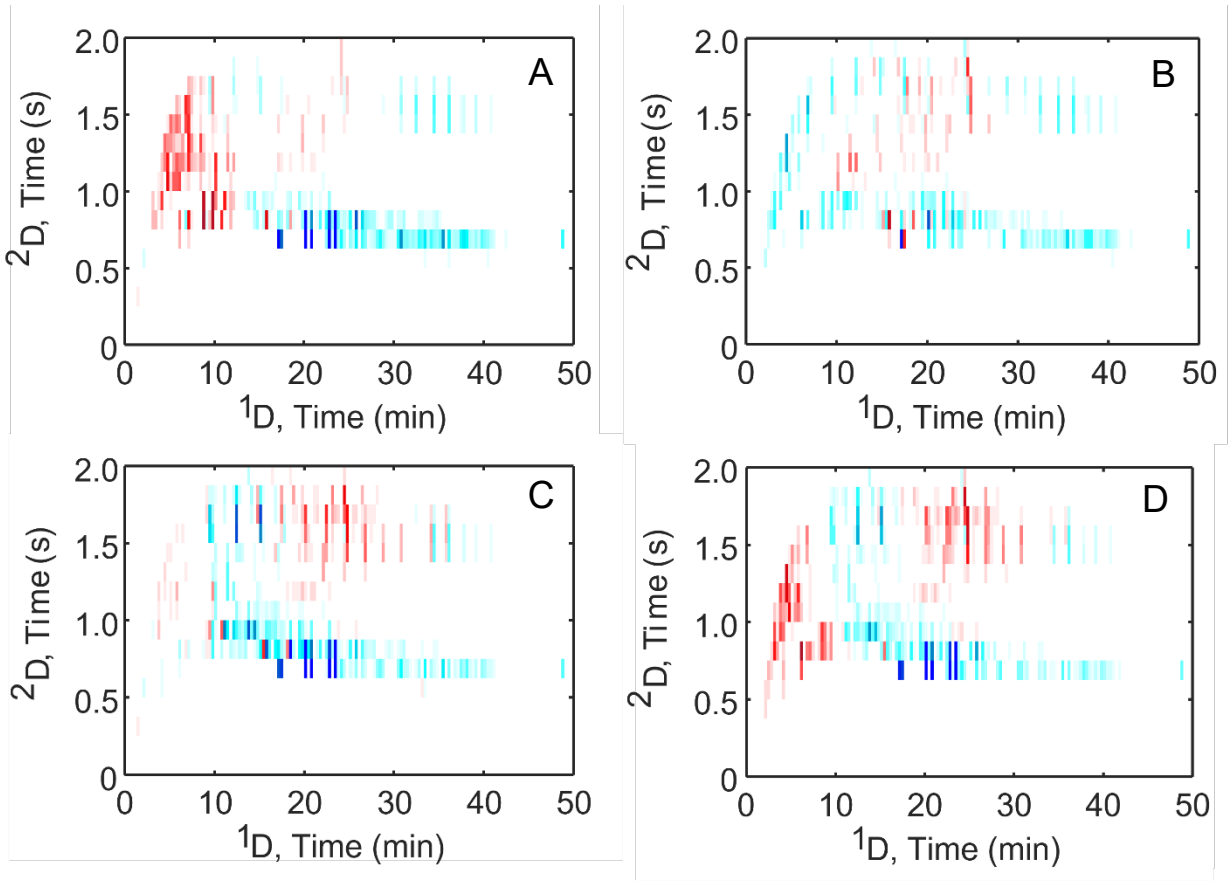

Figure A.2. Subtraction plots (binned to the SOP bin size) using Fuel 2 as a reference point (positive). This fuel has the most positive PC1 scores, so correlations between the PC1 axis and the GC×GC axes can be assessed. (A) Fuel 2 – Fuel 1. (B) Fuel 2 – Fuel 3. (C) Fuel 2 – Fuel 4. (D) Fuel 2 – Fuel 5.

Leave-one-sample class-out study

A leave-one-sample class-out study was performed to provide support of the conclusions, namely that the optimal level of binning prior to PCA depends on the unique chemical differences between two sample classes and thus a single bin size, eg, the SOP bin size, is not optimal for all (or any) of the fuel pairs. Fuels 1-5 were individually excluded from the 110 binned datasets prior to PCA, and DCS was recalculated for the fuel pairs listed in Table 2, except for the fuel pairs involving an excluded fuel. For example, when Fuel 1 was excluded, only fuel pairs B, C and D were compared. No DCS was calculated between Fuels 2 and 5, as there was no comparison of these two fuels initially.

In this round robin fashion, the results when Fuels 1-5 were excluded are shown in Figs. A.3-A.7, respectively, with a summary provided in Fig. A.8. The new scores plots at the SOP bin size are provided in Figs. A.3(A)-A.7(A) for comparison to Fig. 2.7(A). When a given fuel is removed, the remaining fuels maintain their positions in the scores plot with relatively minor shifting along PC1 and PC2. For each excluded fuel, a grid scheme of bin sizes following the heat map format of Fig. 2.2 is provided in Figs. A.3(B)-A.7(B). The original optimal bin sizes for fuels pairs A-E are labeled A-E without superscripts, while the optimal bin sizes when Fuels 1-5 are excluded are labeled A-E with superscripts 1-5, respectively. For example, the optimal bin size for fuel pair A with Fuel 4 omitted is labeled A^4, but without omitting Fuel 4 the optimal bin size is labeled A. Most of the new optimal bin sizes (with leave-on-out) are similar to the original optimal bin sizes (all five fuels), which suggests that the relative amount of ^{1}D versus ^{2}D binning that the fuel pairs can tolerate remains roughly the same (Fig. A.8). Two noteworthy exceptions are the optimal bin sizes for A^4 and E^4, which appear to be significantly different from A and E, respectively. However, when the heat maps for A^4 and E^4 (excluded here for brevity) are

compared to those for A and E (Figs. 2.8(A) and 2.8(E)), the trends in DCS with bin size are preserved. The original optimal bin sizes still produce high DCS values for fuel pairs A and E when Fuel 4 is excluded. These results lend support to the conclusion that DCS between two given sample classes in this study was minimally influenced by other classes in the PCA scores plot, and maximizing DCS relied upon applying an optimal bin size for a given pair of sample classes.

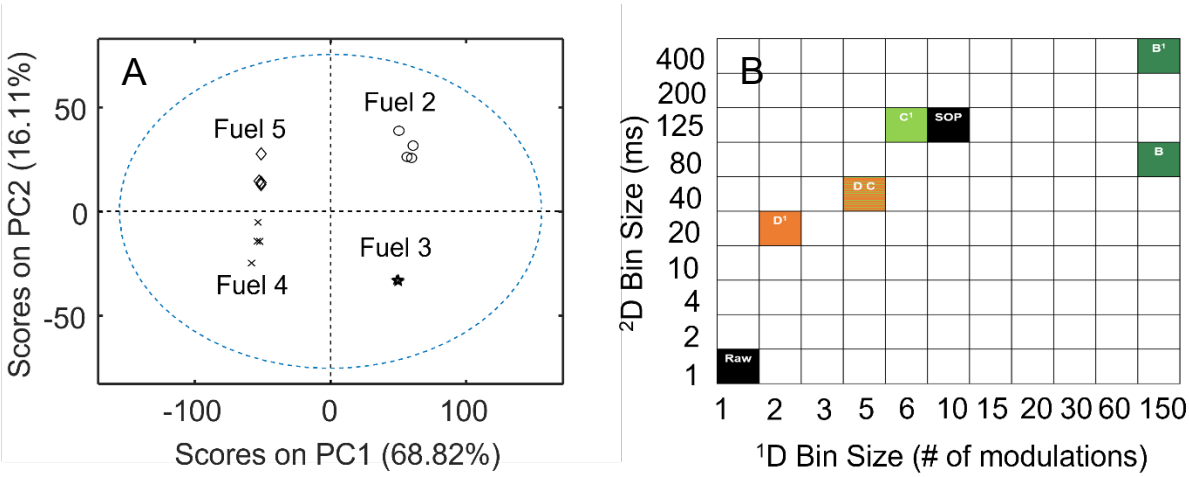

Figure A.3. (A) Scores plot of the SOP-binned data excluding Fuel 1. (B) Grid scheme showing the original optimal bin sizes (without superscripts) compared to the new optimal bin sizes when Fuel 1 is excluded from the PCA model (superscripts of 1 to denote the omission of Fuel 1). The color of the grids is representative of the magnitude of the highest DCS value obtained for each respective fuel pair (see Fig. 2.8 color bar).

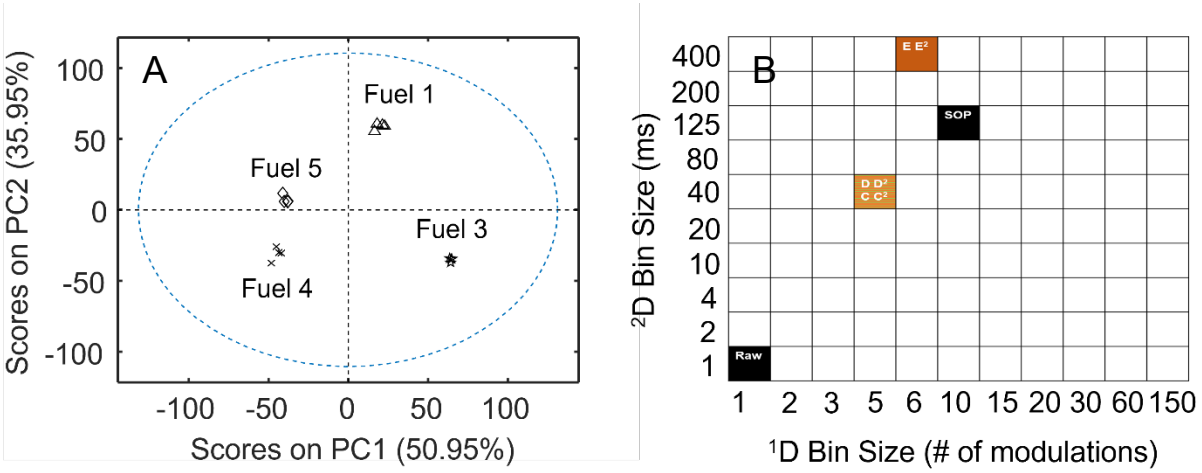

Figure A.4. (A) Scores plot of the SOP-binned data excluding Fuel 2. (B) Grid scheme showing the original optimal bin sizes (without superscripts) for fuel pairs C, D, and E compared to the optimal bin sizes when Fuel 2 is excluded from the PCA model (superscripts of 2).

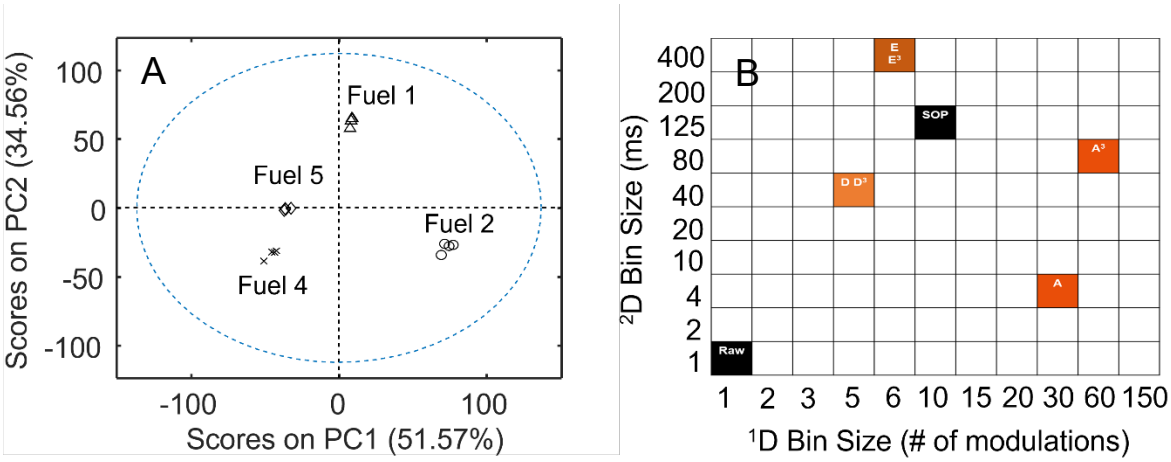

Figure A.5. (A) Scores plot of the SOP-binned data excluding Fuel 3. (B) Grid scheme showing the original optimal bin sizes (without superscripts) for fuel pairs A, D, and E compared to the optimal bin sizes when Fuel 3 is excluded from the PCA model (superscripts of 3).

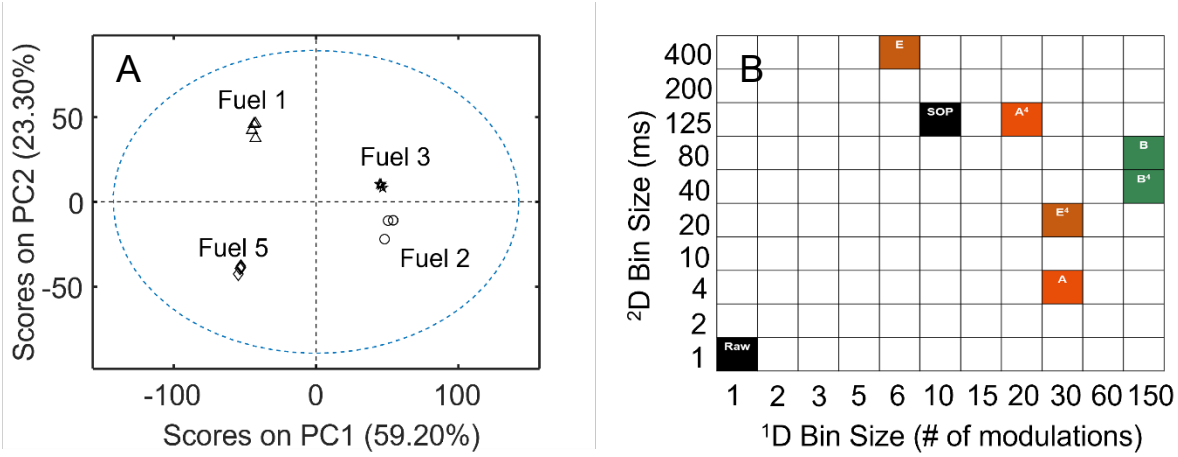

Figure A.6. (A) Scores plot of the SOP-binned data excluding Fuel 4. (B) Grid scheme showing the original optimal bin sizes (without superscripts) for fuel pairs A, B, and E compared to the optimal bin sizes when Fuel 4 is excluded from the PCA model (superscripts of 4).

