2019

Computational Tools for the Untargeted Assignment of FT-MS Metabolomics Datasets

Joshua Merritt Mitchell
University of Kentucky, joshuamerrittmitchell@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0003-1598-1596
Digital Object Identifier: https://doi.org/10.13023/etd.2019.222

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Joshua Merritt Mitchell, Student

Dr. Hunter N. B. Moseley, Major Professor

Dr. Trevor Creamer, Director of Graduate Studies

Computational Tools for the Untargeted Assignment of FT-MS Metabolomics Datasets

---

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Medicine
at the University of Kentucky

By
Joshua Merritt Mitchell
Lexington, Kentucky
Director: Dr. Hunter N.B. Moseley Professor of Molecular and Cellular Biochemistry
Lexington, Kentucky
2019

[https://orcid.org/0000-0003-1598-1596]

ABSTRACT OF DISSERTATION

Computational Tools for the Untargeted Assignment of FT-MS Metabolomics Datasets

Metabolomics is the study of metabolomes, the sets of metabolites observed in living systems. Metabolism interconverts these metabolites to provide the molecules and energy necessary for life processes. Many disease processes, including cancer, have a significant metabolic component that manifests as differences in what metabolites are present and in what quantities they are produced and utilized. Thus, using metabolomics, differences between metabolomes in disease and non-disease states can be detected and these differences improve our understanding of disease processes at the molecular level. Despite the potential benefits of metabolomics, the comprehensive investigation of metabolomes remains difficult.

A popular analytical technique for metabolomics is mass spectrometry. Advances in Fourier transform mass spectrometry (FT-MS) instrumentation have yielded simultaneous improvements in mass resolution, mass accuracy, and detection sensitivity. In the metabolomics field, these advantages permit more complicated, but more informative experimental designs such as the use of multiple isotope-labeled precursors in stable isotope-resolved metabolomics (SIRM) experiments.

However, despite these potential applications, several outstanding problems hamper the use of FT-MS for metabolomics studies. First, artifacts and data quality problems in FT-MS spectra can confound downstream data analyses, confuse machine learning models, and complicate the robust detection and assignment of metabolite features. Second, the assignment of observed spectral features to metabolites remains difficult. Existing targeted approaches for assignment often employ databases of known metabolites; however, metabolite databases are incomplete, thus limiting or biasing assignment results. Additionally, FT-MS provides limited structural information for observed metabolites, which complicates the determination of metabolite class (e.g. lipid, sugar, etc.) for observed metabolite spectral features, a necessary step for many metabolomics experiments.

To address these problems, a set of tools were developed. The first tool identifies artifacts with high peak density observed in many FT-MS spectra and removes them safely. Using this tool, two previously unreported types of high peak density artifact were identified in FT-MS spectra: fuzzy sites and partial ringing. Fuzzy sites were particularly problematic as they confused and reduced the accuracy of machine learning models trained on datasets containing these artifacts. Second, a tool called SMIRFE was developed to assign isotope-

resolved molecular formulas to observed spectral features in an untargeted manner without a database of expected metabolites. This new untargeted method was validated on a gold-standard dataset containing both unlabeled and ^{15}N-labeled compounds and was able to identify 18 of 18 expected spectral features. Third, a collection of machine learning models was constructed to predict if a molecular formula corresponds to one or more lipid categories. These models accurately predict the correct one of eight lipid categories on our training dataset of known lipid and non-lipid molecular formulas with precisions and accuracies over 90% for most categories.

These models were used to predict lipid categories for untargeted SMIRFE-derived assignments in a non-small cell lung cancer dataset. Subsequent differential abundance analysis revealed a sub-population of non-small cell lung cancer samples with a significantly increased abundance in sterol lipids. This finding implies a possible therapeutic role of statins in the treatment and/or prevention of non-small cell lung cancer. Collectively these tools represent a pipeline for FT-MS metabolomics datasets that is compatible with isotope labeling experiments. With these tools, more robust and untargeted metabolic analyses of disease will be possible.

KEYWORDS: Metabolomics, Untargeted Assignments, Spectral Artifacts, NSCLC, Fourier Transform Mass Spectrometry, SMIRFE

Joshua Merritt Mitchell

(Name of Student)

05/08/2019

Date

Computational Tools for the Untargeted Assignment of FT-MS Metabolomics Datasets

By
Joshua Merritt Mitchell

Dr. Hunter N.B. Moseley
Director of Dissertation

Dr. Trevor Creamer
Director of Graduate Studies

05/08/2019
Date

ACKNOWLEDGMENTS

The following dissertation, while an individual work, was only possible thank to the insights and direction of several people. First, my Dissertation Chair, Dr. Hunter Moseley whose guidance and support enabled my work on this project. Second, Dr. Robert Flight who has been my friend and mentor throughout my PhD project. I would also like to thank my committee members and outside examiner: Dr. Trevor Creamer, Dr. David Rodgers, Dr. Sabire Ozcan, Dr. Andrew Lane and Dr. Alan Daugherty. Their feedback has greatly improved my ability to communicate my science more clearly. I would also like to thank Dr. Qing Jun Wang who performed many additional experiments to generate the datasets needed to test components of my algorithms.

Additionally, I received important assistance from family and friends. My wife and co-worker, Shruti Sinha, provided much support throughout the dissertation process as well as assistance with the computer science portions of algorithm development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ADDITIONAL FILES

CHAPTER 1. BACKGROUND

1.1 Metabolism, Metabolites and Metabolomics

Living systems require a large set of chemically diverse molecules to enable the processes of life. These molecules span a large range of chemical heterogeneity consisting of many different elements (including CHONPS (an acronym for carbon, hydrogen, oxygen, nitrogen, phosphorus and sulfur), halogens (Bergmann and Bergmann, 1991) (Schomburg and Köhrle, 2008), silicon (Jugdaohsingh, 2007), selenium (Schomburg and Köhrle, 2008), cobalt (Rickes *et al.*, 1948), calcium (Berridge *et al.*, 1999), magnesium (Lukaski *et al.*, 1996) and various other trace elements (Mertz, 2012)), masses that range from a few Daltons (e.g. H_2 (Tamagnini *et al.*, 2002)) to millions of Daltons (e.g. the protein Titin (Fulton and Isaacs, 1991)) and a wide range of chemical structures and properties (Berdy, 2005) (Bode *et al.*, 2002) (Bourgaud *et al.*, 2001). A subset of these molecules is acquired from the environment, while others are produced through the interconversion of other molecules. All molecules that are acted upon by these chemical processes in living systems can be considered metabolites including large macromolecular entities such as chromosomes and proteins as well as “small” molecules such as glucose, amino acids and lipids. Small is often defined according to an arbitrary mass cutoff (e.g. <1600 Daltons) and some definitions of metabolite are restricted to only these compounds (Moco *et al.*, 2007). While this latter definition is more restrictive, this definition of metabolite will be used throughout the dissertation.

Building upon this definition, the set of chemical processes that interconvert these metabolites constitutes metabolism (Niedenführ *et al.*, 2015). Many of these chemical processes are catalyzed by enzymes that are gene products. These enzymes not only enable chemical processes to occur at rates that are biologically relevant (Lewis and Wolfenden, 2008) (Radzicka and Wolfenden, 1995) , but also serve as important points of regulation (Metallo and Vander Heiden, 2013) that can be affected by changes in signaling or the cellular environment (Saltiel and Kahn, 2001). This regulation affects which metabolites are produced (Mor *et al.*, 2011), when they are produced (Bass and Takahashi, 2010), where they are produced (Ryu *et al.*, 2018) (Alam *et al.*, 2017) and in what quantities. Most metabolites in a living system are produced from other metabolites through these metabolic processes. This chain of derivation starts with one or more exogenously supplied metabolites, which can be referred to as 'feed metabolites' (Menküc *et al.*, 2008). In many cases, these feed metabolites are themselves metabolic products from other living systems. Therefore, metabolism can refer to the set of chemical reactions that occur in a single organism and to the web of reactions that span multiple organisms or even an entire ecosystem (Penuelas and Sardans, 2009) (Goodacre, 2007).

Describing metabolism at these higher levels of biological organization requires introducing two new terms: the metabolome and metabolomics. The metabolome is the entirety of metabolites interconverted by the metabolism of one or more living systems in the biosphere (Fiehn, 2002). Thus, metabolomes

are to metabolites what genomes, transcriptomes, and proteomes are to genes, transcripts, and protein, respectively. Metabolomics is the comprehensive study of metabolomes. In this definition, ‘comprehensive’ seeks to discriminate between metabolite ‘stamp collecting’ and true metabolomics that seeks to integrate patterns of observed metabolites and their respective concentrations (absolute or relative) within the larger system (Wishart, 2008). This definition also discriminates traditional biochemical research from metabolomics research. For example, a biochemist studying a single enzyme or a small subset of metabolites in a pathway is not necessarily performing metabolomics research, just as research regarding a single protein or a small set of related proteins is not proteomics.

Although metabolomics constitutes more than simply identifying metabolites from a living system, the robust identification of metabolites in living systems is an unsolved problem in metabolomics that must be addressed (Wishart, 2011). A major component to this problem is the immense computational complexity of developing algorithms for identifying detected metabolites in metabolomics experiments (Alonso *et al.*, 2015). The algorithms presented in Chapter 3 partially address this problem by enabling the assignment of elemental formulas without databases of known metabolites to observed features from high-resolution mass spectra.

1.2 Why Study the Metabolome?

Interest in metabolomics has grown significantly over the past several decades (Goodacre *et al.*, 2004). This surge can be attributed to many factors, including improved analytical methods for the detection and characterization of metabolites (Hu *et al.*, 2005) (Eliuk and Makarov, 2015) (Roessner and Beckles, 2009) , but also an improved understanding of how better knowledge of the metabolome provides insights into living systems in both healthy and pathological states. Metabolism is the process by which living systems generate many of the small molecules necessary for the synthesis of larger macromolecular entities. Additionally, metabolism provides the energy necessary for their synthesis. Thus, metabolites represent the currency of life and understanding which metabolites are present in a system, in what concentrations, and how they differ between disease and non-disease is extremely helpful in understanding life processes.

Metabolomics provides a window into cellular metabolism. Excluding the very small subset of reactions that occur without catalysis, metabolism, and therefore the metabolome, is a functional output of an organism's genome, transcriptome, and proteome (Gieger *et al.*, 2008). Thus, to understand the relationship between phenotypes of interest and the molecular state of a system, it is imperative to consider not only changes at the genomic, transcriptomic, and proteomic level, but also changes at the metabolic level.

To illustrate this point, consider the following example: the presence of a gene in an organism's genome does not directly indicate when an active gene

product is present, due to various types of regulation. When the metabolome is not considered, an entire layer of possible relevant regulation can be overlooked. The activities of enzymes, for instance, are often regulated by the concentration of one or more other metabolites. One mechanism by which this regulation is achieved is through the binding of metabolites to sites on enzymes to decrease or increase their activity. When these sites are not the active site, this is called allosteric inhibition or activation (Stadtman, 1970). Often the immediate or downstream products of many enzymes bind to inhibit enzyme activity in a process called feedback inhibition (Goyal *et al.*, 2010). Thus, even when there is a large concentration of a particular enzyme, the metabolic state of the system may result in all those enzyme molecules being inactive. Likewise, if a large amount of that inactive enzyme were to be expressed (say due to a constituently active transcription factor), the increase in mRNA and protein levels would result in measurable transcriptomic and proteomic changes without substantial metabolic changes. Without metabolomic information, the transcriptomic and proteomic signature of that regulatory event cannot be functionally interpreted in terms of changes in small molecule phenotype.

Metabolomics also enables insight into the short-term effects of stimuli on a system of interest in ways that observations of higher levels of biological information do not easily allow (Wilson and Nicholson, 2003). Changes at the metabolic level can be sudden and dramatic (Fernie and Stitt, 2012) (Arrivault *et al.*, 2009) . For example, when metabolite concentrations are near the inflection point of the allosteric binding curve for an enzyme or other protein, small relative

changes in a metabolite's concentration can result in dramatic changes in protein activity, especially for heterotrophic effects (van Vugt-Lussenburg *et al.*, 2006). Additionally, it has been observed that allosteric regulation can operate on a faster time scale compared to transcriptional regulation (Link *et al.*, 2013) and that metabolic alteration precedes transcriptional responses to some stimuli (Ralser *et al.*, 2009). While rapid changes in transcription (on the scale of minutes) have been observed in situations such as heat shock (O'Brien and Lis, 1993) (and these transcripts would still need to be translated), in general, metabolic regulation occurs more quickly than transcriptional regulation.

Furthermore, the substantial metabolic changes that can arise from enzyme regulation, either inhibition or activation, can result in dramatic changes in phenotype. For example, Soman, a chemical warfare agent that inhibits acetylcholinesterase can distribute throughout a human and cause neurotoxicity and death via respiratory paralysis in less than five minutes. Inhibition of acetylcholinesterase will have immediate impacts at the metabolic level including the accumulation of acetylcholine, a reduction in acetic acid and choline at the synaptic cleft and an accumulation of pinacolylmethylphosphoric acid (a metabolite of Soman) (Adeyinka and Kondamudi, 2018) (Schulze *et al.*, 2016) . These metabolic changes occur almost immediately upon exposure and although transcriptomic changes are observed in the hours following exposure, the transcriptomic profile continues to change out to 168 hours post exposure. (Dillman III *et al.*, 2009) .The ability for metabolic changes to rapidly impact phenotype are also observed with pharmaceuticals. Glucocorticoids

mechanistically alter gene expression and take hours or days to have a direct impact on an undesirable phenotype (Jin *et al.*, 2003). On the other hand, a pharmaceutical such as Sildenafil inhibits a pre-existing protein to result in the accumulation of cGMP for a relatively rapid phenotypic response (Eardley *et al.*, 2002). As many pharmaceuticals have direct and noticeable impacts on the metabolome, metabolomics provides valuable insights into the mechanisms by which these drugs function as well as ways to quantify their effects at the mechanistic level (Kell and Goodacre, 2014) (James, 2013).

Additionally metabolomics provides great insights into a living system's interaction with the environment (Athersuch, 2012). Small molecules from the environment, either from the diet or other forms of exposure (including pharmaceuticals) are subject to metabolic transformation. Some of these inputs are metabolized to produce the energy needed for cellular function or the precursor molecules needed for macromolecule synthesis. The set of possible fates for an input molecule depends on which metabolic activities are present and the state of the system (Creek *et al.*, 2012). By observing the fates of one or more input compounds to a system, a subset of metabolic activities present in the system can be inferred. For example, if glucose is observed to be converted to glucose-6-phosphate, this implies the activity of a hexokinase and provides insight into the regulatory state of the system (e.g. some hexokinases are sterol sensitive (Foretz *et al.*, 1999), some are insulin sensitive (Kruszynska *et al.*, 1998), the glucokinase isoform is regulated by the glucokinase regulatory protein (Farrelly *et al.*, 1999)). One way in which the fates of metabolites can be

determined is via stable isotope tracing, which will be discussed later. Differences in the fates of the same (or related) molecules under different conditions (disease versus no disease) can provide mechanistic explanations for how different phenotypes differ at the level of metabolism in specific mechanistic ways (e.g. a patient with glucokinase deficiency would have absent or diminished glucose -> glucose-6-phosphate activity).

Finally, although metabolism is the summation of all chemical processes that occur in a living system, metabolism is both an output of a biological system and also an input to the system. As stated previously, metabolism provides living systems with the molecules needed for energy, signaling, and the synthesis of large chemical entities. Thus, metabolites are the currency of life and are both inputs to living systems as well as intermediates and outputs. Therefore, differences in which metabolites are present and in what quantities between disease and non-disease states can be purely diagnostic or can directly drive the formation of disease or be a mixture of both. For example, the change in concentration or presence of a metabolite can imply the presence of a pathological event, but are not directly be involved in the formation of the pathological phenotype. For example, pinacolylmethylphosphoric acid is a biomarker of Soman exposure but is not involved in the mechanism of Soman toxicity. (Schulze *et al.*, 2016). On the other hand, oncometabolites such as 2-Hydroxyglutarate drive the formation of cancer by resulting in upstream epigenetic changes altering gene expression are examples of metabolites that are directly involved in the formation of a pathological phenotype. In many cases,

most disease related metabolites reside somewhere in-between on this spectrum. For example, benzo-[a]-pyrene, a carcinogenic compound found in cigarette smoke requires metabolic modification to become carcinogenic (Sims *et al.*, 1974).

These examples demonstrate the potential benefits and the necessity of studying the metabolome to understand how the small molecules in a system relate to its phenotypes of interest. Through metabolomics, the metabolome can be characterized and more accurate and predictive models of life processes at the systems-level can be constructed. In turn, these models provide researchers the ability to better predict the outcomes of molecular phenotypes and inform possible interventions to interfere or correct pathological phenotypes. However, the construction of accurate models requires high quality metabolomics data. Tools that improve the data quality of metabolomics datasets acquired using Fourier-transform mass spectrometry are presented in Chapter 2. These higher quality datasets enable more robust assignment methods described in Chapter 3 which can then be classified into lipids using tools from Chapter 4.

1.3 Metabolic Reprogramming in Cancer

For cancer cells to proliferate, grow, and survive, cancer cells produce many metabolites in different amounts than their non-transformed counterparts. For example, the production of new cancer cells requires large amounts of lipids, amino acids, and nucleic acids as precursors to produce larger molecular entities such as lipid membranes, proteins, chromosomes etc. in the daughter cell.

Additionally, the production of these large molecular entities and their precursors is energetically demanding, requiring large amounts of reducing equivalents and ATP.

Under normal conditions, the production of these precursors and the production of energy molecules are tightly regulated. However, acquiring these metabolic capabilities is an essential step in the development of cancer. The process by which cancer cells alter their metabolism to acquire these metabolic capabilities is called metabolic reprogramming and represents one of the ten cancer hallmarks (Hanahan and Weinberg, 2011).

The observation that cellular metabolism can differ significantly between cancer and non-cancer cells has been known since Dr. Warburg's studies in 1924 (Warburg *et al.*, 1924). These studies revealed that under normoxic conditions, some cancer cells will preferentially metabolize glucose to lactate (sometimes referred to as "aerobic glycolysis"), a phenomenon typically observed in normal cells only under hypoxic conditions. This property of some cancer cells is now known as the "Warburg Effect" and is just one example of metabolic reprogramming observed in cancer.

The exact roles that metabolic reprogramming serve in the development of cancer remains unclear. For example, Dr. Warburg in 1956 hypothesized that aerobic glycolysis was a byproduct of mitochondrial damage that prevented efficient mitochondrial respiration and that this process was essential in the development of cancer (Warburg, 1956). While his hypothesis was incorrect, metabolic alterations between cancer and non-cancer are common (Ward and

Thompson, 2012). Although some metabolic alterations may simply be a byproduct of disease processes, numerous studies (see examples below) have also demonstrated that the reprogramming of cellular metabolism may confer beneficial properties to cancer cells that enable their malignant proliferation.

For example, several studies have observed that the pyruvate generated by glycolysis in cancer cells is converted to lactate due to the upregulation of lactate dehydrogenase (Rong *et al.*, 2013) (Xiao *et al.*, 2016). Lactate dehydrogenase regenerates NAD+ from NADH which is needed to drive oxidative glycolysis and has been shown to promote tumor growth (Fan *et al.*, 2011). Additionally, hypoxic cells in tumors have been observed to secrete lactate into the tumor microenvironment through monocarboxylate transporters and this lactate is then taken up by oxygenated tumor cells to fuel their oxidative metabolism (Semenza, 2008) (Sonveaux *et al.*, 2008). These same studies demonstrated that abolition of this lactate uptake slows tumor growth and preferentially kills hypoxic cancer cells (Sonveaux *et al.*, 2008). This process is also believed to be important for the maintenance of intracellular pH and the creation of an acidic tumor microenvironment (Pinheiro *et al.*, 2010), which has been shown to inhibit immune function (Huber *et al.*, 2017) and promote cancer metastasis (Riemann *et al.*, 2016).

Furthermore, a shift towards aerobic glycolysis may reduce reactive oxygen species formation (Brand and Hermfisse, 1997); however, other studies indicate that higher levels of reactive oxygen species may be beneficial in cancer development (Mittler, 2017). Finally, previous studies have shown that the

upregulation of lactate production can promote cancer resistance to commonly used chemotherapeutic drugs (Li *et al.*, 2015).

While lactate metabolism may contribute to cancer development in a variety of ways, it is not the only metabolite of glucose that can be produced by glycolysis. Additionally, several of the intermediates of the glycolysis pathway can be shunted to produce other necessary biomolecules, e.g. glucose-6-phosphate is needed for the oxidative branch (NADPH-producing) of the pentose phosphate pathway (PPP) and in glycogen synthesis, fructose-6-phosphate is needed for PPP, and pyruvate is needed for alanine biosynthesis. Therefore, aerobic glycolysis not only produces the energy needed for cancer cells, albeit at 2 ATP at a time, it may also provide a source of necessary precursors for other anabolic processes.

Although these examples demonstrate the potential beneficial role of glycolysis in cancer cells, enhanced glycolysis is also a normal response to hypoxia. So while enhanced glucose uptake via the upregulation of glucose transport has been observed in cancer (Schwartzenberg-Bar-Yoseph *et al.*, 2004) (Kaira *et al.*, 2014) (Pinheiro *et al.*, 2011) and could be a compensatory mechanism for reduced energy production from glycolysis, it is also an expected response to hypoxia (Ouiddir *et al.*, 1999). The relationship between hypoxia related proteins such as HIF-1alpha and cancer implicated growth factors and signaling pathways (such as epithelial growth factor receptor (EGFR) (Lee *et al.*, 2009)) suggest the relationship between hypoxia and cancer development is complex. Regardless, the observation that glucose uptake into cancer cells is

typically much higher than glucose intake into non-cancer cells is the principle underlying positron emission tomography (PET) imaging (Gambhir, 2002). However, a significant number of malignant tumors have been found to be PET-negative (Lieberman *et al.*, 2011) and the glucose uptake measured by PET is not always a useful estimate of cell proliferation or cancer grade (Avril *et al.*, 2001).

Metabolic reprogramming has been observed to occur in other pathways beside glycolysis. Similar reprogramming has also been observed in the TCA cycle. Although a shift towards glycolysis may seem to imply a lower demand for mitochondrially derived ATP, which largely depends on TCA cycle generated reducing equivalents and their reoxidation via ATP-coupled respiration, this may not be the case in all situations. For example, enhanced fatty acid oxidation has been observed in a variety of cancers (Liu, 2006) and is a major source of ATP via the TCA cycle and oxidative phosphorylation. Furthermore, even if TCA-derived ATP is not in demand, the TCA cycle is the source for both reducing equivalents and TCA cycle intermediates needed for biogenesis (Jin *et al.*, 2016). For example, oxaloacetate can be converted to aspartate which is needed for pyrimidine biosynthesis, succinyl-CoA is needed for porphyrin biosynthesis, and citrate from the TCA cycle is the major source of cytosolic acetyl-CoA needed for lipid production. However, removing these intermediates from the TCA cycle cannot occur indefinitely and additional carbon must be added to maintain the cycle. Several studies have observed the upregulation of anapleurotic reactions that produce TCA cycle intermediates (either directly or indirectly) including

glutaminolysis (Lu *et al.*, 2010) and pyruvate carboxylase (Sellers *et al.*, 2015a), which both produce oxaloacetate. Combined with a source of acetyl-CoA (such as beta oxidation or pyruvate dehydrogenase), these reactions enable the TCA cycle to continue even if intermediates are removed for biosynthetic roles.

Additionally, glutamine appears to play a key role in other aspects of cancer development. High extracellular glutamine for example may drive cell transformation (McKeehan, 1982) and intracellular glutamine serves a carbon source for pyrimidine biosynthesis and nitrogen for both purine and pyrimidine biosynthesis (Lacey and Wilmore, 1990). Also the reductive carboxylation of alpha-ketoglutarate (a metabolite of glutamine) has been observed to be necessary for lipid production in hypoxic conditions such as a poorly vascularized tumor or in cells with mitochondrial defects (Mullen *et al.*, 2014). The ability of altered glutamine metabolism to sustain key portions of metabolism under harsh conditions could contribute to the resiliency of cancer cells but also can result in “glutamine addiction”, which could be exploited for therapeutic purposes (Wise and Thompson, 2010) (Bolzoni *et al.*, 2016).

Metabolic reprogramming is not limited to glycolysis or the TCA cycle. Both the PPP and lipid biosynthesis pathways are often upregulated as well. The PPP produces the majority of the pentoses needed for nucleotides (Raïs *et al.*, 1999), and the oxidative branch is a major source of NADPH, which is needed for lipid and cholesterol biosynthesis, the regeneration of glutathione, and thus the neutralization of reactive oxygen species (Rush *et al.*, 1985). The first step in the oxidative branch of the PPP, glucose-6-phosphate dehydrogenase (G6PD) has

been shown to be a regulator of oxidative stress (Nóbrega-Pereira *et al.*, 2016) and G6PD mutants that reduce NADPH production are possibly protective against cancer (Dore *et al.*, 2016), while upregulation of G6PD has been observed in cancer (Jonas *et al.*, 1992). Altered lipid metabolism is also observed in cancer. Lipids serve as the building blocks of cellular membranes and signaling molecules and thus are necessary for malignant proliferation. *De novo* lipid synthesis provides many of these lipids for some cancer cells (Hilvo *et al.*, 2011) and the production of these lipids requires acetyl-CoA, NADPH and ATP. Additionally, sterol production is enhanced in cancer as well and has been correlated with increased invasiveness (Bao *et al.*, 2016) and drug resistance (Janvilisri *et al.*, 2003).

The observation that metabolic reprogramming occurs ubiquitously in cancer has led some researchers to claim that cancer is as much a metabolic disease as it is a genetic disease (Seyfried *et al.*, 2013). While there is truth to this claim, since cancer metabolism is the functional output of a cancer cell's genome (or a tumor's genome given the amount of heterogeneity observed in tumors (Paz-Yaacov *et al.*, 2015)), the relationship between cancer metabolism and genetic alterations observed in cancer are complex and clearly not independent. For example, well known cancer related genes have less well-known roles in regulating cellular metabolism. For instance, the tumor suppressor gene *TP53* not only slows cancer growth through cell-cycle arrest and the activation of apoptosis (Livingstone *et al.*, 1992), but also through the regulation of many metabolic pathways both directly and indirectly. Activated p53 has been

shown to bind and inhibit G6PD to limit flux through the pentose phosphate pathway (Jiang *et al.*, 2011). Additionally, p53 promotes the expression of *TIGAR* which lowers fructose-2,6-bisphosphate levels to inhibit glycolysis (Bensaad *et al.*, 2006), and has been shown to regulate many effectors of lipid metabolism including SREBP-1 (Yahagi *et al.*, 2003), SIRT-1 (Nemoto *et al.*, 2004), aromatase (Wang *et al.*, 2012), Acad11 (Jiang *et al.*, 2015), Lipin1 (Assaily *et al.*, 2011) and caveolin 1 (Bist *et al.*, 2000). Thus, loss of *TP53* not only inhibits cell cycle arrest and apoptosis (Livingstone *et al.*, 1992), it also enables the dysregulation of metabolism in a manner that promotes cancer growth.

Additionally, many oncogenes also have important roles in cancer metabolic reprogramming. *C-Myc* overexpression is correlated with a coordinated change in many aspects of cellular metabolism. C-Myc signaling has been shown to stimulate glycolysis through the upregulation of GLUT1 HK2, PFKM (Osthus *et al.*, 2000) and to promote glutaminolysis (Wise *et al.*, 2008). Similar coordinated alterations in metabolism are observed with *Ras* (Kimmelman, 2015), *EGFR* (Kim *et al.*, 2018), and *ALK* mutations (McDonnell *et al.*, 2013). Furthermore, several metabolic enzymes have been shown to have potential roles in tumor suppression such as fumarate hydratase (Costa *et al.*, 2010), further highlighting the complex relationship between metabolism and cancer development.

Metabolic reprogramming, while necessary for the development of cancer, can also provide potential drug targets to aid in the treatment of cancer. Due to their altered metabolic programming, many cancer cells become dependent on metabolic pathways that are not essential for the survival of healthy cell types.

One example is the dependency of some tumors with *c-Myc* mutations on glutaminolysis (Bolzoni *et al.*, 2016; Wise *et al.*, 2008; Wise and Thompson, 2010). These glutamine-addicted tumors are incapable of surviving without glutaminolysis, making the inhibition of this pathway particularly effective against these cancers (Effenberger *et al.*, 2017). However, in which cancer cell subtypes and under which conditions this process occurs *in vivo* remains unclear (Yuneva *et al.*, 2012) (Marin-Valencia *et al.*, 2012). Similarly, the reliance of some cancers on aerobic glycolysis makes inhibition of one or more glycolytic enzymes a possible approach for anticancer treatment (Sheng and Tang, 2016).

By targeting the downstream metabolic effects of oncogenic mutations, rather than attempting to inhibit the oncogene mutants themselves, several benefits can be had. First, the set of possible oncogenic mutations for any protein is potentially quite large. For example, 44 known point mutations have been observed in *Ras* across human cancers (Prior *et al.*, 2012). Although some overlap is expected for small molecule inhibitors of these mutants, several inhibitors might be necessary to cover the entire range of clinically relevant mutants. However, inhibiting the mutant Ras itself may not be necessary to provide a benefit in cancer patients with Ras mutants. Since many of these mutants result in the same downstream metabolic changes, inhibitors that target the non-mutated enzymes in the downstream pathways allow us to side step the problem of developing inhibitors for each mutant Ras. Additionally, many signaling pathways overlap and converge on the same set of metabolic changes, implying that the same set of metabolic enzyme inhibitors that are used for one

genetic subtype of cancer can likely be reused for another genetic subtype of cancer as well. Therefore, even if a generic Ras inhibitor (that works on all mutant versions of Ras equally well) could be developed and would be useful in the patients with Ras mutants (or closely related proteins), the potential use of metabolic enzyme inhibitors in a variety of cancer subtypes makes their development worthwhile. However, identifying which enzymes make suitable drug targets requires an understanding of how metabolite profiles and metabolism differ between cancer and non-cancer. This problem can be solved using metabolomics techniques. An example of how metabolomics techniques can be applied to improve our understanding of cancer metabolism is shown in Chapter 5.

1.4 Isotope Labeling, Metabolomics and mSIRM

As alluded to in previous examples, information regarding the fate of one or more metabolites in a metabolic network provides significant insight into the molecular state of that system. While conceptually straightforward, performing these experiments can be difficult and require ‘marking’ input metabolites to distinguish them from other metabolites in the system. One mechanism by which metabolites of interest can be marked is through isotopic labeling (Kruger and von Schaewen, 2003) (Wiechert, 2001) (Isin *et al.*, 2012).

Every nucleus of a given element has the same number of protons (e.g. all nuclei with 6 protons are carbon) but not all nuclei with the same number of protons contain the same number of neutrons. Each instance of a nucleus with

the same number of protons but different numbers of neutrons are examples of isotopes of that element. Although the possible number of isotopes for any given element is theoretically large, the relationship between the number of protons and neutrons in a nucleus and the stability of the nucleus greatly restricts the number of naturally occurring isotopes of that element (Erler *et al.*, 2012). For example, carbon has fifteen isotopes that can be produced naturally and artificially, but only three of them: ^{12}C, ^{13}C and ^{14}C occur naturally. A subset of the naturally occurring isotopes for all elements are considered stable and either do not undergo radioactive decay or do so at an extremely small rate to be considered effectively stable (e.g. ^{40}K). Although all isotopes of an element will share many chemical properties such as the number of bonds they form, isotopes of the same element will differ in their atomic masses and can differ in their radioactivity or spectral properties. Therefore, compounds with identical chemical structures but different isotopic compositions can differ in their physical properties (Jue *et al.*, 1989) (Diem *et al.*, 1982) (Wiechert, 2001).

To describe different isotopic forms of the same compound, the terms isotopologue and isotopomer are used. Isotopomer refers to a specific isotoped version of a chemical structure that considers which isotopes are present in a structure and their location within a structure. Isotopomers are often described using a nomenclature that appends which isotopes and where the isotopes are located to the common name or IUPAC name for a compound (e.g. [2-^{13}C]-propanal represents a propanal molecule with a ^{13}C isotope at position 2). When all positions of one element are replaced with an isotope such as [1,2,3-^{13}C]-

propanal, this is called uniform labeling and can be abbreviated as [U-^{13}C]-propanal. Isotopologues represent sets of isotopomers that share the same number and types of isotopes. For example, the three isotopomers of propanal with exactly one ^{13}C: [1-^{13}C]-propanal, [2-^{13}C]-propanal and [3-^{13}C]-propanal collectively constitute one isotopologue of propanal. Isotopologues can be described either using isotope-resolved molecular formulas (IMFs) that are similar to elemental molecular formulas (EMFs) but contain isotope information (e.g. $^{12}C_2{}^{13}C_1{}^{1}H_6{}^{16}O_1$ is the IMF corresponding to the isotopologue of propanal with one ^{13}C) or by using a notation that describes the isotope differences between sets of isotopologues using the lowest mass isotopologue (m) as a reference. Using this nomenclature, the one ^{13}C isotopologue of propanal is described as m+$^{13}C_1$. Examples of isotopologues, isotopomers, and the notation used to describe them is shown in Figure 1.1.

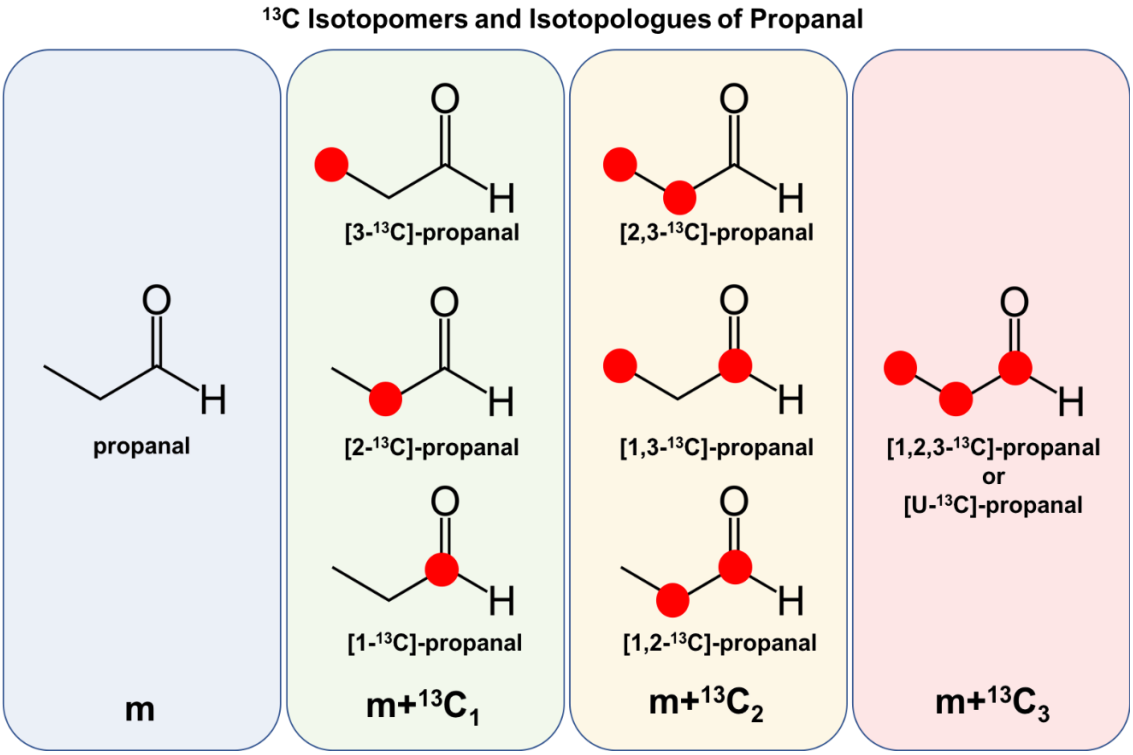

Figure 1.1: 13C Isotopomers and Isotopologues of Propanal
The differences between isotopomers and isotopologues can be demonstrated using propanal as an example. Red dots represent carbons that are 13C. Isotopomers are specific isotope versions of a chemical structure. There are three distinct 13C isotopomers of propanal containing exactly one 13C (green), three 13C isotopomers of propanal with two 13C (yellow) and one 13C isotopomer with three 13C. Isotopologues are sets of isotopomers containing the same set of isotopes. For example, all three 13C isotopomers of propanal containing exactly one 13C are all examples of the m+13C1 isotopologue of propanal. In this case m refers to the lowest mass isotopologue of the compound, in this case that is propanal consisting of only 12C, 1H, and 16O.

Isotopologue and isotopomers can differ from one another in terms of their physical properties. All isotopologues of a compound have different exact masses allowing their disambiguation by mass spectrometry, while different isotopomers can have unique spectral features due to the different positions of potentially spectrally unique isotopes in the structure. These differences in the physical properties of isotopomers and isotopologues determine which analytical

methods are suitable for studying them. By selectively replacing one or more nuclei in a metabolite with a known isotope, isotopically labeled versions of metabolites can be produced that can be distinguished from the non-labeled versions of the same metabolite.

In isotope tracing, isotopically labeled compounds are produced and then introduced into a biological system of interest. Once introduced (e.g. through the diet (Sun *et al.*, 2017) or through an intravenous bolus (Lane *et al.*, 2015) or cannulation (Maher *et al.*, 2012)), metabolism then acts upon this labeled compound to distribute the labeled isotopes throughout the system. By observing into which compounds and which compartments the label is distributed and at what time points, information about the state of the system (i.e. what enzymes are present and active, what regulatory pathways are active, etc.) can be inferred. Additionally, when isotope labeling is considered, enzymatic reactions that represent mappings of substrate compounds to product compounds also represent mappings of substrate isotopomers to product isotopomers and by extension sets of mappings of substrate isotopologues to product isotopologues (Arita, 2012). For example, carbon 1 of glucose becomes carbon 1 of glucose-6-phosphate and therefore represents a mapping between [1-^{13}C]-glucose and [1-^{13}C]-glucose-6-phosphate (isotopomer to isotopomer mapping) and a mapping between m+$^{13}C_1$ glucose and m+$^{13}C_1$ glucose-6-phosphate. Throughout a metabolic network, these mappings are applied at every reaction and therefore determine where isotope labels can ultimately be distributed across the network. Observed labeling patterns can be explained through the actions of enzymes by

tracing the path of reactions that map input isotope labels to their observed locations in labeled derivatives.

To illustrate how this tracing occurs and how it can provide information about the metabolic activities present in a system, consider the following outcomes in a cultured human breast cancer cells when grown on U-^{13}C glucose enriched media, producing a distribution of ^{13}C in lipids (Lane *et al.*, 2009). If many lipids that have an even number of 13C are observed following the bolus, this implies that ^{13}C-acetylCoA was introduced to pre-existing lipids or to pre-formed lipid backbones. The best explanation for this observation is that the ^{13}C glucose was metabolized via glycolysis and the TCA cycle to produced U-^{13}C Acetyl-CoA which was then conjugated to existing lipids or existing lipid backbones. Since Acetyl-CoA has only 2 carbons, the labeled ^{13}C must be added in multiples of two. Pre-existing lipids that were extended can be distinguished from fully de novo synthesized fatty acyls if the total number of carbons (minus carbons in the head group) is larger than the total number of ^{13}C added to the lipids. Alternatively, or additionally, if glycerolipids with an odd number of ^{13}C are enriched, a possible explanation for this observation is that some portion of the glucose was metabolized to yield ^{13}C glycerol-3-phosphate, which was then utilized to form lipid headgroups. Since glycerol has three carbons and acetyl-CoA has two, if glycerol were labeled, the resulting lipids will have an odd number of ^{13}C. This interpretation could be confirmed using NMR to measure enrichment of ^{13}C in the incorporated glycerol. This relatively simple example shows, interpreting these results requires an understanding of metabolism and

can become very labor intensive. One outstanding problem in the isotope tracing field is the automated deduction of which pathways best explain the observed distribution of label from one or more labeled precursors.

Although this example only used a single stable isotope as an example (^{13}C), isotopic labeling and tracing can involve multiple isotopes (Yang *et al.*, 2017b) (Montigon *et al.*, 2001) as well as radioactive isotopes. The earliest isotope labeling experiments employed radioactive isotopes to produce radioactive versions of input molecules. By controlling the total amount of radioactivity and then observing what fraction of the radioactivity was distributed into which compartment, the volume of distribution of those compounds (especially pharmaceuticals) and how they are excreted can be determined (Glass *et al.*, 1980) (Huang *et al.*, 1998) (Lister-James *et al.*, 1996). However, introducing radiolabeled compounds into living systems poses several complications, the least of which is the potential effects of radioactivity on the health of these systems, especially human subjects. Additionally, since radioisotopes are radioactive they are constantly leaving the system through radioactive decay which complicates the determination of kinetic parameters (Williams *et al.*, 1995) and limits the time frames over which experiments can be performed (e.g. some isotopes such as ^{11}C or ^{15}O have half-lives of ~20 minutes and ~2 minutes respectively (Conti and Eriksson, 2016)). This limitation is less relevant for longer lived radioisotopes. Finally, the use of multiple different radiolabels in a single experiment becomes complicated if their decay processes are indistinguishable. An alternative to radiolabeling is stable isotope labeling in

which stable isotopes in a compound are replaced with other stable isotopes. Common substitutions include ^{13}C for ^{12}C and ^{15}N for ^{14}N. When compounds have been substituted with respect to one or more different elements for use in a metabolomics experiment, this technique is referred to as multiple stable isotope resolved metabolomics (mSIRM) (Yang *et al.*, 2017b).

The use of stable isotopes in mSIRM enables their relatively risk-free use in humans and other biological systems but with several significant disadvantages. First to perform tracing, it is necessary to identify not only what metabolites are present, but also what isotopologues and/or isotopomers are present as well. This only complicates the already difficult bioanalytical problem of identifying metabolites in complex biological samples. Our assignment algorithm described in Chapter 3 accounts for isotope labeling and can assign labeled and unlabeled data. Second, most stable isotopes are also found in nature and the mole fraction of an element that is represented by a particular isotope is the natural abundance of that element and for some isotopes this percentage can be small (^{2}H = 0.000115) or relatively large (^{37}Cl = .2424) (Linstrom and Mallard, 2001). The contribution from natural abundance must be considered when attempting to quantify the absolute amount of label present in a derivative (Fernandez *et al.*, 1996) (Moseley, 2010). For example, in the ^{13}C-enriched lipids, some percentage of the lipids will have 1 or more ^{13}Cs even without a contribution from the artificially introduced ^{13}C. The process of correcting for the effects of natural abundance is uncreatively referred to as natural abundance correction, and determining when to correct and how to

correct can dramatically impact how experimental results must be interpreted. Furthermore, any technique for identifying metabolite isotopologues or isotopomers based on their concentrations must consider the effects of natural abundance.

1.5 Analytical Techniques Employed for Metabolomics

A necessary step in all metabolomics experiments is the observation, identification and quantification (absolute or relative) of metabolites from samples. Due to the structural diversity of metabolites and the wide range of chemical properties they possess, the systematic and comprehensive analysis of metabolites represents a significant bioanalytical problem that must be solved to allow more meaningful metabolomics experiments (Doerr, 2016). Although the entire range of analytical chemistry techniques have been applied to metabolomics experiments, nuclear magnetic resonance (NMR) (Fan and Lane, 2011) and mass spectrometry (MS) (often in conjunction with chromatography) (Dettmer *et al.*, 2007b) (Abdelrazig, 2015) (Kanani *et al.*, 2008) (Fang and Gonzalez, 2014) remain the most popular techniques employed for metabolomics.

Both NMR and MS produce spectra as their output where each spectrum consists of multiple spectral features (peaks) that correspond to features of the compounds present in the sample. The two methods differ in how spectra are generated and what physical phenomena they observe to generate spectra. In NMR, spectral features corresponding to sets of magnetically equivalent nuclei

are generated by observing the nuclear magnetic resonance of the nuclei. All nuclei with a nonzero nuclear spin have an intrinsic nuclear magnetic moment which generates a weak magnetic field. These nuclei are “NMR active” and the commonly NMR-detected nuclei include ^{1}H, ^{13}C, ^{15}N. These nuclei are naturally abundant isotopes of elements found in metabolites and ^{13}C and ^{15}N are common isotope labels. The total number of spin states for a nucleus is 2*I + 1, where I is the nuclear spin of that nuclei. All three of these commonly detected “NMR active” nuclei have spin ½ (the lowest spin that is NMR active, other higher spins are possible as well for other nuclei) and as a result each of these nuclei have two spin states: spin up (+½) and spin down (-½). In the absence of a magnetic field, these spin states are equal in energy but in a strong magnetic field, the two spin states are no longer energetically equivalent and the low-energy spin state will become more highly populated than the high spin state at equilibrium. A short pulse of broad-spectrum radio frequency is then applied. The magnetic component of the radio frequency pulse excites the nuclei and results in a non-equilibrium distribution of nuclei between the spin states (i.e. more nuclei will be in the high energy state than at equilibrium). There is now a net magnetization that will decay back to equilibrium with time. As the net magnetization decays, it also precesses (precession occurs whether there is excitation or not) and an electromotive force is generated, causing a decaying induced current flow that can be measured. This measured current flow is called a free induction decay (FID) and can be fitted to a series of exponentially decaying sinusoids and cosinusoids (terms) using the Fourier transform. These

terms represent nuclei in different magnetic and electrical environments and their frequencies can be used to generate an NMR spectrum. Averaging multiple acquisitions yields spectra with good signal-to-noise ratios where each peak in frequency space represents a population of magnetically unique nuclei. By changing which frequency ranges are scanned and what pulse sequences are utilized, a variety of experiments can be performed using NMR; however, the underlying physics of how compounds are detected remains the same.

Unlike NMR where the same physical phenomenon is employed for observing compounds regardless of instrument make or configuration, mass spectrometry refers to any analytical technique that measures the mass-to-charge ratio of ions. To ionize analyte compounds, a variety of techniques are employed each with different advantages and disadvantages. Typically, more harsh ionization techniques such as inductively coupled plasma (Bazilio and Weinrich, 2012) and chemical ionization have better ionization efficiencies but are destructive, while less harsh techniques such electrospray ionization are less efficient but relatively non-destructive. The ionization system is often the first step in a deployed MS system and can be placed in-line after a chromatography system (e.g. GC-MS, LC-MS) or sample can be directly injected into the ionization system (e.g. direct infusion). Some compounds become charged themselves during this process, while others associate with other ions to form adducts that are charged or associate with other molecules to form complexes (Venter *et al.*, 2008) (Cuyckens and Claeys, 2004). The presence of multiple

adducts per ion and complexes contribute to the complexity of assigning mass spectra.

Once ionized, ions can be moved using electromagnetic fields into one or more mass detectors. Different types of mass detectors use different physical properties of ions to infer the mass-to-charge ratio of ions. Common types of mass analyzers include: quadrupoles that are constructed of electrodes that with the appropriate potential applied will select for a specific *m/z* range of ions (March and Londry, 1995) (March, 2006) , time-of-flight analyzers that back calculate the mass-to-charge ratio of an ion based on the amount time needed for the ion to travel a known distance in a vacuum (Wiley and McLaren, 1955) , and magnetic sector analyzers that deflect ions into paths of unique radii based on their mass-to-charge and the strength of the magnetic field (Mattauch, 1936). Various types of mass analyzers are used for MS-based metabolomics, but the highest mass accuracy, resolution, and sensitivities are achieved using the Fourier transform mass spectrometry family of mass analyzers.

The FT-MS mass analyzers consist of two different instrument types: the Makarov trap mass analyzer (often referred to by its trade name Orbitrap) and the ion cyclotron resonance (ICR) mass analyzer (Zubarev and Makarov, 2013). An Orbitrap analyzer consists of two electrodes -an inner and outer electrode – that are separated by a vacuum into which ions can be injected. Between the two electrodes, a large voltage potential is applied, resulting in an electrical field that can trap ions within the volume separating the electrodes. Once trapped, ions orbit the inner electrode in an elliptical trajectory, while simultaneously

experiencing harmonic axial motion. This axial motion has an angular frequency that depends on the mass-to-charge ratio of the ion, but no other parameters of the ions themselves. Thus, if the axial motion of the ions can be detected, the mass-to-charge ratios of the ions can be inferred. Axial oscillations can be induced in trapped ions with radiofrequency pulses that are close in frequency (resonant) with the axial frequencies of the trapped ions. Using properly constructed waveforms, effectively all ions in the trap can be excited simultaneously. Once excited, the oscillations of these ions induce a current in the outer electrode that can be detected (like an FID in NMR). The acquired current is a function of time but can be converted to the frequency domain using the Fourier transform. From these transformed frequencies, the mass-to-charge ratios of ions present as well as the relative intensity of the induced current for that ion species (a function of the number of ions of that species in the trap) can be inferred.

An ICR-type analyzer is in general very similar to an Orbitrap with the main difference being that a magnetic field is used to trap ions in ICR analyzers rather than an electrical field and that the angular frequency of interest for the ions depends upon their ion cyclotronic frequencies (Comisarow and Marshall, 1974). In ICR, application of a radiofrequency pulse induces the trapped ions to a larger cyclotron radius. As in the Orbitrap, the passing of the ions near electrodes in the analyzer induces the formation of a current that is captured as the ions decay from their larger cyclotron radii. This current is an FID that can be Fourier transformed and further processed to yield a mass spectrum. Typically, multiple

FIDs are acquired and transformed to produce a ‘scan’ and multiple scans are combined to generate an aggregate spectrum. On some instruments, the FID for a scan is a summation of multiple microscans, each corresponding to an FID acquired on the same set of ions in the trap. The number of FIDs per scan is controlled by the microscan setting.

The superior analytical capabilities of both orbitraps and ICR instruments are related to their shared use of the Fourier transform to produce spectra. Unlike other mass analyzers where each ion’s mass-to-charge ratio (or a proxy of the ratio) is observed only once per ion, each FID represents multiple measurements of every excited ion’s mass-to-charge ratio. Although each individual measurement has an error component, the Fourier transform effectively averages all these measurements into a single observation, greatly reducing any non-systematic sources of error and increasing signal-to-noise significantly (this effect is called Fellgett’s advantage (Fellgett, 1949)). However, when the mathematical assumptions of the Fourier transform are not met exactly (which is often the case with real-world data) spectral artifacts can be introduced because of this mathematical processing.

Despite their differences, both FT-MS and NMR remain popular choices for metabolomics experiments for several important reasons. First, both instruments are sensitive to a wide-range of chemical structures, which is necessary for observing the entire set of metabolites with as little selection bias as possible. NMR requires the presence of one or more NMR-active nuclei in a compound (effectively all metabolites thanks to ^{1}H and ^{13}C), while mass

spectrometry simply requires that a metabolite be ionizable or already ionized. Second, both instruments are capable of some form of quantification. NMR can achieve absolute quantification with the appropriate standards, while MS can be useful for relative quantification. Third, both NMR and mass spectrometers (although to a less extent FT-MS instruments) are relatively common and have been used extensively for chemical identification and characterization. As a result, large databases of chemical compounds with their corresponding NMR and mass spectra are available to help map spectral features to chemical structures, a process called assignment. However, for mass spectrometry, these databases can be problematic for assigning metabolite features. a problem that will be discussed in more detail in Chapter 3.

While FT-MS and NMR both share many properties that make them ideally suited for metabolomics applications, they also differ significantly in their capabilities. First and foremost, NMR and FT-MS have substantially different detection limits. FT-MS can detect compounds with femtomolar concentrations under ideal circumstances (Eyles and Kaltashov, 2004) (Groenen and van den Heuvel, 2006) while NMR is typically limited to compounds with micromolar concentrations (Gronwald *et al.*, 2008) although in limited cases much better detection limits can be achieved (Spiess, 2008). For metabolomics applications, where the concentration of metabolites of interest are largely outside the control of the researcher, this difference in detection limit can greatly restrict what metabolites can be detected in a living system. For example, signaling molecules and toxins may have significant roles in living systems at very low

concentrations, but would be undetectable by NMR. Metabolites span a large concentration range from the femtomolar range (iodothyronine metabolites in brain tissue (Pinna *et al.*, 2002)) to millimolar concentrations (Cubbon *et al.*, 2010).

However, what NMR lacks in sensitivity it compensates for in terms of the amount and types of information it can provide regarding metabolites. In MS, sets of ions with the same mass-to-charge ratio are indistinguishable. If two ions have the same mass-to-charge ratio only one spectral feature will be produced. While identical nuclei with identical magnetic environments will result in identical signals, an NMR observes many NMR active nuclei per compound and even very similar chemical structures will produce substantially different NMR spectra. Additionally, the NMR spectrum of a compound depends both on the composition and the relative arrangement of isotopes in its structure, thus NMR can distinguish not only isomers from one another but also isotopomers of the same compound. FT-MS on the other hand, cannot resolve isotopomers of the same compound or isomers of the same compound without orthogonal information such as chromatographic retention times (Ferreres *et al.*, 2004), chemoselective adduct formation (Nishikaze *et al.*, 2017) or MS/MS (Menicatti *et al.*, 2016). All three of these techniques require more complicated experimental setups, greater volumes of sample and ultimately greatly increased labor to collective, curate, and process the generated data (Chekmeneva *et al.*, 2017).

Furthermore, while sample preparation for any analytical technique can introduce artifacts or biases that result in analytical samples that do not represent

the biological samples from which they are prepared, FT-MS requires that samples also be ionized. Ionization results in a variety of artifacts that can complicate the interpretation of spectral data. First, not all compounds ionize equally well and the efficiency with which compounds ionize depends on experimental conditions (Rauha *et al.*, 2001). A sample with an equal concentration of many metabolites with differing ionization efficiencies will appear as a mixture of metabolites with different concentrations to the mass analyzer. This effect makes mass spectrometry poorly suited for absolute quantification. Second, some compounds preferentially form ions of a certain charge or specific adducts – this can introduce biases in which compounds are detected and assigned (e.g. positive mode spectra will not observe negatively charged ions). Furthermore, ionization efficiencies and ion / adduct formation can be influenced by the composition of the sample. A common example of this is ion suppression where the presence of high concentrations of one or more compounds 'suppresses' the signal of less abundant compounds (Jessome and Volmer, 2006) (Buhrman *et al.*, 1996). Finally, even the softest of ionization methods are still destructive. Ionization can induce chemical rearrangements to occur or cause metabolites to fragment resulting in chemical species that can be detected but are not representative of the biological samples from which they were prepared (Kingston *et al.*, 1983) (McLafferty, 1959). Even small effects such as the degree of corrosion on electrospray emitters can cause ionization artifacts (Chen and Cook, 2007). Furthermore, FT-MS artifacts can also arise as a byproduct of the mathematical manipulation needed to convert observed FIDs to

actual spectra (Kanawati *et al.*, 2017) (Xian *et al.*, 2013) (Mitchell *et al.*, 2017). These artifacts can lead to large interpretive errors in downstream analyses. Methods for handling a subset of these artifacts is described in Chapter 2.

Due to their relative strengths and weaknesses neither NMR nor FT-MS alone are ideal for the entire range of expected metabolomics experiments. NMR's ability to provide structural information and quantification makes it suitable for chemical structure determination, the characterization of newly discovered metabolites (if they can be isolated and concentrated sufficiently) and for the quantification of specific isotopomers of high abundance metabolites. FT-MS on the other hand is best suited for detecting low concentration metabolites or sets of isotopologues and situations where relative quantification is enough. This makes FT-MS ideal for the discovery of new metabolites. However, the lack of structural information provided by FT-MS to confirm metabolite assignment, the presence of spectral artifacts that can be confused for novel metabolites and the obvious absence of newly-observed metabolites in assignment databases requires new tools for the processing and assignment of FT-MS data. Algorithms for handling spectral artifacts in FT-MS are presented in Chapter 2 and an assignment algorithm capable of assigning molecular formulas to spectral features corresponding to novel metabolites is presented in Chapter 3. These tools are fully automated to handle the large volumes of data that can be produced by FT-MS in high throughput experiments.

1.6 The Assignment Problem and Untargeted versus Targeted Metabolomics

Assignment is the process by which observed spectral features in an FT-MS spectrum are assigned to either molecular formulas or directly to metabolites. Without accurate assignments, the ability to relate observed patterns in FT-MS datasets to phenotypes observed between samples at a biochemical level would effectively be impossible. Obviously, incompletely assigned spectra and incorrectly assigned spectra are undesirable and can result in substantial interpretive errors in downstream data analyses.

Despite the substantial analytical improvements FT-MS provides for the detection of metabolites, the assignment of FT-MS spectral features to elemental molecular formulas that represent compounds (presumably metabolites) in a biological sample remains a difficult step in the FT-MS data analysis pipeline. Some of the reasons why assignment remains difficult are intrinsic to mass spectrometry. First, unlike NMR, mass spectrometry cannot disambiguate isomers or isotopomers of a compound because they have identical mass. Second, each species of ion produces only one signal in an FT-MS spectrum but have the potential to produce multiple signals in NMR that can help cross-validate assignment. Others are related to how these instruments are commonly deployed and used. The increased popularity of high throughput metabolomics experiments has resulted in a substantial increase in the rate of data collection that has outpaced the rate at which the data can be analyzed. Additionally, the use of FT-MS in a direct infusion environment prevents the acquisition of

orthogonal information from chromatography or spectroscopy to aid in metabolite assignment. Even if these shortcomings with existing FT-MS platforms could be addressed, the sheer number of theoretically possible assignments makes the unbiased assignment of FT-MS spectra difficult. Existing assignment methods are divided into two categories: targeted assignment methods and untargeted assignment methods.

