ABSTRACT

Development of Chemical Classification and Annotation Tools
for Metabolomics and Lipidomics MS Analyses

Luke T. Richardson, Ph.D.

Mentor: Touradj Solouki, Ph.D.

Within the context of mass spectrometry (MS), chemical annotation is the process of assigning elements of chemical identity such as structure, stereochemistry, elemental composition (EC), and class to previously unknown features detected by MS. Annotation in metabolomics and lipidomics (i.e., the study of metabolite and lipid populations in biological samples, respectively) allows investigators to analyze samples in terms of relevant labels and infer about the biological and/or chemical significance of the sample based on the character of the data to which the labels are attached.

Annotation strategies vary significantly with respect to the level of specificity with which they describe a feature. A highly specific annotation would account for all structural and stereochemical elements that describe a unique molecule, and a less specific annotation would use terms that describe a group of related molecules (i.e., class) in which the annotated feature resides.

The former is preferable but also often inaccessible in exploratory, untargeted MS workflows; however, class information is readily accessible and can help investigators to make global inferences about their data.

Most MS instrumental classification strategies are dependent on previous assignment of EC to some degree prior to determination of analyte class; often, this process is made trivial using metabolomics or lipidomics databases that contain chemical ontology data about all entries. However, this dissertation documents a classification approach that operates orthogonally to conventional identification workflows and is thus independent of identity assignment by instrumental methods, allowing classification to provide sample information and guide feature identification. Additionally, this dissertation details an instrumental annotation method for MS imaging of lipids that integrates class-based annotation and image filtering for intuitive interrogations of lipid populations.

Chapter two and three discuss the development and application of *In Silico* Fractionation (iSF), a feedforward neural network (FFNN)-based tool that uses neural decision trees (NDT) to classify biological analytes detected by MS. Chapter two demonstrates an application to a wide variety of biomolecules, and Chapter three details a focused application of iSF toward lipid subclassification in lipidomics workflows.

Chapter four details a referenced Kendrick mass defect (RKMD)-based tool developed for integrated annotation and class-based image filtering of lipids

in MS imaging data. This method enables intuitive examination of lipid spatial distributions in MS imaging data via a class data-driven approach.

Chapter five summarizes the work detailed in this dissertation and explores potential future directions to follow this work, including application of iSF to classification in whole x-ome analysis and application of the RKMD-based annotation method to larger scale between-sample MS imaging analyses.

Development of Chemical Classification and Annotation Tools
for Metabolomics and Lipidomics MS Analyses

By

Luke T. Richardson, B.S.

A Dissertation

Approved by the Department of Chemistry and Biochemistry

_____
John L. Wood, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____
Touradj Solouki, Ph.D., Chairperson

_____
C. Kevin Chambliss, Ph.D.

_____
Christopher M. Kearney, Ph.D.

_____
Michael Trakselis, Ph.D.

_____
Elyssia S. Gallagher, Ph.D.

Accepted by the Graduate School
August 2021

_____
J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES x

LIST OF SCHEMES xiii

LIST OF LISTINGS xiv

LIST OF TABLES xv

LIST OF ABBREVIATIONS xvi

ACKNOWLEDGMENTS xix

DEDICATION xxi

ATTRIBUTIONS xxii

CHAPTER ONE 1

Introduction 1

1.1 Topic History, Research Agenda, and Impact 1

1.2 Ionization in Mass Spectrometry: Electrospray and Matrix-assisted Laser Desorption/Ionization 5

1.3 Mass Analyzers in Mass Spectrometry: Time-of-flight and Orbitrap 19

1.4 Ultra-Performance Liquid Chromatography Coupled to Mass Spectrometry 33

1.5 Analyte Annotation in MS Lipidomics and Metabolomics 36

CHAPTER TWO 52

Using Isotopic Envelopes and Neural Decision Tree-based *In Silico Fractionation* for Biomolecule Classification 52

2.1 Abstract 52

2.2 Introduction 53

2.3 Materials and Methods 59

2.4 Results and Discussion 68

2.5 Conclusions 97

2.6 Acknowledgements 98

CHAPTER THREE 99
Chemical Classification for Improved Lipidomics Sample Annotation with In Silico Fractionation 99
3.1 Abstract 99
3.2 Introduction 100
3.3 Materials and Methods 104
3.4 Results and Discussion 111
3.5 Conclusions 121
3.6 Acknowledgements 122

CHAPTER FOUR 123
Referenced Kendrick Mass Defect-Based Annotation and Filtering of Imaging MS Lipidomics Experiments 123
4.1 Abstract 123
4.2 Introduction 124
4.3 Materials and Methods 127
4.4 Results and Discussion 135
4.5 Conclusions 150
4.6 Acknowledgements 151

CHAPTER FIVE 152
Conclusion 152
5.1 Dissertation Overview 152
5.2 Future Directions 155

APPENDIX A 161
Using Isotopic Envelopes and Neural Decision Tree-based *In Silico Fractionation* for Biomolecule Classification 161
A.1 Python Script for Generation of Pseudo-random Polypeptide and Nucleic Acid Sequences and Elemental Compositions 161
A.2 MATLAB Script for Feedfoward Neural Network Training 163
A.3 Binary Classifier Neural Network Training Characteristics 165
A.4 Multi-target Classifier Neural Network Training Characteristics 166
A.5 Mass Resolution Evaluation for Peptide and Lipid MS Peaks 167
A.6 PEPNET Binary Classifier Confusion Matrix 168

A.7 LIPNET Binary Classifier Confusion Matrix 170
A.8 LIPSUBNET Multitarget Classifier Confusion Matrix 172
A.9 Lipid Subclass Coverage in the LIPSUBNET Training Set 174
A.10 PEPNET, LIPNET, and LIPSUBNET Score Outputs for the Multi-Class Sample 176
A.11 PEPNET and LIPNET Score Outputs for the Lipid and Peptide Sample 182
A.12 Neural Network Performance Characteristics with Simulated Low Mass Resolution Data 190
A.13 Partial LC and Full m/z Convolution of Lipids that Differ by One Degree of Unsaturation 191
A.14 Partial m/z Convolution of Lipids 192

APPENDIX B 193
Chemical Classification for Improved Lipidomics Sample Annotation with In Silico Fractionation 193
B.1 Table of Adducts that Modified Lipid Elemental Compositions in the iSF Training Set 193
B.2 iSF Feedforward Neural Network Training Time as a Function of Performance 194
B.3 Proportion of Confident and Unconfident LipiDex MS/MS Lipid Assignments 195
B.4 Representation of each Lipid Class in the LipiDex Validation Positive and Negative Polarity LC-MS Experiments 196

APPENDIX C 197
Referenced Kendrick Mass Defect-Based Annotation and Filtering of Imaging MS Lipidomics Experiments 197
C.1 Representative Schematic of the RKMD-Based Annotation Workflow 197
C.2 Lipid Headgroup Elemental Compositions Used to Calculate RKMD 200
C.3 Charged Adducts that Modified Reference Lipid Headgroup Elemental Compositions 201
C.4 Representative Schematic of the Class-Based Image Filtering Workflow 202

C.5 Heuristic Limits on Radyl Carbon Chain Lengths and Unsaturations for Filtering RKMD Annotations 204
C.6 Relationship Between RKMD δ and PPM Mass Measurement Error 205
C.7 Comparison of LIPIDMAPS Database Searching and RKMD-Based Method Lipid Assignments..... 206
C.8 High Level Lipid Class Images Produced by the RKMD-based Annotation and Image Filtering Method 211

BIBLIOGRAPHY 214

LIST OF FIGURES

Figure 2.1. Histograms of nucleic acid (purple), glycan (green), polypeptide (blue), and lipid (red) compositions as a function of mass percent composition of (a) hydrogen, (b) carbon, (c) nitrogen, and (d) oxygen overlaid with a kernel density estimation line plot..... 70

Figure 2.2. Log-log plots for theoretical MS isotopologue peak ratio, A+1/A+2, as a function of compound neutral mass (Da), show high degrees of separation between classes of biomolecules 74

Figure 2.3. True positive rates for 8 trained networks (each aimed at a specific chemical class identification) for the computationally generated data at confidence thresholds (0 to 1.0) 81

Figure 2.4. Visualization of the true positive rate (TPR) of LIPSUBNET network outputs binned at confidence threshold intervals of 0.1 (e.g., (0.1, 0.2], (0.2,0.3],...(0.9, 1.0], where each plotted point represents the TPR of the outputs which scored in the binned interval range 84

Figure 2.5. A visual representation of the *In Silico Fractionation* approach to a randomly selected portion of an LC-MS dataset that included a total of 106 unknown analytes 88

Figure 2.6. A visual representation of the *In Silico Fractionation* approach to a portion of a reversed-phase LC-MS dataset in which a mixture of HeLa digest peptides (total of 150 ng) and rat brain-extract lipids (from 250 μg rat brain) were separated and analyzed in the same acquisition. Regarding the classification of known components, PEPNET had a 98% TPR and 100% TNR, and LIPNET had a 100% TPR and 99% TNR..... 92

Figure 3.1. Receiver operator characteristic (ROC) curves for fatty acyls (A), glycerolipids (B), glycerophospholipids (C), sphingolipids (D), and sterol lipids (E) showing the performance of each iSF neural network for each class of lipid identified by LipiDex analysis. Optimal score cut-offs are marked and labeled for each class and instrument polarity.....113

Figure 3.2. Scatter plot displaying receiver iSF neural network classifier operator characteristic (ROC) area under curve (AUC) as a function of LipiDex MS/MS Dot Product. iSF inputs were binned in increments of

100 dot product units. For each class, the degree of "agreement" between the iSF predicted classifications and the expected classes as identified by LipiDex increases linearly (as confirmed by the R^2 score of the linear regression for each class) with respect to the LipiDex MS/MS Dot Product.....117

Figure 3.3. Bar charts displaying the predicted class representation for each of the sampled biological sources. Each labeled bracket indicates two biological sample sources that are differentiated by the predicted lipid class representation and the label above each bracket denotes the two biological sources that were differentiated 120

Figure 4.1. Computationally generated $[PG+H]^+$ RKMD plots demonstrate the utility of using data curation parameters, RKMD δ and radyl carbon ε exclusion windows, to enhance specificity and precision of the RKMD-based annotation method in the presence of lipids, peptides, and MALDI matrix clusters.....137

Figure 4.2. A computationally generated summed mass spectrum (A) for an MS dataset included 559 lipids was used to generate the total ion image show in (B) and a series of RKMD-based filtered mass spectrometry images (C-E). Total ion current (TIC) image (84 x 441 pixels) in B depicts the summed intensity for each coordinate. The selected class images were filtered based on molecular class, degrees of unsaturation, and radyl carbon chain length, respectively. 142

Figure 4.3. Labeled renal tissue structures spanning parts of the medulla and cortex region in class composite images depicting saturated (top) and monounsaturated PC (bottom) 144

Figure 4.4. RKMD-based filtering applied to a MALDI-IMS dataset from a human kidney section with medulla and cortex visible in all images: A) saturated PC, B) saturated PE, C) monounsaturated SM, D) monounsaturated PC, E) monounsaturated PE, and F) diunsaturated SM 146

Figure 4.5. RKMD-based filtering applied to a MALDI-IMS dataset from a human kidney section with 34, 36, 38, 40, and 42 radyl carbons in A-E followed by an enlarged region in the cortex for SM with 42 radyl carbons (F) with further classifications by the degree of unsaturation for 1(G), 2 (H), and 3 (I) unsaturations.....148

Figure A.1. Performance (left axis) and training time (right axis) for one and two hidden layers as a function of hidden layer neurons for the binary classifier architecture 165

Figure A.2. Performance (left axis) and training time (right axis) for one and two hidden layers as a function of hidden layer neurons for the multi-target classifier architecture 166

