Metabolomics by GC×GC-TOFMS:

Overcoming Challenges from Sample Preparation to Data Analysis

Towards Standardization of the Workflow

by

Seo Lin Nam

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

Doctor of Philosophy

Department of Chemistry

University of Alberta

Abstract

Metabolomics is an ‘omics’ field, which involves the comprehensive characterization of metabolites using analytical chemistry technologies and statistical methods to interpret the results. The term metabolomics was coined about two decades ago as an analogy to other precedent omics approaches: genomics, transcriptomics, and proteomics. Although the term is relatively new, studies of health and life science have existed since ancient times, with the understanding that the biological samples contain information that can be linked to the health state of an organism.

Advances in analytical technologies and statistical tools have made metabolomics useful for a broad range of biomedical, agricultural, and other applications. The development of technologies, which include novel sampling techniques, sensitive analytical platforms, and computational methods, enables the profiling of thousands of metabolites. With the cutting-edge techniques, metabolomics has evolved into an essential tool providing insights into the molecular complexity of living systems.

Despite the recent improvement in both analytical and data handling technologies, analyzing the metabolome presents many challenges. For genomics or proteomics, using a single instrument is generally sufficient because genome and proteome chemistry is fundamentally established on combinations of four nucleotide bases and 20 amino acids, respectively. Unlike genome and proteome that use a small number of building blocks, metabolome do not purse fixed structural templates. The chemical diversity of metabolites is enormous and there is no single technology that

enables full coverage of the entire metabolome. Indeed, the vast differences in physicochemical properties and abundance amongst metabolites are the bottom-line challenges to metabolomics studies. Accordingly, another major challenge is to extract useful information and interpret complex metabolomics data produced from such high-performance analytical techniques.

Amongst the many analytical platforms that are used for metabolomics studies, comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC×GC-TOFMS) is a well-suited analytical platform to study complex biological samples due to its excellent separation capacity. Although GC×GC-TOFMS has been developed as a powerful separation instrument, considerable challenges associated with it, mainly due to the difficulty of method optimization and data handling, have hindered its adoption by the metabolomics community.

The work presented in this thesis is devoted to making improvements towards the overall metabolomics workflow using GC×GC-TOFMS, from sample preparation to data analysis. The thesis discusses the challenges that remain in the field of metabolomics and GC×GC-TOFMS and suggests improvements to achieve high analytical performance while simplifying the process. The sample preparation methods for GC(×GC)-based metabolomics have been investigated with the particular focus on extraction and derivatization to enhance the coverage of the metabolome while improving the method reproducibility. A new approach to normalize the natural sample variability of biological samples was also explored. In addition, a data analysis strategy to simplify massive GC×GC-TOFMS metabolomics data using scripting filters that classify peaks into their corresponding chemical classes was developed. It is hoped that

the work described in the thesis will contribute towards the standardization of the GC×GC-TOFMS metabolomics workflow, which would allow more widespread use of GC×GC-TOFMS and enhance cross-comparability of results in the metabolomics community.

Preface

A portion of Chapter 2 is being drafted for publication with the collaborators as "The Effects of Benzoic Acid or Feed Enzymes on Gut Microbiota and Metabolome of Weaned Piglets". I was responsible for the GC×GC-TOFMS method development and data collection for pig cecal samples as well as GC×GC-TOFMS data processing and a portion of data analysis involving chemometrics. I also contributed to the manuscript composition and edits. Harynuk, J.J. was the supervisory author and was involved in concept formation, guidance, manuscript composition and edits.

A version of Chapter 3 has been published as "Nam, S.L.; de la Mata, A.P.; Dias, R.P.; Harynuk, J.J. Towards Standardization of Data Normalization Strategies to Improve Urinary Metabolomics Studies by GC×GC-TOFMS. *Metabolites* **2020**, *10*, 376". I was responsible for conceptualization, designing experiments, data collection, data processing, and analysis. I also drafted the full text of the original manuscript and carried out revisions to satisfy reviewer concerns. Harynuk, J.J. was the supervisory author and was involved in concept formation, guidance, manuscript composition and edits.

A version of Chapter 4 has been drafted and prepared for submission as "Nam, S.L.; de la Mata, A.P.; Harynuk, J.J. Automated Screening and Filtering Scripts for GC×GC-TOFMS Metabolomics Data". I was responsible for sample preparation and data collection of standards, urine, blood, and fecal samples. I also was responsible for developing and implementing scripts, evaluation, data analysis, and manuscript preparation. Harynuk, J.J. was the supervisory author and was involved in concept formation, guidance, manuscript composition and edits.

Acknowledgements

I was incredibly fortunate throughout my doctoral studies to be surrounded by many brilliant, inspiring and, most importantly, warm-hearted individuals whom I loved to work and spend time with. The completion of my work in this thesis could not have been accomplished without the support of many amazing individuals and funding agencies. To all of you, I owe a lot of gratitude.

To my supervisor, Dr. James Harynuk, thank you for hosting me in your group and providing such an enjoyable research environment. Thank you for your valuable guidance, support, travel and learning opportunities, and the mentorship you provided throughout my studies.

To my supervisory committee members, Dr. Sarah Styler and Dr. Wolfgang Jaeger, thank you for your continued interest in my progress, guidance, and encouragement that you provided in every meeting. It was always very helpful and a pleasure talking to you.

To current and past members of the Harynuk Group, thank you for all your valuable support and insightful feedback during the group meetings, presentation practice sessions, various discussion times, etc. Thank you for all the great moments and fun memories we have shared as a group.

To Paulina, no words can express how much I am grateful for you. I could not have come this far without your encouragement and support during various stages of my studies and life. I have learned so much from you, not just about science stuff, but on all aspects of life. Your brightness, caring heart, and impressive ability to “make a friend even with a rock” always inspired me and taught me so much. Thank you for being my friend and taking care of us, your lab children. You are truly a wonderful person.

To Dr. Anna Jordan, thank you for your passion, understanding, and willingness to help the students for whom English is a second language. Your help and guidance in writing the thesis tremendously reduced the amount of stress and made it possible for me to complete writing the thesis. More than anything, I was very touched by how much you care for me and how devoted to helping with such a tender heart. Although they were short, the conversations we shared on the third-floor hallway of Chemistry West were always cheerful and heartfelt.

To all the administrative, technical, and support staff of the Department of Chemistry, thank you for your knowledge, skills, and services that helped make this work possible. Thank you for your availability and kindness, centred on professionalism.

To my family, thank you for your unchanging love and continuous support. Thank you for trusting that I can do this. Thank you for trusting my decisions without asking much.

To Jin, thank you for always being on my side. Thank you for your understanding, patience, and hearing my repetitive venting moments without trying to correct me. Thank you for loving me the way I am.

To everyone who has supported me in your prayers, I am very thankful to all of you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2D	Two-Dimensional
CLIC	Computer Language for Identifying Chemicals
CR-FS	Cluster Resolution Feature Selection
CRM	Certified Reference Material
DoE	Design of Experiments
EIC	Extracted Ion Chromatogram
FAME	Fatty Acid Methyl Ester
FID	Flame Ionization Detector
GC	Gas Chromatography
GC×GC	Comprehensive Two-dimensional Gas Chromatography
GCL	Gluconolactone
GUI	Graphical User Interface
HPLC	High Performance Liquid Chromatography
IL	Ionic Liquid
LC	Liquid Chromatography
LOC	Limit of Classification
LOD	Limit of Detection
MS	Mass Spectrometry
MSTFA	N-Methyl-N-trimethylsilyltrifluoroacetamide
MSTUS	Mass Spectral Total Useful Signal
MTBE	Methyl-*tert*-butyl-ether
NIST	National Institute of Standards and Technology
NMR	Nuclear Magnetic Resonance
PAH	Polycyclic Aromatic Hydrocarbon
PC1	The First Principal Component
PC2	The Second Principal Component
PCA	Principal Component Analysis
PCB	Polychlorinated Biphenyl

PEG	Polyethylene Glycols
PQN	Probabilistic Quotient Normalization
R.T.	Retention Time
RI	Retention Index
RSM	Response Surface Methodology
SPME	Solid-Phase Microextraction
SRM	Standard Reference Material
TIC	Total Ion Chromatogram
TMS	Trimethylsilylation
TOFMS	Time-of-Flight Mass Spectrometer
TPA	Total Peak Area
TUPA	Total Useful Peak Area
UV Spectroscopy	Ultraviolet-visible Spectroscopy

CHAPTER 1

General Introduction

1.1 Motivation

Metabolomics is one of the newest members of the "omics" family and has received increasing attention as an emerging field in the last few decades.[1-3] The term 'metabolite' is defined as any small organic molecule detectable in living organisms with a molecular weight of less than 1500 Da.[4,5] The metabolome denotes the complete set of all the metabolites found in biological samples, and metabolomics is the comprehensive study of the metabolome.[3,6] Metabolic changes are downstream of alterations at the genome, transcriptome, and proteome.[7,8] The changes in metabolites represent the result of complex interactions between our genetic inheritance and multiple environmental stimuli. Hence, in the "omics" world, metabolomics is considered as the endpoint of the "omics" cascade.[9] The composition of the metabolome depends on various factors and can reveal valuable information indicative of the host's state of health; such changes may be detected in biological samples, such as urine, plasma, feces, etc. before the appearance of the first sign of disease.[6] Therefore, metabolomics holds great promise to provide a deeper understanding of phenotypic changes as an organism's answer to disease affected by genetic or environmental factors.[10]

Metabolomics studies already have become a vital part of health sciences. There are numerous metabolomics applications used in areas such as nutritional studies and clinical drug safety assessment.[5,11] Many of these involve the identification of biomarkers associated with a disease with the goal of developing a tool for diagnosis, prognosis, or monitoring disease severity.[12] Several studies have reported that the metabolites found in blood, urine, and feces can be useful in disease diagnosis.[13] For example, higher concentrations of saturated fatty acids, amino acids, and ursodeoxycholic acid were detected in fecal samples of colorectal cancer patients.[14,15] The pharmaceutical field also increasingly utilizes metabolomics platforms to

understand and describe the activity of their drugs in biological systems.[16] In recent years, there has been a vast and continuous increase in the application of metabolomics using various biological matrices, including blood, urine, feces, saliva, sweat, breath, and tissue.[4,14,15,17-25] Additionally, metabolomics is not limited to human samples but also extends to include studies of plants, bacteria, yeast, and food.[26-30]

1.2 Conducting Metabolomics Studies

Metabolomics studies have been carried out using various analytical platforms to manage the complex information in metabolome space.[4,31,32] Due to the vast chemical diversity present in metabolomics samples, there is no singular analytical instrument that enables the analysis of "all" metabolites in a sample.[33,34] Complementary technologies are required to capture a "full" picture of the metabolome. At present, nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) hyphenated to chromatographic separation techniques, such as liquid chromatography (LC; LC-MS) or gas chromatography (GC; GC-MS) are the most common analytical platforms supporting metabolomics studies.[4,35,36] Different platforms present their distinct advantages and disadvantages regarding the convenience of sample preparation, detection limits, throughput, coverage, sensitivity, etc.[36] All analytical techniques have improved continuously; however, an MS-based platform provides a powerful combination of sensitivity and selectivity for metabolomics studies.[36-38] The current state-of-the-art MS-based technologies in metabolomics allow detection of thousands of metabolites in a single run.[4,5,10,22,39] Along with the technical advances of analytical instruments and the high amount of data generated, metabolomics studies need unbiased, accurate and reproducible data analysis which is preferably automated.[4,37]

1.2.1 Gas Chromatography (GC)

The goal of chromatography is to separate sample mixtures into their individual components. In GC, a sample is volatilized in a hot injector that is typically 250-300 °C. The inert carrier gas (mobile phase), such as He, H_2, or N_2, transports gas-phase analytes along a narrow, open-tubed column made of fused-silica that has been coated with a stationary phase. This stationary phase is typically a thin layer of highly viscous liquid-

like polymer. Chromatographic separation into individual chemical components is achieved through analytes partitioning between the mobile and stationary phases.

Resolution is defined as a measure of the separation between two adjacent peaks in a chromatogram and can be expressed in an equation to relate the resolution of two compounds (R_s) to the number of theoretical plates (N), retention factor (k), and selectivity factor (α).[40,41] In a resolution equation, selectivity (the separation factor, α) plays a vital role, as it puts the most significant influence on resolution more than any other parameter. The separation factor is defined as the ratio of the retention factors (k) and serves as an indicator of the separation of two peaks.[41]

$$Rs = \frac{\sqrt{N}}{4} \times \frac{\alpha - 1}{\alpha} \times \frac{k}{k+1}$$

In GC, selectivity is governed by the interaction mechanism between the molecules and the base structure (and its pendent groups) of the stationary phase.[41,42] Many selections of GC stationary phases are available commercially and it is important to choose a column chemistry that suits the needs of one's anlaysis.[43-45] The two most widely used base structures of the GC stationary phase are polysiloxane polymers and polyethylene glycols, also known as wax or PEG. One of the major drawbacks of these widely used columns is that the active hydroxyl (-OH) groups at the polymer termini make these phases susceptible to a back-biting reaction if the column is exposed to oxygen or moisture.[41,46] It contributes to column bleed and results in faster phase degradation. In addition, the classical GC polar stationary phases, such as PEG and poly(cyanopropyl)siloxane, are significantly less thermally stable than the non-polar columns.[46] The maximum temperature for these polar phases is limited to 250-280 °C.[43] Besides the thermal degradation of siloxane polymers, another limitation is the ability to modify selectivity. For conventional polysiloxane polymer-based columns, altering selectivity is constrained by the capacity of chemical modifications of the pendent groups.[46]

1.2.2 GC-MS Based Metabolomics

Amongst many different analytical instruments, GC is an analytical platform that is suited for the analysis of volatile compounds.[37,47] In a system where GC is coupled to a mass spectrometer (GC-MS), analytes are separated first on a GC column and then subjected to mass spectrometry.[37,48] GC is the oldest and most robust among the separations techniques coupled to MS.[11] The main advantages of GC for metabolomics are its high chromatographic separation power and high peak capacity. Moreover, the GC-MS technique is reproducible with regard to retention times and mass spectra, which helps compound identification to be relatively convenient through existing commercial spectral libraries (e.g., NIST, Wiley) supplemented by retention indices.[47,49,50] Therefore, the peak assignment via comparison with spectral libraries can be more accurate and reliable for GC-MS than for other MS-based techniques, which can be troubled by ion suppression.[11] Due to its excellent performance, GC-MS has been a routine method in metabolomics studies for analyzing biosamples, especially for small analytes that do not retain well on LC (e.g., short-chain fatty acids, alcohols, and terpenes).

The biggest limitation of GC-based studies is the fact that the analysis is applicable for only the subset of the metabolome that can be volatilized from the sample matrix.[11] Volatile metabolites can be sampled and analyzed without requiring further treatment. However, many metabolites contain polar functional groups and are not sufficiently volatile at the temperature that the GC system can manage.[51,52] To obtain a broader metabolome coverage for GC-based metabolomics, chemical modification of polar functional groups, like –OH, –NH_2, –COOH, –SH, is required to increase volatility.[53,54] Through chemical derivatization, the derivatives generally become less polar, more volatile, and thermally more stable.[55,56] Routinely, this is achieved with a two-step derivatization procedure of methoximation followed by trimethylsilylation.[57] There are other derivatization methods available, such as acylation or esterification, yet these methods have been tailored for particular metabolite classes.[54] Trimethylsylation has been the gold standard suited for untargeted approaches of metabolomics studies because it modifies diverse classes of compounds.[48]

1.2.3 GC×GC-MS as a Platform for Metabolomics

Conventional gas chromatography (GC) uses capillary columns, which offer high peak capacities.[37] However, GC is still unable to separate all compounds in complex samples like those encountered in metabolomics.[22] Inevitably, GC chromatograms of biosamples (which can contain thousands of compounds) will exhibit co-eluting peaks.[15,58] Because compounds in biosamples are present in a wide range of concentrations, it is especially more challenging when a compound of interest exists in a trace concentration and co-elutes with other components in a sample. This adds to complexities in compound identification and quantification. To overcome the limitations of conventional GC-MS, many strategies were attempted for enhanced peak capacity and improved resolution, including an extension of the column length, which turned out to be impractical due to lengthy analysis time.[59]

Multidimensional chromatography is one of the powerful advances in chromatography introduced several decades ago, in which two or more independent separative steps are combined to obtain better separation of complex mixtures.[60] While conventional one-dimensional GC remains popular in various applications, the field of GC has made advancement to heart cutting (GC-GC), leading to the introduction of GC×GC. In 1991, Phillips introduced the first GC×GC experiment result, which was conducted primarily based on the concepts described by Giddings about the coupling of two analytical columns of differing selectivity in series.[61-63] In order to be considered as a "comprehensive" technique, the entire first-column eluate needs to be cut into small sequential fractions and injected incrementally onto the second column for further separation while maintaining the integrity of the first-dimension separation. The process of comprehensive separation over the entire first dimension separation is achieved through the use of a modulator, the interface connecting the two GC columns.[42,64] A modulator is responsible for continuous sampling of the effluent periodically from the first column then releasing the trapped fraction as a narrow band onto the second column for additional separation.[42,65] The modulation process is essentially what allows for successful comprehensive separation of GC×GC; hence, the modulator is considered as the "heart" of the GC×GC instrument.[64]

The instrumentation of GC×GC is composed of the injector, the primary (first-dimension) column, the modulator, the secondary (second-dimension) column and the detector. Compared to the conventional GC system, the injector and columns do not particularly differ for GC×GC; however, a high-speed detection system that is compatible with the narrow secondary peak widths for GC×GC is critical to preserve the achieved chromatographic resolution. The peak widths at the base of the modulated peaks range between 50 to 200 ms and require acquisition rates of at least 100 Hz. Due to this reason, a main requirement of a GC×GC detector is its acquisition speed. Since the advent of GC×GC, the Flame Ionization Detector (FID) has been the most popular detector due to its low cost and simplicity in operation in various fields using GC×GC, such as in petrochemical analysis.[60] However, when structural information is desired, time-of-flight mass spectrometry (TOFMS) is a more effective detector for GC×GC that could achieve high acquisition frequencies. Some currently available TOFMS detectors are capable of performing acquisition rates up to 500 Hz.[66] This technique is particularly advantageous in acquiring full mass spectra at every data point.

Over recent decades, GC×GC-TOFMS has evolved into a robust and powerful separation technique for comprehensive characterization of complex samples, with an order-of-magnitude increase in separation capacity that is capable of detecting thousands of compounds per sample.[4,15,19,22] When compared to conventional one-dimensional GC, three- to ten-fold more peaks are detectable using GC×GC. Additionally, GC×GC techniques provide an ordered structure of the chromatograms, which is useful for the identification of unknown compounds and the structural elucidation of homologs.[42,67] Due to the aforementioned benefits, this instrument is suited ideally for discovery studies where the entire sample is potentially of interest.[4,15,19,22,58] With superior separation power, supported with TOFMS detector and peak deconvolution software, GC×GC-TOFMS has become a potent tool for metabolomics studies, empowering metabolite profiling and biomarker identification.[4,68,69]

1.3 Typical Metabolomics Workflow

The main goal of metabolomics studies is to provide an in-depth understanding of biological questions. A well-defined question is the crucial primary step as it guides establishment of an appropriate study design that is suitable for answering the question. The metabolomics experiment is composed of different stages.[70] The general metabolomics workflow includes sample collection, sample preparation, data acquisition, data processing, and analysis (Figure 1-1).[1,4] When the procedures for analysis consist of several steps, each step potentially affects the metabolites detected in the sample.[31] Inconsistencies in the quantification of metabolites can arise from various sources during sampling, sample storage, extraction, derivatization, analysis, and detection.[48] Some steps in the workflow are common for all the platforms such as storage or extraction of metabolites; however, this thesis puts particular focus on the steps that are involved in metabolomics studies using GC×GC-TOFMS.

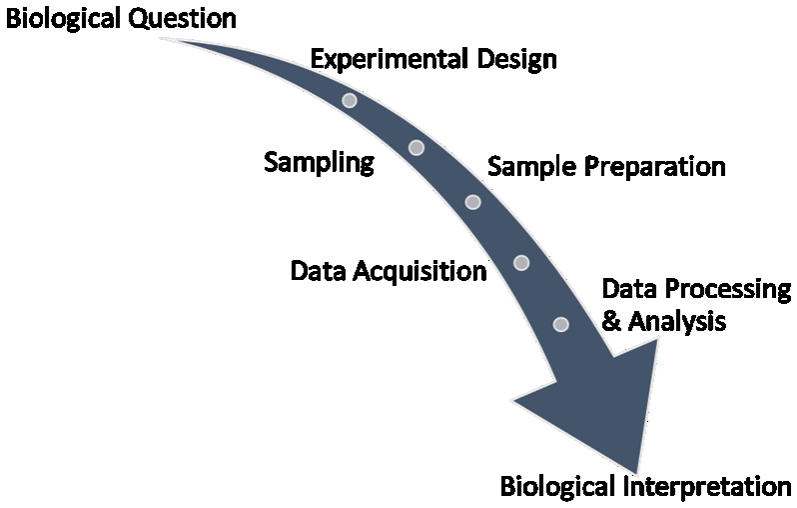

Figure 1-1. Schematic of a typical workflow in metabolomics studies.

One of the major barriers in metabolomics analysis is the lack of standardization in the analytical workflow from sample collection through chemical analysis to data interpretation.[39,71,72] Each step in the workflow matters and can have an impact on the results of the study.[72] Thus, it is essential to establish universal strategies for studying

biological samples that have a good analytical performance, while offering consistency and reproducibility both within a single laboratory to support studies which may last several years, and studies between different laboratories. Raising awareness and putting efforts to standardize metabolomics workflows in order to minimize bias and undesired variations are needed to advance the field of metabolomics using GC×GC-TOFMS.

1.4 Challenges for Metabolomics Studies

As mentioned before, metabolomics studies involve many steps, including pre-analytical work (sample collection, transport, storage), analytical work (sample preparation, data acquisition), and post-analytical work (data processing, interpretation).[11] In order to achieve the correct answer to the question of the study, all these steps need to be planned, conducted, and evaluated carefully. The mistakes made in the earlier stages of metabolomics studies propagate through to later stages; they can lead to significant errors in the outcomes of the study. Each of the variables has the potential to impact the analysis seriously, and variation can occur at any stage of the metabolomics workflow. Failure to correct for systematic variations adequately may result in misleading biological conclusions. While considering all the sources of variations, the metabolomics workflow should minimize the variations in every step. Although much effort has been made to improve metabolomics workflows for analyzing complex biosamples, many challenges still exist. Challenges associated with different stages of metabolomics studies pipeline are discussed in the following sections.

1.4.1 Diverse, Dynamic Samples

Although the number of metabolites in an organism is fewer than the number of genes or proteins in genomics and proteomics, DNA, RNA, and proteins are structured by genetically encoded polymerization of a small number of building blocks.[73] Thus, the task of isolating and characterizing all genetic material or proteins in a biosample is comparatively simple. In contrast, metabolites do not pursue fixed structural templates, and consequently they span a diverse range of physical and chemical properties.[74]

Metabolites are also present over a broad concentration range within the body.[39] For example, in human blood plasma, the normal glucose concentration range is 700–1100 mg/L, while the majority of other metabolites have concentrations ranging from ng/L to mg/L levels.[75] This leads to challenges such as detector saturation caused by abundant metabolites exceeding the dynamic range of the instrumentation[76] and difficulty with accurate identification and quantification of metabolites at low concentrations, possibly complicated by co-elution and peak distortion when these metabolites elute close to highly abundant metabolites.[38]

The human condition is changing continuously with the consumption of food and water, the level of activity, etc. This adds to the variability in the composition and concentrations of metabolites arising from genetic differences. As an example, the sample concentration of urine can vary up to 15-fold, depending on the level of hydration and other various factors.[77,78] For fecal samples, the major component is water (60–80%), but the water content in feces fluctuates depending on fibre intake, causing variations in sample concentrations.[79] Substantial sample-to-sample variation can add difficulty in determining the sample volume or mass to use for the analysis. The concentrations of metabolites vary not only between different individuals, but even from the same host, the concentrations of endogenous metabolites may show significant variations, depending on several factors, including the time of collection. Therefore, when designing the metabolomics studies, factors related to the individual, cohort, and environmental should be considered carefully, and correcting the sample concentration variability with appropriate normalization techniques is needed.

In addition to all these challenges, the metabolic composition of a study subject varies widely, depending on the organism and sample type.[4] Therefore, sample handling and preparation protocols that are specific for a given sample type and instrumental method of analysis have to be developed and utilized. When there is no previous knowledge of sample components, it is possible to overload/underload the instrument.

1.4.2 Sample Preparation

GC-based metabolomics consists of several steps before instrumental analysis, which include sampling, sample storage, extraction, and derivatization. It has been reported that the pre-instrumental analysis process plays a critical role in metabolomics studies. Each step potentially affects the method performance, including metabolite coverage, profile, sensitivity, and reproducibility.[11,31] Numerous studies have revealed that a significant number of errors arise from the sample preparation steps that influence both the observed metabolite content and quantification.[8,11,74] In addition, the variations that occur in these earlier stages propagate through the entire workflow, critically impacting the measurement accuracy. Therefore, the protocols for each step must be examined, optimized, and validated.[11]

An ideal sample preparation procedure aims to obtain an accurate snapshot of metabolite levels while minimizing the changes to sample chemistry.[11] The choice of sample preparation depends on the purpose of the metabolomics study. In targeted studies, the sample preparation should focus on the metabolites of interest.[80] On the contrary, the ambitious goal of the non-targeted metabolomics approach aims to detect as many metabolites as possible, reliably and reproducibly. Ideally, sample preparation for global metabolomics should be as non-selective as possible to provide broad coverage and be as simple and fast as possible to increase throughput and prevent metabolite loss and/or unwanted conversion of analytes during the preparation procedure.[74]

Variation in sample collection and storage may influence the molecular composition and overall quality of the samples significantly.[11] After sampling, the ongoing metabolic activity should be halted to prevent alterations of the metabolic profile.[11,81] Prompt snap freezing immediately after collection by storing the sample at –80 °C until analysis is typically recommended in order to quench any degradation activity, such as enzymatic reactions and oxidation of labile metabolites.[11,74,82]

To gain access to the information enclosed in biosamples, it is necessary to extract the metabolites from the sample matrix. Extraction is the process whereby analytes are separated selectively from undesired compounds present in the sample.[47]

Traditional liquid–liquid extraction commonly is used to extract metabolites. Efficient extraction is the most crucial part of any metabolomics study, regardless of which analytical platform is employed.[52] An ideal metabolite extraction primarily focuses on the efficient and reproducible release of metabolites from the sample, while removing interferences such as proteins and salts that make the analysis difficult.[81] Moreover, the extraction step can benefit the trace metabolites by concentrating them before analysis.[80]

As mentioned previously, the chemical diversity in metabolomics samples is extremely large. Extraction is complicated by the physicochemical diversity and wide range of metabolite concentrations.[38,70] Metabolites can be polar or nonpolar, with diverse chemical characteristics. From a chemical point of view, solubility plays a major role in extraction; water-soluble versus insoluble compounds need to be extracted in different solvents.[83,84] Aside from the solvent's solubility, the sample-to-solvent ratio, pH, and temperature are important factors that must be considered carefully to maximize the metabolite coverage and experimental reproducibility.[11] Without careful consideration of the extraction parameters, inefficient and irreproducible results may arise from problems such as solvent saturation effects or co-precipitation with proteins.[11] The extraction step should be able to produce a stable extract that quantitatively reflects the metabolites present in a given sample.[37] The effectiveness of extraction directly affects the quality of the final data.[74]

Following extraction, the extracted metabolites typically require chemical modification because a significant portion of the metabolites in biosamples are not naturally volatile and not GC-suitable.[37] With the exception of headspace sampling techniques, in most cases, metabolomics samples cannot be analyzed directly by GC without sample derivatization procedures.[52] Trimethylsilylation (TMS) derivatization, which is the gold standard for GC-based metabolomics, also has to be performed carefully. Most importantly, the TMS derivatization reaction can be done only on samples that have been dried completely because the TMS derivatizing reagents are hygroscopic and the presence of water will impede the derivatization reaction, induce variations, and deteriorate reproducibility.[53,85] To eliminate the water from the samples,

a drying step is needed; this step is performed by the evaporation of the extraction solvent that may cause the loss of some volatiles during the process.[37] In addition to the dryness of the sample prior to applying derivatizing reagents, the success and completion of the derivatization are impacted by several other factors, such as the reagent concentration and volume, reaction time, and temperature.[52] Incomplete derivatization results in multiple different TMS derivatives of a single analyte due to not all active protons of a molecule undergoing chemical modification. This leads to multiple chromatographic peaks for a single compound, which affects quantification and adds another layer of complexity in data analysis.[52] Therefore, chemical derivatization must be performed under optimized conditions.

1.4.3 Challenges in Data Analysis

Due to the extreme complexity of metabolomics samples, it is evident that a powerful separation technique is required to obtain comprehensive data from samples that are rich in metabolites content. As a platform, GC×GC-TOFMS is a nearly ideal tool for studies of metabolomics. However, despite the apparent analytical advantages that comprehensive two-dimensional GC proposes, its adoption by the metabolomics community has been reluctant.[4,22] While the progress on the technical part of GC×GC and applications in other fields, such as environmental and petroleum chemistry, is relatively well-reported, the metabolomics community still hesitates to enter the world of two-dimensional GC.[65,75,86-89] The primary barrier has been the difficulty in data handling. GC×GC-TOFMS generates massive, complex datasets.[90,91] In a single chromatographic peak measured by GC×GC-TOFMS, much information is contained; for each peak, information of two retention times from the primary and secondary columns, along with the full spectra collected from the TOFMS system, are obtained.[68,69] Provided that a typical size of metabolomics samples contains thousands of peaks in a single sample and that metabolomics studies should require a large sample size to add statistical power to the studies, a rapid accumulation of large amounts of data is expected for metabolomics studies with GC×GC-TOFMS.[22,92] For metabolomics studies, it is not uncommon to have hundreds of samples, with each

sample containing thousands of peaks. While the ultimate goal of metabolomics analysis is to turn the data collected into useful chemical and biological information, the enormous amount of data generated by this advanced analytical technique poses significant data analysis and interpretation challenges.[22,75] The richness of the data could be understood fully only when a reliable and automated tool for data processing, analyzing, and visualizing is developed to extract meaningful information from the generated data.

Owing to the complexity of the data generated, the non-targeted approach of metabolomics studies often involves the use of multivariate statistical analysis tools. Chemometrics is the data evaluation tool used to extract information from chemical data using multivariate mathematical and statistical methods.[93,94] The development and advances in chemometrics have brought untargeted metabolomics into a new era.[95,96] Although chemometrics plays a vital role in the analytical workflow for the translation of the raw data into useful information, it is essential to preprocess the GC×GC-TOFMS dataset properly to offset the variations that exist in the dataset before applying statistical analysis techniques. The raw metabolomics data contain unwanted biological and technical variations that confound the biological variations of interest.[97] Adequate data normalization is needed to eliminate the chemical and instrumental noise and also to correct for the experimental and biological biases.[97] For example, the vast sample concentration variability that exists in urine samples must be corrected before applying a data reduction technique in urine metabolomics studies.[77] Failure to normalize the data properly and correct between-sample variation can create bias and result in a false discovery or biomarker.

The real issue in metabolomics studies with data normalization is that although there have been several normalization approaches proposed, there has been no consensus on the ideal normalization strategy.[97] Standardization of the normalization technique is demanded in the metabolomics world as the field realizes the impact of normalization techniques on the outcome of the study.[98] Depending on the normalization strategy (or lack of normalization), the result and conclusion obtained in the subsequent data analysis can vary.[97] As data processing and normalization are

crucial in large-scale metabolomics studies, comparison of different normalization methods and working towards the standardization of data normalization strategies to improve metabolomics studies is necessary.

1.5 Scope of the Dissertation

Comprehensive two-dimensional gas chromatography has been developed as a powerful advanced separation analytical instrument in the past few decades. However, it still is considered a relatively novel technique in the metabolomics community and is still far from being fully established. There has been a continuous demand for more reliable and sensitive analytical techniques for analyzing complex metabolomics samples. Additionally, analytical chemists have ambitiously attempted for faster and simpler yet high-performance methods that require minimum human labour and intervention. In this thesis, efforts were made to improve the analytical workflow for GC×GC-TOFMS metabolomics, with attempts to address the challenges that exist in biosamples. Chapter 2 focuses on advancing sample preparation and analytical methods that are common in GC×GC-TOFMS metabolomics studies to obtain faster and better analytical performance. Chapter 3 focuses on improving normalization techniques for data processing, with a particular focus on urinary metabolomics. Chapter 4 focuses on improving data analysis tools for interpreting GC×GC-TOFMS data analysis using mass spectral scripting filters for biosamples. Chapter 5 discusses other applications using scripts that they are not related to human biosamples, which include group type analysis of cannabis and filtering chlorinated species. The thesis attempts to improve the quality in all of the domains of the entire metabolomics workflow using GC×GC-TOFMS from sample preparation to data analysis. In doing so, it is hoped that improvement in metabolomics workflow was made for more widespread use of GC×GC-TOFMS in the metabolomics community.

Chapter 2

Improving Sample Preparation and Analysis Methods for Biosamples

2.1 General Introduction

Metabolites are the final downstream products of complex interactions of metabolism.[70] In contrast to genomics and proteomics, metabolomics is inherently challenging due to the structural diversity, broad concentration range, and differences in metabolites' stability.[81] Currently, there is no standardized universal strategy for the preparation of biological material for global metabolite profiling.[99] In general, the metabolomics pipeline is composed of multiple stages, which include sample collection, preparation, data acquisition, and data analysis.[28] Of the different stages, the sample preparation step is particularly important because the choice of sample preparation method would impact both the resulted metabolite profile and data quality.[74] The sample preparation procedure requires it to be unbiased, robust, and reproducible in order to keep the undesired variability at a minimum and to measure the true biological variability of interest reliably.

Amongst different analytical platforms used for metabolomics studies, GC-MS is a versatile technique that can be applied to a broad spectrum of metabolites.[28] Its strength is the measurement of volatile compounds and those that are not well-retained on LC. However, to obtain a broader metabolome coverage, the metabolites are required to be chemically modified to increase volatility, thermal stability, and chromatographic performance of semi- and non-volatile analytes to make them more suitable for GC analysis.[81] This typically is achieved by extracting metabolites from the sample matrix, followed by chemical derivatization of the dried extracts (Figure 2-1). This chapter focuses on the development and optimization of sample preparation for GC×GC-TOFMS analysis.

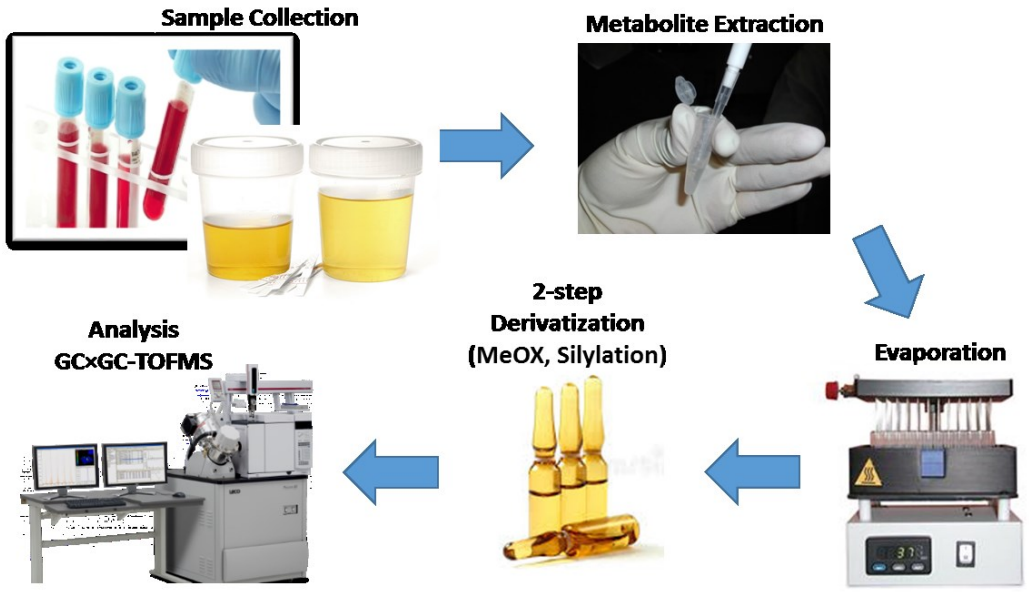

Figure 2-1. General steps to prepare metabolomics samples for GC×GC-TOFMS analysis.

2.2 Extraction Solvent Systems for Metabolomics Studies

2.2.1 Introduction

Metabolomics strives for simultaneous identification and quantification of all metabolites in a biological sample, intending to understand the complex molecular interactions in biological systems.[38] To gain access to the information enclosed in biosamples, a critical step in metabolomics studies, regardless of the analytical platform being employed, is the extraction of the metabolites.[74] Metabolite extraction is commonly done with liquid–liquid extraction using an appropriate choice of a solvent, a solvent mixture, or a series of solvents prior to instrumental analysis.[100] For targeted analysis, the choice of extraction solvent can be customized according to the chemical nature of target compounds.[38] In non-targeted metabolomics studies, the choice of an appropriate solvent system is challenging due to the chemical diversity of the samples, from lipophilic free fatty acids and sterols to hydrophilic sugars and amino acids.[84] The extraction solvent system for global metabolite profiling should be capable of extracting metabolites representative of the sample under study, comprehensively and reproducibly.[74,99]

Until now, it is widely accepted that there is no single extraction solvent that works optimally for all classes.[101] Extraction conditions that favour one metabolite class can eradicate the efficient extraction and stability of other metabolite families.[73] This poses a challenge for accurate quantification of various families of metabolites using the same solvent extraction. In non-targeted analyses, the ultimate goal of extraction is the complete release of metabolites from the sample, aiming to maximize the yield and coverage of metabolites in a reproducible manner. Without acceptable extraction efficiency, even relative quantification would suffer due to variations of metabolite recovery.[4] Sufficient extraction efficiency is required for a reliable estimate of the concentration of metabolites.[73]

Blood plasma contains several thousand metabolites with enormous chemical diversity and widely varying concentrations.[102] In this study, different extraction solvent systems were compared for comprehensive metabolomics profiling of plasma, with the goal of obtaining quantitative yields for as many metabolites as possible. One of the most widely used protocols for GC-based global metabolomics of plasma was published by Dunn et al. and uses methanol as an extraction solvent.[103] The protocol was first tested with the National Institute of Standards and Technology (NIST) Standard Reference Material for human plasma. The performance of different extraction solvent systems was compared with regard to coverage, yields, and reproducibility.

2.2.2 Experimental

2.2.2.1 Plasma

Two plasma samples were used in this study.

1) The National Institute of Standards and Technology (NIST) Standard Reference Material (SRM 1950)–Metabolites in Frozen Human Plasma was used for validating the performance of the protocol for large-scale metabolic profiling of serum and plasma, published in Nature Protocol by Dunn et al., with a particular focus on extraction efficiency for amino acid and fatty acid families.[103] They are the two major classes of primary metabolites, and the standard mixtures were available in the lab. Descriptions

of the method details used in the reported concentrations of analytes in SRM 1950 are provided in its Certificate of Analysis.[104]

2) P9523-5ML, plasma that was prepared from pooled human blood with 4% trisodium citrate as an anticoagulant, purchased from Sigma-Aldrich (St.Louis, MO, USA). When the product was received, the lyophilized plasma was reconstituted with 5 mL of deionized water and was aliquoted into 2 mL Eppendorf tubes. The aliquoted plasma was stored in a –80 °C freezer until the day of analysis. The pooled human plasma was used in the analysis to compare the various extraction solvent systems.

2.2.2.2 Chemicals, Reagents, and Solvents

Chemicals and Solvents

All organic solvents used for plasma metabolite extraction were HPLC or GC grade, purchased from Sigma-Aldrich (St.Louis, MO, USA). Methyl-*tert*-butyl-ether (MTBE), isopropanol (IPA), acetonitrile (ACN), chloroform ($CHCl_3$), methanol (MeOH), and hexane were used as extraction solvents. Deionized water, 18.2 MΩ, was obtained from the Elga PURELAB Flex 3 system (VWR International, Edmonton). HPLC grade pyridine (Millipore-Sigma Canada) was used to dissolve O-methoxylamine hydrochloride (Sigma-Aldrich). HPLC grade toluene (Sigma-Aldrich), dried over anhydrous sodium sulfate (Sigma-Aldrich), was used to ensure the removal of traces of moisture before derivatization.

Derivatization Reagents

A 20 mg/mL solution of methoximation reagent was prepared by dissolving 30 mg of O-methoxylamine hydrochloride in 1.5 mL pyridine. For trimethylsilylation, N-Methyl-N-trimethylsilyltrifluoroacetamide + 1 % chlorotrimethylsilane (MSTFA + 1% TMCS) were purchased from Fisher Scientific, Canada.

Internal and External Standards

The internal standard solution was prepared by dissolving 10 mg of 4-^{13}C methylmalonic acid (Sigma-Aldrich) in 10 mL of methanol and vortex mixing for 1 min. A 10 ppb alkane (C8 – C28) standard mixture was prepared and used as an

external standard solution to check the variations of injection volume and calibrate the retention scale for a retention index calculations.

Standard Addition for Quantification

For the quantification of NIST 1950 plasma, the standard addition method was employed using an amino acid standard mixture (AAS18 10 mL analytical standard, Millipore-Sigma Canada) containing 17 amino acids and a fatty acid standard mixture (GLC-744, Nu-Check, MN, USA) containing 44 fatty acids. Table 2-1 shows the compounds that are included in the mixtures.

Table 2-1. List of Compounds in the Amino Acid and Fatty Acid Standard Mixture

Amino acid	Fatty acid			
L-Alanine	C4:0	Butyric acid	C18:2TT	Linoelaidic acid
L-Arginine	C6:0	Hexanoic acid	C18:2	Linoleic acid
L-Aspartic acid	C8:0	Octanoic acid	C18:3	Gamma linolenic acid
L-Cystine	C10:0	Decanoic acid	C18:3	Linolenic acid
L-Glutamic acid	C12:0	Lauric acid	C20:0	Arachidic acid
Glycine	C13:0	Tridecanoic acid	C20:1	11-Eicosenoic acid
L-Histidine	C14:0	Myristic acid	C21:0	Heneicosanoic acid
L-Isoleucine	C14:1	Myristoleic acid	C20:2	11-14 Eicosadienoic acid
L-Leucine	C14:1T	Myristelaidic acid	C20:3	Homogamma Linolenic acid
L-Lysine	C15:0	Pentadecanoic acid	C20:4	Arachidonic acid
L-Methionine	C15:1	10-Pentadecenoic acid	C20:3	11,14,17 Eicosatrienoic acid
L-Phenylalanine	C16:0	Palmitic acid	C22:0	Behenic acid
L-Proline	C16:1	Palmitoleic acid	C22:1	Erucic acid
L-Serine	C16:1T	Palmitelaidic acid	C20:5	Eicosapentaenoic acid
L-Threonine	C17:0	Heptadecanoic acid	C22:2	Docosadienoic acid
L-Tyrosine	C17:1	10-Heptadecenoic acid	C23:0	Tricosanoic acid
L-Valine	C18:0	Stearic acid	C22:4	Docosatetraenoic acid
	C18:1T	Elaidic acid	C22:5n-6	Docosapentaenoic acid
	C18:1	Oleic acid	C22:5n-3	Docosapentaenoic acid
	C18:1T	Petroelaidic acid	C24:0	Lignoceric acid
	C18:1T	Transvaccenic acid	C24:1	Nervonic acid
	C18:1	Vaccenic acid	C22:6	Docosahexaenoic acid

2.2.2.3 General Sample Preparation Procedure

The following sample preparation steps were conducted for all the experiments, with just a change of the extraction solvent system. An Eppendorf containing a small amount

of plasma was thawed on ice for 1 h on the day of analysis. A 40 µL aliquot of plasma supernatant were pipetted into a new 2 mL Eppendorf tube, and 10 µL of internal standard solution were added. The corresponding amount of extraction solvent indicated in Section 2.2.2.4 was added, and the resulting mixture was vortexed for 3 min, then centrifuged for 15 min at 15800 g and 4 °C. The corresponding amount of supernatant described in Section 2.2.2.4 was aliquoted into a 2 mL GC vial and dried under a gentle stream of nitrogen. Then, 100 µL of toluene dried over anhydrous sodium sulfate was added to the dry residue, mixed gently, and dried again at 80 °C under N_2 for 30 min. Next, 50 µL of methoxyamine HCl were added to each sample and incubated at 80 °C for 30 min. Subsequently, 50 µL MSTFA were added to each sample and incubated at 80 °C for 30 min. The samples were cooled for 5 min and 20 µL alkane standard solution were added. The resulting derivatized sample was transferred into a GC vial with a fused 300 µL insert for GC×GC-TOFMS analysis.

2.2.2.4 Extraction Solvent System Recipes

A) 100% Methanol (Modified Nature Protocol)

The parameters were modified slightly from the protocol reported by Warwick et al;[103] 300 µL of methanol was used as an extraction solvent, and 100 µL of supernatant was withdrawn.

B) Folch

The Folch extraction solvent was prepared by mixing methanol and chloroform in a 1:2 v/v ratio. 300 µL of methanol:chloroform (1:2 v/v) were added to the plasma sample, and 100 µL of the lower (organic) layer were collected.

C) Bligh–Dyer

The Bligh–Dyer extraction solvent was prepared by mixing methanol and chloroform in a 1:1 v/v ratio, 300 µL of the extraction solvent were added to the plasma sample, and 100 µL of the lower (organic) layer were withdrawn.

D) Methanol then Chloroform

To 40 µL of plasma, 300 µL methanol were added, and 100 µL of supernatant were withdrawn. The remaining sample was re-extracted with 300 µL of chloroform, and

100 µL of the lower (organic) phase were withdrawn and combined with the methanol layer previously extracted. Then, the combined extracts were then dried under nitrogen.

E) Methanol then Hexane

The same procedure as the methanol then chloroform solvent series was followed, except for replacing chloroform with hexane. With hexane, the hexane layer is positioned as an upper phase.

F) Acetonitrile/Isopropanol/Water[105]

Acetonitrile, isopropanol, and water were mixed in a ratio of 3:3:2 v/v/v. To 40 µL of plasma, 300 µL of the ternary combination solvent mixture were added, and 100 µL of the supernatant were collected.

G) Matyash[106]

MTBE and methanol were mixed in a ratio of 10:3 v/v. To 40 µL of plasma, 300 µL of the extraction solvent mixture were added, and 100 µL of the upper layer were collected.

H) Modified Matyash

The procedure is the same as the Matyash method described in Section 2.2.2.4G, with a slight modification to the ratio of MTBE and methanol to 5:4 v/v.

I) Methanol Then Chloroform (Solvent System 1 for Section 2.2.3.3)

To 40 µL of plasma, 300 µL of methanol were added and vortexed for 3 min. A 250 µL aliquot of supernatant was withdrawn and placed in a GC-vial. A 600 µL aliquot of chloroform was added to the remaining sample, vortexed for 3 min, and centrifuged at 15,800 g for 5 min. Then, 500 µL of the lower chloroform layer were withdrawn and placed in a separate GC vial. Each methanol and chloroform extract was derivatized and analyzed separately.

J) Methanol, MTBE, Water, Chloroform (Solvent System 2 for Section 2.2.3.3)

To 40 µL of plasma, 300 µL methanol and 1000 µL MTBE were added. The sample was vortexed for 3 min, and 250 µL of water were added. Then, 800 µL of the upper layer were withdrawn for the subsequent drying and derivatization procedure. To the remaining sample, 600 µL of chloroform were added, vortexed for 3 min, and centrifuged at 15,800 g for 5 min. Next, 200 µL of the aqueous (upper) layer were

withdrawn, and 500 µL of the lower (chloroform) layer were taken. Each of the 800 µL of MTBE, 200 µL of aqueous layer, and 500 µL of the chloroform layer were placed in a separate vial and derivatized individually.

2.2.2.5 GC×GC-TOFMS Method

All analyses were conducted using an Agilent 7890A gas chromatograph (Agilent Technologies, Santa Clara, CA) interfaced with Leco Pegasus 4D GC×GC-TOFMS (Leco Instruments, St. Joseph, MI). The columns used for the first and second dimensions were a 60 m × 0.25 mm; 0.25 µm df Rtx-5MS (Chromatographic Specialties), and a 2 m × 0.25 mm, 0.25 µm Rtx-200 (Chromatographic Specialties), respectively. Helium (5.0 grade; Praxair, Edmonton, AB) was used as a carrier gas, with a constant flow at 2 mL/min. The inlet temperature was set at 250 °C, and the injection was done in splitless mode. The GC method began with an initial oven temperature of 80 °C for 4 min, followed by a ramp of 3.5 °C/min to 315 °C, and ending with a 10 min hold in the first oven, which makes a total analysis time of 81.1 min. Relative to the primary oven, the secondary oven was programmed to have a constant offset of +10 ºC and the modulator a constant offset of +15 ºC. The modulation period was 2.5 s (0.60 s hot, 0.65 s cold). The parameters used for mass spectrometry were as follows: electron energy of –70 eV; acquisition rate of 200 Hz; mass range of *m/z* 40–800; detector voltage of –1350 V; ion source temperature of 200 ºC; the MS transfer line temperature was 250 ºC.

2.2.2.6 Data Processing and Analysis

All acquired data were processed using ChromaTOF® (v.4.72; Leco). The baseline offset was set at 0.9, slightly below the top of the noise and above the middle of the noise. The first dimension peak width was set at 15 s, while the second dimension peak width was set at 0.15 s. The peak-finding threshold of S/N was set to 100:1, with the minimum S/N for the sub-peaks to be retained set at 6. A chromatographic region of 0.8 s to 1 s in the second dimension was excluded from data processing due to column bleed. All detected chromatographic peaks were searched against the NIST-MS 2017 Libraries.

The Statistical Compare feature of ChromaTOF® was used to align the analytes across runs. The processed peak tables were aligned based on the parameters of retention time (first- and second-dimension) and mass spectra. Tolerances for retention time shift were set to ±5 modulation period (PM = 2.5 s) in the first dimension, and 0.2 s for the second dimension. For mass spectral matching, the minimum similarity was set at 600 using all *m/z* values with abundances higher than 1%. In the event that a peak located in some samples was not found in other samples from initial peak finding, the region was interrogated with a S/N threshold of 20. The aligned data using the Statistical Compare feature were exported as a .csv file for further data analysis. The standard addition result was analyzed in Excel 2013 (Microsoft).

2.2.3 Results and Discussion

2.2.3.1 Quantification of NIST SRM 1950

Based on the literature review, methanol is a typical extraction solvent used for global metabolic profiling studies.[31,103,107] It provides a straightforward sample preparation, with good extraction efficiency and reproducibility for polar metabolites. The protocol for large-scale metabolic profiling of serum and plasma claimed by Warwick et al. also used methanol for protein precipitation and metabolite extraction purposes.[103] To verify the performance of the protocol for quantitative analysis, a SRM 1950 sample was prepared using the modified Nature Protocol described in Section 2.2.2.3 and 2.2.2.4A. Figure 2-2 shows the chromatogram of SRM 1950, containing various classes of compounds.

For quantitative analysis, the method of standard addition was employed to overcome any matrix effects. The amino acid and fatty acid standard mixture at various concentrations were spiked to the plasma sample to determine the concentration of two major classes of metabolites, amino acids and fatty acids. For each spiked concentration, the analyses were conducted in triplicates. The measured values were compared to the concentrations reported in the certificate of SRM 1950. The standard addition calibration curves for three amino acids and three fatty acids are shown as examples in

Figure 2-3. For both families, adequate reproducibility, with a satisfying linear range was observed.

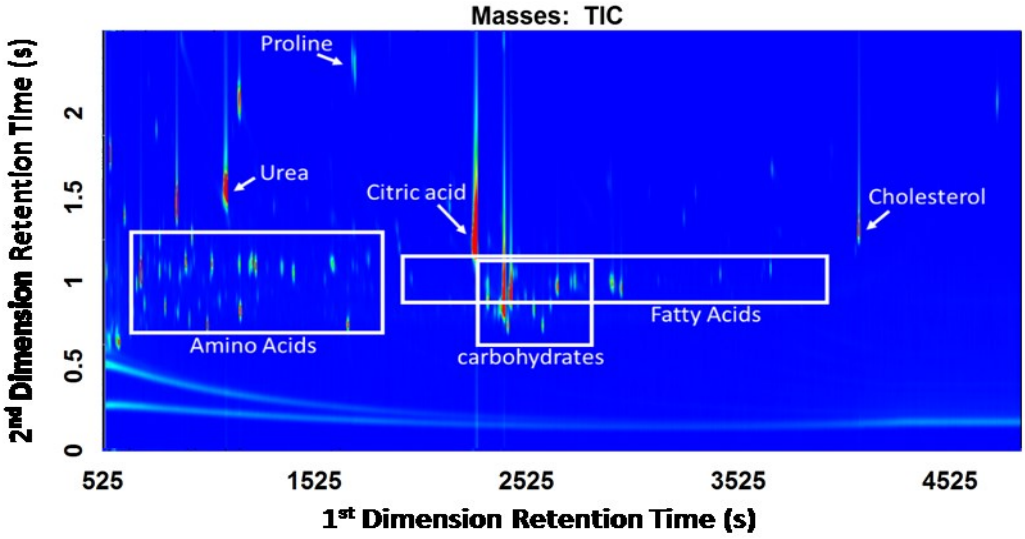

Figure 2-2. A chromatogram of NIST SRM 1950 human plasma.

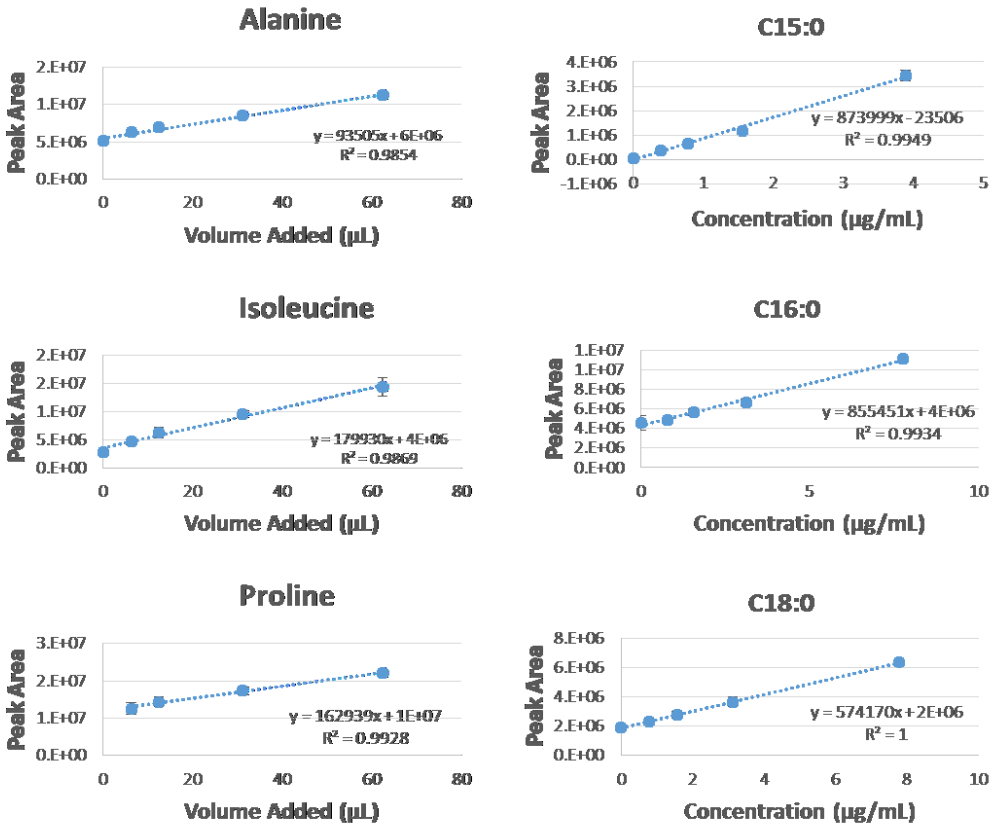

Figure 2-3. Standard addition calibration curve for amino acids (alanine, isoleucine, proline) and saturated fatty acids (pentadecanoic, hexadecanoic, octadecanoic acid).

The experimentally determined concentrations of amino acids and fatty acids were compared to the values in the certificate of analysis for SRM 1950. The uncertainties reported for both NIST certified and experimental values correspond to two standard deviations (95% confidence interval) for each analyte.[104] For amino acids, the experimental values fell in an acceptable range, considering the fact that a different analytical method was used than the one employed in the value assignment for the SRM 1950 certificate by NIST (Table 2-2). However, for fatty acids, the experimentally determined concentrations were significantly smaller than the reported values in the certificate, with the exception of dodecanoic acid (Table 2-3). A 5- to 28-fold difference was observed despite the exceptional reproducibility and linear range obtained in the calibration curve. This result is thought to be due to the poor extraction efficiency of methanol for fatty acids. Polarity is a critical factor affecting solubility in liquid–liquid extraction. Amongst different types of intermolecular forces exist, London dispersion force is the weakest. Methanol may be a satisfactory solvent for polar metabolites; however, it is not suitable to deal with the long aliphatic chain of fatty acids. The only interaction between the nonpolar aliphatic chain of fatty acid and methanol would be London dispersion forces and the energy would be insuffient to break strong hydrogen-bond in methanol. A different solvent system is required for the efficient extraction of lipid classes.

Table 2-2. The Comparison of the Experimentally Determined Values with the Certified Values

	NIST Certified Values	**Experimental**
Amino Acids	**Molar Concentration (µmol/L)**	**Molar Concentration (µmol/L)**
Alanine	300 ± 26	250 ± 32
Isoleucine	55.5 ± 3.4	62.1 ± 7.8
Leucine	100.4 ± 6.3	116.3 ± 15.8
Lysine	140 ± 14	72 ± 7
Methionine	22.3 ± 1.8	28.6 ± 11.8
Proline	177 ± 9	189 ± 39
Serine	95.9 ± 4.3	98.4 ± 16.8
Valine	182.2 ± 10.4	277 ± 32
Phenylalanine	51 ± 7	23 ± 7
Tyrosine	57.3 ± 3.0	13.5 ± 2.2

Table 2-3. The Comparison of the Experimentally Determined Values with the Certified Values for Fatty Acids

Fatty Acids	NIST Certified Values	Experimental
	Mass Fraction (µg/g)	Mass Fraction (µg/g)
C12:0	1.86 ± 0.11	3.83 ± 0.45
C14:0	17.9 ± 3.8	0.8 ± 0.0
C15:0	1.08 ± 0.01	0.11 ± 0.00
C16:0	594 ± 19	21 ± 4
C18:0	179 ± 12	19 ± 1
C20:0	5.5 ± 0.2	1.1 ± 0.3

2.2.3.2 Comparison of Extraction Solvent Systems

Plasma metabolites differ widely in chemical nature and abundances, and to obtain high accuracy and precision for all metabolites is a considerable challenge. The result with the NIST SRM 1950 sample revealed that other solvent systems need to be investigated for apolar metabolites. Amongst various extraction solvent systems that have been proposed for metabolomics studies, the Folch and Bligh–Dyer methods are commonly employed for the extraction of lipids.[101,102,108] They involve the partitioning of metabolites into a binary mixture of chloroform and methanol. Herein, a pooled human plasma, purchased from Sigma-Aldrich, was used to compare the different extraction solvent systems. The same pooled plasma was extracted with the Folch and Bligh–Dyer methods and was compared to the methanol extraction. In addition to the biphasic solvent system, the two-step solvent extractions in a series (methanol, then chloroform and methanol, then hexane) as described in Section 2.2.2.4 were compared.

Figure 2-4 shows the results of the comparison of five different extraction methods. Two amino acids (alanine, glycine), two fatty acids (dodecanoic acid, octanoic acid), and one carbohydrate (galactopyranose) were selected as representative compounds for quantitative comparison. The peak areas of the five compounds in each method were normalized to the peak area of the corresponding compound extracted with methanol. The result shows that both the Folch and Bligh–Dyer methods display higher peak areas obtained for the fatty acids, while for other polar metabolites, both methods showed poor performance. The highest peak areas for various classes of

compounds were achieved from a two-step solvent system of methanol, then chloroform. Ideally, the optimal sample preparation method is aimed to be rapid and straightforward, with a minimum number of steps. A short sample preparation time and a minimal number of steps can facilitate high throughput, while minimizing possible losses of metabolites over the number of steps in the procedure. However, there is a major drawback for the binary mixture extraction solvent methods of Folch and Bligh–Dyer. The biphasic solvent mixture results in a thick cellular and protein debris layer that locates at the interface between the upper phase (methanol/water) and the lower chloroform layer (Figure 2-5A). This insoluble residue complicates withdrawing the bottom organic layer without risking a high chance of contamination (Figure 2-5B). The issue can be lessened when the two solvents are added sequentially. Methanol can be used first to extract polar metabolites, and after the supernatant is withdrawn, chloroform is added to extract non-polar metabolites. Although it appears that an extra step was added, in reality, it makes the process of retrieving metabolites from the lower chloroform layer more convenient and may reduce the technical variability.

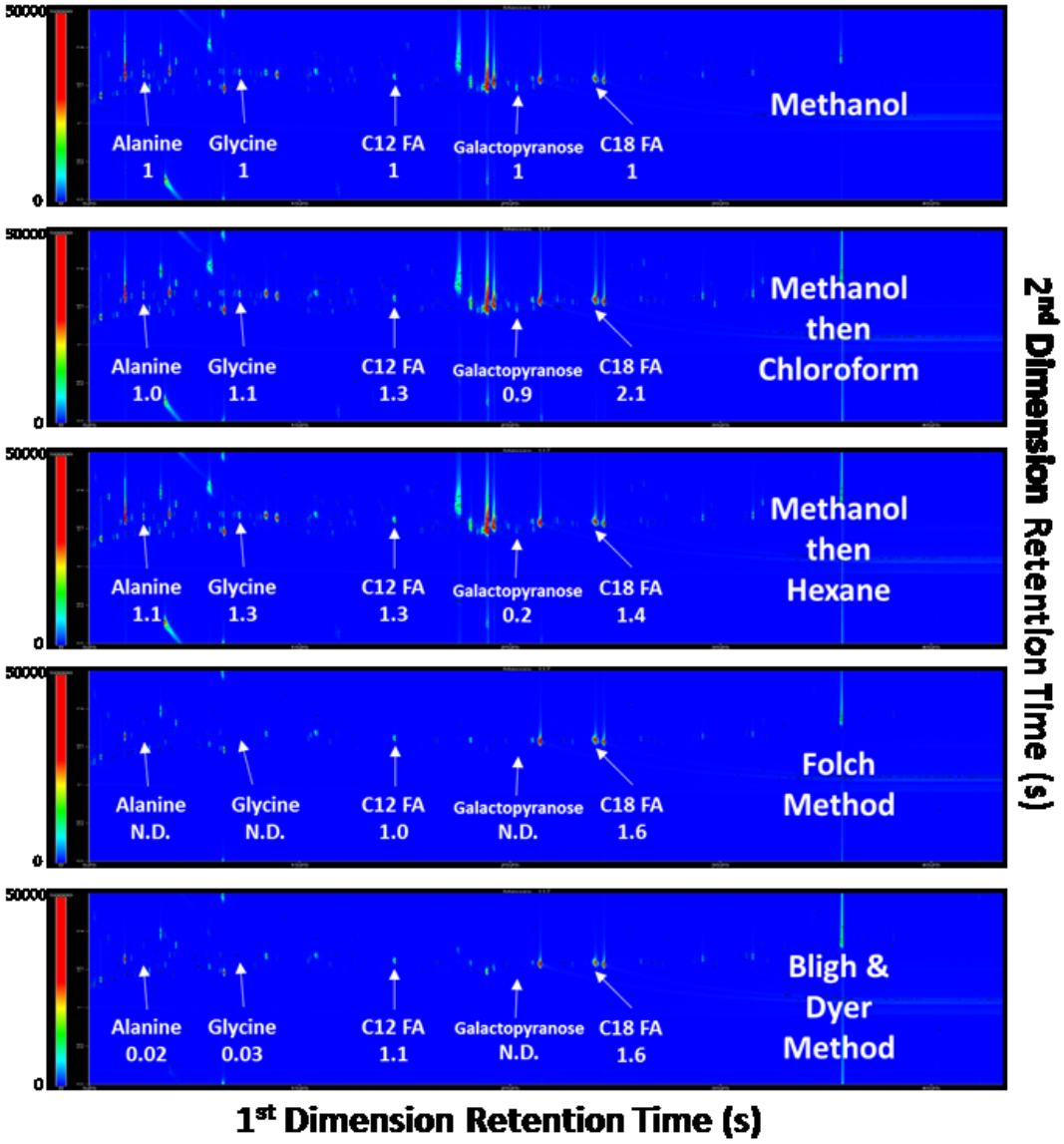

Figure 2-4. Comparison of the five extraction solvent systems, where the peak areas of five representative compounds from various chemical classes in each method were normalized to the peak area of the corresponding compound extracted with methanol. N.D signifies that the peak was not detected.

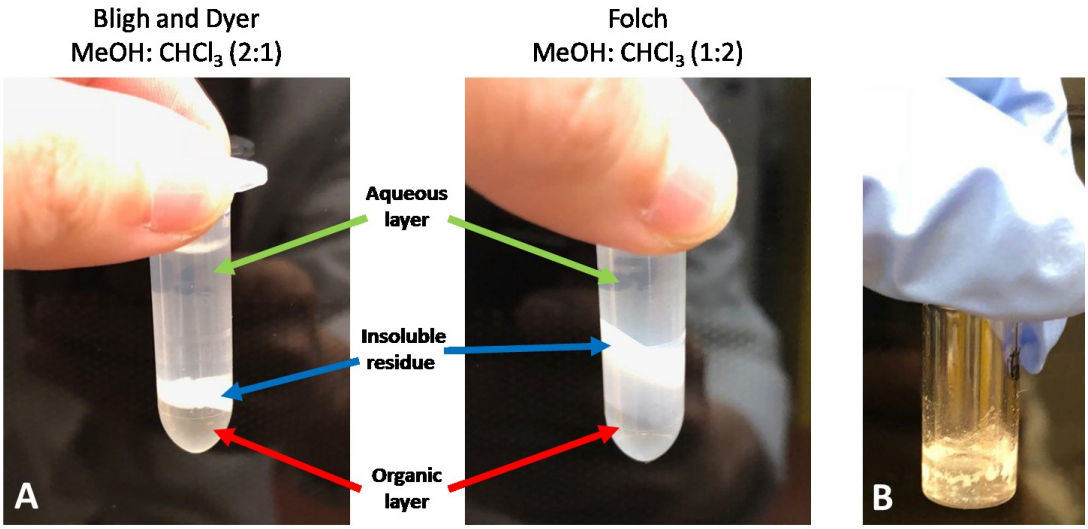

Figure 2-5. A) Phase distributions in the Bligh–Dyer and Folch methods showing the thick, insoluble protein residue being placed between the aqueous and organic layer. B) Retrieved chloroform organic later with contamination from the insoluble interface.