In targeted assignment methods, a set of expected assignments is used as a database against which observed spectral features can be queried. For metabolomics experiments, these databases are usually sets of known metabolites that can be populated from public metabolite databases such as Human Metabolome Database (HMDB) (Wishart *et al.*, 2007), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), LipidMaps Structure Database (LMSD) (Sud *et al.*, 2006), or constructed by hand as is the case with PREMISE (Lorkiewicz *et al.*, 2012). These databases can be tailored for the system of interest and curated to eliminate metabolites that are not expected in the system. Once constructed, a peak can be assigned by querying its observed *m/z* with an *m/z* tolerance against the set of metabolites (correcting for ionization and adduction). These targeted methods have the advantages of being easy to use, computationally cheap and relatively straightforward to implement. Additionally, outside knowledge about the system can be easily incorporated into the assignment process to aid possible assignments. For example, if the sample preparation was selective for lipids, the database of possible assignments needs to include lipids.

The conveniences of targeted assignment come with several significant disadvantages. First, newly discovered metabolites will not reside in existing metabolite databases and all metabolite databases are incomplete (Schrimpe-Rutledge *et al.*, 2016) (Mitchell *et al.*, 2014). As a result, spectral features corresponding to these metabolites will either remain unassigned or worse be misassigned. Second, the use of tailored databases biases assignment towards entries in the database (Moseley, 2013). For example, if it is assumed that a sample consists of only lipids and a lipid-only database is used for assignment, all the generated assignments will be to lipids even if non-lipid compounds were present in the sample. Third, *m/z* matching alone does not provide sufficient information to disambiguate multiple possible assignments and is statistically error-prone due to the lack of cross-validating evidence (Kind and Fiehn, 2006). For example, determining if two peaks approximately 1 Dalton apart are isotopologues of the same compound or two different compounds with *m/z* alone is effectively impossible. All of these limitations become more pronounced when artifactual spectral features are present (Mitchell *et al.*, 2017). For example, artifactual features whose appearance is correlated with sample class or the presence of other metabolites can result in consistent misassignments that can confuse downstream data analyses. Methods for detecting a particularly problematic artifact type that we have discovered with this sample correlation property in FT-MS spectra are described in Chapter 2.

The alternative to targeted assignment is untargeted assignment, where no assumptions or very few assumptions are made regarding the set of possible

assignments. Ideally, no databases of possible metabolites would be used and assignments would be produced using cross-validating information within an MS1 spectrum. Unfortunately, no real untargeted assignment method exists. Many of the tools that are advertised for untargeted assignment are simply targeted tools with very large metabolite databases. The LipidSearch assignment tool is one such example of an “untargeted” assignment method that ultimately relies on a large database of known lipids to generate assignments (Peake *et al.*, 2013). While these large databases can mitigate some of the problems of targeted approaches, they are not a complete solution. Other so-called untargeted tools attempt to calculate an elemental formula from an observed *m/z* directly using heuristics derived from examples of known valid chemical structures. One example of this approach is the “seven golden rules” published by Kind and Fiehn (Kind and Fiehn, 2007). However, these heuristics can only accurately predict the elemental formula of monoisotopic peaks, making them unsuitable for the assignment of isotopologues expected in mSIRM experiments. At high molecular weight, the number of possible elemental formulas remains very large even with these heuristic rules (Watson, 2013).

These compromises in existing untargeted tools exist as a work around for the extremely high computational cost of considering all possible formula assignments for each peak in a spectrum. The number of possible elemental molecular formulas increases exponentially with increasing mass and the number of possible isotope-resolved molecular formulas (IMFs) that must be considered in SIRM experiments is even larger. Simple brute force testing of all possible

formulas against every peak cannot possibly be done efficiently and simple *m/z* matching against such a large set of possible assignments will have an extremely high false positive rate.

However, improvements in the combinatorial mathematics underlying natural abundance correction (Moseley, 2010) (Carreer *et al.*, 2013) enables the calculation of a natural abundance probability (NAP) for any given IMF, even when isotopic labeling is expected. IMFs with low NAPs are unlikely to be observed in an experiment and thus this NAP value can be used to avoid a brute force enumeration of *all* IMFs by allowing a smarter enumeration of only the likely IMFs for an experiment, which is a much smaller set than the set of all IMFs. Furthermore, comparisons between relative peak intensities of assigned peaks and the NAPs of their IMF assignments provides a possible mechanism to reduce false assignment. How NAP can be calculated in both labeled and unlabeled cases and how it can be applied to achieve untargeted IMF assignment is described in detail in Chapter 3.

1.7 Applications of Metabolomics

Although existing methods for processing and assigning metabolomics data have significant deficiencies, metabolomics techniques have already been successfully applied to several human diseases and conditions, improving our understanding of altered metabolism for many biological systems with implications for human health.

One common application of metabolomics is the identification of disease biomarkers for diagnostic or prognostic purposes. By comparing observed metabolomes of healthy and disease samples and identifying overly abundant metabolites present only in disease, or preferentially in disease, robust biomarkers can be identified, typically using a variety of statistical and computational techniques ranging from simple ANOVA (Farshidfar *et al.*, 2018), LASSO (Chan *et al.*, 2015) to machine learning methods (Chen *et al.*, 2013). Biomarkers for a number of cancers have been identified using these techniques: high serum and urinary bile acids, histidine, and inosine levels are indicative of hepatocellular carcinoma (Chen *et al.*, 2011); high urinary acylcarnitines (Ganti *et al.*, 2012), quinolate, 4—hydroxybenzoate, and gentisate (Kim *et al.*, 2011) are indicative of kidney cancer; high serum 2-hydroxybutyrate, aspartate, kynurenine, and cystamine are indicative of colorectal cancer (Nishiumi *et al.*, 2012); and high cell or tissue levels of choline derivatives are indicative of cancer (Glunde *et al.*, 2004) (Beloueche-Babari *et al.*, 2009). Sometimes the change in the concentration of a single or handful of metabolites are insufficient for disease diagnosis. In these cases, machine learning enables the use of many measured metabolites to collectively act as a biomarker signature for a disease (Chen *et al.*, 2013; Zhou *et al.*, 2010). The ability to potentially diagnose clinically asymptomatic cancer patients prior to their progression to advanced disease has the potential to save many lives. Additionally, biomarkers can also inform physicians as to the aggressiveness (Giskeødegård *et al.*, 2013) of a patient's cancer. The use of metabolomics to identify potential biomarkers or metabolic

signatures of disease is not limited to cancer. Differential immune response to influenza with respect to obesity (Milner *et al.*, 2014), markers of cardiovascular disease (Montgomery and Brown, 2013) (Garg, 2011), and markers of air pollution exposure (Jiang *et al.*, 2016) are just a few examples.

In the previous examples, the observed biomarkers were assigned to metabolites or small molecules; however, this level of assignment is not necessary in all cases. Any reliably observed spectral feature that correlates with sample class can serve as a biomarker (Kind *et al.*, 2007). Working at the spectral feature level sidesteps the difficult assignment problem, but ultimately limits what can be inferred about the system. For example, a collection of samples from glutamine-addicted tumors will have a consistent spectral feature corresponding to glutamine; however, inferring that the presence of that signal corresponds to a specific metabolic alteration that could be targeted with a drug or indicative of c-Myc activation is effectively impossible without accurate assignment.

In addition to biomarker discovery, the combination of stable isotope tracing with metabolomics is greatly enhancing our ability to construct accurate, quantitative models of cellular metabolism in multiple ways. First, the combination of labeling and metabolomics enables the disambiguation of multiple pools of a given metabolite in different cellular compartments (Fan *et al.*, 2012). For example, mitochondrial and cytosolic pools of NADPH are compartmentalized but undergo indistinguishable reactions without labeling (Lewis *et al.*, 2014). Distinguishing metabolites pools is particularly important as

the same metabolite in different compartments may have very different metabolic roles.. For example, cytosolic acetyl-CoA is available for lipid biosynthesis, but not mitochondrial acetyl-CoA, despite having the same chemical structure. Similarly, while both pools of acetyl-CoA could be used for protein acetylation, this compartmentalization limits which pool or acetyl-CoA is available to acetylate which proteins (Anderson and Hirschey, 2012). Second, labeling enables metabolites with identical structures, but with different origins to be differentiated. For example, Myc activation results in both enhanced glycolysis and glutaminolysis (Goetzman and Prochownik, 2018), which can result in lactate or TCA cycle intermediates respectively.. Inferring that the primary product of glucose metabolism is lactate and that glutamine is metabolized to create TCA cycle intermediates requires differentiating glucose carbons from glutamine carbons, which can be achieved using ^{13}C labeling (Le *et al.*, 2012). Presence of labeled lactate only when glucose is labeled would imply that glycolysis is the primary source of carbon for lactate biosynthesis, while the presence of labeled TCA intermediates only when glutamine is labeled would imply significant glutaminolysis activity.

In these cases, simply observing a high abundance of isotope label, irrespective of where it is located in the compound is sufficient to differentiate the metabolic fates of glutamine and glucose. However, more powerful inferences can be made when label distributions are studied more closely. Many metabolic pathways can be measured quantitatively using one or more labeled feed metabolites along with observing the concentrations of labeled versions of one or

more readout metabolites (Jang *et al.*, 2018). For example, pyruvate carboxylase activity can be measured using [3-^{13}C glucose] or [U-^{13}C glucose] and measuring m+1 aspartate and m+1 malate levels. This technique was used to identify the necessity of pyruvate carboxylase for non-small cell lung cancer development (Sellers *et al.*, 2015a). Likewise, purine and pyrimidine biosynthesis can be tracked using [1-^{13}C]-bicarbonate, [U-^{15}N]-glutamine, and [U-^{13}C] glutamine by measuring the ^{13}C incorporation into UTP and UDP-Glucose (Strong *et al.*, 1983); protein synthesis can be tracked using $^{2}H_2O$ and [U-^{13}C] amino acids by measuring protein amino acid enrichment in proteins (Busch *et al.*, 2006); etc. Additionally, observed labeling patterns can aid in the discovery of new metabolic pathways and the placement of newly discovered metabolites (Creek *et al.*, 2012) (Higashi *et al.*, 2014).

In many cases, the relative enrichment of various isotopologues or isotopomers at isotopic steady state is enough to infer the relative activities of pathways. However, isotopic labeling can also be used to measure relative flux not at isotopic steady state by observing how quickly pools of metabolites become labeled when labeled precursor is introduced to a system. The higher the relative flux, the faster label accumulates in a pool. While computationally and experimentally challenging, the parameters learned from these experiments can aid in the production of truly quantitative models of cellular metabolism (Leighty and Antoniewicz, 2011). Biochemical models complete with flux parameters (either from dynamic flux analysis or steady state flux analysis) can aid in the identification of drug targets. Enzymatic steps with enhanced relative

flux in disease versus non-disease are prime drug targets (Boros *et al.*, 2004) and changes in metabolic flux through those pathways can be used to quantify drug response (Harris *et al.*, 2012).

Despite these potential advantages, effectively all these analyses require high quality data and trustworthy assignments. Diagnostic biomarkers that are not assigned must still be reproducible and validated, which in turn requires high quality data that is free of obvious data quality problems and artifacts. Confusing spectral artifacts with diagnostic spectral features could at best result in poor diagnostic accuracy or poor reproducibility and, at worst, lead to misdiagnosis. Biomarkers without high quality assignments are ultimately limited in what they tell us about the biological system they represent and more complicated analyses such as metabolic flux modeling or the inference of enzyme activities from labeling patterns requires accurate assignments. Although data quality and assignment remain significant unsolved problems for the metabolomics field, they are particularly pronounced for experiments utilizing direct infusion FT-MS. Direct infusion limits which orthogonal information can be acquired, which in turn makes assignment more difficult. Furthermore, the use of the Fourier transform can result in spectral artifacts in FT-MS (Miladinović *et al.*, 2012) . Addressing these two significant unsolved problems and applying them to improve our understanding of non-small cell lung cancer (NSCLC) metabolism is the main goal of this dissertation.

1.8 Overview of Dissertation

Although each of the different software tools described in this dissertation are stand-alone tools that can be used in a variety of metabolomics experiments, many of them were developed with the goal of analyzing a dataset collected by our collaborators in the Center for Environmental and Systems Biochemistry (CESB). This dataset was comprised of paired disease and non-disease lipid extracts collected from suspected NSCLC patients. The goal was to assign the lipids present in these samples in an untargeted manner and then investigate how the lipid profiles differ between the disease and non-disease samples.

However, this analysis quickly proved to be more complicated than first imagined. First, several significant data quality problems were discovered that complicated the analysis of this data. Some of these data quality problems and how they were addressed is described in Chapter 2. Next, no suitable untargeted assignment tools existed for assigning this dataset. Several years were spent attempting to develop an untargeted method but only until the data quality problems were addressed was SMIRFE able to be developed and assign EMFs and IMFs to spectra generated from these samples. Chapter 3 describes the SMIRFE algorithm and how it was validated.

Although SMIRFE could generate assignments in an untargeted manner, it could only assign EMFs and IMFs to spectra. To explore lipid profile differences, it was necessary to predict lipid categories from these EMFs and IMFs. This was achieved using the machine learning approaches described in Chapter 4. Finally,

with the data quality problems addressed, untargeted assignments generated and lipid categories for these assignments predicted, lipid profile comparisons could be made using the paired NSCLC dataset. The results from this analysis are shown in Chapter 5.

Chapter 6 is the only research product not used directly in the analysis of the NSCLC dataset. In Chapter 6, the potential efficacy of a technique called chemoselective-derivatization is explored *in silico* for disambiguating MS-based assignment of isomeric metabolites. Additionally, the algorithms described in Chapter 6 have proven to have other applications outside of improved MS assignment.

CHAPTER 2. FT-MS DATA QUALITY PROBLEMS AND ARTIFACTS

2.1 Introduction

Due to its potential to provide simultaneous improvements in sensitivity, mass resolution and mass accuracy, Fourier Transform Mass Spectrometry (FT-MS) has the combined analytical capabilities to enable increasingly more complex and informative experiments in the field of metabolomics. In particular, FT-MS is ideally suited for the discovery of new, low concentration metabolites (Dettmer *et al.*, 2007a) and the disambiguation of differently isotopically enriched isotopologues (Higashi *et al.*, 2014). Despite these advantages, several data quality problems have limited the effective application of FT-MS for metabolite discovery and untargeted metabolite assignment. At the forefront of these issues is the presence of spectral artifacts.

Like any other analytical technique, FT-MS can produce artifactual signals that are due to instrumental or data processing limitations. These artifactual features obviously do not represent the underlying biochemistry of a sample and at best complicate data interpretation and at worst lead to incorrect interpretations. These effects are particularly pronounced in untargeted experiments where the set of expected metabolites are not known *a priori;* greatly increasing the likelihood that an artifactual feature can be mistaken for a metabolite feature. As untargeted experiments become increasingly larger and more popular in the field of metabolomics (Goodacre *et al.*, 2004), the ability to distinguish true sample-specific signals from artifactual signals becomes

increasingly more desirable as many existing methods for assigning mass spectra will readily assign both artifactual and non-artifactual features alike (Mahieu and Patti, 2017).

Artifactual peaks observed in FT-MS can be divided into two classes based upon their origin. The first of these is chemical noise resulting from actual but unintended ions introduced during data acquisition. This includes contaminant compounds such as plasticizers and keratin (Keller *et al.*, 2008) as well as ionization byproducts such as molecularly rearranged lipids (McLafferty, 1959). For chemical noise, the artifactual signals represents actual populations of ions observed in the ion trap but these populations of ions are not directly representative of sample analytes. These artifact types are not unique to FT-MS but can occur with any mass spectrometer.

The second class consists of artifactual peaks that do not correspond to actual ions. Often these artifacts arise when the mathematics used to process the raw data from the instrument into a spectrum make incorrect assumptions about that data. For example, the Fourier transform, which is used to convert time domain data (e.g. a decaying FID) into frequency domain data (by fitting to a family of decaying sinusoids and cosinusoids whose sum recapitulates the time domain data) assumes that the observed signal in the time-domain decays to zero continuously. A common way in which this assumption is violated is when the time-domain data is truncated due to insufficient acquisition time. Truncation introduces a discontinuity that will result in the generation of many additional sinusoid and cosinusoids terms after transformation that will appear as peaks in

the spectrum. These additional terms are the result of trying to perfectly fit a discontinuous function with the sum of many continuous functions which is not possible. These peaks are artifactual and are centered around a single intense peak and have a characteristic pattern of decreasing peak intensity as the distance between them and the intense peak increases in the frequency domain. This phenomenon is called peak ringing in FT-MS (Miladinović *et al.*, 2012), sinc wiggles in FT-NMR (Hore, 1985), and side lobes in FT-infrared spectroscopy (Griffiths and Pariente, 1986). An example of peak ringing observed during our study is shown in Figure 2.1.

Although preprocessing techniques such as apodization can be used to force the observed data to conform with the mathematical assumptions of the Fourier transform, preprocessing can introduce more severe artifacts if done improperly and often reduce the analytical capabilities of the instrument (Brenna and Creasy, 1989). Previous studies of specific artifacts in FT-MS have focused on harmonic peaks (i.e. artifactual spectral peaks that arise due to frequency mixing of non-artifactual peaks with one another and radio frequency interference (RFI)) (Mathur and O'Connor, 2009) and peak ringing (Miladinović *et al.*, 2012); however, these artifacts are only a subset of the artifacts observed in FT-MS spectra.

Manual investigation of our FT-MS spectra revealed both the presence of peak ringing artifacts as well as two other artifact types that were previously unreported in the literature that we have named fuzzy sites and partial ringing. All three artifacts (peak ringing, fuzzy sites and partial ringing) result in regions of

spectra with high peak density (HPD) and we collectively refer to them as HPD artifacts. Fuzzy sites are HPD artifacts not centered around an intense peak, but have many peaks in a small *m/z* window. Partial ringing is the subclass of HPD artifacts that show a symmetrical distribution of peaks centered around an intense peak like peak ringing but without the descending intensity pattern observed in peak ringing. Examples of both artifacts are shown in Figure 2.1. Fuzzy sites were the most commonly observed artifact across our spectra and were too numerous to manually identify in all samples. Towards this end, we developed an automated tool to identify a subset of FT-MS artifacts that have statistically higher peak density than surrounding spectra for the detection of these fuzzy sites and to a lesser extent other HPD sites.

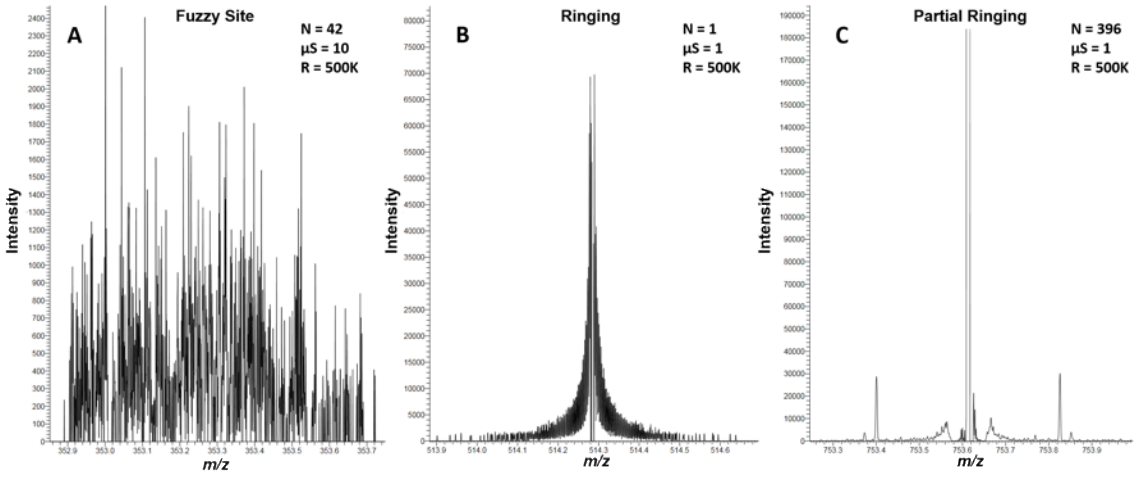

Figure 2.1: Three Types of HPD Artifacts
We observed three subclasses of HPD artifacts. The first is the fuzzy site (A, Sample D), which we believe is a novel artifact type. The second is ringing, a well-known FT-MS artifact where a single intense peak has many side peaks (B, Sample B). We only observed ringing at the scan level. The third artifact is partial ringing which is a ringing-like artifact at the aggregate level (C, Sample A). R is the resolution setting used for data acquisition, μS is the microscan setting which determines how many microscans are per scan, and N is the number of scans aggregated to create the spectrum.

The second major data quality problem facing FT-MS is the high amount of error in observed peak intensities. Peak intensities, measured either by height or by area, should be proportional to the number of ions of that mass-to-charge ratio in the trap when the spectrum was acquired. This by itself is not a problem, except that the reported aggregate spectrum is composed of multiple scans added together, each of which is effectively an entire spectrum corresponding to a separate injection (and each of these scans is composed of microscans). When the scans are replicates averaging multiple scans together increases the signal-to-noise ratio as random noise is “averaged out”. However, due to fluctuations in the electrospray system and inconsistent injections, the composition of ions in each injection can vary between scans and thus each scan is not necessarily a good replicate. For example, if a poor injection only injects 20% as much analyte, the effective concentration of ions in the trap will also be reduced to 20% and the peak intensities will also be scaled a similar amount. While the intensity differences are partially compensated for by the automatic gain control, this issue can still cause some peaks to disappear between scans if the concentration of that ion species falls below the detection limit or the effective dynamic range of the instrument.

These effects result in a large relative standard deviation of intensities for a given peak across all the scans in a spectrum and when non-replicate scans are averaged together, the estimate of the mean peak intensity is skewed towards the most intense scan. This skewing means both the absolute intensities of the peaks and the relative intensities of the peaks have substantial error. An

improved peak picking and characterization method that is aware of these scan-level variations and corrects for them has the potential to mitigate many of these effects and improve the relative intensity of the peaks. Once corrected, the relative intensities between sets of isotopologues can be used to aid metabolite assignment.

2.2 Materials and Methods

2.2.1 *High Peak Density Artifact Detection*

We observed three unique artifact types that share the high peak density (HPD) property in otherwise peak-sparse spectra: fuzzy sites, ringing and partial ringing. Since fuzzy sites were present in most of our spectra, we primarily developed a tool for fuzzy site detection based on this HPD property using the Python programming language (Van Rossum and Drake Jr, 1995) version 3.4 and Numpy (Walt *et al.*, 2011) for accelerating calculations. Starting with a peaklist in a Javascript Object Notation (JSON) format, the detector first parses and sorts the peaks in ascending order of their *m/z* values, needed for binary searching (an efficient mechanism for finding values in a sorted list close to a query value) of the peaklist. A 1 *m/z* window (top black box in Figure 2.2) is then slid across the spectrum in 0.1 *m/z* increments. At each increment, two binary searches are used to find all peaks within the window that are then counted to give a peak density metric (Step 1 in Figure 2.2 that is assigned to the middle *m/z* of the window.

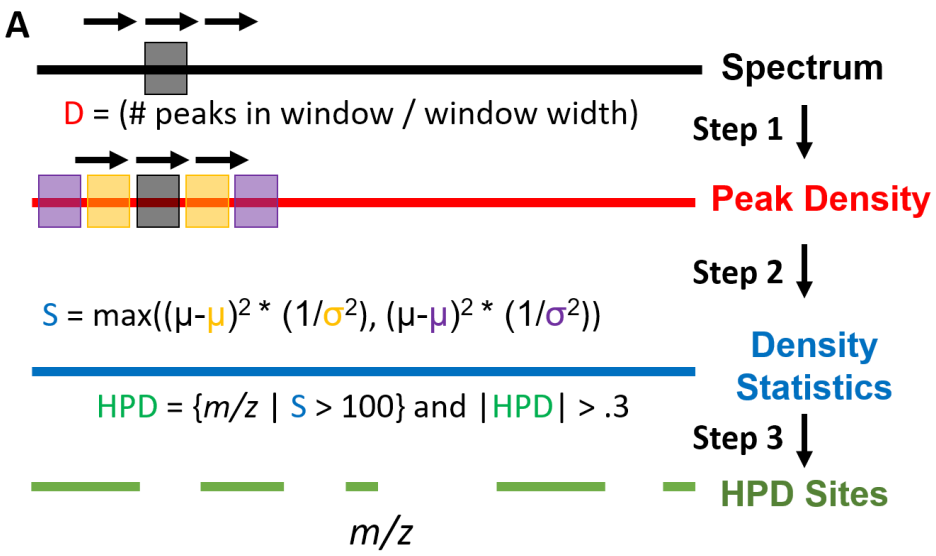

Figure 2.2: Automated HPD-Site Detection
The HPD artifact detection algorithm in three steps. In the first step, a peak density metric (D) is calculated for the peaklist (represented by the black line) using a sliding window method (1 *m/z* window, 0.1 *m/z* increment, represented in black). The peak density for each window is calculated and assigned to the mean *m/z* of the window to give a series of peak density values throughout the peaklist (represented by the red line). In the second step, a set of N+1 windows are slid across the peak density metrics in 0.1 *m/z* increments. In this case N=2 and each pair of "reference" windows (shown in purple and yellow) flank the "test" window (shown in black). For each increment, the peak density statistic value S is calculated. S is calculated using the mean and variance of the peak density in the reference window (black μ and σ respectively) and the corresponding values from the reference windows (yellow μ and σ for the yellow window and purple μ and σ for the purple window). The mean of *m/z* of the test window is then assigned to S to give a series of density statistic values throughout the peaklist represented in blue. This metric flattens out density differences due to signal-to-noise differences or baseline differences and highlights regions of peaklists containing HPD artifacts (Figure 2.3E-H). In the third step, regions of spectra at least 0.3 *m/z* in width with a density statistic value above 100 are reported as possible HPD artifacts.

HPD artifacts can be found through the comparison of the actual peak density at a given *m/z* to the expected peak density derived from surrounding regions (Step 2). In this operation, N pairs of non-overlapping 'reference'

windows (yellow and purple boxes in the Peak Density line of Figure 2.2) distributed symmetrically around a single ‘test’ window (center black box in the Peak Density line of Figure 2.2) are moved across the spectrum in 0.1 *m/z* increments. The test window is the spectral region being tested for HPD phenomena using the reference windows as estimates of “expected” peak density. At each increment, the mean and standard deviation of the peak density is calculated for each pair of reference windows, with each reference window 3 *m/z* in width. The test window is then assigned a density statistic value S (Figure 2.2), a chi-squared inspired metric (i.e., the peak density normalized with respect to expected peak density and variance generated from surrounding regions). By taking the maximum value of S, sensitivity is maximized, enabling detection of HPD even if one pair of reference windows contains HPD. Although higher values of N are theoretically superior, testing demonstrated no significant improvement for N > 2. At the ends of the spectra, only the left and right reference windows are used.

In the final step, the continuous subdomains of *m/z* space at least 0.3 *m/z* in width (smaller than empirically observed HPD artifacts) and with density statistic values over 100 are reported (Step 3 in Figure 2.2). These reported regions very likely contain some form of HPD phenomena as they have significantly higher peak densities as compared to neighboring regions.

2.2.2 Peak Correspondence and Peak Characterization Algorithm Description

The following peak correspondence and peak characterization algorithm was developed by Dr. Robert Flight, a staff scientist in Dr. Hunter Moseley's laboratory at the University of Kentucky (Flight and Moseley, 2018).

Raw Thermo data files were converted to mzML using Proteowizard v 3.0.9205X.X (Chambers *et al.*, 2012) with no modifications to the underlying data. For each sample, a custom data processing pipeline was employed to characterize the peaks detected across scans. mzML data were read into R v 3.5.0 (Team, 2013) using the xcms Bioconductor package v 3.2.0 (Smith *et al.*, 2006), keeping only the primary MS scans. Based on the non-zero *m/z* point - point distances, a model relating *m/z* to *m/z* differences is created using LOESS (locally estimated scatter plot smoothing). Each scan is evaluated for aberrations in this *m/z* difference model as compared to the others, with outlier scans being removed.

The spectrum is divided into overlapping windowed regions of 10 points, where the *m/z* distance is based on the *m/z* distance model and each window is offset by 1 point. The non-zero points within each window are counted and those regions in the 99th percentile are kept and joined together. Within each of these reduced regions, peaks from each scan are detected by simply looking for patterns of two increasing points followed by two decreasing points using the findpeaks function from version 2.1.4 of the PRACMA (Practical Numerical Math Functions) R package (Borchers and Borchers, 2018). Each region may initially contain multiple peaks that need to be split up. Each peak center from peak

picking is binned into tiled regions one point apart, and the number of peaks in each bin counted. Groups of non-zero bins separated by zero count bins (bins with no peaks observed) are used as the true individual “peak regions” for subsequent peak characterization.

A two-pass normalization scheme is used to normalize the scan-specific peak intensities. In the first pass, those peaks present in >= 95th percentile of total number of scans, and >= 70% of the maximum intensity of peaks in a given scan are used to normalize the scans to each other. The scan with the lowest relative median distance to the other scans based on log-intensity differences is used as the reference scan. Then both the peak intensities and raw point intensities are normalized to the reference scan using the median log-ratio differences. Those peaks that show a high correlation (>= 0.5) with scan acquisition order are removed and the normalization scheme is repeated without the scan-correlated peaks. Those scans that do not have at least 25 peaks for normalization are also removed during the normalization procedure.

For each of the peak regions, the full set of normalized raw points are characterized by fitting a quadratic linear model of log-intensity to *m/z*. The peak centroid *m/z* and log-intensity are obtained directly from the linear model. These characteristics are reported for both the scan level peaks and the aggregate peak across scans. Standard deviations of the *m/z* and intensities are reported based on the scan level peak characteristics.

Due to unobserved peaks at relatively low intensities, the standard deviations are corrected in the following fashion based on a truncated normal

distribution, where the distribution is truncated on only one side assuming that measurements are missing because the peaks were below the detection limit (Burkardt, 2014) . A correction factor for the standard deviation is derived from peaks observed in all scans by randomly dropping observations from 5% to 95% of scans, calculating new standard deviations, and fitting a cubic linear model to the ratio of the intensity SDs with the omitted peaks to the original standard deviation. These correction factors can then be applied to the observed standard deviations. The corrected standard deviations can then be used to correct the observed mean, also assuming a truncated normal distribution.

Our new scan-level peak characterization methods result in more descriptive peaklists and peaks. The peak descriptions produced by this method provide direct measurements of peak *m/z*'s and intensities at the aggregate and scan level, the standard deviation and relative standard deviations of these values across scans, and the scans that contain a given peak from the aggregate spectrum. Additionally, reporting only well corresponded peaks across scans greatly reduces the number of artifactual peaks from fuzzy sites and other artifacts and greatly reduces the number of peaks due to noise in the final spectrum. Finally, the scan normalization procedure compensates for fluctuations in the ionization source or injection that can cause peak intensities of corresponded peaks to vary greatly between scans. The resulting normalized peak intensities have lower RSDs than the unnormalized intensities, implying better consistency across scans. An example peaklist generated by this approach is included in Additional Files.

2.2.3 Samples Analyzed by FT-MS

Five different sets of samples were used to investigate HPD artifacts on our FT-MS platforms. Sample set A comprises solvent blanks with and without the Avanti SPLASH™ Lipidomix® Mass Spec Standard added. Sample B contains IC-MS standards prepared from NSG mice livers. Sample C contains ethylchloroformate (ECF) solvent standards. Sample D are paired lipid extracts from human non-small cell lung cancer and non-cancer lung tissue samples. Sample E consisted of human plasma samples. More details regarding these samples, including who collected them, who prepared them, and who analyzed them, are described in greater detail in Appendix 1.

2.2.4 FT-MS Instruments

Several FT-MS instruments were analyzed to determine the instrument dependence of the HPD artifacts. These instruments include three Thermo Tribrid Fusion instruments (named Fusion 1, 2 and 3), a Thermo Fusion Lumos Tribrid (Lumos), a Thermo Q-Exactive+ and a Bruker SolariX instrument. Except for the SolariX, all instruments are orbitraps. Additional instrument details are described in Appendix 1 for some samples.

2.3 Results

2.3.1 Manual Investigation of Artifacts

During FT-MS metabolomics data analysis, we manually inspected several hundred spectra and observed fuzzy sites, peak ringing, and partial

ringing. Fuzzy sites were the most common artifact and were present in nearly all spectra from our Fusion instruments. Partial ringing was more rare than fuzzy sites and was present in only a handful of the spectra examined. Ringing was the rarest of all three artifacts and were only observed in two spectra.

Ringing is a known artifact that has been reported previously in the literature (Miladinović *et al.*, 2012). Fuzzy sites and partial ringing were also identified as likely artifacts. Manual inspection of a selection of fuzzy sites and partial ringing regions consistently failed to identify peak patterns that are explainable by chemical phenomena (*e.g.* isotopologues, different charges, etc.). Additionally, these regions display poor peak correspondence at the scan-level, which also argues against a chemical contaminant explanation of these regions. Furthermore, the location of these artifacts was observed to differ between analytical samples derived from the same biological sample and were observed in blanks and samples with poor injections, which contain little analyte and solvent intensity respectively.

2.3.2 General HPD Detection Across FT-MS Instruments

Using the HPD artifact detector, we generated plots of peak density for a variety of example spectra across various FT-MS instruments (Figure 2.3). The trends in peak density vary between instruments, samples, and across *m/z.* In general, peak density decreases with increasing *m/z* and a monotonic trend was observed for all instruments (Figure 2.3A-D). This observation is partially explained by differences in signal-to-noise and decreasing digitization with increasing *m/z* due to the nonlinear relationship between the observed frequency

and *m/z*. From these plots, there exists no good cutoff based on raw peak density to identify HPD sites. In contrast, our statistical approach (i.e. peak density statistic) compensates for systematic changes in peak density, revealing regions of significantly higher peak densities with respect to average peak density of neighboring regions (Figure 2.3E-H).

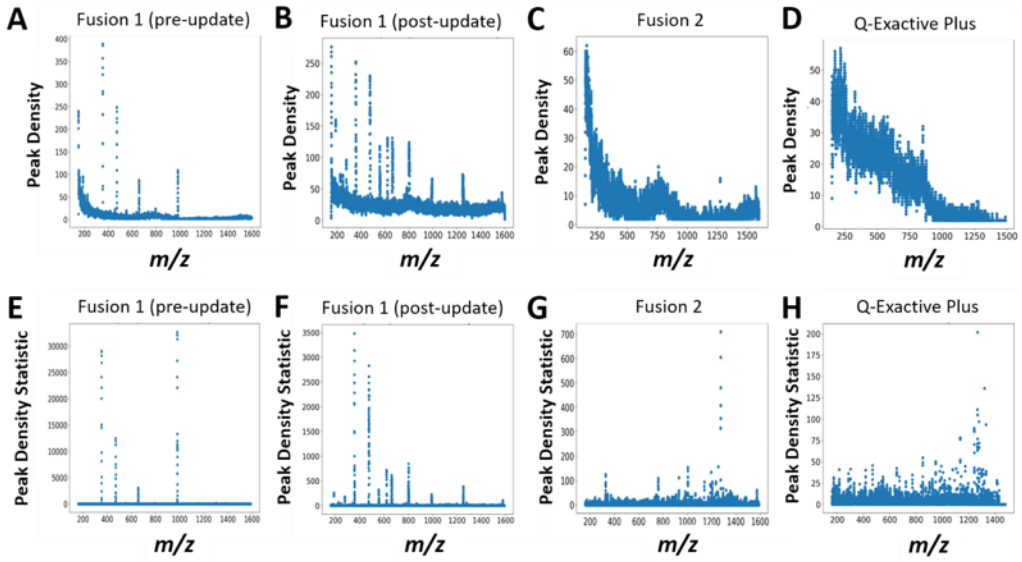

Figure 2.3: Peak Density and Peak Density Statistics.
Peak density metric and statistic plots produced by our HPD-detector tool highlight the impact of the instrument on peak density and HPD artifact location. All instruments have higher peak densities at lower *m/z*, representing trends in signal-to-noise and digitization with respect to *m/z* in FT-MS. The sharp spikes in peak density correspond to HPD artifacts. The locations of these spikes on Fusion 1 are different before and after the firmware update (A, B), suggesting instrument-level data processing is related to HPD generation. E-H show the effectiveness of our peak density statistic metric for flattening the non-constant baseline observed in plots of the raw peak density. Without this correction, identifying HPD regions reliably is difficult. A, B, C, E, F, G were generated from spectra acquired using sample C. D and H were generated from spectra acquired using sample E (Appendix 1).

Although fluctuations in peak densities are expected due to differences in the distribution of compounds in *m/z* space, this fails to explain the large spot

increases (spikes) in the peak density statistic present in the spectra. These spikes in peak density represent HPD artifacts and thus by identifying these spikes, potential HPD artifacts can be detected. Additionally, these results demonstrate that the artifact locations differ before and after a firmware update on the same Fusion 1 machine (Figures 2.3A, 2.3B, 2.3E, 2.3F) and differ significantly between instruments. Together, these findings support an artifactual basis for these HPD regions of spectra and suggest an instrument-level source.

2.3.3 Detection and Characterization of Fuzzy Sites

At the aggregate spectrum (sum of scans) level (Figure 2.1A, 2.5A, 2.5C), fuzzy sites have HPD characteristics and a Gaussian-like distribution of peak intensities between the noise baseline and presumed signal peaks. The intermediate intensities of these peaks make identifying and filtering these regions by intensity alone difficult. Fuzzy sites, like other HPD artifacts, have peak *m/z* differences that are not explainable by isotopologue, charge, or harmonic patterns. Fuzzy sites vary in size from 0.5 *m/z* to up to 3 *m/z,* with larger fuzzy sites found at higher *m/z*. Typically, a spectrum with fuzzy sites contains multiple fuzzy sites. Collectively, these sites can represent a significant portion of the total peaks over a much smaller portion of the total *m/z* range. Fuzzy site location varies between analytical replicates on the same instrument and with sample composition (Figure 2.4). Fuzzy sites have been observed in samples with a failed or no injection as well.

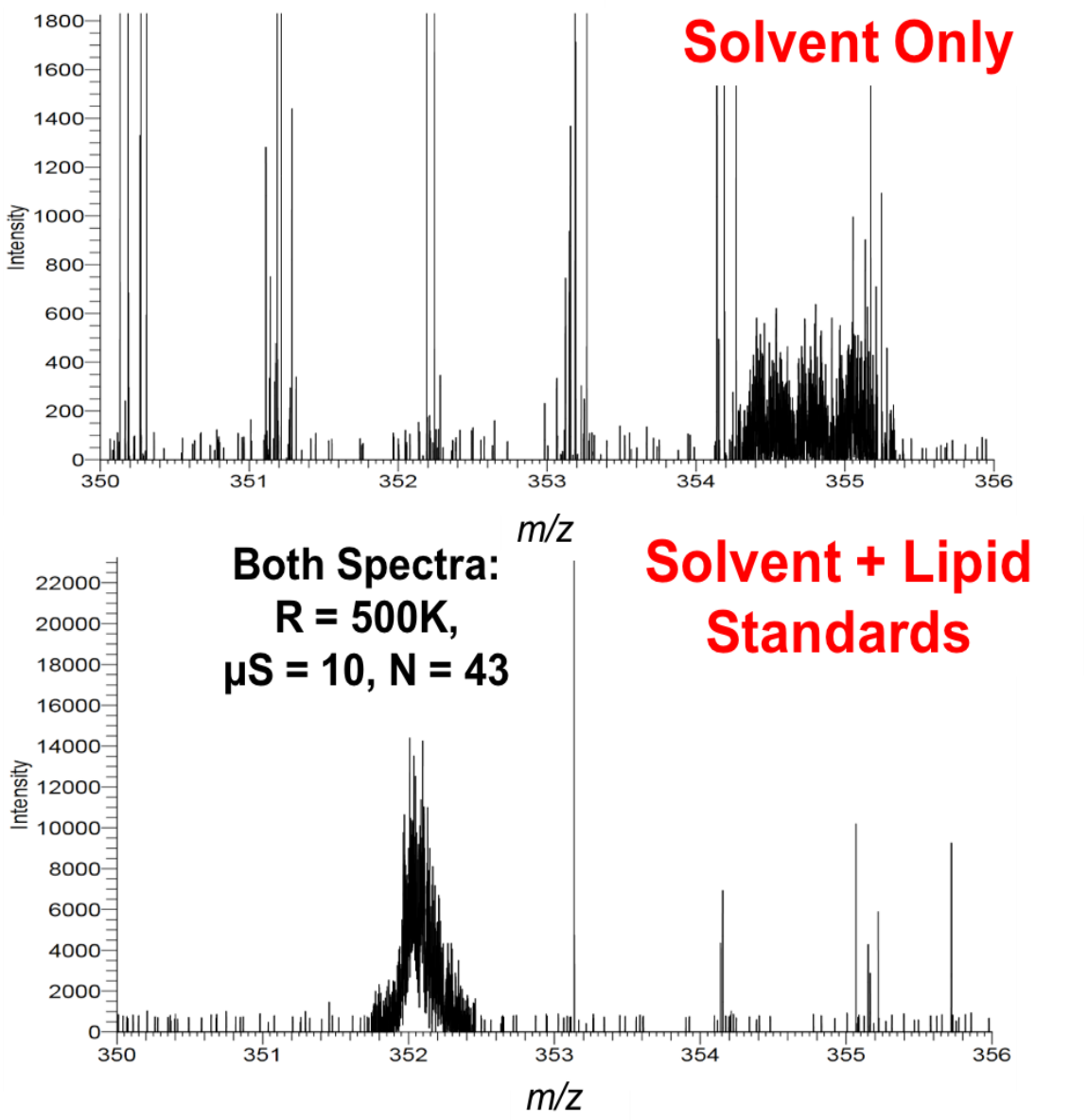

Figure 2.4: Fuzzy Site Location varies with Sample Composition
A small change in chemical composition (Sample A with and without lipid standards) changes fuzzy site location. With only solvent, there is a fuzzy site at 354.8 *m/z*. With lipid standards, this fuzzy site shifts to 352.1 *m/z*. The number of fuzzy sites will remain constant, but will all be shifted by roughly the same *m/z*. R is the resolution setting used for the acquisition, μS is the microscan setting, and N is the number of scans aggregated to create the spectrum.

Fuzzy sites also have interesting properties at the scan-level. While the timing between scans and injection as well as inconsistencies between injections can result in non-perfect scan-to-scan correspondence between peaks (e.g. a peak is present in scan X, but not in scan Y), peaks of the same chemical origin should appear consistently between scans near their true *m/z*, roughly within the resolution of the instrument. However, at the scan level, the peaks in a mass range identified as a fuzzy site at the aggregate level have very low peak correspondence (Figure 2.5B). In any given scan, only sections of the fuzzy site region will have peaks and those sections that are populated with peaks change from scan to scan. However, as increasingly more scans are added together, the Gaussian-like distribution of a fuzzy site at the aggregate level becomes clearer (Figure 2.5A, 2.5C). Furthermore, resolution does not change the presence of fuzzy sites (Figure 2.6). Fuzzy sites appear distinct from either peak ringing or partial peak ringing and represent a novel class of artifact not previously described in the FT-MS literature.

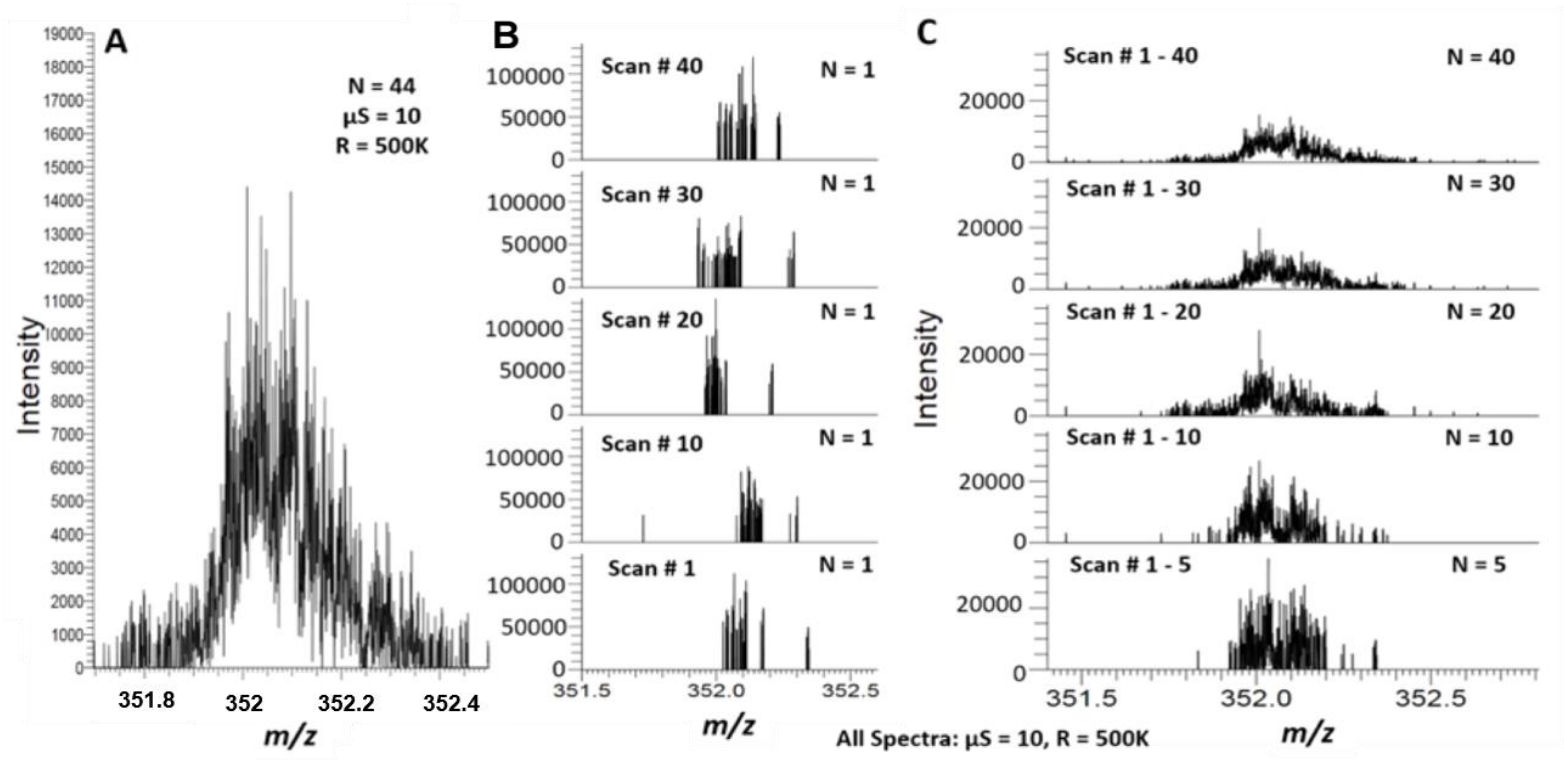

Figure 2.5: Fuzzy sites at the Aggregate and Scan Level
A typical fuzzy site (A) occupies 0.5 to 3 *m/z* at the aggregate level and has a distinct 'fuzzy' appearance due to very high peak density (this image is identical to 2A). At the scan level, only a subdomain of the *m/z* occupied by the fuzzy site contains peaks; the subdomain with peaks varies from scan-to-scan (B). As increasingly more scans are aggregated together, the peak distribution converges to the pattern observed at the aggregate level (C). All panels were generated using Sample A. R is the resolution setting used for the acquisition, µS is the microscan setting, and N is the number of scans aggregated to create the spectrum.

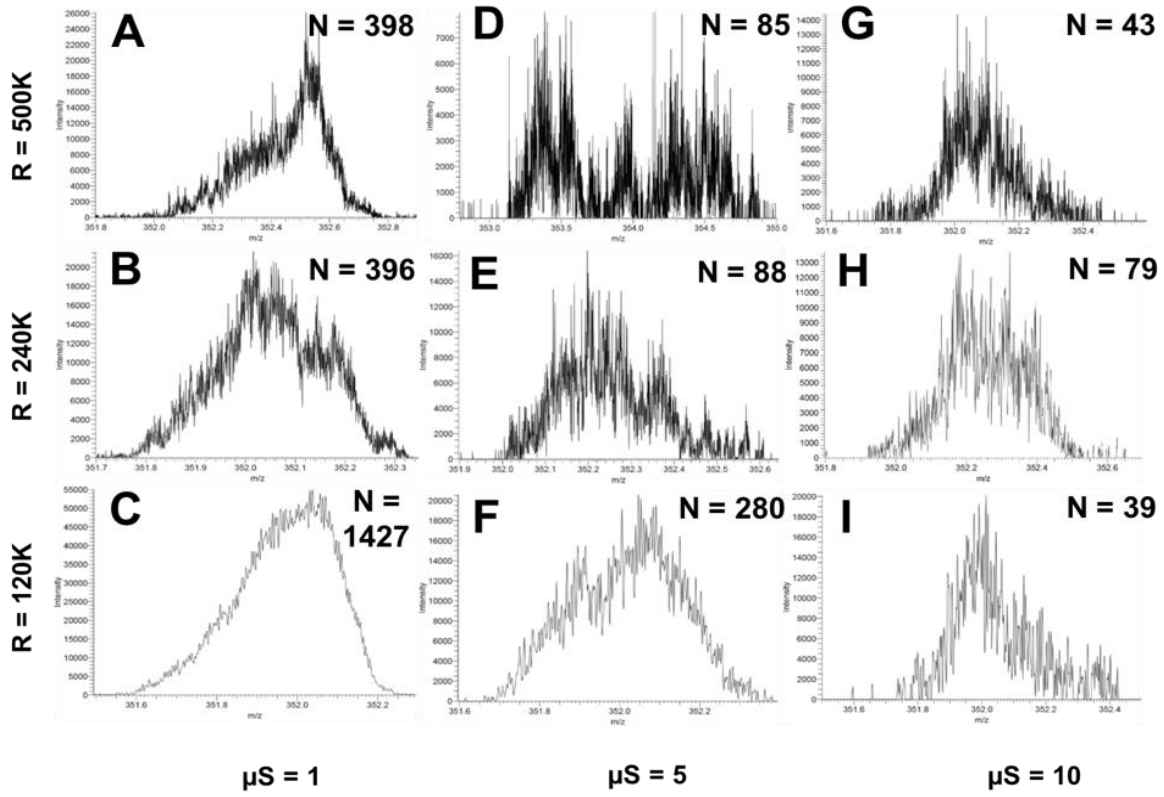

Figure 2.6: Effect of Resolution and Microscan on Fuzzy Site Appearance
Permuting over multiple resolution and μS settings shows that no combination of tested settings eliminated fuzzy sites, but these settings do change their appearance (A-I). Number of scans collected were set so that total acquisition time was constant (7 minutes). Higher μS increases intensity variance with minimal impact on peak density. Increasing resolution increases peak density but has a lesser impact on peak intensity variance. All panels were generated using Sample A on Fusion 1. R is the resolution setting used for the acquisition, μS is the microscan setting, and N is the number of scans aggregated to create the spectrum.

Fuzzy sites were first observed on Fusion 1. After developing our methods using Fusion 1 spectra, spectra from other instruments were examined for fuzzy sites to determine if these artifacts were limited to only one instrument. To date, we have observed fuzzy sites in spectra from every non-Lumos Tribrid Fusion instrument examined. However, we did not find fuzzy sites in spectra from other types of FT-MS instruments (Lumos, Q Exactive+, SolariX).

Using our largest dataset (sample D), we were able to evaluate the robustness of our tool with respect to identifying fuzzy sites in spectra from our Fusion 1 and Fusion 2 instruments. For every region reported by our fuzzy site detector tool, we manually inspected 50 spectra to verify if a fuzzy site was present or not. Additionally, the entire spectrum was completely inspected manually for additional fuzzy sites that were not detected by our tool. We also checked if fuzzy sites were observed for the common fuzzy site locations identified in Fusion 1 (187 *m/z*, 351 *m/z,* 468 *m/z,* 654 *m/z*, 976 *m/z* and 1590 *m/z*) and Fusion 2 (1064 *m/z* and 1275 *m/z*). When an HPD site was predicted but not obvious upon visual inspection, this was counted as a false positive (FP). When an HPD site was not detected in one of the common fuzzy site locations when one was obvious visually, this was counted as a false negative (FN). When an HPD site was predicted in one of the common fuzzy site locations and a fuzzy site was observed, this was counted as a true positive (TP). When an HPD site was not detected in one of the common fuzzy site locations when one was not present, this was counted as a true negative (TN). The detailed results of this analysis are shown in Supplemental Tables 2.1 and 2.2 and across Fusion 1 and

Fusion 2 a sensitivity (#TP / (#TP + #FN)) of 92.1% and a specificity (#TN / (#TN+#FP)) of 33.5% was achieved. This implies that our algorithm tends to overpredict HPD artifacts. However, no HPD sites were consistently missed in all spectra, therefore identifying and removing consistent HPD sites across all samples will side-step the low specificity of our algorithm.

2.3.4 Fuzzy Site Locations are Biological Unit Specific, Class Specific and Instrument Specific

As shown in Figure 2.4, fuzzy site location varies between spectra and appears to shift significantly (a shift far greater than the resolution of the instrument) with changes in sample composition. The dependency of fuzzy site location on sample composition is a potential problem for real metabolomics applications of FT-MS, if fuzzy site features are not eliminated from downstream analyses. To demonstrate this, we compared FT-MS spectra of the lipid samples extracted from paired cancer and non-cancer lung tissue slices (sample D). Due to the differences in the concentrations of various metabolites between cancer and non-cancer tissues, we would anticipate that features assigned to the spectra of different sample classes (i.e. cancer and non-cancer) could be used to distinguish these sample classes. However, if fuzzy sites vary with sample class as well, artifactual features will also distinguish sample class without directly reflecting the underlying biochemical differences between these classes. These samples were ran in a mostly random order with respect to sample class, eliminating the possibility that the artifact patterns are the result of temporal batch effects (Supplemental Table 2.3)

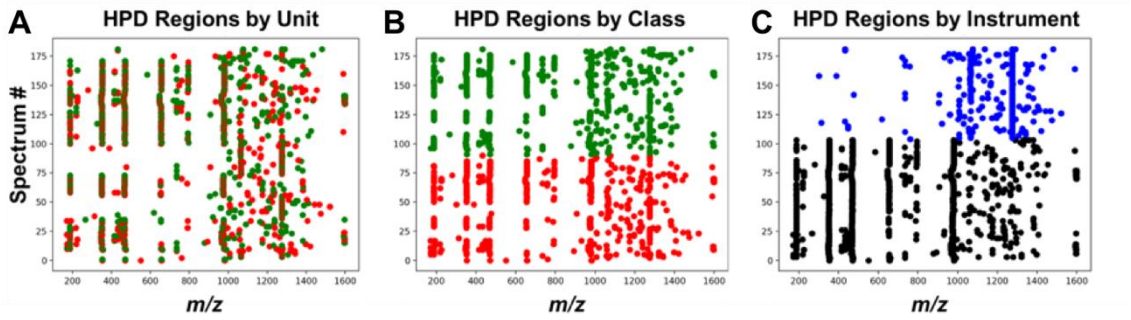

Figure 2.7: HPD Regions Depend on Biological Unit, Sample Class, and Instrument

Using the fuzzy site detector, HPD regions were identified in every spectrum from sample set D. (A) The plot shows that many fuzzy sites are consistent across paired cancer (red) and non-cancer (green) spectra from the same patient. (B) A shifting based on sample class (green versus red) is observed for some consistent fuzzy sites (specific example regions are shown in Figure 2.8). When shifts do occur such as shown in Figure 2.8 between sample classes, the absolute difference between fuzzy sites in *m/z* is too small to be obvious on this plot. (C) Different consistent fuzzy sites are observed for each instrument, Fusion 1 (black) and Fusion 2 (blue). HPD sites differ dramatically between instruments. Across all three plots, the scattering of inconsistent fuzzy sites observed largely at *m/z* > 1000 represent false positive regions. Spectrum number is arbitrary index assigned to each spectrum for each plot.

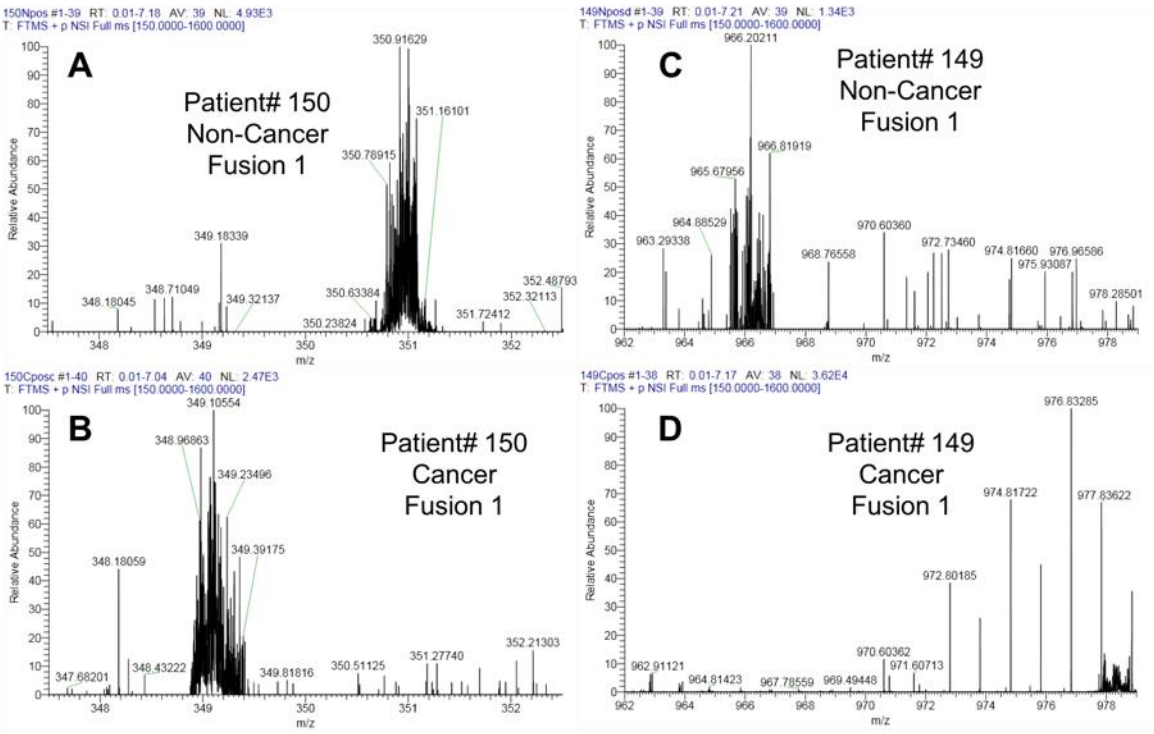

Figure 2.8: Example Fuzzy Site Locations that Vary with Sample Class
The location of fuzzy sites in spectra from the same biological unit can differ significantly based on sample class (cancer versus non-cancer). A and B illustrates one fuzzy site whose location varies by rough 2 *m/z* between sample class. C and D shows a single fuzzy site whose location varies by over 70 *m/z* between sample class.