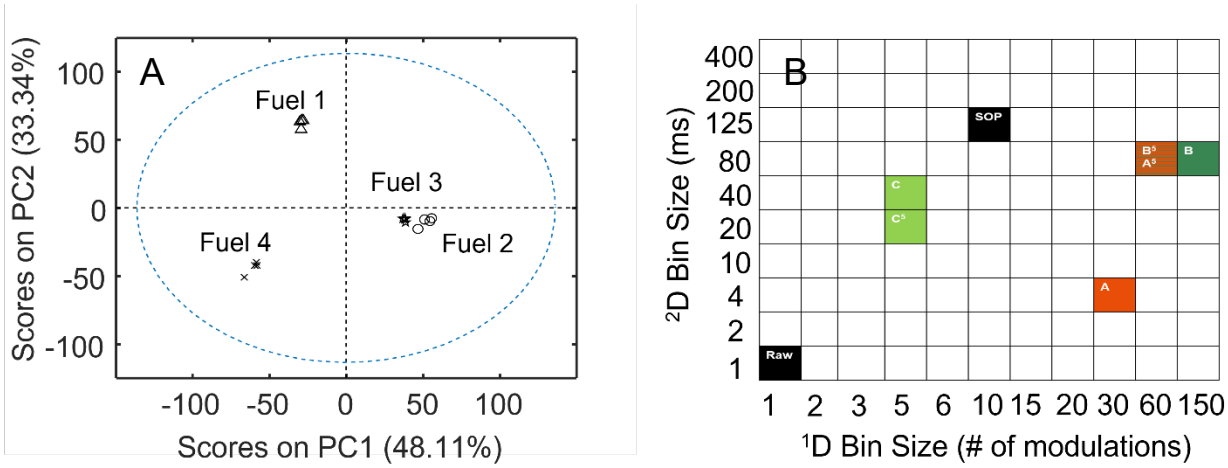

Figure A.7. (A) Scores plot of the SOP-binned data excluding Fuel 5. (B) Grid scheme showing the original optimal bin sizes (without superscripts) for fuel pairs A, B, and C compared to the optimal bin sizes when Fuel 5 is excluded from the PCA model (superscripts of 5).

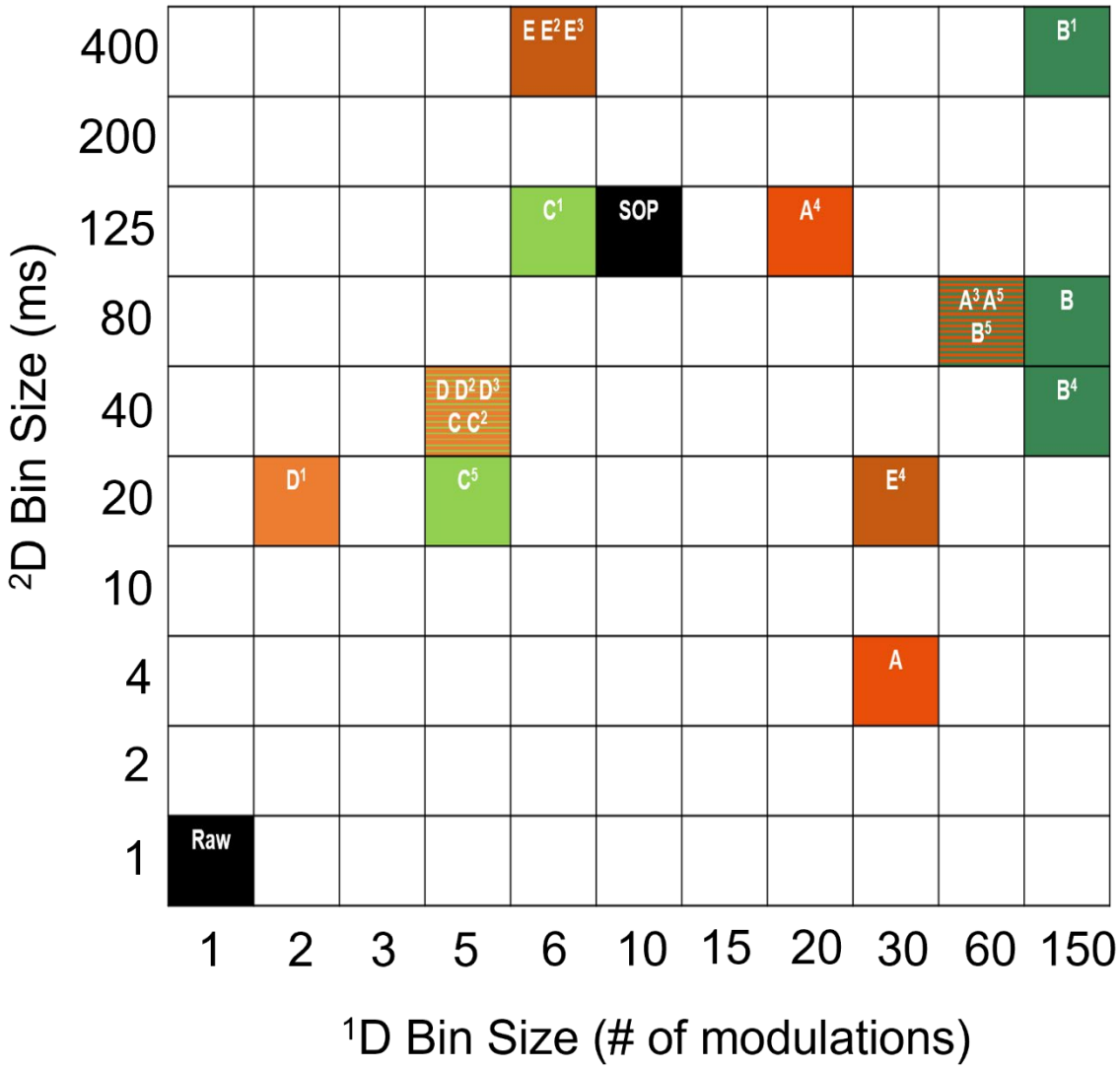

Figure A.8. Summary grid scheme showing the original bin optimal bin sizes (without superscripts) for fuel pairs A-E compared to the optimal bin sizes when Fuels 1-5 are excluded from the PCA model (superscripts of 1-5, respectively).

APPENDIX B

This Appendix is reproduced from the Electronic Supplementary Content of Paige E. Sudol, Grant S. Ochoa, Robert E. Synovec, "Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry" *Journal of Chromatography A* 1644 (2021) 462092.

Table B.1. Prepared concentrations (mg/kg) of sulfur-containing compounds added to JP8 jet fuel in nominal 30 ppm, 15 ppm, 3 ppm, and 1.5 ppm spikes based on mass.

Analyte	Concentration (mg/kg) in nominal 30 ppm samples	Concentration (mg/kg) in nominal 15 ppm samples	Concentration (mg/kg) in nominal 3 ppm samples	Concentration (mg/kg) in nominal 1.5 ppm samples
2-methylthiophene	28.98	14.43	2.89	1.46
3-methylthiophene	27.28	13.58	2.72	1.38
2-propylthiophene	29.68	14.77	2.96	1.50
benzo[b]thiophene	29.99	14.93	2.99	1.51
thiophene	28.05	13.96	2.80	1.42
1,4-oxathiane	29.87	14.87	2.98	1.51
2,5-dimethylthiophene	29.05	14.46	2.90	1.47
3-methylbenzothiophene	29.61	14.74	2.95	1.49
2-butyl-5-ethylthiophene	26.34	13.11	2.63	1.33
tetrahydrothiophene	28.01	13.94	2.79	1.41
2-methylbenzothiophene	29.89	14.87	2.98	1.51
2-chloroethylphenylsulfide	29.25	14.56	2.92	1.48
2-hexylthiophene	28.20	14.04	2.81	1.42
3-acetyl-2,5-dimethylthiophene	30.10	14.98	3.00	1.52

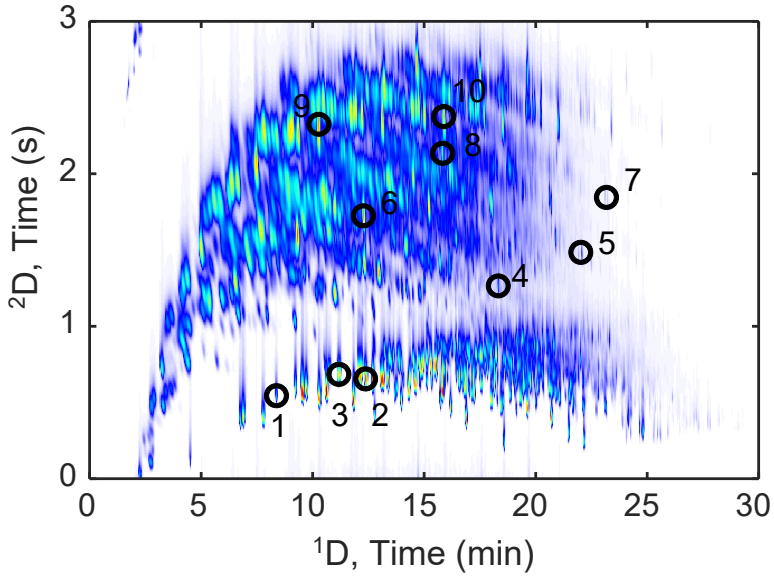

Figure B.1. Individual 15 ppm spiked JP8 jet fuel chromatogram at the total ion current (TIC) signal, with the locations of the ten native fuel peaks used for tile size selection circled in black. The numbers shown correspond to their identities listed in Table B.2 below.

Table B.2. List of ten native fuel peaks used for tile size selection, along with their ^{1}D retention times (^1t_R), ^{2}D retention times (^2t_R), and measured ^{2}D peak widths (^2W_b). The ^2W_b range from 165 to 303 ms, with ^2W_b increasing as a function of ^2k'. To capture the widest peaks, 300 ms was selected as the "optimal" ^{2}D tile dimension. A constant ^1W_b of ~12 s (4 modulations) was observed, so this was selected as the "optimal" ^{1}D tile dimension.

#	Name	^1t_R (min)	^2t_R (s)	^2W_b (ms)
1	1-methylethylbenzene	8.40	0.55	230
2	1-ethyl-3,5-dimethylbenzene	12.40	0.66	220
3	1-methyl-3-(1-methylethyl)-benzene	11.20	0.69	217
4	1,1'-bicyclohexyl	18.35	1.27	165
5	decahydro-4,4,8,9,10-pentamethylnaphthalene	22.05	1.49	170
6	pentylcyclohexane	12.30	1.73	257
7	n-nonylcyclohexane	23.20	1.85	245
8	7-tetradecene	15.85	2.14	253
9	undecane	10.30	2.33	303
10	tridecane	15.90	2.38	283

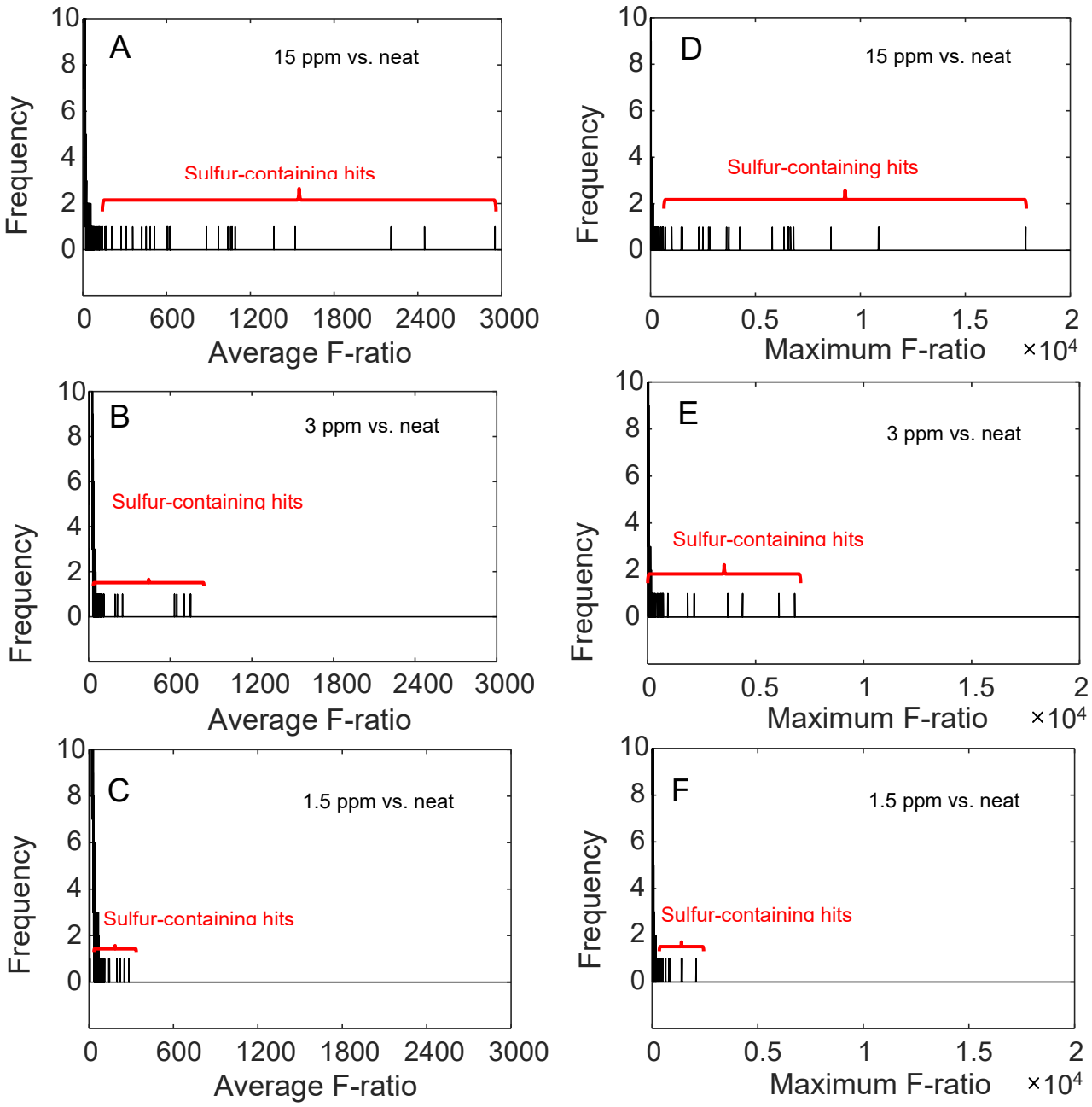

Figure B.2. F-ratio distributions (bin size of 0.2) ranked using standard F-ratio methodology for the (A) 15 ppm versus neat, (B) 3 ppm versus neat, and (C) 1.5 ppm versus neat F-ratio comparisons, compared to the F-ratio distributions ranked using the top F-ratio *m/z* for the (D) 15 ppm versus neat, (E) 3 ppm versus neat, and (F) 1.5 ppm versus neat F-ratio comparisons. The x-axis scales in (A), (B), and (C), are identical to each other, but ~6.7-fold less the x-axis scales in (D), (E) and (F), which are identical to each other, to highlight how F-ratios decrease as concentration decreases in this concentration regime. The magnitude change going from the average F-ratios (A,B,C, maximum of ~3,000) to the top F-ratios (D,E,F, maximum of ~20,000) highlights the diminutive effect of averaging, particularly for analytes at lower concentration with fewer sensitive *m/z*.