Another way to overcome the issue that involves reaching down to the bottom chloroform layer is to use MTBE as an organic solvent to extract non-polar metabolites.[106] Using MTBE in place of chloroform provides the advantage of having the lipid-rich organic phase in an upper layer. The use of MTBE is often called the Matyash method;[106] the modified versions of the method also were developed using different solvent ratios between methanol and MTBE.[109] Another commonly employed solvent system is a ternary combination of acetonitrile, isopropanol, and water, which aims to cover medium-polarity, lipophilic, and hydrophilic metabolites, respectively.[105] The seven different extraction solvent systems described in Section 2.2.2.4 (A-D, F-H) were analyzed in triplicate and evaluated using multivariate statistical analysis. The chromatograms of seven extraction methods were aligned and resulted in 1453 peaks. A total of 21 samples (7 extraction methods × 3 replicates) were analyzed using the Principal Component Analysis (PCA) with cross-validation to evaluate the general clustering of the extraction methods. PCA is an unsupervised method that helps to visualize and interpret the data in exploratory data analysis. The PCA score plot is a useful tool to visualize similarity/dissimilarity between samples on the PCA score space.[110] Figure 2-6 shows the PCA score plot of the plasma extracted with seven

different solvent systems. The result displays three general clusterings of the extraction methods. Acetonitrile/isopropanol/water (3:3:2 v/v/v), 100% methanol, and MTBE:methanol (5:4) were clustered together, whereas Folch, Bligh–Dyer, and MTBE:methanol (10:3) were clustered together. A two-step solvent extraction of methanol and chloroform in a series was separated from the other two clusters, while it placed in the middle of two groups in PC2. This makes sense in that the cluster of Folch, Bligh–Dyer, and a high percentage of MTBE solution are tailored to extract more non-polar metabolites. In comparison, the other cluster of three solvent systems is favored to extract more polar metabolites. Methanol, then chloroform in a series, is capable of extracting both polar and non-polar metabolites. The biplot, which is a useful tool to visualize and interpret samples and variables together, supports this finding that the polar metabolites, such as amino acids and carbohydrates are linked closer to acetonitrile/isopropanol/water (3:3:2 v/v/v), 100% methanol, and MTBE:methanol (5:4). Lipid classes, including fatty acids and sterols are found to be linked closer to Folch, Bligh–Dyer, and MTBE:methanol (10:3). The result implies that the extraction of the same sample by different extraction methods could lead to significantly varying metabolic composition and concentrations and, possibly, steer contradictory biological interpretations.

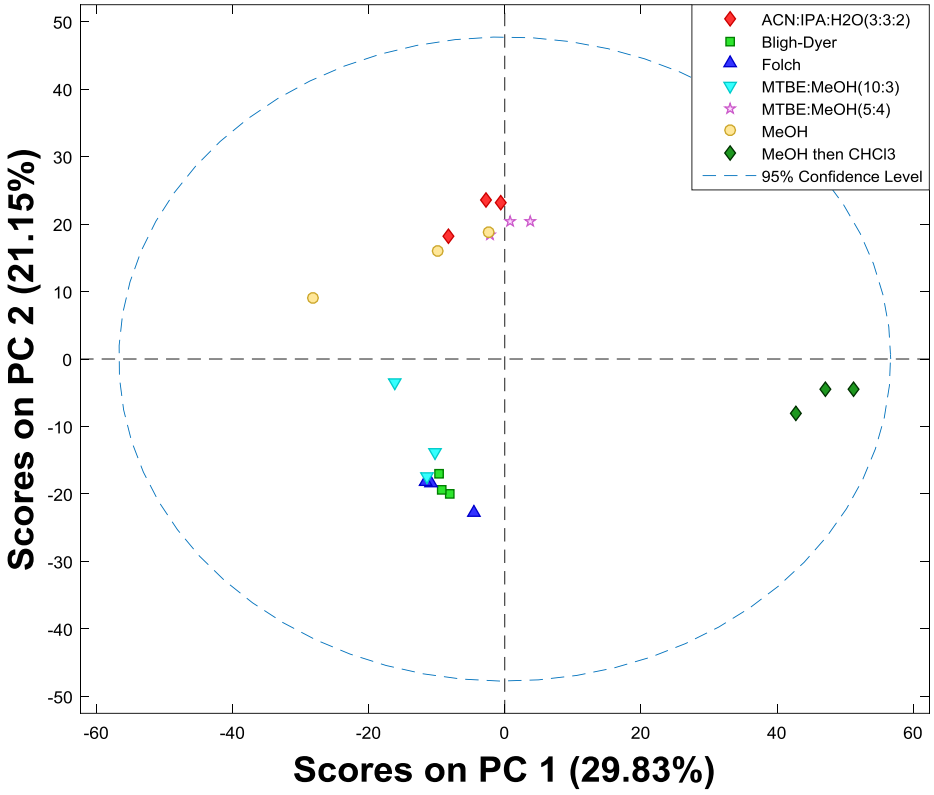

Figure 2-6. PCA score plot of the plasma extracted with seven different methods.

2.2.3.3 Development of a Solvent System for Exhaustive Extraction

Results thus far indicate that a method allowing extraction of as many classes of compounds as possible with high efficiency and reproducibility likely would require a series of extractions using multiple solvents. Two solvent systems involving a series of multiple solvent extractions were proposed (Section 2.2.2.4 I,J). For both solvent systems, the extract of each layer was contained in a different GC-vial and individually derivatized and analyzed to observe the compounds that are extracted in each step. The analyses were done in triplicates to evaluate the reproducibility of the solvent systems.

Figure 2-7 shows the chromatograms obtained from each step of the solvent system. The color scale for peak intensity was set constant for Total Ion Chromatograms (TIC) to facilitate visual comparison. The additional MTBE step of Solvent System 2 contained more lipid classes, such as fatty acids and sterols,

compared to the methanol extract of Solvent System 1. Solvent System 2 also had an addition of water that Solvent System 1 does not have. The methanol/water phase of Solvent System 2 extracted more hydrophilic and polar compounds, such as carbohydrates, carbohydrate derivatives, amino acids, urea and phosphoric acid in greater abundances. The chromatograms of the chloroform layer for the two solvent systems displayed high similarity. They appeared to be equally effective in extracting compounds like N-propylbenzamide and diethylene glycol dibenzoate and remaining fatty acids and carbohydrates in the sample matrix.

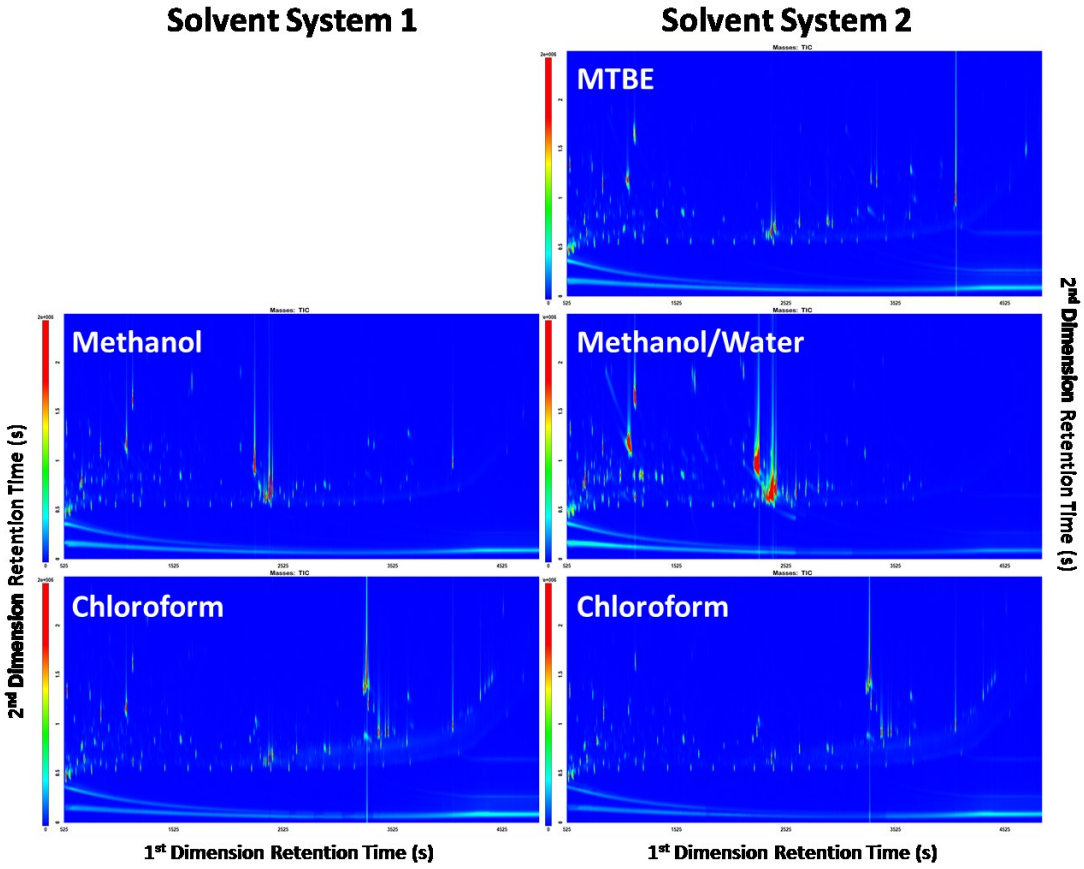

Figure 2-7. Comparison of metabolic profiles extracted in each step of the solvent systems.

A total of 15 chromatograms (5 solvent layers × 3 replicates) were aligned using the Statistical Compare feature. The peak areas of the same compound from the same sample and the same solvent system were summed to calculate the total peak area of

the metabolite extracted with the solvent system. For example, hexadecanoic acid was found in both the methanol and chloroform layers of the Solvent System 1 and was present in all the three layers (MTBE, methanol/water, chloroform) of the Solvent System 2. The peak areas of the hexadecanoic acid in methanol and chloroform layer of the replicate 1 for the Solvent System 1 were summed to represent the hexadecanoic acid of the replicate 1, Solvent System 1. For each solvent system, the mean and relative standard deviation of the peak areas in three replicates for various classes of compounds were calculated and summarized in Table 2-4.

The column "Ratio" in Table 2-4 was calculated by dividing the average peak area of a compound obtained from the Solvent System 2 by the average peak area of the corresponding compound obtained from the Solvent System 1. Of the 42 compounds listed in Table 2-4, 14 compounds were found to have a value less than 1 and, 28 compounds with a value greater than 1. A value less than 1 indicates that the compound was extracted more with the Solvent System 1 whereas a value greater than 1 indicates that the Solvent System 2 had a higher extraction efficiency for the compound. The average value of the ratio (Solvent System 2/Solvent System 1) for 42 compounds was 1.14. Overall, the peak areas obtained from the two solvent systems are similar, which indicates that both polar and apolar metabolites are extracted reliably, with sufficient efficiency for both solvent systems. The Solvent System 2 slightly outperformed the Solvent System 1 with respect to analytical precision and comprehensiveness. The Solvent System 2 displayed a better relative standard deviation with a marginally higher average peak area than the Solvent System 1.

Table 2-4. Comparison of Solvent System 1 and 2

Compounds	Solvent System 1		Solvent System 2		Ratio (S2/S1)
	Average Peak Area	Rel Std (%)	Average Peak Area	Rel Std (%)	
Fatty acids					
Pentanoic acid (C5:0)	1.9E+06	70.8	2.2E+06	5.2	1.2
Hexanoic acid (C6:0)	1.2E+07	10.5	1.3E+07	6.5	1.1
Heptanoic acid (C7:0)	1.1E+06	14.3	1.6E+06	0.9	1.5
Octanoic acid (C8:0)	1.8E+06	6.7	1.9E+06	30.3	1.1
Nonanoic acid (C9:0)	3.3E+06	13.4	9.1E+06	12.7	2.7

Decanoic acid (C10:0)	4.5E+05	4.8	4.1E+05	15.9	0.9
Undecanoic acid (C11:0)	6.7E+04	24.2	4.2E+04	15.2	0.6
Dodecanoic acid (C12:0)	4.3E+06	6.4	3.5E+06	13.5	0.8
Tridecanoic acid (C13:0)	4.8E+04	67.4	1.1E+05	109.8	2.3
Tetradecanoic acid (C14:0)	7.0E+05	5.8	4.1E+05	3.8	0.6
Pentadecanoic acid (C15:0)	3.4E+05	21.9	2.8E+05	15.0	0.8
Hexadecanoic acid (C16:0)	1.9E+07	3.7	1.8E+07	1.7	0.9
Heptadecanoic acid (C17:0)	1.8E+05	27.0	1.2E+05	2.4	0.6
Octadecanoic acid (C18:0)	1.2E+07	6.0	1.2E+07	3.2	1.0
Arachidonic acid (C20:4)	1.0E+06	91.8	5.0E+05	5.0	0.5
Amino acids					
L-Aspartic acid	1.4E+07	10.1	8.5E+06	9.1	0.6
L-Valine	1.5E+07	35.9	1.9E+07	2.3	1.3
L-Leucine	1.1E+07	44.9	1.7E+07	5.3	1.5
L-Glycine	1.6E+07	24.3	1.9E+07	0.9	1.2
L-Lysine	6.6E+06	4.5	7.3E+06	4.2	1.1
L-Proline	5.0E+05	9.5	5.2E+05	12.1	1.0
L-5-Oxoproline	3.3E+07	22.5	3.0E+07	6.5	0.9
L-Serine	3.9E+06	10.7	4.4E+06	3.7	1.1
L-Tyrosine	3.3E+06	20.0	3.3E+06	13.2	1.0
L-Methionine	5.6E+05	28.7	6.1E+05	5.3	1.1
L-Threonine	1.8E+07	12.4	1.9E+07	1.1	1.1
L-Isoleucine	6.8E+06	35.0	9.0E+06	4.1	1.3
Carbohydrates / Keto acids Derivatives					
à-Ketoisovaleric acid	3.0E+04	28.4	3.5E+04	15.8	1.2
4-Ketoglucose	1.8E+06	5.4	1.6E+06	4.1	0.9
2-Keto-gluconic acid	9.6E+05	20.0	1.1E+06	31.8	1.2
Myo-Inositol	8.3E+06	17.4	1.1E+07	6.5	1.3
D-(+)-Xylose	3.5E+07	29.4	4.1E+07	35.8	1.2
Tocopherols / Sterols					
ç-Tocopherol	1.1E+05	5.5	1.6E+05	5.1	1.5
à-Tocopherol	9.4E+05	13.1	1.1E+06	4.9	1.2
Campesterol	8.8E+04	18.9	1.0E+05	3.9	1.2
Cholestan-3-ol	2.5E+05	41.6	1.6E+05	33.1	0.6
Other Organic Compounds					
Phenol	1.8E+07	10.8	2.0E+07	1.7	1.1
m-Cresol	2.6E+06	15.2	2.6E+06	9.6	1.0
Aniline	1.6E+05	60.8	3.4E+05	46.2	2.1
à-Terpineol	7.6E+04	10.7	8.8E+04	3.2	1.2
L-(+)-Tartaric acid	4.6E+05	15.2	5.5E+05	27.1	1.2
Glycerol	3.6E+07	3.8	3.7E+07	5.0	1.0

2.2.4 Conclusions

The chemical diversity of metabolomes is enormous over a wide range of polarities. It has been accepted widely from the literature that there is no single extraction protocol at present that can extract all metabolite classes reproducibly with good efficiency. The aim of the current work was to evaluate the most commonly used extraction systems and to develop a method that can extract "all" metabolite classes in a complex matrix, such as human plasma, simultaneously. The popular extraction systems, including Folch, Bligh–Dyer, and Matyash, were investigated with a pooled human plasma. The solvent system that involves methanol, MTBE, water, and chloroform in a series was developed successfully to be the most promising extraction method for global metabolomics studies with respect to comprehensive coverage and analytical precision. The developed solvent system is applicable for the metabolite classes typically present in a biological matrix and yielded reproducible results for human plasma. However, the increased number of extraction steps requires sacrificing the throughput of the study. Nonetheless, the present study provides options for the extraction solvent system that can be selected, based on the goal and the utmost importance of the analysis, whether it is for targeted analysis, qualitative global profiling, or combined targeted and untargeted methods (i.e. quantitative analysis of as many metabolites as possible in which comprehensive extraction of metabolites would be vital).

2.3 Optimization of Derivatization Parameters

2.3.1 Introduction

The compounds analyzed by GC are required to be volatile, thermally stable, and with low or intermediate polarity.[48] However, the presence of many hydrophilic functional groups in metabolites such as hydroxyl, amine, amide, thiol, and carboxyl groups causes a significant rise in boiling points, making them not suitable for GC analyses.[57] These functional groups commonly are found in most major metabolite classes, such as amino acids, fatty acids, carbohydrates, and organic acids.[53] To achieve the characteristics suitable for GC analysis, the samples have to be treated prior to GC

analysis, following a derivatization procedure to reduce the polarities of the functional groups and render the analytes stable and volatile enough for GC analysis.[10] The three most commonly used derivatization methods for GC analyses are alkylation, acylation, and silylation.[111] Of the several derivatization procedures that exist, silylation has been the gold standard for GC-based metabolomics studies as the most global derivatization method.[57] A two-step derivatization, methoximation followed by silylation, is the most versatile and universally applicable derivatization method that enables the detection of various metabolite families, including sugars, amino acids, organic acids, amines, and alcohols.[111]

2.3.1.1 Methoximation

Methoximation is a reaction in which ketone and aldehyde functional groups are protected from the chemical interconversion between two or more forms, such as keto-enol tautomerization, by replacing the oxygen atom of the carbonyl group to methoxyamino groups (Figure 2-8.1).[10] Methoximation is a critical step prior to trimethylsilylation to enhance quantitative metabolite profiling, especially for the compounds containing of α-keto acid groups. Many metabolites contain α-keto acid groups, and these organic compounds, such as oxaloacetate and pyruvate, are essential in biological systems, as they are involved in the Krebs citric acid cycle and glycolysis.[112] Methoximation helps to protect α-keto acids from decarboxylation by fixing the enolizable keto groups.[53] If the keto group is left unprotected, α-keto acids may undergo chemical loss of the carboxyl group as carbon dioxide and result in unreliable quantification of metabolites.[53,54,57]

In a biosample, carbohydrates typically are present in significant abundances. These carbohydrates are present in a cyclic or open chain form with an aldehyde or a ketone group. The aldehyde or ketone group, along with the carbohydrate hydroxyl groups, can undergo an intramolecular reaction to form a cyclic hemiacetal or hemiketal.[53] Depending on the position of the anomeric hydroxyl groups in carbohydrates (α- or β-), various forms of hemiketals/hemiacetals can be produced, leading to different peaks in the chromatograms. It is difficult to control the ratio of the

different peaks, which adds significant complexity in quantification. Thus, methoximation is crucial to block the carbonyl group of sugars, thereby preventing the ring formation of reducing sugars that otherwise would lead to multiple chromatographic peaks.[1,10,33] In most cases, methoximation results in two distinct peaks from the syn- and the anti-form of the methoxyamine group due to the inhibition of rotation across the C=N bond (Figure 2-8.1).[53]

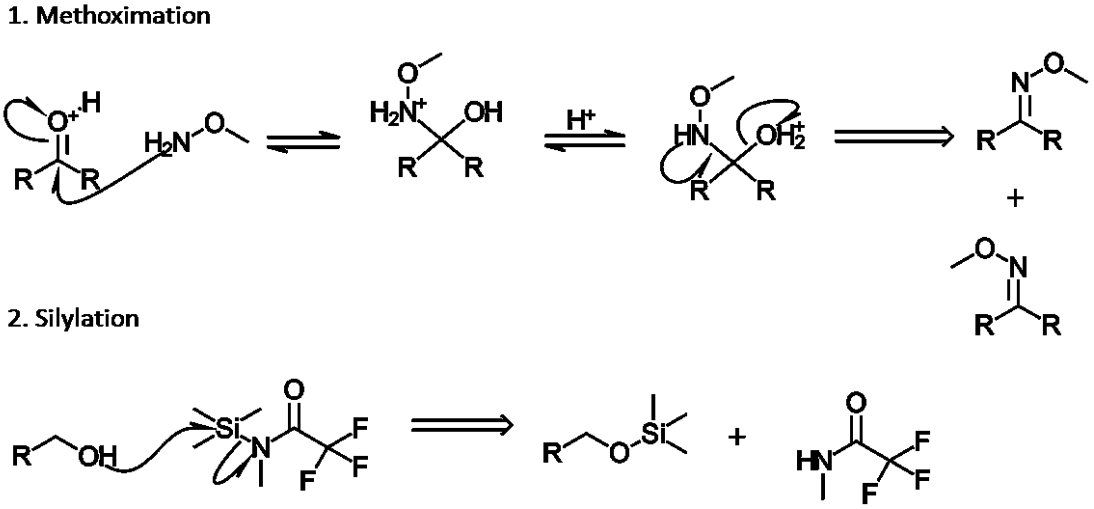

Figure 2-8. The general reaction mechanism for the two-step derivatization. 1.methoximation 2. silylation via MSTFA.

The degree of completion of the methoximation reaction depends on multiple parameters, including reaction time, temperature, and concentration of the methoximation reagent. Methoxyamine hydrochloride dissolved in pyridine is typically used as a methoximation reagent since pyridine, as an aprotic and basic solvent, captures protons and increases the nucleophilicity of methoxyamine.[53] Of various derivatization protocols reported for GC-based metabolomics, the concentration of methoximation reagent does not vary much (20 mg/mL) because generally, it is added in excess; however, there is a wide variation in methoximation time and temperature combinations reported in the literature, from 15 min to 17 hours at different temperatures (Table 2-5).

Table 2-5. Comparison of Derivatization Parameters from Publications

Authors (Year) [Ref]	Methoximation			Trimethylsilylation		
	Concentration, Volume	Temp (°C)	Time (h)	Reagent, Vol	Temp (°C)	Time (h)
Palmas et al. (2018)[113]	20 mg/mL, 30 µL	RT	17	MSTFA 30 µL	RT	1
Yu et al. (2017)[114]	20 mg/mL, 50 µL	30	1.5	MSTFA 70 µL	60	1
Chan et al. (2011)[107]	20 mg/mL, 50 µL	60	2	MSTFA 100 µL	60	1
Liebeke et al. (2019)[115]	20 mg/mL, 80 µL	37	1.5	BSTFA 100 µL	37	0.5
Fiehn (2016)[80]	20 mg/mL, 10 µL	30	1.5	MSTFA 91 µL	37	0.5
Weinert et al. (2015)[116]	20 mg/mL, 50 µL	40	1	MSTFA 200 µL	75	1.5
Ye et al. (2012)[117]	20 mg/mL, 100 µL	40	2	MSTFA 80 µL	40	1
Dunn et al. (2011)[103]	20 mg/mL, 50 µL	80	0.25	MSTFA 50 µL	80	0.25

2.3.1.2 Silylation

The low volatility of compounds commonly can arise from the presence of polar functional groups, with the formation of strong hydrogen bonds.[57] The silylation reaction breaks molecular proton bridge bonding and replaces an active hydrogen with a silyl group, commonly three methylsilylate (TMS) groups.[10] Replacement of an active hydrogen by a silyl group reduces the polarity of the compound and hydrogen bonding; hence, it decreases the boiling point and increases the volatility of polar compounds (Figure 2-8.2). In addition to enhanced physicochemical properties of silylated derivatives for GC purposes, improved chromatographic features, including narrow and symmetric peak shapes, can be observed with the use of siloxane stationary phases.[57]

Silylation is recognized as the most versatile derivatization method currently available for GC analysis that works for various functional groups. The challenge is that the derivatization reaction time varies for the different functional groups on the analytes due to the differences in the ease of formation of TMS derivatives.[118] While the hydroxyl (-OH) and carboxyl (-COOH) groups react almost simultaneously with the silylation reagent, the active hydrogen atoms of the amine ($-NH_2$) groups react slower. In general, greater success can be achieved for compounds containing OH groups, whereas nitrogen-containing functional groups undergo silylation less readily.[57] Steric hindrance also affects the formation of TMS derivatives. Primary groups react more readily than secondary groups, which then react faster than tertiary

groups.[53] Silylation of sterically hindered functional groups may require more vigorous reaction conditions.[57] With the differences in derivatization reaction time required for the complete reaction, it is possible to observe partial silylation for some metabolites containing amine functional groups, which leads to the formation of numerous peaks for a single metabolite. The compounds that resulted in more than one derivative may show different ratios of fully and partially derivatized compounds across samples, which complicates the data analysis.[1,81] For example, glycine can result in three different TMS derivatives, based on the derivatization condition affecting the level of completion (Figure 2-9). The first product is with a TMS replacing the proton of the carboxyl group, but the amine group is left underivatized. The second product is with 2TMS groups replacing the carboxyl group and the less acidic primary amine. The third product is with 3TMS replacing all the active protons. A generation of multiple derivatives for a single compound can create confusion regarding the metabolite concentration or the number of analytes present in the sample, causing inconsistencies in quatification.[85]

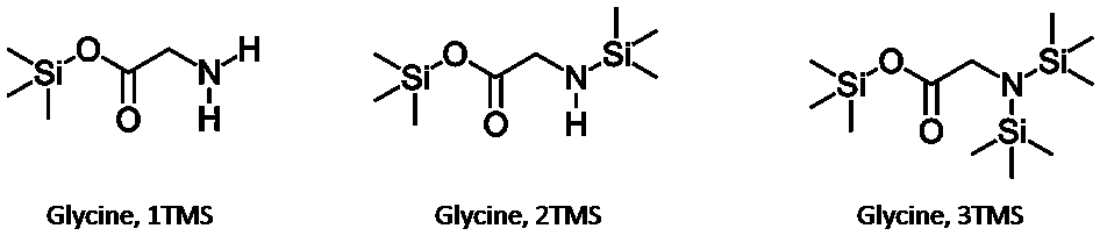

Figure 2-9. Different TMS derivatives of glycine, based on the degree of derivatization completion.

2.3.1.3 Challenges in Standardization of Derivatization Parameters

When derivatization reactions are carried out, the derivatization reagents are generally added in excess to allow all metabolites to be derivatized "sufficiently". The two-step reactions are conducted in a sealed GC glass vial, commonly at elevated temperature to facilitate the chemical transformation.[53] Many derivatization protocols have been published for GC-based metabolomics studies with incubation temperatures and derivatization times varying significantly between protocols practiced by different

laboratories (Table 2-5). Among current published protocols for different biological samples are the global urinary[107] and plasma[103] derivatization methods published in *Nature Protocols.* In the protocol for global urinary metabolic profiling, 60 °C for 2 h was suggested for methoximation, followed by 60 °C for 1 h for silylation;[119] in the protocol for large-scale metabolic profiling of serum and plasma, 80 °C for 15 min was suggested for methoximation, followed by 80 °C for 15 min of silylation.[103] It is not clear why the differences in the derivatization conditions for these two biofluids would be so stark.

While ensuring sufficient completion of derivatization would be beneficial, it would be ideal to keep the preparation time as short as possible to minimize technician labour costs. Additionally, degradation of certain metabolite derivatives or further rearrangement reactions can possibly occur with longer reaction durations.[1] Determining the optimal duration of the derivatization is of great importance for reliable metabolomics studies. In the literature, most silyl derivatization methods required reaction conditions in the range of 60–80 °C and 15–120 min. The laborious and long-hour derivatization procedures are viewed as a great hindrance, but little effort has been placed on developing methods that speed up the reaction time while offering a reproducible yield of derivatives.[51] Despite the versatility of silylation, when it comes to quantification of analytes, consensus on the utility of silylation drops dramatically due to the problem of unsatisfactory reproducibility. Irreproducible derivatization and degradation of derivatized metabolites hinder the reliable quantification and introduce errors in metabolomics studies.[120]

As discussed previously, chemical derivatization is an essential aspect of GC-based metabolomics, and great care needs to be taken in the development and optimization of derivatization conditions.[33] The metabolites can be analyzed reproducibly only when the derivatization efficiency and the stability of derivatized products are sufficiently adequate. The development of silylation based GC-MS methods poses serious challenges, mainly due to the large range of compound classes and the large differences in metabolite concentrations.[37] Establishing an optimal derivatization protocol that ensures complete derivatization and minimizes the

generation of multiple peaks from a single metabolite can enhance both separation and quantification.[121] However, given the complexity of biosamples with metabolites over a broad range of physicochemical properties, finding the best derivatization condition that allows complete intended modification for all metabolites without degradation or rearrangement reactions is challenging.[1]

2.3.1.4 Design of Experiments

A statistical approach referred to as "Design of Experiments" (DoE) is an efficient method for planning and conducting experiments that allow the evaluation of several variables and their interactions simultaneously.[52,122] DoE is useful for planning experiments with the goal of obtaining sufficient information through the minimum number of experiments, while allowing several experimental parameters to be varied systematically and simultaneously.[123] In the context of DoE, independent variables are called factors, and different values for each factor in which experiments are conducted are called levels.[123]

One of the most used DoE approaches for optimization studies is the Response Surface Methodology (RSM).[122,124] RSM is a mathematical technique that is used effectively in exploring the relationships between several explanatory variables and response variables simultaneously.[124] RSM has been recognized as a valuable tool for the development, parameter optimization, and improvement of the response variables.[123,125-127] RSM is an efficient procedure, which primarily aims to yield the maximum amount of information from the minimum amount of work.[122] RSM uses partial factorial designs to reduce the number of experimental points required. The two main types of response surface designs are Central Composite designs and Box-Behnken designs. The Box-Behnken design usually has fewer design points than central composite designs; thus, it can reduce the analysis time and cost. Box-Behnken designs must always have three levels per factor. To understand the Box-Behnken design better, a three-factor Box-Behnken design is illustrated visually in Figure 2-10. It uses middle points instead of corner points, and all experimental points are located on the edges of a cube around the center points.[52] Table 2-6 shows the experimental

runs required for a three-factor Box-Behnken design, where -1, 0, 1 indicate low, medium, high-level, respectively. A full factorial three-level design would require n^3 experiments, where n is the number of factors to be optimized. For the optimization of three variables at three different levels, a full factorial design would require 27 experiments, whereas RSM using Box-Behnken design requires only 13 experiments. It is a massive saving in time and associated experimental analysis costs. Considering the fact that for the full factorial design, the number of experiments required will escalate rapidly as the number of factors or levels increase, RSM is an efficient systematic procedure for studying the interaction effects of various parameters.

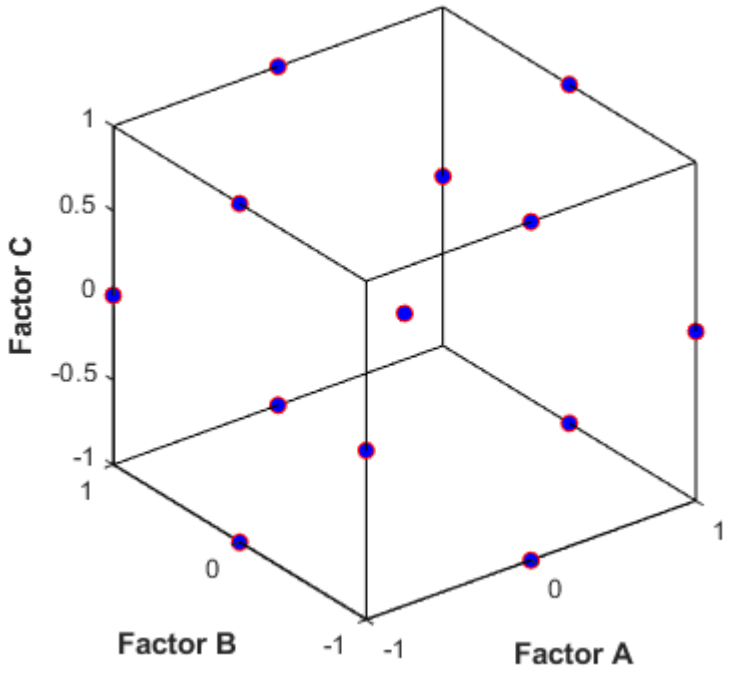

Figure 2-10. Box-Behnken design. The blue dots denote the experimental condition used in the design.

Table 2-6. DoE Run Table with a Three-Factor Box-Behnken Design

Runs	Factor A	Factor B	Factor C
1	-1	-1	0
2	-1	1	0
3	1	-1	0
4	1	1	0
5	-1	0	-1
6	-1	0	1
7	1	0	-1
8	1	0	1
9	0	-1	-1
10	0	-1	1
11	0	1	-1
12	0	1	1
13	0	0	0

Herein, different variables for derivatization were investigated collectively for urine metabolomics. The four factors (temperature and duration of the methoximation and silylation) were investigated for the optimization of derivatization conditions using DoE. The Box-Behnken design was employed to vary the levels systematically in a relatively limited number of experiments, and response surface modelling was used to explore the relationship of derivatization factors and to obtain a greater understanding of the collective effects of different variables in a mathematical system. The objective of this study is to find the derivatization conditions (time and temperature) that maximize the completion of the reaction in the shortest possible time.

2.3.2 Experimental

2.3.2.1 Sample and Chemicals

The same pooled urine sample was stored at –80 °C until the analysis was used for both the DoE study and the validation. Urease (Millipore-Sigma Canada) suspensions of approximately 40 mg per 250 µL of 18.2 MΩ deionized water (Elga PURELAB flex 3 system, VWR International, Edmonton) were prepared on the day of derivatization. HPLC grade methanol (>99.9%) was purchased from Millipore-Sigma Canada, and HPLC grade toluene (Millipore-Sigma Canada) was dried with anhydrous sodium

sulfate (Millipore-Sigma Canada) prior to use. Methoxyamine hydrochloride (Millipore-Sigma Canada) was dissolved in HPLC grade pyridine (Millipore-Sigma Canada) to make a 20 mg/mL solution, and 1 mL ampoules of N-methyl-N-trimethylsilyltrifluoroacetamide + 1% chlorotrimethylsilane (MSTFA + 1% TMCS) were purchased from Fisher Scientific Canada. A 100 µg/mL solution of 4-^{13}C methylmalonic acid was prepared in deionized water to use as an internal standard.

2.3.2.2 Sample Preparation

The frozen urine samples were thawed on ice on the day of analysis and vortexed for 1 min. Then, 40 µL of each urine sample was transferred into a 2 mL centrifuge tube, and 20 µL of internal standard were added, along with 10 µL of urease suspension. The samples were vortexed for 3 min, then incubated at 37 °C for 1 h. Next, 960 µL of methanol were added to each of the samples, which were vortexed again for 5 min, then centrifuged for 10 min at 10,000 g and 4 °C. A 500 µL aliquot of the supernatant was transferred to a 2 mL GC vial and dried under a gentle stream of nitrogen at 50 °C. A corresponding incubation time and temperature was used for methoximation and silylation (Table 2-4). Methoximation was performed using 50 µL of 20 mg/mL methoxyamine hydrochloride in pyridine, and silylation with 100 µL MSTFA.

2.3.2.3 Design of Experiment and Response Surface Methodology

A Box-Behnken design with four factors and three levels was used for DoE (Table 2-7). The RSM optimization was performed using Minitab® 19.2020.1 (64-bit), and 27 experimental conditions are listed in Table 2-8. Normalized peak areas of derivatized metabolites were used as response variables for optimization.

Table 2-7. Factors and Levels Used in the Box-Behnken Design for Derivatization Optimization

	Factors (unit)	Low	Middle	High
1	Methoximation temperature (°C)	60	70	80
2	Methoximation duration (min)	15	67.5	120
3	Silylation temperature (°C)	60	70	80
4	Silylation duration (min)	15	37.5	60

Table 2-8. Experimental setup for the optimization of derivatization

Exp #	Factor 1	Factor 2	Factor 3	Factor 4
	MeOX Temp (°C)	MeOX Time	Silylation Temp (°C)	Silylation Time (min)
1	60	15	70	37.5
2	60	120	70	37.5
3	80	15	70	37.5
4	80	120	70	37.5
5	70	67.5	60	15
6	70	67.5	60	60
7	70	67.5	80	15
8	70	67.5	80	60
9	60	67.5	70	15
10	60	67.5	70	60
11	80	67.5	70	15
12	80	67.5	70	60
13	70	15	60	37.5
14	70	15	80	37.5
15	70	120	60	37.5
16	70	120	80	37.5
17	60	67.5	60	37.5
18	60	67.5	80	37.5
19	80	67.5	60	37.5
20	80	67.5	80	37.5
21	70	15	70	15
22	70	15	70	60
23	70	120	70	15
24	70	120	70	60
25	70	67.5	70	37.5
26	70	67.5	70	37.5
27	70	67.5	70	37.5

2.3.2.4 GC×GC-TOFMS Conditions

All GC×GC-TOFMS analyses were performed on a LECO Pegasus 4D system (Leco Instruments, St. Joseph, MI) equipped with a four-jet dual-stage modulator. A 60 m × 0.25 mm; 0.25 µm df Rxi-5SilMS was used as the first-dimension column and a 1.6 m × 0.25 mm; 0.25 µm df Rtx-200MS as the second-dimension column (Chromatographic Specialties). Two-dimensional chromatographic separations were performed with a

constant flow rate of 2.0 mL/min using helium as the carrier gas and a modulation period of 2.5 s. A GERSTEL MPS Autosampler was used for automated injection of 1 µL aliquots of samples. The oven was held initially at 80 °C for 4 min and heated to 315 °C at a ramping rate of 3.5 °C/min. The final temperature was held for 10 min. The secondary oven and modulator temperature offset were constant at +10 °C relative to the GC oven temperature and +15 °C relative to the secondary oven temperature, respectively. Mass spectra were collected at an acquisition rate of 200 Hz over a mass range between 40 and 800 *m/z*. A relative voltage offset of 200 V was selected as the optimized detector voltage, with an electron impact energy of −70 eV. The ion source temperature was 200 °C, with a transfer line temperature of 250 °C. The total analysis time for each run was 81.1 min. All samples were analyzed in randomized order. The first sample was analyzed by GC×GC-TOFMS immediately after the derivatization was finished and the last sample was analyzed within 48 h of the end of the derivatization.

2.3.2.5 Data Processing and Analysis

All data were processed using the LECO ChromaTOF® (v4.72) software. Automatic peak detection and mass spectral deconvolution were performed with a peak width set to 15 s in the first dimension and 0.18 s in the second dimension. Peaks with a signal-to-noise ratio lower than 100 were rejected. Spectral searching was performed using the NIST 2017 and Wiley 08 library database that was added to the software. The peaks of different samples were aligned using the Statistical Compare feature in ChromaTOF®, with the retention time shift tolerance of 5 modulation periods in the first dimension and 0.2 s in the second dimension to account for the possible retention time shifts. Peak areas were calculated using the selected quantification masses for each metabolite. Each compound was normalized against the Total Useful Peak Area (TUPA) of the corresponding sample (Chapter 3) prior to the RSM optimization data analysis using Minitab® 19.2020.1 (64-bit).

2.3.3 Results and Discussion

The effects of temperature and time of oximation and silylation on the derivatization of the urinary metabolites were explored over a broad range of these parameters in 27

experimental runs (Table 2-8). The 27 samples were aligned using the Statistical Compare feature on ChromaTOF® and resulted in 1731 aligned peaks, with 540 compounds present in all the samples. All metabolites in each sample were normalized to the corresponding TUPA (the sum of the 540 common compounds of the corresponding sample).

To identify the optimal derivatization conditions, RSM was employed to evaluate the data mathematically. Urine is a complex matrix that contains a great number of compounds with highly diverse chemistry. Theoretically, each metabolite with its distinct chemistry would be affected differently by different derivatization conditions. However, it is not possible to investigate each metabolite separately. Therefore, the classes of metabolites that are typically of interest, such as amino acids, fatty acids, and sugars, were selected, and the determination of optimum values was examined separately by the chemical classes. RSM establishes a relationship between control variables (time and temperature for oximation and silylation) and responses (normalized peak area) and determines the significance of the variables through computational statistical testing. RSM uses a polynomial equation to fit the response (peak area) and finds the optimum combination of the parameters that will result in the maximum peak area for a number of derivatized metabolites that belong to the same class. A multi-response desirability function was applied computationally to search for optimal derivatization parameters that simultaneously maximize the peak areas of numerous metabolites of the same family. The Multiple Response Prediction results for amino acids and glucosamine derivatives were shown as examples in Table 2-9 and Table 2-10. The results include the standard error of the fit (SE fit), 95% confidence interval (95% CI) and 95% prediction interval (95% PI).

Multiple Response Prediction

Variable	Setting
MeOX Temp	60
MeOX Time	27.7273
TMS Temp	80
TMS time	15

Response	Fit	SE Fit	95% CI	95% PI
Glycine	0.0663	0.0178	(0.0283, 0.1044)	(0.0182, 0.1145)
Cysteine	0.000828	0.000249	(0.000294, 0.001363)	(0.000153, 0.001504)
Homoserine	0.000676	0.000699	(-0.000824, 0.002176)	(-0.001220, 0.002572)
Tyrosine	0.00833	0.00194	(0.00418, 0.01248)	(0.00308, 0.01358)
Lysine	0.002060	0.000444	(0.001108, 0.003013)	(0.000857, 0.003264)
Histidine	0.00317	0.00130	(0.00037, 0.00596)	(-0.00036, 0.00670)
Phenylalanine	0.002043	0.000384	(0.001220, 0.002865)	(0.001003, 0.003082)
Glutamic acid	0.000319	0.000091	(0.000124, 0.000514)	(0.000072, 0.000566)
Aspartic acid	0.000076	0.000021	(0.000031, 0.000121)	(0.000019, 0.000133)
Methionine	0.000271	0.000094	(0.000069, 0.000474)	(0.000015, 0.000527)
Proline	0.003205	0.000619	(0.001877, 0.004533)	(0.001526, 0.004883)
Threonine	0.003205	0.000619	(0.001877, 0.004533)	(0.001526, 0.004883)
Serine	0.00795	0.00216	(0.00332, 0.01259)	(0.00210, 0.01381)

Table 2-9. The RSM result for the class of amino acid TMS derivatives. The regression equation is used where SE fit stands for the standard error of the fit. 95% CI and 95% PI represents 95% confidence interval and 95% prediction interval, respectively.

Multiple Response Prediction

Variable	Setting
MeOX Temp	60
MeOX Time	120
TMS Temp	80
TMS Time	15

Response	Fit	SE Fit	95% CI	95% PI
G5	0.000442	0.000127	(0.000171, 0.000714)	(0.000115, 0.000769)
G4	0.001104	0.000294	(0.000474, 0.001734)	(0.000345, 0.001863)
G3	0.000316	0.000077	(0.000151, 0.000481)	(0.000117, 0.000515)
G2	0.000163	0.000051	(0.000054, 0.000272)	(0.000032, 0.000294)
G1	0.000500	0.000148	(0.000183, 0.000817)	(0.000117, 0.000883)

Table 2-10. The RSM result for the N-Acetyl glucosamine methoxime TMS derivatives.

The optimal conditions for each class of compounds are summarized in Table 2-11. The optimal condition that was determined from the RSM was 28 min and 15 min methoximation at 60 °C for amino acids and fatty acids (both short-chain fatty acids (C6, C7) and long-chain fatty acids (C14, C16, C18)), respectively. Amino acids contain amine (NH_2) and carboxyl (COOH) functional groups that are subject to

trimethylsilylation. The carbonyl group adjacent to oxygen in the ester group is not electropositive enough for nucleophilic attack by the methoxyamine reagent; therefore, the methoximation step would not impact the amino acids significantly. The same thing applies to the family of fatty acids whose carbonyl group is not a target of methoximation reaction, which can explain the short oximation time determined by the response optimization. For both amino acids and fatty acids, a short silylation (15 min) at high temperature (80 °C for amino acids and long-chain fatty acids, 70 °C for short-chain fatty acids) was selected as the optimal condition. It was observed from the desirability function that for some compounds, such as cysteine, tyrosine, and lysine, long term silylation negatively affected and decreased the responses of the derivatized compounds (Figure 2-11).

Table 2-11. Results of the Optimized Conditions for Each Class from DoE Analysis

	MeOX Temp (°C)	MeOX Time (min)	TMS Temp (°C)	TMS Time (min)
Short Chain FA	60	15	70	15
Long Chain FA	60	15	80	15
Sugars	60	120	80	60
Glucosamine-related	60	120	80	15
Amino Acids	60	28	80	15
Organic Acids	60	120	80	60

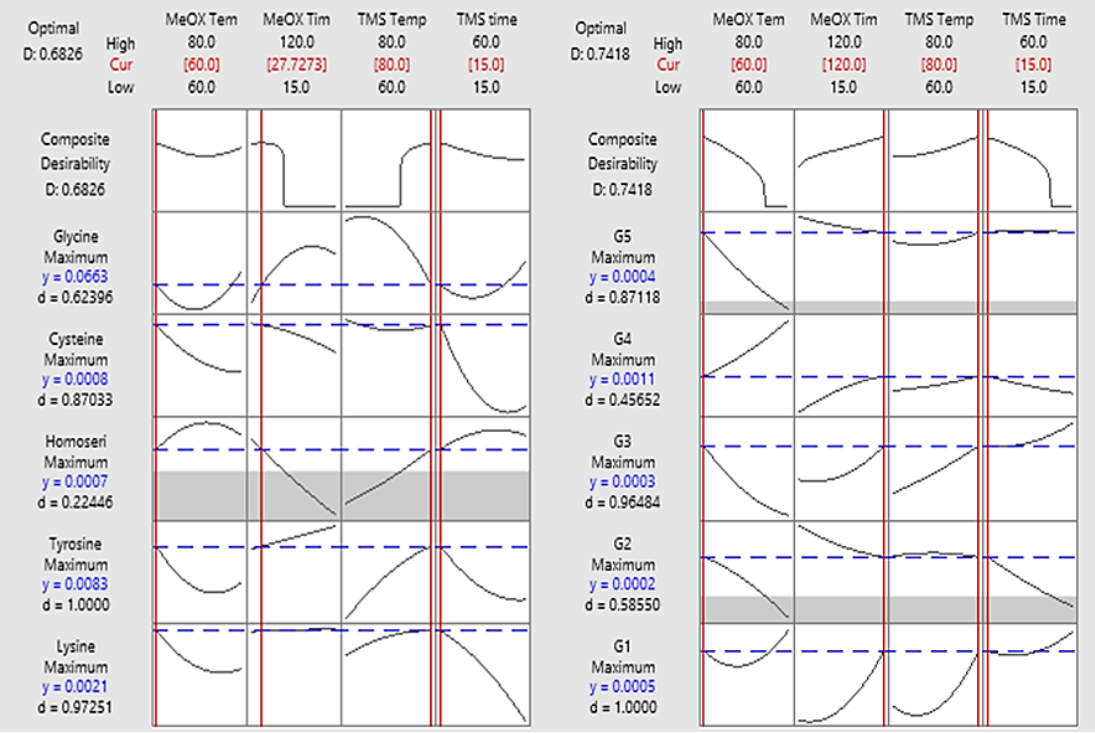

Figure 2-11. Composite desirability Left: amino acids Right: Glucosamine-related compounds.

On the other hand, the result showed that a longer oximation time (120 min) at 60 °C increased the formation of TMS-derivatized products for sugars, N-acetyl-glucosamine, and other organic compounds. Of the sugar compounds, glucosamine-related compounds were selected separately for RSM optimization because Gullberg et al. reported that the oximation must be done for a long time and/or at high temperature to ensure complete derivatization for glucosamine.[100] Other organic acids included various organic compounds that are subject to methoximation, such as α-ketoglutaric acid, and also the compounds that are not affected by methoximation, such as tartaric acid, cresol, and lactic acid. Methoximation is a critical step, especially required for reliable detection for TMS derivatives of sugars by blocking the carbonyl group and preventing ring formation that otherwise would lead to multiple chromatographic peaks. Glucosamine-related compounds have one site for methoximation to take place with five potential silylation sites. For the compounds that contain target sites for methoximation, such as sugars and glucosamine-related compounds, the optimization results indicated a long oximation time. In addition to the number of possible sites for

oximation and silylation, the concentrations of the compounds present in a sample also play a significant role in determining the derivatization time required. Carbohydrates usually are present in high abundance, with a large isomer diversity. Also, relatively bulky compounds, such as disaccharides (e.g., maltose) may require long derivatization at elevated temperature due to steric hindrance. Therefore, the determined condition of long oximation and silylation time for carbohydrates from RMS seems to be consistent with the understanding of the derivatization mechanism.

The DoE results confounded choosing a single adequate derivatization protocol for all the compounds across a sample, as the optimum derivatization conditions differ widely for various classes of compounds. To evaluate the reproducibility of the derivatized products from varying derivatization conditions, three methods were selected, along with the original protocol that was published in *Nature Protocol*, for further comparison and were analyzed in triplicates (Table 2-12). Method 1 and 2 are with a long oximation time (120 min) at 60 °C, with varying silylation time. Method 3 has a short oximation time, with a short silylation time.

Table 2-12. Four Conditions Selected for Reproducibility Validation

Method	Methoximation Temp (°C)	Methoximation Time (min)	Silylation Temp (°C)	Silylation Time (min)
Method 1 (M1)	60	120	80	60
Method 2 (M2)	60	120	80	15
Method 3 (M3)	60	15	80	15
Nature Protocol (NP)	60	120	60	60

The compounds that have oximation sites, thereby would be affected by the methoximation condition, and they were investigated primarily in detail. Gullberg et al. reported that the temperature and/or time of the oximation had impacted glucosamine and α-ketoglutaric acid the most.[100] The α-ketoglutaric acid resulted in three derivatized products, which include α-ketoglutaric acid, 2TMS and two isomers of 2-oxoglutaric acid, O-methyloxime, 2TMS. Methoximation typically leads to two products, syn- and anti-forms, when the methoximation replaces the oxygen atom of the carbonyl group by methoxyamine (=N-OCH_3). When methoximation did not occur fully, and the two

carboxyl groups were trimethylsilylated, it resulted in α-ketoglutaric acid, 2TMS. Figure 2-12 is a zoomed-in EIC at *m/z* 89, displaying four chromatographic peaks, where peak A and C are the two isomers of 2-oxoglutaric acid, O-methyloxime, 2TMS. Figure 2-13 shows the ratio between syn- and anti- forms of 2-oxoglutaric acid, O-methyloxime, 2TMS. The advantage of methoximation is that it produces two methoxime isomers, and the ratio of syn- and anti- form is stable for each metabolite because it is dependent only on the internal energy of the molecule.[53] The ratio between syn- and anti-form is similar for all the four derivatization conditions; however, the most considerable variations were found with Method 3, which has the shortest methoximation time. Method 1 and the *Nature Protocol* method have a longer silylation time than Method 2, which can possibly allow further methoximation to occur and lead to a more reproducible ratio between syn/anti. In addition to the least stability of the syn/anti ratio of peak A and C in Figure 2-12, peak B (2-heptenoic acid, (E)-, TMS derivative) is present in the least abundance with Method 3. In the following figures, the abbreviations were used for each method that was analyzed in triplicates as M1, M2, M3, and NP representing Method 1, Method 2, Method 3, and *Nature Protocol* method, respectively.

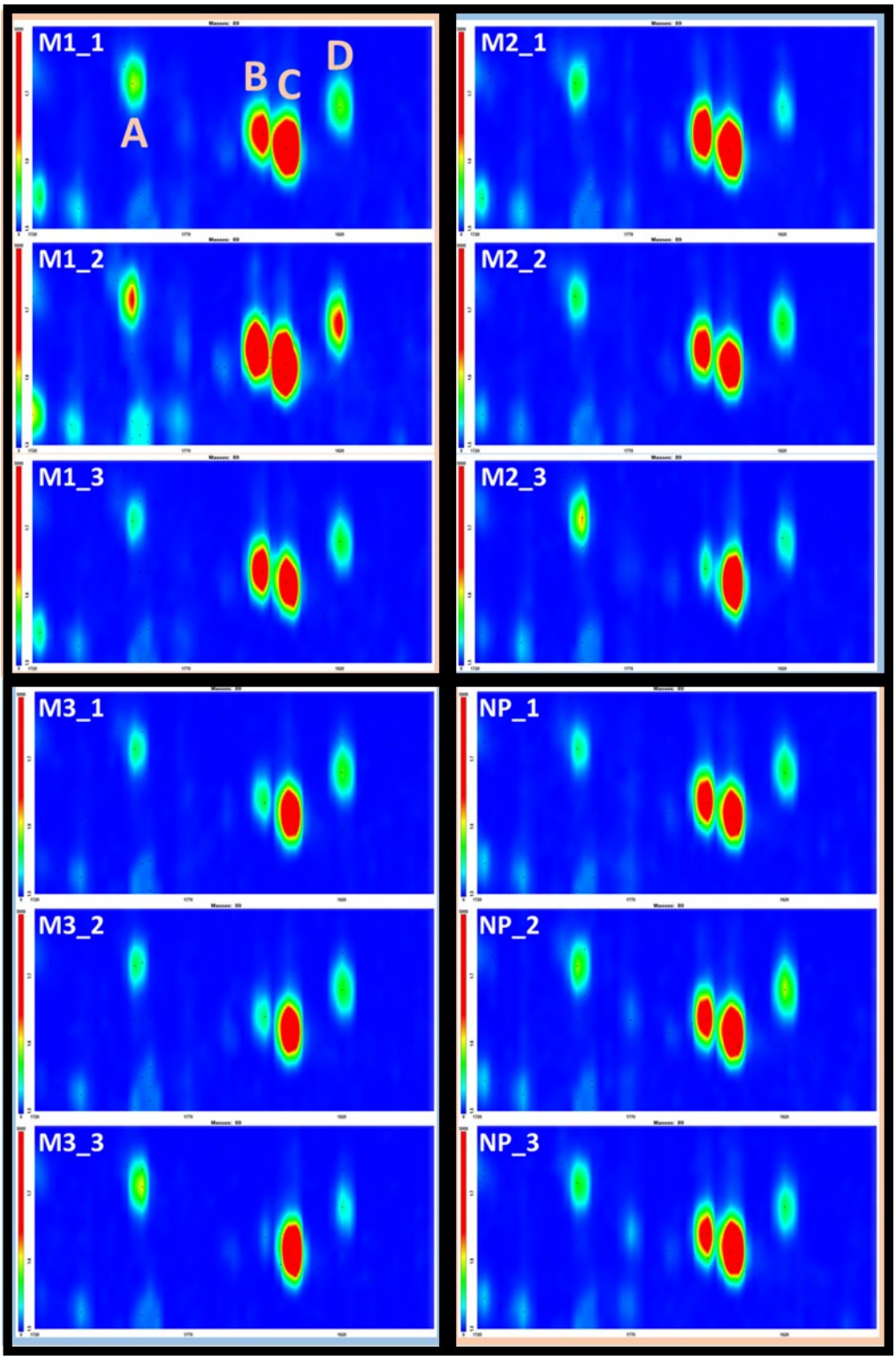

Figure 2-12. A zoomed-in EIC at *m/z*89 showing four derivatized products. A) 2-Oxoglutaric acid, O-methyloxime, 2TMS. B) 2-Heptenoic acid, (E)-, TMS derivative. C) 2-Oxoglutaric acid, O-methyloxime, 2TMS. D) 1,2-Dithiane-4,5-diol, (Z)-, 2TMS derivative.

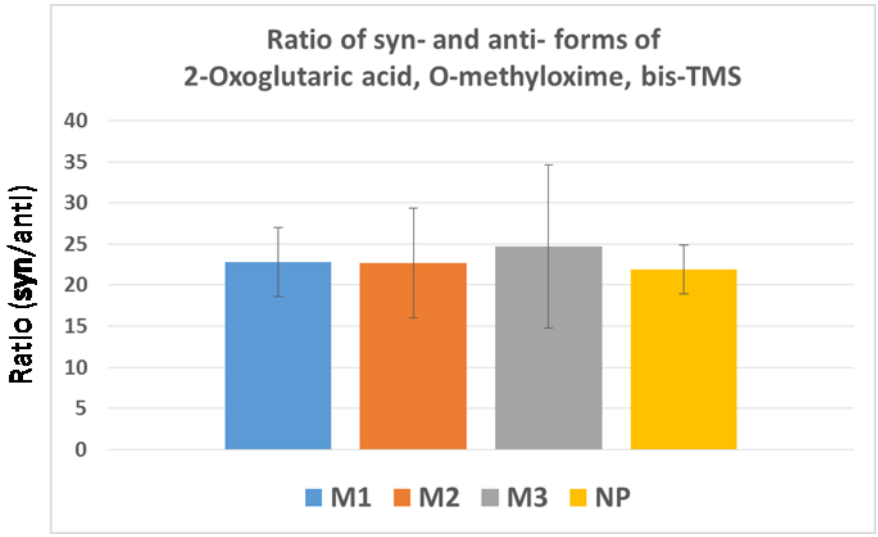

Figure 2-13. The ratio of syn- and anti- forms of 2-oxoglutaric acid, O-methyloxime, 2TMS. Error bars represent the standard deviation (n=3).

The derivatization efficiency was represented quantitatively via calculating the percentage of incomplete methoximation by dividing the peak area of α-ketoglutaric acid, 2TMS by the sum of the peak areas of the three derivatized products of α-ketoglutaric acid for the corresponding sample (Figure 2-14). Method 3 (the shortest methoximation time) resulted in the largest production of the incompletely methoximated α-ketoglutaric acid, 2TMS compounds.

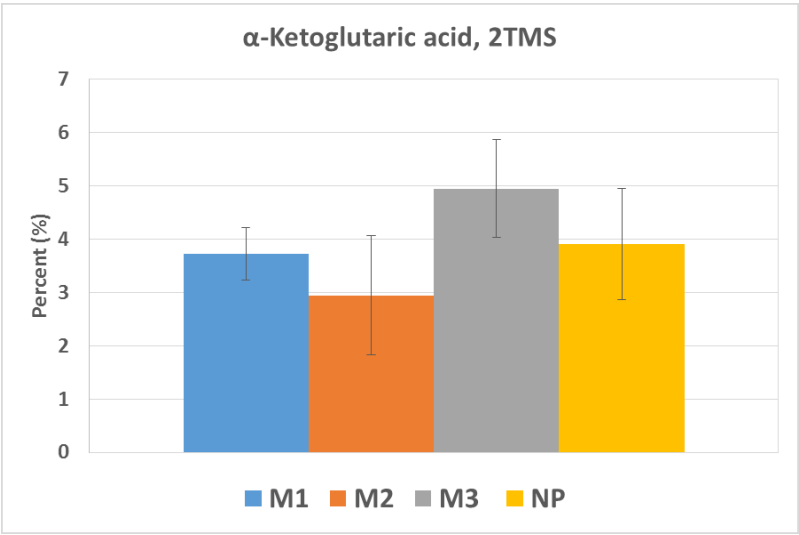

Figure 2-14. The percentage of α-ketoglutaric acid, 2TMS, a silylated derivative without methoximation. Error bars represent the standard deviation (n=3).

4-ketoglucose is a sugar compound that contains two potential sites for methoximation and four hydroxyls for silylation. In Figure 2-15, it is evident that a long oximation time is required for successful methoximation of 4-ketoglucose. From the triplicate analyses done with Method 3, no 4-ketoglucose, bis(O-methyloxime), tetrakis(trimethylsilyl) was detected at all. A longer methoximation time at higher temperature appears to be necessary for the analysis of compounds like 4-ketoglucose; however, this might raise problems for some compounds that would be at risk of decomposition.

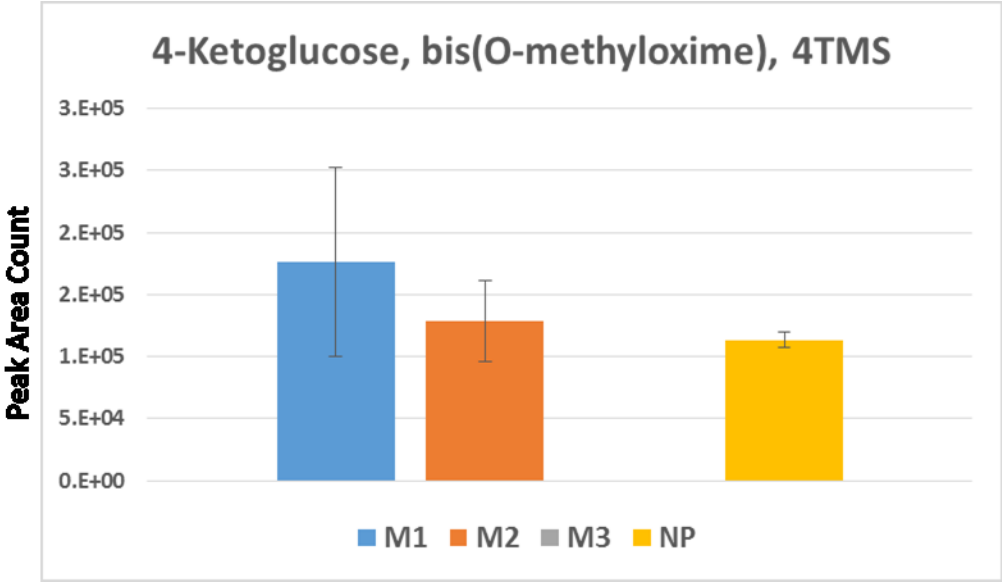

Figure 2-15. The peak area comparison of 4-ketoglucose, bis(O-methyloxime), tetrakis(trimethylsilyl). Error bars represent the standard deviation (n=3).

N-acetyl-D-(+)-glucosamine has one oximation site with five potential silylation sites. Figure 2-16 and Figure 2-17 show one peak and five peaks, respectively, that were identified as N-Acetyl-D-glucosamine, tetrakis(trimethylsilyl) ether, methyloxime with the MS library match score ranging 65–85 %. Although the true identity of these six peaks would not be confirmed without the standards, they are considered to be structurally similar compounds, such as mannose or galactose. In both Figure 2-16 and Figure 2-17, the derivatized N-acetyl-D-(+)-glucosamine peaks are present in the least abundance in Method 3. Figure 2-18 shows the quantitative

comparison of the three most abundant peaks in Figure 2-17, peak 2,4, and 5. The comparison of the peak areas of the three peaks that were identified as the derivatized products of N-acetyl-D-glucosamine is in agreement that Method 3 results in the least abundant methoximated products.

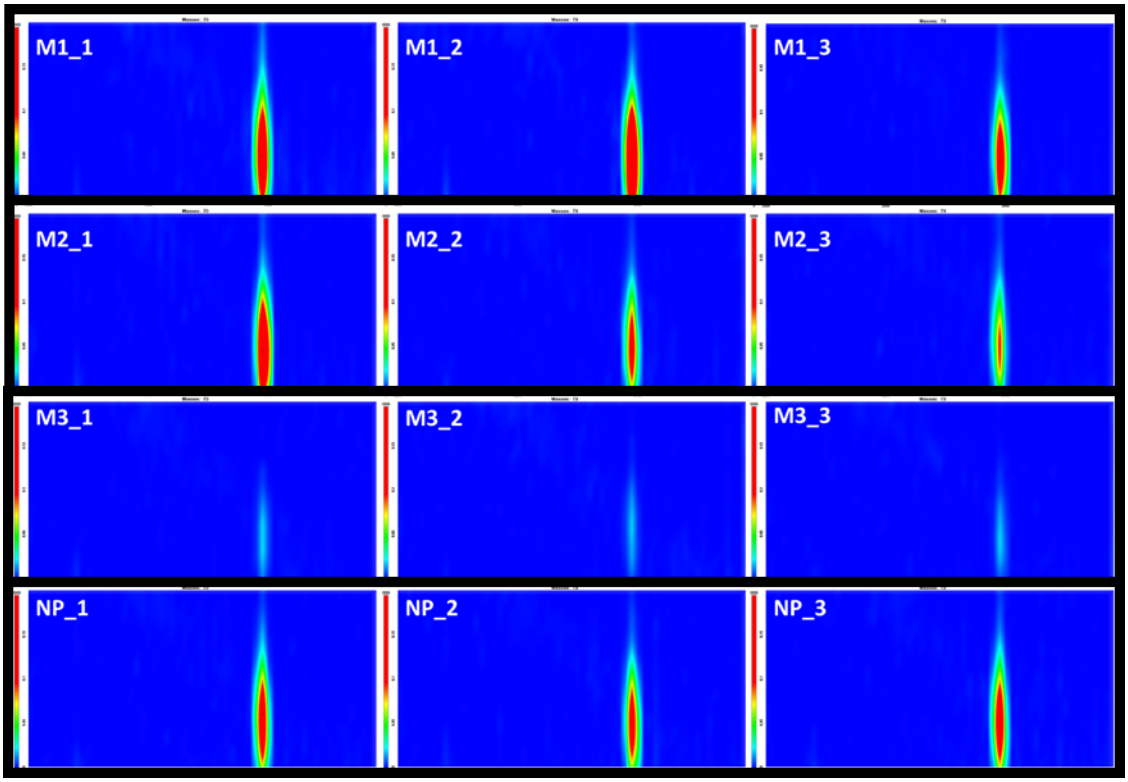

Figure 2-16. N-acetyl-D-glucosamine, tetrakis(trimethylsilyl) ether, methyloxime (syn).

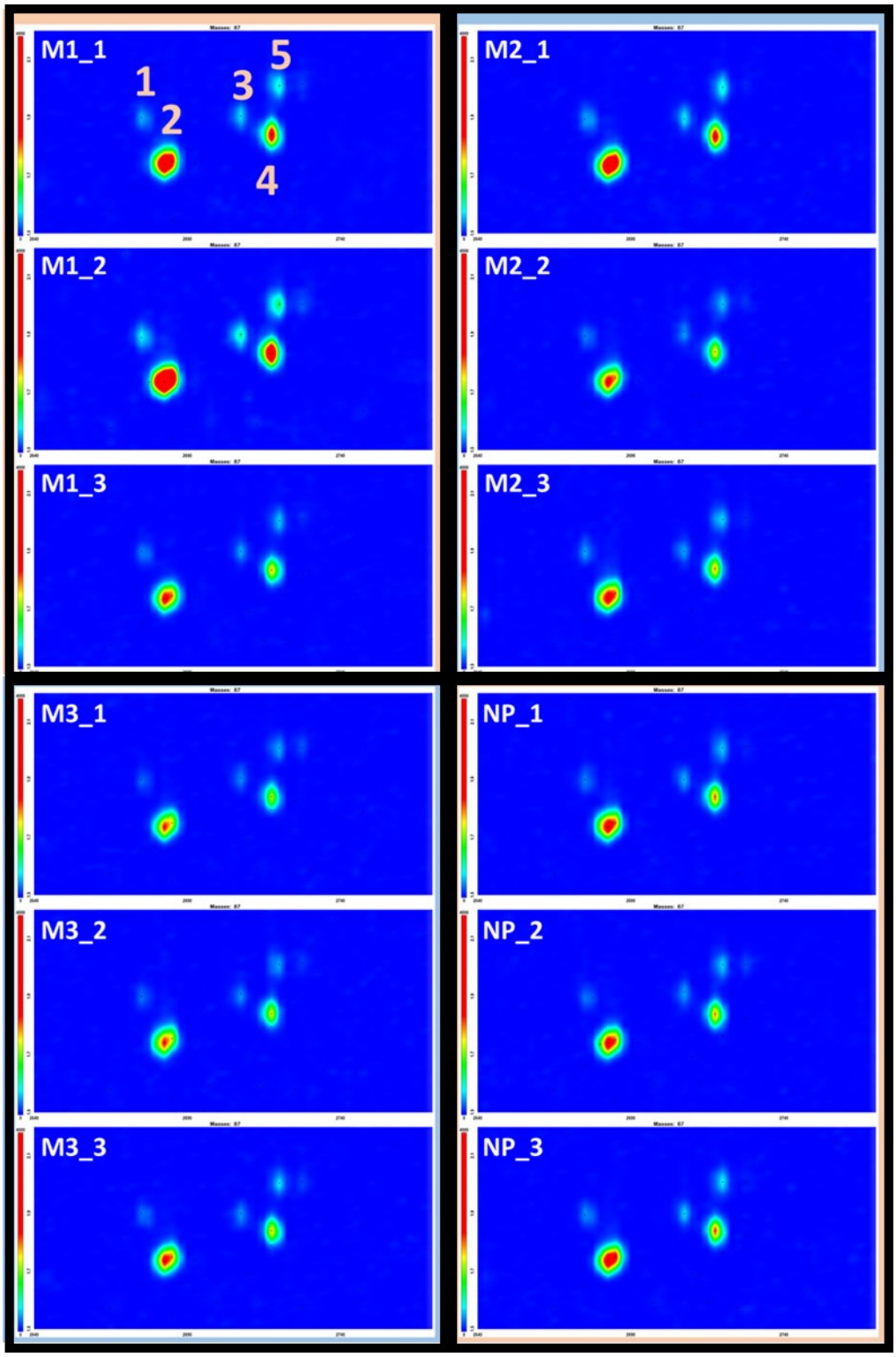

Figure 2-17. Five peaks of N-acetyl glucosamine methoxime, 4TMS.