Figure 2.5 shows the location of the detected HPD regions for every spectrum in sample set D, illustrating a clear relationship between consistent HPD site location and sample origin (biological unit, i.e. patient, Fig 2.7A), sample class (Fig 2.7B) and instrument (Fig 2.7C). HPD regions with less consistency are spurious false positives. Figure 2.8 shows clear shifts in fuzzy site locations within cancer and non-cancer spectra derived from the same patient. In one example, the fuzzy site appears to have shifted by over 12 *m/z* units.

This creates a potential confounding factor or batch effect in all downstream statistical analyses if spurious changes in sample conditions introduce new sample-specific artifacts. For example, machine learning methods such as Random Forest (Breiman, 2001) can use samples of known class to build classifiers that classify samples of unknown class based on spectral features identified in the training set of samples (more about how the Random Forest algorithm builds these classifiers is included in Chapter 4). However, these techniques rely upon an important assumption that detectable artifacts are not confounded with sample class. The large number of sample-specific artifactual peaks produced by fuzzy sites can hijack the classifier training, anchoring the classification to artifacts that may change due to unforeseen sample conditions in unknown samples or unforeseen changes in the analytical instrument.

These hypothetical problems arising from artifact batch effects are clearly demonstrated in Table 2.1 and in Supplemental Tables 2.4, 2.5 and 2.6. Twenty random forests were trained on LipidSearch assignment features derived from spectra covering a mass range of 150 – 1600 *m/z* acquired from non-polar extracts of paired cancer and non-cancer lung tissue samples (sample D). We used only spectra from Fusion 1 to eliminate the instrument as a potential bias and we dropped features present in less than 25% of either the cancer or non-cancer classes. Each classifier was built using the R randomForest package (version 4.6-14) from CRAN. Default package options were used except with the number of trees set at 1000. The top features based on mean importance of

each feature across the 20 random forests are reported. High mean importance means that the classifiers heavily used this feature when predicting if a sample is a cancer or non-cancer sample. Consistent HPD regions are identified first by determining the *m/z* range that contain an HPD site in at least 10% of the samples. Features with an *m/z* in one of the consistent HPD regions are considered HPD-tainted features and are presumed artifacts. If artifact location is not confounded with sample class, then artifactual features should not be present in the importance lists for the classifiers.

Without HPD feature removal, 7 out of the top 30 features are HPD features (Table 2.1 – No Artifact Removal). Removing features observed only in a sample's HPD regions results in a reduction but not complete abolition of HPD-features in the models (Table 2.1 – Per-Spectrum Artifact Removal), since this removal indirectly encodes sample class via the inflation of the importance of features in *m/z* regions that overlapped with HPD sites in the other class. Only when the consistent HPD regions are removed from every sample does no HPD feature makes it into the top 30 mean importance list (Table 2.1 – Consistent Artifact Removal), either directly or indirectly. With sample specific and no artifact removal, classification errors of 13.5% and 3.8% are achieved for cancer and non-cancer respectively and consistent HPD removal achieved 11.5% and 3.8%, representing a minor improvement in this example. The ability to retain classification accuracy with artifact removal implies that we are not significantly removing any important information through our artifact removal process, while increasing the likelihood that the classification is based on true biological

variance. The similarity between the important non-HPD features for all three removal methods supports this hypothesis. The mean importance lists for each type of artifact removal are shown in Supplemental Tables 2.4 (no artifact removal), 2.5 (per-spectrum artifact removal) and 2.6 (consistent artifact removal).

Table 2.1 – Effects of HPD-Artifact Removal

	Number of Artifact Features in Top 30 Important Features	Cancer Samples		Non-Cancer Samples	
		Number Classified as Cancer	Number Classified as Non-Cancer	Number Classified as Non-Cancer	Number Classified as Cancer
No Artifact Removal	7	45 (86.5%)	7 (13.5%)	51 (96.2%)	2 (3.8%)
Per-Spectrum Artifact Removal	3	45 (86.5%)	7 (13.5%)	52 (96.2%)	2 (3.8%)
Consistent Artifact Removal	0	**46 (88.5%)**	**6 (11.5%)**	53 (96.2%)	2 (3.8%)

Table 2.1: Effects of HPD-Artifact Removal - The number of artifactual features used by the models and the classification results are summarized for models trained using various types of artifact removal. Correct assignments are highlighted in the green columns, incorrect in red columns. Without artifact removal 7 out of the top 30 features (by importance) used by the models are artifactual. This implies that the models are using artifactual features to distinguish cancer and non-cancer. With per-spectrum artifact removal, 3 out of the top 30 features are artifactual and have relatively low rank in the importance list compared to artifactual features without artifact removal (Supplemental Tables 2.4 and 2.5). However, no improvement in assignment accuracy is achieved for either cancer or non-cancer samples. With consistent artifact removal, there are no HPD features in the top 30 features (Supplemental Table 2.6) and a slight improvement in accuracy is achieved (one additional cancer sample is correctly identified).

2.3.5 Peak Characterization Significantly Improves Relative Peak Intensities

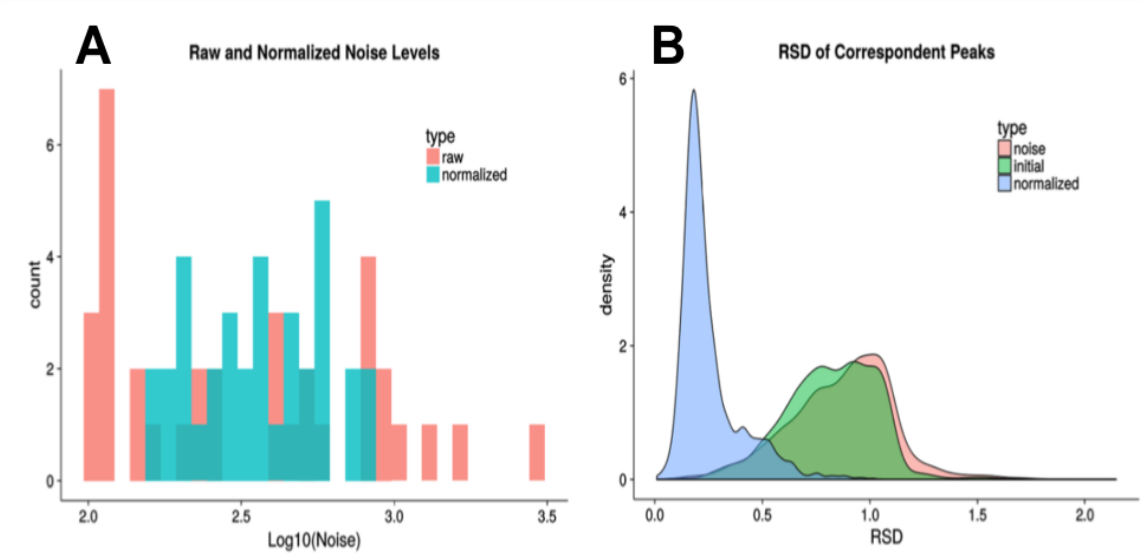

Figure 2.9: The effects of normalization on noise and peak intensity
Without normalization, the estimated noise level across all scans in a spectrum spans nearly two orders of magnitude (Figure 2.9A). This massive variance across scans in the noise ratio results in an intensity RSD for noise peaks of more than 1.0 (Figure 2.9B). The observed RSDs for all peaks prior to normalization is less than the RSD observed for only noise but still unacceptably high. After normalization, the range of observed noise values spans less than 1 order of magnitude and the intensity RSDs for most peaks is greatly reduced (blue trace in Figure 2.9B) and effectively no peaks have an intensity RSD greater than 1. Algorithm, data, and figure generated by Dr. Robert Flight (Flight and Moseley, 2018).

The effects of scan-level variance in peak intensities is shown in Figure 2.9 and both noise and non-noise peaks have their intensities significantly affected. Without normalization, the estimated noise level spans nearly two orders of magnitude and both noise and non-noise peaks have intensity RSDs near 1.0, implying that the standard deviation of peak intensities for each correspondent peak is on the same scale as their observed intensities. After normalization, the observed RSDs for both noise and non-noise peaks are

greatly reduced, as is the range of noise cutoffs observed across the scans. Both findings imply that our peak characterization and normalization methods correct some sources of scan-to-scan variance in peak intensities and thus making each scan more comparable to each other scan. Finally, it was also observed that the most intense scans tend to have the highest noise levels.

2.4 Discussion

2.4.1 *Origin of FT-MS Artifacts*

The scan and aggregate level properties of HPD sites strongly support an artifactual origin for HPD features; however, their exact origin has not been confirmed.

The detectors used in FT-MS instruments are highly sensitive, which introduces the possibility that radio frequency interference (RFI) can result in signals that are misinterpreted by the instrument to produce artifactual peaks in a spectrum. However, the correlation between fuzzy site location and sample class is surprising. The change in fuzzy site location after a firmware update and the presence of fuzzy sites in every Fusion instrument we have tested could indicate an internal source of the RFI. An internal source could produce class-specific fuzzy sites if the state of the components producing the RFI changes in some manner with sample composition. However, a more likely hypothesis is that the source of the RFI is the nanoelectrospray system (or other external device that is part of the deployed FT-MS system). The nanoelectrospray system is in direct contact with the sample and operates at high voltage as part of its normal

operation. Any interaction between the voltage of the electrospray system, the pulses used to ionize the sample and the sample itself could result in the generation of RFI signals that result in sample-specific fuzzy sites. Unfortunately, robustly investigating these hypotheses is difficult and ultimately not relevant to fixing these artifacts in already acquired data. Discussions with Thermo staff support an RFI origin for fuzzy sites, although the source of the interference remains unclear.

The origin of partial ringing is even less clear. Since partial ringing does not occur in most spectra and across multiple instruments, if it is an RFI-based phenomenon, the source of that RFI is likely transient and external. Due to its rarity and with no obvious relationship to sample class, partial ringing is less of a burden to data analysis. Likewise, rarity of ringing largely mitigates its harmful effect on data analysis. ThermoFisher Scientific had no explanation for the partial ringing artifacts, but did confirm that ringing was due to insufficient digitization of the FIDs due to a combination of hardware and software limitations regarding 32-bit versus 64-bit encoding that results in a truncation of the FID. (personal communication with Mike Senko, ThermoFisher Scientific).

Eliminating the source of these artifacts, particularly fuzzy sites, is challenging. First, the radio environment that an instrument observes constantly changes and is often beyond the control of the instrument's operator. Second, these artifacts may not be recognized until after spectral acquisition and the original samples are no longer available due to either consumption or degradation. In this case, the only option is to clean up the acquired data before

downstream analyses. Third, access to the raw FID data is not always available for a given instrument. In this case, the only option is to remove the artifact peaks in the *m/z* domain.

LC-MS and other approaches for gathering orthogonal information about features observed in MS offer an approach to mitigate the effects of these artifacts, but are obviously not applicable to direct injection FT-MS experiments. Retention times combined with *m/z* can often unambiguously identify metabolites where *m/z* cannot and artifactual features that are ambiguous by *m/z* alone will lack supporting chromatographic information. As a result, LC-MS is less vulnerable to the artifacts, but is not always suitable for the same set of experiments that direct injection FT-MS can be applied to.

2.4.2 *Mitigating the Effects of Fuzzy Sites on Downstream Data Analyses*

As shown in Table 2.1, HPD artifacts, namely fuzzy sites, impact data analysis. The resulting classifiers can effectively predict sample class; however, these classifiers make extensive use of peaks present in fuzzy sites and thus have no direct molecular interpretation. Although LipidSearch can use orthogonal information (chromatographic retention time, MS2), which presumably could prevent assignment to artifactual peaks, this information is only available for a small fraction of peaks (typically less than 10%) in the direct-infusion FT-MS spectra collected.

Building robust classifiers based on true biological variance therefore requires eliminating fuzzy site assignments from the feature list prior to classifier

training to mitigate any fuzzy site / sample class confounds. Removing consistent HPD regions from all spectra safely removes artifactual features efficiently and consistently. Additionally, a check on the consistency of the HPD sites minimizes the impact of false positives returned by our tool and effectively increases the sensitivity of our method. Since we could not find a fuzzy site that was consistently missed by our detector, the bagging process of random forest makes the method very robust against artifacts that do not show a clear batch effect.

In this example, models capable of disambiguating cancer and non-cancer samples were achieved (Table 2.1) regardless of HPD artifact removal. While this might seem to imply that artifact removal is not necessary, this interpretation is not the case. The presence of artifactual features in the importance lists for the classifiers without consistent HPD site removal implies that these models are using artifactual information to determine cancer or non-cancer status. Since these artifacts are known to vary between instruments (Figure 2.7C) and in some cases between sample class (Figure 2.8), a change in sample preparation or in which instrument is used to acquire a spectrum will limit the effectiveness of these artifact-informed models. Furthermore, these results highlight that when artifacts and other systematic errors are present, classification accuracy is not necessarily a complete indication of classifier quality or robustness. In this particular dataset, the artifactual features do not rise to the very top of the importance lists for any of the models, but in other datasets we have investigated, highly important features have been artifacts.

Although the removal of consistent fuzzy site regions is efficient, safe and consistent, it does have shortcomings. First, it will remove larger amounts of spectrum as the number of samples and classes increases. Second, when fuzzy site location consistently overlaps with non-artifactual peaks, these peaks will be discarded even if they are not artifactual. Without a per-peak metric for determining if a peak is a likely artifact or not, this is an unavoidable but undesired side effect. Third, not all consistent HPD regions are fuzzy sites. We have observed consistent HPD regions that do not contain obvious fuzzy sites, at least not in the spectra we manually inspected. These regions do contain high numbers of peaks that are too consistent across spectra, have too many peaks to be random noise, and have poor scan-to-scan consistency. Examples of these regions are illustrated in Figure 2.10. We hypothesize that these sites do contain an HPD artifact of some sort, but thresholding at the instrument level has removed all but the most intense of these artifactual peaks.

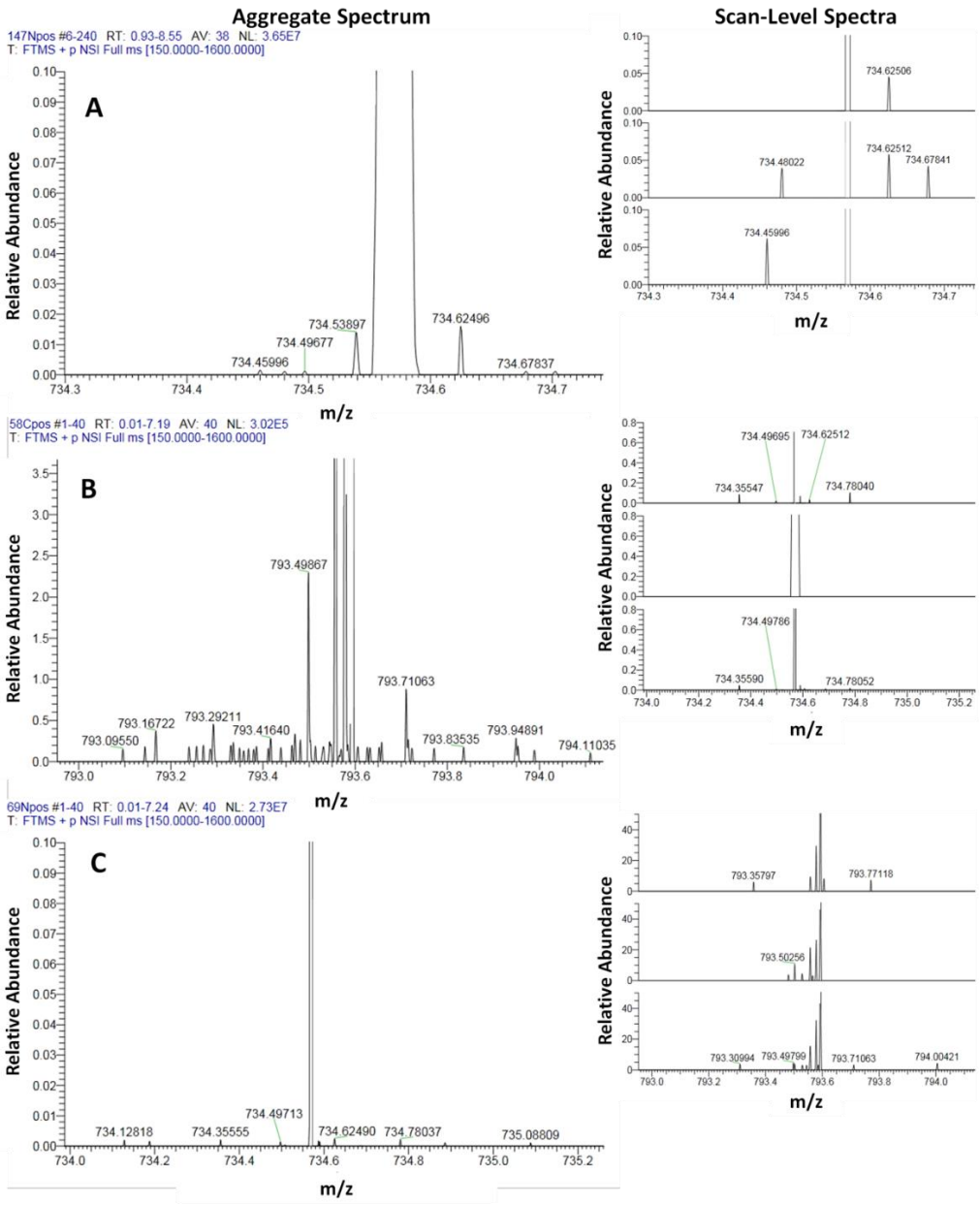

Figure 2.10: Example False Positive HPD Sites
Not all HPD positive regions contain fuzzy sites; however, the false positive sites do contain abnormal regions of peak density. At the aggregate level, there is no clear fuzzy site, partial ringing, or ringing in these examples. However, a scan-level analysis shows a characteristic pattern of poor scan-to-scan correspondence for many of the peaks similar to the peaks found in HPD artifacts. For example, Panel A has a scattering of side peaks resembling filtered partial ringing. Panel B appears to be a thresholded fuzzy site with some non-artifactual peaks as well. Panel C has a few low intensity peaks, but the central peak has a variable width across scans. All spectra acquired using μS =10, R = 500K

2.4.3 *Peak Characterization Generates High Quality Peak Lists*

Dr. Robert Flight's scan-level peak characterization algorithm addresses both the high peak intensity variance observed between scans and results in better described peaks than other peak picking software (Flight and Moseley, 2018).

First, normalizing each scan prior to aggregation to yield the final spectrum provides a substantial improvement in peak height RSD and reduces the ranges of observed noise levels between scans. Without scan normalization, peak height RSDs can exceed 1.0 implying that the standard deviation in observed peak heights across all scans is equal to or greater than the mean peak height. Without reducing this RSD, meaningful statistical comparisons of relative peak height between isotopologues of the same compound (which could be used to confirm metabolite assignments) would be effectively impossible.

Second, by making full use of scan-level information peak, correspondence can produce more informative descriptions of peaks at the aggregate level. For example, this method returns the measured standard deviation and relative standard deviations of peak intensities and *m/z*'s across scans. Furthermore, this method identifies which scans contain a given peak from the aggregate spectrum. These improved descriptions of peaks will enable more sophisticated and informative downstream analyses. Having access to more complete statistical descriptions of these parameters will enable more

sophisticated and informative downstream analyses including more statistically robust metabolite assignment.

2.5 Conclusions

With our HPD site detection methods and manual investigation of many spectra, we have identified and characterized three distinct types of artifacts that produce large numbers of peaks (up to 1/3 of the peaks in a spectrum): fuzzy sites, ringing and partial ringing.

Although peaks resulting from any of these artifacts could have negative implications for experimental interpretation, our study focused primarily on fuzzy sites and their detection, because they were novel artifacts that were prevalent in our datasets and are particularly problematic for classification studies. Fuzzy sites are likely the result of radio frequency interference and their location correlates with sample class, increasing the probability of class-specific misassignment and the introduction of sample-specific artifactual features. Ultimately, the presence of these artifacts produces brittle classifiers and complicates the characterization of true biological variance between sample classes using direct-injection FT-MS. The results presented in this study highlight how peaks resulting from artifacts can be misassigned by tools such as LipidSearch and how these misassignments can impact downstream analyses. The methods and tools presented in this study detect and remove fuzzy sites in a non-encoding manner from direct-injection FT-MS spectra (i.e. without encoding

sample class as absence of peaks), while providing sufficient protection from fuzzy site artifacts with existing assignment methods.

In addition to the artifact detection and removal tools, a peak characterization method was developed that can significantly improve peak intensity RSDs and produce more informative peaklists than other methods without human intervention. These more informative peaklists describe each peak parameter statistically and can be generated very efficiently, making our approach compatible with large high-throughput metabolomics experiments. Combined with our artifact detection and removal tools, the ability to provide accurate statistical descriptions of detected peaks and to correct for scan-to-scan variances will provide the high quality peaklists necessary for more complicated but more capable assignment pipelines.

CHAPTER 3. SMALL MOLECULE ISOTOPE RESOLVED FORMULA ENUMERATOR (SMIRFE) – A TOOL FOR UNTARGETED MOLECULAR FORMULA ASSIGNMENT

3.1 Introduction

Advances in Fourier-transform mass spectrometry (FT-MS) provide substantial simultaneous improvements in mass accuracy, mass resolution, and sensitivity (Eliuk and Makarov, 2015). Theoretically, these combined capabilities provide several analytical and interpretive improvements including the ability to resolve distinct isotopologues of metabolites and perform multi-isotope natural abundance correction (Carreer *et al.*, 2013) (Moseley, 2010), improved assignment accuracy (Kind and Fiehn, 2006), and the detection of low concentration metabolites (Eyles and Kaltashov, 2004). For the metabolomics field, these improvements permit the use of multi-isotope labeled precursors in multiple stable isotope resolved metabolomics (mSIRM) experiments (Yang *et al.*, 2017b). These experiments provide richer information that be used to elucidate unknown metabolic pathways (Creek *et al.*, 2012) (Higashi *et al.*, 2014), quantify relative metabolic fluxes through connected pathways (Hiller *et al.*, 2010), identify multiple pools of metabolites in different compartments (Fan *et al.*, 2012), and identify the active metabolic pathways under various cellular conditions such as cancer versus non-cancer (Sellers *et al.*, 2015b).

However, these more informative experiments require that the spectral features representing isotopologues of metabolites observed in these experiments are accurately assigned. As discussed in the background, the

assignment of FT-MS spectra remains an unsolved problem in metabolomics (Wishart, 2011). Existing spectral assignment tools are largely targeted and rely upon databases of known metabolites as sources of possible assignments. Although orthogonal information from tandem-MS or chromatography can be used to aid in assignment, many of these targeted tools use simple *m/z*-based queries against metabolite databases to generate assignments. Assignments generated from *m/z* queries alone are poorly cross-validated even with high mass resolution (Kind and Fiehn, 2006) and acquiring orthogonal information in direct infusion experiments is challenging. Furthermore, existing metabolite databases are incomplete (Mitchell *et al.*, 2014), which limits the assignment of unknown metabolites and biases assignments towards the set of metabolites represented in a given database (Moseley, 2013).

An alternative approach is untargeted assignment, where assignments are generated without databases of known metabolites. This would enable the assignment of unknown metabolites and reduce assignment bias; however, enumerating all possible assignments for an observed spectral feature is an immense computational challenge. This limitation is especially pronounced in SIRM experiments where the enrichment of labeled isotopes makes otherwise low abundant isotopologues detectable. Furthermore, untargeted assignments must still be cross-validated in some manner to reduce the set of possible assignments for a spectral feature to a likely set of possible assignments. One approach for this is to compare the intensities of observed spectral features to one another to identify likely isotopologues of the same metabolite. However, the

data quality problems discussed in Chapter 2 result in artifactual peaks that when confused for isotopologue peaks can complicate this process and the high variances in observed peak intensities makes robust intensity comparisons difficult.

Using Dr. Flight's improved peak characterization methods from Chapter 2, these data quality problems in FT-MS metabolomics data sets have been addressed (Flight and Moseley, 2018). The improved peaklists generated by these methods combined with the combinatorial mathematics underlying natural abundance probabilities (NAPs) used for natural abundance correction in SIRM experiments provides an avenue for developing an untargeted assignment method that can use sets of isotopologues within an MS1 as cross-validating information. We have developed a method based on this principle called Small Molecule Isotope Resolved Formula Enumerator (SMIRFE) that integrates this methodology to provide cross-validated and statistically evaluated elemental molecular formula (EMF) and isotopologue-resolved molecular formula (IMF) assignments for FT-MS spectra without additional orthogonal information nor databases of known metabolites.

3.2 Materials and Methods

3.2.1 SMIRFE Algorithm Overview

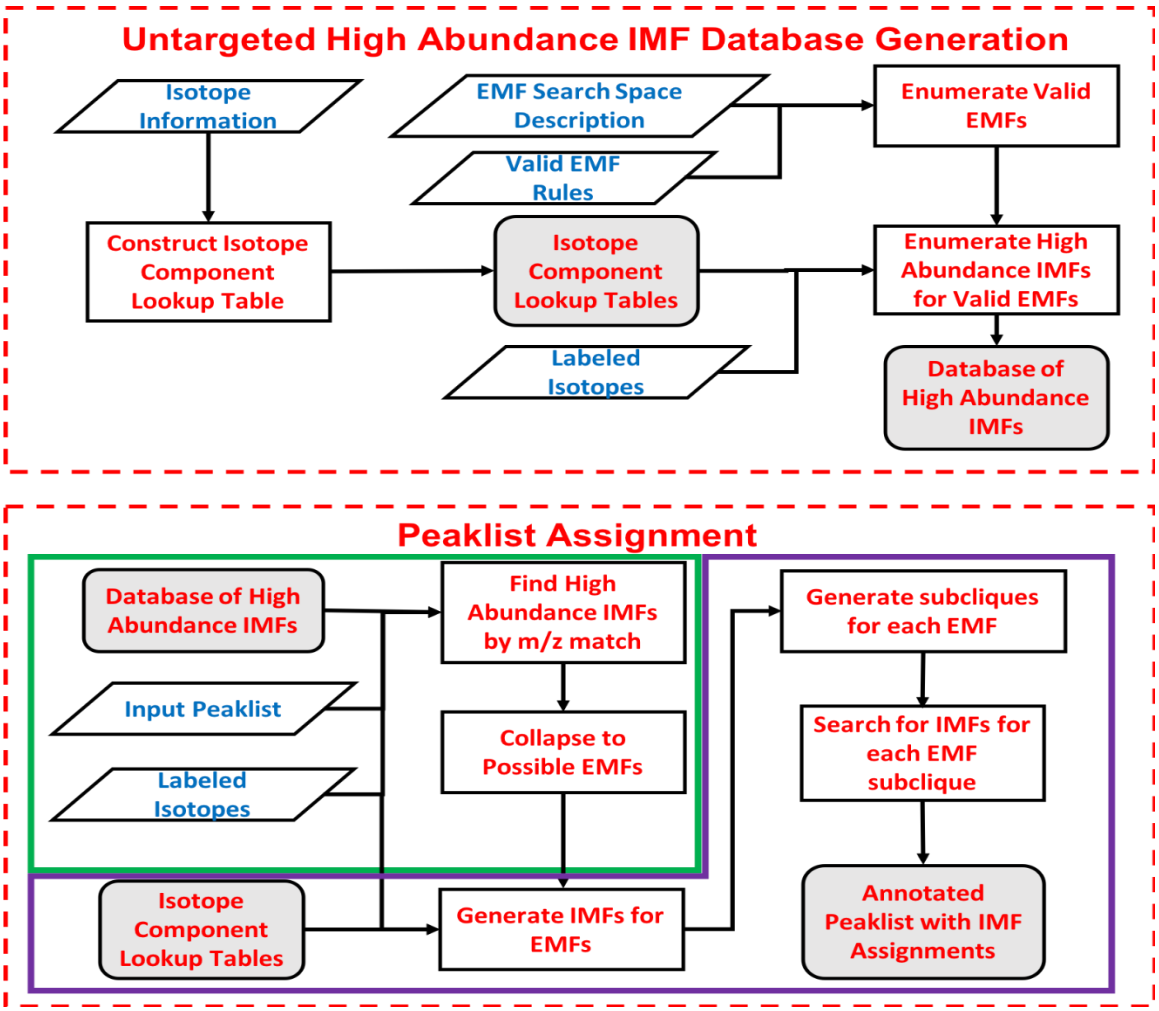

Figure 3.1: Flowchart of the SMIRFE Algorithm

SMIRFE can be divided into two main components. The first is the combined generation of an EMF search space and an isotope component lookup table. The second is a peaklist assignment algorithm. The component lookup table stores the NAPs and masses of all isotope combinations for each element to be used during assignment and IMF enumeration. The high abundance IMF database represents all likely IMFs in the EMF search space. Peaklist assignment uses this database of IMFs to generate possible EMFs present in the spectrum which are then used to inform a targeted IMF search for each EMF. The untargeted component of the search is highlighted in green and the targeted component is highlighted in purple.

The number of elemental molecular formulas (EMFs) and corresponding isotope-resolved molecular formulas (IMFs) for an EMF search space is prohibitively large to recalculate for every spectrum to be assigned. For CHONPS (carbon, hydrogen, oxygen, nitrogen, phosphorus and sulfur) stable isotopes alone, very large IMF search spaces greater than 10^{29} are typical for molecules of 1400 Daltons or less. Instead, SMIRFE first generates a representation of EMFs in the search space given a set of possible labeled isotopes and a user-provided description of the search space. The resulting EMF search space, which can still be quite large, is specific for that search space and labeled isotopes but can be reused between similar experiments. With this EMF search space, untargeted EMF assignment followed by targeted IMF assignment on an input peaklist can be performed efficiently. The organization of these steps is represented in Fig 3.1 and will be discussed in more detail.

3.2.2 SMIRFE Nomenclature

The output from SMIRFE is a mapping of peaks in a spectrum to distinct IMFs and a measurement of the statistical reliability of that assignment. Throughout SMIRFE, it is necessary to keep track of which groups of IMFs belong to which EMFs. On the surface, this seems straightforward as each IMF belongs to exactly one EMF, but this becomes more complicated when multiply labeled forms of multiply adducted forms of EMFs are present. To organize IMFs during SMIRFE analysis, IMFs are grouped into EMF supercliques, EMF cliques and EMF subcliques as shown in Figure 3.2. An EMF superclique is defined as all IMFs that have the same EMF. For example, the EMF superclique for $C_6H_{12}O_6$

contains all IMFs sharing the EMF $C_6H_{12}O_6$. An EMF clique is defined as all IMFs that have the same EMF and the same adduct. EMF cliques subdivide an EMF superclique. For example, all IMFs of $C_6H_{12}O_6$ that are adducted with sodium belong to the $[Na^+]$-$C_6H_{12}O_6$ EMF clique. An EMF subclique is defined as all IMFs that have the same EMF, the same adduct, and have the same number of labeled isotopes (i.e. are 'identically-labeled'). EMF subcliques subdivide EMF cliques. For example, all IMFs of $C_6H_{12}O_6$ that are adducted to sodium and contain two ^{13}C due to labeling belong to the $^{13}C_2$ subclique of the $[Na^+]$-$C_6H_{12}O_6$ EMF clique.

Because EMF supercliques, EMF cliques and EMF subcliques are defined in terms of EMFs, IMFs of isomeric compounds (which by definition have the same EMF) are grouped together into EMF supercliques, EMF cliques and EMF subcliques. For example, although all IMFs of glucose will belong to the $C_6H_{12}O_6$, EMF superclique, all IMFs of all isomers of glucose will also belong to the $C_6H_{12}O_6$, EMF superclique

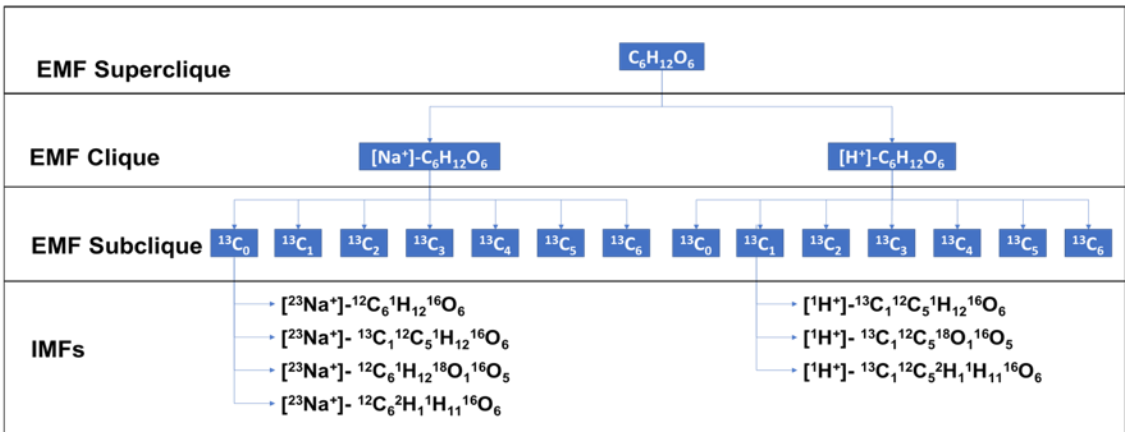

Figure 3.2: Organization of EMF Supercliques, Cliques and IMFs
All IMFs of a given EMF belong to one superclique which is further divided by adduct into cliques which are further subdivided into subcliques based on labeling. For example, all glucose IMFs belong to the same superclique, but will be subdivided into one or more cliques based on which adducts of those IMFs are observed (e.g. Na+ IMFs in one clique, H+ adducts in the other). The number of subcliques changes based on which isotopes are presumed labeled. In the unlabeled case, there is only one subclique for each clique, but when one or more isotopes are potentially labeled, the number of subcliques increases quickly. If 13C is labelable in C6H12O6, there will be 7 possible subcliques representing sets of IMFs with 0 to 6 13Cs from labeling and their isotopologues from natural abundance.

SMIRFE uses the natural abundance probability (NAP) and relative natural abundance probability (relative or relNAP) of IMFs both for IMF generation and for assignment scoring. The NAP of an IMF is the probability of observing the set of isotopes present in that IMF by random chance alone assuming that all isotopes have a probability of occurring equal to their natural abundance, i.e. their expected relative frequency in the biosphere. Labeling introduces a non-natural source of specific isotopes and thus relative NAP must be used to quantify the probability of observing a labeled IMF when one or more isotopes can be labeled. These values are calculated as follows (Moseley, 2010) (Carreer *et al.*, 2013):

$$[Eq\ 3.1] - NAP_E(k_1, k_2, \ldots, k_m) = \binom{E_{max}}{k_1, k_2, \ldots, k_m} \prod_{x=1}^{m} NA_{E_{1x}}^{k_x}$$

$$[Eq\ 3.2] - NAP_{IMF} = \prod_{j=1}^{n} NAP_{E_j}$$

Where for a given number of nuclei (E_{max}) for an element E with a set of stable isotopes (k_i) with known natural abundances (NA), the elemental NAP for a given combination of isotopes is calculated with Equation 3.1. The NAP of an IMF (Equation 3.2) is the product of each element's NAP in the IMF. Without labeling, this NAP is referred to as the absolute NAP. When labeling is present, labeled isotopes must be removed from the calculation of NAP for the element(s) of the labeling isotope(s). This effectively changes E_{max} and the isotope counts for a labeled IMF. NAP values for isotope combinations can be calculated and stored in a lookup table for later use. Therefore, within SMIRFE, any isotope combination has its NAP calculated only one time. Also, while absolute NAP and relative NAP are interchangeable without labeling, comparisons between relative NAP is always used when comparing IMFs.

3.2.3 Isotope Component Lookup Table Creation

Throughout SMIRFE, highly accurate absolute NAPs and masses for the set of isotopes for any given element are needed to calculate various NAPs and masses of IMFs. Although the total number of isotope combinations for each element is large, it is much smaller than the number of possible IMFs and can be

calculated once and stored for future use. The first step in both the untargeted search space generation and in the actual assignment algorithm is to generate this lookup table. The lookup table is sub-divided by element and for each element every possible combination of isotopes for that element is stored along with the mass of that component, it's absolute NAP, and the number of atoms it contains. The masses in this table are calculated using the mpmath high-precision math library (Fredrick Johansson, 2018) in Python and the absolute NAPs are calculated using the multinomial calculation shown in Equation 3.1. The number of components stored in this table grows exponentially as the search size increases, but the table is only megabytes in size and can be stored in JavaScript Object Notation (JSON) format.

3.2.4 Untargeted EMF Search Space Creation

SMIRFE requires an internal representation of an EMF search space that can be queried to produce probable high NAP IMF assignments for a given peak. This search space consists of several components. First is a high level description of the search space provided by the end user that specifies the max count for each element in the space, the maximum mass to consider (max_mass), the minimum NAP (min_NAP) to consider (i.e. IMFs with NAP below the min_NAP do not need to be enumerated), and the set of labeled elements in the search space (e.g. C:100, H:230, O:40, max_mass:1600, min_NAP:0.4, labeled:['C']). From this description, SMIRFE constructs an N-dimensional integer lattice where each unique N-tuple of integers represents exactly one EMF in the space. For example, (6,12,6) would represent the EMF

$C_6H_{12}O_6$ in the above search space. The number of tuples in a possible search space grows exponentially with increasing dimensionality, but the enumeration of integer lattices can be done quickly using existing functions from the Numpy Python package (Walt *et al.*, 2011). The set of all unique tuples is equal to the set of all possible EMFs in the search space; however, many of these EMFs are not consistent with known rules for valid chemical structures. For example (1,100,0) corresponds to the EMF C_1H_{100}, which very likely does not exist as the number of hydrogens exceeds the maximum number of valence electrons from non-hydrogen atoms in EMF (i.e. 1 carbon contributes 4 valence electrons max which is not enough to make covalent bonds to 100 hydrogens). As EMFs are generated during lattice enumeration, EMFs that fail to comply with known hard rules about valence electrons are removed. Heuristics based on patterns observed in the HMDB and KEGG further restrict possible EMFs.

The second component of the search space representation is the set of high abundance IMFs for each possibly valid EMF. The IMF generation procedure can be modeled as a depth first search on the complete tree of possible isotope components possible with that EMF, with each level in the tree representing the components of a given element. Components of labeled elements are treated as having a NAP of 1, while components of unlabeled elements below the min_NAP are pruned prior to traversal. A depth first search is then performed to find all paths through the tree for which the product of the NAP of all nodes in the path exceeds the min_NAP and the sum of the mass of the nodes in the path is below the max_mass. Every such path represents exactly

one IMF for the given EMF, The IMF and its mass is then stored in an SQLite (Owens, 2006) database for use in SMIRFE. As NAP can only decrease along a path and mass can only increase along the path, traversal can be short-circuited whenever a nascent path has a NAP below the min_NAP or mass above the max_mass. Once all IMFs have been enumerated for all EMFs, an index is built for the IMF database on IMF mass. These two components combined with the adduct description provided to the assignment algorithm completes the internal representation of the search space needed for SMIRFE.

3.2.5 Preliminary EMF Search

Searching all peaks in a spectrum against all possible IMFs for any sufficiently large search space quickly becomes computationally intractable. For example, $C_6H_{12}O_6$ has 1,119,744 possible isotopologues; however, without extensive labeling, most of these isotopologues are too low in abundance to be detected based on natural abundance probabilities. Searching for all these isotopologues would not only be wasteful, but would also generate many false assignments. This search can be constrained by observing that for a given labeled version of an EMF, there is only one most-abundant IMF that must be present if that labeled EMF exists in the sample. For example, without labeling, the monoisotopic version of $C_6H_{12}O_6$ is the most abundant IMF of glucose and must be observed if unlabeled glucose is present in the sample. If only $m+{}^{13}C_1$ glucose was observed and no ^{13}C labeling source is present, it can be inferred that this assignment is spurious and should be ignored. However, if a ^{13}C labeling

source is present, detection of the m+$^{13}C_1$ glucose isotopologue is possible without observing the monoisotopic version.

By searching each peak in a spectrum against the untargeted EMF search space generated previously and testing if each returned IMF is a most abundant IMF of a possible labeled EMF based on the user-provided labeling pattern, a set of possible EMFs in the spectrum can be generated. The set of possible adducts is user specified (e.g. Na^+, H^+, K^+, NH_4^+), and, for every peak, the database query is performed for each adduct as the adduct's mass must be accounted for during searching. This set of possible EMFs informs the more targeted IMF search in the next step of SMIRFE assignment. To further eliminate spurious assignments, only peaks present in most scans (>90%) are used for building this possible EMF set. This effectively eliminates the possibility that noise is included in this search and restricts the search to peaks corresponding to high-abundance IMFs. IMFs that do not correspond to the most abundant are stored as a source of random assignments needed for statistical testing (i.e. even false positives are useful in the right statistical situation). Limiting this search to only peaks that are present in 90% of scans has the potential to reduce sensitivity; however, peaks that are present in fewer than 90% scans are still assignable to IMFs of these possible EMFs in the next step of the algorithm.

3.2.6 Targeted IMF Searching

With possible EMFs identified, the next step is a more targeted search for the IMFs of those EMFs. This is done at the subclique level and for every possible subclique for every clique for that EMF, the possible IMFs are

enumerated in order of descending relative NAP (to a minimum relative NAP). For each IMF, peaks with matching *m/z* are tentatively assigned to that IMF. If there are multiple possible matching peaks (very rare), the peak with the closest matching mass is assigned.

Given the limitations of *m/z*-based matching, IMFs after the first IMF must meet additional requirements to be tentatively assigned. First, the nth IMF may only be assigned if the scans in which the nth IMF was observed overlap significantly (>90%) with the scans in which the n-1th IMF was observed. Second, since labeling may only increase the relative abundance of an IMF but never decrease it, IMFs should be observed in the order in which they are enumerated (i.e. it does not make sense to see an IMF with a lower relative NAP but not see an IMF of the same subclique with a higher relative NAP). Third, the nth IMF must have at least one intensity ratio with another IMF in the subclique with a score greater than the 95th percentile of the random intensity ratio scores of the non-matching, non-most abundant IMFs stored from the EMF search step. The third rule is omitted in the case where an IMF may be more heavily labeled. For example, if ^{13}C is presumed labeled and the m+$^{13}C_2$ subclique of Na-glucose is being enumerated, the m+$^{13}C_3$ isotopologue might be observed but does not have to have a matching intensity ratio as it could represent either $^{13}C_2$ from labeling and $^{13}C_1$ from natural abundance (in which case it should have a matching intensity ratio) or $^{13}C_3$ from labeling and no ^{13}C from natural abundance (in which case it may not have a matching intensity ratio). If at any point one of these rules is broken, further enumeration of the subclique is terminated.

Anytime a subclique contains two or more IMFs, it is possible to evaluate the likelihood of observing a set of IMFs that match the intensity ratios observed for that subclique at random. This is reported as an e-value and is calculated using Equation 3.4.

$$[Eq\ 3.3\text{A}]: \mathbf{R} = \left\{\text{Log10}\left(\text{I}_{\text{A,j}}\right) - \text{Log10}(\text{I}_{\text{B,j}}) \middle| \forall\ \text{j} \in \text{scans}, \text{I}_{\text{A}} > 0,\ \text{I}_{\text{B}} > 0\right\}$$

$$[Eq\ 3.3\text{B}]: \text{M}_{\text{log\_ratios}} = \text{median}(\mathbf{R})$$

$$[Eq\ 3.3\text{C}]: \sigma^2_{\text{log\_ratios}} = \text{variance}(\mathbf{R})$$

$$[Eq\ 3.3\text{D}]: \chi_{\text{A,B}} = \frac{\left(\text{log10}(\text{NAP}_{\text{A}}) - \text{log10}(\text{NAP}_{\text{B}}) - \text{M}_{\text{log\_ratios}}\right)^2}{\sigma^2_{\text{log\_ratios}}}$$

$$[Eq\ 3.3\text{E}]:\ S_{A,B} = 1 - chi2cdf\left(\chi_{A,B}, df = 1\right)$$

Equation 3.3 - The score for a pair of isotopologues A, B ($S_{A,B}$) in the same subclique can be calculated using a chi-squared statistic $X_{A,B}$. $\text{M}_{\text{log\_ratios}}$ is the median observed ratio of the log10 intensities of A and B across all scans containing both A and B. $\sigma^2_{\text{log\_ratios}}$ is the variance of the observed scan level log10 intensity ratios for A and B. These log ratios of observed peak intensities is compared to the log10 ratio of the NAPs of A and B.

$$[\text{Eq. } 3.4\ A]: \boldsymbol{S} = \left\{\frac{S_{\boldsymbol{A,B}}}{{\sigma_{random\ pair\ scores}}^2} \middle| \forall A, B \in \boldsymbol{I}, A \neq B\right\}$$

$$[Eq.\ 3.4\ B]: E = 1\ - chi2cdf\left(\boldsymbol{S}\ * \boldsymbol{S}^T,\ df = |\boldsymbol{S}|\right)$$

Equation 3.4 - Every subclique containing at least two isotopologues can be assigned an E value *E,* which represents the probability that random pairs of isotopologues would have the same distribution of chi squared statistics as the observed subclique. This calculation requires constructing the score stat vector S for all unique pairs of isotopologues in the subclique (***I***). The variance for these statistics is the variance of the random pairs of isotopologues constructed earlier from the invalid isotopologue assignments. A combined chi-squared statistic is then calculated for the subclique by multiplying S by its transpose with which the E value can be calculated. This formulation enables the future correction of potential correlation between isotopologue pair statistic values and does not require modification if the underlying isotopologue pair scoring function is changed.

3.2.7 Dataset for Validation

The method for generating our validation dataset was adapted from previously published method for performing ethylchloroformate (ECF) amino acid derivatization (Yang *et al.*, 2017b). Two replicate samples were prepared and spectra were obtained for both samples using a Tribrid Fusion Orbitrap at 500k resolution and from 150 to 1000*m/z*. Additional details for the preparation of these samples is included in Appendix 1 (Sample C). Spectra were characterized using the peak characterization methods described in 2.2.2.

3.2.8 Manual Inspection of Spectra

Both reference spectra were examined to verify the presence of peaks corresponding to expected derivatives of the amino acids known to be present in the sample and the absence of any large spectral defects. Although fuzzy site FT-MS artifacts are present in the spectrum (Mitchell *et al.*, 2017), these artifacts do not overlap with any of the important non-artifactual peaks. Peaks corresponding to derivatives of 19 of 20 amino acids (or 18 of 19 as isoleucine and leucine are isomers) were observed. No peaks were observed that would correspond to the expected Na+ or H+ adducts of cysteine. The regions of spectra where peaks corresponding to these amino acids are expected to occur are shown in Figure 3.3.

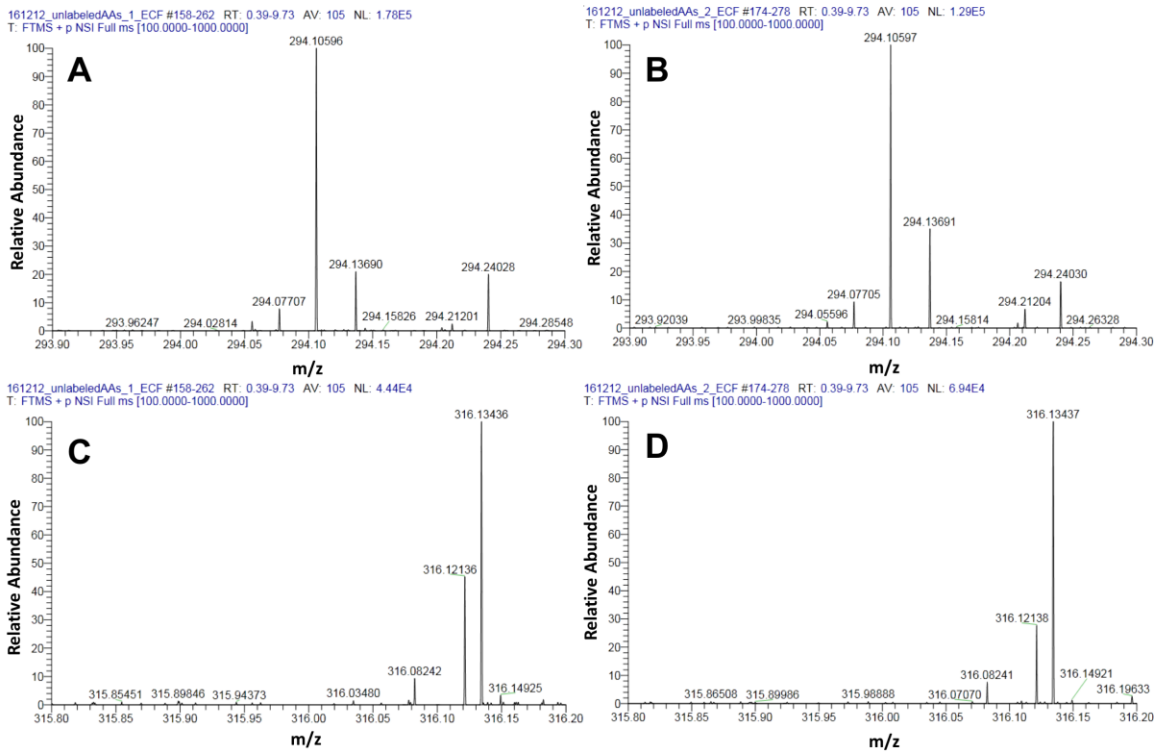

Figure 3.3: Regions of Spectra Showing Absence of Cysteine
The H+ adduct of ECF derivatized cysteine has a theoretical monoisotopic mass of 294.100586, while the Na+ adduct has a theoretical monoisotopic mass of 316.082531. Shown in panels A, B are the regions of spectra corresponding to the hydrogen-adducted cysteine derivative in the amino acid (AA) samples 1 and 2 respectively. Panels C, D show the sodium-adduct region in AA samples 1 and 2 respectively. For the hydrogen adduct, in both A and B the peaks at 294.10598 or 294.10595 are intense but are too far from the expected *m/z* to be reasonably assigned to a cysteine derivative. In C and D, the peak at 316.082 has an *m/z* very close to the expected *m/z* of the sodium adduct but this peak is not very intense. If this peak was the monoisotopic isotopologue of the cysteine derivative, the instrument has insufficient dynamic range to observe the less abundant isotopologues

3.3 Results

3.3.1 *EMF Search Space Growth*

The number of possible IMFs in a given EMF search space, without limitations on NAP, valid EMF, and mass, quickly makes a brute force enumeration of all IMFs impossible. For example, the search space C: 100, N: 7, O: 40, H: 230, P: 3, S: 3, contains 1.021×10^{29} possible IMFs! However, since the partial EMF search space must only contain highly abundant IMFs for those partial EMFs, the number of possibilities can be greatly reduced by enforcing a minimum relative NAP for the IMF entries in the database. When labeling is considered, the effect of the NAP filter is reduced, but the database remains a manageable size and has enough performance to be queried efficiently with observed *m/z*'s from a spectrum. For example, the search space used to assign our reference spectra with a min_NAP of 0.4 contains 13,045,817 IMFs with no labeling, but increases to 53,102,054 with ^{15}N-labeling. The size of SMIRFE search spaces with respect to min_NAP, mass and labeling is shown in Figure 3.4.

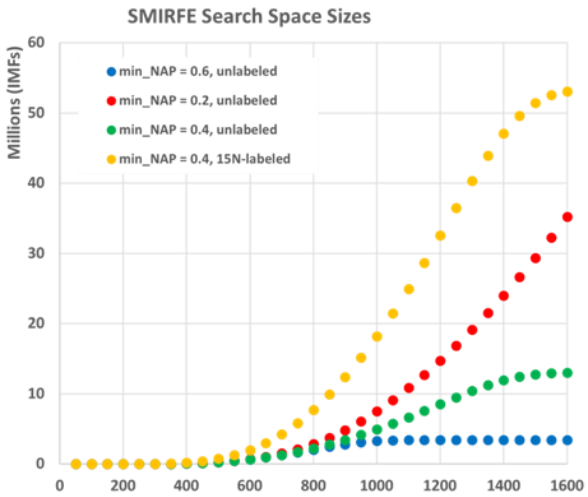

Figure 3.4: NAP and Labeling Effects on SMIRFE Search Space Size
The size of an EMF search space in terms of IMFs grows with increasing mass. When min-NAP is low, the number of IMFs grows exponentially (red) and many EMFs have many IMFs in the database. When min_NAP is too high, databases are smaller but incomplete. Full enumeration of the entire mass range is impossible (blue) due to the relative abundance of m and m+$^{13}C_1$ decreasing with increasing mass. For our spectra, which have a max *m/z* of 1000, a min_NAP of .4 (green) is sufficient. Database size remains small as less probable IMFs are not enumerated but each EMF has its most abundant IMF. Labeling can dramatically increase database size (yellow) as the min_NAP restriction is lifted on labeled elements. Rules on valid molecular formulas keeps database generation tractable but are more permissive than NAP based rules.

3.3.2 *Assignments for the ECF Derivatized Spectra*

Using a C:130, N:7, O:40, P:3, S:3, H:230, max_mass: 1600, min_NAP: 0.4, labeled_isotopes: ^{15}N, the EMF search space was constructed using SMIRFE for the ECF derivatized spectra. This database was used to identify likely EMFs in the spectrum, which were then followed by a targeted IMF search for those EMFs. Na^+ and H^+ adducts were considered for every peak. From previous studies, the expected derivatives of each amino acid with ECF are known (Yang *et al.*, 2017b). Of the 20 amino acids present in the solution

(representing 19 derivatives with distinct formulas – leucine and isoleucine are isomers), the expected derivatives of 18 out of 19 of those formula-distinct amino acid species were assigned by SMIRFE. No assignments were made to cysteine which was confirmed to be absent in the spectrum and likely was lost during sample processing. The ^{15}N-labeled versions of these derivatives were also identified as well as multiple adducted versions of both labeled and unlabeled versions. The distinct IMFs identified for each amino acid are summarized in Supplemental Tables 3.1 and 3.2

3.3.3 *Assignment Ambiguity*

Intensity ratio scoring combined with scan subsetting and a tight *m/z* match tolerance significantly narrows the number of possible assignments for a peak. Without these filters, the number of possible assignments for a peak can be large even with high mass accuracy (Figure 3.5). At low *m/z*, when the number of possible EMFs and therefore IMFs is relatively small, this can result in unique IMF assignments for those peaks; however, at higher *m/z* when the number of possible assignments is very large, unique assignments are rare. Multiple assignments can be disambiguated by their e-value in many cases. In our test case, every expected EMF has at least one assignment that had the best e-value for a peak and many of the observed IMFs for the expected EMFs were the best assignment for their peak (82.4% for sample 1 and 82.6% for sample 2).

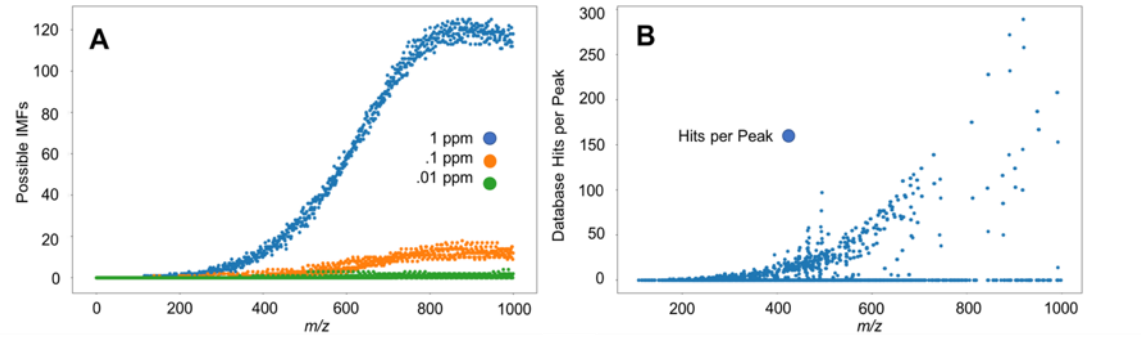

Figure 3.5: Theoretical and Experimental Possible Assignment Ambiguity - By scanning the 15N-labeled database described previously for masses between 0 and 1000 in .1 *m/z* increments assuming three different levels of accuracy, 1ppm (blue), .1ppm (orange) and .01ppm (green), the relationship between the number of possible assignments and mass accuracy can be explored. For all levels of accuracy, increasing mass increases the number of possible assignments. This is obvious for the 1ppm result, but all three levels of accuracy tested have this behavior. This behavior is partly due to the presence of more EMFs at higher *m/z* but also because a ppm definition of mass error results in a higher absolute mass error for larger masses. A similar behavior was observed in the assignments for our experimental spectra where peaks with higher *m/z* have more possible assignments (panel B). The number of possible assignments for a peak in real world spectra can be considerably higher than expected from the results in panel A where no adduction and a relative NAP filter are applied. Peaks with zero possible assignments either correspond to noise, artifacts, compounds whose elemental composition includes elements not included in the search space, or compounds with low relative abundances for which only the most abundant isotopologue is detectable.