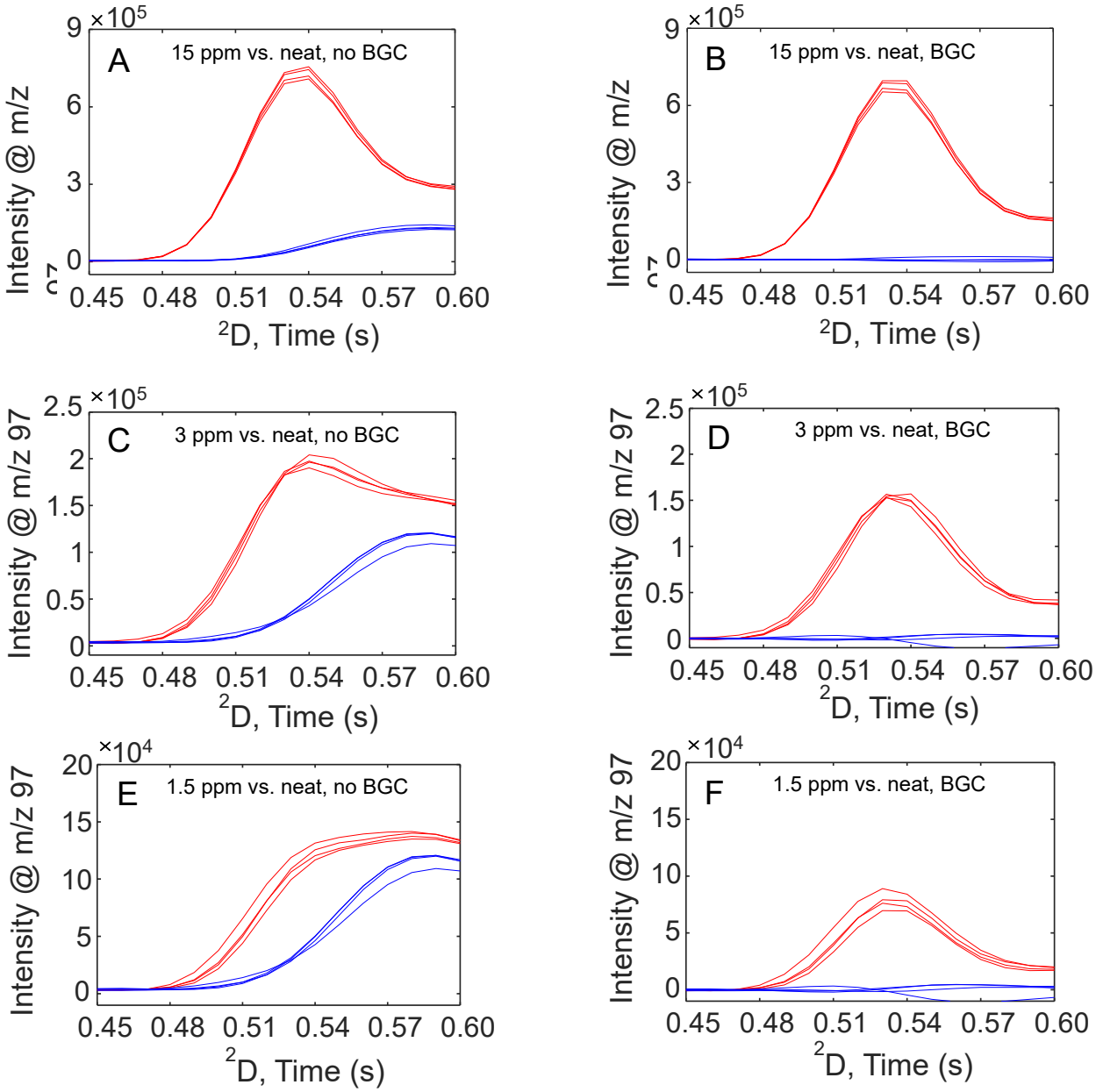

Figure B.3. Summed ^{2}D peaks for the replicates of 2-propylthiophene: spiked (red) and neat (blue). A ^{2}D window of 150 ms was used to prepare all summed ^{2}D peaks for this analyte. (A) Original summed ^{2}D peaks for the 15 ppm (red) and neat (blue). (B) Background corrected (BGC) summed ^{2}D peaks for the 15 ppm (red) and neat (blue). (C) Original summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (D) BGC summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (E) Original summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). (F) BGC summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). At 1.5 ppm, the average BGC peak height of 2-propylthiophene greatly exceeds its *LOD* (0.19 ppm) by 8-fold, but only exceeds its *LOQ* (0.64 ppm) by approximately 2-fold, which is why it suffers a decrease in hit number from the 3 ppm versus neat to the 1.5 ppm versus neat F-ratio comparison (hit 1 to hit 17 in Table 3.3).

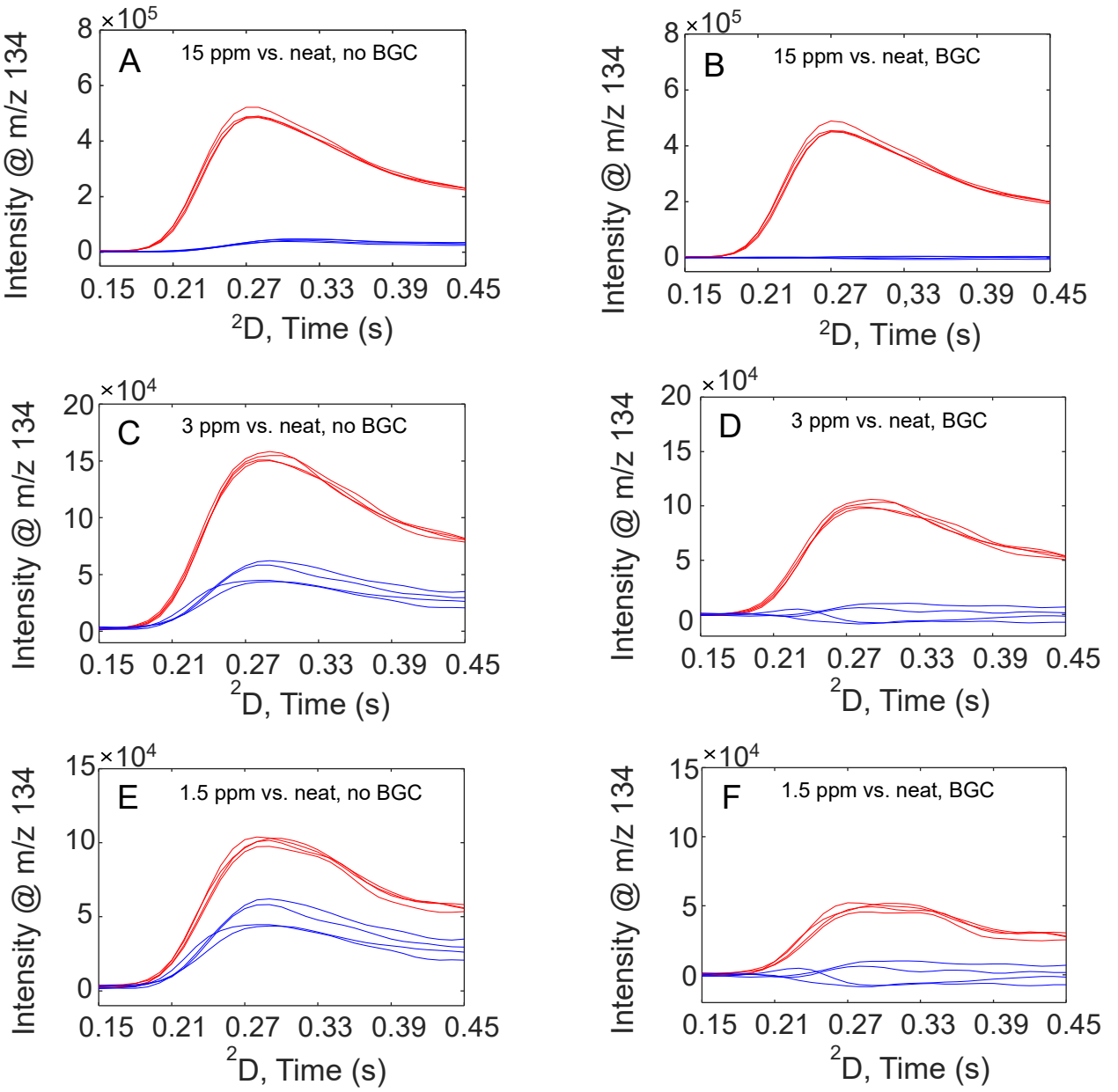

Figure B.4. Summed ^{2}D peaks for the replicates of benzo[b]thiophene: spiked (red) and neat (blue). A ^{2}D window of 300 ms was used to prepare all summed ^{2}D peaks for this analyte. (A) Original summed ^{2}D peaks for the 15 ppm (red) and neat (blue). (B) BGC summed ^{2}D peaks for the 15 ppm (red) and neat (blue). (C) Original summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (D) BGC summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (E) Original summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). (F) BGC summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). At 1.5 ppm, the average BGC peak height of benzo[b]thiophene greatly exceeds its *LOD* (0.34 ppm) by approximately 5-fold, but just barely exceeds its *LOQ* (1.1 ppm) by a factor of 1.3, which is why it suffers a drastic decrease in hit number from the 3 ppm versus neat to the 1.5 ppm versus neat F-ratio comparison (hit number 7 to 28 in Table 3.3).

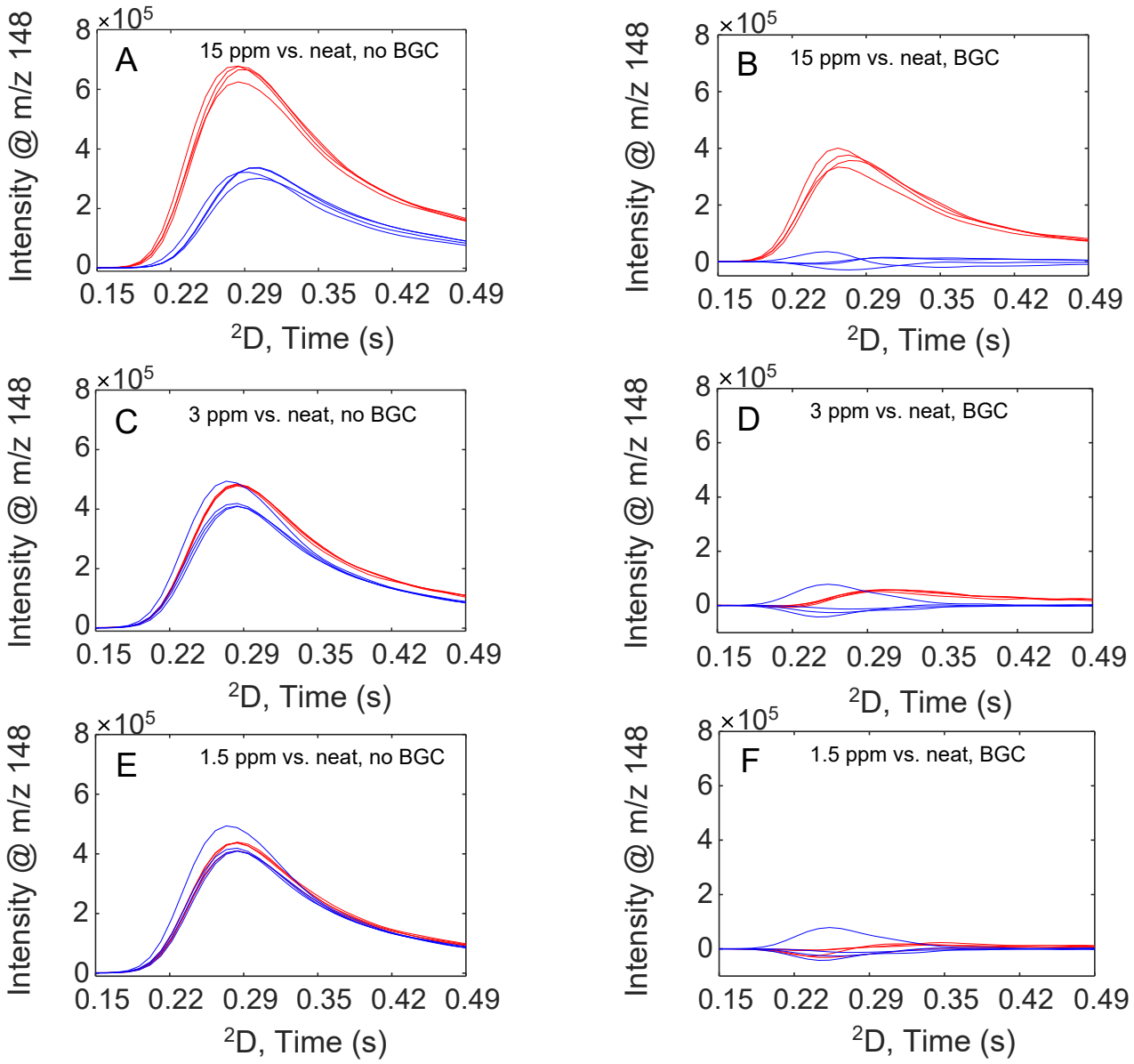

Figure B.5. Summed ^{2}D peaks for the replicates of 3-methylbenzothiophene: spiked (red) and neat (blue). A ^{2}D window of 340 ms was used to prepare all summed ^{2}D peaks for this analyte. (A) Original summed ^{2}D peaks for the 15 ppm (red) and neat (blue). (B) Background corrected (BGC) summed ^{2}D peaks for the 15 ppm (red) and neat (blue). (C) Original summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (D) BGC summed ^{2}D peaks for the 3 ppm (red) and neat (blue). (E) Original summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). (F) BGC summed ^{2}D peaks for the 1.5 ppm (red) and neat (blue). At 3 ppm, the average BGC peak height of 3-methylbenzothiophene exceeds its *LOD* (1.3 ppm) by a factor of 1.1 but falls below its *LOQ* (4.2 ppm) by 3-fold, which is why it is not discovered in the 3 ppm versus neat and 1.5 ppm versus neat F-ratio hitlists (Table 3.3).