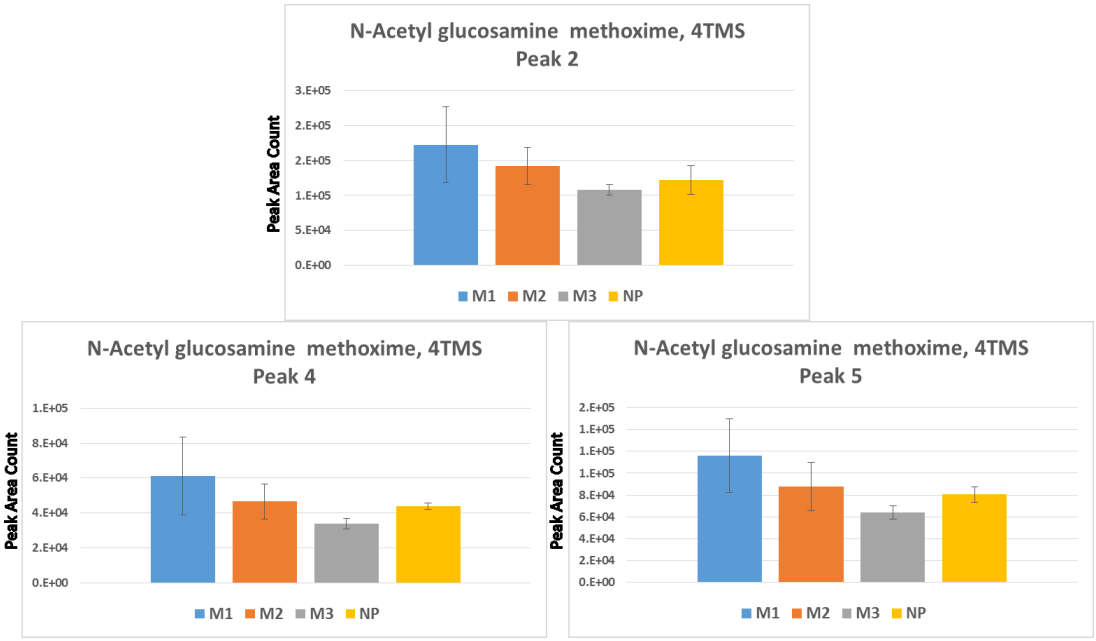

Figure 2-18. Comparison of the derivatized N-acetyl-D-(+)-glucosamine peaks. Error bars represent the standard deviation (n=3).

One of the major limitations in quantitative GC-MS metabolomics via silylation would be the difficulty in preserving a constant degree of conversion of the metabolite to its silylated derivatives. In all the derivatization conditions, it was observed that hydroxyl and carboxyl groups are derivatized effectively in almost all cases. In contrast, some compounds were found to be derivatized only partially, leaving the amine/amide group non-silylated. A longer silylation duration may be required to result in a complete derivatization product for compounds containing amine/amide groups to prevent the formation of multiple derivatized peaks of a single analyte. As an example, N-acetylglutamic acid is a compound that contains two hydroxyl groups and a secondary amide. Secondary amides are less reactive than hydroxyl groups; thus, the protons at both hydroxyl groups will be replaced faster than the secondary amide. In all four derivatization conditions, N-acetylglutamic acid with 2TMS groups replacing two hydroxyl groups was found in significant abundances (Figure 2-19); however, a 3TMS derivative of N-acetylglutamic acid was not detected. This may be explained by the study conducted by Bekele et al. that longer than four hours of silyation time may be needed to ensure sufficient derivatization of slowly reacting analytes.[52] One hour of

silylation time appeared to be insufficient to convert the slowly reacting amide and would require a longer time for silylation.

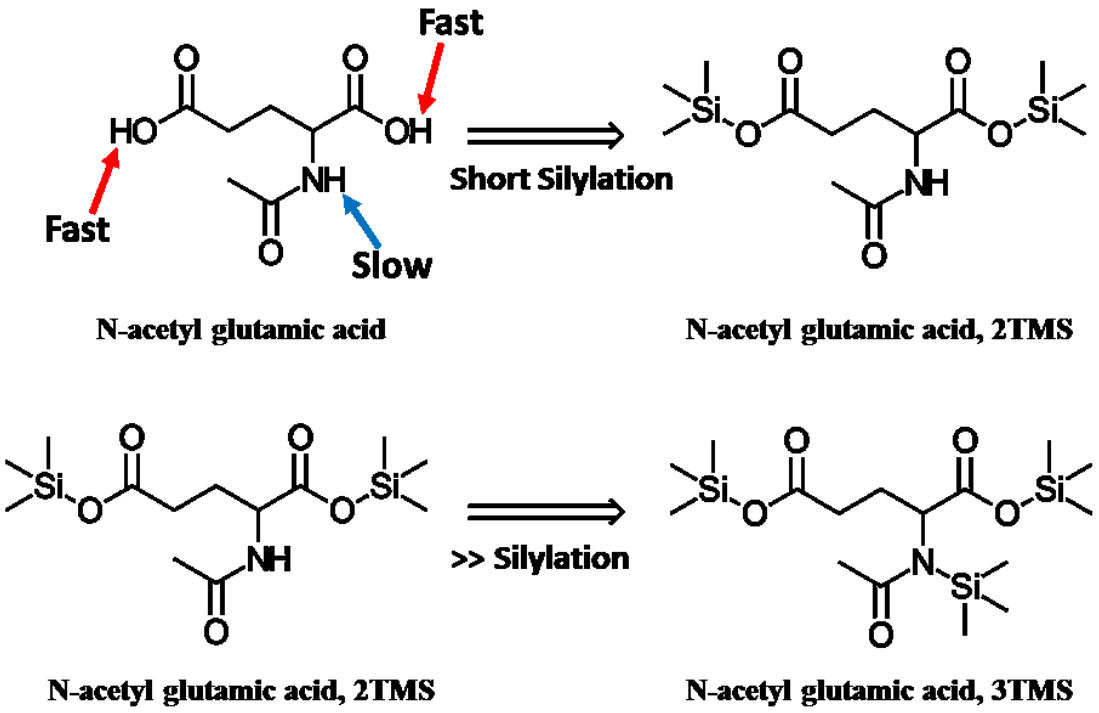

Figure 2-19. Silylation of N-acetyl glutamic acid to 2TMS and 3TMS with two hydroxyl groups and the secondary amide being derivatized.

A higher temperature and a longer derivatization time may drive the reaction and contribute to the completion of chemical conversion; however, this may risk the progressive degradation of heat-labile and unstable metabolites. For example, two products resulted from the derivatization of glycine, 2TMS and glycine, 3TMS. Figure 2-20 displays the ratio of the peak areas of glycine, 3TMS and glycine, 2TMS for the four different derivatization conditions. It is noteworthy that Method 3 and the *Nature Protocol* conditions resulted in higher glycine, 3TMS contents than the other two methods. Method 3 has a higher temperature (80 °C) with short silylation time (15 min), whereas the Nature Protocol uses a lower temperature (60 °C) with a long silylation time (60 min). To avoid decomposition of some compounds, it appears to be keeping at least one condition mild (time or temperature) is needed. Method 1 has a higher temperature (80 °C) and a long silylation time (60 min) and resulted in lower glycine,

3TMS contents compared to Method 3 and *Nature Protocol*. Method 2 has the same silylation condition as Method 3. The reason for low glycine, 3TMS content is not clearly understood yet; however, based on a large relative standard deviation, it is suspected that the degradation reaction might have proceeded while the samples were waiting for injection. The finding would require further verification, yet temperature and time during the silylation did not seem to have much dramatic effect on the compounds as methoximation.

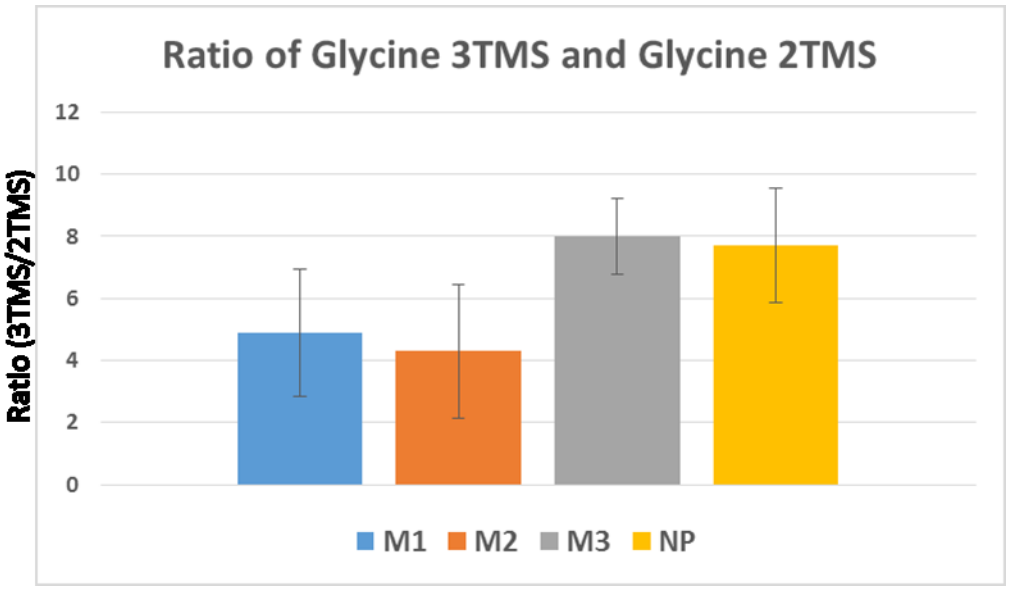

Figure 2-20. Comparison of the ratio between glycine, 3TMS and glycine, 2TMS. Error bars represent the standard deviation (n=3).

Gluconolactone (GCL) is commonly used as a food additive and is a lactone of gluconic acid. GCL hydrolyzes in water to gluconic acid, establishing chemical equilibrium between the lactone form and the acid form. Pocker et al. reported that the rate of hydrolysis is increased by the addition of heat.[128] With the two-step derivatization procedures, GCL was converted to gluconolactone, 4TMS derivative, while gluconic acid was derivatized to gluconic acid, 6TMS. Figure 2-21 shows the ratio of the TMS derivatives of gluconic acid and gluconolactone. The result indicated that gluconic acid was present 2.0–2.8 times more than gluconolactone in the four

different derivatization methods. The derivatization condition did not impact the kinetic equilibrium between the two compounds significantly.

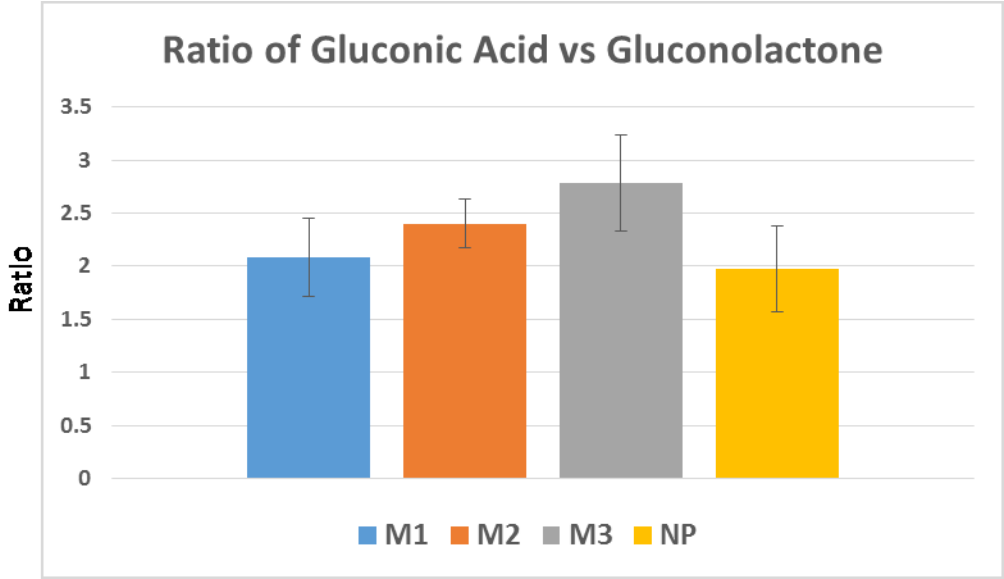

Figure 2-21. The ratio between gluconic acid and gluconolactone. Error bars represent the standard deviation (n=3).

2.3.4 Conclusions

Despite the complexity and challenges that exist in sample preparation, GC(×GC)-MS is an indispensable analytical platform for untargeted analysis. Due to the highly contrasting physicochemical properties and widely differing metabolites, finding a single optimum set of derivatization parameters that satisfies all classes of compounds may not be achievable. Using DoE and RSM for multi-response optimization, the optimal set of derivatization parameters for numerous classes of urinary metabolites were reported. Four derivatization conditions were selected for a more detailed performance comparison of the derivatization methods, with particular focus on the abundance of (in)completely derivatized products and their reproducibility. Oximation at 60 °C for 2 h seemed to be beneficial to ensure sufficient methoximation for compounds that are potentially subject to the reaction. Silyation seemed to be almost instantaneous (acceptable with 15 min at elevated temperature) for carboxyl and hydroxyl groups; however, for metabolites containing amine/amide functional groups, a longer silylation time would be required.

2.4 Investigation on Different Column Combinations

2.4.1 Introduction

GC×GC uses two independent separation mechanisms to provide high peak capacity and structured chromatograms.[22] These advantages of GC×GC are extremely beneficial for the analysis of highly complex biosamples, such as urine, blood, and feces. To take the full advantages of the GC×GC and achieve maximum component separation, the selection of column combinations plays a significant role.[42]

The significance of proper column selection is well-recognized in GC-world as selectivity has the greatest impact on resolution. Several selections of GC stationary phases are available commercially, and the vendors have put efforts to guide users for column selection and optimizing separations.[43,44] The typical base structures of the GC stationary phase have been polysiloxane polymers and polyethylene glycols since their first introduction in the 1950s. The major disadvantages of the traditional GC polar phases have been low thermal stability and limited tunability for altering selectivity, which has been discussed more in detail in Chapter 1 (Section 1.2.1). Despite many continuous improvements made on the GC phase platforms, the conventional polysiloxane polymer-based stationary phase platforms have remained almost unchanged since their development and still retain the fundamental limitations.[46]

To overcome the limitation of the conventional GC polar phases, ionic liquid columns, which have entirely different chemistries from the traditional columns that are based on polysiloxane or polyethylene glycols (PEG), have been developed.[46] In terms of chemistry, ionic liquids generally consist of two or more organic cations joined by a linkage and charge balancing anions (either inorganic or organic). Ionic liquids do not have active hydroxyl groups, and they are much smaller physically compared to bulky polysiloxane and polyethylene glycol phases. These properties of ionic liquids lead them to have greater stability even in the presence of oxygen and/or moisture. In addition, ionic liquids have high thermal stability, which potentially can replace the conventional polar GC stationary phases with their increased maximum temperature and lower column bleed over the traditional columns of similar polarity. Since the

column selectivity is dependent on the chemical composition, many modifications are possible, with numerous choices of cations and anions to make columns with the desired selectivity.[129,130] With all these advantages and potentials, ionic liquids columns hold great promise in the world of GC columns as an option for polar stationary phases that can provide a higher operatable temperature limit.[46]

Table 2-13 shows various commercial non-ionic liquid (polysiloxane-based and PEG) columns and ionic liquid columns with their temperature limits. For non-ionic liquid columns, in general, it is listed from non-polar to more-polar. It is improper to compare polarity on a one-way scale since different types of polarity exist, making it more complicated than a one-dimension polarity scale. However, it is sufficient to demonstrate the general trend of the maximum temperature decreasing as the stationary phase becomes more polar for the conventional polysiloxane and PEG columns. The three ionic liquid columns have selectivity slightly more polar than the PEG phase according to the GC column polarity scale provided by Supelco.[45] While the polarity is similar, ionic liquid columns display higher maximum temperatures.

Table 2-13. Various GC column Stationary Phases and Their Temperature Limits

	Phase Composition	**Max Temp**
Non-Ionic Liquid	100% Dimethyl polysiloxane	350 °C
	5% Diphenyl/95% dimethyl polysiloxane	350 °C
	50% Diphenyl/50% dimethyl polysiloxane	320 °C
	14% Cyanopropylphenyl/86% dimethyl polysiloxane	280 °C
	50% Cyanopropyl methyl/50% phenylmethyl polysiloxane	240 °C
	Polyethylene glycol	250 °C
	Biscyanopropyl polysiloxane	250 °C
Ionic Liquid	SLB-IL59; 1,12-Di(tripropylphosphonium)dodecane bis(trifluoromethanesulfonyl)imide	300 °C
	SLB-IL60; 1,12-Di(tripropylphosphonium)dodecane bis(trifluoromethylsulfonyl)imide	300 °C
	SLB-IL61; 1,12-Di(tripropylphosphonium)dodecane bis(trifluoromethanesulfonyl)imide trifluoromethanesulfonate	290 °C

In comprehensive GC×GC, two different separation mechanisms (with the degree of difference in the chemistries being optimized for particular samples) are required.[129,131] Typically, in the "normal" column combination, a non-polar column is

used in the first dimension, whose separation mechanism is based on dispersive forces. The first dimension column is coupled to a more selective polar second dimension that separates the analytes by another mechanism of intermolecular forces, such as π-π interactions and hydrogen-bonding.[132] The emergence of two-dimensional GC coupled to a fast acquisition TOFMS has enhanced the metabolic space coverage of conventional GC-MS greatly.[58,66] Despite the immense advantages of GC×GC-TOFMS, two-dimensional GC separation in fecal metabolomics has been conducted only in a few studies.[133-135] Aura et al. used a non-polar RTX-5 column (10 m × 0.18 mm × 0.20 μm (Restek Corp., USA)), coupled to a polar BPX-50 column (50% phenyl polysilphenylene-siloxane, 1.50 m × 0.10 mm × 0.10 μm (SGE, Australia)).[134] Shi et al. used a 60 m × 0.25 mm; 0.25 μm df (DB-5ms, phenyl arylene polymer virtually equivalent to a 5%-phenyl)-methylpolysiloxane (RTX-5) and 1 m × 0.25 mm; 0.25 μm df DB17 ms (50%-phenyl)-methylpolysiloxane, virtually equivalent to BPX-50) as a secondary column.[133] Trošt et al. did not specify which column combination they used for the analysis. The two papers mentioned above (Aura et al. and Shi et al.) used the column combination of virtually the same selectivity, with widely varying column dimensions.

Herein, a GC×GC-TOFMS method for fecal/cecal samples was developed, with a particular focus on investigating GC column combinations of different chemistry, in search of the best column combination for global metabolic profiling of fecal/cecal samples. In this work, cecal metabolites were extracted, derivatized, and analyzed by GC×GC-TOFMS using different GC×GC column configurations. A cecum is a pouch within the peritoneum, which is at the beginning of the large intestine.[136] Cecal samples were considered to be similar to fecal samples and treated in the same way as fecal samples. The sample preparation process from sample collection until the instrumental analysis, which includes sample handling, storage, extraction, and derivatization for GC-based fecal metabolomics, has been covered in many works of literature.[31,35,79,82,137-142] To the best of my knowledge, it is the first time that extensive discussion of different column chemistries for GC×GC-TOFMS untargeted metabolomics of fecal sample has been made.

2.4.2 Experimental

2.4.2.1 Chemicals and Materials

HPLC grade methanol (>99.9%, Millipore-Sigma Canada) was mixed with 18.2 MΩ deionized MilliQ water (Elga PURELAB flex 3 system, VWR International Edmonton) in a 4:1 ratio. HPLC grade toluene (Millipore-Sigma Canada), was dried with anhydrous sodium sulfate (Millipore-Sigma Canada) before use. A 20 mg/mL solution of methoxyamine hydrochloride (Sigma Canada) was prepared in HPLC grade pyridine, and 1 mL ampoules of N,O-Bis(trimethylsilyl)trifluoroacetamide + 1% chlorotrimethylsilane (BSTFA + 1% TMCS) were purchased from Fisher Scientific, Canada. Fatty acid methyl esters standard mixture (SUPELCO 37 Component FAME Mix) was purchased from Millipore-Sigma Canada. Samples were treated with heat and nitrogen using a 099A EV2412S Glas-Col Heated Analytical Evaporator (Cole-Parmer Canada) using pre-purified nitrogen (Praxair Canada Inc., Edmonton).

2.4.2.2 Sampling and Sample Storage

The samples were obtained from the University of Alberta laboratory animals directly from their cecum after sacrifice by the collaborator from the department of Agricultural, Food and Nutritional Science, University of Alberta. All animal procedures were conducted in accordance with the guidelines of the Canadian Council on Animal Care (CCAC, 2009). All protocols were approved by the University of Alberta Animal Care and Use Committee for Livestock and conducted at the Swine Research and Technology Center at the University of Alberta (Study ID: AUP00000157). Immediately after collection, the samples were snap frozen in liquid nitrogen and stored at –80 °C. When the samples were received for GC×GC-TOFMS analysis, approximately 250 mg of each wet stool was lyophilized for 72 h. After lyophilization, the freeze-dried feces continued to be stored at –80 °C until the analysis.

2.4.2.3 Metabolite Extraction and Derivatization

The extraction solvent, 80% methanol, that was suggested from the study conducted by Yang et al. was prepared by mixing methanol and deionized water in a 4:1 ratio

(methanol:water 4:1 v/v).[138] Then, 40 mg (±2) of the lyophilized sample was weighed and put into 2 mL Eppendorf tubes, 1 mL of 80% methanol extraction solvent was added and vortexed for 3 min, followed by centrifugation for 15 min at 10,000 rpm. A 500 µL aliquot of the supernatant was transferred to a 2 mL GC vial and dried under nitrogen at 50 °C. When the sample was dried, 100 µL of toluene were added to the sample vial and further dried under nitrogen to remove the residual moisture. To the dried metabolite extracts, 50 µL of 20 mg/mL methoximation reagent were added, and the samples were incubated at 60 ºC for 60 min. Subsequently, 100 µL of BSTFA were added to each sample, and the samples were incubated again at 60 ºC for 30 min. The TMS derivatives were transferred into a GC vial with an insert and were subjected to the GC×GC-TOFMS analysis.

2.4.2.4 GC×GC-TOFMS Experimental Conditions

Table 2-14 shows the details of the four GC×GC-TOFMS methods employed to explore the different GC×GC column combinations. Slightly different GC×GC methods with respect to the temperature programming were used for the different column combinations due to the differences in maximum working temperature of the secondary column.

Table 2-14. GC×GC-TOFMS Conditions of the Tested GC×GC Columns and the Associated Temperature Programs

Column combinations explored	A	B	C	D
1° column geometry	60 m × 0.25 mm × 0.25 µm	30 m × 0.25 mm × 0.25 µm	30 m × 0.25 mm × 0.25 µm	30 m × 0.25 mm × 0.25 µm
1° stationary phase and type	Rtx-5MS (Restek) Crossbond diphenyl dimethyl polysiloxane (350°C)			
2° column geometry	1.3 m × 0.25 mm × 0.25 µm	1.16 m × 0.25 mm × 0.25 µm	1.6 m × 0.25 mm × 0.20 µm	1.7 m × 0.25 mm × 0.20 µm
2° stationary phase, type, max temperature	Rtx-200 (Restek) Crossbond trifluoropropylmethyl polysiloxane (340 °C)	Stabilwax (Restek) Polyethylene glycol (260 °C)	SLB-IL111 (Supelco) Non-bonded; 1,5-Di(2,3-dimethylimidazolium)pentane bis(trifluoromethanesulfonyl)imide (270 °C)	SLB-IL59 (Supelco) Non-bonded; 1,12-Di(tripropylphosphonium)dodecane bis(trifluoromethanesulfonyl)imide (300 °C)
GC×GC method	70 °C (5 min), ramp at 5 °C/min to 300 °C (10 min) 2 ° oven: 5 °C offset to the GC oven Modulator: 15 °C offset to the 2 ° oven Transfer line: 240 °C Modulation: 2.5 s (0.6 hot, 0.65 cold)	70 °C (5 min), ramp at 5 °C/min to 250 °C (10 min) 2 ° oven: 5 °C offset to the GC oven Modulator: 10 °C offset to the 2 ° oven Transfer line: 240 °C Modulation: 2.0 s (0.4 hot, 0.6 cold)	70 °C (5 min), ramp at 5 °C/min to 245 °C (15 min) 2 ° oven: 10 °C offset to the GC oven Modulator: 15 °C offset to the 2 ° oven Transfer line: 240 °C Modulation: 2.0 s (0.4 hot, 0.6 cold)	70 °C (5 min), ramp at 9.7 °C/min to 280 °C (15 min) 2 ° oven: 10 °C offset to the GC oven Modulator: 10 °C offset to the 2 ° oven Transfer line: 270 °C Modulation: 2.0 s (0.4 hot, 0.6 cold)
MS method	m/z 25-900 acquisition rate: 200 spectra/second optimized voltage offset: 200 electron energy: -70 eV ion source: 200 °C			
Total Analysis Time	43.7 minutes	61 minutes	55 minutes	41.6 minutes

2.4.2.5 GC×GC-TOFMS Final Analysis Condition

A Leco Pegasus 4D GC×GC-TOFMS (Leco Instruments, St. Joseph, MI) with a 30 m × 0.25 mm; 0.25 µm df Rtx-5MS (Chromatographic Specialties) column as the first dimension and a 1.7 m × 0.25 mm × 0.20 µm SLB-IL59 (Supelco) column as the second dimension was used. Helium (5.0 grade; Praxair, Edmonton, AB) was used as the carrier gas at a constant flow of 2 mL/min. A 1 µL aliquot of the sample was injected in both the splitless (inlet purge time of 80 s) and split (1:40) mode, with an inlet temperature set at 270 °C. The primary GC oven was programmed from 70 °C, held for 5 min, and then ramped at 9.7 °C/min to 280 °C, with a final hold of 15 min, which resulted in the total analysis time for each run to be 41.65 min. The secondary oven was programmed to have a constant offset of +10 ºC relative to the primary oven, and the modulator was set at a constant offset of +10 ºC relative to the secondary oven temperature. The modulation period was 2.0 s, with 0.4 s hot pulse and 0.6 s cold time. The MS transfer line temperature was set at 270 °C. The parameters used for mass spectrometry are as follows: electron energy of –70 eV; acquisition rate of 200 Hz; mass range of *m/z* 25–900; detector voltage with the optimized voltage offset of 200 V; ion source temperature of 200 ºC; the solvent delay time of 300 s.

2.4.2.6 Data Processing

All GC×GC-TOFMS data were processed using ChromaTOF® (v.4.72; Leco) equipped with the US National Institutes of Standards and Technology (NIST) MS database (NIST 2011) and Wiley 08. The baseline offset was set to 0.9 above the middle of the noise. The data were processed with a peak finding threshold of S/N 100:1, and the minimum S/N ratio for sub-peaks to be retained was set at 6, and the mass spectral match required for the subpeaks to be combined was set at 650. The peak widths for peak picking criteria was set to 12 s in the first dimension and 0.15 s in the second dimension.

2.4.3 Results and Discussion

In the present work, all the samples were stored and prepared following a well-established method from the papers previously published discussing sample

preparation for feces in detail.[141,143] The lyophilized cecal samples were extracted, derivatized, and analyzed using different column combinations. As the first column combination to explore, 60 m × 0.25 mm; 0.25 μm df (Rtx-5MS, Restek) was selected for the first dimension, and 1.3 m × 0.25 mm; 0.25 μm df (Rtx-200, Restek) was used in the second dimension. In the GC×GC separation technique, the first column is much longer than the second column and is where most of the separation is performed. The second column, helps to separate the coeluting compounds from the first dimension; however, it is kept short to retain the separation produced by the primary column. Therefore, the first column selection is critical. Rtx 5MS is 5% diphenyl dimethylpolysiloxane. It is a low polarity general-purpose column and is suitable for analyzing semi-volatiles, including phenols, amines, PAH, PCBs, etc. For the primary column, the diphenyl stationary phase was selected for its low polarity. Typically, a 30 m column is used for the first dimension, but the longer length, 60 m, was used to increase the theoretical plate, as it was suggested in the study of dioxins and PCBs by Focant et al. that a 60 m column was used to alleviate most of the coelution problems in the first dimension in a highly complex sample.[144] For the secondary column, Rtx-200 stationary phase, which is a mid-polarity phase with crossbonded trifluoropropylmethyl polysiloxane was used. This stationary phase is highly selective for analytes containing lone electron pairs, such as nitrogen and carbonyl groups. It is an excellent column for environmental applications such as organochlorine pesticide analysis.[145] The speed-optimized flow of 2 mL/ min constant flow, and the optimal heating rate of 9.7 °C (10 °C/t_m) were used. An example chromatogram of a cecal sample using the column combination is shown in Figure 2-22A. A total of 2756 compounds were detected, which include amino acids, fatty acids and FAMEs. The separation along the first dimension was satisfactory, but improved separation in the second dimension would be recommended because the second dimension was used ineffectively. Nonetheless, it is a large gain compared to the result from Trošt et al. that 107 metabolites were identified using GC-TOFMS.[135]

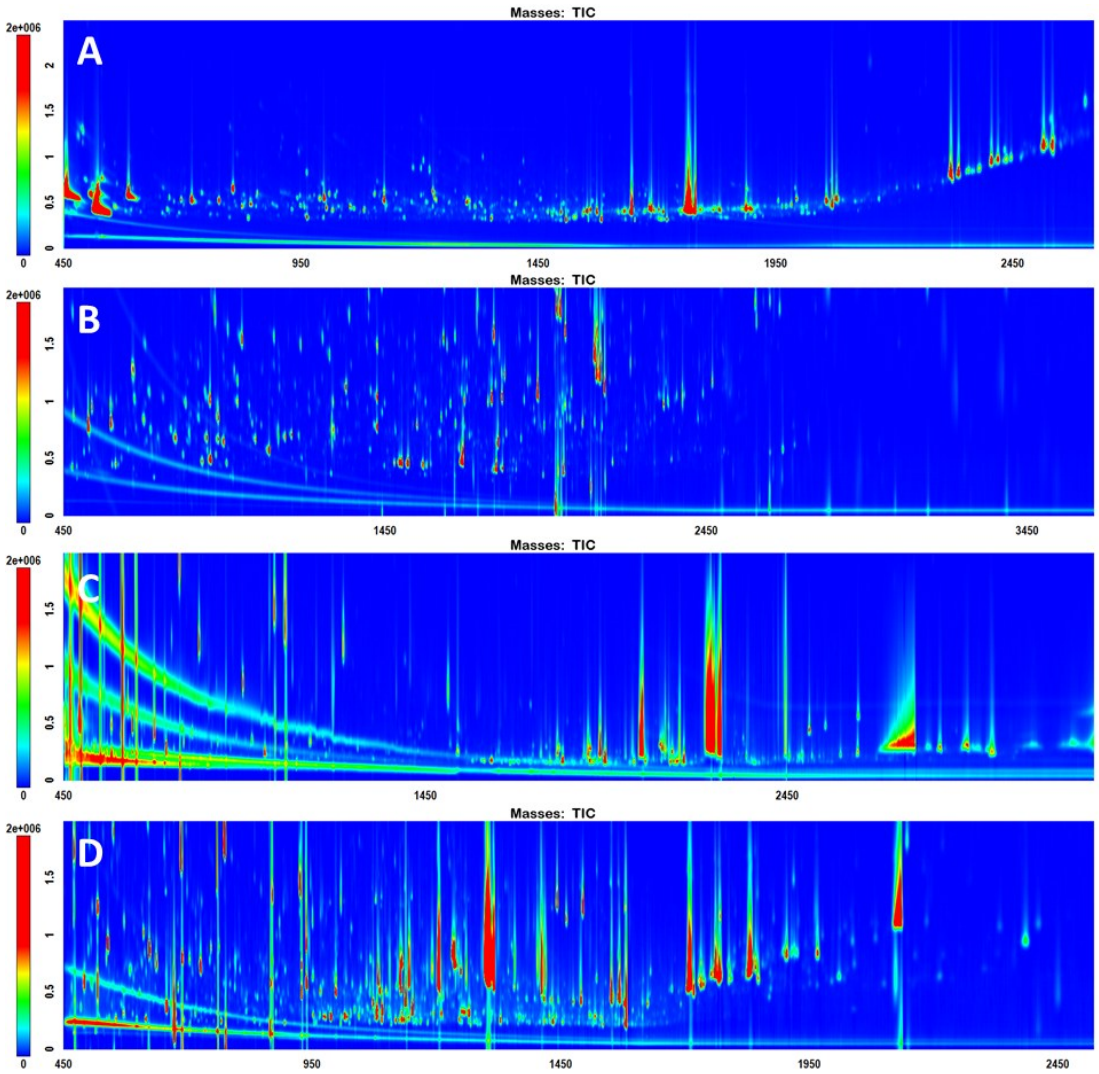

Figure 2-22. Chromatograms resulted from four different GC×GC-TOFMS conditions in described in Table 2-14.

It is true that a longer column offers higher column efficiency; however, as a trade-off, it costs more in analysis time, and the price of the column increases as well as the cost associated with the extended analysis time. For the next column combination to explore, a 30 m Rtx-5MS in the first column was considered, with PEG phase in the second dimension. The first dimension was shortened to half, which is a more commonly used size for the first dimension. The PEG was chosen as the second dimension column to attempt. PEG is a highly polar stationary phase that is used widely for analyzing alcohols, aldehydes, free fatty acids, fatty acid methyl esters, and aromatics. The resulted chromatogram (Figure 2-22B) displayed an excellent

separation in both the first and second dimensions. However, with this column set, a greater retention time shift than usual was observed from sample to sample. With more literature search, it was found that the TMS derivatized sample is not compatible with PEG columns because the excess TMS derivatization reagent would interact with PEG columns and result in irreversible damage to the stationary phase.[146] This would impact the column performance and result in poor reproducibility. Thus, despite the excellent separation gained, a PEG phase cannot be considered as an option for TMS-derivatized samples.

In a place of the highly polar PEG column, using polar ionic liquid columns as the second-dimension was evaluated. Two ionic liquid columns (SLB-IL111 and SLB-IL59) were investigated. SLB-IL111 (Non-bonded; 1,5-Di(2,3-dimethylimidazolium) pentane bis(trifluoromethylsulfonyl)imide) is an extremely polar column that ranks the highest out of all the ionic liquid columns, according to the GC column polarity scale from Supelco.[45] SLB-IL111 is specified to be a great column choice for the analysis of fatty acid methyl esters (FAMEs), one of the major classes of interest in fecal metabolomics.[147] The high polarity of the SLB-IL 111 phase is useful to resolve cis/trans isomers of FAMEs.[148] This column was selected as the second dimension column coupled to 30 m 5% diphenyl/95% dimethyl polysiloxane phase (the same first dimension column that was used with PEG) to analyze the cecal derivatized samples. The result presented a poor chromatographic performance, displaying massive solvent bleed in the beginning region of the chromatogram and poor peak shapes (Figure 2-22C).

Another available ionic liquid stationary phase was SLB-IL59. It has a polarity similar to a PEG phase, according to the polarity scale that compares the relative polarity of the ionic liquid phases to classical GC phases. SLB-IL59 has improved thermal stability, with a maximum temperature of 300 °C, compared to a PEG column of 250 °C. The increased maximum temperature of the ionic liquid column can be beneficial for bulky semi-volatile compounds, such as tocopherols and sterols. The resulted chromatogram using SLB-IL59 as the second dimension column achieved satisfactory separation for derivatized fecal metabolites, although tailing peak shapes

for some overloaded peaks were present. The 37 components of FAMEs standard mixture also were analyzed using the column combination to evaluate the performance on the separation of isomers. The column set was successful in separating all 37 compounds of the standard mixture (Figure 2-23). The excellent performance of the selected column set for FAMEs suggests that the given column combination is promising for fecal metabolomics studies since FAME is a critical family group for fecal metabolomes.[147,149]

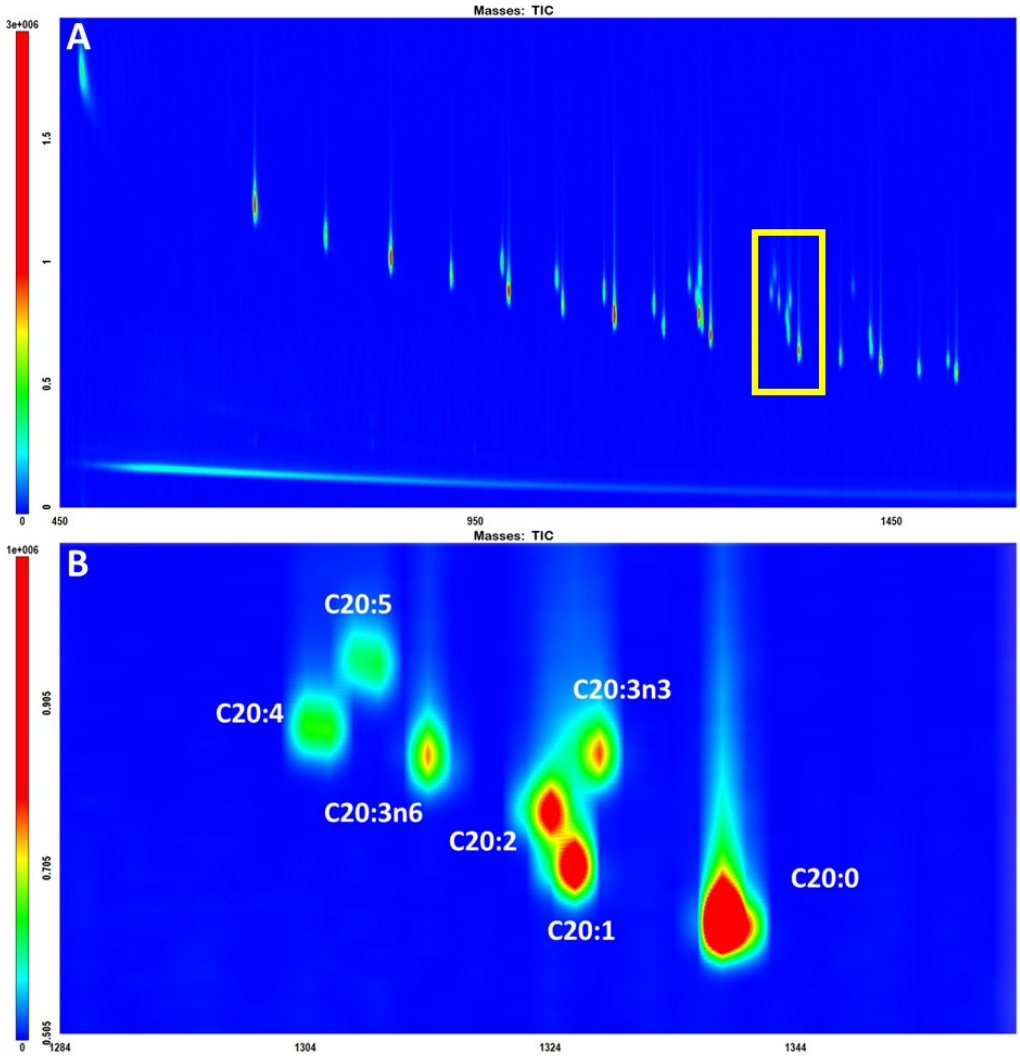

Figure 2-23. A) chromatogram obtained by the separation of 37 FAMEs standard mixture on Rtx-5 column × IL-59 column combination B) zoomed-in of C20 FAMEs (yellow box of A).

The global GC-TOFMS analysis that Phua et al. reported used 80 mg of lyophilized feces with a split ratio of 1:10.[141] While the remaining sample preparation method was kept similar to the process that was used in the study, 268 chromatographic peaks were detected, and 107 peaks were putatively identified with a conventional GC-TOFMS. For the current GC×GC-TOFMS study, half of the sample weight (40 mg of lyophilized feces) was used with splitless injection. From a cecal sample that was analyzed with SLB-IL59 as the second dimension column, more than 2000 peaks were detected, with a signal-to-noise ratio greater than 100 (Figure 2-22D). The analytes were putatively identified as belonging to various families by matching the MS spectra with the NIST library. Table 2-15 contains the subset of the entire compound list, showing the predominant groups identified, which includes TMS derivatives of fatty acids, FAMEs, TMS derivatives of carbohydrates and amino acids.

Table 2-15. Partial List of Tentatively Identified Compounds from the Pig Cecal Using GC×GC-TOFMS (Rtx-5 × SLB-IL59)

	Name	R.T.(s)	Area	Similarity	S/N	Exp RI	Library RI	$\lvert Ri_{exp}-Ri_{Lib}\rvert$
Carbohydrates	D-(+)-Xylose	962 , 1.230	471345	869	877	1693	1738	45
	D-(-)-Arabinose	964 , 0.255	111791	889	157	1696	1748	52
	D-(-)-Ribose	968 , 0.235	17282777	854	17945	1702	1651	51
	D-(-)-Lyxose	978 , 0.240	17214047	841	19935	1718	1651	67
	D-(-)-Fructose	1106 , 0.220	9269122	874	15519	1906	1875	31
	D-(+)-Glucose	1120 , 0.230	635238	926	163	1923	1970	47
	D-(+)-Galactose	1130 , 0.225	1907552	915	4372	1936	1970	34
	D-(+)-Turanose	1514 , 0.190	11872422	822	4925	2541	2693	153
	D-(+)-Maltose	1558 , 0.435	106059	861	815	2595	2693	98
	Sucrose	1578 , 0.185	266886	833	1639	2616	2712	96
FAMEs	Tetradecanoic acid,methyll ester	990 , 0.575	120116	838	358	1731	1725	6
	Pentadecanoic acid, methyl ester	1056 , 0.540	568256	882	2009	1832	1820	12
	Hexadecanoic acid, methyl ester	1118 , 0.520	15905774	683	15934	1927	1920	7
	Heptadecanoic acid, methyl ester	1178 , 0.485	262241	789	959	2013	2006	7
	Octadecanoic acid, methyl ester	1234 , 0.465	1638010	906	5868	2103	2123	20
	Eicosanoic acid, methyl ester	1340 , 0.425	111526	751	500	2275	2307	32

	9-Octadecenoic acid (Z)-, methyl ester	1220 , 0.515	3480121	721	3883	2081	2085	4
	cis-11-Eicosenoic acid, methyl ester	1328 , 0.465	120486	849	423	2254	n/a	n/a
Fatty acid TMS	Pentanoic acid	314 , 0.965	2430611	863	1502	998	974	24
	Hexanoic acid	428 , 0.770	2700227	940	7913	1082	1073	9
	Heptanoic acid	534 , 0.650	268901	813	487	1175	1093	82
	Octanoic acid	630 , 0.585	108288	908	376	1271	1265	6
	Nonanoic acid	718 , 0.530	77329	786	544	1370	1369	1
	Decanoic acid	798 , 0.495	260088	826	361	1468	1460	8
	Undecanoic acid	872 , 0.465	10162	586	55	1563	1490	73
	Dodecanoic acid	942 , 0.460	259565	811	965	1653	1657	4
	Tridecanoic acid	1008 , 0.430	23817	811	149	1755	1755	0
	Tetradecanoic acid	1072 , 0.415	575341	880	855	1819	1851	32
	Pentadecanoic acid	1132 , 0.420	2471184	711	1763	1921	1943	22
	Hexadecanoic acid	1190 , 0.475	26948319	895	10536	2034	2046	12
	Octadecanoic acid	1298 , 0.550	9067698	805	4308	2197	2243	46
	(9E)-Octadecenoic acid	1288 , 0.785	84565	878	448	2185	2218	33
Sterol & Tocopherol	Coprostan-3-ol	1682 , 0.410	12625614	771	3797	2791	2810	19
	Cholesterol	1738 , 0.490	5278223	791	3319	2897	3145	248
	Campesterol	1810 , 0.530	1638479	826	2289	3000	3253	253
	á-Sitosterol	1882 , 0.595	5014310	755	1915	3103	3344	241
	ç-Tocopherol	1650 , 0.375	545848	771	2919	2760	2900	140
Amino acids	Isoleucine	666 , 0.525	10393323	816	25743	1309	1299	11
	Glycine	674 , 0.415	438081	859	2583	1319	1315	4
	Serine	728 , 0.420	1881279	869	6141	1381	1370	11
	Threonine	750 , 0.365	6157255	882	6426	1408	1393	15
	Alanine	778 , 0.420	880197	846	3189	1443	1438	5
	Methionine	852 , 0.635	4500522	865	9112	1537	1514	23
	Phenylalanine	932 , 0.675	6960550	842	6850	1642	1636	6
	Lysine	1126 , 0.315	1731170	805	6816	1938	1933	4
Other compounds	Glycolic acid	436 , 1.210	14659263	906	6912	1088	1080	8
	Oxalic acid	464 , 1.035	11029942	821	12377	1111	1122	11
	Methylmalonic acid	496 , 0.785	45853	706	175	1140	1210	70
	p-Cresol	514 , 0.865	26530	827	4	1156	1137	19
	Lactic acid	532 , 0.605	2642334	855	2070	1173	1086	87
	Benzoic acid	612 , 1.370	10709379	909	5525	1252	1248	4
	Catechol	682 , 0.565	1829498	839	8153	1328	1330	2
	Creatinine	880 , 0.610	17605513	817	15979	1574	1548	25
	Adenosine	1502 , 0.425	628756	884	4059	2529	2626	97

Figure 2-24 displays a region in a chromatogram where the power of GC×GC is well represented. The mass channel 73 is a TMS fragment and is common to all TMS derivatives. The TMS derivatives of 2-oleoylglycerol and D-lactose were coeluting from the first dimension and would not have been separated without help from the second dimension. In addition, the TMS derivative of octadecanoic acid, which is present at low abundance, could be obscured by a more abundant maltose peak in 1D GC and not be well-resolved without the second dimension.

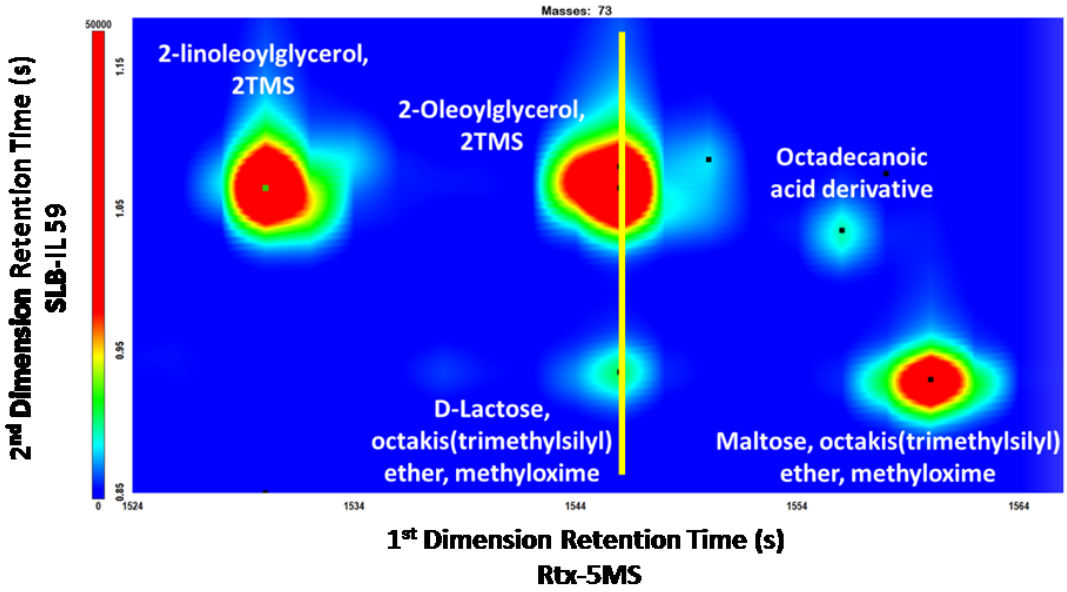

Figure 2-24. A zoomed-in GC×GC-TOMFS chromatogram showing the separation power.

Due to the broad concentration range of the metabolites in a cecal sample, considerable variations in peak size were observed (Figure 2-22D). While splitless injection could be adequate for low abundance analytes, the data analysis struggled for high abundant peaks. The high concentration compounds overloaded the column, which led to tailing peak shapes, and the poor chromatographic peak shape limited the software to integrate the peaks correctly. Multiple peak entries were generated for a single analyte and consequently resulted in the wrong integration of peak areas. To address the problem associated with overloaded peaks, the cecal sample was injected with a split ratio of 1:40 (Figure 2-25). The split injection brought massive peaks, such as carbohydrates, fatty acids, and sterols into reasonable sizes, with good Gaussian peak

shape in both the first and second dimensions. Reproducibility was assessed by analyzing the two different cecal samples, each in duplicates, with a split injection. The two replicates of the same sample, overlayed in 1D, displayed excellent analytical performance (Figure 2-25A). For both samples, the total number of peaks detected and the total peak area, which is the summation of the peak areas of all compounds, were compared. The total peak area (TPA) was normalized by the respective sample weight (Table 2-16). The percent difference between replicates for the TPA was 0.4% and 10.7% for sample A and sample B, respectively. The percent difference between replicates for the TPA normalized by weight was 9.7% and 3.5% for sample A and sample B, respectively. For both samples, adequate reproducibility was achieved between replicates (<10%) for the normalized total peak area. While splitless injection is useful for capturing all metabolites as possible, including trace analytes, the additional split injection can complement the splitless injection to bring high abundant compounds in reasonable peak shape and size for more accurate quantification.

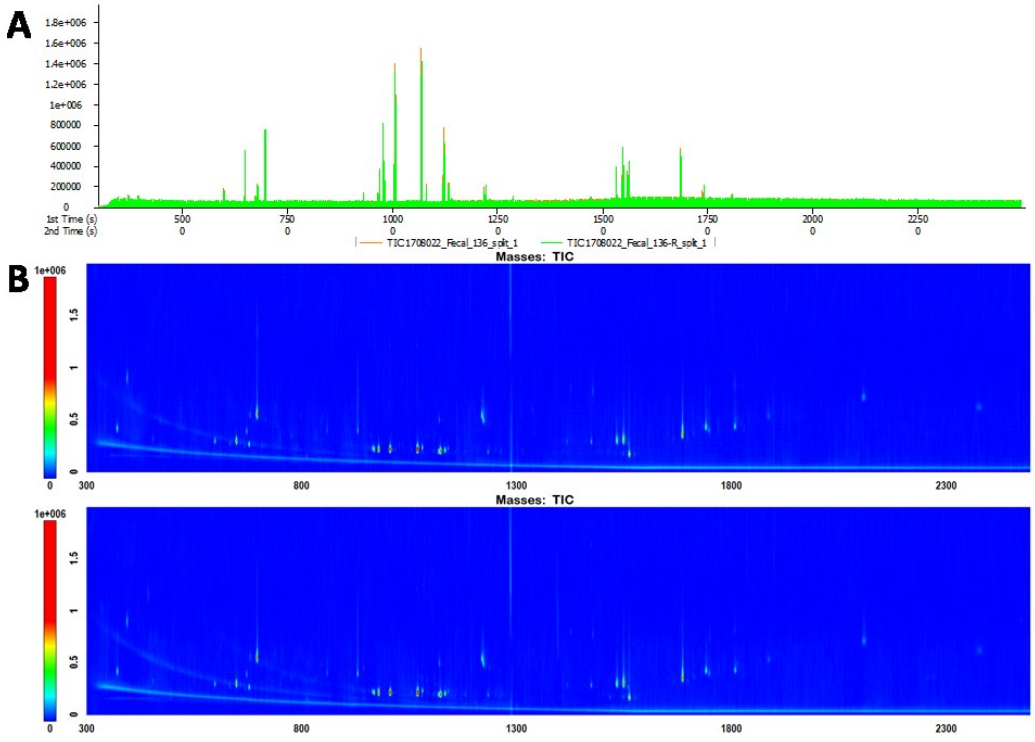

Figure 2-25. The total ion current chromatogram of a cecal sample analyzed on GC×GC-TOFMS with a split ratio of 1:40 A) Overlay of the two replicates in 1D B) 2D contour plots of the two replicates.

Table 2-16. Comparison of Two Cecal Samples in Duplicates

	Sample A Replicate 1	Sample A Replicate 2	Sample B Replicate 1	Sample B Replicate 2
Dry Weight of the sample (mg)	39.5	43.4	43.2	39.8
Number of peaks detected	578	753	724	574
Total peak area	5.67E+07	5.65E+07	8.58E+07	7.66E+07
Normalized TPA by weight	1.44E+06	1.30E+06	1.99E+06	1.92E+06

2.4.4 Conclusions

GC×GC-TOFMS is a powerful tool for global metabolomics profiling of fecal material to study the interactions between the gut microbiome and the host health. The 2D GC allows a high level of separation that is not possible with conventional one-dimensional gas chromatography. To exploit the advantages of GC×GC, the maximum use of the chromatographic space is critical; this can be achieved by having the right choice of column combination. In this work, the GC×GC-TOFMS methodology was developed for fecal/cecal samples, and the results of exploring different column combinations, including relatively new ionic liquid stationary phases, were discussed. Of the tested columns, the combination of the traditional Rtx-5MS in the first dimension and the SLB-IL59 ionic liquid phase in the second dimension was determined to be the column set that utilizes the advantages of the 2D GC fully with the increased thermal stability from the ionic liquid column while using the chromatographic space efficiently. The use of the column set allowed the detection of over 2000 compounds in a cecal sample. The method was enhanced further by the split injection complementing the splitless injection for high abundance and overloaded compounds, which resulted in a good reproducible analytical performance.

CHAPTER 3

Improving Normalization Methods

3.1 Introduction

Metabolomics studies strive for comprehensive monitoring of metabolites in biological systems.[80] While metabolomics involves measurements of hundreds to thousands of metabolites at various concentrations in biosamples, there exist analytical challenges arising from sources of variations and potential biases at every step involved in metabolomics studies.[11] Each step of the metabolomics workflow (sample collection, preparation, instrumental analysis, and data analysis) is susceptible to introducing variability into the data.[150] To achieve reliable conclusions from metabolomics studies, it is crucial to understand the factors that introduce variability, and adequate normalization is essential to minimize unwanted variations.[97] One of the major sources of variation in the metabolomics pipeline is related directly to the sample concentration and the biological variability existing in the sample.[77] The natural variability of metabolomics samples due to the combined genetic and environmental factors adds an extra layer of complexity to metabolomics studies. In addition to differences between the samples from different individuals, the sample compositions and concentrations of the same individual may vary over time due to external factors, such as diet and hydration level. Therefore, it is important to follow the controlled procedures for a range of pre-analytical factors, including sample collection, storage, and preparation, to reduce the variability.[79] Moreover, proper data normalization that can adequately address both biological and technical variations must be applied.[150]

3.1.1 Biological Variability in Urine

Of the various biological matrices, urine has been a popular matrix for metabolomics studies due to its rich metabolite content, and ability to non-invasively collect it in large volumes. Due to these advantages, it has been reported as a precious clinical sample for early non-invasive disease diagnosis.[151-154] Despite the convenience of urine as a diagnostic biofluid, there exists a major drawback for its use as an analytical sample:

urine concentration may vary significantly, based on the level of hydration, and also may be influenced by external factors, such as diet, excretion via sweating, or kidney disease.[155] For urine samples, it has been reported that the total concentration of urinary metabolites can vary by more than 15-fold between samples.[151] The fluctuations of urine sample concentration cause significant challenges for data interpretation and biomarker discovery because the random biological variation could mask any biological variations due to disease.[37,156] Consequently, normalization is critical in metabolomics studies to correct for different sample concentrations and must be incorporated into a metabolomics workflow.[97,156] Once the sample-to-sample variation of urine is considered and addressed adequately, then truly significant and relevant metabolomic changes of interest can be observed.[157] If the metabolomics data is not treated with an appropriate normalization technique, the result can lead to an incorrect conclusion in comparative studies.[97]

3.1.2 Normalization Methods for Urinary Metabolomics

Sample normalization is a strategy to correct the sample-to-sample variability, either by adjusting the sample volume or the concentration prior to data acquisition or by adjusting acquired signals after data acquisition.[77,150] Various normalization strategies have been reported to account for fluctuations in urine concentrations.[98,158,159]

3.1.2.1 Normalization to Creatinine

The most common method is to use creatinine concentration as a normalization factor.[158,160,161] Creatinine is a waste product of muscle metabolism; in the absence of kidney malfunction, the rate of urinary creatinine excretion is relatively constant between different individuals and within an individual over time.[77,155,162-164] Using creatinine as a normalization factor has been considered the gold standard for urinary metabolomics, with the underlying assumption that creatinine concentration reflects the urine sample concentration.[161,165] Creatinine has become an important biomarker for kidney related disease, and the concentrations of endogenous metabolites are commonly expressed as ratios relative to creatinine.[161,162]

However, it has been reported that creatinine excretion is highly dependent on many factors, and thus it may not provide a reliable reference for sample normalization.[77,78,151,155,161,166] The variation of creatinine excretion in human subjects may occur due to differences in age, sex, ethnicity, level of physical activity, muscle mass, hydration, diet, diurnal rhythms, emotional stress, disease state, body mass, etc. Recent publications suggest that the use of creatinine as a single reference compound to normalize the wide concentration ranges of all urinary metabolites present in the sample is not advisable.[77,78,155,167] For many other biological sample types that suffer from a wide range of sample concentration variabilities, such as feces and sweat, there is no known compound yet that is commonly accepted as a reference compound like creatinine in urinary metabolomics.[31]

3.1.2.2 Other Normalization Methods

Another method is the use of osmolality, a measure of total urinary solute concentration, as a normalization factor.[157,159] The procedure of measuring osmolality; however, is not straightforward and requires a separate analysis to measure freezing-point depression.[158] Another well-known approach for global urinary metabolic profiling is the "constant sum normalization," which is to use the sum of all signals from metabolites detected in a sample as a normalization factor.[105,107] As a post-acquisition normalization technique, this carries a great advantage of not requiring any additional experimental procedures.[77] However, the major drawback of this approach is that the number of compounds detected in each sample can vary widely from one sample to the next. Therefore, using the total signal of all the metabolites as a normalization factor may not be an accurate reflection of the urine concentration. Thus, "MS Total Useful Signal" (MSTUS) as a normalization factor, has been proposed and is routinely used in multiple MS-based metabolomics studies coupled to liquid chromatography (LC).[78,168,169] This method involves computational data processing to identify component mass spectral signals, associate related ions into molecular components, and then sum all peaks that are common to all samples to obtain the MSTUS normalization factor.[78] By incorporating only "useful signals" that are common to all

samples, contributions from xenobiotics and artifacts can be diminished.[151] From several non-targeted metabolomic profiling studies, it has been reported that MSTUS normalization yielded better results when compared to other common normalization approaches.[77,78] The limiting factor of this approach is that it can be applied only to MS data, and the method involves a proprietary algorithm to calculate the normalization factor.[77,78,151]

Some metabolomics studies perform normalization using an internal standard, a traditional analytical approach for normalization.[97] For untargeted metabolomics studies, due to differences in extraction and derivatization efficiencies of chemically diverse metabolites, normalization using multiple internal standards (commonly one or two per class of compounds) has been employed.[103,107] However, even compounds within the same homologous series may exhibit different efficiencies, and ideally, an internal standard (typically isotope-labelled) for each metabolite is preferred.[77] Given the complexity of biological samples, it is not possible to add stable isotope-labelled internal standards for all detected compounds.[100] For LC-based metabolomics, Wu et al. developed a normalization method through chemical labelling of metabolites across the sample.[170] A relatively simple derivatization method of dansylation was used to attach a UV absorbing dansyl moiety to amines and phenols.[170] The dansyl moiety is a good chromophore that unifies the absorptivity of labeled metabolites, and UV absorption can be measured at 338 nm.[170] A rapid step-gradient is used to allow co-elution of all labelled metabolites, and the dansylated metabolites can be measured.[77,170] Normalization using dansylation was developed further to be used in LC-MS metabolome profiling by derivatizing the pooled sample and using the chemically labeled metabolites as the internal standards for quantification of a broad range of metabolites.[77,170] However, this method is applicable only with LC-MS analyses.

An analogous idea of using chemical labelling to obtain labelled internal standards across the metabolome is not currently feasible for GC.[100,171] Lien et al. attempted to carry out similar work by derivatizing a mixture of standards with a deuterated derivatization reagent, d9-MSTFA, to synthesize the labelled internal standards.[111] An equal volume of d9-MSTFA derivatized standard solution was spiked

into the MSTFA-derivatized sample prior to GC analysis. The deuterated counterparts of individual metabolites enabled the individual correction of the metabolites by normalizing the responses of the TMS derivatives by the responses of their d9-TMS counterparts.[111] From this work, Lien et al. claimed that the individual correction improved the analytical performance more than the group correction, which refers to normalizing responses of a compound class by using the response of one representative labelled compound belonging to that class.[111] In addition to the quantitative improvement, this d9-MSTFA also can enhance compound identification by increasing the molecular mass and the fragment ions by nine mass units for each derivatized functional group. Although good performance results were reported, this method of using a labelled derivatization reagent for GC is practically impossible due to the incredibly high cost. Aside from the fact that this product not currently commercially available, the last-checked price for d9-MSTFA from the Sigma-Aldrich website was approximately $3000 CAD for 500 μL, which can cover only five samples.

Although the importance of sample normalization has been well-received and recognized by the metabolomics community, to date there is no standardized method for the normalization of urine concentration.[77,172] Unlike other omics work, sample normalization in metabolomics is much more complicated because of the greater diversity of chemical structures and concentration ranges. Nonetheless, normalization is an indispensable step before any further advanced data analysis to ensure that the data being mined is an accurate reflection of the relative concentrations of the samples so that true metabolomic variances of interest can be found. Failure to correct for such variation may lead to bias in the data that further results in misclassification and false interpretation of results.[159] Despite the awareness of the necessity of sample normalization, it has been ignored in many metabolomics studies due to the lack of a suitable means of performing sample normalization.

3.1.2.3 Scope of the Chapter

Herein, three sample concentration normalization methods are discussed with a particular focus on urine. Urine samples from two different groups, females and males,

were analyzed by two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC×GC-TOFMS). The acquired data were normalized by the three different methods: 1) creatinine, a conventional method of urine normalization, 2) Total Peak Area (TPA), the sum of the areas of all detected peaks, and 3) "Total Useful Peak Area" (TUPA), a strategy that is being proposed as a modification to MSTUS. The proposed TUPA uses the sum of chromatographic peak areas (instead of MS signals) of the peaks that are present in all samples as the normalization factor in GC×GC-TOFMS data. The greatest advantage of this method is that it can be used more universally with any chromatographic separation technique, regardless of the detector type. Contrary to MSTUS, TUPA does not mandate the use of an expensive MS instrument nor sophisticated computational data processing. While it is common in metabolomics studies to have MS detection, TUPA could be applied equally to normalize GC×GC-FID signals, LC-fluorescence, or any other chromatographic data. TUPA only requires a table of analyte identifiers (retention coordinates) and response data (peak areas) from a set of samples analyzed and processed with a consistent method. The performance of the three different normalization methods was evaluated by a manual review of the data after normalization, along with principal component analysis (PCA). The aim of this work is to examine the impact of normalization methods on the result of chemometrics analysis and to propose a reliable sample normalization method for chromatographic data in metabolomics studies.

Due to a large amount of metabolomics data, the use of chemometrics tools to simplify the data analysis has been common. Although it is accepted widely that normalization is essential before applying data reduction techniques, to date, there is no unified standard normalization technique. Many different normalization methods have been discussed in the literature; however, the best and most appropriate method remains elusive, and the normalization process still perplexes researches. Only after proper data treatment can statistical tools be used to find discriminative features.

3.2 Experimental

3.2.1 Subjects and Sample Collection

Two urine samples from each subject were collected for this study from a group of healthy volunteers (i.e., volunteers without any known preconditions). The instructions given to each participant were to collect the first urine of the day in the morning (M) to obtain highly concentrated urine and another sample during the same evening (E) with enough hydration to obtain relatively less-concentrated urine from the same individual. A total of 24 urine samples from males and 30 urine samples from females were collected from the 12 male and 15 female participants for the study. The instructions for the study participants also included directives to store the sample approximately at 4 °C (in a refrigerator or in an ice bag without freezing the sample) upon collection and to bring the sample to the lab within at most 12 h. Upon receiving the sample, aliquots were taken to store the samples in smaller portions to avoid excessive freeze/thaw cycles followed by storage at −80 °C until sample preparation. A pooled QC sample was made by taking equal aliquots (500 µL) from each sample to check for analytical variability. All samples were analyzed in random order with adequate blanks, replicates and quality control (QC) samples. Informed consent was obtained from all individual participants included in the study. The study was approved by the University of Alberta Research Ethics Board (Study ID: Pro00071285).

3.2.2 Chemicals, Reagents, and Solvents

Urease (Millipore-Sigma Canada) suspensions of approximately 40 mg per 250 µL of 18.2 MΩ deionized water (Elga PURELAB flex 3 system, VWR International, Edmonton) were prepared on the day of derivatization. HPLC grade methanol (>99.9%) was purchased from Millipore-Sigma Canada. HPLC grade toluene (Millipore-Sigma Canada) was dried with anhydrous sodium sulfate (Millipore-Sigma Canada) prior to use. Methoxyamine hydrochloride (Millipore Sigma Canada) was dissolved in HPLC grade pyridine (Millipore-Sigma Canada) to make a 20 mg/mL solution. 1 mL ampoules of N-methyl-N-trimethylsilyltrifluoroacetamide + 1% chlorotrimethylsilane (MSTFA + 1% TMCS) were purchased from Fisher Scientific, Canada. 100 µg/mL

solution of 4-^{13}C methylmalonic acid was prepared in deionized water to use as an internal standard.

3.2.3 Sample Preparation

All urine samples were prepared according to a modified protocol for global derivatization of urinary metabolites, which involves a typical two-step derivatization process of methoximation followed by subsequent trimethylsilylation.[28,107] The frozen urine samples were thawed on ice on the day of analysis, then vortexed for 1 min. Then, 40 µL of each urine sample was transferred into a 2-mL centrifuge tube. Also, 20 µL of internal standard was added along with 10 µL of urease suspension. Samples were vortexed for three minutes then incubated at 37 °C for 1 h. Next, 960 µL of methanol was added to each of the samples. Samples were vortexed again for 5 minutes, then centrifuged for 10 min at 10,000× g and 4 °C. Then, 500 µL of the supernatant was transferred to a 2-mL GC vial, and dried under a gentle stream of nitrogen at 50 °C until dry. Next, 50 µL of 20 mg/mL methoxyamine hydrochloride in pyridine was added to the dried residue and incubated at 60 °C for 2 h. Subsequently, 100 µL of MSTFA was added and incubated again at 60 °C for 1 h, and 100 µL of the derivatized metabolite extract was transferred to GC vials with 300-µL inserts for analysis by GC×GC-TOFMS.

3.2.4 GC×GC-TOFMS Conditions

All GC×GC-TOFMS analyses were performed on a LECO Pegasus 4D system (Leco Instruments, St. Joseph, MI) equipped with a four-jet dual-stage modulator. A 60 m × 0.25 mm; 0.25 µm df Rxi-5SilMS was used as the first-dimension column and 1.6 m × 0.25 mm; 0.25 µm df Rtx-200MS as the second-dimension column (Chromatographic Specialties). Two-dimensional chromatographic separations were performed with a constant flow rate of 2.0 mL/min using helium as the carrier gas and a modulation period of 2.5 s. A GERSTEL MPS autosampler was used for automated injection of 1 µL aliquots of sample. The oven initially was held at 80 °C for 4 min and heated to 315 °C at a ramping rate of 3.5 °C/min. The final temperature was held for 10 min. The secondary oven and modulator temperature offset were constant at +10 °C relative to

the GC oven temperature and +15 °C relative to the secondary oven temperature, respectively. Mass spectra were collected at an acquisition rate of 200 Hz over a mass range between 40 and 800 *m/z*. A relative voltage offset of 200 V was selected as the optimized detector voltage with an electron impact energy of –70 eV. The ion source temperature was 200 °C, with a transfer line temperature of 250 °C. Total analysis time for each run was 81.1 min.

3.2.5 Data Processing and Analysis

Data were processed using ChromaTOF® (v.4.72; LECO). Each of the 54 urine samples was processed with two different data processing methods, one for general purpose data processing and one for targeted integration of the creatinine peak. Omics data sets exhibit differences in analyte concentrations that span several orders of magnitude within a sample. The general data processing method was developed to detect as many compounds as possible while minimizing artifacts such as split peaks and spurious signals, while the creatinine method was optimized to integrate specifically only the creatinine peak.

For the general data processing method, the baseline offset was set to 0.9, and the expected peak widths throughout the entire chromatographic run were set to 10 s for the first dimension and 0.15 s for the second dimension. The peak-finding threshold of S/N was set to 100:1, with the minimum S/N ratio for sub-peaks to be retained set at 10. A chromatographic region of 3530 s to 3640 s in the first dimension and 0.8 s to 1 s in the second dimension was excluded from data processing due to signal saturation from residual urease and siloxane column bleed. All chromatographic peaks were searched against the NIST-MS 2017 Libraries.