3.3.4 *Assignment m/z Error*

From the set of IMF assignments for expected EMFs in our example dataset, the patterns in mass error can be investigated. Both spectra were acquired using the same Thermo Tribrid Fusion instrument for which the stated mass accuracy is 1 ppm. The observed mass errors fall within this specification for 100% of the assigned IMFs for the validated EMFs. For the set of all assignments, validated plus unvalidated assignments and including all possible assignment for each peak, only 34 of 24215 fall outside the 1 ppm error specification. When only the best assignment for each peak is considered (lowest

E-value), only 1 of 975 assignments have a ppm error over 1ppm. These results are consistent with previous findings that high mass accuracy alone cannot provide unambiguous metabolite assignments (Kind and Fiehn, 2006).

The mass error across *m/z* is not constant (as might be expected from miscalibration) and instead changes with *m/z*. During manual investigation of these spectra, it was observed that peaks near intense peaks in *m/z* space often had a larger mass error than peaks far away from other intense peaks. This effect can be attributed to coulombic interactions between ions and can be quantified with Equation 3.5:

$$[Eq.\,3.5]: Repulsion = \sum_{j=0}^{|P|} I_j * \frac{sign(f_i - f_j)}{|f_i - f_j|^2}$$

The repulsion expression is derived from Coulomb's law but as the distance between ion clouds in the orbitrap is not known, the average distance between two clouds can be estimated by the different in their precession frequencies f_i and f_j. In our example, we cannot directly obtain the precession frequencies either, but a proxy frequency can be derived from the observed *m/z* and the estimated digital resolution. The force on an ion in cloud *i* depends also on the number of ions in cloud *j* (assuming they are all charge z=1). Summing this force across all other ion clouds for each ion cloud *i* gives a net repulsion term for each cloud. Ions with higher repulsion experience more of a net interaction with other ions that can potentially introduce error into their observed *m/z*'s. These repulsive mass errors can exceed 1 ppm in size (Philip Remes *et al.*, 2016). The mass

error with respect to *m/z* and repulsion on the validated IMF assignments are shown in Figure 3.6.

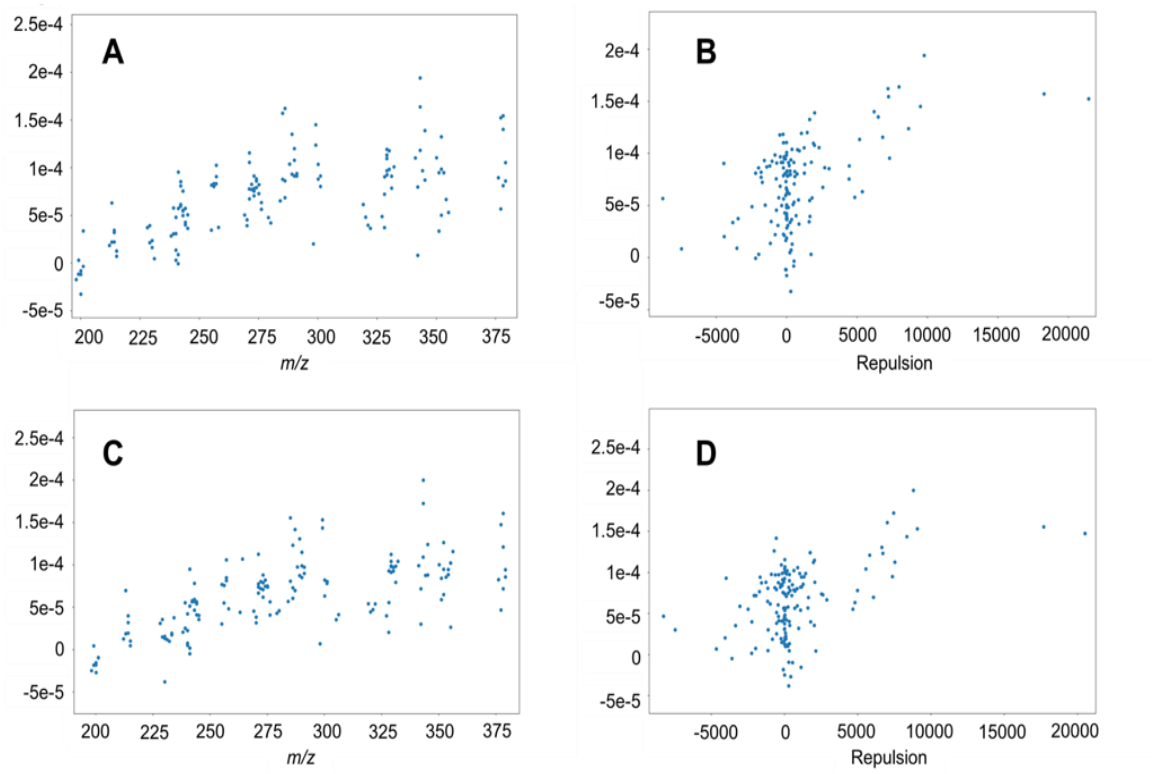

Figure 3.6: Mass Error versus *m/z* and Repulsion - For the set of IMFs corresponding to expected EMFs in our test case, the mass error of those assignments can be calculated. The Fusion instrument has a stated 1ppm accuracy and most of the observed mass errors fall within this specification. The observed mass errors are not constant (and thus not easily correctable) and they change with respect to *m/z* (A – ECF Replicate 1, C – ECF Replicate 2). Mass error does correlate with the measured repulsion (see eq. 3.5) for the peak which estimates the net columbic forces on the ion cloud corresponding to that IMF (B – ECF Replicate 1, D – ECF Replicate 2). Correcting for both an *m/z* dependent and repulsion dependent *m/z* error without standards in an untargeted context, while difficult, could substantially improve the ability to differentiate plausible IMF assignments.

3.4 Discussion

3.4.1 *SMIRFE Algorithm*

Although more untargeted than existing assignment methods, SMIRFE is not a completely untargeted assignment tool (i.e. can search all possible EMFs and all possible IMFs for any number of any elements for all possible labelling patterns). Enumerating all possible IMFs even for relatively small search spaces quickly becomes intractable both in terms of runtime and storage space. The hybrid untargeted / targeted approach employed by SMIRFE leverages the best of both methodologies with the main limitation being the need of the end-user to specify the EMF search space to search. Given that most biomolecules are CHONPS-based, a search space for CHONPS plus expected adducts (in this case $Na^{=}$ and H^{+} adducts) only with element counts large enough to cover 99%+ of known biomolecules in the desired mass range will be applicable for most experiments. When an additional element is needed, this is usually known ahead of time and can be trivially added (i.e. if you are looking for vanadium containing metabolites, you know to add vanadium to the search space).

In these more complicated search spaces, the hybrid untargeted / targeted assignment method employed by SMIRFE becomes advantageous. Since the untargeted EMF database must only contain the most abundant IMFs for the EMFs in the search space, the runtime and storage cost of expanding a search space is minimized. Additionally, since these databases are reusable, the cost of building a large database is amortized across all experiments that can utilize it.

Despite these improvements in performance, large heavily-labeled EMF search spaces will take considerable resources to generate and triply- or quadruply-labeled full-size search spaces are likely impractical to enumerate.

As mentioned previously, *m/z*-based matching alone is error prone without cross-validating information supporting the assignment of an IMF. SMIRFE utilizes the relative intensity ratios observed between pairs of assigned IMF peaks and compares these ratios to the ratio of the relative NAP of the assigned IMFs as a source of cross-validating information. While absolute peak intensities are not necessarily trustworthy in mass spectrometry, the relative peak intensities can be made trustworthy through proper preprocessing. Scan subsetting effectively enforces that IMFs of the same EMF appear in the same scans. This additional rule filters out nonsense groups of peaks that happen to match on intensity and mass.

Unique to SMIRFE is the ability to quantify statistically the fit of the generated assignments. This is achieved using the e-value calculation that comes from the clique score. Although the ‘correct’ assignment is not always the best scoring assignment, the e-value still captures how likely an assignment that is compatible with the peaks could have occurred at random. These e-values are useful in downstream analyses built upon SMIRFE assignments.

3.4.2 *Implications for Experimental Design*

The ability to assign a FT-MS spectrum depends highly on the quality of the spectrum. As SMIRFE assigns FT-MS spectra in a fundamentally different

manner than targeted tools, spectra can be collected in a way that optimizes them for assignment by SMIRFE. First, additional MS1 scans should be acquired whenever possible. Additional MS1 scans improve aggregate peak characteristics that in turn enable better intensity ratio comparisons, but also improve the likelihood that scan subsetting will eliminate spurious assignments. Since SMIRFE cannot currently use MS2 data in any manner to improve assignment accuracy, MS2 time can be used for additional MS1 scans. However, this recommendation may change, if future versions of SMIRFE are developed that use intensity comparisons between observed MS2 peaks to reduce MS1 EMF assignment ambiguity.

Second, since SMIRFE attempts to assign sets of IMFs of an EMF, it is necessary to have enough dynamic range to observe at least two or more IMFs for each EMF. A dynamic range of at least 1000 is sufficient for many examples, but higher dynamic range is always advantageous. While not always possible, ensuring that one or more peaks do not dominate the spectrum ensures that the effective dynamic range of a spectrum is maximized. However, because overloading of ion traps and orbitraps can result in space charge effects that limit mass accuracy and resolution, effective dynamic range must be balanced against effective mass resolution. Alternatively, rather than acquiring a spectrum for the entire *m/z* domain in a single acquisition, multiple overlapping spectra for subdomains of *m/z* can be acquired, assigned and reassembled to yield an assigned spectrum for the desired *m/z* domain. This splitting should provide each sub-spectrum with improved effective dynamic range at the cost of increased

processing times (Southam *et al.*, 2016). The improved quality of the MS1 data provided by this technique could be combined with MS2 to improve assignment accuracy without requiring all MS2 time being allocated for additional MS1 scans.

Additionally, the design of metabolomics experiments should keep in mind that, strictly speaking, SMIRFE does not assign metabolites to spectra but rather EMFs and IMFs that correspond to metabolites. Simply looking up these assigned formulas in a metabolite database can produce metabolite "assignments", but with many of the same caveats of targeted assignment. Machine learning and/or experimentally validated rules can enable the mapping of assigned formulas to metabolite classes, potentially circumventing this limitation in a less targeted manner. Once assigned to classes, differential abundances of metabolites classes could be determined using SMIRFE assignments. However, existing tools for predicting metabolite classes typically require metabolite structure information, which SMIRFE cannot provide.

3.4.3 *Mass Accuracy and SMIRFE*

Despite the use of NAP-based cross-validating information to support SMIRFE generated IMF assignments, the *m/z* of a peak is still important for determining which IMFs it could possibly match. For targeted assignment tools, a mass accuracy of 1 ppm is typically sufficient to achieve unique assignments (although not necessarily correct assignments) up to approximately 800 *m/z*. Due to the large number of possible assignments that SMIRFE considers, a 1 ppm accuracy does not always translate into unambiguous assignments even when additional information is considered.

Furthermore, a 1 ppm mass error while small in absolute terms is very large in relation to the resolving power of FT-MS instruments. If the mass error and resolution were on similar scales, the score of a clique could incorporate the "goodness" of the *m/z* matching to further disambiguate multiply assigned peaks. Our investigation of the mass error trends in our example spectra revealed that the mass error depends both on *m/z* and on proximity of peaks to other intense peaks in a spectrum. Thus, the 1 ppm designation of "accuracy" is illusionary, since multiple sources of error can sum above 1 ppm. Further studies are needed to develop robust methods of correcting both the *m/z* and repulsion component of the mass error for the IMF assignments. Ideally this correction needs to be feasible with limited knowledge about the composition of the sample to enable its use in an untargeted experiment.

3.5 Conclusions

SMIRFE is a novel assignment algorithm that enables an unprecedented level of untargeted assignment for direct-infusion FT-MS spectra. SMIRFE uses a hybrid untargeted / targeted assignment approach that considers effectively all EMFs in a specified search space in a preliminary untargeted search that informs a more targeted IMF search. This hybrid approach minimizes the computational cost of assignment, saves considerable storage space while enabling the relatively untargeted search of all reasonable EMFs. Additionally, SMIRFE can assign labeled spectra and disambiguate between labeled sub-populations of IMFs within a spectrum containing unlabeled and labeled species.

At higher *m/z*, the number of possible assignments considered by SMIRFE results in multiple assignments for many peaks. E-values can aid in disambiguating multiple assignments but ultimately improvements in mass accuracy will greatly facilitate disambiguation by limiting the number of possible assignments that are considered in the first place. These improvements in mass accuracy will likely come from both improvements in mass spectrometry instrumentation and better methods for correcting mass error due to *m/z* shifts and ionic repulsion.

The IMF and EMF assignments provided by SMIRFE will ultimately require classification into metabolite classes for class-level enrichment studies to be performed. Additional tools that can perform these predictions from assigned formulas must be developed and is the focus of the next chapter.

CHAPTER 4. MACHINE LEARNING METHODS FOR LIPID CATEGORY AND CLASS PREDICTION

4.1 Introduction

Lipidomics is the subdiscipline of metabolomics focused on the study of the lipidome - the set of lipid metabolites and their roles within the metabolome (Wenk, 2005). Unlike other categories of metabolites that are largely grouped based on their structures, lipids are defined by their low solubility in water and collectively represent a structurally and chemically diverse set of metabolites with various roles in healthy and pathological cellular function. By virtue of their diversity, seemingly every life process involves lipids including but not limited to: maintenance of cellular structure, membrane fluidity (Singer and Nicolson, 1972), intracellular, extracellular and hormonal signaling (Zechner *et al.*, 2012) (Morrison and Farmer, 2000), energy metabolism (Adeva-Andany *et al.*, 2018) (J R Neely and Morgan, 1974), and disease processes (De Pablo and De Cienfuegos, 2000) including cancer (Zhang and Du, 2012) (Ray and Roy, 2018). Thus, through lipidomics, more complete modeling of cellular metabolism and a better understanding of physiological and pathological processes at the mechanistic level can be achieved (Lydic and Goo, 2018).

Despite the potential benefits, the rigorous analytical investigation of the lipidome in real-world biological samples remains difficult. Lipid features must first be reliably observed in samples and these features must be accurately assigned to a lipid structure and/or lipid category. As was the case with the

detection and assignment of metabolites in general, this represents a significant bioanalytical chemistry problem due to the high structural diversity of lipids, their wide range of observed concentrations, and differences in lipid profiles between different compartments and with time (Horvath and Daum, 2013) (Aviram *et al.*, 2016) (Fahy *et al.*, 2005). The analytical capabilities of mass spectrometry, particularly FT-MS, make it suitable for the resolved detection of lipid features in lipidomics studies (Köfeler *et al.*, 2012). Although untargeted assignment tools such as SMIRFE can assign observed lipid features to elemental molecular formulas, inferring to which category of lipid these features belong remains difficult, especially for direct infusion experiments, where orthogonal chromatographic data is unavailable to confirm molecular formula assignments or disambiguate lipid isomers. Knowing what lipid categories are represented in a sample is necessary for lipidomics experiments concerned with differences at the lipid category level.

Targeted assignment tools attempt to solve the lipid categorization problem indirectly. Rather than assigning spectral features to formulas and then to lipid categories, tools such as LipidSearch (Köfeler *et al.*, 2012) and PREMISE (Lorkiewicz *et al.*, 2012) assign spectral features to metabolites of known lipid categories. As discussed in Chapter 3, targeted assignment methods have several shortcomings that our untargeted formula assignment tool SMIRFE circumvents (Mitchell *et al.*, 2019). However, untargeted formula assignment does not provide assignments to metabolites, from which lipid category can be inferred. Instead a method for predicting lipid category from untargeted molecular

formula assignments from SMIRFE would provide the necessary lipid category information while offering the benefits of untargeted assignment.

Although lipids as a whole are defined by their physical properties, each category of lipid is a set of lipids that share certain chemical substructural features. The prediction of lipid category from known lipid structures is straightforward. Automated tools such as ClassyFire (Djoumbou Feunang *et al.*, 2016) use machine learning methods to automatically assign lipid category to provided structures; however, structural information cannot be directly acquired from SMIRFE. Ultimately, the classification of molecular formulas into specific lipid classes remains an unsolved problem that can prevent the effective biochemical interpretation of SMIRFE-generated formulas derived from lipidomics-focused FT-MS spectra.

Manually constructing rules that map elemental molecular formulas to one or more lipid categories is a daunting proposition that would most likely result in rules that are fragile, incomplete, and incorrect. Previous success at predicting potential lipid features automatically using ratios of heteroatoms (Brockman *et al.*, 2018) strongly suggests that a similar technique could be applied to predict lipid categories from assigned molecular formulas. Fortunately, the prediction of lipid category from molecular formulas can be stated as a supervised machine learning problem.

In supervised machine learning, models are constructed that can predict a ‘label’ for a given input. Inputs are represented by feature vectors which are a mathematical representation of features or properties of the input. For example,

predicting biological sex (i.e. male or female) from the height and weight of an individual can be stated as a supervised machine learning problem. In this example, the biological sex of the person is the label to be predicted and the height and weight are features represented in a feature vector (e.g. <5’11’’, 190lbs>).

The first step in solving many supervised machine learning problems is the collection of a training dataset, which is composed of many examples of inputs of known label. From each entry in this training dataset, a feature vector must be constructed. Continuing with the sex prediction example, this would mean obtaining records (e.g. medical or DMV records) that record the height, weight and gender of many people. Which features are included in the feature vector representation of the entry must be decided based on what information can be readily acquired about an entry (e.g. we cannot include structural information in our feature vectors as we cannot get structural information easily from SMIRFE assignments) and which information is expected to aid in classification (e.g. blood type could be included in our feature vectors for predicting biological sex but is unlikely to be helpful). Second, this training dataset is then used to fit a mathematical model that can predict label from any feature vector, based on patterns between features in the training feature vectors and their known labels. Different machine learning models vary dramatically in how they approach the fitting of this model to the training data. Once trained, the performance of these models can be evaluated and then applied to new datasets to make predictions (Kotsiantis *et al.*, 2007).

Although many algorithms exist for generating models that solve supervised machine learning problems, one of the most popular algorithms is the Random Forest algorithm (Breiman, 2001). In Random Forest, anywhere from 10 to hundreds or thousands of decision tree classifiers (Breiman *et al.*, 1984) are constructed and each decision tree considers only a subset of the feature vectors present in the training dataset. This subsetting of the training dataset is referred to as bagging. Individual decision trees can be thought of as flow charts that provide a set of rules that enable label prediction from an input vector (e.g. if height < 5'3'' and weight < 140lbs predict label "female"). The rules learned by each tree may differ as each tree was trained using different bags of feature vectors and different subsets of features. The maximum complexity of the rules represented by each tree is governed by a set of user-provided parameters, which limit how many features each tree can use and the maximum size (depth) of a tree. These parameters that control model generation are called hyperparameters. When presented with a new feature vector to classify, a Random Forest model applies the rules represented by each tree to the new feature vector. Each tree then votes for the label it predicts and the label with the most votes is the label predicted by the Random Forest model.

Random Forest has been successfully applied to other supervised machine learning problems in metabolomics (Chen and Yu, 1994) (Wang-Sattler *et al.*, 2012) and has several properties that make it ideal for our use-case. First, the bagging process implemented by Random Forest protects against overfitting and enables the direct measurement of classifier accuracy similar to explicit

cross-validation (Svetnik *et al.*, 2003) using a metric called out-of-bag accuracy (described in more detail in the materials and methods). Second, bagging and the construction of many classifiers mitigates problems that arise when training datasets are unbalanced with respect to the number of examples for each label (see Table 4.1 Number of Examples, each lipid category is not equally represented) (Svetnik *et al.*, 2003). Finally, the individual decision trees constructed by Random Forest are expected to excel at learning classification rules based on discrete data such as atom counts, which are expected in our feature vectors. For these reasons, Random Forest is an ideal candidate for the generation of models that can predict lipid category and class for SMIRFE assigned EMFs.

4.2 Materials and Methods

4.2.1 Structure of Chemically-Descriptive Feature Vectors

In our application each feature vector represents an elemental molecular formula from a metabolite database or an assigned EMF from SMIRFE. The features that comprise the feature vector was selected to be computable from only an EMF assignment (i.e. no information from MS2 or chromatography to provide structural information). These features include the atom count for each CHONPS element, the sum of atom counts for other elements, and the theoretical monoisotopic mass and individual decimal places from this mass (e.g. the tenths place, the ones place, and the tens place. These individual digit places are included in case there are patterns in the distribution of these digits that could

inform lipid category or class. To ensure that all molecular weights for all entries were comparable, every entry had its theoretical monoisotopic molecular mass re-calculated using isotope molecular masses from NIST (Wieser *et al.*, 2013) (Berglund and Wieser, 2011) . Each element atom count is an integer, but for different elements the expected atom count range can vary significantly. For example, only a handful of phosphorus or sulfur atoms are expected for any given lipid, but a fully saturated hydrocarbon of mass 1600 Daltons would have 114 carbon and 230 hydrogen. The theoretical monoisotopic mass is a floating-point number between zero and a few thousand Daltons, while each digit will be represented as an integer between 0 and 9. As a result, each feature in our feature vector will be on a different scale. Although these could be normalized to remedy the differences in scale, which is a requirement for some machine learning algorithms, the Random Forest algorithm does not have this limitation. An example feature vector representing an EMF is shown in Figure 4.1.

Palmitic Acid $C_{16}H_{32}O_2$

1. Calculate Theoretical Monoisotopic Mass

$$\text{Monoisotopic Mass} = \#C * M_{12C} + \#H * M_{1H} + \#O * M_{16O} + \#N * M_{14N} + \#P * M_{31P} + \#S * M_{32S}$$
$$16 * 12 + 32 * 1.0078250321 + 2 * 15.9949146196$$
$$\mathbf{256.240230266}$$

2. Calculate Unsaturation

$$\text{Unsaturation} = 4*\#C + 3*\#N + 2*\#O + 6*\#P + 6*\#S - (\#H + \#X) - 2*(\#C+\#N+\#O+\#P+\#S-1)$$
$$4*16+2*2 - (32) - 2*(16+2-1)$$
$$2$$

3. Build Feature Vector

Feature = <mass, tens, ones, tenths, unsaturation, #X, #H+#X, #C, #N, #O, #P, #S, #H>
= <**256.240230266**, 5, 6, 2, 2, 0, 32+0, 16, 0, 2, 0, 0, 32>
= <**256.240230266**, 5, 6, 2, 2, 0, 32, 16, 0, 2, 0, 0, 32>

Figure 4.1: Example Feature Vector

Example construction of a feature vector for the EMF $C_{16}H_{32}O_2$, corresponding to palmitic acid - In a real-world application, the EMF would be provided from an assignment method such as SMIRFE and the compound it represents may not be known. The first step in constructing the feature vector is to calculate the theoretical monoisotopic mass for that EMF. Calculating the theoretical mass for an EMF rather than relying upon the observed mass for the corresponding spectral feature, eliminates the potential confound of mass error at the classification step. Using the monoisotopic mass is necessary so that isotopologues of the same EMF can be classified using the same classifiers. In the second step, the number of hydrogens missing in the formula due to unsaturation is calculated. Finally, the theoretical monoisotopic mass, the number of missing hydrogens, and the EMF are used to construct the feature vector. The coloring and bolding of the numbers in the example feature vector reflect the sources of these values.

4.2.2 Derivation and Organization of Training Datasets

In addition to the selection of proper feature vectors and the selection of an appropriate machine learning algorithm, the quality of a machine learning model depends heavily on the quality of the training data from which the model is constructed. Training datasets should be large, contain examples of both true positives and true negatives, and cover most of the expected feature space. Additionally, training data must be organized in the appropriate manner. In this case, the training data should have the training inputs mapped to both high-level lipid categories (e.g. glycerolipid, phospholipid, etc.) and further subdivided into more specific “main classes” (e.g. monoradylglycerols, eicosanoids, secosteroids, etc.).

The LIPIDMAPS database is the largest lipid-specific repository of metabolite structures and every entry in LIPIDMAPS is assigned to both a high-level lipid category and a lower-level lipid class. There are 7 lipid categories, which are further subdivided into 79 distinct classes. Each entry represents either an observed lipid or a predicted lipid and contains an elemental formula for that lipid and its assigned lipid category and lipid class. Therefore, entries from LIPID MAPS are sources of true positives for our training dataset. LIPID MAPS is also subdivided into two databases: the LIPID MAPS Structure Database (LMSD) and the LIPID MAPS In-Silico Structure Database (LMISSD). The LMSD contains both manually verified and computationally generated lipids and is freely

available for download, while the LMISSD is composed completely of computationally generated lipids. Unlike the LMSD, the LMISSD is not directly downloadable and a web scraper written in R (Ihaka and Gentleman, 1996) using the RSelenium package (Harrison, 2016) was used to extract every LMISSD entry with its lipid category, lipid class, and molecular formula. We downloaded the LMSD in September, 2018, which contained 42,004 entries. We webscraped the LMISSD in September, 2018, obtaining 1,131,106 entries.

However, true positives are only one half of a training dataset. True negatives are also needed for the construction of robust models. In this case, a true negative is a biological formula that is not a lipid. The human metabolome database (HMDB) (Wishart *et al.*, 2007) contains many examples of biological formulas of known class and is freely downloadable. By filtering out and removing known lipids from the HMDB, a set of false negatives were constructed. These entries, of course, do not have a lipid category or lipid class assigned to them, so an extra category and class called ‘non_lipid’ was assigned to these entries. The downloaded version 4.0 of the HMDB on September, 2018, which contained 114,089 entries, with 22,657 entries being non-lipids.

Since in-silico generated lipids may not necessarily exist in biological systems, it is prudent to construct two example training datasets: HMDB non_lipids + LMSD (referred to as LMSD training set) and HMDB non_lipids + LMSD + LMISSD (referred to as LMISSD training set). Since isomers of lipids can have the same molecular formula but have a different structure that can even belong to different lipid categories and lipid classes, each training dataset was

deduplicated by mapping each formula to all observed lipid categories and classes for each formula. A large portion of the entries in both the LMSD and LMISSD are isomers of other entries of the same lipid class and category. The final LMSD + HMDB non_lipid training dataset resulted in 16,215 unique entries as compared to 30,692 for the LMSD + LMISSD + HMDB non_lipid training dataset.

4.2.3 HMDB-Derived Molecular Formula Convex Hull Construction

Given a set of points in an N-dimensional space, there exists exactly one unique polygon of minimal volume that encloses all the points in the space. In 2-dimenions, a convex hull is the shape formed by stretching a rubber band around a collection of points. In three dimensions, it is the same as tightly wrapping a collection of points in wrapping paper. A convex hull can be used to aid in the generation of formulas that are similar to known metabolite formulas in elemental composition, but are not known metabolite formulas.

For every formula in the HMDB composed only of CHONPS elements and with a molecular weight below 1600 Daltons, each formula was mapped to a single 6-dimensional integer tuple that represents that entry's EMF (e.g. (6,12,6,0,0,0) represents the EMF $C_6H_{12}O_6$). This is identical to the integer lattice constructed to enumerate EMFs in the SMIRFE algorithm described in Chapter 3. The convex hull enclosing these points was then calculated using the Python implementation of the qhull algorithm (Barber *et al.*, 1996). Each point enclosed by the convex hull represents an EMF that is similar to a known metabolite's

EMF. All these enclosed points were enumerated and converted back to EMFs to generate a set of EMFs that are similar to known metabolite EMFs. Feature vector representations of each EMF were then generated that can be classified using our machine learning models.

4.2.4 Experimentally-Derived Molecular Formulas from Human Lung Cancer Samples

Paired cancer and non-cancer tissue samples used for this analysis are described in Appendix 1. SMIRFE was used to assign all samples in the dataset. SMIRFE uses patterns in the intensity ratios of suspected isotopologues of the same elemental molecular formula and how these patterns compare to predicted intensity ratios based on isotope natural abundances. SMIRFE assignments were generated for 192 samples up to 1600 *m/z*, representing 127,338 unique EMFs. Feature vectors were generated for each of these EMFs that can be classified using our machine learning models.

4.2.5 Machine Learning Classifier Construction

Two approaches were explored in regards to the organization of models trained for lipid category and lipid class prediction. The first of these was the construction of a pair of monolithic models for the lipid categories and lipid classes. In this formulation, each model takes each feature vector and attempts to classify it into one of the many lipid categories or classes *at once* (*i.e.* attempts a multi-class classification - is this feature vector a glycerolipid or a sterol or a sphingolipid or a…). The alternative approach was to train a collection of many

binary classifiers for each lipid category and class (*i.e.* attempts to classify into only one category or class at a time - is this feature vector a sterol – yes or no?). In this formulation, models can be arranged as a hierarchy, with class models organized under their respective category models (e.g. the steroid class is under the sterol category). The differences in the organization of these models is shown in Figure 4.2.

Both the monolithic models and the hierarchical models were constructed using the Random Forest implementation from sklearn (Pedregosa *et al.*, 2011), a well-implemented and commonly used machine learning library in Python. Both sets of models were trained using the default hyperparameters for Random Forest, except the number of trees was set to 500. Training the monolithic models was straightforward. For the category model, a Random Forest model was trained using all feature vectors in the training data and their lipid categories. For the class model, a Random Forest model was trained using all feature vectors in the training data and their lipid classes.

The training of the hierarchical models was more complex. For a given category or class, all feature vectors in the training dataset were divided into two sets. One set are all feature vectors that map to that category or class (true positives), the other were all other feature vectors that did not map to that category or class (true negatives). Using these examples of true positives and true negatives, a Random Forest model is trained to perform this binary classification (i.e. for a class X is the given feature vector an instance of X – yes

or no). This process is repeated for all categories and classes in the training dataset.

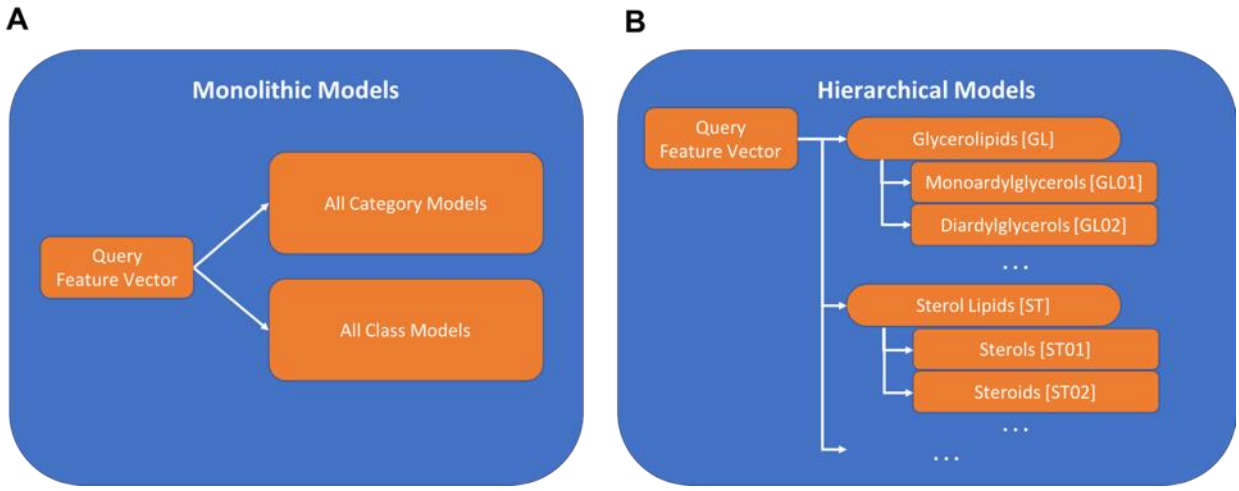

Figure 4.2: Organization of Hierarchical and Monolithic Models
In a monolithic organization, there exists one model for classifying feature vectors into lipid categories and another model for lipid classes (Panel A). This organization is simpler with fewer models to train as compared to the hierarchical organization of models (Panel B). In the hierarchical organization, there are more models in total, but the class models are organized under their respective category model.

4.2.6 Evaluation of Lipid Classification Performance

Although Random Forest attempts to build a model that accurately predicts labels (i.e. in this case lipid categories or classes) from input feature vectors, these models are not perfect. Sometimes the predicted label for an input vector is the correct label (a true positive) and at other times the predicted label for an input feature vector is incorrect (a false positive). The number of true positive and false positive results for a category or class can be used to calculate a model's precision for that class where precision equals the number of true

positives divided by the number of true positives plus the number of false positives. This can be evaluated by classifying the feature vectors in the training data and recording the number of true positives and false positives returned by the model. A high precision means that when a label was returned by a model, it was likely the correct label. Precision was calculated in this manner for every model in the hierarchical models.

Another metric for evaluating model performance is the out-of-bag accuracy. This value is calculated automatically by the Random Forest implementation in sklearn during training. Recall that each tree in the Random Forest is not trained using the entire training dataset. Instead each tree is only trained on a subset of the entire training dataset (this is the “bag” of samples referred to by “out-of-bag”). The out-of-bag accuracy is the average accuracy for all trees in the Random Forest model on each training feature vector considering **only** the trees in the Random Forest model that did not use that feature vector during training. In this case, accuracy is simply the percentage of predicted labels that matched the known label (e.g. what percentage of feature vectors representing Glycerolipids [GL] that were classified correctly). Out-of-bag accuracy was calculated for both the monolithic models and the hierarchical models.

By these metrics, models that have high precision produce very few incorrect classifications, but may not produce many classifications total. Models with high out-of-bag accuracy produce many correct classifications at a global level but may do poorly on any individual label. Therefore, models with both high

precision and high out-of-bag accuracy are desirable as they produce many classifications with few false positives.

4.3 Results

4.3.1 Monolithic Classifier Performance on Training Datasets

Using the LMSD + HMDB_non_lipid dataset, the performance of a monolithic classifier for lipid category and lipid class was tested. Even with 500 trees, the monolithic random forest models were only able to achieve an out-of-bag accuracy of 74.9% for lipid category and 87.3% for lipid class. Including the LMISSD resulted in an out-of-bag accuracy of 83.1% for lipid categories and 80.1% for lipid class. In both datasets, the presence of many non-lipid entries inflates the lipid class accuracy as all non-lipid entries map to the non-lipid class. Although monolithic classifiers may have the theoretical advantage of being simpler to implement, train, and deploy, their usefulness is limited by their relatively lower classification performance.

4.3.2 Hierarchical Classifier Performance on Training Datasets

For both training datasets (LMSD + HMDB_non_lipid and LMSD + LMISSD + HMDB_non_lipid), the out-of-bag accuracy and precision for each lipid category are shown in Table 4.1, while the class level results are shown in Supplemental Table 4.1. For all categories, the LMSD + HMDB_non_lipid trained models achieved high precision and high accuracy for all lipid categories. Classification performance for lipid class varies between classes but is in general

excellent for classes with enough examples. The LMISSD-trained models achieved similar precision and accuracy for all categories (Table 4.2) and classes (Supplemental Table 4.2) compared to the LMSD-trained models. Although individually high accuracy or high precision would not necessarily indicate a well-trained model, the combination of high accuracy and precision across the models implies that the combined classification performance is robust and can be effectively applied to experimentally-derived molecular formulas.

Table 4.1 LMSD + HMDB non_lipid Model Performance (Category)

LMSD + HMDB_non_Lipid Model Performance (Category)			
Category	**Precision**	**Out of Bag Accuracy**	**Number of Examples**
Fatty Acyls [FA]	0.841	0.901	2031
Glycerolipids [GL]	0.996	0.995	532
Glycerophospholipids[GP]	0.995	0.996	1886
Polyketides [PK]	0.767	0.885	1376
Prenol Lipids [PR]	0.989	0.971	473
Saccharolipids [SL]	1.000	0.998	102
Sphingolipids [SP]	0.999	0.993	1404
Sterol Lipids [ST]	0.934	0.972	824
non_lipid	0.928	0.799	7587

Table 4.2 LMSD + LMISSD + HMDB_non_Lipid Model Performance (Category)

LMSD + LMISSD + HMDB_non_Lipid Model Performance (Category)			
Category	**Precision**	**Out of Bag Accuracy**	**Number of Examples**
Fatty Acyls [FA]	0.838	0.939	2031
Glycerolipids [GL]	0.996	0.993	2715
Glycerophospholipids [GP]	0.979	0.980	9766
Polyketides [PK]	0.773	0.934	1376
Prenol Lipids [PR]	0.989	0.983	473
Saccharolipids [SL]	1.000	0.999	102
Sphingolipids [SP]	0.976	0.976	3089
Sterol Lipids [ST]	0.931	0.983	824
non_lipid	0.930	0.882	7587

Tables 4.1 and 4.2: Classifier Accuracy and Precision Analyses – Most lipid category and class models achieved excellent accuracy above 90% and excellent accuracy above 93%. Polyketides had worse accuracy and precision in both models (88.5% and 76.7% accuracy in precision in the LMSD models) and fatty acyls had high accuracy but lower precision (90.1% accuracy and 84.1% precision). The polyketides represent a very diverse set of structures compared to other lipid classes which explains this discrepancy. The number of examples of each category highlights the unbalanced nature of this dataset and motivated the use of Random Forest for these models. Inclusion of the LMISSD (Table 4.2) provided no additional examples of Fatty Acyls, Polyketides, Saccharolipids, or Sterol Lipids and had minimal effect on the precision and accuracy of the models.

4.3.3 Hierarchical Classifier Performance on Theoretical Molecular Formulas

Brute force enumeration and testing of all points within the convex hull constructed around all CHONPS-only molecular formulas in the HMDB identified 110,857,519 formulas. While a brute force approach was computationally expensive, requiring several thousand hours of CPU time, it was necessary due to memory constraints with more complex methods. Classifying these formulas with the LMSD + HMDB_non_lipid models resulted in 8.34% assigned to a lipid category or class, 67.31% assigned to non_lipid and 26.34% receiving no classification. Results for each category are summarized in Table 4.3. The LMISSD models predicted lipid categories for 13.44% of the provided formulas, 66.36% were assigned to non-lipids and 25.08% were assigned to no category (Table 4.4). Note, due to the ability to assign formulas to multiple categories, the percentages do not sum to 100%. The LMISSD models predicted 4 of the 7 categories more frequently than the LMSD models.

Table 4.3 LMSD + HMDB non_lipid Model Performance for Convex Hull (Category)

LMSD + HMDB_non_lipid Model Performance for Convex Hull (Category)		
Category	**Predictions**	**% of Hull Formulas**
Fatty Acyls [FA]	475,516	0.429
Glycerolipids [GL]	8,205	0.007
Glycerophospholipids [GP]	1,145,418	1.033
Polyketides [PK]	84,333	0.076
Prenol Lipids [PR]	18,684	0.016
Saccharolipids [SL]	6,708	0.006
Sphingolipids [SP]	7,494,579	6.761
Sterol Lipids [ST]	18,643	0.017
non_lipid	74,621,680	67.31
no category	29,202,459	26.34

Table 4.4 LMSD + LMISSD +HMDB_non_lipid Model Performance for Convex Hull (Category)

LMSD + LMISSD +HMDB_non_lipid Model Performance for Convex Hull (Category)		
Category	**Predictions**	**% of Hull Formulas**
Fatty Acyls [FA]	393,314	0.354
Glycerolipids [GL]	56,116	0.051
Glycerophospholipids [GP]	1,735,925	1.566
Polyketides [PK]	118,968	0.107
Prenol Lipids [PR]	15,881	0.014
Saccharolipids [SL]	2,795	0.002
Sphingolipids [SP]	12,568,226	11.34
Sterol Lipids [ST]	15,670	0.014
non_lipid	73,562,707	66.36
no category	27808607	25.08

Tables 4.3 and 4.4: Model Performance for Theoretical Molecular Formulas - The formulas within the convex hull surrounding the HMDB metabolites represent a very large set of possible metabolites formulas. Lipid categories were predicted for every formula within the hull. For all categories, more formulas were predicted for each category than existed in the training dataset, indicating that the models have generalized beyond the training dataset. The

extent of this generalization varied depending on the training dataset. For example, saccharolipids (the category with the smallest number of examples in the training dataset) was predicted more frequently in the LMSD based models than in the LMISSD models, while sphingolipids were more frequently predicted in the LMISSD trained models. Although the distribution of predicted lipid categories varies slightly between the two models, the overall trends are comparable.

4.3.4 Hierarchical Classifier Performance on Experimentally-Observed Molecular Formulas

The distribution of the assigned lipid categories on molecular formulas enumerated from a human lung cancer FT-MS dataset is shown in Tables 4.7 and 4.8 for the LMSD based classifier. SMIRFE generates many possible assignments for each peak at higher *m/z* as the number of possible formulas increases dramatically with increasing *m/z*. As a result, a relatively small percentage of formulas are assigned to a lipid category but many peaks have at least one formula that was assigned to a lipid category. For the LMSD models, the ability to predict lipid category and class drops substantially after about 1200 *m/z*. This is due to the low number of entries in the LMSD at higher *m/z*.

When the masses of the peaks are shifted by +21 *m/z* to mimic a gross miscalibration error, the number of SMIRFE assignments is increased, from 127,338 to 131,690 total formulas and the number of predicted lipids increases as well from 32,688 to 34,755 (Tables 4.5 and 4.6). This result implies that the lipid classifier cannot be used alone to screen out bad assignments when lipids are expected, instead other orthogonal data must be used to verify the quality of the assignments and select the correct assignments prior to classification.

Table 4.5 LMSD + HMDB_non_lipid Model Performance for Unshifted Assignments

LMSD + HMDB_non_lipid Model Performance for Unshifted Assignments		
Category	**Predictions**	**% of Assigned Formulas**
Fatty Acyls [FA]	**639**	0.502
Glycerolipids [GL]	795	0.624
Glycerophospholipids [GP]	8062	6.331
Polyketides [PK]	28	0.022
Prenol Lipids [PR]	1054	0.827
Saccharolipids [SL]	166	0.130
Sphingolipids [SP]	21586	16.952
Sterol Lipids [ST]	358	0.281
non_lipid	54389	42.71
no category	40683	31.95

Table 4.6 LMSD + HMDB_non_lipid Model Performance for Shifted Assignments

LMSD + HMDB_non_lipid Model Performance for Shifted Assignments		
Category	**Predictions**	**% of Assigned Formulas**
Fatty Acyls [FA]	**258**	0.1951
Glycerolipids [GL]	923	0.7001
Glycerophospholipids [GP]	9517	7.227
Polyketides [PK]	37	0.0281
Prenol Lipids [PR]	1160	0.8808
Saccharolipids [SL]	233	0.1769
Sphingolipids [SP]	22370	16.99
Sterol Lipids [ST]	257	0.1952
non_lipid	51863	39.38
no category	45663	34.67

Tables 4.5 and 4.6: Model Performance for Experimental Molecular Formulas- SMIRFE assignments were generated for the NSCLC dataset described in Appendix 1 (Sample D). SMIRFE assignments are generated in an untargeted manner without using a database of known lipids. For the peak masses across all peaklists, 127,338 total EMFs were assigned and then classified. 32,688 total lipid category classifications were made with the most commonly assigned categories being glycerophospholipids and sphingolipids. A similar result was observed in the convex hull results as well, potentially indicating that this is an artifact of the classification method or possibly that these lipid categories are

much more diverse than other categories. When each peak was shifted by 21 *m/z*, roughly 3% more formulas were assigned and 6% more lipids were classified. This small relative increase in SMIRFE assignments is likely due to the increased search space density with a 21 *m/z* shift. More importantly, the large number of artifactual assignments reflects the necessity of high-quality data prior to classification. Methods that can predict high quality assignments correctly are not necessarily protected from the effects of low-quality spectral data that can cause misassignment.

4.3.5 Cross-Sample Assignment Correspondence Improves Assignment Quality

Limits in mass resolution and intensity resolution and the immense size of the search space considered by SMIRFE at high masses leads to ambiguous assignments for many peaks. When mass error is present, ambiguous and incorrect assignments can be generated. However, the correct assignment for a peak should be assigned more consistently for a consistently observed feature in the dataset. Therefore, how well an assignment corresponds across samples in a dataset is a potential avenue for selecting high quality assignments. Figure 4.3 shows histograms of assignment correspondence for elemental molecular formulas derived the spectra of the lung cancer dataset. Higher correspondence is observed in the unshifted assignments versus the shifted assignments and the number of shifted high-correspondence assignments are fewer at lower *m/z*.

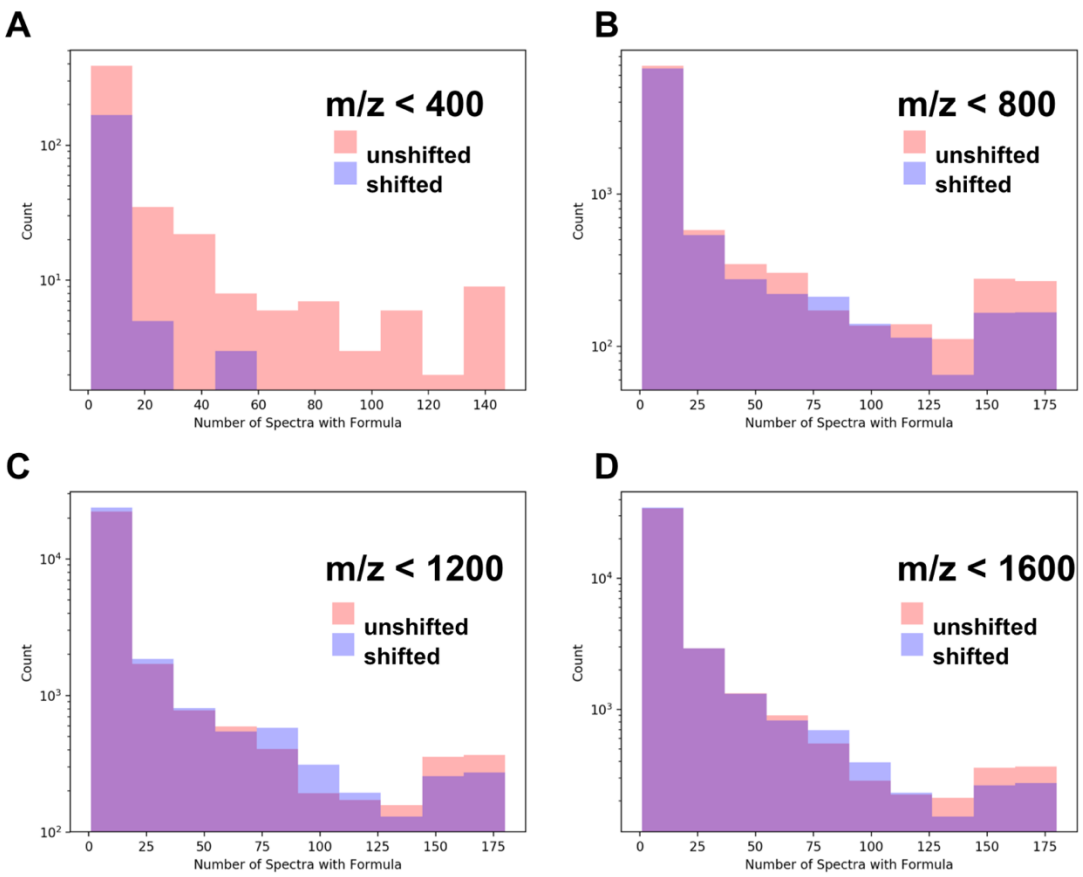

Figure 4.3: Assignment Correspondence Histograms
Correct assignments are expected to occur more consistently within a set of samples than incorrect assignments. As shown in Panel A, below 400 *m/z*, very few assignments are made in the shifted spectra and none of the assignments correspond. As *m/z* increases (Panels B-D), shifted spectra have increasingly more assignments and by chance some of these assignments correspond in multiple samples. However, at up to 1200 *m/z*, there are more well corresponding formulas in the unshifted assignments than in the shifted assignments. These results imply that assignment correspondence can be used to select high correct assignments up to a given mass cutoff. In this dataset, this appears to be less than 1200 *m/z*.

4.4 Discussion

4.4.1 *Classifier Organization and Performance*

As shown in the results, a monolithic, multi-lipid-class predictive model failed to achieve top performance for the task of classifying assigned molecular formulas into lipid categories and classes. I hypothesize that this is due to the inability of a single classifier to represent all these boundaries completely and accurately. This single classifier must not only learn how to separate lipids from non-lipids, but it must also subdivide the lipid feature space into discrete spaces representing each category and further subdivide these category spaces into class spaces. Much of this subdivision can be done explicitly during training. For example, the diacylglycerols are a sub-class of the larger category of glycerolipids and a less powerful classifier can easily identify the diacylglycerols from other glycerolipids when it must only learn that single decision boundary. As a result, our organization of more specialized predictive models had superior performance. Initially, this behavior can seem counterintuitive but is consistent with the concept of ensemble learning from machine learning where collections of more specialized classifier models often outperform fewer larger classifier models when properly organized. The hierarchy of models that are constructed mirror how a human would approach the classification problem (Fahy *et al.*, 2005). For example, if a molecular formula is known not to be of the sphingolipid category, a human will not attempt to assign this formula to a sphingolipid class; however, a monolithic model will attempt to do so. This wastes computational

power and increases the likelihood of incorrect prediction of both class and category.

The final models produced by our tool achieved both high accuracy and high sensitivity on the training dataset. Of course, performance on training data does not paint a complete picture of model performance, but for Random Forest which implements bagging, these metrics predict performance on inputs similar to the training data (Janitza and Hornung, 2018). Models with both high accuracy and high sensitivity are unlikely to produce incorrect lipid assignments, but may be overly conservative and fail to generate a non_lipid assignment for some inputs. While this behavior is undesirable, it is preferable to less conservative models that will yield many incorrect lipid category and class predictions.

4.4.2 *LMSD Versus LMISSD Trained Models*

One method for improving the performance of a machine learning model is to provide larger amounts of training data, which in turn enables more informed and more accurate decision boundaries to be determined. For this reason, models were trained using both the LMSD and LMISSD, which has nearly 25 times the number of entries as the LMSD. However, LMISSD trained models did not offer substantially improved performance as compared to the LMSD-only models on the training datasets. Although the LMISSD contained many entries, the input training set only doubled in size after isomeric entries were removed, implying that little information was added regarding the distribution of formulas in lipid category or lipid class space. Another possible explanation for this observation is that the LMISSD contains substantially more entries, but for only 4

out of the 7 categories in the LIPIDMAPS database and that the decision boundaries for these categories were already well-determined by the LMSD only models.

4.4.3 *Classifier Generalization*

A benefit that machine learning models have over traditional database lookups are their ability to infer rules that can be applied to never observed inputs to make accurate predictions. This ability was demonstrated with both the LMSD- and LMISSD-constructed models. Both models produced lipid category and class predictions for experimental and theoretical molecular formulas not present in the training dataset. However, the generalizability of the models depends heavily on the quality and size of the input dataset (Figure 4.4).

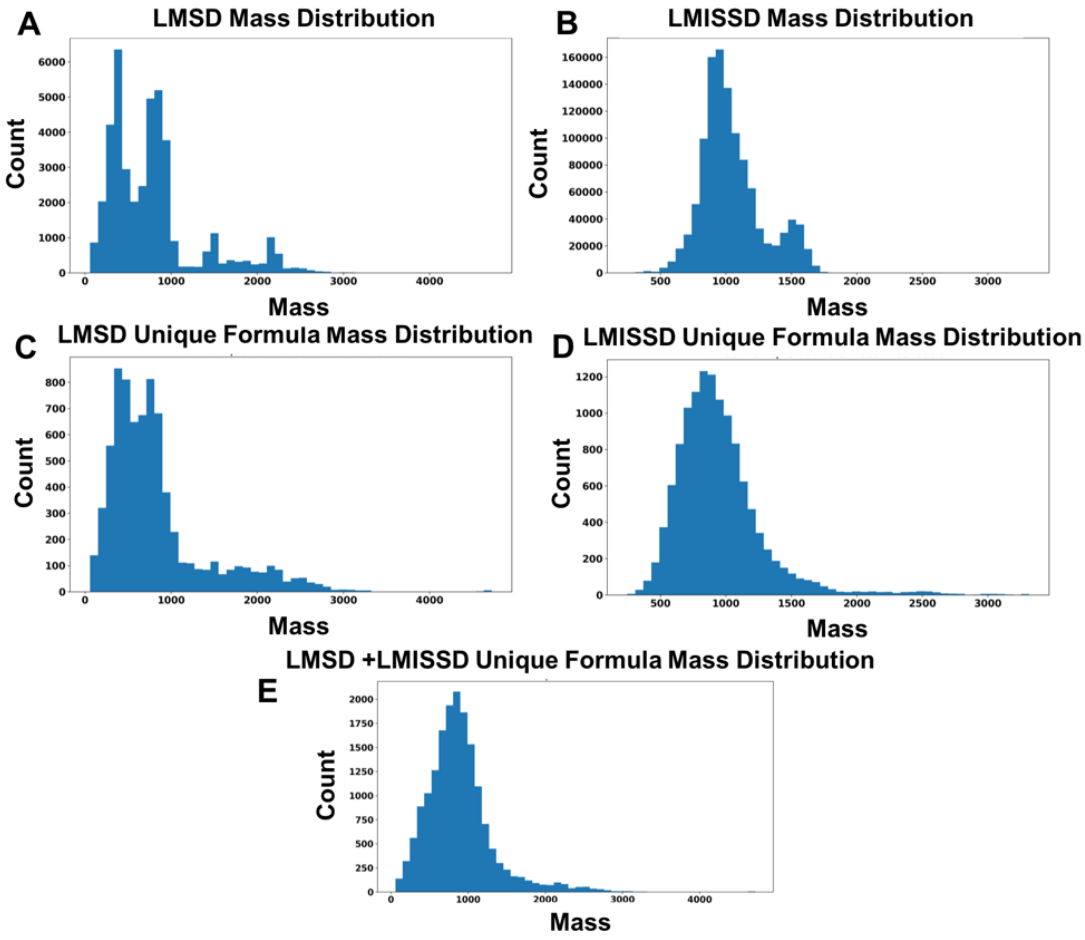

Figure 4.4: Mass Limitations of Training Datasets
Both the LMSD and LMISSD are heavily biased towards lipids with a mass below 1200 Daltons (Panels A and B respectively). This effect becomes clearer once entries are deduplicated to yield only unique formulas (Panels C and D). Some entries exist out to 3000+ Daltons, but the bulk of the formulas still reside in the sub 1200 mass range (these high mass entries may represent lipid conjugates that are not lipids for our purposes since they will not be observed in mass spectra up to 1600 *m/z* unless multiply charged but still represent a difference in the training datasets). When combined, the bias is still present in the combined set of unique formulas (Panel E). Additionally, the LMISSD does not have entries from all the lipid categories.

Despite having similar performance on the training datasets, the LMISSD and LMSD trained models showed distinct behavior on the convex hull metabolites. The LMISSD assigned many more sphingolipids than the LMSD

models and in general, the categories with more examples in the training dataset were more frequently predicted. This could be due to a bias in the trained models from the unbalanced training data or could reflect the relative amount of structural diversity possible within each class, i.e. the number of possible sphingolipid formulas might truly be larger than the number of possible sterol formulas. However, the percentage of hull formulas predicted for each category was similar between the two models, implying that they are overall very similar. Both models predicted roughly the same number of not lipid formulas implying that the overall lipid versus non-lipid decision boundaries of the two models are very similar. Discrepancies between the two models can also be attributed to the presence of predicted lipids in the LMISSD that do not exist – this could confuse classifiers if the predicted lipids and the validated lipids suggest different decision boundaries.

Ultimately the ability of both models to make accurate predictions will be improved with larger training datasets. With more examples that more exhaustively span lipid formula space, the more accurate and generalizable the models constructed using these same methods will become. However, given the marginal improvement with a doubling of the training dataset from the LMSD versus LMISSD, improvements may be marginal without a vastly larger training dataset.

4.4.4 *Mass Error and Classification Results*

Ideally, a substantial mass error would result in no formulas being assigned by SMIRFE or that the assigned formulas fail to classify. As shown with

our NSCLC dataset, a large mass error does not eliminate all assignments nor completely abolish our ability to classify the resulting, almost certainly incorrect, assigned formulas.

Given the very large search space that an untargeted tool must search to generate assignments, almost any *m/z* has many possible assignments, given the theoretical molecular formula search space. Since a systematic error does not change the mass difference between isotopologues, patterns of isotopologues for these incorrect formulas can still be identified and assigned. Thus, without extremely high mass resolution to restrict the set of possible assignments considerably, which still may not be effective (Kind and Fiehn, 2006), a constant mass error will still produce assignments. Furthermore, current variance in peak intensities is not low enough to prevent artifactual assignment at higher *m/z*.

As was seen from the classification results of the convex hull formulas, approximately a quarter of the generated formulas appeared to be lipids to the models. This could reflect the true distribution of lipids in possible formula space, but more likely it represents limitations of our models. Nonsense formulas that can arise from *m/z* error or from the convex hull method cannot be properly learned as they are very different from the training set data. Although the ability of our models to produce no classification for an input feature vector protects against this effect, it is not perfect. The same models that learn **real** (biochemically relevant) metabolite formulas correctly may fail to properly handle

nonsense formulas that SMIRFE can assign to peaks with high mass error and noise or artifactual peaks.

Therefore, lipid classification alone should not be used to filter out features in datasets. Information such as how many times a formula is observed across a dataset and observed correlation between features classified to the same lipid category and/or class should be included as well. Features considered trustworthy by other methods can then be classified for further analysis. Since relatively small mass errors can introduce correct assignments (Kind and Fiehn, 2006), high cross-sample formula correspondence is an indication (though not definitive proof) of accurate (or at the very least consistent) assignments.