Figure A.3. Evaluation of mass resolution of peptide and lipid MS peaks.....167

Figure A.4. Confusion matrix of PEPNET binary classifier.....168

Figure A.5. Confusion matrix of LIPNET binary classifier 170

Figure A.6. Confusion matrix of LIPSUBNET multi-target classifier 172

Figure A.7. Line plot show lipid subclass coverage (left axis) and the coverage standard error of the mean (SEM, right axis) between subclasses as a function of Confidence Threshold 174

Figure A.8. Select ion chromatograms (top) and mass spectrum (bottom) of two partially convolved lipids that differ by one degree of saturation.....191

Figure A.9. Mass spectrum of two partially *m/z* convolved lipids 192

Figure B.1. Scatter plot displaying FFNN training time as a function of performance (incorrect predictions/total predictions) 194

Figure B.2. Pie chart depicting the proportions of confident (orange) and unconfident (blue) assignments as defined by a MS/MS dot product score threshold of 700.....195

Figure C.1. Line plot displaying the inverse power function that relates ppm mass measurement error to RKMD δ as a function of *m/z*.....205

Figure C.2. High level class images from the MALDI-IMS analysis of human kidney tissue.....211

LIST OF SCHEMES

Scheme 2.1. Representative *In Silico Fractionation* neural decision tree workflow diagram of a biological sample dataset containing polypeptide, lipid, and polar metabolite components 86

Scheme C.1. A representative schematic of the RKMD-based annotation workflow 197

Scheme C.2. A representative schematic of the class-based image filtering workflow 202

LIST OF LISTINGS

Listing A.1. Code for generating random peptide and nucleic sequences and elemental compositions 161

Listing A.2. MATLAB code for training feedforward neural networks 163

LIST OF TABLES

Table 2.1. LIPSUBNET performance metrics for subclassification of lipids in experimental data..... 90

Table 3.1. FFNN input vector structure.....105

Table 3.2. iSF network performance metrics in application to inputs with high confidence LipiDex identifications (MS/MS Dot Product > 700), with high isotopic purity parameters ($R^2_1 > 0.9$, $R^2_2 > 0.9$, $|\Delta\Delta m/z| < 0.1$), and with a score cutoff at 0.5. 115

Table A.1. FFNN output scores for PEPNET, LIPNET, and LIPSUBNET 176

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation 182

Table A.3. Neural network performance metrics for PEPNET, LIPNET, and LIPSUBNET with simulated low mass resolution MS data 190

Table B.1. Table displaying positive and negative polarity adducts commonly generated in ESI-MS experiments that were used to modify lipid elemental compositions of each class for iSF training set generation 193

Table C.1. Elemental compositions of lipid headgroups for each lipid class used calculate RKMD.....200

Table C.2. Adducts that were used to modify reference lipid headgroup elemental compositions to calculate RKMD for each lipid class.....201

Table C.3. Heuristic limits on radyl carbon chain lengths and unsaturations for filtering RKMD annotations 204

Table C.4. LIPIDMAPS and RKMD-based lipid assignments from kidney tissue MALDI-IMS analysis.....206

LIST OF ABBREVIATIONS

Abbreviation	Definition
A	analyte
A (in Chapter 2)	monoisotope peak
A+1	first heavy isotopologue peak
A+2	second heavy isotopologue peak
ADC	analog-to-digital converter
APCI	atmospheric pressure chemical ionziation
APPI	atmospheric pressure photoionization ionziation
AUC	area under curve
CANOPUS	Class Assignment aNd Ontology Prediction Using mass Spectrometry
CCS	collision cross section
CEM	chain ejection model
CHCA	a-cyano-4-hydroxycinnamic acid
CID	collision-induced dissociation
CRM	charge residue model
Da	dalton
DAN	1,5-diaminonaphthalene
DC	direct current
DDA	data-dependent analysis
DESI	desorption electrospray ionization
DG	diacylglycerol
DHB	2,5-dihydroxybenzoic acid
DIA	data-independent analysis
DNA	deoxyribonucleic acid
EC	elemental composition
EI	electron ionization
EP	exciton pooling
ESI	electrospray ionization

ESPT	excited-state proton transfer
FA	fatty acyls
FFNN	feedforward neural network
FNR	false-negative rate
FPR	false-positive rate
FT	Fourier transform
FT-ICR	Fourier transform ion cyclotron resonance
GDX	gradient descent with momentum and adaptive learning rate
GL	glycerolipids
GPL	glycerophospholipids
H	high
IEM	ion ejection model
IFS	isotopic fine structure
IM	ion mobility
iSF	*In Silico* Fractionation
ITO	indium tin-oxide
KM	Kendrick mass
KMD	Kendrick mass defect
L	low
LC	liquid chromatography
LMSD	LIPIDMAPS structure database
m	slope
M (in Chapter one)	matrix
m/z	mass-to-charge ratio
$M^{+\cdot}$	radical cation
MAD	median absolute deviation
MALDI	matrix-assisted laser desorption/ionization
MCP	microchannel plate
MD	molecular dynamics
MME	mass measurement error
MPI	multiphoton ionization

MRP	mass resolving power
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MS1	single-stage mass spectrometry
MW	molecular weight
N	number of neural network inputs
NA	natural abundance
NCE	normalized collision energy
NDT	neural decision tree
O/P-	ether-linked
oa	orthogonal acceleration
PA	phosphatidic acid
PC	phosphatidylcholine
PE	phosphatidylethanolamine
PF	polar fluid
PG	phosphatidylglycerol
PK	polyketides
PPM	parts per million
PPV	positive predictive value
PR	prenol lipids
QC	quality control
QqTOF	double quadrupole time-of-flight
Q-TOF	quadrupole time-of-flight
R^2	Pearson correlation coefficient squared
RF	radio frequency
RKMD	referenced Kendrick mass defect
RNA	ribonucleic acid
ROC	receiver operator characteristic
RP	reversed-phase
RP-UPLC	reversed-phase ultra-performance liquid chromatography

ACKNOWLEDGMENTS

I thank my Ph.D. advisor, Dr. Touradj Solouki, for his constant support of my research goals and his warm encouragement through good and hard times. He enthusiastically sought opportunities for me to learn and grow as a scientist, and I value all of the conversations we had about new and old ideas, interesting research findings, family, and faith-driven research.

I would also like to thank my committee members, Drs. C. Kevin Chambliss, Elyssia S. Gallagher, Michael Trakselis, and Christopher M. Kearney. Throughout my time as a Ph.D. student, you have each supported my growth and efforts as a scientist. I would also offer a special thanks to Dr. Christopher M. Kearney who fostered my early love for research as my undergraduate research advisor.

My time at Baylor was made all the more valuable by the friendship, support, constructive criticism, and friendly banter provided by my peers. I would like to thank Drs. Brett Harper, Michael Pettit, and Ian Anthony and future Drs. Amy Schnelle, Raul Villacob, Kyle Wilhem, Drew Stolpman, Carter Lantz, and Brooke Brown. I offer a special thanks to Dr. Matthew Brantley who taught me, among myriad lessons, the beauty of asking the right questions and using the right tool for the job. I would also be remiss if I did not thank the bright and talented undergraduate members of the Solouki lab (many of whom

are graduated and onto bigger and better things!): Shubhneet Warar, Christina A. Gaw, Adam R. Floyd, Nathan Dunkerley, and Surya Chivukula.

I am likewise grateful for the fruitful collaboration with the members of Dr. Kermit Murray's lab at LSU. In particular, I would like to thank Drs. Kermit Murray, Fabrizio Donnarumma, and Remi Lawal and future Drs. Chao Dong and Chisom Egbejiogu for all of their hard work and the opportunity to let me contribute to their work during my Ph.D.

I would like to thank my family and dear friends that have supported and loved me through this long endeavor. It would not have been possible without you and your consistent reminders that the best is yet to come. Finally, I wish to thank my fiancé, Dr. Katie Adair, for her love and encouragement through this last mighty push to bring this chapter to a close.

I have found this time in graduate school the most rewarding in my life, rich in friends, teachers, mentors, and mentees. The conversations shared, whether scientific discussions, political arguments, philosophical speculations, or debates on the nature of bugs and features, honed me in ways at which I can only guess. I consider it all a blessing. Thank you all for making it so.

DEDICATION

To Dr. Katie

ATTRIBUTIONS

Chapters two, three, and four were collaborative works. For Chapter two, Luke T. Richardson collected roughly seventy-five percent of the data, performed ninety percent of the data analysis, and wrote ninety-five percent of the text. Matthew R. Brantley performed ten percent of the data analysis, wrote the other five percent of the text, and provided consistent conceptual support. Touradj Solouki provided project leadership and edited the chapter. Twenty-five percent of the data was acquired through a mass spectrometry data repository, Chorus, and was provided courtesy of Dr. Fecundo Fernandez.

For Chapter three, Luke T. Richardson performed ninety percent of the data analysis and wrote all of the text. Shubhneet Warar performed the other ten percent of the data analysis and provided edits for the chapter. Touradj Solouki provided project leadership and edited the chapter. One hundred percent of the data was acquired through a mass spectrometry data repository, Chorus, and was provided courtesy of Dr. Joshua Coon.

For Chapter four, Luke T. Richardson performed ninety percent of the data analysis and wrote ninety percent of the text. Elizabeth K. Neumann collected all of the data, performed the other ten percent of the data analysis, and wrote the other ten percent of the text. Touradj Solouki provided project leadership and edited the chapter.

In Chapter five, all preliminary data analysis was performed by Luke T. Richardson. The whole x-omics data discussed in Chapter 5.1 was provided by Austin Quach. The MALDI-MS image data discussed in Chapter 5.2 was provided by Elizabeth K. Neumann.

CHAPTER ONE

Introduction

1.1 Topic History, Research Agenda, and Impact

Advances in exploratory mass spectrometry (MS) metabolomics and lipidomics have greatly enhanced our understanding of cellular metabolic pathways[1, 2]. The goal of untargeted, exploratory MS approaches is to determine, with the greatest level of specificity possible, the identity and abundance of each detected analyte[3, 4]. The process of assigning structural identity, class, or other chemical information to analyte features in MS is often referred to as "annotation" and is indispensable for characterization of unknown analyte populations in biological samples[3, 5-9]. Annotation by chemical classification of analytes in MS has proved invaluable in cases where identification of discrete structures is hindered by sample complexity and instrumental limitations; analyzing data in terms of class labels can elucidate broad changes in sample populations[10-12]. Additionally, if a chemical classification method is orthogonal to an employed identification method, it can refine lists of putative identifications to produce high confidence identifications[13, 14]. The research presented in Chapters two and three demonstrates an *In Silico* Fractionation (iSF) method[15] for classification and annotation of varied classes of biological analytes and an application of iSF for classification of lipid analytes. Chapter four describes a

lipid annotation method tailored for class-based image filtering in MS images of lipids in tissues. In short, this dissertation aims to provide helpful chemical classification and annotation tools to enhance metabolomics and lipidomics analysis workflows.

Chemical classification workflows for MS data were first utilized in application to highly complex sample analyses of natural organic matter (e.g., coal, bitumen, petroleum, etc.) with ultrahigh resolving power MS instrumentation[10-12, 16, 17]. The high mass resolving power and mass measurement accuracy of such high-end instruments enabled determination of elemental composition (EC) and then visualization and discrimination of classes of molecules based on stoichiometric ratios in van Krevelen diagrams[10]. Additionally, accurate measurement of mass defect (i.e., the decimal value of exact mass measurements) allowed for identification of chemically related families of compounds in Kendrick mass defect (KMD) visualization where related compounds cluster in linear series of data points[18-20]. KMD was particularly useful for families of polymers with repeating structural elements but has also been extended to analyses of biological molecules[18, 21-23]. Notably, classification in van Krevelen-based workflows is dependent on correct EC assignment and conventional KMD workflows require structural identification of one element of a family cluster to classify each data point. The quality of chemical classifications produced by these methods are determined by the quality of the EC assignment made by mass accuracy.

