Metabolomics and the Development of Nontarget Discovery Analysis Methods for Two-dimensional Gas Chromatography Time-of-Flight Mass Spectrometry

Luke C. Marney

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Robert Synovec, Chair
Bo Zhang
Munira Khalil

Program Authorized to Offer Degree:
Chemistry

University of Washington

Abstract

Metabolomics and the Development of Nontarget
Discovery Analysis Methods for GC x GC – TOFMS Data

Luke C. Marney

Chair of the Supervisory Committee:
Robert Synovec
Chemistry

Two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) is a highly capable instrumental platform that produces complex and information-rich multi-dimensional chemical data. The complex data is computationally overwhelming, especially when many samples are analyzed with multiple injections for each sample. The highly information rich data from GC x GC – TOFMS benefits from elegant and comprehensive target and nontarget algorithmic methods. The development of a novel tile-based Fisher ratio method that greatly decreases the false-positive rate is described thoroughly in this dissertation. Also, the initially application of currently available instrumental and data analysis methods to the optimization of the preparation of mouse heart tissue metabolomics and the investigation of the pathophysiology of pressure overload hypertrophy is described herein. Then, novel algorithmic methods to solve issues seen with the difficult pixel-based analysis are described. Overall, GC x GC – TOFMS and the related data analysis tools investigated here aim to reduce the complex chemical data to glean the most meaningful information from an experiment.

Table of Contents

Ch 1: Introduction

1.1 Overview

Two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) is a highly capable instrumental platform that produces complex and information-rich multi-dimensional chemical data. The complex data can be initially overwhelming, especially when many samples (of various sample classes) are analyzed with multiple injections for each sample. Thus, the data must be analyzed in such a way as to extract the most meaningful information.

The investigation of naturally volatile and derivatized metabolites in biological tissues by comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) can provide highly complex and information rich data for comprehensive metabolomics analysis. Metabolomics is the comprehensive measurement and analysis of 100's to 1000's of metabolites in a biological sample. The addition of a second separation dimension with different chemical selectivity and fast scanning time-of-flight mass spectrometry offers benefits in chemical selectivity and overall peak capacity compared to traditional one-dimensional gas chromatography. Further, methods of derivatization, the most prominent being the silation of organic compounds, have extended the use of gas chromatography as an important metabolomics tool.

The highly information rich data from GC x GC – TOFMS benefits from elegant and comprehensive target and nontarget algorithmic methods. The thorough discussion and use of Fisher ratios, a nontarget data analysis method, is contained in this dissertation. Overall, the goal of GC x GC – TOFMS and the related data analysis tools is to reduce the complex chemical data resulting from GC x GC –TOFMS analysis to glean the most meaningful information from an experiment.

The demonstration and optimization of preparatory technique and previously available software methods are discussed in Chapters 2 and 3. Then specific software limitations are improved upon in novel ways in Chapters 4 and 5. In the end, the future application of GC x GC – TOFMS and the improved tile-based Fisher ratio software will be highly beneficial to future metabolomics studies.

1.2 Instrumentation, Experimental Design and Specimen Preparation

Two-dimensional gas chromatography coupled with fast scanning time-of-flight mass spectrometry (GC x GC – TOFMS) has become an important tool for comprehensive chemical analysis because it provides unparalleled separation ability and chemical selectivity. Advances in the development and routine use of derivatization reagents to enhance the volatility of biologically derived chemicals, such as metabolites, have raised the prominence of gas chromatography as an important tool in metabolomics. The addition of a second chromatographic dimension and coupling to fast scanning mass spectrometry (GC x GC – TOFMS) provides a remarkable increase in peak capacity as well as selectivity with the use of two unique stationary phases, allowing the separation of compounds that would otherwise co-elute in traditional one-dimensional gas chromatography (Fig. 1.1).

Gas chromatographic analysis requires that analytes be volatile, a prerequisite that traditionally has provided some methodological constraints to applications in metabolomic studies. GC x GC - TOFMS has been used in the analysis of naturally volatile metabolites and those metabolites made amenable to gas chromatography analysis via derivatization [1-7]. The metabolomics investigations of many sample types by GC × GC – TOFMS has been very fruitful, including, but not limited to; bacteria [8], yeast [9, 10], plant [11], brain [1, 2], urine [3],

blood [4, 5], spleen [6] and heart [7] as well as head space analysis of plant material [12, 13]. Liquid chromatography coupled to mass spectrometry (LC-MS) does not face the same issues as GC based metabolomics associated with making analytes volatile; however, there are other challenges of the technique applied to metabolomics such as limitations in chromatographically separating complex mixtures, the inherent retention time precision and minimization of unpredictable sample-based matrix effects such as ion-suppression caused by large macromolecules in liquid chromatography-mass spectrometry [14] Metabolomic separations conducted using gas chromatography coupled with the highly complex and information rich GC x GC - TOFMS data structure, reveal that the platform is extremely powerful for the analysis of metabolomics samples.

Owing to resolution, selectivity, and peak capacity, GC x GC-TOFMS has been used for targeted and nontargeted metabolomics. These two approaches to metabolomics vary somewhat in requirements and experimental outcomes. In targeted analysis, metabolites of interest are chosen prior to experimentation in order to test a particular metabolic hypothesis and chemical standards exist for sure identification. Traditional methods of quantification are used and the statistical comparison of samples is performed per metabolite, such as regression analysis or Student's t-test. In nontargeted analysis, metabolites of interest are discovered during instrumental and data analysis. A metabolic hypothesis may be proposed, but the specific metabolic phenotype is often entirely, or partly, unknown. A nontargeted approach creates unique challenges for data analysis and this new data centered approach to metabolic science (i.e. metabolomics) has the potential to decipher highly complex and new information from biological samples. New methods of comprehensive software tools are being developed so that useful information can be gleaned from the highly complex data.

Because metabolomic studies can be conducted on a variety of materials (e.g. cultured cells, minced tissue, urine) it is very important that the sample preparation is well controlled before proceeding with subsequent steps. Stability of samples due to degradation, changes in pH, clotting or other biologic or enzymatic mechanisms are important to understand and limit variation. For the biological applications discussed in subsequent chapters, careful weighing and uniform homogenization of solid tissue is of primary importance.

In chromatographic studies, cellular metabolism must be quenched at some point prior to injection of the sample. Quenching refers to the rapid inactivation of metabolism. In GC x GC - TOFMS quenching must occur prior to derivatiation and ideally should be fast and efficient in stopping metabolism and yet compatible with subsequent steps. Typically, quenching protocols are designed to stop cellular metabolism while simultaneously limiting degradation of metabolites. An optimal timeframe for quenching is on the order of 1 second, a process that typically occurs at low temperatures [15].

Derivatization via methoximation and trimethylsilyl (TMS) or tert-butyldimethylsilyl (TBDMS) reagents has been successfully applied to many different biological sample types and allows for the global derivatization of metabolites in comprehensive GC x GC – TOFMS metabolomics analysis. Silylation reagents and the derivatization products are sensitive to hydrolysis by water, therefore preparing derivatized samples requires all samples to be totally dried and tightly sealed to prevent inadvertent interaction with even small amounts of water vapor.

The derivatization method described herein is a two-step procedure, involving the methoximation of ketone groups by methoxyamine HCl and the trimethylsilation with N,O-bis(trimethylsilyl) trifluoroacetamide and trimethylchlorosilane (BSTFA + TMCS). In the first

step, methoxamine hydrochloride (HCl) is dissolved in an aprotic polar solvent such as pyridine, where it reacts with ketone groups, primarily sugars, in metabolomics samples. While methoximation helps to increase the volatility of ketone containing metabolites, it also opens the ring structure of sugars, stabilizes the open chain form while protecting intramolecular conversion of hemiacetal sugars to acetal sugars during subsequent silation [16]. This results in two chromatographic peaks that correspond to the unfolding of alpha and beta anomers of the sugar. The two peaks for each sugar, while making the separation slightly more complicated, can be an important tool for the identification of sugars by standard addition, because the ratio between the two methoxamine derivatives is constant [17].

In the second step of the derivitization, the silation reagent BSTFA is catalyzed by TMCS to replace reactive hydrogens with trimethylsilyl (TMS) groups. This reaction increases the vapor pressure of the metabolite, removes possible hydrogen bonding sites, and also increasing the thermal stability [18]. Hydroxyl, carboxyl, and amine functional groups are the main target of this step, with the amine TMS reaction the slowest of the three [17].

Derivatization chemistry for gas chromatography, particularly by silation, has been characterized for some time [18]. For metabolomics analysis, ketones are derivatized with methoxyamine HCl and silation reagents BSTFA, N-Methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA), or (MTBSTFA) are used to derivatize hydroxyl, carboxyl, and primary amine groups. In general, TBDMS derivatized metabolites (MTBSTFA reagent) are more resistant to hydrolysis than TMS, and are occasionally been used when a special focus on amino acid metabolism [19] is needed or isotopic uptake experiments are performed [20, 21].

The complexity of any resulting derivatized metabolomics mixture is most thoroughly separated by two-dimensional gas chromatography. A representative metabolomics

chromatogram (bacterial sample) of m/z 73, the mass of a single TMS group and thus the majority of metabolites, is shown in Fig. 1.1A, with a section of the chromatogram expanded for detail shown in Fig 1.1B. Fig 1.1C is a special view of this same data as it would appear in traditional one dimensional chromatography demonstrating the decrease in resolution and sensitivity of many of the small peaks shown separated by the second dimension in Fig. 1.1B.

Peak capacity is a metric of separation performance that can be used for evaluating different separation technology methods for applications such as metabolomics. It is a measure of how many peaks can be theoretically separated in a given chromatographic run, at a resolution of one. In traditional single dimensional gas chromatography (e,g, a typical metabolomics gas chromatogram with standard run time and 20 m x 180 um I.D. column), with commercial autoinjectors, a peak capacity of 450 in 45 min is common [22] The complexity of metabolomic samples typically warrants the adoption of various methods to maximize the peak capacity, the addition of a second chromatographic dimension being a powerful option. Attempts to increase the peak capacity by minimizing the extra-column band broadening due to injection via valve-based, or thermal injection systems can increase this for traditional single dimensional gas chromatography [23]; however, the use of a second column connected by a thermal modulator system for two-dimensional gas chromatography provides additional chemical selectivity as well as vastly increased peak capacity for comprehensive separation of metabolites. The peak capacity in Figure 1.1A is 4500.

Thermal modulators can be viewed as a real time thermal injection system, where all effluent from column 1 can be cryogenically trapped and thermally desorbed onto a second column without any loss, referred to as total transfer modulation [24]. The same 45 min metabolite separation, with a modulation period of 1.5 seconds, results in a peak capacity >4000

(the multiplication of the peak capacity in both dimensions), an order of magnitude higher than single dimensional gas chromatography. Fast scanning time-of-flight mass spectrometry is needed for comprehensive two-dimensional gas chromatography separation, because second dimension peak widths are narrow. A data collection frequency of 100-500 spectra per second is possible with the LECO Pegasus III TOFMS, which is sufficient for sampling the narrow peaks (30 ms at base) seen on the second column [23].

1.3 Comprehensive Nontarget Data Analysis

Nontarget techniques aim to comprehensively analyze entire complex chromatograms to discover important analytes or chemical fingerprints while requiring few user inputs and minimizing the need for prior information about the samples, as opposed to target analysis, where the analyte(s) of interest are known prior to data collection and analysis. Nontarget techniques can be supervised or unsupervised, where supervision refers to external calibration or prior classification of chromatograms as they relate to the experimental design.

The 2D misalignment of GC × GC – TOFMS peaks across different samples can make non-targeted analysis challenging [25]. The two approaches used are can be categorized as either pixel-based or peak table-based. Pixel-based analysis refers to the analysis of the chromatographic data without any initial preprocessing first (other than baseline correction and/or normalization which are required) and the individual data points (i.e. pixels) are compared across samples. Peak-table based analysis of GC × GC –TOFMS data refers to the use of deconvolution, peak spectral matching, and arrangement of results into a table first, then comparisons of table entries of the same metabolite are compared. The advantage of pixel-based analysis is that it is unbiased to mass spectral characteristics or peak shape, but inadvertent

misalignment of peaks across multiple GC × GC chromatograms makes these algorithms challenging.

Nontarget algorithms often differ from targeted algorithms in their scope, not their method. For example, the targeted mass spectral matching or deconvolution algorithms that exist for a small group of target analytes are developed into nontarget algorithms by becoming more automated, being reprogrammed to increase computational speed, and being redesigned to comprehensively process the entire chromatogram rather than just a few targeted sub-regions. An example of broadening the scope of a target method is development of nontarget PARAFAC. An automated algorithm was developed that selected an appropriate number of factors for PARAFAC models applied to select sub-regions in GC × GC–TOFMS chromatograms, thus making the process of factor selection an objective operation [26]. This is an important advancement in converting PARAFAC into an objective data analysis method requiring fewer user inputs than prior PARAFAC algorithms. Furthermore, this PARAFAC algorithm was converted into an automated comprehensive algorithm that automatically resolves all peaks in an *entire* GC × GC–TOFMS chromatogram [27]. PARAFAC was already a powerful chemometric technique in that it improves S/N extremely well for selected sub-regions, but being converted into an algorithm that can process entire multidimensional chromatograms with fewer user inputs makes it even more powerful. However, the analysis of entire chromatograms by PARAFAC is computationally intense for some applications.

The LECO ChromaTOF software, a peak table-based approach provided by the instrument manufacturer LECO, can comprehensively analyze entire GC × GC–TOFMS chromatograms in keeping with nontarget analysis goals. Newer versions of ChromaTOF allow

the built-in use of chemometrics such as principal component analysis (PCA) and Fisher ratio analysis applied to the peak tables. Recently, ChromaTOF software was used for peak identification followed by Excel spreadsheet organization and submission to an analysis of variance (ANOVA) algorithm written in SAS (SAS, Cary, NC) [28]. ChromaTOF peak tables for glycerolysis products of animal and plant oils were imported into SAS for normalization and submission to ANOVA, revealing sources of variation in the data [29]. These peak tables were also submitted to PCA in Unscrambler (CAMO Software AS, Oslo, Norway) to reveal that monoglycerides and diglycerides differentiated the sample classes. ChromaTOF peak tables for honeys were submitted to linear discriminant analysis (LDA) and PLSDA with support vector machines and Pearson VII universal kernel to reveal chemical similarities and differences between the honey samples [30]. Peak signal volumes from ChromaTOF peak tables of fatty acid methyl esters (FAMEs) from six different animals were submitted to PCA in an effort to model biological variability of FAMEs among individual animals [31]. ChromaTOF peak tables for metabolic extracts from two lines of transgenic *Artemisia annua* L. were imported into SIMCA and submitted to PLSDA with analysis of variable importance in projection scores for feature selection and to determine metabolites that differentiated the classes [32].

Unsupervised cluster classification algorithms indicate the similarities and dissimilarities of entire sample profiles. Researchers used PCA to discover correlations among different manufacturing processes of dense non-aqueous phase liquids (DNAPL) with GC × GC–TOFMS chromatograms of the DNAPL, revealing a single platform alternative to previous traditional multiplatform approaches [33]. Chemical fingerprinting algorithms such as hierarchical cluster analysis (HCA) can analyze data in terms of Euclidian distance between clusters of sample chromatograms that often include calibration chromatograms [34], [35] and [33]. Unsupervised

classification algorithms are subject to overfitting, and clustering of samples in unsupervised algorithms such as PCA is informative only when the user can assume that the clustering is correct. Validation methods for assessing clustering and avoiding over-fitting are generally beneficial. One method is ANOVA simultaneous component analysis (ASCA), which is a semi-unsupervised technique that limits rotational ambiguity in PCA with prior knowledge about the experimental design. ASCA was shown to diminish the overfitting problem of PCA models built for low-sample size, high-dimensionality data sets [36]. Various classification methods provided by the publicly available WEKA data mining and machine learning software (M. Hall, WEKA, http://www.cs.waikato.ac.nz.offcampus.lib.washington.edu/ml/weka/, 2010) were compared and validated by researchers who classified breast-cancer tumors based on unsupervised and supervised analysis of GC × GC–HRMS separations of tissue sample metabolites [37].

Supervised machine learning algorithms and PLS models are subject to the same potential overfitting shortcomings as unsupervised methods. The limitations of both supervised and unsupervised techniques allow for combination approaches that leverage the strength of both types of techniques while avoiding their pitfalls. Combining unsupervised and supervised techniques into a fully automated comprehensive analytical platform for 2D separations resulted in the discovery of meaningful variations in bacterial metabolomics studies [38] and yeast metabolomics studies [39]. In these two reports, the combined methods included Fisher-ratio data reduction, PCA, comparative signal-ratio analysis, PARAFAC deconvolution and quantification, qualification via mass spectral matching, ANOVA statistical analysis, and *t*-testing. This interplay between supervised and unsupervised techniques was also applied to discover biomarkers of uroepithilieal tumors by utilizing PCA and distance-to-model (DModX)

algorithms for outlier determination, followed by PLSDA for identification of metabolite biomarkers [40].

1.4 Fisher Ratio Feature Selection

Two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) is a highly capable instrumental platform that produces complex and information-rich multi-dimensional chemical data. The complex data can be initially overwhelming, especially when many samples (of various sample classes) are analyzed with multiple injections for each sample. Thus, the data must be analyzed in such a way as to extract the most meaningful information. The pixel-based and peak table-based algorithmic use of Fisher ratios has been used successfully in the past to reduce the multi-dimensional data down to those chemical compounds that are changing between sample classes relative to those that are not.

Herein, I report on the development of a computationally fast novel tile-based Fisher-ratio software that addresses challenges due to 2D retention time misalignment without explicitly aligning the data, which is a problem for both pixel-based and peak table-based methods. In Chapter 1, the important extraction and experimental optimization is performed so that pixel-based F-ratio analysis is able to detect metabolites of interest that change in small concentrations due to desired experimental conditions. Concurrently, in Chapter 2, the Fisher-ratio software was refined based on knowledge gained during this project to maximize the sensitivity contrast of true positives against a background of potential false positives and noise. Lastly, in Chapter 4, further development of a redundant hit removal algorithm and fisher ratio threshold are demonstrated allowing the automation and further enhancement of contrast between true positives and potential false positive and noise.

References

[1] Snyder, L. R., Hoggard, J. C., Montine, T. J, Synovec, R. E. J. Chromatogr. A 1217 (2010) 4639–4647.

[2] Snyder, L. R., Cruz-Aguado, R., Sadilek, M., Galasko, D., Shaw, C. A., Montine, T. J. Toxicol. Appl. Pharmacol. 240 (2009) 180–188.

[3] Rocha, S. M., Caldeira, M., Carrola, J., Santos, M., Cruz, N., Duarte, I. F. J. Chromatogr. A 1252 (2012) 155–163.

[4] Beckstrom, A., Humston, E., Snyder, L., Synovec, R., Juul, S. J. Chromatogr. A 1218 (2011) 1899-1906.

[5] Beckstrom, A. C., Tanya, P., Humston, E. M., Snyder, L. R., Synovec, R. E., Juul, S. E. Pediatr. Res. 71 (2012) 338–344.

[6] Welthagen, W., Shellie, R. A., Spranger, J., Ristow, M., Zimmermann, R., Fiehn, O. Metabolomics 1 (2005) 65–73.

[7] Marney, L.C., Kolwicz, S. C., Tian, R., Synovec, R. E. Talanta 108 (2013) 123–130.

[8] Yang, S., Sadilek, M., Synovec, R.E., Lidstrom, M.E. J. Chromatogr. A 1216 (2009) 3280–3289.

[9] Mohler, R.E., Dombek, K.M., Hoggard, J.C., Pierce, K.M., Young, E.T., Synovec, R.E. Analyst 132 (2007) 756-767.

[10] Humston, E.M., Dombek, K.M., Tu, B.P., Young, E.T., Synovec, R.E. Anal. Bioanal. Chem. 401 (2011) 2387–2402.

[11] Pierce, K.M., Hope, J.L., Hoggard, J.C., Synovec, R.E. Talanta 70 (2006) 797–804.

[12] Humston, E.M., Knowles, J.D., McShea, A., Synovec, R.E. J. Chromatogr. A 1217 (2010) 1963–1970.

[13] Humston, E.M., Zhang, Y., Brabeck, G.F., McShea, A., Synovec, R.E. J. Sep. Sci. 32 (2009) 2289–2295.

[14] Annesley, T.M. Clin. Chem. 49 (2003) 1041–1044.

[15] Ewald, J.C., Heux, S., Zamboni, N. Anal. Chem. 81 (2009) 3623−3629.

[16] Little, J.L. J. Chromatogr. A 844 (1999). 1–22.

[17] Kanani, H., Chrysanthopoulos, P.K., Klapa, M.I. J. Chromatogr. B. 871 (2008) 191–201.

[18] Pierce, E. (1968) Silylation of Organic Compounds, *Pierce Chemical Co.* Rockford, IL.

[19] van-der-Klift, E. J.C., Vivó-Truyols, G., Claassen, F.W., van-Holthoon, F.L., van-Beek, T.A. J. Chromatogr. A 1178 (2008) 43−55.

[20] Hutschenreuther, A., Kiontke, A., Kirkenmeier, G., Birkemeyer, C. Anal. Methods 4 (2012) 1953−1063.

[21] Huang, X.D., Regnier, F.E. Anal. Chem. 80 (2008) 107-14.

[22] Wilson, R.B., Hoggard, J.C., Synovec, R.E. Anal. Chem. 84 (2012) 4167–4173.

[23] Fitz, B.D., Wilson, R.B., Parsons, B.A., Hoggard, J.C., Synovec, R.E. J. Chromatogr. A 1266 (2012) 116–123.

[24] Liu, Z., Zhang, M., Phillips, J. B. J. *Chromatogr. Sci.* 28 (1990) 567−571.

[25] Fraga, C.G., Prazen, B.J., Synovec, R.E. Anal. Chem. 73 (2001) 5833–5840.

[26] Hoggard, J.C., Synovec, R.E. Anal. Chem. 80 (2008) 6677.

[27] Hoggard, J.C., Siegler, W.C., Synovec, R.E., J. Chemom. 23 (2009) 421.

[28] Johanningsmeier, S.D., McFeeters, R.F. J. Food Sci. 76 (2011) C168.

[29] Indrasti, D., Che Man, Y.B., Chin, S.T., Mustafa, S., Mat Hashim, D., Abdul Manaf, M., J. Am. Oil Chem. Soc. 87 (11) (2010) 1255.

[30] Stanimirova, I., Üstün, B., Cajka, T., Riddelova, K., Hajslova, J., Buydens, L.M.C., Walczak, B. Food Chem. 118 (2010) 171.

[31] Chin, S.-T., Che Man, Y.B., Tan, C.P., Hashim, D.M. J. Am. Oil Chem. Soc. 86 (2009) 949.

[32] Ma, C., Wang, H., Lu, X., Wang, H., Xu, G., Liu, B. Metabolomics 5 (2009) 497.

[33] Cordero, C., Liberto, E., Bicchi, C., Rubiolo, P., Schieberle, P., Reichenbach, S.E., Taod, Q. J. Chromatogr. A 1217 (2010) 5848.

[34] Reichenbach, S.E., Tian, X., Tao, Q., Stoll, D.R., Carr, P.W. J. Sep. Sci. 33 (2010) 1365.

[35] Gröger, T., Schäffer, M., Pütz, M., Ahrens, B., Drew, K., Eschner, M., Zimmermann, R. J. Chromatogr. A 1200 (2008) 8.

[36] Verouden, M.P.H., Westerhuis, J.A., van-der-Werf, M.J., Smilde, A.K. Chemom. Intell. Lab. Syst. 98 (2009) 88.

[37] Reichenbach, S.E., Tian, X., Tao, Q., Ledford, E.B., Wu, Z., Fiehn, O. Talanta 83 (2011) 1279.

[38] Yang, S., Sadilek, M., Synovec, R.E., Lidstrom, M.E. J. Chromatogr. A 1216 (2009) 3280.

[39] Humston, E.M., Dombek, K.M., Tu, B.P., Young, E.T., Synovec, R.E. Anal. Bioanal. Chem. 401 (2011) 2387.

[40] Pasikanti, K., Norasmara, J., Cai, S., Mahendran, R., Esuvaranathan, K., Ho, P., Chan, E. Anal. Bioanal. Chem. 398 (2010) 1285.

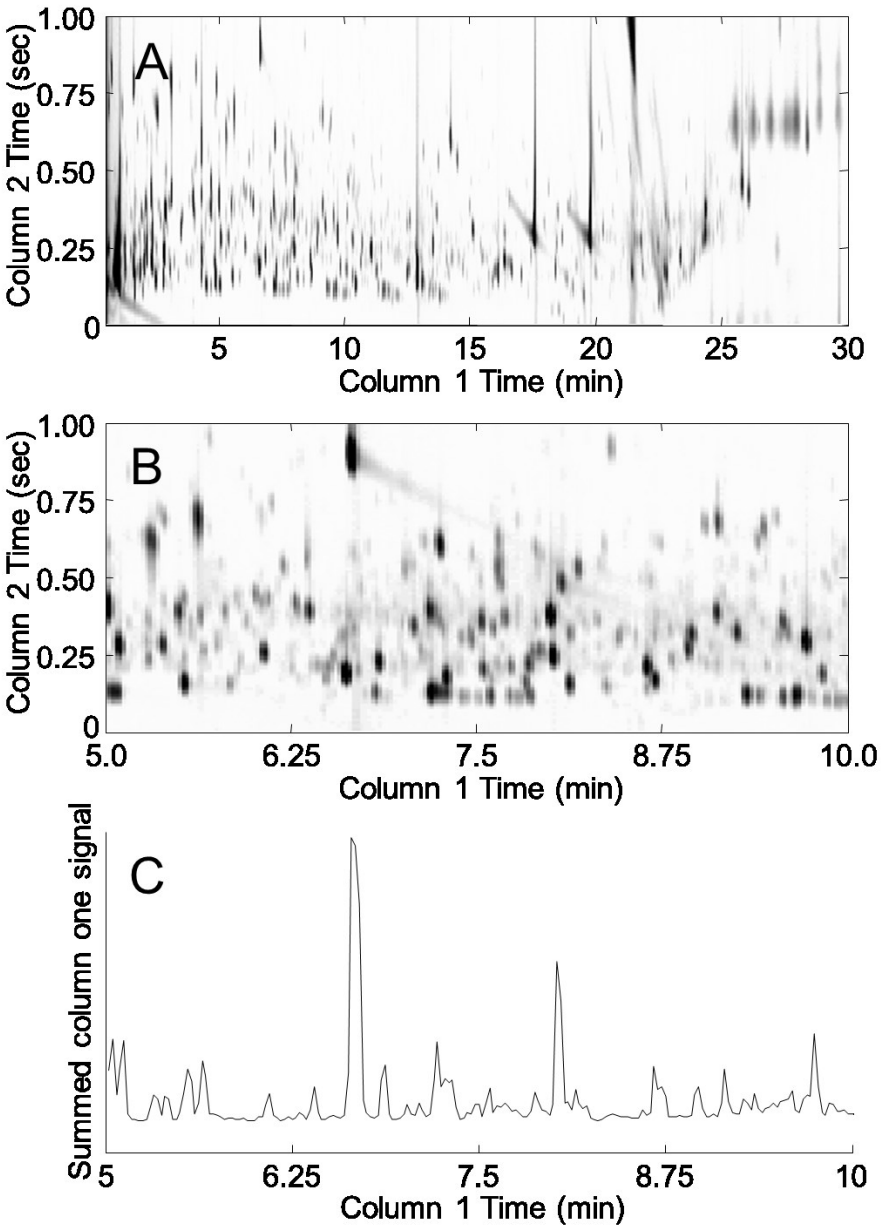


Figure 1. (A) A representative chromatogram from the derivatization of a methanol extraction of bacteria is shown (*Clostridium acetobutylicum*, chromatogram produced by an Agilent 6890 gas chromatograph modified with a 4D thermal modulator and a LECO Pegasus III TOFMS). The selective *m/z* of 73 is shown because it corresponds to the ionization of a TMS group and thus the majority of metabolites in the sample. (B) A zoomed in section is shown highlighting the benefit of the second column separation. Peaks eluting at the same time on the first column are separated by the second column. (C) The first column summed signal, summing up all data points in the second dimension, is shown to simulate what this metabolomics sample would look like with traditional one-dimensional GC. Many metabolites are overlapped in the one-dimension view or suffer from poor signal-to-noise.

Chapter 2:

Sample Preparation Methodology for Mouse Heart Metabolomics using Comprehensive Two-Dimensional Gas Chromatography coupled with Time-of-Flight Mass Spectrometry

The investigation of naturally volatile and derivatized metabolites in mammalian tissues by comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) can provide the data for a comprehensive analysis of the pathophysiology of disease processes. When relative quantification is needed for hypothesis testing, the preparation of sample tissue must provide clear evidence that a quantitative relationship exists between the final detected signal and the amount of metabolite in the tissue. In this chapter, I report the optimization of a metabolite extraction method for mouse heart tissue for GC × GC – TOFMS analysis. A recursive extraction experiment was initially performed to measure the extraction efficiency of representative target metabolites (sugars, tricarboxylic acid cycle metabolites, amino acids, lipid and signaling molecules) in the aqueous fraction of a three-phase extraction system involving tissue, methanol:water, and chloroform. Some metabolites suffered from incomplete extraction with a single extraction of ~ 40 mg in 600 µl organic and 400 µl aqueous phases, possibly caused by saturation effects. Subsequent experiments, calibrating resulting metabolite signal to the mass of heart tissue extracted, demonstrated that doubling the solvent volumes and a lower tissue mass was needed to provide accurate relative quantification of the derivatized mouse heart metabolome. I demonstrate quantitative extraction of metabolites from ~ 20 mg of heart tissue using 1200 µl organic phase (chloroform) and 800 µl aqueous phase (methanol : water in equal parts by volume).

2.1 Introduction

The coupling of comprehensive two-dimensional (2D) gas chromatography with time-of-flight mass spectrometry (GC × GC – TOFMS) for metabolomics investigations combines the high peak capacity of GC × GC with fast scanning and highly selective detection of TOFMS for the analysis of naturally volatile metabolites and those metabolites made amenable to GC

analysis via derivatization [1-11]. Metabolomics investigations into many mammalian tissue types by GC × GC – TOFMS has been very fruitful for a variety of sample types, including brain [1,2], urine [3], blood [4,5], and spleen [6]. The metabolomics analysis of the heart tissue of transgenic mice is of significant interest [12-18], but the detailed sample preparation of heart tissue for derivatized metabolomics analysis via GC × GC – TOFMS has not yet been reported in the literature. There are critical issues and challenges in the sample preparation of heart tissue for metabolomics with the GC × GC – TOFMS instrumental platform, specifically the mass of the tissue used and the volume of the extraction solvent(s). Investigations into other tissue types, mainly brain [1], spleen [6], and muscle tissue [13], vary in the tissue masses analyzed from a low of ~ 5 mg to a high of ~ 100 mg extracted into a 1 to 2 ml total extraction solvent (total volume of two-phase liquid extractions). To address these challenges, reported here is the investigation and development of a sample preparation method to facilitate the analysis of the heart metabolome in mice by GC × GC – TOFMS.

The selective knockout of genes in mammalian tissues enables the tissue specific testing of metabolic hypotheses [12, 15-18]. However, when relative quantification is needed for hypothesis testing, preparation of the tissue must provide clear evidence that a quantitative relationship exists between the final detected signal and the amount of metabolite in the tissue. This is difficult with the analysis of metabolites in solid tissues such as the heart, because hearts cannot be synthesized with known concentration levels of metabolites within them, and applying the standard addition method is impractical for comprehensive analysis. Unless complete extraction of all metabolites can be achieved, one can only be certain of the quantification of the metabolites contained in the extraction solvent. For accurate quantification there needs to be a

reproducible relationship between the amount of metabolite in the tissue, the amount extracted, and the amount detected.

The advantage of using a mouse model to test hypotheses about specific genotype to phenotype interactions is the ability to reflect upon human physiology without using human subjects. The development and optimization of this extraction method has been applied to the analysis of heart tissue from mice with a cardiac-specific deletion of acetyl-CoA carboxylase 2 during pressure overload hypertrophy [12]. However, the details of the sample preparation methodology as they relate to the analytical chemistry issues of reproducibility, saturation, and tissue mass calibration should be discussed. I explore first the details regarding the sample preparation methodology, with primary focus on the sample mass-to-solvent volume ratio of the extraction conditions. Later, I will demonstrate the metabolomics application of this extraction technique, by applying a pixel-based Fisher ratio method and targeted PARAFAC of the mouse heart metabolome during pressure overload hypertrophy.

Three specimen preparatory optimization experiments were performed in this study. A recursive extraction experiment was initially performed to measure the extraction efficiency of representative target metabolites (sugars, tricarboxylic acid cycle metabolites, amino acids, lipid and signaling molecules) in the aqueous fraction of a three-phase extraction system involving heart tissue, methanol:water, and chloroform. Only metabolites in the aqueous fraction of the two-phase liquid extraction were dried, derivatized with methoxyamine hydrochloride and N,O-bis-trimethylsilyl-trifluoroacetamide (BSTFA) plus trimethylchlorosilane (TMCS), analyzed by GC × GC – TOFMS, and quantified utilizing chemometric software via PARAFAC deconvolution [19, 20]. Some metabolites will be shown to suffer from incomplete extraction

with a single extraction of ~ 40 mg heart tissue in 600 µl organic and 400 µl aqueous phases, possibly caused by saturation effects. Therefore, subsequent to the recursive extraction experiment, a tissue mass calibration experiment was performed using 1200 µl organic and 800 µl aqueous phases to optimize the mass-to-volume ratio used for future investigations of mouse heart tissue. To achieve the goal of relative quantification of the mouse heart metabolome, as measured by derivatization and GC x GC –TOFMS, I demonstrate quantitative extraction of metabolites from ~ 20 mg of heart tissue in 1200 µl organic and 800 µl aqueous phases. Finally, in a third experiment I use this optimized method to assess the biological variation across ten hearts extracted with ~ 20 mg of heart tissue in 1200 µl organic and 800 µl aqueous phases and quantifying each metabolite peak signal after chemometric deconvolution and relative quantification of targeted metabolites.

2.2. Experimental

In the experimental approach taken to optimize the extraction of metabolites from mouse heart tissue, a few commonalities exist with all three experiments: (1) recursive extraction experiment, (2) a tissue mass calibration experiment, and (3) a reproducibility experiment. Briefly, the approach to extract metabolites from heart tissue involves removing the heart from the mouse, pulverizing the tissue in liquid nitrogen with a morter and pestle, adding 2:1 chloroform:methanol (by volume) with actual volumes as specified for each experiment, homogenizing the mixture on ice for 30 s with a Tissue Tearor homogenizer (Biospec Products Inc. Bartlesville, OK) and sonicating for 20 s at 50 J. Equal parts of water and methanol (the actual volumes used are detailed below for each experiment) are added and the sample is centrifuged at 4 ^{o}C for phase separation of organic and aqueous layers. The aqueous layer is

dried under nitrogen stream and the dried metabolites are derivatized with methoxyamine hydrochloride (Sigma-Aldrich, St. Louis, MO) and BSTFA + TMCS (Supelco, Bellefonte, PA).

Removal of heart tissue, metabolite derivatization, and GC × GC – TOFMS analysis are the same for all experiments. Metabolite derivatization and GC × GC – TOFMS analysis is detailed in the Metabolite Derivatization and Instrumental Analysis section. Prior to any metabolite extraction, mice were fasted for 4-5 hours prior to harvesting and at the same time of day. The hearts were removed from the mice by cutting below the auricles so that only ventricles were left. The heart was immediately placed on a gauze pad to remove blood from the exterior. Modified Wollenberger tongs (pre-cooled in liquid nitrogen) were used to freeze-clamp and rapidly submerge the heart in liquid nitrogen. Heart tissue was stored at -80 ^{o}C until the day of extraction. On the day of extraction, heart tissue was weighed with an analytical balance and pulverized under liquid nitrogen with a mortar and pestle.

Two experiments were performed for the optimization of heart tissue extraction. The first is the recursive extraction experiment of ~ 40 mg of heart tissue. A schematic of the experimental procedure is shown in Fig. 2.1. Briefly, a 600 µl volume of cold 2:1 chloroform:methanol was added to 40.5 mg powdered and frozen heart tissue (ventricles). This mixture was homogenized on ice for 30 s with a Tissue Tearor homogenizer (Biospec Products Inc., Bartlesville, OK). The homogenate was sonicated for 20 s using a pulse sonicator at 50 J. A volume of 200 µl chloroform and 200 µl of deionized water were added and the sample was centrifuged at 3220 g for 20 min at 4°C (Eppendorf, 5810R, Hauppauge, NY). After phase separation, the aqueous (methanol:water) layer was removed, dried under nitrogen gas stream, with the metabolites derivatized and analyzed in duplicate by GC × GC – TOFMS. Then an

additional 200 µl methanol and 200 µl deionized water was added to the chloroform and tissue pellet. The sample was vortexed, centrifuged, and the aqueous layer was removed for derivatization. Additional volumes of 200 µl methanol and 200 µl deionized water were added recursively, resulting in five total extractions from a single piece of heart tissue. For ~ 40 mg of heart tissue in 1 ml total volume, some metabolites are completely extracted by the first extraction, reflected in little-to-no metabolite extracted during the recursive extractions. However, some metabolites are not completely extracted from the first extraction, indicated as a large amount of metabolite extracted by the successive extractions.

We desired to achieve an extraction method that completely extracts all metabolites important for future biochemical investigations in a single extraction. The low volume extraction method (600 µl organic and 400 µl aqueous) does not provide complete extraction of a few representative important metabolites with ~ 40 mg sample in 1 ml total volume. Additionally, the incomplete extraction of these metabolites could cause incomplete or variable partitioning of other metabolites, Therefore, I performed a tissue mass calibration experiment to find an appropriate mass-to-volume ratio of the extraction with a higher total amount of total volume used, as compared to the recursive extraction experiment. I hypothesized that by doubling the solvent volumes used in the recursive extraction experiment, we would not significantly dilute low signal-to-noise analytes to an undetectable level, but would dilute previously saturated metabolites sufficiently for accurate relative quantification (Fig. 2.2). Three heart tissue samples were cut into four pieces: the masses of tissue were nominally 40 mg, 20 mg, 10 mg, and 5 mg (the exact mass of a given sample to 3 significant figures is plotted in subsequent figures). A volume of 1200 µl of 2:1 chloroform:methanol was added to each weighed, powdered, and frozen heart tissue sample (ventricles). Each mixture was homogenized on ice for ~ 30-45 s. The

homogenate was sonicated for 20 s using a pulse sonicator at 50 J. 400 µl of chloroform and 400 µl of distilled water were added to each sample, and the samples were centrifuged at 3220 g for 20 min at 4°C. After phase separation, aqueous (methanol:water) layers were removed, dried under nitrogen gas stream, derivatized and analyzed in duplicate by GC × GC – TOFMS. Linear regression with Microsoft Excel 2010 (Microsoft Corp. Redmond, WA) was used to evaluate the calibration of tissue mass to the mathematically resolved (deconvoluted) peak signal, obtained using the chemometric data analysis tool PARAFAC [19, 20].