Table B.3. Hitlists obtained when a larger ^{2}D tile dimension of 600 ms is used, relative to the optimum ^{2}D tile dimension determined in Table B.2. The top F-ratio *m/z* was used to rank the resulting hitlists. The analytes are listed top to bottom based on the ordering shown in Table 3.2 in the manuscript, for the 15 ppm versus neat comparison.

Analyte	Hit Number (Top F-ratio), 15 ppm vs. neat	Hit Number (Top F-ratio), 3 ppm vs. neat	Hit Number (Top F-ratio), 1.5 ppm vs. neat
2-methylthiophene	3 (7568)	3 (2653)	3 (678)
1,4-oxathiane	1 (8684)	44 (133)	*NF (N/A)*
benzo[b]thiophene	6 (4483)	9 (892)	*12 (235)*
thiophene	4 (6407)	6 (1940)	4 (435)
2-propylthiophene	2 (8292)	2 (4345)	10 (256)
2,5-dimethylthiophene	8 (3048)	35 (159)	111 (84)
3-methylthiophene	7 (3936)	7 (1279)	2 (1142)
tetrahydrothiophene	14 (901)	*NF (N/A)*	*NF (N/A)*
3-methylbenzothiophene	12 (1033)	*NF (N/A)*	*NF (N/A)*
2-chloroethylphenylsulfide	10 (1691)	28 (178)	*NF (N/A)*
2-butyl-5-ethylthiophene	9 (1727)	1 (5204)	1 (2933)
2-methylbenzothiophene	11 (1180)	*NF (N/A)*	*NF (N/A)*
2-hexylthiophene	16 (518)	4 (2506)	19 (209)
3-acetyl-2,5-dimethylthiophene	20 (309)	*NF (N/A)*	*NF (N/A)*

In Table B.3 above, the most dramatic changes in hit numbers relative to those provided in Table 3.3 in the manuscript are highlighted in grey. It appears that the larger ^{2}D tile dimension of 600 ms is better than that used for the 15 ppm versus neat comparison, as the last hit improves from hit 34 (Table 3.3) to hit 20 (Table B.3). This is because a few of the spiked sulfur-containing compounds have larger ^2W_b relative to the native fuel peaks at 15 ppm. In the 3 ppm versus neat hitlist above, the rankings of 1,4-oxathiane, 2,5-dimethylthiophene, and 2-chloroethyl phenyl sulfide worsen from hit 25 to 44, hit 19 to hit 35, and hit 16 to hit 28, respectively, relative to Table 3.3. In the 1.5 ppm versus neat hitlist above, the ranking of 2,5-dimethylthiophene worsens from hit 39 to 111, and 2-chloroethyl phenyl sulfide can no longer be found ("*NF*"). These results highlight how too large of a tile will drown out the signal from analytes of interest and thus worsen their discoverability.

Table B.4. Hitlists obtained when a smaller ^{2}D tile dimension of 100 ms is used, relative to the optimum ^{2}D tile dimension determined in Table B.2. The top F-ratio *m/z* was used to rank the resulting hitlists. The analytes are listed top to bottom based on the ordering shown in Table 3.2 in the manuscript, for the 15 ppm versus neat comparison.

Analyte	Hit Number (Top F-ratio), 15 ppm vs. neat	Hit Number (Top F-ratio), 3 ppm vs. neat	Hit Number (Top F-ratio), 1.5 ppm vs. neat
2-methylthiophene	4 (25610)	3 (6391)	1 (2368)
1,4-oxathiane	1 (55974)	51 (353)	*NF (N/A)*
benzo[b]thiophene	8 (18243)	14 (2054)	13 (621)
thiophene	21 (7568)	7 (4927)	4 (1234)
2-propylthiophene	15 (10638)	1 (15185)	16 (560)
2,5-dimethylthiophene	7 (21183)	36 (510)	47 (269)
3-methylthiophene	2 (51285)	2 (8229)	3 (1276)
tetrahydrothiophene	23 (7020)	*NF (N/A)*	*NF (N/A)*
3-methylbenzothiophene	82 (983)	*NF (N/A)*	*NF (N/A)*
2-chloroethylphenylsulfide	39 (3283)	21 (1151)	39 (294)
2-butyl-5-ethylthiophene	36 (3644)	4 (6243)	5 (1222)
2-methylbenzothiophene	65 (1368)	*NF (N/A)*	*NF (N/A)*
2-hexylthiophene	66 (1304)	5 (6021)	11 (761)
3-acetyl-2,5-dimethylthiophene	93 (649)	*NF (N/A)*	*NF (N/A)*

In Table B.4 above, the most dramatic changes in hit numbers relative to those provided in Table 3.3 in the manuscript are highlighted in grey. The ranking of the last hit in the 15 ppm versus neat hitlist has worsened from 34 (Table 3.3) to 93 (Table B.4) due to the increased number of redundant hits per analyte. Given the especially wide ^2W_b observed at 15 ppm for some of the analytes (up to 700 ms), this effect is more pronounced at 15 ppm than in the 3 ppm versus neat and 1.5 ppm versus neat F-ratio hitlists. However, noticeable differences in rankings can still be observed. In the 3 ppm versus neat hitlist above, the rankings of 1,4-oxathiane and 2,5-dimethylthiophene worsen from hit 25 to 51 and hit 19 to 36, respectively, relative to Table 3.3. In the 1.5 ppm versus neat hitlist above, the rankings of 2,5-dimethylthiophene and 2-chloroethyl phenyl sulfide worsen from hit 39 to 47 and hit 5 to 39, respectively. These results highlight how discoverability can worsen when too small of a tile is used.

APPENDIX C

This Appendix is reproduced from the Electronic Supplementary Content of Paige E. Sudol[†], Micaela Galletta[†], Peter Q. Tranchida, Mariosimone Zoccali, Luigi Mondello, Robert E. Synovec, "Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis" *Journal of Chromatography A* 1662 (2022) 462735.

[†]These authors contributed equally to this work.

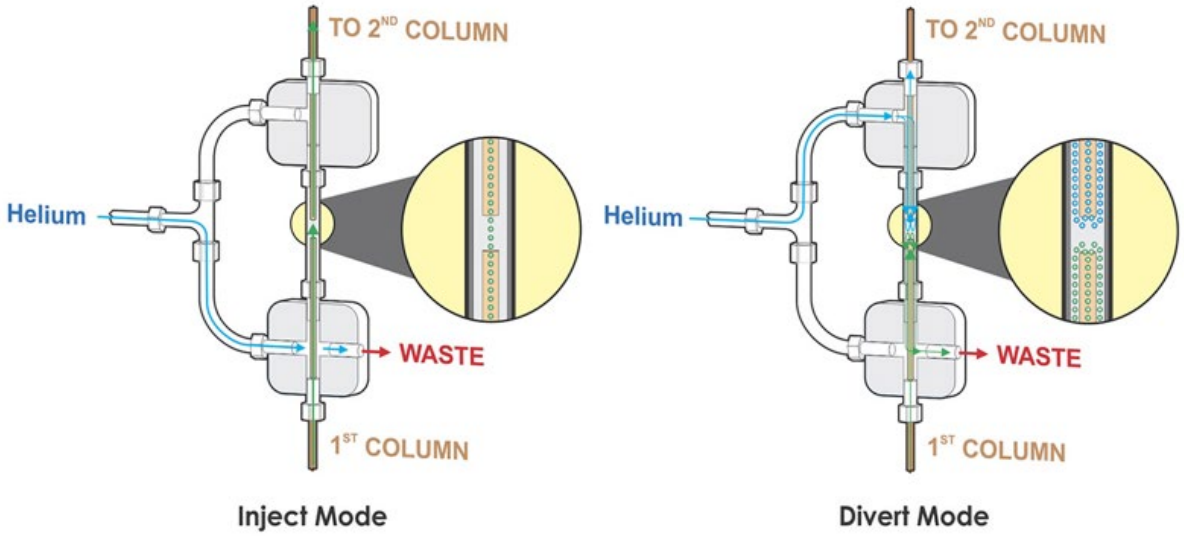

Figure C.1. Schematic representations of the low duty cycle flow modulator in the re-inject and divert modes. Reproduced with permission from LECO Corp.

Table C.1. 220 hits which were manually selected from the ChromaTOF tile hitlist. Analytes which had a one-way *ANOVA* p-value > 0.05 are highlighted in pink, and the additional analytes which had a one-way *ANOVA* p-value > 0.01 are highlighted in purple.

Hit #	^1t_R (min)	^2t_R (s)	F-ratio	*m/z*	Identity	p-value
1	8.17	0.27	3434.2	152	ethyl trans-4-decenoate	2.9E-12
2	6.69	0.51	1608.4	87	cis-Ocimenol	1.6E-12
3	8.24	0.01	1594.0	136	α-Terpineol	6.9E-09
4	7.29	0.64	1385.7	72	Linalool	1.9E-12
5	7.07	0.35	1370.9	41	geranyl vinyl ether	9.7E-12
6	8.09	0.63	1323.6	174	Succinate <diethyl->	5.9E-12
7	6.92	0.18	1301.6	55	7-Octenoic acid, ethyl ester	9.4E-06
8	5.75	0.45	1217.2	94	linalool ethyl ether	6.4E-12
9	7.47	0.59	917.3	70	isoamyl lactate	2.1E-11
10	6.86	0.35	806.6	93	geranyl isobutyrate	3.6E-10
11	5.96	0.52	689.6	75	Lactate <ethyl->	1.1E-10
12	4.45	0.37	631.6	93	Myrcene	1.1E-08
13	5.21	0.32	623.0	93	Carene <delta-3->	4.1E-07
14	7.68	0.61	610.4	71	3,7-Octadiene-2,6-diol, 2,6-dimethyl-	2.2E-09
15	5.07	0.33	551.6	93	Hexanoic acid, ethyl ester	7.5E-07
16	12.95	0.07	529.2	149	Phthalate <di-isobutyl>	9.3E-11
17	8.58	0.61	503.9	123	Citronellol	1.2E-09
18	4.45	0.54	475.1	93	α-phellandrene	2.4E-09
19	9.26	0.47	472.4	91	benzyl alcohol	1.2E-10
20	9.35	0.04	442.9	129	Butanedioic acid, ethyl 3-methylbutyl ester	5.7E-10
21	6.83	0.19	412.7	68	nerol oxide	2.7E-10
22	4.76	0.52	373.0	92	limonene	2.8E-09
23	8.79	0.60	361.6	69	nerol	1.2E-09
24	9.42	0.44	353.7	70	Undecenoate <isopentyl->	1.6E-07
25	10.95	0.28	353.5	221	1H-Indene, 2,3-dihydro-1,1,3-trimethyl-3-phenyl-	5.3E-09
26	6.09	0.60	318.6	99	allyl isothiocyanate	5.6E-07
27	5.56	0.37	317.1	70	Pentanoic acid, 3-methylbutyl ester	3.3E-07
28	3.02	0.13	305.2	55	Acetate <isobutyl->	8.4E-08
29	9.04	0.59	301.0	68	Geraniol	3.6E-06
30	9.31	0.34	300.2	88	Undec-10-enoate <ethyl->	1.2E-07
31	7.14	0.45	294.8	136	2-carene	1.7E-06
32	9.55	0.24	292.2	183	methyl 2-oxooctadecanoate	5.0E-08
33	6.19	0.31	285.8	99	*unidentified 1*	2.1E-09
34	8.25	0.36	285.2	163	*unidentified 2*	1.5E-08
35	3.92	0.61	283.4	139	Bois de Rose oxide	2.2E-08
36	10.93	0.11	270.9	81	*unidentified 3*	5.7E-12
37	4.24	0.21	264.0	106	5-(1-methylethylidene)-1,3-cyclopentadiene	1.5E-11
38	8.93	0.68	258.8	106	phenethyl acetate	1.7E-08
39	4.33	0.53	257.4	56	butyl alcohol	5.7E-09
40	5.47	0.45	253.6	93	terpinolene	1.1E-06
41	8.69	0.67	250.0	143	diethyl glutarate	1.0E-07
42	3.61	0.15	247.3	73	butyl acetate	1.0E-08
43	8.43	0.37	240.2	163	*unidentified 4*	8.7E-09
44	9.02	0.44	231.5	63	ethyl dodecanoate	2.1E-07
45	8.98	0.13	220.0	122	(E)-β-damascenone	2.9E-07
46	9.21	0.09	215.0	123	butyl benzoate	4.0E-08
47	8.44	0.16	214.9	85	vinyl decanoate	2.3E-08
48	8.44	0.63	211.8	90	benzyl acetate	3.5E-08
49	6.37	0.15	211.0	95	*unidentified 5*	4.9E-07
50	6.86	0.51	203.2	96	furfural	6.7E-08
51	8.31	0.55	177.6	111	*unidentified 6*	4.1E-08
52	8.37	0.15	162.0	68	neryl phenylacetate	7.8E-04
53	6.96	0.42	161.1	93	trans-(-)-5-methyl-3-(1-methylethenyl)-cyclohexene	5.2E-07
54	10.86	0.51	153.0	135	4-vinyl guaiacol	3.5E-07
55	8.60	0.48	152.2	218	*unidentified 7*	3.3E-07
56	5.54	0.12	151.3	68	ethyl (3Z)-3-hexenoate	2.1E-05
57	7.24	0.47	145.0	151	1,1,4a-trimethyl-3,4,4a,5,6,7-hexahydro-2(1H)-naphthalenone	1.9E-07