The scope of this research is to explore development of annotation and orthogonal chemical classification workflows. In its usage here, "orthogonal" describes the independence of the chemical classification workflow from other conventional identification workflows, namely EC determination by mass accuracy or fragmentation spectral matching by tandem MS (MS/MS). For instance, Chapter two discusses the use of a supervised feedforward neural network (FFNN) tool that allows for classification of features in MS spectra utilizing the feature's *m/z*, isotopic ratios, and KMD information. Because chemical classification is independent of identification workflows in this application, the information that iSF provides is available to refine compound identifications made by mass accuracy or tandem MS fragmentation spectral matching and describe sample composition from a "bird's eye" view. Chapter three demonstrates an application of iSF to a liquid chromatography (LC) MS/MS dataset for analysis of lipids from several different biological sources; it is shown that iSF can be used to differentiate those biological sources by their observed differential lipid class compositions. However, some MS analyses, namely those that do not incorporate pre-MS separations (e.g., LC and ion mobility (IM)), are incompatible with iSF because significant convolution of isotopologue peaks in the *m/z* domain severely degrades iSF performance. This challenge is particularly present in MS imaging of lipids in cells and tissues as most MS imaging instruments lack pre-MS IM separation. To address this challenge, Chapter four discusses an enhanced implementation of a KMD

chemical classification workflow, called referenced KMD (RKMD)[23], that provides complete sum composition annotation of lipids and integrates class-based image filtering to give investigators an intuitive analytical workflow progressing from lesser to greater specificity.

The importance of this research is twofold. First, orthogonal chemical classification methods such as iSF will improve exploratory metabolomics and lipidomics workflows by providing more comprehensive sample annotation at varying levels of specificity. Second, analyses at the class-level of specificity allow investigators access to global inferences about their data. This is especially useful when assessing large scale differences between sample conditions, organisms, and tissue structures to mention just a few applicable cases.

The remainder of Chapter one is devoted to introducing and discussing topics that are helpful to understanding the research that follows. Although the research presented in this dissertation focuses on novel chemical classification, annotation, and data handling methods, these methods were tested and demonstrated on MS data acquired from a variety of different MS instrumental configurations that incorporated electrospray ionization and matrix-assisted laser desorption/ionization using Orbitrap and time-of-flight (TOF) mass analyzers equipped with LC pre-separation. Background for each ionization method and mass analyzer as well separation methods is provided to assist the reader in understanding the data types used for developing the chemical classification and annotation tools discussed in this dissertation.

1.2 Ionization in Mass Spectrometry: Electrospray and Matrix-assisted Laser Desorption/Ionization

1.2.1 Introduction to Mass Spectrometry

Mass spectrometry (MS) is a powerful chemical characterization technique that utilizes electric and/or magnetic fields for separation of ionized molecules in gas phase. Ion separation is based on the mass-to-charge ratio (*m/z*) of ionized molecules and, hence, a mass spectrum provides *m/z* and ion abundance values in x and y axes, respectively[24]. Generally, all mass spectrometers include three major components (an ionization source, a mass analyzer, and a detector) that are housed in a vacuum chamber[24]. The ionization source is used to convert neutral analytes, either in the solid, liquid, or gas phase, to positively- or negatively-charged, gas-phase species that can then be separated in the mass analyzer[24]. The mass analyzer allows for temporal and/or spatial separation of the ionized molecules and determination of their *m/z* values in the gas-phase. Separation of the ions can be accomplished via ion acceleration and utilization of either an electric, a magnetic, or a combined electromagnetic field and subsequent measurement of ion's physical properties such as natural frequency in a magnetic or electric fields, time-of-flight in a defined field-free space, and/or other complex ion trajectories. The detector indicates the presence of gas-phase ions via transduction of ion currents to electrical currents to yield information about ion abundances. Thus, primary

data output in MS is the mass spectrum which correlates the quantity of ion current (i.e., ion abundance on the y-axis) as a function of ion's *m/z* value (on the x-axis). Mass spectrometers can also be used as gas-phase devices to carry out chemical reactions or monitor dissociation or fragmentation of ions. Therefore, mass spectrometry analysis can provide a wealth of information about ion's identity (*e.g.*, *m/z* value, ionization energy, appearance energies of its unimolecular dissociation products or fragments, fragmentation pathways), affinity for adduct formation (e.g., protons, electrons, *etc.*), and abundances.

The work described in this dissertation utilizes data acquired from MS characterization of biological tissue samples employing both liquid- and solid-phase ionization methods. Specifically, electrospray ionization (ESI) was utilized for characterization of rat brain samples in solution (data presented in Chapters two and three) and matrix-assisted laser desorption ionization (MALDI) was employed to acquire data from solid-phase kidney tissue samples (data presented in Chapter four). Mass spectrometers utilized in this work included TOF and Fourier transform (FT) MS instruments. The following section provides a brief survey of the MS field and currently utilized modern ionization sources, mass analyzers, and detectors that are relevant to the work presented in this dissertation.

1.2.2 Introduction to MS Ionization Sources

MS analytes that originate in biological samples are introduced to MS ionization sources predominantly in the liquid- or solid-phases. Ionization sources have two primary functions: 1) conversion of molecules in the liquid- or solid-phase to gas-phase species and 2) conversion of neutrally charged molecules into an ionized form[25]. Ionization sources can generally be classified as "hard" or "soft". Hard ionization techniques are characterized by a propensity to generate molecular fragment ions in the ionization source prior to MS measurement and detection.[24] An exemplary hard ionization technique is electron ionization (often also referred to as electron impact ionization (EI) which uses high energy electrons to abstract electrons from vaporized analytes[24]. EI produces both radical cations (termed "$M^{+\cdot}$") and fragment ions as some portion of radical molecular ions undergo molecular dissociation within the ionization source. However, EI has limited applications to large analytes (> 1000 Da) and to biological compounds which are generally thermally unstable, polar, and thus difficult to vaporize in the high temperatures at which EI operates[24, 26]. Conversely, soft ionization techniques are characterized by their predominant generation and conservation of molecular ion species in a wide range of ionization source conditions (e.g., under large pressure ranges, from gas-, liquid-, or solid-phase sources) and in application to a wide variety of polar and thermally unstable biological analytes[27, 28].

Several different types of soft ionization sources have been developed for the ionization of liquid-phase analytes under ambient pressure conditions, including ESI[29], atmospheric pressure chemical ionization (APCI)[30], and atmospheric pressure photoionization (APPI)[31]. In general, ambient pressure liquid-phase ionization sources operate via rapid aerosolization of analytes[25, 32]. In the process of aerosolization, such ionization techniques produce ionic species which are then selectively extracted into the MS system. Cationic analyte species are produced primarily via attachment of positively-charged chemical groups (e.g., H^+, Na^+, K^+, NH_4^+)[33] whereas anionic analyte species are produced via abstraction of positively charged chemical groups or attachment of negatively charged chemical groups (e.g., Cl^-, Li^-)[34, 35]. ESI is a commonly used ionization source for MS analysis and is broadly suitable for a wide range of polar analytes; however, ESI is less suitable for ionization of non-polar organic molecules, such as steroids, non-polar lipids, and synthetic polymers (more details on ESI are provided in section 1.2.3). Alternatively, APCI and APPI are often utilized to ionize less polar compounds that are incompatible with ESI. APPI excels in ionization of highly non-polar molecules such as naphthalene and acridine[31] that are incompatible with APCI. APCI involves a series of gas-phase reactions that include: 1) ionization of nitrogen gas by corona discharge, 2) charge transfer from charged nitrogen to solvent, and 3) charge transfer from the solvent to analyte (to produce cationic analyte forms) or from analyte to solvent (to produce anionic analyte forms)[30]. In the positive-ion mode, the

above-mentioned APCI reactions produce $[M+H]^+$, and, in the negative-ion mode $[M-H]^-$ or $[M+X]^-$, in which X is an attached anion species, are formed[36]. Ionization in APPI is driven by direct ultraviolet photon excitation of analytes to yield radical cation $M^{+\cdot}$ ions or by excitation and ionization of dopant solvent additives (e.g., toluene) which can then transfer charges (i.e., predominantly in the form of H^+) to analytes[37]. A combination of the direct and dopant-assisted APPI processes yield both $M^{+\cdot}$ and $[M+H]^+$ species[37].

Soft ionization sources purposed for ionization of analytes from solid-phase samples have also been developed. Among these ionization sources, there are three primary means of sampling: 1) laser desorption as in MALDI [REF], 2) droplet-based desorption as in desorption ESI (DESI) [REF], and 3) ion beam desorption as in secondary ion MS (SIMS) [REF] using a primary ion beam. Laser irradiation ionizes analytes principally through electronic excitation of sigma bonds in the molecules of the substrate, resulting in the gain or loss of electrons[38]. To improve ionization efficiency of analytes specifically in MALDI experiments, a UV- or IR-absorbent matrix, depending on the incident laser wavelength, is used to mediate the transfer of charge from ionized matrix to analytes in the plume created by laser irradiation[27, 39] (additional details on MALDI is provided in section 1.2.4).

1.2.3 Electrospray Ionization

Electrospray ionization (ESI) was utilized to acquire data presented in Chapters two and three, and it is the most prevalent type of ionization for use with MS, given its wide compatibility with many types of polar and thermally labile biomolecules and operation at ambient pressure[32]. Additionally, ESI is easily coupled to the outlet of LC systems which enables MS analysis of LC separated analytes in real-time and from complex mixtures[40]. Generally, ESI generates ions by application of a high voltage to a solution fed into a capillary (usually made of steel or silica) that leads to the MS inlet; in this process, a charged aerosol is created. Typical solution flow rates depend on the application but range widely from tens of nL/min to several hundred μL/min[40]. At the tip of the charged capillary, the solution is deformed from a shape dependent on surface tension into a Taylor cone because of the effect of the electric field[28, 41-43]. At the end of the Taylor cone, a fine mist of charged droplets is produced; these initial droplets usually have radii in the order of micrometers[28]. Following emission from the Taylor cone, the charged droplets undergo a process of rapid evaporation and fission. As droplets evaporate and shrink in size, their charge density increases until the surface tension is matched by the Coulombic repulsion of the charges. This balanced state is achieved at the Rayleigh limit at which the number z_R of charges e is given by

$$z_R = \frac{8\pi}{e}\sqrt{\varepsilon_0 \gamma R^3} \qquad 1\text{-}1$$

where ε_0 is the vacuum permittivity physical constant, γ is the surface tension, and R is the droplet radius[28, 44]. At the Rayleigh limit, droplets emit smaller highly charged droplets via a process called jet fission. The repeated process of evaporation and jet fission continues until the size of the droplets are reduced to a few nanometers. These highly charged nanodroplets generate the gas-phase ions detectable by MS[29, 43, 45, 46]. At this point, analytes contained in these highly charged nanodroplets undergo one of three theoretical ion release mechanisms that generate gas-phase analyte ions: the ion evaporation model (IEM), the charge residue model (CRM), and the chain-ejection model (CEM)[28].