A third experiment was performed to determine the reproducibility of the preparation method in combination with the biological variation across multiple mice. Ten hearts were removed and 20 mg pieces were extracted with the 1200 µl organic and 800 µl aqueous phase combination. Samples were dried, derivatized, and analyzed with the same preparation method as all other experiments. Precision was calculated from the PARAFAC deconvoluted peak signals for each metabolite.

2.2.1 Metabolite Derivatization and Instrumental Analysis:

With a Hamilton syringe, 30 µl of 20 mg/ml methoxyamine hydrochloride (Sigma Aldrich) dissolved in pyridine was added to the dried aqueous metabolites. The heart tissue samples were vortexed and placed in a 30°C oven for 90 min. Then 70 µl of BSTFA + TMCS in pyridine was added and the samples were vortexed and placed in a 60 °C oven for 60 min.

For GC × GC – TOFMS analysis, 1 µl of sample extract was injected without a split by an autosampler equipped Agilent 6890 gas chromatograph (Agilent Technologies, Palo Alto, CA, USA). The gas chromatograph was modified with a 4D thermal modulator (LECO Corp., St. Joseph, MI, USA) that transfers effluent from the primary column (20 m x 250 µm i.d. x 0.5 µm

Rtx-5MS, Restek, Bellefonte, PA, USA) to the secondary column (2 m x 180 µm i.d. x 2 µm Rtx 200MS, Restek, Bellefonte, PA, USA). The modulation period, the time between two episodes that effluent from the primary column is cryogenically frozen, thermally desorbed, injected, and separated in secondary column, was 1.5 s. The temperature program for primary column began at 60 °C for 0.25 min and then ramped at 8 °C/min to a final temperature of 280 °C. The temperature program for the secondary column was analogous, except that it was consistently held 10 °C higher than primary column. The carrier gas was helium at a volumetric flow rate of 1 ml/min. The GC × GC was coupled to a Leco Pegasus III TOFMS (Leco Corp., St. Joseph, MI, USA). Ions were generated within the TOFMS by electron impact ionization and collected from mass channels of *m*/*z* 40 to *m*/*z* 600 at an acquisition rate of 100 spectra/s. Data were collected by LECO ChromaTOF software version 3.32 (LECO Corp., St. Joseph, MI, USA).

2.2.2 Data Analysis

LECO's ChromaTOF software v 3.32 (St. Joseph, MI, USA) was used to collect GC × GC–TOFMS data. Metabolite identification was determined by mass spectral match value and retention time similarity with metabolite standards. Peak signals for relative quantification and precision analysis were attained for each metabolite using a target PARAFAC GUI [19, 20] developed in-house. The in-house software imports the raw data collected with ChromaTOF v 3.32 and deconvolutes the pure component chromatographic peak profile and the pure mass spectrum of an individual metabolite from overlapping peaks and background noise for quantification. The PARAFAC software provides baseline correction, because the baseline noise and the chromatographic peak signal profiles of target metabolites as well as any interference (in both chromatographic dimensions) are deconvoluted.

2.3 Results and Discussion

A representative GC × GC – TOFMS chromatogram from the aqueous fraction of 20 mg of heart tissue extracted by 1200 µl chloroform and 800 µl equal parts methanol:water (by volume) is shown in Fig. 2.3. The single ion 2D chromatogram at *m*/*z* 73 is used to show those metabolites that contained trimethylsilyl groups from derivatization. Hundreds of metabolites are detected. Evaluation of the complexity of this type of sample benefits from using the two separation dimensions provided by GC × GC. There are many metabolites that would be overlapped if only a single dimension of GC were used. Because of the secondary column, a larger number of metabolites can be separated in the 2D space, relative to only one dimension.

In early experiments with heart tissue, I encountered questionable quantitative results and suspected that the extraction solvent conditions were suffering from saturation of some metabolites due to the small extraction volumes used. I designed an experiment to recursively extract mouse heart tissue to see how much metabolite remained unextracted after initial extraction, and found that a few metabolites remained in the two phases (chloroform and tissue pellet) after initial extraction (i.e., the recursive extraction experiment). Indeed, extraction of fumarate from ~ 40 mg of heart tissue gave the same peak signal as extracting ~ 20 mg of heart tissue when extracted in 600 µl chloroform and 400 µl equal parts methanol:water.

The metabolite signals for eight representative metabolites from PARAFAC signal deconvolution of each aqueous extract by applying the recursive extraction experiment (Fig. 2.1) are shown in Fig. 2.4. Fumarate, glycerol, and citric acid all show that ~ 50% of the peak signal recovered from the initial extraction (40 mg in 600 µl organic and 400 µl aqueous phases) was still left in the heart tissue or organic layer after the second addition of methanol:water. Many analytes did not show this dramatic under-extraction with the lower solvent volume conditions. Approximately 10% of the metabolite signal recovered from the initial extraction was recovered

with a subsequent extraction of the metabolites succinyl-CoA, myo-inositol, glutamate, malate, and glycerol-3-phosphate. These results, combined with pilot experimental observation, confirmed our hypothesis that metabolite extraction was generally incomplete for the conditions employed in the recursive extraction experiment. The cause for incomplete extraction was possibly due to saturation of the aqueous solvent, and required further experimentation to establish a reproducible quantitative extraction protocol. The use of ~ 40 mg in 600 µl organic and 400 µl aqueous was quantitatively adequate for certain metabolites (providing acceptable precision and accuracy). However, these initial solvent extraction conditions would not be acceptable for experiments aimed at the relative quantification and concentration comparison of fumarate, glycerol, and citric acid in multiple heart samples, because the extent of extraction, and thus metabolite signal, was limited by the amount of solvent used, and not reflective of the metabolite concentration in the tissue.

Extracting ~ 40 mg of heart sample in 600 µl organic and 400 µl aqueous was not the appropriate heart mass and/or solvent volume conditions to provide a comprehensive extraction of all metabolites for relative quantification—even though it was acceptable for some metabolites. Therefore, I performed a tissue mass calibration experiment, employing heart mass and extraction solvent conditions whereby the amount of metabolite extracted could be shown to be representative of the amount of metabolite in the heart tissue. Based on the recursive extraction experiment I hypothesized that doubling the solvent volumes used would sufficiently dilute previously saturated metabolites, while not diluting too many low signal-to-noise analytes as to make them undetectable. Therefore, following the tissue mass calibration experimental procedure outlined in Fig. 2.2, three heart samples were cut into four decreasing mass pieces (nominally 40 mg, 20 mg, 10 mg, and 5 mg), and extracted with 1200 µl chloroform and 800 µl

equal parts methanol:water (twice the amount of solvent used relative to the recursive extraction experiment).

Quantitative results for the eight representative metabolites following the tissue mass calibration experiment (Fig. 2.2) are shown in Fig. 2.5. Four different masses of heart tissue from three different hearts are shown (exact masses plotted, with duplicate injections of each piece of heart). Signal for each metabolite was determined based upon PARAFAC signal deconvolution software [19 ,20] and a linear relationship between the mass of heart extracted and the signal detected for each metabolite was evaluated using the Pearson's correlation coefficient (by Microsoft Excel 2010). The linearity of response of metabolite signal after PARAFAC deconvolution of the eight representative metabolites is of paramount importance for the future relative quantification of these metabolites. The relative standard deviation (RSD) as a percent is reported for the four mass groupings across the three heart tissue samples, and is dominated by the biological variation for a given metabolite at each mass. For each of the eight representative metabolites, doubling the amount of solvent used previously is demonstrated in Fig. 2.5 to provide good linear calibration and thus confidence that the metabolite signal observed by extraction, derivatization, and GC x GC –TOFMS analysis is indicative of the concentration of metabolite in heart tissue. Notably, the previously saturated fumarate, glycerol, and citric acid now calibrate signal to mass of the heart tissue, solving our initial problem outlined in the recursive extraction experiment. The calibration plots in Fig. 2.5 of signal as a function of the heart mass suggest that no saturation effects are seen in the extracts utilizing 2 ml (1200 µl organic phase (chloroform) and 800 µl aqueous phase (methanol : water in equal parts by volume)). In addition, the total amount of metabolite extracted from 40 mg of heart tissue analyzed in the mass calibration experiment is equal or greater to the amount extracted by the

sum of all the recursive extractions, as inferred by a comparison of signals in Figures 2.4 and 2.5. For example, the total summed PARAFAC signal in Fig. 2.4(A) for fumarate from every extraction of a 40 mg portion of heart in the recursive extraction experiment is about 6 (arbitrary signal units), while the extraction of 40 mg of heart in 2 ml using the optimized condition in the mass calibration experiment in Fig. 2.5(A) results in a PARAFAC signal ranging from about 6 to10 (with the range a result of the biological variation). We were confident that by doubling the amount of solvent used (to 1200 µl chloroform and 800 µl equal parts methanol:water), all other metabolites should also be quantitative, because fumarate, glycerol, and citric acid are of similar signal to other high, medium, and low (respectively) intensity metabolite signals in the full 2D chromatogram.

For the mass tissue calibration experiment, the Pearson's correlation coefficient is 0.71 and 0.67 for succinyl-CoA and glycerol respectively, because of large residuals caused by biological variation, where the injection variation for the method is demonstrated by duplicate injections in Fig. 2.5, and was determined to be ~ 10%RSD which is consistent with similar methodology applied with this instrumentation [10]. The biological variation observed across most metabolites (20-30% RSD) is comparable to GC × GC – TOFMS analysis of other mammalian tissues, but should be analyzed for more metabolites to determine if use of this sample preparation method in the future will detect statistically significant changes in metabolism [1-6]. Overall, the biological variation dominates over the method variation.

Finally, a reproducibility experiment for the extraction method was performed with ten mice, to assess the added effect of biological variation, by extracting ~ 20 mg of heart tissue in 1200 µl organic and 800 µl aqueous phases and quantifying each metabolite peak signal after PARAFAC deconvolution and relative quantification of targeted metabolites. Here I applied the

sample preparation method outlined in Fig. 2.2 that yielded the acceptable results in Fig. 2.5. A mass of ~ 20 mg was used for two reasons. It represents an adequate sample from the heart tissue, as 20 mg falls in the middle of the mass range used for the tissue mass calibration experiment and thus is demonstrated to not suffer from saturation since signal linearly correlates with mass (see plots in Fig. 2.5), and many metabolite signals are not readily visible when only 10 mg of tissue are used (see Fig. 6). For the GC × GC – TOFMS separations, the primary column retention times and secondary column retention times are reported in Table 2.1, along with the PARAFAC deconvoluted peak signal precision (RSD) of each metabolite across ten different mice. An average RSD of 20-30% is typical for mammalian tissues and is sufficiently small to find small (biologically relevant) differences between samples with adequate sample size. Next, this method is used for the evaluation of metabolomic changes under pressure overload hypertrophy in a cardiac-specific acetyl-CoA carboxylase isoform 2 knockout mouse model [12].

2.4 Conclusions for Specimen Preparatory Optimization

The investigation of mammalian metabolism in a variety of tissues by GC × GC – TOFMS has been successfully performed and reported in the literature, but the preparation and metabolomics analysis of heart tissue by GC x GC – TOFMS has not yet been reported. I uncovered and addressed a few critical issues and challenges in the sample preparation of heart tissue for derivatized GC × GC – TOFMS. In particular, I discovered that particular metabolites were saturating the extraction solvent and thus I needed to calibrate the tissue mass and extraction solvent volume to the resulting metabolite signal.

The experimental optimization of mouse heart tissue extraction was performed. A recursive extraction experiment demonstrated that the metabolites fumarate, glycerol, and citric

acid suffer from incomplete extraction and possible extraction solvent saturation if ~ 40 mg of heart tissue is used in a total volume of 1 ml of extraction solvent (using 600 µl chloroform and 400 µl 1:1 methanol:water). A tissue mass calibration experiment using twice the amount of solvent used in primary extraction of the recursive extraction experiment demonstrated the effective dilution and subsequent successful relative quantification of eight metabolites, including the previously saturated fumarate, glycerol, and citric acid. Finally, the use of ~ 20 mg of heart tissue extracted in 1200 µl chloroform and 800 µl 1:1 methanol:water was successful for the comprehensive relative quantification of the heart tissue metabolome. The method's variation including biological variation, extraction variation, and comprehensive GC x GC – TOFMS instrumental variation, is reported herein (an average RSD of 20-30%). This RSD level will allow us to detect biologically relevant metabolite changes due to genetic differences in the response to pressure overload hypertrophy [12].

2.5 Metabolomics Application of Optimized Specimen Preparatory Technique, Pixel-based F-ratio Software, and Targeted Analysis with PARAFAC

The heart is capable of oxidizing a wide range of carbon substrates and this metabolic ability is necessary to maintain an intense, variable workload in the midst of hormonal and nutritional changes. Shifts in substrate utilization in the heart occur in response to chromic stresses (both hemodynamic and metabolic) [21-24]. It is unknown how metabolic shifts contribute to the pathogenesis of contractile dysfunction. For instance, diabetes mellitus results in an increased dependence on fatty acids for oxidative energy production in the heart and this increase in lipid metabolism may contribute to impaired cardiac function [25]. Also, increased rates of fatty acid oxidation after an ischemic event have been suspected of exacerbating

reperfusion injury [26]. The direct causal relationship between contractile abnormalities and metabolic deregulation has yet to be reported.

In response to pressure overload cardiac hypertrophy, the heart decreases fatty acid oxidation which results in an increased reliance on carbohydrate substrates. This substrate shift contributes to the progression to contractile failure [27]. Many attempts have been made to provide insight regarding the impact that substrate shifts have on pressure overload remodeling after the ischemic cardiac event.

The shift toward carbohydrate metabolism following ischemic reperfusion and pressure overload injury is inefficient in utilizing carbon substrates for ATP production during increased energy demand, leading to impaired myocardial energetics and depletion of contractile reserve [28, 29]. Enhancing fatty acid oxidation during the development of pathological cardiac hypertrophy could possibly prevent metabolic remodeling and preserve myocardial energetics and function. The genetic knockout of acetyl-CoA carboxylase isoform 2 (ACC2) prevents metabolic remodeling during injury by preventing the production of malonyl-CoA, the presence of which is the major metabolic signal pathway for a decrease in fatty acid oxidation. The lack of production of malonyl-CoA could potentially allow maintenance of beneficial myocardial energetics and thus contractile function.

2.5.1 Method for GC x GC – TOFMS Metabolomics

Metabolic profiling was performed on frozen ventricular tissue extracted in 2:1 chloroform:methanol as detailed previously in this chapter. Aqueous and organic fractions were separated, dried down under nitrogen and derivitized with methoxyamine-HCl and BSTFA + TMCS. Samples were separated and detected by comprehensive two-dimensional gas

chromatography coupled with time-of-flight mass spectrometry GC x GC-TOFMS [33]. Pixel-based Fisher ratio algorithm analysis was performed on GC x GC-TOFMS chromatograms as previously described [34, 35] to find locations in the data that contained up- or down-regulated metabolite concentrations. Briefly, the algorithm calculates a Fisher ratio (between-class variation divided by within-class variation) in and across two groups of samples, for all mass channels collected at each point in the two-dimensional chromatograms. F-ratio values are then summed across all mass channels for each chromatographic point (Sum of F-ratios). Two algorithm improvements were utilized that reduce the number of false positive hits from baseline noise, via a signal-to-noise mask, and remove redundant F-ratio values from wide or tailing peaks, by analyzing the similarity of the mass spectrum of neighboring chromatographic locations. F-ratio values were plotted against a randomized group of an equal number of controls and ACC2H-/- to produce a theoretical null distribution. Initial detection of metabolic features was made via Peak Table analysis (LECO ChromaTOF software version 3.2, LECO Corp., St. Joseph, MI, USA). GC x GC-TOFMS data for each relevant metabolite hit from the F-ratio analysis were subjected to Parallel Factor Analysis (PARAFAC) to deconvolute the metabolite peaks to produce 3D information (purified mass spectrum, metabolite retention time on the first GC column, and metabolite retention time on the second GC column). Additionally, PARAFAC provided the peak signal volume which is used to determine the relative concentration of the individual metabolites normalized to the total ion current of each chromatogram [36]. Peak signal values were then normalized to a log2 fold change over control animal-sham surgery (CON-sham, aortic constriction not performed, but anesthetic was given).

2.5.2 Metabolomics Application Results

We performed metabolomics profiling in control (CON) and knockout (ACC2H-/-) hearts to explore potential changes in the metabolic network in response to chronic elevation of fatty acid oxidation under normal and pressure overload conditions. Heart extracts were derivatized, analyzed by GC x GC – TOFMS and chromatographic data were processed by the pixel-based Fisher ratio algorithm [39]. Two dimensional Sum of F-ratio plots were produced to compare the 4 groups (Fig. 2.7) and more than 800 hits were sorted in descending order and plotted against a null-distribution (Fig 2.8). With this approach, we observed minimal effects of transverse aortic constriction (TAC, pressure overload condition) and moderate effects of ACC2 deletion on the metabolomic profile. Identification of the top 15 hits yielded relatively few known metabolites. PARAFAC analysis did not show significant differences in nearly half of the hits (Tables 2.2 – 2.5). These findings suggest that there was no global cardiac metabolite shift during pressure-overload hypertrophy and/or with increased fatty acid oxidation as a result of ACC2 deletion.

We also incorporated a targeted approach specifically examining metabolites involved in fatty acid, glucose, and amino acid metabolism as well as tricarboxylic acid (TCA) cycle intermediates. We found significant decreases in a number of amino acids and several metabolites in glucose metabolism with minor changes in TCA cycle intermediates in ACC2H-/- sham hearts (Fig. 2.9), suggestive of decreased glucose reliance and increased amino acid consumption in ACC2H-/- hearts. These results indicate that relative amounts of metabolites are maintained during the early development of pressure-overload hypertrophy despite a shift toward increased glucose utilization. Furthermore, adaptations to increased fatty acid oxidation in ACC2H-/- probably include reduced glucose uptake and utilization with increased consumption of amino acids, which may contribute to the resistance to metabolic remodeling during pathological hypertrophy.

References

[1] Snyder, L. R., Hoggard, J. C., Montine, T. J., Synovec, R. E. J. Chromatogr. A 1217 (2010) 4639–4647.

[2] Snyder, L. R., Cruz-Aguado, R., Sadilek, M., Galasko, D., Shaw, C. A., Montine, T. J. Toxicol. Appl. Pharmacol. 240 (2009) 180–188.

[3] Rocha, S. M., Caldeira, M., Carrola, J., Santos, M., Cruz, N., Duarte, I. F. J. Chromatogr. A 1252 (2012) 155–163.

[4] Beckstrom, A., Humston, E., Snyder, L., Synovec, R., Juul, S. J. Chromatogr. A 1218 (2011) 1899-1906.

[5] Beckstrom, A. C., Tanya, P., Humston, E. M., Snyder, L. R., Synovec, R. E., Juul, S. E. Pediatr. Res. 71 (2012) 338–344.

[6] Welthagen, W., Shellie, R. A., Spranger, J., Ristow, M., Zimmermann, R., Fiehn, O. Metabolomics 1 (2005) 65–73.

[7] Humston, E. M., Dombek, K. M., Hoggard, J. C., Young, E. T., Synovec, R. E. Anal. Chem. 80 (2008) 8002–8011.

[8] Humston, E. M., Dombek, K. M., Tu, B. P., Young, E. T., Synovec, R. E. Anal. Bioanal. Chem. 401 (2011) 2387–2402.

[9] Mohler, R. E., Tu, B. P., Dombek, K. M., Hoggard, J. C., Young, E. T., Synovec, R. E. J. Chromatogr. A 1186 (2008) 401–411.

[10] R. E. Mohler, K. M. Dombek, J. C. Hoggard, K. M. Pierce, E. T. Young, R. E. Synovec, Analyst 132 (2007) 756-767.

[11] Yang, S., Sadilek, M., Synovec, R. E., Lidstrom, M. E., J. Chromatogr. A 1216 (2009) 3280–3289.

[12] Kolwicz, S. C., Olson, D. P., Marney, L. C., Garcia-Menendez, L., Synovec, R. E., Tian, R. Circ. Res. 111 (2012) 728-738.

[13] Dyck, J. R. B., Hopkins, T. A., Bonnet, S., Michelakis, E. D., Young, M. E., Watanabe, M., Kawase, Y., Kawase, Y., Jishage, K., Lopaschuk, G. Circulation 114 (2006) 1721–1728.

[14] Koves, T. R., Ussher, J. R., Noland, R. C., Slentz, D., Mosedale, M., Ilkayeva, O., Bain, J., Stevens, R., Dyck, J., Newgard, C., Lopaschuk, G., Muoio D. Cell Metab. 7 (2008) 45–56.

[15] Lopaschuk, G. D., Wambolt, R. B., Barr, R. L. J. Pharmacol. Exp. Ther. 264 (1993) 135–144.

[16] Lopaschuk, G. D., Witters, L. A., Itoi, T., Barr, R., Barr, A. J. Biol. Chem. 269 (1994) 25871–25878.

[17] Saddik, M., Gamble, J., Witters, L. A., Lopaschuk, G. D. Journal of Biological chemistry 268 (1993) 25836–25845.

[18] Dyck, J. R. B., Hopkins, T. A., Bonnet, S., Michelakis, E. D., Young, M. E., Watanabe, M., Kawase, Y., Jishage, K., Lopaschuk, G. Circulation 114 (2006) 1721–1728.

[19] Hoggard, J. C., Siegler, W. C., Synovec, R. E. J. Chemom. 23 (2009) 421–431.

[20] Hoggard, J. C., Synovec, R. E. Anal. Chem. 79 (2007) 1611–1619.

[21] Lopaschuk GD, Ussher JR, Folmes CD, Jaswal JS, Stanley WC. Myocardial fatty acid metabolism in health and disease. Physiol Rev. 2010;90:207–258.

[22] Allard M, Schonekess B, Henning S, English D, Lopaschuk G. Am J Physiol. 267 (1994) H742–H750.

[23] Abel E, Litwin S, Sweeney G. Cardiac remodeling in obesity. Physiol Rev. 88 (2008) 389–419.

[24] Taegtmeyer H, McNulty P, Young ME. Circulation 105 (2002) 1727–1733.

[25] Larsen TS, Aasum E. Cardiovasc. Drugs Ther. 22 (2008) 91–95.

[26] Lopaschuk GD, Wambolt RB, Barr RL. J Pharmacol. Exp. Ther. 264 (1993) 135–144.

[27] Taegtmeyer H, Sen S, Vela D. Ann. N. Y. Acad Sci. 1188 (2010) 191–198.

[28] Ingwall JS. Cardiovasc. Res. 81 (2009) 412-419.

[29] Neubauer S. N. Engl. J. Med. 356 (2007) 1140-1151.

[33] Humston, EM, Dombek, KM, Hoggard, JC, Young, ET Synovec, RE Anal. Chem.80 (2008) 8002-11.

[34] Mohler, RE, Dombek, KM, Hoggard, JC, Pierce, KM, Young, ET Synovec, RE Analyst 132 (2007) 756-767.

[35] Pierce, KM, Hoggard, JC, Hope, JL, Rainey, PM, Hoofnagle, AN, Jack, RM, Wright, BW Synovec, RE Anal. Chem.78 (2006) 5068-5075.

[36] Hoggard, JC, Siegler, WCSynovec, RE J. Chemometrics 23 (2009) 421-431.

[39] Koves TR, Ussher JR, Noland RC, Slentz D, Mosedale M, Ilkayeva O, Bain J, Stevens R, Dyck JR, Newgard CB, Lopaschuk GD, Muoio DM. Cell. Metab.7 (2008) 45-56.

Table Captions

Table 2.1. For the reproducibility experiment with the analysis of ten mice, a subset of metabolite identifications are provided in the extraction of ~ 20 mg of heart tissue in 1200 µl chloroform and 800 µl methanol:water (1:1 by volume). Metabolite signals for the determination of quantitative precision (RSD) were calculated after signal deconvolution by PARAFAC.

Table 2.2. Presented are the top 15 hits from the calculated F-ratio between CON-sham and ACC2H-/--sham, sorted in descending order. Column 1 (RT1) and column 2 (RT2) retention times (measured in seconds) were used to identify the hits from an in-house library. Ratio was determined by dividing the normalized peak volume of the CON-sham by ACC2H-/--sham. P-value was determined by t-test. ND, PARAFAC analysis not determined for these unknown analytes.

Table 2.3. Presented are the top 15 hits from the calculated F-ratio between CON-sham and CON-TAC, sorted in descending order. Column 1 (RT1) and column 2 (RT2) retention times (measured in seconds) were used to identify the hits from an in-house library. Ratio was determined by dividing the normalized peak volume of the CON-sham by CON-TAC. P-value was determined by t-test. * $P < 0.05$. ND, PARAFAC analysis not determined for these unknown analytes.

Table 2.4. Presented are the top 15 hits from the calculated F-ratio between CON-TAC and ACC2H-/--TAC, sorted in descending order. Column 1 (RT1) and column 2 (RT2) retention times (measured in seconds) were used to identify the hits from an in-house library. Ratio was determined by dividing the normalized peak volume of the CON-TAC by ACC2H-/--TAC. P-value was determined by t-test. * $P < 0.05$. ND, PARAFAC analysis not determined for these unknown analytes.

Table 2.5. Presented are the top 15 hits from the calculated F-ratio between ACC2H-/--sham and ACC2H-/--TAC, sorted in descending order. Column 1 (RT1) and column 2 (RT2) retention

times (measured in seconds) were used to identify the hits from an in-house library. Ratio was determined by dividing the normalized peak volume of the ACC2H-/--sham by ACC2H-/--TAC. P-value was determined by t-test. * $P < 0.05$. ND, PARAFAC analysis not determined for these unknown analytes.

Figure Captions

Figure 2.1. Recursive extraction experiment procedure to evaluate the effectiveness of extracting ~ 40 mg of heart tissue in 600 µl chloroform and 400 µl methanol:water (1:1 by volume).

Figure 2.2. The tissue mass calibration experiment schematic is shown detailing the procedure used to evaluate the linearity of response of the metabolite signal relative to mass of heart extracted for targeted metabolites. Four nominal masses of heart tissue were taken from three different mice: ~ 40 mg, ~ 20 mg, ~ 10 mg, and ~ 5 mg.

Figure 2.3. A representative GC × GC – TOFMS chromatogram at *m*/*z* 73 from a single heart tissue sample is presented, showing all trimethylsilated metabolites from the derivatization of the aqueous layer of the extraction of 20 mg of mouse heart tissue.

Figure 2.4. Representative quantitative results of the recursive extraction experiment (using procedure outlined in Fig. 1) of ~ 40 mg of heart tissue are shown. Replicate injections of each recursive extraction are shown. Signal for each metabolite was determined using PARAFAC signal deconvolution software [20, 21]. The actual PARAFAC signals have been scaled down by a factor of 100,000 for clarity.

Figure 2.5. Quantitative results for the eight representative metabolites following the tissue mass calibration experiment (using procedure outlined in Fig. 2) based upon extracting four different masses of heart tissue from three different hearts is shown. Different hearts are shown as different symbols. Duplicate injections of each piece of heart were analyzed. Signal for each metabolite was determined using PARAFAC software [20, 21]. The linear relationship between

the mass of heart extracted and the signal detected for each metabolite is provided. The PARAFAC signals have been scaled down by 100,000 for clarity. The biological variation across six injections (three mice) is indicated as the RSD for each grouping of heart tissue masses (masses varied less than 4% RSD for each group). Linear regression is shown with Pearson's correlation coefficient, with results indicating good linearity. The high residuals and low correlation coefficient observed is dominated by the biological variation. Linearity was also observed for many other metabolites (not shown for brevity).

Figure 2.6. A zoomed in section of two GC × GC chromatograms from the mass calibration experiment is shown, plotting *m/z* 73 showing all trimethylsilated metabolites from the derivatization of the aqueous layer of two pieces of heart tissue from the same mouse. The two GC × GC chromatograms are from (A) 20 mg of heart tissue and (B) 10 mg of heart tissue from the same mouse extracted in 1200 µl chloroform and 800 µl methanol:water (1:1 by volume). Some important metabolite signals are not readily visible if only 10 mg is used.

Figure 2.7. 2D Sum of Fisher Ratio plots from heart tissue extracts comparing CON and ACC2H-/- hearts with and without hypertrophy. The points in the plot are used to identify locations of chromatographic peaks of interest and to create a list of chemical components that are potentially different between groups: A) CON-sham vs. ACC2H-/--sham; B) CON-sham vs. CON-TAC; C) CON-TAC vs. ACC2H-/--TAC; D) ACC2H-/--sham vs. ACC2H-/--TAC. RT, retention time.

Figure 2.8. F-ratio values were plotted in descending order for each hit number obtained between the 4 group comparisons (Blue, CON-sham vs. CON-TAC; Purple, CON-sham vs. ACC2H-/--sham; Green, ACC2H-/--sham vs. ACC2H-/--TAC; Light Blue, CON-TAC vs. ACC2H-/--TAC, n = 6 each group). Samples were randomized into two classes with an equal number of each experimental group and again submitted for F-ratio analysis. This randomization models what a theoretical null-change between samples would be (Red, null distribution, n=12 each group). The inset enlarges the plot for the top 15 hits as determined by the F-ratio analysis.

Figure 2.9. Heat map represents metabolites from glucose metabolism, TCA cycle intermediates, amino acids, and other significantly different metabolites in CON and ACC2H-/- hearts 4 wks after TAC or sham surgery. Color coding for each metabolite was assigned using a log2 fold

change versus the mean value of CON (n = 6 each group). *P < 0.05 vs. CON-sham; **P < 0.05 vs. CON-TAC,

Metabolite	Column 1 Retention Time (s)	Column 2 Retention Time (s)	Quantitative Precision Detected Signal (RSD)
lactate	481.5	0.41	21 %
glucose	1278	0.21	40 %
fumarate	778.5	0.69	15 %
malate	922.5	0.48	19 %
succinate	711	0.90	47 %
succinyl-CoA	747	0.62	39 %
myo-inositol	1422	0.28	26 %
coenzyme A	633	1.20	11 %
glycerol	633	0.21	26 %
glycerol-3-phosphate	1084.5	0.77	25 %
glucose-1-phosphate	1194	0.20	27 %
glutamate	877	1.41	26 %
citric acid	1130	0.47	24 %
serine	615	0.71	27%
thiamine diphosphate	747	0.71	14%
nicotinamide adenine dinucleotide	1281	0.85	39%
uracil	699	0.40	24%

Table 2.1

Hit #	Sum of F-Ratios	RT1	RT2	Identification	Ratio	P value
1	2493	711.0	0.94	CoA Fragment	1.58	0.0054
2	1509	462.0	0.71	unknown	1.10	0.3614
3	1222	451.5	0.32	unknown	1.48	0.0001
4	1025	570.0	0.60	unknown	1.44	0.0008
5	1007	604.5	0.77	unknown	2.30	0.0006
6	915	991.5	0.39	creatinine enol	1.06	0.2451
7	821	481.5	0.41	Lactic Acid	4.47	0.7451
8	802	892.5	0.16	unknown	ND	ND
9	778	742.5	0.42	unknown	ND	ND
10	740	744.0	0.17	unknown	ND	ND
11	720	1422.0	0.28	myo-inositol	2.51	0.0212
12	650	448.5	0.45	unknown	ND	ND
13	611	796.5	0.51	unknown	ND	ND
14	551	612.0	0.43	unknown	ND	ND
15	508	897.0	0.32	unknown	ND	ND

Table 2.2

Hit #	Sum of F-Ratios	RT1	RT2	Identification	Ratio	P-value
1	237	661.5	0.20	unknown	1.42	0.0375
2	226	1422.0	0.29	myo-inositol	2.08	0.0287
3	218	711.0	0.91	CoA-Fragment	1.12	0.0961
4	194	892.5	0.15	unknown	0.66	0.5826
5	173	744.0	0.16	unknown	ND	ND
6	156	451.5	0.32	unknown	1.26	0.0034
7	149	612.0	0.44	unknown	ND	ND
8	133	568.5	0.63	unknown	1.18	0.2559
9	131	1141.5	0.13	unknown	0.44	0.1988
10	129	462.0	0.71	unknown	1.16	0.0046
11	119	358.5	0.20	unknown	ND	ND
12	105	382.5	0.25	unknown	ND	ND
13	86	648.0	0.72	unknown	ND	ND
14	83	922.5	0.48	malate	1.26	0.1509
15	77	300	0.15	unknown	ND	ND

Table 2.3

Hit #	Sum of F-Ratios	RT1	RT2	Identification	Ratio	P-value
1	3156	711.0	0.91	CoA Fragment	2.23	0.0001
2	1818	604.5	0.77	unknown	4.61	0.0001
3	1529	382.5	0.25	unknown	ND	ND
4	1222	597.0	0.54	Unknown	1.34	0.0813
5	623	481.5	0.41	Lactic Acid	1.49	0.0419
6	563	958.5	1.02	5-oxoproline	35.17	0.0001
7	502	567.0	0.63	unknown	1.08	0.0013
8	442	586	0.19	unknown	ND	ND
9	433	451.5	0.32	unknown	1.19	0.0763
10	414	897.0	0.32	unknown	ND	ND
11	388	796.5	0.51	unknown	ND	ND
12	330	991.5	0.39	creatinine enol	1.17	0.3176
13	278	475.5	0.19	unkown	ND	ND
14	267	603	0.53	unknown	ND	ND
15	265	711.0	1.13	unknown	ND	ND

Table 2.4

Hit #	Sum of F-Ratios	RT1	RT2	Identification	Ratio	P-value
1	346	711.0	0.91	CoA Fragment	1.58	0.0396
2	141	1140.0	0.13	unknown	2.41	0.1378
3	151	568.5	0.65	unknown	0.91	0.3532
4	116	1063.5	0.20	unknown	4.73	0.0878
5	95	1420.5	0.29	myo-inositol	2.31	0.1272
6	92	922.5	0.48	malic acid	1.35	0.1424
7	86	936.0	0.42	unknown	ND	ND
8	72	382.5	0.25	unknown	ND	ND
9	63	357.0	0.20	unknown	ND	ND
10	53	460.5	0.73	unknown	0.99	0.8728
11	47	660.0	0.20	unknown	0.82	0.6834
12	37	547.5	0.43	unknown	ND	ND
13	35	447.0	0.46	unknown	ND	ND
14	33	568.5	0.61	unknown	0.91	0.9616
15	26	796.5	0.50	unknown	ND	ND

Table 2.5

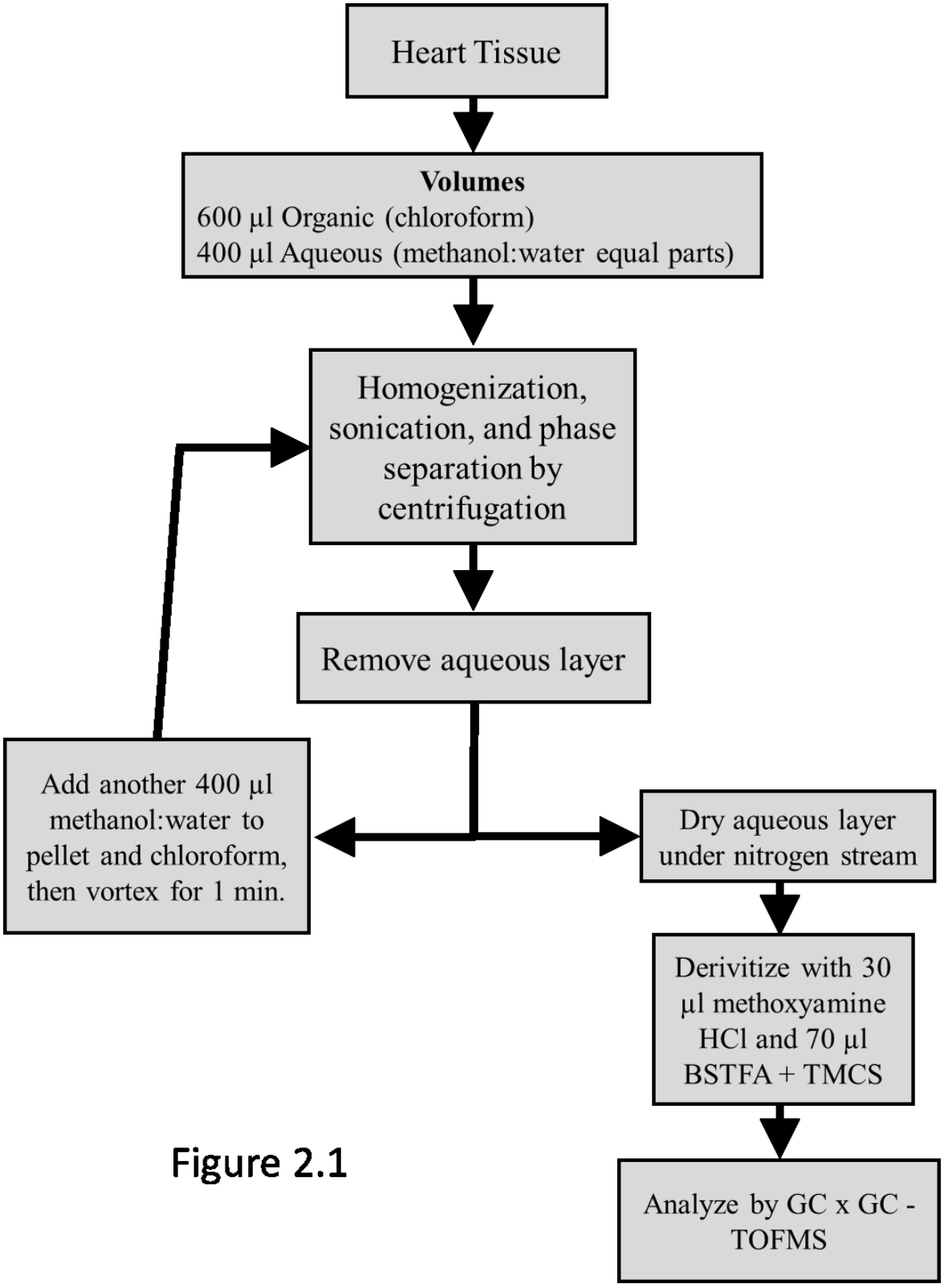

Heart Tissue
Volumes
600 µl Organic (chloroform)
400 µl Aqueous (methanol:water equal parts)
Homogenization, sonication, and phase separation by centrifugation
Remove aqueous layer
Add another 400 µl methanol:water to pellet and chloroform, then vortex for 1 min.
Dry aqueous layer under nitrogen stream
Derivitize with 30 µl methoxyamine HCl and 70 µl BSTFA + TMCS
Analyze by GC x GC - TOFMS