58	3.42	0.25	144.7	57	ethyl 2-methylbutyrate	2.6E-07
59	8.00	0.21	143.6	123	citronellyl acetate	3.3E-07
60	4.59	0.52	141.4	121	*unidentified 8*	1.7E-07
61	3.56	0.24	141.3	85	ethyl isovalerate	5.2E-07
62	7.35	0.43	141.3	127	isobutyl octanoate	9.5E-06
63	7.72	0.41	140.7	127	butyl lactate	8.1E-08
64	9.23	0.34	138.9	71	2-methyl-1-(1,1-dimethylethyl)-2-methylpropanoic acid, 1,3-propanediyl ester	3.7E-05
65	5.86	0.25	133.6	113	ethyl heptanoate	1.8E-05
66	2.57	0.19	129.4	116	ethyl isobutyrate	8.1E-07
67	5.32	0.21	127.2	105	1-ethyl-3-methyl-benzene	1.7E-06
68	8.83	0.26	127.1	107	*unidentified 9*	5.2E-10
69	6.93	0.33	122.6	137	sulfur dioxide	2.3E-06
70	8.37	0.42	122.5	173	*unidentified 10*	7.1E-07
71	8.56	0.13	122.0	68	(R)-lavandulyl acetate	9.5E-06
72	6.26	0.21	121.5	129	methyl octanoate	7.7E-07
73	6.15	0.19	120.7	56	heptyl acetate	4.8E-07
74	7.67	0.69	120.3	132	7-methylbenzofuran	3.9E-06
75	5.48	0.22	119.0	106	mesitylene	8.8E-08
76	10.06	0.05	118.2	172	*unidentified 11*	6.6E-07
77	7.89	0.37	118.2	51	ethyl decanoate	2.2E-06
78	6.43	0.31	115.3	99	*unidentified 12*	4.5E-08
79	6.01	0.33	115.1	99	isobutyl hexanoate	2.3E-05
80	8.84	0.47	111.1	70	*unidentified 13*	2.9E-07
81	10.41	0.59	110.9	88	ethyl pentadecanoate	2.1E-06
82	7.15	0.17	106.3	113	*unidentified 14*	3.4E-06
83	8.86	0.60	106.2	55	9-decen-1-ol	7.0E-05
84	11.75	0.40	105.7	120	4-vinylphenol	7.2E-06
85	4.54	0.14	105.5	61	amyl acetate	3.9E-06
86	9.13	0.55	104.1	70	isoamyl decanoate	1.1E-05
87	5.01	0.24	101.5	105	2,4-nonadiyne	8.4E-06
88	9.93	0.23	101.2	157	*unidentified 15*	2.1E-06
89	8.52	0.39	99.9	68	*unidentified 16*	9.7E-04
90	6.20	0.50	99.5	45	*unidentified 17*	1.9E-07
91	4.06	0.19	99.4	101	isoamyl acetate	7.9E-07
92	8.55	0.49	96.2	155	isobutyl decanoate	1.3E-06
93	5.62	0.50	96.2	45	acetoin	3.9E-06
94	10.07	0.55	95.9	93	ethyl tetradecanoate	1.5E-05
95	9.38	0.49	92.7	88	ethyl tridecanoate	1.6E-04
96	5.67	0.24	89.4	125	*unidentified 18*	2.2E-07
97	10.75	0.43	87.5	107	*unidentified 19*	5.7E-08
98	10.03	0.08	87.3	67	nerolidol	2.4E-04
99	5.16	0.48	87.2	93	*unidentified 20*	1.8E-08
100	11.81	0.40	86.6	81	n-decanoic acid	5.9E-06
101	2.34	0.11	86.5	78	ethanol	8.4E-06
102	7.60	0.27	83.4	43	methyl decanoate	2.9E-04
103	6.76	0.47	82.9	45	*unidentified 21*	6.2E-06
104	7.21	0.09	81.7	43	methyl 2-oxononanoate	1.0E-06
105	5.72	0.09	81.7	68	(4Z)-4-hexenyl acetate	1.3E-05
106	8.51	0.19	80.2	57	*unidentified 22*	6.1E-06
107	11.02	0.21	78.6	183	guaiazulene	1.9E-06
108	7.58	0.54	76.6	109	*unidentified 23*	2.2E-06
109	10.22	0.57	76.1	101	methyl 2,8-dimethyl-undecanoate	1.5E-05
110	8.92	0.28	74.4	190	*unidentified 24*	6.3E-04
111	6.03	0.01	71.3	86	*unidentified 25*	5.8E-06
112	10.10	0.65	68.3	85	*unidentified 26*	1.4E-05
113	9.47	0.02	67.7	131	phenylethyl alcohol	2.6E-03
114	10.20	0.45	67.3	111	ethyl E-11-hexadecenoate	6.8E-06
115	8.00	0.45	66.7	101	isoamyl octanoate	3.5E-05
116	5.39	0.17	66.3	35	hexyl acetate	7.1E-06
117	9.61	0.03	65.9	163	*unidentified 27*	2.3E-05
118	5.93	0.19	65.5	105	*unidentified 28*	2.7E-05
119	10.67	0.52	63.7	85	*unidentified 29*	1.7E-05
120	4.66	0.17	63.2	106	methyl hexanoate	2.9E-04
121	10.44	0.69	63.2	126	*unidentified 30*	2.9E-05
122	5.30	0.05	62.6	104	styrene	3.6E-05
123	3.25	0.19	61.3	61	ethyl butyrate	5.5E-03
124	7.22	0.35	59.9	93	ethyl nonanoate	4.7E-06
125	6.69	0.44	59.9	174	α-ionene	1.8E-04

126	7.06	0.59	59.1	84	(S)-3-ethyl-4-methylpentanol	6.7E-06
127	5.93	0.53	59.0	159	*unidentified 31*	3.7E-05
128	6.53	0.47	58.9	174	*unidentified 32*	1.3E-03
129	7.46	0.67	57.2	61	ethyl 3-(methylthio)propionate	5.2E-06
130	10.17	0.66	57.0	70	*unidentified 33*	8.4E-05
131	13.44	0.49	56.1	82	2,6-bis(1,1-dimethylethyl)-1,4-benzenediol	3.7E-05
132	7.37	0.20	55.8	99	*unidentified 34*	1.7E-06
133	7.37	0.67	54.5	109	*unidentified 35*	6.6E-02
134	2.65	0.09	54.0	61	n-propyl acetate	1.6E-04
135	6.75	0.36	53.9	43	isopentyl hexanoate	2.5E-02
136	12.87	0.41	53.8	129	n-dodecanoic acid	1.2E-05
137	6.45	0.51	52.5	57	*unidentified 36*	8.2E-08
138	5.75	0.29	50.9	155	*unidentified 37*	1.0E-01
139	4.81	0.53	49.7	93	2-methyl-1-butanol	1.0E-03
140	9.20	0.49	49.3	101	*unidentified 38*	1.5E-04
141	8.56	0.27	48.8	155	dehydro-ar-ionene	6.8E-05
142	7.63	0.61	48.4	113	n-hexadecane	8.3E-05
143	6.33	0.53	47.2	127	tetradecane	1.4E-04
144	8.35	0.25	46.4	88	ethyl 9-decenoate	1.2E-04
145	4.74	0.05	46.4	68	*unidentified 39*	1.3E-05
146	5.95	0.12	45.9	99	ethyl hex-(2E)-enoate	5.6E-04
147	9.50	0.38	45.7	157	α-calacorene	7.0E-05
148	8.72	0.53	45.1	220	*unidentified 40*	2.4E-06
149	10.44	0.50	43.9	123	*unidentified 41*	2.8E-05
150	8.66	0.55	43.8	93	*unidentified 42*	5.0E-05
151	10.84	0.39	42.1	81	octanoic acid	2.0E-04
152	7.70	0.09	41.5	93	*unidentified 43*	1.2E-04
153	5.68	0.55	41.2	56	4-methyl-pentan-1-ol	1.0E-04
154	7.35	0.60	41.0	41	n-octyl acrylate	2.6E-04
155	3.81	0.55	40.7	74	2-methyl-1-propanol	5.7E-03
156	6.92	0.59	40.2	57	2-ethyl-1-hexanol	1.6E-04
157	5.86	0.07	39.6	67	(2Z)-2-hexenyl acetate	3.4E-06
158	6.32	0.23	38.9	84	*unidentified 44*	4.6E-04
159	6.69	0.15	38.4	91	*unidentified 45*	7.3E-04
160	5.83	0.49	38.3	43	*unidentified 46*	2.1E-03
161	5.16	0.27	38.2	109	1,2,3-trimethylbenzene	2.9E-02
162	9.42	0.16	38.1	205	butylated hydroxytoluene	9.8E-05
163	5.89	0.05	37.3	108	6-methyl-5-hepten-2-one	3.2E-05
164	10.63	0.23	36.6	197	*unidentified 47*	1.1E-03
165	11.24	0.56	36.3	91	*unidentified 48*	6.1E-05
166	7.42	0.22	35.3	68	*unidentified 49*	1.9E-05
167	8.38	0.49	35.1	59	3-(methylthio)-1-propanol	3.3E-02
168	6.03	0.23	33.9	119	*unidentified 50*	3.3E-04
169	11.06	0.25	33.7	197	*unidentified 51*	3.7E-04
170	3.16	0.46	33.0	151	*unidentified 52*	1.0E-02
171	12.49	0.67	31.7	111	*unidentified 53*	3.8E-04
172	7.14	0.32	31.5	61	propyl octanoate	1.1E-02
173	10.25	0.38	29.4	269	*unidentified 54*	3.6E-02
174	11.03	0.69	29.4	251	ethyl palmitate	8.8E-02
175	9.52	0.17	26.4	126	*unidentified 55*	2.8E-05
176	9.72	0.54	26.2	88	*unidentified 56*	2.1E-04
177	9.87	0.46	23.5	87	*unidentified 57*	5.1E-04
178	3.48	0.32	23.4	237	*unidentified 58*	2.4E-02
179	10.26	0.53	23.2	85	*unidentified 59*	4.8E-05
180	6.59	0.30	22.7	141	ethyl octanoate	1.2E-02
181	9.01	0.28	22.3	221	*unidentified 60*	6.7E-04
182	12.40	0.61	21.9	87	*unidentified 61*	1.2E-03
183	7.91	0.62	21.9	336	*unidentified 62*	9.1E-02
184	10.00	0.29	19.7	216	acetic acid	1.2E-03
185	7.10	0.65	19.5	45	2-nonanol	3.5E-02
186	5.40	0.48	19.3	78	phenacyl formate	6.7E-03
187	9.55	0.04	18.9	159	*unidentified 63*	3.8E-03
188	2.03	0.08	18.9	71	ethyl acetate	5.1E-04
189	7.80	0.57	18.1	140	ethyl 2-furoate	2.6E-03
190	10.81	0.64	17.1	88	*unidentified 64*	1.9E-04
191	7.31	0.19	17.0	146	2,3-dihydro-4,7-dimethyl-1H-indene	6.1E-06
192	3.58	0.48	15.9	40	*unidentified 65*	1.7E-02
193	7.66	0.19	15.6	71	*unidentified 66*	1.3E-05

194	3.01	0.43	15.3	248	(2-aziridinylethyl)amine	1.6E-01
195	11.92	0.02	15.1	101	ethyl stearate	6.4E-01
196	13.19	0.61	14.9	73	*unidentified 67*	6.4E-03
197	5.98	0.67	14.6	54	tetrahydro-3,6-dimethyl-2H-pyran-2-one	8.0E-06
198	6.71	0.13	14.1	134	cosmene	8.1E-02
199	13.75	0.56	14.0	89	metaldehyde isomer IV	1.9E-04
200	8.07	0.47	13.4	159	*unidentified 68*	8.7E-03
201	7.86	0.15	13.1	94	*unidentified 69*	1.8E-07
202	11.06	0.49	13.0	86	*unidentified 70*	6.6E-01
203	6.85	0.05	11.6	59	*unidentified 71*	3.6E-03
204	8.03	0.63	11.3	105	acetophenone	6.4E-02
205	6.27	0.52	11.2	121	*unidentified 72*	7.8E-02
206	9.11	0.15	10.7	43	neryl acetone	2.5E-01
207	12.83	0.15	10.6	245	*unidentified 73*	8.8E-02
208	5.76	0.17	9.7	59	*unidentified 74*	5.8E-03
209	5.96	0.41	9.6	334	*unidentified 75*	3.9E-01
210	5.19	0.45	7.8	85	*unidentified 76*	2.1E-02
211	13.90	0.19	7.3	89	metaldehyde isomer I	2.7E-01
212	11.98	0.56	6.8	157	*unidentified 77*	4.0E-01
213	12.73	0.01	6.0	89	metaldehyde isomer II	6.3E-03
214	13.12	0.53	6.0	121	*unidentified 78*	8.4E-02
215	13.03	0.53	5.9	121	*unidentified 79*	1.5E-01
216	10.41	0.11	5.6	217	*unidentified 80*	2.4E-01
217	4.53	0.41	5.1	343	*unidentified 81*	2.7E-01
218	11.54	0.69	4.9	73	*unidentified 82*	8.9E-01
219	4.87	0.09	4.6	67	isopentyl alcohol	1.2E-01
220	4.98	0.06	4.2	68	*unidentified 83*	5.1E-01

Table C.2. Sensory characteristics of 220 hits selected from ChromaTOF Tile [39]. For analytes which could not be identified, the sensory profile is indicated as "n/a". Analytes which had a one-way *ANOVA p*-value > 0.05 are highlighted in pink, and the additional analytes which had a one-way *ANOVA p*-value > 0.01 are highlighted in purple.