Small retention time shifts are possible in GC×GC and common in metabolomics studies. The Statistical Compare feature of ChromaTOF® aligned the peak tables across runs and ensured that the same ion was chosen to quantify a given peak throughout the data set. Tolerances for retention time shift were ±5 s modulation period (P_M = 2.5 s) in the first dimension, and 0.2 s for the second dimension to account for the possible retention time shift across all samples. The minimum similarity for

mass spectral match to combine sub-peaks was set at 600 using all *m/z* values with abundances greater than 1%. After aligning the peak tables with the set parameters, the Statistical Compare result was exported as a .csv file for further data analysis in MATLAB® R2017a, Windows 64-bit version (The Mathworks Inc., Natick, MA, USA).

3.2.6 Multivariate Analysis

Further multivariate statistical analysis was performed using PLS Toolbox (R8.5.2; Eigenvector Research Inc., Wenatchee, WA, USA). The data were normalized using the three different normalization approaches for comparison (creatinine, TPA, TUPA). The details of the normalization procedures are discussed in Section 3.3. Principal component analysis (PCA) was applied with cross-validation to evaluate general clustering of normalized data by two different methods. An in-house feature selection algorithm, Cluster Resolution Feature Selection (CR-FS), was used for variable selection. This approach is based on the evaluation of ranked variables via a sequential backward elimination and a forward selection mechanism to maximize cluster resolution between comparative groups, where cluster resolution is defined as the maximum confidence interval over which two confidence ellipses do not intersect.[173] The feature selection was permuted over the data set 100×, with the samples being partitioned randomly between training and optimization sets to avoid possible spurious results due to optimization with any given partitioning of the data. Complete information on this feature selection tool is beyond the scope of this thesis, and more details regarding the approach can be found in other literature.[173-175]

3.3 Summary of Normalization Procedures

3.3.1. Normalization to Creatinine

The abundance of creatinine can be determined from the GC×GC-TOFMS signal, provided that the signal is not outside of the dynamic range. With the general data processing method described in Section 3.2.5, the single compound of creatinine was integrated as multiple peaks due to the higher concentration of creatinine compared to the majority of metabolites in the sample. The highly abundant creatinine resulted in a wider chromatographic peak width than the 10 s of the expected peak width that was

set in the general data processing method. This led to multiple chromatographic peaks, as seen in Figure 3-1. The mass spectra of peak A, B, and C are similar in both masses and the relative abundance ratios of fragment ions; this confirms that these are spectra of the one compound.

Integrating the creatinine peak area accurately is extremely important in order to use its peak area as a normalization factor for urinary metabolomics. The data processing method for targeted creatinine integration was used to account for variable creatinine peak sizes/shapes arising from concentration differences in various samples. The processing method for creatinine limited the search region to the window of 1720 s to 1850 s in the first dimension, with an expected peak width of 50 s and 0.5 s in the first and second dimensions, respectively. The creatinine peak was found in the chromatographic region of 1775 ± 2.5 s in the first dimension and 1.57 ± 0.02 s in the second dimension. The *m/z* of 329, which is the molecular ion of creatinine-3TMS (derivatized creatinine with three trimethylsilyl groups), was used as a quantification mass. The rest of the parameters were the same as for the general data processing method. The creatinine peak data processing method was applied to all 54 chromatograms, and the peak tables were aligned using the statistical compare feature in ChromaTOF®. The aligned peak table was imported to MATLAB® R2017a, and the peak areas of all metabolites in each sample obtained from Section 3.2.5 were normalized to the creatinine response for that particular sample.

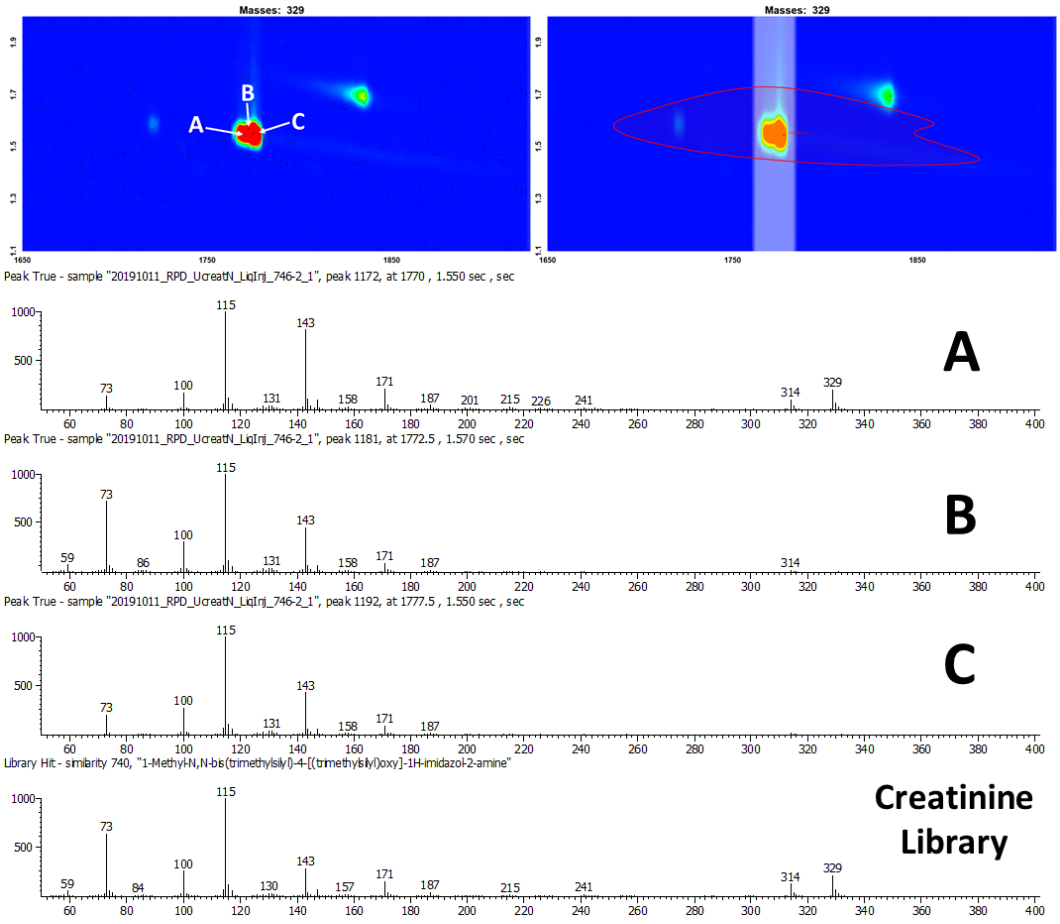

Figure 3-1. Integration of the creatinine peak to yield an accurate peak area. Top left: creatinine peak processed with the general data processing method, Top right: creatinine peak of the same sample processed with the data processing method that is specific for creatinine peak area integration. Bottom: mass spectra of the peak A,B, and C compared to the creatinine library spectra.

3.3.2 Normalization to TPA and TUPA

The Statistical Compare in ChromaTOF® is a useful feature for analyte alignment and obtaining basic statistics. Analyte alignment is critical when dealing with a number of samples for the accumulation of multiple peak tables. This Statistical Compare feature aligns the same analyte peak over multiple samples, based on the retention times (1st and 2nd dimensions) and mass spectral similarity (Figure 3-2). The peak areas of the aligned peaks are calculated using the same mass channel for consistency in quantification. With the Statistical Compare feature, the analytes in 54 samples were aligned and ChromaTOF® generated a peak table containing basic statistics, including average, standard deviation, and minimum and maximum responses, for each

compound across the dataset. The table also included a count column, which records the number of samples in which the corresponding peak appeared.

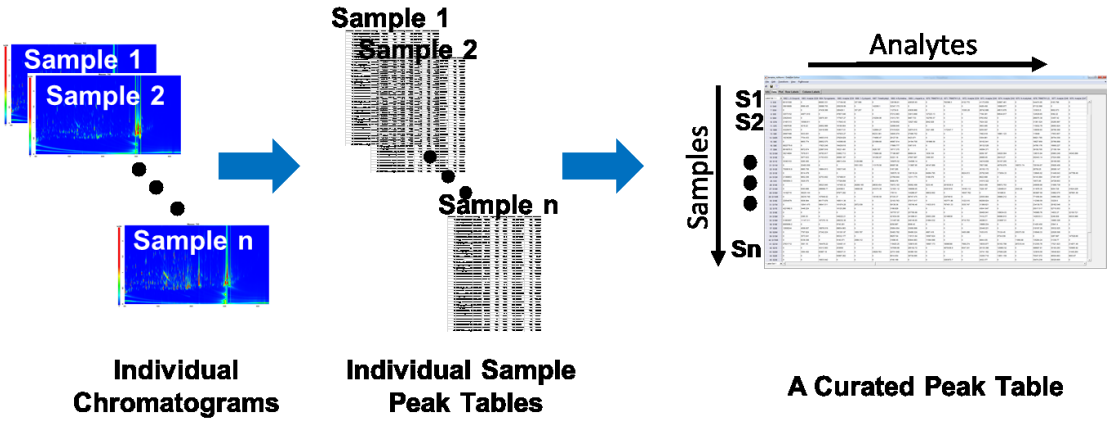

Figure 3-2. Schematic illustration of the software process for the alignment of analytes across the samples and generation of the final peak table.

For TPA normalization, the total peak area for each sample was calculated by summing the areas of all aligned peaks as the normalization factor. The TUPA normalization factor includes only peaks that were present in all 54 samples across the full dataset in the summation. Figure 3-3 shows the graphic illustration of the method used to obtain the TUPA normalization factor. For simplicity of explanation, a matrix size of 15 samples containing 20 aligned analytes was simulated and illustrated in Figure 3-3. The light blue color indicates the analytes that are present in the samples, whereas the white blanks indicate none-detected peaks. Out of 20 total analytes, 7 of them (A1, A5, A10, A11, A14, A17, A19) were present in all 15 samples and were shaded in gray. These 7 analytes are considered as "useful peaks". The corresponding peak areas of those 7 analytes for each sample were summed to obtain the TUPA that serves as a normalization factor for the corresponding sample. All metabolites processed and aligned according to Section 3.2.5 were normalized to the corresponding TPA and TUPA of the respective sample by dividing the original peak areas of metabolites by either the TPA or TUPA normalization factors.

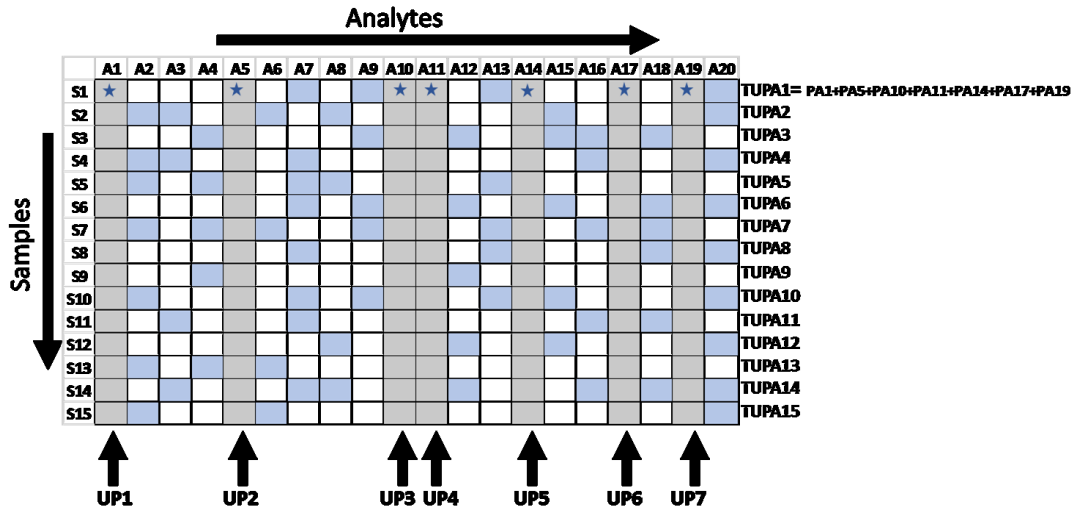

Figure 3-3 Graphic illustration for obtaining the TUPA normalization factor.

3.4 Results and Discussion

Two samples, S15E and S26E, were removed due to an apparent instrument malfunction (frozen cold jet) in the data. Aligned urine GC×GC-TOFMS data of 54 urine samples including replicates of one male and female samples (S11MR, S17MR), were then normalized using the three different strategies: urinary creatinine concentration, TPA, and TUPA. The effects of different normalization strategies on the chemometric analysis were evaluated by visualizing the effects of normalization on the data in PCA score space both pre- and post-feature selection.

The creatinine peak area, TPA and TUPA for each of the 54 urine samples are listed in Table 3-1. The minimum creatinine peak area out of all 54 samples was found to be 1.33×10^6, whereas the maximum creatinine peak area was 22.08×10^6 (a 16.6-fold difference). The average creatinine value of 54 urine samples was 8.61×10^6, with a relative standard deviation of 65%. Creatinine levels were compared based on the sex of the donor, but there were no statistically significant differences in this respect. For female samples, 2.07×10^6 was the minimum creatinine peak area, and the maximum creatinine peak area was 22.08×10^6. The average creatinine peak area value of 30 female urine samples was 8.53×10^6 , with a relative standard deviation of 65%. For male samples, 1.33×10^6 was the minimum creatinine peak area, and the maximum

creatinine peak area was 20.49×10^6. The average creatinine peak area value of 24 male urine samples was 8.72×10^6, with a relative standard deviation of 66%.

The Total Peak Area (TPA) and Total Useful Peak Area (TUPA) were calculated based on the aligned peak table. From the 54 urine chromatograms, 5572 peaks were aligned using the statistical compare feature on ChromaTOF®. For TPA, the areas of all aligned 5572 peaks were summed to generate a TPA normalization factor for each sample. Out of 54 samples, the minimum TPA was found to be 5.37×10^8, while the maximum TPA was 3.61×10^9. For calculation of TUPA, the sum of all peak areas for those peaks that are present in all samples under study was taken as the normalization factor. Out of 5572 aligned peaks, 470 peaks were found to be present in all 54 samples, covering about 8.4% of the total aligned analytes. The peak areas of these 470 compounds in each sample were summed, and the total peak area of these "useful" compounds was used as a normalizing factor for the respective sample. Since these are the compounds that are common in all samples and the peak area was determined using the same mass channel for the same analyte in all samples, it was considered that the total peak area of these compounds could be regarded as a realistic representation of urine sample concentration. The creatinine peak area, TPA and TUPA for each of the 54 urine samples are listed in Table 3-1 for comparison with the corresponding sample. The smallest TUPA out of all 54 samples is 1.67×10^8 , whereas the largest TUPA is 8.48×10^8. It is noteworthy that the total variation in TUPA is the smallest amongst the three different normalization methods at ~1/3 the variation in creatinine concentration. For creatinine measurmenet, no additional work was done to verify the accuracy of the creatinine measurement obtained from GC×GC, such as using colorimetry for creatinine measurement. While it would be beneficial to validate the creatinine measurement, it also reveales the disadvantages of using creatinine as a normalization factor; it either requires an additional analysis or the creatinine measurment that is directly from GC×GC as a normalization factor may be impacted more significatly by variations from sample preparation, data acquisition, and/or data integration.

Table 3-1. Comparison of Creatinine Peak Areas, TPA, and TUPA of 54 Urine Samples

Sample	Sex	CT	TPA	TUPA	Sample	Sex	CT	TPA	TUPA
S1M	M	8.42E+06	1.34E+09	3.27E+08	**S14M**	F	1.75E+07	2.41E+09	7.01E+08
S1E	M	3.57E+06	8.28E+08	2.48E+08	**S14E**	F	1.32E+07	1.83E+09	5.99E+08
S2M	F	3.83E+06	1.81E+09	5.02E+08	**S15M**	F	1.58E+07	2.79E+09	6.85E+08
S2E	F	2.14E+06	7.14E+08	2.53E+08	**S16M**	F	1.47E+07	2.16E+09	7.05E+08
S3M	M	1.28E+07	1.67E+09	4.31E+08	**S16E**	F	6.64E+06	1.19E+09	3.65E+08
S3E	M	8.77E+06	2.68E+09	6.48E+08	**S17M**	F	1.02E+07	1.75E+09	5.71E+08
S4M	F	1.46E+07	1.67E+09	5.41E+08	**S17MR**	F	5.47E+06	1.05E+09	3.35E+08
S4E	F	3.04E+06	1.38E+09	3.66E+08	**S17E**	F	2.31E+06	6.06E+08	2.48E+08
S5M*	M	1.33E+06	2.39E+09	7.46E+08	**S18M**	M	2.05E+07	2.71E+09	8.30E+08
S5E*	M	1.65E+07	1.95E+09	5.15E+08	**S18E**	M	1.65E+07	1.86E+09	6.28E+08
S6M	F	5.74E+06	1.70E+09	8.48E+08	**S19M**	F	2.21E+07	2.28E+09	6.91E+08
S6E	F	1.19E+07	1.53E+09	4.71E+08	**S19E**	F	2.07E+06	5.37E+08	1.67E+08
S7M	M	3.76E+06	1.78E+09	4.61E+08	**S20M**	F	3.60E+06	1.26E+09	3.54E+08
S7E	M	4.06E+06	9.66E+08	3.33E+08	**S20E**	F	4.71E+06	8.62E+08	2.78E+08
S8M	M	4.59E+06	3.61E+09	4.31E+08	**S21M**	F	4.38E+06	1.80E+09	4.79E+08
S8E	M	1.48E+07	1.55E+09	4.20E+08	**S21E**	F	7.29E+06	1.42E+09	4.55E+08
S9M	M	7.16E+06	9.16E+08	2.70E+08	**S22M**	M	6.00E+06	1.22E+09	3.40E+08
S9E	M	3.61E+06	6.77E+08	2.19E+08	**S22E**	M	1.72E+06	8.50E+08	2.35E+08
S10M	F	1.36E+07	1.67E+09	4.47E+08	**S23M**	F	7.29E+06	9.50E+08	2.57E+08
S10E	F	1.38E+07	1.71E+09	3.88E+08	**S23E**	F	2.42E+06	5.38E+08	1.80E+08
S11M	M	1.35E+07	1.88E+09	5.96E+08	**S24M**	M	2.75E+06	8.70E+08	2.43E+08
S11MR	M	1.32E+07	1.87E+09	5.49E+08	**S24E**	M	9.66E+06	1.39E+09	3.82E+08
S11E	M	1.17E+07	1.64E+09	4.44E+08	**S25M**	F	1.51E+07	3.01E+09	7.31E+08
S12M	F	7.71E+06	1.46E+09	3.87E+08	**S25E**	F	5.27E+06	1.04E+09	2.77E+08
S12E	F	4.68E+06	1.04E+09	3.25E+08	**S26M**	M	4.26E+06	1.29E+09	3.45E+08
S13M	M	2.72E+06	1.54E+09	4.25E+08	**S27M**	F	1.17E+07	2.18E+09	6.17E+08
S13E	M	1.74E+07	2.07E+09	6.09E+08	**S27E**	F	3.16E+06	6.40E+08	2.21E+08

***** denotes the samples that are used for further detailed comparison.
R denotes replicate samples.

Two urine samples (S5M and S5E) from the same individual at different concentrations were selected randomly for a more detailed and thorough comparison of the different normalization methods. S5M is expected to be more concentrated than S5E because urine produced in the morning is generally more concentrated than urine produced throughout the day due to overnight dehydration.[176-178] The region of interest in the TIC chromatograms (150–3100 s ${}^{1}t_R$ and 1.0–2.5 s ${}^{2}t_R$) and the zoomed-in regions for creatinine (plotted as ion 329 EIC) are illustrated in Figure 3-4. The color scales are

constant for the TIC and EIC to facilitate visual comparison. From Figure 3-4, it is evident that S5M produced slightly more intense signals than S5E in general. The TPA were 2.39×10^9 and 1.95×10^9, and the TUPA were 7.46×10^8 and 5.15×10^8 for S5M and S5E, respectively. The peak areas of creatinine obtained from the creatinine-specific processing method also were compared. The creatinine peak area of S5M was found to be 1.33×10^6, whereas the creatinine peak area was 16.49×10^6 for S5E. According to the TUPA, S5M is approximately 1.5 times more concentrated than S5E, which agrees with a visual inspection. Conversely, the creatinine peak suggests that S5E is about 12 times more concentrated than S5M, which is obviously incorrect by visual inspection of the data.

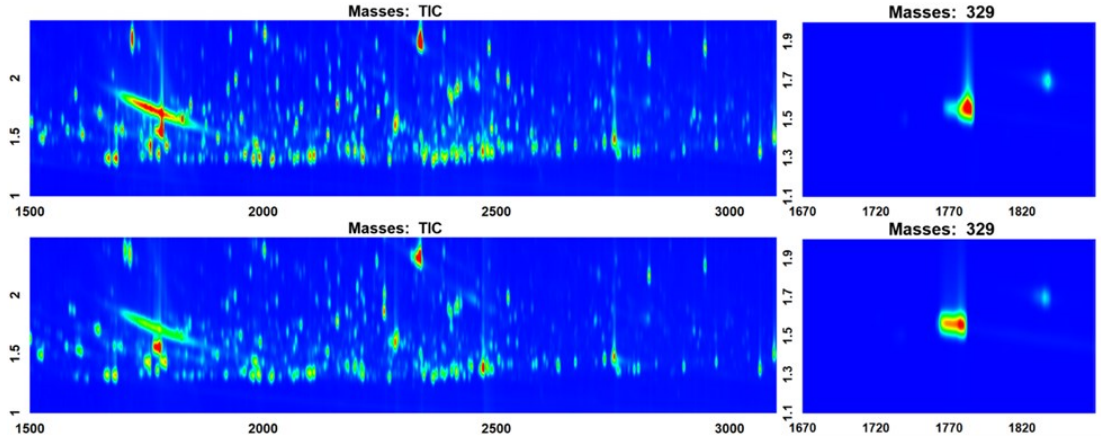

Figure 3-4. Comparison of two samples from the same subject (Top: S5M, Bottom: S5E). Left: TIC (Total Ion Chromatogram), Right: Creatinine peak.

To evaluate the impact of the different normalization strategies on a chemometric interpretation of the results, the three normalization methods were compared in PCA space in addition to the dataset without normalization. PCA is an unsupervised method that helps visualizing and interpreting the data in exploratory data analysis. In this study, PCA was used for dimensionality reduction with the aim to project the data using the first two principal components. PCA as an unsupervised method was not used to classify male versus female, but PCA was employed to visualize similarity/dissimilarity between samples on the PCA score space.[179]

The PCA score plots of each dataset normalized with different methods, are shown in Figure 3-5. Red diamonds refer to female samples and green sqaures refer to male samples. The urine dataset is very rich in the information collected from the high dimensional GC×GC-TOFMS. With all 5572 variables included, no distinct separation was possible for any of the normalization. Feature selection was applied to each normalized version of the data to discover which variables were of value to separate the two sample groups while removing noise and irrelevant variables. This was achieved using an in-house algorithm for cluster-resolution guided feature selection (CR-FS). In applying the CR-FS algorithm, the 54 urine samples were split randomly into a training set for feature selection and model construction (36 samples, 2/3 of the dataset) and a test set for model validation (18 samples, the remaining 1/3).[173] Assignment into either set was random, and the process was permuted 100 times to diminish the possibility of overfitting or obtaining a good model by chance due to a particular division of samples into training, optimization, and validation sets. A survival rate of 50% was chosen for the feature selection algorithm (i.e. variables that were selected more than 50% of the time were included in the list of selected variables). Since the starting data matrices differ in terms of the values of the analyte responses depending on the normalization strategy (or lack of normalization), the number of variables selected for each data set was different, although the same automated feature selection algorithm was used. After feature selection, PCA was performed on the data (here aiming to separate male vs. female subjects) (Figure 3-6).

Without normalization, the optimal model required 223 variables; explaining 11.95% of total variance by PC 1 and 7.84% by PC 2. With normalization to creatinine, 168 variables were chosen, and 10.84% of the variance was explained by PC 1 and 7.34% explained by PC 2. In neither case is there a clear distinction between the two groups in the first two components of PCA space. When normalized with TPA, 285 variables were selected, and 13.89% of the variance was explained by PC1 and 7.05% explained by PC2, with a slight overlap of 95% confidence level ellipse. When TUPA was used, 306 variables were selected that clearly could distinguish between male and female samples in PCA space. It is noteworthy that for all four PCA score plots,

approximately 20% of the variance was explained with the first two principal components. It is also evident from the PCA score plots that TUPA displays tighter clusters (smaller within-group variation), especially for the male group, and a more distinct separation between the two classes (large between-group variation).

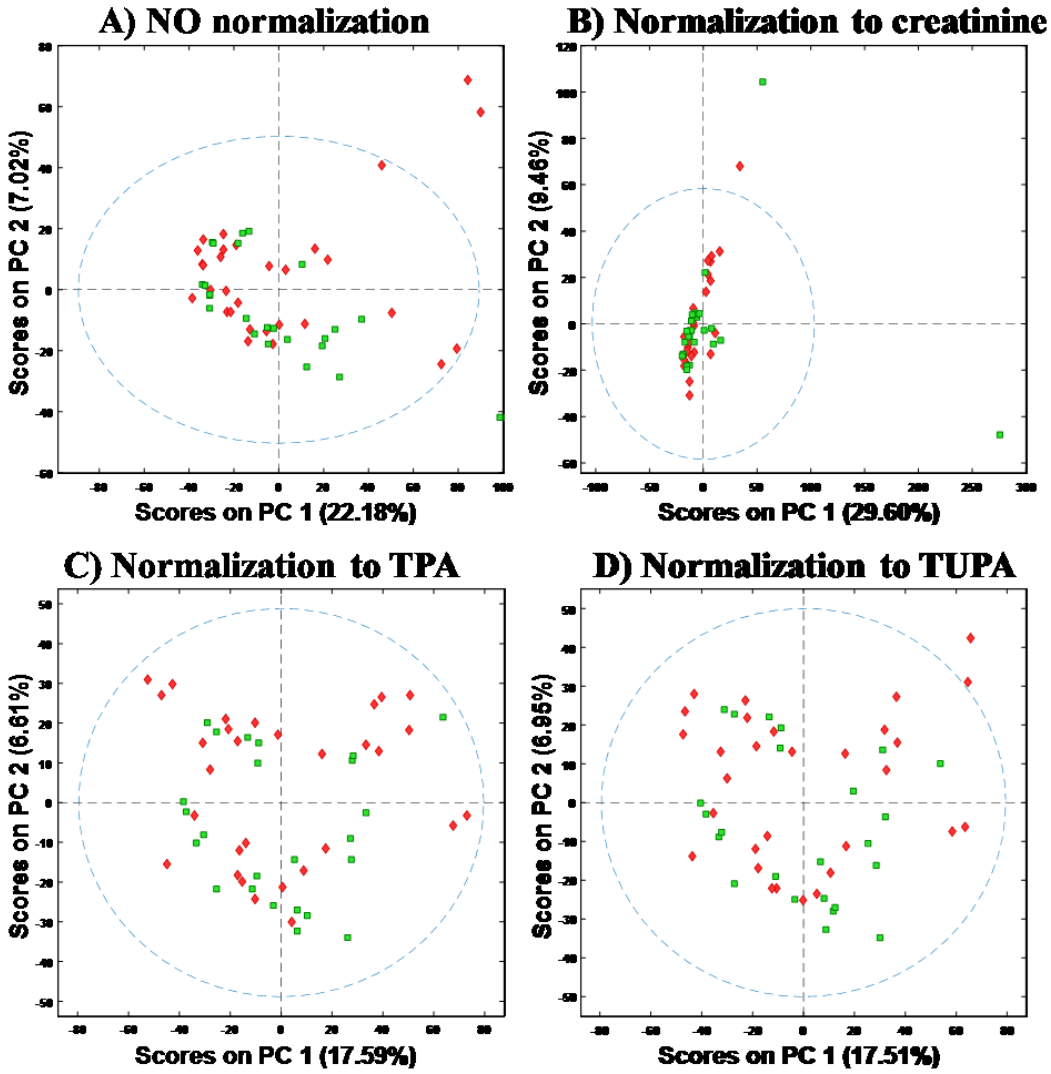

Figure 3-5. PCA score plots for different normalization methods before applying any feature selection. Each dataset was 54 samples × 5572 compounds.

Although the PCA score plot of the urine data with no normalization technique applied (Figure 3-6A) shows a moderate separation between the two groups, this result is analytically invalid because the variations in sample concentrations were not corrected. It has been extensively studied in the metabolomics community that urinary

metabolite profiles vary widely based upon physiological and external factors such as diet, BMI, age, stress, exercise and etc.[155] For this particular study, it was intended to collect samples of widely varying concentrations from the donors of diverse backgrounds to obtain a dataset that contains sufficient diversity for investigating normalization approaches. With the collected widely varying samples for the study, neglecting normalization is irrational regardless of the performance of the resulted model. The PCA score plot that was normalized to creatinine (Figure 3-6B) shows the worst separation between the two groups out of all four PCA score plots. The normalization to creatinine made the separation worse than when no normalization was applied. This is likely due to creatinine normalization contributing nonsensical variations into the data (Figure 3-4). The use of creatinine for normalization was based on the assumption of constant creatinine excretion.[160-162,165,166] The poor performance of the creatinine-normalized model and the conclusion drawn from it that creatinine may not be a reliable normalization method is in agreement with the recent study that compared three normalizations—creatinine, specific gravity, and probabilistic quotient normalization (PQN).[98]

The TPA-normalized PCA score plot (Figure 3-6C) shows a significantly improved separation between the two classes. However, this method would include all the insignificant peaks from the instrument and sample preparation in the normalization factor. The best separation was achieved by TUPA normalization. The whole dataset was normalized to a sum of only the peaks that are present in all samples; thus, the contributions from non-desired signals, such as false peaks, noises, and column bleed, and uncommon exogenous compounds could be eliminated. Many normalization techniques commonly focus on addressing either one of either the biological or technical variability. For example, normalization to creatinine may adjust the biological variability but does not address the technical variability. The good performance of TUPA compared to other normalization methods possibly could be due to its effectiveness to address both biological and technical variations. TUPA is used to adjust the differences in metabolites concentrations arising from different sample concentration levels while effectively eliminating the contributions from xenobiotics

or possible contaminations. On the other hand, technical variations, such as derivatization artifacts, noises, and column bleeds also can be eliminated adequately.

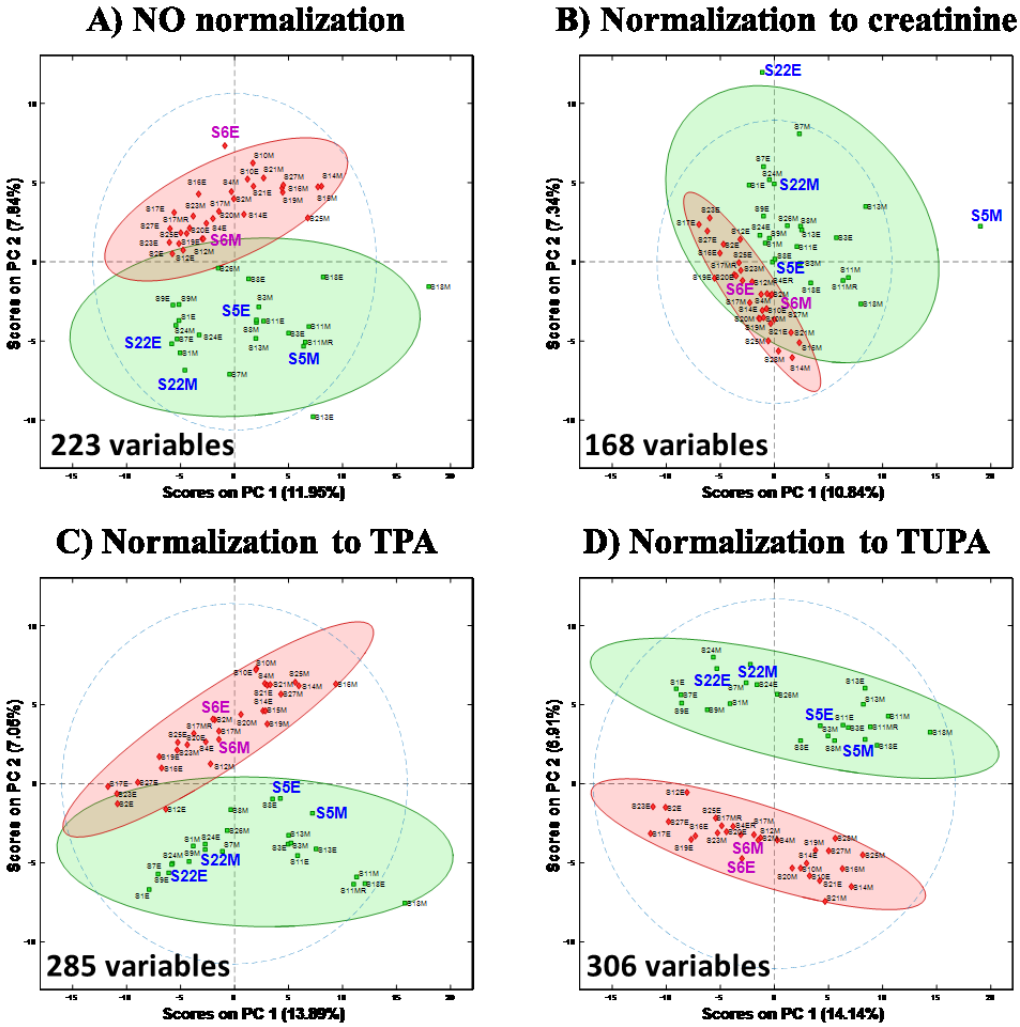

Figure 3-6. PCA score plots with 95% confidence ellipse for (A) without normalization (B) normalized to creatinine concentration (C) normalized to TPA (D) normalization to TUPA.

The major challenge of urine analysis is the variability of urine concentration, which can depend on the time of sample collection, diet, exercise, and level of hydration, etc.[77,172,180] Ideally, the differences between samples from the same subject at different times during the same day without any major change in diet, disease state, etc. should be relatively small, with effective normalization. In PCA score space, one would expect that samples coming from the same subject should project to be relatively

closer together with an optimal sample normalization method. To show how the normalization method would affect these results, the Euclidean distance between the projected pairs of points for three subjects was calculated for each PCA model. To reflect distances in the original data, 20 PCs were selected to build each model, explaining 71–74% of the variance. PCA score plots were set at the same scale, therefore, the distances between samples can be measured easily from the plots directly, and similarities and differences among the samples can be assessed quickly.[179] Figure 3-6 also depicts the points for the pairs of samples from three subjects, labelled in a bigger font to show the sample locations in the PCA score space. In order of Figure 3-6A-D, the Euclidean distance between 5M and 5E is 18.4, 26.1, 13.0, 18.1, respectively. The distance between 6M and 6E is 19.9, 13.1, 17.5, 12.2, respectively, and the distance between 22M and 22E is 8.3, 11.9, 7.1, 6.6, respectively. It is observed that by using TUPA, samples from a single subject tend to project closer together in PCA space. In addition to the aforementioned distinct between-group separation, this suggests that TUPA outperforms the other normalization strategies to correct for the urine sample concentration.

To further demonstrate the need for proper normalization, the variables that were selected to model the data after TUPA normalization (Figure 3-6D) were used to model the system both without normalization, and with normalization to creatinine. (Figure 3-7). This was to demonstrate that the separation is not just based on the selected variables, but on the combination of proper normalization of analyte responses and the appropriate choices of variables. In the case of no-normalization, there was some differentiation between classes, but with creatinine normalization, there was no separation. This result confirms the need to use the appropriate normalization method.

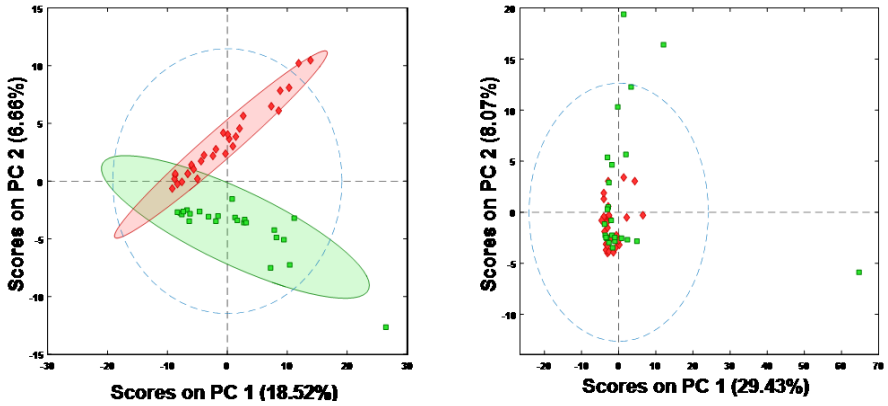

Figure 3-7. PCA score plots with 306 variables that were selected for TUPA modelled on the dataset (A) without normalization (B) normalization to creatinine.

The present study adds weight to the current consensus in the metabolomics literature that failure to normalize the data appropriately and correct between-sample variation can create bias and result in a false discovery of biomarkers. A considerable variation in urine sample concentration, which can differ by up to an order of magnitude, must be managed by using a suitable normalization technique to minimize biases. Of note, this study was conducted with a relatively small sample size and the question of the study may be viewed as too simple compared to more challenging and complex questions in metabolomics that involve very subtle differences between comparative groups. Yet, TUPA was the only normalization method that was able to produce data which could be separated into two distinct groups in this study. A thorough review of the data reveals TUPA to be a more reasonable method to account for urine sample concentration variability than using creatinine concentration.

3.5 Conclusions

Urine contains an enormous amount of metabolic information and carries considerable value as a convenient diagnostic biofluid. An appropriate normalization method, however, is vital to prevent the biological information of the study from being masked by variation in urine sample concentration and to achieve a reliable determination of features that are truly responsible for discriminating the comparative groups of interest. This work demonstrated conclusively that normalization to creatinine response can be

misleading as the creatinine response (and thus normalization factor) can be at odds with changes in overall concentration that are self-evident upon manual inspection of the raw data. This result is in agreement with other recent literature. The present study demonstrated that normalization using TUPA appears to be a more suitable way to account for the variations of urine concentration when performing GC×GC-MS-based metabolomics studies as opposed to using creatinine concentration or TPA as a normalization factor. This post-acquisition normalization method is easy and convenient to apply without requiring any additional experiments. These findings suggest that TUPA may be an effective and feasible alternative to normalize urine data. Furthermore, it may be potentially useful as a post-acquisition normalization method for other metabolomics sample types where data normalization is also a great challenge.

CHAPTER 4

Mass Spectral Filtering Scripts for GC×GC-TOFMS Data – Applications to Biosamples

4.1 Introduction

Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOFMS) as a technology platform offers unparalleled capacity for non-target screening or hybrid target/non-target analysis of complex mixtures. The higher and more effective use of peak capacity, when compared to one-dimensional GC methods, results in improved signal-to-noise ratios due to increased signal (focusing/band compression at modulator) and decreased noise (separation of analytes from primary column bleed and coeluting analytes). Consequently, the spectra are cleaner, allowing improved compound identification. When compared to LC-MS methods, matrix effects are less in GC-MS, and the technique offers a broad dynamic range. Additionally, GC×GC techniques provide an ordered structure of the chromatograms, which is useful for the identification of unknown compounds. Moreover, this technique is advantageous through the possibility of "seeing everything"; the TOFMS allows capturing complete mass spectra at every point. The acquisition of full-range mass spectra for all the compounds present in the sample opens the door to hybrid target/non-target techniques where some known compounds of interest are quantified, while simultaneously detecting unexpected compounds in the sample.

Due to the above benefits, this instrument is extremely useful for non-targeted analysis and discovery studies, where the entire sample is potentially of interest, such as encountered in metabolomics. GC×GC-TOFMS is being used increasingly for non-target metabolomics studies of various sample types, including urine, blood, breath, plant extracts, etc.[22,58,181] However, the amount of data generated from such comprehensive techniques is massive and nearly impossible to handle manually. The complexity of data analysis, rather than the usual culprits of sample preparation or

instrumental time, serves as the major bottleneck in GC×GC-TOFMS analysis. The challenge of data processing is holding GC×GC-TOFMS back from more widespread use.

4.1.1 Motivation

Most of the GC-based metabolomics applications combine GC with MS detection to help the identification of unknown analytes. GC(×GC)-MS uses the electron impact ionization (EI) source, and it generates highly reproducible fragmentation patterns in both *m/z* values and the relative abundance of the corresponding ions. This helps the construction of databases of searchable mass spectral libraries. When a chromatogram is processed with a MS library database, the final peak table contains a list of tentatively identified analytes with the library match similarity factors.

However, despite the use of such databases, a manual review of thousands of peaks in a sample can be a tedious task. Each analyte in the peak table needs to be verified by comparing the retention index and the MS library for a higher degree assurance of compound identification. Unfortunately, in many studies with unit-mass resolution mass spectrometers, many (sometimes most) detected compounds have spectra that do not match well with libraries. While high-speed, high-resolution mass spectrometers, or ion sources with softer ionization than 70 eV electron impact may help with some of this, there are many GC×GC systems in use which use low resolution mass spectrometers. With the process of manually examining the peak table, compounds with low spectral match quality can be reviewed along with library retention indices to obtain the final list of provisional identifications. The issue is that when analytes are searched against a library, it is possible to have multiple compounds with the same library hit for structurally similar compounds. It is also common that a detected analyte is not registered in the mass spectral library database. Especially for TMS/methoximated derivatives, mass spectral and retention index libraries are currently limited and consequently, the resulting peak table suffers from incorrect/ambiguous name assignment.[50] These compounds must be verified through a manual process by knowledgeable personnel capable of interpreting the mass spectrum

and elution order. With a complex sample, where the list of analytes can reach several thousand, this manual process is a significant burden.

In order to speed up and simplify data analysis, script-based filtering of peaks is a promising tool. Scripting involves programming a series of logic rules based on mass spectrometric and/or retention properties for target compounds to determine whether the compound belongs to a specific class.[87,88,182] The extensive and reproducible fragmentation patterns from EI are advantageous for creating mass spectral filtering scripts. The scripts work as a data reduction filter by enabling the classification of chromatographic peaks, based on distinguishable features in mass spectral information. Scripting tools that apply filters to GC×GC-TOFMS data were originally used for environmental and petroleum samples. Numerous scripts have been published to aid with, for example, the identification of halogenated species.[87,183-186] It was evident that the scripting tool greatly assists with automated and rapid classification of the compounds in GC×GC chromatograms.[87,88,187-190] The speed and convenience of data analysis contributed to the more widespread use of GC×GC in the environmental and petroleum fields. This chapter presents progress to develop scripts for GC×GC-TOFMS metabolomics data for rapid screening of complex biological samples, which typically have thousands of compounds with a large variety of compound classes. To the best of my knowledge, this scripting tool was the first time that automated screening by filtering scripts for handling GC×GC-TOFMS data in metabolomics applications had been reported.

4.1.2 Generation of Mass Spectra and Scripting

When a compound elutes from the GC column into an electron impact ion source, it is bombarded by a stream of electrons.[191] These electrons have energy that exceeds the ionization energy of the molecule and thus can extract an electron from an organic molecule to form a positive ion, called the molecular ion.[192] EI typically creates singly charged positive ions by the ejection of one electron out of the neutral molecule.[193] The molecular ion corresponds to an analyte molecule that has not encountered

fragmentation. The neutral form of a molecule is denoted as M, whereas the molecular ion formed upon EI is denoted as a positive radical ion $M^{+\bullet}$ (odd-electron).[192]

$$M + e^- \rightarrow M^{+\bullet} + 2e^-$$

The molecular ion peak can provide valuable information about the analyte and serves as an important reference point in identifying the other fragments to determine the unknown compound.[87] Therefore, one of the main approaches in scripting is to determine the molecular ion in the mass spectrum as the first step. Hilton et al. published an algorithm to find the molecular ion, based on the theoretical criteria of M^+.[87] The molecular ion, in theory, must be the ion of highest *m/z* in the mass spectrum other than the corresponding isotopic clusters. However, in the real world, the peak of the highest *m/z* may not necessarily represent the molecular ion due to coelution, or unstable molecular ions and contributions from noise. Therefore, the undesired signals in mass spectra have to be distinguished from real signals in order to reliably identify the molecular ion. The algorithms by Hilton et al. enable searches from high mass to low mass, seeking for the first occurrence of an ion with significant intensity (at least six standard deviations greater than the mean of previously examined signals).[87]

The energy of the EI ion source (70 eV) exceeds the ionization energies of most molecules that are typically in the range of 7-15 eV.[193] The excess energy that was not expended in the ionization of the molecule is distributed over various internal degrees of freedom: vibrational and rotational levels of the electronic ground state of the ion.[194] The singly charged positive molecular ion that is of low symmetry and energetically unstable, can break up into smaller fragments.[194] Two significant fragmentation types are direct bond cleavage and rearrangement reactions.

Direct cleavage: $ABCD^{+\bullet} \rightarrow AB^+ + CD^\bullet$

Rearrangement reaction: $ABCD^{+\bullet} \rightarrow BCDA^{+\bullet} \rightarrow BC^{+\bullet} + AD$

As to which reaction will take place and dominates, both thermodynamic and kinetic factors, such as the ion's internal energy and structure, bond strength, activation energy, and the thermodynamic stability of the products during the process, are the determining features in the preferred fragmentation/rearrangement routes.[193,194] These also affect the abundance of a fragmentation product. In general, a strong abundance will be observed for the fragments of an increased stability.[193] These characteristic fragmentation patterns can be written as logic rules in scripting. Because the molecular ion fragments in predictable ways, the peaks at the next lowest m/z need to be explicable in terms of reasonable neutral losses. Signals at $[M-5]^+$ to $[M-14]^+$ and at $[M-21]^+$ to $[M-25]^+$ should not exist, as they are irrational fragments. The presence of irrational fragment ions serves as an indicator of a true M^+ with a higher *m/z* than the presumed M^+. On the other hand, the presence of certain fragmentation peaks with intelligible losses helps to confirm that a particular peak is a molecular ion. For example, the existence of $[M-15]^+$ and $[M-18]^+$ peaks indicate the loss of methyl (CH_3) and water (H_2O), respectively. The presence of $[M-28]^+$ peak implies the loss of ethylene (C_2H_4), carbon monoxide (CO) or (rarely) nitrogen (N_2).[191]

Scripting based on the detection of the molecular ion may lead to good results for compounds with a significant molecular ion peak. However, compounds such as tertiary alcohols, carboxylic acids, and esters may lack a visible molecular ion as a result of an unstable molecular ion undergoing rapid fragmentation.[192] In cases of a molecular ion that is hard to identify due to weak signals, inevitably, scripts must rely heavily on characteristic ions in the fragmentation pathway. Although the molecular ion can be absent, the presence of the loss of small neutral molecules, such as H_2O, HOR, HX, CO_2, and CH_3O, from the expected molecular ion can be used as possible diagnostic peaks and help to piece together the possible structure of the compound. In the case of a homologous family of compounds, searching for the ion series can be helpful. For example, linear hydrocarbons contain the ion series of $C_nH_{2n+1}^+$.[185]

On the other hand, the molecular ion of aromatic compounds is stable because of the availability of π-bonding electrons which delocalize the charge.[191] This results in less fragmentation and the absence of a noticeable fragmentation series. Therefore,

scripting based on fragmentation patterns may not be suitable, but the molecular ion can be useful.[87] The accuracy of scripting could be enhanced by using the combinations of molecular ion information, specific neutral losses, common ions and ion series in fragmentation patterns, and isotope abundances of the molecular ion deliberately. Examining multiple spectra of compounds within the target class may provide certain patterns and characteristic features in the mass spectra that would be useful for developing scripts.

4.1.3 Challenges in Scripting for Metabolomics Samples

When developing scripts, finding the molecular ion is beneficial, as the subsequent expected neutral losses can be deduced from the molecular ion peak. The primary reason why scripts could be developed and used widely for environmental studies is due to the convenience in locating the molecular ion for the major compound classes of interest, such as halogenated species and aromatic compounds. Investigation of the molecular ion was performed as the fundamental step in many of the previously reported scripts for environmental samples.[87-89] Writing scripts become more challenging for compounds that generate a weak or absent molecular ion, where the molecular ion cannot be determined readily using the automated algorithm. The TMS derivatives generally produce undetectable molecular ion peaks due to the fast elimination of the substituent radical from the silicon of the molecular ion.[57,192] Of non-TMS derivatives, a compound class of alcohols also fragment quickly due to the strong electronegativity of oxygen, which makes it susceptible to alpha cleavage of the alkyl group.[191]

In this work, the scripts were developed using the scripting feature in the LECO ChromaTOF® software, the dominant commercial software in the world of GC×GC-TOFMS. The greatest advantage of the written scripts presented is their use of mass spectral information only (i.e., without retention information). This makes the scripts versatile and applicable to any GC×GC-TOFMS data, regardless of the conditions used in the analytical run. The scripts were applied to standard mixtures of four different major classes of metabolites (amino acids, fatty acids, fatty acid methyl esters, and

carbohydrates) at different concentrations. After validating the performance of classifying scripts with standards at low and high concentrations, the automated filtering scripts were applied to various derivatized and non-derivatized biosamples to evaluate their performance.

4.2 Experimental

4.2.1 Derivatization Materials

HPLC grade methanol, HPLC grade toluene, and 99.9% pyridine were purchased from Millipore-Sigma Canada. Toluene was dried over anhydrous sodium sulfate (Millipore-Sigma Canada). Methoxyamine hydrochloride (Millipore-Sigma, Canada) solution was prepared in pyridine at a concentration of 20 mg/mL. Ampoules of N-methyl-N-trimethylsilyl fluoroacetamide + 1% trichloromethylsilane (MSTFA + 1% TMCS), purchased from Fisher Scientific Canada and opened immediately prior to use. Safe-Lock amber centrifuge tubes were purchased from Eppendorf Canada Ltd, while 2 mL glass GC vials, glass vials with 300 μL inserts, and GC vial caps (PTFE, silicon) were purchased from Chromatographic Specialities Inc (Canada).

4.2.2 Standard Mixtures

To test the performance of the developed scripts at various concentrations (from LOD level to overload-the-column high), a mixture of amino acid standard mixture (AAS18-10mL analytical standard, Millipore-Sigma Canada), fatty acid standard mixture (GLC-744, Nu-Check, MN, USA), fatty acid methyl esters standard mixture (SUPELCO 37 Component FAME Mix, Millipore-Sigma Canada), and a carbohydrates standard mixture (Carbohydrates kit, Millipore-Sigma Canada) were mixed at 18 different concentrations. The details of how the mixtures were prepared are included in Table A-1 in Appendix A. The compounds of the standard mixture used in the experiment are listed in Table 4-6. A total of 108 compounds are in the mixture of standards, including 17 amino acids, 44 fatty acids, 37 fatty acid methyl esters, and 10 carbohydrates.

Table 4-1. Compounds Included in the Standards Mixture

Amino acid	Carbohydrate
L-Alanine	D-(−)-Arabinose
L-Arginine	D-(−)-Fructose
L-Aspartic acid	D-(+)-Galactose
L-Cystine	D-(+)-Glucose
L-Glutamic acid	α-Lactose monohydrate
Glycine	D-(+)-Maltose monohydrate
L-Histidine	D-(+)-Mannose
L-Isoleucine	D-(−)-Ribose
L-Leucine	Sucrose
L-Lysine	D-(+)-Xylose
L-Methionine	
L-Phenylalanine	
L-Proline	
L-Serine	
L-Threonine	
L-Tyrosine	
L-Valine	

Fatty acid		Fatty acid methyl ester	
C4:0	Butyric acid	C4:0	Methyl butyrate
C6:0	Hexanoic acid	C6:0	Methyl hexanoate
C8:0	Octanoic acid	C8:0	Methyl octanoate
C10:0	Decanoic acid	C10:0	Methyl decanoate
C12:0	Lauric acid	C11:0	Methyl undecanoate
C13:0	Tridecanoic acid	C12:0	Methyl laurate
C14:0	Myristic acid	C13:0	Methyl tridecanoate
C14:1	Myristoleic acid	C14:0	Methyl myristate
C14:1T	Myristelaidic acid	C14:1	Methyl myristoleate
C15:0	Pentadecanoic acid	C15:0	Methyl pentadecanoate
C15:1	10-Pentadecenoic acid	C15:1	Methyl cis-10-pentadecenoate
C16:0	Palmitic acid	C16:0	Methyl palmitate
C16:1	Palmitoleic acid	C16:1	Methyl palmitoleate
C16:1T	Palmitelaidic acid	C17:0	Methyl heptadecanoate
C17:0	Heptadecanoic acid	C17:1	cis-10-Heptadecanoic acid methyl ester
C17:1	10-Heptadecenoic acid	C18:0	Methyl stearate
C18:0	Stearic acid	C18:1T	trans-9-Elaidic acid methyl ester
C18:1T	Elaidic acid	C18:1	cis-9-Oleic acid methyl ester
C18:1	Oleic acid	C18:2T	Methyl linolelaidate
C18:1T	Petroelaidic acid	C18:2	Methyl linoleate

C18:1T	Transvaccenic acid	C20:0	Methyl arachidate
C18:1	Vaccenic acid	C18:3	Methyl γ-linolenate
C18:2TT	Linoelaidic acid	C20:1	Methyl cis-11-eicosenoate
C18:2	Linoleic acid	C18:3	Methyl linolenate
C18:3	Gamma linolenic acid	C21:0	Methyl heneicosanoate
C18:3	Linolenic acid	C20:2	cis-11,14-Eicosadienoic acid methyl ester
C20:0	Arachidic acid	C22:0	Methyl behenate
C20:1	11-Eicosenoic acid	C20:3	cis-8,11,14-Eicosatrienoic acid methyl ester
C21:0	Heneicosanoic acid	C22:1	Methyl erucate
C20:2	11-14 Eicosadienoic acid	C20:3	cis-11,14,17-Eicosatrienoic acid methyl ester
C20:3	Homogamma Linolenic acid	C20:4	cis-5,8,11,14-Eicosatetraenoic acid methyl ester
C20:4	Arachidonic acid	C23:0	Methyl tricosanoate
C20:3	11,14,17 Eicosatrienoic acid	C22:2	cis-13,16-Docosadienoic acid methyl ester
C22:0	Behenic acid	C24:0	Methyl lignocerate
C22:1	Erucic acid	C20:5	cis-5,8,11,14,17-Eicosapentaenoic acid methyl ester
C20:5	Eicosapentaenoic acid	C24:1	Methyl nervonate
C22:2	Docosadienoic acid	C22:6	cis-4,7,10,13,16,19-Docosahexaenoic acid methyl ester
C23:0	Tricosanoic acid		
C22:4	Docosatetraenoic acid		
C22:5n-6	Docosapentaenoic acid		
C22:5n-3	Docosapentaenoic acid		
C24:0	Lignoceric acid		
C24:1	Nervonic acid		
C22:6	Docosahexaenoic acid		

4.2.3 Sample Preparation

The performance of the scripts was evaluated using previously acquired data from a variety of sample types, including urine, plasma, algae, feces, and sweat.[195]

Derivatization

Urine, plasma, algae samples, and standard mixtures were prepared by a typical two-step derivatization process of methoximation, followed by subsequent trimethylsylation.[103] The details of sample preparation for plasma and urine are described in Section 2.2.2.3 and Section 2.3.2.2, respectively. For algae, the samples were prepared by my colleague; I was provided with chromatograms, where I was

responsible for analyzing the data with scripts. In brief, the sample was extracted with methanol and dried under nitrogen. To the dried residue, 50 µL of 20 mg/mL methoxyamine hydrochloride in pyridine were added for methoximation and incubated at 60 ºC for 2h. Subsequently, 100 µL of MSTFA were added and incubated again at 60 ºC for 1 h.

SPME

The volatiles from feces and sweat samples were extracted using a three-phase SPME fibre (CAR/DVB/PDMS).[195,196]

4.2.4 GC×GC-TOFMS Analysis

All GC×GC-TOFMS analyses were performed on the LECO Pegasus 4D system (Leco Instruments, St. Joseph, MI), with an Agilent Technologies 7890 gas chromatograph (Palo Alto, CA) equipped with a four-jet dual-stage modulator. The samples were analyzed with different column combinations and different GC×GC and MS methods.

4.2.5 Data Processing and Automated Classification

All GC×GC-TOFMS data were processed using ChromaTOF® (v.4.72; LECO), commercial software from LECO. The baseline offset was set to 0.9, and the expected peak widths throughout the entire chromatographic run were set to 10 s for the first dimension and 0.15 s for the second dimension. The data were processed with a peak finding threshold of S/N 30:1. Peak finding and deconvolution of mass spectra were performed automatically as an embedded function of ChromaTOF®. All chromatographic peaks were searched against the NIST MS Search v.2.3 (2017) and Wiley 08. The scripting option was enabled in the ChromaTOF® software, allowing user-written scripts to be applied over the entire chromatographic spaces. The scripts were written with Microsoft VBScript language, a Visual Basic dialect.

4.2.6 Scripting-based Classifications and Evaluation

The developed scripts used mass spectral information without any retention time information to locate the members of target compounds/classes. The scripts were written as a set of logical operations, incorporating the knowledge about mass spectral fragmentation for the class of compounds of interest. In general, the scripts presented in this study involve the following steps: the expected molecular ion of the family of compounds was calculated based on the molecular structure, probable neutral losses were subtracted from the calculated molecular ion, and then other prominent features (e.g., abundance of major fragments and low intensities for specific regions in the mass spectrum) were evaluated. For metabolites that are mostly non-halogenated species, isotopic ratios are not as useful as they are for the halogenated species in environmental studies.

For a homologous series of metabolites, the expected molecular ion (Expected_MW) was calculated using the number of carbons in the alkyl chain. For example, for the class of normal saturated fatty acid methyl esters, the theoretical molecular ion was determined with Equation E1. From the calculated molecular ion, the subsequent fragment losses, such as $[M-31]^+$, representing a loss of a methoxy group, as well as $[M-43]^+$, representing a loss of a C_3 unit via a rearrangement, were investigated; this confirms that it is indeed a methyl ester (Figure 4-1A). $[M-29]^+$ is also a diagnostic fragment ion for FAMEs, which is formed from the rearrangement of C_2 and C_3. The ion at m/z 74 is the McLafferty rearrangement ion, which is the base peak for FAMEs (Figure 4-1B). For the case of TMS derivatives of saturated fatty acids, the molecular ion also was calculated, based on the number of carbons in the alkyl chain, with Equation E2. For TMS derivatives of fatty acids, the molecular ion is generally weak or absent due to its susceptibility to hydrolysis. Instead, $[M-15]^+$, which represents the loss of a methyl group, is significantly abundant. In addition, *m/z* 73 and 75 are common to all TMS derivatives and usually are considerably abundant.

$$\textbf{E1}: \text{Expected\_MW} = 14 \times \text{Carbon\_number} + 46$$

$$\textbf{E2}: \text{Expected\_MW} = 14 \times \text{Carbon\_number} + 104$$

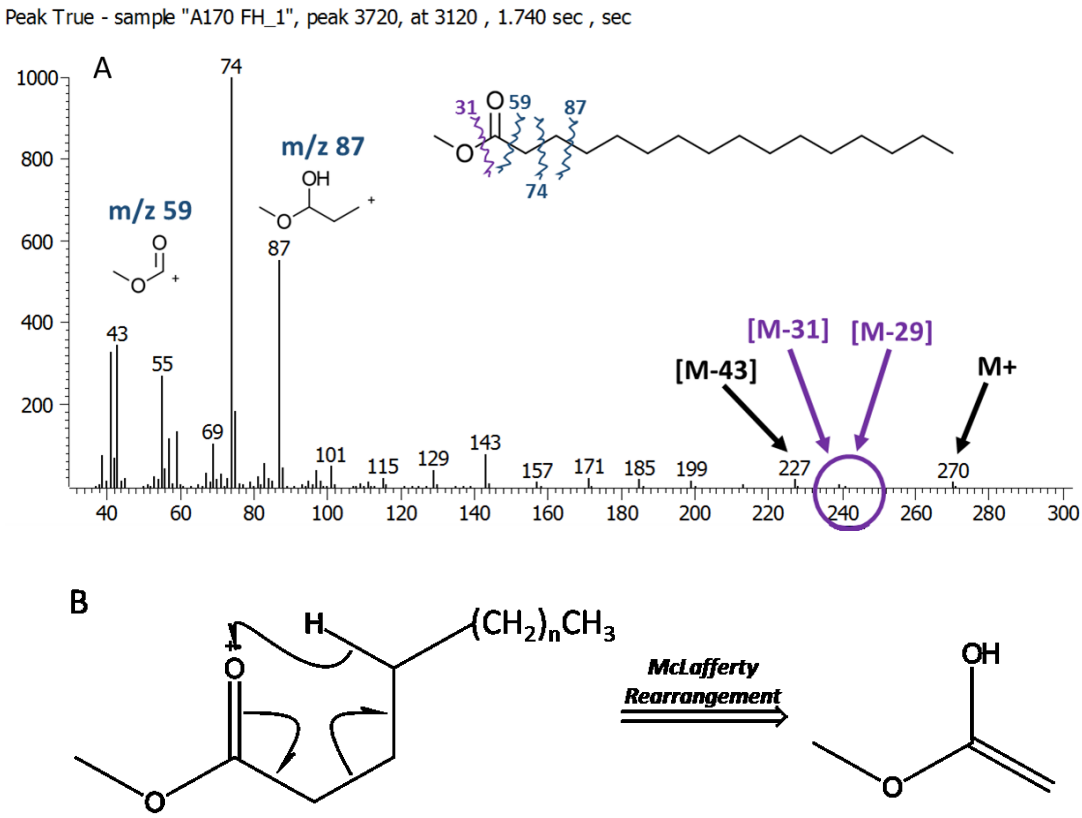

Figure 4-1. A: EI mass spectra of hexadecanoic acid methyl ester with characteristic fragmentations labelled. B: A product ion of the McLafferty rearrangement, which contributes to *m/z* 74 base peak of FAME mass spectrum.

In theory, the molecular ion should be the highest *m/z* aside from its isotopic cluster. To address the limitation of compounds with weak/nonexistent molecular ion peaks, the strategy of calculating the molecular ion, instead of searching for the appearance of the molecular ion from the mass spectra, was used. With this approach, the spectra that have a considerable number of high mass ions bigger than the expected molecular ion with significant intensities were programmed to be excluded from the classification. With logical operations, the intensities (ion counts) in the mass channel region above $[M+2]^+$ to the end mass of the mass spectrum were checked if they fall within the set tolerance for noise (i.e. 1% relative abundance of the base peak). This was done to reduce the chance of bigger molecules being falsely filtered due to common fragment ions or random luck.

A script for the class of saturated fatty acid methyl esters, written using the prominent features described earlier, are outlined below as an example. The other developed scripts that are used in this thesis are included in the Appendix.

```
'Fatty Acid Methyl Esters
'Normal Saturated Fatty Acids Methyl Esters
'1. base peak is m/z 74
'2. Intense m/z 87 peak; either the second or the third most intense peak
'3. molecular ion at reasonable abundance
'4. 150 < m/z 55 abundance < 400
'5. long homologous series of ions general formula [CH3OCO(CH2)n]+ 14 amu apart at m/z =
87, 101, 115, 129, 143, 157, 199
'6. high abundance: [M-31] & [M-43] & [M-29]
'7. m/z 69 <200
'8. m/z 59 >0

Function LinearSaturatedFAME() as boolean
Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
```

```
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5

bed1 = Rank(1)=74
bed2 = Rank(2)=87 or Rank(3)=87
bed3 = (Abundance(55)>150 And Abundance(55)<400)
bed4 = abundance(57)<350
bed5 = abundance(59)>50
'Checks for ion series of general formula [CH3OCO(CH2)n]+
fragment_counter = 0
for n = 2 to 30  step 1
        fragment_test =n*14+59
        if abundance(fragment_test)>5 then
                fragment_counter= fragment_counter +1
        else
                LinearSaturatedFAME = false
        end if
```

```
next
Carbon_number = 5
Do while Carbon_number < 30
Expected_MW = 14*(Carbon_number-1) + 60
M29 = Expected_MW -29
M31 = Expected_MW -31
M43 = Expected_MW -43
bed6 = Abundance(Expected_MW)>5
bed7 = Abundance(M29)>0 or Abundance(M31)>0 or Abundance(M43)>0
noise_counter1 = 0
noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and fragment_counter
>= (Carbon_number-10) and noise_counter1<=5 and noise_counter2<5 then
LinearSaturatedFAME = true
Carbon_number = Carbon_number + 1
Loop
end function
```

To evaluate the performance of the developed scripting, standard mixtures of four different classes of compounds (amino acids, fatty acids, fatty acids methyl esters, and carbohydrates) were prepared at various concentrations. Since the scripts rely entirely on the mass spectral information, the spectral quality is crucial for scripts to work reliably. Various concentrations of standard mixtures were prepared to test both ends of concentrations: 1) at extremely high concentrations, where the peaks are overloaded and saturating the detector, causing ion ratios to be erroneous and 2) at very low concentrations, where the signals may be close to the noise level, consequently boosting the noise in mass spectra, leading to inaccurate ion ratios and fragment patterns. Eighteen mixtures of the standards at different concentrations were prepared by mixing the four different classes of standard mixtures at various concentrations. These standard mixtures were derivatized following the two-step methoximation and trimethylsilylation derivatization procedure.

4.3 Results and Discussion

Since GC×GC-TOFMS data provide both chromatographic and mass spectral information, scripts can be written using either or a combination of retention and spectral information. While mass spectral information is independent of the GC×GC-TOFMS method used, retention information changes depending on the choice of the column combination and the GC×GC method (temperature programming) used for the analysis. Some scripts use retention information in the search algorithms in order to increase accuracy of the scripts for the compound classes that are challenging to distinguish with the mass spectral information alone. The developed scripts presented herein were written using only mass spectral information. This provides a significant advantage because these scripts are independent of separation parameters and can be applied to any GC or GC×GC-TOFMS chromatograms, regardless of the column combinations and GC and MS conditions used.

4.3.1 Evaluation of Scripts

The same data processing method with the in-house written scripts were applied to all 18 chromatograms using the LECO ChromaTOF®. A peak table for each chromatogram was generated automatically from the software after the data processing was finished. The family group for the compounds that are classified by the developed scripts were displayed in the classification column in the peak table. The peak tables were sorted to prioritize the compounds that are classified (Table 4-2, 4-3). The detailed results of how many compounds in the standard mixtures at various concentrations were classified for each group are included in Table A-2 in Appendix A. Figure 4-2 shows the classified peaks using the "bubbles" feature, where the radii of the bubbles correspond to the relative area of the peak represented. Each class of compounds was assigned with a different colour. It is visually evident that the scripts seemed to struggle more at low concentrations (Figure 4-2A) rather than at high concentrations (Figure 4-2C). Figure 4-2C showed that some peaks were overloaded, yet they did not affect the performance of the classifying scripts significantly.