Figure 2.1

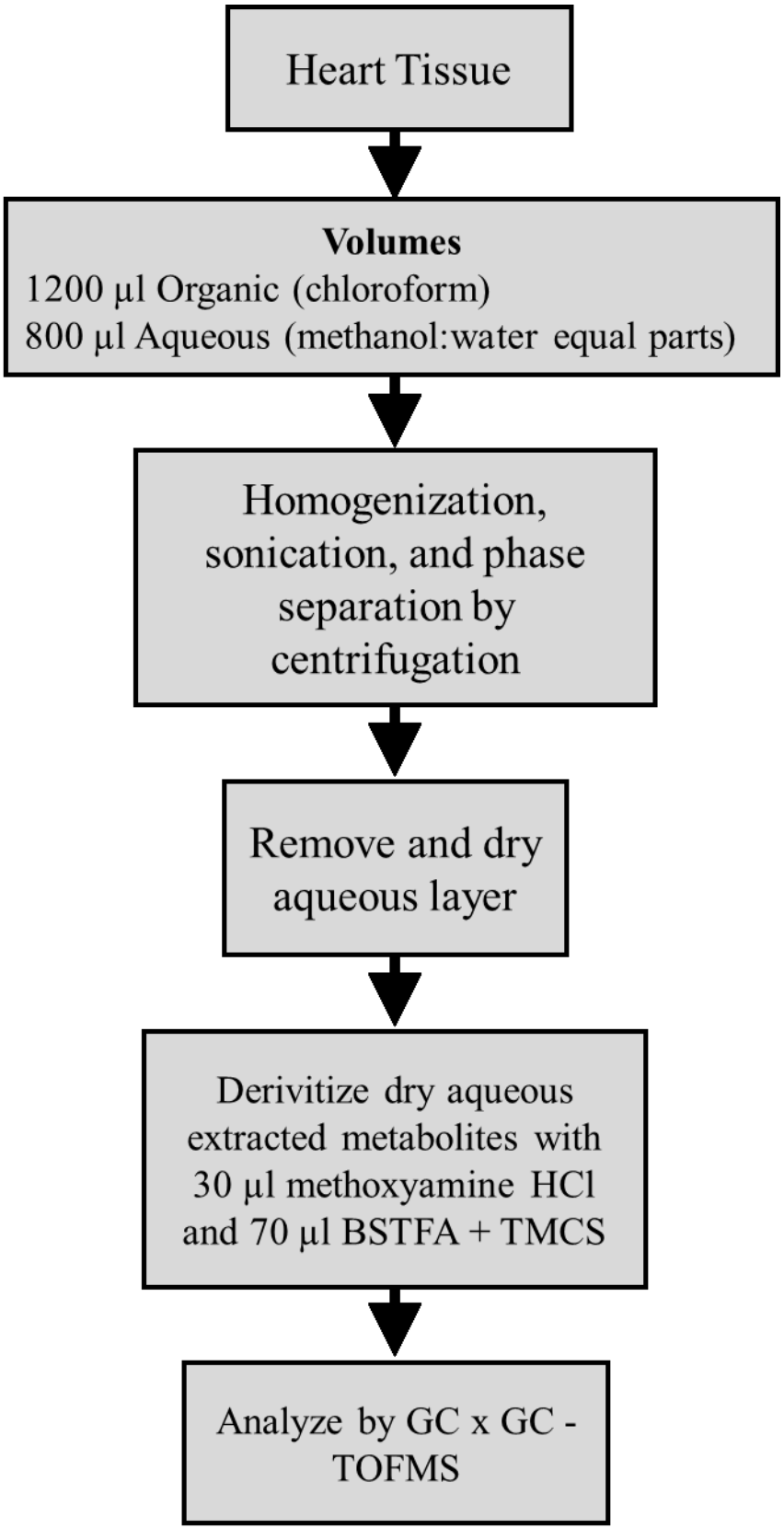

Heart Tissue
Volumes
1200 µl Organic (chloroform)
800 µl Aqueous (methanol:water equal parts)
Homogenization, sonication, and phase separation by centrifugation
Remove and dry aqueous layer
Derivitize dry aqueous extracted metabolites with 30 µl methoxyamine HCl and 70 µl BSTFA + TMCS
Analyze by GC x GC - TOFMS

Figure 2.2

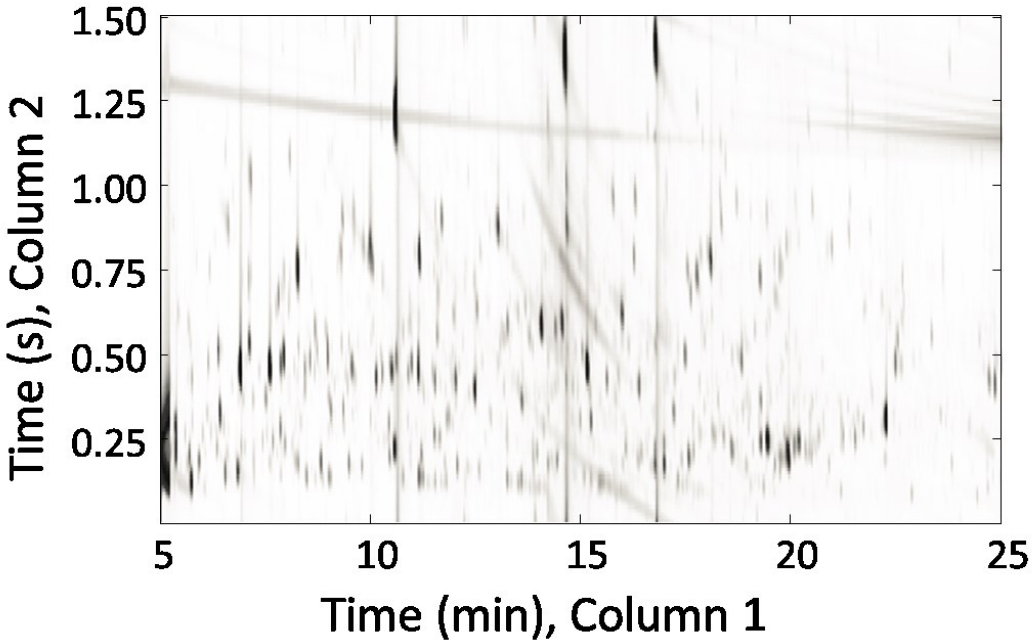

1.50
1.25
1.00
0.75
0.50
0.25
Time (s), Column 2
5
10
15
20
25
Time (min), Column 1

Figure 2.3

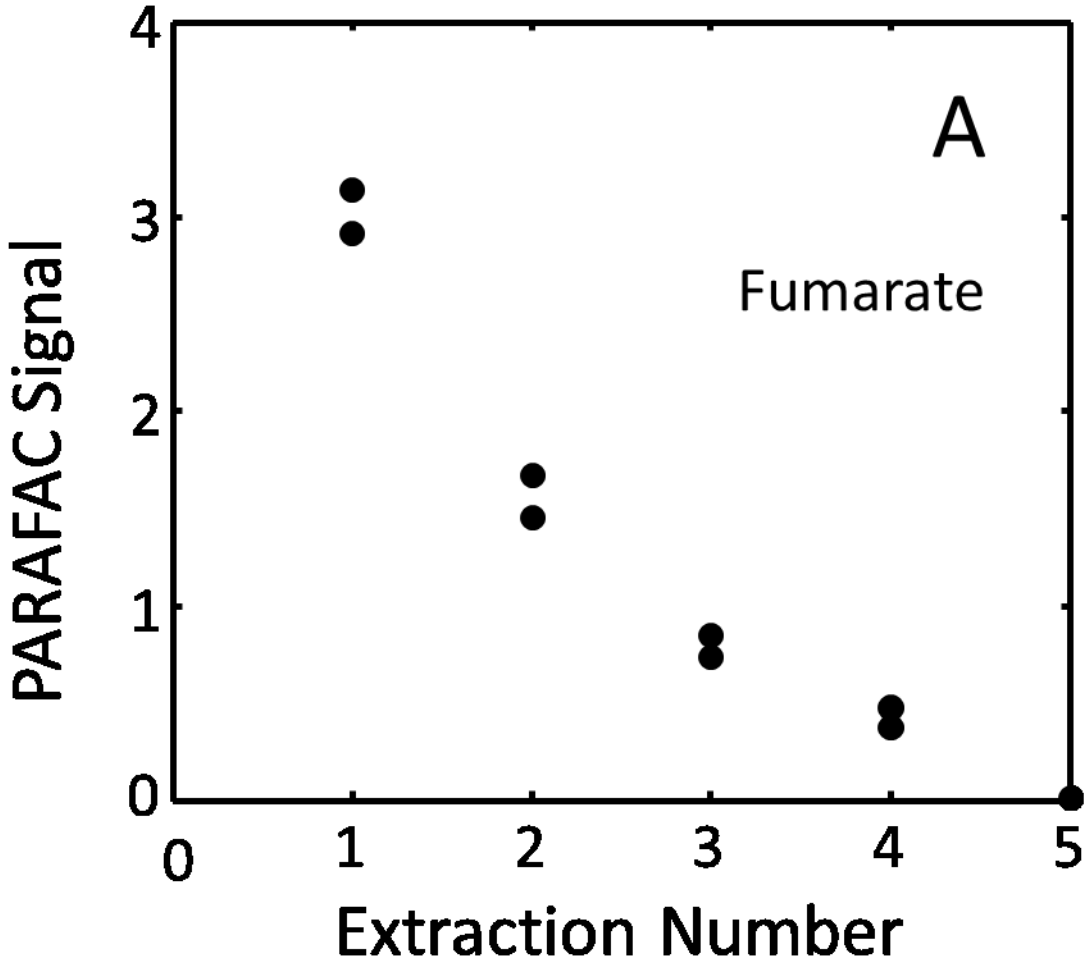


Figure 2.4 (A)

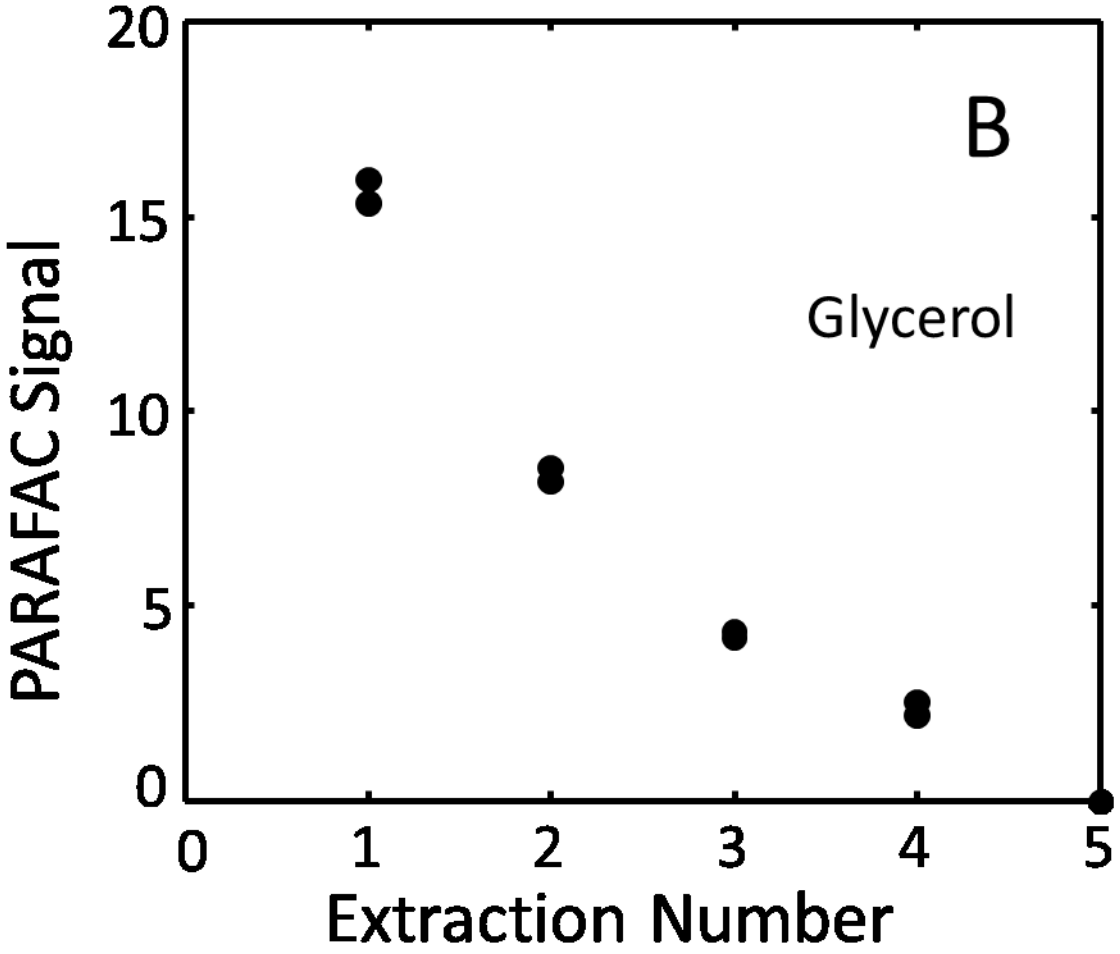

B
Glycerol
PARAFAC Signal
Extraction Number
0
5
10
15
20
0
1
2
3
4
5

Figure 2.4 (B)

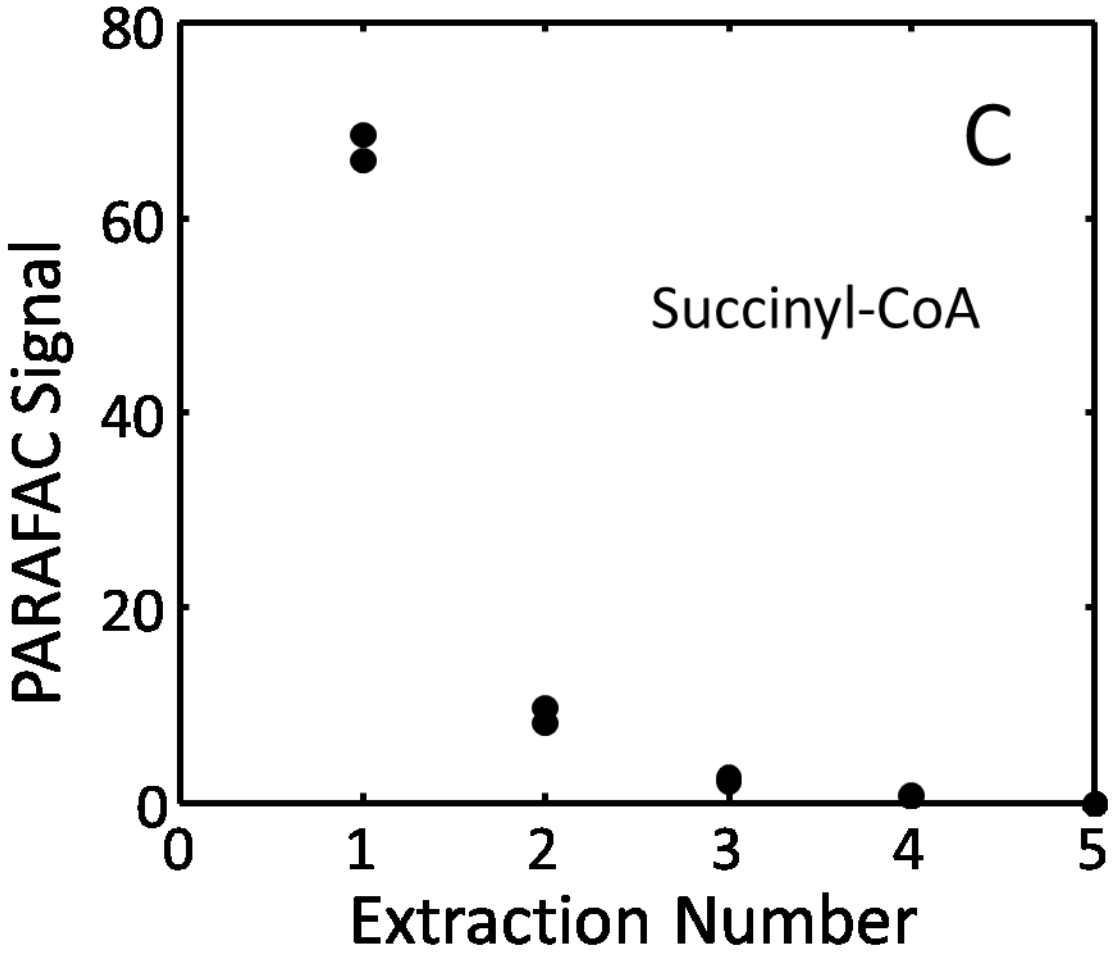

C
Succinyl-CoA
PARAFAC Signal
Extraction Number
0
20
40
60
80
0
1
2
3
4
5

Figure 2.4 (C)

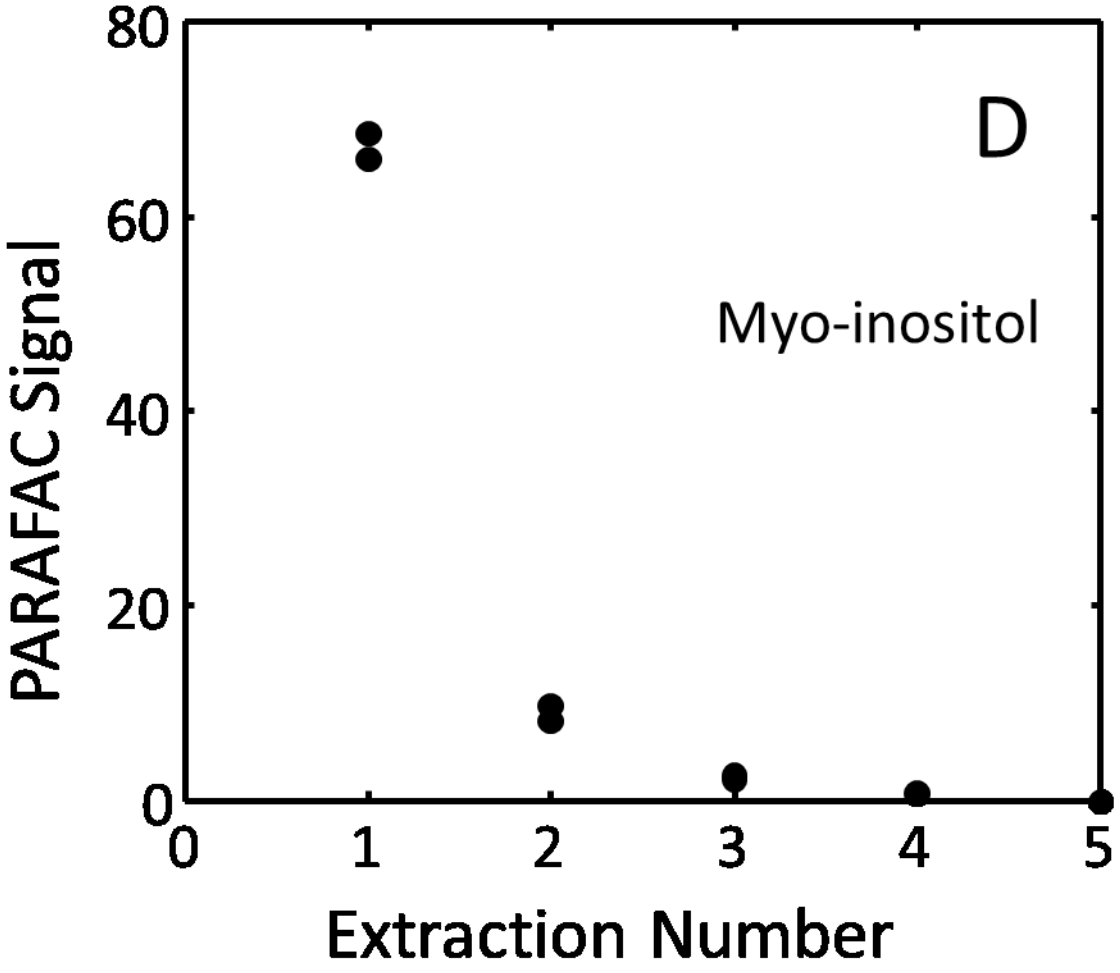

D
Myo-inositol
PARAFAC Signal
Extraction Number
0
20
40
60
80
0
1
2
3
4
5

Figure 2.4 (D)

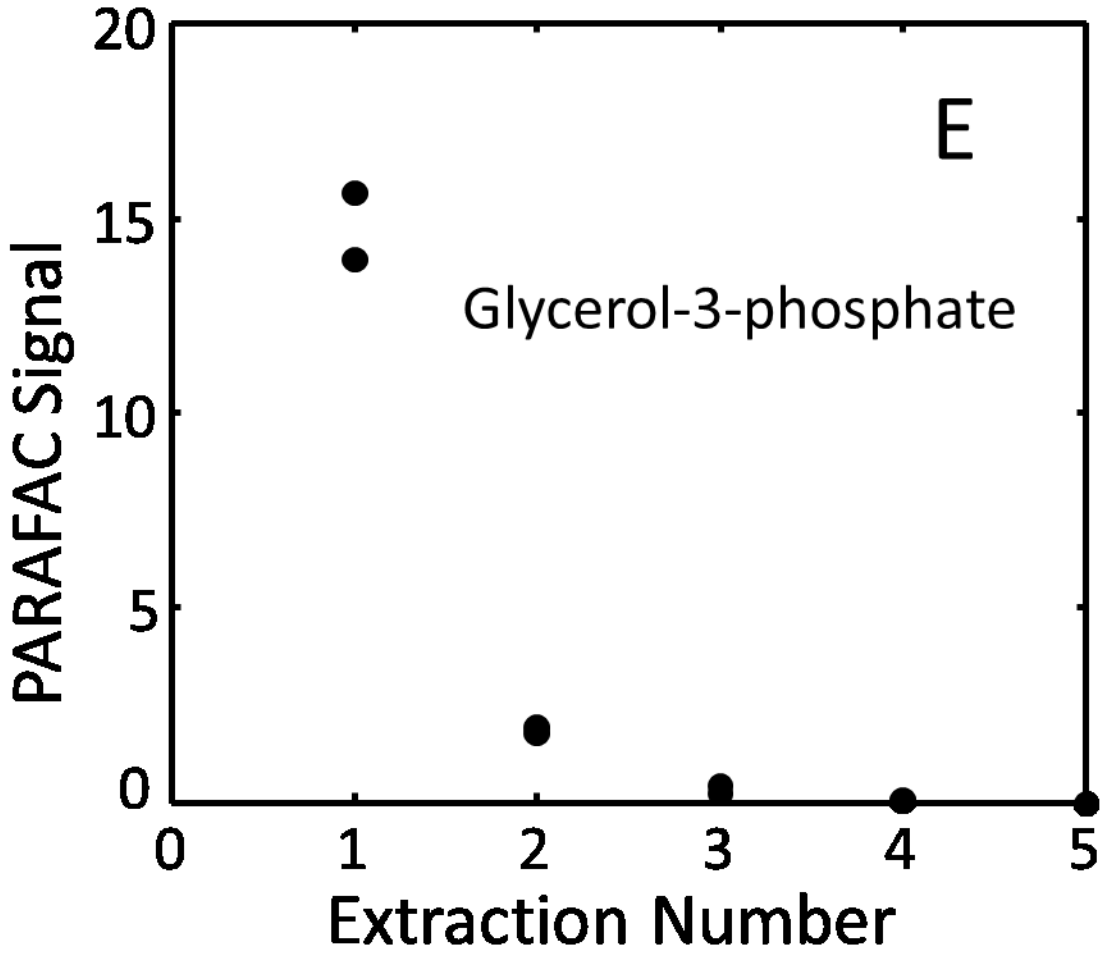

E
Glycerol-3-phosphate
PARAFAC Signal
Extraction Number
0
5
10
15
20
0
1
2
3
4
5

Figure 2.4 (E)

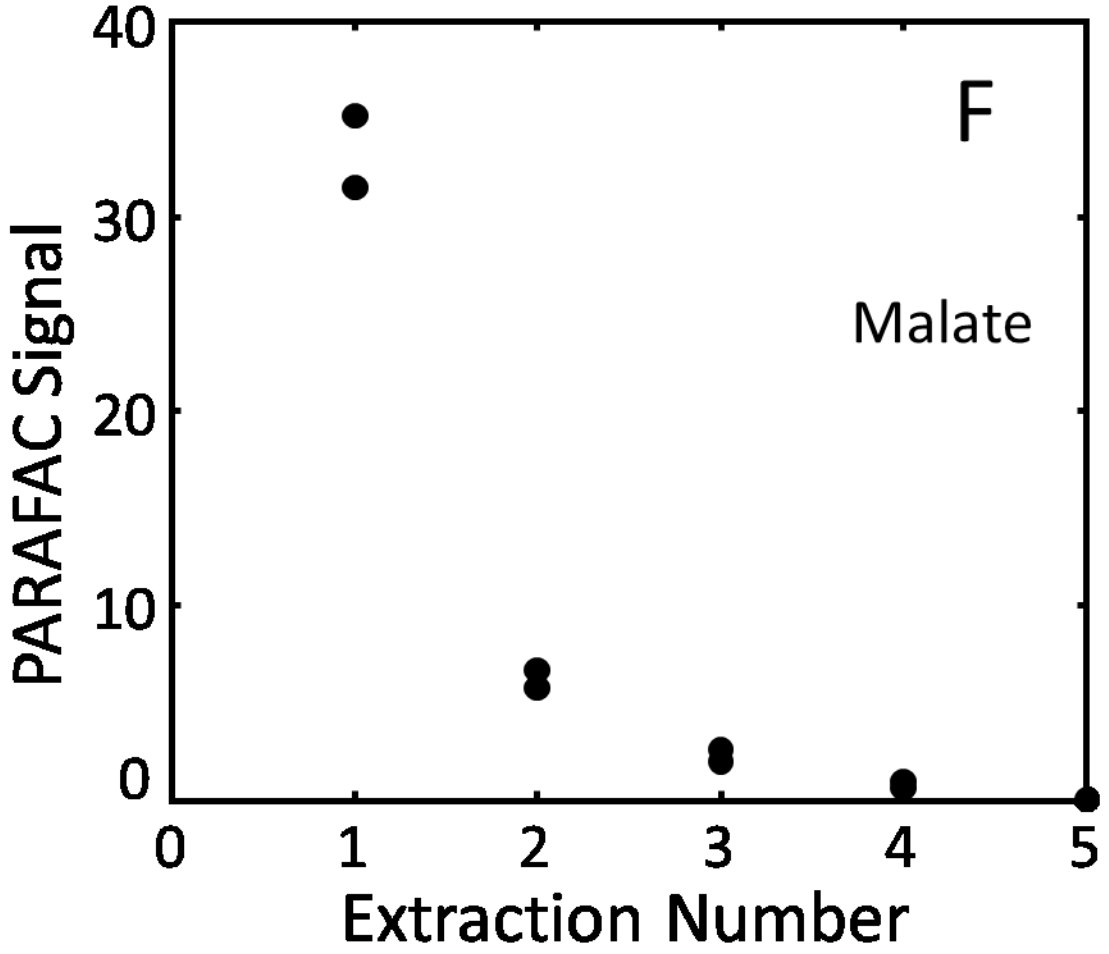

F
Malate
PARAFAC Signal
Extraction Number
0
10
20
30
40
0
1
2
3
4
5

Figure 2.4 (F)

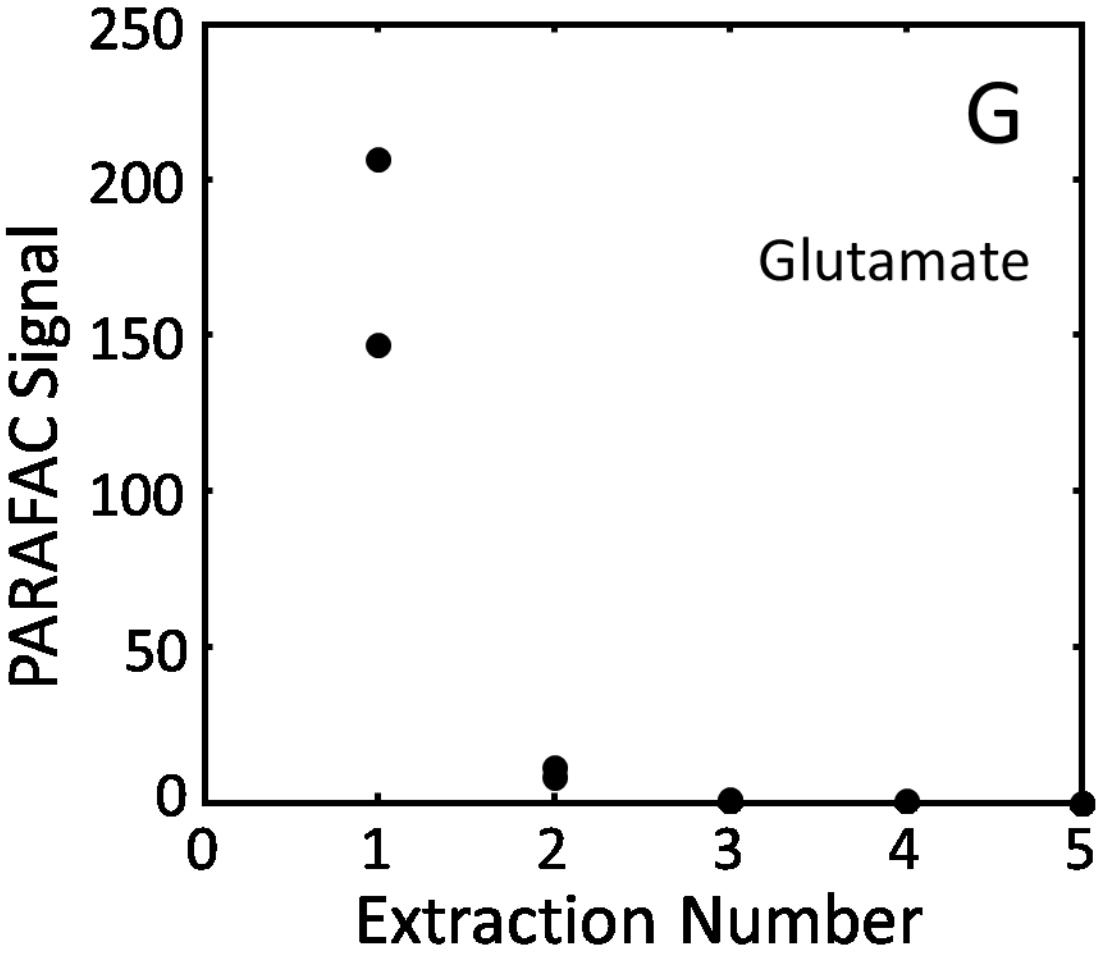

G
Glutamate
PARAFAC Signal
250
200
150
100
50
0
0
1
2
3
4
5
Extraction Number

Figure 2.4 (G)

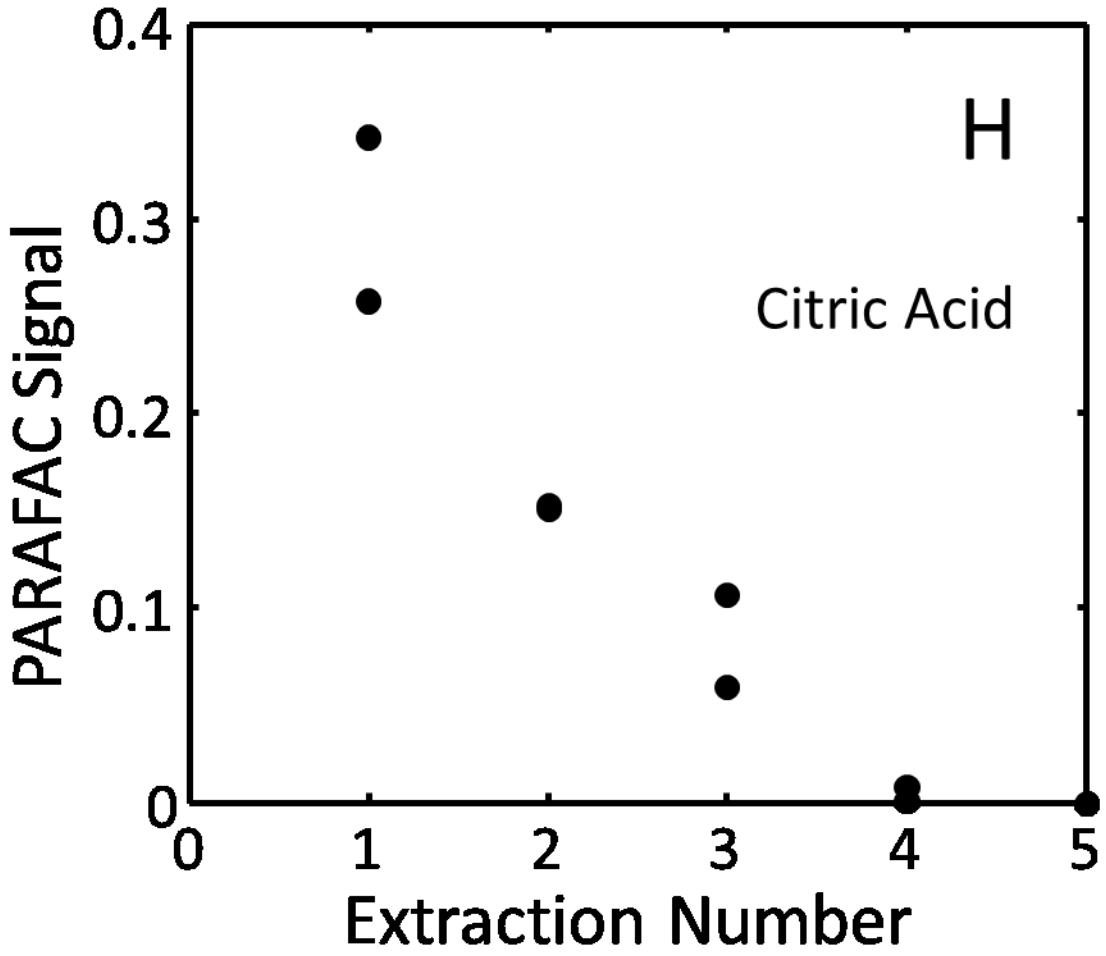

H
Citric Acid
PARAFAC Signal
Extraction Number
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5

Figure 2.4 (H)

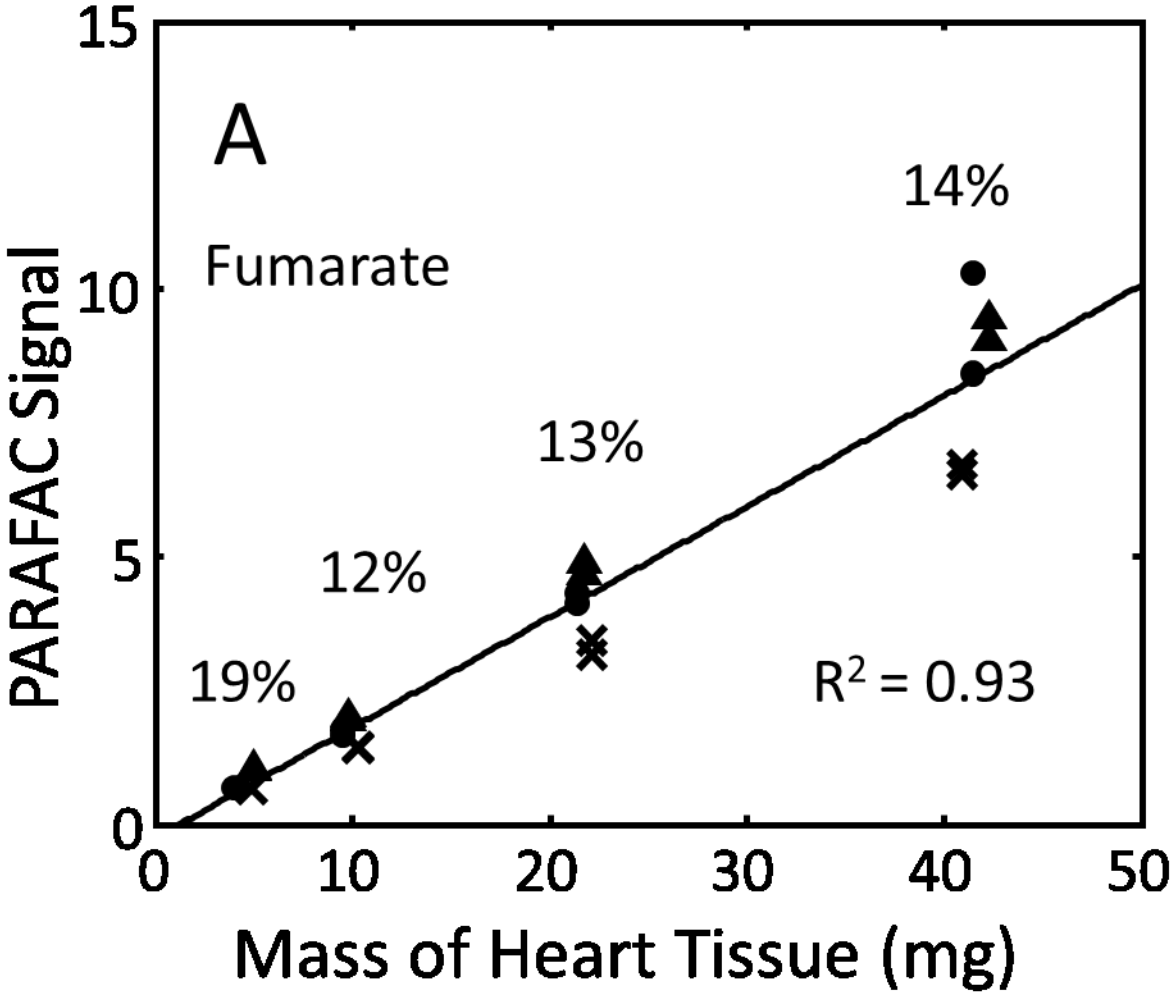

A
Fumarate
14%
13%
12%
19%
$R^2 = 0.93$
PARAFAC Signal
Mass of Heart Tissue (mg)
0
5
10
15
0
10
20
30
40
50

Figure 2.5 (A)