Hit #	Identity	Wine, highest concentration	Sensory Profile
1	ethyl trans-4-decenoate	1	fatty, waxy, green, pineapple, pear
2	cis-Ocimenol	1	**unknown**
3	α-Terpineol	1	citrus, woody, lemon lie, soapy
4	Linalool	1	orange, lemon, floral, waxy, aldehydic, woody
5	geranyl vinyl ether	1	**unknown**
6	Succinate <diethyl->	5	fruity, tart, floral, tropical, passion fruit
7	7-Octenoic acid, ethyl ester	1	**unknown**
8	linalool ethyl ether	1	floral
9	isoamyl lactate	5	fruity, creamy, nutty
10	geranyl isobutyrate	1	sweet, floral, citrus, fruity
11	Lactate <ethyl->	5	sweet, fruity, creamy, pineapple, caramellic
12	Myrcene	1	woody, vegetable, citrus, fruity, tropical, minty
13	Carene <delta-3->	1	citrus, pine, terpenic, tropical, juniper, wasabi
14	3,7-Octadiene-2,6-diol, 2,6-dimethyl-	1	**unknown**
15	Hexanoic acid, ethyl ester	1	sweet, pineapple, fruity, waxy, banana, green
16	Phthalate <di-isobutyl>	1	**unknown**
17	Citronellol	1	floral, rose, sweet, green, fruity, citrus
18	α-phellandrene	1	terpenic, citrus, lime, fresh, green
19	benzyl alcohol	3	chemical, fruity, balsamic
20	Butanedioic acid, ethyl 3-methylbutyl ester	5	**unknown**
21	nerol oxide	1	green, vegetable, floral, waxy, herbal, minty
22	limonene	1	citrus, herbal, terpenic, camphoreous
23	nerol	1	lemon, bitter, green, fruity, terpenic
24	Undecenoate <isopentyl->	1	floral, rose, waxy
25	1H-Indene, 2,3-dihydro-1,1,3-trimethyl-3-phenyl-	2	**unknown**
26	allyl isothiocyanate	1	mustard, horseradish, wasabi
27	Pentanoic acid, 3-methylbutyl ester	1	strawberry
28	Acetate <isobutyl->	1	sweet, fruity, banana
29	Geraniol	1	floral, rose, waxy, fruity, peach
30	Undec-10-enoate <ethyl->	1	fatty, waxy, green, fruity
31	2-carene	1	**unknown**
32	methyl 2-oxooctadecanoate	5	**unknown**
33	*unidentified 1*	1	*n/a*
34	*unidentified 2*	5	*n/a*
35	Bois de Rose oxide	1	sweet, camphoreous, woody, cooling, floral
36	*unidentified 3*	5	*n/a*
37	5-(1-methylethylidene)-1,3-cyclopentadiene	1	**unknown**
38	phenethyl acetate	1	sweet, honey, floral, rose, green, fruity
39	butyl alcohol	2	banana, fusel
40	terpinolene	1	woody, terpenic, lemon, lime, herbal, floral
41	diethyl glutarate	5	**Unknown**
42	butyl acetate	2	sweet, ripe, banana, tropical, candy, green
43	*unidentified 4*	5	*n/a*
44	ethyl dodecanoate	5	waxy, soapy, floral, creamy, dairy, fruity
45	(E)-β-damascenone	2	apple, rose, honey, tobacco, sweet
46	butyl benzoate	1	amber, balsamic, fruity
47	vinyl decanoate	3	**unknown**
48	benzyl acetate	2	fruity, sweet, balsamic, jasmin, floral
49	*unidentified 5*	1	*n/a*
50	furfural	5	brown, sweet, woody, bready, nutty, burnt
51	*unidentified 6*	3	*n/a*
52	neryl phenylacetate	1	honey, rose, honeysuckle
53	trans-(-)-5-methyl-3-(1-methylethenyl)-cyclohexene	1	**unknown**
54	4-vinyl guaiacol	3	bacon, smoky, spicy, clove, phenolic, woody
55	*unidentified 7*	3	*n/a*
56	ethyl (3Z)-3-hexenoate	1	green, pear, apple, tropical
57	1,1,4a-trimethyl-3,4,4a,5,6,7-hexahydro-2(1H)-naphthalenone	3	**unknown**

58	ethyl 2-methylbutyrate	5	fruity, fresh, berry, grape, pineapple, mango
59	citronellyl acetate	2	floral, waxy, aldehydic, green, fruity, pear, apple
60	*unidentified 8*	1	*n/a*
61	ethyl isovalerate	5	sweet, fruity, spicy, metallic, green, pineapple
62	isobutyl octanoate	1	fruity, green, oily, floral
63	butyl lactate	2	dairy, creamy, milky, coconut, nutty
64	2-methyl-1-(1,1-dimethylethyl)-2-methylpropanoic acid, 1,3-propanediyl ester	1	**unknown**
65	ethyl heptanoate	1	fruity, pineapple, banana, strawberry, spicy, oily
66	ethyl isobutyrate	5	pungent, ethereal, fruity, alliaceous, egg nog
67	1-ethyl-3-methyl-benzene	1	**unknown**
68	*unidentified 9*	5	*n/a*
69	sulfur dioxide	1	**unknown**
70	*unidentified 10*	5	*n/a*
71	(R)-lavandulyl acetate	2	**unknown**
72	methyl octanoate	1	green, fruity, waxy, citrus, aldehydic, fatty
73	heptyl acetate	1	green, fatty, spicy, citrus, soapy, aldehydic, floral
74	7-methylbenzofuran	3	earthy, mushroom, hazelnut
75	mesitylene	1	**unknown**
76	*unidentified 11*	3	*n/a*
77	ethyl decanoate	5	waxy, fruity, sweet, apple
78	*unidentified 12*	2	*n/a*
79	isobutyl hexanoate	1	sweet, fruity, pineapple, green, tropical, estery
80	*unidentified 13*	1	*n/a*
81	ethyl pentadecanoate	1	honey, sweet
82	*unidentified 14*	1	*n/a*
83	9-decen-1-ol	1	fresh, waxy, metallic, cilantro, watery, oily, fatty
84	4-vinylphenol	1	phenolic, medicinal, spicy
85	amyl acetate	2	fruity, pear, banana, sweet
86	isoamyl decanoate	5	waxy, fruity, banana, green, creamy, cheesy, fatty
87	2,4-nonadiyne	1	**unknown**
88	*unidentified 15*	5	*n/a*
89	*unidentified 16*	1	*n/a*
90	*unidentified 17*	5	*n/a*
91	isoamyl acetate	2	sweet, fruity, banana, green, ripe
92	isobutyl decanoate	5	oily, sweet, brandy, apricot, fermented, cognac
93	acetoin	5	creamy, dairy, sweet, oily, buttery, yogurt
94	ethyl tetradecanoate	1	sweet, waxy, creamy
95	ethyl tridecanoate	1	**unknown**
96	*unidentified 18*	1	*n/a*
97	*unidentified 19*	4	*n/a*
98	nerolidol	1	green, floral, woody, fruity, citrus, melon
99	*unidentified 20*	1	*n/a*
100	n-decanoic acid	5	soapy, waxy, fruity
101	ethanol	4	alcoholic, ethereal, medicinal
102	methyl decanoate	1	fatty, oily, fruity
103	*unidentified 21*	5	*n/a*
104	methyl 2-oxononanoate	3	**unknown**
105	(4Z)-4-hexenyl acetate	2	**unknown**
106	*unidentified 22*	5	*n/a*
107	guaiazulene	2	**unknown**
108	*unidentified 23*	5	*n/a*
109	methyl 2,8-dimethyl-undecanoate	1	**unknown**
110	*unidentified 24*	2	*n/a*
111	*unidentified 25*	4	*n/a*
112	*unidentified 26*	1	*n/a*
113	phenylethyl alcohol	5	floral, sweet, rose, bready
114	ethyl E-11-hexadecenoate	1	**unknown**
115	isoamyl octanoate	5	sweet, fruity, waxy, pineapple, green, coconut
116	hexyl acetate	2	fruity, green, fresh, sweet, banana, apple, pear
117	*unidentified 27*	5	*n/a*
118	*unidentified 28*	1	*n/a*
119	*unidentified 29*	5	*n/a*
120	methyl hexanoate	1	fruity, fatty, banana, pineapple, apple, creamy
121	*unidentified 30*	3	*n/a*
122	styrene	2	sweet, balsamic, floral, plastic, almond
123	ethyl butyrate	4	fruity, sweet, apple, fresh, ethereal

124	ethyl nonanoate	3	waxy, soapy, cognac, estery, fruity, grape
125	α-ionene	3	**unknown**
126	(S)-3-ethyl-4-methylpentanol	1	**unknown**
127	*unidentified 31*	3	*n/a*
128	*unidentified 32*	3	*n/a*
129	ethyl 3-(methylthio)propionate	5	sulfurous, onion, garlic, pineapple, rummy
130	*unidentified 33*	2	*n/a*
131	2,6-bis(1,1-dimethylethyl)-1,4-benzenediol	1	**unknown**
132	*unidentified 34*	1	*n/a*
133	*unidentified 35*	3	*n/a*
134	n-propyl acetate	4	estery, fruity, ethereal, banana, honey
135	isopentyl hexanoate	1	fruity, green, pineapple, waxy
136	n-dodecanoic acid	2	fatty, coconut, bay
137	*unidentified 36*	5	*n/a*
138	*unidentified 37*	1	*n/a*
139	2-methyl-1-butanol	1	ethereal, alcoholic, fatty, cocoa, whiskey, leathery
140	*unidentified 38*	1	*n/a*
141	dehydro-ar-ionene	3	licorice
142	n-hexadecane	1	**unknown**
143	tetradecane	1	mild, waxy
144	ethyl 9-decenoate	1	fruity, fatty
145	*unidentified 39*	1	*n/a*
146	ethyl hex-(2E)-enoate	5	fruity, green, sweet, juicy
147	α-calacorene	1	woody
148	*unidentified 40*	3	*n/a*
149	*unidentified 41*	1	*n/a*
150	*unidentified 42*	5	*n/a*
151	octanoic acid	5	rancid, soapy, cheesy, fatty, brandy
152	*unidentified 43*	1	*n/a*
153	4-methyl-pentan-1-ol	5	nutty
154	n-octyl acrylate	1	**unknown**
155	2-methyl-1-propanol	1	ethereal, fusel, whiskey
156	2-ethyl-1-hexanol	3	sweet, fatty, fruity
157	(2Z)-2-hexenyl acetate	2	**unknown**
158	*unidentified 44*	1	*n/a*
159	*unidentified 45*	1	*n/a*
160	*unidentified 46*	5	*n/a*
161	1,2,3-trimethylbenzene	1	**unknown**
162	butylated hydroxytoluene	2	phenolic, camphoreous
163	6-methyl-5-hepten-2-one	3	green, vegetable, musty, apple, banana, bean
164	*unidentified 47*	2	*n/a*
165	*unidentified 48*	5	*n/a*
166	*unidentified 49*	1	*n/a*
167	3-(methylthio)-1-propanol	1	meaty, onion, garlic, bouillon, sweet, soup
168	*unidentified 50*	4	*n/a*
169	*unidentified 51*	2	*n/a*
170	*unidentified 52*	4	*n/a*
171	*unidentified 53*	5	*n/a*
172	propyl octanoate	5	coconut, cocoa, cognac, winey, fatty
173	*unidentified 54*	1	*n/a*
174	ethyl palmitate	1	waxy, fruity, creamy, fermented, vanilla
175	*unidentified 55*	5	*n/a*
176	*unidentified 56*	1	*n/a*
177	*unidentified 57*	1	*n/a*
178	*unidentified 58*	4	*n/a*
179	*unidentified 59*	1	*n/a*
180	ethyl octanoate	4	sweet, waxy, fruity, pineapple, creamy, fatty
181	*unidentified 60*	5	*n/a*
182	*unidentified 61*	2	*n/a*
183	*unidentified 62*	1	*n/a*
184	acetic acid	3	pungent, sour, overripe fruit, vinegar
185	2-nonanol	5	waxy, soapy, musty, green, fruity, dairy
186	phenacyl formate	5	**unknown**
187	*unidentified 63*	5	*n/a*
188	ethyl acetate	1	ethereal, fruity, sweet, grape, cherry
189	ethyl 2-furoate	3	burnt
190	*unidentified 64*	1	*n/a*

191	2,3-dihydro-4,7-dimethyl-1H-indene	3	**unknown**
192	*unidentified 65*	4	*n/a*
193	*unidentified 66*	1	*n/a*
194	(2-aziridinylethyl)amine	4	**unknown**
195	ethyl stearate	5	mild, waxy
196	*unidentified 67*	1	*n/a*
197	tetrahydro-3,6-dimethyl-2H-pyran-2-one	5	**unknown**
198	cosmene	1	**unknown**
199	metaldehyde isomer IV	1	**unknown**
200	*unidentified 68*	2	*n/a*
201	*unidentified 69*	5	*n/a*
202	*unidentified 70*	2	*n/a*
203	*unidentified 71*	1	*n/a*
204	acetophenone	3	powdery, bitter almond, cherry, coumarinic, fruity
205	*unidentified 72*	5	*n/a*
206	neryl acetone	3	fatty, metallic
207	*unidentified 73*	5	*n/a*
208	*unidentified 74*	4	*n/a*
209	*unidentified 75*	5	*n/a*
210	*unidentified 76*	4	*n/a*
211	metaldehyde isomer I	1	**unknown**
212	*unidentified 77*	5	*n/a*
213	metaldehyde isomer II	1	**unknown**
214	*unidentified 78*	1	*n/a*
215	*unidentified 79*	1	*n/a*
216	*unidentified 80*	5	*n/a*
217	*unidentified 81*	4	*n/a*
218	*unidentified 82*	1	*n/a*
219	isopentyl alcohol	4	fusel, fermented, fruity, banana, ethereal, cognac
220	*unidentified 83*	3	*n/a*

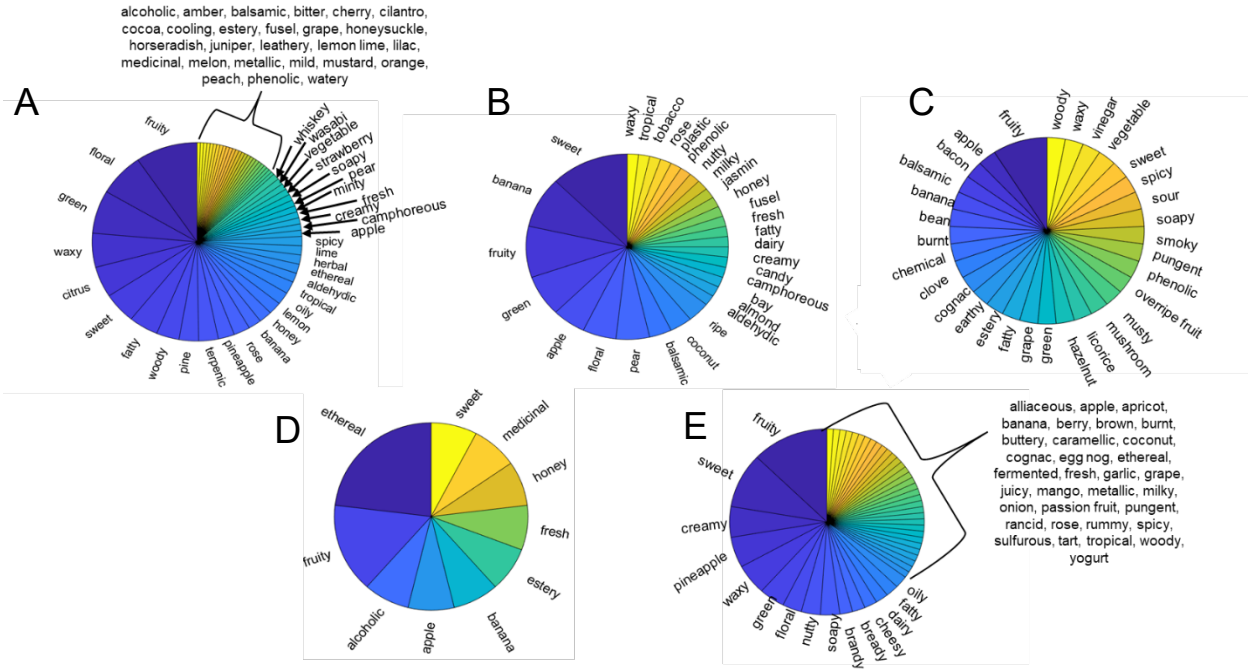

Figure C.2. Pie charts displaying the distribution of sensory descriptors for the analytes with the highest concentrations in Wines 1-5 (A-E, respectively), with a p-value < 0.01 (wine 1, 90 analytes; wine 2, 24 analytes; wine 3, 21 analytes; wine 4, 7 analytes; wine 5, 45 analytes).