4.4.5 *Implications for Experimental Design*

The ability to predict lipid category and class from molecular formula assignments without the need for cross-validated *metabolite* assignments, enables simpler experimental designs as the volume of information needed to perform class or category level comparisons is lessened. As molecular formula can be assigned from direct infusion FT-MS MS1 spectra directly and in a cross-validating manner, chromatography and other cross-validation information is not necessary for class or category level comparisons when using these models. However, the quality of the analyses will depend heavily on the quality of the assigned molecular formulas.

SMIRFE leverages patterns in the relative heights of isotopologue peaks for the same elemental molecular formula to determine what molecular formulas

best explain features observed in a spectrum. Although SMIRE is not necessarily limited to only high-end mass spectrometers such as FT-MS instruments, only these instruments provide enough mass accuracy and resolution to observe and characterize relevant sets of isotopologues. This restriction is becoming increasingly less relevant as high-performance spectrometers become more available. Additionally, SMIRFE and subsequent lipid prediction does not enable the robust assignment of metabolite structures to spectral features and this will still require additional information from orthogonal experiments.

4.5 Conclusions

Untargeted lipidomics has the potential to produce more informative datasets that will aid in the construction of more complete models of cellular metabolism. This in turn enables a better understanding of both healthy and disease processes. A necessary step in many of these analyses is the assignment of lipid category or class to an observed lipid feature. When multiple orthogonal sources of information are available (i.e., MS + chromatography, NMR + chromatography, MS/MS), lipid category and class assignment can be inferred from trustworthy metabolite assignments.

The application of machine learning algorithms enables the construction of models that can accurately and precisely assign lipid labels to observed spectral features that have been assigned to a molecular formula. Unlike other approaches that leverage metabolite databases directly for lipid assignment, these models have the capacity to infer lipid category and class for entries not

present in existing databases. This capacity is essential for untargeted metabolomics experiments as database incompleteness can lead to a biasing of lipid classification and in turn biological interpretation. Since these models are informed by the existing metabolite databases during training, their capacity to compensate for database incompleteness is not unlimited as observed with our LMSD informed models having limited efficacy at higher mass ranges. The inclusion of additional sources of empirically observed lipids in these mass ranges may extend the useful mass range of this methodology. LMISSD-informed models did not suffer from this limitation, but had decreased accuracy and specificity, potentially attributable to unrealistic entries in the LMISSD.

The high out-of-bag accuracy and precision achieved by these models on the training datasets suggest that the predicted lipid categories generated by these models on experimentally-derived molecular formulas are likely to be trustworthy. The distribution of predicted lipid categories on the convex hull formulas, the majority of which are nonsensical formulas, demonstrates the conservativeness of our models for all lipid categories except sphingolipids. Thus, machine learning-based approaches will allow for more untargeted lipid profiling analyses than existing database-centric methods, even with the more limited data that can be acquired using direct injection MS1 alone. Similar methods could be applied to the classification of other major types of biomolecules or to identify potential contaminants or non-biological compounds detected in complex biological samples. However, the quality of the predictions made by such methods remains limited by the ability to generate high quality

assignments in an unbiased manner for high *m/z* ranges that a relevant to lipid profiling. Methods such as SMIRFE combined with cross-sample correspondence provides a potential avenue to generating such assignments.

CHAPTER 5. CLINICAL IMPLICATIONS OF DIFFERENTIAL LIPID EXPRESSION IN NON-SMALL CELL LUNG CANCER (NSCLC)

5.1 Introduction

Lung cancer remains the most common cause of cancer death worldwide (Kanitkar *et al.*, 2018) with approximately 85% of newly diagnosed lung cancers belonging to the non-small cell lung cancer subtype (Molina *et al.*, 2008). The high mortality of lung cancer and NSCLC is partially explained by the relative asymptomatic progression of disease resulting in most patients presenting with advanced-stage disease at the initial diagnosis. Although improvements have been made both in the treatment, prevention, and early diagnosis of NSCLC, 5-year survival rates for NSCLC remain low (16.4%) (Kanitkar *et al.*, 2018).

Current treatment options for NSCLC vary depending on disease stage at diagnosis and the presence of one or more genetic markers. For low stage disease, surgery remains the most common and most effective treatment option (Uramoto and Tanaka, 2014), especially when combined with chemotherapeutic drugs (Betticher, 2005). For advanced stage disease, chemotherapy (Sandler *et al.*, 2000) (Pirker *et al.*, 1995) (Wozniak *et al.*, 1998) represents the primary first-line treatment option despite relatively poor improvements in patient survival, partially due to the development of chemotherapy resistance (Chang, 2011). Additionally, radiotherapy is also often used for inoperable NSCLC (Vansteenkiste *et al.*, 2013).

Many of the recent advances in the treatment of NSCLC have focused on the development of targeted therapies that inactivate one or more specific

biomarkers expressed by certain subtypes of NSCLC. Among these targets are epidermal growth factor receptor (EGFR) (Paez *et al.*, 2004) (Shepherd *et al.*, 2004), which is often mutated in NSCLC tumors in patents of East-Asian origin (Chang *et al.*, 2006), vascular endothelial growth factor (VEGF) (Piperdi *et al.*, 2014) (Ferrara *et al.*, 2005), and anaplastic lymphoma kinase (ALK) also known as ALK tyrosine kinase receptor (Crino *et al.*, 2011) (Kim *et al.*, 2014). These targeted therapies offer more effective treatment options as compared to traditional chemotherapy and with decreased side-effects, but not all NSCLC patients express these biomarkers and thus are not candidates for these treatments. In addition to these targeted therapies, drugs that target the PD-1/PD-L1 immune checkpoint to enhance anti-cancer T-cell response are a promising treatment for advanced NSCLC (Sunshine and Taube, 2015). While not side effect free, this approach offers improvements in survival and much fewer side-effects than traditional chemotherapy (Sgambato *et al.*, 2016). However, overall survival of advanced NSCLC remains poor (Spigel *et al.*, 2015).

Although many NSCLC patients fail to express these particular biomarkers, the observation that metabolic alterations are common in all human cancers implies that there are measurable and perhaps drug targetable differences between the metabolism of NSCLC cells and their non-transformed counterparts (Hanahan and Weinberg, 2011) (Ray and Roy, 2018) (Sellers *et al.*, 2015b). The process by which cancer cells acquire these altered metabolic phenotypes is known as metabolic reprogramming and is now considered to be a hallmark in the development of cancer (Hanahan and Weinberg, 2011). Previous

studies have observed metabolic reprogramming in NSCLC (Sellers *et al.*, 2015b) (Hensley *et al.*, 2016).

First, glucose metabolism is markedly altered in NSCLC. Labeling studies have shown that both glycolysis and the TCA cycle are highly active in NSCLC tumors and that the rate of glucose oxidation in NSCLC tumors exceeds the rate of oxidation in surrounding non-cancer tissue (Hensley *et al.*, 2016). Maintaining these high rates of oxidation requires replenishing TCA cycle intermediates that are often shunted towards anabolic processes. Previous studies have demonstrated that this is achieved, in part, through enhanced pyruvate carboxylation (Sellers *et al.*, 2015b). Studies have also shown that NSCLC tumors also oxidize a variety of other substrates with preference for non-glucose substrates increasing with higher perfusion rates (Hensley *et al.*, 2016). The relatively low amount of ^{13}C enrichment observed in NSCLC intracellular acetyl-CoA from U-^{13}C glucose labeling experiments implies that the majority of the flux through PDH comes from non-labeled precursors (i.e. not derived from U-^{13}C glucose) such as lactate from the tumor microenvironment (Faubert *et al.*, 2017).

Additionally, lipid metabolism is known to be altered in NSCLC as well. Key enzymes for lipid metabolism such as ATP citrate lyase (ACLY) (Osugi *et al.*, 2015), fatty acid synthase (FASN) (Uramoto *et al.*, 1999) and Stearoyl CoA desaturase 1 (SCD) (Huang *et al.*, 2016) are all known to be differentially expressed in NSCLC as compared to non-cancer tissue. Increased activity of these enzymes enables the enhanced production of many lipid classes in NSCLC tumors and their overexpression correlates with clinical outcomes. For

example, FASN expression in early stage NSCLC is indicative of poor outcomes and tumor aggressiveness (Wang *et al.*, 2002) (Visca *et al.*, 2004), SCD expression promotes tumor development and aggressiveness (Noto *et al.*, 2013) (Huang *et al.*, 2016), and ACLY expression is associated with survival in young NSCLC patients but poorer survival in older populations (Csanadi *et al.*, 2015). Each of these enzymes is required for one or more metabolic steps needed to produce lipids and inhibition of one or more of these enzymes remains a promising drug target in NSCLC patients.

Sterols are the subset of lipids that share a common four ring structure and are essential for many biological processes including: membrane fluidity, hormonal signaling, inflammation, and the regulation of cellular metabolism. All endogenously produced sterols in humans are produced by the mevalonate pathway. The first step of this pathway is the condensation of acetyl-CoA with acetoacetyl-CoA to form 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA). HMG-CoA is then reduced to mevalonate by the enzyme HMG-CoA reductase and this step is the rate limiting step for the pathway. Statins are inhibitors of HMG-CoA reductase that are used clinically for the reduction of plasma cholesterol levels. Although statins are not currently recommended for the treatment of cancer, some epidemiological studies have observed that patients prescribed statins have better overall survival for a variety of cancers (Nielsen *et al.*, 2012) including NSCLC (Hung *et al.*, 2017), even at late stages (Lin *et al.*, 2016). However other studies have failed to recapitulate these findings in controlled double-blinded studies (Seckl *et al.*, 2017). Contributing to this controversary are

the many effects of statins that could potentially be therapeutic including a dose-dependent effect on angiogenesis (Weis *et al.*, 2002), anti-inflammatory effects (Diomede *et al.*, 2001), and immunomodulating effects (Sadeghi *et al.*, 2001).

Previous studies have observed that sterol production is often upregulated in many cancers and high serum cholesterol levels are correlated with the development of some cancers (Jamnagerwalla *et al.*, 2018) (Kitahara *et al.*, 2011). Increased sterol production has been attributed to both upregulation of HMG-CoA reductase (Gustbée *et al.*, 2015) and SREBP-2 (Li *et al.*, 2016b), a master regulator of sterol biosynthesis. Due to their relatively limited side effects, a potential chemotherapeutic role for statins in NSCLC could represent a significant improvement when combined with existing treatment options. Also, drugs such as nitrogenous bisphosphonates that inhibit later steps in the mevalonate pathway (Tsoumpra *et al.*, 2015) could have similar cancer chemotherapeutic potential in NSCLC.

Overall, understanding the metabolic differences between NSCLC and non-cancerous lung tissue represents a major first step in constructing more complete models of NSCLC development and ultimately the development of more effective therapeutics. Advances in mass spectrometry, namely Fourier transform mass spectrometry (FT-MS), provides simultaneous improvements in terms of sensitivity, resolution, and mass accuracy. These combined analytical capabilities provide significant analytical improvements, including the ability to resolve distinct isotopologues of detected metabolites and the detection of lower concentration metabolites. However, leveraging these analytical improvements to

generate tangible and biologically-interpretable results remains difficult. The lack of highly descriptive peak characterization, the presence of spectral artifacts in FT-MS spectra (Mitchell *et al.*, 2017), and the lack of untargeted assignment methods that are not biased due to the incompleteness of existing metabolic databases (Mitchell *et al.*, 2014) (Schrimpe-Rutledge *et al.*, 2016) have limited the ability to capitalize on the analytical advantages of FT-MS to improve our understand of NSCLC lipid metabolism. However, using methodologies presented in Chapter 2, many of these data quality problems have been minimized and the SMIRFE algorithm from Chapter 3 enables elemental molecular formula assignment of FT-MS spectra without relying upon existing metabolic databases. Additionally, the lipid classifier tool from Chapter 4 allows the prediction of lipid category to these assigned formulas.

Using our improved data processing tools, SMIRFE (Mitchell *et al.*, 2019) and the lipid category predictor, we have assigned and classified consistently observed spectral features from a large set of paired disease and non-disease samples from suspected NSCLC patients into lipid categories. Differential abundance analysis of categorized lipid features reveals a significant increase in sets of sterols and glycerolipids and a significant decrease in glycerophospholipids and sphingolipids in disease versus non-disease tissue. The increase in sterol lipids was most interesting as it suggests a possible role of existing pharmaceutical such as statins and nitrogenous bisphosphonates in the treatment of NSCLC that is consistent with previous epidemiological studies regarding the use of the drugs in NSCLC.

5.2 Materials and Methods

5.2.1 Description of Paired Human Suspected NSCLC and Non-Disease Tissue Samples

Paired disease and non-disease tissue samples were acquired from patients with suspected stage I or IIa primary NSCLC at both the University of Kentucky and the University of Louisville. Lipid extracts were prepared from these samples and analyzed by direct infusion ultrahigh resolution mass spectrometry on two Thermo Orbitrap Tribrid Fusion instruments (Fusion 1 and Fusion 2). Additional details are provided in Appendix 1 (Sample D).

5.2.2 Peak Characterization and Assignment

Raw Thermo data files were processed using the peak characterization method described in 2.2.2. Peak characterization resulted in 100 spectra (50 patients) from Fusion 1 and 76 spectra (38 patients) from Fusion 2. SMIRFE was used to generate assignments for all spectra.

5.2.3 Categorization of Assigned Formulas

Assigned molecular formulas were classified into one or more lipid categories using the lipid classifier tool from Chapter 4. This tool uses a family of Random Forest classifiers trained on entries from the LipidMaps (Sud *et al.*, 2006) and HMDB (Wishart *et al.*, 2017) databases. Entries from LipidMaps represent examples of known lipids and their categories, while the non-lipid entries from the HMDB were used as examples of known non-lipid biological compounds. For every category in LipidMaps, using the Random Forest

implementation from sklearn (Pedregosa *et al.*, 2011), a machine learning library in Python, a classifier was trained. Each classifier was trained with default settings except that the number of trees for each forest was increased to 500. The feature vectors for each formula consisted of the element counts in the formula for CHONPS elements, the number of halogens in the formula, the number of hydrogen equivalents (number of hydrogens + number of halogens), number of unsaturation sites, the monoisotopic mass of the formula and the tens, ones, and tenths place of the monoisotopic mass. For each category, classifier accuracy and precision were estimated on the training dataset. For all categories, good accuracy and precision were achieved and are summarized in Table 4.1

5.2.4 Consistently Assigned Spectral Feature (corresponded-peak) Generation and Differential Abundance Analysis

First, corresponded peaks were identified by finding the set of peaks across all samples that represent a disjoint set of isotopologue assignments. Any two peaks that share at least one IMF assignment can be collapsed into a corresponded peak. This can be done recursively (i.e. corresponded peaks with sharing IMFs can be collapsed) until all possible corresponded peaks that can be collapsed have been collapsed. When no more collapsing can occur, each resulting corresponded peak represents a disjoint set of isotopologue assignments to one consistently assigned spectral feature. In Fusion 1, 3242 such corresponded-peaks were identified in total across all 100 spectra, with 756 corresponded-peaks present in 25% or more disease or non-disease samples. In Fusion 2, 4257 corresponded-peaks were identified across all 76 spectra, with

733 corresponded-peaks present in 25% or more disease or non-disease samples. LIMMA (Smyth, 2005) and SDAMS (Li *et al.*, 2019) were both used to identify corresponded-peaks that are differentially abundant between the disease and non-disease samples. For Fusion 1, LIMMA and SDAMS identified 250 and 179 significant corresponded-peaks respectively, with 161 significant peaks in common and 268 significant peaks combined. Of these 268 significant peaks, 211 were classified into at least one lipid category and 11 were classified into multiple lipid categories. The 11 multi-classified peaks are dropped from further differential analyses. In Fusion 2, 200 and 140 corresponded-peaks were identified for LIMMA and SDAMS respectively, with 125 significant peaks in common and 215 significant peaks combined. Of these 215 significant peaks, 185 were classified into at least one lipid category and 9 were classified into multiple lipid categories. The log2 fold changes for each categorized corresponded-peak between disease and non-disease was then calculated. P-values were calculated using a hypergeometric test and adjusted for multiple testing using the Benjamini-Hochberg technique (Benjamini and Hochberg, 1995).

5.3 Results

5.3.1 PCA and Correlation Shows Separation of Disease and Non-Disease Samples

Principal component analysis (PCA) is a technique that represents high dimensionality data in a lower dimensional space where each dimension

represents an orthogonal ‘component’ of the variance between the input data. PCA was performed using the normalized intensities of the corresponded-peaks present in each sample for each peak present at least 25% of the samples for each class, revealed a clear, if imperfect, boundary between disease and non-disease samples along principal component (PC) 2 (Figure 5.1). This implies that a significant portion of the variance observed in the second principal component (PC2, 13% explained variance in Fusion 1, 12.3% in Fusion 2) represents biological variance between the two classes of samples. Correlation heatmaps between the samples further supports this claim (Figure 5.2). Normal samples cluster well together with high correlation observed between most normal samples and other normal samples. Correlation within the disease class were less strong but still differentiated disease samples from normal. Lower correlation between cancer is somewhat expected given the large amount of heterogeneity observed in cancer versus non-cancer as has been previous shown (Paz-Yaacov *et al.*, 2015). In both groups, samples that appear to correlate with neither disease nor non-disease were detected. These samples likely have data quality or spectrum acquisition issues.

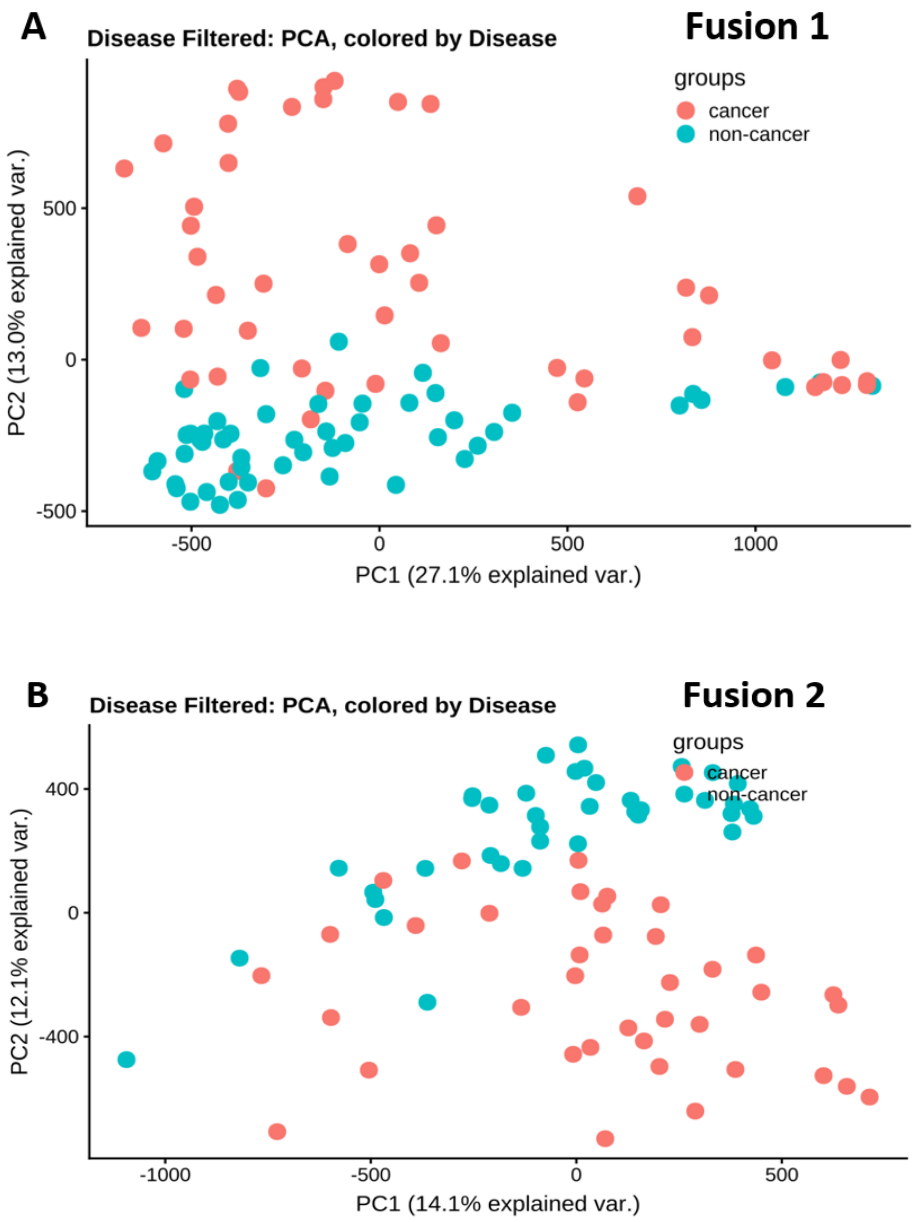

Figure 5.1: PCA by Disease
PCA performed on the normalized intensities of the corresponded-peaks present in at least 25% of a sample class shows a clear separation between disease and non-disease that is most obvious along PC2 in both Fusion 1 (Panel A) and Fusion 2 (Panel B). This separation is not present along PC1. The grouping of samples by disease class in PCA space clear indicate that the corresponded-peak intensities capture some of the biological variance between disease classes. Note that cancer and non-cancer appear flipped between Fusion 1 and Fusion 2. This is normal with PCA as different PCA plots do not have the exact same principal components (PC1 in Fusion 1 is not the same as PC1 in Fusion 2), causing the sign of principle components to arbitrarily flip.

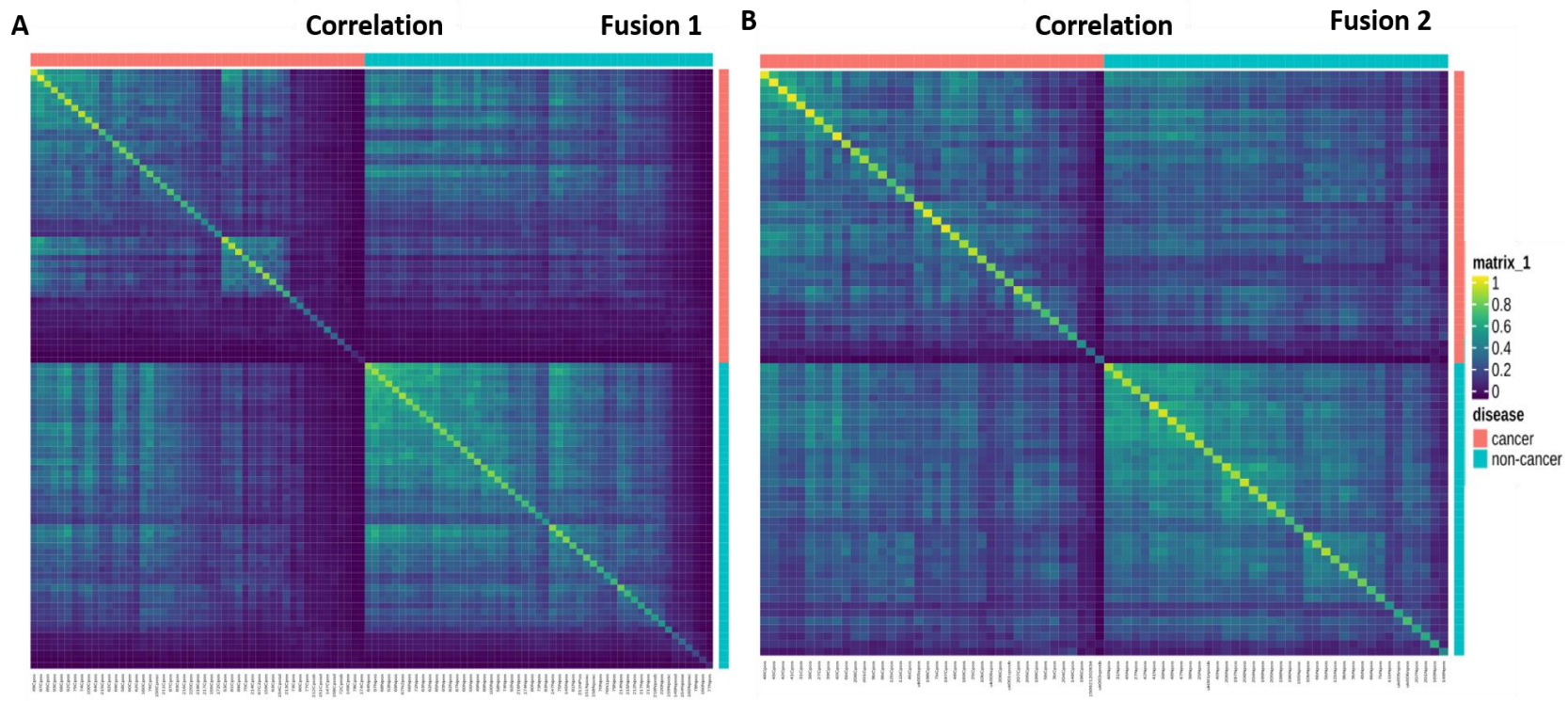

Figure 5.2: Correlation Heatmaps by Disease
High correlation is observed between most non-disease samples and other non-disease samples. Good correlation is also observed amongst most cancer samples, but with two obvious groupings of samples present in both Fusion 1 and Fusion 2 datasets. In both disease classes, a smaller number of samples have poor correlation with samples from either class. Overall, the correlation between cancer samples is less strong than the correlation between non-disease samples.

5.3.2 Differential Abundance of Lipid Categories Between Disease and Non-Disease Lung Tissue

Four distinct categories of lipids were identified with differentially abundant assignments between disease and non-disease: Glycerolipids [GL], Glycerophospholipids [GP], Sphingolipids [SP] and Sterols [ST]. Statistically significant fold changes were observed for glycerolipids, sphingolipids, and sterols in both Fusion 1 and Fusion 2. In Fusion 1, 59 consistently assigned sterol IMFs were observed, 46 of which were upchanged in disease (p=2.06e-18); 79 of 146 glycerolipid IMFs were upchanged in disease (p=1.69e-18) and 36 of 53 sphingolipid IMFs were downchanged in disease (p=1.41e-14). In Fusion 2, 33 of 50 sterol IMFs were upchanged in disease (p=5.99e-9), 62 of 125 glycerolipids were upchanged in disease (p=6.67e-10) and 33 of 49 sphingolipids were downchanged in disease (4.11e-17). Although differential fold changes were observed for one fatty acyls and some glycerophospholipids, as a group, they were statistically insignificant (p=1). These fold changes patterns are shown in Figures 3A-D at the category level and summarized in Supplemental Tables 5.1 and 5.2. In cases where an IMF is observed in disease but not in non-disease, the fold changes are very high due to imputation of missing values. For lipid categories that are mixed up and down, the up and down sub-populations correlate with *m/z*. (Figure 5.3E, F).

At the class level, 44 of 46 sterol categorized lipids in Fusion 1 and 33 of 33 of sterol categorized lipids were of the sterol [ST01] class. Querying the top 5 most abundant unique EMFs of the upchanged sterol IMFs from Fusion 1 and

Fusion 2 against PubChem (6 total unique EMFs, 4 in common) indicate that these formulas may correspond to known unsaturated sterol esters. Sterol esters are included in the [ST01] sterol class in LipidMaps.

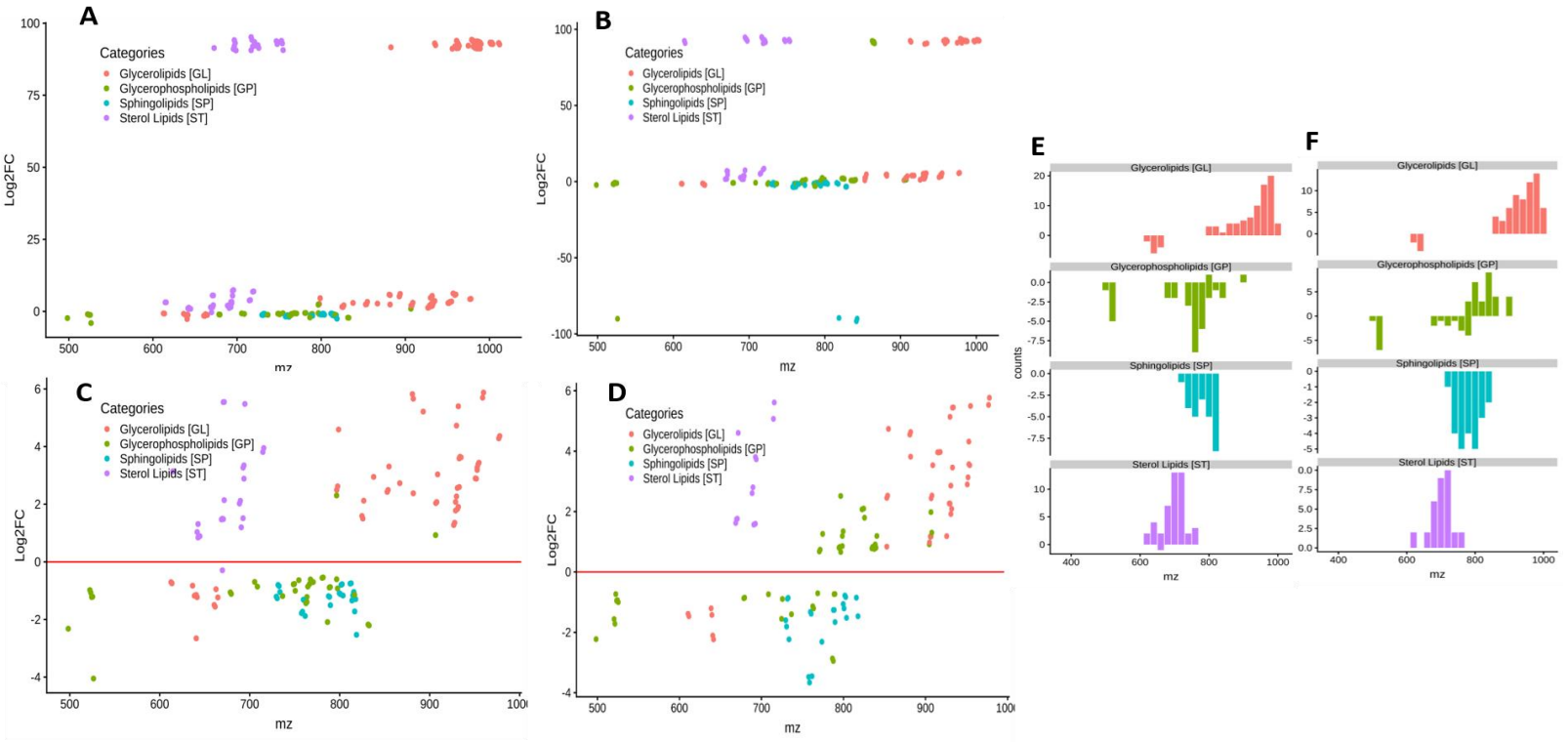

Figure 5.3: Log Fold Change by Category and *m/z*

The Log 2-fold change of consistently assigned, classified, and significantly changed lipids are shown in panels A and C with respect to *m/z*. From panel A, there is a population of glycerolipids and sterols that are highly up changed in cancer. A similar pattern was observed in Fusion 2 spectra as well (panels B and D). Extremely high fold changes are observed for some IMFs as seen in panels A and C. This is artifactual and occurs when an IMF is observed in disease but not in non-disease. Zooming in on the features that have Log2FC less than 6, these artifactual fold-changes are ignored, and the same trend can be observed in sterols and glycerolipids (Panel C). The distribution of lipids with significantly changed concentrations between disease and non-disease with respect to *m/z* is shown in Panels E and F for Fusion 1 and Fusion 2 respectively.

5.3.3 Lipid Class Correlation and Co-Occurrence Heatmaps

In addition to log fold changes observed in corresponded-peaks at the lipid category level, correlations between corresponded-peak normalized intensities were examined across all samples (Figure 5.4). Specifically, correlation values between features were calculated using samples with non-zero values for both features. The resulting correlation values were normalized by the information content, i.e. the fraction of samples used in the correlation calculation. At the lipid category level, all corresponded-peaks of a category correlate strongly with at least a subset of the other corresponded-peaks of the same category. Strong intra-category correlations are expected, if regulation of these lipids is controlled at the category level and if the corresponded-peaks are consistently and accurately assigned to lipid categories. Multi-classified corresponded-peaks were dropped from these analyses for this reason. Within the glycerolipid category, two subgroups of strongly correlated glycerolipids are observed and in sterols, one strongly correlated and one weakly correlated subgroup is observed. The sterols and the glycerolipids of the triradylglycerol class were strongly correlated suggesting a potential shared regulatory mechanism between these two types of lipids.

Patterns observed in the co-occurrence heatmaps of the lipids by class and category revealed more interesting findings. Although the sterols were correlated with one another, two distinct groups of sterols are observed in the co-occurrence plots. This suggests that there are two distinct populations of samples with different sterol abundances. Additionally, these two sterol groups

have distinct patterns of co-occurrence with other lipid categories. One sterol group has some co-occurrence with glycerolipids. The other sterol group has co-occurrence with the sphingolipids.

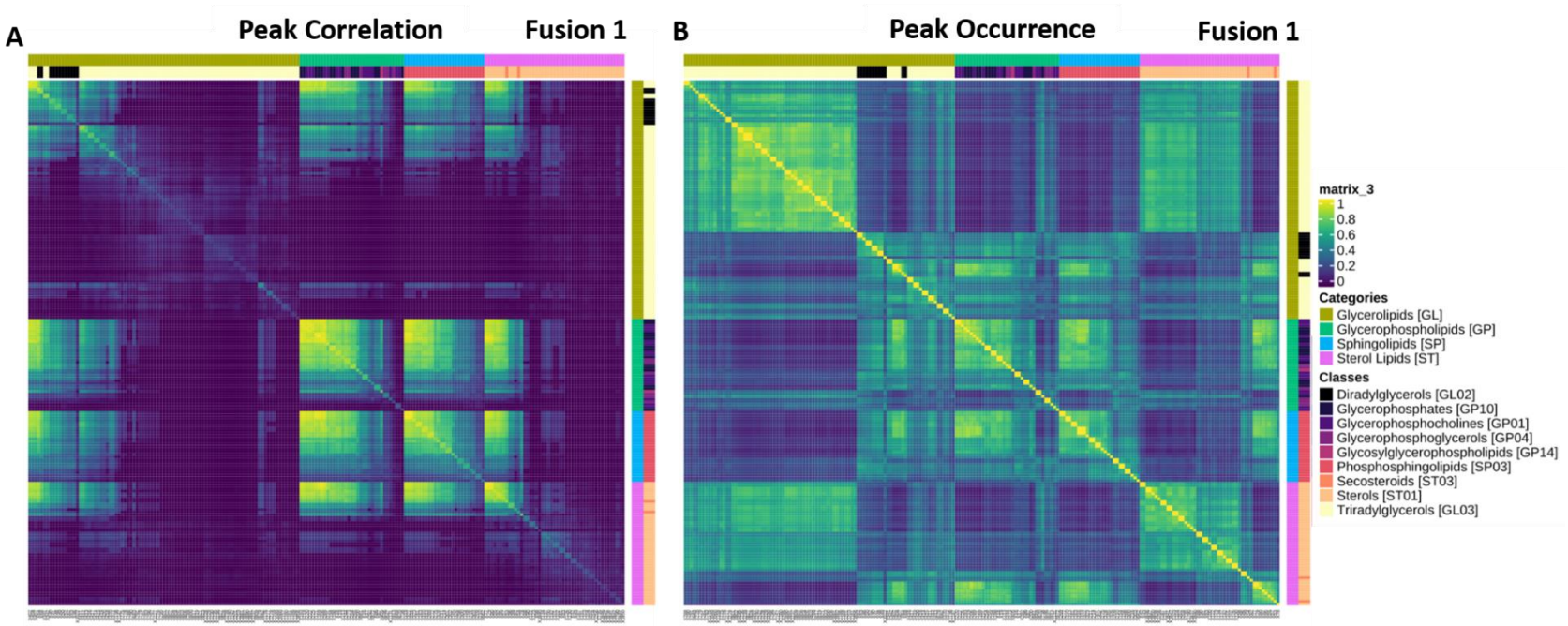
A
Peak Correlation
Fusion 1
B
Peak Occurrence
Fusion 1
matrix_3
1
0.8
0.6
0.4
0.2
0
Categories
Glycerolipids [GL]
Glycerophospholipids [GP]
Sphingolipids [SP]
Sterol Lipids [ST]
Classes
Diradylglycerols [GL02]
Glycerophosphates [GP10]
Glycerophosphocholines [GP01]
Glycerophosphoglycerols [GP04]
Glycosylglycerophospholipids [GP14]
Phosphosphingolipids [SP03]
Secosteroids [ST03]
Sterols [ST01]
Triradylglycerols [GL03]

Figure 5.4 Continued

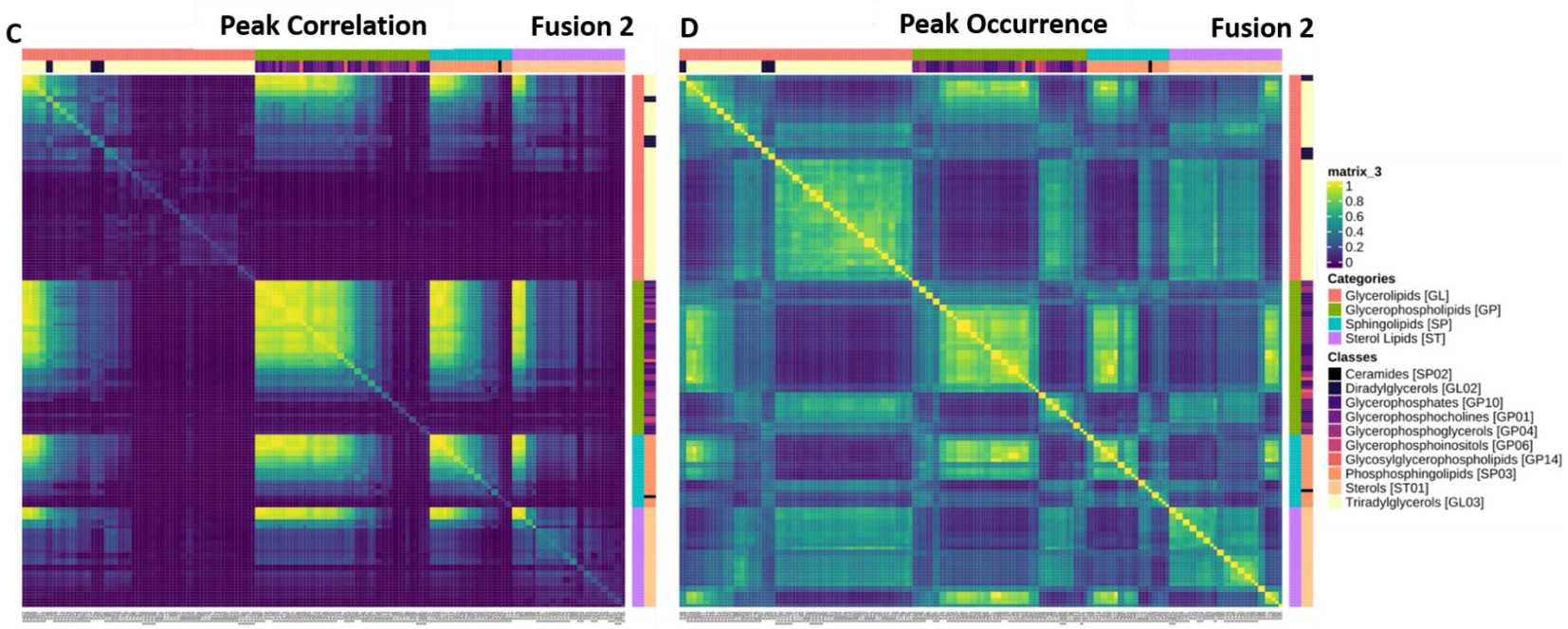

Figure 5.4: Peak Correlation and Peak Occurrence Similarity
Shown in panel A is an information-informed correlation between corresponded-peak normalized intensities grouped by lipid category for Fusion 1. Obvious strong correlations are seen between members of the same lipid category for many members of the same lipid category. Correlation however is only part of the picture. Shown in panel C is the occurrence similiarity of lipids by lipid category and class for Fusion 1 (i.e. how often lipids are observed together). Although all sterols were correlated with one another two distinct sets of sterols are observed to co-occur with one another. These subsets of co-occuring sterols also co-occur with sets of lipids of other categories. Visually similar patterns of intra-category correlation and co-occurrence are observed in Fusion 2 (panels B and D).

5.4 Discussion

5.4.1 *Assignment Ambiguity, Sparsity and Corresponded-Peak Generation*

Although SMIRFE can generate assignments in a completely untargeted manner and rule out many incorrect assignments using ratios of relative peak intensities, SMIRFE often does not provide unique unambiguous assignments for observed peaks, especially at higher *m/z*. The ambiguous assignments from SMIRFE are a largely unavoidable consequence of the massive search space considered by SMIRFE and current limitations on peak intensity variances that can be achieved with existing instruments. In many cases, the set of assignments provided for a peak at high *m/z* is likely to contain the correct assignment, but also many incorrect assignments as well.

Furthermore, even if all peaks were unambiguously assigned, a second significant problem arises in that the set of features observed between samples can change significantly. Any given feature is unlikely to be observed across all samples in a dataset even if that feature is non-artifactual. This high variance in observed features between samples can be attributed to variance between biological units from which samples are acquired (in our case human patients), variance in sample preparation, and spectrum acquisition that can lead to the loss and gain of features.

Corresponded-peak generation and filtering corresponded-peaks to those present in 25% of samples for a given class provides a partial solution to both problems by selecting only consistently assigned features present in multiple

samples. This subsequent reduction in feature sparsity allows for the meaningful use of more traditional statistical approaches such as PCA that do not handle high data sparsity well.

5.4.2 *Sample Correlation Analysis Shows Evidence of Metabolic Reprogramming in NSCLC*

The dysregulated proliferation of cancer cells requires metabolic reprogramming to occur, enabling the production of metabolites necessary for biological processes important for cancer development. The number of distinct dysregulations of metabolic programs that can lead to a pro-cancer state is likely large and, even among a cohort of patients with the same type of cancer, metabolic reprogramming is expected to vary significantly. As a result, the lipid profiles of non-disease samples are expected to be more similar to one another than the lipid profiles of disease samples from the same patients.

This hypothesis is supported by the correlation patterns observed in Figure 5.4. Although individual genetic, metabolic, and environmental variance leads to some differences between lipid profiles of non-disease samples, the majority of the non-disease samples show higher correlation to one another with respect to the normalized intensities of shared corresponded-peaks as compared to the correlation between disease samples. Additionally, the number of common corresponded-peaks in non-disease was higher than the number of common corresponded-peaks in the disease samples, arguing that in general, the non-disease samples have similar lipid profiles to one another. The disease samples on the other hand displayed weaker correlation to one another with two distinct

clusters of correlation observed. The top left most cluster corresponds to primary lung tumors of various subtypes and the other cluster was mixed between primary lung tumors, metastases from other organs and granulomas. Weaker correlation is expected amongst the disease class given that lipid profiles in cancer are likely to display higher variance (given the many ways that metabolic reprogramming can occur).

Although metabolic reprogramming will perturb the lipid profiles of cancer as compared to non-cancer, not all components of cellular metabolism are likely to be affected and thus some correlation should be observed between disease and non-disease samples. This is observed in our analysis as well, suggesting that there are lipid profile similarities between disease and non-disease as expected for samples derived from the same tissue and the same patients. Conversely, random incorrect assignments would not result in these observed correlation patterns, thus strongly implying that the SMIRFE assignments and corresponded-peak generation selects for highly consistent assignments that capture the biological variance between disease classes.

5.4.3 *Lipid Category Correlation and Correspondence*

The *de novo* synthesis of many lipid categories is controlled by a single rate limiting enzyme early in their biosynthetic pathway (e.g. sterols – HMG-CoA reductase, sphingolipids – serine palmitoyltransferase (Rütti *et al.*, 2009), glycerolipids – glycerol-3-phosphate acyltransferase (Wendel *et al.*, 2009)). Precursors from these pathways are further metabolized to yield subclasses of these lipids and are similarly regulated largely at the class level. As a result,

changes in the activities of these rate-limiting enzymes result in large changes in lipid concentrations at the class and category level. This effect is observed in Figure 5.4A where high correlation is observed within lipid categories and classes (i.e. if one sterol is upchanged, most sterols are upchanged). This correlation pattern is most strongly observed with sterols and the triradylglycerolipid subclass of glycerolipids (the triradylglycerolipid class contains mostly triacylglycerides). Furthermore, the triradylglycerolipid subclass and the sterol categories are strongly correlated with one another, implying that these lipid categories and classes are co-regulated. A possible mechanism for this co-regulation is through steroid response element binding proteins (SREBPs), which are involved in regulating key enzymes in the sterol and glycerolipid biosynthetic pathway (Ericsson *et al.*, 1997). Both SREBP-1 and SREBP-2 have been implicated in the control of lipid biosynthesis (Wen *et al.*, 2018) and previous studies have shown that altered SREBP-1 signaling due to B7-H3 (aka CD276, an immune checkpoint protein) overexpression is correlated with aggressive NSCLC and increased glycerolipid production (Luo *et al.*, 2017).

Strong correlation patterns within groups of predicted lipid categories would not be expected from random assignments or from incorrectly categorized lipids. The weaker correlation patterns observed among sphingolipids and glycerophospholipids could imply more incorrect assignments for these lipid categories (which is expected for sphingolipids due to our machine learning model's tendency to overpredict that lipid category) or could represent more

complex regulatory mechanisms or a mixture of scavenging and *de novo* synthesis.

Correlation however is only part of the picture. Although all lipid categories were strongly correlated to other members of their same category or class, not all members of each class are observed to co-occur with one another. Two distinct populations of sterols are clear in Figure 4B, one large and one small population of sterols that co-occur with one another. Additionally, each population of sterols co-occurs with other lipids from the other categories, implying that there are two distinct lipid profiles observed across the samples. The origin and clinical implications of these two lipid profiles is unclear, but one population could correspond to lung cancer that is known to respond to statins.

5.4.4 *Reproducibility across Instruments and Clinical Environments*

In this study, two sets of samples were analyzed using two different Thermo Tribrid Fusion instruments. Samples from Fusion 1 were derived only from patients undergoing resections at the University of Louisville while Fusion 2 samples were derived from patients undergoing resections at the University of Kentucky and the University of Louisville. In both sets of samples, the same gross changes in lipid profile between disease and non-disease were observed. Furthermore, as shown in Figure 5.5, many differentially abundant IMFs were observed in both Fusion 1 and Fusion 2 (110 IMFs representing 48% of all IMFs observed in both instruments). Although the sets of abundant IMFs do not completely overlap between the two instruments, the assignment and lipid profile

similarities across biological and analytical replicates suggest the reproducibility and consistency of these results.

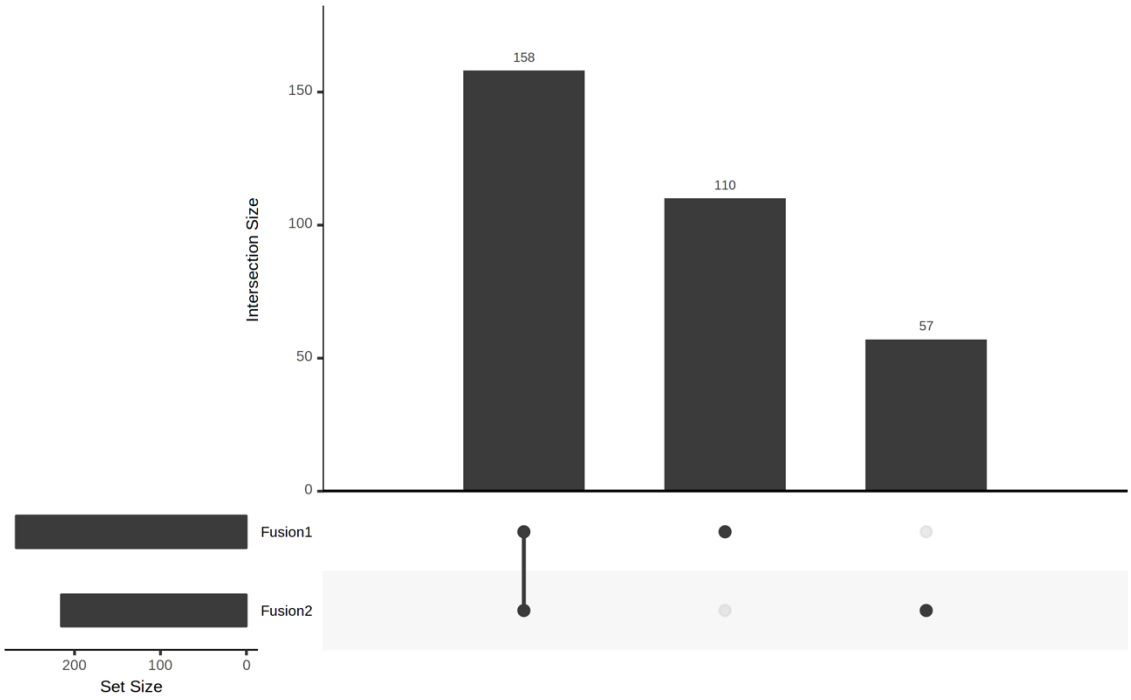

Figure 5.5: Differentially Abundant IMF Overlap Between Instruments
325 unique differentially abundant IMFs were observed across Fusion 1 and Fusion 2 samples. 158 (48%) of these IMFs were shared between Fusion 1 and Fusion 2, 110 were observed only in Fusion 1, and 57 were observed only in Fusion 2. In our experience, Fusion 1 is the more sensitive of the two instruments and a difference in sensitivity between instruments is a hypothesis explaining some of these discrepancies.

5.4.5 *Potential Clinical Implications*

Our study identified clear differences in the relative concentrations between disease and non-disease samples derived from the same patients. Most notably, the relative concentrations of a subset of glycerolipids and sterols (Figure 5.4B) were significantly and consistently higher in the disease samples as compared to control. Although the disease class contains several granuloma

samples, most of the samples are cancer. Given the key roles that these lipids have in membrane structure and function and are needed for cancer cells to divide, the observation that these lipid categories are upchanged in disease is not unexpected. Alterations in lipid metabolism are common in cancer and increased sterol concentrations are observed in many cancers.

Although interfering with the production of either class of lipids could potentially have therapeutic roles in the treatment of NSCLC, sterol metabolism is the easiest to target with existing pharmaceuticals. Two classes of drugs that are already used clinically, statins and nitrogenous bisphosphonates, inhibit enzymes in the mevalonate pathway from which all endogenous sterols are derived (Wiemer *et al.*, 2007) (Istvan and Deisenhofer, 2001). Although neither class inhibits the uptake of exogenous sterols, bile acid sequestrants, which reduce the amount of available exogenous sterols, have not been demonstrated to improve NSCLC outcomes in the same manner as statins. The previous epidemiological studies that have suggested a possible protective or therapeutic role for statins in NSCLC and other cancers have not provided direct evidence that statins improve outcomes by targeting the mevalonate pathway. Statins are known to have other off-target effects such as inhibition of angiogenesis (Garjani *et al.*, 2012), anti-inflammatory effects (Blake and Ridker, 2000), and these off-target effects differ with high and low dosing strategies (Skaletz-Rorowski and Walsh, 2003) and may not directly involve the inhibition of sterol biosynthesis (Dulak and Jozkowicz, 2005). While our results do not directly support a claim either way that statins have an anti-cancer effect through the inhibition of the

mevalonate pathway, increased sterol production in cancer versus non-cancer is a prerequisite for such a hypothesis to be true. More importantly, our methods provide a potential avenue to explore the effects of statins and other pharmaceuticals on lipid profiles in an untargeted and comprehensive manner.

The mechanisms resulting in the observed lipid profile changes from our study remain unclear and could be the object of future studies. Increased sterol and glycerolipid concentrations are consistent with direct (constitutively active mutants) or indirect (downstream signaling) activation of EGFR, which promotes glycerolipid and sterol biosynthesis (Sukhanova *et al.*, 2013) (Gabitova *et al.*, 2013); SCD, which is required for some glycerolipids and sterol esters (Bené *et al.*, 2001); and ACLY, which is essential for sterol biosynthesis (Zaidi *et al.*, 2012). As mentioned previously, SREBPs might also be involved in these lipid profile changes (Luo *et al.*, 2017) (Ericsson *et al.*, 1997). Furthermore, if the presence of upchanged sterol esters can be verified, the enzyme ACAT-1 becomes a promising target in NSCLC. Previous studies have shown in pancreatic cancer that inhibition of ACAT-1 prevents the conversion of cholesterol to sterol esters, which results in apoptosis due to elevated endoplasmic reticulum stress (Li *et al.*, 2016a).

One possible hypothesis explaining the observed differences is that acquisition of a high concentration sterol and glycerolipid metabolic phenotype may be necessary for the development of NSCLC. Therefore multiple distinct genetic lesions may result in this metabolic phenotype. Thus, pharmaceutical intervention that directly targets the sterol component of this phenotype may be

beneficial for a variety of genetically distinct NSCLC subtypes. Alternatively, these lipid profile differences may be a byproduct of other disease processes and not directly contributory to the development of disease. In this case, these metabolic changes would still be useful as biomarkers for genetic subtypes of NSCLC.

5.5 Conclusion

The use of untargeted assignment tools such as SMIRFE combined with machine learning models for the prediction of lipid categories enables comprehensive lipid profiling in human derived samples from molecular formula assignment alone. Ambiguous assignments from SMIRFE and the high feature sparsity intrinsic to such studies can be largely mitigated using corresponded-peak generation and filtering out inconsistently observed corresponded-peaks. Corresponded-peak generation improves assignment consistency and accuracy, which can be confirmed through patterns of observed correlation between categories and classes of assigned lipids.

Subsequent differential abundance analysis, performed using the filtered corresponded-peaks, identified a consistent and significant difference in lipid profiles between disease samples and controls. Most notably, both glycerolipids and sterols are significantly and consistently observed in higher relative concentrations in disease tissue than in neighboring non-disease tissue. These findings are consistent with known genetic lesions observed in NSCLC and with common metabolic alterations observed in cancer metabolic reprogramming.

Furthermore, the change in sterol profile between disease and non-disease has the potential to be clinically translatable. Statins and other mevalonate pathway altering drugs, which are known to improve outcomes in NSCLC patients, could be altering these lipid profiles as part of their mechanism of action. Although further studies are necessary to confirm this hypothesis, the ability to detect and quantify lipid profiles in an untargeted manner provides the capability to directly observe the metabolic effect of pharmacological interventions targeting the mevalonate, with respect to NSCLC lipid profiles.

Which genetic markers correlate to our observed molecular phenotype remains unclear, but future genomic or transcriptomics datasets from a similar cohort of patients combined with our lipid profiling approach could identify the genetic markers for this metabolic phenotype. The possibility that multiple genetic lesions could result in the same or similar metabolic phenotype implies that anti-mevalonate pathway therapies could have a direct chemotherapeutic or adjuvant role in the treatment of many genetically distinct subtypes of NSCLC.

CHAPTER 6. CHEMICALLY AWARE SUBSTRUCTURE SEARCH (CASS)

6.1 Introduction

The high volumes of complex spectral data generated by large scale FT-MS metabolomics experiments require automated assignment tools such as to assign these datasets in an efficient manner. Untargeted tools such as SMIRFE are desirable to reduce assignment bias and for the assignment of previously unreported metabolites. However, both SMIRFE and targeted tools ultimately cannot assign spectral features to metabolite structures directly. At best, assignment tools such can only provide trustworthy isotopically resolved or elemental molecular formula assignments to peaks. When these assignments are ambiguous the set of possible metabolites corresponding to that peak is possibly very large but even an unambiguous assignment may represent more than one metabolite structure due to structural isomerism.

To overcome this difficulty, additional information must be obtained to accurately assign metabolite mass spectra to metabolite structures. Tandem mass spectrometry is often used to obtain chemical substructures of a given metabolite via its fragmentation pattern. Unfortunately, the data produced by tandem-MS requires very complicated, predictive algorithms for metabolite assignment and differences in fragmentation patterns generated by different instruments, in algorithms used for data analysis, and in data interpretation hampers the reproducibility and accuracy of these methods (Nesvizhskii *et al.*, 2007). However, libraries do exist for interpreting metabolomics tandem-MS data,

especially when high resolution is used (Yang *et al.*, 2017a) (Huan *et al.*, 2015). Chromatographic retention times either from GC or LC can also aid in metabolite assignment but cannot be obtained in a direct infusion experiment. An approach for obtaining structural or substructural information that is compatible with direct infusion MS1 experiments is therefore desirable.

Chemoselective adduct formation (i.e. CS-tagging) refers to a family of techniques where metabolites are derivatized in a predictable manner depending upon their functional group composition. Which functional groups are derivatized and what derivatives are formed depends upon which chemoselective reagent or reagents were added to the sample. Chemoselective reagents exist for many functional groups present in metabolites, such as carboxylate (Ye *et al.*, 2009), carbonyl (Fu *et al.*, 2011) (Mattingly *et al.*, 2012) amino (Guo and Li, 2009) and sulfhydryl (Gori *et al.*, 2014). This derivatization process produces an identifiable pattern in the mass spectra that can be used to infer the presence of one or more of these functional groups in a metabolite corresponding to an observed spectral feature. These reagents can also be isotopically labeled to produce more unique and identifiable patterns. Although all structural isomers of the same formula will have the same mass, not all these isomers will have the same functional group composition and thus functional group composition derived from CS-tagging can provide limitations on which metabolites could be assigned to that feature.

Using untargeted assignment methods such as SMIRFE in combination with CS-tagging effectively provides both an assigned formula and a predicted functional group composition for potential metabolite spectral features in an FT-

MS spectrum. However, the lack of functional group composition annotations in existing metabolite databases makes utilizing this information to aid in assignment difficult. The ideal solution to this problem is to add these annotations to existing databases which requires designing an algorithm to identify functional groups in database representations of metabolites (typically .mol format (Haider, 2010)]). Tools such as CheckMol (Haider, 2016) can partially solve this problem but are hard-coded to only identify a pre-determined set of functional groups. If a functional group of interest (or another substructure) is not present in this pre-determined list, the actual CheckMol program must be modified to include it. Making these modifications without introducing error limits the usefulness of this tool.