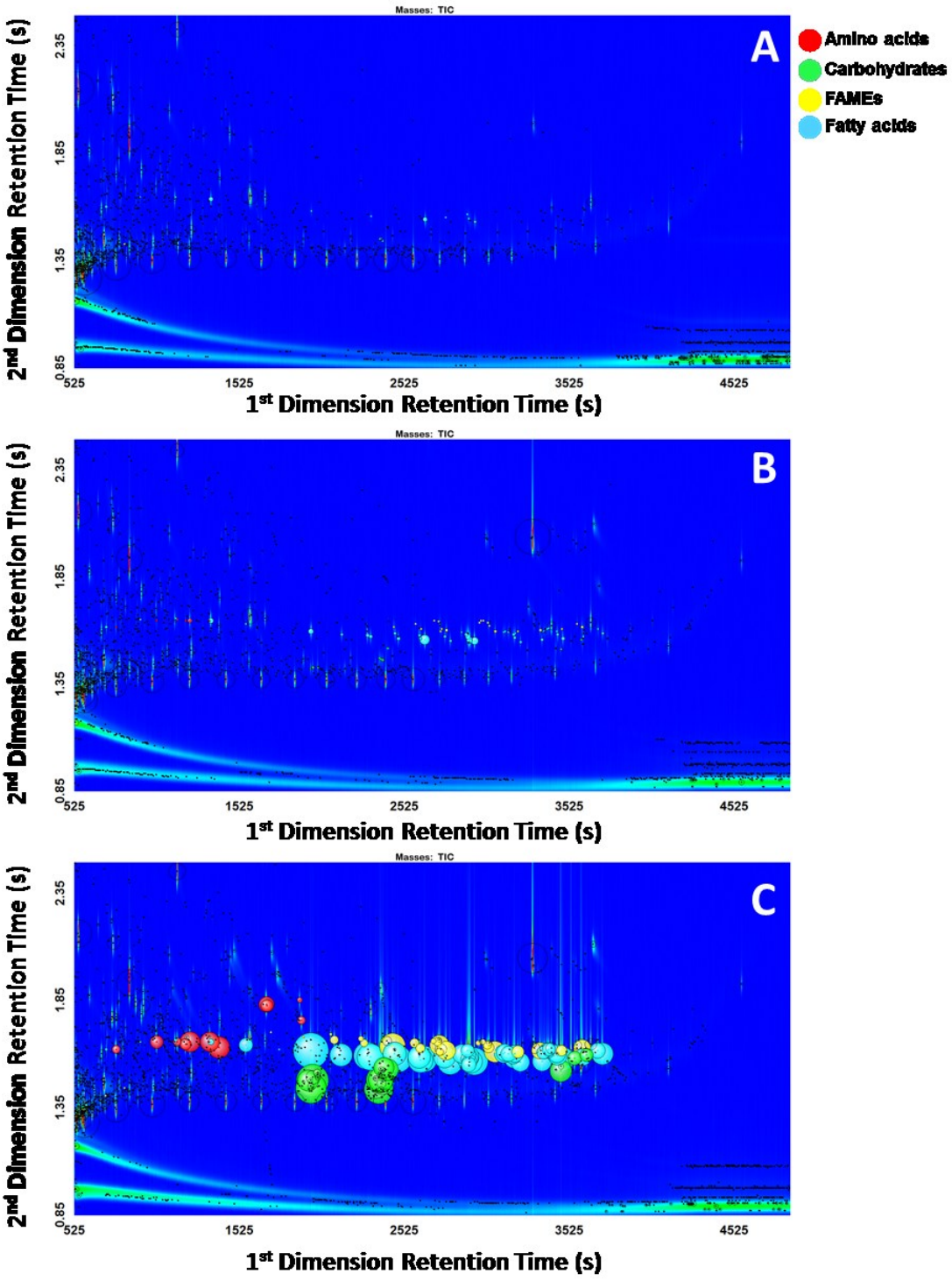

Figure 4-2. A) Low concentration **B)** Intermediate concentration **C)** High concentration

Table 4-2. 14 Compounds Filtered from the Standard Mixture at the Lowest Concentration

Name	Classifications	R.T. (s)	Quant S/N	Similarity	Reverse
d-Mannose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, o-methyloxyme, (1Z)-	sugar_5TMS	2375 , 1.455	220.08	747	882
d-Glucose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, o-methyloxyme, (1Z)-	sugar_5TMS	2382.5 , 1.455	185.3	680	866
d-Galactose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, o-methyloxyme, (1E)-	sugar_5TMS	2392.5 , 1.445	380.89	823	887
Palmitoleic acid 1TMS	monoenoicFA_TMS	2775 , 1.555	52.982	529	763
trans-9-Octadecenoic acid, trimethylsilyl ester	monoenoicFA_TMS	2912.5 , 1.550	497.37	844	870
9,12-Octadecadienoic acid (Z,Z)-, trimethylsilyl ester	dienoicFA_TMS	2905 , 1.570	176.76	726	852
Myristic acid, TMS derivative	SatFA_TMS	2322.5 , 1.560	348.5	789	855
Dodecanoic acid, trimethylsilyl ester	SatFA_TMS	1960 , 1.585	863.29	860	912
Trimethyl palmitate	SatFA_TMS	2650 , 1.545	3103	906	932
Trimethyl stearate	SatFA_TMS	2952.5 , 1.535	1711.2	883	917
Trimethylsilyl hexanoate	SatFA_TMS	717.5 , 1.605	2417.3	928	949
heptanoic acid TMS	SatFA_TMS	922.5 , 1.650	488.14	811	870
Octanoic acid, trimethylsilyl ester	SatFA_TMS	1137.5 , 1.650	1037.2	914	932
Trimethylsilyl nonanoate	SatFA_TMS	1352.5 , 1.640	3152.1	897	951

Table 4-3. 95 Compounds Filtered from the Standard Mixture at the Highest Concentration

Name	Classifications	R.T. (s)	Quant S/N	Similarity	Reverse
L-Valine, N-(trimethylsilyl)-, trimethylsilyl ester	valine	1022.5 , 1.660	27807	909	915
L-Threonine, 3TMS derivative	threonine_3TMS	1400 , 1.635	13900	937	938
D-(+)-Turanose, octakis(trimethylsilyl) ether	sugar_8TMS	3470 , 1.525	13774	848	852
D-Lactose, octakis(trimethylsilyl) ether, methyloxime (isomer 2)	sugar_8TMS	3555 , 1.585	9778.9	912	912
Maltose, octakis(trimethylsilyl) ether, methyloxime (isomer 2)	sugar_8TMS	3627.5 , 1.600	6294.7	913	917
d-Glucose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, o-methyloxyme, (1E)-	sugar_5TMS	2372.5 , 1.485	30046	930	948
d-Galactose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, o-methyloxyme, (1E)-	sugar_5TMS	2380 , 1.485	163 93	924	941
d-Glucose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, o-methyloxyme, (1Z)-	sugar_5TMS	2400 , 1.540	19742	936	955

d-Galactose, 2,3,4,5,6-pentakis-O-(trimethylsilyl)-, o-methyloxyme, (1Z)-	sugar_5TMS	2420 , 1.530	10441	949	969
D-(-)-Fructose, pentakis(trimethylsilyl) ether, methyloxime (syn)	sugar_5TMS	2367.5 , 1.440	30674	922	922
D-Arabinose, tetrakis(trimethylsilyl) ether, ethyloxime (isomer 1)	sugar_4TMS	1940 , 1.480	24484	916	916
D-Arabinose, tetrakis(trimethylsilyl) ether, ethyloxime (isomer 2)	sugar_4TMS	1957.5 , 1.445	42525	915	915
D-Arabinose, tetrakis(trimethylsilyl) ether, ethyloxime (isomer 1)	sugar_4TMS	1967.5 , 1.480	30778	919	919
D-Arabinose, tetrakis(trimethylsilyl) ether, ethyloxime (isomer 1)	sugar_4TMS	1995 , 1.515	56.38	907	907
Trimethylsilyl 2-[(Trimethylsilyl)amino]-3-[Trimethylsilyl)oxy]propanoate	serine_3TMS	1332.5 , 1.680	68.207	489	929
Serine, 3TMS derivative	serine_3TMS	1345 , 1.660	21854	933	934
Phenylalanine, 2TMS derivative	phenylalanine_2TMS	1900 , 1.760	9406.5	923	924
Linolenic acid, trimethylsilyl ester	multienoicFA_TMS	2870 , 1.605	4904	897	898
à-Linolenic acid, TMS derivative	multienoicFA_TMS	2907.5 , 1.620	2745.4	885	886
Arachidonic acid, TMS derivative	multienoicFA_TMS	3122.5 , 1.605	5789.9	918	918
Eicosapentaenoic Acid, TMS derivative	multienoicFA_TMS	3132.5 , 1.610	6779.1	930	930
Norlinolenicacid TMS	multienoicFA_TMS	3155 , 1.605	5041.1	839	856
à-Linolenic acid, trimethylsilyl ester	multienoicFA_TMS	3192.5 , 1.605	5088.9	839	897
Arachidonic acid, trimethylsilyl ester	multienoicFA_TMS	3365 , 1.610	4977.9	860	907
Doconexent, TMS derivative	multienoicFA_TMS	3375 , 1.615	3782.9	914	914
7,10,13,16-Docosatetraenoic acid, (Z)-, TMS derivative	multienoicFA_TMS	3390 , 1.620	4778.1	896	897
Eicosapentaenoic Acid, TMS derivative	multienoicFA_TMS	3400 , 1.625	4578.1	876	877
Arachidonic acid, trimethylsilyl ester	multienoicFA_TMS	3432.5 , 1.595	54.074	581	769
9-Tetradecenoic acid, (E)-, TMS derivative	monoenoicFA_TMS	2297.5 , 1.595	9267.5	932	948
13-Methyltetradec-9-enoic acid, TMS derivative	monoenoicFA_TMS	2465 , 1.605	7969.6	813	855
cis-9-Hexadecenoic acid, trimethylsilyl ester	monoenoicFA_TMS	2612.5 , 1.590	11490	899	900
cis-9-Hexadecenoic acid, trimethylsilyl ester	monoenoicFA_TMS	2620 , 1.575	12246	910	910
10-Heptadecenoic acid, (Z)-, TMS derivative	monoenoicFA_TMS	2767.5 , 1.590	9970.3	876	877
Trimethylsilyl (9E)-9-octadecenoate	monoenoicFA_TMS	2907.5 , 1.590	3780.5	920	937
trans-9-Octadecenoic acid, trimethylsilyl ester	monoenoicFA_TMS	2925 , 1.570	10067	912	929

11-Eicosenoic acid, (E)-,TMS derivative	monoenoicFA_TMS	3190 , 1.585	8551.5	883	898
13-Docosenoic acid, (Z)-, TMS derivative	monoenoicFA_TMS	3452.5 , 1.595	7640.5	891	892
15-Tetracosenoic acid, (Z)-, TMS derivative	monoenoicFA_TMS	3695 , 1.620	5686	874	876
L-Leucine, N-(trimethylsilyl)-, trimethylsilyl ester	leucine_2tms	1147.5 , 1.660	18073	903	905
L-Isoleucine, N-(trimethylsilyl)-, trimethylsilyl ester	isoleucine_2TMS	1197.5 , 1.650	22457	894	916
Glycine, N,N-bis(trimethylsilyl)-, trimethylsilyl ester	glycine_TMS	1225 , 1.660	39151	889	891
Trimethylsilyl (9E,12E)-9,12-octadecadienoate	dienoicFA_TMS	2897.5 , 1.600	6814.3	923	947
9,12-Octadecadienoic acid (Z,Z)-, trimethylsilyl ester	dienoicFA_TMS	2912.5 , 1.555	4319.8	923	946
11,14-Eicosadienoic acid, TMS derivative	dienoicFA_TMS	3182.5 , 1.600	5420.5	890	908
13,16-Docasadienoic acid, (Z)-, TMS derivative	dienoicFA_TMS	3447.5 , 1.605	5593.5	853	854
Silane, (dodecyloxy)trimethyl-	alcohol_TMS	1792.5 , 1.410	254.14	823	890
Trimethylsilyl 2-[(trimethylsilyl)amino]propanoate	alanine_2TMS	777.5 , 1.625	19604	910	916
Bis(Trimethylsilyl) 2-[(Trimethylsilyl)amino]succinate	acidicAA	1687.5 , 1.835	16566	882	914
L-Aspartic acid, 3TMS derivative	acidicAA	1690 , 1.830	14313	909	924
L-Glutamic acid, 3TMS derivative	acidicAA	1890 , 1.855	4360.8	870	872
Trimethyl stearate	SatFA_TMS	2972.5 , 1.535	138.31	648	854
Docosanoic acid, trimethylsilyl ester	SatFA_TMS	3517.5 , 1.545	80.095	467	705
Dodecanoic acid, trimethylsilyl ester	SatFA_TMS	1957.5 , 1.620	31098	924	959
Tridecanoic acid, TMS derivative	SatFA_TMS	2140 , 1.600	20595	920	921
Tetradecanoic acid, trimethylsilyl ester	SatFA_TMS	2317.5 , 1.590	23511	921	939
Pentadecanoic acid, TMS derivative	SatFA_TMS	2485 , 1.580	14080	921	922
Trimethyl palmitate	SatFA_TMS	2645 , 1.580	7568.3	928	963
Heptadecanoic acid, TMS derivative	SatFA_TMS	2800 , 1.565	21580	895	920
Trimethyl stearate	SatFA_TMS	2947.5 , 1.570	9388.6	917	960
Arachidic acid, TMS derivative	SatFA_TMS	3225 , 1.565	15089	857	918
Heneicosanoic acid, TMS derivative	SatFA_TMS	3357.5 , 1.570	15982	841	843
Docosanoic acid, trimethylsilyl ester	SatFA_TMS	3482.5 , 1.580	12521	868	886
Trimethylsilyl tricosanoate	SatFA_TMS	3605 , 1.590	13168	846	861
Trimethylsilyl ester of tetracosanoic acid	SatFA_TMS	3722.5 , 1.605	16162	878	914
Trimethylsilyl Hexanoate	SatFA_TMS	717.5 , 1.625	1416.5	917	933
Heptanoic acid, TMS derivative	SatFA_TMS	925 , 1.660	446.39	842	878

Octanoic acid, trimethylsilyl ester	SatFA_TMS	1137.5 , 1.670	1681.1	917	927
Nonanoic acid, TMS derivative	SatFA_TMS	1352.5 , 1.660	4337.4	889	890
8-Methylnonanoic acid, trimethylsilyl ester	SatFA_TMS	1562.5 , 1.645	11862	920	920
Undecanoic acid, TMS derivative	SatFA_TMS	1765 , 1.625	166.1	631	778
Methyl ç-linolenate	LinearTrienoicFAME	2700 , 1.665	2039.4	906	908
8,11,14-Eicosatrienoic acid, methyl ester, (Z,Z,Z)-	LinearTrienoicFAME	3000 , 1.660	2075.6	893	894
Undecanoic acid, methyl ester	LinearSaturatedFAME	1715 , 1.705	236.43	653	837
Tridecanoic acid, methyl ester	LinearSaturatedFAME	1912.5 , 1.685	1516.5	912	951
Tetradecanoic acid, methyl ester	LinearSaturatedFAME	2100 , 1.670	11892	935	938
Pentadecanoic acid, methyl ester	LinearSaturatedFAME	2280 , 1.655	10751	925	936
Hexadecanoic acid, methyl ester	LinearSaturatedFAME	2452.5 , 1.640	35352	930	930
Heptadecanoic acid, methyl ester	LinearSaturatedFAME	2620 , 1.630	10789	911	912
Octadecanoic acid, methyl ester	LinearSaturatedFAME	2777.5 , 1.620	23629	916	919
Methyl icosanoate	LinearSaturatedFAME	3072.5 , 1.615	28121	924	925
Heneicosanoic acid, methyl ester	LinearSaturatedFAME	3212.5 , 1.615	14766	909	912
Docosanoic acid, methyl ester	LinearSaturatedFAME	3345 , 1.620	18749	913	916
Octadecanoic acid, methyl ester	LinearSaturatedFAME	3475 , 1.620	11248	898	941
Tetracosanoic acid, methyl ester	LinearSaturatedFAME	3600 , 1.630	19398	921	940
5,8,11,14-Eicosatetraenoic acid, methyl ester, (all-Z)-	LinearMultienoicFAME	2972.5 , 1.655	1964.8	888	889
Methyl myristoleate	LinearMonoenoicFAME	2077.5 , 1.695	1179	900	901
9-Octadecenoic acid (Z)-, methyl ester	LinearMonoenoicFAME	2260 , 1.675	2455.4	874	880
cis-10-Heptadecenoic acid, methyl ester	LinearMonoenoicFAME	2585 , 1.655	4075.1	914	914
9-Octadecenoic acid (Z)-, methyl ester	LinearMonoenoicFAME	2735 , 1.645	8660.7	923	925
cis-Methyl 11-eicosenoate	LinearMonoenoicFAME	3035 , 1.635	4759	893	893
Cyclohexene, 1-butyl-	LinearDienoicFAME	712.5 , 1.410	197.92	653	777
Naphthalene, decahydro-2-methyl-	LinearDienoicFAME	827.5 , 1.445	88.903	584	807
9,12-Octadecadienoic acid, methyl ester	LinearDienoicFAME	2725 , 1.660	4462.6	915	936
9,12-Octadecadienoic acid, methyl ester	LinearDienoicFAME	2737.5 , 1.615	2530.8	888	908
cis-11,14-Eicosadienoic acid, methyl ester	LinearDienoicFAME	3027.5 , 1.645	4465.8	916	916

For the low concentration, of 108 compounds in the standard mixture, only 14 compounds (8 saturated fatty acids, 2 monoenoic fatty acids, 1 dienoic fatty acid, and 3 carbohydrates) were classified using the scripts. In this work, the term limit of classification (LOC) is used to describe the lowest concentration where the scripts could correctly classify the compound. The LOC varied widely for different classes and even for different compounds within the same class due to the differences in the complexity of characteristic mass spectra features that can be useful in developing scripts. Depending on how unique the fragmentation of the target compound is, scripting can filter out the compounds of interest more effectively using the distinct mass spectral features. As an example, the LOC for a TMS derivative of arachidic acid was determined to be 85.1 pg on column. The compound was detected with a signal-to-noise ratio of 387. The concentration of 85.1 pg on column was the first occurrence of the compound that was classified correctly as a TMS derivative of a saturated fatty acid. The mass spectral match score for this compound at the lowest concentration (21.3 pg on column, S/N 64.22) was 464 for similarity and 732 for reverse (Figure 4-3B), which was significantly lower than 857 for similarity and 918 for reverse in a higher concentration (2.13 ng on column, S/N 15089) standard (Figure 4-3A).

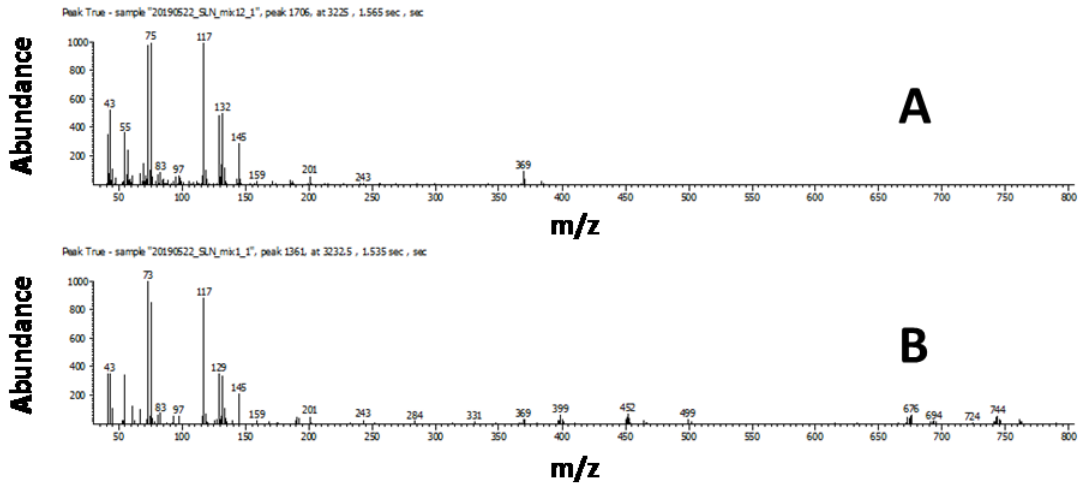

Figure 4-3. Arachidic acid, TMS derivative A) high concentration B) low concentration.

Low concentrations resulted in low quality spectra with higher noise, which hindered the ability of the classifying scripts and resulted in some false negatives. To alleviate this issue, modifications were made to the scripts, which used the logic of eliminating any compound that has more than five prominent peaks (abundance greater than 1% of the base peak) beyond the $[M+2]^+$ peak. The threshold for tolerating noise in the higher masses above $[M+2]^+$ was calculated by taking the average of all the signals of masses above $[M+2]^+$, and four standard deviations above the average was set as the threshold to discriminate the real signals versus noise. This small alteration slightly improved the result of classifying scripts at the lower concentrations; however, it increased the possibility of false-positives due to the increased flexibility for lower intensity peaks at the higher end of the mass spectrum. Nonetheless, as a proof-of-concept, testing the scripts on the standard mixtures at various concentrations revealed its fairly robust performance as an automated, convenient, and rapid screening tool. Once the scripts are written, the accuracy and leniency of such scripts can be improved easily by adding or removing specific features and adjusting the tolerance level for ion ratios, based on need. Furthermore, inserting retention time information in the case of absolute necessity may increase the accuracy of the scripts further, though it sacrifices their versatility.

4.3.2 Versatility of Scripts

The greatest advantage of the proposed scripts is their flexibility; they can be applied to any GC×GC-TOFMS chromatograms, regardless of the conditions used. To validate this benefit, a standard mixture of 37 FAMEs (Table 4-1) was analyzed using two different GC×GC-TOFMS conditions. The chromatograms are shown in Figure 4-4A and 4-4B. The chromatogram in figure 4-4C is a FAME extract from algae, which was analyzed using the same conditions as Figure 4-4B. The chromatograms were processed using the same data processing method with scripts for FAMEs. Table 4-4 shows the results for the two conditions and the algae extracts.

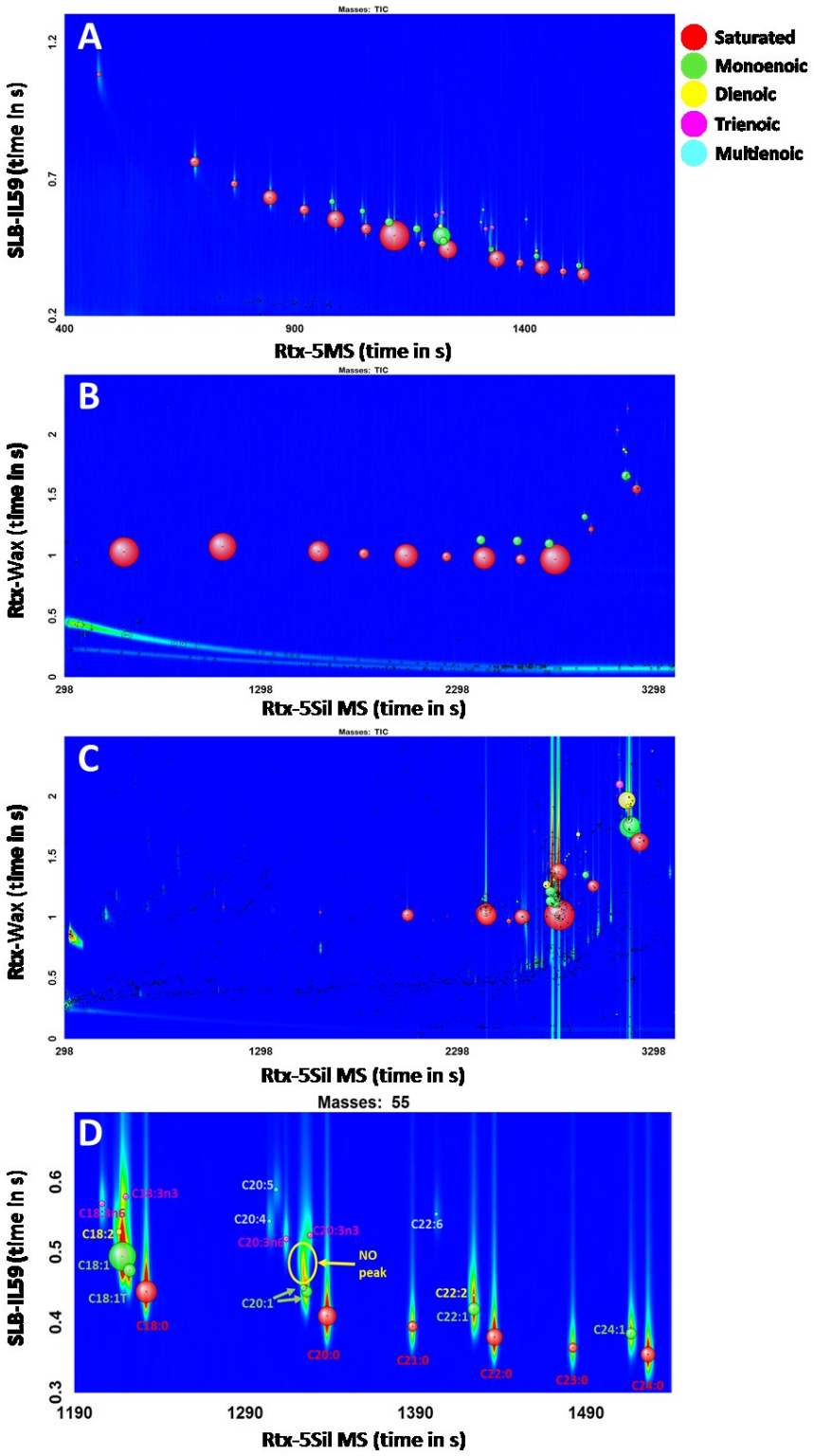

Figure 4-4. A) SUPELCO 37 Component FAME Mix ran with GC×GC-TOFMS Condition 1 **B)** GC×GC-TOFMS Condition 2 **C)** Algae extract with GC×GC-TOFMS Condition 2 **D)** zoomed-in chromatogram of Figure 4-4A. Red bubbles for saturated, light green for monoenoic, yellow for dienoic, pink for trienoic, and light blue multienoic.

Table 4-4. Standard FAMEs Results Using Two Different Conditions and One FAME Algae Extract

FAMEs	Condition 1	Condition 2	Algae Extract
Total	34	21	49
Saturaed	15	11	18
Monoenoic	10	6	11
Dienoic	2	2	11
Trienoic	4	2	6
Multienoic	3	0	3

In Condition 1, thirty-four FAMEs were identified (Figure 4-4A). However, methyl butyrate and methyl hexanoate (short-chain FAMEs) were missed in classification because they eluted before the solvent delay time of the GC×GC-TOFMS conditions that were used to acquire the chromatogram. One isomer of C18:2 and C20:2 were not detected, as the peaks were not well resolved in the region of the chromatogram with other nearby abundant peaks (Figure 4-4D). Both *cis* and *trans* forms of methyl eicosenoate (C20:1) were detected and classified correctly as monoenoic FAMEs, although the SUPELCO certificate only mentions the presence of methyl cis-11-eicosenoate (Figure 4-4D). Twenty-one FAMEs were identified in Condition 2 correctly from C6 to C18; the last compound to be able to elute with the given GC×GC-TOFMS condition was C18 FAME due to the temperature limitation with an insufficient hold time at the end of the run. The use of the PEG phase in the second-dimension restricted the maximum temperature to 230 °C in the ^{1}D and 245 °C in the ^{2}D. Among the detected peaks in the range of C6 to C18 FAMEs, all twenty-one FAMEs that are in the standard mixture were correctly classified into the corresponding groups.

The scripts applied to algae extracts classified forty-nine FAMEs from 981 peaks in approximately 10 min of data processing time. All 49 peaks were identified correctly, despite the evidently overloaded peaks towards the end of the chromatogram (C16 and C18 FAMEs). Algae extract results showed that the scripting tool allowed rapid screening and provided a general understanding of the composition of a given sample in a short time. The scripting tool enables quick visualization of the location of

members of target classes of compounds, while simultaneously offering a rough visual estimation of the concentration of the compounds.

4.3.3 Filtering of Peak Table by Scripts

After the evaluation of the scripts, the developed scripts were applied to four different real-world samples that were prepared with two major sample preparation methods for metabolomics studies, SPME and TMS derivatization. Sweat and fecal samples were prepared with SPME, a method for volatile analysis without derivatization, whereas plasma and urine were prepared with a two-step methoximation / silylation derivatization. The four different samples were analyzed with different GC×GC-TOFMS conditions, each with different column configurations. All four acquired chromatograms (Figure 4-5) were processed with the same scripts, without any special treatment to the data, such as artifact removal (column bleed), in order to fully assess the power of the scripts. The classes of metabolites that were used in the scripts were aldehydes, alcohols, ketones, free fatty acids, fatty acid methyl esters, fatty acid ethyl esters, and isopropyl esters to target for non-derivatized compounds and trimethylsilyl esters of amino acids, fatty acids, sugars, other organic acids, and sterols for the TMS derivatives.

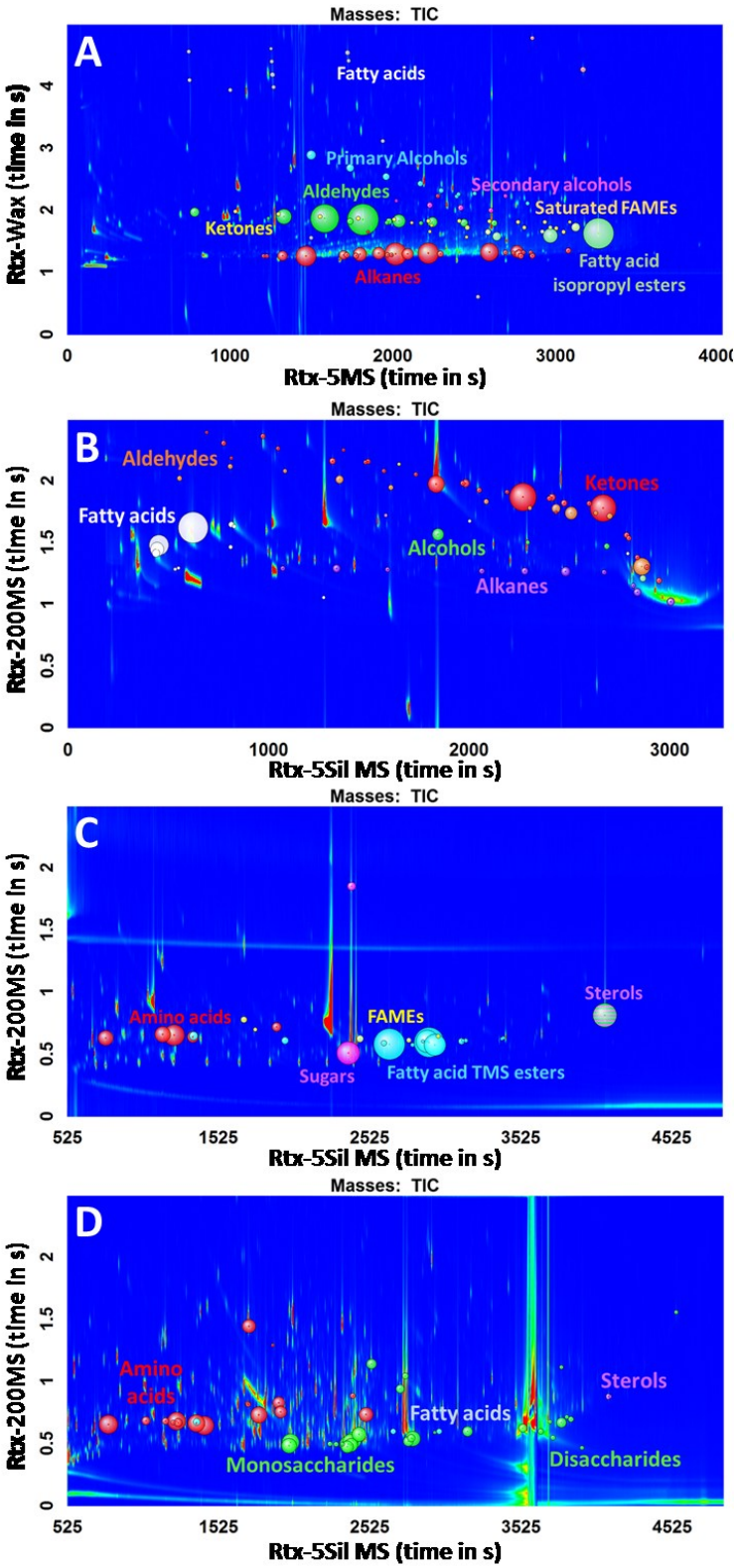

Figure 4-5. Chromatograms of different sample types with the same scripts applied. **A)** Sweat **B)** Feces **C)** Plasma **D**) Urine.

The raw chromatograms of each sample contained thousands of peaks, which make reviewing the data and getting useful information and interpretation of the data a challenge. After the scripting filters were applied, the peak tables were reduced from the original thousands of peaks detected to a few dozen classified compounds, which makes the manual revision of the data more realistic and convenient. The classified peaks for each sample were reviewed manually to verify the accuracy of the scripts. In the process of the manual revision, true-positives (TPs) and false-positives (FPs) for classification were determined. TPs indicate the compounds that are correctly classified to the corresponding class, whereas FPs represent the compounds that should not be flagged but are incorrectly assigned to the class. The number of TPs and FPs for each class of compound was counted for the evaluation of the scripts. To confirm the identity of a peak, the entire mass spectrum of each classified analyte was examined against a library. The ordered structure in GC×GC also helped the compound identification by diagnosing the relative position of the compound in chromatographic space, especially for a homologous series of compounds.

Table 4-5 shows the results of the scripts applied to four different samples. The number of compounds that were classified correctly into the corresponding class was counted, and TPs were recorded, with the number of FPs recorded in parenthesis. Overall, the scripts displayed high accuracy, given that the samples were analyzed with different GC×GC-TOFMS methods. It is noteworthy that for the samples that were extracted with SPME, no compounds were classified as TMS derivatives. On the other hand, for derivatized samples, no compounds were classified as alcohols or free fatty acids, which would have been trimethylsilylated. Although it is not practical to examine every single peak in the sample that contains thousands of peaks to assess the occurrence of true negatives and false negatives, the fact that no compound classified as TMS derivatives for SPME and vice versa for TMS derivatized samples provides some degree of assurance that the scripts can work reliably.

Table 4-5. Results of Sweat, Feces, Plasma, and Urine Samples with Scripting Filters Applied

	SPME		Derivatization	
Sample	**Sweat**	**Feces**	**Plasma**	**Urine**
Total Number of Peaks	**3995**	**1685**	**5104**	**11097**
Aldehydes	12 (1)	12	0	0
1° alcohols	9	5 (2)	0	0
2° alcohols	5	3	0	0
Ketones	15	28	0	0
Free fatty acids	14	12	0	0
FAMEs	5 (1)	0 (1)	3	0
FAEEs	2	1	0	0
Isopropylesters	3	1	0	0
Amino acids (TMS)	0	0	8	18
Free fatty acids (TMS)	0	0	14 (1)	4
Sugars (TMS)	0	0	3	25
Sterol (TMS)	0	0	2	1
Others (TMS)	0	0	0	5
Total Classified Peaks	65 (1)	62 (3)	30 (1)	53

Even with the powerful separation efficiency of GC×GC, coelution is inevitable for complex biological samples. As an example, in the plasma sample, methionine 2TMS and aspartic acid 3TMS coeluted almost perfectly in both first- and second-dimensions (Figure 4-6A). Because both are TMS derivatives (i.e., *m/z* 73 is a common ion in the derivatized products), it must have been difficult to distinguish them as two different peaks even with EIC of m/z 73 and would have been easily missed without careful examination of the data. However, with the scripting tool, they were identified as two distinct compounds, despite their huge difference in intensities, where methionine 2TMS could be obscured by aspartic acid 3TMS (Figure 4-6B).

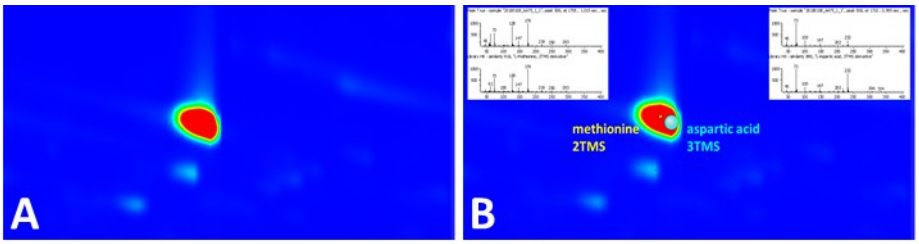

Fig 4-6. A) At *m/z* 73, the region of the chromatogram appeared to have a single intense peak due to complete coelution of two peaks in both 1st and 2nd dimensions. B) Coeluted peaks were classified as two different peaks by script.

4.3.4 Applying Cached Scripts

There are two ways that scripts can be applied to GC×GC data in ChromaTOF®. Scripts so far were used to classify chromatographic peaks that match specific spectral criteria in a peak table. Another way of how the scripts can be applied in ChromaTOF® is with cached scripts. The script function with a cached script returns a numeric value. These calculated "ion traces" are cached as they are created during data processing. After the data is processed, the cached ion traces are plotted and the responses of only the target analytes that match the particular spectral criteria are plotted.

As an example, cached scripts for TMS esters of saturated fatty acids were applied to a derivatized standard mixture. The standard mixture contained over 1000 peaks. Figure 4-7 presents the comparison between TIC (Figure 4-7A) and the cached script result (Figure 4-7B). The EIC using mass channel *m/z* 73 would not be sufficiently selective in derivatized biological samples since *m/z* 73 is a common mass in derivatized products. Using the cached scripts, the TOFMS has been turned into a selective detector for spectra with the desired characteristics, and the surface plot showing only the target analytes allows rapid screening of the sample.

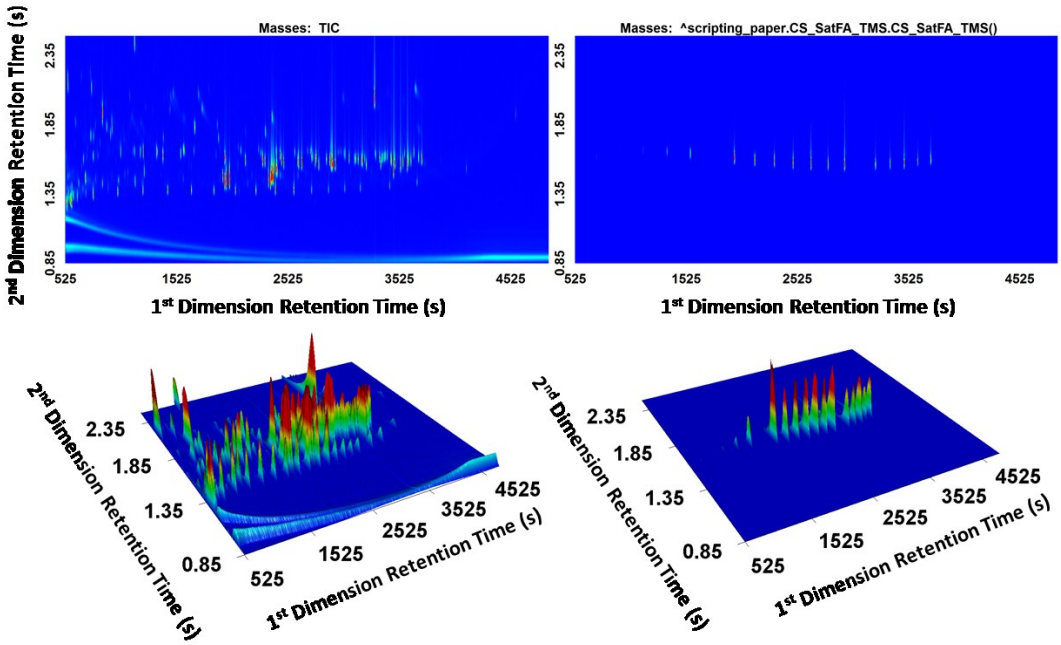

Figure 4-7. Cached script result for generating a selective response for TMS esters of saturated fatty acids.

4.4 Conclusions

Due to the complexity and the amount of data acquired from GC×GC-TOFMS analyses, data handling has been a significant challenge, especially with metabolomics data. In this work, scripting algorithms for numerous classes of metabolites were written using the scripting feature in the LECO ChromaTOF®. To the best of my knowledge, this scripting tool was the first time that automated screening by filtering scripts for handling GC×GC-TOFMS data in metabolomics applications had been reported. The objective of filtering scripts was to visualize quickly the members of different classes of metabolites in samples from thousands of peaks in the GC×GC-TOFMS data and to reduce the size of peak tables for further manual review. It could be considered as the semi non-targeted detection of classes of compounds and offered quick visualization of the members of multiple classes of compounds. The most significant advantage of the developed scripting algorithms was that they are agnostic to the GC×GC-TOFMS configuration and methods used. No retention time information was used in the scripts, which made the scripting approach versatile for many

applications. Further updating and modifying of classification scripts may be needed due to some factors such as the discrepancy of the detector condition for the instrument, which affects the abundance ratio of ions, etc. However, the developed scripts are sufficient to provide the basic skeleton of mass spectral information for the different classes of metabolites, while avoiding over-specifying the filters. Although room for improvement remains, the presented automated scripting proved to be a useful tool for the classification and visualization of different classes of compounds.

CHAPTER 5

Other Applications Using Scripting

5.1 Automated Group-Type Terpene Profiling of Cannabis Using Scripting

5.1.1 Group-Type Analysis

Even with the superior separation power of GC×GC, it may not be possible to differentiate and individually identify chemically similar compounds. A group-type analysis is a useful alternative to classify structurally related compounds, where it is impractical to identify every component in complex samples, but the information on the overall composition of the sample can be valuable.[88] Although determining a particular compound's identity can be challenging, if that compound can possibly be classified as belonging to a compound class, quantitation of group-types may provide adequate compositional information of a sample. Thus, the group-type analysis aims not to identify or quantify individual components but to classify and quantify families of compounds. A group-type analysis of hydrocarbons has been commonly employed in complex petroleum samples to link the compositional information obtained from group-type quantitation to fuel properties.[197,198]

In addition to the enhanced separation provided by GC×GC, it offers chromatograms that are highly structured. The generation of ordered patterns of structurally related compounds is undoubtedly beneficial in determining compound classes. However, compound groups may still overlap to a certain degree, and spectral properties are required to classify the compounds that appear in the same chromatographic regions. Along with the ordered structure of GC×GC chromatograms, data processing with filtering scripts can be an efficient method, particularly for complex samples, for rapid fingerprinting and sample profiling.

5.1.2 Terpene Profiling of Cannabis

Since ancestral times, natural products have received continuous attention in plant science and medicine for their therapeutic and psychological effects.[199] With the

legalization of cannabis use for medicinal or recreational purposes in many jurisdictions, studies into the chemical composition in cannabis and the associated health effects are becoming more prevalent.[200] Cannabis contains various families of compounds, including terpenes, cannabinoids, flavonoids, etc.[201] Different strains provide unique terpene profiles, and the total composition of cannabis is critical in determining the potency and medical effects.[201] The differences in the chemical composition produce different aromatic impressions, which may influence the medicinal properties and are an important factor in consumer preference.[201] Therefore, an efficient way to characterize cannabis is essential for growers, consumers, and industrial monitoring and control.

Terpenes constitute one of the largest and most diverse groups of naturally occurring organic compounds.[202,203] Terpenes are derived fundamentally from a five-carbon isoprene structure.[203] The enzyme called terpene synthase catalyzes complex reactions, such as rearrangements, cyclization, and elimination of carbon skeleton precursors, to synthesize various types of terpenes and their derivatives. As a result, thousands of unique terpenes are yielded, ranging from the simple five-carbon isoprene unit to lengthy polymers.[204] Due to this reason, terpene synthase enzymes are claimed to be the metabolic gatekeepers in generating chemical diversity for the terpene related compounds.[205]

The broad class of terpenes is subdivided primarily on the basis of the number of isoprene units. The subgroups include hemiterpenes (C_5H_8), monoterpenes($C_5H_8)_2$, sesquiterpenes($C_5H_8)_3$, diterpenes($C_5H_8)_4$, sesterterpenes ($C_5H_8)_5$, triterpenes ($C_5H_8)_6$, etc.[199,206] The compounds of terpene origin are predominantly monoterpenes and sesquiterpenes. Terpenes may oxidize over time into terpene alcohols. Terpene alcohols also are present abundantly in nature and make a significant part of plant volatiles, together with monoterpenes and sesquiterpenes.[207] Other oxygenated derivatives of terpenes may be found in the form of acids, aldehydes, ketones, esters, ethers, and phenols due to oxidation or thermal- or UV-induced rearrangements during processing or storage.[201,208] These non-enzymatic modifications add another layer of chemical variations in addition to the variations from the plant genome and

biochemistry.[201] In this work, the term "terpene" was applied to include all terpene-related compounds, including terpenoids, compounds comprised of oxygen-containing functional groups.

The biggest challenges in terpene profiling are huge terpene variations that exist and the lack of standards for many of the terpene compounds. Due to the difficulty of having standards, it is common for terpene analysis reports to rely on a tentative identification or, often, incomplete terpene profile presentations, inevitably leaving a significant portion of the sample unidentified.[201] The use of scripts for group-type analysis can alleviate some of the problems by classifying the structurally similar compounds into the same group. Even the compounds that are not registered in the library database can be classified if they fulfill the search criteria. Thus, it offers a higher potential to provide more complete terpene profiles that can address the great diversity of terpenes in cannabis.

Due to the high volatility, terpene analyses are commonly conducted by gas chromatography (GC), with a headspace sampling technique. Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOFMS) is a more advanced technique that offers high separation efficiency with the additional separation from the second column. It is emerging as one of the preferred tools for the characterization of cannabis terpenes due to its ability to provide superior separations of isomeric species.[200]

In jurisdictions where cannabis use is legalized, cannabis has been consumed for mainly two different purposes: medical or recreational. The different end-use of the product has led to different breeding goals.[201,209] Medicinal strains typically are focused more on developing a specific cannabinoid profile that is more tightly controlled so that therapeutic effects from one dose to the next are as consistent as possible.[210] Recreational strains are bred to enhance flavour, fragrance, and potency. Presented herein is the recent progress towards the development of a series of scripts for the rapid classification of GC×GC-TOFMS peaks for automated group-type terpene profiling. Cannabis samples of different strains (<5 mg) were analyzed by direct thermal desorption, followed by GC×GC-TOFMS. The group-type analysis was conducted

using the scripting approach for medicinal and recreational cannabis comprised of hundreds of terpenes and cannabinoids.

5.1.3 Experimental

5.1.3.1 Standards and Samples

Terpene Standard: 19-components Cannabis Terpenes Standard #1 was purchased from Restek. The 19 compounds (and CAS numbers) include (-)-α-bisabolol (23089-26-1), camphene (79-92-5), δ-3-carene (13466-78-9), β-caryophyllene (87-44-5), geraniol (106-24-1), (-)-guaiol (489-86-1), α-humulene (6753-98-6), p-isopropyltoluene (p-cymene) (99-87-6), (-)-isopulegol (89-79-2), d-limonene (5989-27-5), linalool (78-70-6), β-myrcene (123-35-3), nerolidol (7212-44-4), ocimene (13877-91-3), α-pinene (80-56-8), (-)-β-pinene (18172-67-3), α-terpinene (99-86-5), and γ-terpinene (99-85-4), terpinolene (586-62-9) formulated in isopropanol.

Cannabis Samples: Five different strains of cannabis (three medicinal, two recreational) were analyzed during this study. Three samples of medicinal strains were denoted as M1, M2, and M3, and two recreational cannabis strains were designated as R1 and R2. A minuscule mass (i.e., <5 mg) of unperturbed plant material was added to glass "cup" inserts for thermal desorption tubes in duplicate. For each strain, a pistil (hairlike projection) and a calyx (flower) were separated carefully from bulk vegetal mass prior to TD-GC×GC-TOFMS analysis.

5.1.3.2 Instrumentation

Thermal Desorption (TD) of Cannabis

Thermal desorption of cannabis samples was achieved using a Gerstel Thermal Desorption Unit (TDU2) and a Cooled Injection System-Programmable Temperature Vaporizing (CIS-PTV) inlet for cryogenic focusing of analytes prior to GC×GC separation. Desorption was accomplished in splitless mode, with the following parameters: initial temperature 30 °C (no hold), ramped to 265 °C at 720 °C/min, and held for 10 min with desorption flow of 50 mL/min. The TDU transfer temperature was set to 320 °C. Analytes were focused in the CIS-PTV inlet in solvent vent mode using

the following conditions: initial temperature −30 °C, ramped to 265 °C at 12 °C/min (hold 5 min).

GC×GC-TOFMS Condition

A Leco Pegasus 4D GC×GC-TOFMS (Leco Instruments, St. Joseph, MI) with a 30 m × 0.25 mm; 0.25 μm df Rtx-5MS (Chromatographic Specialties) column as the first dimension and a 1.7 m × 0.25 mm; 0.25 μm df (Stabilwax) column as the second dimension were used. Helium (5.0 grade; Praxair, Edmonton, AB) was used as the carrier gas at a constant flow of 2 mL/min. The primary GC oven was programmed from 40 °C, held for 3 min, and ramped at 12.88 °C/min to 255 °C, with a final hold of 2.5 min. Relative to the primary oven, the secondary oven was programmed to have a constant offset of +15 ºC and the modulator a constant offset of +15 ºC. The modulation period was 1.3 s (0.2 s hot, 0.45 s cold). The parameters used for mass spectrometry are as follows: electron energy of –70 eV; acquisition rate of 200 Hz; mass range of *m/z* 25–500; detector voltage of –1756 V, with the optimized voltage offset of 200 V; ion source temperature of 200 ºC; the MS transfer line temperature was 240 ºC. The total GC run time per analysis was 22.193 min. The sample was injected with a split ratio of 1:20.

5.1.3.3 Data Processing and Automated Classification Using Filtering Scripts

GC×GC-TOFMS data were processed with ChromaTOF® (v.4.72; LECO), commercial software from LECO. Automated peak finding with subsequent deconvolution of mass spectra was performed by ChromaTOF®, with a peak-finding threshold of S/N 50. The baseline offset was set to 0.9, and the peak widths throughout the entire chromatographic run were set to 10 s for the first dimension and 0.12 s for the second dimension. All chromatographic peaks were searched against the NIST-MS 2017 libraries. The optional scripting feature was enabled in the ChromaTOF®, and the scripts written for the different terpene classes and cannabinoids were added in classifications. The scripts were applied to the entire chromatographic space without selecting a specific region. The calculations were performed using Excel 2013, Microsoft Office 2013.

5.1.4 Scripts for Terpenes

Terpenes isomerize very easily.[211-213] For example, there are more than 20 isomers of monoterpene hydrocarbons that have a chemical formula of $C_{10}H_{16}$.[214] Monoterpenes may be linear (acyclic) or may contain rings. Whether they are acyclic or cyclic monoterpenes, all $C_{10}H_{16}$ monoterpene hydrocarbons have the molecular ion of *m/z* 136, with very similar fragment ion compositions (Figure 5-1).[214] The m/z of fragment ions are identical with trivial differences in intensities.

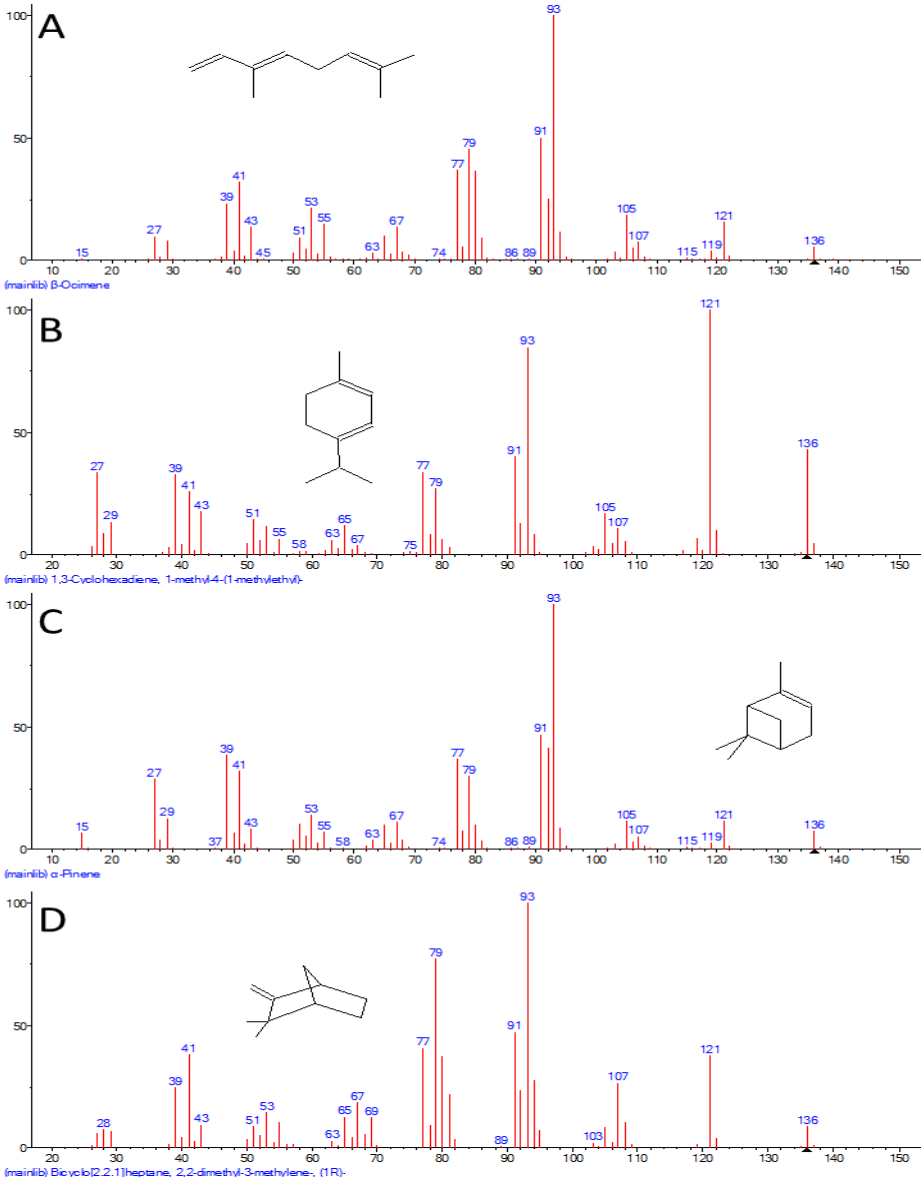

Figure 5-1. Mass spectrum of different $C_{10}H_{16}$ monoterpene hydrocarbons from the NIST 2017 library. A: β-Ocimene, B: α-Terpinene, C: α-Pinene, D: Camphene.

Monoterpenes of $C_{10}H_{16}$ can be characterized by the major fragment ion at *m/z* 93, which results from the loss of $CH(CH_3)_2$ either from hydrogen ion (H^+) of the methyl group migrating to C1 or via the sequential dissociation of CH_3 and C_2H_4 (Figure 5-2). The resulting $C_7H_9^+$ is a stable cation and gives a high rise of the *m/z* 93 peak; often, it is a base peak in the mass spectra of the monoterpene hydrocarbons.[214] The cleavage of H_2 from the highly abundant *m/z* 93 fragment ion ($C_7H_9^+$) results in the formation of a benzyl carbocation. The benzyl carbocation rearranges to give a tropylium carbocation with *m/z* 91($C_7H_7^+$), which also is present in substantial abundance. In addition to these major fragment ions, $C_{10}H_{16}$ monoterpene compounds also give rise to characteristics peaks at *m/z* 39 ($C_3H_3^+$), 51($C_4H_3^+$), 65($C_5H_5^+$), 77($C_6H_5^+$), 79($C_6H_7^+$), 105($C_8H_9^+$), 107($C_8H_{11}^+$), 119($C_9H_{11}^+$), and 121($C_9H_{13}^+$). The oxidation products of terpenes have a characteristic neutral loss of OH or H_2O, which leads to fragment ions of $[M-17]^+$ or $[M-18]^+$. Sesquiterpenes containing 15 carbon atoms can be characterized by *m/z* 43, 53, 55, 67, 77, 79, 119, 133, 134, 136, 147, 148, 161, 175, 189, and 204, according to the work of Hirose.[215]

In this work, due to the lack of terpene standards that can cover the diversity of terpenoids, instead of developing the terpene scripts from the experimental mass spectra of the terpene standards, the selection rules were determined for each terpene family, based on library mass spectra. The development of the terpene scripts was aided by an understanding of the general fragmentation pathways giving rise to the characteristic fragment ions observed in the spectra. As a general rule in developing scripts, over-specifying the filter by setting the tolerance range too narrow was avoided. The scripts were evaluated first with the terpene standard and the marijuana smoke data that was published by Graves et al.[200] and then were adjusted to filter out undesired peaks that were accepted with the initial implementation of the scripts. After examination of the terpene scripts, they were applied to the chromatograms of five different strains of cannabis for a group-type terpene profiling comparison.

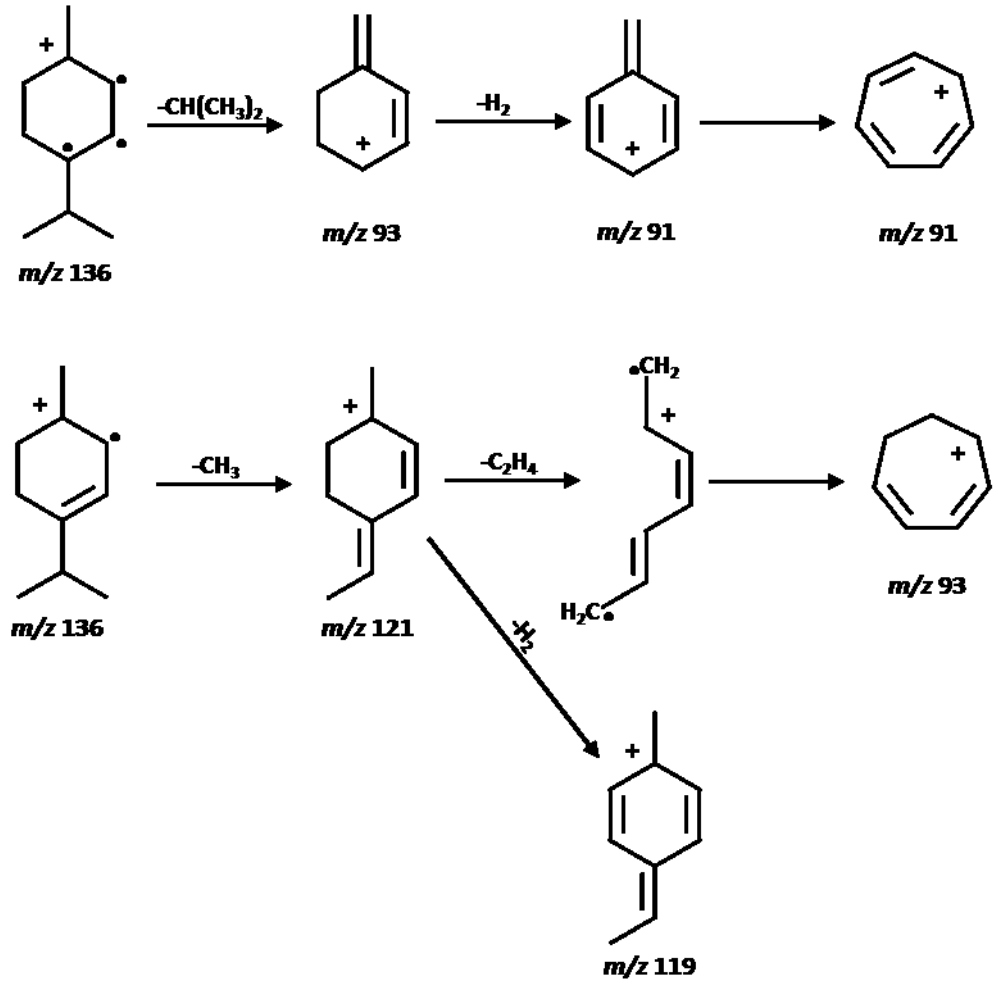

Figure 5-2. Fragmentation pathway of monocyclic terpene (α-terpinene, $C_{10}H_{16}$). The mechanism in the fragmentation pathway was borrowed from Yermakov et al. and the figure was reconstructed.[214]

5.1.5 Results and Discussion

5.1.5.1 Evaluation of Scripts

To evaluate the performance of the developed scripts, a 19-component cannabis terpenes standard mixture was analyzed by GC×GC-TOFMS, and the developed scripts for structurally related terpene groups were applied to the chromatogram. Each terpene subgroup was assigned with a different color-code when the scripts were applied for classification. The processed chromatogram was displayed as a contour plot, with the "bubbles" feature enabled (Figure 5-3). With this feature, the classified compounds were displayed with the color corresponding to the class, and the radii of the bubbles

correspond to the area of the peak. This visual representation of the data allows rapid screening of a sample and quickly provides information about the presence of compounds/classes of interest, with a rough estimation of their abundance.

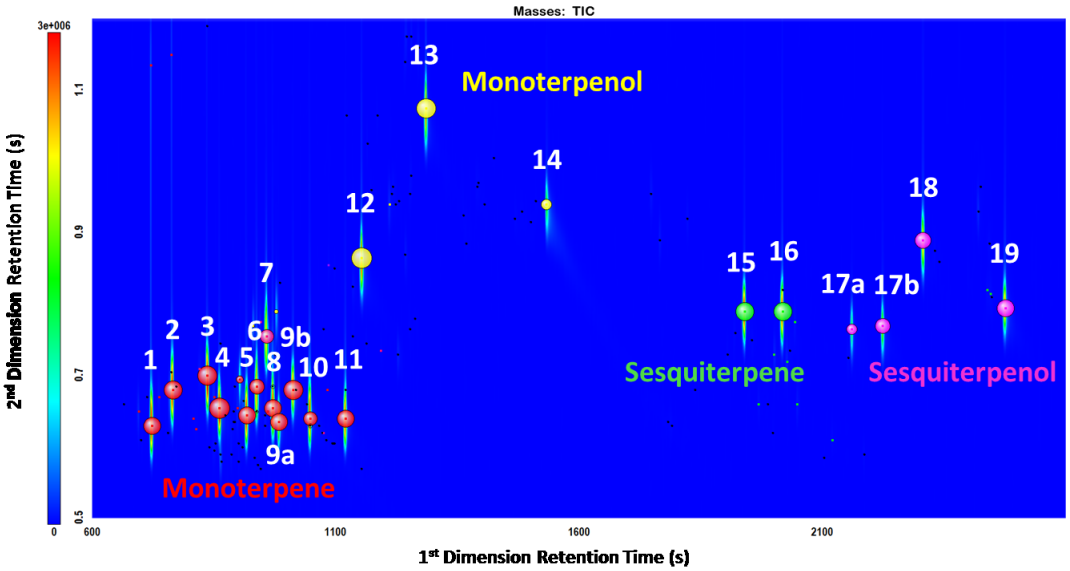

Figure 5-3. A chromatogram of a 19 component cannabis terpene standard mixture resulted from processing with the developed terpene scripts. All 21 compounds, including two isomers, were classified correctly by the scripts.

The data processing with the scripts also generates a peak table containing information such as the analyte names, retention times, and peak area. The peak table was sorted to prioritize the compounds that are classified as terpene groups through the scripts. The result showed that the written scripts classified all 21 terpenes (including isomers of Ocimene and Nerolidol) successfully into the respective subgroup (Table 5-1). This revealed the advantage of scripts as a promising tool for finding and filtering even unknown compounds that carry similar mass spectral properties. This is helpful when screening samples without the need to have a priori knowledge.

Table 5-1. A Peak Table of a 19 Component Cannabis Terpene Standard Mixture

Peak #	Analyte (CAS)	Formula	Classification	tR1, tR2	Unique Mass	Similarity
1	alpha-Pinene (80-56-8)	C10H16	monoterpenes	722.5 , 0.630	93	943
2	Camphene (79-92-5)	C10H16	monoterpenes	765 , 0.680	93	946
3	(-)-beta-Pinene (18172-67-3)	C10H16	monoterpenes	835 , 0.700	93	934
4	beta-Myrcene (123-35-3)	C10H16	monoterpenes	860 , 0.655	69	943
5	delta-3-Carene (13466-78-9)	C10H16	monoterpenes	917.5 , 0.645	93	921
6	alpha-Terpinene (99-86-5)	C10H16	monoterpenes	937.5 , 0.685	121	944
7	p-Isopropyltoluene (p-cymene) (99-87-6)	C10H14	monoterpenes	957.5 , 0.755	119	929
8	d-Limonene (5989-27-5)	C10H16	monoterpenes	970 , 0.655	68	942
9a	Ocimene (13877-91-3)	C10H16	monoterpenes	982.5 , 0.635	93	918
9b			monoterpenes	1012.5 , 0.680	93	941
10	gamma-Terpinene (99-85-4)	C10H16	monoterpenes	1047.5 , 0.640	93	953
11	Terpinolene (586-62-9)	C10H16	monoterpenes	1120 , 0.640	121	931
12	Linalool (78-70-6)	C10H18O	monoterpenol	1152.5 , 0.865	71	940
13	(-)-Isopulegol (89-79-2)	C10H18O	monoterpenol	1285 , 1.075	71	934
14	Geraniol (106-24-1)	C10H18O	monoterpenol	1532.5 , 0.940	41	922
15	beta-Caryophyllene (87-44-5)	C15H24	sesquiterpenes	1940 , 0.790	93	944
16	alpha-Humulene (6753-98-6)	C15H24	sesquiterpenes	2017.5 , 0.790	93	933
17a	Nerolidol (7212-44-4)	C15H26O	sesquiterpenol	2160 , 0.765	69	938
17b			sesquiterpenol	2222.5 , 0.770	69	935
18	(-)-Guaiol (489-86-1)	C15H26O	sesquiterpenol	2305 , 0.890	59	923
19	(-)-alpha-Bisabolol (23089-26-1)	C15H26O	sesquiterpenol	2475 , 0.795	69	929

5.1.5.2 Application of the Scripts to Marijuana Smoke and Essential Oil

To validate their performance, the developed scripts were applied to a GC×GC-TOFMS chromatogram of marijuana smoke obtained from the study that was conducted in our lab and published by Graves et al;[200] the details of sample preparation were discussed in the paper.[200] In brief, the mainstream marijuana smoke was obtained using a smoke machine and following the puff routine of Health Canada Intense (55 mL puff of 2 s duration, every 30 s). A total of 5913 peaks were detected when processed with S/N 50. In Figure 5-4, each black marker represents a chromatographic peak, and the compounds that were classified through the scripts were color-coded with the assigned color for the respective class for better visualization. The developed scripts classified the detected analytes into various terpene subgroups: monoterpenes, monoterpenols, monoterpene aldehydes, monoterpene ketones, sesquiterpenes, sesquiterpenols, and cannabinoids. With the scripting filters, the peak table was reduced to 183 compounds, classifying them as a terpenes family. This corresponds to 3.1% of the size of the original peak table.

The same scripts were also applied to a GC×GC-TOFMS chromatogram acquired from a home remedy essential oil mix for indigestion. From a total of 2911

peaks detected, 140 peaks were classified as terpene groups; this corresponds to 4.8% of the original data. A significantly reduced number of compounds allows for a much more manageable manual review of the data. Scripting is an effective approach for the visualization of the presence of target chemical classes, providing a rough understanding of the chemical composition quickly.

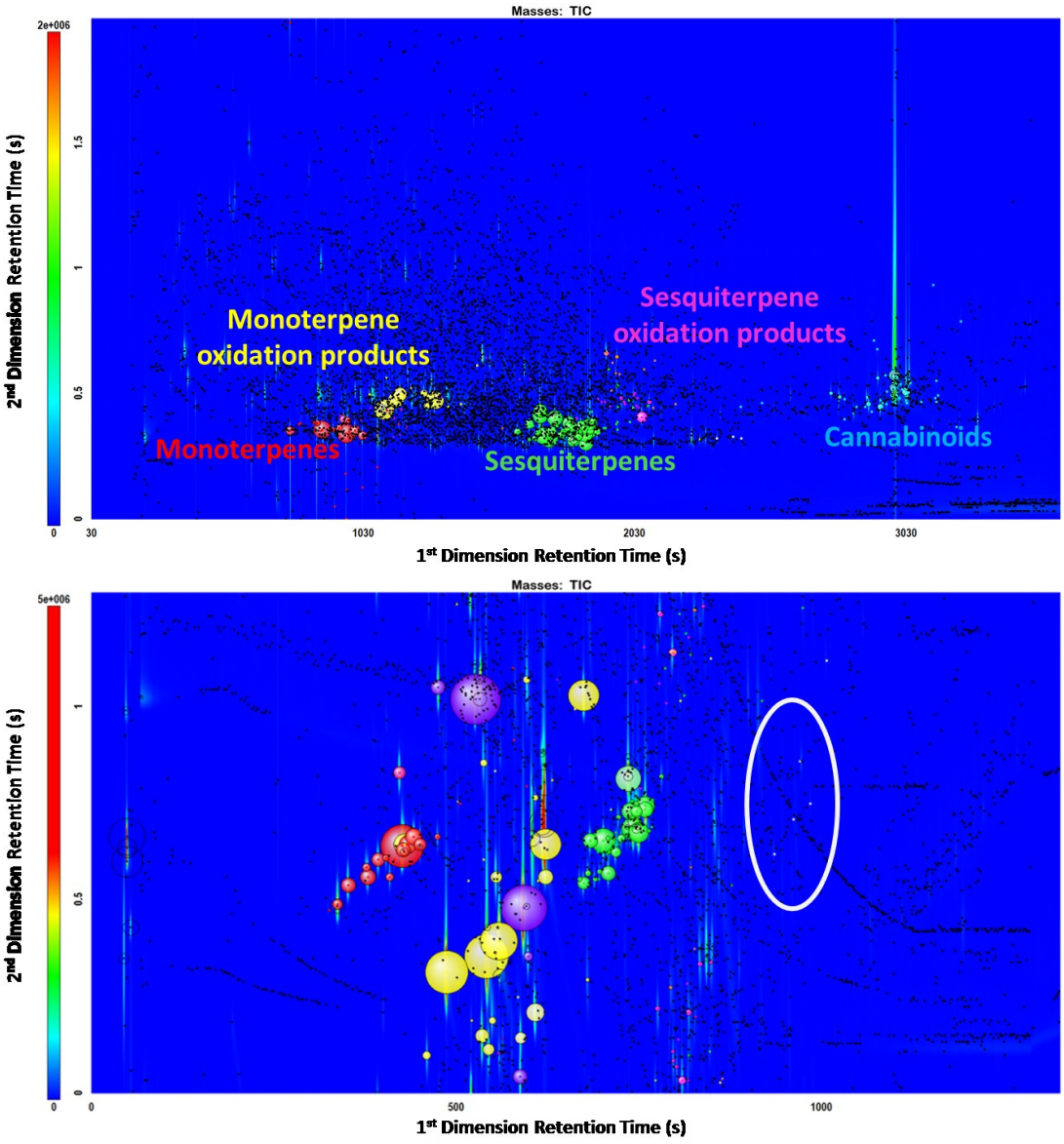

Figure 5-4. GC×GC-TOFMS chromatogram of a marijuana smoke sample (top) and essential oil remedy mix for indigestion (bottom) shown here as a contour plot with bubbles.