APPENDIX D

This Appendix is reproduced from the Electronic Supplementary Content of Paige E. Sudol, Grant S. Ochoa, Caitlin N. Cain, Robert E. Synovec, “Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry” Submitted to *Analytica Chimica Acta* (2021).

Table D.1. Identity of sulfur-containing compounds in the equal-mass mixture that was spiked into JP8 jet fuel at varying concentration levels. The $^{1}t_{R}$, $^{2}t_{R}$, and top ten most intense m/z per analyte are listed.

	Analyte	$^{1}t_{R}$ (min)	$^{2}t_{R}$ (s)	Top 10 most intense *m/z* (highest to lowest)
1	thiophene	3.25	2.82	84, 58, 45, 57,50, 69, 83, 85, 51, 81
2	2-methylthiophene	5.00	0.34	97, 98, 45, 69, 50, 53, 63, 58, 99, 57
3	3-methylthiophene	5.20	0.37	97, 98, 45, 91, 65, 63, 92, 69, 51, 50
4	tetrahydrothiophene	6.30	0.53	60, 88, 45, 46, 59, 47, 87, 54, 58, 55
5	2,5-dimethylthiophene	7.25	0.69	111, 112, 97, 45, 59, 51, 77, 50, 69, 53
6	1,4-oxathiane	9.40	0.35	46, 104, 61, 45, 74, 47, 59, 60, 48, 58
7	2-propylthiophene	9.75	0.53	97, 126, 45, 53, 98, 99, 69, 58, 51, 71
8	2-butyl-5-ethylthiophene	17.80	0.76	125, 168, 97, 126, 91, 123, 110, 153, 77, 45
9	2-hexylthiophene	18.45	0.80	97, 98, 168, 45, 53, 99, 111, 84, 85, 110
10	benzothiophene	19.30	0.27	134, 89, 135, 63, 90, 69, 45, 67, 108, 50
11	3-acetyl-2,5-dimethylthiophene	20.45	0.55	139, 154, 111, 59, 67, 140, 141, 77, 45, 51
12	2-methylbenzothiophene	21.8	0.35	147, 148, 149, 69, 45, 115, 63, 74, 103, 77
13	3-methylbenzothiophene	22.45	0.26	147, 148, 149, 69, 45, 74, 115, 103, 77, 63
14	2-chloroethyl phenyl sulfide	23.85	0.35	123, 45, 172, 109, 65, 51, 110, 174, 69, 50

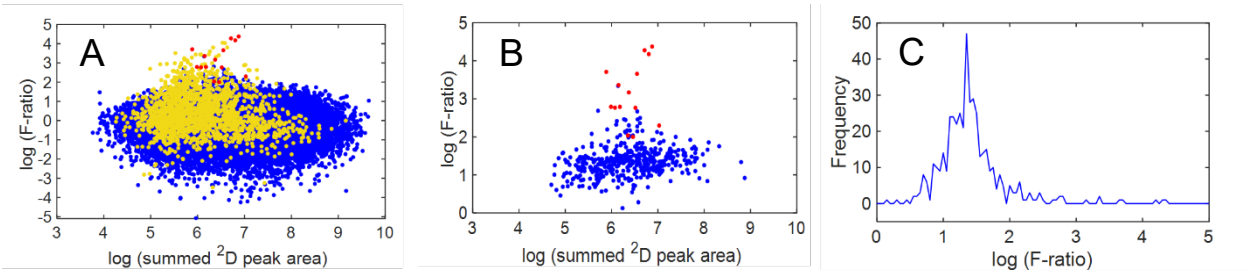

Figure D.1. Result of tile-based F-ratio analysis of the 30-ppm versus 15-ppm versus neat JP8 jet fuel samples. The same tile-based software parameters used for VRI-USI (see Experimental) were used for F-ratio analysis.

(A) Scatter plot of log (F-ratio) versus log (summed ^{2}D peak area) for all 443 hits in the F-ratio hitlist, wherein the top F-ratio *m/z* for the 14 spiked sulfur hits (i.e., true positives) are colored red, the secondary F-ratio *m/z* for the true positives are colored gold, and the false positive F-ratio *m/z* are colored blue. Just as was observed with VRI-USI in Fig. 5.3(A), the red dots largely fall above the cloud of gold dots, highlighting the advantage of ranking the hitlist with the top F-ratio *m/z*.
(B) Scatter plot of log (F-ratio) versus log (summed ^{2}D peak area) for just the top F-ratio *m/z*, with the top true positive *m/z* colored red and the top false positive *m/z* colored blue. Interestingly, compared to Fig. 5.3(B), the 14 spiked sulfur hits are more intermingled with false positive hits in the F-ratio hitlist relative to the VRI-USI hitlist, with one spiked analyte being found as hit 39 (log (F-ratio) ~ 2).
(C) Distribution of log (F-ratio) values for the top F-ratio *m/z* using a bin size = 0.05.

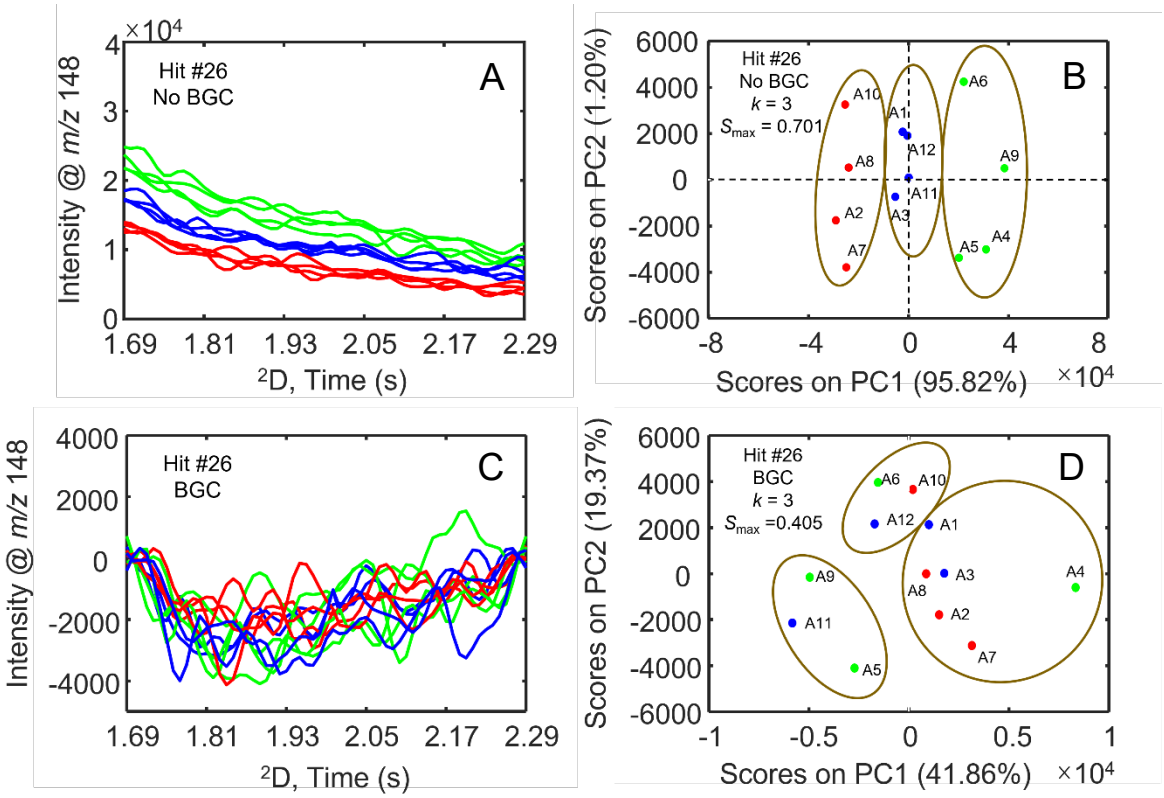

Figure D.2. Illustration of the secondary baseline correction procedure on a typical redundant hit in the RSD^2 ranked hitlist for Comparison (1), hit 26, which is a redundant hit on the tail of the peak from the spiked analyte 3-methylbenzothiophene.

(A) Original summed ^{2}D peaks for hit 26 at its top RSD^2 m/z = 148. The 30-ppm samples are green; the 15-ppm samples are blue; and the neat samples are red. It is interesting to note that a distinct concentration change is observed, even though this hit is relatively low signal and simply the tail of the 3-methylbezothiophene peak (hit 13 in Table 5.1).
(B) PCA scores plot of the summed ^{2}D peaks in (A), with the same color-coding by spike level. The k at S_{max} and S_{max} = 0.701 are listed, and the index assignments at S_{max} are circled accordingly: samples 2,7,8,10, samples 1,3,11,12, and samples 4-6,9 at k = 3. Ideally, only the 14 spiked hits will cluster according to these sample index assignments. The secondary baseline correction is thus necessary to remove misleading sample index assignments from the hitlist.
(C) Summed ^{2}D peaks for hit 26 following secondary baseline correction of the peaks in (A). Notably, because distinct approximately Gaussian "peaks" were not visible in (A), the concentration difference has been removed by baseline correction, leaving seemingly random (i.e., not spike level concentration-dependent) noise in (C).
(D) PCA scores plot of the summed ^{2}D peaks in (C), with the same color-coding by spike level. The index assignments at S_{max} are as follows: samples 5, 9, 11; samples 6, 10, 12; and samples 1-4, 7, 8 at k = 3. It is interesting to compare the very small S_{max} = 0.405 in (D) to the larger S_{max} = 0.701 in (B), which underscores the apparent randomness of the index assignments generated after baseline correction of these summed ^{2}D peaks.

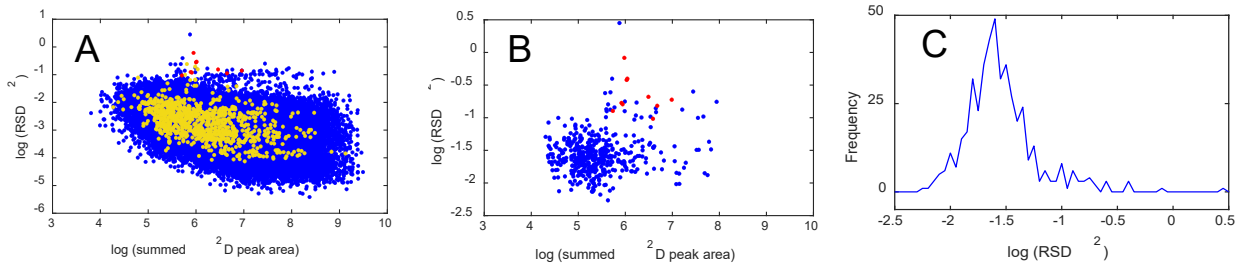

Figure D.3. Result of VRI-USI analysis of Comparison (2), 3-ppm versus neat JP8 jet fuel samples. The same tile-based software parameters used in Comparison (1) (Fig. 5.3) were used for Comparison (2) herein.

(A) Scatter plot of log (RSD^2) versus log (summed ^{2}D peak area) for all 460 hits in the RSD^2 ranked hitlist, wherein the top RSD^2 *m/z* for the spiked sulfur hits (i.e., true positives) are colored red, the secondary RSD^2 *m/z* for the true positives are gold, and the false positive RSD^2 *m/z* are blue. Interestingly, compared to the analogous plot for Comparison (1) in Fig. 5.3(A), the cloud of gold dots falls below a considerable portion of the blue dots, which reinforces the fact that fewer selective *m/z* will be present at lower concentrations. Correspondingly the red dots are intermingled amongst more blue dots compared to Fig. 5.3(A), but they are still largely distinguishable relative to the gold dots, which highlights the advantage of ranking with just these top RSD^2 *m/z* [1].
(B) Scatter plot of log (RSD^2) versus log (summed ^{2}D peak area) for just the top RSD^2 *m/z*, with the top true positive *m/z* are red and the top false positive *m/z* are blue. Note that only 10/14 spiked sulfur-containing compounds are discovered at this concentration via the VRI-USI method. The reason for this is two-fold; in our previous work, four analytes (tetrahydrothiophene, 3-acetyl-2,5-dimethylthiophene, 2-methylbenzothiophene, and 3-methylbenzothiopene) were undiscoverable in a tile-based F-ratio comparison of the 3-ppm vs. neat samples. However, using new neat sample files, 2-methylbenzothiophene and 3-methylbenzothiophene were discovered using F-ratio analysis herein (no shown for brevity). They were still not discovered with VRI-USI, however, due to the presence of spurious variation at a higher RSD^2 *m/z* at the same ^1t_R and ^2t_R. Future work will involve exploring all the RSD^2 *m/z* per hit pin location.
(C) Distribution of log (RSD^2) values for the top RSD^2 *m/z* using a bin size = 0.05.

Table D.2. List of top 35 hits obtained from the 3-ppm vs. neat VRI-USI comparison, including retention time information and S values at $k = 2$, 3, and 4 and the corresponding index assignments at S_{max}. The common index assignments are color coded, with green shading for clustering by spike level at $k = 2$ (samples 1-4, samples 5-8).