Ideally, a generalized solution that does not rely on pre-determined lists of functional groups is needed. Such a tool would also be useful for other applications beyond just functional group identification. By restating the functional group identification problem as a graph theory problem, in which a chemical structure is represented as a collection of nodes representing atoms and vertices representing bonds, the problem of finding functional groups becomes identical to the maximum common subgraph isomorphism problem in graph theory (i.e. given a two graphs Ga and Gb determine if Gb is a subgraph of Ga) (Raymond and Willett, 2002). Our algorithms called Chemically Aware Substructure Search (Mitchell *et al.*, 2014) and its improved version Biochemically Aware Substructure Search solves the MCSI problem efficiently for chemical structures. Using these algorithms, functional group annotated versions of KEGG Compound and the

HMDB were constructed that can be queried with molecular formula and functional group composition.

Additionally, the atom coloring method developed for reducing the computational complexity of the CASS algorithm has potential applications for identifying NMR equivalent nuclei and the construction of useful metrics of chemical similarity (Trainor *et al.*, 2018).

6.2 Materials and Methods

6.2.1 Database Access and Parsers

A copy of the HMDB database was downloaded in May 2014 as a flat file of concatenated and annotated .mol files (.sdf format). The KEGG Compound database was not available for download in any consolidated format but was instead web scraped using a Python program through the KEGG REST interface. The output from this scraping was a collection of .mol files.

Because we could not find a recent functional group database that fit our design criteria (Haider, 2010; Kotera *et al.*, 2008),we created one from scratch. .mol files corresponding to each functional group from CheckMol were created manually using JChem. Since functional group structures often have positions that can be occupied by a set of elements (i.e. an alkyl halide has a position that can be ‘any halogen’) a nomenclature for describing these conditions was added for our functional group mol files. Lists of valid elements for a position can be specified using ‘|’ while ‘!’ designates an invalid element type (e.g. “H|O|N” means hydrogen or oxygen or nitrogen at that position, “!H” means any element

but hydrogen). Additionally, functional groups often refer to specific chemical substructures within a larger chemical context. The carbonyl substructure -C=O for example is either an aldehyde or a ketone due to its surrounding chemical structure. These contextual atoms must be present for a chemical substructure to be identified but are not part of the substructure. Contextual atoms are designated by an asterisk (e.g. “C*” is a contextual carbon). The ability to designate contextual atoms allow the clean determination of overlapping functional groups. The descriptive abilities of this nomenclature exceed the simplistic wild-card atom designations in previously published chemoinformatics toolkits such as SMARTS (Daylight Chemical Information Systems, 2008) but render the mol files non-compliant with the standard.

Our non-compliant mol files and the non-compliant mol files from the HMDB required developing a dedicated parser. Our parser returns for every mol file a molecule object which consists of multiple atom and bond objects. Internally atoms are uniquely identified by an integer index value. In many database mol files, implicit hydrogens are often excluded to reduce the size of the files. Since hydrogens are often part of the functional group descriptions, these implicit hydrogens are added in accordance to standard connectivity and valence methods (Weininger, 1988). This procedure does not account for pH or pK in these calculations and hydrogens are added to produced non-charged molecules unless the .mol file specifies a charge.

6.2.2 Adjacency Matrix Representations

The graph theory algorithms in CASS and BASS require numerical representations of each chemical structure. The two most common representations for graph-like objects are adjacency lists and adjacency matrices. Both representations describe which vertices (bonds) connect which nodes (atoms) and they differ in their runtime and memory requirements. The list representation saves space but the adjacency matrix is faster for determining if two nodes are connected by a vertex. In the adjacency matrix representation, the value of the element i,j in the matrix represents the order of the bond connecting atoms i and j (by index) or zero if no bond exists. Thus, testing if two atoms are connected or not can be done in constant time by retrieving the element i, j from the matrix and testing if it is non-zero. Examples of adjacency matrices are shown in Figure 6.1.

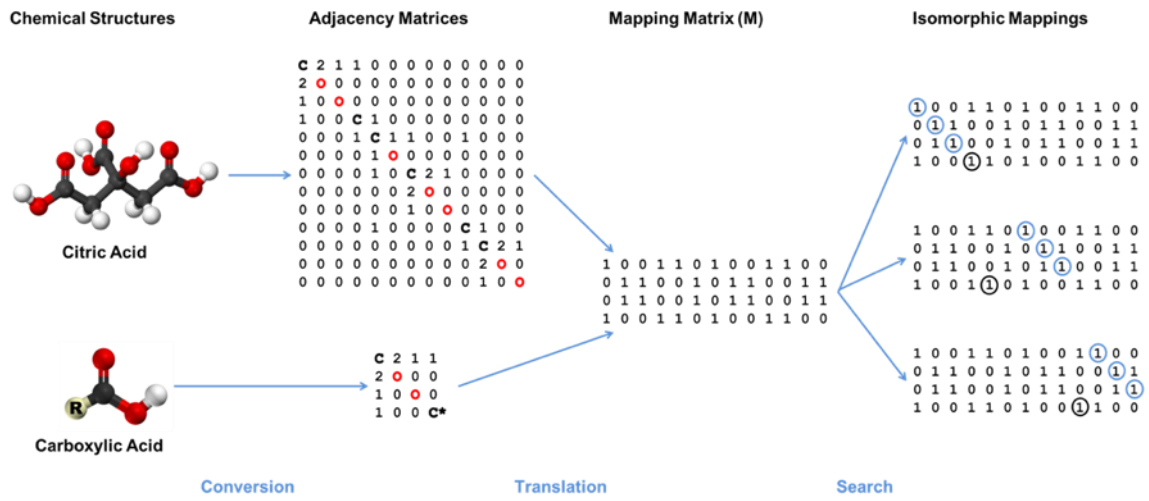

Figure 6.1: Algorithm Overview and Example Matrices
Chemical structures can be represented as colored graphs, which in turn can be converted into adjacency matrices. The diagonal of an adjacency matrix can be loaded with information about the nodes. A mapping matrix (M) represents which atoms in the functional group or substructure are mappable to the atoms of the larger query compound. The mapping matrix represents all possible mappings and thus M must be enumerated or searched to generate the specific mappings that represent isomorphisms (i.e. map the functional group completely onto the compound chemical structure).

6.2.3 Substructure Search Algorithm Description

The starting point for the development of our algorithm was the original Ullmann algorithm for solving the MCSI problem (Ullmann, 1976). The presence of a typo in the description of the algorithm in the original manuscript and the presence of numerous ‘goto’ statements in the original pseudocode complicated our translation of the original algorithm Figure 6.2 into a modern control flow representation shown in Figure 6.3. We deviated significantly from the Ullmann algorithm during the development of CASS and BASS.

A

```
# Variable Initialization
P_alpha = number of atoms in A
P_beta = number of atoms in B
d = 0
k = -1
For all i = 1,.., P_alpha: F_i = 0
For all i = 1,.., P_beta: H_i = -1
For all i = 1,.., P_beta: M_d = M

Step 1:
M = M^0; d = 1; H_1 = 0
For all i= 1,..,P_alpha, F_i = 0
Step 2:
If there is no value of j such that M_dj == 1 and F_j == 0 then goto step 7
M_d = M
If d == 1 then k = H_1 else k = 0
Step 3:
k := k+1
If M_dk == 0 or F_k == 1 then goto step 3
For all j != k set M_dj := 0
Step 4:
If d < P_alpha then goto step 6 else check for isomorphism
Step 5:
If there is no j > k such that M_dj == 1 and F_j  == 0 then goto step 7
M := M_d
goto step 3
Step 6:
H_d = k;  F_k = 1; d = d + 1
goto step 2
Step 7:
If d == 1 then terminate algorithm
F_k = 0; d = d - 1; M = M_d; k = H_d
goto step 5
```

B

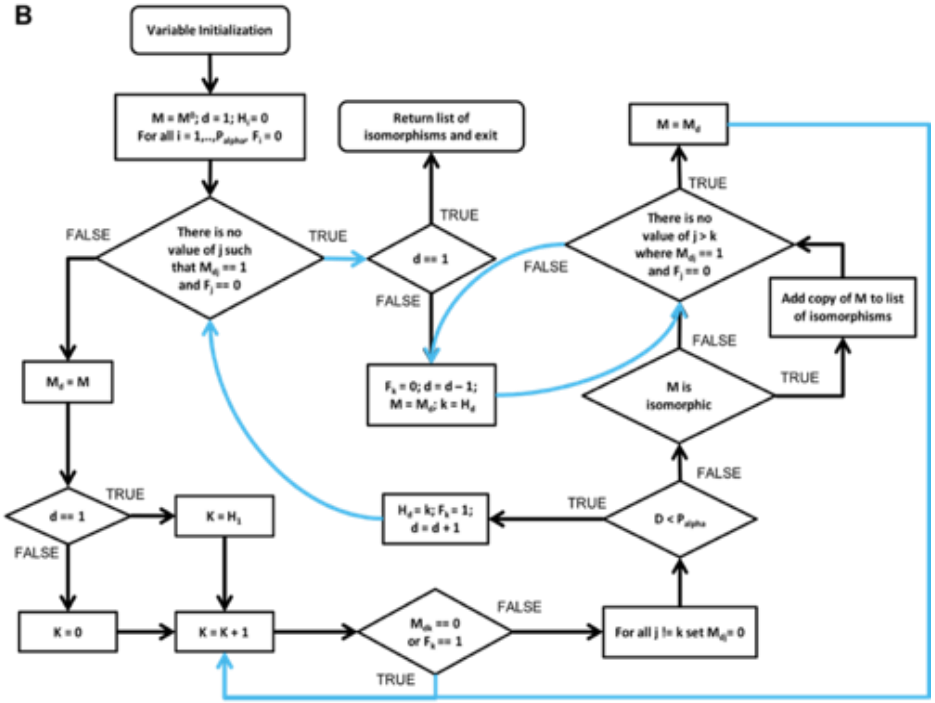

Figure 6.2: Pseudocode and Control Flow of the Original Algorithm

Panel A shows the original pseudocode for the Ullmann simple enumeration algorithm. ":=" denotes assignment. Panel B shows the control flow diagram corresponding to this pseudocode. Blue lines represent 'goto' statements. In modern computer programming, explicit goto statements are discouraged both to improve readability and maintainability but also to prevent errors. The variable "SKIP2B" does not exist in the original pseudocode but was needed for control flow purposes.

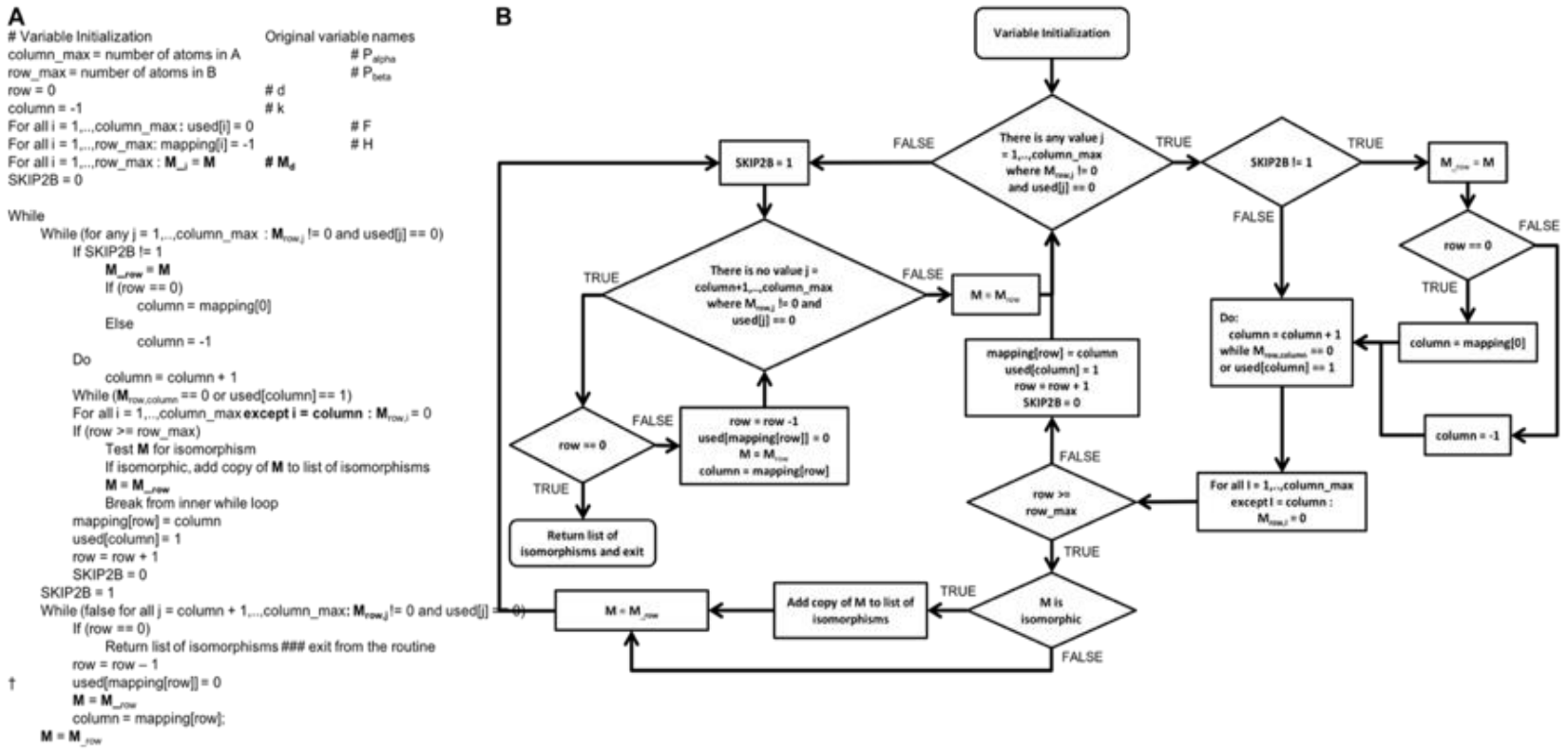

Figure 6.3: Modernized Pseudocode and Control Flow for the Ullmann Algorithm
Panel A shows the modernized pseudocode with better named variables and no “goto” statements. While modernizing the pseudocode, we identified a typographical error in line 2, step 7 in the Ullmann algorithm on page 33 of the original publication. This line is marked † in the pseudocode. Panel B shows the corresponding control flow structure.

Given two adjacency matrix representations (Aa and Ab) of the graphs Ga and Gb representing a query metabolite structure A and a generic substructure B, the first step in our algorithms is the construction of the mapping matrix M (Figure 6.1). M has dimensions b x a where a and b are the number of atoms in A and B. Since B must be substructure of A, a > b. M[i,j] = 1 if atom i in B has a compatible element type with atom j in A and if atom j in A has as many or more bonds than atom i in B. M[i,j] = 0 if either condition is false. These restrictions greatly reduce the number of 1s in M which has a direct effect on runtime. It is this step where our expanded element types for our .mol files is important. If an entire row of M contains only zero, this means there is at least one atom in B that cannot be mapped to any atom in A and thus B cannot be a subgraph of A.

The matrix M represents all possible mappings of the atoms of B to the atoms A. Only a subset of the mappings of B to A represent actual isomorphisms. Each isomorphism corresponds to a M' matrix that contains only one non-zero value per row. Possible isomorphisms can be generated by enumerating M to yield many M' where each M' represents a version of M where there is exactly one 1 per row. In the original Ullmann algorithm each M' is made explicitly by copying the entirety of M and then modifying the elements of the copy. Instead, our algorithm uses two 1 dimensional vectors v and u. v records which atoms in B are mapped to atoms in A (v[2] = 3 means atom 2 in B is mapped to atom 3 in A), u keeps track of which atoms in A are currently used in a mapping and cannot be mapped to another atom. This approach saves

considerable runtime by eliminating all copying and manipulation of M. Every unique v where each M[v[i],j] is 1 represents a unique M'.

Each M' is tested to represent an isomorphism by comparing Aa to Ab using the information stored in v. If ($\forall$ 0 =< i =< |**v**|, $\forall$ 0 =< j =< |**v**|) ($\mathbf{A}_{B\,i,j} = \mathbf{A}_{A\mathbf{v}[i],\mathbf{v}[j]}$) then v represents an isomorphism between B and A and a copy of v, v' is stored in a list of known isomorphisms. This approach requires a linear number of operations compared to the original Ullmann algorithm that used matrix multiplication and thus a polynomial number of operations to test for isomorphism. In some applications, simply knowing that a single isomorphism exists is enough. In this case, our algorithm can be short-circuited to return the first isomorphism found. The complete pseudocode and control flow for our algorithm is shown in Figure 6.4.

A

```
# Variable Initialization
column_max = number of atoms in A
row_max = number of atoms in B
row = 0
For all i = 0,..,row_max : mapping[i] = -1
For all i = 0,..,column_max : used[i] = 0
For all i in excluded_list : used[i] = 1
M⁰ is the mapping matrix between A and B.

While (row >= 0)
    Do
        mapping[row] = mapping[row] + 1
    While (mapping[row] <= column_max and (M⁰ row, mapping[row] == 0 or used[mapping[row]] == 1))
    If (mapping[row] > column_max)
        Next
    If (row == row_max)
        Test mapping for isomorphism
        Add a copy of mapping to list of isomorphisms if mapping is isomorphic
†       Return copy of mapping and exit algorithm if short-circuiting enabled.
Δ   Else If (row == 0 or mapping[0,..,row] contains an invalid mapping)
        used[mapping[row]] = 1
        row = row + 1
Continue
    While (row >= 0 and mapping[row] >= column_max)
        mapping[row] = -1;
        row = row – 1
        If (row >= 0 and mapping[row] not in excluded_list)
            used[mapping[row]] = 0
```

B

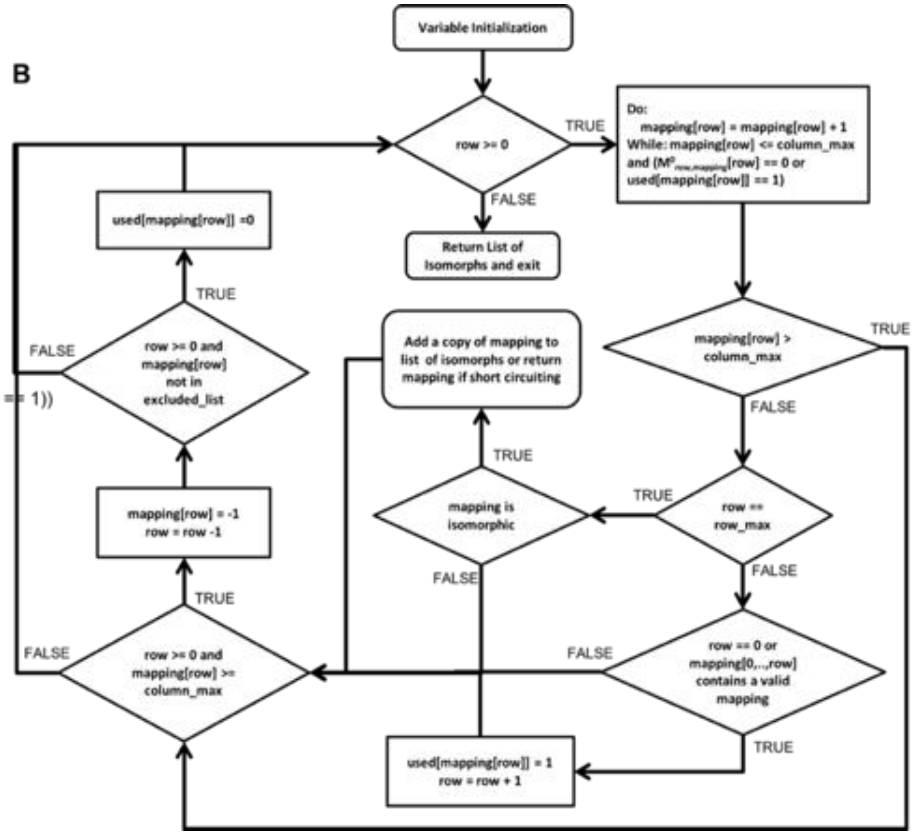

Figure 6.4: Pseudocode and Control Flow Diagram for CASS

Panel A shows the pseudocode for our CASS (and BASS) algorithm. The lack of matrix copies and multiplication saves both runtime and memory. Panel B shows the control flow for CASS which is cleaner than the modernized Ullmann algorithm. The short-circuiting step is marked with a †. Short circuiting allows the algorithm to terminate early when only one isomorphism must be found. Also, partial mappings are tested during enumeration as shown in step Δ. This testing of partial mappings greatly reduces the amount of enumeration that occurs as most partial mappings are invalid and not further enumerated.

Once the set of all v' is known, overlaps between found isomorphisms are then determined. For two isomorphisms Ve and Vf, the set O = Ve union Vf represents the indices of A to which atoms in B are mapped. If |O| = |Ve| = |Vf| then Ve and Vf are equivalent isomorphisms representing mirror images of the same substructure. If |O| = |Vf| and |O| > 0 then F overlaps with E and if |O| = |Ve| and |O| > 0 then E is a subgraph of F. Understanding which functional groups overlap has implications for the chemical properties of the detected functional groups. For example, all carboxylic acids contain a hydroxyl group but that hydroxyl group does not have the same properties of a hydroxyl group of an alcohol. In this case the hydroxyl group of a carboxylic acid would be identified as a subgraph alcohol and this information can be carried into other analyses. Additionally, functional groups that have planes of symmetry such as anhydrides must have these mirror images accounted for to arrive at the correct counts of each functional group. Functional groups that do not represent a single specific chemical structure such as Alkyl halides are excluded from these comparisons. These "super" functional groups were manually identified.

The set of functional groups for every metabolite complete with "overlapping" and "subset" designations must only be calculated once. These results are stored in an SQLite database with a column for every functional group that contains the number of those groups found per database entry and the molecular formula for that entry. This database can be queried using information expected from CS-tagging FT-MS experiments.

6.2.4 Aromaticity and Resonance Detection

A feature added to the original CASS algorithm in BASS was the ability to describe resonant and aromatic bonds in chemical structures. Resonant and aromatic bonds are inferred from the KEGG atom types of the atoms. The KEGG atom type system assigns a string to each atom in a chemical structure based on the atoms to which it is bonded and the context that contains that substructure. For example, the oxygen of a carboxylate has the type “O6a” and the carbon of a carboxylate has the type “C6a” and bonds between atoms with the types “O6a” and “C6a” are therefore known to be resonant. Similarly, atoms of type C8x and N5x are within an aromatic ring so bonds between atoms of these types are aromatic. Resonant and aromatic bonds are represented by “R” and “A” respectively in a molecule’s adjacency matrix.

Inferring resonant and aromatic bonds in this manner is straightforward for entries from KEGG but entries from the HMDB and other databases do not have these annotations. The solution to this problem is to use CASS/BASS and a set of reference structures corresponding to KEGG atom types to “apply” these KEGG atom types to unannotated structures. Each reference structure represents the chemical substructure that corresponds to one or more KEGG atom types. Regions of a query compound that are isomorphic to a reference structure share the KEGG atom types of the reference structure. Thus, adding KEGG atom type annotations simply requires a set of reference structures that are properly annotated. For most KEGG atom types these reference structures were constructed by hand, but for aromatic KEGG atom types reference

structures were created programmatically. This was achieved by first finding all cycles in all compounds from KEGG containing only aromatically annotated atoms. Second any overlapping cycles in the same compound were merged until the set of distinct aromatic cycles are found for each compound. Third, each substructure in the set of aromatic cycles is then “extended” out by one bond to capture the local structure around the aromatic substructure. This is to account for electron withdrawing or adding effects of additions to the aromatic cycle. This process yields a set of aromatic reference graphs complete with aromatic KEGG atom type annotations.

6.2.5 Node Coloring and Stereoisomer Detection

The CASS/BASS algorithm can also be used to determine if two structures are stereoisomeric by observing that a stereoisomer of a given structure A must be isomorphic to B when the stereochemistry of bonds are ignored. Testing for isomorphism is computationally very costly as effectively all substructures of A must be compared for isomorphism to all substructures in B. To minimize this cost a technique called node coloring was developed to filter out A, B pairs that cannot be stereoisomers of one another.

Node coloring assigns to every atom in a structure a ‘color’ that describes the local substructure around that atom out to a depth of d bonds. The depth d-color of an atom is expressed in terms of the d-1 colors of the neighboring atoms and the bond order of the connecting bonds. The d0 color of an atom is defined as the element type of that atom. All atom colors can therefore be defined

recursively from this base case. Optionally stereoisomer information and charge information can be added to these colors.

If A and B are stereoisomers, the set of all dk colors observed in A must be observed in B since all stereoisomers contain the same substructures. If even a single color at depth k is observed in A but not in B, then A and B cannot be stereoisomers. Additionally, when constructing the mapping matrix, atoms must have identical dk colors to be mappable to one another. Thus, through node coloring many A, B pairs can be rejected outright for stereoisomerism and the cost of testing remaining pairs can be greatly reduced.

6.2.6 Optimal CS-Tagging Strategy Analysis

Although currently many CS-reagents are available to researchers, what the ideal set of functional groups to identify to best enable the disambiguation of isomeric metabolites remains unknown. Using our functional group annotated metabolite database, the ideal sets of functional groups to identify can be constructed using an iterative approach.

Starting with the set of all one-functional group strategies, the performance of that strategy is calculated by determining what percentage of the database entries could be uniquely identified if their molecular formula and some information about the number of that functional group were known. The top 50 best-performing strategies are kept in the first iteration to create the first generation of parent strategies. The parent strategies are then expanded to generate all pairs of parent strategies with each new functional group to generate

new child strategies. All child strategies that do not exceed the performance of the parent strategy by a user-defined amount are eliminated and surviving strategies become the parents for the next iteration. This continues for a set number of iterations or until no new child strategies exceed the performance cutoff. As two functional groups can perform synergistically, (wherein a strategy with functional groups 1 and 2 performs much better than expected from functional groups 1 and 2 individually), the number of child strategies kept in each iteration must be sufficiently high to enable synergistic strategies to be encountered.

Additionally, functional group adducts may not be formed stoichiometrically and the ideal strategy should take this into account. For example, although citrate contains three carboxylates and any one (or possible two) of these carboxylates could be derivatized, derivatizing all three carboxylates is unlikely due to steric constraints. Therefore, strategy analysis can be performed in one of three modes: stoichiometrically – CS-tagging determines the exact number of a functional group in a compound, pseudostoichiometrically – CS-tagging can distinguish one or two instances of a functional group but three or more are indistinguishable, and non-stoichiometrically – where CS-tagging determines if at least one instance of a functional group is present or not. Furthermore, strategy analysis was performed considering various combinations of distinct, overlapping and subgraph functional groups to account for the possibility that CS-reagents react with chemical substructures similar to, but not exactly, their intended targets. The optimal strategies generated by this analysis

can help inform which commercially available reagents should be used in metabolite experiments and guide development of new reagents.

6.2.7 Computational Platforms Used

All timed analyses were done on three identical machines with dual Xeon X5650 processors @ 2.67GHz and 24 GB of 1333 MHz ECC memory running Fedora 18 “Spherical Cow”. All three algorithms were implemented in Perl 5.16.3 and SQLite v3.7.13 with DBI 1.631 was used in all programs interacting with an SQLite database.

6.3 Results

6.3.1 Computational Performance of CASS/BASS

The performance of the Ullmann algorithm and our algorithm for a set of representative database compounds and functional groups are shown in Figures 6.5 and 6.6. The runtime required to identify instances of functional groups in these compounds depends largely on the number of possible node mapping count (m, the sum of all entries in the M matrix) which is summarized in Table 6.1. In all examples, the Ullmann algorithm is slower than our algorithm for the same inputs. Additionally, the runtime of the Ullmann algorithm depends on m in an exponential manner while our algorithm has a pseudo-linear dependency on m. This pseudo-linear dependency becomes clearly polynomial (of degree 2 with an R2 of .99944) when the time needed to identify alkenes in each compound in the HMDB is plotted versus m for each compound. This polynomial behavior, if

rigorously and mathematically confirmed, could have implications for the ongoing debate regarding the theoretically best computational complexity achievable for the common subgraph isomorphism problem (Schöning, 1987) (Torán, 2004).

Short circuiting improves the runtime of our algorithm's performance sporadically but sometimes dramatically (Figure 6.7). The benefit of short circuiting is difficult to predict as the order in which M is enumerated determines when short circuiting can occur. Short circuiting can have a dramatic improvement on the time needed to confirm that two structures are stereoisomers compared to the non-short-circuiting variant (Supplemental Table 6.1). For stereoisomers, node coloring out to a depth of 4 bonds substantially reduces the possible node mapping count further enhancing the performance of CASS for stereoisomer detection.

Table 6.1 Possible Node Mapping Counts for Test Compounds and Functional Groups

	Possible node mapping count (*m*)			
Database compounds	**Functional group**			
	CA	**Epoxide**	**Alkene**	**Alcohol**
Deoxycytidine	46	22	18	26
R-3-Hydroxybutyric acid	25	11	8	15
2-Hydroxybutyric acid	25	11	8	15
Deoxyuridine	47	23	18	26
1-Methylhistidine	34	16	14	20
Cortexolone	84	46	42	155
2-Methoxyestrone	71	41	38	46
Deoxycorticosterone	70	45	42	43
1,3-Diaminopropane	18	6	6	13
2-Ketobutyric acid	23	11	8	13

Table 6.1: Possible Node Mapping Counts for Test Compounds and Functional Groups – the value of the possible node mapping count is shown for every compound, functional group pair. Larger values of m represent more possible mappings of functional group atoms to compound atoms that must be tested.

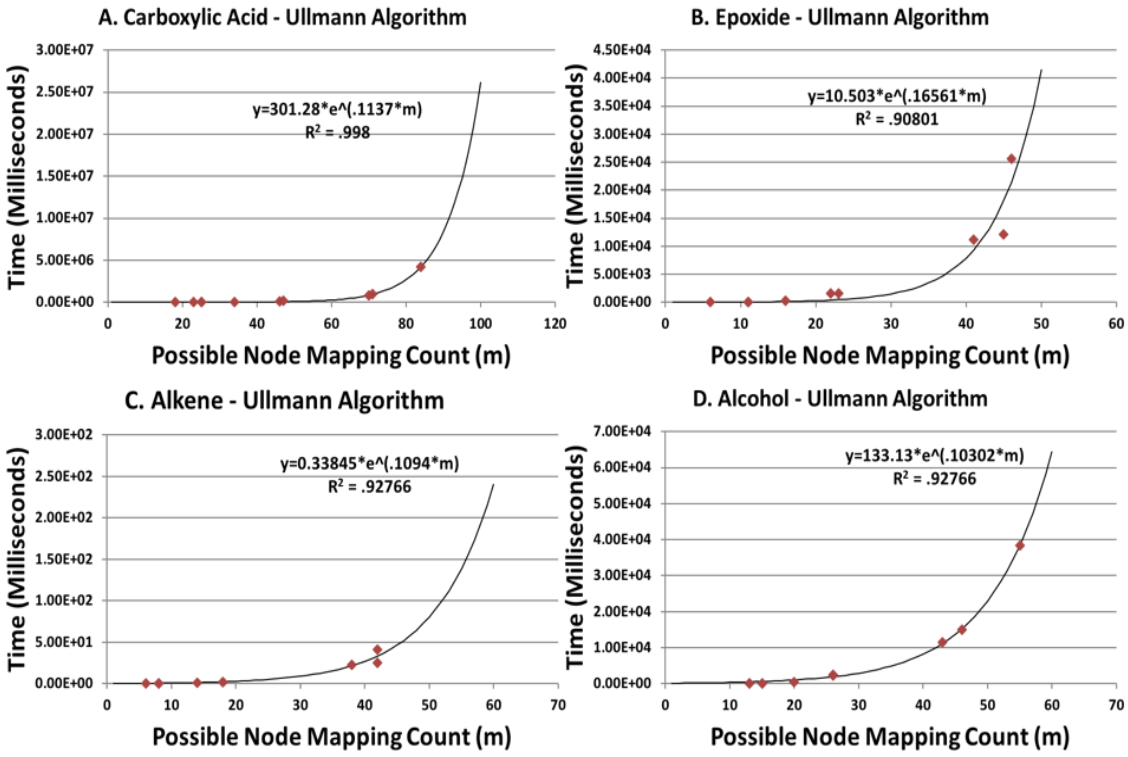

Figure 6.5: Runtime Analysis of the Ullmann Algorithm

In all cases, the time needed for the Ullmann algorithm to find all instances of a functional group in a query compound requires exponential time with respect to m. Alkenes (C) were relatively quick to identify compared to the other functional groups (A,B,D) but still had an exponential trend. These finding imply that the Ullmann algorithm's runtime does not scale sufficiently well to enable the comparison of large compounds or database quantities of compounds.

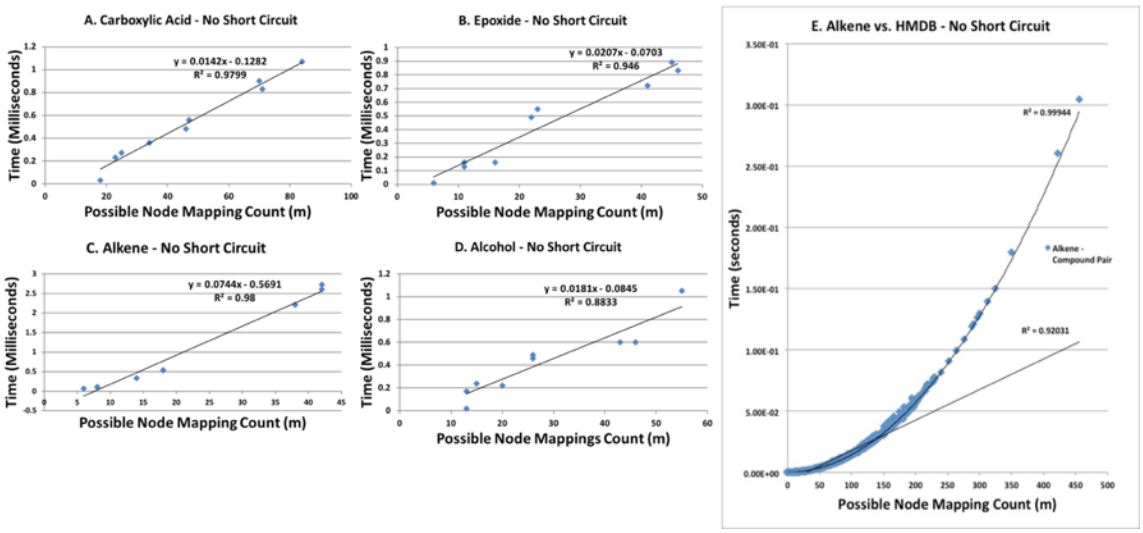

Figure 6.6: Runtime Analysis of the CASS Algorithm

The time needed for CASS to identify the instances of functional groups in the test compounds appears pseudolinear for small values of m (A, B, C, D). Pseudolinear scaling represents a significant improvement over the Ullmann algorithm's exponential scaling. Additionally, the CASS algorithm was significantly faster in absolute time than the Ullmann algorithm for the same task. The ability of the CASS algorithm to scale to large comparisons is shown in panel E. The time needed to find all alkenes in the HMDB demonstrates that CASS actually has polynomial scaling with respect to m but will remain sufficiently fast for even the largest queries.

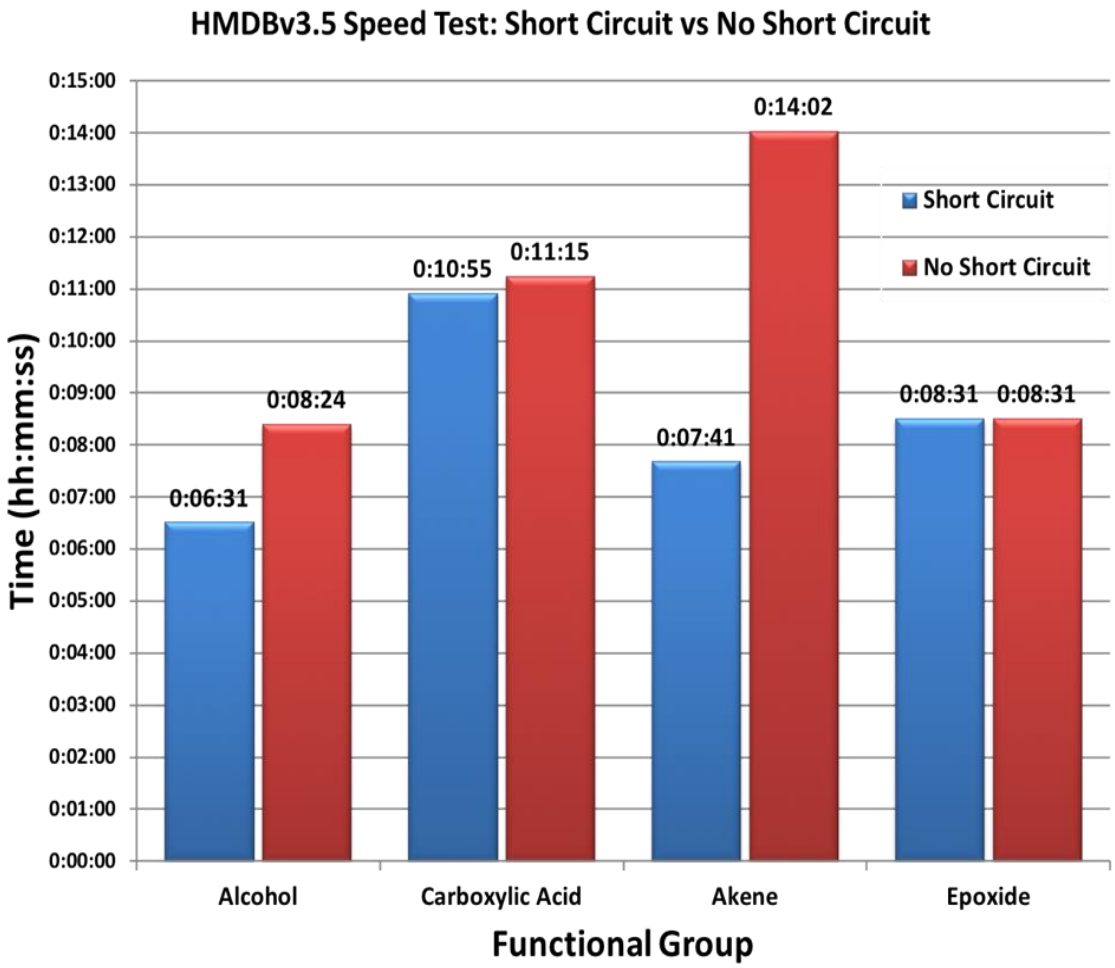

Figure 6.7: Effects of Short Circuiting on CASS Runtime

Short circuiting allows CASS to terminate the search for isomorphisms when the first isomorphism is found. When the number of possible isomorphisms is large, short circuiting can save a large amount of time, but when they are relatively rare, the effect is minimal. This effect explains why common functional groups such as Alkenes show a substantial time improvement with short circuiting but epoxides do not. Short circuiting is most suitable for determining stereoisomerism.

6.3.2 Systematic Isomer and Stereoisomer Analysis

Using our combined HMDB and KEGG metabolite database the number of distinct molecular formulas in each database and shared between both databases were determined. This identified 6094 formulas unique to the HMDB,

5521 unique to KEGG and 3788 formulas shared between the two databases. These shared formulas were then tested to determine if they were isomers in neither, both or only of those databases. This identified that 39% of these formulas were isomers in neither, 32% were isomers in both and 29% were isomers in only one database.

Additionally, the trend in isomeric content over time in both databases and our combined database were determined. The HMDB's isomeric formulas content and our combined database has plateaued at 43 and 46% respectively. KEGG has reached a 28% isomeric content. The inclusion of pharmaceutical and synthetic compounds in KEGG possibly explains this observation. Additionally, in all three databases, the number of isomeric entries is much larger than the number of isomeric formulas, implying that on average a single isomeric formula is represented by several isomeric entries in the databases (Figure 6.8).

The composition of the databases with respect to stereoisomerism was determined as well. Entries with duplicate names were excluded from this analysis as they likely represent identical compounds. The HMDB and KEGG both include stereoisomers with 1.14% and 9.43% of each database consisting of stereoisomers. The combined database has a percent stereoisomerism of 8.3%. This difference can again be explained by the portions of the metabolome represented by each database. The HMDB contains mostly lipids which have many structural isomers but few stereoisomers, while KEGG has many sugars which have many stereoisomers.

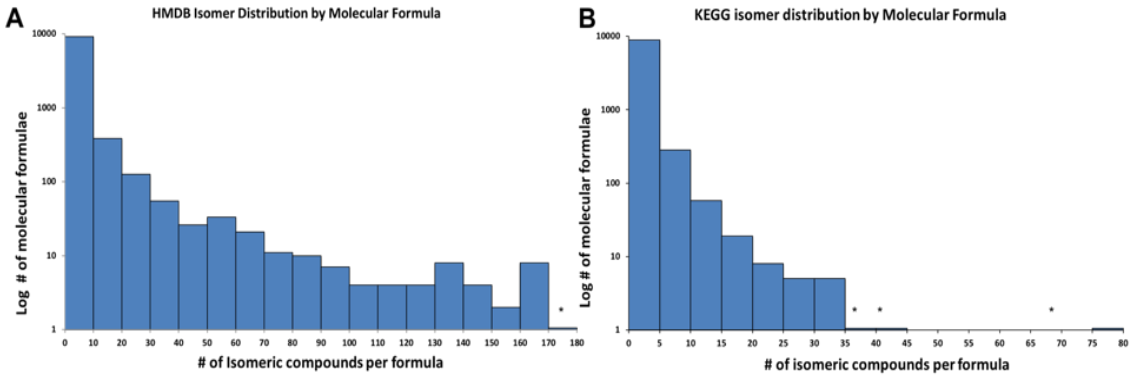

Figure 6.8: Isomer Distribution in HMDB and KEGG
Shown in panel A is the distribution of isomeric compounds per formula in the HMDB. In the HMDB, many isomeric formulas correspond to more than 40 compounds and some have more than 100! This is largely due to the presence of many isomeric lipids in the HMDB that share the same formula. The distribution is very different in KEGG as shown in panel B. Only three formulas in KEGG map to more than 35 formulas. This implies that compounds in KEGG will be easier to disambiguate than compounds in HMDB using CS-tagging and elemental molecular formula. This also implies that lipids as a category could be more difficult to disambiguate with CS-tagging than other classes of metabolites.

6.3.3 CS-Tagging Strategy Analysis

The optimal strategies for 3,5, 10 or 15 functional groups under stoichiometric, pseudostoichiometric and non-stoichiometric conditions with different allowed forms of overlap were determined. In all analyses, a maximum of 15 iterations were performed with the top 15 strategies kept in subsequent iterations and a performance cutoff of 0.1%. Without CS-adduct formation, only 17.13% of compounds in the combined database and 40.98% of compounds in the KEGG database can be identified from molecular formula alone. This represents the baseline performance against which different strategies can be compared. Furthermore, functional group information will be unable to resolve stereoisomers and thus the maximum performance of any strategy is the percent

of compounds in each database that have no stereoisomer which represents 91.7% of the combined database and 90.6% of KEGG.

Stoichiometric CS-tagging consistently generated the best increases in percent unambiguously assignable compounds with the ideal strategy of 3 functional groups improving assignment coverage from 17.13% to 30.35% in the combined database and from 40.98 to 61.63% in the KEGG database. Strategies with 15 functional groups achieved 36.67% in the combined database and 69.13% in the KEGG database. Allowing for the detection of overlapping, subgraph or super functional groups only offers minimal improvement; less than 1% for 3 functional groups and less than 2.5% for strategies of 15 functional groups.

Non-stoichiometric strategies unsurprisingly provide the worst improvements in assignment accuracy. The 3 functional group strategy under non-stoichiometric conditions enables 23.18% unique assignment in the combined database compared to 30.35% for stoichiometric. A similar decrease occurs in the KEGG database as well. For non-stoichiometric strategies detection of overlapping groups offers only marginal improvement as well. As with stoichiometric strategies a point of diminishing returns is reached after 10 functional groups.

Neither stoichiometric nor non-stoichiometric CS-tagging is expected. Pseudostoichiometric CS-tagging likely best represents expected patterns of CS-tagging. In general, pseudostoichiometric strategies perform slightly worse than stoichiometric strategies and much better than non-stoichiometric ones. The

optimal strategy of 3 groups allows for the unique assignment of 28.37% and 59.32% of compounds in the combined and KEGG databases and increases to 35.83% and 68.13% respectively with 15 groups. Again, overlap detection provides few improvements.

Strategies using only the super functional groups under stoichiometric, pseudostoichiometric, and non-stoichiometric conditions all performed poorly. During enumeration the algorithm terminated early in most cases due to the performance cutoff.

The most common functional groups in strategies with five or fewer functional groups are alkene, methyl, ketone, carboxylic acid, dialkyl ether and enol. The performance of these functional groups alone is shown in Supplemental Tables 6.2, 6.3 and 6.4 for stoichiometric, pseudostoichiometric and nonstoichiometric adduct formation for the KEGG and combined database. Reagents capable of derivatizing these functional groups in a detectable manner are optimal for disambiguating metabolite assignments and reagents already exist for most of these functional groups. No CS-tagging reagent exists for methyl groups nor can one be easily developed due to the groups lack of reactivity. This strongly suggests that direct infusion mass spectrometry combined with CS-tagging cannot use methyl group counts to help disambiguate ambiguous metabolites.

6.3.4 BASS Reliably Identifies KEGG Atom Types

Using the set of manually constructed and generated reference graphs for the KEGG atom types, KEGG atom types can be predicted. Recapitulating the KEGG atom types on all the KEGG entries resulted in a 97.5% agreement between the database annotated types and the predictions. Manual inspection of a subset of the entries where disagreement was observed revealed that the database atom types were inconsistent with their descriptions from the KEGG website. Three common errors were observed explaining these discrepancies: failure to detect a cycle in a compound, confusion between esters, carboxylic acids and their sulfur analogues, and inconsistent application of atom types to aromatic atoms. The agreement between the BASS-assigned atom types and the KEGG-assigned types are shown in Figure 6.9.

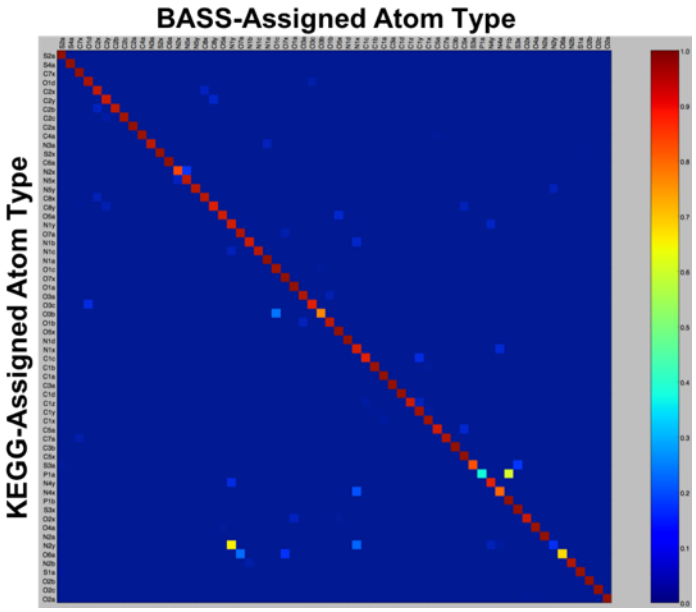

Figure 6.9: BASS-predicted KEGG Atom Types versus KEGG Atom Types
BASS predicted KEGG atom types agree with the database assigned atom types for the majority of the atoms in KEGG database. Where disagreements are seen, they are the result of KEGG incorrectly applying their rules to the structures. Notably, many cyclic structures are missed by the KEGG entries and many atoms have annotations corresponding to other element types.

6.3.5 Atom Coloring Identifies Possible NMR Equivalent Nuclei

When the optional charge, stereochemistry and isotope fields are added to atom colors, the resulting colors are highly descriptive of the substructure centered at that atom. As a result, atom coloring is useful whenever comparisons between local structure are needed and this use case extends beyond CASS/BASS.

One application is the need to identify potentially magnetically equivalent nuclei within a chemical structure. Nuclei located within the same chemical substructure are highly likely to be magnetically equivalent. For the case of

hydrogen and J-coupling the local structure out to three bonds (i.e. corresponding to the d2 color) of the nuclei describes its relevant substructure. Atoms with identical colors reside in the same chemical substructure and excluding effects from the 3D arraignment of the nuclei, are likely to represent a single spectral feature. For example, the hydrogens of the red methyl groups in Figure 6.10C will likely produce identical sets of ^{1}H-NMR signals with the same relative intensity and splitting patterns. Thus, by coloring all nuclei in a molecule at d2 then finding sets of identically colored nuclei, a subset of nuclei with similar J-coupling patterns can be identified; however, additional work is needed to account for effects beyond 2D substructure similarity (e.g. this will not detect diastereotopic protons). This use of atom coloring has been incorporated into Andrey Smelter's Isotopic Enumerator package (Smelter and Moseley, 2018).

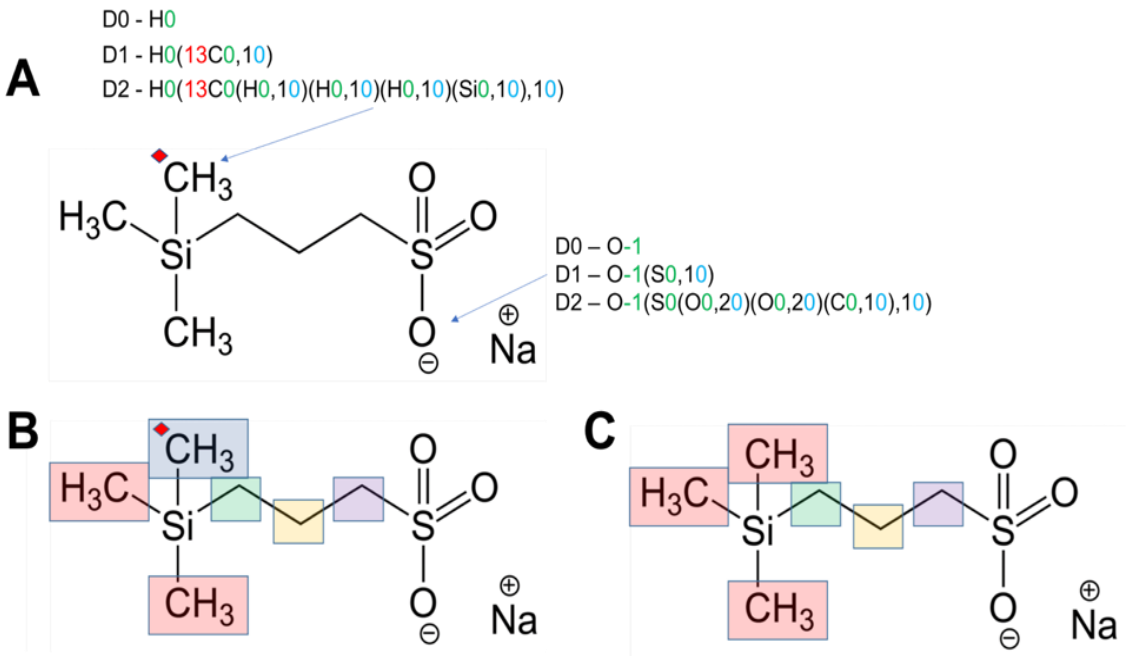

Figure 6.10: Possible NMR Equivalent Nuclei Predicted by Atom Coloring
The d0-d2 colors of selected nuclei are shown in panel A. Each color contains charge information (green) and isotope information (red) if the nuclei is not the most abundant isotope of that element. Panel B shows groups of equivalent nuclei identified by atom coloring when one methyl carbon is labeled with ^{13}C. Panel C shows the same result when that methyl group contains ^{12}C. Not all identically colored atoms are equivalent due to effects that extend beyond simple substructure but atom coloring does make an efficient screening test to identify possible equivalent nuclei automatically.

6.3.6 Application of Atom Coloring Based Structural Similarity Metrics

Additionally, atom coloring can provide a mechanism to estimate the structural similarity between a pair of compounds. First all atom colors out to a depth of N are generated for all atoms in the pair of compounds. Second, all atom colors at depth N are collected for all atoms in the compound to give a list of atom colors. Third, the Tanimoto coefficient between the list of colors in compound A and compound B are calculated(Chen and Reynolds, 2002). This returns a value between 0 and 1 that estimates overall structural similarity by

comparing the substructure similarity between two compounds. This structural similarity metric has been successfully applied for the construction of a reference metabolite interaction network in coronary artery disease (Trainor *et al.*, 2018).

6.4 Discussion

Our new algorithms, CASS and its more capable version BASS, significantly outperform the Ullmann algorithm in finding subgraph isomorphisms in chemical structures. The stark difference in the performance of CASS/BASS compared to the Ullmann algorithm highlights that despite being the prototypical solution to the common subgraph isomorphism problem, its performance makes it unsuitable for solving this problem on all but trivial structures. Additionally, in the process of modernizing the Ullmann algorithm, a typographical mistake was found in the original publication.

Using our algorithm, a functional group annotated versions of KEGG and the HMDB were created that can be queried with information expected from FT-MS CS-tagging. Storing these databases as SQLite provides several advantages such as portability, easy of query and improvements in database access speed. Furthermore, the custom mol file format created for functional group descriptions enables the description of complex substructures within a chemical context. Furthermore, these substructure descriptions can be created on the fly and searched for using CASS without any changes to CASS's code.

Analyzing the functional group annotated databases revealed a high percentage of isomeric molecular formulas in both databases, highlighting the

need for techniques such as CS-tagging for FT-MS. Additionally, a high percentage of stereoisomers (9%) were observed in both databases. These compounds are unresolvable using current CS-tagging and FT-MS methods and thus limit the maximum percentage of compounds that can be unambiguously assigned using these methods. However, stereospecific CS-reagents are not inconceivable – chiral resolving agents already exist for disambiguating enantiomers (Vincent and Vigh, 1998). Our analysis of CS-tagging strategies reveals that a set of only three or five reagents when multiplexed can substantially increase the ability to unambiguously assign metabolites provided their elemental molecular formulas are known accurately. Tools such as SMIRFE can provide these formula assignments. CS-tagging was substantially less effective in the HMDB than in KEGG largely due to the large number of isomeric lipids in the HMDB. Resolving isomers of lipids will require additional methods that identify more specific chemical substructures in this class of molecules.

The dramatic differences in the formula content of the HMDB and KEGG illustrate the potential biases of targeted FT-MS assignment. Only 24% of the total formulas between the HMDB and KEGG exist in both databases, implying that features that are not assignable by one database may be assignable by another. Additionally, the observation that some of the shared formulas are isomers in one database but not the other suggests that falsely unambiguous assignments can be generated if only one database is used for assignment. These findings not only strengthen the case for untargeted tools such as

SMIRFE but also that if targeted assignment is unavoidable, multiple databases should be used to prevent misassignment.

While originally designed only to accelerate stereochemistry calculations, the atom coloring methods found several additional applications outside of the original CASS algorithm. The ability to describe chemical substructures succinctly as an atom color string rather than as a graph, enables easy comparisons between different atom colors as a proxy for substructure comparisons. The ability to automatically identify potentially NMR equivalent nuclei will greatly improve the automaticity of NMR analysis tools while the ability to provide numerical estimates of structural similarity through substructure similarity has a variety of applications. The atom coloring structural priors that were generated for Bayesian graphical models could be applied to any modeling approach where a similarity score is needed for a pair of compounds.

6.5 Conclusions

CASS and BASS represent a substantial improvement both computationally and in terms of their capabilities over the original Ullmann algorithm for identifying subgraph isomorphisms. Our algorithms were employed to identify and quantify instances of functional groups in every metabolite entry from the HMDB and KEGG as well as stereoisomeric entries in these databases. The results of this analysis were used to construct a functional group-annotated version of both databases as well as a combined HMDB and KEGG database.

From this combined database it was determined that a large percentage of the metabolome represented by the database was isomeric and not unambiguously assignable from elemental formula alone. Using just 3 or 5 CS-reagents in multiplex to partially determine the functional group composition of the compounds combined with accurate elemental molecular formula assignments significantly increases the percent of unambiguously assignable metabolites. Stereoisomers, which represent 9% of the metabolome, cannot be disambiguated using existing CS-reagents. The differences in the formula content between the HMDB and KEGG were dramatic, highlighting the need for untargeted tools that are not biased towards their databases for assignment of FT-MS spectra.

Finally, the atom coloring methodology developed to accelerate certain calculations in CASS and BASS has applications outside of this algorithm. The atom coloring techniques are useful for describing and comparing chemical substructures and have been used to identify NMR equivalent nuclei and to inform Bayesian methods for metabolite interaction network generation.

CHAPTER 7. CONCLUSION

The untargeted assignment of FT-MS metabolomics datasets represents a significant unsolved problem in the metabolomics community. Existing assignment tools are largely targeted methods that rely upon databases of metabolites for assignment and are therefore ill-suited for discovering new metabolites. Additionally, the incompleteness of and inconsistency between databases can introduce significant assignment bias that can have significant effects on downstream data analysis. Existing untargeted assignment tools on the other hand are either not untargeted in a meaningful sense or rely upon heuristics and other rules that render them unusable in a stable-isotope tracing context. Furthermore, the presence of spectral artifacts and poor data quality in many FT-MS datasets make both assignment methods prone to misassignment and significant interpretative errors. Towards this end, a collection of tools was developed that address these data quality issues and the short comings of existing assignment tools.