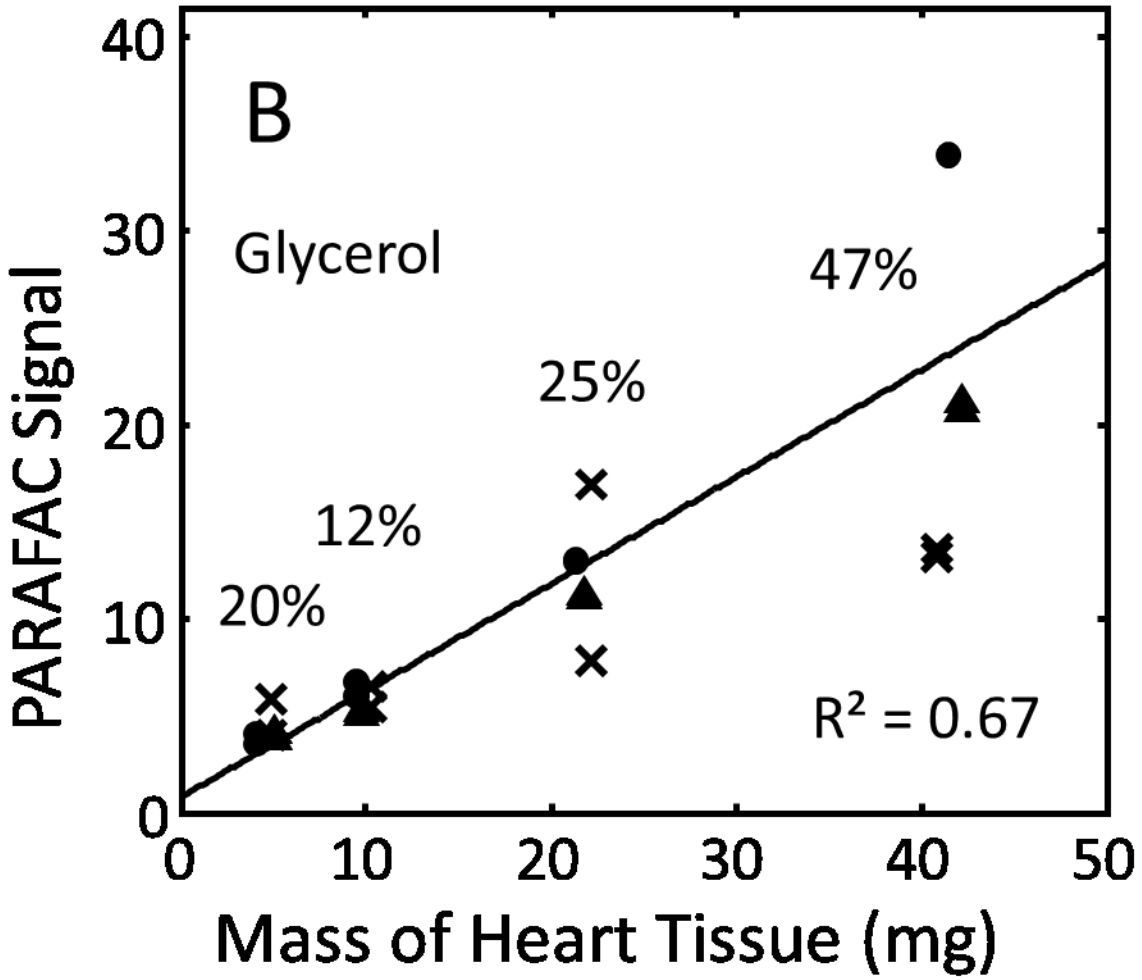

B
Glycerol
47%
25%
12%
20%
R^2 = 0.67
PARAFAC Signal
Mass of Heart Tissue (mg)
0
10
20
30
40
50

Figure 2.5 (B)

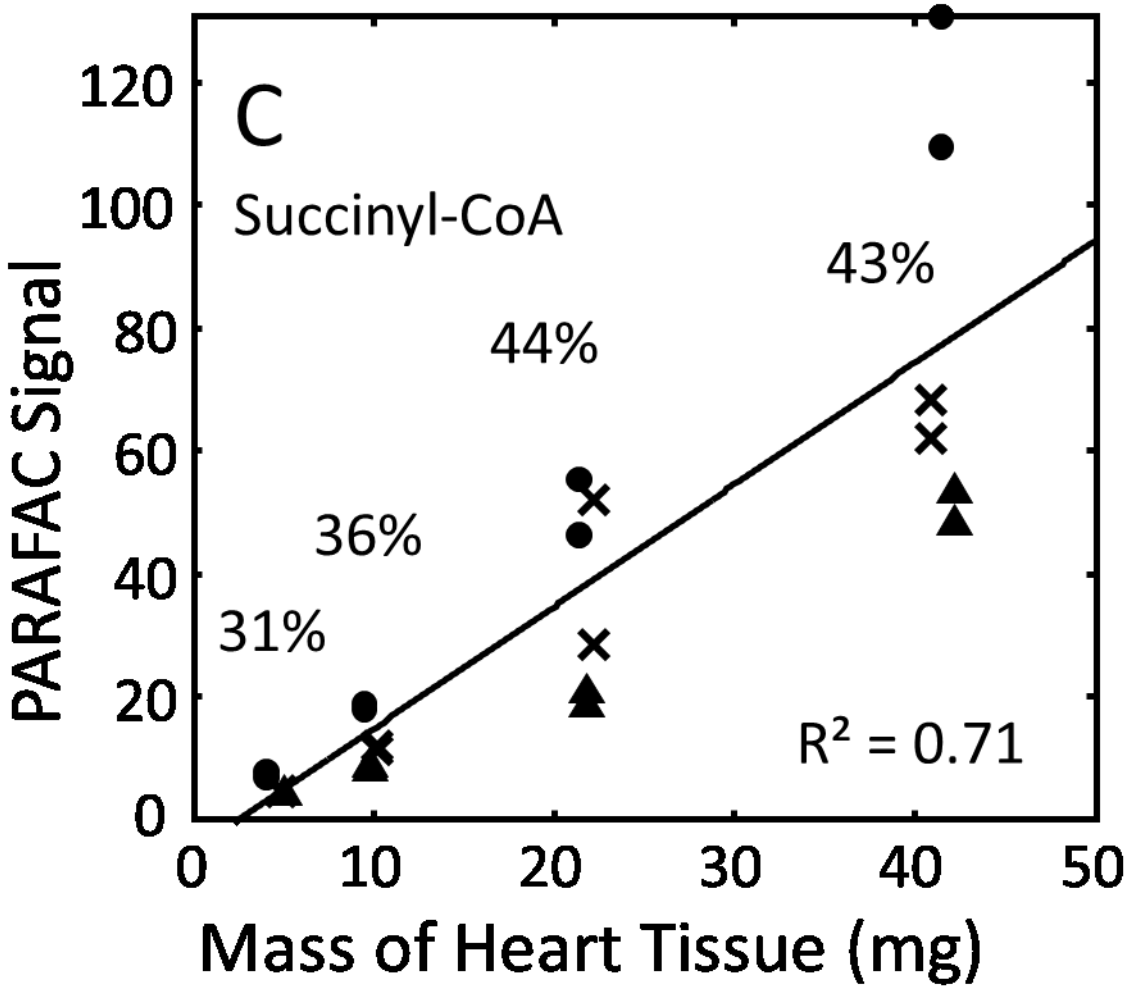

C
Succinyl-CoA
PARAFAC Signal
Mass of Heart Tissue (mg)
31%
36%
44%
43%
R² = 0.71
0
20
40
60
80
100
120
0
10
20
30
40
50

Figure 2.5 (C)

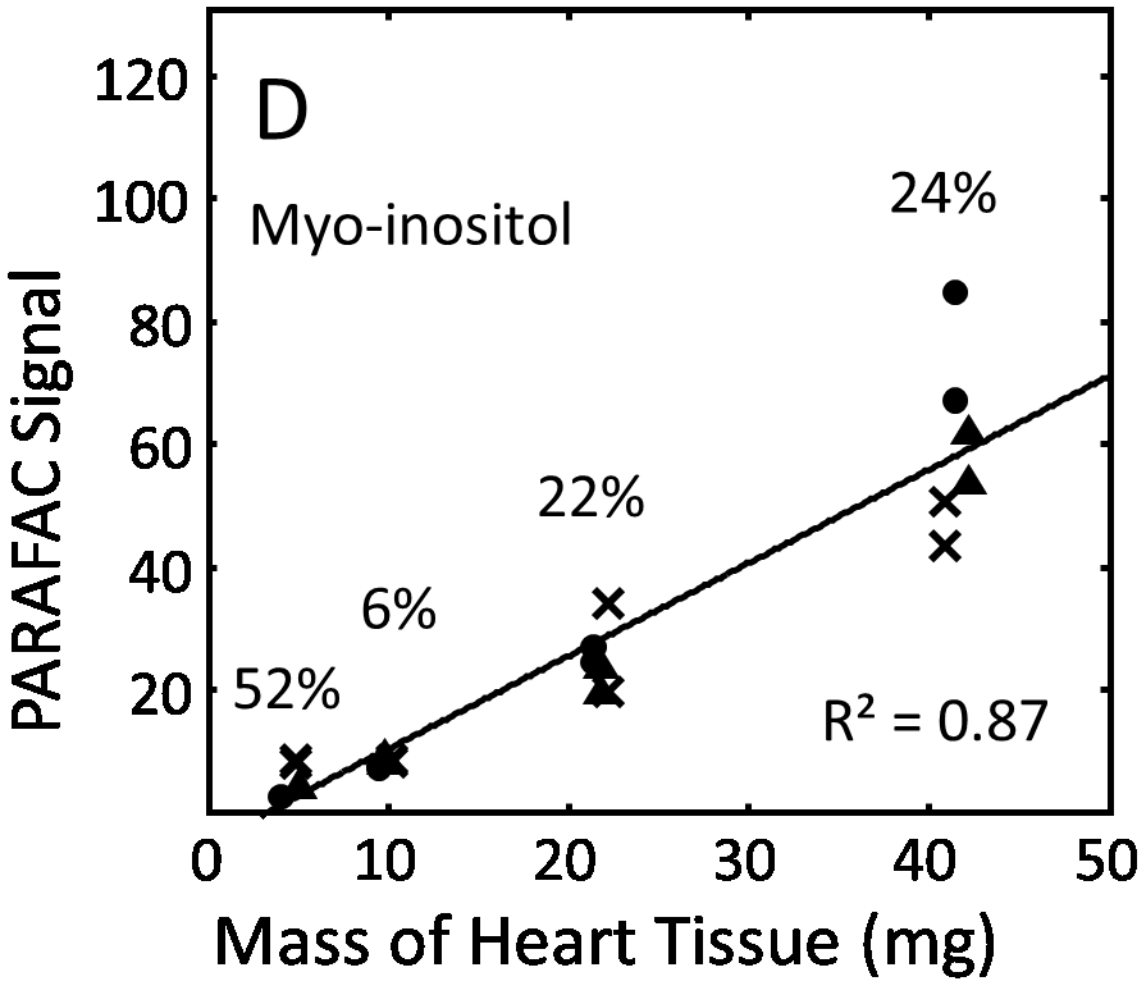

D
Myo-inositol
PARAFAC Signal
Mass of Heart Tissue (mg)
120
100
80
60
40
20
0
10
20
30
40
50
52%
6%
22%
24%
R² = 0.87

Figure 2.5 (D)

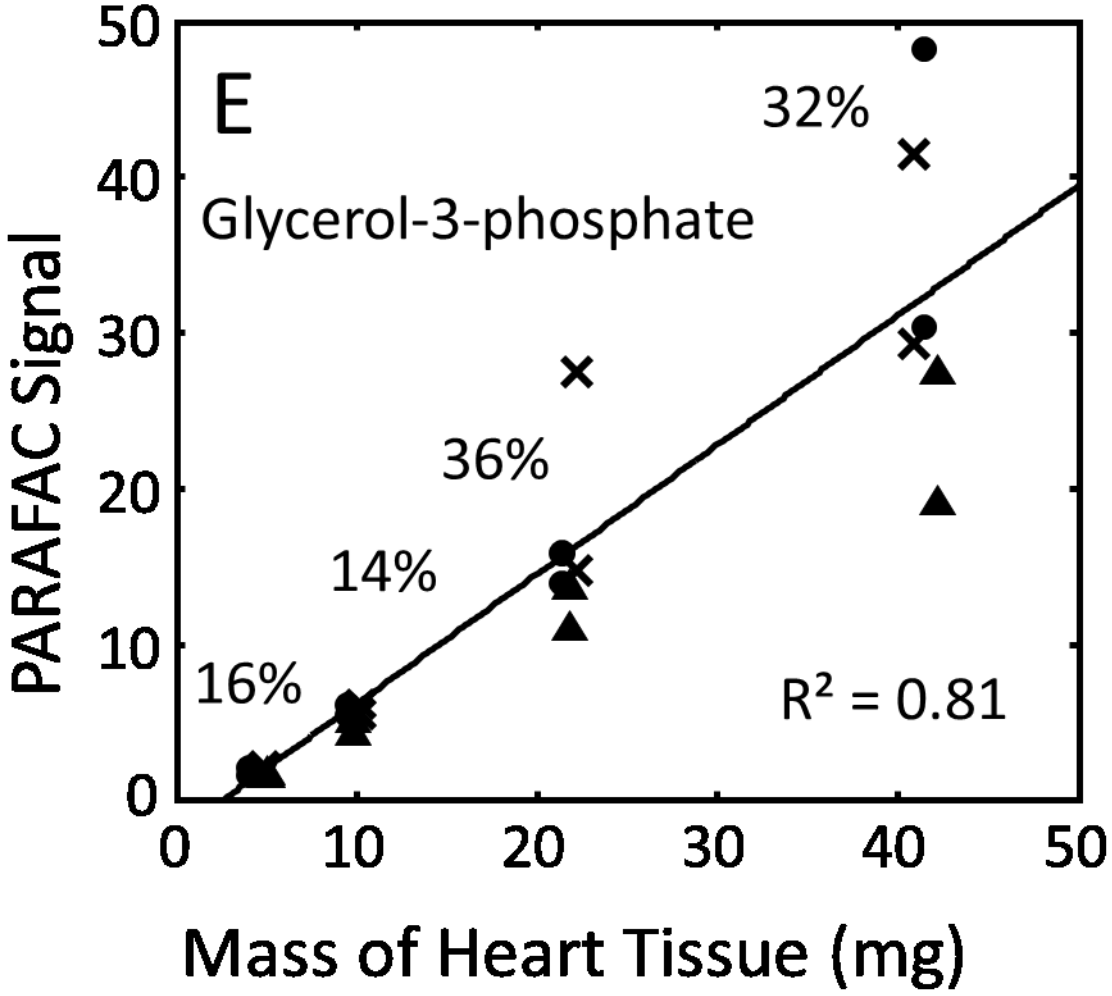

E
Glycerol-3-phosphate
32%
36%
14%
16%
R² = 0.81
PARAFAC Signal
Mass of Heart Tissue (mg)
0
10
20
30
40
50

Figure 2.5 (E)

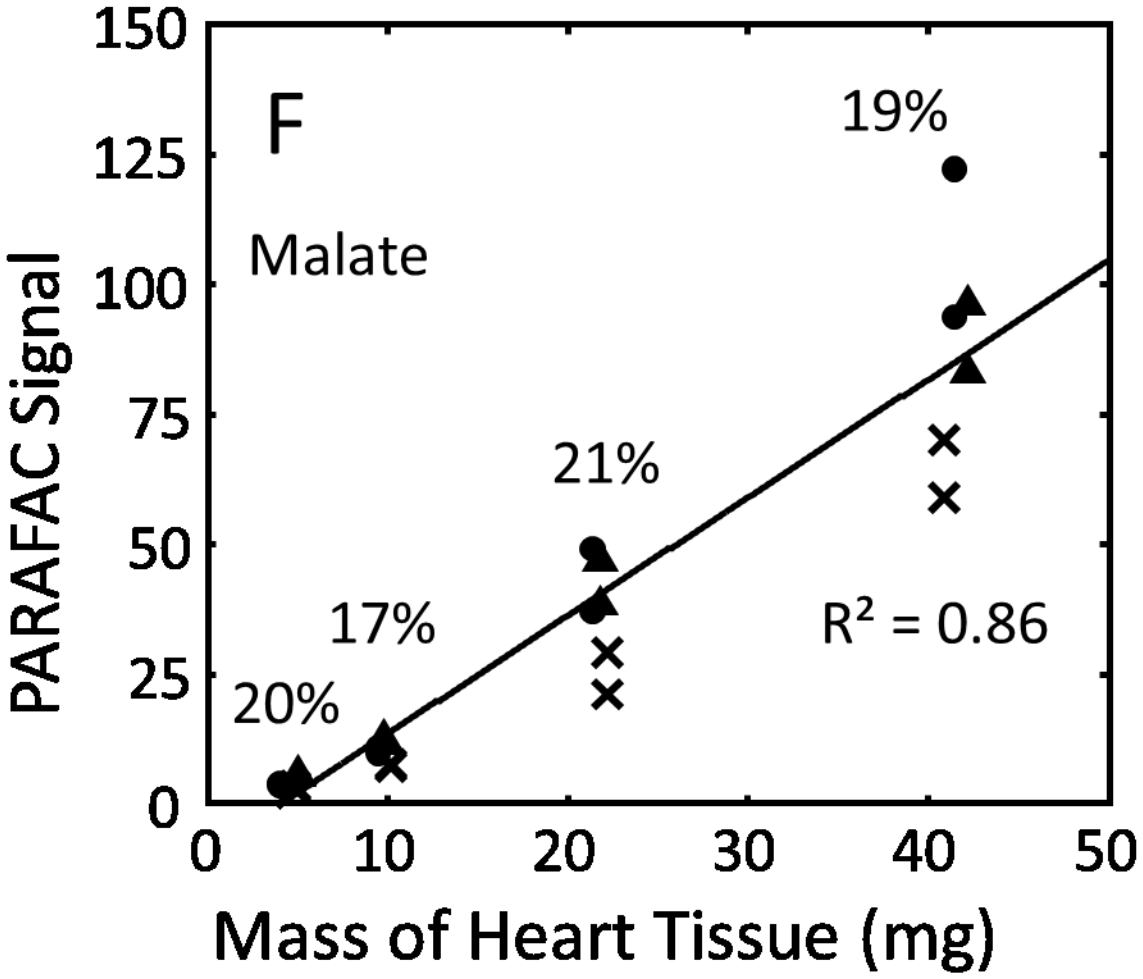

F
Malate
19%
21%
17%
20%
R² = 0.86
PARAFAC Signal
Mass of Heart Tissue (mg)
0
25
50
75
100
125
150
0
10
20
30
40
50

Figure 2.5 (F)

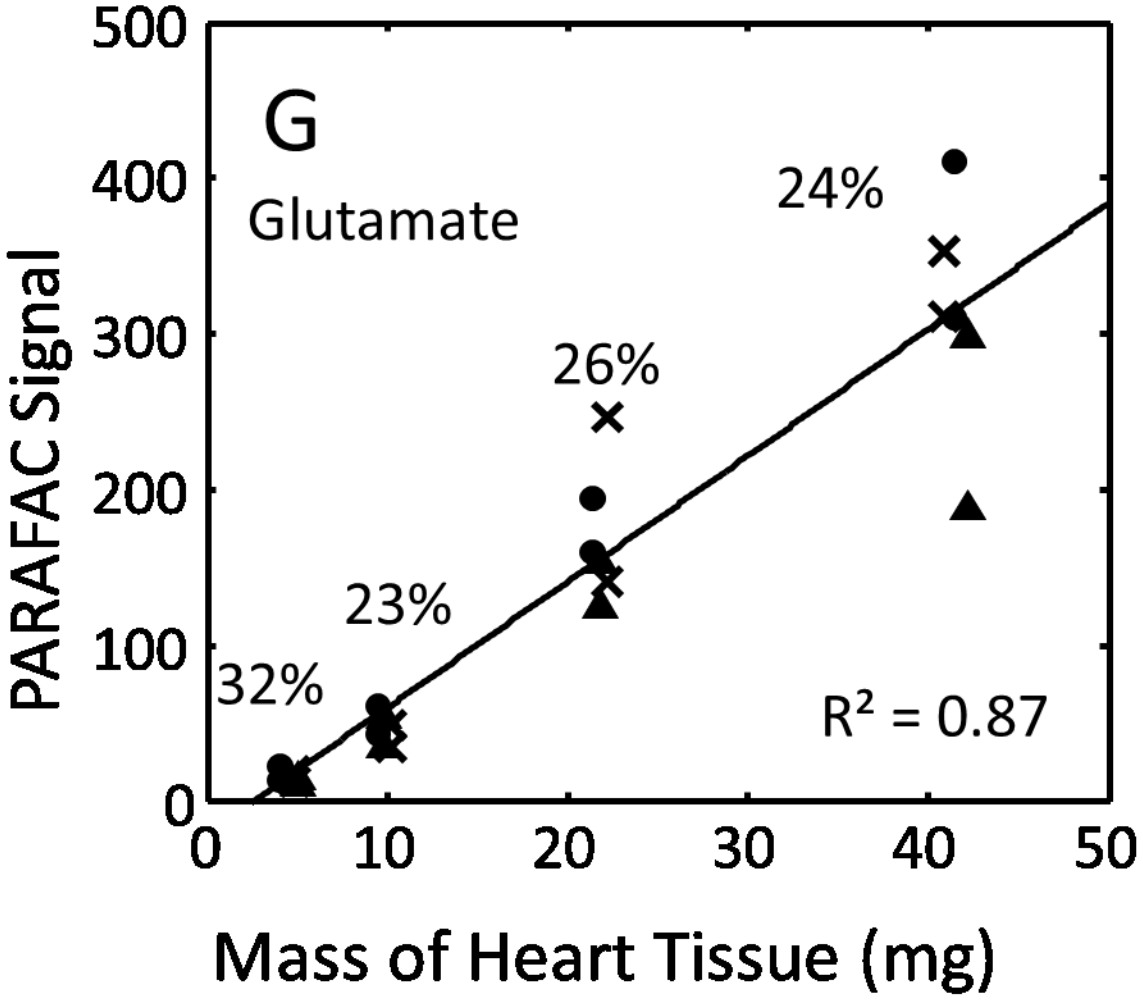

G
Glutamate
PARAFAC Signal
Mass of Heart Tissue (mg)
32%
23%
26%
24%
R² = 0.87
0
100
200
300
400
500
0
10
20
30
40
50

Figure 2.5 (G)

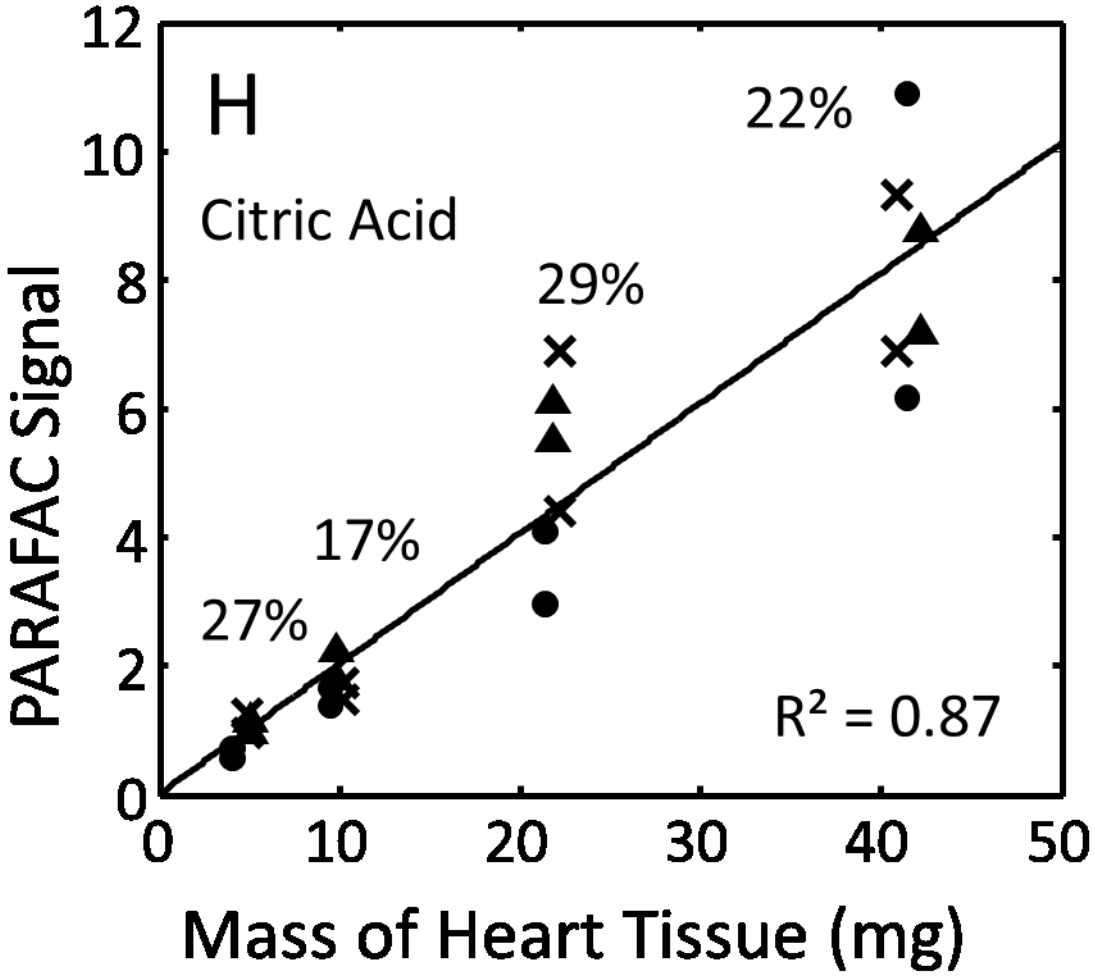

H
Citric Acid
22%
29%
17%
27%
R² = 0.87
PARAFAC Signal
Mass of Heart Tissue (mg)
0
2
4
6
8
10
12
0
10
20
30
40
50

Figure 2.5 (H)

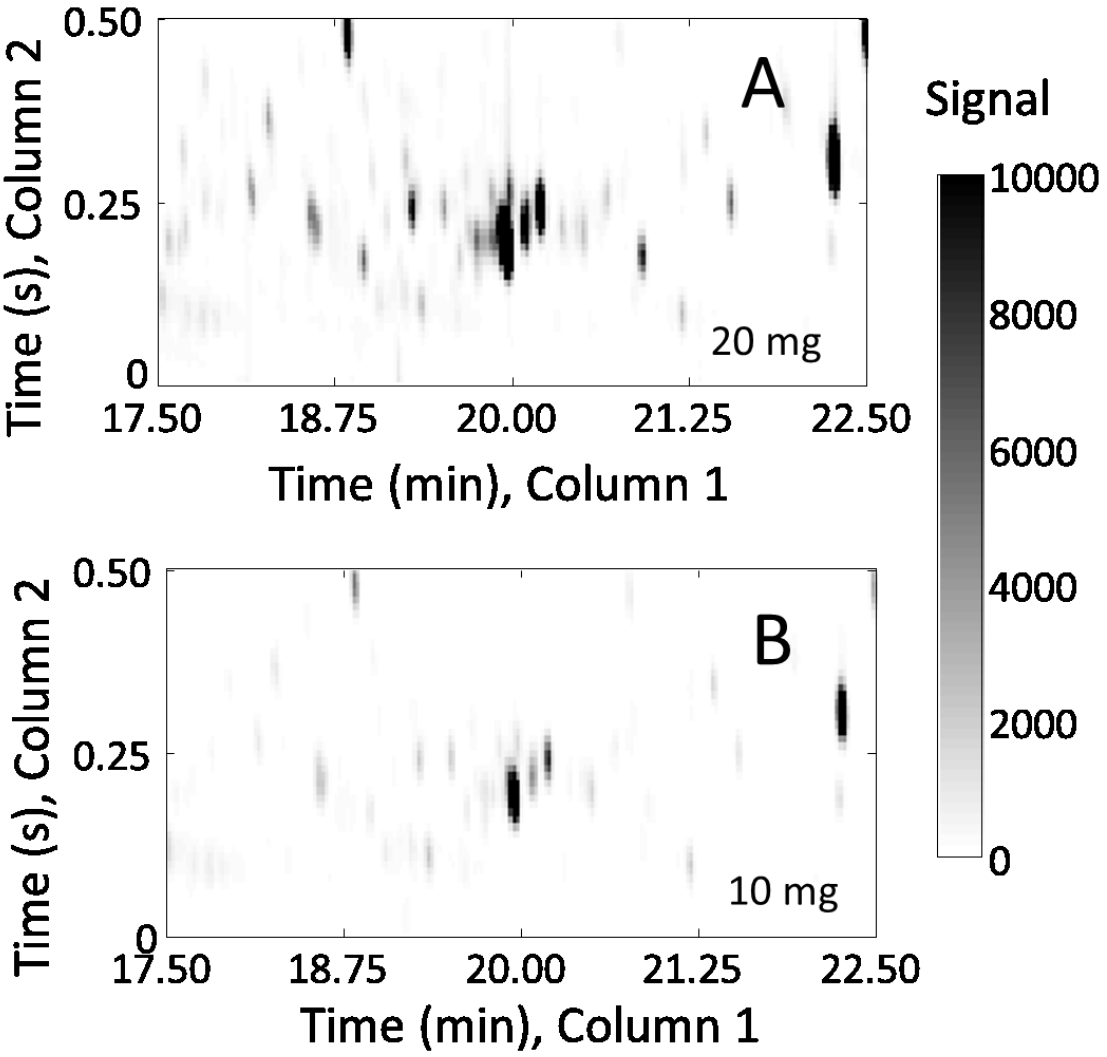

A
20 mg
Time (s), Column 2
0.50
0.25
0
17.50
18.75
20.00
21.25
22.50
Time (min), Column 1
B
10 mg
Time (s), Column 2
0.50
0.25
0
17.50
18.75
20.00
21.25
22.50
Time (min), Column 1
Signal
10000
8000
6000
4000
2000
0

Figure 2.6

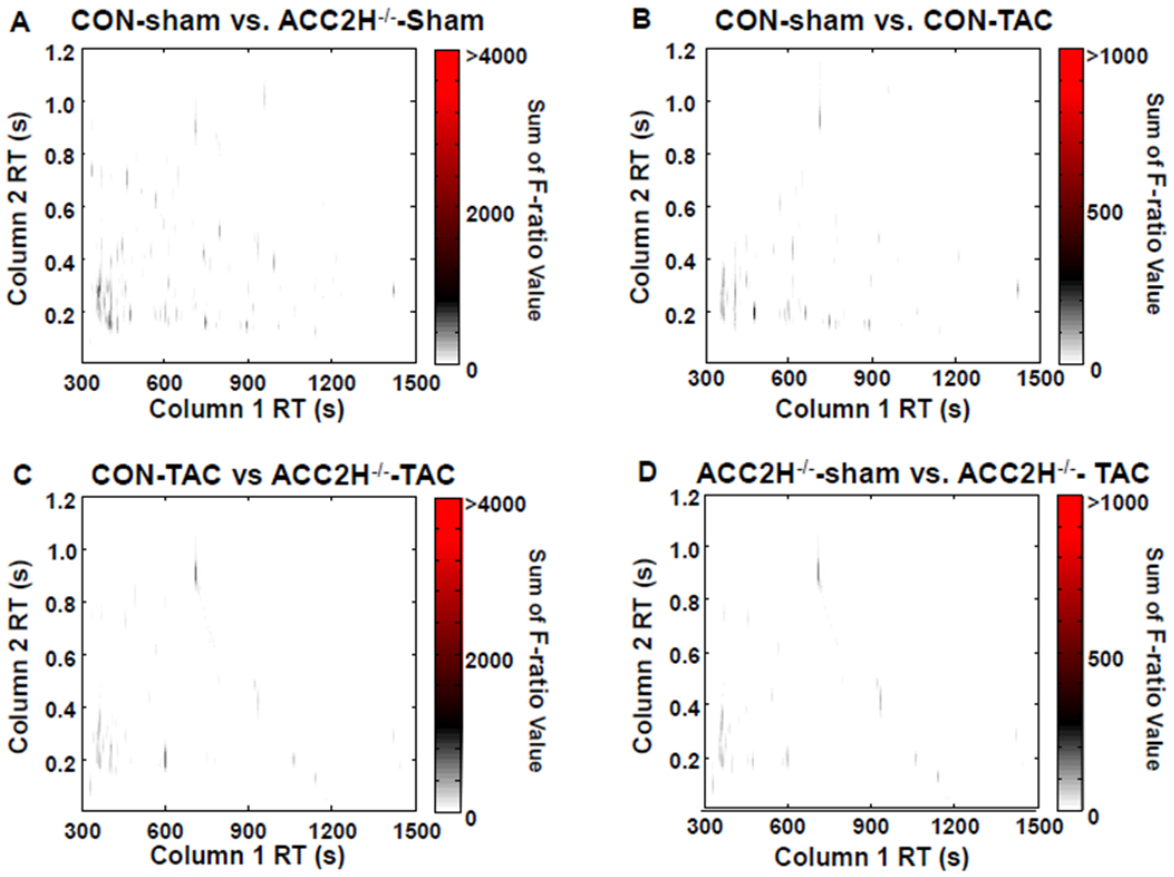

A
CON-sham vs. ACC2H-/--Sham
B
CON-sham vs. CON-TAC
C
CON-TAC vs ACC2H-/--TAC
D
ACC2H-/--sham vs. ACC2H-/-- TAC
Column 2 RT (s)
Column 1 RT (s)
Sum of F-ratio Value
>4000
2000
0
>1000
500
1.2
1.0
0.8
0.6
0.4
0.2
300
600
900
1200
1500

Figure 2.7

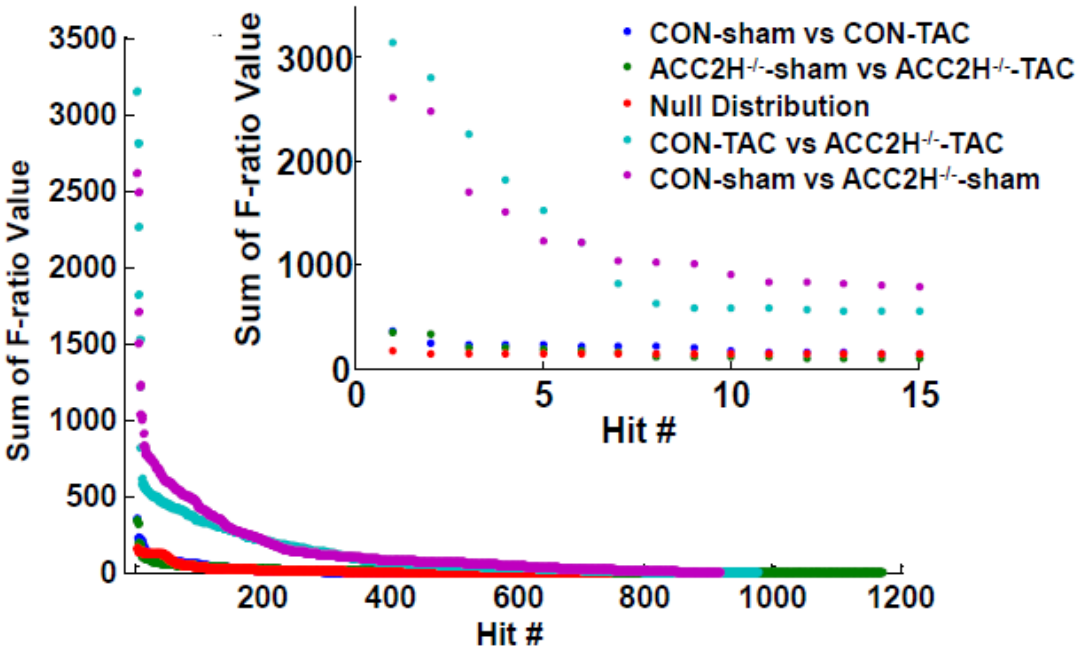

Sum of F-ratio Value
3500
3000
2500
2000
1500
1000
500
0
200
400
600
800
1000
1200
Hit #
Sum of F-ratio Value
3000
2000
1000
0
0
5
10
15
Hit #
CON-sham vs CON-TAC
ACC2H-/--sham vs ACC2H-/--TAC
Null Distribution
CON-TAC vs ACC2H-/--TAC
CON-sham vs ACC2H-/--sham

Figure 2.8

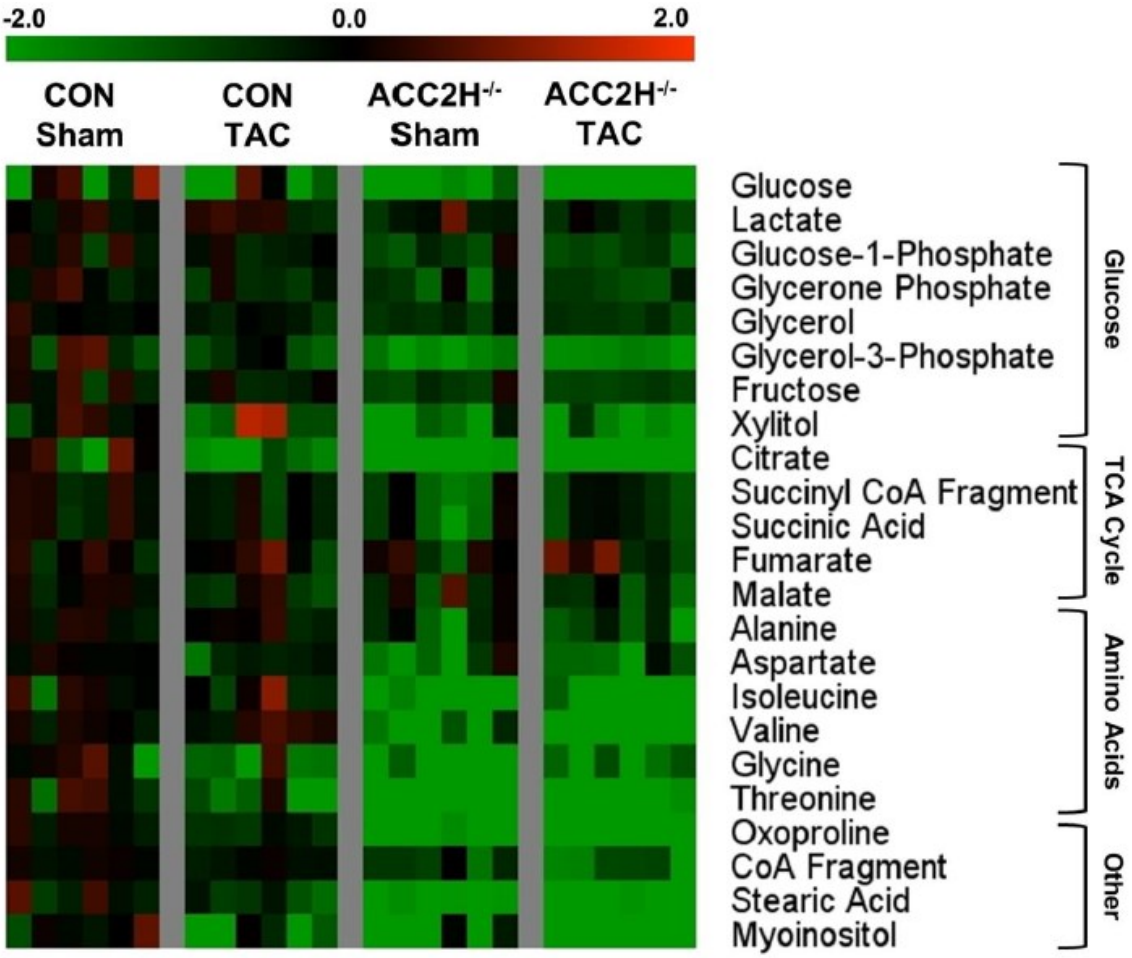

-2.0
0.0
2.0
CON Sham
CON TAC
ACC2H-/- Sham
ACC2H-/- TAC
Glucose
Lactate
Glucose-1-Phosphate
Glycerone Phosphate
Glycerol
Glycerol-3-Phosphate
Fructose
Xylitol
Citrate
Succinyl CoA Fragment
Succinic Acid
Fumarate
Malate
Alanine
Aspartate
Isoleucine
Valine
Glycine
Threonine
Oxoproline
CoA Fragment
Stearic Acid
Myoinositol
Glucose
TCA Cycle
Amino Acids
Other

Figure 2.9

Chapter 3:

Tile-Based Fisher-Ratio Software for Improved Feature Selection Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data

Two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) is a highly capable instrumental platform that produces complex and information-rich multi-dimensional chemical data. The complex data can be initially overwhelming, especially when many samples (of various sample classes) are analyzed with multiple injections for each sample. Thus, the data must be analyzed in such a way as to extract the most meaningful information. The pixel-based and peak table-based algorithmic use of Fisher ratios has been used successfully in the past to reduce the multi-dimensional data down to those chemical compounds that are changing between sample classes relative to those that are not (i.e., chemical feature selection). I detail here the development of a computationally fast novel tile-based Fisher-ratio software that addresses challenges due to 2D retention time misalignment without explicitly aligning the data, which is a problem for both pixel-based and peak table-based methods is described in this chapter. Concurrently, the tile-based Fisher-ratio software maximizes the sensitivity contrast of true positives against a background of potential false positives and noise. To study this software, eight compounds, plus one internal standard, were spiked into diesel at various concentrations. The tile-based F-ratio software was able to discover all spiked analytes, within the complex diesel sample matrix with thousands of potential false positives, in each possible concentration comparison, even at the lowest absolute spiked analyte concentration ratio of 1.06.