Most gas-phase ion production of small molecules in ESI can be explained by IEM. IEM is theorized to occur as low MW ionic species evaporate from the nanodroplet; the prevailing hypothesis is that the IEM process is driven by the electric field produced by surface charge accumulation on the surface of the nanodroplet. When at a sufficiently small radius (R < 10 nm), the electric field of the charged droplet (i.e., at the Rayleigh limit) causes the ejection of ions along with a small solvation sphere. The ejection rate constant k can be expressed as

$$k = \frac{k_B T}{h} exp\left(\frac{-\Delta G^*}{k_B T}\right) \qquad 1\text{-}2$$

where k_B is the Boltzmann constant, T is the temperature, h is Planck's constant, and -ΔG* is the height of the activation free energy barrier[28]. Molecular dynamics (MD) simulations of this process have shown that the ejecting solvated

ion remains “tethered” to the droplet by a string of solvent molecules, which elongates until it ruptures and releases the solvated ion. The energy barrier required for the ejection of the solvated ion is ~32 kJ/mol, and MD simulations have shown that ions do not always adopt a trajectory that overcomes this barrier and results in ejection[28]. However, Fenn et al. showed that the enthalpy of condensation released from the bonding of solvent vapor molecules to the Rayleigh-charged droplet surface can greatly assist this process via 1) drastically raising the droplet temperature and/or 2) providing the necessary energy for evaporation of nearby molecules. In the first case, the adsorption of the incident vapor molecule causes a localized “hot spot” as the vapor molecules bond noncovalently with the droplet solvent molecules (i.e., in an exothermic reaction)[43]. The enthalpy of condensation is then conducted through the entire droplet. In the second case, the condensation enthalpy is absorbed locally and results in the evaporation of a nearby molecule, molecules, or the incident vapor molecule itself. Fenn et al. conclude that, whether either of the two processes are predominant, the condensation enthalpy of incident vapor molecules increases the flux of ions from the droplet surface into the surrounding gas[43]. After release and extraction into the MS inlet, the solvated ion sheds the remaining solvent clusters through successive collisions with gasses[42].

The CRM is widely accepted as the model for the release of large globular proteins from charged ESI droplets. In CRM, single analytes are retained in the Rayleigh-charged nanodroplets that evaporate over time, and, as the last

solvation shell evaporates, the solvent charges transfer to the analyte[47-49]. Throughout the process, the droplet sheds small ions as well as solvated protons via IEM ejection to remain close to the Rayleigh limit as the nanodroplet radius shrinks[47, 49]. Although evidence from MD simulations is lacking due to the long (> 1 μs) timescales and computational demands, there is strong experimental evidence for CRM protein ion formation in that observed protein charge states are close in composition to $[M + z_R H]^{z_R+}$, where z_R is the Rayleigh charge[28, 29, 50-52]. CRM also accounts for the attachment of non-proton charge carriers such as Na^+ or K^+ when they are present in the nanodroplet. Formation of sodium and/or potassium adducts occur when the protein terminal or residue carboxylates or other high affinity functional groups bind Na^+ or K^+ just prior to complete solvent evaporation. If the population of adducted charge carriers is heterogeneous, the protein ion count is spread over multiple adduct peaks (e.g., $[M + (z_R - i - j)H + iNa + jK]^{z_R+}$) with reduced signal-to-noise ratio (S/N) as opposed to a single $[M + nX]^{n+}$ peak where X is any single charge carrier with maximum possible S/N[28, 53].

Proteins that unfold during or before the ESI process are desolvated and ionized according to the CEM. Protein folding and adoption of protein conformations are mediated by pH-dependent electrostatic interactions; when subject to acidic solvent conditions commonly used for ESI of proteins, electrostatic interactions are disrupted via protonation of basic amino acid residues[28, 54]. These proteins then adopt a highly disordered structure, and the

hydrophobic amino acid residues that reside in the hydrophobic core of physiological protein conformers are exposed to the charged solvent environment of the ESI nanodroplet[55]. Due to their hydrophobic character, unfolded protein (or perhaps any partly nonpolar polymer) chains are pushed to the surface of a Rayleigh-charged nanodroplet as indicated by MD simulations[28]. Starting with one terminus, the polymer chain is ejected from the nanodroplet to the gas-phase until the chain is completely expelled. The unfolded and disordered polymer chain can carry many charges, protons, and other charge carriers as each amino acid residue is ejected. On a residue-by-residue basis, the CEM resembles the IEM for charge-attachment and ejection of small organic molecules[28]. The short timescale of CEM is also similar to that of IEM (i.e., on the scale of ns) rather than the longer timescale of CRM (i.e., on the scale of μs)[28, 54, 56, 57]. The faster kinetics of CEM results in better desolvation, ionization efficiency, and extraction of protein molecular ions into the MS inlet result in greater detected ion intensity, as compared to signals from MS experiments that constrain ionization mechanisms to CRM[28, 56]. Given the kinetics and timescale of CRM, many proteins remain trapped in the ESI nanodroplet when they reach the interface region of the MS inlet resulting in poor extraction into the MS inlet. However, the analytes that were analyzed in this work (e.g., polypeptides, lipids, and metabolites) were likely generated by the IEM (for lipids, metabolites, and polar peptides) and CRM (for nonpolar peptides).

1.2.4 Ultraviolet Matrix-Assisted Laser Desorption/Ionization

UV-MALDI is a common method for MS surface sampling and imaging of biological substrates as it has been well-adapted for analysis of a wide range of biological analytes, including proteins[58-60], lipids[61-63], and polar metabolites[64, 65]. As the name denotes, MALDI utilizes an added matrix compound to assist in the ionization of analytes upon irradiation by a UV or IR laser. Several models for ionization have been proposed and defended[39, 66-71], but there is a general consensus that the ionization process occurs in two steps: primary ion formation followed by secondary reactions in the expanding plume[39, 66, 72]. The first step begins with the absorption of laser radiation by the matrix and conversion of the laser energy to heat; this rapid heating causes for gas-phase phase expansion and disintegration of the sample into a rapidly expanding and cooling plume[72]. The primary ion formation step occurs on the nanosecond timescale; this process begins with the irradiation of a laser that, depending on the laser system, has a pulse width of ~3-7 ns. Following the laser irradiation event photons are absorbed to initiate the condense phase ionization of the matrix molecules, which is followed by relaxation of the matrix molecules' excited states over the course of tens of nanoseconds[39, 73-76]. Following primary ionization in the condensed phase, the expansion of the irradiated matrix into the gas phase occurs over a much longer period (i.e., over tens of microseconds)[39, 77-81]. Ion pair generation stops during this phase as the energy density of the system is decreased by physical expansion and heat release[39].

Several models describe the process of primary ion formation in UV MALDI, and they can generally be divided into two categories: excited and ground electronic state models[66, 67, 76]. Among the excited electronic state models, there are the exciton pooling (EP), multiphoton ionization (MPI), and excited-state proton transfer (ESPT) models[66, 67, 76]. The cluster model, autoprotolysis/polar fluid (PF) model, and preformed ion emission constitute the ground electronic state models[66, 67].

In the MPI model, primary ions are generated through, as the name denotes, multiple photon absorption events that raise the molecular energy state above its ionization threshold. Generally, molecules used as matrix have ionization energies of ~8 eV or higher (e.g., ionization energy of dihydroxy benzoic acid (DHB), a commonly used MALDI matrix, is 8.054 eV[39, 71]), requiring 3 UV photons for photoionization. However, given the generally low laser pulse energies and ns laser pulse widths used in MALDI, this event has a relatively low chance of occurring[82]. Although matrix-matrix interactions do not significantly lower ionization potentials, analyte-matrix interactions have been demonstrated to lower ionization potentials significantly, especially if proton or electron accepting groups are present[83, 84]. For example, the interaction between DHB and proline lowered the ionization potential of DHB to 7 eV[83]. In most of such cases, favorable interactions reduce the photon absorption requirement to 2 UV photons; however, the generally large matrix:analyte ratio (~1000:1) necessitates that the majority of laser light is absorbed through matrix-only mechanisms and

that these mechanisms are responsible for the generation of the majority of ions[39]. Moreover, Knochenmuss asserts that the MPI model is incompatible with the fact that most analyte ions are formed through secondary reactions with ionized matrix molecules and not direct ionization of matrix-analyte complexes[39].

The EP model overcomes some of the shortcomings of the MPI model. The EP model asserts that photon energy, once absorbed by the matrix-rich solid, migrates within the solid as transferred excitation energy in the form of pseudo-particles called excitons[85]. Following a period of migration, excitons concentrate or "pool" in which electronic excited state energy is redistributed between two molecules. Three or more localized excitons in a molecule—presumably matrix—can pool energy to make a transition to an excited-state quantum level above its ionization energy threshold[81]. The migration and pooling effect remove the need for direct 2 or 3 photon photoionization events to generate primary matrix ions. However, upon ionization, radical matrix cations ($M^{+\cdot}$) are generated thus necessitating a downstream mechanism for generation of protonated $[M+H]^+$ and deprotonated $[M-H]^-$ matrix ions observed by positive and negative polarity MS modes, respectively[86].

The ESPT model describes proton transfer from excited-state matrix M* to analyte (A) and then to matrix molecules (M) to produce $[A+H]^+$ and $[M+H]^+$, respectively[39, 67, 87]. However attractive ESPT might be given its 1 UV photon absorption requirement and the large increase in acidity of some molecules

upon excitation[88], Knochenmuss disapproves of the model primarily because molecules that engage in ESPT efficiently serve poorly as MALDI matrixes[39]. Secondarily, ESPT systems are usually only active in solvating water or amine environments, and the MALDI plume, though it does adopt some properties of a polar fluid, likely does not equally solvate[39].

Among the ground electronic state models, the autoprotolysis model asserts the simultaneous production of [M+H]+ and [M-H]- via proton abstraction between M + M[67]. The polar fluid model expanded on this model to include dielectric stabilization of cations and anions; this is theorized to occur through the generation of a MALDI plume that acts as a dense polar fluid that has solvating properties[89, 90]. However, Knochenmuss contends that the plume fluid is not polar enough to cause significant separation of counterions, especially in the hot plume environment in which dielectric constants would decrease[39]. Lastly, the pKa of matrix autoionization ($M + M \rightarrow [M+H]^+ + [M-H]^-$) should determine ion yield, but does not[39, 89].

Preformed ion emission assumes that some of both matrix and analyte molecules exist as ionic species in the solid and that both are emitted upon laser irradiation and desorption/ablation[67, 91-93]. It also assumes some degree of autoprotolysis for generation of $[M-H]^-$ in the solution phase[66].

Under laser ablation conditions, the cluster model describes generation of large gas-phase aggregates comprised of analyte and matrix molecules either in neutral or ionic forms[68-70]. In this case, charge separation occurs by

mechanical disruption and rapid disintegration of the solid; it is possible then that large amounts of positive or negative charges can accumulate on the aggregates. Following aggregate ejection, free ions must either be generated by evaporation of neutral matrix or by ion ejection from the aggregate. As a note, analyte ions may also be generated by intra-cluster charge transfer from matrix, but that process falls under secondary reactions. Using a variable repulsive potential before extraction to the MS, the Tabet group detected large aggregates comprised of mostly matrix molecules up to 50 kDa in mass[94, 95]. Further evidence has also given support for "hard" and "soft" evaporative pathways[74]. Hard pathways are characterized by evaporation of matrix ions resulting in only low charge state analytes. Soft pathways are characterized by loss of neutral matrix molecules, often resulting in high charge state analytes.

1.3 Mass Analyzers in Mass Spectrometry: Time-of-flight and Orbitrap

Following ionization, the next two major components of MS instruments are the mass analyzer and detector. Mass analyzers separate ions in space and time such that ion populations of different mass-to-charge ratios (m/z) can be detected distinctly from one another. A requirement for all mass analyzers is high vacuum to varying degrees depending on the analyzer. Lower pressures are required in systems that require longer ion flight trajectories from which ions might deviate if colliding with gaseous species. Currently, some of the commonly used mass analyzers include transmission quadrupoles, quadrupole

ion traps, Fourier transform ion cyclotron resonance (FT-ICR), time-of-flight (TOF), and Orbitrap devices. In general, transmission quadrupoles separate ions by scanning through radiofrequency (RF) and direct current offset voltages applied to metal rods that allow for selection of *m/z* ranges with stable ion flight trajectories through the analyzer to the detector.[96, 97] Ions with unstable trajectories collide with the metal rods and are lost. Quadrupole ion traps operate in a similar manner; ions are confined in 2 or 3 dimensions (in linear and 3D ion traps, respectively) with combined RF and DC fields that can be selectively ejected from the ion trap based on the *m/z* dependent ion trajectory stabilities.[96, 97] Linear quadrupole ion traps confine ions in the 3rd dimension in a voltage well by applied DC voltages to each end of the trap. Quadrupole-based mass analyzers are high throughput due to rapid scan rates but are limited in their mass resolution (measured as $m/\Delta m_{50\%}$, the peak width at 50% of peak height divided by the *m/z*) and mass accuracy. Quadrupole mass analyzers are often coupled with higher resolution mass analyzers and operate as ion guides or mass filters; however, if quadrupoles are the primary mass analyzer in a system, they are often paired with discrete-dynode or continuous-dynode electron multiplier detectors[97].