The accuracy of scripting algorithms was quantified by manual revision of the filtered peak table for each sample. The mass spectra of each filtered chromatographic peak were examined one-by-one, along with the relative position of the compound in the chromatogram to determine if the assignment to the class by the classifying scripts was correct or not. For example, there were five compounds that were classified as diterpenes for the essential oil. They all appeared in the region that was circled in Figure 5-4. The data processing method was set up to display the top 10 library matches for each peak. The hit table for one of the five compounds that are classified as diterpene in the essential oil sample is shown as an example in Table 5-2. Due to the absence of a distinct molecular ion peak in the mass spectra and common fragmentation ions with just slight differences in abundances among sesquiterpene oxidation products, diterpene oxidation products, and diterpenes, the hit table displays the possibility of the compound being either of three groups with similar library match scores. In this case, the location of the compound in the chromatographic space was used to determine the class of the compounds. GC×GC offers a structured chromatogram, and since these compounds elute in the region that is expected to be a diterpene region with the elution pattern, they were considered that it was correctly classified as diterpene.

Table 5-2. List of Top 10 Library Match for a Single Chromatographic Peak Classified as a Diterpene

Hit	Compound Name	Formula	Similarity	Reverse	CAS
1	Nerolidol	C15H26O	790	859	0-00-0
2	Geranyl Linaloo Isomer B	C20H34O	787	813	0-00-0
3	1,6,10,14-Hexadecatetraen-3-ol, 3,7,11,15-tetramethyl-, (E,E)-	C20H34O	785	802	1113-21-9
4	1,6,10,14-Hexadecatetraen-3-ol, 3,7,11,15-tetramethyl-, (E,E)-	C20H34O	785	802	1113-21-9
5	(6Z)-3,7,11-Trimethyl-1,6,10-dodecatrien-3-ol	C15H26O	784	852	142-50-7
6	(6E)-3,7,11-Trimethyl-1,6,10-dodecatrien-3-ol	C15H26O	780	848	7212-44-4
7	(E,E,E)-3,7,11,15-Tetramethylhexadeca-1,3,6,10,14-pentaene	C20H32	779	834	77898-97-6
8	(E,E,E)-3,7,11,15-Tetramethylhexadeca-1,3,6,10,14-pentaene	C20H32	779	834	77898-97-6
9	(E,E,E)-3,7,11,15-Tetramethylhexadeca-1,3,6,10,14-pentaene	C20H32	778	833	77898-97-6
10	1,6,10-Dodecatrien-3-ol, 3,7,11-trimethyl-, [S-(Z)]-	C15H26O	777	838	142-50-7

Through the manual revision, true-positives (TPs) and false-positives (FPs) were identified. As mentioned previously in Chapter 4, TPs are meant to indicate the compounds that are classified correctly, whereas FPs represent the compounds that should not be assigned to the corresponding class. The number of TPs and FPs for both samples were counted to evaluate the accuracy of the scripts. The accuracy of the scripts is highly dependent on the tolerance of the filter. The tighter the logic rules of the filter specified, the more peaks may be excluded. On the other hand, overly flexible rules result in included false peaks. The tolerance level of scripts would depend on the choice of the user, based on the goal. For this study, the scripts were written without over-specifying the filters to include all potential peaks. This was done in view that the uninteresting peaks that are included by the filter may be removed later through the manual revision with a reduced size of the peak table to review. It is noteworthy that despite the attempts to avoid over-specifying, there were no false positives for cannabinoids in the essential oil. The performance of the developed scripting algorithms is summarized in Table 5-3. Overall, the developed scripts displayed high accuracy (85% and 93%) in two different samples that were analyzed with different GC×GC-TOFMS conditions.

Table 5-3. Summary of Compound Classified in Marijuana Smoke and Essential Oil by Scripts

Compound Class	Marijuana Smoke			Essential Oil		
	Compounds found by scripts	False Positives	True Positives (accuracy)	Compounds found by scripts	False Positives	True Positives (accuracy)
Monoterpenes	32	0	32 (100%)	21	1	20 (95%)
Monoterpene oxidation products	26	5	21 (81%)	41	1	40 (98%)
Sesquiterpenes	51	6	45 (88%)	65	7	58 (89%)
Sesquiterpene oxidation products	29	2	27 (93%)	8	1	7 (88%)
Diterpenes	3	0	3 (100%)	5	0	5 (100%)
Cannabinoids	42	12	30 (71%)	0	0	N/A
Total	183	25	158 (85%)	140	10	140 (93%)

5.1.5.3 Comparison of Terpene Profiles for Different Cannabis Strains

Five different strains of cannabis (three medicinal and two recreational) were analyzed using TD-GC×GC-TOFMS, and the acquired chromatograms were processed with the developed scripts for the profiling of terpene compositions. On average, approximately

3000 peaks were detected for each sample. With the scripting filters, the compounds that meet the search criteria were classified into the corresponding classes. From the five different strains of cannabis, the scripts developed for terpene profiling enabled the classification of 142 compounds of the terpene family, which include 29 monoterpene hydrocarbons, 15 monoterpene alcohols, 13 monoterpene ketones and oxides, 5 monoterpene aldehydes, 37 sesquiterpene hydrocarbons, 27 sesquiterpene oxidation products, 3 diterpenes, and 13 cannabinoids. The list of the compounds that are classified through the scripts is included in Appendix A Table A-2. The chromatograms of 5 cannabis samples with filtered scripts are shown in Figure 5-5.

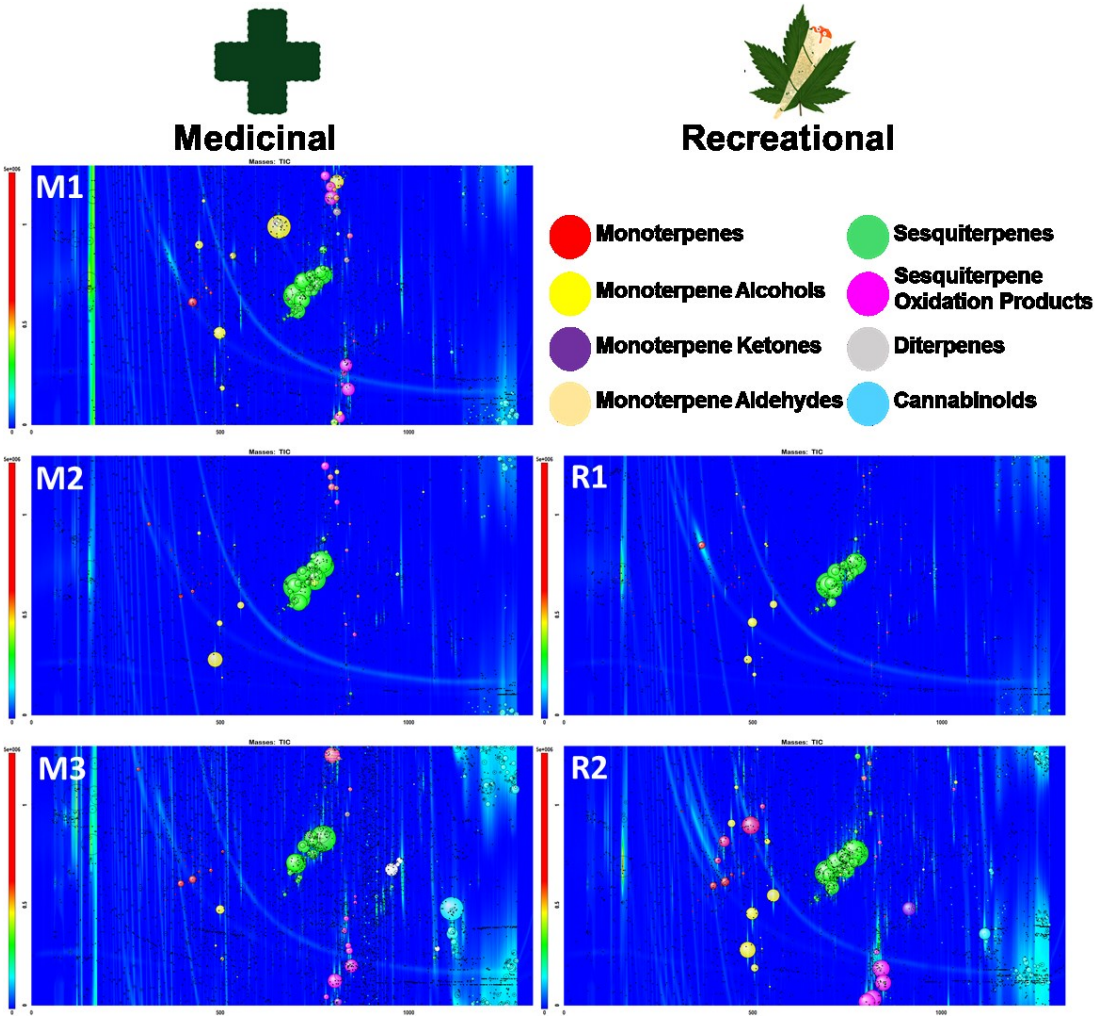

Figure 5-5. Comparison of terpene compositions of cannabis from different strains analyzed by TD-GC×GC-TOFMS. Scripts

The cannabis industry, and in particular producers for the recreational market, is in need of an analytical method for profiling terpenes, both qualitatively and quantitatively. In attempts to obtain a group-type quantification of the terpene composition for different cannabis strains, the peak areas of the compounds that belong to the same class were summed, then divided by the total peak area (the sum of peak areas of all the peaks that were detected in a sample). This calculation yielded the percent contents of the compound class, which is useful to gain a more in-depth insight into the quantitative distribution of different compound classes in a sample. The results that compare terpene compositions for different cannabis strains are illustrated in Figure 5-6. It is evident that different cannabis samples display varying chemical compositions. It is noteworthy how recreational cannabis samples (R1, R2) contain more monoterpenes and sesquiterpenes compared to medicinal cannabis (M1, M2, M3). Cannabinoids are the compounds that are primarily responsible for psychoactive and medicinal effects in cannabis, while other terpene volatiles, such as monoterpenes and sesquiterpenes, are main flavor compounds, which are more important for the recreational cannabis market, as they contribute to different fragrance and aroma attributes, which are essential for consumer preferences. The group-type quantification results using the scripts, revealing higher contents of monoterpenes and sesquiterpenes in recreational cannabis, are in agreement with the general knowledge about the samples.

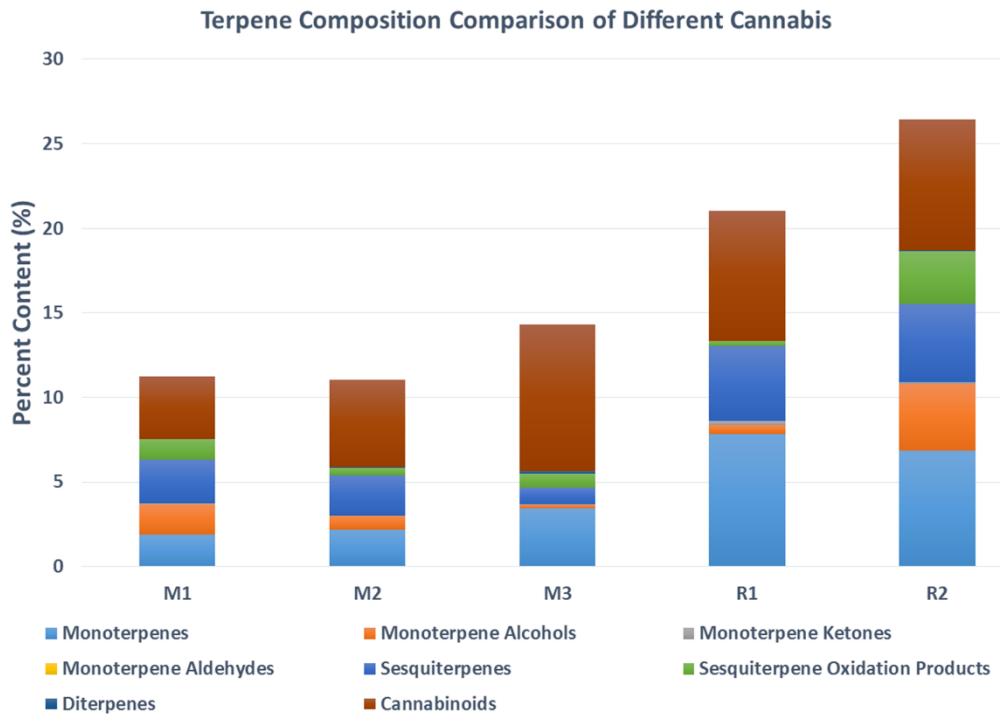

Figure 5-6. Comparison of terpene compositions of cannabis from different strains analyzed by TD-GC×GC-TOFMS. Scripts were used to classify the compounds, and the percent content was calculated using peak areas.

5.1.6 Conclusions

The young legal cannabis industry requires the development of robust technologies to support, for example breeding programs and quality control for flavour/aroma compounds. One challenge facing the industry is its lack of standardization. However, in order to standardize strains of cannabis, tools are required for effective phenotyping of the strains. Likewise, tools for phenotyping are needed to support breeding programs and the development of new strains. GC×GC-TOFMS, in conjunction with scripting, provides an automated, rapid data screening and visualization tool that meets these needs. The developed scripts displayed a high accuracy of classification for various classes of terpenoids. The peak areas of the classified compounds for each target class were summed to examine the terpene contents in different cannabis samples quantitatively. The results showed that different cannabis strains exhibit a wide variability of terpene contents. It also was observed that for the samples investigated, recreational cannabis contained higher terpene levels, especially monoterpenes and

sesquiterpenes, which are major contributors to flavors and aroma. Scripting, in conjunction with GC×GC-TOFMS, is a promising tool for an automated group-type profiling of cannabis.

5.2 Scripting Using a Different Software for Chlorinated Species

5.2.1 Extending Scripting to a Different Platform

The works described until now in this thesis have shown that the GC×GC-TOFMS with scripting approach facilitates a rapid and convenient screening of samples and useful as an investigative tool to evaluate samples. All the scripting works discussed earlier in Chapter 4 and Chapter 5.1 have been performed using the ChromaTOF®, a commercial software developed by LECO. ChromaTOF® is the first commercial mass spectrometry data system designed for comprehensive two-dimensional GC data. After releasing one of the first GC-TOFMS instruments in the market approximately 20 years ago.[216] LECO continuously improved not just for hardware but also the software to aid with data handling. In addition to the basic features of the software, classification and scripting features were introduced as a special package that allows users to write and apply scripts conveniently. These added features in the software reduced the data evaluation process and made data analysis convenient when implemented appropriately. In ChromaTOF®, the algorithms are written in Microsoft VBScript language.

LECO and its proprietary software, ChromaTOF®, have been dominating in the field of comprehensive two-dimensional GC over the decades. Recently, more commercial platforms for GC×GC-TOFMS data have emerged. Scripting, programming of a series of logical rules and constraints based on fragmentation patterns and/or retention information, is not restricted to specific software. Scripting of search parameters can be performed in various computer programs and programming languages (e.g. MATLAB®, Java).[187,217,218] Reichenbach et al. reported development and implementation of the Computer Language for Identifying Chemicals (CLIC) in software with a calculator-like graphical user interface (GUI), applied over

chromatograms produced by GC Image™ software.[219] Despite the development of many computer platforms to aid with the analysis of complex GC×GC-TOFMS data, still, the most common way to implement scripting is using ChromaTOF®. The primary reason for the persistent popularity of the software is that ChromaTOF® generates deconvoluted data with retention times and mass spectrometric information that scripts can be applied to, providing full integration in a single software platform.

With scripting gaining increasing attention as a powerful tool for data analysis, Chromspace® (Sepsolve), a relatively newer commercial GC×GC-TOFMS software package, also incorporates the scripting feature. In order to learn the skills needed for writing scripts before developing a series of scripts to support omics analyses, as a positive control of my ability to write scripts, I started with implementing scripts for chlorinated species on Chromspace®, following the work of Hilton et al.[87] performed on ChromaTOF®.

5.2.2 Scripting for Chlorinated Species

A few studies have reported automated screening approaches for halogenated compounds in various samples using scripting.[145,182,220] The scripts for halogenated species are relatively straightforward due to the distinctive isotope cluster distributions resulting from the natural isotope abundances of chlorine and bromine. Chlorine and Bromine have two major naturally occurring isotopes, and the abundance ratio of these isotopes provides a particularly useful characteristic to classify halogenated species in a sample.

The scripts that have been developed for chlorinated species commonly are based on locating a molecular ion and testing it to determine if it matches the abundance ratio patterns expected for compounds containing specific numbers of chlorine atoms. Chlorine mainly exists as ^{35}Cl and ^{37}Cl, with an isotopic ratio of 3.129 to 1.000.[221] When there are multiple chlorine atoms, there will be peaks from all combinations of each isotope (Figure 5-7). The isotopic cluster pattern begins with the molecular ion, M, and includes isotopic peaks separated by 2 u (*m/z* M+2, M+4, M+6, …), based on a binomial expansion.[87,222] Figure 5-7 shows a graphical form of such an expansion for

the calculation of the coefficients of the isotope patterns of chlorinated species. For example, compounds containing one chlorine atom have a M+2 peak that is one-third of the peak height of the molecular ion. For compounds containing two chlorine atoms, the peak height ratio of M:M+2:M+4 would be 9:6:1. Additionally, masses between the M, M+2, M+4, etc. (i.e., minor M+1, M+3, etc.) peaks must be less abundant than the adjacent masses. The calculations in Figure 5-7 did not account for the effects of other isotopes entirely, with the assumption that the relative abundance of the M, M+2, M+4, … ions were attributable primarily to the isotopes of chlorine, and the contributions of other isotopes to these molecular ion peak cluster intensities were assumed to be insignificant for the purpose of identifying chlorinated compounds.

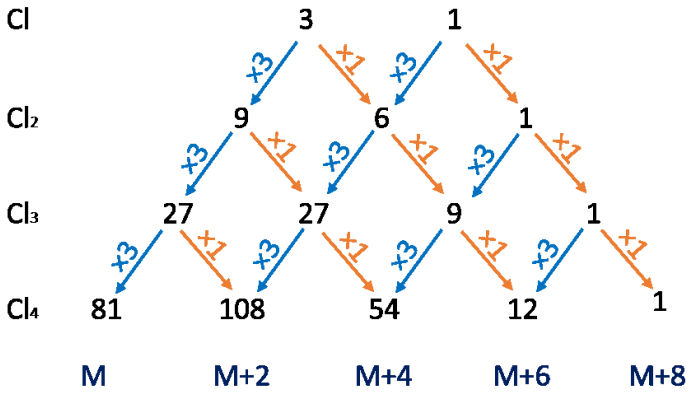

Figure 5-7. Isotopic distributions of chlorine-containing species using the binomial expansion.

The 70 eV EI mass spectrum of methylene chloride, a dichloro compound, is shown as an example for interpreting mass spectral information in Figure 5-8. The critical features useful for the scripts include the three-ion cluster of the masses 84, 86, and 88, with the ratio of 9:6:1. The base peak of *m/z* 49 is due to the loss of one chlorine atom (Cl^{35}) from M^+ through simple bond cleavage. $[M-33]^+$ and $[M-37]^+$ ion clusters also are present due to the loss of one chlorine atom from the molecular ion cluster. The presence of characteristic isotopic clusters separated by 2 nominal atomic mass units and knowledge of the relative abundances are helpful features for writing the scripts.

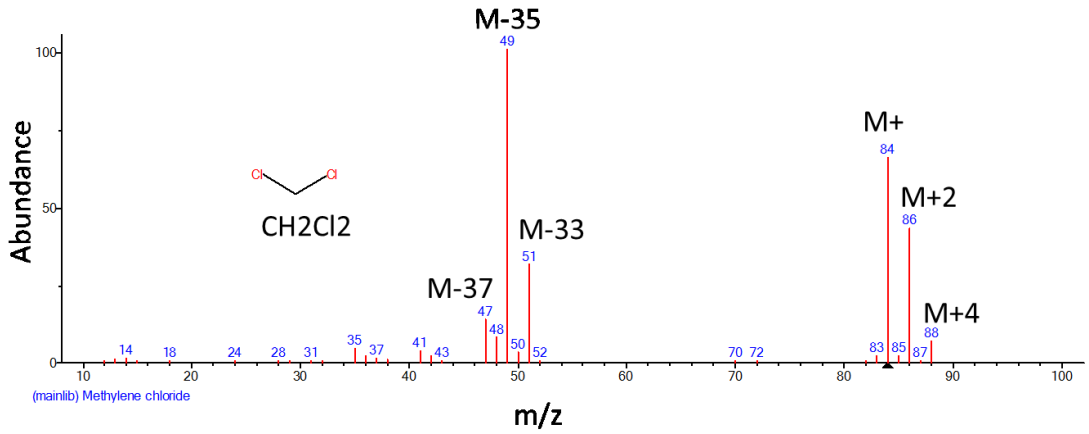

Figure 5-8. Mass spectrum of methylene chloride and important mass spectral features, including the chlorine isotope abundance pattern.

5.2.3 Experimental

5.2.3.1 Standards and Samples

Two standards purchased from Absolute Standards, Inc. were used in the study:

Standard A: 78-components EPA Method 524.2 (CRM 33003). Of the 78 components, 40 compounds contain at least one chlorine atom and are given in Table 5-4.

Standard B: 22-components EPA Method 8121 Analytes (CRM 81202), given in Table 5-5.

Table 5-4. List of 40 chlorine-containing compounds in the CRM 33003 (Absolute Standards, Inc)

#	Name	CAS	#	Name	CAS
1	1-Chlorobutane	109-69-3	21	1,2-Dichloroethane	107-06-2
2	1,1-Dichloropropanone-2	513-88-2	22	1,2-Dichloropropane	78-87-5
3	3-Chloropropene (allyl chloride)	107-05-1	23	1,3-Dichloropropane	142-28-9
4	Chloroacetonitrile	107-14-2	24	1,1-Dichloropropene	563-58-6
5	Hexachloroethane	67-72-1	25	cis-1,3-Dichloropropene	10061-01-5
6	Pentachloroethane	76-01-7	26	trans-1,3-Dichloropropene	10061-02-6
7	Bromodichloromethane	75-27-4	27	Hexachloro-1,3-butadiene	87-68-3
8	Dibromochloromethane	124-48-1	28	1,1,1,2-Tetrachloroethane	630-20-6
9	cis-1,2-Dichloroethene	156-59-2	29	1,1,2,2-Tetrachloroethane	79-34-5
10	trans-1,2-Dichloroethene	156-60-5	30	1,1,2-Trichloroethane	79-00-5
11	Methylene chloride	75-09-2	31	Trichloroethene	79-01-6
12	1,1-Dichloroethene	75-35-4	32	1,2,3-Trichloropropane	96-18-4
13	Bromochloromethane	74-97-5	33	1,2,3-Trichlorobenzene	87-61-6
14	Carbon tetrachloride	56-23-5	34	1,2,4-Trichlorobenzene	120-82-1
15	Chloroform	67-66-3	35	Chlorobenzene	108-90-7
16	1,1-Dichloroethane	75-34-3	36	2-Chlorotoluene	95-49-8
17	2,2-Dichloropropane	594-20-7	37	4-Chlorotoluene	106-43-4
18	Tetrachloroethene	127-18-4	38	1,2-Dichlorobenzene	95-50-1
19	1,1,1-Trichloroethane	71-55-6	39	1,3-Dichlorobenzene	541-73-1
20	1,2-Dibromo-3-chloropropane	96-12-8	40	1,4-Dichlorobenzene	106-46-7

Table 5-5. List of 22 chlorine-containing compounds in the CRM 81202 (Absolute Standards, Inc)

#	Name	CAS	#	Name	CAS
1	Benzyl dichloride	98-87-3	12	d-BHC	319-86-8
2	a,a,a-Trichlorotoluene	98-07-7	13	g-BHC (Lindane)	58-89-9
3	Benzyl chloride	100-44-7	14	Hexachlorocyclopentadiene	77-47-4
4	2-Chloronaphthalene	91-58-7	15	Hexachloroethane	67-72-1
5	1,2-Dichlorobenzene	95-50-1	16	Pentachlorobenzene	608-93-5
6	1,3-Dichlorobenzene	541-73-1	17	1,2,3,4-Tetrachlorobenzene	634-66-2
7	1,4-Dichlorobenzene	106-46-7	18	1,2,4,5-Tetrachlorobenzene	95-94-3
8	Hexachlorobenzene	118-74-1	19	1,2,3,5-Tetrachlorobenzene	634-90-2
9	Hexachloro-1,3-butadiene	87-68-3	20	1,2,3-Trichlorobenzene	87-61-6
10	a-BHC	319-84-6	21	1,2,4-Trichlorobenzene	120-82-1
11	b-BHC	319-85-7	22	1,3,5-Trichlorobenzene	108-70-3

5.2.3.2 GC×GC-TOFMS Conditions

Table 5-6 describes the various conditions employed for the acquisition of GC×GC chromatograms.

Table 5-6. Chromatographic conditions employed for the acquisition of GC×GC chromatograms

Description	Condition
Injection	Temperature: 325 °C Mode: Splitless
Column	Primary: 20 m × 0.18 mm; 0.18 µm df (Rxi-5MS) Secondary: 5 m × 0.18 mm; 0.18 µm df (Rxi-17MS) Carrier gas: He Flow rate: 20 mL/min constant flow
Temperature Ramp	Initial temp: 40 °C Duration: 10 min Ramp: 3 °C/min Final temp: 335 °C Duration: 10 min
Modulator	INSIGHT™ flow modulator (SepSolve Analytical, Peterborough, UK) P_M: 2.5 s
Mass Spectrometer	Instrument: BenchTOF-Select™ Ionization® mode at 70 eV and 14 eV Acquisition mass range: 30–594 u Acquisition rate: 200 Hz Source temperature: 250 °C Transfer line temperature: 315 °C
Software	ChromSpace® GC×GC software (Markes International, Llantrisant, UK) for full instrument control and data processing

5.2.3.3 Building Scripts

ChromaTOF® uses the Microsoft VBScript language to write functions, which adds flexibility in developing customized programming tasks. Existing useful syntaxes in VBScript such as "for loop" or "select case" statements are possible with ChromaTOF®. On contrary, ChromSpace® software has a feature called parametric filtering, where the functions can be developed and applied to generate a new file, which only includes data points that meet the specific conditions. The scripts can be constructed using the built in functions in the GUI, which is called "the expression builder" (Figure 5-9). This expression builder makes writing scripts easier and more convenient for the beginners and less skilled users, however, tasks are restricted by the functions present in the GUI and more advanced tasks such as "for loop" are not possible with the expression builder. In this work, scripts were written using the ChromSpace® expression builder to classify compounds containing specific numbers of chlorine atoms, referring to the published work by Hilton et al. and the fundamental chlorine rule for isotopic distribution patterns.[221]

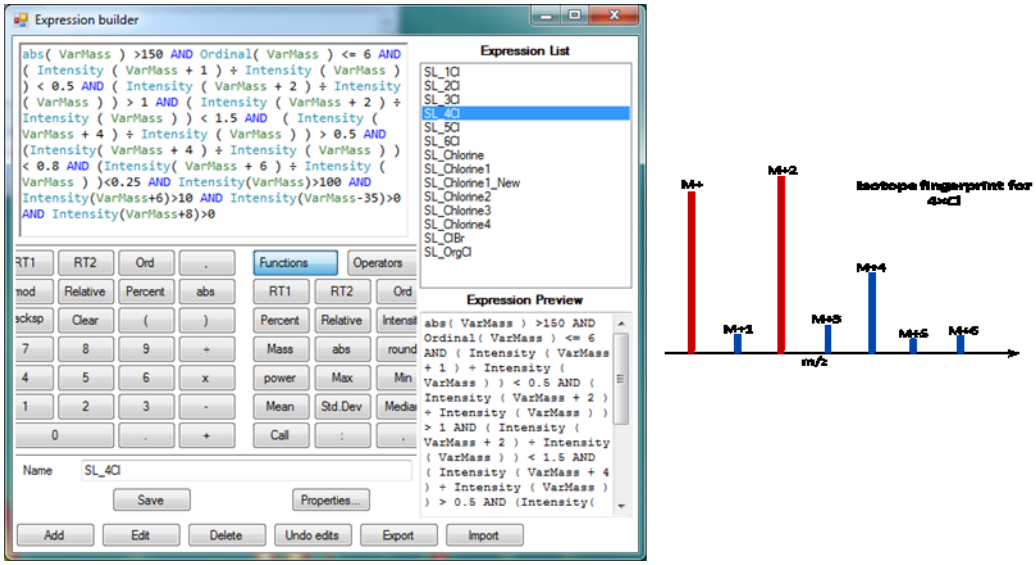

Figure 5-9. Scripting expressions used to classify compounds containing four chlorine atoms are shown as an example. The developed scripts are based on the distinct isotopic distribution of the molecular ion cluster.

5.2.4 Results and Discussion

Chlorinated species were selected as target compounds for the purpose of testing the scripts due to the relatively simple mass spectral features with distinct isotopic patterns, which makes the writing of scripts for them less complicated. The developed chlorine filtering scripts were applied to the entire chromatographic space of the chlorine standard mixture that was generated using SPME-GC×GC-TOFMS. From the analysis of the 78 component standard mix, a total of 166 total peaks were detected. All 40 chlorine-containing compounds (mono- through hexachloro) have been filtered successfully in just a few minutes of processing time using the developed chlorine scripts. The filtered chromatogram with scripts only shows peaks that pass the qualifier expression (Figure 5-10. Top).

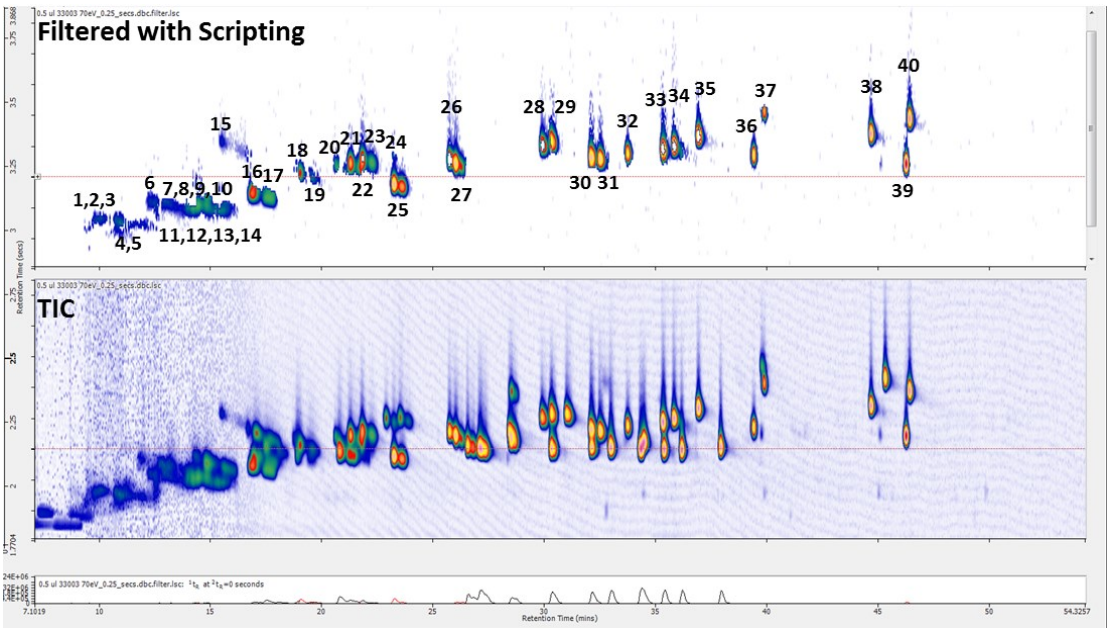

Figure 5-10. A chromatogram of a 33003 standard mix. Top: filtered image using scripts, Bottom: TIC.

The analytes that are filtered through the scripts are listed in ascending eluting order in Table 5-7. The 14 early eluting compounds (peak # 1-14 in Table 5-7) struggled to achieve a sufficient resolution at the beginning of the chromatogram. However, the deconvolution algorithm embedded in the ChromSpace® software package was able to provide library searchable mass spectra even with closely eluting compounds, and the chlorine scripts were able to filter the region of chlorine-containing atoms. Through the manual inspection of this region, 14 filtered compounds were confirmed as chlorinated species, despite the low mass spectral quality due to the coelution of peaks.

Table 5-7. List of 40 Chloride Compounds that were Filtered with Scripts

Peak #	Name	CAS
1	3-Chloropropene (allyl chloride)	107-05-1
2	Methylene chloride	75-09-2
3	1,1-Dichloroethene	75-35-4
4	trans-1,2-Dichloroethene	156-60-5
5	1,1-Dichloroethane	75-34-3
6	cis-1,2-Dichloroethene	156-59-2
7	2,2-Dichloropropane	594-20-7
8	Bromochloromethane	74-97-5
9	Chloroform	67-66-3
10	1-Chlorobutane	109-69-3
11	1,1,1-Trichloroethane	71-55-6
12	1,1-Dichloropropene	563-58-6
13	Carbon tetrachloride	56-23-5
14	1,2-Dichloroethane	107-06-2
15	Chloroacetonitrile	107-14-2
16	Trichloroethene	79-01-6
17	Bromodichloromethane	75-27-4
18	1,1-Dichloropropanone-2	513-88-2
19	cis-1,3-Dichloropropene	10061-01-5
20	trans-1,3-Dichloropropene	10061-02-6
21	1,1,2-Trichloroethane	79-00-5
22	1,2-Dichloropropane	78-87-5
23	1,3-Dichloropropane	142-28-9
24	Dibromochloromethane	124-48-1
25	Tetrachloroethene	127-18-4
26	Chlorobenzene	108-90-7
27	1,1,1,2-Tetrachloroethane	630-20-6
28	1,1,2,2-Tetrachloroethane	79-34-5
29	1,2,3-Trichloropropane	96-18-4
30	2-Chlorotoluene	95-49-8
31	4-Chlorotoluene	106-43-4
32	Pentachloroethane	76-01-7
33	1,3-Dichlorobenzene	541-73-1
34	1,4-Dichlorobenzene	106-46-7
35	1,2-Dichlorobenzene	95-50-1
36	Hexachloroethane	67-72-1
37	1,2-Dibromo-3-chloropropane	96-12-8
38	1,2,4-Trichlorobenzene	120-82-1
39	Hexachloro-1,3-butadiene	87-68-3
40	1,2,3-Trichlorobenzene	87-61-6

The same scripts were applied to a diesel sample spiked with a 22-component Cl standards mixture. Diesel is a complex mixture, and it was observed that some compounds still can coelute, even with the enhanced separation power of GC×GC-TOFMS. A total of 911 peaks were detected. The TIC (Total Ion Chromatogram), shown in Figure 5-11, is of little use for identifying chlorine-containing compounds in the mixture quickly and conveniently.

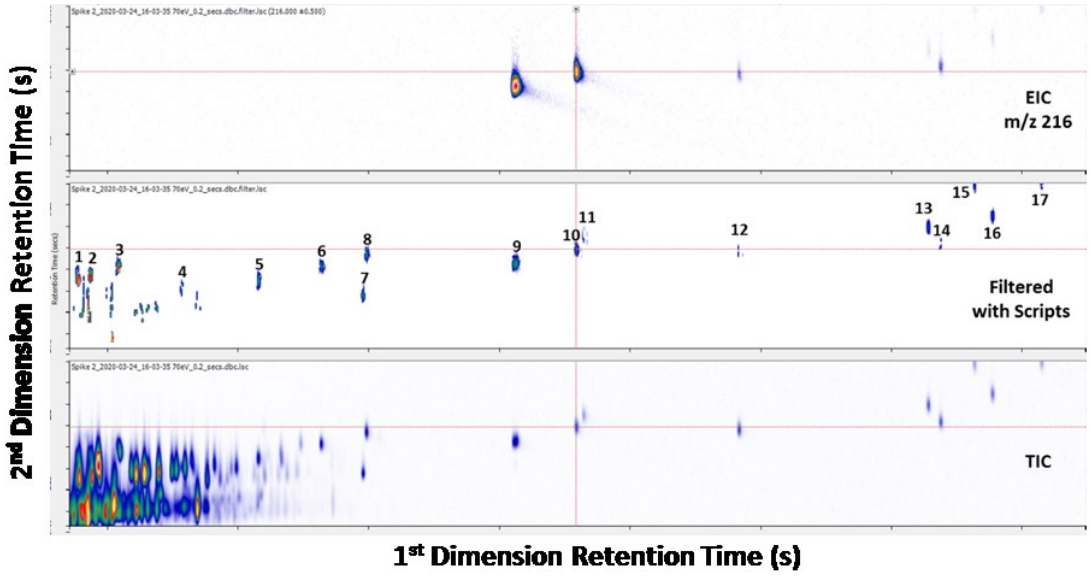

Figure 5-11. A diesel sample spiked with a 22-components Cl standard. Top: EIC *m/z* 216, the base peak for tetrachlorobenzene, Middle: Filtered with scripts, Bottom: TIC.

From the 22 chloride compounds spiked in, 17 compounds were filtered successfully using the scripts (Figure 5-11). The filtering scripts demonstrated increased selectivity compared to the Extracted Ion Chromatogram (EIC) because they only show peaks that pass the search criteria, whereas EIC shows all the peaks with that ion in common. The *m/z* 216 is a base peak for tetrachlorobenzene, and the EIC at *m/z* 216 showed only four compounds that contain significantly abundant *m/z* 216 ions (M+2 ion of tetrachlorobenzene). On the other hand, the scripts filtered 17 compounds (of 911 peaks in the chromatogram), which corresponds to 77% of the target analyte list. The five compounds that were missed in classification by the scripting filters were benzyl dichloride, a,a,a-trichlorotoluene, benzyl chloride, hexachlorocyclopentadiene, and 1,2,3,4-tetrachlorobenzene. Out of these five compounds, three compounds

(a,a,a-trichlorotoluene, benzyl chloride, 1,2,3,4-tetrachlorobenzene) were not detected at all due to unknown reasons.

Table 5-8. A List of 17 Chloride Compounds that Were Filtered with Scripts

Peak #	Analyte	CAS
1	1,3-Dichlorobenzene	541-73-1
2	1,4-Dichlorobenzene	106-46-7
3	1,2-Dichlorobenzene	95-50-1
4	Hexachloroethane	67-72-1
5	1,3,5-Trichlorobenzene	108-70-3
6	1,2,4-Trichlorobenzene	120-82-1
7	Hexachloro-1,3-butadiene	87-68-3
8	1,2,3-Trichlorobenzene	87-61-6
9	1,2,4,5-Tetrachlorobenzene	95-94-3
10	1,2,3,5-Tetrachlorobenzene	634-90-2
11	2-Chloronaphthalene	91-58-7
12	Pentachlorobenzene	608-93-5
13	a-BHC	319-84-6
14	Hexachlorobenzene	118-74-1
15	g-BHC (Lindane)	58-89-9
16	b-BHC	319-85-7
17	d-BHC	319-86-8

The two compounds (benzyl dichloride and hexachlorocyclopentadiene) evidently were detected by the TOFMS but missed in the classification via scripts (Figure 5-12, 5-14). The false-negatives in classification were likely due to the poor spectral quality, which was caused by the low concentration level of the compounds and coelution with other compounds. The mass spectral quality is a critical factor that affects the performance of the scripts directly. For the case of benzyl dichloride, the molecular ion is at *m/z* 160 with a base peak of *m/z* 125. Compared to other compounds in the standard mix (peak 5-8), the benzyl dichloride is present at a lower intensity (Figure 5-12). In the EIC of *m/z* 125 and *m/z* 160, the benzyl dichloride peak was evident with the selective mass channels, but in the TIC, the intensity of the peak was

significantly lower than other peaks. The low concentration level of benzyl dichloride resulted in a poor quality of spectra, with increased noise measurements in the high *m/z* region (beyond the molecular ion), indicated with a red box in Figure 5-13. The signature isotope cluster pattern also was demolished by contributions from other compounds, which resulted in the overall library match of only 15.58% (Figure 5-13). The false positives at the beginning of the chromatogram in Figure 5-11 are due to the poor resolution in the region and coelution with other analytes in diesel, causing spectral overlapping, which ruins the mass spectral quality.

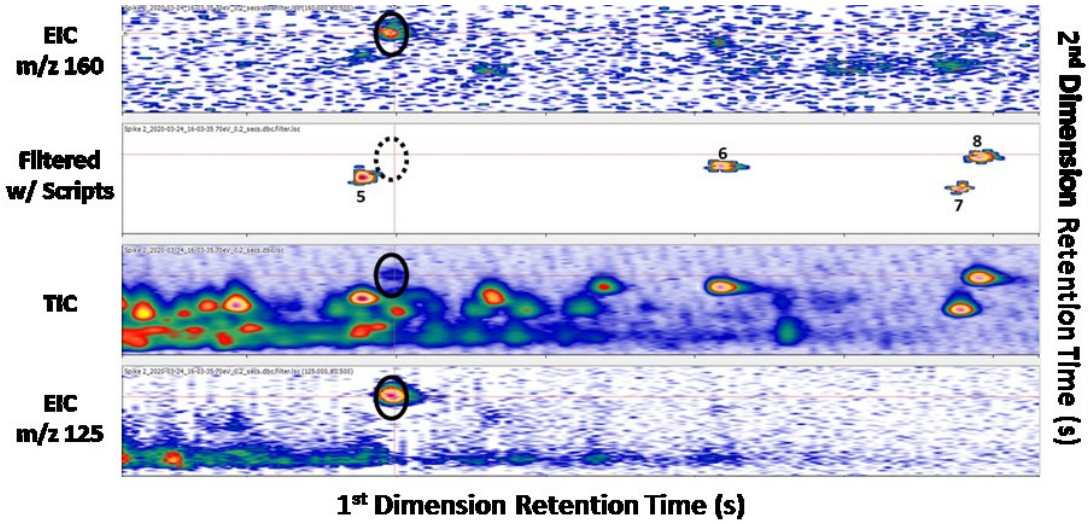

Figure 5-12. A zoomed-in region of benzyl dichloride. Benzyl dichloride was missed in classification by the scripts, although the TOFMS detected it.

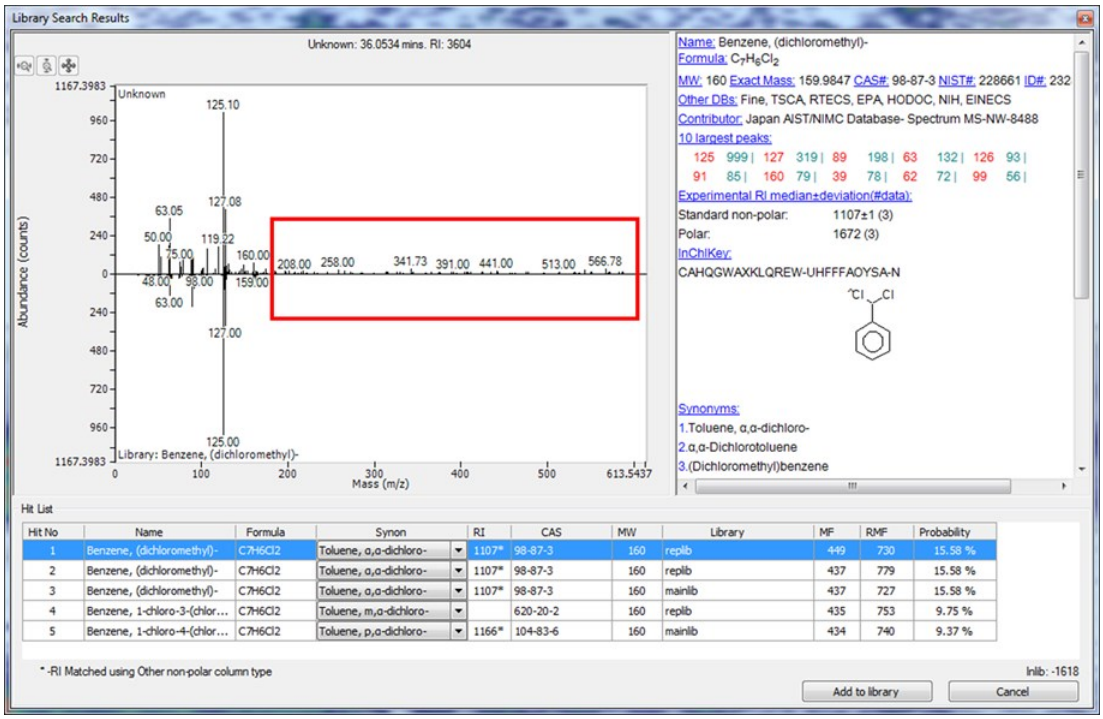

Figure 5-13. The peak true mass spectrum of benzyl dichloride (Top) with the library spectrum (Bottom) for comparison. The mass spectrum of the detected benzyl dichloride appeared to be noisy, resulting in poor spectral quality.

Figure 5-14, the zoomed-in chromatogram that focuses on the region of tetrachlorobenzene and hexachlorocyclopentadiene, illustrates well the separation power of 2D GC. With just a conventional one-dimensional GC, the two compounds (1,2,4,5-tetrachlorobenzene and hexachlorocyclopentadiene) would coelute completely and would not be resolved (Figure 5-14). The *m/z* 216 is a base peak for 1,2,4,5-tetrachlorobenzene, whereas the *m/z* 237 is a base peak for hexachlorocyclopentadiene. The red cross on the stacked chromatograms in Figure 5-14 indicates where the hexachlorocyclopentadiene peak is present in each chromatogram. The compound tetrachlorobenzene elutes slightly later than hexachlorocyclopentadiene in the second dimension (Figure 5-14). With the EIC of each base peak, it is evident that two distinct compounds exist with slightly different retention times in the second-dimension. However, tetrachlorobenzene is present at a more dominant concentration level than hexachlorocyclopentadiene, and the Cl scripts were able to filter tetrachlorobenzene but failed to filter out hexachlorocyclopentadiene.

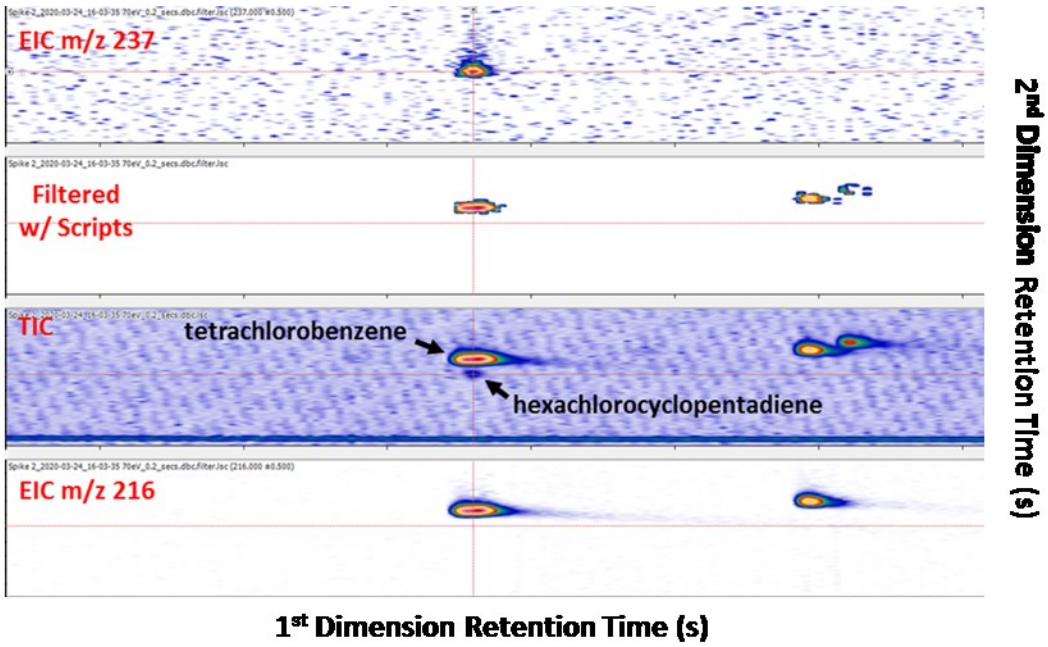

Figure 5-14. A zoomed-in region of tetrachlorobenzene and hexachlorocyclopentadiene**.**

5.2.5 Conclusions

Scripting tools that apply filters to GC×GC-TOFMS data based on logical operations applied to spectral and/or retention data have been reported in the literature for environmental and petroleum studies using the dominating software, ChromaTOF®. In this work, scripts were developed for chlorine-containing compounds in a relatively new GC×GC software called Chromspace®. The developed scripts were applied and filtered chlorine-containing compounds successfully within a few minutes of processing time. The scripting approach allowed automated screening of the presence of mono- through hexachloro compounds in a diesel sample that was spiked with a Cl standard mixture. It was confirmed that spectral quality plays a significant role in the performance of the scripts. The scripting approach proved to be a powerful tool for rapid and convenient screening of a sample.

CHAPTER 6

Conclusions and Future Works

6.1 Conclusions

Metabolomics is one of the latest omics technologies in the "omics" family, yet the field has seen rapid growth in a wide range of applications, spanning biological sciences, pharmacology, agriculture, food safety, etc. The growing number of data and publications generated using metabolomics technologies over time shows the great importance of achieving a better understanding of metabolism and related mechanisms of biological systems, given the highly increased incidence of human diseases. From an analytical chemistry perspective, GC×GC-TOFMS is a nearly ideal analytical platform for global metabolomics profiling due to its superior separation capacity for complex samples.

Over the years, our research group has focused on studies involving GC×GC from theoretical research, such as creating a thermodynamic-based predictive retention model to numerous applications-based studies in the petrochemical fields and in the environmental and forensic sciences. When I joined the research group, metabolomics was a relatively new area for our group. I soon realized that only a handful of groups in the metabolomics community utilize GC×GC-TOFMS. The major obstacles preventing the widespread usage of GC×GC-TOFMS in metabolomics include tedious sample preparation steps required to convert the metabolites as GC-analyzable, challenges in reliable quantitation due to many steps involved in sample preparation, and the difficulty in data analysis due to lack of advanced data handling tools for a large amount of GC×GC-TOFMS data. It is found that sample preparation and analysis methods vary significantly within the literature, and there is little standardization. This thesis explored the challenges and opportunities for improvement towards the standardization of the GC×GC-TOFMS metabolomics workflow. Although the general analytical workflow for GC-based metabolomics has been established, many limitations and challenges still need to be addressed and improved for more widespread

use of GC×GC-TOFMS. The work described in this thesis contains several significant contributions that improve the metabolomics pipeline using GC×GC-TOFMS.

The pre-analytical part of the metabolomics pipeline is pivotal in metabolomics studies because the observed metabolite content can vary depending on how the sample is prepared. Consequently, the choice of sample preparation method can alter the resulting interpretation of the data. The biggest challenge in GC-based metabolomics studies is the large chemical diversity in a biological sample, which is complicated further by oximation/silylation to obtain volatile derivatives. In Chapter 2, challenges related to sample preparation and analysis for GC×GC-TOFMS metabolomics studies have been discussed, with a particular focus on extraction, derivatization, and GC×GC column combination selection. The work demonstrated in Chapter 2 was completed in three different parts: the first focused on a comparison of different extraction solvent systems, the second focused on the optimization of derivatization parameters, and the third focused on exploring different chemistries of the GC×GC column combinations. Unfortunately, despite the efforts made to find the best sample preparation method, even with aid from the advanced chemometrics optimization method, finding a universal method that works for "all" classes of metabolites was not possible due to the highly contrasting physicochemical properties of the metabolome. Nonetheless, Chapter 2 provides an in-depth discussion for carrying out sample preparation, providing an understanding of how the choice of the method affects the resulting metabolic profile of a sample and the analytical performance.

Metabolomics studies have shown great potential in biomedical research, with an ability to find metabolites that can be used as an indicator that separates comparative groups. In a clinical setting, comparing the levels of metabolites in a patient's sample against a normal range is performed routinely. For an accurate and reliable interpretation of the result, the normalization of sample concentration variability is critical before applying any statistical modelling. To date, there has been no consensus on the normalization method to address the biological sample variability, and the lack of standardization of normalization techniques perplexes researchers. In addition, it often is observed that normalization is neglected due to the lack of a convenient means

of performing sample normalization. In Chapter 3, a new normalization approach using TUPA, which involves the summation of the peaks that are common to all samples as a normalization factor, was proposed. Using urine as a sample matrix, the performance of the new strategy was evaluated against the two most common normalization techniques for urine, creatinine adjustment and constant sum (TPA). Chapter 3 demonstrated an improvement in the data normalization using the proposed post-acquisition normalization method. The new method is easy to apply, and it outperformed alternative normalization methods. The convenience in applying the method and the excellent performance results suggest that TUPA may be an effective and feasible alternative for standardization of data normalization in untargeted metabolomics studies.

The development and improvement made in the chemometrics tools and software allows users to extract useful information conveniently from large data and has brought untargeted metabolomics into a new era. On the other hand, many researchers unnecessarily turn to chemometrics to simplify the data analysis, without sufficient information about the samples or skills to handle the statistical tools adequately. The reliance on chemometrics tools is mainly due to the lack of available tools to aid with reviewing a large amount of data. In Chapter 4, the scripts for various classes of metabolites were developed for the rapid and automated classification of compounds in biosamples. The scripting algorithms were created using the characteristic mass spectral features of major metabolite classes and were applied to various metabolomics sample types. With the use of a scripting tool, a peak table generated from GC×GC-TOFMS, which contains thousands of analytes, was reduced to tens or hundreds of compounds, making the manual review of a filtered peak table much more plausible. In Chapter 5, other applications using scripting were discussed. Scripts were developed for the subgroups of the terpene family. Automated group-type analysis of terpene was successfully performed on different strains of cannabis samples using scripting, which enabled rapid screening and quantitation of sample composition. The scripts for chlorinated species were also developed using another software platform for GC×GC data analysis, which demonstrated that scripting is not bound to specific

software, but applicable independent of software as long as deconvoluted peaks with quality spectral data are provided. The presented scripts in Chapters 4 and 5 proved to be a useful tool for the rapid classification of different classes of compounds and the visualization of the chemical composition of the sample.

6.2 Perspectives for Future Work

The field of metabolomics, in conjunction with a high separation power platform, GC×GC-TOFMS, holds great promise as the metabolome contains a wealth of information. The potential of GC×GC-TOFMS in metabolomics can be increased even more with the recent development of advanced techniques that could be coupled to GC×GC. The advancement of technology platforms for sample introduction and tandem ionization, which has recently become available in our lab, has sprung up some research ideas to be explored.

6.2.1 Expand Coverage Using Headspace Techniques

A significant number of metabolites from biological samples are organic volatiles, which may contain clinically useful diagnostic information. The potential of volatile organic compounds (VOCs) as biomarkers has received increasing attention in the metabolomics community.[223-231] Several ketones, including 4-heptanone, were found to be linked to diabetes mellitus and tentatively related to more specific stages of the disease.[232]

As discussed in Chapter 2 of this thesis, the two-step methoximation, followed by silylation derivatization, has been the gold standard for GC-based metabolomics studies. This approach is incredibly useful for molecules such as carbohydrates, amino acids, and fatty acids. However, not only is this method laborious and time-consuming, it also carries a critical drawback for the analysis of VOCs. The derivatization procedure involves evaporation/drying steps to remove the water because the presence of moisture quenches the derivatization chemistry due to the hygroscopic nature of a TMS derivatizing reagent. The drying steps result in the inevitable loss of volatile metabolites.

A headspace sampling technique would be a brilliant choice for the determination of volatiles and obtaining valuable information that VOCs provide.[233] In headspace analysis, volatile components are isolated from non-volatile sample components in the headspace of a vial. Solid-phase microextraction (SPME) is an example of a headspace technique, and it has seen an increase in use for the determination of volatile metabolites from biosamples.[21,234,235] SPME involves the use of a fibre coated with an extraction phase. In the headspace SPME (HS-SPME), the fibre is exposed to the headspace for a fixed duration, and the molecules from the sample are sorbed onto the fibre coating. Subsequently, the fibre with concentrated extracts is desorbed directly into an analytical instrument. The HS-SPME technique has substantial advantages, such as avoiding solvent consumption, generating a cleaner background, reducing the sample preparation to a minimum, and lowering the analysis cost.[235] The goal of the SPME technique is not the exhaustive extraction, but instead, it has strength in convenience and speed. Some disadvantages of the technique include the restricted sample loading due to the limited capacity of the fibre and limitations in the selections of commercially available extraction phases. Moreover, since the technique is an equilibrium-based, non-exhaustive approach, it poses significant challenges for quantitative analysis.

Dynamic headspace sampling (DHS) is an alternative headspace technique, which involves the passing of carrier gas through the headspace of a sample, followed by trapping of purged volatiles on an adsorbent-filled trap. In contrast to non-exhaustive sample extraction with SPME, DHS can be tuned to achieve nearly exhaustive extraction of the volatiles and extract the less volatile components more efficiently with dynamic purging.[236,237] This may lead to the generation of different VOC profiles, possibly with more comprehensive sample information. Until now, the studies about VOCs in metabolomics samples have been undertaken mostly with the SPME technique.

Recently, an automated sample introduction system for both SPME and DHS has become available in our lab. One of the core issues with the derivatization sample preparation method is the difficulty of automating the system. The automation of

sample introduction for SPME and DHS is very appealing, especially when the sample preparation for the headspace technique is typically quick and easy, with the use of minimal to no solvent. With the automated system, once the sample is prepared, the extraction and injection are automated. With the additionally available sample introduction techniques, the metabolic profiles of biological samples can be presented in three ways: conventional derivatization and two approaches of volatiles analysis, SPME and DHS. The headspace techniques can complement the data obtainable from derivatization methods to provide a more holistic picture of metabolic profiles and expand the metabolome coverage.

6.2.2 Tandem Ionization

Despite the superior separation power afforded by GC×GC, one of the major drawbacks of traditional GC×GC-MS systems is that the conventional hard electron ionization (EI; 70 eV) results in extensive mass spectral fragmentation. This is a particular challenge for metabolomics, especially studies relying on derivatization with TMS. In these studies, 70 eV ionization can make compound identification challenging as molecular ions are often missing, fragmentation patterns for similar compounds (isomers, homologues, etc.) can be difficult to distinguish, and the dominant ion is often the TMS ion. Consequently, it can be challenging to develop effective scripting tools to aid in data processing.

To overcome this limitation, lower energy EI has been recently introduced, where EI is conducted at energies in the range of 12-15 eV. These low-energy (soft) ionization energies result in mass spectra that are biased towards larger fragments (compared with 70 eV spectra) and increased intensity for the molecular ions. This provides new opportunities for compound identification. In tandem ionization mode, spectra are alternately collected at 70 eV and at a lower voltage throughout the entire analysis, generating complementary spectra.

The conventional approach of reliance on retention indices and 70 eV library mass spectra for compound identification can often be insufficient for challenging analytes such as terpenes with a high degree of isomerism. The hope is that with the

increased dimensionality of the analysis from tandem ionization and the ability to provide an enhanced characterization, analyte speciation that must have been difficult or impossible to identify by 70 eV alone can now be possible.

Our lab recently purchased a BenchTOF-Select (Markes International). This new instrument is equipped with a time-of-flight mass spectrometer for GC(×GC) with variable-energy electron ionization. The use of this instrument for TMS derivatized metabolomics samples would be beneficial for obtaining mass spectra with enhanced molecular ion signals from soft ionization. In Chapter 4, it was discussed how finding the molecular ion is extremely helpful in writing scripts because subsequent characteristic fragment losses can be deduced once the molecular ion peak is found correctly. The increased molecular ion peak for TMS derivatives and additional mass spectral features delivered by soft ionization would provide more valuable ingredients that can be used in scripts. This, in turn, would enhance the accuracy of scripting, and the reliable performance of scripts would speed up data analysis significantly and amplify the throughput of analytical approaches.

References

1. Kanani, H.; Chrysanthopoulos, P. K.; Klapa, M. I. Standardizing GC–MS metabolomics. *Journal of Chromatography B* **2008***, 871*, 191-201.

2. Putri, S. P.; Nakayama, Y.; Matsuda, F.; Uchikata, T.; Kobayashi, S.; Matsubara, A.; Fukusaki, E. Current metabolomics: Practical applications. *Journal of Bioscience and Bioengineering* **2013***, 115*, 579-589.

3. Ryan, D.; Robards, K. Metabolomics: The Greatest Omics of Them All? *Anal. Chem.* **2006***, 78*, 7954-7958.

4. Almstetter, M.; Oefner, P.; Dettmer, K. Comprehensive two-dimensional gas chromatography in metabolomics. *Anal Bioanal Chem* **2012***, 402*, 1993-2013.

5. Becker, S.; Kortz, L.; Helmschrodt, C.; Thiery, J.; Ceglarek, U. LC–MS-based metabolomics in the clinical laboratory. *Journal of Chromatography B* **2012***, 883-884*, 68-75.

6. Khodadadi, M.; Pourfarzam, M. A review of strategies for untargeted urinary metabolomic analysis using gas chromatography–mass spectrometry. *Metabolomics* **2020***, 16*.

7. Dunn, W. B.; Ellis, D. I. Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry* **2005***, 24*, 285-294.

8. Wang, H.; Muehlbauer, J. M.; O'Neal, K. S.; Newgard, B. C.; Hauser, R. E.; Bain, R. J.; Shah, H. S. Recommendations for Improving Identification and Quantification in Non-Targeted, GC-MS-Based Metabolomic Profiling of Human Plasma. *Metabolites* **2017***, 7*.

9. Nishiumi, S.; Suzuki, M.; Kobayashi, T.; Matsubara, A.; Azuma, T.; Yoshida, M. Metabolomics for biomarker discovery in gastroenterological cancer. *Metabolites* **2014***, 4*, 547-571.

10. Mastrangelo, A.; Ferrarini, A.; Rey-Stolle, F.; García, A.; Barbas, C. From sample treatment to biomarker discovery: A tutorial for untargeted metabolomics based on GC-(EI)-Q-MS. *Analytica Chimica Acta* **2015***, 900*, 21-35.

11. Dudzik, D.; Barbas-Bernardos, C.; García, A.; Barbas, C. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. *J. Pharm. Biomed. Anal.* **2018***, 147*, 149-173.

12. Manzi, M.; Riquelme, G.; Zabalegui, N.; Monge, M. E. Improving diagnosis of genitourinary cancers: Biomarker discovery strategies through mass

spectrometry-based metabolomics. *Journal of Pharmaceutical and Biomedical Analysis* **2020***, 178*, 112905.

13. Gowda, G. A. N.; Zhang, S.; Gu, H.; Asiago, V.; Shanaiah, N.; Raftery, D. Metabolomics-based methods for early disease diagnostics. *Expert review of molecular diagnostics* **2008***, 8*, 617-633.

14. Weir, T. L.; Manter, D. K.; Sheflin, A. M.; Barnett, B. A.; Heuberger, A. L.; Ryan, E. P. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One* **2013***, 8*, e70803.

15. Seeley, J. V.; Seeley, S. K. Multidimensional Gas Chromatography: Fundamental Advances and New Applications. *Analytical chemistry* **2013***, 85*, 557-578.

16. Olivier, K. J.; Krejsa, C. M.; Freeman, J. A. Chapter 9 - Biomarkers. *Information Resources in Toxicology (Fifth Edition)* **2020**, 97-107.

17. Karu, N.; Deng, L.; Slae, M.; Guo, A. C.; Sajed, T.; Huynh, H.; Wine, E.; Wishart, D. S. A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. *Analytica Chimica Acta* **2018***, 1030*, 1-24.

18. Wishart, D. S.; Lewis, M. J.; Morrissey, J. A.; Flegel, M. D.; Jeroncic, K.; Xiong, Y.; Cheng, D.; Eisner, R.; Gautam, B.; Tzur, D.; Sawhney, S.; Bamforth, F.; Greiner, R.; Li, L. The human cerebrospinal fluid metabolome. *Journal of Chromatography B* **2008***, 871*, 164-173.

19. Murray, J. A. Qualitative and quantitative approaches in comprehensive two-dimensional gas chromatography. *Journal of Chromatography A* **2012***, 1261*, 58-68.

20. Vasquez, N. P.; Crosnier de bellaistre-Bonose, M.; Lévêque, N.; Thioulouse, E.; Doummar, D.; Billette de Villemeur, T.; Rodriguez, D.; Couderc, R.; Robin, S.; Courderot-Masuyer, C.; Moussa, F. Advances in the metabolic profiling of acidic compounds in children's urines achieved by comprehensive two-dimensional gas chromatography. *Journal of Chromatography B* **2015***, 1002*, 130-138.

21. de la Mata, A. Paulina; McQueen, R. H.; Nam, S. L.; Harynuk, J. J. Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics. *Analytical and Bioanalytical Chemistry* **2017***, 409*, 1905-1913.

22. Higgins Keppler, E. A.; Jenkins, C. L.; Davis, T. J.; Bean, H. D. Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics. *Trends in Analytical Chemistry* **2018***, 109*, 275-286.

23. Dettmer, K.; Almstetter, M. F.; Appel, I. J.; Nürnberger, N.; Schlamberger, G.; Gronwald, W.; Meyer, H. H. D.; Oefner, P. J. Comparison of serum versus plasma collection in gas chromatography--mass spectrometry-based metabolomics. *Electrophoresis* **2010***, 31*, 2365-2373.

24. Das, M. K.; Bishwal, S. C.; Das, A.; Dabral, D.; Varshney, A.; Badireddy, V. K.; Nanda, R. Investigation of Gender-Specific Exhaled Breath Volatome in Humans by GCxGC-TOF-MS. *Analytical Chemistry* **2014***, 86*, 1229-1237.

25. Peralbo-Molina, A.; Calderón-Santiago, M.; Priego-Capote, F.; Jurado-Gámez, B.; Luque de Castro, M. D. Development of a method for metabolomic analysis of human exhaled breath condensate by gas chromatography–mass spectrometry in high resolution mode. *Anal. Chim. Acta* **2015***, 887*, 118-126.

26. Hall, R.; Beale, M.; Fiehn, O.; Hardy, N.; Sumner, L.; Bino, R. Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell* **2002***, 14*, 1437-1440.

27. Hong, J.; Yang, L.; Zhang, D.; Shi, J. Plant Metabolomics: An Indispensable System Biology Tool for Plant Science. *International journal of molecular sciences* **2016***, 17*, 767.

28. Almstetter, M.; Oefner, P.; Dettmer, K. Comprehensive two-dimensional gas chromatography in metabolomics. *Anal Bioanal Chem* **2012***, 402*, 1993-2013.