The FT-MS spectral artifact detection tools identified a family of FT-MS artifacts that we named high peak density artifacts. These artifacts result in regions of spectra that have significantly higher peak density than the rest of the spectrum. The poor correspondence of peaks in these regions across multiple scans (peaks in these regions appear, disappear, and move around considerably in *m/z* between scans), the *m/z* difference patterns between these peaks, and the observation that the location of these regions depends both on sample

composition and instrument and firmware versions all strongly support our claim that these features are artifactual. Spectral artifacts by themselves are undesirable, but not necessarily problematic, if they are not assigned; however, we observed that targeted assignment tools such as LipidSearch will assign these features. Because the location of these artifacts is sample-specific, these artifactual features allow for the classification of spectra into sample classes using machine learning models such as Random Forest. Ultimately classifiers that utilize these assigned artifactual features are both brittle and do not represent the real biochemical variance between the sample classes.

To eliminate these artifactual features, two methods were developed. The first of these was to remove regions of spectra that consistently contain these features across all spectra in the dataset. A cross-sample methodology for artifact removal was necessary for two reasons. First, our high-peak density artifact detection method was highly sensitive, but not very specific. Removing only regions of spectra with consistently identified artifacts minimizes the amount of unaffected spectra that is removed. Second, when artifactual regions of spectra are removed from each sample individually, this process will encode the presence of artifacts as the absence of peaks in parts of the spectra. In turn, this encoded information can trick the machine learning models in a similar to manner to the presence of the artifacts themselves. Thus, by removing consistent artifact regions across all samples, the specificity of our approach is effectively increased and the accidental encoding of spectra is prevented.

The second method for eliminating these artifacts was the improved peak correspondence and peak characterization. Since artifactual peaks in fuzzy sites have low peak correspondence across scans, many of them can be eliminated using the peak correspondence algorithm (i.e. by detecting their poor correspondence). Peak correspondence and characterization also provide much better estimates of peak intensity by correcting for inconsistent injections and ionization efficiencies between scans. This has the notable effect of greatly reducing the variance in observed peak intensities at the aggregate peak list level. This approach does not result in absolute peak intensities that are comparable to one another, but it is necessary for comparisons of relative peak intensity between peaks – which was necessary for our SMIRFE assignment tool. Additionally, unlike other peak picking methods that only return an observed *m/z* and intensity for each detected peak, peak characterization provides statistical estimates of a peak's *m/z* and intensity, along with the variances of those values across scans as well as in which scans a peak is present.

With many of the data quality problems in our FT-MS spectra resolved with peak correspondence and characterization, a novel algorithm for the untargeted assignment of FT-MS metabolomics datasets called SMIRFE was developed. SMIRFE does not use a database of known metabolites or even known compounds to generate assignments, rather SMIRFE searches a nearly exhaustively large EMF search space that represents all possible formulas within a defined region of theoretical chemical space. Other assignment tools typically query their databases with observed *m/z* values and a tolerance to generate

assignments for a peak, which is inherently error prone due to the lack of cross validating evidence supporting a returned assignment. SMIRFE uses *m/z*-based queries to generate possible assignments for peaks, but then compares the observed peak intensities between potential isotopologues of the same EMF to eliminate unlikely assignments and to assign statistical estimates of assignment quality. The relative intensity ratio between isotopologues can be predicted using the natural abundance probabilities of the isotopologues and these predictions can be calculated, taking the effects of isotopic labeling into consideration. As a result, SMIRFE can assign datasets from stable-isotope tracing experiments with no changes to the underlying algorithm. This methodology could be added to existing targeted tools based on metabolite databases to improve their assignment accuracy.

In our test case consisting of a mixture of isotopically labeled and unlabeled amino acids derivatized with ECF, SMIRFE was able to assign 18 of 19 predicted amino acid derivatives. Manual examination of the spectra demonstrated that the unassigned derivative corresponding to cysteine was not present in the spectra for those samples, implying that SMIRFE assigned all the predicted amino acid derivatives present in each spectrum. From this set of assignments, the *m/z* accuracy of the peak lists generated by our peak characterization method was examined and found to be below the 1ppm error as advertised by the instrument vendor. However, the mass error pattern is not constant with respect to *m/z* and appears to change based on a peak's proximity to one or more other intense peaks. As shown in our simulations, if mass

accuracy can be increased from 1ppm to 0.1ppm or even 0.01ppm, the number of possible assignments for each peak can be greatly reduced. Improved mass accuracy will be necessary for generating unambiguous assignments at high *m/z*. By correcting both the *m/z* dependent and the repulsion dependent mass error we observed, improved mass accuracy can be achieved.

As SMIRFE requires well characterized peaks with low mass error in order to generate less ambiguous assignments, experiments can be optimized to maximize their compatibility with the SMIRFE algorithm. Future experiments seeking to use SMIRFE should dedicate more acquisition time towards MS1, as MS2 data is not used currently in SMIRFE, and the number of scans acquired should be maximized. The optimal settings for microscan settings remain unknown. High dynamic range is desirable to observe the most isotopologues possible for each EMF, but only if the tradeoffs in mass and intensity accuracy that accompany higher dynamic range are minimal. Finally, when isotope labeling experiments are performed, unlabeled controls should be included in replicate for each population.

While these changes in experimental design can improve the accuracy of SMIRFE assigned IMFs and EMFs to observed spectral features, SMIRFE still suffers from one significant limitation – it does not assign metabolite structures to spectral features. Even if IMF and EMF assignments could be made with perfect accuracy, the prevalence of isomers for many metabolites implies that formula assignments will not map uniquely to metabolite structures. Fundamentally, this limitation arises from the inability of mass spectrometry alone to differentiate

isomers and if metabolite structure assignment is necessary, orthogonal information will be required.

The inability to assign metabolite structures to observed spectral features assigned by SMIRFE originally limited our ability to perform differential lipid analysis in our non-small cell lung cancer data set. Traditionally, lipid metabolite features are identified by querying databases of known lipids with observed *m/z*'s from experiments then inferring the class or category from the database. Obviously, this approach is not suitable for SMIRFE as it undermines much of the untargeted nature of the approach. To solve this problem, machine learning models were built to predict lipid category and class from assigned formulas directly. For all categories, these models achieved excellent accuracy and precision implying that for distributions of formulas similar to those observed in existing metabolite databases, trustworthy predictions of lipid category and to a lesser extent, lipid class, can be generated.

Applying these models to large sets of formulas generated from a convex hull of HMDB formulas demonstrated that these models have difficulty generating trustworthy assignments for formulas that are significantly dissimilar from the training dataset. For example, formulas with multiple sulfur and phosphorus can classify but also are unlikely to exist. Similarly, large mass errors in spectra will result in incorrect assignments that still classify for much the same reason. Due to the large search space of SMIRFE almost any pair of peaks corresponding to a pair of isotopologues will be assigned to IMFs, but the assigned IMFs will only be limited to the correct IMFs when the mass error is relatively small. Eliminating

the extra untrustworthy assignments can be achieved through a combination of heuristics and cross-sample assignment correspondence. Incorrect formula assignments tend to correspond less well across samples than correct assignments, but inferring this pattern requires many samples. Improvements in peak characterization and / or instrumentation that result in improved peak intensity measurements will also reduce the inclusion of incorrect formula assignments.

This sample correspondence method was successfully used in our lung cancer dataset (which contains ~180 samples) to filter out many inconsistent assigned features. Differential abundance analysis on lipid classified consistently assigned features revealed a subpopulation of NSCLC samples with substantially up-changed relative concentrations of sterol categorized features. These findings were consistent with previously reported observations in the scientific literature that have reported that NSCLC survival is improved in patients using statins, a class of drugs that inhibit the mevalonate pathway and interfere endogenous sterol production. To-date these findings have simply been observational. Statins have no clinically indicated use in cancer but due to the large number of patients prescribed statins for the treatment of high cholesterol or prophylactically in patients with a history or at risk for cardiovascular vascular disease, a significant number of lung cancer patients happen to be prescribed statins. While further testing is necessary to confirm our findings, if sterol upregulation is crucial to the development of a subpopulation of NSCLC, which currently lacks effective

treatments for late stage disease, statins provide a potential therapeutic option that should be explored.

CS-tagging represents one method by which additional structural information could be obtained in direct infusion FT-MS experiments that is compatible with SMIRFE or other assignment methods. Through CS-tagging, the partial functional group composition of compounds can be determined and this information used to limit possible assignments to metabolite structures with those functional groups. The lack of a functional group annotated metabolite database however made leveraging the information that could be gained from CS-tagging for metabolite assignment effectively impossible. Towards this end, we developed algorithms that can identify substructures within chemical structures efficiently and used these algorithms to build a functional group annotated database of metabolites. Additionally, with these databases, the set of ideal functional groups to target for CS-tagging was identified. The mathematical formulation of the chemical substructure identification problem called the maximum common subgraph isomorphism problem is a widely encountered problem in graph theory and our algorithms provide best in class performance for this problem. Furthermore, the mathematical tools developed to tackle this problem have uses in a much wider context of chemoinformatics problems. The atom coloring algorithms in particular have applications in NMR and in the construction of metabolite interaction networks. Ultimately, these mathematical tools will be helpful in the construction of atom-resolved metabolic networks that

will allow the automated deduction of atom tracing through the metabolic network.

In conclusion, these tools form a prototype FT-MS data analysis pipeline that is suitable for large scale metabolomics experiments utilizing stable-isotope tracing. Better quality peaklists with more reliable spectral features can be generated and then assigned in an untargeted manner. These formulas can be assigned to biologically meaningful metabolite categories and classes using machine learning and then used in differential abundance analyses to identify molecular phenotypes of interest.

CHAPTER 8. FUTURE DIRECTIONS

In their current state, the tools described herein form a prototype FT-MS data analysis pipeline that could be used to power untargeted analyses of disease and non-disease metabolism. However, improvements in several key areas for each tool could dramatically improve the overall performance of this pipeline and expand the set of experiments to which it can be applied.

8.1 Determining the Origin of Fuzzy Site Artifacts

Although our methods for handling spectral artifacts identify and eliminate fuzzy sites from FT-MS spectra, the source of the fuzzy site artifacts remain largely unknown. Our leading hypothesis for the cause of fuzzy site artifacts is radio frequency interference from the nanoelectrospray system. This would explain why the artifact locations are different between instruments (i.e., they have different ESI systems) and why the artifacts differ based on sample composition, as the electrospray system changes the pulses used for ionization depending on sample conditions. One approach to testing this hypothesis is to swap the electrospray systems on our two instruments and test if the artifact locations also “swap” on a solvent-only sample. If the fuzzy site locations appear to “swap”, this strongly suggests that the root cause of these artifacts is the electrospray system. If the fuzzy site locations remain the same, this suggests that the instruments themselves may be the cause. Additionally, since the appearance of these artifacts change with both microscan and resolution parameters on the instruments, these could be optimized to identify the settings

that minimize the severity of these artifacts. Although both peak characterization and artifact removal largely eliminate these artifacts, preventing these artifacts in the first place is desirable.

8.2 Peak Characterization

Currently, peak characterization greatly improves the consistency of reported peak intensities and is a necessary step to assign peaklists using SMIRFE. However, the ability of SMIRFE to generate unambiguous assignments at high *m/z* is largely limited by the mass accuracy of the instrument. A 1ppm mass error, while small compared to other mass spectrometry techniques, translates into a large range of possible assignments at high *m/z* that cannot be disambiguated without substantial improvements in intensity accuracy or a modest improvement in mass accuracy. Peak characterization could potentially be modified to improve mass accuracy in one of two non-mutually exclusive ways.

First, if replicate spectra are collected across multiple instruments that have slightly different mass accuracy profiles, this information could be used to shift peaks toward their true *m/z*. For example, if instrument one has an increasing positive mass error across *m/z* and instrument two has an increasing negative mass error across *m/z*, a corresponded peak found in spectra from both instruments has a true *m/z* in between the two reported *m/z*'s from both instruments. Determining this mass error profile could be achieved using large amounts of replicate spectra and correspondence or with standards added to

each sample. Standards introduce potential artifacts but is the most straightforward way to determine the mass accuracy pattern on an instrument.

Alternatively, common contaminants such as plasticizers and keratin could be used as internal calibrants and used to calibrate a spectrum after acquisition. Some forms of chemical contamination are effectively impossible to eliminate and are likely to behave similarly in samples of similar composition. This approach could be used in conjunction with peak correspondence in lieu of traditional standards, but would require an iterative process of assigning spectra to identify contaminants and correcting based on assigned contaminant peaks before reassigning.

Additionally, if peak correspondence could be adapted to work with GC or LC-MS experiments, additional measurements of peak intensity could be obtained from the chromatographic peak areas. Noise and artifactual peaks could be eliminated by observing that they do not correspond to a chromatographic peak as well. This would have the additional benefit of opening up our methods to the broader metabolomics community who largely utilize GC or LC-MS experiments instead of direct infusion.

8.3 Optimizing Experimental Designs for Peak Characterization and SMIRFE

Although improvements in peak characterization will ultimately provide more descriptive peaklists that will improve SMIRFE assignments, there is no substitute for better quality raw data. Currently, SMIRFE and the related tools

have only been applied to spectra that were acquired without the limitations of our methods in mind and thus the spectra we have worked with so far were collected using various numbers of scans, microscans, resolutions, etc. and the number of replicates in each experimental dataset has varied significantly. Quantifying the effects of changing each of these parameters with respect to assignment accuracy and cross-sample assignment correspondence provides a mechanism by which optimal values for these parameters can be determined.

Therefore, a set of experiments should be performed to identify the ideal values of microscan and total ion target for a target spectrum acquisition time. Blood plasma lipid extracts would be an ideal candidate for these experiments given that it is biological in nature, reasonably complex and cheap. These spectra would be peak characterized and assigned with SMIRFE. The settings that produce the best e-values for validated assignments and produce the best intensity relative standard deviations (RSDs - standard deviation divided by mean) and smallest *m/z* errors represent the optimal settings. Validated assignments could be confirmed using CS-tagging. If the settings that are optimal for intensity RSD are not optimal for *m/z* accuracy (or vice versa), a family of optimal settings could be identified and applied on a per-experiment basis. Additionally, the effect on cross-sample assignment correspondence could be investigated with a sufficiently large number of replicates.

8.4 Improving SMIRFE Improvements through Improved Scoring

The high mass error with respect to the mass resolution currently limits the statistical comparisons that SMIRFE can use to disambiguate possible assignments. If the mass error was relatively small compared to the mass resolution, the window of possible assignments could be tested by comparing the absolute mass difference between a peak's *m/z* and a possible assignment's *m/z* based on the resolution of the peak. Combined with existing tests on the relative intensity of peaks versus the expected ratio based on natural abundance probabilities, the field of possible assignments for a peak could be greatly reduced.

Scoring could also be improved by including a metric that compares how close an assigned formula resides to a known metabolite formula in EMF space. Formulas that are very similar to known metabolites will have better scores in this system, while formulas very different from known metabolites will have smaller scores. This would bias results towards formulas that are similar to known metabolites but could significantly disambiguate assignments. Furthermore, when sufficient evidence is present, the components of the score from *m/z* matching and intensity ratio matching could overcome this weighting to enable the assignment of novel metabolites when there is sufficient spectral evidence.

8.5 Investigating Lipid Profiles observed in NSCLC Samples

Our findings that sterols and glycerolipids were significantly up-changed in a sub-population of NSCLC implies that statins or other sterol altering drugs could have potential therapeutic effects in some NSCLC patients. However, it is not clear from which upstream mutations this metabolic phenotype results. Understanding the genetic drivers of these metabolic differences will be necessary to identify these patients clinically using genomics methods. Identifying these mutations will require genomic or transcriptomic data that could be acquired from additional patient samples. Of interest would be if the observed mutation patterns are mutually exclusive with other known mutation patterns in subtypes of NSCLC. This would suggest that the mutations that are responsible for this molecular phenotype are in fact drivers of NSCLC rather than just a byproduct of other upstream genomic changes. Enzymes and proteins in these pathways are potential drug targets if the upchanged sterols and glycerolipids are important for cancer development.

Additionally, our hypothesis that statins could be used therapeutically in this subtype of NSCLC can be tested experimentally. Patient tumors could be analyzed using the existing methodology to identify cancer samples with this molecular phenotype. These samples could then be cultured with and without various statins to observe if an effect is observed on the growth of the cancer cells derived from these tumors. However, differences between cell culture versus in vivo cancer could limit the accuracy of cell culture testing of statins with respect to clinical performance. 3D cell culture may be a potential remedy to

some of these discrepancies (Riedl *et al.*, 2017). Alternatively, patient-derived xenograft mice models could be used, although these are likely to have other different limitations (Hidalgo *et al.*, 2014). Additionally, using stable isotope tracing, changes of ^{13}C incorporation into sterols would confirm if statin treatment influences sterol biosynthesis in the samples when cholesterol uptake from the diet is controlled. If an effect is observed with statin exposure on tumor growth discriminating between a cytotoxic or cytostatic effect will be necessary. These can be distinguished from one another using measures of cellular respiration (Cummings and Schnellmann, 2004), cell morphology (Bortner and Cidlowski, 2001) (Zhang *et al.*, 1999), or the presence of cellular components in the extracellular space.

A potential prophylactic role of statins could be investigated using chemically inducible lung cancer models in mice (Safari and Meuwissen, 2015). While the chemically inducible forms of lung cancer likely differ from the cancers present in human patients, if statin exposure prolongs survival in the mouse models significantly, this suggest a possible prophylactic role. While this is a long shot, if statins have a prophylactic role in at risk NSCLC patients (namely former smokers), the potential healthcare implications would be enormous.

One further aspect of the lipid profile differences to investigate is the possibility of upchanged sterol esters between cancer and non-cancer. Presence of sterol esters in an NSCLC tumor sample could be verified using NMR or tandem-MS. If sterol esterification is enhanced in NSCLC cancer cells, ACAT-1 inhibitors could be tested using a similar design to that proposed for examining

the effects of statins. Determining from which cells in the tumor these sterol esters are produced, foam cells, cancer cells, or both, is challenging. One approach is to attempt to correlate the number of foam cells to sterol ester levels. No correlation or weak correlation suggests the major source is another cell type. Alternatively, inhibitors of foam cell differentiation could also be employed such as IL-33 (McLaren *et al.*, 2010); however, these may have unexpected effects on the cancer cells as well.

8.6 Aromaticity and Tautomer Detection for BASS

The improvements introduced in BASS enable the original CASS algorithm to handle chemical structures that contain resonant and aromatic bonds; however, the BASS methodology does not directly detect these features, rather they are inferred from KEGG atom type annotations that come from the database entries or added using BASS. A general approach for detecting resonance and aromaticity would eliminate the method's dependency on KEGG atom types and provide more accurate predictions on chemical structures that contain substructures not described by KEGG atom types. Additionally, tautomeric structures are currently not handled correctly by BASS as there is no single graph that can represent a tautomer correctly.

A solution to all these problems is to create an algorithm that mimics the electron "pushing" that a chemist would perform on paper to identify resonant, aromatic and tautomeric structures. Valid structures that represent the movement of a pair of electrons would allow resonant bonds to be detected directly.

Aromaticity can be tested by checking each valid structure against Huckel's rules. Tautomers could be generated if the rearrangement of hydrogens was considered. Metrics for the likelihood of hydrogen rearrangements would be necessary to prevent non-labile hydrogens from rearranging. Handling these edge cases will be necessary to handle the full range of chemical features present in biological compounds.

APPENDICES

APPENDIX 1. SAMPLE DESCRIPTIONS

A1.1 Preparation of Solvent Blanks with and without Standards (Sample A)

The solvent blank was composed of Isopropanol:Methanol:Chloroform 800 µl:344 µl:200 µl. The solvent blank was mixed with 28 µl 1 M ammonium formate (final ~20 mM; Aldrich #516961), and without or with 70 µl 1:10 diluted Avanti SPLASH™ Lipidomix® Mass Spec Standard (cat# 330707) in MeOH. The solvent blank without or with lipid standards was loaded onto a 96-well polypropylene PCR plate (USA Scientific cat# 1402-9800) and 15 µl was injected into Fusion 1 by direct infusion through an Advion nanomate. Various resolution and microscan settings were tested in positive mode with 7 min acquisition, normal mass range between 150-1600 *m/z*, S-lens RF level 60%, AGC target 1e5, maximum injection time 100 ms, and Easy-IC on.

This sample was prepared by Dr. Qing Jun Wang at the University of Kentucky. Mass spectrometric analysis was performed by Dr. Wang using the Fusion instruments at the Center for Environmental and Systems Biochemistry under the direction of Drs. Teresa Fan, Andrew Lane and Richard Higashi.

A1.2 Preparation of IC-MS standards from NSG Mice Liver (Sample B)

The NOD/SCID gamma (NSG) mouse colony was maintained by the Center for Environmental and Systems Biochemistry within Division of Laboratory

Animal Resources at the University of Kentucky. The initial breeding pairs were purchased from The Jackson Laboratory in Bar Harbor, ME. The mice were housed in a climate-controlled environment with a 1410 hours light / dark cycle and lights-on at 0600 hours. The mice had free access to food and water. The mice were feed a liquid diet base containing casein, L-cystine, soy oil, cellulose, mineral mix (AIN-93G-MX), calcium phosphate, vitamin mix (AIN-93-VX), choline bitartrate, tert-butylhydroquinone and xanthan gum purchased from Harlan Laboratories (Madison WI).

Mice were euthanized by spinal dislocation and livers excised and flash frozen in liquid nitrogen within 5 minutes of euthanization. Frozen tissues were ground into powder under liquid nitrogen to <10 μm particles using a Spex freezer mill. Approximately 0.5 g of the powder was extracted with 50 ml acetonitrile:water (6:4, v/v). After centrifugation at 22 kg and 4°C for 20 min, the supernatant containing polar extracts was distributed into aliquots and lyophilized for long-term storage at -80 °C.

Immediately before IC-FTMS analysis, the lyophilized powder was reconstituted with water and 10 μl was injected onto an ICS5000+ system (Dionex) interfaced to the FT-MS (Fusion 2). Data were acquired in negative mode at a resolving power of 500,000 (at *m/z*=200) over 52 min of chromatography. The mass range was set between *m/z* 80 and 700, maximum injection time was 100 ms with 1 microscan, AGC target was 2e5, S-lens RF level was 60%, and Easy-IC was turned on for internal mass calibration. The chromatograph was outfitted with a DionexIonPac AG11-HC-4μm RFIC&HPIC

guard (2x50mm) guard column upstream of a DionexIonPack AS11-HC-4μM RFIC HPIC (2x250mm) column (Sun *et al.*, 2017). This analysis was selective for negatively charged ions only.

These samples were prepared and analyzed by mass spectrometry by Dr. Qiushi Sun at the Center for Environmental and Systems Biochemistry at the University of Kentucky under the direction of Drs. Teresa Fan, Andrew Lane and Richard Higashi.

A1.3 Preparation of Ethylchloroformate Solvent and Derived Amino Acids (Sample C)

The ECF solvent blank was composed of acetonitrile:water 9:1 (v/v) with a concentration of 20 μM NaCl to convert positively charged ions into sodium adducts (Yang *et al.*, 2017c).

Unlabeled amino acid standards were purchased from Sigma Aldrich as a mixture of acidic and neutral amino acids and basic amino acids (A6407, A6282). The ^{15}N-labeled amino acids mixture was purchased from Cambridge Isotope Laboratories (NLM-6695). Ethylchloroformate was also purchased from Sigma Aldrich.

2 μL of 2.5 mM Sigma A6407 was added to 2 μL of 2.5mM Sigma A6282 and 5 μL of mM unlabeled glutamine were combined to yield a sample containing unlabeled and labeled amino acids. In the hood, 100 μL of H_2O/EtOH/Pyridine (6:3:1 by volume) was added to each sample and vortexed to let the sample dissolve. 5uL of ECF was then added to sample. The samples were then

vortexed for 30 seconds and then spun down. Under the hood, 100 µL of $CHCl_3$ (RT) were added and shaken vigorously at 3000rpm for 3 minutes with the Disruptor Genie. The samples were then centrifuged at 21100g on a Thermo Scientific Legend Micro 21R for 10 minutes 4C. The bottom $CHCl_3$ layer was then transferred to a new glass vial and capped. 10 µL of 7M NaOH was added into the remaining aqueous phase to adjust the pH to 10. An additional 5 µL of ECF was then added and the samples centrifuged and extract again to ensure complete derivatization. The samples were then vortexed again and then diluted 10x in 90% acetonitrile (in water) with 0.2 µM of tetraethylammonium and 2) µM NaCl in 0.5mL snap-cap tubes. These were then loaded onto 384-well plate that was pre-washed twice with ddH_2O and 1x with acetonitrile for direct infusion FTMS.

FTMS analysis was performed using a Tribrid Fusion Orbitrap interfaced with an Advion Triversa Nanomate. The Nanomate's operating voltage was 1.5kV and the head pressure was 0.5 psi. Spectra were acquired in positive mode. The maximum ion time for automatic gain control was set to 100ms, 5 microscans were acquired per scan and total acquisition time was approximately 5 minutes, resulting in approximately 100 total scans. Spectra were acquired for *m/z*'s in the 100 to 1000 range selected using quadrupole isolation.

These samples were prepared and analyzed by mass spectrometry by Drs. Qing Jun Wang, and Dr. Ye Yang at the Center for Environmental and Systems Biochemistry at the University of Kentucky under the direction of Drs. Teresa Fan, Andrew Lane and Richard Higashi.

A1.4 Preparation of Paired Lipid Extracts from Suspected Human Non-Small Cell Lung Cancer and Non-Cancer Lung Tissue Samples (Sample D)

Eighty-six patients with suspected resectable stage I or IIa primary non-small cell lung cancer (NSCLC) and without diagnosed diabetes were recruited based on their surgical eligibility. The extent of resection was determined by the surgeon in accordance with clinical criteria. Many of the specimens were obtained from wedge resections which minimizes surgery time while the other specimens were acquired in less than 5 minutes after the pulmonary vein was clamped. Both techniques minimize ischemia in the resected tissues. Immediately after resection, the tumor was transected and section of cancerous tissue and surrounding non-cancer tissue at least 5 cm away from the tumor were immediately flash frozen in liquid nitrogen and stored at <80°C. On-site pathologists confirmed the diagnosis and cancer-free margins on parallel tissue samples. All samples were collected under a University of Louisville approved Internal Review Board (IRB) protocol and written informed consent was obtained from all subjects prior to inclusion in the study (Sellers *et al.*, 2015b).

The frozen samples were pulverized under liquid nitrogen to <10 μm particles using a Spex freezer mill, and extracted using a modified Folch method as previously described (Ren *et al.*, 2014). The lipid fraction was supplemented with 1 mM butylated hydroxytolune and then dried by vacuum centrifugation at room temperature. Samples for FT-MS analysis were redissolved 200-500 μl chloroform/methanol (2:1) supplemented with 1 mM butylated hydroxytolune. Reconsitituted lipids samples were diluted in in isopropanol/methanol/chloroform

4/2/1 (v/v/v) with 20 mM ammonium formate (95 µl of solvent for 5 µl of sample) before direct infusion.

Ultrahigh resolution (UHR) mass spectrometry was performed on the paired samples using a Thermo Orbitrap Fusion interfaced to an Advion Nanomate nanoelectrospray source using the Advion “type A” chip, also from Advion, inc. (chip p/n HD_A_384). The nanospray conditions on the Advion Nanomate were as follows: sample volume in wells in 96 well plate – 50 µl, sample volume taken up by tip for analysis – 15 µl, delivery time – 16 minutes, gas pressure – 0.4 psi, voltage applied – 1.5 kV, polarity – positive, pre-piercing depth – 10 mm. The Orbitrap Fusion Mass Spectrometer method duration was 15 minutes, and the MS conditions during the first 7 minutes were as follows: scan type – MS, detector type – Orbitrap, resolution – 450,000, lock mass with internal calibrant turned on, scan range (*m/z*) – 150-1600, S-Lens RF Level (%) – 60, AGC Target – 1e5, maximum injection time (ms) – 100, microscans – 10, data type – profile, polarity – positive. For the next 8 minutes, the conditions were as follows for the MS/MS analysis: MS properties: detector type – Orbitrap, resolution – 120,000, scan range (*m/z*) – 150-1600, AGC Target – 2e5, maximum injection time (ms) – 100, microscans – 2, data type – profile, polarity – negative; monoisotopic precursor selection – applied, top 500 most intense peaks evaluated with minimum intensity of 5e3 counts; data dependent MS^n scan properties: MS^n level – 2, isolation mode – quadrupole, isolation window (*m/z*) – 1, activation type – HCD, HCD collision energy (%) – 25, collision gas – Nitrogen, detector – Orbitrap, scan range mode – auto *m/z* normal, Orbitrap resolution – 120,000, first mass (*m/z*) – 120, maximum injection time (ms)

– 500, AGC target – 5e4, data type – profile, polarity – positive. The ion transfer tube temperature was 275^{o}C. (Yang *et al.*, 2017b).

These samples were collected and prepared under the direction of Dr. Andrew Lane, Dr. Teresa Fan, and Dr. Richard Higashi at the University of Kentucky and at the University of Louisville. Mass spectrometry was performed by Drs. Timothy Fahrenholz and Woo-Young Kang at the Center for Environmental and Systems Biochemistry at the University of Kentucky.

A1.5 Preparation of Human Plasma Samples (Samples E)

Mixed gender, unfiltered pooled lithium heparin treated plasma (Seralab Catalog# HMPLLIHP, Lot# BRH1049783) from healthy donors was extracted using previously published protocols (Acharjee *et al.*, 2017). Briefly, 15 µl of plasma was extracted with 100 µl of ultra-pure H_2O in a glass vial (2 ml). 250 µl of MeOH was added, and lipids were partitioned into 500 µl of methyl-tertiary-butyl ether. Following centrifugation (13,000 rpm, 4°C, 4 min), a 20 µl aliquot of the organic layer was transferred to a 96-well glass coated plate (Thermo Fisher). 95 µl of a solution containing 7.5 mM ammonium acetate in Isopropanol:Methanol (2:1 v/v) was also added to the well. Direct infusion high-resolution mass spectrometry was performed using on a Q-Exactive+ Orbitrap (Thermo), equipped with a Triversa Nanomate (Advion). The Nanomate infusion mandrel was used to pierce the seal of each well before analysis, after which, with a fresh tip, 5 µl of sample was aspirated, followed by a 1.5 µl air gap.

These samples were prepared and spectra were collected by Dr. Thomas Wilson and the High Resolution Metabolomics Laboratory at the Institute of Biological, Environmental and Rural Sciences at Aberystwyth University in the United Kingdom.

APPENDIX 2. COMMONLY USED ABBREVIATIONS / TERMINOLOGY

- BASS – Biochemically Aware Substructure Search. An improvement over the CASS tool that is the focus of Chapter 6.
- CASS – Chemically Aware Substructure Search. The chemical substructure search tool that is the focus of Chapter 6.
- CHONPS – an acronym for Carbon, Hydrogen, Oxygen, Nitrogen, Phosphorus and Sulfur. These are the most common elements in molecules found in living systems.
- EMF – Elemental Molecular Formula. An EMF is like a molecular formula in the traditional sense, it is what elements constitute the compound (e.g. $C_6H_{12}O_6$ is the EMF of glucose).
- EMF Clique - The set of all IMFs adducted to the same ion sharing the same EMF. For example, all labeled forms of $^{12}C_6{}^{1}H_{12}{}^{16}O_6$ adducted with sodium belong to the $[Na^+]$-$C_6H_{12}O_6$ clique.
- EMF Subclique - The set of all IMFs adducted to the same ion with the same number and type of labeled isotopes sharing the same EMF. For example, all forms of $^{12}C_6{}^{1}H_{12}{}^{16}O_6$ adducted with sodium containing two ^{13}C due to labeling belong to the $m+^{13}C_2$ $[Na^+]$-$C_6H_{12}O_6$ subclique.
- EMF Superclique – The set of all IMFs sharing the same EMF. For example, all adducted and labeled forms of $^{12}C_6{}^{1}H_{12}{}^{16}O_6$ belong to the $C_6H_{12}O_6$ superclique.
- IMF – Isotope-resolved Molecular Formula. An IMF is like a molecular formula but is isotope-resolved (e.g. $^{12}C_6{}^{1}H_{12}{}^{16}O_6$ is an IMF of glucose).

- m, m+$^{13}C_1$, m+$^{15}N_1$ – shorthand for describing isotopologues. Here m represents the lowest mass isotopologue for a given compound (or EMF). m+$^{N}E_x$ represents an isotopologue where a X instances of the most abundant isotope of the element E has been replaced with the isotope ^{N}E. So for m+$^{13}C_1$ represents the isotopologue where one ^{12}C has been replaced with one ^{13}C.
- NAP – Natural Abundance Probability. The NAP value for an IMF is a floating-point number between 0 and 1 and represents the probability of observing that combination of isotopes at random assuming natural abundance. When the effects of labeling are not considered, these are referred to as absolute NAPs. When labeling is considered, it is a relative NAP as the isotope contribution from labeling is ignored.
- SMIRFE – Small Molecule Isotope Resolved Formula Enumerator. The untargeted assignment tool that assigns FT-MS spectra without a database of expected metabolites that is the focus of Chapter 4.

REFERENCES

Abdelrazig, S. (2015) Mass spectrometry for high-throughput metabolomics analysis of urine, University of Nottingham.

Acharjee, A., Prentice, P., Acerini, C., Smith, J., Hughes, I.A., Ong, K., Griffin, J.L., Dunger, D. and Koulman, A. (2017) The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. *Metabolomics* **13,** 25.

Adeva-Andany, M.M., Carneiro-Freire, N., Seco-Filgueira, M., Fernández-Fernández, C. and Mouriño-Bayolo, D. (2018) Mitochondrial β-oxidation of saturated fatty acids in humans. *Mitochondrion*.

Adeyinka, A. and Kondamudi, N.P. (2018) Cholinergic Crisis.

Alam, M.T., Olin-Sandoval, V., Stincone, A., Keller, M.A., Zelezniak, A., Luisi, B.F. and Ralser, M. (2017) The self-inhibitory nature of metabolic networks and its alleviation through compartmentalization. *Nature Communications* **8,** 16018.

Alonso, A., Marsal, S. and Julià, A. (2015) Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology* **3,** 23.

Anderson, K.A. and Hirschey, M.D. (2012) Mitochondrial protein acetylation regulates metabolism. *Essays in biochemistry* **52,** 23-35.

Arita, M. (2012) From metabolic reactions to networks and pathways, *Bacterial Molecular Networks*, Springer. pp. 93-106.

Arrivault, S., Guenther, M., Ivakov, A., Feil, R., Vosloh, D., Van Dongen, J.T., Sulpice, R. and Stitt, M. (2009) Use of reverse-phase liquid chromatography, linked to tandem mass spectrometry, to profile the Calvin cycle and other metabolic intermediates in Arabidopsis rosettes at different carbon dioxide concentrations. *The Plant Journal* **59,** 826-839.

Assaily, W., Rubinger, D.A., Wheaton, K., Lin, Y., Ma, W., Xuan, W., Brown-Endres, L., Tsuchihara, K., Mak, T.W. and Benchimol, S. (2011) ROS-mediated p53 induction of Lpin1 regulates fatty acid oxidation in response to nutritional stress. *Molecular cell* **44,** 491-501.

Athersuch, T.J. (2012) The role of metabolomics in characterizing the human exposome. *Bioanalysis* **4,** 2207-2212.

Aviram, R., Manella, G., Kopelman, N., Neufeld-Cohen, A., Zwighaft, Z., Elimelech, M., Adamovich, Y., Golik, M., Wang, C., Han, X. and Asher, G. (2016) Lipidomics Analyses Reveal Temporal and Spatial Lipid Organization and Uncover Daily Oscillations in Intracellular Organelles. *Molecular Cell* **62,** 636-648.

Avril, N., Menzel, M., Dose, J., Schelling, M., Weber, W., Jänicke, F., Nathrath, W. and Schwaiger, M. (2001) Glucose metabolism of breast cancer assessed by 18F-FDG PET: histologic and immunohistochemical tissue analysis. *Journal of Nuclear Medicine* **42,** 9-16.

Bao, J., Zhu, L., Zhu, Q., Su, J., Liu, M. and Huang, W. (2016) SREBP-1 is an independent prognostic marker and promotes invasion and migration in breast cancer. *Oncology letters* **12,** 2409-2416.

Barber, C.B., Dobkin, D.P., Dobkin, D.P. and Huhdanpaa, H. (1996) The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)* **22,** 469-483.

Bass, J. and Takahashi, J.S. (2010) Circadian integration of metabolism and energetics. *Science* **330,** 1349-1354.

Bazilio, A. and Weinrich, J. (2012) The Easy Guide to: Inductively Coupled Plasma-Mass Spectrometry.

Beloueche-Babari, M., Peak, J.C., Jackson, L.E., Tiet, M.-Y., Leach, M.O. and Eccles, S.A. (2009) Changes in choline metabolism as potential biomarkers of phospholipase Cγ1 inhibition in human prostate cancer cells. *Molecular Cancer Therapeutics* **8,** 1305-1311.

Bené, H., Lasky, D. and Ntambi, J.M. (2001) Cloning and characterization of the human stearoyl-CoA desaturase gene promoter: transcriptional activation by sterol regulatory element binding protein and repression by polyunsaturated fatty acids and cholesterol. *Biochemical and biophysical research communications* **284,** 1194-1198.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289-300.

Bensaad, K., Tsuruta, A., Selak, M.A., Vidal, M.N.C., Nakano, K., Bartrons, R., Gottlieb, E. and Vousden, K.H. (2006) TIGAR, a p53-inducible regulator of glycolysis and apoptosis. *Cell* **126,** 107-120.

Berdy, J. (2005) Bioactive microbial metabolites. *The Journal of antibiotics* **58,** 1.

Berglund, M. and Wieser, M.E. (2011) Isotopic compositions of the elements 2009 (IUPAC Technical Report). *Pure and applied chemistry* **83,** 397-410.

Bergmann, R. and Bergmann, K. (1991). Fluoride nutrition in infancy--is there a biological role of fluoride for growth? *Nestle nutrition workshop series*.

Berridge, M., Lipp, P. and Bootman, M. (1999) Calcium signalling. *Current biology* **9,** R157-R159.

Betticher, D.C. (2005) Adjuvant and neoadjuvant chemotherapy in NSCLC: a paradigm shift. *Lung Cancer* **50,** S9-S16.

Bist, A., Fielding, C.J. and Fielding, P.E. (2000) p53 regulates caveolin gene transcription, cell cholesterol, and growth by a novel mechanism. *Biochemistry* **39,** 1966-1972.

Blake, G.J. and Ridker, P.M. (2000) Are statins anti-inflammatory? *Current controlled trials in cardiovascular medicine* **1,** 161-165.

Bode, H.B., Bethe, B., Höfs, R. and Zeeck, A. (2002) Big effects from small changes: possible ways to explore nature's chemical diversity. *ChemBioChem* **3,** 619-627.

Bolzoni, M., Chiu, M., Accardi, F., Vescovini, R., Airoldi, I., Storti, P., Todoerti, K., Agnelli, L., Missale, G. and Andreoli, R. (2016) Dependence on glutamine uptake and glutamine addiction characterize myeloma cells: a new attractive target. *Blood***,** blood-2016-01-690743.

Borchers, H.W. and Borchers, M.H.W. (2018) Package ‘pracma’.

Boros, L.G., Serkova, N.J., Cascante, M.S. and Lee, W.-N.P. (2004) Use of metabolic pathway flux information in targeted cancer drug design. *Drug Discovery Today: Therapeutic Strategies* **1,** 435-443.

Bortner, C.D. and Cidlowski, J.A. (2001) Flow cytometric analysis of cell shrinkage and monovalent ions during apoptosis. *Methods in cell biology* **66,** 49-67.

Bourgaud, F., Gravot, A., Milesi, S. and Gontier, E. (2001) Production of plant secondary metabolites: a historical perspective. *Plant science* **161,** 839-851.

Brand, K.A. and Hermfisse, U. (1997) Aerobic glycolysis by proliferating cells: a protective strategy against reactive oxygen species. *Faseb j* **11,** 388-95.

Breiman, L. (2001) Random forests. *Machine learning* **45,** 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and regression trees. Belmont, CA: Wadsworth. *International Group***,** 432.

Brenna, J. and Creasy, W.R. (1989) Experimental evaluation of apodization functions for quantitative Fourier transform mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes* **90,** 151-166.

Brockman, S.A., Roden, E.V. and Hegeman, A.D. (2018) Van Krevelen diagram visualization of high resolution-mass spectrometry metabolomics data with OpenVanKrevelen. *Metabolomics* **14,** 48.

Buhrman, D.L., Price, P.I. and Rudewiczcor, P.J. (1996) Quantitation of SR 27417 in human plasma using electrospray liquid chromatography-tandem mass spectrometry: a study of ion suppression. *Journal of the American Society for Mass Spectrometry* **7,** 1099-1105.

Burkardt, J. (2014) The truncated normal distribution.

Busch, R., Kim, Y.-K., Neese, R.A., Schade-Serin, V., Collins, M., Awada, M., Gardner, J.L., Beysen, C., Marino, M.E. and Misell, L.M. (2006) Measurement of protein turnover rates by heavy water labeling of nonessential amino acids. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1760,** 730-744.

Carreer, W., Flight, R. and Moseley, H. (2013) A Computational Framework for High-Throughput Isotopic Natural Abundance Correction of Omics-Level Ultra-High Resolution FT-MS Datasets. *Metabolites* **3,** 853.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B. and Egertson, J. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* **30,** 918.

Chan, A.W., Mercier, P., Schiller, D., Bailey, R., Robbins, S., Eurich, D.T., Sawyer, M.B. and Broadhurst, D. (2015) 1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. *British Journal Of Cancer* **114,** 59.

Chang, A. (2011) Chemotherapy, chemoresistance and the changing treatment landscape for NSCLC. *Lung cancer* **71,** 3-10.

Chang, A., Parikh, P., Thongprasert, S., Tan, E.H., Perng, R.-P., Ganzon, D., Yang, C.-H., Tsao, C.-J., Watkins, C. and Botwood, N. (2006) Gefitinib (IRESSA) in patients of Asian origin with refractory advanced non-small cell lung cancer: subset analysis from the ISEL study. *Journal of thoracic oncology* **1,** 847-855.

Chekmeneva, E., dos Santos Correia, G.a., Chan, Q., Wijeyesekera, A., Tin, A., Young, J.H., Elliott, P., Nicholson, J.K. and Holmes, E. (2017) Optimization and application of direct infusion nanoelectrospray HRMS method for large-scale urinary metabolic phenotyping in molecular epidemiology. *Journal of proteome research* **16,** 1646-1658.

Chen, J.J. and Yu, B.P. (1994) Alterations in mitochondrial membrane fluidity by lipid peroxidation products. *Free Radical Biology and Medicine* **17,** 411-418.

Chen, M. and Cook, K.D. (2007) Oxidation artifacts in the electrospray mass spectrometry of Aβ peptide. *Analytical chemistry* **79,** 2031-2036.

Chen, T., Cao, Y., Zhang, Y., Liu, J., Bao, Y., Wang, C., Jia, W. and Zhao, A. (2013) Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-Based Complementary and Alternative Medicine* **2013**.

Chen, T., Xie, G., Wang, X., Fan, J., Qiu, Y., Zheng, X., Qi, X., Cao, Y., Su, M. and Wang, X. (2011) Serum and urine metabolite profiling reveals potential

biomarkers of human hepatocellular carcinoma. *Molecular & Cellular Proteomics*, mcp. M110. 004945.

Chen, X. and Reynolds, C.H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of chemical information and computer sciences* **42,** 1407-1414.

Comisarow, M.B. and Marshall, A.G. (1974) Fourier transform ion cyclotron resonance spectroscopy. *Chemical physics letters* **25,** 282-283.

Conti, M. and Eriksson, L. (2016) Physics of pure and non-pure positron emitters for PET: a review and a discussion. *EJNMMI physics* **3,** 8-8.

Costa, B., Dettori, D., Lorenzato, A., Bardella, C., Coltella, N., Martino, C., Cammarata, C., Carmeliet, P., Olivero, M. and Di Renzo, M.F. (2010) Fumarase tumor suppressor gene and MET oncogene cooperate in upholding transformation and tumorigenesis. *The FASEB Journal* **24,** 2680-2688.

Creek, D.J., Chokkathukalam, A., Jankevics, A., Burgess, K.E., Breitling, R. and Barrett, M.P. (2012) Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation. *Analytical chemistry* **84,** 8442-8447.

Crino, L., Kim, D., Riely, G., Janne, P., Blackhall, F., Camidge, D., Hirsh, V., Mok, T., Solomon, B. and Park, K. (2011) Initial phase II results with crizotinib in advanced ALK-positive non-small cell lung cancer (NSCLC): PROFILE 1005. *Journal of Clinical Oncology* **29,** 7514-7514.

Csanadi, A., Kayser, C., Donauer, M., Gumpp, V., Aumann, K., Rawluk, J., Prasse, A., zur Hausen, A., Wiesemann, S. and Werner, M. (2015) Prognostic value of malic enzyme and ATP-citrate lyase in non-small cell lung cancer of the young and the elderly. *PloS one* **10,** e0126357.

Cubbon, S., Antonio, C., Wilson, J. and Thomas-Oates, J. (2010) Metabolomic applications of hilic–lc–ms. *Mass spectrometry reviews* **29,** 671-684.

Cummings, B.S. and Schnellmann, R.G. (2004) Measurement of cell death in mammalian cells. *Current protocols in pharmacology* **25,** 12.8. 1-12.8. 22.

Cuyckens, F. and Claeys, M. (2004) Mass spectrometry in the structural analysis of flavonoids. *Journal of Mass spectrometry* **39,** 1-15.

Daylight Chemical Information Systems, I. (2008) 4. SMARTS - A Language for Describing Molecular Patterns.

De Pablo, M.A. and De Cienfuegos, G.Á. (2000) Modulatory effects of dietary lipids on immune system functions. *Immunology & Cell Biology* **78,** 31-39.

Dettmer, K., Aronov, P.A. and Hammock, B.D. (2007a) MASS SPECTROMETRY-BASED METABOLOMICS. *Mass spectrometry reviews* **26,** 51-78.

Dettmer, K., Aronov, P.A. and Hammock, B.D. (2007b) Mass spectrometry-based metabolomics. *Mass spectrometry reviews* **26,** 51-78.

Diem, M., Polavarapu, P.L., Oboodi, M. and Nafie, L.A. (1982) Vibrational circular dichroism in amino acids and peptides. 4. Vibrational analysis, assignments, and solution-phase Raman spectra of deuterated isotopomers of alanine. *Journal of the American Chemical Society* **104,** 3329-3336.

Dillman III, J.F., Phillips, C.S., Kniffin, D.M., Tompkins, C.P., Hamilton, T.A. and Kan, R.K. (2009) Gene expression profiling of rat hippocampus following exposure to the acetylcholinesterase inhibitor soman. *Chemical research in toxicology* **22,** 633-638.

Diomede, L., Albani, D., Sottocorno, M., Donati, M.B., Bianchi, M., Fruscella, P. and Salmona, M. (2001) In vivo anti-inflammatory effect of statins is mediated by nonsterol mevalonate products. *Arteriosclerosis, thrombosis, and vascular biology* **21,** 1327-1332.

Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., Greiner, R. and Wishart, D.S. (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **8,** 61.

Doerr, A. (2016) Global metabolomics. *Nature Methods* **14,** 32.

Dore, M.P., Davoli, A., Longo, N., Marras, G. and Pes, G.M. (2016) Glucose-6-phosphate dehydrogenase deficiency and risk of colorectal cancer in Northern Sardinia: A retrospective observational study. *Medicine* **95**.

Dulak, J. and Jozkowicz, A. (2005) Anti-angiogenic and anti-inflammatory effects of statins: relevance to anti-cancer therapy. *Curr Cancer Drug Targets* **5,** 579-94.

Eardley, I., Ellis, P., Boolell, M. and Wulff, M. (2002) Onset and duration of action of sildenafil for the treatment of erectile dysfunction. *British journal of clinical pharmacology* **53,** 61S-65S.

Effenberger, M., Bommert, K.S., Kunz, V., Kruk, J., Leich, E., Rudelius, M., Bargou, R. and Bommert, K. (2017) Glutaminase inhibition in multiple myeloma induces apoptosis via MYC degradation. *Oncotarget* **8,** 85858.

Eliuk, S. and Makarov, A. (2015) Evolution of orbitrap mass spectrometry instrumentation. *Annual Review of Analytical Chemistry* **8,** 61-80.

Ericsson, J., Jackson, S.M., Kim, J.B., Spiegelman, B.M. and Edwards, P.A. (1997) Identification of glycerol-3-phosphate acyltransferase as an adipocyte determination and differentiation factor 1- and sterol regulatory element-binding protein-responsive gene. *J Biol Chem* **272,** 7298-305.

Erler, J., Birge, N., Kortelainen, M., Nazarewicz, W., Olsen, E., Perhac, A.M. and Stoitsov, M. (2012) The limits of the nuclear landscape. *Nature* **486,** 509.

Eyles, S.J. and Kaltashov, I.A. (2004) Methods to study protein dynamics and folding by mass spectrometry. *Methods* **34,** 88-99.

Fahy, E., Subramaniam, S., Brown, H.A., Glass, C.K., Merrill Jr., A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M.S., White, S.H., Witztum, J.L. and Dennis, E.A. (2005) A comprehensive classification system for lipids. *European Journal of Lipid Science and Technology* **107,** 337-364.

Fan, J., Hitosugi, T., Chung, T.-W., Xie, J., Ge, Q., Gu, T.-L., Polakiewicz, R.D., Chen, G.Z., Boggon, T.J., Lonial, S., Khuri, F.R., Kang, S. and Chen, J. (2011) Tyrosine Phosphorylation of Lactate Dehydrogenase A Is Important for NADH/NAD⁺ Redox Homeostasis in Cancer Cells. *Molecular and Cellular Biology* **31,** 4938-4950.

Fan, T.W. and Lane, A.N. (2011) NMR-based stable isotope resolved metabolomics in systems biochemistry. *Journal of biomolecular NMR* **49,** 267-280.

Fan, T.W.M., Lorkiewicz, P., Sellers, K., Moseley, H.N.B., Higashi, R.M. and Lane, A.N. (2012) Stable isotope-resolved metabolomics and applications for drug development. *Pharmacology & therapeutics* **133,** 366-391.

Fang, Z.-Z. and Gonzalez, F.J. (2014) LC–MS-based metabolomics: an update. *Archives of toxicology* **88,** 1491-1502.

Farrelly, D., Brown, K.S., Tieman, A., Ren, J., Lira, S.A., Hagan, D., Gregg, R., Mookhtiar, K.A. and Hariharan, N. (1999) Mice mutant for glucokinase regulatory protein exhibit decreased liver glucokinase: a sequestration mechanism in metabolic regulation. *Proceedings of the National Academy of Sciences* **96,** 14511-14516.

Farshidfar, F., Kopciuk, K.A., Hilsden, R., McGregor, S.E., Mazurak, V.C., Buie, W.D., MacLean, A., Vogel, H.J. and Bathe, O.F. (2018) A quantitative multimodal metabolomic assay for colorectal cancer. *BMC Cancer* **18,** 26.

Faubert, B., Li, K.Y., Cai, L., Hensley, C.T., Kim, J., Zacharias, L.G., Yang, C., Do, Q.N., Doucette, S. and Burguete, D. (2017) Lactate metabolism in human lung tumors. *Cell* **171,** 358-371. e9.

Fellgett, P.B. (1949) On the Ultimate Sensitivity and Practical Performance of Radiation Detectors. *Journal of the Optical Society of America* **39,** 970-976.

Fernandez, C.A., Des Rosiers, C., Previs, S.F., David, F. and Brunengraber, H. (1996) Correction of 13C mass isotopomer distributions for natural stable isotope abundance. *Journal of Mass Spectrometry* **31,** 255-262.

Fernie, A.R. and Stitt, M. (2012) On the discordance of metabolomics with proteomics and transcriptomics: coping with increasing complexity in logic, chemistry, and network interactions scientific correspondence. *Plant Physiology* **158,** 1139-1145.

Ferrara, N., Hillan, K.J. and Novotny, W. (2005) Bevacizumab (Avastin), a humanized anti-VEGF monoclonal antibody for cancer therapy. *Biochemical and biophysical research communications* **333,** 328-335.

Ferreres, F., Llorach, R. and Gil-Izquierdo, A. (2004) Characterization of the interglycosidic linkage in di-, tri-, tetra-and pentaglycosylated flavonoids and differentiation of positional isomers by liquid chromatography/electrospray

ionization tandem mass spectrometry. *Journal of Mass Spectrometry* **39,** 312-321.

Fiehn, O. (2002) Metabolomics—the link between genotypes and phenotypes, *Functional genomics*, Springer. pp. 155-171.

Flight, R.M. and Moseley, H.N.B. (2018) Scan-Level Peak Correspondence and Characterization, University of Kentucky.

Foretz, M., Guichard, C., Ferré, P. and Foufelle, F. (1999) Sterol regulatory element binding protein-1c is a major mediator of insulin action on the hepatic expression of glucokinase and lipogenesis-related genes. *Proceedings of the National Academy of Sciences* **96,** 12737-12742.

Fredrick Johansson, e.a. (2018) mpmath: a Python library for arbitrary-precision floating-point arithmetic.

Fu, X., Li, M., Biswas, S., Nantz, M.H. and Higashi, R.M. (2011) A novel microreactor approach for analysis of ketones and aldehydes in breath *Analyst* **136,** 4662-4666.

Fulton, A.B. and Isaacs, W.B. (1991) Titin, a huge, elastic sarcomeric protein with a probable role in morphogenesis. *Bioessays* **13,** 157-161.

Gabitova, L., Gorin, A. and Astsaturov, I. (2013) Molecular pathways: sterols and receptor signaling in cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **20,** 28-34.

Gambhir, S.S. (2002) Molecular imaging of cancer with positron emission tomography. *Nature Reviews Cancer* **2,** 683.

Ganti, S., Taylor, S.L., Kim, K., Hoppel, C.L., Guo, L., Yang, J., Evans, C. and Weiss, R.H. (2012) Urinary acylcarnitines are altered in human kidney cancer. *International journal of cancer* **130,** 2791-2800.

Garg, A. (2011) What is the role of alternative biomarkers for coronary heart disease? *Clinical endocrinology* **75,** 289-293.

Garjani, A., Rezazadeh, H., Andalib, S., Ziaee, M., Doustar, Y., Soraya, H., Garjani, M., Khorrami, A., Asadpoor, M. and Maleki-Dizaji, N. (2012) Ambivalent

effects of atorvastatin on angiogenesis, epidermal cell proliferation and tumorgenesis in animal models. *Iranian biomedical journal* **16,** 59-67.

Gieger, C., Geistlinger, L., Altmaier, E., De Angelis, M.H., Kronenberg, F., Meitinger, T., Mewes, H.-W., Wichmann, H.-E., Weinberger, K.M. and Adamski, J. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics* **4,** e1000282.

Giskeødegård, G.F., Bertilsson, H., Selnæs, K.M., Wright, A.J., Bathen, T.F., Viset, T., Halgunset, J., Angelsen, A., Gribbestad, I.S. and Tessem, M.-B. (2013) Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness. *PloS one* **8,** e62375.

Glass, J.M., Stephen, R.L. and Jacobson, S.C. (1980) The quantity and distribution of radiolabeled dexamethasone delivered to tissue by iontophoresis. *International journal of dermatology* **19,** 519-525.

Glunde, K., Jie, C. and Bhujwalla, Z.M. (2004) Molecular causes of the aberrant choline phospholipid metabolism in breast cancer. *Cancer research* **64,** 4270-4276.

Goetzman, E.S. and Prochownik, E.V. (2018) The Role for Myc in Coordinating Glycolysis, Oxidative Phosphorylation, Glutaminolysis, and Fatty Acid Metabolism in Normal and Neoplastic Tissues. *Frontiers in endocrinology* **9,** 129-129.

Goodacre, R. (2007) Metabolomics of a superorganism. *The Journal of nutrition* **137,** 259S-266S.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in biotechnology* **22,** 245-252.

Gori, S.S., Lorkiewicz, P., Ehringer, D.S., Belshoff, A., Higashi, R.M., Fan, T.W.-M. and Nantz, M.H. (2014) Profiling Thiol Metabolites and Quantification of Cellular Glutathione using FT-ICR-MS Spectrometry, *Analytical and Bioanalytical Chemistry*.

Goyal, S., Yuan, J., Chen, T., Rabinowitz, J.D. and Wingreen, N.S. (2010) Achieving Optimal Growth through Product Feedback Inhibition in Metabolism. *PLOS Computational Biology* **6,** e1000802.

Griffiths, P.R. and Pariente, G.L. (1986) Introduction to spectral deconvolution. *TrAC Trends in Analytical Chemistry* **5,** 209-215.

Groenen, P.J. and van den Heuvel, L.P. (2006) Teaching molecular genetics: Chapter 3–Proteomics in nephrology. *Pediatric Nephrology* **21,** 611-618.

Gronwald, W., Klein, M.S., Kaspar, H., Fagerer, S.R., Nürnberger, N., Dettmer, K., Bertsch, T. and Oefner, P.J. (2008) Urinary metabolite quantification employing 2D NMR spectroscopy. *Analytical chemistry* **80,** 9288-9297.

Guo, K. and Li, L. (2009) Differential 12C-/13C-Isotope Dansylation Labeling and Fast Liquid Chromatography/Mass Spectrometry for Absolute and Relative Quantification of the Metabolome. *Analytical Chemistry* **81,** 3919-3932.

Gustbée, E., Tryggvadottir, H., Markkula, A., Simonsson, M., Nodin, B., Jirström, K., Rose, C., Ingvar, C., Borgquist, S. and Jernström, H. (2015) Tumor-specific expression of HMG-CoA reductase in a population-based cohort of breast cancer patients. *BMC clinical pathology* **15,** 8-8.