FT-ICR confines and measures ions in a Penning trap which utilizes a strong homogenous magnetic field to radially confine ions in a cyclotron motion via the exertion of the Lorentz Force on the ions in a cubic or cylindrical cell comprised of two excitation and two detection plates[98]. The frequency of the

cyclotron motion is directly related to the charge and inversely related to the mass of the ion, and, following an RF excitation that increases the radii of the ions flight trajectory at resonant frequencies, the "coherent" cyclotron motion is detected by image current detection[98]. Therefore, in FT-ICR cells, lower mass ions have higher cyclotron frequencies than that the higher mass ions. Finally, the detected complex time-domain waveform signal, often called the transient or free induction decay, is converted to frequency domain via Fourier transform conversion of the data followed by conversion to *m/z*[98] values that represent these ICR frequencies. As the two specific types of mass analyzers used in this work, details of TOF and Orbitrap instruments are provided in the following section.

1.3.1 Time-of-flight and Microchannel Plate Detectors

For this work, several TOF-based systems were used for LC/ESI-MS analysis of lipid and peptide mixtures extracted from rat brain and MALDI-MS imaging analysis of lipids from a human kidney section; a brief review of TOF development and theory is discussed below. In TOF-MS analysis, ions are separated by accelerating them across a distance, or field-free region, to a detector. The time required for ions to traverse the field-free region is directly related to square root of its mass and inversely related to square root of its charge. After the ion extraction from the source, ions are accelerated through a

potential difference, and their m/z dependent potential energies are converted to kinetic energies according to the equation 1-3:

$$\frac{1}{2}mv^2 = zeV \qquad 1\text{-}3$$

where m is the mass of an ion in kg, v is its velocity in m/s, e is the charge of an electron, z is the number of charges, and V is the ion acceleration potential. Ions entering the field-free region have different velocities (v)

$$v = \sqrt{\frac{2zeV}{m}} \qquad 1\text{-}4$$

that depend on their mass. Hence, the time (t) required for ions to traverse the field-free region

$$t = \frac{d}{\sqrt{\frac{2zeV}{m}}} \qquad 1\text{-}5$$

(where d is the length of the field-free drift region) is dependent on the mass of the ion and can be used to calculate ions' m/z values by the rearranged equation 1-6:

$$\frac{m}{z} = 2eV\left(\frac{t}{d}\right)^2 \qquad 1\text{-}6$$

Early TOF instruments were linear and ions generally flew in a straight path from ion source to detector. In competition with other early magnetic sector, quadrupole, and FT-ICR instruments, TOF offered a number of

advantages that still remain as compelling rationale for their use today: 1) fast acquisition rates at 10s of microseconds per spectrum, 2) theoretically infinite mass range, 3) capacity for broadband *m/z* detection in the same period, and 4) mass-independent resolving power[99]. The early TOF instruments used pulsed ion sources such as laser desorption, ^{252}Cf plasma desorption, and field desorption to generate discrete ion packets that were immediately extracted by application of a large potential (25-30 kV). Under these high ion acceleration fields, ion packets with high velocities and relatively small kinetic energy distributions were produced to yield high ion detection efficiencies that minimized adverse effects on mass resolution[100]. Further efforts to reduce initial velocity, temporal, and spatial distributions of ions led to implementation of reflectron ion mirrors which improved mass resolution by lengthening the ion flight path (and thus separation time) and narrowing ion packet widths at the detector[101]. Additionally, methods for delayed source extraction were implemented to the same effect; instead of applying a static potential that would immediately and continuously accelerate ions into the field-free region, a delayed potential pulse would follow ~200-500 ns after desorption to extract ions. Delayed extraction compensates for differences in initial velocities of ion packets from the source and accelerates them differentially based on their position relative to the extraction plate. A detector placed at certain place in the flight path would detect simultaneous ion impacts as faster ions traverse a greater distance to catch up with the slower ions at the detector plane.

Reflectrons compensate for these distributions in initial velocities in a similar way; ions with greater initial velocities travel further into the reflectron field than slower ions and the packet of ions coalesces at the detector plane for simultaneous detection.

TOF instrumentation were designed first with pulsed ion sources in mind; early implementations utilized very short laser pulses[100] (1-5 ns) or simultaneous detection events (^{252}Cf plasma desorption[102]) as a start time after which ion's time of flight was recorded. However, continuous ion sources such as ESI were less compatible and required gating of ion beams[103]. The implementation orthogonal acceleration (oa) demonstrated by Guilhaus and Dawson marked a departure from previous methods in ways that increased mass resolution, sensitivity, and compatibility with continuous ion sources[104]. Generally, oa-TOF-MS instruments utilize ion optics to focus and transmit an ion beam from the ion source to the oa region. Once full, the oa, which is usually a series of conducting electrodes, rapidly pushes the ion beam orthogonally to its velocity into the field-free region. The "push" is applied by rapidly generating a potential between one or more pairs of electrodes in the oa. One large advantage of the orthogonal geometry was its capability to reduce the average initial velocity component of the ion beam in the TOF direction to zero and, therefore, narrow the distribution of initial velocities in the TOF direction[105]. Additionally, the kinetic energies of the ion beam form the source and the drift ions could be controlled independently such that the oa region of the TOF could be filled in roughly the

same time as was necessary for the largest m/z ions to traverse the field-free region[105]. The accumulating ion beam was therefore made as parallel as possible to narrow and reduce the average velocity in the TOF direction. For continuous ion sources in ambient conditions, this was best achieved via collisional cooling of the ion beam with inert gas molecules in RF-only ion guides, which reduced the average energy of the ion beam to that of the inert gas[105].

The oa-TOF design has been implemented in many hybrid MS systems in which two or more mass analyzers are coupled in series. In each, the ion beam passes through multiple stages of mass analyzers and ion optics that can perform collisional cooling, m/z filtering, and/or ion fragmentation. In particular, the quadrupole TOF (Q-TOF) configuration has found use in many analyses of biological molecules due to its high sensitivity and mass accuracy. The most common Q-TOF configuration utilizes two transmission quadrupoles in series followed by the oa-TOF mass analyzer (QqTOF). In general, the first transmission quadrupole collisionally cools ions extracted from the ion source acting as ion guide or an *m/z* filter, and the second quadrupole acts as a collision cell for collision-induced dissociation of precursor ions to collect fragmentation data in the subsequent oa-TOF analyzer. One of the TOF instruments utilized in this work is a Waters Synapt G2-S HDMS which is a variant of Q-TOF and includes a resolving quadrupole and a series of stacked ring ion guides (SRIG) that are used to trap, transmit, and fragment ions as well as perform ion mobility separations. SRIG consist of stacks of ring electrode pairs on which

radially confining RF voltages are applied with DC offsets to push or trap ions[106]. The first SRIG following the resolving quadrupole is the trap cell which guides and accumulates ions before periodically gating them into the ion mobility cell by modulating a superimposed DC voltage on the final electrode. The second SRIG is the travelling wave ion mobility cell that, when in operation, separates ions by ion collisional cross section as they are periodically pushed by a "travelling" half-sinusoidal DC offset potential against a drag force produced by collisions with inert gas molecules. The third SRIG is the transfer cell which transmits ions from the ion mobility cell to the oa region of the TOF. However, for this work, the resolving quadrupole and all SRIGs were operating as ion guides with no ion mobility separation or collisional activation. Additionally, MS data from a Waters Xevo G2 Q-TOF was used; the Xevo G2 Q-TOF is very similar in configuration to the Waters Synapt G2-S except without the travelling wave ion mobility cell and transfer cell. Instead, one travelling wave ion guide extends from a resolving quadrupole to the entrance of the TOF analyzer.

The most common type of detector used in TOF instrumentation is the microchannel plate (MCP) detector. In principle, it acts much like an electron multiplier which transduces ion current to electric current via secondary emission of electrons. However, MCPs have many small tubes or microchannels that traverse the plate at a small angle from the normal of the surface that each act like electron multipliers when a large potential is applied across the plate. The microchannel offset angle guarantees that ions that enter the channel

impact on the walls. When a large potential is applied across the plate, ion impact in a channel causes a cascade of emitted electrons that cause multiplication of electrons with each subsequent impact through the channel until they reach the detector anode. The multiplication effect causes a large amount of signal gain that allows sensitive detection of low numbers of ions in MS. Following ion impact and electron cascade, MCPs require a rest period for charge replenishment from an external voltage source[107]. Furthermore, the detector gain can be increased by stacking two (or more) MCP and rotating the second 180° from the first such that aligned microchannels form a chevron (or zig zag) pattern. So called "Chevron MCP" are commonly used in TOF instruments today. Importantly for oa-TOF instruments, MCP detectors can be made large enough to detect the full width of the ion beam on the axis orthogonal to the oa-TOF acceleration, which spreads after acceleration into the field-free region proportionately to the distribution of kinetic energies in the ion packet[105]. Notably, some arrival time spread can be introduced by variable degrees of penetration into MCP microchannels, and therefore the smallest diameter microchannels should be used[108].

Following ion impact, the electrical current collected by the anode is converted to digital signal in one of two ways in TOF system: 1) analog-to-digital converters (ADC) or 2) time-to-digital converters (TDC). TOF systems with pulsed ion sources such as MALDI typically use ADC to digitize signal for their wide dynamic range as pulsed ion sources produce a large number of ions per

pulse[108]. ADC are not generally used with oa-TOF systems due to the background noise produced by analog signal detection. Alternatively, oa-TOF systems use TDC which have a smaller dynamic range but are much more sensitive and can be used to record single ion impact events. This increased sensitivity is important in ion fragmentation experiments conducted with hybrid oa-TOF systems as ion counts of fragment ions can often be very low. When an ion strikes the MCP, a pulse of electrons is generated and subsequently amplified and used to generate a timing pulse to the TDC. The TDC registers the arrival time of the pulse relative to the initial time point when the ions were accelerated into the field-free region, producing a series of triggered arrival times. Over the course of many TOF detection events in an acquisition (or scan) period, the triggered arrival times are summed in memory to produce a mass spectrum. Following a TDC timing pulse, there is a certain deadtime that follows over several nanoseconds in which no ion impacts can be registered. The length of TDC deadtime defines the dynamic range of TDC detectors; for example, if two ions that follow each other in rapid succession, the first ion impact event will register on the TDC, but the second will not if it arrives during the deadtime. TDC detectors have been improved over time by decreasing detector deadtime and by using MCP with multiple collection anodes connected to separate TDC channels.