29. Beckles, D. M.; Roessner, U. 5 - Plant metabolomics: Applications and opportunities for agricultural biotechnology. In *Plant Biotechnology and Agriculture*; Altman, A., Hasegawa, P. M., Eds.; Academic Press: San Diego, 2012; pp 67-81.

30. Kim, S.; Kim, J.; Yun, E. J.; Kim, K. H. Food metabolomics: from farm to human. *Curr. Opin. Biotechnol.* **2016***, 37*, 16-23.

31. Karu, N.; Deng, L.; Slae, M.; Guo, A. C.; Sajed, T.; Huynh, H.; Wine, E.; Wishart, D. S. A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. *Analytica Chimica Acta* **2018***, 1030*, 1-24.

32. Psychogios, N.; Hau, D. D.; Peng, J.; Guo, A. C.; Mandal, R.; Bouatra, S.; Sinelnikov, I.; Krishnamurthy, R.; Eisner, R.; Gautam, B.; Young, N.; Xia, J.; Knox, C.; Dong, E.; Huang, P.; Hollander, Z.; Pedersen, T. L.; Smith, S. R.; Bamforth, F.; Greiner, R.; McManus, B.; Newman, J. W.; Goodfriend, T.; Wishart, D. The human serum metabolome. *PLoS ONE* **2011***, 6*.

33. Danielsson, A. P. H.; Moritz, T.; Mulder, H.; Spégel, P. Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics* **2012***, 8*, 50-63.

34. Winnike, J. H.; Wei, X.; Knagge, K. J.; Colman, S. D.; Gregory, S. G.; Zhang, X. Comparison of GC-MS and GC×GC-MS in the analysis of human serum samples for biomarker discovery. *Journal of proteome research* **2015***, 14*, 1810.

35. Gratton, J.; Phetcharaburanin, J.; Mullish, B. H.; Williams, H. R. T.; Thursz, M.; Nicholson, J. K.; Holmes, E.; Marchesi, J. R.; Li, J. V. Optimized Sample Handling Strategy for Metabolic Profiling of Human Feces. *Anal. Chem.* **2016***, 88*, 4661-4668.

36. Emwas, A. M. The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research. In *Metabonomics: Methods and Protocols*; Bjerrum, J. T., Ed.; Springer New York: New York, NY, 2015; pp 161-193.

37. Koek, M. M.; Jellema, R. H.; van, d. G.; Tas, A. C.; Hankemeier, T. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* **2010***, 7*, 307-328.

38. Cajka, T.; Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal. Chem.* **2016***, 88*, 524-545.

39. Álvarez-Sánchez, B.; Priego-Capote, F.; Luque de Castro, M. D. Metabolomics analysis I. Selection of biological samples and practical aspects preceding sample preparation. *TrAC Trends in Analytical Chemistry* **2010***, 29*, 111-119.

40. Calvert, J. G. Glossary of atmospheric chemistry terms (Recommendations 1990). *Pure and Applied Chemistry* **1990***, 62*, 2167-2219.

41. Broekaert, J. A. C. Daniel C. Harris: Quantitative chemical analysis, 9th ed. *Analytical and Bioanalytical Chemistry* **2015***, 407*, 8943-8944.

42. Mostafa, A.; Edwards, M.; Górecki, T. Optimization aspects of comprehensive two-dimensional gas chromatography. *Journal of Chromatography A* **2012***, 1255*, 38-55.

43. Restek Guide to GC Column Selection and Optimizing Separations.

44. Supelco GC column selection guide. *LC GC North America* **2012***, 30*, 523.

45. Millipore Sigma GC Column Selection Guide. *https://www.sigmaaldrich.com/technical-documents/articles/analytical/gc-column-selection-guide.html* (accessed Nov 25, 2020).

46. Rahman, M. M.; Abd El-Aty, A. M.; Choi, J.; Shin, H.; Shin, S. C.; Shim, J. Basic Overview on Gas Chromatography Columns. *Analytical Separation Science* **2015**, 823-834.

47. Fiehn, O. Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Curr Protoc Mol Biol* **2016***, 114*, 30.4.1-30.4.32.

48. Koek, M.; Jellema, R.; van der Greef, J.; Tas, A.; Hankemeier, T. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* **2011***, 7*, 307-328.

49. Halket, J. M.; Waterman, D.; Przyborowska, A. M.; Patel, R. K. P.; Fraser, P. D.; Bramley, P. M. Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J. Exp. Bot.* **2004***, 56*, 219-243.

50. Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. *Anal. Chem.* **2009***, 81*, 10038-10048.

51. Bowden, J. A.; Colosi, D. M.; Mora-Montero, D.; Garrett, T. J.; Yost, R. A. Evaluation of Derivatization Strategies for the Comprehensive Analysis of Endocrine Disrupting Compounds using GC/MS. *J. Chromatogr. Sci.* **2009***, 47*, 44-51.

52. Bekele, E. A.; Annaratone, C. E. P.; Hertog, Maarten L. A. T. M.; Nicolai, B. M.; Geeraerd, A. H. Multi-response optimization of the extraction and derivatization protocol of selected polar metabolites from apple fruit tissue for GC–MS analysis. *Anal. Chim. Acta* **2014***, 824*, 42-56.

53. Agilent, G.; Fiehn, G.; Metabolomics, R.; Library Agilent G1676AA Fiehn GC/MS Metabolomics RTL Library.

54. Orata, F. Derivatization Reactions and Reagents for Gas Chromatography Analysis. *Advanced Gas Chromatography - Progress in Agricultural, Biomedical and Industrial Applications* **2012**.

55. Abbiss, H.; Rawlinson, C.; Maker, G. L.; Trengove, R. Assessment of automated trimethylsilyl derivatization protocols for GC–MS-based untargeted metabolomic analysis of urine. *Metabolomics* **2015***, 11*, 1908-1921.

56. Zarate, E.; Boyle, V.; Rupprecht, U.; Green, S.; Villas-Boas, G.; Baker, P.; Pinu, R. F. Fully Automated Trimethylsilyl (TMS) Derivatisation Protocol for Metabolite Profiling by GC-MS. *Metabolites* **2017***, 7*.

57. Halket, J. M.; Zaikin, V. G. Derivatization in Mass Spectrometry—1. Silylation. *Eur J Mass Spectrom (Chichester)* **2003***, 9*, 1-21.

58. Winnike, J. H.; Wei, X.; Knagge, K. J.; Colman, S. D.; Gregory, S. G.; Zhang, X. Comparison of GC-MS and GC×GC-MS in the Analysis of Human Serum Samples for Biomarker Discovery. *Journal of proteome research* **2015***, 14*, 1810-1817.

59. Berger, T. A. Separation of a gasoline on an open tubular column with 1.3 million effective plates. *Chromatographia* **1996***, 42*, 63-71.

60. Marriott, P.; Shellie, R. Principles and applications of comprehensive two-dimensional gas chromatography. *TrAC Trends in Analytical Chemistry* **2002***, 21*, 573-583.

61. Liu, Z.; Phillips, J. B. Comprehensive Two-Dimensional Gas Chromatography using an On-Column Thermal Modulator Interface. *J. Chromatogr. Sci.* **1991***, 29*, 227-231.

62. Giddings, J. C. Two-dimensional separations: concept and promise. *Anal. Chem.* **1984***, 56*, 1258A-1270A.

63. Giddings, J. C. Concepts and comparisons in multidimensional separation. *J. High Resol. Chromatogr.* **1987***, 10*, 319-323.

64. Bahaghighat, H. D.; Freye, C. E.; Synovec, R. E. Recent advances in modulator technology for comprehensive two dimensional gas chromatography. *TrAC Trends in Analytical Chemistry* **2019***, 113*, 379-391.

65. Koek, M. M.; van der Kloet, Frans M.; Kleemann, R.; Kooistra, T.; Verheij, E. R.; Hankemeier, T. Semi-automated non-target processing in GC × GC–MS metabolomics analysis: applicability for biomedical studies. *Metabolomics* **2011***, 7*, 1-14.

66. Higgins Keppler, E. A.; Jenkins, C. L.; Davis, T. J.; Bean, H. D. Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics. *Trends in Analytical Chemistry* **2018***, 109*, 275-286.

67. Adahchour, M.; Beens, J.; Vreuls, R. J. J.; Brinkman, U. A. T. Recent developments in comprehensive two-dimensional gas chromatography (GC ×

GC). III. Applications for petrochemicals and organohalogens. *TrAC Trends Anal. Chem.* **2006***, 25*, 726-741.

68. Welthagen, W.; Shellie, R. A.; Spranger, J.; Ristow, M.; Zimmermann, R.; Fiehn, O. Comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC-TOF) for high resolution metabolomics: biomarker discovery on spleen tissue extracts of obese NZO compared to lean C57BL/6 mice. *Metabolomics* **2005***, 1*, 65-73.

69. Li, X.; Xu, Z.; Lu, X.; Yang, X.; Yin, P.; Kong, H.; Yu, Y.; Xu, G. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: Biomarker discovery for diabetes mellitus. *Anal. Chim. Acta* **2009***, 633*, 257-262.

70. Dunn, W. B.; Bailey, N. J. C.; Johnson, H. E. Measuring the metabolome: current analytical technologies. *Analyst* **2005***, 130*, 606-625.

71. Deda, O.; Gika, H. G.; Wilson, I. D.; Theodoridis, G. A. An overview of fecal sample preparation for global metabolic profiling. *Journal of Pharmaceutical and Biomedical Analysis* **2015***, 113*, 137-150.

72. González-Domínguez, R.; González-Domínguez, Á; Sayago, A.; Fernández-Recamales, Á Recommendations and Best Practices for Standardizing the Pre-Analytical Processing of Blood and Urine Samples in Metabolomics. *Metabolites* **2020***, 10*, 229.

73. Lu, W.; Su, X.; Klein, M. S.; Lewis, I. A.; Fiehn, O.; Rabinowitz, J. D. Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *Annu. Rev. Biochem.* **2017***, 86*, 277-304.

74. Vuckovic, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. *Analytical and Bioanalytical Chemistry* **2012***, 403*, 1523-1548.

75. Johnson, C. H.; Gonzalez, F. J. Challenges and opportunities of metabolomics. *J. Cell. Physiol.* **2012***, 227*, 2975-2981.

76. Bilbao, A.; Gibbons, B. C.; Slysz, G. W.; Crowell, K. L.; Monroe, M. E.; Ibrahim, Y. M.; Smith, R. D.; Payne, S. H.; Baker, E. S. An algorithm to correct saturated mass spectrometry ion abundances for enhanced quantitation and mass accuracy in omic studies. *International journal of mass spectrometry* **2018***, 427*, 91-99.

77. Wu, Y.; Li, L. Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A* **2016***, 1430*, 80-95.

78. Warrack, B. M.; Hnatyshyn, S.; Ott, K.; Reily, M. D.; Sanders, M.; Zhang, H.; Drexler, D. M. Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B* **2009***, 877*, 547-552.

79. Smith, L.; Villaret-Cazadamont, J.; Claus, S. P.; Canlet, C.; Guillou, H.; Cabaton, N. J.; Ellero-Simatos, S. Important Considerations for Sample Collection in Metabolomics Studies with a Special Focus on Applications to Liver Functions. *Metabolites* **2020***, 10*, 104.

80. Fiehn, O. Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Current protocols in molecular biology* **2016***, 114*, 30.4.1-30.4.32.

81. Lu, W.; Su, X.; Klein, M. S.; Lewis, I. A.; Fiehn, O.; Rabinowitz, J. D. Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *Annu. Rev. Biochem.* **2017***, 86*, 277-304.

82. Flores, R.; Shi, J.; Yu, G.; Ma, B.; Ravel, J.; Goedert, J. J.; Sinha, R. Collection media and delayed freezing effects on microbial composition of human stool. *Microbiome* **2015***, 3*, 33.

83. Reis, A.; Rudnitskaya, A.; Blackburn, G. J.; Mohd Fauzi, N.; Pitt, A. R.; Spickett, C. M. A comparison of five lipid extraction solvent systems for lipidomic studies of human LDL. *Journal of lipid research* **2013***, 54*, 1812.

84. Gullberg, J.; Johansson, A.; Moritz, T.; Marklund, S. L.; Jiye, A.; Antti, H.; Trygg, J.; Jonsson, P. Extraction and GC/MS analysis of the human blood plasma metabolome. *Analytical chemistry* **2005***, 77*, 8086.

85. Little, J. L. Artifacts in trimethylsilyl derivatization reactions and ways to avoid them. *Journal of Chromatography A* **1999***, 844*, 1-22.

86. Higgins Keppler, E. A.; Jenkins, C. L.; Davis, T. J.; Bean, H. D. Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics. *TrAC Trends in Analytical Chemistry* **2018***, 109*, 275-286.

87. Hilton, D. C.; Jones, R. S.; Sjödin, A. A method for rapid, non-targeted screening for environmental contaminants in household dust. *Journal of Chromatography A* **2010***, 1217*, 6851-6856.

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

89. Titaley, I. A.; Ogba, O. M.; Chibwe, L.; Hoh, E.; Cheong, P. H. -.; Simonich, S. L. M. Automating data analysis for two-dimensional gas chromatography/time-of-flight mass spectrometry non-targeted analysis of comparative samples. *Journal of Chromatography A* **2018***, 1541*, 57-62.

90. Moayedpour, S.; Parastar, H. RMet: An automated R based software for analyzing GC-MS and GC×GC-MS untargeted metabolomic data. *Chemometrics and Intelligent Laboratory Systems* **2019***, 194*, 103866.

91. Sinha, A. E.; Hope, J. L.; Prazen, B. J.; Nilsson, E. J.; Jack, R. M.; Synovec, R. E. Algorithm for locating analytes of interest based on mass spectral similarity in GC × GC–TOF-MS data: analysis of metabolites in human infant urine. *Journal of Chromatography A* **2004***, 1058*, 209-215.

92. Billoir, E.; Navratil, V.; Blaise, B. J. Sample size calculation in metabolic phenotyping studies. *Brief Bioinform* **2015***, 16*, 813-819.

93. Sarker, S. D.; Nahar, L. Chapter 19 - Applications of High Performance Liquid Chromatography in the Analysis of Herbal Products. In *Evidence-Based Validation of Herbal Medicine*; Mukherjee, P. K., Ed.; Elsevier: Boston, 2015; pp 405-425.

94. Héberger, K. Chapter 7 - Chemoinformatics—multivariate mathematical–statistical methods for data evaluation. In *Medical Applications of Mass Spectrometry*; Vékey, K., Telekes, A. and Vertes, A., Eds.; Elsevier: Amsterdam, 2008; pp 141-169.

95. Hassan, H. A.; Ammar, N. M.; Serag, A.; Shaker, O. G.; El Gendy, A. N.; Abdel-Hamid, A. Z. Metabolomics driven analysis of obesity-linked colorectal cancer patients via GC-MS and chemometrics: A pilot study. *Microchemical Journal* **2020***, 155*, 104742.

96. Madsen, R.; Lundstedt, T.; Trygg, J. Chemometrics in metabolomics—A review in human disease diagnosis. *Anal. Chim. Acta* **2010***, 659*, 23-33.

97. Han, T.; Yang, Y.; Zhang, H.; Law, K. P. Analytical challenges of untargeted GC-MS-based metabolomics and the critical issues in selecting the data processing strategy. *F1000Research* **2017***, 6*, 967.

98. Rosen Vollmar, A.,K.; Rattray, N. J. W.; Cai, Y.; Santos-Neto, Á,J.; Deziel, N. C.; Jukic, A. M. Z.; Johnson, C. H. Normalizing Untargeted Periconceptional Urinary Metabolomics Data: A Comparison of Approaches. *Metabolites* **2019***, 9*, 198.

99. Yang, Y.; Yin, Y.; Chen, X.; Chen, C.; Xia, Y.; Qi, H.; Baker, P. N.; Zhang, H.; Han, T. Evaluating different extraction solvents for GC-MS based metabolomic analysis of the fecal metabolome of adult and baby giant pandas. *Scientific Reports* **2019***, 9*, 12017.

100. Gullberg, J.; Jonsson, P.; Nordström, A.; Sjöström, M.; Moritz, T. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal. Biochem.* **2004***, 331*, 283-295.

101. Gil, A.; Zhang, W.; Wolters, J. C.; Permentier, H.; Boer, T.; Horvatovich, P.; Heiner-Fokkema, M.; Reijngoud, D.; Bischoff, R. One- vs two-phase extraction: re-evaluation of sample preparation procedures for untargeted lipidomics in plasma samples. *Analytical and Bioanalytical Chemistry* **2018***, 410*, 5859-5870.

102. A, J.; Trygg, J.; Gullberg, J.; Johansson, A. I.; Jonsson, P. ä; Antti, H.; Marklund, S. L.; Moritz, T. Extraction and GC/MS Analysis of the Human Blood Plasma Metabolome. *Anal. Chem.* **2005***, 77*, 8086-8094.

103. Warwick B Dunn; David Broadhurst; Paul Begley; Eva Zelena; Sue Francis-mcintyre; Nadine Anderson; Marie Brown; Joshau D Knowles; Antony Halsall; John N Haselden; Andrew W Nicholls; Ian D Wilson; Douglas B Kell; Royston Goodacre Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols* **2011***, 6*, 1060.

104. Technology, Panel on Review of the Material Measurement Laboratory at the National Institute of Standards and; Board, L. A.; Sciences, Division on Engineering and Physical; Council, N. R. *An Assessment of the National Institute of Standards and Technology Material Measurement Laboratory;* National Academies Press: Washington, D.C, 2015; .

105. Fiehn, O. Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Curr. Protoc. Mol. Biol.* **2016***, 114*, 30.4.1-30.4.32.

106. Matyash, V.; Liebisch, G.; Kurzchalia, T. V.; Shevchenko, A.; Schwudke, D. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of Lipid Research* **2008***, 49*, 1137-1146.

107. Chan, E. C. Y.; Pasikanti, K. K.; Nicholson, J. K. Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. *Nature Protocols* **2011***, 6*, 1483-1499.

108. Bligh, E. G.; Dyer, W. J. A rapid method of total lipid extraction and purification. *Can J Biochem Physiol* **1959***, 37*, 911-917.

109. Sostare, J.; Di Guida, R.; Kirwan, J.; Chalal, K.; Palmer, E.; Dunn, W. B.; Viant, M. R. Comparison of modified Matyash method to conventional solvent systems for polar metabolite and lipid extractions. *Analytica Chimica Acta* **2018***, 1037*, 301-315.

110. Bro, R.; Smilde, A. K. Principal component analysis. *Anal. Methods* **2014***, 6*, 2812-2831.

111. Lien, S. K.; Kvitvang, H. F. N.; Bruheim, P. Utilization of a deuterated derivatization agent to synthesize internal standards for gas chromatography–tandem mass spectrometry quantification of silylated metabolites. *Journal of Chromatography A* **2012***, 1247*, 118-124.

112. Butterworth, P. J. Lehninger: principles of biochemistry (4th edn) D. L. Nelson and M. C. Cox, W. H. Freeman & Co., New York, 1119 pp (plus 17 pp glossary), ISBN 0-7167-4339-6 (2004). *Cell Biochem. Funct.* **2005***, 23*, 293-294.

113. Palmas, F.; Mussap, M.; Fattuoni, C. Urine metabolome analysis by gas chromatography–mass spectrometry (GC–MS): Standardization and optimization of protocols for urea removal and short-term sample storage. *Clinica Chimica Acta* **2018***, 485*, 236-242.

114. Yu, Z.; Huang, H.; Reim, A.; Charles, P. D.; Northage, A.; Jackson, D.; Parry, I.; Kessler, B. M. Optimizing 2D gas chromatography mass spectrometry for robust tissue, serum and urine metabolite profiling. *Talanta* **2017***, 165*, 685-691.

115. Liebeke, M.; Puskás, E. Drying Enhances Signal Intensities for Global GC–MS Metabolomics. *Metabolites* **2019***, 9*.

116. Weinert, C. H.; Egert, B.; Kulling, S. E. On the applicability of comprehensive two-dimensional gas chromatography combined with a fast-scanning quadrupole mass spectrometer for untargeted large-scale metabolomics. *Journal of Chromatography A* **2015***, 1405*, 156-167.

117. Ye, G.; Zhu, B.; Yao, Z.; Yin, P.; Lu, X.; Kong, H.; Fan, F.; Jiao, B.; Xu, G. Analysis of Urinary Metabolic Signatures of Early Hepatocellular Carcinoma Recurrence after Surgical Removal Using Gas Chromatography–Mass Spectrometry. *J. Proteome Res.* **2012***, 11*, 4361-4372.

118. Parkinson, D. R. Analytical Derivatization Techniques☆. In *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*Elsevier: 2014; .

119. Chan, E. C. Y.; Pasikanti, K. K.; Nicholson, J. K. Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. **2011**.

120. Koek, M. M.; Muilwijk, B.; van der Werf, Mariët J.; Hankemeier, T. Microbial Metabolomics with Gas Chromatography/Mass Spectrometry. *Anal. Chem.* **2006**, *78*, 1272-1281.

121. Fiehn, O.; Kopka, J.; Trethewey, R. N.; Willmitzer, L. Identification of Uncommon Plant Metabolites Based on Calculation of Elemental Compositions Using Gas Chromatography and Quadrupole Mass Spectrometry. *Anal. Chem.* **2000**, *72*, 3573-3580.

122. Geiger, E. O. Chapter 21 - Statistical Methods for Fermentation Optimization. *Fermentation and Biochemical Engineering Handbook (Third Edition)* **2014**, 415-422.

123. Jacyna, J.; Kordalewska, M.; Markuszewski, M. J. Design of Experiments in metabolomics-related studies: An overview. *J. Pharm. Biomed. Anal.* **2019**, *164*, 598-606.

124. Eren Şenaras, A. Chapter 8 - Parameter optimization using the surface response technique in automated guided vehicles. In *Sustainable Engineering Products and Manufacturing Technologies*; Kumar, K., Zindani, D. and Davim, P., Eds.; Academic Press: 2019; pp 187-197.

125. Mumtaz, M. W.; Adnan, A.; Mukhtar, H.; Rashid, U.; Danish, M. Chapter Fifteen - Biodiesel Production Through Chemical and Biochemical Transesterification: Trends, Technicalities, and Future Perspectives. In *Clean Energy for Sustainable Development*; Rasul, M. G., Azad, A. k. and Sharma, S. C., Eds.; Academic Press: 2017; pp 465-485.

126. Das, A. K.; Dewanjee, S. Chapter 3 - Optimization of Extraction Using Mathematical Models and Computation. *Computational Phytochemistry* **2018**, 75-106.

127. Yuangyai, C.; Nembhard, H. B. Chapter 8 - Design of Experiments: A Key to Innovation in Nanotechnology. *Emerging Nanotechnologies for Manufacturing* **2010**, 207-234.

128. Pocker, Y.; Green, E. Hydrolysis of D-glucono-.delta.-lactone. I. General acid-base catalysis, solvent deuterium isotope effects, and transition state characterization. *J. Am. Chem. Soc.* **1973**, *95*, 113-119.

129. Zapadlo, M.; Krupčík, J.; Májek, P.; Armstrong, D. W.; Sandra, P. Use of a polar ionic liquid as second column for the comprehensive two-dimensional GC separation of PCBs. *Journal of Chromatography A* **2010***, 1217*, 5859-5867.

130. Cagliero, C.; Bicchi, C.; Cordero, C.; Liberto, E.; Rubiolo, P.; Sgorbini, B. Analysis of essential oils and fragrances with a new generation of highly inert gas chromatographic columns coated with ionic liquids. *Journal of Chromatography A* **2017***, 1495*, 64-75.

131. Harynuk, J.; Marriott, P. Comprehensive multidimensional separations. *Analytical and Bioanalytical Chemistry* **2011***, 401*, 2333-2334.

132. Murray, J. A. Qualitative and quantitative approaches in comprehensive two-dimensional gas chromatography. *Journal of Chromatography A* **2012***, 1261*, 58-68.

133. Shi, X.; Wei, X.; Yin, X.; Wang, Y.; Zhang, M.; Zhao, C.; Zhao, H.; McClain, C. J.; Feng, W.; Zhang, X. Hepatic and Fecal Metabolomic Analysis of the Effects of Lactobacillus rhamnosus GG on Alcoholic Fatty Liver Disease in Mice. *J. Proteome Res.* **2015***, 14*, 1174-1182.

134. Aura, A.; Mattila, I.; Hyötyläinen, T.; Gopalacharyulu, P.; Bounsaythip, C.; Orešič, M.; Oksman-Caldentey, K. Drug metabolome of the Simvastatin formed by human intestinal microbiota in vitro. *Mol. BioSyst.* **2011***, 7*, 437-446.

135. Trošt, K.; Ahonen, L.; Suvitaival, T.; Christiansen, N.; Nielsen, T.; Thiele, M.; Jacobsen, S.; Krag, A.; Rossing, P.; Hansen, T.; Dragsted, L. O.; Legido-Quigley, C. Describing the fecal metabolome in cryogenically collected samples from healthy participants. *Scientific Reports* **2020***, 10*, 885.

136. Zhang, C.; Yu, M.; Yang, Y.; Mu, C.; Su, Y.; Zhu, W. Effect of early antibiotic administration on cecal bacterial communities and their metabolic profiles in pigs fed diets with different protein levels. *Anaerobe* **2016***, 42*, 188-196.

137. Gao, X.; Pujos-Guillot, E.; Sébédio, J. Development of a quantitative metabolomic approach to study clinical human fecal water metabolome based on trimethylsilylation derivatization and GC/MS analysis. *Anal. Chem.* **2010***, 82*, 6447-6456.

138. Yang, Y.; Yin, Y.; Chen, X.; Chen, C.; Xia, Y.; Qi, H.; Baker, P. N.; Zhang, H.; Han, T. Evaluating different extraction solvents for GC-MS based metabolomic analysis of the fecal metabolome of adult and baby giant pandas. *Scientific Reports* **2019***, 9*, 12017.

139. Deda, O.; Gika, H. G.; Wilson, I. D.; Theodoridis, G. A. An overview of fecal sample preparation for global metabolic profiling. *Journal of Pharmaceutical and Biomedical Analysis* **2015***, 113*, 137-150.

140. Deda, O.; Chatziioannou, A. C.; Fasoula, S.; Palachanis, D.; Raikos, N; Theodoridis, G. A.; Gika, H. G. Sample preparation optimization in fecal metabolic profiling. *Journal of Chromatography B* **2017***, 1047*, 115-123.

141. Phua, L. C.; Koh, P. K.; Cheah, P. Y.; Ho, H. K.; Chan, E. C. Y. Global gas chromatography/time-of-flight mass spectrometry (GC/TOFMS)-based metabonomic profiling of lyophilized human feces. *Journal of Chromatography B* **2013***, 937*, 103-113.

142. Choo, J. M.; Leong, L. E. X.; Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Scientific reports* **2015***, 5*, 16350.

143. Ng, J. S. Y.; Ryan, U.; Trengove, R. D.; Maker, G. L. Development of an untargeted metabolomics method for the analysis of human faecal samples using Cryptosporidium-infected samples. *Molecular and Biochemical Parasitology* **2012***, 185*, 145-150.

144. Focant, J.; Reiner, E. J.; MacPherson, K.; Kolic, T.; Sjödin, A.; Patterson, D. G.; Reese, S. L.; Dorman, F. L.; Cochran, J. Measurement of PCDDs, PCDFs, and non-ortho-PCBs by comprehensive two-dimensional gas chromatography-isotope dilution time-of-flight mass spectrometry (GC × GC-IDTOFMS). *Talanta* **2004***, 63*, 1231-1240.

145. de Vos, J.; Dixon, R.; Vermeulen, G.; Gorst-Allman, P.; Cochran, J.; Rohwer, E.; Focant, J. Comprehensive two-dimensional gas chromatography time of flight mass spectrometry (GC×GC-TOFMS) for environmental forensic investigations in developing countries. *Chemosphere* **2011***, 82*, 1230-1239.

146. AnonymousGas Chromatography Problem Solving and Troubleshooting. *J. Chromatogr. Sci.* **1996***, 34*, 470-471.

147. Duran, M.; Haznedaroğlu, B. Z.; Zitomer, D. H. Microbial source tracking using host specific FAME profiles of fecal coliforms. *Water Res.* **2006***, 40*, 67-74.

148. Zeng, A. X.; Chin, S.; Nolvachai, Y.; Kulsing, C.; Sidisky, L. M.; Marriott, P. J. Characterisation of capillary ionic liquid columns for gas chromatography–mass spectrometry analysis of fatty acid methyl esters. *Anal. Chim. Acta* **2013***, 803*, 166-173.

149. Raman, M.; Ahmed, I.; Gillevet, P. M.; Probert, C. S.; Ratcliffe, N. M.; Smith, S.; Greenwood, R.; Sikaroodi, M.; Lam, V.; Crotty, P.; Bailey, J.; Myers, R. P.; Rioux, K. P. Fecal microbiome and volatile organic compound metabolome in obese humans with nonalcoholic fatty liver disease. *Clin. Gastroenterol. Hepatol.* **2013***, 11*, 868-3.

150. Misra, B. B. Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *Eur J Mass Spectrom (Chichester)* **2020***, 26*, 165-174.

151. Chen, Y.; Shen, G.; Zhang, R.; He, J.; Zhang, Y.; Xu, J.; Yang, W.; Chen, X.; Song, Y.; Abliz, Z. Combination of Injection Volume Calibration by Creatinine and MS Signals' Normalization to Overcome Urine Variability in LC-MS-Based Metabolomics Studies. *Anal. Chem.* **2013***, 85*, 7659-7665.

152. Cook, T.; Ma, Y.; Gamagedara, S. Evaluation of statistical techniques to normalize mass spectrometry-based urinary metabolomics data. *Journal of Pharmaceutical and Biomedical Analysis* **2020***, 177*, 112854.

153. Bidin, M. Z.; Shah, A. M.; Stanslas, J.; Seong, C. L. T. Blood and urine biomarkers in chronic kidney disease: An update. *Clinica Chimica Acta* **2019***, 495*, 239-250.

154. Herman-Saffar, O.; Boger, Z.; Libson, S.; Lieberman, D.; Gonen, R.; Zeiri, Y. Early non-invasive detection of breast cancer using exhaled breath and urine analysis. *Computers in Biology and Medicine* **2018***, 96*, 227-232.

155. Ryan, D.; Robards, K.; Prenzler, P. D.; Kendall, M. Recent and potential developments in the analysis of urine: A review. *Analytica Chimica Acta* **2011***, 684*, 17-29.

156. De Livera, A. M.; Dias, D. A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P. Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **2012***, 84*, 10768-10776.

157. Chetwynd, A. J.; Abdul-Sada, A.; Holt, S. G.; Hill, E. M. Use of a pre-analysis osmolality normalisation method to correct for variable urine concentrations and for improved metabolomic analyses. *Journal of Chromatography A* **2016***, 1431*, 103-110.

158. Chadha, V.; Garg, U.; Alon, U. S. Measurement of urinary concentration: A critical appraisal of methodologies. *Pediatr. Nephrol.* **2001***, 16*, 374-382.

159. Gagnebin, Y.; Tonoli, D.; Lescuyer, P.; Ponte, B.; de Seigneux, S.; Martin, P.; Schappler, J.; Boccard, J.; Rudaz, S. Metabolomic analysis of urine samples by

UHPLC-QTOF-MS: Impact of normalization strategies. *Analytica Chimica Acta* **2017***, 955*, 27-35.

160. Wagner, B. D.; Accurso, F. J.; Laguna, T. A. The applicability of urinary creatinine as a method of specimen normalization in the cystic fibrosis population. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society* **2010***, 9*, 212-216.

161. Tang, K. W. A.; Toh, Q. C.; Teo, B. W. Normalisation of urinary biomarkers to creatinine for clinical practice and research--when and why. *Singapore Med. J.* **2015***, 56*, 7-10.

162. Waikar, S. S.; Sabbisetti, V. S.; Bonventre, J. V. Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate. *Kidney International* **2010***, 78*, 486-494.

163. Boeniger, M. F.; Lowry, L. K.; Rosenberg, J. Interpretation of urine results used to assess chemical exposure with emphasis on creatinine adjustments: a review. *Am Ind Hyg Assoc J* **1993***, 54*, 615-627.

164. Garde, A. H.; Hansen, Å M.; Kristiansen, J.; Knudsen, L. E. Comparison of Uncertainties Related to Standardization of Urine Samples with Volume and Creatinine Concentration. *Ann Occup Hyg* **2004***, 48*, 171-179.

165. Barr Dana, B.; Wilder Lynn, C.; Caudill Samuel, P.; Gonzalez Amanda, J.; Needham Lance, L.; Pirkle James, L. Urinary Creatinine Concentrations in the U.S. Population: Implications for Urinary Biologic Monitoring Measurements. *Environ. Health Perspect.* **2005***, 113*, 192-200.

166. Miller, R. C.; Brindle, E.; Holman, D. J.; Shofer, J.; Klein, N. A.; Soules, M. R.; O'Connor, K.,A. Comparison of Specific Gravity and Creatinine for Normalizing Urinary Reproductive Hormone Concentrations. *Clin. Chem.* **2004***, 50*, 924-932.

167. Zhang, T.; Watson, D. G. A short review of applications of liquid chromatography mass spectrometry based metabolomics techniques to the analysis of human urine. *Analyst* **2015***, 140*, 2907-2915.

168. Mattarucchi, E.; Guillou, C. Critical aspects of urine profiling for the selection of potential biomarkers using UPLC-TOF-MS. *Biomed. Chromatogr.* **2012***, 26*, 512-517.

169. Vogl, F. C.; Mehrl, S.; Heizinger, L.; Schlecht, I.; Zacharias, H. U.; Ellmann, L.; Nürnberger, N.; Gronwald, W.; Leitzmann, M. F.; Rossert, J.; Eckardt, K.; Dettmer, K.; Oefner, P. J.; GCKD, S. I. Evaluation of dilution and normalization

strategies to correct for urinary output in HPLC-HRTOFMS metabolomics. *Analytical and Bioanalytical Chemistry* **2016***, 408*, 8483-8493.

170. Wu, Y.; Li, L. Determination of Total Concentration of Chemically Labeled Metabolites as a Means of Metabolome Sample Normalization and Sample Loading Optimization in Mass Spectrometry-Based Metabolomics. *Anal. Chem.* **2012***, 84*, 10723-10731.

171. AnonymousPeer Reviewed: Isotopically Labeled Analogues for Drug Quantitation. *Anal. Chem.* **2002***, 74*, 618 A-626 A.

172. Chen, J.; Zhang, P.; Lv, M.; Guo, H.; Huang, Y.; Zhang, Z.; Xu, F. Influences of Normalization Method on Biomarker Discovery in Gas Chromatography–Mass Spectrometry-Based Untargeted Metabolomics: What Should Be Considered? *Anal. Chem.* **2017***, 89*, 5342-5348.

173. Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011***, 83*, 1079-1087.

174. Sinkov, N. A.; Harynuk, J. J. Three-dimensional cluster resolution for guiding automatic chemometric model optimization. *Talanta* **2013***, 103*, 252-259.

175. Sinkov, N. A.; Sandercock, P. M. L.; Harynuk, J. J. Chemometric classification of casework arson samples based on gasoline content. *Forensic Sci. Int.* **2014***, 235*, 24-31.

176. Kawamura, M.; Ohmoto, A.; Hashimoto, T.; Yagami, F.; Owada, M.; Sugawara, T. Second morning urine method is superior to the casual urine method for estimating daily salt intake in patients with hypertension. *Hypertension Research* **2012***, 35*, 611-616.

177. Kavouras, S. A.; Johnson, E. C.; Bougatsas, D.; Arnaoutis, G.; Panagiotakos, D. B.; Perrier, E.; Klein, A. Validation of a urine color scale for assessment of urine osmolality in healthy children. *Eur. J. Nutr.* **2016***, 55*, 907-915.

178. Witte, E. C.; Lambers Heerspink, H.,J.; de Zeeuw, D.; Bakker, S. J. L.; de Jong, P.,E.; Gansevoort, R. First morning voids are more reliable than spot urine samples to assess microalbuminuria. *Journal of the American Society of Nephrology : JASN* **2009***, 20*, 436-443.

179. Bro, R.; Smilde, A. K. Principal component analysis. *Anal. Methods* **2014***, 6*, 2812-2831.

180. Emwas, A.; Saccenti, E.; Gao, X.; McKay, R. T.; Dos Santos, V.,A.P.Martins; Roy, R.; Wishart, D. S. Recommended strategies for spectral processing and post-processing of 1D (1)H-NMR data of biofluids with a particular focus on urine. *Metabolomics : Official journal of the Metabolomic Society* **2018***, 14*, 31.

181. Caldeira, M.; Perestrelo, R.; Barros, A. S.; Bilelo, M. J.; Morête, A.; Câmara, J. S.; Rocha, S. M. Allergic asthma exhaled breath metabolome: A challenge for comprehensive two-dimensional gas chromatography. *Journal of Chromatography A* **2012***, 1254*, 87-97.

182. Pena-Abaurrea, M.; Jobst, K. J.; Ruffolo, R.; Shen, L.; McCrindle, R.; Helm, P. A.; Reiner, E. J. Identification of Potential Novel Bioaccumulative and Persistent Chemicals in Sediments from Ontario (Canada) Using Scripting Approaches with GC×GC-TOF MS Analysis. *Environ. Sci. Technol.* **2014***, 48*, 9591-9599.

183. Hashimoto, S.; Takazawa, Y.; Fushimi, A.; Tanabe, K.; Shibata, Y.; Ieda, T.; Ochiai, N.; Kanda, H.; Ohura, T.; Tao, Q.; Reichenbach, S. E. Global and selective detection of organohalogens in environmental samples by comprehensive two-dimensional gas chromatography–tandem mass spectrometry and high-resolution time-of-flight mass spectrometry. *Journal of Chromatography A* **2011***, 1218*, 3799-3810.

184. S Haglund, P.; Löfstrand, K.; Siek, K.; Asplund, L. Powerful GC-TOF-MS Techniques for Screening, Identification and Quantification of Halogenated Natural Products. *Mass spectrometry (Tokyo, Japan)* **2013***, 2*, S0018.

185. Weggler, B. A.; Gröger, T.; Zimmermann, R. Advanced scripting for the automated profiling of two-dimensional gas chromatography-time-of-flight mass spectrometry data from combustion aerosol. *Journal of Chromatography A* **2014***, 1364*, 241-248.

186. Gómez, M. J.; Herrera, S.; Solé, D.; García-Calvo, E.; Fernández-Alba, A. R. Automatic Searching and Evaluation of Priority and Emerging Contaminants in Wastewater and River Water by Stir Bar Sorptive Extraction followed by Comprehensive Two-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry. *Anal. Chem.* **2011***, 83*, 2638-2647.

187. Welthagen, W.; Schnelle-Kreis, J.; Zimmermann, R. Search criteria and rules for comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry analysis of airborne particulate matter. *Journal of Chromatography A* **2003***, 1019*, 233-249.

188. Reichenbach, S. E.; Kottapalli, V.; Ni, M.; Visvanathan, A. Computer language for identifying chemicals with comprehensive two-dimensional gas chromatography and mass spectrometry. *Journal of Chromatography A* **2005***, 1071*, 263-269.

189. Brokl, M.; Bishop, L.; Wright, C. G.; Liu, C.; McAdam, K.; Focant, J. Analysis of mainstream tobacco smoke particulate phase using comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *Journal of Separation Science* **2013***, 36*, 1037-1044.

190. Gröger, T.; Welthagen, W.; Mitschke, S.; Schäffer, M.; Zimmermann, R. Application of comprehensive two-dimensional gas chromatography mass spectrometry and different types of data analysis for the investigation of cigarette particulate matter. *Journal of Separation Science* **2008***, 31*, 3366-3374.

191. AnonymousInterpretation of Mass Spectra, 4th ed. (McLafferty, Fred W.; Turecek, Frantisek). *J. Chem. Educ.* **1994***, 71*, A54.

192. Nicolescu, T. O. Interpretation of Mass Spectra. *Mass Spectrometry* **2017**.

193. Gross, J. H.; Roepstorff, P. *Mass Spectrometry: A Textbook;* Springer Berlin Heidelberg: 2011; .

194. Dass, C. *Fundamentals of Contemporary Mass Spectrometry;* Wiley: 2007; .

195. Mukhtar Abdul-Bari, M.; McQueen, R. H.; Paulina de la Mata, A; Batcheller, J. C.; Harynuk, J. J. Retention and release of odorants in cotton and polyester fabrics following multiple soil/wash procedures. *Text. Res. J.* **2020**, 0040517520914411.

196. Dixon, E.; Clubb, C.; Pittman, S.; Ammann, L.; Rasheed, Z.; Kazmi, N.; Keshavarzian, A.; Gillevet, P.; Rangwala, H.; Couch, R. D. Solid-phase microextraction and the human fecal VOC metabolome. *PloS one* **2011***, 6*, e18471.

197. Weber, B. M.; Walsh, P.; Harynuk, J. J. Determination of Hydrocarbon Group-Type of Diesel Fuels by Gas Chromatography with Vacuum Ultraviolet Detection. *Anal. Chem.* **2016***, 88*, 5809-5817.

198. Striebich, R. C.; Shafer, L. M.; Adams, R. K.; West, Z. J.; DeWitt, M. J.; Zabarnick, S. Hydrocarbon Group-Type Analysis of Petroleum-Derived and Synthetic Fuels Using Two-Dimensional Gas Chromatography. *Energy Fuels* **2014***, 28*, 5696-5706.

199. Rubulotta, G.; Quadrelli, E. A. Chapter 11 - Terpenes: A Valuable Family of Compounds for the Production of Fine Chemicals. *Studies in Surface Science and Catalysis* **2019***, 178*, 215-229.

200. Graves, B. M.; Johnson, T. J.; Nishida, R. T.; Dias, R. P.; Savareear, B.; Harynuk, J. J.; Kazemimanesh, M.; Olfert, J. S.; Boies, A. M. Comprehensive

characterization of mainstream marijuana and tobacco smoke. *Scientific Reports* **2020***, 10*, 7160.

201. Booth, J. K.; Bohlmann, J. Terpenes in Cannabis sativa – From plant genome to humans. *Plant Science* **2019***, 284*, 67-72.

202. Tetali, S. D. Terpenes and isoprenoids: a wealth of compounds for global use. *Planta* **2019***, 249*, 1-8.

203. Jiang, Z.; Kempinski, C.; Chappell, J. Extraction and Analysis of Terpenes/Terpenoids. *Current protocols in plant biology* **2016***, 1*, 345-358.

204. Davies, F. K.; Jinkerson, R. E.; Posewitz, M. C. Toward a photosynthetic microbial platform for terpenoid engineering. *Photosynthesis Res.* **2015***, 123*, 265-284.

205. Karunanithi, P. S.; Zerbe, P. Terpene Synthases as Metabolic Gatekeepers in the Evolution of Plant Terpenoid Chemical Diversity. *Frontiers in Plant Science* **2019***, 10*, 1166.

206. Awouafack, M. D.; Tane, P.; Kuete, V.; Eloff, J. N. 2 - Sesquiterpenes from the Medicinal Plants of Africa. *Medicinal Plant Research in Africa* **2013**, 33-103.

207. Szczepaniak, L.; Isidorov, V. A. Gas chromatographic retention indices of trimethylsilyl derivatives of terpene alcohols. *Journal of Chromatography A* **2011***, 1218*, 7061-7064.

208. Kupska, M.; Wasilewski, T.; Jędrkiewicz, R.; Gromadzka, J.; Namieśnik, J. Determination of Terpene Profiles in Potential Superfruits. *Int. J. Food Prop.* **2016***, 19*, 2726-2738.

209. Justin, F.; Frank, V. D. K.; Robert, V. Cannabinoid Receptor 1 Binding Activity and Quantitative Analysis of Cannabis sativa L. Smoke and Vapor. *Chemical & Pharmaceutical Bulletin* **2010***, 58*, 201-208.

210. Bernstein, N.; Gorelick, J.; Koch, S. Interplay between chemistry and morphology in medical cannabis (Cannabis sativa L.). *Industrial Crops and Products* **2019***, 129*, 185-194.

211. Chowdhury, F. T.; Islam, M. R.; Islam, M. R.; Khan, H. Diversity of Plant Endophytic Volatile Organic Compound (VOC) and Their Potential Applications. In *Endophytes and Secondary Metabolites*; Jha, S., Ed.; Springer International Publishing: Cham, 2018; pp 1-27.

212. Gershenzon, J.; Dudareva, N. The function of terpene natural products in the natural world. *Nature Chemical Biology* **2007***, 3*, 408-414.

213. Jackson, R. S. 6 - Chemical Constituents of Grapes and Wine. *Wine Science (Third Edition)* **2008**, 270-331.

214. Yermakov, A. I.; Khlaifat, A. L. Characteristics Of The GC-MS Mass Spectra Of Terpenoids (C10H16). *Chemical sciences journal* **2010***, 1*.

215. HIROSE, Y. Mass Spectra of Sesquiterpenes. *Journal of the Mass Spectrometry Society of Japan* **1967***, 15*, 162-178.

216. LECO About LECO Corporation. *https://www.leco.com/about* (accessed Nov 25, 2020).

217. Ruiz-Jiménez, J.; Parshintsev, J.; Laitinen, T.; Hartonen, K.; Riekkola, M.; Petäjä, T.; Kulmala, M. Comprehensive two-dimensional gas chromatography, a valuable technique for screening and semiquantitation of different chemical compounds in ultrafine 30 nm and 50 nm aerosol particles. *J. Environ. Monit.* **2011***, 13*, 2994-3003.

218. Castillo, S.; Mattila, I.; Miettinen, J.; Orešič, M.; Hyötyläinen, T. Data Analysis Tool for Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry. *Anal. Chem.* **2011***, 83*, 3058-3067.

219. Reichenbach, S. E.; Kottapalli, V.; Ni, M.; Visvanathan, A. Computer language for identifying chemicals with comprehensive two-dimensional gas chromatography and mass spectrometry. *Journal of Chromatography A* **2005***, 1071*, 263-269.

220. S Haglund, P.; Löfstrand, K.; Siek, K.; Asplund, L. Powerful GC-TOF-MS Techniques for Screening, Identification and Quantification of Halogenated Natural Products. *Mass spectrometry (Tokyo, Japan)* **2013***, 2*, S0018.

221. Gross, R. A. A Mass Spectral Chlorine Rule for Use in Structure Determinations in Sophomore Organic Chemistry. *J. Chem. Educ.* **2004***, 81*, 1161.

222. Meija, J. Understanding Isotopic Distributions in Mass Spectrometry. *J. Chem. Educ.* **2006***, 83*, 1761.

223. Liu, D.; Zhao, N.; Wang, M.; Pi, X.; Feng, Y.; Wang, Y.; Tong, H.; Zhu, L.; Wang, C.; Li, E. Urine volatile organic compounds as biomarkers for minimal change type nephrotic syndrome. *Biochemical and Biophysical Research Communications* **2018***, 496*, 58-63.

224. Martinez-Vernon, A.; Covington, J. A.; Arasaradnam, R. P.; Esfahani, S.; O'Connell, N.; Kyrou, I.; Savage, R. S. An improved machine learning pipeline for urinary volatiles disease detection: Diagnosing diabetes. *PloS one* **2018***, 13*, e0204425.

225. Zlatkis, A.; Brazell, R. S.; Poole, C. F. The role of organic volatile profiles in clinical diagnosis. *Clin. Chem.* **2019***, 27*, 789-797.

226. Monteiro, M.; Carvalho, M.; Henrique, R.; Jerónimo, C.; Moreira, N.; de Lourdes Bastos, M.; de Pinho, P. G. Analysis of volatile human urinary metabolome by solid-phase microextraction in combination with gas chromatography–mass spectrometry for biomarker discovery: Application in a pilot study to discriminate patients with renal cell carcinoma. *European Journal of Cancer* **2014***, 50*, 1993-2002.

227. Monteiro, M.; Moreira, N.; Pinto, J.; Pires-Luís, A.,S.; Henrique, R.; Jerónimo, C.; Bastos, M. d. L.; Gil, A. M.; Carvalho, M.; Guedes de Pinho, P. GC-MS metabolomics-based approach for the identification of a potential VOC-biomarker panel in the urine of renal cell carcinoma patients. *J. Cell. Mol. Med.* **2017***, 21*, 2092-2105.

228. Erb, A.; Marsan, P.; Burgart, M.; Remy, A.; Lambert-Xolin, A.; Jeandel, F.; Hanser, O.; Robert, A. Simultaneous determination of aromatic and chlorinated compounds in urine of exposed workers by dynamic headspace and gas chromatography coupled to mass spectrometry (dHS-GC–MS). *Journal of Chromatography B* **2019***, 1125*, 121724.

229. Wright, H.; Bannaga, A. S.; Iriarte, R.; Mahmoud, M.; Arasaradnam, R. P. Utility of volatile organic compounds as a diagnostic tool in preterm infants. *Pediatr. Res.* **2020**.

230. Di Lena, M.; Porcelli, F.; Altomare, D. F. Volatile organic compounds as new biomarkers for colorectal cancer: a review. *Colorectal Disease* **2016***, 18*, 654-663.

231. van Keulen, K. E.; Jansen, M. E.; Schrauwen, R. W. M.; Kolkman, J. J.; Siersema, P. D. Volatile organic compounds in breath can serve as a non-invasive diagnostic biomarker for the detection of advanced adenomas and colorectal cancer. *Alimentary Pharmacology and Therapeutics* **2020***, 51*, 334-346.

232. Laffel, L. Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes. *Diabetes Metab. Res. Rev.* **1999***, 15*, 412-426.

233. Woollam, M.; Teli, M.; Angarita-Rivera, P.; Liu, S.; Siegel, A. P.; Yokota, H.; Agarwal, M. Detection of Volatile Organic Compounds (VOCs) in Urine via Gas Chromatography-Mass Spectrometry QTOF to Differentiate Between Localized and Metastatic Models of Breast Cancer. *Scientific Reports* **2019***, 9*, 2526.

234. Huang, S.; Chen, G.; Ye, N.; Kou, X.; Zhu, F.; Shen, J.; Ouyang, G. Solid-phase microextraction: An appealing alternative for the determination of endogenous substances - A review. *Analytica Chimica Acta* **2019***, 1077*, 67-86.

235. Kataoka, H.; Lord, H. L.; Pawliszyn, J. SOLID-PHASE MICROEXTRACTION | Biomedical Applications. *Encyclopedia of Separation Science* **2000**, 4153-4169.

236. de Koning, S.; Janssen, H.; Brinkman, U. A. T. Modern Methods of Sample Preparation for GC Analysis. *Chromatographia* **2009***, 69*, 33.

237. Liberto, E.; Bicchi, C.; Cagliero, C.; Cordero, C.; Rubiolo, P.; Sgorbini, B. Chapter 1 Headspace Sampling: An “Evergreen” Method in Constant Evolution to Characterize Food Flavors through their Volatile Fraction. In *Advanced Gas Chromatography in Food Analysis*The Royal Society of Chemistry: 2020; pp 1-37.

Appendix A

Preparation of Standard Mixtures for Evaluation of the Scripts

For the standard mixture of each class of compounds, high and low concentration solutions were prepared with a serial dilution from the stock solution.

A) Amino acid standard

Stock solution: 2.5 µmoles/mL for each component (AAS18-10mL analytical standard, Millipore-Sigma Canada)

- 100 fold dilution (25 nmoles/mL) was made in methanol to make a high-concentration standard solution.
- 1000 fold dilution (2.5 nmoles/mL) was made in methanol to make a low-concentration standard solution.

B) Fatty acid standard

Stock solution: 35.1 mg of total fatty acids (GLC-744, Nu-Check, MN, USA) were dissolved in 2.2 mL of methanol and isopropanol solution (2 mL MeOH and 0.2 mL IPA). This corresponds to 319 µg/mL for each component.

- 100 fold dilution (3 µg/g) was made in methanol to make a high-concentration standard solution.
- 1000 fold dilution (0.3 µg/g) was made in methanol to make a low-concentration standard solution.

C) Fatty acid methyl ester standard

Stock solution: 400 µg/g for each component (SUPELCO 37 Component FAME Mix, Millipore-Sigma Canada)

- 100 fold dilution (4 µg/g) was made in dichloromethane to make a high-concentration standard solution.
- 1000 fold dilution (0.4 µg/g) was made in dichloromethane to make a low-concentration standard solution.

D) Carbohydrates

Stock solution: 100 µg/mL for each component (Carbohydrates kit, Millipore-Sigma Canada) was made in methanol.

- 100 fold dilution (1 µg/mL) was made in methanol to make a high-concentration standard solution.
- 1000 fold dilution (0.1 µg/mL) was made in methanol to make a low-concentration standard solution.

The indicated volumes (µL) in Table A-1 were withdrawn from high or low concentrations of the corresponding standard solution into a GC vial for derivatization.

Table A-1. The Details of How the Standard Mixtures were Prepared

Vial #	Conc	AA	FA	FAME	Carbohydrate
1	Low Con	10	10	10	10
2	Low Con	20	20	20	20
3	Low Con	40	40	40	40
4	Low Con	60	60	60	60
5	Low Con	80	80	80	80
6	Low Con	100	100	100	100
7	High Conc	20	20	20	20
8	High Conc	40	40	40	40
9	High Conc	60	60	60	60
10	High Conc	80	80	80	80
11	High Conc	100	100	100	100
12	Mix	L20	L20	H20	H20
13	Mix	H20	H20	L20	L20
14	Mix	L40	H40	L40	H40
15	Mix	H40	L40	H40	L40
16	Mix	L40	H40	H40	L40
17	Mix	L20	L20	L20	H40
18	Mix	L20	L20	L20	H100

Table A-2. The Detailed Results of How Many Compounds in the Standard Mixtures at Various Concentrations were Classified for Each Group

Classes	Subgroups	mix 1	mix 2	mix 3	mix 4	mix 5	mix 6
Carbohydrates	4TMS	0	0	0	3	1	3
	5TMS	3	4	6	4	7	5
	8TMS	0	0	1	1	2	2
Fatty acids	saturated	8	11	15	17	17	17
	monoenoic	2	6	9	11	11	11
	dienoic	1	0	2	2	1	3
	multienoic	0	1	4	6	9	8
Fatty acid methyl esters	saturated	0	1	5	8	8	8
	monoenoic	0	1	3	4	4	4
	dienoic	0	0	0	2	2	3
	trienoic	0	0	0	1	2	2
	multienoic	0	0	0	0	1	2
Amino acids		0	1	4	3	6	6
Total Number Correct		**14**	**25**	**49**	**62**	**71**	**74**
Total Number Incorrect		**1**	**0**	**0**	**2**	**1**	**0**
Total Number Classified		**15**	**25**	**49**	**64**	**72**	**74**
Classes	**Subgroups**	**mix 7**	**mix 8**	**mix 9**	**mix 10**	**mix 11**	**mix 12**
Carbohydrates	4TMS	4	4	6	3	4	4
	5TMS	7	10	8	7	5	8
	8TMS	5	5	5	4	3	5
Fatty acids	saturated	17	17	18	20	20	11
	monoenoic	11	11	12	10	10	5
	dienoic	3	3	3	3	4	0
	multienoic	9	9	9	9	11	0
Fatty acid methyl esters	saturated	9	10	10	11	12	8
	monoenoic	7	5	7	6	5	7
	dienoic	2	3	3	3	3	2
	trienoic	3	2	2	2	2	1
	multienoic	2	2	3	2	1	2
Amino acids		7	10	11	12	13	1
Total Number Correct		**86**	**91**	**97**	**92**	**93**	**54**
Total Number Incorrect		**0**	**0**	**0**	**1**	**2**	**0**
Total Number Classified		**86**	**91**	**97**	**93**	**95**	**54**
Classes	**Subgroups**	**mix 13**	**mix 14**	**mix 15**	**mix 16**	**mix 17**	**mix 18**
Carbohydrates	4TMS	2	5	2	3	4	3
	5TMS	3	8	2	4	9	9
	8TMS	0	5	0	0	5	5
Fatty acids	saturated	16	17	15	16	10	12
	monoenoic	9	10	9	10	4	3
	dienoic	3	3	0	3	0	0
	multienoic	9	9	5	10	0	1
Fatty acid methyl esters	saturated	3	4	10	10	0	0
	monoenoic	2	3	7	6	2	0
	dienoic	0	2	3	3	0	0

	trienoic	0	0	3	2	0	0
	multienoic	0	0	3	2	0	0
Amino acids		8	3	9	3	3	8
Total Number Correct		**55**	**69**	**68**	**72**	**37**	**41**
Total Number Incorrect		**0**	**0**	**0**	**0**	**0**	**1**
Total Number Classified		**55**	**69**	**68**	**72**	**37**	**42**

Table A-3. Compounds Classified for Each Terpene Group

Classification	Compound	Formula	CAS
Monoterpene hydrocarbons	p-Cymene	C10H14	99-87-6
	o-Cymene	C10H14	527-84-4
	Cosmene	C10H14	460-01-5
	1,3,8-p-Menthatriene	C10H14	18368-95-1
	p-Mentha-1,5,8-triene	C10H14	21195-59-5
	Cyclene	C10H16	508-32-7
	β-Thujene	C10H16	28634-89-1
	α-Pinene	C10H16	80-56-8
	β-Pinene	C10H16	127-91-3
	Camphene	C10H16	79-92-5
	Sabinene	C10H16	3387-41-5
	2-Bornene	C10H16	464-17-5
	β-Myrcene	C10H16	123-35-3
	α-Phellandrene	C10H16	99-83-2
	α-Terpinene	C10H16	99-86-5
	Limonene	C10H16	138-86-3
	L-Limonene	C10H16	5989-54-8
	trans-β-Ocimene	C10H16	3779-61-1
	cis-β-Ocimene	C10H16	3338-55-4
	β-Ocimene	C10H16	13877-91-3
	(4E,6E)-Allocimene	C10H16	3016-19-1
	Allo-Ocimene	C10H16	673-84-7
	γ-Terpinene	C10H16	99-85-4
	Terpinolene	C10H16	586-62-9
	2-Carene	C10H16	554-61-0
	3-Carene	C10H16	13466-78-9
	Cyclohexene, 4-methyl-3-(1-methylethylidene)-	C10H16	99805-90-0
	p-Mentha-2,4(8)-diene	C10H16	586-63-0
	β-Phellandrene	C10H16	555-10-2
Monoterpene alcohols	cis-Carveol	C10H16O	1197-06-4
	δ-Terpineol	C10H18O	7299-42-5
	endo-Borneol	C10H18O	507-70-0
	Terpinen-4-ol	C10H18O	562-74-3
	4-Thujanol	C10H18O	546-79-2
	Linalool	C10H18O	78-70-6
	p-Menth-2-en-1-ol	C10H18O	619-62-5
	cis-2-p-Menthen-1-ol	C10H18O	29803-82-5

	Isopulegol	C10H18O	89-79-2
	Eucalyptol	C10H18O	470-82-6
	Geraniol	C10H18O	106-24-1
	p-Menthan-3-ol	C10H20O	1490-04-6
	p-Menthan-1-ol	C10H20O	21129-27-1
	3-p-Menthol	C10H20O	15356-70-4
	2-Methylisoborneol	C11H20O	2371-42-8
Monoterpene ketones / oxidation products	Carvone	C10H14O	99-49-0
	cis-Ocimenone	C10H14O	33746-71-3
	Fenchone	C10H16O	1195-79-5
	Pulegone	C10H16O	89-82-7
	Piperitone	C10H16O	89-81-6
	(-)-Camphor	C10H16O	464-48-2
	p-Menthone	C10H18O	89-80-5
	Menthyl acetate	C12H22O2	16409-45-3
	Isomenthyl acetate	C12H22O2	20777-45-1
	Nerol acetate	C12H20O2	141-12-8
	Geranyl acetate	C12H20O2	105-87-3
	Limonene dioxide	C10H16O2	96-08-2
	Limonene oxide, cis-	C10H16O	13837-75-7
Monoterpene aldehydes	α-Citral	C10H16O	141-27-5
	β-Citral	C10H16O	106-26-3
	Perillal	C10H14O	2111-75-3
	Anethole	C10H12O	104-46-1
	Estragole	C10H12O	140-67-0
Sesquiterpene hydrocarbons	α-Corocalene	C15H20	20129-39-9
	Isolongifolene, 4,5,9,10-dehydro-	C15H20	156747-45-4
	α-Curcumene	C15H22	644-30-4
	β-Elemene	C15H24	515-13-9
	γ-Elemene	C15H24	29873-99-2
	Caryophyllene	C15H24	87-44-5
	α-Bergamotene	C15H24	17699-05-7
	cis-à-Bergamotene	C15H24	18252-46-5
	Alloaromadendrene	C15H24	25246-27-9
	β-Chamigrene	C15H24	18431-82-8
	Germacrene B	C15H24	15423-57-1
	à-Guaiene	C15H24	3691-12
	4,11,11-trimethyl-8-methylenebicyclo[7.2.0]undec-3-ene	C15H24	889360-49-0
	7-epi-Sesquithujene	C15H24	159407-35-9

	γ-Amorphene	C15H24	6980-46-7
	Bicyclogermacrene	C15H24	24703-35-3
	β-Sesquiphellandrene	C15H24	20307-83-9
	Copaene	C15H24	3856-25-5
	(-)-β-Bourbonene	C15H24	5208-59-3
	β-Gurjunene	C15H24	17334-55-3
	α-Farnesene	C15H24	502-61-4
	cis-β-Farnesene	C15H24	28973-97-9
	trans-β-Farnesene	C15H24	18794-84-8
	Germacrene D	C15H24	23986-74-5
	α-Zingiberene	C15H24	495-60-3
	β-Bisabolene	C15H24	495-61-4
	β-Sesquiphellandrene	C15H24	20307-83-9
	trans-γ-Bisabolene	C15H24	53585-13-0
	α-Muurolene	C15H24	31983-22-9
	α-Gurjunene	C15H24	489-40-7
	α-Selinene	C15H24	473-13-2
	γ-Gurjunene	C15H24	22567-17-5
	Humulene	C15H24	6753-98-6
	γ-Cadinene	C15H24	39029-41-9
	Isocaryophyllene	C15H24	118-65-0
	cis-α-Bisabolene	C15H24	29837-07-8
	Ylangene	C15H24	14912-44-8
Sesquiterpene oxidation products	Zerumbone	C15H22O	471-05-6
	Farnesene epoxide, E-	C15H24O	83637-40-5
	Caryophyllene oxide	C15H24O	1139-30-6
	2,6,10-Dodecatrienal, 3,7,11-trimethyl-	C15H24O	19317-11-4
	Eudesma-4(15),7-dien-1β -ol	C15H24O	119120-23-9
	(-)-Spathulenol	C15H24O	77171-55-2
	6,10-Dodecadien-1-yn-3-ol, 3,7,11-trimethyl-	C15H24O	2387-68-0
	Humulene epoxide I	C15H24O	19888-33-6
	Ylangenol	C15H24O	41610-69-9
	α-Bisabolol	C15H26O	515-69-5
	trans-Nerolidol	C15H26O	40716-66-3
	Ledol	C15H26O	577-27-5
	Epicubebol	C15H26O	38230-60-3
	Cubebol	C15H26O	23445-02-5
	Elemol	C15H26O	639-99-6
	trans-Sesquisabinene hydrate	C15H26O	145512-84-1
	Pogostole	C15H26O	21698-41-9

	Zingiberenol	C15H26O	58334-55-7
	β-Eudesmol	C15H26O	473-15-4
	γ-Eudesmol	C15H26O	1209-71-8
	α-Cadinol	C15H26O	481-34-5
	4aH-cycloprop[e]azulen-4a-ol, decahydro-1,1,4,7-tetramethyl-	C15H26O	95975-84-1
	trans-Sesquisabinene hydrate	C15H26O	145512-84-1
	α-Bisabolol	C15H26O	515-69-5
	β-Bisabolol	C15H26O	15352-77-9
	Nerolidol	C15H26O	7212-44-4
	(E)-Farnesyl acetate	C17H28O2	4128-17-0
Diterpenes	(E,E)-7,11,15-Trimethyl-3-methylene-hexadeca-1,6,10,14-tetraene	C20H32	70901-63-2
	m-Camphorene	C20H32	20016-73-3
	α-Springene	C20H32	77898-97-6
Cannabinoids	Cannabichromene	C21H30O2	20675-51-8
	Cannabicyclol	C21H30O2	21366-63-2
	DELTA.8-Tetrahydrocannabinol	C21H30O2	5957-75-5
	Dronabinol	C21H30O2	1972-08
	Cannabidivarol	C19H26O2	24274-48-4
	Podocarpa-8,11,13-triene-1,3-dione, 14-isopropyl-13-methoxy-	C21H28O3	15372-53-9
	.DELTA.-1(2)-TETRAHYDROCANNABINOL	C21H30O2	16849-50-6
	Cannabidiol	C21H30O2	13956-29-1
	δ9-Tetrahydrocannabivarin	C19H26O2	31262-37-0
	.DELTA.9-Tetrahydrocannabinol, TMS derivative	C24H38O2Si	55449-68-8
	(-)-trans-Cannabidiol	C21H30O2	521-37-9
	Cannabigerol	C21H32O2	25654-31-3
	Methoxy-THC	C22H32O2	N/A