8/10 of the discovered spiked sulfur-containing analytes exhibit matching index assignments at S_{max}, specifically clustering by spike level at $k = 2$, along with one false positive hit further down the hitlist. The additional two spiked analytes have index assignments with resemblances to spike level clustering, though, with thiophene (hit 2) isolating sample 4 in its own cluster at $k = 3$ and 2-hexylthiophene (hit 10) splitting up the neat samples at $k = 3$. Thus VRI-USI appears highly robust for identifying low concentration differences.

In terms of a probabilistic argument, 9 matching index assignments (spike level clustering) out of 460 total hits corresponds to a random chance probability of 8.85×10^{-2} via Eq. (5.6) ($N = 70$ using Eq. (5.5)). The most frequently occurring set of index assignments in the hitlist is the blue shaded group, which occurs 38 total times and corresponds to a smaller random chance probability of 9.86×10^{-4} ($N = 8$). Thus, for Comparison (2), the true sample class assignments were not the predominant index assignments identified with VRI-USI. However, clustering by spike level is the second most frequently occurring set of index assignments, and manual examination of the blue shaded hits would reveal that sample 4 has spurious signal at a select few *m/z*. This example reveals an additional utility of VRI-USI analysis in identifying outliers in a dataset.

Hit Number	^1t_R (min)	^2t_R (s)	RSD^2 (top *m/z*)	Analyte	S, k=2	S, k=3	S, k=4	S_{max}	Index assignment at S_{max}
1	3.10	2.32	2.82 (45)	unknown	0.854	0.467	0.252	2	samples 1-3,5-8; sample 4
2	3.25	2.84	0.826 (45)	thiophene	0.667	0.725	0.610	3	samples 1,2,3; sample 4; samples 5,6,7,8
3	5.00	0.33	0.396 (97)	2-methylthiophene	0.935	0.730	0.540	2	samples 1-4; samples 5-8
4	3.05	1.28	0.396 (45)	unknown	0.551	0.505	0.362	2	samples 1,4,5,7; samples 2,3,6,8
5	5.20	0.37	0.379 (97)	3-methylthiophene	0.897	0.583	0.392	2	samples 1-4; samples 5-8
6	14.10	2.81	0.252 (154)	unknown	0.768	0.560	0.479	2	samples 1-4,5,7,8; sample 6
7	2.55	2.54	0.245 (58)	unknown	0.741	0.478	0.306	2	samples 1-3,5-8; sample 4
8	2.80	2.46	0.239 (58)	unknown	0.767	0.379	0.380	2	samples 1-3,5-8; sample 4
9	9.75	0.53	0.209 (97)	2-propylthiophene	0.929	0.795	0.475	2	samples 1-4; samples 5-8
10	18.40	0.82	0.189 (97)	2-hexylthiophene	0.739	0.816	0.781	3	samples 1-4; samples 5,8; samples 6,7
11	14.10	0.11	0.177 (198)	unknown	0.733	0.550	0.572	2	samples 1-4,5,7,8; sample 6
12	14.10	1.95	0.175 (168)	unknown	0.790	0.579	0.477	2	samples 1-4,5,7,8; sample 6
13	3.50	2.30	0.171 (45)	unknown	0.737	0.540	0.461	2	samples 1-3,5-8; sample 4
14	7.25	0.71	0.169 (112)	2,5-dimethylthiophene	0.740	0.645	0.451	2	samples 1-4; samples 5-8
15	9.40	0.53	0.161 (46)	1,4-oxathiane	0.804	0.531	0.341	2	samples 1-4; samples 5-8
16	16.75	2.82	0.160 (177)	unknown	0.364	0.467	0.361	3	samples 1,3,6,7; samples 2,5,8; sample 4
17	19.30	0.27	0.152 (134)	benzo[b]thiophene	0.902	0.668	0.402	2	samples 1-4; samples 5-8
18	3.05	2.88	0.144 (45)	unknown	0.508	0.289	0.152	2	samples 1-3,5-8; sample 4
19	3.00	0.80	0.140 (45)	unknown	0.564	0.423	0.470	2	samples 1,4,8; samples 2,3,5-7
20	3.05	1.82	0.135 (45)	unknown	0.291	0.339	0.356	4	samples 1,2,4; sample 8; samples 5-7; sample 3
21	14.10	2.51	0.133 (181)	unknown	0.774	0.625	0.347	2	samples 1-4,5,7,8; sample 6
22	16.90	0.03	0.133 (147)	unknown	0.555	0.531	0.371	2	samples 1-4,6,7; samples 5,8
23	23.85	0.35	0.127 (123)	2-chloroethyl phenyl sulfide	0.706	0.627	0.500	2	samples 1-4; samples 5-8
24	3.55	2.85	0.126 (45)	unknown	0.354	0.290	0.397	4	samples 1,8; samples 2,3,5; sample 4; samples 6,7

25	16.75	0.45	0.126 (177)	unknown	0.597	0.299	0.145	2	samples 1-3,5-8; sample 4
26	17.50	2.81	0.122 (133)	unknown	0.574	0.195	0.250	2	samples 1-3,5-8; sample 4
27	3.80	2.61	0.108 (45)	unknown	0.761	0.288	0.221	2	samples 1-3,5-8; sample 4
28	17.15	0.37	0.104 (119)	unknown	0.699	0.317	0.214	2	samples 1-3,5-8; sample 4
29	17.25	2.81	0.104 (133)	unknown	0.562	0.428	0.501	2	samples 1-3,5-8; sample 4
30	17.40	0.21	0.102 (177)	unknown	0.360	0.377	0.324	3	samples 1,7; samples 2-6; sample 8
31	6.30	1.00	0.101 (60)	unknown	0.366	0.460	0.476	4	samples 1,2; samples 3,4,8; sample 5; samples 6,7
32	14.10	0.71	0.0993 (146)	unknown	0.765	0.624	0.365	2	samples 1-4,5,7,8; sample 6
33	17.30	0.76	0.0993 (197)	unknown	0.539	0.447	0.268	2	samples 1-3,5-8; sample 4
34	16.20	0.87	0.0964 (196)	unknown	0.364	0.440	0.419	3	samples 1,2,3; sample 4; samples 5,6,7,8
35	17.80	0.75	0.0962 (125)	2-butyl-5-ethylthiophene	0.922	0.713	0.658	2	samples 1-4; samples 5-8

References

[1] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1644 (2021) 462092. https://doi.org/10.1016/j.chroma.2021.462092.

APPENDIX E

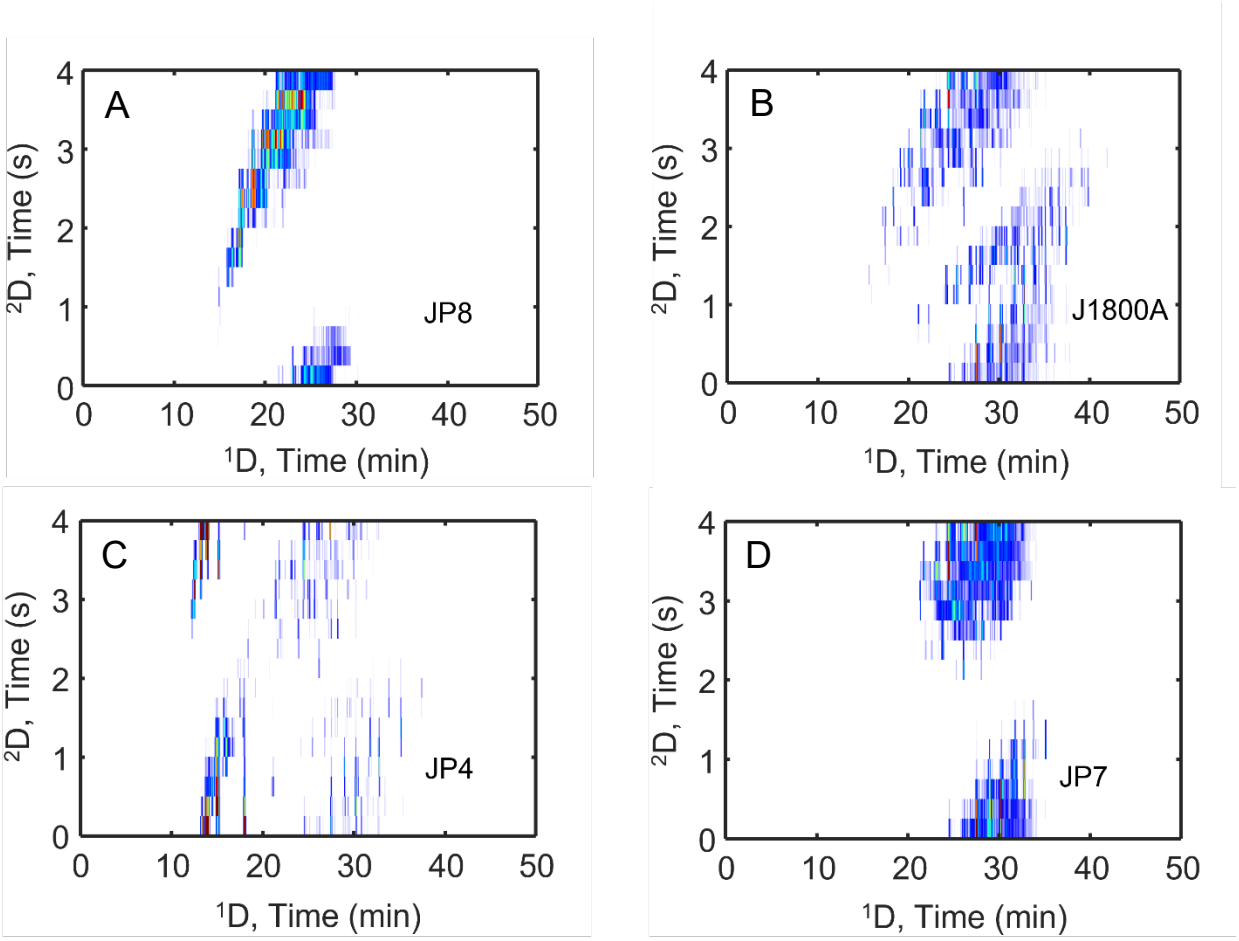

Figure E.1. ^{1}D × ^{2}D (^{3}D summed away) total ion current (TIC) chromatograms of four jet fuels prior to ^{2}D re-registration. (A) JP8. (B) J1800A. (C) JP4. (D) JP7.

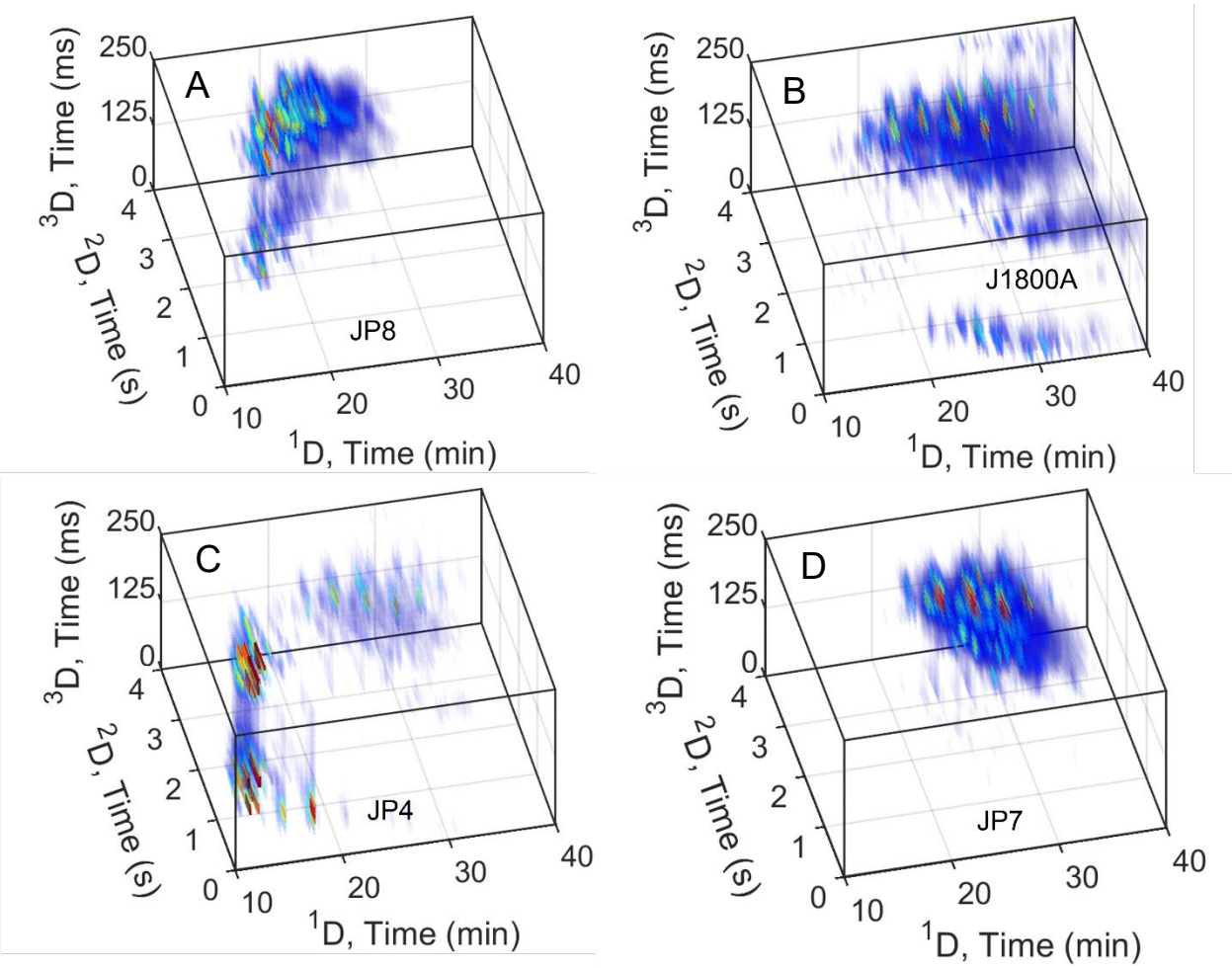

Figure E.2. 3D TIC chromatograms of four jet fuels after ^{2}D re-registration, but prior to ^{3}D re-regisration. (A) JP8. (B) J1800A. (C) JP4. (D) JP7. Wrap-around on ^{3}D can be clearly observed in (A) and (C) for peaks eluting beween 10 and 20 min on ^{1}D, and in (B) the aromatic region experiences noticseable wrap-around.