1.3.2 Orbitrap and Image Current Detection

Beyond TOF data, this work also utilized data from an Orbitrap Q-Exactive HF MS system for acquisition of LC/MS lipidomics data. A brief review of the theory behind Orbitrap mass analyzers and detection systems is provided below. Similar to FT-ICR, Orbitrap mass spectrometers trap ions and measure the *m/z* of ions based on their ion motions frequencies in the trap. Orbitrap utilize an improved implementation of the Kingdon trap which was an early ion storage device that trapped ions in a pure electrostatic fields (as opposed to Penning traps that utilize magnetic fields as well)[109]. The early Kingdon trap used a thin wire cathode run through the center of a cylindrical anode with two electrodes capping the volume on both ends of the cylinder. DC voltage were applied between the wire and cylinder to generate a radial logarithmic potential (Φ) given by

$$\Phi = A \ln r + B \qquad 1\text{-}7$$

(where A and B are voltage-dependent constants[110] and r is the radial coordinate), and ions, when injected into the volume with a sufficient velocity perpendicular to the wire electrode, adopted a stable orbit around the wire if the potential difference between the wire and cylinder electrodes is great that the value given by

$$qV = \frac{1}{2} m v^2 \left(\frac{R}{r}\right) \qquad 1\text{-}8$$

(where R is the radius of the cylinder anode, r is the radius of the wire cathode, m is the mass of the ion, q is the charge of the ion, and v is the initial velocity of the injected ion). Application of DC voltages on the end-cap electrodes confined the ions in the axial direction. Later, Knight modified the outer cylinder electrode of the Kingdon trap to produce a quadratic axial potential given by

$$\Phi = A\left(z^2 - \frac{r^2}{2}\right) \qquad 1\text{-}9$$

(where r is the radial coordinate, z is the axial coordinate, and A is a voltage and geometry dependent constant) which produced harmonic oscillations of ions along central electrode. The combination of the quadratic axial potential and the logarithmic radial potential produced a cylindrically symmetric electrostatic potential given by

$$\Phi = A\left(z^2 - \frac{r^2}{2} + B\,ln(r)\right) \qquad 1\text{-}10$$

(where B is an additional geometry and voltage dependent constant)[109]. The radial logarithmic potential provided radial confinement of ions, and the quadratic axial potential provided the harmonic oscillation of ions along the inner electrode; however, Knight observed that resonances were weaker and frequency shifted from expected values for quadrupolar fields and surmised that they may be distorted by the logarithmic radial field[109]. These distortions were confirmed by simulation experiments, and ideal Kingdon trap was demonstrated to have spindle-shaped equipotential lines in order to achieve a

harmonic potential field[111]. The ideal Kingdon trap was demonstrated experimentally first by setting voltages on a system of electrodes to match the equipotential lines produced from equation 1-10 by the Russel group where m/z ratio was determined by ICR frequency[111, 112]. Makarov then demonstrated an ideal Kingdon trap by shaping the central electrode to match the equipotential lines to create the orbitrap mass spectrometer[113].

In summary, the orbitrap mass analyzer is a Knight-style Kingdon trap with a central spindle-shaped electrode inside of a barrel-shaped outer electrode. A DC potential is generated between to the two electrodes to produce a potential field gradient (U) given by

$$U(r,z) = \frac{k}{2}\left(z^2 - \frac{r^2}{2}\right) + \frac{k}{2}(R_m)^2 \, ln\left[\frac{r}{R_m}\right] + C \qquad 1\text{-}11$$

(where r is the radial coordinate, z is the axial coordinate, k is the axial restoring force, R_m is the characteristic radius, and C is a constant). Only ions with orbital radii less than Rm will enter stable trapping trajectories. Notably, the stability of ion trajectories is dependent on both orbital motion around the central spindle electrode and on the harmonic oscillations in the axial direction[113], and, given the lack of cross terms in equation 1-11, orbital motion and harmonic oscillations are independent of each other[114]. Although it was initially thought that *m/z* of ions might be measured by the frequency of orbital motion, these measurements were sensitive to small changes in ion properties upon injection

into the trap[113]. Orbitrap, however, uses the axial harmonic frequency of ion motion along the z-direction given by

$$z(t) = z_0 \cos\left(\sqrt{\frac{kq}{m}}t\right) + \left(\frac{2E_z}{k}\right)^{1/2} \sin\left(\sqrt{\frac{kq}{m}}t\right) \qquad 1\text{-}12$$

(where z_0 is the initial axial amplitude, E_z is the initial ion kinetic energy, and m and q are the mass and charge, respectively) to determine the m/z of ions because it is independent of the initial injection velocity and radius[113].

In contrast to TOF and similarly to FT-ICR, mass resolution is dependent on m/z. However, compared to FT-ICR, the resolving power of orbitrap decreases more slowly given the inverse proportionality of the frequency of axial oscillations with $(m/q)^{1/2}$ in orbitrap and the inverse proportionality of ICR frequency with m/q. Mass resolution and resolving power are primarily affected by the length of the transient but are also impacted by instability of voltage supplies, machining imperfections, and necessary concessions in orbitrap cell design which include the gap between the two outer electrodes and the injection slot[114, 115]. Additionally, mass resolution can decrease as a function of molecular collisional cross section; ion impacts with inert gas molecules can cause fragmentation, loss of ion packet coherence, or ejection from the trap and thus necessitate the use of ultrahigh vacuum to minimize such events[115].

Ions are detected by image current at the outer electrodes, which are divided at the center of axial plane (z = 0). The image current from each half of

the outer electrodes are differentially amplified and sampled by a digitizer to convert it to a time-domain digital signal transient. The time-domain signal is converted to a frequency spectrum by fast Fourier-transform[116] and then to *m/z*. More recently, Makarov et al. introduced enhanced Fourier transform that utilizes the phase information of the Fourier transform (specifically the "absorption" spectrum from the real component), the magnitude spectrum (i.e., the root sum squared of real and imaginary components), and finite-impulse-response filtering to increase Orbitrap mass resolution up to two-fold over significantly shorter detection periods[117].

1.4 Ultra-Performance Liquid Chromatography Coupled to Mass Spectrometry

The total peak capacity of MS analyses of complex mixtures has been greatly improved via the coupling of MS with in-line chemical separation devices. In this context, "in-line" is describing the connection of the chromatography outlet directly to the MS ion source interface. Chemical species in mixtures that are separated in-line with the MS are sampled continuously, and individual fractions are not collected. Ultra-performance liquid chromatography (UPLC) is a common chemical separation method that is coupled with MS systems for detection[40, 118, 119]. Given that chemical separations with UPLC occur in solution phase, the outlet of UPLC systems is coupled to a liquid phase MS ionization source that operates at ambient pressure such as those described in Chapter 2.2 and 2.3[30, 31, 36, 37, 120]. ESI is the most common

ionization interface used for its wide compatibility with various UPLC flowrate regimes (e.g., ~0.2 μL/min – 500 μL/min) and with medium polar to polar biological analytes[33, 40, 118, 119].

In UPLC separations, analytes are dissolved in a mobile phase liquid that is being forced through an immiscible stationary phase that is fixed within a column[40]. The mobile phase and stationary phase are specifically chosen to mediate interactions with analytes such that analytes will distribute between the mobile and stationary phases depending on some chemical property[40]. Analytes that distribute preferentially into the stationary phase will move through the column more slowly relative to the flow of the mobile phase[40]. Analytes that distribute preferentially into the mobile phase will travel quickly through the column. The differential retention of analytes on the stationary phase and migration rates through the column result separation of homogenous analyte mixtures into discrete migrating bands. Upon injection of a sample mixture, analytes distribute into the mobile and stationary phases, and, with introduction of fresh mobile phase into the column, the eluent—the part of the sample contained in the mobile phase—moves through the column and continually partitions between the mobile and stationary phase[40]. Given that analytes can only move through the column in the mobile phase, analyte migration rate depends on the average fraction of time spent in the mobile phase[40]. With continuous mobile phase flow, the eluent will eventually exit, or elute from, the

column. The goal of UPLC is to elute pure analyte bands for collection or detection by MS or some other suitable detector.

1.4.1 Reversed-Phase Chromatography

Many types of intermolecular interactions between analytes and mobile and stationary phases are utilized in UPLC applications, and various UPLC methods take advantage of a variety of mechanisms that mediate analyte partitioning between mobile and stationary phases. Most UPLC applications are conducted with a nonpolar reversed-phase (RP) stationary phase and a polar mobile phase; RP-UPLC is therefore most suitable for separation slightly polar to very polar and ionic chemicals[40, 97, 118, 120]. Some applications may use a static mobile phase, or isocratic, separation in which the composition of the mobile phase is not changed for the duration of the separation; however, it is far more common to use a gradient separation in which the composition of the mobile phase is changed as a function of time[40, 97]. For RP separations, the polarity of the mobile phase is adjusted by mixing solvents of different polarities to make the mobile phase more nonpolar with respect to time[97]. A common pair of solvents for gradient mobile phase separations are water and acetonitrile[97]. Changing the polarity of the mobile phase is useful to modulate the retention factor of analytes and increase the resolution of chromatographic separations[40, 97].

1.4.2 Considerations for Coupling with ESI-MS

The coupling of the UPLC outlet to MS for detection introduces several factors for consideration to prevent degradation of MS signal. Firstly, MS is sensitive to background contaminants and dissolved gasses that may not be an issue for other detectors; therefore, it is necessary to use ultrahigh purity solvents that have been degassed prior to use[33, 40, 97]. Secondly, depending on the ion source interface used, the mobile phase may have to be modified to allow efficient analyte ionization and detection[33]. For ESI, it is often necessary to add formic or acetic acid to small percentage (~0.1-1%, v/v%) to provide free protons for generation of protonated molecular ion species in positive-ion mode[33]. Thirdly, signal suppression by matrix effects can significantly decrease or even eliminate signal from analytes if UPLC or sample preparation parameters are not optimized[120].

1.5 Analyte Annotation in MS Lipidomics and Metabolomics

1.5.1 Introduction to Analyte Annotation Strategies in MS

Mass spectrometry has been proven a powerful tool for untargeted identification and quantitation of unknown chemical species in biological systems[4]. To date, MS has been broadly applied in characterizations of large populations of lipids and metabolites[4, 6, 121]. The process of assigning chemical information to detected MS features is called “annotation”. Annotations can vary in their level of detail and can include class, structure, stereochemistry, and/or

definitive identity information[3, 5, 7]. In order to acquire this information, MS systems are designed to operate in two general modes: 1) single-stage MS which measures the *m/z* and abundance of molecular ion species and 2) multi-stage, or tandem, MS which measures the *m/z* and abundance of fragment ions generated from precursor ions by some form of energetic activation.

This dissertation describes two analytical tools that are purposed towards annotation of MS features, which are discussed at length in Chapters two, three, and four. To test the performance of these feature annotation tools, MS peaks in LC/ESI-MS and MALDI-MS acquisitions were assigned by mass accuracy, relative isotopic peak intensity, and tandem MS fragmentation spectra matching, which are discussed in the following section.

1.5.2 Elemental Composition Determination by Mass Accuracy

In untargeted metabolomics and lipidomics, single-stage MS provides chemical information primarily by measurement of mass accurate monoisotopologue peaks which is directly related to the elemental composition of the compound[122]. The term "monoisotopologue" refers to a molecule that contains in its elemental composition the most abundant naturally occurring isotopes of each element. Accurate mass measurements benefit greatly from low mass measurement error and high mass resolving power of the modern instruments. Much like percent error, mass measurement error (MME) in parts-per-million (ppm) can be calculated dividing the difference between the

experimentally measured value and its theoretical value to theoretical values multiplied by one million:

$$MME = \frac{M_E - M_T}{M_T} * 10^6 \qquad 1\text{-}13$$

where M_T is the theoretical *m/z* of the putative identification and the M_E value is the experimentally assigned *m/z* values for the detected feature. Although different conventions can be used to define mass resolving power (MRP)[123], in this dissertation we will use the following definition for MRP:

$$MRP = \frac{m}{\Delta m_{50\%}} \qquad 1\text{-}14$$

where m is the m/z value of the peak maxima (or centroid) and $\Delta m_{50\%}$ is the full width of the peak at half its maximum intensity.