3.1 Introduction

Multi-dimensional chromatographic instrumentation produces information rich, and chemically complex multi-dimensional data containing meaningful chemical signals, often buried in a background of less meaningful chemical signal and noise. Experiments can be designed to analyze the similarities and differences between multiple injections of different

samples, producing a data set of even higher dimensionality. With the aid of computer software, scientists need to be able to quickly, easily and comprehensively analyze multi-dimensional data sets, so important analytes and/or chemical fingerprints can be gleaned during discovery-based experimentation. Comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC – TOFMS) is a prominent multi-dimensional separation technique that has been used extensively for discovery-based experimentation, especially when chemical species of interest are sufficiently volatile or amenable to derivatization [1-8]. To address the challenges, chemometric software for analyzing GC × GC – TOFMS data, as well as other multi-dimensional separation techniques, are available and continue to be developed [9,10].

For discovery-based experimentation with GC × GC – TOFMS, the 2D misalignment of peaks across different samples makes the non-targeted analysis difficult [11]. Some alignment algorithms have been developed for "point by point" pixel-level data [11-14], while others have been developed for peak table-based data. Current use of these algorithms has been recently reviewed [10]. Briefly, data warping and interpolation are used to stretch and compress data in order to objectively optimize the match between analyte peaks in a "target" GC × GC – TOFMS separation and the analyte peaks in a "sample" separation. The application of 2D alignment strategies for comprehensive non-targeted GC × GC – TOFMS is computationally expensive and preserving the peak signal "volumes" of every analyte during interpolation and warping is difficult and often exhibits shortcomings. In general, it is nearly impossible to preserve both signal intensity and signal area (or volume) while applying such alignment. Additionally, for non-targeted approaches, it is highly beneficial to maximize peak capacity of the 2D separations while maintaining a tri-linear data structure for subsequent deconvolution and quantification [15, 16]. With GC × GC – TOFMS, because of the cryogenically focused injection onto the

secondary column with thermal modulators, increasing the modulation period (separation run time in the second dimension) generally results in higher peak capacities as well as better chemical selectivity. Optimization for higher peak capacities preferred for non-targeted discovery-based experimentation often leads to a small number of modulations acquired per first dimension peak (~ 2 to 4 modulations commonly applied), and thus makes efforts to apply 2D alignment problematic due to a limited data density in the first separation dimension.

Non-targeted discovery-based experimentation generally aims to comprehensively analyze complex chromatograms to bring to light important analytes and/or chemical fingerprints, representing the chemistry that is significantly changing in the context of the experimental design. Non-targeted approaches can be either supervised or unsupervised, where supervision refers to either external calibration or prior classification of chromatograms as they relate to the experimental design. Popular supervised methods are Fisher ratio (referred to herein as F-ratio) and partial least squares discriminant analysis (PLS-DA). F-ratio methods have been applied to GC × GC – TOFMS data at the pixel-level (Chapter 2) [17-20] and at the peak table-level (LECO Fisher Ratio ChromaTOF 2009). PLS-DA has also been recently applied successfully to peak table-based GC × GC – TOFMS data [21-24].

Peak table-based data is acquired by peak finding, deconvolution, and alignment across peak tables. The steps for peak table preparation can be computationally demanding, especially deconvolution, and the majority of the signals processed are often not of interest to the experimental design being implemented. Application of a pixel-level based F-ratio approach, performed prior to any peak detection or deconvolution has been proposed previously as a method to reduce the initial GC × GC – TOFMS data set down to only those 2D separation locations which change significantly between sample classes per the experimental design.

Following the pixel-level F-ratio determinations, only those 2D locations undergo deconvolution and peak quantification, improving the efficiency of the discovery-based analysis [17-20].

While the previously reported pixel-level based F-ratio software has the aforementioned distinct intrinsic advantages over a peak table-based approach for discovery-based studies, the prior F-ratio software for application to GC × GC – TOFMS data has some shortcomings that need to be addressed in a sufficiently peak-based way. In order for this powerful software approach to be optimally implemented, it is essential to (1) reduce the number of false positives that are generated, while (2) optimizing the sensitivity "contrast" of finding true positives. First dimension misalignment can severely impact the sensitivity contrast for the F-ratio determination of true positives with pixel-level data. First dimension misalignment in GC × GC is a result of desynchronized modulation of a peak from one sample to another, which we refer to herein as "phasing." Phasing causes supervised non-target algorithms based in the pixel-level data analysis mode to have high false-positive rates while concurrently reducing the sensitivity contrast.

In this chapter, I demonstrate a solution that is unbiased to peak shape or specific mass spectral fragmentations, like pixel-level based algorithms, as well as sufficiently peak based to avoid the challenges of GC × GC phasing. Relative to our groups prior F-ratio software that functioned strictly on the pixel-level data [17-20], the proposed software algorithm significantly reduces false positives while concurrently optimizing the sensitivity contrast by which true positives are identified. The proposed algorithm, referred to as the tile-based F-ratio algorithm, works by creating a 2D grid (encompassing the entire GC × GC separation) composed of 2D tiles, whereby each tile is wide enough to capture the retention time variation in both the first and second separation dimensions and sums all signals within the tile as a function of each mass channel. Thus, the tile-based F-ratio algorithm is designed to eliminate the need to have a

rigorous bilinear data structure from the separation [15]. Four adjacent, but overlapping 2D grids, are used so that one grid will optimally capture any given peak in the GC × GC separation. Use of four grids, combined with summing of all chromatographic signal within a tile (at a given *m/z*), provides data reduction and added sensitivity, similar to peak-based algorithms. Sensitivity is improved because a single grid (of four) that captures a given analyte peak the best in one of its tiles will have the least amount of interference. The signal-to-noise (S/N), i.e., the sensitivity contrast, for a given analyte peak is improved roughly by the square root of the number of points summed, and by minimizing the impact of the GC × GC phasing effect. However, if the tile size is inadvisably too big (as discussed in the data analysis section), summation of noise will adversely impact the S/N benefit achieved through summation. I use the standard addition method by spiking both native and non-native chemicals into a diesel fuel to demonstrate the tile-based F-ratio algorithm in the context of a challenging complex sample matrix.

3.2 Experimental

3.2.1 Sample Preparation

Eight analyte compounds (four are non-native and four native to diesel) were spiked at the following nominal concentrations into an ultra-low sulfur diesel (ULSD) fuel: 1000, 750, 500, 250, 100 and 0 parts-per-million by mass (ppm). An internal standard, 1-bromoheptane (also non-native to the diesel fuel), was spiked at a concentration of 1 part-per-thousand (ppt) into each diesel sample. The four, non-native spiked compounds were bromobenzene, 1-chlorohexane, 5-decyne, 3-octanone and the four native spiked compounds were butylcyclohexane, cyclohexylbenzene, heptane and o-xylene. All diesel samples were prepared gravimetrically using a 5-place analytical balance, and the actual (not nominal) concentration for each analyte at each spike level is provided in Table 3.1. The actual concentration of each

spiked analyte is used as much as possible in this report. However, for clarity in certain instances, we use the nominal spike concentrations when referring to a particular concentration comparison.

3.2.2 Instrumental Parameters

The GC × GC – TOFMS instrumental platform consisted of an Agilent 6890N gas chromatograph equipped with an Agilent 7683 autoinjector (Agilent Technologies, Palo Alto, CA, USA) coupled with a Leco Pegasus III time-of-flight mass spectrometer, and equipped with the 4D thermal modulator upgrade (Leco, St. Joseph, MI, USA). The GC × GC – TOFMS instrument was used to analyze the diesel fuel samples at all spike levels. The primary column of the GC × GC (column 1) was a 20 m x 250 μm inside diameter x 0.5 μm DB-5 film (J&W Scientific/Agilent Technologies, Santa Clara, CA, USA), producing the first dimension separation. The secondary column (column 2) was a 2m x 180 μm inside diameter x 0.2 μm RTX-200MS film (Restek, Bellefonte, PA, USA), producing the second dimension separation. The GC instrument inlet was set at 275 °C and the transfer line was set at 305 °C. Column 1 was held at 50 °C for 0.25 min and then increased at 5 °C/min to 300 °C, where it was held for 5 min. Column 2 was initially set at 55 °C and followed the same temperature program as column 1 giving a total run time of 55.25 min. The modulator was kept 20 °C higher than column 1, and the modulation period was 1 s. The GC instrument was set to maintain a constant (ambient temperature and pressure corrected) flow rate of 2 ml/min at the outlet of column 2, with helium used as the carrier gas. The ion source was set to 300 °C and the detector voltage was set to 1600 V. Mass channels, *m/z* 41-340, were collected at 100 spectra/s after a 6 s solvent delay. A 1 μl injection of each diesel sample was made in split mode with a split ratio of 200:1. All diesel samples were injected in quadruplicate, producing a total of 24 runs (6 spike levels injected in

quadruplicate). An example separation run is presented in Fig. 3.1, with the GC × GC – TOFMS base ten logarithm contour plot of the TIC of a nominal 100 ppm spiked diesel sample. The locations of all nine spiked compounds are indicated in the figure (eight analytes and the internal standard). Note that the GC × GC separation conditions were selected to allow some wrap-around to more fully utilize the 2D peak capacity, without having compounds wrap-around too much. Therefore, compounds in a given second dimension separation are not allowed to wrap-around physically into compounds eluting in a subsequent second dimension separation. Figure 3.1 was not adjusted to hide the wrap around, which is commonly practiced but not necessary.

3.2.3. Data Analysis Methodology

GC × GC – TOFMS data from all of the 24 runs were imported from LECO's ChromaTOF software v 3.32 (LECO, St. Joseph, MI) to Matlab v 7.0.4 via in-house written software that utilized a peg2mat function. The imported GC × GC – TOFMS data was then analyzed with the in-house developed tile-based F-ratio algorithm. For software validation, an in-house developed target-PARAFAC GUI was used to target and extract the quantitative signal "volume" at each of the spiked analyte chromatographic locations and to obtain the normalization values for the internal standard [16]. We refer to each analyte signal as the signal volume because it represents the signal area for a given analyte along the column 2 time axis summed over all column 1 modulations in the GC × GC separation. The determination of absolute concentrations of each nominal spike level was determined via the standard addition method (SAM) and linear regression, with examples presented in Fig. 3.2 for bromobenzene in Fig. 3.2(A) and butylcyclohexane in Fig. 3.2(B). For bromobenzene the y-intercept is essentially zero, indicating it is not present in the unspiked fuel. In contrast, the y-intercept for butylcyclohexane is reasonably large, and by application of the SAM, the concentration in the

unspiked fuel was determined to be 1569 ppm. Accordingly, a complete list of concentrations is presented in Table 3.1 with the corresponding nominal spike level. The concentration ratios comparing all nominal spike concentration levels are summarized in Table 3.2 for all analytes.

At the pixel-level, a F-ratio is calculated at every point in the 2D separation, as a function of *m/z*. A F-ratio is the class-to-class variation of the detected signal divided by the sum of the within-class variations of the signal [17, 25, 26]. The class-to-class variation is calculated as

$$\sigma_{\text{cl}}^2 = \frac{\sum(\bar{x}_i - \bar{x})^2 n_i}{(k - 1)} \quad (1)$$

where n_i is the number of measurements in the ith class, $\bar{x}_i$ is the mean of the ith class, $\bar{x}$ is the overall mean, and k is the number of classes. The within-class variation is calculated as

$$\sigma_{\text{err}}^2 = \frac{\sum\left(\sum(\bar{x}_{ij} - \bar{x})^2\right) - \left(\sum(\bar{x}_i - \bar{x})^2 n_i\right)}{(N - k)} \quad (2)$$

where $\bar{x}_{ij}$ is the ith measurement of the jth class, and N is the total number of sample profiles. A F-ratio is then calculated as the ratio between the two variances,

$$\text{Fisher ratio} = \frac{\sigma_{\text{cl}}^2}{\sigma_{\text{err}}^2} \quad (3)$$

For the pixel-level F-ratio algorithm, the F-ratio in Eq. (3) is initially calculated as a function of the 2D time location of the pixel-level data, at each *m/z*. Finally, with the pixel-level F-ratio algorithm, the F-ratios at each *m/z* are summed together to give a single sum of F-ratio at each 2D pixel location.

In contrast, for the tile-based F-ratio algorithm, the GC × GC pixel-level data is summed using a 2D grid of 2D tiles, as a function of m/z, which provides data reduction in the 2D time

domain. Equations (1) through (3) are then applied as a function of 2D tile location, instead of the 2D pixel location. Then, like the pixel-level algorithm, the F-ratios at each *m/z* are summed to give a single sum of F-ratio at each 2D tile location. Four 2D grids of tiles are applied to optimally capture each peak. For added information, F-ratio "spectra" are also reported, whereby the F-ratios as a function of *m/z* are provided for tiles of interest.

All tile-based F-ratio analyses were conducted with the entire *m/z* range (41-340). Prior to tile-based F-ratio analysis, the data was baseline corrected and normalized to the internal standard signal for each sample (normalization values provided by quantification using target-PARAFAC, [16]). A S/N threshold was applied by computationally ignoring detector signal that was less than three times the standard deviation of a noise region. The noise region for this study was the first ten seconds of detector signal during a representative chromatographic run. F-ratio values for each mass channel within a given tile were summed, giving a singular sum of F-ratio value that is used to rank results in a hit list. Each row in the hit list is called an individual hit. For clarity, in some figures, a smaller *m/z* range may be plotted (less than 41-340), if there was not signal above the S/N threshold at higher *m/z*.

3.3. Results and Discussion

An advantage of pixel-based feature selection and classification algorithms is that no assumptions are made regarding peak shape or spectral features and do not rely on computationally expensive deconvolution algorithms, as do peak table-based methods. However, there are two major shortcomings to strictly pixel-based F-ratio algorithms. First, there often are many false positives, and second, the true positives may have low F-ratio values that do not provide sufficient contrast from false positives and true negatives. To address these two shortcomings, I investigated the causes of observing low F-ratio values for true positives (in this

study using spiked analytes), and I investigated the causes of observing high values for locations that do not correspond with the spiked analytes (false positives). I developed new software to implement the solutions to address these shortcomings.

Sensitivity contrast is the most important feature that the F-ratio software should provide the analyst. Increasing contrast is achieved by optimizing the chemical selectivity in discovering statistically significant changes between the two (or more) sample classes. Ideally, true positive hits should have vastly higher F-ratio values than false-positives to ensure that small changes in the concentration ratio for a given true positive hit can be detected. To illustrate and address these issues, I begin with the comparison of quadruplicate injections of a 113 ppm bromobenzene spike in diesel versus quadruplicate injections of unspiked diesel, Fig. 3.3(A), with the bromobenzene peak from a zoomed-in section of the separation also indicated (an eight modulation time window along the column 1 time axis, 8s, since the modulation period is 1 s). Gaussian peaks have been superimposed showing the nominal peak width for the first separation dimension peak as it is modulated onto the shorter second dimension separation (Figure 3.3(B) and 3(C)). The misalignment in the first dimension that is observed in the two superimposed Gaussian peaks in Figure 3.3B generally causes significant peak height variation (referred to as "phasing") for each modulated second dimension peak across the four separate sample replicate injections of the nominal 100 ppm spiked diesel (actual concentration of bromobenzene is 113 ppm). Phasing can result in widely differing peak heights (for a given modulated peak on the second separation dimension) from one injection replicate to the next, even when the total peak areas from the sum of all modulations of an analyte should be, in principle, identical. In this example in Fig. 3.3, the highest F-ratio obtained by a pixel-based F-ratio algorithm for bromobenzene was found to be 216 (for the selective mass channel *m/z* 77). As we develop the

tile-based F-ratio algorithm, we shall compare the F-ratio obtained to this pixel-based F-ratio of 216, which serves as a benchmark metric. Figure 3.3(D) shows a library mass spectrum (National Institute for Standards in Technology 2011) from the bromobenzene peak shown in Fig. 3.3 (A), (B), and (C). Mass channel *m/z* 77 was used for each chromatographic data plot in Fig. 3.3. The mass spectrum for each spiked analyte, in conjunction with the 2D retention time, is used to identify the analyte associated with each F-ratio spectra in subsequent figures.

Another significant shortcoming of applying pixel-based F-ratio analysis is the common occurrence of false positives, with a numerically very substantial false positive hit shown in Figure 3.4. This false positive hit was on the shoulder of a large native compound peak (eicosane), with the large peak not changing in concentration from one class to the other class. The actual pixel that resulted in a high F-ratio is shown in the zoomed-in selection at 1404 s along the first dimension and 71 ms along the second dimension in Fig. 3.4(B), in which we overlay four runs from each samples class (100 ppm spiked versus unspiked). The raw data in Fig. 3.4(A) along the first dimension time axis are plotted in Fig. 3.4(B) and plotted along the second dimension time axis for clarity, with it understood that we are zooming in on the modulation initiated at 1404 s. Randomly occurring covariance of detected signal (above the S/N threshold) resulted in a pixel-based F-ratio value of 249 at mass channel 95 (other mass channels not shown for brevity also indicated the same result). This result is of great significance because this pixel-based F-ratio value for a single mass channel of a known false positive is higher (i.e., more prominent) than the pixel-based F-ratio obtained for the true positive bromobenzene at a single selective mass channel (Table 3.3).

To address these shortcomings (low sensitivity contrast and high false positive rate) in the pixel-based F-ratio software performance, I developed the tile-based F-ratio algorithm

utilizing the positive attributes of the pixel-based and peak-table based approaches. The algorithmic method functions by creating a 2D tile (i.e., a 2D time window) that is wide enough to capture the retention time variation in both the first and second separation dimensions, and sums all signals at a given *m/z* within the tile. The size of the tile is chosen to contain ~ 99% of a Gaussian peak in both the first and second dimensions, plus two extra modulations in the first dimension, and additional time greater than the peak width in the second dimension, in order to fully contain retention time variation in each dimension (i.e. eliminating the negative impact of phasing from one sample injection to the next). For this data set, the window size was set at eight modulations in the first dimension, the time window of the data in Fig. 3.3(A), and 20 ms in the second dimension, which is the inset caliper in Fig. 3.3(C). This tile-based signal-summation approach removes false positives caused by random covariance of detector noise within and across sample classes (as illustrated in Fig. 3.4, because summing all signals within a given tile very effectively averages out false positive pixels.

One set of tiles, referred to as a tile grid, covering the entire 2D chromatogram is not sufficient, because with only one tile grid some peaks will be split into two or more tiles. Thus, four overlapping rectangular tile grids are required. Tile grids are shifted by half a tile length in the first separation dimension, or shifted half a tile length in the second separation dimension, or shifted half a tile length in both dimensions (illustrated in Fig. 3.5 (A)). Naming of each of the tile grids is done in relation to the Cartesian quadrant system. Tile grid 3 (t3) is anchored with the lower left corner at the origin of the 2D chromatogram (lower-left quadrant) while tile grid 1 (t1) is shifted right and up (upper-right quadrant), with the other tile grids named with respect to the direction of the grid shifting. With this tile-based F-ratio algorithm, baseline corrected signal

at all of the 2D points (or pixels) in each tile is summed to provide a single signal value per tile, and per each *m/z*.

The four tile grids are used to ensure that one of the four grids will optimally capture the entire signal for a given analyte peak hit, and thus, that one tile out of the four tile grids will have the highest F-ratio value for a given true positive analyte hit. The other three tile grids inevitably split the analyte peak. Figure 3.5 illustrates the tile grid approach with an illustration of the tile grids (Fig. 3.5(A)) and contour plots of zoomed portions of the chromatograms from the 113 ppm bromobenzene quadruplicate injections (Fig. 3.5(B)). The bromobenzene peak is misaligned slightly in both dimensions (also shown in Fig 3.3), but alignment is not needed because summing all data points within the tile obviates the need to align.

As shown in Fig. 3.5(B), tile grid t4 optimally captures the bromobenzene peak, and minimizes retention time variation of the four injections, resulting in the highest F-ratio of 9103 at *m/z* 77 (40-fold greater than by the pixel-based F-ratio software, as per Fig. 3.3). This example illustrates how the tile-based F-ratio algorithm substantially increases the sensitivity contrast of the hit discovery by decreasing the within-class variance (denominator of the F-ratio per Eq. (3)), while at the same time improving the between class variance by boosting S/N with summation (numerator of F-ratio per Eq. (3)). The F-ratio spectra (as a function of *m/z*) for adjacent tiles for bromobenzene are shown in Figure 3.6. The mass spectrum shown in Fig. 3.3 (D) can be compared with each tile's F-ratio spectrum to assist in the identification of bromobenzene as the analyte changing between the two classes. The sum of F-ratio values for the three most adjacent tiles are included; they are much less than the bromobenzene tile in tile grid t4.

Application of the four tile grids to a larger section of 2D data containing multiple peaks is shown in Figure 3.7. Each individual peak is optimally sampled by only one of the four tile

grids and that tile grid will have the highest F-ratio value for all selective *m/z*, because that is where the within-class variance will be the smallest. However, the software at the stage reported herein, produces some "redundant" hits initially for each analyte peak, because of the different grid schemes, but one grid scheme will be optimal for each analyte. In a future software development stage, it is envisaged that the software will readily allow for removing redundant hits, but this is not discussed here for brevity.

To summarize (see Table 3.3) with these two representative hit examples (the true positive for bromobenzene in Fig. 3.3, and the false positive in Fig. 3.4), utilizing a novel tile-based algorithm with four tile grids has significantly increased sensitivity contrast for the sum of F-ratio value for the true positive by two orders of magnitude (an increase in chemical selectivity across sample classes), relative to the previously reported pixel-based F-ratio software [17-20]. Concurrently, the sum of F-ratio value for the highest false positive by pixel-based analysis has decreased by two orders of magnitude. Thus, the tile-based F-ratio software provides a substantial improvement in reducing false positives, concurrent with enhancing the sensitivity contrast to discover true positives.

Application of the tile-based F-ratio software to the entire 2D chromatogram for the nominal 100 ppm spiked diesel versus 0 ppm (unspiked) diesel, summarized in Table 3.4, indicates that the eight spiked analytes are the top eight hits. Note that at this point in the software development, the redundant hits caused by peak splitting in adjacent tile grids have been removed by analyzing the spectrum for each tile. The first false positive seen after the eight spiked compounds has a sum of F-ratio value of 643 (provided in Table 3.4), with only half the sum of F-ratio value obtained for cyclohexylbenzene, which had an absolute concentration ratio of only 1.16 (since it was present in the unspiked fuel at a concentration of 789 ppm). Of further

note, the tile-based F-ratio software detected o-xylene with a sum of F-ratio value of 3464 at an absolute concentration ratio of 1.06, suggesting that very small concentration ratios can likely be found for some compounds of interest in future studies. Thus, the up or down change of analytes of interest that can be confidently discerned, from one sample class to the next, appears to be approaching very small concentration ratios. This is very important for studying complex samples in which only small concentration changes may be occurring. Overall, the tile-based F-ratio software now provides outstanding sensitivity contrast to discover true positives relative to false positives.

3.4 Conclusions

A major challenge for non-targeted chemometric algorithm methods for 2D separations is inadvertent misalignment of chromatographic peaks from one sample run to the next. I have developed novel software that addresses this challenge by taking a quick, sufficiently peak-based approach (tiling), while maintaining the advantages of pixel-based methods. The tile-based F-ratio software provides effective data reduction, detection of chemical patterns and is robust to misalignment.

The data reduction strategy of summing signal within a defined tile region is beneficial for GC × GC – TOFMS and is computationally fast. Utilizing four grids ensures that the highest F-ratio value calculated for a given up or down changing concentration ratio, will represent the most selective chromatographic data cube for a given discovered analyte. The further analysis of this data cube by deconvolution software (i.e., PARAFAC) is simple and the mass channels associated with separation of sample class via the F-ratio analysis can be used to readily and confidently identify which PARAFAC loading (i.e., peak profile) represents the discovered chemical component.

The entire tile-based F-ratio analysis (up to the production of a hit list) for the class comparison of four samples per class was carried out on a desktop personal computer in twenty minutes. The application of this quick method for the analysis of complex samples will reduce the bottleneck caused by overwhelming amounts of data. With the aid of the tile-based F-ratio software, scientists will be able to quickly and easily analyze multi-dimensional data sets comprehensively so that important analytes or chemical fingerprints that glean useful information from a particular experiment are discovered.

References

[1] Z. Liu, J.B. Phillips, J. Chromatogr. Sci. 29 (1991) 227-231.

[2] M. Adahchour, L.L. van Stee, J. Beens, R.J. Vreuls, M.A. Batenburg, U.A.Th. Brinkman, J. Chromatogr. A 1019 (2003) 157-172.

[3] J. Beens, M. Adahchour, R.J. Vreuls, K. van Altena, U.A.Th. Brinkman, J. Chromatogr. A 919 (2001) 127-132.

[4] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796-2804.

[5] J. Dalluge, M. van Rijn, J. Beens, R.J. Vreuls, U.A.Th. Brinkman, J. Chromatogr. A 965 (2002) 207-217.

[6] R.M. Kinghorn, P.J. Marriott, JHRC 21 (1998) 620-622.

[7] J.V. Seeley, F. Kramp, C.J. Hicks, Anal. Chem. 72 (2000) 4346-4352.

[8] R. Shellie, L. Mondello, P. Marriott, J. Chromatogr. A 970 (2002) 225-234.

[9] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, J. Chromatogr. A 1184 (2008) 341–352.

[10] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, J Chromatogr. A 1255 (2012) 3–11.

[11] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 73 (2001) 5833–5840.

[12] C. G. Fraga, B. J. Prazen, R. E. Synovec, Anal. Chem. 72 (2000) 4154-4162.

[13] K. M. Pierce, L. F. Wood, B. W. Wright, R. E. Synovec, Anal. Chem. 77 (2005) 7735-7743.

[14] T. Skov, J. C. Hoggard, R. Bro, R. E. Synovec, J. Chromatogr. A. 1216 (2009) 4020-4029.

[15] K. J. Johnson, B. J. Prazen, R. K. Olund, R. E. Synovec, J. Sep. Sci. 25 (2002) 297-303.

[16] J.C. Hoggard, R.E. Synovec, Anal. Chem. 79 (2007) 1611–1619.

[17] K. J. Johnson, R. E. Synovec, J. Chemom. Intell. Lab. Syst. 60 (2002) 225-237.

[18] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756.

[19] E.M. Humston, K.M. Dombek, B.P. Tu, E.T. Young, R.E. Synovec, Anal. Bioanal. Chem. 401 (2011) 2387–2402.

[20] A.C. Beckstrom, E.M. Humston, L.R. Snyder, R.E. Synovec, S.E. Juul, J. Chromatogr. A 1218 (2011) 1899–1906.

[21] K.K. Pasikanti, J. Norasmara, S. Cai, R. Mahendran, K. Esuvaranathan, P.C. Ho, E. Chan, Anal. Bioanal. Chem. 398 (2010) 1285–1293.

[22] X. Li, X. Lu, J. Tian, P. Gao, H. Kong, G. Xu, Anal. Chem. 81 (2009) 4468–4475.

[23] C. Ma, H. Wang, X. Lu, H. Wang, G. Xu, B. Liu, Metabolomics 5 (2009) 497–506.

[24] Y. Qiu, X. Lu, T. Pang, C. Ma, X. Li, G. Xu, J. Sep. Sci. 31 (2008) 3451–3457.

[25] D. L. Massart, Chemometrics: a Textbook, Elsevier Sciences Ltd, New York, 1988.

[26] R. O. Duda, P. E. Hart, Pattern Classifications and Scene Analysis, Wiley, New York, 1973.

Table Captions

Table 3.1. The actual concentrations are shown in ppm for each nominal spike concentration (first column) for all spiked analytes. The nominal concentrations are used occasionally in the text for brevity and clarity.

Table 3.2. The concentration ratios for each spiked analyte for each paired sample class are reported. The nominal 0 ppm versus 100 ppm comparison is the most challenging comparison. (a) The concentration ratio is not defined since the denominator is essentially zero.

Table 3.3. Results showing the increased sensitivity contrast achieved by using the tile-based F-ratio software versus the previous pixel-based F-ratio software. The known pixel-based false positive F-ratio values (Figure 3.4) have been decreased by two orders of magnitude, while the known true positive F-ratio values have increased by two orders of magnitude.

Table 3.4. The hit list resulting from the tile-based F-ratio software comparison of a nominal 100 ppm spiked diesel with the 0 ppm unspiked diesel. The first false positive in the hit list is the last entry in the table. At this stage of the software development, redundant hits caused by splitting of each of the eight analytes were removed by analyzing the mass spectrum of each in adjacent grid tiles. Each tile # on ^{1}D and ^{2}D is defined for each of the four grid schemes, so each of the ^{1}D and ^{2}D locations per each grid scheme is at a slightly different 2D locations in the GC × GC separation (as illustrated in Fig. 3.7). For example, for Hit #1 there will be a tile location #34 on ^{1}D and tile location #5 on ^{2}D for each of the four grids, but the 1-chlorohexane peak is centered optimally in only one grid (t1 in this case).

Figure Captions

Figure 3.1. The GC × GC – TOFMS base ten logarithm (log10) contour plot of the TIC of a nominal 100 ppm spiked diesel sample and the locations of all nine spiked compounds (eight analytes and the internal standard). The first dimension elution order and column 1 retention time of the nine spiked analytes is as follows: 1: heptane at 2.2 min; 2: 1-chlorohexane at 4.6 min; 3: o-xylene at 5.5 min; 4: bromobenzene at 6.4 min; 5: 3-octanone at 7.8 min; 6: butylcyclohexane at 9.0 min; 7: 1-bromoheptane at 9.2 min; 8: 5-decyne at 9.3 min and 9: cyclohexylbenzene at 17.2 min.

Figure 3.2. (A) A calibration curve is plotted using PARAFAC signal volumes as a function of the spike concentration for a non-native analyte, bromobenzene. (B) A calibration curve is plotted using PARAFAC signal volumes as a function of the spike concentration for a native analyte, butylcyclohexane. The PARAFAC signal volumes correlate linearly with the standard concentration spike levels. From a linear fit of the data, based on the standard addition method, the native concentration of butylcyclohexane in the diesel was calculated, and found to be 1,569 ppm. This SAM approach was performed to determine the concentration for all native analytes in the unspiked diesel.

Figure 3.3. (A) The unfolded raw 2D chromatogram at *m/z* 77, zoomed in on the bromobenzene peak shows four separate injections of a 113 ppm bromobenzene spiked diesel and four separate injections of a non-spike diesel as labeled in the zoomed in box. The further zoomed in second

dimension peak in the dashed box, produced the highest pixel-based F-ratio value for bromobenzene (max F-ratio equal to 216). (B) Two Gaussian peaks have been superimposed, showing the first dimension bromobenzene peak sampled by the GC × GC modulator, for the two sample injections with the most retention time shifting. (C) A zoom in view of (B) shows the effect of phasing on the within-class variation of peak height (eg. the modulated peak at 386.2 s). The black bracket indicates the size (20 ms) of the tile in the second dimension (subsequently used with the tile-based F-ratio software). (D) The average chemical spectra for the bromobenzene peak.

Figure 3.4. (A) The chromatographic location in the raw 2D data of a high hit number (from application of previous pixel-based F-ratio software) at the selective mass channel *m/z* 95, that appears on the shoulder of a peak (native compound in diesel). Four injections of a 113 ppm bromobenzene spiked diesel (solid) and four injections of a unspiked diesel (dashed) are shown. (B) Random fluctuation of the detector coincidently co-vary with the sample classes at 1404.71 s, producing a large false positive hit for this location. The pixel-based F-ratio value at *m/z* 95 for this hit was 249.

Figure 3.5. (A) A graphical representation of the differences in position of the four tile grids used in concert with the tile-based F-ratio software. All the points within each of the grey rectangle tiles (t1, t2, t3, t4) are summed to create the four binned tile grids. Tile grid t3 is centered at the origin of the 2D chromatogram. (B) Four 113 ppm bromobenzene spiked diesel injections are shown, zoomed in upon bromobenzene at *m/z* 77. The contour lines for the signal are at ± 1 SD (67% of signal intensity, inner ring) and ± 2 SD (95% of signal intensity, outer ring) of the bromobenzene peak are shown, where SD is the standard deviation of the peak based upon a Gaussian peak shape definition. Tile grid t4 (shaded) optimally captures the bromobenzene peak.

Figure 3.6. The F-ratio spectra (value at each *m/z*) for each tile grid (t1, t2, t3, and t4) for bromobenzene are indicated in each plot. The F-ratios from the comparison of bromobenzene at 113 ppm versus 0 ppm at each *m/z* are provided, along with the F-ratio sum at all *m/z*, at the same ^{1}D and ^{2}D tile number for the four tile grid schemes that each optimally contains the bromobenzene peak. Tile grid t4 optimally captures the retention time alignment and phasing effects, and thus, provided the highest F-ratio values, indicating it contains the bromobenzene peak to the greatest extent of the tile grids. A S/N mask has been applied removing any result from mass channels without statistically significant detector signal.

Figure 3.7. A contour plot from a subset of a nominal 100 ppm spiked diesel chromatogram is shown to illustrate the four tile grids (t1, t2, t3, t4) with a width of 8 modulations along the first dimension time axis (^{1}D), and a height of 20 ms along the second dimension time axis (^{2}D). Each peak is optimally contained by only one of the tile grids (more than ~ 99 % of the peak area for typical 2D peaks).