Appendix B

Scripts used for classification of compounds

Non-derivatized

```
'*************************************************************************
function straightchain_aldehyde() as Boolean
'high abundance m/z41, 55
Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
```

```
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed1 = (Rank(1) = 41 or Rank(1) = 44 or rank(1)=57)
bed2 = abundance(41)>600 and abundance(42)<500 and abundance(44)>200 and
abundance(45)>50
bed3 = abundance(55)>150 and abundance(57)>400
bed4 = abundance(74)<50
bed5 = (abundance(57)/abundance(55)>0.95)
bed6 = abundance(82)>abundance(83)

ionseriescounter1 = 0
for k = 47 to 52 step 1
        if abundance(k)>60 then
                ionseriescounter1 = ionseriescounter1+1
        end if
next
ionseriescounter2 = 0
for m = 74 to 79 step 1
        if abundance(m)>50 then
                ionseriescounter2 = ionseriescounter2+1
        end if
next
ionseriescounter3 = 0
for n = 87 to 94 step 1
```

```
        if abundance(n)>60 then
                ionseriescounter3 = ionseriescounter3+1
        end if
next

Carbon_number = 4
Do while Carbon_number<=30
        Expected_MW = 14*Carbon_number +16
        m18 = Expected_MW -18
        m28 = Expected_MW -28
        m43 = Expected_MW -43
        m44 = Expected_MW -44
        noise_counter = 0
        for noisecheck = Expected_MW+2 to Em step 1
                If Intensity(noisecheck) <= (ave_intensity + 6*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter = noise_counter + 1
                end if
        next
If noise_counter < 5 and abundance(m28)>0 and abundance(m43)>0 and
abundance(m44)>10 and ionseriescounter1<=1 and ionseriescounter2<=1 and
ionseriescounter3<=1 and bed1 and bed2 and bed3 and bed4 and bed5 and bed6 then
straightchain_aldehyde =true
Carbon_number = Carbon_number+1
Loop
```

```
end function
‘*************************************************************************
Function Ketones()as Boolean
'base peak 43
'high abundance peak ion series m/z 71,85,99
Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity
Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
```

```
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5


bed1 = rank(1) = 43 or rank(1) = 58
bed2 = abundance(43)>700 and abundance(71)>50 and abundance(57)<200 and
abundance(59)>50 and abundance(58)>200


fragment_counter = 0
for n = 5 to 30  step 1
        fragment_test =n*14+1
        if abundance(fragment_test)>10 then
                fragment_counter= fragment_counter +1
        end if
next


Carbon_number = 5
Do while Carbon_number<=30
Expected_MW = 14*Carbon_number+16
M43 = Expected_MW - 43
noise_counter1 = 0
noise_counter2 = 0
        for noisecheck = Expected_MW+2 to Em step 1
                If Intensity(noisecheck) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck)
                end if
```

```
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
If bed1 and bed2 and abundance(M43)>0 and fragment_counter >= (Carbon_number-5) and noise_counter1 <=5 and noise_counter2<5 then Ketones = true
Carbon_number = Carbon_number+1
Loop
end function
‘*************************************************************************
function alcohols2() as Boolean
'2-alcohol
‘base peak m/z45
'ion series 55,69,83,97,111,125
'aldehyde 'ion series 68,82,96,110,124,138

Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity
Em = Endmass()
sum_intensity = 0
```

```
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5

bed1 = Rank(1)=45
bed2 = abundance(41)>100 and abundance(43)>100
bed3 = abundance(55)>100 and abundance(56)>30 and abundance(57)>10

fragment_counter = 0
for n = 3 to 30  step 1
        fragment_alcohol =n*14-1
        if abundance(fragment_alcohol)>0 then
                fragment_counter= fragment_counter +1
        end if
next
```

```
Carbon_number = 4
Do while Carbon_number<=30
        Expected_MW = 14*Carbon_number+18
        m18 = Expected_MW -18
        noise_counter1 = 0
        noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
        noise_counter3 = 0
        for noisecheck2 = 100 to Em step 1
                If abundance(noisecheck2)> 100 then
                        noise_counter3 = noise_counter3 + 1
                end if
        next
If bed1 and bed2 and bed3 and abundance(m18)>5 and fragment_counter >=
(Carbon_number-5) and noise_counter1<=5 and noise_counter2<5 and noise_counter3<1
then Alcohols2 = true
Carbon_number = Carbon_number+1
```

```
Loop
End function
'*************************************************************************
function 1-alcohols() as Boolean
'alcohol  'ion series 55,69,83,97,111,125
'aldehyde 'ion series 68,82,96,110,124,138

Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity
Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
```

```
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed1 = Rank(1)=55 or Rank(1)=43 or Rank(1)=41
bed2 = abundance(55)>700 and abundance(56)>400 and abundance(57)>200
bed3 = abundance(69)>400 and abundance(83)>200

fragment_counter = 0
for n = 3 to 30  step 1
        fragment_alcohol =n*14-1
        fragment_aldehyde =n*14-2
        fragment_next = n*14+1
        if abundance(fragment_alcohol)>abundance(fragment_aldehyde) and
abundance(fragment_alcohol)>abundance(fragment_next) and
abundance(fragment_alcohol)>10 then
                fragment_counter= fragment_counter +1
        end if
next
Carbon_number = 4
Do while Carbon_number<=30
        Expected_MW = 14*Carbon_number+18
        m18 = Expected_MW -18
        m46 = Expected_MW-46
        noise_counter1 = 0
        noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
```

```
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
        noise_counter3 = 0
        for noisecheck2 = 100 to Em step 1
                If abundance(noisecheck2)> 100 then
                        noise_counter3 = noise_counter3 + 1
                end if
        next
If bed1 and bed2 and bed3 and abundance(m18)>0 and abundance(m46)>10 and
fragment_counter >= (Carbon_number-5) and noise_counter1<=5 and noise_counter2<5
and noise_counter3<1 then 1-alcohols = true
Carbon_number = Carbon_number+1
Loop
End function
'*************************************************************************
function linearSC_FFA() as Boolean
'[HOOC(CH2)n]+ from m/z = 115 to 255
'Checks for ion series of general formula [CH3OCO(CH2)n]+
'base peak at m/z=60
'high abundance of m/z73
```

```
'[M-17] loss of OH-
'[M-29] & [M-43]

Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = EndMass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
```

```
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed1 = (Rank(1)=60 and Rank(2)=73)
bed2 = abundance(60)>500 and abundance(61)>10 and abundance(73)>300 and
abundance(55)<500
bed3 = abundance(41)>100 and abundance(42)>50 and abundance(43)>50 and
abundance(45)>100
Carbon_number = 4
Do while Carbon_number < 15
        Expected_MW = 14*(Carbon_number-1)+46
        m17 = Expected_MW -17
        m29 = Expected_MW -29
        m43 = Expected_MW -43
        bed4 = abundance(m17)>0 or abundance(m29)>0 or abundance(m43)>0
        noise_counter1 = 0
        noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
```

```
        if bed1 and bed2 and bed3 and bed4 and noise_counter1<=5 and noise_counter2<5
then linearSC_FFA =true
        Carbon_number = Carbon_number + 1
        Loop
end function
'*************************************************************************
function linearLC_FFA()as Boolean
Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = EndMass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
```

```
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed1 = (Rank(1)=60 or Rank(1)=73)
bed2 = Abundance(73)>700 and abundance(60)>700
bed3 = abundance(55)>400 and abundance(57)>200
bed4 = abundance(83)>50 and abundance(85)>30 and abundance(87)>50 and
abundance(129)>100
bed5 = abundance(41)>500 and abundance(43)>500
bed6 = abundance(97)>10 and abundance(61)>100 and abundance(61)<400 and
abundance(69)>100

'ion series [HOOC(CH2)n]+ from m/z = 115 to 255
fragment_counter = 0
for n = 8 to 30  step 1
        fragment_test =n*14+3
        if abundance(fragment_test)>5 then
                fragment_counter= fragment_counter +1
        end if
next
Carbon_number = 8
Do while Carbon_number < 30
        Expected_MW = 14*(Carbon_number-1)+46
        m29 = Expected_MW -29
        m43 = Expected_MW -43
        noise_counter1 = 0
```

```
        noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
if abundance(Expected_MW)>5 and abundance(m29)>0 and abundance(m43)>0 and bed1
and bed2 and bed3 and bed4 and bed5 and bed6 and noise_counter1<=5 and
noise_counter2<5 and fragment_counter>=Carbon_number-9 then linearLC_FFA =true

Carbon_number = Carbon_number + 1
Loop
end function
‘*************************************************************************
'm/z 43 base peak
'm/z 60 rank(2)
'high abundance(102)
'ion series 45,59,73,87,101,115,129,143,157,171
function isopropylester() as Boolean
Dim Carbon_number
Dim Expected_MW
Dim Em
```

```
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed3 = abundance(102)>300 and abundance(97)>50
bed4 = Rank(1)=43
bed5 = (Rank(2)=60 or Rank(3)=60)
```

```
fragment_counter1 = 0
for n = 3 to 30  step 1
        fragment_test =n*14+3
        if abundance(fragment_test)>15 then
                fragment_counter1= fragment_counter1+1
        end if
next

Carbon_number = 4
Do while Carbon_number < 30
        Expected_MW = 14*Carbon_number+74
        m40 = Expected_MW -40
        m41 = Expected_MW -41
        m42 = Expected_MW -42
        m59 = Expected_MW -59
        bed1 = abundance(m40)>1 and abundance(m41)>30 and abundance(m42)>50
        bed2 = abundance(m59)>30
        noise_counter1 = 0
        noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
```

```
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and fragment_counter1>=(Carbon_number-6) and noise_counter1<=5 and noise_counter2<5 then isopropylester = true
Carbon_number = Carbon_number + 1
Loop
End function
‘*************************************************************************
'm/z 88 base peak instead of m/z 74
‘high abundance m/z 101
‘high abundance of [M-45] & [M-43] (loss of the ethoxide)
‘[M-29] loss of the ethyl group
'ion series 101,115,129,143,157,171,185,199
‘m/z 41,55,60,73
function FA_EthylEsters() as Boolean
Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
```

```
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed1 = Rank(1)=88
bed2 = abundance(41)>200 and abundance(43)>200
bed3 = abundance(55)>100 and abundance(57)>100 and abundance(60)>50 and
abundance(61)>50
bed4 = abundance(73)>100 and abundance(101)>200

fragment_counter = 0
for n = 3 to 30  step 1
        fragment_test =n*14+3
        if abundance(fragment_test)>3 then
                fragment_counter= fragment_counter+1
        end if
next
```

```
Carbon_number = 4

Do while Carbon_number < 30

        Expected_MW = 14*Carbon_number+60

        m27 = Expected_MW -27

        m29 = Expected_MW -29

        m43 = Expected_MW -43

        m45 = Expected_MW -45

        bed5 = abundance(m27)>0 or abundance(m29)>0

        bed6 = abundance(Expected_MW)>0 and abundance(m43)>3 and
abundance(m45)>3

        noise_counter1 = 0

        noise_counter2 = 0

        for noisecheck1 = Expected_MW + 2 to Em step 1

                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then

                        set_intensity = 0

                else

                        set_intensity = intensity(noisecheck1)

                end if

                If set_intensity/intensity(Rank(1))> 0.005 then

                noise_counter1 = noise_counter1 + 1

                else if abundance(noisecheck1)>30 then

                noise_counter2 = noise_counter2 + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and
fragment_counter>=(Carbon_number-6) and noise_counter1<=5 and noise_counter2<5
then FA_EthylEsters = true

Carbon_number = Carbon_number + 1
```

```
Loop
End function
'*************************************************************************
Function LinearSaturatedFAME_v2() as boolean

Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
```

```
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5

bed1 = Rank(1)=74
bed2 = Rank(2)=87 or Rank(3)=87
bed3 = (Abundance(55)>150 And Abundance(55)<400)
bed4 = abundance(57)<350
bed5 = abundance(59)>50

'Checks for ion series of general formula [CH3OCO(CH2)n]+

fragment_counter = 0
for n = 2 to 30  step 1
        fragment_test =n*14+59
        if abundance(fragment_test)>5 then
                fragment_counter= fragment_counter +1
        else
                LinearSaturatedFAME_v2 = false
        end if
next

Carbon_number = 5
Do while Carbon_number < 30
Expected_MW = 14*(Carbon_number-1) + 60
M29 = Expected_MW -29
```

```
M31 = Expected_MW -31

M43 = Expected_MW -43

bed6 = Abundance(Expected_MW)>5

bed7 = Abundance(M29)>0 or Abundance(M31)>0 or Abundance(M43)>0

noise_counter1 = 0

        noise_counter2 = 0

        for noisecheck1 = Expected_MW + 2 to Em step 1

                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then

                        set_intensity = 0

                else

                        set_intensity = intensity(noisecheck1)

                end if

                If set_intensity/intensity(Rank(1))> 0.005 then

                noise_counter1 = noise_counter1 + 1

                else if abundance(noisecheck1)>30 then

                noise_counter2 = noise_counter2 + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and fragment_counter
>= (Carbon_number-10) and noise_counter1<=5 and noise_counter2<5 then
LinearSaturatedFAME_v2 = true

Carbon_number = Carbon_number + 1

Loop

end function


'*************************************************************************

'Linear Monoenoic Fatty Acids Methyl Esters

'1. base peak: rank(1) = 55

'2. m/z69 in high abundance, compare to saturated FAMEs
```

```
'm/z69>300
'3. molecular ion at reasonable abundance
'4. m/z74in lower abundance, compare to saturated FAMEs
'm/z74<600
'5. m/z83 in higher abundance, compare to saturated FAMEs
'm/z83>200
'6. long homologous series (general formula [CnH2n 1]+)  of ions 14 amu apart at m/z = 83, 97, 111, 125, 139, 153, 167
'7. high abundance: [M-32] & [M-74]& [M-116]
'8. Presence of [M-60] & [M-61] and [M-49] & [M-50]
'*************************************************************************
function Linear_monoenoicFAME_v2() as boolean

Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
```

```
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5

bed1 = Rank(1)=55
bed2 = Abundance(69)>300
bed3 = abundance(74)<600
bed4 = abundance(83)>150
bed5 = abundance(74)/abundance(69)>0.8 and abundance(74)/abundance(69)<1.6

'Checks for ion series of general formula [CnH2n-1]+)
'Prominent features and must have fragments

fragment_counter = 0
for n = 6 to 30  step 1
        fragment_test =n*14- 1
        if abundance(fragment_test)>5 then
                fragment_counter= fragment_counter +1
        end if
```

```
next


Carbon_number = 5
Do while Carbon_number < 30
Expected_MW = 14*(Carbon_number-1) + 58
M32 = Expected_MW -32
M74 = Expected_MW -74
M116 = Expected_MW -116
M49 = Expected_MW -49
M50 = Expected_MW -50
M60 = Expected_MW -60
M61 = Expected_MW -61


bed6 = Abundance(M32)>10 and Abundance(M74)>10 and Abundance(M116)>10
bed7 =  Abundance(M49)>0 and Abundance(M50)>0
bed8 = Abundance(M60)>0 and Abundance(M61)>0
noise_counter1 = 0
noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
```

```
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
fragment_counter >= (Carbon_number-8) and noise_counter1<=5 and noise_counter2<5
then Linear_monoenoicFAME_v2= true
Carbon_number = Carbon_number + 1
Loop
End function
'*************************************************************************
function Linear_dienoicFAME_v2() as boolean

Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity
Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
```

```
avg_intensity = sum_intensity/noisecounter1

sum_sq = 0

for masscheck2 = 500 to Em step 1

        if intensity(masscheck2)>0 then

                sq = (intensity(masscheck2)-avg_intensity)^2

                sum_sq = sum_sq + sq

        end if

next

stdev_intensity =(sum_sq/(noisecounter1-1))^0.5


bed1 = Rank(1)=67

bed2 = Abundance(74)<250

bed3 = Abundance(81)>500 and abundance(81)>abundance(79)

bed4 = abundance(68)<500 and abundance(82)<500 and abundance(73)<200 and
abundance(75)<200

bed5 = (abundance(95)>abundance(96)) and (abundance(109)>abundance(110))

'Checks for Ion Series CnH2n-3

'm/z 67,81,95,109,123,

fragment_counter = 0

for n = 5 to 30 step 1

        fragment_test1 = n*14-3

        if abundance(fragment_test1)>0 then

                fragment_counter =  fragment_counter +1

        end if

next

'prominent features and must have fragments

'base peak 67 or 81

Carbon_number = 5

Do while Carbon_number < 30
```

```
Expected_MW = 14*(Carbon_number-1) + 56
M31 = Expected_MW -31
M32 = Expected_MW -32
bed6 = abundance(M31)>0 and Abundance(M32)>0
noise_counter1 = 0
noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and fragment_counter >=
(Carbon_number-6) and abundance(Expected_MW)>0 and noise_counter1<=5 and
noise_counter2<5 then Linear_dienoicFAME_v2= true
Carbon_number = Carbon_number + 1
Loop
end function
'*************************************************************************
function Linear_trienoicFAME_v2() as Boolean
Dim Carbon_number
Dim Expected_MW
Dim Em
```

```
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed1 = Rank(1)= 79
bed2 = Abundance(74)<250 and abundance(59)>100
ratio91=abundance(91)/abundance(93)
```

```
bed3 = ratio91<1.3

bed4 = (Rank(2)=41 or Rank(2)=55 or Rank(2) =67)

bed5 = abundance(55)>400 and abundance(67)>500

bed6 = abundance(95)>100 and abundance(108)>30 and abundance(121)>50 and
abundance(135)>30

'ion series m/z 65,79,93,107,121,135,149,163,177

fragment_counter = 0

for n = 5 to 30 step 1

        fragment_test = n*14-5

        if abundance(fragment_test)>0 then

                fragment_counter =  fragment_counter +1

        end if

next

Carbon_number = 5

Do while Carbon_number < 30

Expected_MW = 14*(Carbon_number-1) + 54

'M29 = Expected_MW -29

M31 = Expected_MW -31

bed7 = Abundance(M31)>0

noise_counter1 = 0

noise_counter2 = 0

        for noisecheck1 = Expected_MW + 2 to Em step 1

                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then

                        set_intensity = 0

                else

                        set_intensity = intensity(noisecheck1)

                end if

                If set_intensity/intensity(Rank(1))> 0.005 then

                noise_counter1 = noise_counter1 + 1
```

```
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and fragment_counter >= (Carbon_number-6) and abundance(Expected_MW)>0 and noise_counter1<=5 and noise_counter2<5 then Linear_trienoicFAME_v2= true
Carbon_number = Carbon_number + 1
Loop
end function

'*************************************************************************
function Linear_multienoicFAME_v2() as Boolean
Dim Carbon_number
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
```

```
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed1 = Rank(1)= 79
bed2 = (rank(2)=91 or rank(3)=91)
bed3 = abundance(105)>200 and abundance(105)<400
bed4 = abundance(67)>500 and abundance(67)<900
bed5 = abundance(87)<100 and abundance(55)>200
bed6 = (abundance(105)/abundance(108))>2.5
bed7 = abundance(74)>50 and abundance(77)>100 and abundance(78)>100 and
abundance(80)>100 and abundance(81)>100
'ion series m/z 75,89,103,117,131,145,159,173
fragment_counter = 0
for n = 5 to 30 step 1
        fragment_test = n*14+5
        if abundance(fragment_test)>0 then
                fragment_counter =  fragment_counter +1
        end if
next
```

```
Carbon_number = 5
Do while Carbon_number < 30
        Expected_MW = 14*Carbon_number+60
m27 = Expected_MW -27
m29 = Expected_MW -29
        m43 = Expected_MW -43
        m45 = Expected_MW -45
bed5 = abundance(m27)>0 or abundance(m29)>0
bed6 = abundance(Expected_MW)>0 and abundance(m43)>3 and abundance(m45)>3
noise_counter1 = 0
        noise_counter2 = 0
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and fragment_counter>5
then Linear_multienoicFAME_v2= true
end function
'*************************************************************************
TMS derivatized
'*************************************************************************
```

```
function Valine_2TMS_v2() as Boolean

'molecular weight 261

mass = 261

M15 = mass-15

M43 = mass-43

bed1 = (Rank(1)=73 and Rank(2)=144)

bed2 = abundance(M43)>100 and abundance(M15)>0

bed3 = abundance(73)>500 and abundance(74)>10 and abundance(75)>10

bed4 = abundance(144)>400 and abundance(145)>10 and abundance(147)>10

bed5 = abundance(45)>0 and abundance(45)<400

bed6 = abundance(59)>0 and abundance(59)<250

bed7 = abundance(85)>0 and abundance(100)>10 and abundance(128)>0

bed8 = abundance(218)>30 and abundance(219)>0

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 then
Valine_2TMS_v2 = true

end function

‘*************************************************************************

function proline_2TMS() as boolean

Dim MW

Dim M15

Dim M29

Dim M43


MW=259

M15 = MW-15

M29 = MW-29

M43 = MW -43

bed1 = Rank(1)=73

bed2 = Rank(2)=142
```

bed3 = abundance(MW)>0 and abundance(M15)>0 and abundance(M29)>0 and abundance(M43)>30

bed4 = abundance(217)>0 and abundance(218)>0

bed5 = abundance(84)>5 and abundance(100)>10 and abundance(147)>10 and abundance(170)>0

bed6 = abundance(45)>100 and abundance(59)>10 and abundance(66)>10 and abundance(75)>10

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 then proline_2TMS = true

end function

‘**********

function oxoproline_2TMS() as Boolean

‘MW = 273

bed1 = Rank(1)=73 or Rank(2)= 73

bed2 = Rank(1)=156 or Rank(2)=156

bed3 = abundance(273)>0 and abundance(258)>10 and abundance(230)>10

bed4 = abundance(147)>50 and abundance(133)>5 and abundance(156)>500

bed5 = abundance(45)>100 and abundance(59)>10 and abundance(84)>3 and abundance(86)>3 and abundance(100)>3

If bed1 and bed2 and bed3 and bed4 and bed5 then oxoproline_2TMS = true

end function

‘**********

function tyrosine() as boolean

bed1 = Rank(1)=73

bed2 = Rank(2)=218

bed3 = abundance(218)>400 and abundance(219)>50 and abundance(220)>10 and abundance(179)>30

bed4 = abundance(100)>100 and abundance(147)>50

```
bed5 = abundance(45)>100 and abundance(59)>10

If bed1 and bed2 and bed3 and bed4 and bed5 then tyrosine_3TMS = true


bed6 = Rank(1)=73

bed7 = Rank(2)=179

bed8 = abundance(179)>500 and abundance(180)>100 and abundance(181)>30 and
abundance(182)>0

bed9 = abundance(208)>50 and abundance(219)>30

If bed6 and bed7 and bed8 and bed9 and bed5 then tyrosine_2TMS = true


If tyrosine_3TMS or tyrosine_2TMS then tyrosine = true

end function

'*************************************************************************

Function Lysine_4TMS() as boolean

Dim Em

Em = Endmass()

'TMS-NH=CH-COOTMS m/z 218

bed1 = abundance(434)>0 and abundance(435)>0 'M & [M+1]

bed2 = abundance(329)>0         'M-105

bed3 = abundance(59)>100 and abundance(86)>50 and abundance(100)>100 and
abundance(115)>0 and abundance(128)>50 and abundance(147)>0

bed4 = (Rank(1)=73)

bed5 = (Rank(2)=156 or Rank(2)=174)

bed6 = (Rank(3)=156 or Rank(3)=174)

bed7 = abundance(186)>0 and abundance(200)>0 and abundance(218)>0 and
abundance(230)>10 and abundance(273)>0 and abundance(317)>30

noise_counter = 0

for noisecheck = 435 to Em step 1

        If abundance(noisecheck)>10 then
```

```
                noise_counter = noise_counter + 1
        end if
next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and noise_counter<5
then Lysine_4TMS=true
End function
‘*************************************************************************
‘M+ = 276
Function Asparagine_2TMS() as Boolean
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity
Expected_MW = 276
Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
```

```
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5


bed1 = Rank(1)=44
bed2 = (Rank(2)=73 or Rank(2)=75 or Rank(2)=159)
bed3 = abundance(73)>200 and abundance(75)>200
bed4 = abundance(86)>5 and abundance(100>10 and abundance(116)>50
bed5 = abundance(130)>20 and abundance(147)>20 and abundance(159)>100 and
abundance(186)>3 and abundance(244)>3
        for noisecheck1 = Expected_MW + 2 to Em step 1
                If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then
                        set_intensity = 0
                else
                        set_intensity = intensity(noisecheck1)
                end if
                If set_intensity/intensity(Rank(1))> 0.005 then
                noise_counter1 = noise_counter1 + 1
                else if abundance(noisecheck1)>30 then
                noise_counter2 = noise_counter2 + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and noise_counter1<=5 and
noise_counter2<5 then Asparagine_2TMS = true
End function
```

```
'*************************************************************************
Function glycine_2TMS
Dim MW_MonoTMS_glycine
Dim MW_TriTMS_glycine
Dim Em
Em = Endmass()
bed1 = (Rank(1)=73 or Rank(2)=73)
bed2 = (Rank(1)=102 or Rank(2)=102)
bed2 = abundance(45)>100 and abundance(59)>100 and abundance(86)>100 and
abundance(100)>100 and abundance(133)>50 and abundance(159)>0
bed3 = abundance(174)>400
bed4 = abundance(117)>0 and abundance(147)>200
bed5 = abundance(276)>10 'M-15
bed6 = abundance(248)>50 'M-43

'not significant picks from m/z179 to m/z 245
significant_counter1 = 0
for sig_ion_check1 = 179 to 245 step 1
        If abundance(sig_ion_check1)>10 then
                significant_counter1 = significant_counter1 + 1
        end if
next

significant_counter2 = 0
for sig_ion_check2 = 252 to 275 step 1
        If abundance(sig_ion_check2)>5 then
                significant_counter2 = significant_counter2 + 1
        end if
next
```

```
noise_counter = 0
for noisecheck = MW_TriTMS_glycine+1 to Em step 1
        If abundance(noisecheck)>10 then
                noise_counter = noise_counter + 1
        end if
next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and significant_counter1<3 and
significant_counter2<2 and noise_counter<5 then TriTMS_glycine=true
End Function
‘*************************************************************************
function MonoenoicFA_TMS() as boolean
Dim Carbon_number
Dim Expected_MW
'ion series m/z
bed1 = (Rank(1)=75 or rank(2)=75)
bed2 = Abundance(73)>500
bed3 = abundance(117)>300 and abundance(129)>200 and abundance(145)>50
bed4 = abundance(41)>300 and abundance(55)>300
bed5 = abundance(74)<150 and abundance(91)<100
bed6 = abundance(81)>100 and abundance(96)>100 and abundance(110)>20

'from m/z145 to M-15
'ion series m/z 157,171,185,199,213,227,241,255,269,283,297 etc
fragment_counter1 = 0
for n = 11 to 30  step 1
        fragment_test =n*14+ 3
        if abundance(fragment_test)>1 then
```

```
                fragment_counter1= fragment_counter1 +1
        end if
next

for Carbon_number = 5 to 30 step 1
Expected_MW = 14*(Carbon_number) + 102
M15 = Expected_MW -15
M16 = Expected_MW -16
bed7 = abundance(Expected_MW)>0 and Abundance(M15)>30
fragment_counter2 = 0
        for k = 146 to M16
                if abundance(k)>50 then
                        fragment_counter2= fragment_counter2 +1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and fragment_counter1
>= (Carbon_number-10) and fragment_counter2<3 then MonoenoicFA_TMS = true
next
end function

'*************************************************************************

function DienoicFA_TMS() as boolean
Dim Carbon_number
Dim Expected_MW
for Carbon_number = 5 to 30 step 1
Expected_MW = 14*(Carbon_number) + 100
M15 = Expected_MW -15
M16 = Expected_MW -16
```

```
bed1 = (abundance(M15)/abundance(Expected_MW))>3

bed2 = (Rank(1)=73 or Rank(1)=75 or Rank(2)=73 or Rank(2)=75)

bed3 = abundance(117)>100 and abundance(129)>100 and abundance(132)<100

bed4 = abundance(41)>200 and abundance(55)>200

bed5 = abundance(67)>200

bed6 = abundance(73)>500 and abundance(75)>500 and abundance(79)<500

bed7 = Abundance(M15)>20 and Abundance(Expected_MW)>0

bed8 = abundance(150)>3 and abundance(164)>3

counter1 = 0

        for k=130 to M16 step 1

                if abundance(k)>100 then

                        counter1 = counter1+1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
counter1<3 then DienoicFA_TMS = true

next

end function


'*************************************************************************

'multidouble bonds in aliphatic chain

'high abundance of m/z 79


function multienoicFA_TMS() as boolean

Dim Em

Em = Endmass()

bed1 = (Rank(1)=73 or Rank(1)=75 or rank(1)=79)

bed2 = abundance(79)>500

bed3 = abundance(108)>10 and abundance(129)>50 and abundance(135)>10
```

```
bed4 = abundance(41)>300 and abundance(55)>200 and abundance(67)>400

bed5 = abundance(91)>150 and abundance(93)>100 and abundance(95)>50

bed6 = abundance(73)>500 and abundance(75)>500

bed7 = abundance(105)>50 and abundance(80)>100 and abundance(81)>30

bed8 = abundance(107)>10 and abundance(108)>10

counter1 = 0

        for k=160 to Em step 1

                if abundance(k)>100 then

                        counter1 = counter1+1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
counter1<3 then multienoicFA_TMS = true

end function

'*************************************************************************

function SatFA_TMS() as boolean

Dim Carbon_number

Dim Expected_MW

Dim Em

Em = Endmass()

for Carbon_number = 5 to 30 step 1

Expected_MW = 14*(Carbon_number) + 104

M16 = Expected_MW -16

M15 = Expected_MW -15

M14 = Expected_MW -14

M13 = Expected_MW -13

bed1 = abundance(M15)>50 and abundance(M14)>0 and abundance(M13)>0

bed2 = (Rank(1)=73 or Rank(1)=75 or Rank(1)=117)
```

```
bed3 = abundance(129)>50 and abundance(130)>0 and abundance(131)>50 and
abundance(132)>50

bed4 = abundance(41)>50 and abundance(45)>50 and abundance(55)>50

bed5 = abundance(145)>20

bed6 = abundance(73)>500 and abundance(75)>500 and abundance(117)>200

bed7 = abundance(159)>3

counter1 = 0
        for n=76 to 116 step 1
                if abundance(n)>90 then
                        counter1 = counter1+1
                end if
        next
counter2 = 0
        for k=146 to M16 step 1
                if abundance(k)>50 then
                        counter2 = counter2+1
                end if
        next
noise_counter = 0
for noisecheck = Expected_MW+1 to Em step 1
        If abundance(noisecheck)>15 then
                noise_counter = noise_counter + 1
        end if
next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and counter1<3 and
counter2<3 and noise_counter<5 then SatFA_TMS = true
next
end function
'*************************************************************************
```

```
Function sugars_TMS()as Boolean

bed1 = rank(1) = 73

bed2 = abundance(59)>0

bed3 = abundance(147)>0 and abundance(160)>0

bed4 = abundance(103)>50 and abundance(117)>0 and abundance(129)>0 and
abundance(133)>0

bed5 = abundance(204)>0 and abundance(205)>50 and abundance(217)>50 and
abundance(229)>0

bed6 = abundance(319)>100 and abundance(320)>0

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 then sugars_TMS = true

End Function

'*************************************************************************

Function sugar_4TMS() as boolean

Dim Em

Em = Endmass()

bed1 = (Rank(1)=73 and Rank(2)=103)

bed2 = abundance(45)>50 and abundance(59)>30

bed3 = abundance(89)>10 and abundance(103)>300 and abundance(117)>10 and
abundance(133)>10 and abundance(147)>50

bed4 = abundance(160)>10 and abundance(172)>0 and abundance(189)>10

bed5 = abundance(204)>0 and abundance(205)>0 and abundance(217)>50 and
abundance(233)>0

bed6 = abundance(262)>0 and abundance(277)>0 and abundance(307)>10

bed7 = abundance(74)>0 and abundance(75)>0


noise_counter = 0

for noisecheck = 401 to Em step 1

        If abundance(noisecheck)>10 then

                noise_counter = noise_counter + 1

        end if
```

```
next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and noise_counter<5
then sugar_4TMS=true

End function


'*************************************************************************

Function sugar_8TMS() as boolean

Dim Em

Em = Endmass()

bed1 = Rank(1)=73

bed2 = abundance(45)>30 and abundance(59)>10

bed3 = abundance(81)>5 and abundance(89)>5 and abundance(103)>50

bed4 = abundance(117)>20 and abundance(129)>50 and abundance(131)>5 and
abundance(133)>10 and abundance(147)>100

bed5 = abundance(157)>10 and abundance(169)>20 and abundance(191)>10

bed6 = abundance(204)>5 and abundance(205)>0 and abundance(217)>50 and
abundance(231)>0 and abundance(243)>10

bed7 = abundance(259)>0 and abundance(271)>10 and abundance(319)>5 and
abundance(361)>30

noise_counter = 0

for noisecheck = 401 to Em step 1

        If abundance(noisecheck)>10 then

                noise_counter = noise_counter + 1

        end if

next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and noise_counter<5
then sugar_8TMS=true

End function


'*************************************************************************
```

```
Function sugar_5TMS()as Boolean

Dim Em

Em = Endmass()

bed1 = rank(1) = 73

bed2 = abundance(45)>30 and abundance(59)>10 and abundance(89)>30

bed3 = abundance(103)>50 and abundance(117)>30 and abundance(129)>20 and
abundance(131)>5 and abundance(133)>20

bed4 = abundance(147)>100 and abundance(189)>10

bed5 = abundance(204)>0 and abundance(205)>10 and abundance(217)>10 and
abundance(229)>0

bed6 = abundance(307)>30

bed7 = abundance(319)>50 and abundance(320)>5 and abundance(320)>1

bed8 = bed6 or bed7

noise_counter = 0

for noisecheck = 465 to Em step 1

        If abundance(noisecheck)>10 then

                noise_counter = noise_counter + 1

        end if

next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed8 and noise_counter<5 then
sugar_5TMS=true

End function

‘*************************************************************************

function sterol_TMS() as Boolean

'MW=458

'm/z 41,43,55,57,,73,75

'm/z 81,95,105,119,129,145,159,173,185,199,213,255,275,329,354,368,443,458

'high abundance of m/z 73,129

Dim Carbon_number
```

```
Dim Expected_MW
Dim Em
Dim masscheck1
Dim masscheck2
Dim avg_intensity
Dim sum_intensity
Dim sum_sq
Dim set_intensity

Em = Endmass()
sum_intensity = 0
noisecounter1 = 0
for masscheck1 = 500 to Em step 1 'get average noise from m/z 500 to Endnum
        if intensity(masscheck1)>0 then
                sum_intensity = sum_intensity + intensity(masscheck1)
                noisecounter1 = noisecounter1 + 1
        end if
next
avg_intensity = sum_intensity/noisecounter1
sum_sq = 0
for masscheck2 = 500 to Em step 1
        if intensity(masscheck2)>0 then
                sq = (intensity(masscheck2)-avg_intensity)^2
                sum_sq = sum_sq + sq
        end if
next
stdev_intensity =(sum_sq/(noisecounter1-1))^0.5
bed2 = abundance(41)>200 and abundance(43)>300 and abundance(55)>100 and
abundance(57)>50
```

bed3 = abundance(73)>300 or abundance(75)>300

bed4 = abundance(91)>100 and abundance(105)>50 and abundance(119)>10

bed5 = abundance(129)>50 and abundance(145)>10 and abundance(159)>10 and abundance(173)>10

bed6 = abundance(215)>0 and abundance(233)>0

bed7 = abundance(247)>3 and abundance(255)>10

MW=456

Do while MW<=500

noise_counter1 = 0

noise_counter2 = 0

bed1 = abundance(MW)>0 and abundance(MW-15)>0 and abundance(MW-90)>10 and abundance(105)>5

for noisecheck1 = MW + 2 to Em step 1

If Intensity(noisecheck1) <= (ave_intensity + 4*stdev_intensity) then

set_intensity = 0

else

set_intensity = intensity(noisecheck1)

end if

If set_intensity/intensity(Rank(1))> 0.005 then

noise_counter1 = noise_counter1 + 1

else if abundance(noisecheck1)>30 then

noise_counter2 = noise_counter2 + 1

end if

next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and noise_counter1<=5 and noise_counter2<5 then sterol_TMS=true

MW = MW+2

Loop

```
End function
‘*************************************************************************
'reference: Characteristics of the GC-MS Mass Spectra of Terpenoids (C10H16)
'base peak 93
'high abundance peak m/z 136
'abundance m/z137>0
'm/z 77,79, 91,93, 105,107,119,121,136

Function TerpenoidsC10H16()as Boolean
dim mass                        ' New variable mass
dim threshold           ' New variable threshold (not used actual)
dim ip as parent_return ' the function ip belongs to the parentreturn substructure this is to get the mass of the molecular ion
ip = first_parent()         ' ip is the function firstparent
mass = Ip.mass              ' the variable mass is the mass from the molecular ion
threshold=Ip.noise          ' the variable threshold is the noise from firstparent
Dim Em
Em= Endmass()
bed1 = rank(1) = 93 or rank(1) = 121
bed2 = abundance(91)>100 and abundance(93)>500
bed3 = abundance(119)>30 and abundance(121)>100
bed4 = abundance(136)>50 and abundance(137)>0
bed5 = abundance(77)>100 and abundance(79)>100
bed6 = abundance(105)>30 and abundance(107)>30
bed7 = abundance(mass)>50
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 then TerpenoidsC10H16 = true
End Function
```

```
'*************************************************************************

'high abundance peak m/z 136

'abundance m/z137>0

'm/z 77,79, 91,93, 105,107,119,121,136

'm/z 39 (C3H3+) 51 (C4H3+) 65(C5H5+), 77 (C6H5+), 91 (C7H7+)


Function bicyclicC10H16()as Boolean

dim mass                        ' New variable mass

dim threshold           ' New variable threshold (not used actual)

dim ip as parent_return ' the function ip belongs to the parentreturn substructure this is to get the mass of the molecular ion

ip = first_parent()         ' ip is the function firstparent

mass = Ip.mass              ' the variable mass is the mass from the molecular ion

threshold=Ip.noise          ' the variable threshold is the noise from firstparent

Dim Em

Em= Endmass()

bed1 = rank(1) = 93

bed2 = abundance(91)>100 and abundance(93)>500

bed3 = abundance(121)>10

bed4 = abundance(136)>50

bed5 = abundance(77)>100 and abundance(79)>100

bed6 = abundance(105)>30 and abundance(107)>30

bed7 = abundance(mass)>50

bed8 = abundance(51)>10 and abundance(65)>100

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 then bicyclicC10H16= true

End Function


'*************************************************************************
```

```
'loss of [M-15]+ and [M-18]+

Function Terpenol()as Boolean

dim mass                        ' New variable mass

dim threshold           ' New variable threshold (not used actual)

dim ip as parent_return ' the function ip belongs to the parentreturn substructure this is to get the mass of the molecular ion

ip = first_parent()         ' ip is the function firstparent

mass = Ip.mass              ' the variable mass is the mass from the molecular ion

threshold=Ip.noise          ' the variable threshold is the noise from firstparent


Dim M15

Dim M18

M15 = mass-15 'loss of CH3

M18 = mass-18 'loss of H2O

M17 = mass-17 'loss of OH

bed1 = abundance(41)>100 and abundance(43)>100

bed2 = abundance(91)>10 and abundance(93)>50

bed3 = abundance(121)>10

bed4 = abundance(136)>10

bed5 = abundance(77)>10 and abundance(79)>10

bed6 = abundance(105)>0 and abundance(107)>0

bed7 = abundance(mass)>0 and abundance(M15)>3 and abundance(M18) >3 and abundance(M17)>0

bed8 = abundance(51)>10 and abundance(65)>10

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 then Terpenol = true

End Function


'*************************************************************************
```

```
'loss of [M-15]+ and [M-18]+

Function C10H18O()as Boolean

Dim Em

Dim Expected_mass

carbon_number = 10

hydrogen_number = 18

oxygen_number = 1

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3

M18 = Expected_mass -18 'loss of H2O

M1 = Expected_mass -1 'irrational fragment

M43 = Expected_mass -43 'loss of CH(CH3)2

bed1 = abundance(41)>100 and abundance(43)>50 and abundance(53)>30 and
abundance(55)>50

bed2 = abundance(91)>10 and abundance(93)>50 and abundance(95)>0

bed3 = abundance(107)>5 or abundance(111)>5

bed4 = abundance(121)>5 and abundance(136)>3

bed5 = abundance(77)>10 and abundance(79)>10 and abundance(67)>50 and
abundance(69)>10

bed6 = abundance(105)>0 or abundance(106)>0

bed7 = abundance(M15)>3 and abundance(M18) >3 and abundance(M1)<3 and
abundance(M43)>0

bed8 = abundance(51)>10 and abundance(65)>10 and abundance(154)>=abundance(155)

bed9 = abundance(205)<2 and abundance(220)<2 and abundance(158)<20

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 5 then

                        noise_counter = noise_counter + 1
```

```
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and noise_counter<3 then C10H18O = true
End Function

'*************************************************************************
Function C10H16_Terpenoids()as Boolean
Dim Em
Dim Expected_mass
carbon_number = 10
hydrogen_number = 16
oxygen_number = 0
Em=Endmass()
Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number
M13 = Expected_mass-13
M3 = Expected_mass-3

bed1 = rank(1) = 93 or rank(2) = 93 or rank(3) = 93 or rank(4) = 93
bed2 = abundance(91)>50 and abundance(93)>400
bed3 = abundance(121)>10
bed4 = abundance(67)>10 or abundance(68)>10
bed5 = abundance(77)>100 and abundance(79)>100
bed6 = abundance(105)>10 and abundance(107)>10
bed7 = abundance(Expected_mass)>10
bed8 = abundance(41)>100 and abundance(51)>10 and abundance(65)>30
bed9 = abundance(147)<80 and abundance(154)<5 and abundance(161)<20 and abundance(189)<30 and abundance(204)<40
```

```
bed10 = abundance(135)<10 and (abundance(137)/abundance(136))<0.25 and
abundance(139)<20 and abundance(137)<100

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 50 then

                        noise_counter = noise_counter + 1

                end if

        next

irrational_counter = 0

        for irrational_fragment = M13 to M3 step 1

                If abundance(irrational_fragment)> 30 then

                        irrational_counter = irrational_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and noise_counter<=5 and irrational_counter<=1 then C10H16_Terpenoids = true

End Function


'*************************************************************************

'm/z 147,148,161,134,133,136 (from Japanese paper)

'm/z 175,189,204 (MW)

'high abundance m/z 41,53,55,67,79,77,79,

Function sesquiterpenes()as Boolean

Dim Em

Em=Endmass()

carbon_number = 15

hydrogen_number = 24

oxygen_number = 0

Em=Endmass()
```

```
Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3

M43 = Expected_mass -43 'loss of CH(CH3)2

M1 = Expected_mass +1

M18 = Expected_mass +18

bed1 = abundance(Expected_mass)>3 and abundance(M15)>3

bed2 = abundance(M43)>3 or abundance(162)>3 or (abundance(175)>0 and
abundance(176)>0)

bed3 = abundance(43)>50 and abundance(53)>50 and abundance(55)>50

bed4 = abundance(67)>50 and abundance(69)>10

bed5 = abundance(77)>0 and abundance(79)>50 and abundance(81)>10

bed6 = abundance(91)>50 and abundance(93)>100

bed7 = abundance(107)>10 and abundance(109)>0

bed8 = abundance(119)>10

bed9 = abundance(136)>0 or abundance(134)>0

bed10 = abundance(147)>10 or abundance(148)>5

bed11 = abundance(M18)<6 and abundance(207)<1 and abundance(272)<5

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 10 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and bed11 and noise_counter<5 then sesquiterpenes = true

End Function

'**************************************************************************

Function sesquiterpenols()as Boolean

Dim Em
```

```
Em=Endmass()

carbon_number = 15

hydrogen_number = 26

oxygen_number = 1

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3

M18 = Expected_mass -18 'loss of H2O

M33 = Expected_mass -33 'loss of CH3OH

bed0 = abundance(Expected_mass)>0 or abundance(M15)>0

bed1 = abundance(M18)>0 and abundance(M33)>5

bed2 = abundance(41)>100 and abundance(43)>50

bed3 = abundance(53)>50 and abundance(55)>50

bed4 = abundance(67)>50 and abundance(69)>10

bed5 = abundance(77)>0 and abundance(79)>50 and abundance(81)>10 and
abundance(80)>1

bed6 = abundance(91)>50 and abundance(93)>100

bed7 = abundance(105)>0 and abundance(107)>10 and abundance(109)>0

bed8 = abundance(119)>10 and abundance(121)>10

bed9 = abundance(133)>10 and abundance(136)>0 and abundance(134)>0

bed10 = abundance(147)>5 and abundance(148)>5

bed11 = abundance(161)> 10 and abundance(175)>0 and
abundance(222)>=abundance(220)

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 10 then

                        noise_counter = noise_counter + 1

                end if

        next
```

```
If bed0 and bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
bed9 and bed10 and bed11 and noise_counter<5 then sesquiterpenols = true

End Function

'*************************************************************************

Function C10H14 ()as boolean

Dim Em

Em = Endmass()

carbon_number = 10

hydrogen_number = 14

oxygen_number = 0

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M1 = Expected_mass +1

bed1 = rank(1)=119 or rank(2)=119

bed2 = abundance(41)>30 and abundance(51)>20 and abundance(58)>0 and
abundance(63)>10 and abundance(65)>20

bed3 = abundance(77)>40 and abundance(91)>100

bed4 = abundance(103)>10

bed5 = abundance(115)>10 and abundance(117)>10

bed6 = abundance(Expected_mass)>50 and abundance(M1)>0 'Molecular ion and M+1

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 30 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and noise_counter<5 Then
C10H14= True

End Function
```

```
'*************************************************************************
'MW272
Function diterpeneC20H32()as boolean
Dim Em
Em = Endmass()
carbon_number = 20
hydrogen_number = 32
oxygen_number = 0
Em=Endmass()
Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number
M15 = Expected_mass -15
M43 = Expected_mass -43
bed1 = rank(1)=69 or rank(2)=69 or rank(3)=69
bed2 = abundance(41)>100 and abundance(53)>30 and abundance(55)>50 and
abundance(65)>10 and abundance(67)>50
bed3 = abundance(77)>50 and abundance(79)>50 and abundance(81)>50
bed4 = abundance(91)>50 and abundance(93)>50 and abundance(105)>30 and
abundance(107)>30 and abundance(109)>5
bed5 = abundance(119)>10 and abundance(121)>5 and abundance(133)>3 and
abundance(147)>3 and  abundance(161)>3 and abundance(187)>2 and abundance(203)>0
and abundance(255)<2
bed6 = abundance(M43)>0
noise_counter = 0
        for noisecheck = Expected_mass+2 to Em step 1
                If abundance(noisecheck)> 30 then
                        noise_counter = noise_counter + 1
                end if
        next
```

```
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and noise_counter<5 Then
diterpeneC20H32= True

End Function

'*************************************************************************

'MW202

Function Sesquiterpenes_C15H22()as boolean

Dim Em

Em = Endmass()

carbon_number = 15

hydrogen_number = 22

oxygen_number = 0

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number


M1 = Expected_mass +1

M15 = Expected_mass -15

M29 = Expected_mass -29

M43 = Expected_mass -43

bed1 = abundance(41)>100 and abundance(43)>0 and abundance(51)>10 and
abundance(53)>20 and abundance(55)>50

bed2 = abundance(65)>10 and abundance(67)>10

bed3 = abundance(77)>50 and abundance(79)>30 and abundance(81)>3

bed4 = abundance(91)>50 and abundance(93)>30 and abundance(105)>30 and
abundance(107)>0

bed5 = abundance(115)>20 and abundance(117)>30 and abundance(119)>10 and
abundance(121)>3

bed6 = abundance(133)>3 and abundance(145)>10 and  abundance(159)>0

bed7 = abundance(Expected_mass)>50 and abundance(M1)>5 and abundance(M43)>0

bed8 = abundance(205)<10 and abundance(220)<10

noise_counter = 0
```

```
        for noisecheck = Expected_mass+2 to Em step 1
                If abundance(noisecheck)> 10 then
                        noise_counter = noise_counter + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
noise_counter<5 Then Sesquiterpenes_C15H22= True
End Function

'*************************************************************************
'loss of [M-15]+ and [M-18]+
'MW 152
Function monoterpenealdehydes_C10H16O()as Boolean
Dim Em
Dim Expected_mass
carbon_number = 10
hydrogen_number = 16
oxygen_number = 1
Em=Endmass()
Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number
M15 = Expected_mass -15 'loss of CH3
M18 = Expected_mass -18 'loss of H2O
M1 = Expected_mass -1 'irrational fragment
M43 = Expected_mass -43 'loss of CH(CH3)2
bed1 = rank(1) = 69 or rank(2) = 69 or rank(3) = 69
bed2 = rank(1) = 41 or rank(2) = 41 or rank(3) = 41 or rank(4) = 41 or rank(5) = 41
bed3 = abundance(41)>100 and abundance(43)>20 and abundance(53)>30 and
abundance(55)>30
```

```
bed4 = abundance(84)>50 and abundance(91)>10 and abundance(93)>10 and
abundance(94)>0

bed5 = abundance(109)>10 and abundance(123)>10

bed6 = abundance(77)>10 and abundance(79)>10 and abundance(67)>50 and
abundance(69)>50

bed7 = abundance(M15)>3 and abundance(M18) >3 and abundance(M1)<3 and
abundance(M43)>3

bed8 = abundance(51)>10 and abundance(65)>10 and abundance(152)>=abundance(154)

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 5 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
noise_counter<3 then monoterpenealdehydes_C10H16O = true

End Function


'*************************************************************************

'm/z 147,162,177,187,205

'm/z 220 (MW)

'high abundance m/z 41,53,55,67,79,77,79,

Function terpeneoxide_C15H24O() as Boolean

Dim Em

Em=Endmass()

carbon_number = 15

hydrogen_number = 24

oxygen_number = 1

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number
```

```
M15 = Expected_mass -15 'loss of CH3

M18 = Expected_mass -18 'loss of CH(CH3)2

M43 = Expected_mass -43 'loss of CH(CH3)2

bed0 = abundance(Expected_mass)>0 or abundance(M18)>0

bed1 = abundance(M15)>0 and abundance(M43)>0

bed2 = abundance(41)>300 and abundance(43)>50

bed3 = abundance(53)>50 and abundance(55)>50

bed4 = abundance(67)>50 and abundance(69)>10

bed5 = abundance(77)>0 and abundance(79)>50 and abundance(81)>10 and
abundance(80)>10

bed6 = abundance(91)>50 and abundance(93)>100

bed7 = abundance(105)>0 and abundance(107)>10 and abundance(109)>0

bed8 = abundance(119)>10 and abundance(121)>10

bed9 = abundance(136)>0 and abundance(134)>0 and abundance(250)<30

bed10 = abundance(147)>10 and abundance(149)>3 and abundance(177)>0 and
abundance(220)>abundance(219) and abundance(220)>abundance(222)

bed11 = abundance(186)>0 or abundance(187)>0

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 20 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed0 and bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
bed9 and bed10 and bed11 and noise_counter<4 then terpeneoxide_C15H24O = true

End Function


'*************************************************************************

'loss of [M-15]+ and [M-18]+
```

```
'MW 150

Function C10H14O()as Boolean

Dim Em

Dim Expected_mass

carbon_number = 10

hydrogen_number = 14

oxygen_number = 1

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3

M18 = Expected_mass -18 'loss of H2O

M1 = Expected_mass -1 'irrational fragment

M43 = Expected_mass -43 'loss of CH(CH3)2

bed1 = abundance(43)>100 and abundance(51)>10 and abundance(53)>5

bed2 = abundance(65)>10 and abundance(67)>5 and abundance(77)>5 and
abundance(79)>5

bed3 = abundance(91)>10 and abundance(105)>5 and abundance(107)>5

bed4 = abundance(117)>5 and abundance(119)>5 and abundance(136)>0

bed5 = abundance(M15)>50 and abundance(M18) >0 and abundance(M1)<3 and
abundance(M43)>3

bed6 = abundance(150)> abundance(151)

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 5 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and noise_counter<3 then
C10H14O = true
```

```
End Function

'*************************************************************************

'm/z 147,162,177,187,205

'm/z 220 (MW)

'high abundance m/z 41,53,55,67,79,77,79,

Function sesquiterpenol_C15H24O() as Boolean

Dim Em

Em=Endmass()

carbon_number = 15

hydrogen_number = 24

oxygen_number = 1

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3

M18 = Expected_mass -18 'loss of CH(CH3)2

M43 = Expected_mass -43 'loss of CH(CH3)2

bed1 = abundance(Expected_mass)>0 and abundance(M18)>0  and abundance(M15)>0 and
abundance(M43)>0

bed2 = abundance(41)>50 and abundance(43)>50

bed3 = abundance(53)>10 and abundance(55)>50 and abundance(59)>50

bed4 = abundance(67)>30 and abundance(69)>10

bed5 = abundance(77)>0 and abundance(79)>10 and abundance(81)>10 and
abundance(80)>0

bed6 = abundance(91)>50 and abundance(93)>30 and abundance(95)>10

bed7 = abundance(105)>10 and abundance(107)>10 and abundance(109)>0

bed8 = abundance(119)>10 and abundance(121)>10

bed9 = abundance(131)>0 and abundance(136)>0 and abundance(134)>0
```

```
bed10 = abundance(147)>10 and abundance(149)>3 and abundance(159)>3 and
abundance(177)>0 and abundance(220)>abundance(219) and
abundance(220)>abundance(222)

bed11 = abundance(187)>0

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 20 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and bed11 and noise_counter<4 then sesquiterpenol_C15H24O = true

End Function


'************************************************************************

Function C10H16Terpenes()as Boolean

Dim Em

Dim Expected_mass

carbon_number = 10

hydrogen_number = 16

oxygen_number = 0

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M13 = Expected_mass-13

M3 = Expected_mass-3


bed1 = rank(1) = 93 or rank(2) = 93 or rank(3) = 93 or rank(4) = 93 or rank(5) = 93 or rank(6)
=93 or rank(7) = 93

bed2 = abundance(91)>50 or abundance(92)>50

bed3 = abundance(121)>10 and abundance(93)>300
```

```
bed4 = abundance(67)>10 or abundance(68)>1

bed5 = abundance(77)>100 and (abundance(79)>100 or abundance(80)>80)

bed6 = abundance(105)>10 and abundance(107)>10

bed7 = abundance(Expected_mass)>10

bed8 = abundance(65)>10 and abundance(330)<30

bed9 = abundance(147)<80 and abundance(154)<5 and abundance(161)<20 and
abundance(189)<30 and abundance(204)<40

bed10 = abundance(135)<10 and (abundance(137)/abundance(136))<0.25 and
abundance(139)<20 and abundance(137)<100

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 50 then

                        noise_counter = noise_counter + 1

                end if

        next

irrational_counter = 0

        for irrational_fragment = M13 to M3 step 1

If abundance(irrational_fragment)> 30 then

                irrational_counter = irrational_counter + 1

        end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and noise_counter<=5 and irrational_counter<=1 then C10H16Terpenes= true

End Function

'*************************************************************************

Function MonoterpeneKetones()as Boolean

Dim Em

Dim Expected_mass

Em=Endmass()
```

```
For Expected_mass = 150 to 154 step 2

M15 = Expected_mass-15

M18 = Expected_mass-18

M17 = Expected_mass-17

M43 = Expected_mass-43

bed1 = abundance(Expected_mass)>20 and abundance(M15)>10 and abundance(M18)<5
and abundance(M17)<5 and abundance(M43)>20

bed2 = abundance(41)>30 or abundance(43)>10

bed3 = abundance(53)>30 and abundance(55)>10

bed4 = abundance(67)>10 or abundance(69)>10

bed5 = abundance(81)>10 and abundance(83)> 0

bed6 = abundance(91)>5 and abundance(93)>5 and abundance(95)>5 and abundance(97)>
0

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 20 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and noise_counter<3 then
MonoterpeneKetones = true

next

End Function


'*************************************************************************

'loss of [M-15]+ and [M-18]+

'MW 156

Function C10H20O()as Boolean
```

```
Dim Em

Dim Expected_mass

carbon_number = 10

hydrogen_number = 20

oxygen_number = 1

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3

M18 = Expected_mass -18 'loss of H2O

M17 = Expected_mass -17

M1 = Expected_mass -1 'irrational fragment

M43 = Expected_mass -43 'loss of CH(CH3)2

bed1 = abundance(43)>100 and abundance(43)>100 and abundance(53)>10 and
abundance(55)>100 and abundance(57)>10

bed2 = abundance(67)>50 and abundance(69)>50 and abundance(71)>50

bed3 = abundance(80)>10 and abundance(81)>50 and abundance(82)>5

bed4 = abundance(95)>50 and abundance(96)>5 and abundance(109)>10 and
abundance(123)>10

bed5 = abundance(M15)> 0 and abundance(M18) >10 and abundance(M17)>0 and
abundance(M1)<5 and abundance(M43)>3

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 5 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and noise_counter<3 then C10H20O = true

End Function
```

```
'*************************************************************************

Function Monoterpenol()as Boolean

Dim Em

Dim Expected_mass

Em=Endmass()


For Expected_mass = 148 to 156 step 2

M1 = Expected_mass +1

M15 = Expected_mass-15

M18 = Expected_mass-18

M17 = Expected_mass-17

M43 = Expected_mass-43

bed1 = abundance(Expected_mass)>2 and abundance(M15)>10 and abundance(M18)>3 and abundance(M17)>0 and abundance(M43)>20

bed2 = abundance(41)>30 or abundance(43)>2

bed3 = abundance(51)>10 and abundance(53)>10 and abundance(55)>10

bed4 = abundance(65)>10 and abundance(67)>0 or abundance(77)>10

bed5 = abundance(91)>5 and abundance(93)>0

bed6 = abundance(121)> 20 and abundance(158)<20 and abundance(130)<500

bed7 = (abundance(Expected_mass)/abundance(M1))>1.2

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 5 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and noise_counter<4 then Monoterpenol = true

next
```

```
End Function

'*************************************************************************

'm/z 147,148,161,134,133,136 (from Japanese paper)

'm/z 175,189,204 (MW)

'high abundance m/z 41,53,55,67,79,77,79,

Function sesquiterpenes_test()as Boolean

Dim Em

Em=Endmass()

carbon_number = 15

hydrogen_number = 24

oxygen_number = 0

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3

M43 = Expected_mass -43 'loss of CH(CH3)2

M1 = Expected_mass +1

M18 = Expected_mass +18

bed1 = abundance(Expected_mass)>3 and abundance(M15)>1

bed2 = abundance(M43)>3 or abundance(162)>3 or (abundance(175)>0 and
abundance(176)>0)

bed3 = abundance(43)>30 and abundance(53)>50 and abundance(55)>30

bed4 = abundance(67)>30 and abundance(69)>5

bed5 = abundance(77)>0 and abundance(79)>10 and abundance(81)>5

bed6 = abundance(91)>50 or abundance(93)>50

bed7 = abundance(107)>2 and abundance(109)>0

bed8 = abundance(119)>10 or abundance(120)>10

bed9 = abundance(136)>0 or abundance(134)>0

bed10 = abundance(147)>10 or abundance(148)>2
```

```
bed11 = abundance(M18)<6 and abundance(207)<1

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 10 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and bed11 and noise_counter<5 then sesquiterpenes_test = true

End Function

'*************************************************************************

Function Sesquiterpenes2()as Boolean

Dim Em

Dim Expected_mass

Em=Endmass()


For Expected_mass = 202 to 204  step 2

M15 = Expected_mass-15

M18 = Expected_mass-18

M43 = Expected_mass-43

M3 = Expected_mass+3

bed1 = abundance(Expected_mass)>2 and abundance(M15)>1 and abundance(M3)<10

bed2 = abundance(M43)>3 or abundance(162)>3 or (abundance(175)>0 and
abundance(176)>0)

bed3 = abundance(41)>10 and abundance(43)>1 and abundance(53)>30 and
abundance(55)>30

bed4 = abundance(67)>30 and (abundance(68)>5 or abundance(69)>5)

bed5 = abundance(77)>0 and abundance(79)>10 and abundance(81)>5

bed6 = abundance(91)>50 or abundance(93)>50
```

```
bed7 = abundance(107)>2 and abundance(109)>0

bed8 = abundance(119)>10 or abundance(120)>10

bed9 = abundance(136)>0 or abundance(134)>0

bed10 = abundance(147)>10 or abundance(148)>2

bed11 = abundance(M18)<6 and abundance(207)<1 and abundance(215)<1 and
abundance(220)<2 and abundance(221)<0.5 and abundance(222)<2 and abundance(209)<1
and abundance(272)<5 and abundance(229)<8 and abundance(165)<500

bed12 = abundance(204)>abundance(205) or abundance(203)>0 or abundance(202)>5

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 10 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and bed11 and bed12 and noise_counter<4 then Sesquiterpenes2= true

next

End Function

'*************************************************************************

Function C15H20 ()as boolean

Dim Em

Em = Endmass()

carbon_number = 15

hydrogen_number = 20

oxygen_number = 0

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15

bed1 = abundance(128)>50 and abundance(143)>50 and abundance(157)>50  and
abundance(186)>5
```

```
bed2 = abundance(Expected_mass)>50 and abundance(M15)>100
noise_counter = 0
        for noisecheck = Expected_mass+2 to Em step 1
                If abundance(noisecheck)> 30 then
                        noise_counter = noise_counter + 1
                end if
        next
If bed1 and bed2 and noise_counter<5 Then C15H20= True
End Function
'*************************************************************************
Function Sesquiterpenol2()as Boolean
Dim Em
Dim Expected_mass
Em=Endmass()

For Expected_mass = 220 to 222  step 2
M15 = Expected_mass-15
M33 = Expected_mass-33
M43 = Expected_mass-43
M42 = Expected_mass-42
M2 = Expected_mass-2
bed1 = abundance(M15)>0  and abundance(M33)>5 and (abundance(M43)>2 or
abundance(M42)>2)
bed2 = abundance(41)>100
bed3 = abundance(53)>50 and abundance(55)>50
bed4 = abundance(67)>50 and abundance(69)>10
bed5 = abundance(77)>0 and abundance(79)>50 and abundance(81)>10 and
abundance(80)>1
bed6 = abundance(91)>50 and abundance(93)>100
```

```
bed7 = abundance(105)>0 and abundance(107)>10 and abundance(109)>0

bed8 = abundance(119)>10 and abundance(121)>10

bed9 = abundance(133)>10 and abundance(136)>0 and abundance(134)>0

bed10 = abundance(147)>5 and (abundance(148)>5 or abundance(149)>5)

bed11 = abundance(161)> 10 and abundance(177)>0 and abundance(204)<200 and
abundance(189)<200 and abundance(Expected_mass)>=abundance(M2)

bed12 = abundance(272)<5 and abundance(229)<8

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 10 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and bed11 and bed12 and noise_counter<4 then Sesquiterpenol2= true

next

End Function


'*************************************************************************

'MW 222

Function C15H26O()as Boolean

Dim Em

Em=Endmass()

carbon_number = 15

hydrogen_number = 26

oxygen_number = 1

Em=Endmass()

Expected_mass = 12*carbon_number +1*hydrogen_number+16*oxygen_number

M15 = Expected_mass -15 'loss of CH3
```

```
M18 = Expected_mass-18 'loss of H2O

M32 = Expected_mass-32

M33 = Expected_mass -33 'loss of CH3OH

bed0 = abundance(Expected_mass)>0 or abundance(M15)>0

bed1 = abundance(M18)>0 and (abundance(M33)>5 or abundance(M32)>5)

bed2 = abundance(41)>100 and abundance(43)>50

bed3 = abundance(53)>30 and abundance(55)>50

bed4 = abundance(67)>50 and (abundance(68)>10 or abundance(69)>10)

bed5 = abundance(77)>0 and abundance(79)>50 and abundance(81)>10 and
abundance(80)>1

bed6 = abundance(91)>50 and abundance(93)>100

bed7 = abundance(105)>0 and abundance(107)>10 and abundance(109)>0

bed8 = abundance(119)>10 and abundance(121)>10

bed9 = abundance(133)>10 and ((abundance(136)>0 and abundance(134)>0) or
(abundance(137)>10 and abundance(138)>10))

bed10 = abundance(147)>5 and abundance(148)>5

bed11 = abundance(161)> 10 and abundance(222)>=abundance(220)

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 10 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed0 and bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
bed9 and bed10 and bed11 and noise_counter<5 then C15H26O= true

End Function

'*************************************************************************

Function Sesquiterpenes3()as Boolean

Dim Em
```

```
Dim Expected_mass

Em=Endmass()


For Expected_mass = 202 to 204  step 2

M15 = Expected_mass-15

M18 = Expected_mass-18

M43 = Expected_mass-43

M3 = Expected_mass+3

bed1 = abundance(Expected_mass)>2 and abundance(M15)>1 and abundance(M3)<10

bed2 = abundance(M43)>2 or abundance(162)>3 or (abundance(175)>0 and
abundance(176)>0)

bed3 = abundance(43)>1

bed4 = abundance(67)>30

bed5 = abundance(77)>0 and abundance(79)>10 and abundance(81)>5

bed6 = abundance(91)>50 or abundance(93)>50

bed7 = abundance(107)>2 and abundance(109)>0

bed8 = abundance(119)>10 or abundance(120)>10

bed9 = abundance(136)>0 or abundance(134)>0

bed10 = abundance(147)>10 or abundance(148)>2

bed11 = abundance(M18)<6 and abundance(207)<1 and abundance(215)<1 and
abundance(220)<2 and abundance(221)<0.5 and abundance(222)<2 and abundance(209)<1
and abundance(272)<5 and abundance(229)<8 and abundance(165)<500


noise_counter = 0
        for noisecheck = Expected_mass+2 to Em step 1
                If abundance(noisecheck)> 10 then
                        noise_counter = noise_counter + 1
                end if
        next
```

```
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and bed9 and
bed10 and bed11 and noise_counter<4 then Sesquiterpenes3= true

next

End Function

'*************************************************************************

'm/z 67, 77, 79, 81, 91, 93, 105, 107, 119, 121, 133, 147

Function Sesquiterpenes4()as Boolean

Dim Em

Dim Expected_mass

Em=Endmass()


For Expected_mass = 202 to 204  step 2

M15 = Expected_mass-15

M18 = Expected_mass-18

M43 = Expected_mass-43

M3 = Expected_mass+3

bed1 = abundance(M15)>3 or abundance(M43)>3

bed2 = abundance(67)>30 and abundance(77)>200 and abundance(79)>100 and
abundance(81)>20

bed3 = abundance(91)>500 and  abundance(93)>500

bed4 = abundance(105)>50 and abundance(107)>10

bed5 = abundance(119)>200 and abundance(121)>10

bed6 = abundance(133)>5 and abundance(134)>2

bed7 = abundance(147)>2 and abundance(148)>2

bed8 = abundance(M18)<6 and abundance(207)<1 and abundance(215)<1 and
abundance(220)<2 and abundance(221)<2 and abundance(222)<2 and abundance(209)<1
and abundance(272)<5 and abundance(229)<8 and abundance(165)<500


noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1
```

```
                If abundance(noisecheck)> 10 then
                        noise_counter = noise_counter + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and bed7 and bed8 and
noise_counter<4 then Sesquiterpenes4= true
next
End Function

'*************************************************************************
'MW272, 270
Function diterpenes()as boolean
Dim Em
Dim Expected_mass
Em=Endmass()

For Expected_mass = 270 to 272  step 2
M43 = Expected_mass -43
M42 = Expected_mass -42
bed1 = rank(1)=69 or rank(2)=69 or rank(3)=69 or rank(1) = 119 or rank(2) = 119
bed2 = abundance(41)>100 and abundance(53)>30 and abundance(55)>50 and
abundance(65)>10 and abundance(67)>50
bed3 = abundance(77)>50 and abundance(79)>50 and abundance(81)>50
bed4 = abundance(91)>50 and (abundance(93)>50 or abundance(95)>50) and
abundance(105)>30 and (abundance(107)>30 or abundance(109)>30)
bed5 = abundance(119)>10 and abundance(121)>5 and abundance(133)>3 and
abundance(147)>2 and  (abundance(161)>3 or abundance(159)>3) and (abundance(187)>2
or abundance(185)>2) and abundance(203)>0 and abundance(255)<2
bed6 = abundance(M43)>0 or abundance(M42)>0
noise_counter = 0
```

```
        for noisecheck = Expected_mass+2 to Em step 1
                If abundance(noisecheck)> 30 then
                        noise_counter = noise_counter + 1
                end if
        next
If bed1 and bed2 and bed3 and bed4 and bed5 and bed6 and noise_counter<5 Then
diterpenes = True
next
End Function
‘*************************************************************************
'm/z 41,43,55,67,69,77,79,81,91,115,128,193
Function Cannabinoids()as Boolean
dim mass                        ' New variable mass
dim threshold           ' New variable threshold (not used actual)
dim ip as parent_return ' the function ip belongs to the parentreturn substructure this is to
get the mass of the molecular ion
ip = first_parent()     ' ip is the function firstparent
mass = Ip.mass          ' the variable mass is the mass from the molecular ion
threshold=Ip.noise      ' the variable threshold is the noise from firstparent

bed1 = abundance(mass)>10 and abundance(mass-15)>0 and abundance(mass-43)>3 and
abundance(mass-56)>3 and abundance(mass-71)>5
bed2 = abundance(41)>50 and abundance(43)>50 and abundance(55)>10
bed3 = abundance(67)>10 and abundance(69)>10 and abundance(77)>10 and
abundance(79)>10 and abundance(81)>5
bed4 = abundance(91)>10 and abundance(115)>10 and abundance(128)>10 and
abundance(193)>10
bed5 = abundance(231)>10 and abundance(174)>5
If bed1 and bed2 and bed3 and bed4 and bed5 then Cannabinoids = true
End Function
```

```
'*************************************************************************

Function Cannabinoids2 () as Boolean

Dim Em

Dim Expected_mass

Em=Endmass()


For Expected_mass = 280 to 400  step 2

M1 = Expected_mass +1

M15 = Expected_mass-15

M43 = Expected_mass-43

M56 = Expected_mass-56

M71 = Expected_mass-71

bed1 =(abundance(M1)>0 and abundance(Expected_mass)>10 and abundance(M15)>1 and
abundance(M43)>2 and abundance(M56)>3 and abundance(M71)>3) or Rank(1)=231

bed2 = abundance(91)>10 and abundance(128)>0

noise_counter = 0

        for noisecheck = Expected_mass+2 to Em step 1

                If abundance(noisecheck)> 15 then

                        noise_counter = noise_counter + 1

                end if

        next

If bed1 and bed2 and noise_counter<4 then Cannabinoids2 = true

next

End Function


'*************************************************************************

Function THC()as Boolean
```

bed1 = Rank(1) = 231 or Rank(1) = 299 or Rank(1) = 314 or Rank(2) = 231 or Rank(2) = 299 or Rank(2) = 314 or Rank(1) = 41

bed2 = abundance(217)>50 and abundance(231)>400 and abundance(243)>100 and abundance(258)>100 and abundance(271)>100 and abundance(285)>10 and abundance(299)>400 and abundance(314)>400

If bed1 and bed2 then THC = true

End function

'**********

Function CBD()as Boolean

bed1 = Rank(1) = 231

bed2 = abundance(174)>50 and abundance(193)>10 and abundance(232)>10 and abundance(246)>10 and abundance(271)<200 and abundance(271)>10 and abundance(314)<200 and abundance(314)>10

If bed1 and bed2 then CBD = true

End function