TOF and hybrid Q-TOF instruments are particularly well suited for untargeted metabolomics and lipidomics for their fast duty cycle with scan rates, that are generally between 1-5 Hz and therefore easy to couple with UPLC systems. However, the resolving power of hybrid Q-TOF systems used in LC/MS experiments is limited (generally < 40,000) and their achievable mass accuracy with internal calibration is usually within 5 ppm[124]. Although suitable in many applications, Q-TOF mass accuracy tends to produce ambiguous identifications in metabolomics and lipidomics because there are often many elemental

compositions that produce *m/z* within 5 ppm error[125, 126]. Additionally, insufficient mass resolving power can have deleterious effects on mass accuracy due to *m/z* convolution of isobaric species[127]. Isobaric species are those which have the same nominal mass (denoted by an *m/z* integer value) but have a different monoisotopic exact mass. With insufficient mass resolving power, MS peaks from isobaric ions can be convolved such that they cannot be discriminated in the *m/z* domain. However, due to the difference in exact mass, convolution of isobaric ions is likely to shift the peak centroid away from the exact mass of each convolved species, resulting in increased MME and ambiguity[127-129].

To address this problem, the MS community develops and employs ultra-high MRP instrumentation. Orbitrap instruments now consistently achieve MRP greater that 60,000 up to 1,000,000, and FT-ICR has achieve greater than 1,000,000 mass resolving power and will inevitably increase further as the field strengths of superconducting magnets continue to increase[114, 117, 130-132]. With this increase in MRP, FT-MS instruments can routinely produce mass accurate measurements below 1 ppm, and high field FT-ICR can achieve routine mass accuracy below 0.1 ppm; this is particularly advantageous in lipidomics in which mass differences between monoisotopic peaks of isobaric lipid molecular ions of less than 10 mDa are commonplace[130].

Accurate mass measurements are generally purposed towards determination of elemental composition (EC)[133, 134]. Of course, multiple

structurally distinct compounds may share the same EC, but its determination is an important step in definitively identifying a compound. In general, as MME decreases, the number of potential EC that have exact mass values within the error window also decrease, which increases the confidence associated with the best assignment[126]. Ultimately, the goal is to determine a unique EC. However, even with high mass accuracy (MME < 1 ppm), many conceivable EC can be possible matches for acquired *m/z* in mid to high mass regions (*m/z* > 300) assuming EC are limited to C, H, N, S, O, and P. EC assignment with 1 ppm MME becomes more unrealistic for molecules with complex ECs that include diverse heteroatoms[125]. In efforts to address this ambiguity in metabolomics and lipidomics analyses, it is common practice to search experimental *m/z* against metabolomics or lipidomics databases that include characterized or theorized metabolites or lipids, respectively[135]. Searching against more refined lists of compounds limits potential ECs but also increases the potential for discovering uncharacterized chemical species not included in chemical databases. Therefore, there is a compelling interest in providing additional orthogonal evidence for EC assignment in MS analyses.

1.5.3 Isotopic Envelope Analysis

In addition to accurate mass measurements, high resolution MS^1 measurements yield information about the abundance of different isotopes in a molecule[134, 136-139]. The degree to which the relative intensities of heavy

isotopologue MS peaks agree with the natural abundance of heavy isotopes in a molecule is referred to as "spectral accuracy"[140]. Similar to the exact monoisotopic mass of a molecule, the relative intensity of isotopologue peaks is directly related to elemental composition. For example, the relative abundance of carbon isotopes ^{12}C and ^{13}C are 98.94% and 1.08%, respectively. Therefore, each carbon position in a molecule has an independent probability of ~1.08% of being ^{13}C. The probability of at least one carbon position containing ^{13}C is determined by the addition rule of probability, and the relative intensity of the $^{13}C_1$ peak is ideally proportional to the sum of probabilities. For instance, C_{60} "Buckyball" has a $^{13}C_1$ isotopologue peak that is ~64.8% (1.08% x 60) of the intensity of the monoisotopic peak. Moreover, there is a lesser probability that two ($^{13}C_2$) or more (up to the number of carbons in the elemental composition) will be present in the compound; the $^{13}C_2$ isotopologue peak would be two nominal m/z units heavier than the monoisotopologue peak for a singly charged ion. In this way, the natural isotopic abundance of each element in a compound is represented in the isotopic envelope.

At Q-TOF and the low-end of Orbitrap MRPs (< 300,000), the isotopic envelopes of singly charged organic molecular ions (consisting of primarily C, H, N, O, P, and/or S) may be represented by a series of peaks at increasing nominal *m/z* units that are each convolutions of unique isotopologue peaks. At greater MRP (especially > 500,000) that can be attained with high-end Orbitrap and FT-ICR, distributions of individual isotopologues, or isotopic fine structures

(IFS), are resolved and can be attributed. Nominal m/z isotopic envelopes are useful in lipidomic and metabolomics workflows to refine potential identifications made by mass accuracy by, in general, comparing the simulated isotopic peaks of each potential identification to experimental isotopic envelopes[125, 136, 139, 140]. In fact, Feihn et al. suggested that, after a comprehensive treatment of potential metabolite structures, maintaining spectral accuracy of MS systems (i.e., limited spectral error < 2%) was of greater importance for determination of metabolite EC than decreasing MME < 0.1 ppm without regard for spectral accuracy[125]. Quantitative measurement of IFS greatly improves this method by allowing comparisons of individual isotopologue peaks (rather than convolved clusters) and especially those of low abundance heteroatoms[133, 134, 137, 138]. Moreover, the work in this dissertation shows that the relative intensities of isotopologue peaks can be used to derive chemical class information about organic biological molecules even when isotopic fine structures are convoluted. The chemometric analysis tool presented in this dissertation that performs this function, *In Silico* Fractionation, harnesses the defined distributions of elemental compositions, and therefore relative isotopologue intensities, within classes of molecules to make class predictions about unknown analytes. The background and principles for *In Silico* Fractionation operation will be discussed in Chapters two and three.

1.5.4 Tandem Mass Spectrometry Fragment Spectral Matching

Given the ambiguity that can be introduced to identification results in metabolomics and lipidomics by mass accuracy-based workflows, tandem MS methods are often employed to increase confidence in identifications[141]. The terms "Tandem MS" and "MS/MS" refer to a multi-stage MS acquisition mode in which molecular ions are fragmented by energetic ion activation mechanisms followed by separation and detection of fragment ions. Tandem MS applications in metabolomics and lipidomics can be conducted in targeted or untargeted modes. Targeted MS/MS strategies are useful for quantifying the abundance of known analytes and are therefore not suitable for annotation of unknown chemical features in LC/MS[142]. Untargeted MS/MS strategies are designed to determine the identity of unknowns principally through continuous acquisition of fragmentation spectra from LC eluting species. Acquired fragment ion spectra are compared to fragment ion spectra of standards in MS/MS spectral libraries to determine the identity of the precursor; the quality of match between experimental and library spectra are generally determined by calculation of a similarity score[141, 142]. In data-dependent analysis (DDA), fragment ions are associated with each other and their precursor via a scanning *m/z* window that targets the most abundant precursor ions for fragmentation[143, 144]. In data-independent analysis (DIA), precursor ions are continuously fragmented often with no *m/z* selection and fragment ions are grouped by the correlation of their LC (and IM, if utilized) retention times [144, 145].

1.5.5 Class-based Annotation Strategies in Metabolomics and Lipidomics

Introduction to class-based annotation. Although definitive compound identification is the goal in any untargeted MS metabolomics or lipidomics analysis, this is often precluded by instrumental limitations and sample complexity. For example, mass accuracy can discriminate between different ECs given a sufficiently small error tolerance; however, isobars with *m/z* values that fall within said error limit, including structural isomers, cannot be uniquely identified[125]. Additionally, tandem MS experiments yield fragment ion spectra of varying quality depending on the quality of pre-MS separations and the abundance of analytes, which can vary over orders of magnitude; low quality fragmentation spectra may result in incorrect assignments or compel assignment by precursor mass accuracy in lieu of fragmentation spectral matching[144]. Finally, many metabolite and lipid species may be completely uncharacterized (as members of the so-called "dark metabolome"[146]) or may not be included in tandem MS fragmentation spectral libraries. In cases that preclude definitive assignment, it is still worthwhile to characterize unknown compounds with the greatest level of specificity possible, usually by identifying the chemical classification to which they belong.

Chemical classes describe groups of related compounds based on shared physical properties, structural homology, and/or bioactivity[8, 9]. Chemical classification systems are arrayed in hierarchical ontologies, which are similar

to taxonomies in other scientific disciplines (e.g., biology) in their hierarchical structure, but describe a more complex "web" of relationships[9]. Generally, higher levels in these structures hold groups with greater diversity and lower levels hold groups with less diversity and more homology between individual species. In lieu of definitive identification, chemical classification of unknowns is important because classification enables predictions about metabolic function and structure[8].

Including the work described in this dissertation, several chemical classification methods have been developed in the field of mass spectrometry to elucidate groups of chemically related molecules or directly classify them into ontological classification structures. The methods developed in this dissertation fall into two general categories: machine learning and instrumental. Machine learning classification methods use machine learning tools to make inferences about chemical class by describing patterns in instrumental measurements that are not apparent to the user. In contrast, instrumental classification methods utilize instrumental measurements followed by relatively simple mathematical transformations and user inferences to group molecules or assign classifications. Chapters two and three describe a machine learning classification tool, *In Silico* Fractionation, for classifying biomolecules in LC/MS data, and Chapter four describes an implementation of an instrumental classification tool for lipids in MS imaging data. The state of the field of

chemical classification in MS and relevant background for each tool presented in this dissertation will be discussed in Chapters two, three, and four.

van Krevelen diagrams. EC determination useful for identity assignment and annotation in metabolomics and lipidomics especially when using databases that link ECs with metabolite and lipid structures. However, in cases that preclude structure assignment[138], van Krevelen diagrams have utility in grouping analytes by the atomic ratios determined by their experimental EC. In its early implementation towards analysis of environmental samples, MS features from organic samples are visualized based on their H/C, O/C, and N/C ratios[10, 147]. In van Krevelen visualization, related compounds cluster in sometimes complex patterns with series of data points that are oriented linearly with different slopes; the orientation of these linear series are often due to reaction products of various chemical reactions[10]. It was found that various sources natural organic matter and fossil fuel samples could be differentiated based on their O/C and N/C ratios[10, 147]. In 2018, a more comprehensive implementation of the van Krevelen Diagram as a classification tool was demonstrated in which biological organic compounds were classified using a set of class-specific constraints on H/C, O/C, N/C, P/C, and N:P ratios[148].

Kendrick mass defect methods. Like van Krevelen diagrams, Kendrick mass defect (KMD) was first implemented as a high-resolution MS data visualization method[19]. In KMD visualization, the atomic mass unit reference is converted

from ^{12}C to $^{12}C^{1}H_2$ or a different group that is repeated in polymer chain elongation and thus eliminates mass defect contributions from that group[19]. The conversion from the measured mass (M) on the ^{12}C reference to the Kendrick mass (KM) on the $^{12}C^{1}H_2$ reference is given by

$$KM = M * \frac{14.00000}{14.01565} \quad \text{1-15}$$

(where 14.00000 is the newly defined mass of $^{12}C^{1}H_2$ and 14.01565 is the ^{12}C reference mass of $^{12}C^{1}H_2$). The calculation of KMD is then given by

$$KMD = RoundFunc(KM) - KM \quad \text{1-16}$$

where the "RoundFunc" function can be a round, floor, or ceiling function. The work in this dissertation employs a floor function to return the KMD value based on subtracting the KM value from its corresponding integer value (less than or equal KM value). Molecules that differ by numbers of repeating units that define the mass scale have the same KMD as a function of mass, which is shown by conventional KMD visualization that plots KMD as a function of nominal KM[18]. Series of "horizontal" data points in KMD visualization are indicative of chemically related polymer series that only differ with respect to the number of repeated units[18]. KMD analysis is therefore very effective in indicating groups of closely related organic molecules such as lipids that are often differentiated within classes by numbers of fatty acid chain carbons and degrees of

unsaturation[21, 23]. Notably, differences in degrees of unsaturation are readily apparent as they exhibit KMD differences of 0.01335 per unsaturation, which corresponds to the KMD of H_2. Clusters of data points can be intuitively defined by KMD and to allow users infer classifications about the whole cluster if only a fraction could be identified. In the work discussed in Chapters two and three, KMD was determined to be a relevant feature of MS peaks for the *In Silico* Fractionation neural networks to discriminate classes of biomolecules.