Nominal Spike (ppm)	1-chlorohexane (ppm)	bromobenzene (ppm)	3-octanone (ppm)	5-decyne (ppm)	Heptane (ppm)	o-xylene (ppm)	butylcyclohexane (ppm)	cyclohexylbenzene (ppm)
1000	1172	938	933	1210	1562	3680	2517	1830
750	889	712	708	918	1332	3393	2288	1579
500	608	487	484	628	1104	3107	2060	1329
250	308	247	245	318	860	2802	1818	1063
100	141	113	112	146	725	2632	1683	914
0	0	0	0	0	610	2489	1569	789

Table 3.1

Concentration Comparison	1-chlorohexane	bromobenzene	3-octanone	5-decyne	heptane	o-xylene	butylcyclohexane	cyclohexylbenzene
1000 to 0	_[a]	_[a]	_[a]	_[a]	2.56	1.48	1.60	2.32
1000 to 100	8.30	8.30	8.30	8.30	2.15	1.40	1.50	2.00
750 to 0	_[a]	_[a]	_[a]	_[a]	2.18	1.36	1.46	2.00
1000 to 250	3.80	3.80	3.80	3.80	1.82	1.31	1.38	1.72
750 to 100	6.29	6.29	6.29	6.29	1.84	1.29	1.36	1.73
500 to 0	_[a]	_[a]	_[a]	_[a]	1.81	1.25	1.31	1.69
750 to 250	2.89	2.89	2.89	2.89	1.55	1.21	1.26	1.49
500 to 100	4.30	4.30	4.30	4.30	1.52	1.18	1.22	1.45
1000 to 500	1.93	1.93	1.93	1.93	1.42	1.18	1.22	1.38
250 to 0	_[a]	_[a]	_[a]	_[a]	1.41	1.13	1.16	1.35
500 to 250	1.97	1.97	1.97	1.97	1.28	1.11	1.13	1.25
750 to 500	1.46	1.46	1.46	1.46	1.21	1.09	1.11	1.19
1000 to 750	1.32	1.32	1.32	1.32	1.17	1.08	1.10	1.16
250 to 100	2.18	2.18	2.18	2.18	1.19	1.06	1.08	1.16
100 to 0	_[a]	_[a]	_[a]	_[a]	1.19	1.06	1.07	1.16

Table 3.2

	Pixel-Based	Tile Algorithm
F-ratio (*m/z* 95) false positive	249	1.1
F-ratio (*m/z* 77) bromobenzene	216	9103
Sum of F-ratio false positive	1068	88
Sum of F-ratio bromobenzene	2799	213712

Table 3.3

F-ratio Hit #	Sum of F-ratio	Tile #, ^{1}D	Tile #, ^{2}D	Grid	Compound
1	807000	34	5	t1	1-chlorohexane
2	214000	48	1	t4	bromobenzene
3	84700	70	4	t4	5-decyne
4	67300	58	5	t4	3-octanone
5	3460	41	1	t3	o-xylene
6	3180	16	2	t4	heptane
7	2070	67	5	t2	butylcyclohexane
8	1270	128	2	t2	cyclohexylbenzene
First False Positive	643	175	1	4	eicosane

Table 3.4

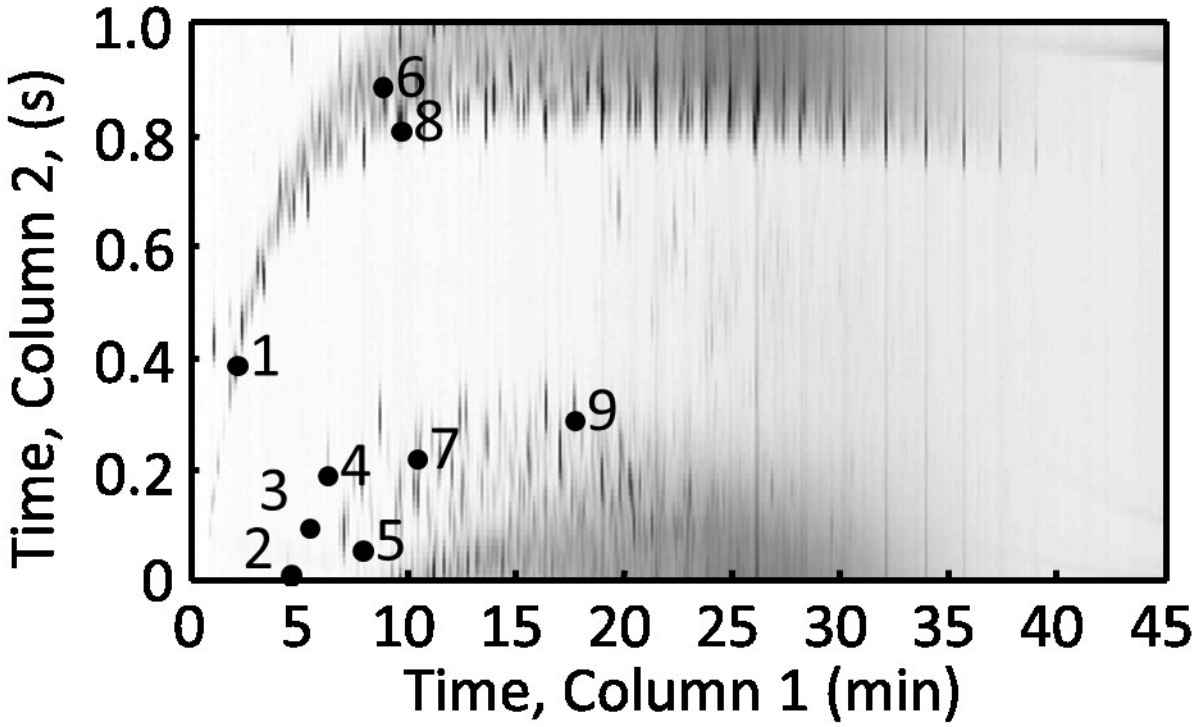

Time, Column 2, (s)
1.0
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30
35
40
45
Time, Column 1 (min)
1
2
3
4
5
6
7
8
9

Figure 3.1

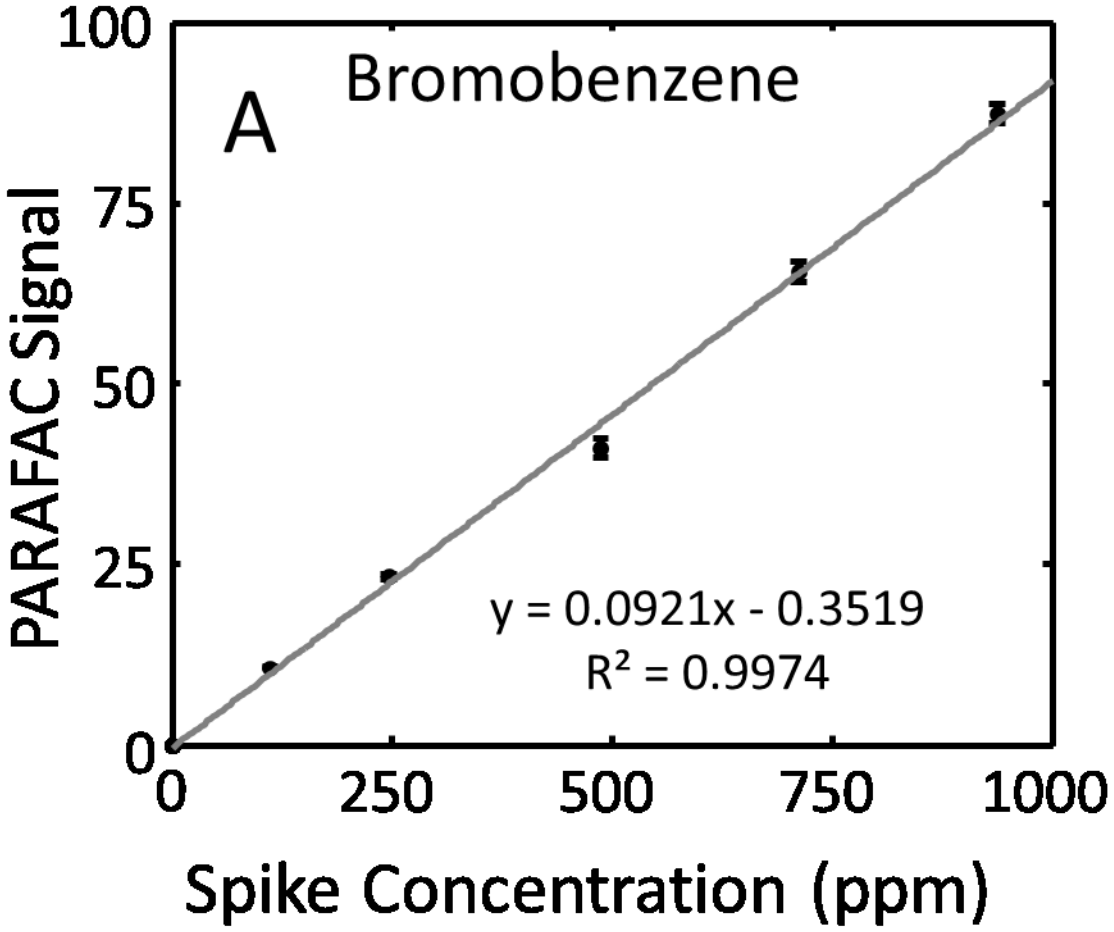

A
Bromobenzene
PARAFAC Signal
100
75
50
25
0
0
250
500
750
1000
Spike Concentration (ppm)
y = 0.0921x - 0.3519
R² = 0.9974

Figure 3.2 (A)

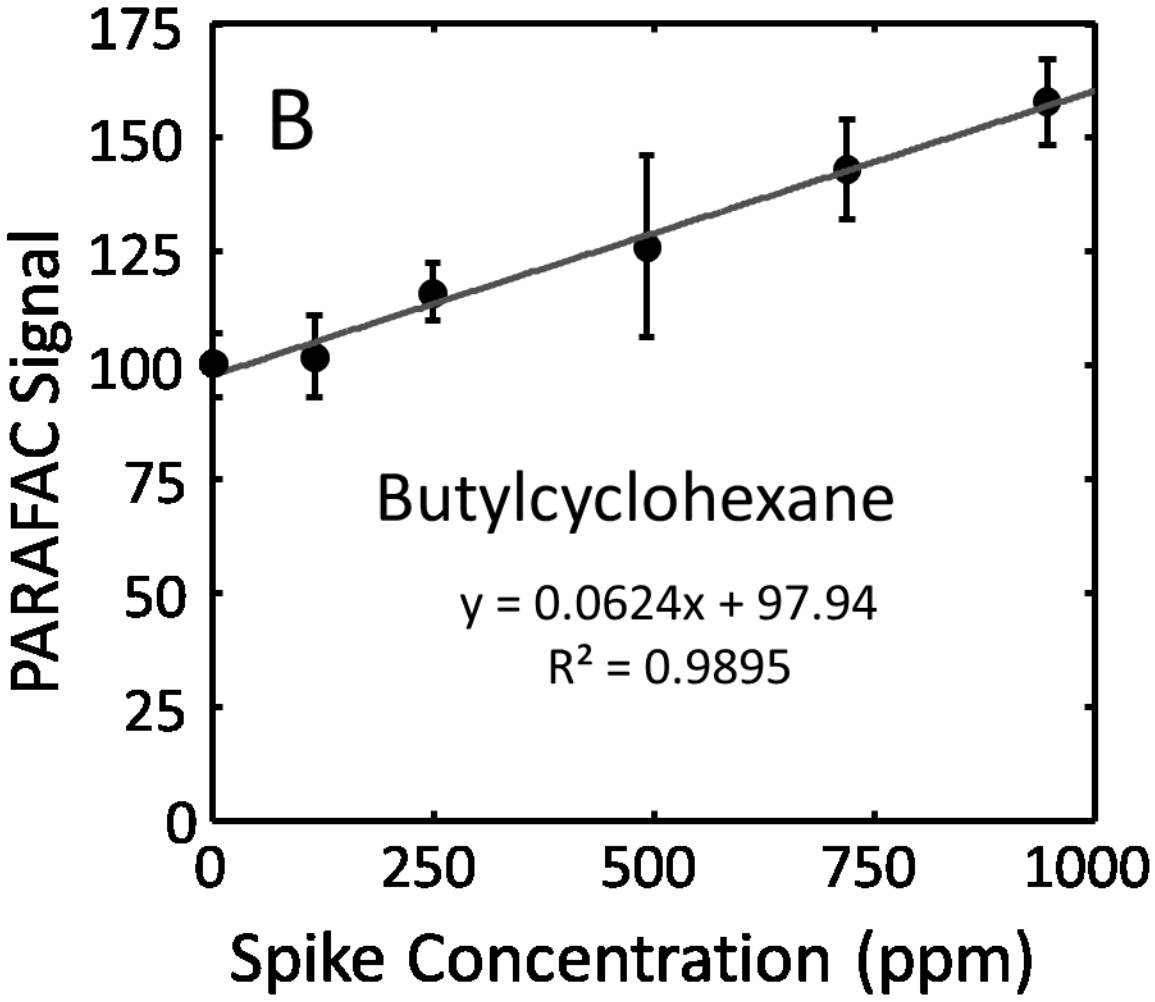

B
PARAFAC Signal
175
150
125
100
75
50
25
0
0
250
500
750
1000
Spike Concentration (ppm)
Butylcyclohexane
y = 0.0624x + 97.94
R² = 0.9895

Figure 3.2 (B)

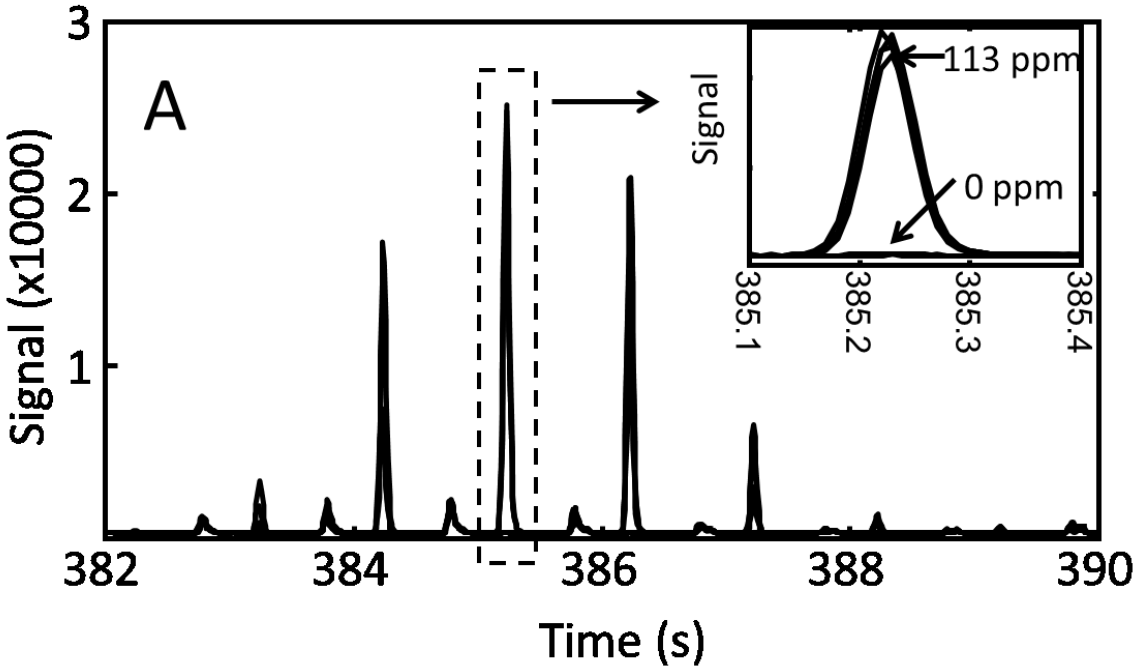

A
Signal (x10000)
3
2
1
382
384
386
388
390
Time (s)
Signal
113 ppm
0 ppm
385.1
385.2
385.3
385.4

Figure 3.3 (A)

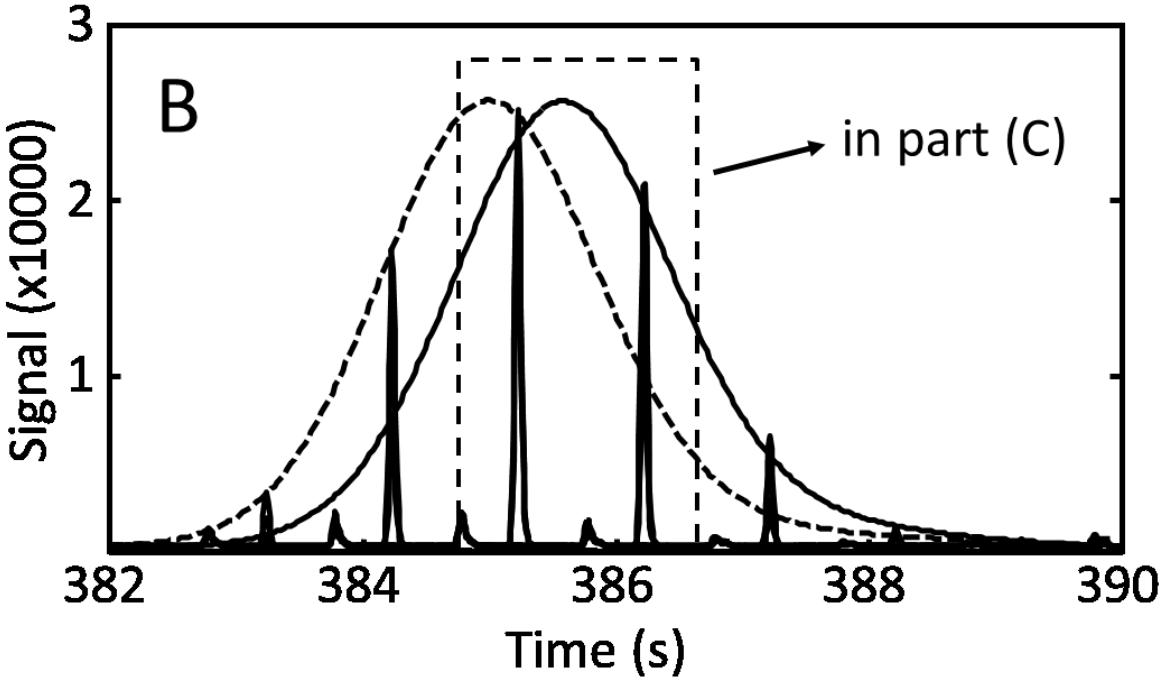


Figure 3.3 (B)

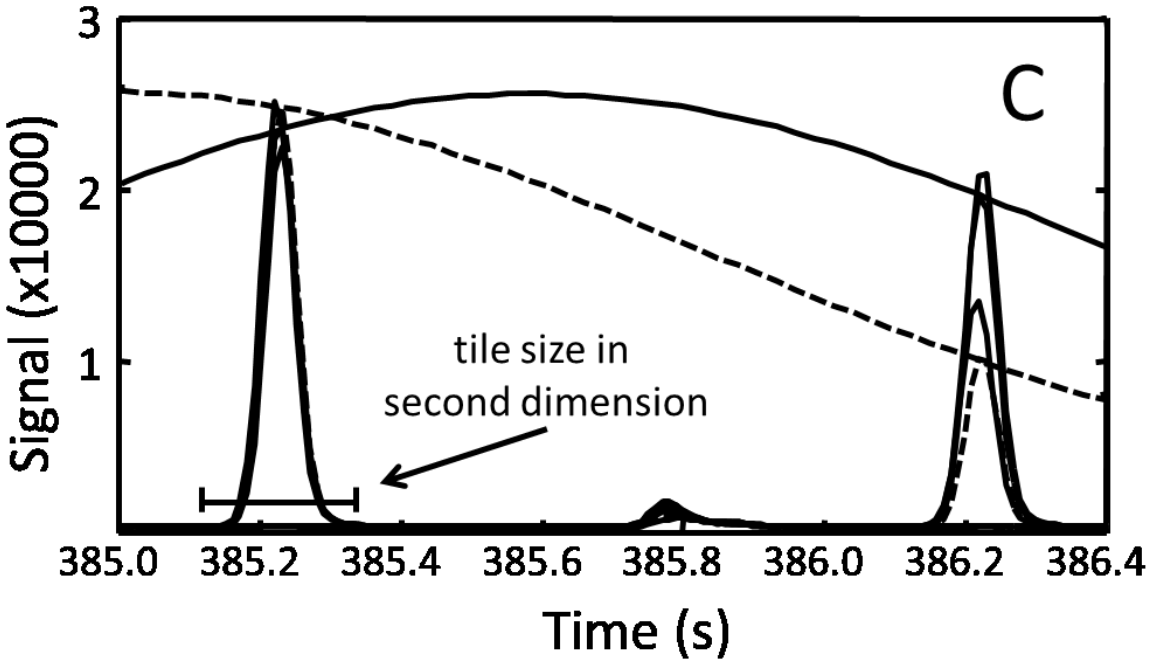

C
tile size in
second dimension
Signal (x10000)
3
2
1
385.0
385.2
385.4
385.6
385.8
386.0
386.2
386.4
Time (s)

Figure 3.3 (C)

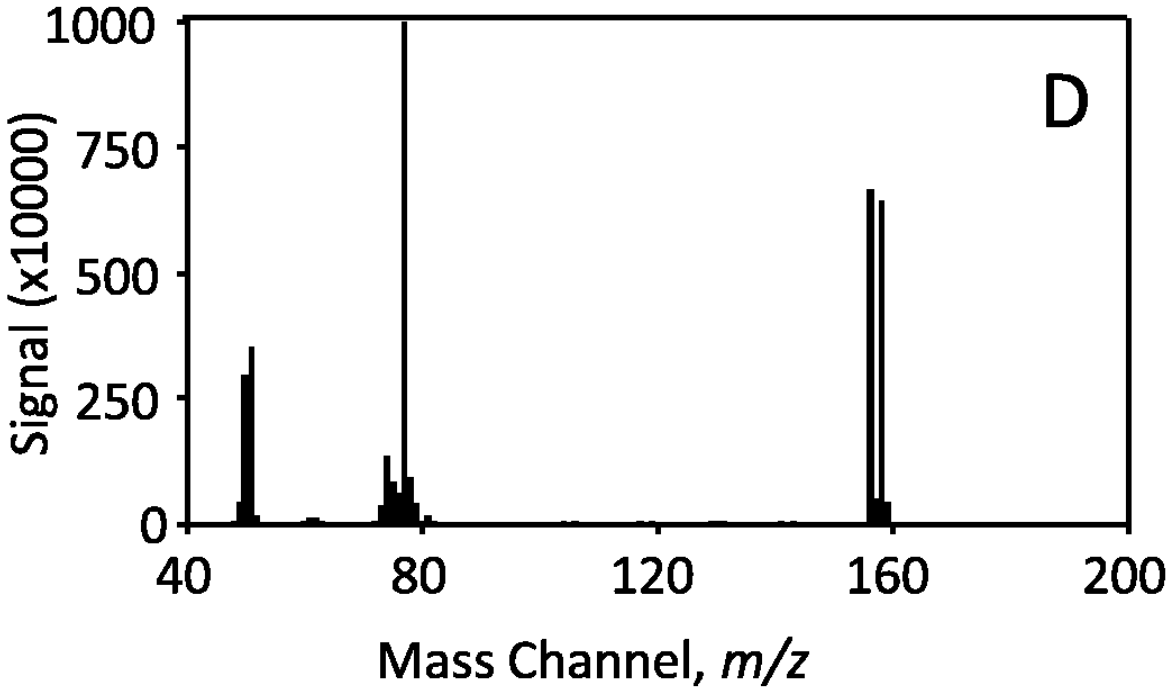

D
1000
750
500
250
0
Signal (x10000)
40
80
120
160
200
Mass Channel, m/z

Figure 3.3 (D)

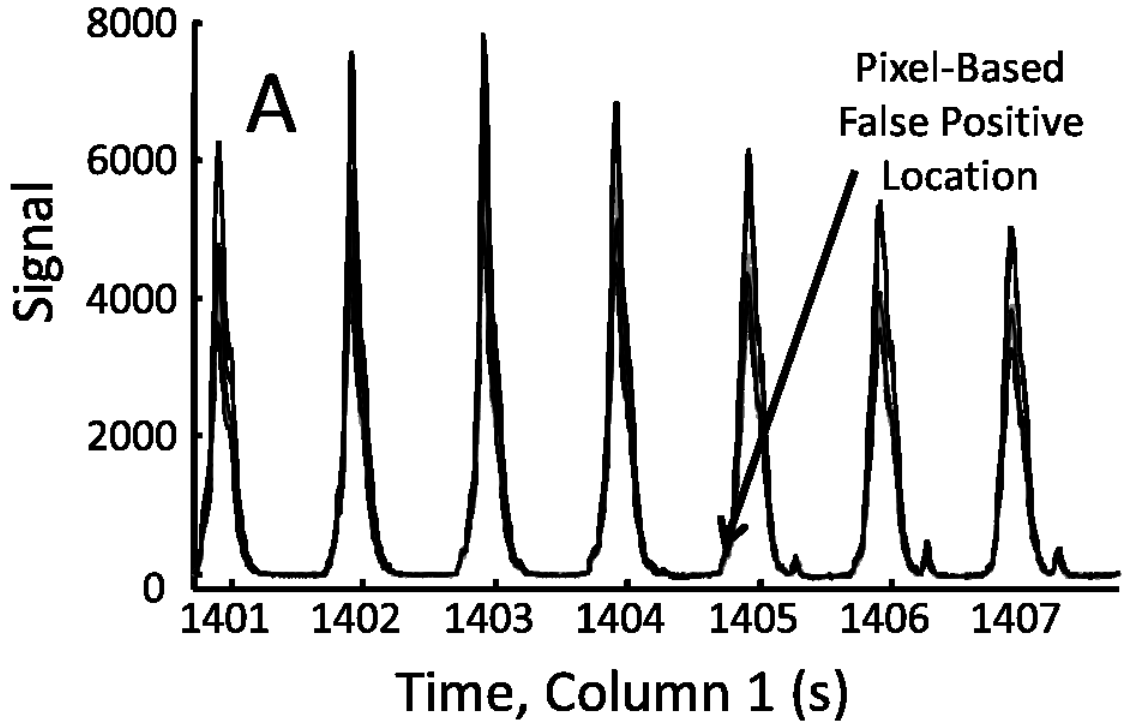

A
Pixel-Based
False Positive
Location
Signal
8000
6000
4000
2000
0
1401
1402
1403
1404
1405
1406
1407
Time, Column 1 (s)

Figure 3.4 (A)

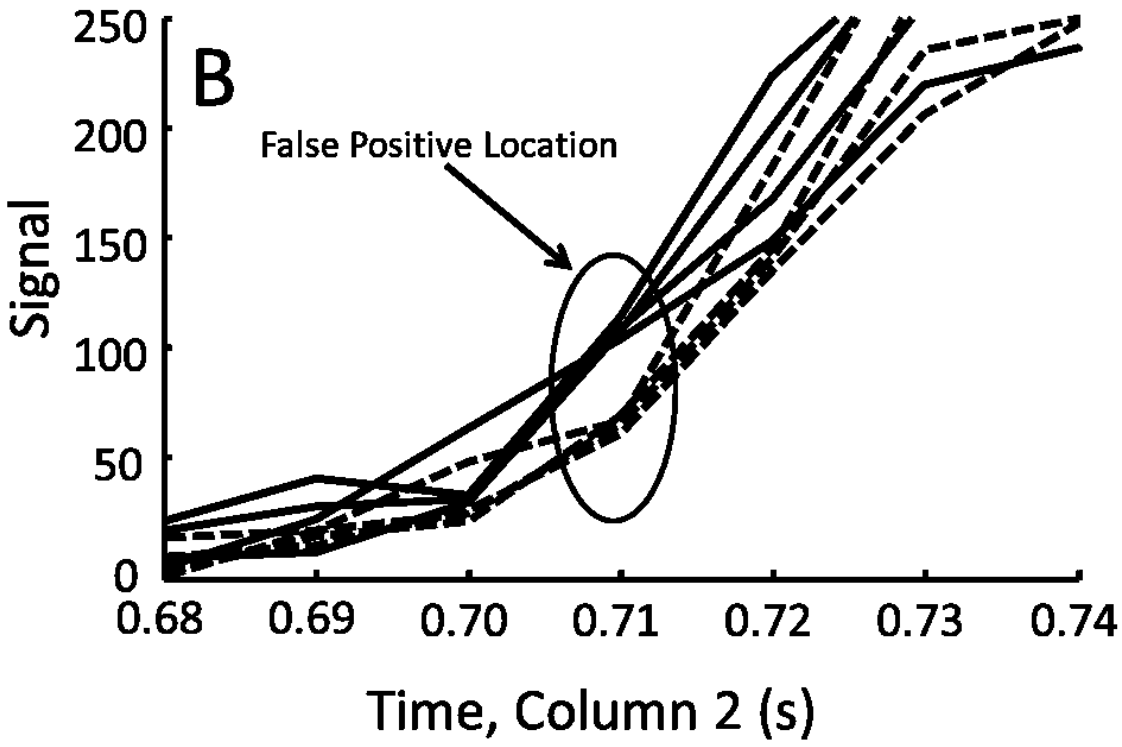
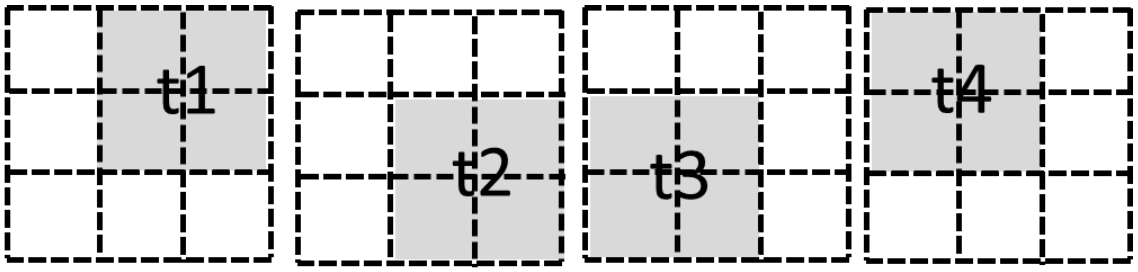

B
False Positive Location
Signal
250
200
150
100
50
0
0.68
0.69
0.70
0.71
0.72
0.73
0.74
Time, Column 2 (s)

Figure 3.4 (B)



A

Figure 3.5 (A)

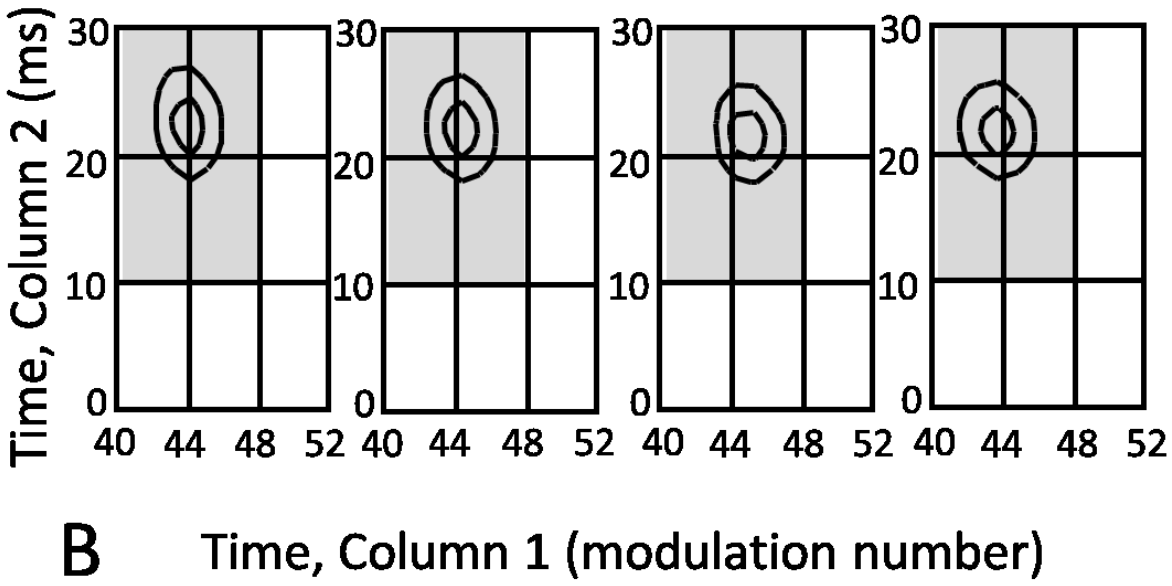

Time, Column 2 (ms)
30
20
10
0
40 44 48 52
B
Time, Column 1 (modulation number)

Figure 3.5 (B)

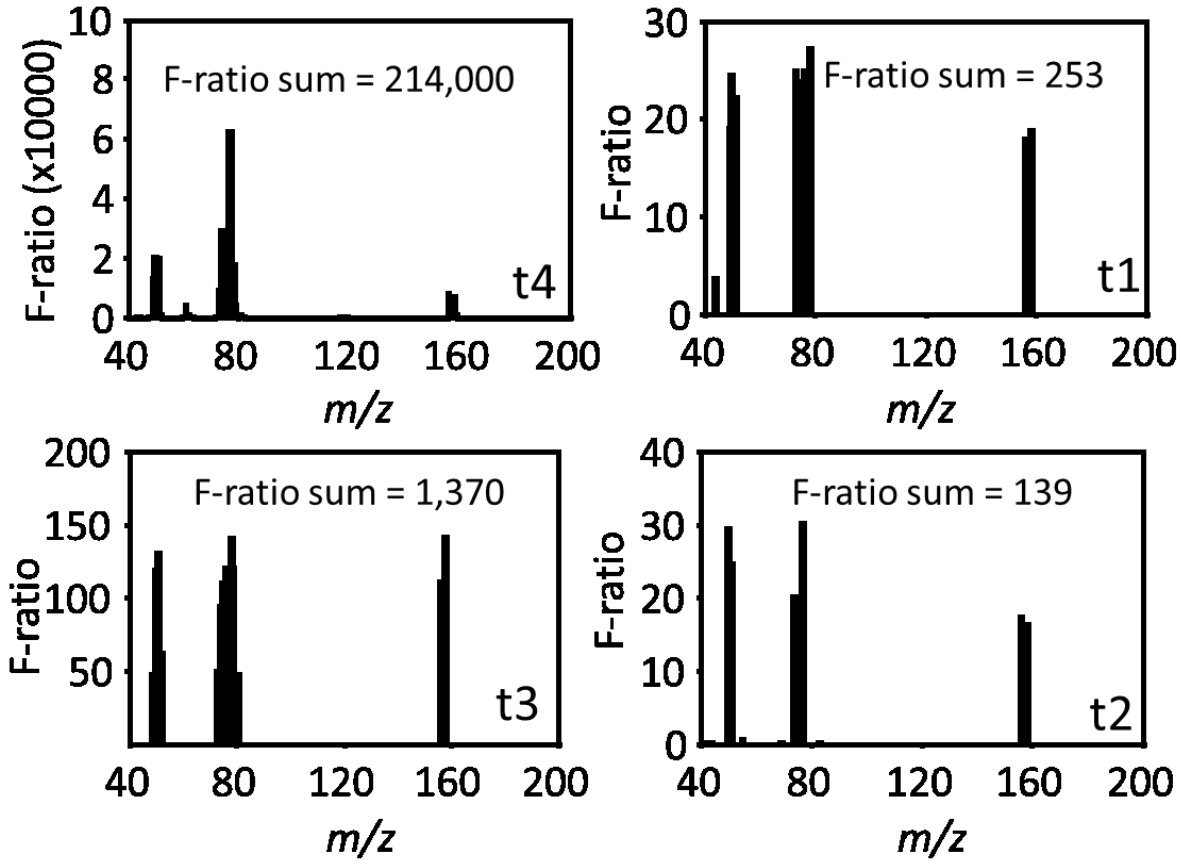

F-ratio sum = 214,000
t4
F-ratio (x10000)
m/z
F-ratio sum = 253
t1
F-ratio
m/z
F-ratio sum = 1,370
t3
F-ratio
m/z
F-ratio sum = 139
t2
F-ratio
m/z

Figure 3.6

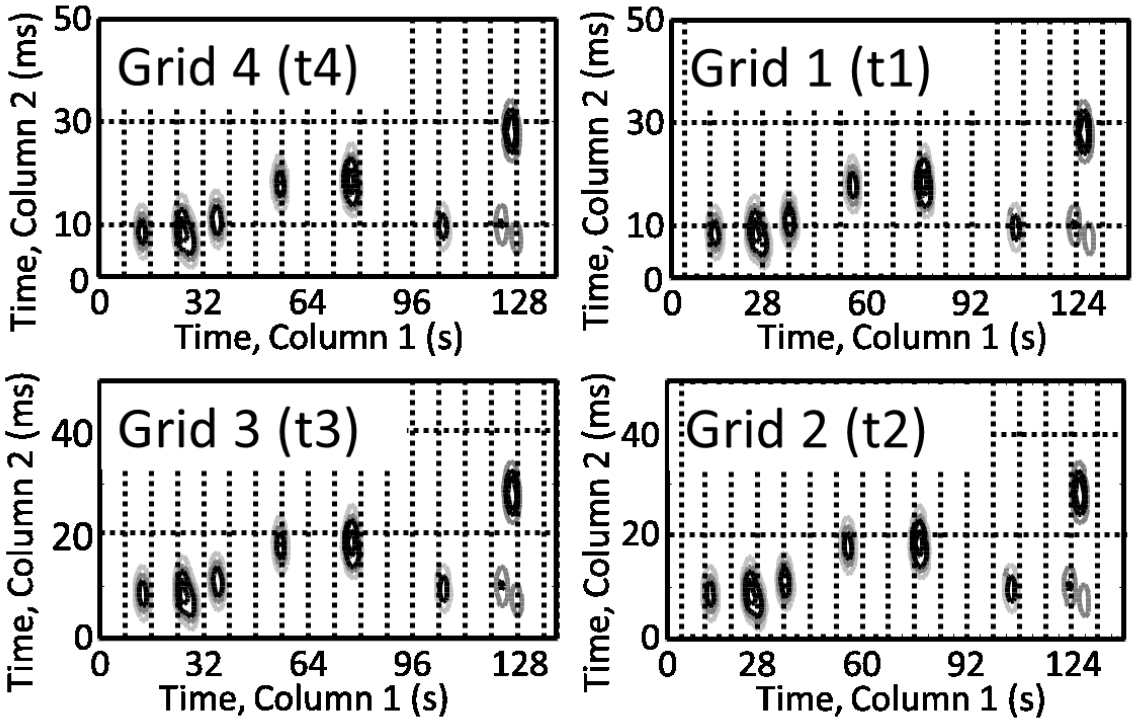

Grid 4 (t4)
Grid 1 (t1)
Grid 3 (t3)
Grid 2 (t2)
Time, Column 2 (ms)
Time, Column 1 (s)
0
10
30
50
20
40
32
64
96
128
28
60
92
124

Figure 3.7

Chapter 4:

Automation of a Novel Tile Fisher Ratio Algorithm

and the Reduction in False-Positives by Null Distribution Analysis

4.1 Introduction

Two-dimensional gas chromatography time-of-flight mass spectrometry (GC x GC – TOFMS) produces chemically complex multi-dimensional data. The experimental use of this technique and the associated data analysis strategies aim to uncover meaningful chemical signals or chemical patterns, however these are buried in a background of less meaningful chemical signal and noise. Making the challenge even more difficult, experimentation is optimally designed to analyze multiple injections of different samples, producing yet higher dimensionality data. Feature selection software then become increasingly important. For routine use, automated software is needed to comprehensively analyze GC x GC – TOFMS data so that important analytes and/or chemical fingerprints can be ascertained during non-target discovery-based experimentation.

Targeted GC x GC – TOFMS investigations include experiments where the chemical compounds of interest are known prior to instrumental and data analysis and chemical standards exist for the targeted chemical species. In non-target discovery GC x GC – TOFMS investigations, the compounds of interest are not known and the complex chemical data is mined for representative chemical features of a particular experiment. Non-targeted approaches can be either supervised or unsupervised, where supervision refers to either external calibration or prior classification of chromatograms as they relate to the experimental design. The use of these methods has been reviewed [1]. Generally these supervised or unsupervised methods are applied on a pixel-based level or peak table-based level. Pixel-based analysis compares every 3D data point (i.e. pixel) in each sample chromatogram with limited preprocessing (baseline correction or background subtraction are common necessary preprocessing methods). Peak table-based methods utilize baseline correction, peak deconvolution and identification, peak integration, and

arrangement of every peak in a single chromatogram into a table (i.e. peak table). Recently, we proposed another form of analysis, tile-based analysis, that solves problems with misalignment of multiple chromatograms without explicitly aligning the data, a strength of peak table-based methods, while remaining sufficiently unbiased to particular chromatographic peak shapes or mass spectral signal characteristics like pixel-based analysis (cite first F-ratio paper).

The 2D misalignment of GC x GC – TOFMS chromatographic data across different samples can make non-targeted analysis difficult [2]. Alignment algorithms are used in cases where there is significant retention time variation due to injection timing or inadvertent carrier gas pressure issues. If the misalignment is bad enough, this limits the scope of the data analysis methods to targeted methods. However, many algorithms have been developed to correct for small systematic retention time imprecision. Pixel-based [2-5] and peak table-based approaches (often provided by the instrument software (e.g. ChromaTOF) [6-10]) have been developed and their current use has been recently reviewed [1]. The limitation of these alignment strategies for comprehensive non-targeted GC × GC – TOFMS analysis is that alignment is computationally expensive and preserving the peak signal "volumes" of every analyte during peak shifiting is difficult and often exhibits shortcomings. Therefore, I we were motivated to develop a suitable, computationally fast approach utilizing data summation within tile grids (Chapter 3).

In experimental design where sample class information is known and the data is sufficiently aligned, supervised data analysis methods can be used for nontarget feature selection analysis. The two most popular supervised methods include Fisher ratio (referred to herein as F-ratio) and partial least squares discriminant analysis (PLS-DA). F-ratio methods have been applied to GC × GC – TOFMS data previously at the pixel-level [11-14] and at the peak table-

level (LECO Fisher Ratio ChromaTOF 2009). PLS-DA has also been recently applied successfully to peak table-based GC × GC – TOFMS data [15-18].

In the previous chapter, I detailed a computationally fast tile-based F-ratio method that decreases the number of false-positives normally detected by pixel-based analysis. Most of the false-positives were due to systematic first dimension misalignment in GC × GC as a result of desynchronized modulation of a peak from one sample to another, referred to as "phasing." Briefly, the approach works by creating a 2D grid (encompassing the entire GC × GC separation) composed of 2D tiles, where each tile is large enough to capture the retention time variation in both the first (phasing) and second separation dimensions. The software then sums all signals within the tile as a function of each mass channel. One set of tiles in a chromatogram is not sufficient because some peaks will become split between two tiles, thus four adjacent, but overlapping 2D grids of tiles are used so that one tile within a single grid will optimally capture any given peak in the GC × GC separation. The use of four grids, combined with summing of all chromatographic signal within a tile (at a given *m/z*), provides data reduction and added sensitivity, similar to peak-based algorithms. Sensitivity is improved because a single grid (of four) that captures a given analyte peak the best in one of its tiles will have the least amount of interference. The signal-to-noise (S/N), i.e., the sensitivity contrast, for a given analyte peak is improved roughly by the square root of the number of points summed, and by minimizing the impact of the GC × GC phasing effect. The F-ratio of each mass channel for each tile in the four grids is calculated and an average F-ratio is used to rank 2D tile locations as they relate in importance to the experimental design.