Haider, N. (2010) Creating a Web-based, Searchable Molecular Structure Database Using Free Software.

Haider, N. (2016) The checkmol/matchmol Homepage.

Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *cell* **144,** 646-674.

Harris, D.M., Li, L., Chen, M., Lagunero, F.T., Go, V.L.W. and Boros, L.G. (2012) Diverse mechanisms of growth inhibition by luteolin, resveratrol, and quercetin in MIA PaCa-2 cells: a comparative glucose tracer study with the fatty acid synthase inhibitor C75. *Metabolomics* **8,** 201-210.

Harrison, J. (2016) RSelenium: R Bindings for Selenium WebDriver. *R package version* **1**.

Hensley, C.T., Faubert, B., Yuan, Q., Lev-Cohain, N., Jin, E., Kim, J., Jiang, L., Ko, B., Skelton, R. and Loudat, L. (2016) Metabolic heterogeneity in human lung tumors. *Cell* **164,** 681-694.

Hidalgo, M., Amant, F., Biankin, A.V., Budinská, E., Byrne, A.T., Caldas, C., Clarke, R.B., de Jong, S., Jonkers, J. and Mælandsmo, G.M. (2014) Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer discovery* **4,** 998-1013.

Higashi, R.M., Fan, T.W.-M., Lorkiewicz, P.K., Moseley, H.N.B. and Lane, A.N. (2014) Stable Isotope-Labeled Tracers for Metabolic Pathway Elucidation by GC-MS and FT-MS in Raftery, D. (Ed), *Mass Spectrometry in Metabolomics: Methods and Protocols*, Springer New York, New York, NY. pp. 147-167.

Hiller, K., Metallo, C.M., Kelleher, J.K. and Stephanopoulos, G. (2010) Nontargeted elucidation of metabolic pathways using stable-isotope tracers and mass spectrometry. *Anal Chem* **82,** 6621-8.

Hilvo, M., Denkert, C., Lehtinen, L., Muller, B., Brockmoller, S., Seppanen-Laakso, T., Budczies, J., Bucher, E., Yetukuri, L. and Castillo, S. (2011) Novel theranostic opportunities offered by characterization of altered membrane lipid metabolism in breast cancer progression. *Cancer research***,** canres. 3894.2010.

Hore, P.J. (1985) NMR data processing using the maximum entropy method. *Journal of Magnetic Resonance (1969)* **62,** 561-567.

Horvath, S.E. and Daum, G. (2013) Lipids of mitochondria. *Progress in lipid research* **52,** 590-614.

Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M. and Graham Cooks, R. (2005) The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry* **40,** 430-443.

Huan, T., Tang, C., Li, R., Shi, Y., Lin, G. and Li, L. (2015) MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites. *Analytical Chemistry* **87,** 10619-10626.

Huang, J., Fan, X.-X., He, J., Pan, H., Li, R.-Z., Huang, L., Jiang, Z., Yao, X.-J., Liu, L. and Leung, E.L.-H. (2016) SCD1 is associated with tumor promotion, late stage and poor survival in lung adenocarcinoma. *Oncotarget* **7,** 39970.

Huang, S.-C., Stout, D.B., Yee, R.E., Satyamurthy, N. and Barrio, J.R. (1998) Distribution volume of radiolabeled large neutral amino acids in brain tissue. *Journal of Cerebral Blood Flow & Metabolism* **18,** 1288-1293.

Huber, V., Camisaschi, C., Berzi, A., Ferro, S., Lugini, L., Triulzi, T., Tuccitto, A., Tagliabue, E., Castelli, C. and Rivoltini, L. (2017) Cancer acidity: An ultimate frontier of tumor immune escape and a novel target of immunomodulation. *Seminars in Cancer Biology* **43,** 74-89.

Hung, M.-S., Chen, I.C., Lee, C.-P., Huang, R.-J., Chen, P.-C., Tsai, Y.-H. and Yang, Y.-H. (2017) Statin improves survival in patients with EGFR-TKI lung cancer: A nationwide population-based study. *PloS one* **12,** e0171137-e0171137.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of computational and graphical statistics* **5,** 299-314.

Isin, E.M., Elmore, C.S., Nilsson, G.r.N., Thompson, R.A. and Weidolf, L. (2012) Use of radiolabeled compounds in drug metabolism and pharmacokinetic studies. *Chemical research in toxicology* **25,** 532-542.

Istvan, E.S. and Deisenhofer, J. (2001) Structural mechanism for statin inhibition of HMG-CoA reductase. *Science* **292,** 1160-1164.

J R Neely, a. and Morgan, H.E. (1974) Relationship Between Carbohydrate and Lipid Metabolism and the Energy Balance of Heart Muscle. *Annual Review of Physiology* **36,** 413-459.

James, L. (2013) Metabolomics: integration of a new “omics” with clinical pharmacology. *Clinical Pharmacology & Therapeutics* **94,** 547-551.

Jamnagerwalla, J., Howard, L.E., Allott, E.H., Vidal, A.C., Moreira, D.M., Castro-Santamaria, R., Andriole, G.L., Freeman, M.R. and Freedland, S.J. (2018) Serum cholesterol and risk of high-grade prostate cancer: results from the REDUCE study. *Prostate cancer and prostatic diseases* **21,** 252.

Jang, C., Chen, L. and Rabinowitz, J.D. (2018) Metabolomics and isotope tracing. *Cell* **173,** 822-837.

Janitza, S. and Hornung, R. (2018) On the overestimation of random forest’s out-of-bag error. *PloS one* **13,** e0201904.

Janvilisri, T., Venter, H., Shahi, S., Reuter, G., Balakrishnan, L. and van Veen, H.W. (2003) Sterol transport by the human breast cancer resistance protein

(ABCG2) expressed in Lactococcus lactis. *Journal of Biological Chemistry* **278,** 20645-20651.

Jessome, L.L. and Volmer, D.A. (2006) Ion suppression: a major concern in mass spectrometry. *Lc Gc North America* **24,** 498.

Jiang, D., LaGory, E.L., Brož, D.K., Bieging, K.T., Brady, C.A., Link, N., Abrams, J.M., Giaccia, A.J. and Attardi, L.D. (2015) Analysis of p53 transactivation domain mutants reveals Acad11 as a metabolic target important for p53 pro-survival function. *Cell reports* **10,** 1096-1109.

Jiang, P., Du, W., Wang, X., Mancuso, A., Gao, X., Wu, M. and Yang, X. (2011) p53 regulates biosynthesis through direct inactivation of glucose-6-phosphate dehydrogenase. *Nature cell biology* **13,** 310.

Jiang, S., Bo, L., Gong, C., Du, X., Kan, H., Xie, Y., Song, W. and Zhao, J. (2016) Traffic-related air pollution is associated with cardio-metabolic biomarkers in general residents. *International archives of occupational and environmental health* **89,** 911-921.

Jin, J.Y., Almon, R.R., DuBois, D.C. and Jusko, W.J. (2003) Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *Journal of Pharmacology and Experimental Therapeutics* **307,** 93-109.

Jin, L., Alesi, G. and Kang, S. (2016) Glutaminolysis as a target for cancer therapy. *Oncogene* **35,** 3619.

Jonas, S., Benedetto, C., Flatman, A., Hammond, R., Micheletti, L., Riley, C., Riley, P., Spargo, D., Zonca, M. and Slater, T. (1992) Increased activity of 6-phosphogluconate dehydrogenase and glucose-6-phosphate dehydrogenase in purified cell suspensions and single cells from the uterine cervix in cervical intraepithelial neoplasia. *British journal of cancer* **66,** 185.

Jue, T., Rothman, D.L., Shulman, G.I., Tavitian, B.A., DeFronzo, R.A. and Shulman, R.G. (1989) Direct observation of glycogen synthesis in human muscle with 13C NMR. *Proceedings of the National Academy of Sciences* **86,** 4489-4491.

Jugdaohsingh, R. (2007) Silicon and bone health. *The journal of nutrition, health & aging* **11,** 99.

Kaira, K., Serizawa, M., Koh, Y., Takahashi, T., Yamaguchi, A., Hanaoka, H., Oriuchi, N., Endo, M., Ohde, Y. and Nakajima, T. (2014) Biological significance of 18F-FDG uptake on PET in patients with non-small-cell lung cancer. *Lung Cancer* **83,** 197-204.

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

Kanawati, B., Bader, T.M., Wanczek, K.P., Li, Y. and Schmitt-Kopplin, P. (2017) Fourier transform (FT)-artifacts and power-function resolution filter in Fourier transform mass spectrometry. *Rapid Communications in Mass Spectrometry* **31,** 1607-1615.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27-30.

Kanitkar, A.A., Schwartz, A.G., George, J. and Soubani, A.O. (2018) Causes of death in long-term survivors of non-small cell lung cancer: A regional Surveillance, Epidemiology, and End Results study. *Annals of thoracic medicine* **13,** 76.

Kell, D.B. and Goodacre, R. (2014) Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. *Drug Discovery Today* **19,** 171-182.

Keller, B.O., Sui, J., Young, A.B. and Whittal, R.M. (2008) Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta* **627,** 71-81.

Kim, D.-W., Mehra, R., Tan, D.S.-W., Felip, E., Chow, L.Q.M., Camidge, D.R., Vansteenkiste, J.F., Sharma, S., De Pas, T. and Riely, G.J. (2014) Ceritinib in advanced anaplastic lymphoma kinase (ALK)-rearranged (ALK+) non-small cell lung cancer (NSCLC): Results of the ASCEND-1 trial, American Society of Clinical Oncology.

Kim, J.H., Nam, B., Choi, Y.J., Kim, S.Y., Lee, J.-E., Sung, K.J., Kim, W.S., Choi, C.-M., Chang, E.-J. and Koh, J.S. (2018) Enhanced glycolysis supports cell survival in EGFR-mutant lung adenocarcinoma by inhibiting autophagy-mediated EGFR degradation. *Cancer research* **78,** 4482-4496.

Kim, K., Taylor, S.L., Ganti, S., Guo, L., Osier, M.V. and Weiss, R.H. (2011) Urine metabolomic analysis identifies potential biomarkers and pathogenic pathways in kidney cancer. *Omics: a journal of integrative biology* **15,** 293-303.

Kimmelman, A.C. (2015) Metabolic dependencies in RAS-driven cancers, AACR.

Kind, T. and Fiehn, O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics* **7,** 234.

Kind, T. and Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8,** 105.

Kind, T., Tolstikov, V., Fiehn, O. and Weiss, R.H. (2007) A comprehensive urinary metabolomic approach for identifying kidney cancer. *Analytical biochemistry* **363,** 185-195.

Kingston, E.E., Shannon, J.S. and Lacey, M.J. (1983) Rearrangements in chemical ionization mass spectra. *Organic Mass Spectrometry* **18,** 183-192.

Kitahara, C.M., de González, A.B., Freedman, N.D., Huxley, R., Mok, Y., Jee, S.H. and Samet, J.M. (2011) Total cholesterol and cancer risk in a large prospective study in Korea. *Journal of Clinical Oncology* **29,** 1592.

Köfeler, H.C., Fauland, A., Rechberger, G.N. and Trötzmüller, M. (2012) Mass spectrometry based lipidomics: an overview of technological platforms. *Metabolites* **2,** 19-38.

Kotera, M., McDonald, A.G., Boyce, S. and Tipton, K.F. (2008) Functional Group and Substructure Searching as a Tool in Metabolomics. *PLoS ONE* **3,** e1537.

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P. (2007) Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* **160,** 3-24.

Kruger, N.J. and von Schaewen, A. (2003) The oxidative pentose phosphate pathway: structure and organisation. *Current opinion in plant biology* **6,** 236-246.

Kruszynska, Y.T., Mulford, M.I., Baloga, J., Yu, J.G. and Olefsky, J.M. (1998) Regulation of skeletal muscle hexokinase II by insulin in nondiabetic and NIDDM subjects. *Diabetes* **47,** 1107-13.

Lacey, J.M. and Wilmore, D.W. (1990) Is glutamine a conditionally essential amino acid? *Nutrition reviews* **48,** 297-309.

Lane, A.N., Fan, T.W.-M., Xie, Z., Moseley, H.N. and Higashi, R.M. (2009) Isotopomer analysis of lipid biosynthesis by high resolution mass spectrometry and NMR. *Analytica chimica acta* **651,** 201-208.

Lane, A.N., Yan, J. and Fan, T.W. (2015) 13C tracer studies of metabolism in mouse tumor xenografts. *Bio-protocol* **5**.

Le, A., Lane, A.N., Hamaker, M., Bose, S., Gouw, A., Barbi, J., Tsukamoto, T., Rojas, C.J., Slusher, B.S. and Zhang, H. (2012) Glucose-independent glutamine metabolism via TCA cycling for proliferation and survival in B cells. *Cell metabolism* **15,** 110-121.

Lee, S.H., Koo, K.H., Park, J.W., Kim, H.J., Ye, S.K., Park, J.B., Park, B.K. and Kim, Y.N. (2009) HIF-1 is induced via EGFR activation and mediates resistance to anoikis-like cell death under lipid rafts/caveolae-disrupting stress. *Carcinogenesis* **30,** 1997-2004.

Leighty, R.W. and Antoniewicz, M.R. (2011) Dynamic metabolic flux analysis (DMFA): a framework for determining fluxes at metabolic non-steady state. *Metabolic engineering* **13,** 745-755.

Lewis, C.A., Parker, S.J., Fiske, B.P., McCloskey, D., Gui, D.Y., Green, C.R., Vokes, N.I., Feist, A.M., Vander Heiden, M.G. and Metallo, C.M. (2014) Tracing compartmentalized NADPH metabolism in the cytosol and mitochondria of mammalian cells. *Molecular cell* **55,** 253-263.

Lewis, C.A. and Wolfenden, R. (2008) Uroporphyrinogen decarboxylation as a benchmark for the catalytic proficiency of enzymes. *Proceedings of the National Academy of Sciences* **105,** 17328-17333.

Li, J., Gu, D., Lee, S.S.Y., Song, B., Bandyopadhyay, S., Chen, S., Konieczny, S.F., Ratliff, T.L., Liu, X., Xie, J. and Cheng, J.X. (2016a) Abrogating cholesterol esterification suppresses growth and metastasis of pancreatic cancer. *Oncogene* **35,** 6378.

Li, X., Wu, J.B., Li, Q., Shigemura, K., Chung, L.W. and Huang, W.-C. (2016b) SREBP-2 promotes stem cell-like properties and metastasis by transcriptional activation of c-Myc in prostate cancer. *Oncotarget* **7,** 12869.

Li, X., Zhao, H., Zhou, X. and Song, L. (2015) Inhibition of lactate dehydrogenase A by microRNA-34a resensitizes colon cancer cells to 5-fluorouracil. *Molecular medicine reports* **11,** 577-582.

Li, Y., Wang, C. and Chen, L. (2019) The SDAMS package.

Lieberman, B.P., Ploessl, K., Wang, L., Qu, W., Zha, Z., Wise, D.R., Chodosh, L.A., Belka, G., Thompson, C.B. and Kung, H.F. (2011) PET imaging of glutaminolysis in tumors by 18F-(2S,4R)4-fluoroglutamine. *J Nucl Med* **52,** 1947-55.

Lin, J.J., Ezer, N., Sigel, K., Mhango, G. and Wisnivesky, J.P. (2016) The effect of statins on survival in patients with stage IV lung cancer. *Lung Cancer* **99,** 137-142.

Link, H., Kochanowski, K. and Sauer, U. (2013) Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nature biotechnology* **31,** 357.

Linstrom, P.J. and Mallard, W.G. (2001) The NIST Chemistry WebBook: A chemical data resource on the internet. *Journal of Chemical & Engineering Data* **46,** 1059-1063.

Lister-James, J., Moyer, B. and Dean, T. (1996) Small peptides radiolabeled with 99mTc. *The quarterly journal of nuclear medicine: official publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR)* **40,** 221-233.

Liu, Y. (2006) Fatty acid oxidation is a dominant bioenergetic pathway in prostate cancer. *Prostate cancer and prostatic diseases* **9,** 230.

Livingstone, L.R., White, A., Sprouse, J., Livanos, E., Jacks, T. and Tlsty, T.D. (1992) Altered cell cycle arrest and gene amplification potential accompany loss of wild-type p53. *Cell* **70,** 923-935.

Lorkiewicz, P., Higashi, R.M., Lane, A.N. and Fan, T.W.-M. (2012) High information throughput analysis of nucleotides and their isotopically enriched isotopologues by direct-infusion FTICR-MS. *Metabolomics* **8,** 930-939.

Lu, W., Pelicano, H. and Huang, P. (2010) Cancer metabolism: is glutamine sweeter than glucose? *Cancer cell* **18,** 199-200.

Lukaski, H., Siders, W., Hoverson, B. and Gallagher, S. (1996) Iron, copper, magnesium and zinc status as predictors of swimming performance. *International journal of sports medicine* **17,** 535-540.

Luo, D., Xiao, H., Dong, J., Li, Y., Feng, G., Cui, M. and Fan, S. (2017) B7-H3 regulates lipid metabolism of lung cancer through SREBP1-mediated expression of FASN. *Biochemical and biophysical research communications* **482,** 1246-1251.

Lydic, T.A. and Goo, Y.-H. (2018) Lipidomics unveils the complexity of the lipidome in metabolic diseases. *Clinical and Translational Medicine* **7,** 4.

Maher, E.A., Marin-Valencia, I., Bachoo, R.M., Mashimo, T., Raisanen, J., Hatanpaa, K.J., Jindal, A., Jeffrey, F.M., Choi, C. and Madden, C. (2012) Metabolism of [U-13C] glucose in human brain tumors in vivo. *NMR in biomedicine* **25,** 1234-1244.

Mahieu, N.G. and Patti, G.J. (2017) Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites. *Analytical chemistry* **89,** 10397-10406.

March, R.E. (2006) Quadrupole ion trap mass spectrometer. *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*.

March, R.E. and Londry, F.A. (1995) Theory of quadrupole mass spectrometry. *Practical aspects of ion trap mass spectrometry* **1,** 25-48.

Marin-Valencia, I., Yang, C., Mashimo, T., Cho, S., Baek, H., Yang, X.L., Rajagopalan, K.N., Maddie, M., Vemireddy, V., Zhao, Z., Cai, L., Good, L., Tu, B.P., Hatanpaa, K.J., Mickey, B.E., Mates, J.M., Pascual, J.M., Maher, E.A., Malloy, C.R., Deberardinis, R.J. and Bachoo, R.M. (2012) Analysis of tumor metabolism reveals mitochondrial glucose oxidation in genetically diverse human glioblastomas in the mouse brain in vivo. *Cell Metab* **15,** 827-37.

Mathur, R. and O'Connor, P.B. (2009) Artifacts in Fourier transform mass spectrometry. *Rapid communications in mass spectrometry : RCM* **23,** 523-529.

Mattauch, J. (1936) A double-focusing mass spectrograph and the masses of N 15 and O 18. *Physical Review* **50,** 617.

Mattingly, S.J., Xu, T., Nantz, M.H., Higashi, R.M. and Fan, T.W.M. (2012) A carbonyl capture approach for profiling oxidized metabolites in cell extracts. *Metabolomics* **8,** 989-996.

McDonnell, S.R., Hwang, S.R., Rolland, D., Murga-Zamalloa, C., Basrur, V., Conlon, K.P., Fermin, D., Wolfe, T., Raskind, A. and Ruan, C. (2013) Integrated phosphoproteomic and metabolomic profiling reveals NPM-ALK-mediated phosphorylation of PKM2 and metabolic reprogramming in anaplastic large cell lymphoma. *Blood***,** blood-2013-01-482026.

McKeehan, W.L. (1982) Glycolysis, glutaminolysis and cell proliferation. *Cell Biol Int Rep* **6,** 635-50.

McLafferty, F.W. (1959) Mass spectrometric analysis. Molecular rearrangements. *Analytical Chemistry* **31,** 82-87.

McLaren, J.E., Michael, D.R., Salter, R.C., Ashlin, T.G., Calder, C.J., Miller, A.M., Liew, F.Y. and Ramji, D.P. (2010) IL-33 reduces macrophage foam cell formation. *The Journal of Immunology* **185,** 1222-1229.

Menicatti, M., Guandalini, L., Dei, S., Floriddia, E., Teodori, E., Traldi, P. and Bartolucci, G. (2016) The power of energy-resolved tandem mass spectrometry experiments for resolution of isomers: the case of drug plasma stability investigation of multidrug resistance inhibitors. *Rapid Communications in Mass Spectrometry* **30,** 423-432.

Menküc, B.S., Gille, C. and HOLZHÜTTER, H.-G. (2008) Computer aided optimization of carbon atom labeling for tracer experiments. *Genome Informatics* **20,** 270-276.

Mertz, W. (2012) *Trace elements in human and animal nutrition*. Elsevier.

Metallo, Christian M. and Vander Heiden, Matthew G. (2013) Understanding Metabolic Regulation and Its Influence on Cell Physiology. *Molecular Cell* **49,** 388-398.

Miladinović, S.M., Kozhinov, A.N., Tsybin, O.Y. and Tsybin, Y.O. (2012) Sidebands in Fourier transform ion cyclotron resonance mass spectra. *International Journal of Mass Spectrometry* **325-327,** 10-18.

Milner, J.J., Wang, J., Sheridan, P.A., Ebbels, T., Beck, M.A. and Saric, J. (2014) 1H NMR-based profiling reveals differential immune-metabolic networks during influenza virus infection in obese mice. *PloS one* **9,** e97238.

Mitchell, J., Flight, R. and Moseley, H. (2019) *Small Molecule Isotope Resolved Formula Enumerator: a Tool for Assigning Isotopologues and Metabolites in Fourier Transform Mass Spectra*.

Mitchell, J.M., Fan, T.W.M., Lane, A.N. and Moseley, H.N.B. (2014) Development and in silico evaluation of large-scale metabolite identification methods using functional group detection for metabolomics. *Frontiers in Genetics* **5,** 237.

Mitchell, J.M., Flight, R.M., Wang, Q., Kang, W.-Y., Higashi, R.M., Fan, T.W., Lane, A.N. and Moseley, H.N. (2017) High Peak Density Artifacts in Fourier Transform Mass Spectra and their Effects on Data Analysis. *bioRxiv***,** 191205.

Mittler, R. (2017) ROS are good. *Trends in plant science* **22,** 11-19.

Moco, S., Vervoort, J., Bino, R.J., De Vos, R.C. and Bino, R. (2007) Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry* **26,** 855-866.

Molina, J.R., Yang, P., Cassivi, S.D., Schild, S.E. and Adjei, A.A. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clinic Proceedings*, pp. 584-594.

Montgomery, J.E. and Brown, J.R. (2013) Metabolic biomarkers for predicting cardiovascular disease. *Vascular health and risk management* **9,** 37.

Montigon, F., Boza, J. and Fay, L. (2001) Determination of 13C-and 15N-enrichment of glutamine by gas chromatography/mass spectrometry and gas chromatography/combustion/isotope ratio mass spectrometry after N (O, S)-ethoxycarbonyl ethyl ester derivatisation. *Rapid Communications in Mass Spectrometry* **15,** 116-123.

Mor, I., Cheung, E. and Vousden, K. (2011). Control of glycolysis through regulation of PFK1: old friends and recent additions. *Cold Spring Harbor symposia on quantitative biology*, pp. a010868.

Morrison, R.F. and Farmer, S.R. (2000) Hormonal Signaling and Transcriptional Control of Adipocyte Differentiation. *The Journal of Nutrition* **130,** 3116S-3121S.

Moseley, H.N. (2010) Correcting for the effects of natural abundance in stable isotope resolved metabolomics experiments involving ultra-high resolution mass spectrometry. *BMC bioinformatics* **11,** 139.

Moseley, H.N.B. (2013) ERROR ANALYSIS AND PROPAGATION IN METABOLOMICS DATA ANALYSIS. *Computational and Structural Biotechnology Journal* **4,** e201301006.

Mullen, A.R., Hu, Z., Shi, X., Jiang, L., Boroughs, L.K., Kovacs, Z., Boriack, R., Rakheja, D., Sullivan, L.B. and Linehan, W.M. (2014) Oxidation of alpha-ketoglutarate is required for reductive carboxylation in cancer cells with mitochondrial defects. *Cell reports* **7,** 1679-1690.

Nemoto, S., Fergusson, M.M. and Finkel, T. (2004) Nutrient availability regulates SIRT1 through a forkhead-dependent pathway. *Science* **306,** 2105-2108.

Nesvizhskii, A.I., Vitek, O. and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* **4,** 787.

Niedenführ, S., Wiechert, W. and Nöh, K. (2015) How to measure metabolic fluxes: a taxonomic guide for 13C fluxomics. *Current opinion in biotechnology* **34,** 82-90.

Nielsen, S.F., Nordestgaard, B.G. and Bojesen, S.E. (2012) Statin use and reduced cancer-related mortality. *New England Journal of Medicine* **367,** 1792-1802.

Nishikaze, T., Tsumoto, H., Sekiya, S., Iwamoto, S., Miura, Y. and Tanaka, K. (2017) Differentiation of sialyl linkage isomers by one-pot sialic acid derivatization for mass spectrometry-based glycan profiling. *Analytical chemistry* **89,** 2353-2360.

Nishiumi, S., Kobayashi, T., Ikeda, A., Yoshie, T., Kibi, M., Izumi, Y., Okuno, T., Hayashi, N., Kawano, S. and Takenawa, T. (2012) A novel serum metabolomics-based diagnostic approach for colorectal cancer. *PloS one* **7,** e40459.

Nóbrega-Pereira, S., Fernandez-Marcos, P.J., Brioche, T., Gomez-Cabrera, M.C., Salvador-Pascual, A., Flores, J.M., Viña, J. and Serrano, M. (2016) G6PD protects from oxidative damage and improves healthspan in mice. *Nature communications* **7,** 10894-10894.

Noto, A., Raffa, S., De Vitis, C., Roscilli, G., Malpicci, D., Coluccia, P., Di Napoli, A., Ricci, A., Giovagnoli, M. and Aurisicchio, L. (2013) Stearoyl-CoA desaturase-1 is a key factor for lung cancer-initiating cells. *Cell death & disease* **4,** e947.

O'Brien, T. and Lis, J.T. (1993) Rapid changes in Drosophila transcription after an instantaneous heat shock. *Mol Cell Biol* **13,** 3456-63.

Osthus, R.C., Shim, H., Kim, S., Li, Q., Reddy, R., Mukherjee, M., Xu, Y., Wonsey, D., Lee, L.A. and Dang, C.V. (2000) Deregulation of glucose transporter 1 and glycolytic gene expression by c-Myc. *Journal of Biological Chemistry* **275,** 21797-21800.

Osugi, J., Yamaura, T., Muto, S., Okabe, N., Matsumura, Y., Hoshino, M., Higuchi, M., Suzuki, H. and Gotoh, M. (2015) Prognostic impact of the combination of glucose transporter 1 and ATP citrate lyase in node-negative patients with non-small lung cancer. *Lung cancer* **88,** 310-318.

Ouiddir, A., Planes, C., Fernandes, I., VanHesse, A. and Clerici, C. (1999) Hypoxia upregulates activity and expression of the glucose transporter GLUT1 in alveolar epithelial cells. *Am J Respir Cell Mol Biol* **21,** 710-8.

Owens, M. (2006) *The definitive guide to SQLite*. Apress.

Paez, J.G., Jänne, P.A., Lee, J.C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F.J., Lindeman, N. and Boggon, T.J. (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304,** 1497-1500.

Paz-Yaacov, N., Bazak, L., Buchumenski, I., Porath, Hagit T., Danan-Gotthold, M., Knisbacher, Binyamin A., Eisenberg, E. and Levanon, Erez Y. (2015) Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. *Cell Reports* **13,** 267-276.

Peake, D.A., Yokoi, Y., Wang, J. and Yingying, H. (2013) A New Lipid Software Workflow for Processing Orbitrap-based Global Lipidomics Data in Translational and Systems Biology Research in Scientific, T.F. (Ed).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12,** 2825-2830.

Penuelas, J. and Sardans, J. (2009) Ecological metabolomics. *Chemistry and Ecology* **25,** 305-309.

Philip Remes, J., Senko, M., Huguet, R., Zabrouskov, V., Soltero, N. and Eliuk, S. (2016) Improved Control of Ion Population for Orbitrap Mass Analysis. *Thermo Fisher Scientific*.

Pinheiro, C., Reis, R.M., Ricardo, S., Longatto-Filho, A., Schmitt, F. and Baltazar, F. (2010) Expression of monocarboxylate transporters 1, 2, and 4 in human tumours and their association with CD147 and CD44. *BioMed Research International* **2010**.

Pinheiro, C., Sousa, B., Albergaria, A., Paredes, J., Dufloth, R., Vieira, D., Schmitt, F.C. and Baltazar, F. (2011) GLUT1 and CAIX expression profiles in breast cancer correlate with adverse prognostic factors and MCT1 overexpression. *Histology and histopathology* **26,** 1279-1286.

Pinna, G., Brodel, O., Visser, T., Jeitner, A., Grau, H., Eravci, M., Meinhold, H. and Baumgartner, A. (2002) Concentrations of seven iodothyronine metabolites in brain regions and the liver of the adult rat. *Endocrinology* **143,** 1789-800.

Piperdi, B., Merla, A. and Perez-Soler, R. (2014) Targeting angiogenesis in squamous non-small cell lung cancer. *Drugs* **74,** 403-413.

Pirker, R., Krajnik, G., Zöchbauer, S., Malayeri, R., Kneussl, M. and Huber, H. (1995) Paclitaxel/cisplatin in advanced non-small-cell lung cancer (NSCLC). *Annals of oncology* **6,** 833-835.

Prior, I.A., Lewis, P.D. and Mattos, C. (2012) A comprehensive survey of Ras mutations in cancer. *Cancer research* **72,** 2457-2467.

Radzicka, A. and Wolfenden, R. (1995) A proficient enzyme. *Science* **267,** 90-93.

Raïs, B., Comin, B., Puigjaner, J., Brandes, J.L., Creppy, E., Saboureau, D., Ennamany, R., Paul Lee, W.-N., Boros, L.G. and Cascante, M. (1999) Oxythiamine and dehydroepiandrosterone induce a G1 phase cycle arrest in Ehrlich's tumor cells through inhibition of the pentose cycle. *FEBS letters* **456,** 113-118.

Ralser, M., Wamelink, M.M., Latkolik, S., Jansen, E.E., Lehrach, H. and Jakobs, C. (2009) Metabolic reconfiguration precedes transcriptional regulation in the antioxidant response. *Nature biotechnology* **27,** 604.

Rauha, J.P., Vuorela, H. and Kostiainen, R. (2001) Effect of eluent on the ionization efficiency of flavonoids by ion spray, atmospheric pressure chemical ionization, and atmospheric pressure photoionization mass spectrometry. *Journal of mass spectrometry* **36,** 1269-1280.

Ray, U. and Roy, S.S. (2018) Aberrant lipid metabolism in cancer cells – the role of oncolipid-activated signaling. *The FEBS Journal* **285,** 432-443.

Raymond, J.W. and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of computer-aided molecular design* **16,** 521-533.

Ren, J.-G., Seth, P., Clish, C.B., Lorkiewicz, P.K., Higashi, R.M., Lane, A.N., Fan, T.W.M. and Sukhatme, V.P. (2014) Knockdown of malic enzyme 2 suppresses lung tumor growth, induces differentiation and impacts PI3K/AKT signaling. *Scientific reports* **4**.

Rickes, E.L., Brink, N.G., Koniuszy, P., Wood, T.R. and Folkers, K. (1948) Vitamin B12, a cobalt complex. *Science (Washington)* **108**.

Riedl, A., Schlederer, M., Pudelko, K., Stadler, M., Walter, S., Unterleuthner, D., Unger, C., Kramer, N., Hengstschläger, M., Kenner, L., Pfeiffer, D., Krupitza, G. and Dolznig, H. (2017) Comparison of cancer cells in 2D vs 3D culture reveals differences in AKT–mTOR–S6K signaling and drug responses. *Journal of Cell Science* **130,** 203-218.

Riemann, A., Schneider, B., Gundel, D., Stock, C., Gekle, M. and Thews, O. (2016) Acidosis Promotes Metastasis Formation by Enhancing Tumor Cell Motility. *Adv Exp Med Biol* **876,** 215-220.

Roessner, U. and Beckles, D.M. (2009) Metabolite measurements, *Plant metabolic networks*, Springer. pp. 39-69.

Rong, Y., Wu, W., Ni, X., Kuang, T., Jin, D., Wang, D. and Lou, W. (2013) Lactate dehydrogenase A is overexpressed in pancreatic cancer and promotes the growth of pancreatic cancer cells. *Tumor Biology* **34,** 1523-1530.

Rush, G.F., Gorski, J.R., Ripple, M.G., Sowinski, J., Bugelski, P. and Hewitt, W.R. (1985) Organic hydroperoxide-induced lipid peroxidation and cell death in isolated hepatocytes. *Toxicology and applied pharmacology* **78,** 473-483.

Rütti, M.F., Richard, S., Penno, A., von Eckardstein, A. and Hornemann, T. (2009) An improved method to determine serine palmitoyltransferase activity. *Journal of lipid research* **50,** 1237-1244.

Ryu, K.W., Nandu, T., Kim, J., Challa, S., DeBerardinis, R.J. and Kraus, W.L. (2018) Metabolic regulation of transcription through compartmentalized NAD(+) biosynthesis. *Science* **360**.

Sadeghi, M.M., Tiglio, A., Sadigh, K., O'Donnell, L., Collinge, M., Pardi, R. and Bender, J.R. (2001) INHIBITION OF INTERFERON-γ–MEDIATED MICROVASCULAR ENDOTHELIAL CELL MAJOR HISTOCOMPATIBILITY COMPLEX CLASS II GENE ACTIVATION BY HMG-COA REDUCTASE INHIBITORS1. *Transplantation* **71,** 1262-1268.

Safari, R. and Meuwissen, R. (2015) Practical use of advanced mouse models for lung cancer, *Mouse Models of Cancer*, Springer. pp. 93-124.

Saltiel, A.R. and Kahn, C.R. (2001) Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414,** 799.

Sandler, A.B., Nemunaitis, J., Denham, C., Von Pawel, J., Cormier, Y., Gatzemeier, U., Mattson, K., Manegold, C., Palmer, M. and Gregor, A. (2000) Phase III trial of gemcitabine plus cisplatin versus cisplatin alone in patients with locally advanced or metastatic non–small-cell lung cancer. *Journal of Clinical Oncology* **18,** 122-122.

Schomburg, L. and Köhrle, J. (2008) On the importance of selenium and iodine metabolism for thyroid hormone biosynthesis and human health. *Molecular nutrition & food research* **52,** 1235-1246.

Schöning, U. (1987). Graph isomorphism is in the low hierarchy. *Annual Symposium on Theoretical Aspects of Computer Science*, pp. 114-124.

Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D. and McLean, J.A. (2016) Untargeted metabolomics strategies – Challenges and Emerging Directions. *Journal of the American Society for Mass Spectrometry* **27,** 1897-1905.

Schulze, N.D., Hamelin, E.I., Winkeljohn, W.R., Shaner, R.L., Basden, B.J., deCastro, B.R., Pantazides, B.G., Thomas, J.D. and Johnson, R.C. (2016) Evaluation of Multiple Blood Matrices for Assessment of Human Exposure to Nerve Agents. *Journal of analytical toxicology* **40,** 229-235.

Schwartzenberg-Bar-Yoseph, F., Armoni, M. and Karnieli, E. (2004) The tumor suppressor p53 down-regulates glucose transporters GLUT1 and GLUT4 gene expression. *Cancer research* **64,** 2627-2633.

Seckl, M.J., Ottensmeier, C.H., Cullen, M., Schmid, P., Ngai, Y., Muthukumar, D., Thompson, J., Harden, S., Middleton, G., Fife, K.M., Crosse, B., Taylor, P., Nash, S. and Hackshaw, A. (2017) Multicenter, Phase III, Randomized, Double-Blind, Placebo-Controlled Trial of Pravastatin Added to First-Line Standard Chemotherapy in Small-Cell Lung Cancer (LUNGSTAR). *J Clin Oncol* **35,** 1506-1514.

Sellers, K., Fox, M.P., Bousamra, M., Slone, S.P., Higashi, R.M., Miller, D.M., Wang, Y., Yan, J., Yuneva, M.O. and Deshpande, R. (2015a) Pyruvate carboxylase is critical for non–small-cell lung cancer proliferation. *The Journal of clinical investigation* **125,** 687-698.

Sellers, K., Fox, M.P., Michael Bousamra, II, Slone, S.P., Higashi, R.M., Miller, D.M., Wang, Y., Yan, J., Yuneva, M.O. and Deshpande, R. (2015b) Pyruvate carboxylase is critical for non–small-cell lung cancer proliferation. *The Journal of clinical investigation* **125,** 687.

Semenza, G.L. (2008) Tumor metabolism: cancer cells give and take lactate. *The Journal of clinical investigation* **118,** 3835-3837.

Seyfried, T.N., Flores, R.E., Poff, A.M. and D'agostino, D.P. (2013) Cancer as a metabolic disease: implications for novel therapeutics. *Carcinogenesis* **35,** 515-527.

Sgambato, A., Casaluce, F., C Sacco, P., Palazzolo, G., Maione, P., Rossi, A., Ciardiello, F. and Gridelli, C. (2016) Anti PD-1 and PDL-1 immunotherapy in the treatment of advanced non-small cell lung cancer (NSCLC): a review on toxicity profile and its management. *Current drug safety* **11,** 62-68.

Sheng, H. and Tang, W. (2016) Glycolysis inhibitors for anticancer therapy: a review of recent patents. *Recent patents on anti-cancer drug discovery* **11,** 297-308.

Shepherd, F., Pereira, J., Ciuleanu, T., Tan, E., Hirsh, V., Thongprasert, S., Bezjak, A., Tu, D., Santabarbara, P. and Seymour, L. (2004) A randomized placebo-controlled trial of erlotinib in patients with advanced non-small cell lung cancer (NSCLC) following failure of 1st line or 2nd line chemotherapy. A National Cancer Institute of Canada Clinical Trials Group (NCIC CTG) trial. *Journal of Clinical Oncology* **22,** 7022-7022.

Sims, P., Grover, P., Swaisland, A., Pal, K. and Hewer, A. (1974) Metabolic activation of benzo (a) pyrene proceeds by a diol-epoxide. *Nature* **252,** 326.

Singer, S.J. and Nicolson, G.L. (1972) The Fluid Mosaic Model of the Structure of Cell Membranes. *Science* **175,** 720-731.

Skaletz-Rorowski, A. and Walsh, K. (2003) Statin therapy and angiogenesis. *Curr Opin Lipidol* **14,** 599-603.

Smelter, A. and Moseley, H. (2018) Isotopic Enumerator.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry* **78,** 779-787.

Smyth, G.K. (2005) Limma: linear models for microarray data, *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer. pp. 397-420.

Sonveaux, P., Vegran, F., Schroeder, T., Wergin, M.C., Verrax, J., Rabbani, Z.N., De Saedeleer, C.J., Kennedy, K.M., Diepart, C., Jordan, B.F., Kelley, M.J., Gallez, B., Wahl, M.L., Feron, O. and Dewhirst, M.W. (2008) Targeting lactate-

fueled respiration selectively kills hypoxic tumor cells in mice. *J Clin Invest* **118,** 3930-42.

Southam, A.D., Weber, R.J., Engel, J., Jones, M.R. and Viant, M.R. (2016) A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics. *Nat Protoc* **12,** 310-328.

Spiess, H.W. (2008) NMR spectroscopy: pushing the limits of sensitivity. *Angewandte Chemie International Edition* **47,** 639-642.

Spigel, D.R., Reckamp, K.L., Rizvi, N.A., Poddubskaya, E., West, H.J., Eberhardt, W.E.E., Baas, P., Antonia, S.J., Pluzanski, A. and Vokes, E.E. (2015) A phase III study (CheckMate 017) of nivolumab (NIVO; anti-programmed death-1 [PD-1]) vs docetaxel (DOC) in previously treated advanced or metastatic squamous (SQ) cell non-small cell lung cancer (NSCLC), American Society of Clinical Oncology.

Stadtman, E.R. (1970) 8 Mechanisms of Enzyme Regulation in Metabolism in Boyer, P.D. (Ed), *The Enzymes*, Academic Press. pp. 397-459.

Strong, J.M., Anderson, L.W., Monks, A., Chisena, C.A. and Cysyk, R.L. (1983) A 13C tracer method for quantitating de novo pyrimidine biosynthesis in vitro and in vivo. *Analytical biochemistry* **132,** 243-253.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill Jr, A.H., Murphy, R.C., Raetz, C.R. and Russell, D.W. (2006) Lmsd: Lipid maps structure database. *Nucleic acids research* **35,** D527-D532.

Sukhanova, A., Gorin, A., Serebriiskii, I.G., Gabitova, L., Zheng, H., Restifo, D., Egleston, B.L., Cunningham, D., Bagnyukova, T. and Liu, H. (2013) Targeting C4-demethylating genes in the cholesterol pathway sensitizes cancer cells to EGF receptor inhibitors via increased EGF receptor degradation. *Cancer discovery* **3,** 96-111.

Sun, R.C., Fan, T.W.-M., Deng, P., Higashi, R.M., Lane, A.N., Le, A.-T., Scott, T.L., Sun, Q., Warmoes, M.O. and Yang, Y. (2017) Noninvasive liquid diet delivery of stable isotopes into mouse models for deep metabolic network tracing. *Nature communications* **8,** 1646.

Sunshine, J. and Taube, J.M. (2015) Pd-1/Pd-L1 Inhibitors. *Current opinion in pharmacology* **23,** 32-38.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* **43,** 1947-58.

Tamagnini, P., Axelsson, R., Lindberg, P., Oxelfelt, F., Wünschiers, R. and Lindblad, P. (2002) Hydrogenases and hydrogen metabolism of cyanobacteria. *Microbiology and Molecular Biology Reviews* **66,** 1-20.

Team, R.C. (2013) R: A language and environment for statistical computing.

Torán, J. (2004) On the hardness of graph isomorphism. *SIAM Journal on Computing* **33,** 1093-1108.

Trainor, P.J., Mitchell, J.M., Carlisle, S.M., Moseley, H.N., DeFilippis, A.P. and Rai, S.N. (2018) Inferring metabolite interactomes via molecular structure informed Bayesian graphical model selection with an application to coronary artery disease. *bioRxiv***,** 386409.

Tsoumpra, M.K., Muniz, J.R., Barnett, B.L., Kwaasi, A.A., Pilka, E.S., Kavanagh, K.L., Evdokimov, A., Walter, R.L., Von Delft, F. and Ebetino, F.H. (2015) The inhibition of human farnesyl pyrophosphate synthase by nitrogen-containing bisphosphonates. Elucidating the role of active site threonine 201 and tyrosine 204 residues using enzyme mutants. *Bone* **81,** 478-486.

Ullmann, J.R. (1976) An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)* **23,** 31-42.

Uramoto, H., Osaki, T., Inoue, M., Taga, S., Takenoyama, M., Hanagiri, T., Yoshino, I., Nakanishi, R., Ichiyoshi, Y. and Yasumoto, K. (1999) Fas expression in non-small cell lung cancer: its prognostic effect in completely resected stage III patients. *European Journal of Cancer* **35,** 1462-1465.

Uramoto, H. and Tanaka, F. (2014) Recurrence after surgery in patients with NSCLC. *Translational lung cancer research* **3,** 242.

Van Rossum, G. and Drake Jr, F.L. (1995) *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

van Vugt-Lussenburg, B.M., Damsten, M.C., Maasdijk, D.M., Vermeulen, N.P. and Commandeur, J.N. (2006) Heterotropic and homotropic cooperativity by a drug-metabolising mutant of cytochrome P450 BM3. *Biochemical and biophysical research communications* **346,** 810-818.

Vansteenkiste, J., De Ruysscher, D., Eberhardt, W., Lim, E., Senan, S., Felip, E., Peters, S. and Group, E.G.W. (2013) Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology* **24,** vi89-vi98.

Venter, A., Nefliu, M. and Cooks, R.G. (2008) Ambient desorption ionization mass spectrometry. *TrAC Trends in Analytical Chemistry* **27,** 284-290.

Vincent, J.B. and Vigh, G. (1998) Nonaqueous capillary electrophoretic separation of enantiomers using the single-isomer heptakis (2, 3-diacetyl-6-sulfato)-β-cyclodextrin as chiral resolving agent. *Journal of Chromatography A* **816,** 233-241.

Visca, P., Sebastiani, V., Botti, C., Diodoro, M.G., Lasagni, R.P., Romagnoli, F., Brenna, A., De Joannon, B.C., Donnorso, R.P. and Lombardi, G. (2004) Fatty acid synthase (FAS) is a marker of increased risk of recurrence in lung carcinoma. *Anticancer research* **24,** 4169-4174.

Walt, S.v.d., Colbert, S.C. and Varoquaux, G. (2011) The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* **13,** 22-30.

Wang, X., Zhao, X., Gao, X., Mei, Y. and Wu, M. (2012) A new role of p53 in regulating lipid metabolism. *Journal of molecular cell biology* **5,** 147-150.

Wang, Y., Zhang, X., Tan, W., Fu, J. and Zhang, W. (2002) Significance of fatty acid synthase expression in non-small cell lung cancer. *Zhonghua zhong liu za zhi [Chinese journal of oncology]* **24,** 271-273.

Wang-Sattler, R., Yu, Z., Herder, C., Messias, A.C., Floegel, A., He, Y., Heim, K., Campillos, M., Holzapfel, C., Thorand, B., Grallert, H., Xu, T., Bader, E., Huth, C., Mittelstrass, K., Döring, A., Meisinger, C., Gieger, C., Prehn, C., Roemisch-Margl, W., Carstensen, M., Xie, L., Yamanaka-Okumura, H., Xing, G., Ceglarek, U., Thiery, J., Giani, G., Lickert, H., Lin, X., Li, Y., Boeing, H., Joost, H.-G., de Angelis, M.H., Rathmann, W., Suhre, K., Prokisch, H., Peters, A., Meitinger, T., Roden, M., Wichmann, H.E., Pischon, T., Adamski, J. and Illig, T. (2012) Novel

biomarkers for pre-diabetes identified by metabolomics. *Molecular systems biology* **8,** 615-615.

Warburg, O. (1956) On the origin of cancer cells. *Science* **123,** 309-14.

Warburg, O., Posener, K. and Negelein, E. (1924) The metabolism of cancer cells. *Biochem Z* **152,** 319-44.

Ward, P.S. and Thompson, C.B. (2012) Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell* **21,** 297-308.

Watson, D.G. (2013) A Rough Guide to Metabolite Identification Using High Resolution Liquid Chromatography Mass Spectrometry in Metabolomic Profiling in Metazoans. *Computational and Structural Biotechnology Journal* **4,** e201301005.

Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28,** 31-36.

Weis, M., Heeschen, C., Glassford, A.J. and Cooke, J.P. (2002) Statins have biphasic effects on angiogenesis. *Circulation* **105,** 739-45.

Wen, Y.-A., Xiong, X., Zaytseva, Y.Y., Napier, D.L., Vallee, E., Li, A.T., Wang, C., Weiss, H.L., Evers, B.M. and Gao, T. (2018) Downregulation of SREBP inhibits tumor growth and initiation by altering cellular metabolism in colon cancer. *Cell Death & Disease* **9,** 265.

Wendel, A.A., Lewin, T.M. and Coleman, R.A. (2009) Glycerol-3-phosphate acyltransferases: rate limiting enzymes of triacylglycerol biosynthesis. *Biochimica et biophysica acta* **1791,** 501-506.

Wenk, M.R. (2005) The emerging field of lipidomics. *Nature reviews Drug discovery* **4,** 594.

Wiechert, W. (2001) 13C metabolic flux analysis. *Metabolic engineering* **3,** 195-206.

Wiemer, A.J., Tong, H., Swanson, K.M. and Hohl, R.J. (2007) Digeranyl bisphosphonate inhibits geranylgeranyl pyrophosphate synthase. *Biochemical and biophysical research communications* **353,** 921-925.

Wieser, M.E., Holden, N., Coplen, T.B., Böhlke, J.K., Berglund, M., Brand, W.A., De Bièvre, P., Gröning, M., Loss, R.D. and Meija, J. (2013) Atomic weights of the elements 2011 (IUPAC Technical Report). *Pure and Applied Chemistry* **85,** 1047-1078.

Wiley, W. and McLaren, I.H. (1955) Time-of-flight mass spectrometer with improved resolution. *Review of scientific instruments* **26,** 1150-1157.

Williams, L.E., Odom-Maryon, T.L., Liu, A., Chai, A., Raubitschek, A.A., Wong, J.Y. and D'Argenio, D.Z. (1995) On the correction for radioactive decay in pharmacokinetic modeling. *Medical physics* **22,** 1619-1626.

Wilson, I. and Nicholson, J. (2003) Topics in xenobiochemistry: do metabolic pathways exist for xenobiotics? The micro-metabolism hypothesis. *Xenobiotica* **33,** 887-901.

Wise, D.R., DeBerardinis, R.J., Mancuso, A., Sayed, N., Zhang, X.-Y., Pfeiffer, H.K., Nissim, I., Daikhin, E., Yudkoff, M. and McMahon, S.B. (2008) Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. *Proceedings of the National Academy of Sciences* **105,** 18782-18787.

Wise, D.R. and Thompson, C.B. (2010) Glutamine addiction: a new therapeutic target in cancer. *Trends in biochemical sciences* **35,** 427-433.

Wishart, D.S. (2008) Metabolomics: applications to food science and nutrition research. *Trends in food science & technology* **19,** 482-493.

Wishart, D.S. (2011) Advances in metabolite identification. *Bioanalysis* **3,** 1769-1782.

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C. and Karu, N. (2017) HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research* **46,** D608-D617.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D. and Sawhney, S. (2007) HMDB: the human metabolome database. *Nucleic acids research* **35,** D521-D526.

Wozniak, A.J., Crowley, J.J., Balcerzak, S.P., Weiss, G.R., Spiridonidis, C.H., Baker, L.H., Albain, K.S., Kelly, K., Taylor, S.A. and Gandara, D.R. (1998) Randomized trial comparing cisplatin with cisplatin plus vinorelbine in the treatment of advanced non-small-cell lung cancer: a Southwest Oncology Group study. *Journal of Clinical Oncology* **16,** 2459-2465.

Xian, F., Valeja, S.G., Beu, S.C., Hendrickson, C.L. and Marshall, A.G. (2013) Artifacts induced by selective blanking of time-domain data in Fourier transform mass spectrometry. *Journal of The American Society for Mass Spectrometry* **24,** 1722-1726.

Xiao, X., Huang, X., Ye, F., Chen, B., Song, C., Wen, J., Zhang, Z., Zheng, G., Tang, H. and Xie, X. (2016) The miR-34a-LDHA axis regulates glucose metabolism and tumor growth in breast cancer. *Scientific reports* **6,** 21735.

Yahagi, N., Shimano, H., Matsuzaka, T., Najima, Y., Sekiya, M., Nakagawa, Y., Ide, T., Tomita, S., Okazaki, H. and Tamura, Y. (2003) p53 Activation in adipocytes of obese mice. *Journal of Biological Chemistry*.

Yang, X., Neta, P. and Stein, S.E. (2017a) Extending a Tandem Mass Spectral Library to Include MS 2 Spectra of Fragment Ions Produced In-Source and MS n Spectra. *Journal of The American Society for Mass Spectrometry* **28,** 2280-2287.

Yang, Y., Fan, T.W., Lane, A.N. and Higashi, R.M. (2017b) Chloroformate derivatization for tracing the fate of Amino acids in cells and tissues by multiple stable isotope resolved metabolomics (mSIRM). *Analytica chimica acta* **976,** 63-73.

Yang, Y., Fan, T.W.M., Lane, A.N. and Higashi, R.M. (2017c) Chloroformate derivatization for tracing the fate of Amino acids in cells and tissues by multiple stable isotope resolved metabolomics (mSIRM). *Analytica Chimica Acta* **976,** 63-73.

Ye, T., Mo, H.P., Shanaiah, N., Gowda, G.A.N., Zhang, S.C. and Raftery, D. (2009) Chemoselective N-15 Tag for Sensitive and High-Resolution Nuclear Magnetic Resonance Profiling of the Carboxyl-Containing Metabolome. *Analytical Chemistry* **81,** 4882-4888.

Yuneva, M.O., Fan, T.W., Allen, T.D., Higashi, R.M., Ferraris, D.V., Tsukamoto, T., Mates, J.M., Alonso, F.J., Wang, C., Seo, Y., Chen, X. and Bishop, J.M. (2012) The metabolic profile of tumors depends on both the responsible genetic lesion and tissue type. *Cell Metab* **15,** 157-70.

Zaidi, N., Royaux, I., Swinnen, J.V. and Smans, K. (2012) ATP citrate lyase knockdown induces growth arrest and apoptosis through different cell- and environment-dependent mechanisms. *Mol Cancer Ther* **11,** 1925-35.

Zechner, R., Zimmermann, R., Eichmann, Thomas O., Kohlwein, Sepp D., Haemmerle, G., Lass, A. and Madeo, F. (2012) FAT SIGNALS - Lipases and Lipolysis in Lipid Metabolism and Signaling. *Cell Metabolism* **15,** 279-291.

Zhang, F. and Du, G. (2012) Dysregulated lipid metabolism in cancer. *World journal of biological chemistry* **3,** 167-174.

Zhang, J., Reedy, M.C., Hannun, Y.A. and Obeid, L.M. (1999) Inhibition of caspases inhibits the release of apoptotic bodies: Bcl-2 inhibits the initiation of formation of apoptotic bodies in chemotherapeutic agent-induced apoptosis. *The Journal of cell biology* **145,** 99-108.

Zhou, M., Guan, W., Walker, L.D., Mezencev, R., Benigno, B.B., Gray, A., Fernández, F.M. and McDonald, J.F. (2010) Rapid mass spectrometric metabolic profiling of blood sera detects ovarian cancer with high accuracy. *Cancer Epidemiology and Prevention Biomarkers***,** 1055-9965. EPI-10-0126.

Zubarev, R.A. and Makarov, A. (2013) Orbitrap mass spectrometry, ACS Publications.

VITA

Joshua Merritt Mitchell

EDUCATION

BS in Chemistry from University of Louisville, Louisville KY, Aug 2008 – May 2012

MD (Incomplete), University of Louisville, Louisville KY, Aug 2012 – Aug 2014

ACADEMIC EMPLOYMENT

Graduate Teaching Assistant, Department of Molecular and Cellular Biochemistry at University of Kentucky, Jan 2015 – Aug 2015

Graduate Research Assistant to Dr. Hunter Moseley, Department of Molecular and Cellular Biochemistry at University of Kentucky, Aug 2014 – Present

Student Research Assistant to Dr. Hunter Moseley, Department of Chemistry at University of Louisville, Dec 2010 – Aug 2014

SCHOLASTIC AND PROFESSIONAL HONORS

ASBMB Graduate Student Travel Award, 2015

Course Representative for Genetics and Molecular Medicine, 2013

Honors in Microscopic Anatomy, Human Embryology, Neurosciences, Genetics and Molecular Medicine, Clinical Neurosciences, Cancer Immunology, and Biochemistry Research, 2012 – 2014

Participant in R25 NIH / NCI Cancer Education Program (R25-CA134283, David W. Hein), 2012 and 2013

“Outstanding Winner” and SIAM Prize Winner for the Mathematical Competition in Modeling for Discrete Mathematics, 2012

PROFESSIONAL PUBLICATIONS

Mitchell JM, Flight RM (2019) “Deriving Accurate Lipid Classification based on Molecular Formula” In Preparation.

Mitchell, Joshua; Flight, Robert M and Hunter Moseley (2019): Small Molecule Isotope Resolved Formula Enumerator: a Tool for Assigning Isotopologues and Metabolites in Fourier Transform Mass Spectra. ChemRxiv. Preprint. Submitted to Analytical Chemistry

Trainor, P. J., Mitchell, J. M., Carlisle, S. M., Moseley, H. N., DeFilippis, A. P., & Rai, S. N. (2018). Inferring metabolite interactomes via molecular structure

informed Bayesian graphical model selection with an application to coronary artery disease. *bioRxiv*, 386409.

Mitchell JM, Flight RM, Wang QJ, et al. New methods to identify high peak density artifacts in Fourier transform mass spectra and to mitigate their effects on high-throughput metabolomic data analysis. *Metabolomics*. 2018;14(10):125.

Mitchell JM, Fan TW, Lane AN and Moseley HN 2014. ”Development and in silico Eval- uation of Large-Scale Metabolite Identification Methods using Functional Group Detection for Metabolomics”. Frontier in Genetics 5:237. doi: 10.3389/fgene.2014.00237

Jones James, Suraj Kannan and Joshua Mitchell 2013. ”Dynamic Scheduling of White Water Rafting”. Harvard College Mathematics Review 6.1: 96-112. Web 13 Jul 2013

PUBLISHED ABSTRACTS

Rouchka, E. C., Chariker, J. H., Tieri, D. A., Park, J. W., Rajurkar, S., Singh, V., ... & Moore, N. (2017, September). Proceedings of the 16th Annual UT-KBRIN Bioinformatics Summit 2016: bioinformatics. In *BMC bioinformatics* (Vol. 18, No. 9, p. 377). BioMed Central.

Mitchell JM, Fan TW, Lane AN and Moseley HN 2014. ”Development and in silico Develop- ment of large-scale metabolite identification for metabolomics”. BMC Bioinformatics 15(Suppl 10): P36 5:237. doi: 10.1186/1471-2105-15-S10-P36

PATENT APPLICATIONS

Hunter N. Moseley, William J. Carreer, Robert M. Flight, Joshua Mitchell 2016 “Method and System for Identification of Metabolites” U.S. Provisional No. 62/187,901