While the tile-based F-ratio approach provided a computationally fast way to solve misalignment problems and at the same time improved sensitivity contrast, it somewhat

complicated the final analysis because there are multiple 2D tile locations per chemical compound that complicate the final analysis of the data. Herein, I discuss a simple algorithmic method that focuses the multiple 2D tile locations back to the original high resolution 3D data. Briefly, the 2D location of the maximum signal-difference in each tile between the two F-ratio classes is calculated and every F-ratio value (over all mass channels), the average F-ratio value, and the original tile and grid locations, is pinned (i.e. assigned) to that location. A simple cluster analysis algorithm is used to remove redundant 2D pin locations, preserving only the pinned tile information with the highest average F-ratio value (average over all mass channels) at a single 2D location. Resulting is a list of 2D pins and retention time windows ranked by average F-ratio value and includes an indexed array of F-ratio values at each mass channel for that location (an F-ratio "spectra"). The window locations can be easily used for further deconvolution or identification, and the F-ratio spectra for that 2D window location provides the particular mass channels that are the most important for the comparison of the two sample classes and are the most chemically selective in the chromatographic separation for the chemical compound of interest.

While the previously reported tile-based F-ratio software significantly reduced the number of false positives found with pixel-based analysis, there are still instances where there is significant random covariance of tile-summed chemical signal with sample class. Because of the large amount of this type of statistical hypotheses tested in the entire tiled GC x GC – TOFMS chromatogram, there is a finite, significantly high number (as opposed to a probability) of those statistical tests will falsely reject the null hypothesis when it is true, i.e. false positives. This type of problem has been discussed extensively in the statistics literature, particularly in application to genetic microarray analysis [19-26]. We can, however, leverage the thousands of measurements

in a comprehensive detection platform to experimentally determine the distribution of potential false positive results in an experiment. No assumptions need to be made about the underlying distribution, because the frequency of random covariance of signal with sample class can be experimentally determined. Our approach is to algorithmically mix the chromatographic data into two classes that represent the null hypothesis we are trying to reject in F-ratio analysis (i.e. that there is no difference between the two classes). Calculating F-ratios for every tile in the null class comparison produces a distribution of F-ratios that represent the magnitude and frequency of F-ratio values from an experiment where there is inherently no difference between the two classes. With this approach, the probability of detecting a single false F-ratio event at a given magnitude in our non-null class comparison experiment is known and an F-ratio threshold can be used to remove 2D locations that are suspected to be false-positives. Herein we apply the tile-based F-ratio analysis software to the standard addition method by spiking non-native chemicals into a diesel fuel at low concentrations to demonstrate the tile-based F-ratio algorithm and the reduction in false-positives by null distribution analysis in the context of a challenging complex sample matrix and low signal-to-noise.

4.2 Experimental

4.2.1 Sample Preparation

Four non-native chemical compounds were spiked at the following nominal concentrations into an ultra-low sulfur diesel (ULSD) fuel: 100, 50, 25, 12.5, 6.2, 3.2, 1.6 and 0 parts-per-million by mass (ppm). An internal standard, 1-bromoheptane (also non-native to the diesel fuel), was spiked at a concentration of 1 part-per-thousand (ppt) into each diesel sample. The four spiked compounds were bromobenzene, 1-chlorohexane, 5-decyne, and 3-octanone. All diesel samples were prepared gravimetrically using a 5-place analytical balance, and the actual

(not nominal) concentration for each analyte at each spike level is provided in Table 1 and the mole amount injected on column is provided in Table 2. Actual concentration of spiked analytes are used as much as possible in this manuscript; however, it is more clear occasionally to use the nominal spike concentrations when referring to a particular concentration comparison with many analytes.

4.2.2 Instrumental Parameters

The GC × GC – TOFMS instrumental platform consisted of an Agilent 6890N gas chromatograph equipped with an Agilent 7683 autoinjector (Agilent Technologies, Palo Alto, CA) coupled with a Leco Pegasus III time-of-flight mass spectrometer, and equipped with a 4D thermal modulator upgrade (Leco, St. Joseph, MI). The GC × GC – TOFMS instrument was used to analyze the diesel fuel samples at all spike levels. The primary column of the GC × GC (column 1) was a 20 m x 250 μm i.d. x 0.5 μm DB-5 film (J&W Scientific/Agilent Technologies, Santa Clara, CA, USA) and the secondary column (column 2) was a 2m x 180 μm i.d. x 0.2 μm RTX-200MS film (Restek, Bellefonte, PA). The GC instrument inlet was set at 275 °C and the transfer line was set at 305 °C. Column 1 was held at 50 °C for 0.25 min and then increased at 5 °C/min to 300 °C, where it was held for 5 min. Column 2 was initially set at 55 °C and followed the same temperature program as column 1 giving a total run time of 55.25 min. The modulator was kept 20 °C higher than column 1, and the modulation period was 1 s. The GC instrument was set to maintain a constant (ambient temperature and pressure corrected) flow rate of 2 ml/min at the outlet of column 2, with helium was used as the carrier gas. The ion source was set to 300 °C and the detector voltage was set to 1600 V. Mass channels, *m/z* 41-340, were collected at 100 spectra/s after a 6 s solvent delay. A 1 μl injection of each diesel sample was made in split mode with a split ratio of 200:1. All diesel samples were injected in quadruplicate

on the instrument. An example separation run is presented in Fig. 1, with the GC × GC–TOFMS log10 contour plot of the TIC of a nominal 100 ppm spiked diesel sample. The locations of all spiked compounds are indicated in Fig. 1 (four analytes and the internal standard).

4.2.3 Data Analysis Methodology

GC × GC – TOFMS data from all of the 32 runs were imported from LECO's ChromaTOF software v 3.32 (LECO, St. Joseph, MI) to Matlab v 7.0.4 via an in-house data converter. The imported GC × GC – TOFMS data was then analyzed with the in-house developed Tile-Based F-ratio algorithm. For software validation, an in-house developed target-PARAFAC GUI was used to extract the quantitative signal "volume" at each of the spiked analyte chromatographic locations and to obtain the normalization values for the internal standard [4]. We refer to each analyte signal as the signal volume because it represents the signal area for a given analyte along the column 2 time axis summed over all column 1 modulations in the GC × GC separation.

An F-ratio is the class-to-class variation of the detected signal divided by the sum of the within-class variations of the signal [11, 27, 28]. The class-to-class variation is calculated as

$$\sigma_{\text{cl}}^2 = \frac{\sum(\bar{x}_i - \bar{x})^2 n_i}{(k - 1)} \quad (1)$$

where n_i is the number of measurements in the ith class, $\bar{x}_i$ is the mean of the ith class, $\bar{x}$ is the overall mean, and k is the number of classes. The within-class variation is calculated as

$$\sigma_{\text{err}}^2 = \frac{\sum\left(\sum\left(\bar{x}_{ij} - \bar{x}\right)^2\right) - \left(\sum(\bar{x}_i - \bar{x})^2 n_i\right)}{(N - k)} \quad (2)$$

where $\bar{x}_{ij}$ is the *i*th measurement of the *j*th class, and *N* is the total number of sample profiles. A F-ratio is then calculated as the ratio between the two variances,

$$\text{Fisher ratio} = \frac{\sigma_{cl}^2}{\sigma_{err}^2} \quad (3)$$

The GC × GC pixel-level data is summed using a 2D grid of 2D tiles, as a function of m/z, which provides data reduction in the 2D time domain. Four 2D grids of tiles are applied to optimally capture each peak. Equations (1) through (3) are then applied as a function of 2D tile location. The F-ratios at each *m/z* are averaged to give a single average F-ratio at each 2D tile location for ranking of 2D chromatographic locations in importance (arranged in a table, or hit list) and the mass channel specific F-ratio values indexed to each location provide additional information for deconvolution, identification, and quantification.

All tile-based F-ratio analyses were conducted with the entire *m/z* range (41-340). Prior to tile-based F-ratio analysis, the data was baseline corrected and normalized to the internal standard signal for each sample (normalization values provided by quantification with t-PARAFAC, 4). A signal-to-noise threshold was applied where detector signal that is less than three times the standard deviation of a noise region (the noise region is the first ten seconds of detector signal during a representative chromatographic run) is left out of the analysis.

For the null class comparison experiments, many possible mixing of samples into two classes that would represent the null case are possible and performing a tile-based F-ratio analysis on all of them becomes extremely computationally intense. Therefore, for this study only four of those possible arrangements of samples were used. It is important however, that the null classes be balanced, such that there are an equal number of samples from the original classes within them. For instance, to produce null classes in the comparison of four injections of a spiked

diesel and four injections of a blank diesel, there would need to always be two injections of a spiked diesel and two injections of a blank diesel in each class, otherwise the null distribution would be biased toward the alternative hypothesis (i.e. the hypothesis that the classes are in fact different) not the null hypothesis.

To remove redundant hits caused by the tile-based approach I utilized novel "pinning" and "clustering" algorithm. As redundant hits are due to the same analyte being sampled multiple times by the tile-based approach utilized in the F-ratio software, redundant hits have very similar, if not the same, two-dimensional chromatographic peak location. The pinning algorithm analyzes each hit found by the initial tile-based F-ratio approach and locates the maximum signal difference (between the sample classes) observed in any mass channel. These locations are then denoted in the 2D chromatographic space, analogous to how locations may be pinned on a map. Since redundant hits are attributable to the same analyte peak eluting at the same, or similar, 2D retention times multiple pins are consolidated into small regions. With the use of a simple cluster algorithm we can remove the pin location with redundant F-ratio information in an automated fashion. The cluster algorithm ranks the pins by their associated average F-ratios, and then removes those that are within a user-specified 2D chromatographic distance from one another, based on the hypothesis that the pin having the highest average F-ratio best locates the peak maximum of the class-distinguishing analyte. The highest F-ratio pin in a given cluster is preserved and assigned as a hit in the final hit list, while the lesser pins are removed.

4.3 Results

A schematic of the algorithm is shown in Fig. 2. For F-ratio analysis (a supervised nontarget analysis technique) the experimental design must include prior knowledge of two different sample classes. Data is collected by GC x GC – TOFMS to produce highly complex

data with high peak capacity to ensure that discovery of chemical compounds as they relate to the experimental design is possible. Once the data is collected and the data has been imported preprocessing can begin. The preprocessing step is aimed at reducing the inherent instrumental variation in the data set that contributes to the within-class variation (Eq. 2, and the denominator in Eq. 3). Baseline correction corrects for background mass spectrometry noise, while tiling corrects the inadvertent dissynchronous modulation from one sample to the next (small chromatographic misalignment, i.e. phasing). Baseline correction techniques have been thoroughly reviewed elsewhere so have not been included here for brevity (cite reviews). Tiling is a new approach to chromatographic alignment (cite paper), where four grids of tiles are calculated and the data is summed within each tile to capture the small retention time misalignment seen.

The tile-based approach assists in the discovery of changing analytes and reduces the number of high F-ratio false positives due to retention time misalignment and covariance of detector noise with the sample classes. However, a single grid of tiles is not sufficient to ensure discovery of changing analytes regardless of their chromatographic location, it is necessary to use multiple grids (Fig. 3). As a consequence of the multiple grids, each analyte is sampled via the tile-based F-ratio software multiple times, which can lead to multiple tile hits for an analyte which is significantly changing between sample classes (Table 3). In complicated sample matrices, such as diesel, the various grid schemes also sample neighboring peaks (chlorohexane and neighboring peak in grid 2, Fig. 4) precluding the possibility of simply using only mass spectral matching to eliminate the redundant hits for a given analyte. To effectively remove redundant hits, it is necessary to transform the tile-based F-ratio results from tile grid-based

space back to chromatographic pixel-based space, taking advantage of the resolution originally present in the two-dimensional (2D) chromatographic data.

To remove redundant hits from the tile hit list, we have developed an algorithm which uses the outputs of the tile-based F-ratio approach and the original pixel-based 2D chromatographic data to locate the 2D chromatographic locations associated with significant F-ratio values. Using the class information and the tile-base F-ratio results for the analysis, the pinning portion of the algorithm finds the 2D location at which the signal difference for the classes is most different for a given analyte. The 2D location is indexed for the given tile, and the process repeated for the entirety of the tile-based hit list. Since redundant hits resulting from the use of multiple grid schemes can be attributable to a given analyte (and their 2D chromatographic locations), the 2D locations corresponding to redundant hits are the same, or very similar and they can be consolidated with a simple clustering approach. The unique points, or pins, are then able to be displayed on a representative 2D chromatogram according to their locations. Indices are maintained such that in addition to location information, the pins also have the average F-ratio value, tile and grid location and the associated F-ratio spectra for choosing the m/z to use in the visualization and quantification of the pinned class distinguishing analyte peak.

The distance by which the function searches for redundant pins is a user-defined parameter. Minimal parameter values, such as +/- 2 chromatographic data collection points in the first and second column dimensions will primarily remove redundant signal difference locations that are due to imprecision in the peak maximum of the given class-distinguishing analyte discovered by the tile-based F-ratio software. Larger parameter values, similar in size to the tiles used for initially binning the data, will accomplish the goal of eliminating redundant tile F-ratios,

while conserving the tile F-ratio spectrum of the highest average F-ratio (average across all m/z) and pin locations. Overall, increasing the distance limits of the cluster algorithm will reduce the occurrence of redundant hits, but eliminates entries in the final hit list table for class-distinguishing analytes which closely elute. It is important than to keep the size of the cluster window similar to the original tile size, so that the associated spectral information in the F-ratio spectra of any closely eluting class distinguishing compounds will be seen in the final analysis and identification of each hit in the final hit list. The results of applying this redundant hit removal to Table 3 are seen in Table 4. Each spike analyte is seen only once in the final list of class distinguishing compounds and the exact location of the highest signal difference is pinpointed (columns labeled pin location 1D and pin location 2D, Table 4) while at the same time the F-ratio value for each m/z is preserved to allow for quick identification, deconvolution, and will allow for multiple identifications per location if, in the future, the detection of closely eluting class distinguishing compounds becomes imperative (last column, # of m/z passing, is shown for easy visualization in the table rather than the actual F-ratio value at each m/z for each hit).

Table 4 shows the existence of many false positives, i.e. non-spiked compounds in this experimental design. With the high number of F-ratio statistical hypotheses tested in a tiled GC x GC – TOFMS chromatogram, the simple probability of observing single high F-ratio events becomes an observable fact (often referred to as the multiple hypothesis testing problem). Many strategies aimed at mitigating this problem have been proposed in other detection platforms (mainly genetic microarray analysis), but none have been applied to GC x GC – TOFMS data.

The advantage of having a separation that contains thousands of chemical measurements is that we can experimentally determine the magnitude and frequency of F-ratio values for those

compounds that do not change between the F-ratio comparison classes. By shuffling our chromatograms into a null class comparison (in this experiment, balanced classes containing equal numbers of spiked and unspiked diesel chromatograms) this is computationally easy. However, the number of possible ways to shuffle the chromatograms is large so the current approach uses four possible arrangements of chromatograms into null classes, because the shuffling and calculation of all F-ratio values for each chromatographic tile can be done in a reasonable time frame with current personal computer computational abilities and provides sufficient distribution data density.

Our initial hypothesis is that there would be particular m/z's, such as column bleed masses or representative atmospheric detecting m/z's, that would show a particularly different distribution of F-ratio values. Fig. 5 shows the m/z dependence of the highest observable F-raito value in the distribution of F-ratio's per m/z for the four different null class comparisons. If there were particular m/z's that were ubiquitously causing a shift in the distribution of F-ratios, the sum of the four plots in Fig. 5A would show the constructive increase in those m/z's in Fig. 5B; however, this is not the case. For the comparison of replicate injections of spiked diesel and non-spiked diesel, there does not seem to be a bias in the distribution of null F-ratio values that is dependent on m/z.

Because there is no bias in the distribution of null F-ratio values as a function of m/z, we combine all m/z and the four different null class comparisons into a single distribution of null F-ratio values. That distribution is shown in Fig. 6A for the null classes in the comparison of 6.2 ppm nominal spike level to blank diesel. Most null F-ratio values are approximately 1, with a few single occurrences of high values, up to 250. This distribution allows us to algorithmically

define an allowable number of null F-ratio detection events acceptable as a fraction of the total number of observable F-ratio statistics.

However, for GC x GC – TOFMS analysis, we are interested in chromatographic locations that have a representative spectra associated with a single chemical component. The identification of a chemical compound is most thoroughly evaluated by the presence of at least three selective ions, a primary selective ion and two confirmatory ions. Therefore, our tile F-ratio approach requires that at least three m/z's be present in a tile to be considered a potential class distinguishing location containing a class distinguishing chemical compound (Fig. 6B shows the frequency of average F-ratio values calculated with a minimum of three m/z for the null comparison in the 6.2 ppm spike versus blank diesel). Table 5 begins to summarize the capability of identifying all spiked analytes in a comparison to blank diesel, by showing the third highest F-ratio value for each spiked compound in each comparison (at the optimum tile and grid location). At low signal-to-noise, 3-octanone is not capable of being identified with the tile F-ratio approach at or below a concentration of 3.7 ppm, while 5-decyne is not identified at or below 2.11 ppm.

In detecting small changes in concentration between sample classes the contrast with which the tile F-ratio approach can find changes in spike concentration relative to the magnitude of false positives (modeled by the distribution of null F-ratio values and null average F-ratio values) becomes difficult below the comparison of a nominal 12.5 ppm spiked diesel versus blank diesel. The F-ratio value seen for the third m/z for a spiked compound has a significant probability that it could have resulted from a random null event at or below the nominal spike concentration of 12.5 ppm, especially for 5-decyne and 3-octanone. For instance, the third highest F-ratio value for 3-octanone in the 12.5 ppm comparison has a 0.0008% probability of

occurring randomly, meaning the fraction of null F-ratio events seen at or above this F-ratio value in the null distribution contains 0.0008% of total observed null F-ratio events, Fig. 6A). With approximately 2.5-million tile locations (four grid schemes and three-hundred mass channels) that means that this false positive should occur twenty times (2.5 million multiplied by 0.0008%).

The effect of applying an F-ratio threshold such that any chromatographic location with an average F-ratio value below the threshold is removed is shown in Fig. 6. For this spiked diesel validation study, our goal is to understand what F-ratio threshold value removes the most false positives, while retaining the detection of the four spiked compounds. For the nominal 6.2 ppm comparison versus blank diesel, any F-ratio threshold value above zero removes some false positives and the higher the threshold the more false positives will be removed from the final table of class distinguishing locations (final hit list). However, from Table 5 we know that any F-ratio threshold value above three for the comparison of 6.2 ppm versus blank diesel will result in not identifying 3-octanone as a class distinguishing compound. There is a trade-off between removing false positives and retaining detection and identification of class distinguishing compounds. The 12.5 nominal spike level versus blank diesel is summarized in Tables 6, 7, and 8, showing the final software output with three different F-ratio thresholds, 20, 40, and 50 respectively. At 12.5 ppm, high F-ratio thresholds below 73 do not result in loss of identifying all four spiked compounds. The false positive rate for an F-ratio threshold of 20 is 3%, calculated by dividing the number of false positives seen in the final output by the total number of possible false positives calculated with the null class comparisons. A summary of the total number of possible false positives as calculated by the null class comparisons is shown in Table 7, indicating that the total number of false possible false positives is independent of the different

spike levels in this experiment. Additionally, for the 12.5 nominal spike comparison, the F-ratio threshold would have to be above 73 in order to not identify 3-octanone as a true class distinguishing compound. Table 6, 7, and 8 show that the final software output can be optimized to remove all false positives while retaining all true positives (spike compounds). This drastically simplifies the final analysis of class distinguishing chromatographic locations.

For the routine analysis of complex chromatographic data with the tile F-ratio software, null distribuition analysis allows analysts to pick the optimum contrast needed to simplify the final analysis and statistically validate that threshold experimentally. The contrast of each spiked compound's average F-ratio value versus the maximum null average F-ratio for every comparison in this study is summarized in Fig. 8. Fig. 8 shows that for many absolute mole quantities injected on column, there is significant contrast compared to the maximum average null F-ratio value. For instance, bromobenzene, at 140 pmol on column, has an average F-ratio value ~5 times that of the highest null average F-ratio value, while chlorohexane, at 80 pmol on column, has an average F-ratio value of ~2 times that of the highest null average F-ratio value. In order to identify 3-octanone and 5-decyne in this study at extremely low concentrations, less than 12.5 ppm or 550 pmol on column, requires the acceptance and thus manual (or other algorithmic) approaches to processing of false positives before these compounds are discovered to be class distinguishing.

Using the Fisher ratio statistic to find class distinguishing chromatographic locations and mass channels on tiled GC x GC - TOFMS data works by maximizing selectivity, not necessarily sensitivity. The deconvolution and quantification of chlorohexane (Fig. 9) and 5-decyne (Fig. 10) with PARAFAC using the m/z found at the most optimum tile location for the spiked compound at different nominal concentrations (found with the tile-based F-ratio approach) demonstrates the

selectivity advantage of this algorithmic approach. Especially for 5-decyne, the 3 m/z's found in the 6.2 ppm nominal comparison versus blank diesel has lower sensitivity (because of the summation of few signal m/z's) but is able to deconvolute, identify, and quantify 5-decyne to lower concentration because those m/z are the most selective and do not make deconvolution by PARAFAC impossible for the spiked compound in the small data cube found by tile-based F-ratio software approach.

4.4 Conclusion

GC x GC – TOFMS instrumentation provides highly dense chemically complex data that requires the use of sophisticated data mining techniques to discover only those chemical compounds that change due to experimental conditions amongst a background of other chemical signal and noise. The use of the supervised nontarget analysis technique of F-ratios along with a novel tile-based algorithmic approach allows for detection of small concentration changes between sample classes. With the addition of the novel refocusing techniques of pinning and clustering, redundant hits originally caused by the tile-based approach are removed and the original high resolution chemical data is preserved. Additionally the addition of an F-ratio threshold provides analysts with a tool to reduce the number of false-positives seen in nontarget discovery analysis with F-ratio, thus simplifying the final deconvolution, identification, and quantification of class distinguishing chemical compounds.

References

[1] Pierce, K.M., Kehimkar, B., Marney, L.C., Hoggard, J.C., Synovec, R.E. J Chromatogr. A 1255 (2012) 3–11.

[2] Fraga, C.G., Prazen, B.J., Synovec, R.E. Anal. Chem. 73 (2001) 5833–5840.

[3] Fraga, C. G., Prazen, B. J., Synovec, R. E. Anal. Chem. 72 (2000) 4154-4162.

[4] Pierce, K. M., Wood, L. F., Wright, B. W., Synovec, R. E. Anal. Chem. 77 (2005) 7735-7743.

[5] Skov, T., Hoggard, J. C., Bro, R., Synovec, R. E. J. Chromatogr. A. 1216 (2009) 4020-4029.

[6] Hoggard, J.C., Synovec, R.E. Anal. Chem. 80 (2008) 6677–6688.

[7] Wang, B., Fang, A., Heim, J., Bogdanov, B., Pugh, S. Libardoni, M., Zhang, X. Anal. Chem. 82 (2010) 5069–5081

[8] Kim, S., Koo, I., Fang, A., Zhang, X. BMC Bioinf. 12 (2011) 235–246.

[9] Oh, C., Huang, X., Regnier, F., Buck, C., Zhang, X. J. Chromatogr. A 1179 (2008) 205–215.

[10] Kempa, S., Hummel, J., Schwemmer, T., Pietzke, M., Strehmel, N., Wienkoop, S., Kopka, J., Weckwerth, W. J. Basic Microbiol. 49 (2009) 82–91.

[11] K. J. Johnson, R. E. Synovec, J. Chemom. Intell. Lab. Syst. 60 (2002) 225-237.

[12] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756.

[13] E.M. Humston, K.M. Dombek, B.P. Tu, E.T. Young, R.E. Synovec, Anal. Bioanal. Chem. 401 (2011) 2387–2402.

[14] Beckstrom, A.C., Humston, E.M., Snyder, L.R., Synovec, R.E., Juul, S.E. J. Chromatogr. A 1218 (2011) 1899–1906.

[15] Pasikanti, K.K., Norasmara, J., Cai, S., Mahendran, R., Esuvaranathan, K., Ho, P.C., Chan, E. Anal. Bioanal. Chem. 398 (2010) 1285–1293.

[16] Li, X., Lu, X., Tian, J., Gao, P., Kong, H., Xu, G. Anal. Chem. 81 (2009) 4468–4475.

[17] Ma, C., Wang, H., Lu, X., Wang, H., Xu, G., Liu, B. Metabolomics 5 (2009) 497–506.

[18] Qiu, Y., Lu, X., Pang, T., Ma, C., Li, X., Xu, G. J. Sep. Sci. 31 (2008) 3451–3457.

[19] Efron, B. JASA 100 (2005) 1–5.

[20] Efron, B. JASA 102 (2007) 93–103.

[21] Efron, B. JASA 105 (2010) 1042–1055.

[22] Efron, B., Tibshirani, R. Genet. Epidemiol. 23 (2002) 70–86.

[23] Leek, J.T., Storey, J.D. Stat. Appl. Genet. Molec. 10 (2011) 1–22.

[24] Storey, J.D. J. R. Stat. Soc. B (Statistical Methodology) 64 (2002) 479–498.

[25] Storey, J.D. J. R. Stat. Soc. B (Statistical Methodology) 69 (2007) 347–368.

[26] Woo, S., Leek, J.T., Storey, J.D. Bioinformatics 27 (2011) 509–515.

[27] D. L. Massart, Chemometrics: a Textbook, Elsevier Sciences Ltd, New York, 1988.

[28] R. O. Duda, P. E. Hart, Pattern Classifications and Scene Analysis, Wiley, New York, 1973.

Table Captions

Table 1. The actual concentrations are shown in ppm for each nominal spike concentration (first column) for all spiked analytes. The nominal concentrations will be used occasionally in the text for brevity and clarity.

Table 2. The mole quantity injected on column is shown in nanomoles (nmol) for each nominal spike concentration (first column) for all spiked analytes. A 1 μl injection of each diesel sample was made in split mode with a split ratio of 200:1.

Table 3. The top twenty entries in the initial list of results from the tile-based algorithm approach applied to the comparison of the nominal 12.5 ppm spike level versus blank diesel. There are many redundant tiles associated with each spiked analyte that will be removed by pinning and clustering. There are a total of 5859 entries in the list.

Table 4. The top twenty entries in the final list of results after pinning and clustering, of note is the complete removal of all redundant hits and the presence of many false positives. There are a total of 1151 entries in the list, nearly six times smaller than the list in Table 3.

Table 5. Table showing the F-ratio values for the third highest F-ratio per spike compound and per comparison. The concentration entries denote which concentration was compared with the blank diesel (i.e. first column represents 1.6 ppm nominal spike level versus blank diesel). As the signal-to-noise decreases for each analyte the ability to identify these analytes decreases as a function of the decreasing mass spectral sensitivity and thus selectivity. (-a) Three selective m/z's (from library) were not detected with F-ratio software.

Table 6. The final result from the application of the entire tile F-ratio software including pinning, clustering, and an F-ratio threshold of 20 is shown. The false positive rate is 3%.

Table 7. The final result from the application of the entire tile F-ratio software including pinning, clustering, and an F-ratio threshold of 40 is shown. The false positive rate is 0.03%.

Table 8. The final result from the application of the entire tile F-ratio software including pinning, clustering, and an F-ratio threshold of 50 is shown. The false positive rate is 0%.

Table 9. The mean and standard deviation of the total number of null hits (i.e. no F-ratio threshold) from the four different null classes are shown. The nominal spike level is shown indicating which chromatograms were used for producing the null classes. In each row, the same blank diesel injections were used.

Figure Captions

Figure 1. The two-dimensional GC × GC–TOFMS log10 plot of the TIC of a 100 ppm spiked diesel sample and the locations of all four spiked analytes and the internal standard. The column 1 elution order is as follows: 1-chlorohexane; bromobenzene; 3-octanone; 5-decyne; 1-bromoheptane (internal standard).

Figure 2. A schematic for the Tile-based F-ratio Software showing all steps in the data analysis is shown.

Figure 3. A zoom in on 1-chlorohexane with the sum of m/z's 91 and 55. The contour lines for the signal are ± 1 SD (67% of signal intensity, inner ring) and ± 2 SD (95% of signal intensity, outer ring). The SD is the standard deviation of the peak based upon a Gaussian peak shape definition. Tile grid t3 (shaded) optimally captures the 1-chlorohexane peak.

Figure 4. The tile algorithm is shown applied to a single small two-dimensional subsection of the overall separation. The four tile grids are grid 1 (A), grid 2 (B), grid 3 (C), grid 4 (D). One grid of tiles, grid 3, is shown capturing 1-chlorohexane within a single tile. While grid 4 does include chlorohexane well, a small portion of the signal is lost in the adjacent tile. Grid 1 (A) and grid 4 (D) were designed to include 0.1 s from the beginning of a modulation concatenated onto the end of the dimension, so that peaks that are split directly between modulations can be included in a single tile properly.

Figure 5. (A) Per-m/z F-ratio thresholds for the four null comparisons for the 100 ppm versus 50 ppm comparison. (B) Sum of the four F-ratio thresholds, indicating that none of the m/z have consistently high threshold values for the four null comparisons. The x-axis only goes to 280, as many m/z above do not have enough signals above the S/N filter to produce a good histogram (analysis required >40 locations to be included).

Figure 6. (A) The histogram of all null F-ratio events observed for the 6.2 ppm versus blank diesel comparison. This histogram combines all of the m/z and null class comparisons. The y-axis is log10 scaled. The majority of counts are F-ratio events of ~1, though there are occasional single events with values up to 250. (B) The frequency of detecting an average null F-ratio from tile locations with three or more F-ratio m/z for the null class shuffles of a 6.2 ppm versus blank diesel comparison.

Figure 7. (A) The number of false positives (log scale) as a function of F ratio threshold is plotted for the 6.2 ppm nominal comparison versus blank diesel. (B) The number of true positives (i.e. discovered and identified spike compounds with F-ratio software) as a function of F-ratio threshold is shown for all comparisons.

Figure 8. Plotted is the average F-ratio from multiple spike comparisons versus blank diesel without using an F-ratio threshold normalized to the first false positive average F-ratio in the corresponding comparison. (A) The tile F-ratio algorithm is able to discover the minute difference of 80 pmol of 1-chlorohexane and 140 pmol of bromobenzene in a diesel GC x GC – TOFMS injection with few false positives. (B) The more chromatographically interfered spikes of 5-decyne and 3-octanone are plotted the same way. The tile F-ratio algorithm was able to discover the small difference of 0.56 nmol of 5-decyne and 0.56 nmol of 3-octanone injected with diesel with fewer false positives, but much less contrast than the less interfered bromobenzene and chlorohexane. The addition of an F-ratio threshold does not explicitely increase the F-ratio signal contrast, but reduces the number of false positives discovered.

Figure 9. (A) PARAFAC was performed to quantify 1-chlorohexane for each spike level using the 7 highest F-ratio m/z above a F-ratio threshold of 20, for the 12.5 ppm versus blank diesel

comparison and 3 of the highest F-ratio m/z from the 6 ppm versus blank diesel comparison. (B) The 7 F-ratio m/z's and the 3 F-ratio m/z's (plus one noise m/z) is able to identify and quantify 1-chlorohexane down to ~2 ppm. Using the entire library spectra (51 m/z's), identification and quantification can reach ~2 ppm. The slope of the best fit line is 2742 for library m/z, 2566 for 7 m/z, and 1193 for 3 m/z.

Figure 10. (A) PARAFAC was performed to quantify 5-decyne for each spike level using the 20 highest F-ratio m/z above an F-ratio threshold of 20, for the 50 ppm versus blank diesel comparison and 3 of the highest F-ratio m/z from the 6 ppm versus blank diesel comparison. (B) The 20 F-ratio m/z's and the 3 F-ratio m/z's (plus one noise m/z) is able to identify and quantify 5-decyne down to ~4 ppm. Using the entire library spectra (51 m/z's), identification and quantification can only reach ~7 ppm. The slope of the best fit line is 2293 for library m/z, 2123 for 20 m/z, and 944 for 3 m/z.

Nominal Spike Level (ppm)	3-octanone (ppm)	5-decyne (ppm)	bromobenzene (ppm)	1-chlorohexane (ppm)
100	111	121	133	110
50	55.5	60.5	66.6	54.9
25	27.8	30.3	33.4	27.5
12.5	14.3	15.6	17.2	14.2
6.2	6.89	7.51	8.27	6.81
3.2	3.67	4.00	4.40	3.63
1.6	1.93	2.11	2.32	1.91

Table 1

Nominal Spike Level (ppm)	3-octanone (nmol)	5-decyne (nmol)	Bromobenzene (nmol)	1-chlorohexane (nmol)
100	4.33	4.38	4.24	4.56
50	2.16	2.19	2.12	2.28
25	1.08	1.10	1.06	1.14
12.5	0.558	0.564	0.548	0.589
6.2	0.269	0.272	0.263	0.282
3.2	0.143	0.145	0.140	0.150
1.6	0.0753	0.0763	0.0739	0.0792

Table 2

Identity	Average F-ratio	Tile 1D	Tile 2D	Grid #
1-chlorohexane	1138	32	5	4
1-chlorohexane	1061	32	5	2
1-chlorohexane	1055	32	5	1
1-chlorohexane	968	32	5	3
bromobenzene	612	45	1	2
bromobenzene	572	45	1	1
bromobenzene	143	45	1	4
bromobenzene	92	46	1	4
bromobenzene	81	45	1	3
5-decyne	65	67	4	3
bromobenzene	53	46	1	3
5-decyne	34	67	4	2
3-octanone	26	56	5	4
bromobenzene	24	45	5	1
false positive	21	14	1	1
false positive	20	23	2	1
false positive	20	23	2	4
false positive	19	15	1	4
3-octanone	19	55	5	2
false positive	19	6	3	3
…	…	…	…	…

Table 3

Identity	Pin Location 1D	Pin Location 2D	Average F-ratio	Tile 1D	Tile 2D	Grid #	# of m/z
1-chlorohexane	93	255	1138	32	5	4	10
bromobenzene	13	360	612	45	1	2	14
5-decyne	73	534	65.1	67	4	3	34
3-octanone	91	441	25.8	56	5	4	16
false positive	30	109	20.7	14	1	1	11
false positive	50	186	19.9	23	2	1	29
false positive	30	116	19.3	15	1	4	6
false positive	60	47	19.1	6	3	3	4
false positive	70	414	18.0	52	3	4	30
false positive	59	239	17.6	30	3	3	37
false positive	7	69	15.9	9	5	1	3
false positive	40	140	15.3	18	2	3	25
false positive	70	1202	12.5	150	3	1	3
false positive	60	253	12.2	32	3	3	26
false positive	38	128	11.6	16	2	3	5
false positive	74	709	11.6	89	4	3	37
false positive	91	48	10.7	6	5	3	4
false positive	70	399	9.6	50	3	4	39
false positive	70	496	9.5	62	3	4	3
false positive	50	193	9.5	25	2	4	6
...	...	...	...	...	...	...	...

Table 4

Spike Compound	1.6 ppm	3.2 ppm	6.2 ppm	12.5 ppm	25 ppm	50 ppm	100 ppm
bromobenzene	0.5	155	261	996	5018	1179	935
1-chlorohexane	26	145	252	2628	2020	1632	1703
5-decyne	-[a]	35	21	310	238	545	544
3-octanone	-[a]	-[a]	3	73	242	810	832

Table 5

Table 6

Identity	Average F-ratio	Tile 1D	Tile 2D	Grid #	# m/z above Threshold
1-chlorohexane	1623	32	5	4	7
bromobenzene	857	45	1	2	10
5-decyne	159	67	4	3	13
3-octanone	132	55	5	1	3
false positive	94	103	1	2	3
false positive	89	52	3	4	3
false positive	47	186	2	3	4
false positive	44	35	3	1	4
false positive	42	201	2	3	3
false positive	35	149	2	3	4
---	---	---	---	---	---

Table 7

Identity	Average F-ratio	Tile 1D	Tile 2D	Grid #	# m/z above Threshold
1-chlorohexane	1623	32	5	4	7
bromobenzene	857	45	1	2	10
5-decyne	170	67	4	3	12
3-octanone	132	55	5	1	3
false positive	47	23	2	4	3
false positive	46	52	5	3	3
false positive	46	30	3	3	3

Table 8

Identity	Average F-ratio	Tile 1D	Tile 2D	Grid #	# m/z above Threshold
1-chlorohexane	1623	32	5	4	7
bromobenzene	947	45	1	2	9
5-decyne	181	67	4	3	11
3-octanone	132	55	5	1	3

Nominal Spike (vs Blank Diesel)	Total Number Null Hits (mean +/- SD)
100	1151 +/- 26
50	1145 +/- 41
25	1144 +/- 33
12.5	1158 +/- 41
6.2	1185 +/- 10
3.2	1170 +/- 23
1.6	1166 +/- 26

Table 9

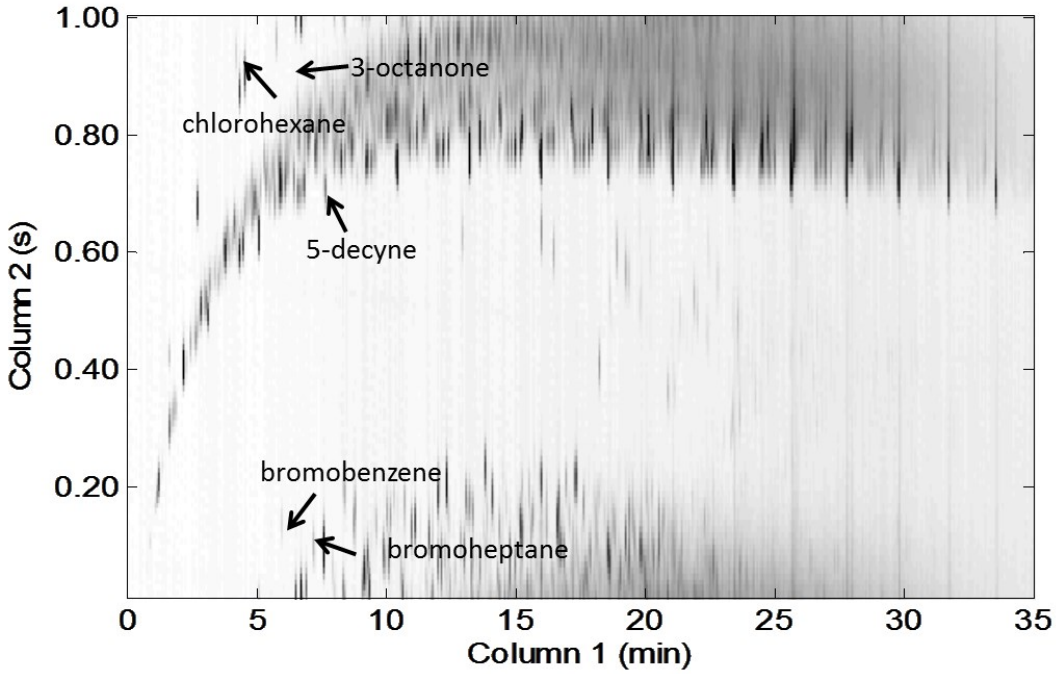
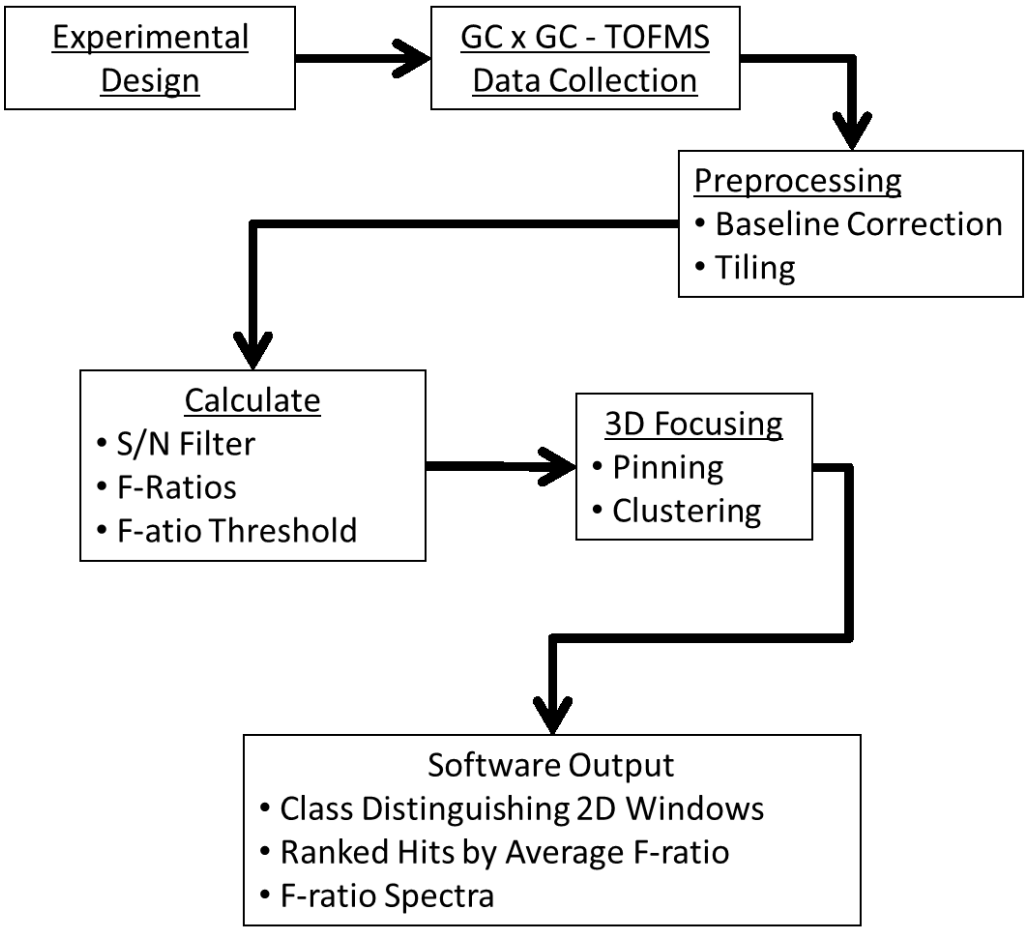

1.00
0.80
0.60
0.40
0.20
Column 2 (s)
3-octanone
chlorohexane
5-decyne
bromobenzene
bromoheptane
0
5
10
15
20
25
30
35
Column 1 (min)

Figure 1

Figure 2


Experimental Design
GC x GC - TOFMS Data Collection
Preprocessing
• Baseline Correction
• Tiling
Calculate
• S/N Filter
• F-Ratios
• F-atio Threshold
3D Focusing
• Pinning
• Clustering
Software Output
• Class Distinguishing 2D Windows
• Ranked Hits by Average F-ratio
• F-ratio Spectra

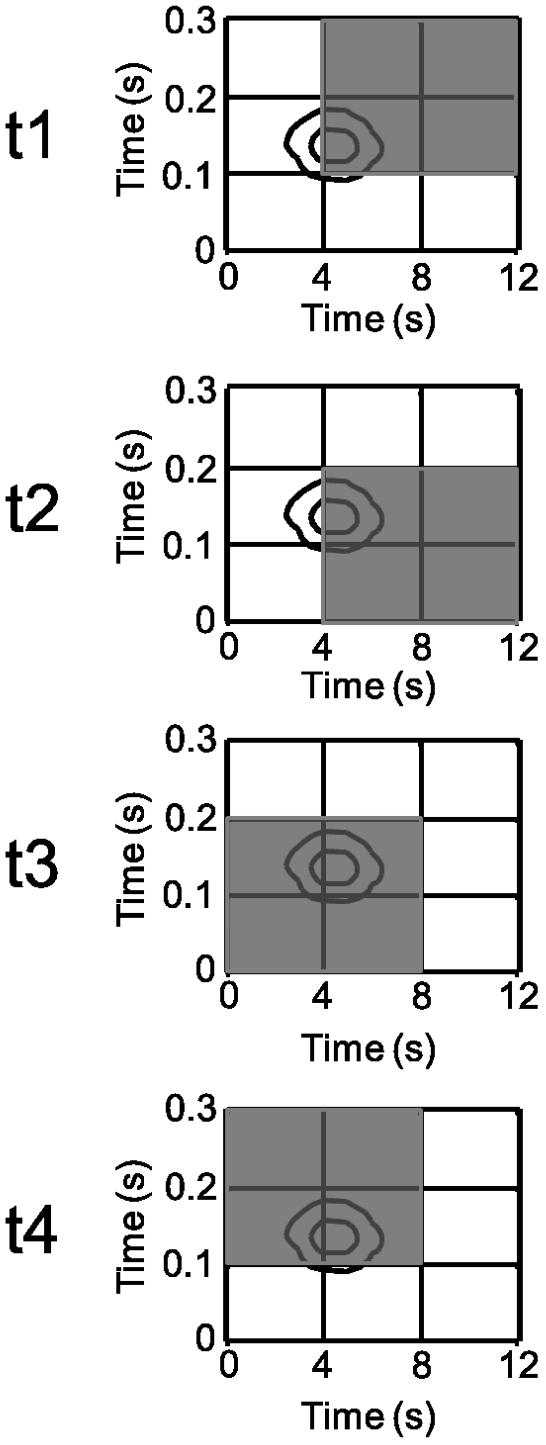


Figure 3

Figure 4

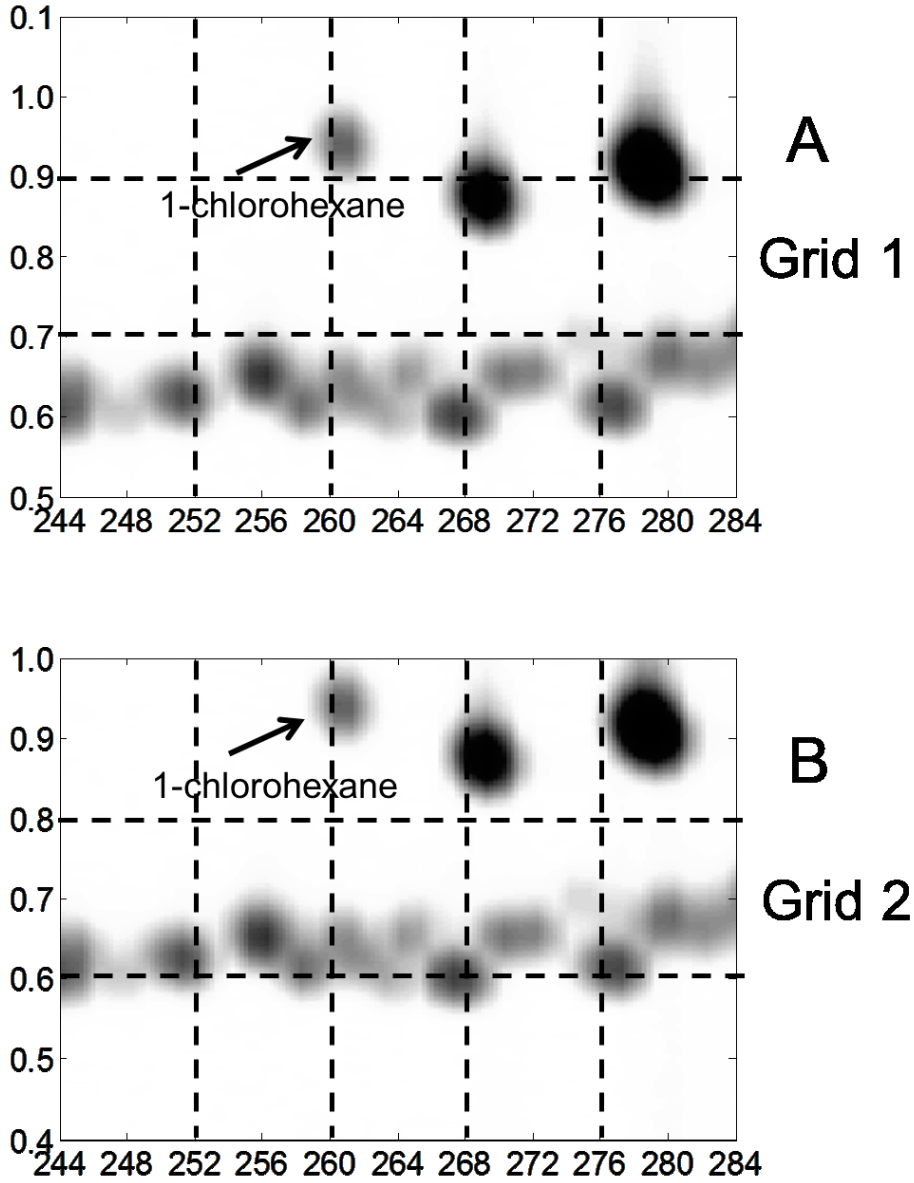

0.1
1.0
0.9
0.8
0.7
0.6
0.5
1-chlorohexane
244 248 252 256 260 264 268 272 276 280 284
A
Grid 1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
1-chlorohexane
244 248 252 256 260 264 268 272 276 280 284
B
Grid 2

Figure 4

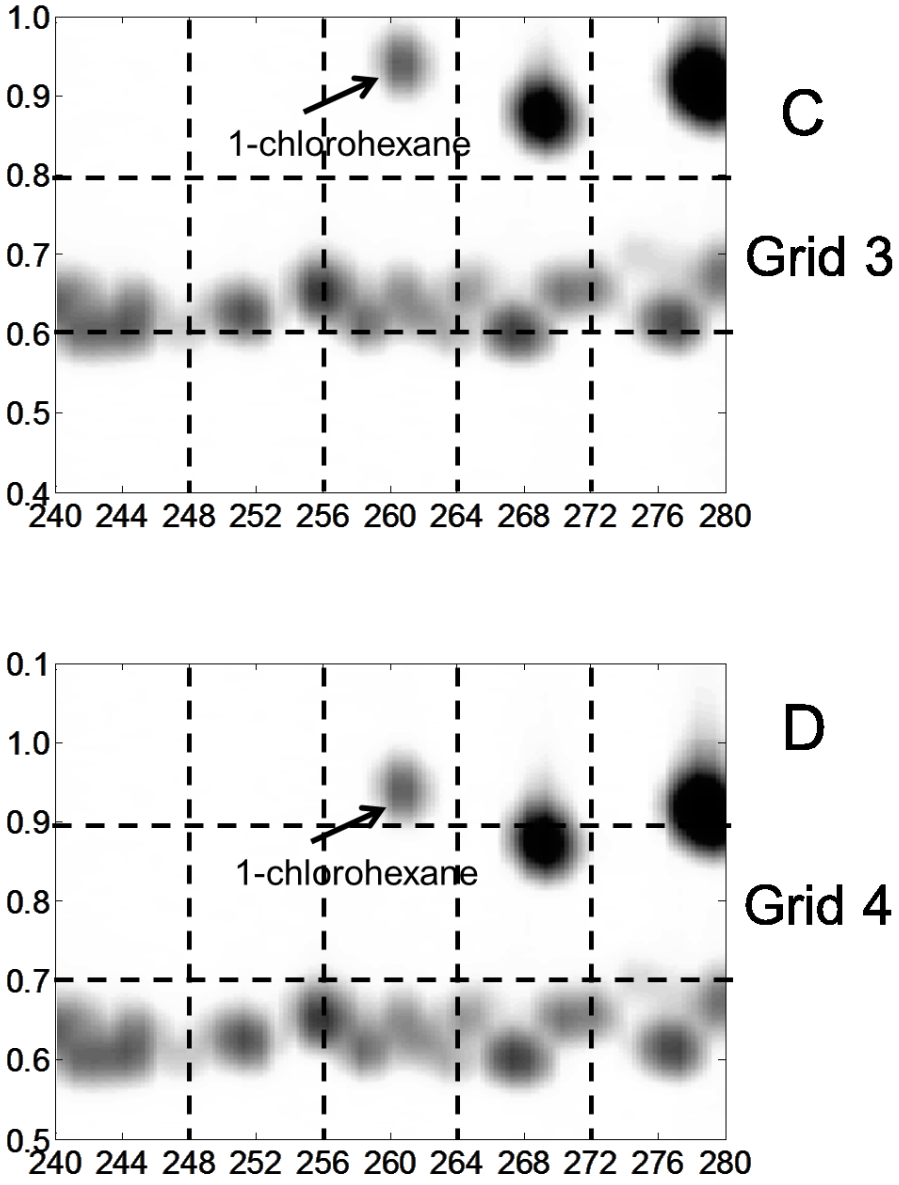

1.0
0.9
0.8
0.7
0.6
0.5
0.4
240 244 248 252 256 260 264 268 272 276 280
1-chlorohexane
C
Grid 3
0.1
1.0
0.9
0.8
0.7
0.6
0.5
240 244 248 252 256 260 264 268 272 276 280
1-chlorohexane
D
Grid 4

Figure 5

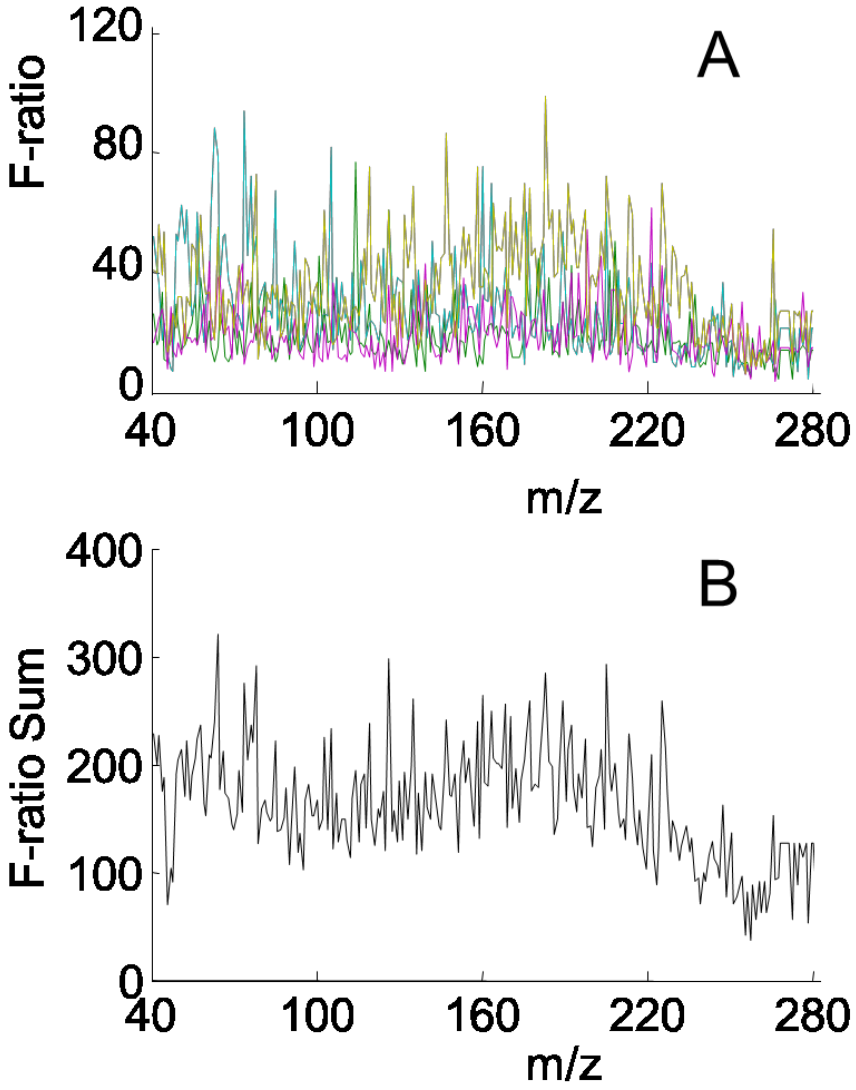
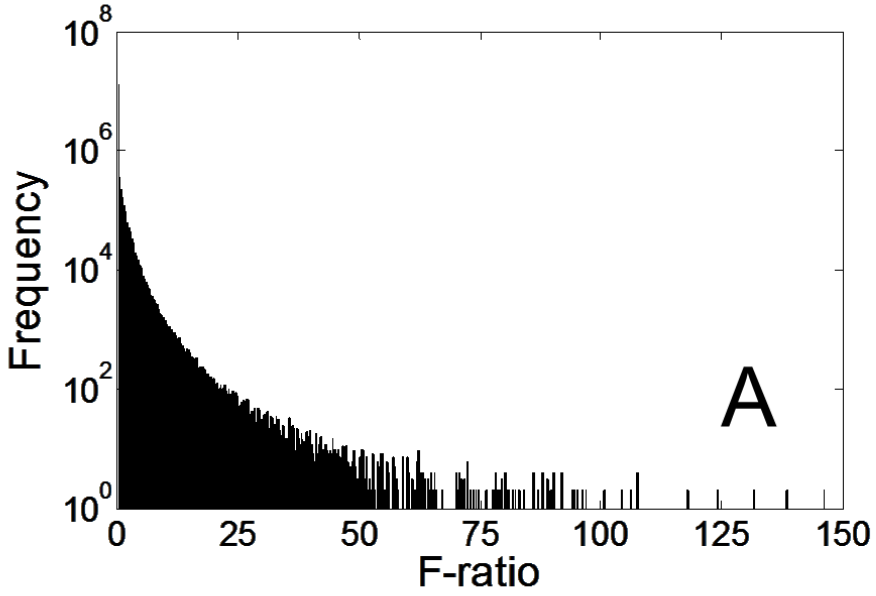

A
F-ratio
120
80
40
0
40
100
160
220
280
m/z
B
F-ratio Sum
400
300
200
100
0
40
100
160
220
280
m/z

Figure 6


Frequency
10^8
10^6
10^4
10^2
10^0
0
25
50
75
100
125
150
F-ratio
A

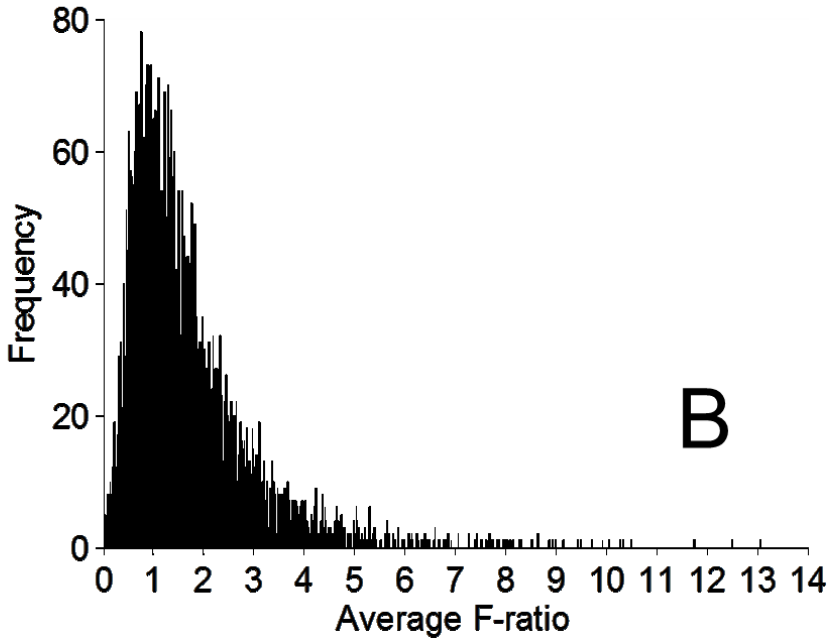

Frequency
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Average F-ratio
B

Figure 7

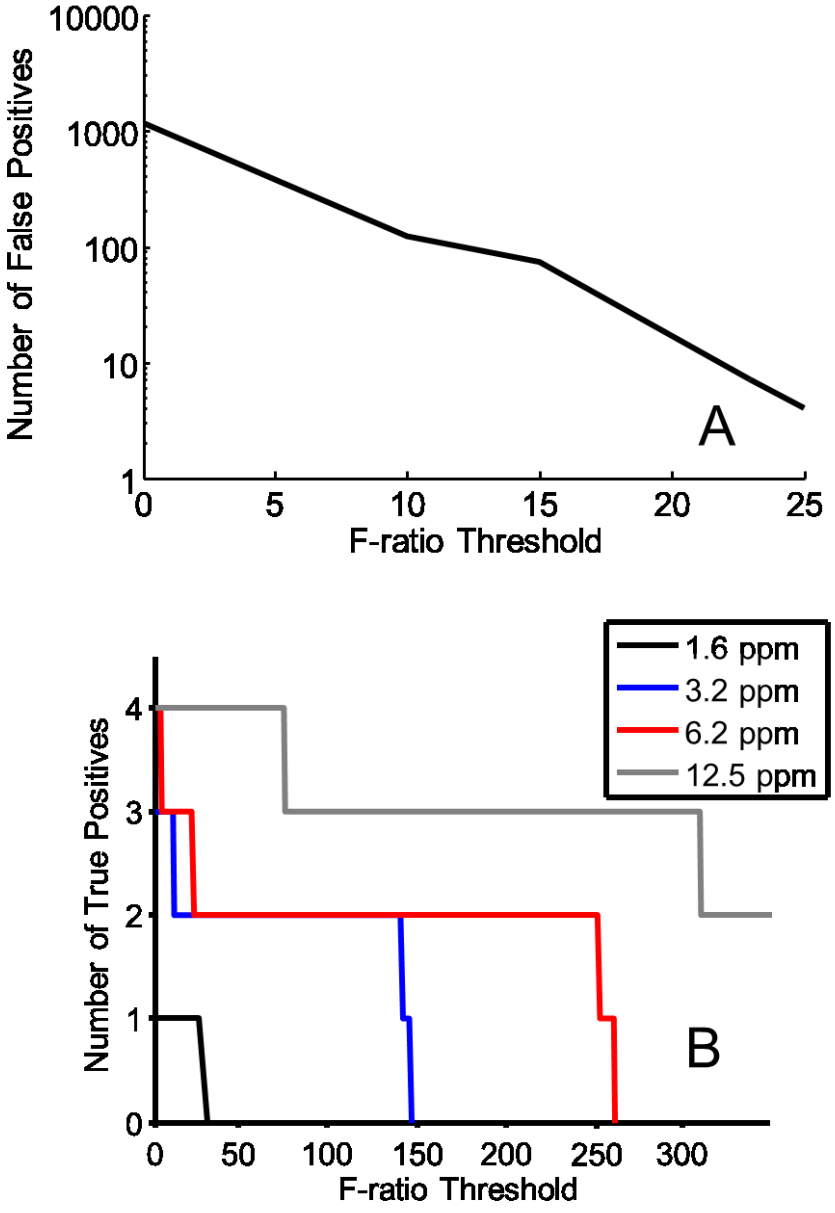

10000
1000
100
10
1
Number of False Positives
0
5
10
15
20
25
F-ratio Threshold
A
1.6 ppm
3.2 ppm
6.2 ppm
12.5 ppm
4
3
2
1
0
Number of True Positives
0
50
100
150
200
250
300
F-ratio Threshold
B

Figure 8

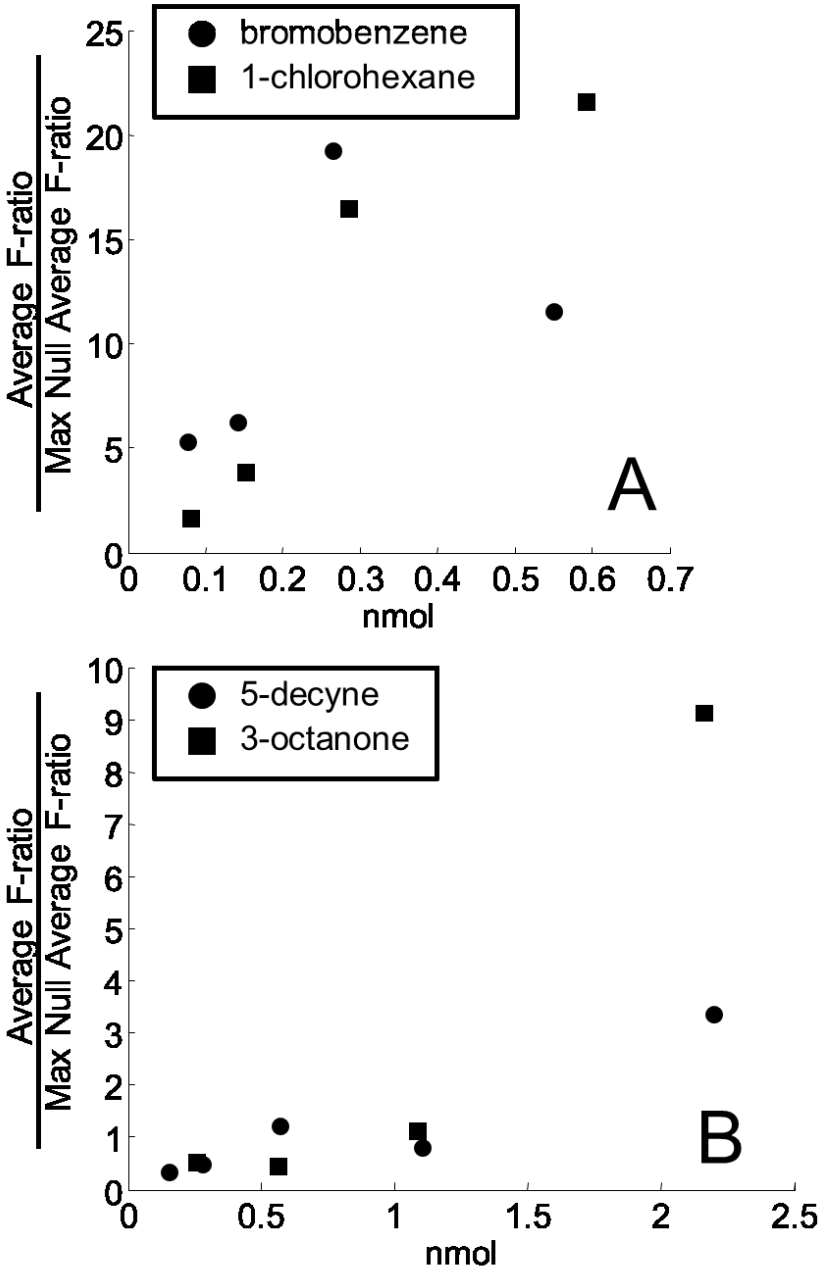

bromobenzene
1-chlorohexane
Average F-ratio
Max Null Average F-ratio
nmol
A
5-decyne
3-octanone
Average F-ratio
Max Null Average F-ratio
nmol
B

Figure 9

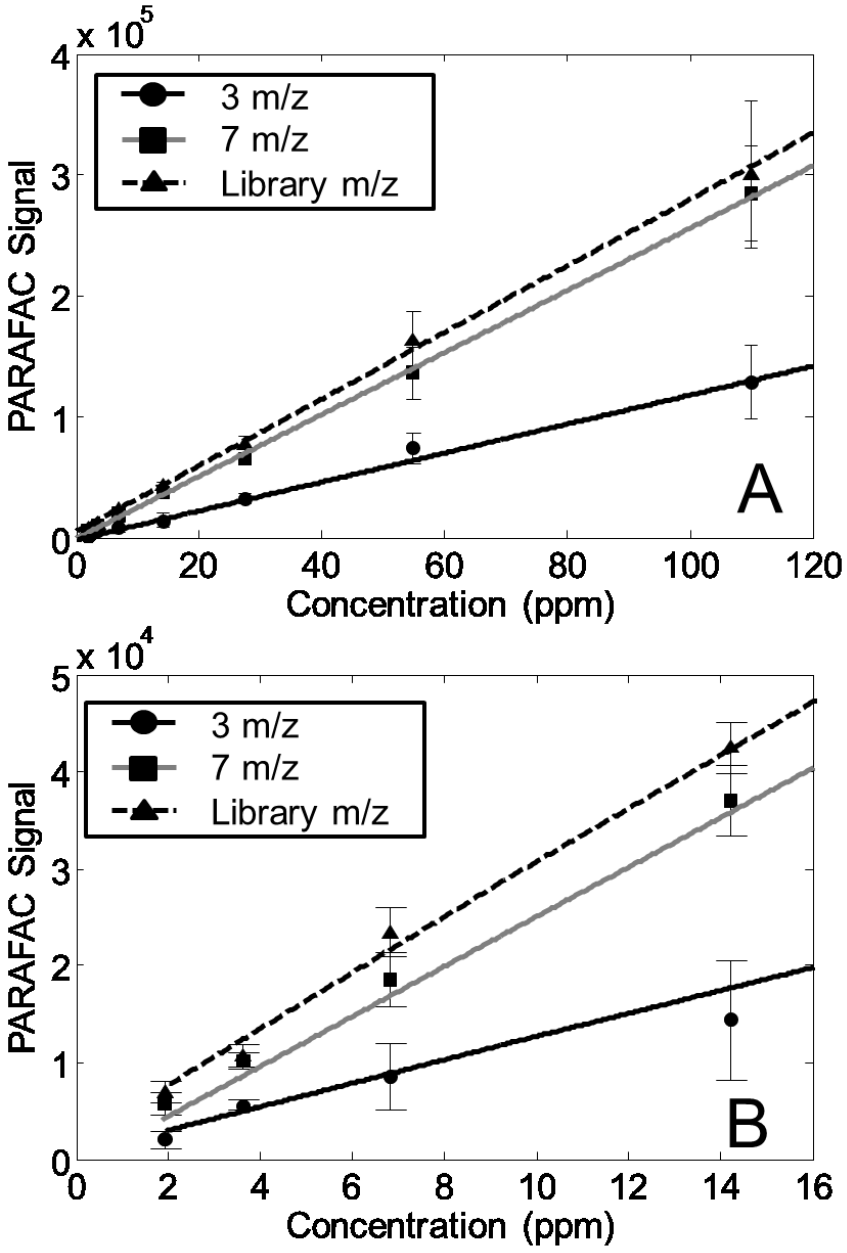
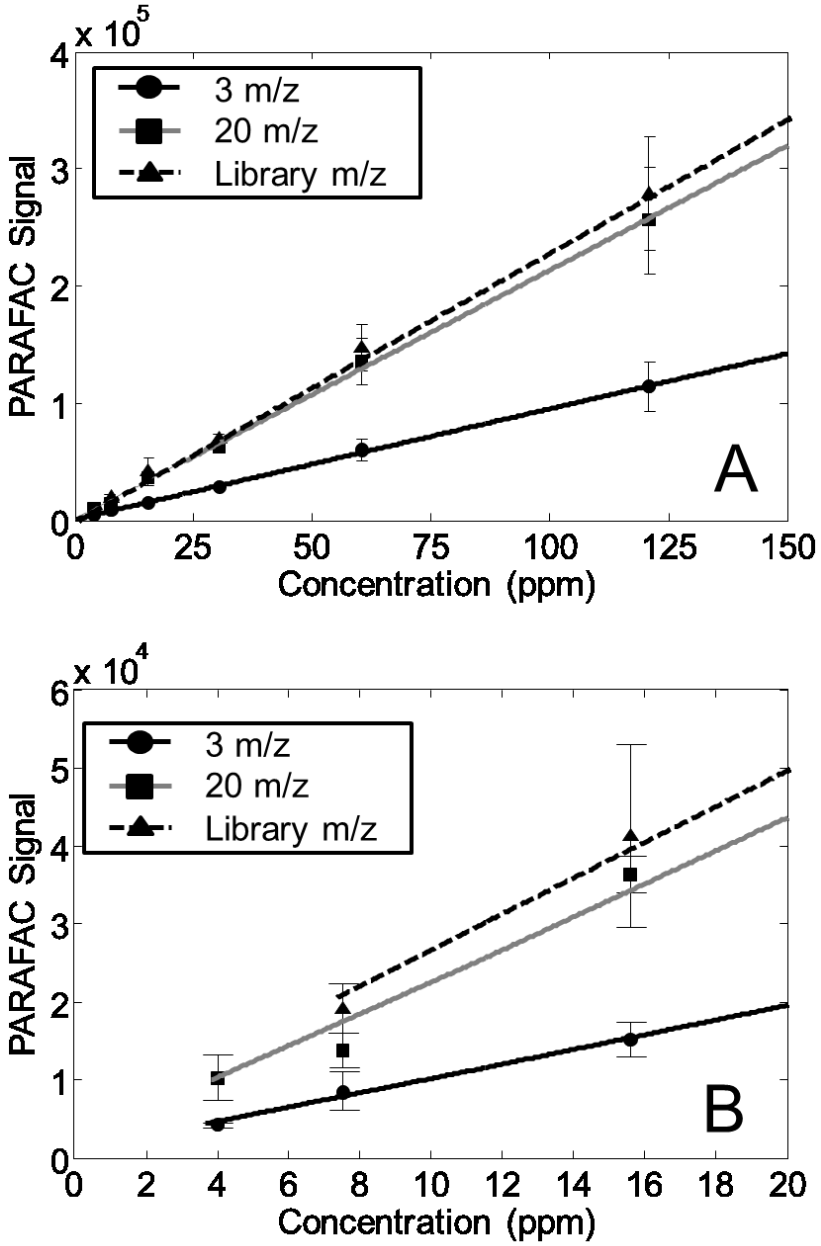

x 10^5
PARAFAC Signal
3 m/z
7 m/z
Library m/z
A
Concentration (ppm)
x 10^4
PARAFAC Signal
3 m/z
7 m/z
Library m/z
B
Concentration (ppm)

Figure 10


x 10^5
PARAFAC Signal
3 m/z
20 m/z
Library m/z
A
Concentration (ppm)
x 10^4
PARAFAC Signal
3 m/z
20 m/z
Library m/z
B
Concentration (ppm)

Ch 5: Conclusion

GC x GC –TOFMS is uniquely suited for comprehensive chemical analysis, such as metabolomics, because it provides extremely high peak capacity and good selectivity for thousands of chemicals at once. However, because it is able to separate and detect thousands of chemicals at once, it has provided researchers with a unique data analysis challenge. The data resulting from the instrument is highly complex and thus nontarget discovery experimentation of biological samples is a prominante area of application of this technology.

Before the full benefits of comprehensive GC x GC –TOFMS analysis are realized for biological samples, chemicals of interest must be extracted and made more volatile via derivatization. The extraction of biological tissues presents a unique challenge, one that requires rigorous attention to potential saturation and homogeneity issues in liquid extraction techniques. In the Chapter 2 I showed that by quantifying a few representative metabolites in heart tissue (targeted approach) I was able to test modification to the extraction technique to optimize the extraction of metabolites for comprehensive nontarget analysis. This method was then used to gain new knowledge about cardiac physiology, specifically related to the metabolic remodeling in pressure overload hypertrophy. However, mining the complex data and making biological interpretations was extremely time consuming and challenging.

There are currently a variety of analysis methods aimed at reducing the initial complex GC x GC – TOFMS data down to those chemicals that are important to a particular experiment. Generally, they reduce the data via a peak table-based approach or a pixel-based approach and use various statistical or machine learning algorithms to rank chemicals in importance to a particular experiment. The use of Fisher ratio with a novel tile-based approach takes advantage of the objectiveness of pixel-based approaches and the simplicity of peak table-based

approaches. In Chapter 2, I demonstrated the successful use of a pixel-based approach to GC x GC –TOFMS metabolomics analysis.

The pixel-based approach to F-ratio analysis of GC x GC – TOFMS data has some limitations, mostly due to the inadvertent misalignment of two-dimensional (2D) chromatographic data (phasing). The tiling of GC x GC – TOFMS data into four grid schemes and subsequent calculation of an F-ratio per tile solves this misalignment problem as well as increases the sensitivity contrast for supervised detection of chemicals that are changing between the classes within a highly complex background of non-changing chemical signal and noise. Prior to the tile-based approach, pixel-level analysis via Fisher ratio suffered from poor sensitivity contrast and high false-positive rates.

With the initial successes seen with the tile-based approach to Fisher ratio data reduction, Chapter 3, the software was refined to remove redundant hits cause by the multiple grid schemes used in the underlying algorithm, described in Chapter 4. By pinning each tile's average F-ratio to the highest signal difference location associated with the high average F-ratio, many redundant F-ratios were consolidated to a single 2D chromatographic location. With the further use of a simple clustering algorithm, the high resolution information can be simplified for the user as to focus on a representative 2D window location that can be analyzed by deconvolution algorithms such as PARAFAC and the relative quantification of specific compounds can be performed.

As 2D locations are found it becomes necessary to evaluate our statistical confidence in the resulting hit list from each experiment. By preparing a null distribution of F-ratios from the mixing of samples between two new balanced null classes, we can experimentally determine what the distribution of F-ratio values are for an experiment were no chromatographic location

should vary between the two classes significantly. This enables the selection of an F-ratio threshold based on actual knowledge of the level of random covariance of signal with sample class. This has provided an even further increase in sensitivity contrast and reduction of false-positives by F-ratio analysis.

The entire software has been written to be routinely used for the analysis of various experimental data sets where the prior knowledge of sample class is known. This software is a substantial contribution to the GC x GC –TOFMS field as well as the field of metabolomics. In roughly 30 min, depending on personal computer processing speed, many gigabites of highly complex metabolomics GC x GC –TOFMS data can be reduced to an ordered list of the most interesting metabolites. However, the identity of which is still dependent on the availability of suitable mass spectral libraries and chemical standards. With this software contribution, the optimization of identifying unknowns will become even more important and pressing.

In the future, analysis of other biological systems can be performed with tile-based F-ratio analysis of GC x GC –TOFMS data. The challenge of such investigations is no longer the complexity of the data. Rather, the complexity is initially reduced via tiling and reorganized so as to fully take advantage of the complex richness of the data. Additionally, a major bottleneck, the time it takes to perform GC x GC – TOFMS data analysis, during the metabolomics investigation of biological samples has been improved drastically, so we are able now to push our experimental design to ask more complicated and interesting questions about fundamental biology and pathophysiology.