Chapter four of this work demonstrates an enhanced implementation of a KMD method that was adapted specifically for direct lipid classification (as opposed to analyte grouping/clustering) called "referenced KMD" (RKMD)[23]. Lipids are amenable to the RKMD workflow because most lipids are differentiated by the composition of their headgroups (e.g., phosphatidylcholine, phosphatidylethanolamine, glycerol, etc.). In RKMD analysis, the KMD of a specified lipid headgroup (i.e., the reference KMD) is subtracted from the KMD of the experimental MS peak, and the difference is divided by 0.0134; the RKMD for a given feature is given by

$$RKMD = \frac{KMD_m - KMD_{hg}}{0.0134} \qquad 1\text{-}17$$

(where KMD_m is the KMD of an MS peak and KMD_{hg} is the reference KMD of a lipid headgroup)[23]. In a theoretical case, if the experimental feature has the same headgroup as the reference KMD, the RKMD value will be equal to 0 or a

negative integer value that corresponds to the degree of unsaturation of the lipid. However, MME often precludes this theoretical case, and RKMD values approach integer values as a function of decreasing MME[23]. In cases where the class of the experimental feature does not match the reference headgroup class, RKMD values can deviate greatly from integer values, produce positive integer values, or produce negative integer values that do not correspond with degrees of unsaturation. The latter possibility introduced a non-negligible possibility for false-positive classifications, and, therefore, Lerno et al. implemented a heuristic bound on the degrees of unsaturation that would be accepted. Chapter four discusses techniques that were implemented in this work that greatly reduced the false- positive rate and increased the specificity of the presented RKMD-based method.

Supervised machine learning. Machine learning methods are designed to build models for complex and usually high dimensional patterns in data[149, 150]. In a limited number of dimensions, the human "eye" can effectively discern patterns and attribute group labels to associated data points; however, this task is made much more difficult when the dimensionality of the data exceeds those that we are able to graphically represent. Supervised machine learning methods use external help, a set of class labels associated with predictive features, to model the distribution of class labels in terms of those features[150]. The resulting model then receives the predictive feature values as input and outputs the class

label. The predictive features that a supervised machine learning model receives for training is called the "training set".[151] As the model trains, the performance of the model is tested on a separate dataset called the "test set"; performance is determined by a function that calculates the cumulative error of the model in its current state. It is important that machine learning models, including supervised types, are generalized. A generalized model approximates relationships between predictive features in the training set but does not fit them perfectly. A model that fits the training set data too well is said to be "overfit".

Chapters two and three utilize artificial feedforward neural networks (FFNN) to make predictions about analyte biological class. In general, neural networks are computational devices that mimic the structure and learning capacity of neural structures in the brain[151, 152]. FFNN are based on the early perceptron model that consists of an input layer, hidden layer, and output layer. The input and output layers receive inputs and give outputs, respectively. The hidden layer processes the inputs by weighing, summing, and submitting input values to an activation function that then produces an output[152]. The error of the output is then calculated by comparison to a target value indicating a class label, which is then used to adjust the weights applied to the inputs in the next iteration. This iterative training process continues until the error is minimized between predicted and target values. However, the FFNNs utilized in Chapters two and three utilize a multi-layer perceptron architecture with multiple inputs,

more than one hidden layer, and a variable number of nodes in each hidden layer. The more complex architecture operates with the same basic principles as the simple perceptron but is capable of modelling more complex non-linear relationships and is resistant to input outliers and noise[151].

CHAPTER TWO

Using Isotopic Envelopes and Neural Decision Tree-based *In Silico Fractionation* for Biomolecule Classification

This chapter is published as: Richardson, L. T.; Brantley, M. R.; Solouki, T., Using isotopic envelopes and neural decision tree-based in silico fractionation for biomolecule classification. Analytica Chimica Acta 2020, 1112, 34-45.

2.1 Abstract

Untargeted mass spectrometry (MS) workflows are more suitable than targeted workflows for high throughput characterization of complex biological samples. However, analysis workflows for untargeted methods are inadequate for characterization of complex samples that contain multiple classes of compounds as each chemical class might require a different type of data processing approach. To increase the feasibility of analyzing MS data for multi-class/component complex mixtures (*i.e.*, mixtures containing more than one major class of biomolecules), we developed a neural network-based approach for classification of MS data. In our *In Silico Fractionation* (iSF) approach, we utilize a neural decision tree to sequentially classify biomolecules based on their MS-detected isotopic patterns. In the presented demonstration, the neural decision tree consisted of two supervised binary classifiers utilized to positively classify polypeptides and lipids, respectively, and a third supervised network trained to classify lipids into the eight main sub-categories of lipids. The two binary classifiers assigned polypeptide and lipid experimental components with

100% sensitivity and 100% specificity; however, the 8-target classifier assigned lipids into their respective subclasses with 95% sensitivity and 99% specificity. Here, we discuss important relationships between class-specific chemical properties and MS isotopic envelopes that enable analyte classification. Moreover, we evaluate the performance characteristics of the utilized networks.

2.2 Introduction

Mass spectrometry (MS)-based profiling strategies have been developed to handle increased throughput[153-155] and sample complexity[156-158]. Combined analysis of multiple classes of biomolecule (*i.e.*, integrated/multiomics) can be advantageous and enable evaluation of correlations between different, but related, biological systems (*e.g.*, the proteome and lipidome)[159-161]. Conventionally, each class of compounds in biological samples are interrogated separately and then the results from multiple MS analyses are integrated. However, these approaches are not ideal for high throughput workflows as they often require time-intensive, off-line (as opposed to "in-line" with MS injection) sample fractionation methods. Off-line fractionation methods reduce sample complexity, prior to MS analysis, by isolating classes or groups of compounds based on their polarity, size, charge, or other physicochemical properties. In conventional analysis of biological samples, off-line sample fractionation allows researchers to make reasonable assumptions about classes of analytes to be

detected and, hence, determine the most appropriate MS operational modes and sample-specific post-acquisition data analysis workflows.

In contrast to conventional class-specific characterization of multi-component samples, simultaneous characterization of these complex samples can increase throughput and reduce biases associated with targeted sample fractionation techniques. Currently, combined sampling is complicated by biases inherent to ionization methods (*e.g.*, basicity and proton transfer kinetics [162], solvent types and ionization efficiency/proton affinity differences based on analyte polarity in electrospray ionization (ESI)[163] and matrix dependent ionization efficiencies of different classes of molecules in matrix assisted laser desorption ionization (MALDI)[76, 164-166]) and by a historical lack of in-line physical separation and sample preparation methods compatible with multi-class analyses in the literature. Recently introduced "integrated" liquid-, gas-, and solid-phase ion source technologies seek to eliminate ionization biases and accommodate samples of greater biological diversity[157, 158, 167-175]. Numerous MS-based techniques, focused on the increasingly popular field of multi-omics analyses[176-180], could directly benefit from the presented *In Silico* Fractionation (iSF) approach herein which employs a post-data acquisition tactic for analyte classifications of multi-class component samples. Despite matrix dependent ionization biases for different classes of biomolecule exhibited by MALDI matrixes, MALDI surface sampling coupled with ion mobility (IM)-MS has been used for simultaneous analysis of multi-class mixtures of small biomolecules

directly from tissue[181-183]; iSF could be applied to such MALDI generated complex data sets. Moreover, a preliminary demonstration of Omni-MS showed a sample preparation and separation strategy for concurrent LC-MS analysis of electrolytes, small molecules, lipids, polypeptides, nucleic acids, and polysaccharides[184]. We expect that technologies for simultaneous, multi-omics analyses will continue to develop in pursuit of higher throughput and information density per data acquisition. However, downstream analysis of such multi-class acquisitions will remain a challenge given that: (a) assumptions regarding the class of detected ions may not be reliable, and (b) different classes may have different requisite MS operational modes and/or analytical workflows. Thus, multi-omics MS analyses require some method for discriminating between (and identifying) classes of analytes. We have previously shown that MS isotopic envelope of molecular ion signals contain sufficient information to discriminate between compounds containing different functional groups or specific elements [185]; here, we aim to demonstrate the strength of combining neural networks and isotopic pattern-based biomolecule class designation for multi-omics characterization of multi-class sample mixtures.

The isotopic envelope (*i.e., m/z* range containing all isotopologues of a specific compound) is a readily observed feature in high resolution mass spectra[186]. In theory, each elemental composition has a unique MS isotopic fine structure[187] that is dependent on the mass contributions of each element and relative abundances of heavy isotopes. Given that chemical homology is

conserved (often to different degrees) within classes of biomolecules, the elemental compositions, and thus the MS isotopic envelopes, of biomolecules reflect this similarity. This idea of intra-class chemical homology is at least tacitly understood in untargeted MS experiments that utilize a combination of sub-ppm error (exact mass measurement assignments for presumably resolved peaks with > 30,000 mass resolving power (MRP) and isotopic envelope information to generate compound elemental compositions and cluster analytes roughly by class in van Krevelen visualization (though without means for definitive classification)[10, 188]. However, high (H) and ultrahigh (UH) MRP MS instruments often lack the necessary MRP and mass measurement accuracy to consistently yield accurate elemental compositions for large molecules (molecular weight (MW) > 500 Da) containing elements beyond carbon, hydrogen, nitrogen, and oxygen[188, 189]. Moreover, UHMRP FT-MS instruments (> 250,000) are unfavorable for high throughput applications in which high scan rate, HMRP instruments (i.e., time-of-flight (TOF) instruments with ~10,000-100,000 MRP) are preferred. *A priori* knowledge of analyte class is often necessary to determine the elements allowed for use in the generation of elemental compositions in order to shorten the often-lengthy list of possible elemental compositions produced within a margin of experimental error[190]. Without *a priori* knowledge, determination of the elemental composition becomes exponentially more difficult due to the possibility of having myriad different elemental compositions that can yield the measured mass and isotopic

envelope[126]. As expected, the length of the produced "list" of elemental compositions generally decreases with increased MRP, resulting in higher confidence compound identifications and/or classifications; however, it is often impossible to reduce the number of possibilities to a single, high confidence determination[126]. Elemental composition, and therefore compound class, could theoretically be determined via computationally intensive solutions to the polynomial model that describes the summation of elemental contributions[191] to isotopologue peaks given error-free conditions. In general, the effect of non-ideal experimental conditions on exact mass determinations of elemental composition, and inapplicability of UHMRP MS instrumentation in high throughput applications necessitate an automated approach for definitive analyte classification in multi-omics analyses that is less dependent on MRP through utilization of other features in the isotopic envelope.

Feedforward neural networks (FFNNs, described by Bishop[192]) offer a potential solution to this problem. Trained FFNNs excel in estimating non-linear, high dimensional relationships to discern hidden patterns and classify network inputs. FFNNs are based on the early perceptron model in which network inputs are recursively weighted, summed, and submitted to an activation function; in this simple scenario, training stops once the output of an activation function exceeds a specified value, producing a linear function that serves as a binary classifier[193]. FFNNs utilize an expanded hidden perceptron layer architecture to more effectively map complex systems and solve problems

with high dimensionality (*i.e.*, those with large sets of inputs and multiple output functions)[194]. We hypothesized that FFNNs could effectively estimate the non-linear relationships that describe the features of the MS isotopic envelope to discriminate between classes of biomolecules in MS data (without the painstaking task of mathematically accounting for the discrete contributions of each element). To test our hypothesis, we examined both theoretical and experimentally acquired MS data. Additionally, given a sufficiently large and varied training set and an appropriate network architecture, we hypothesized that FFNNs could be sufficiently generalized such that experimental error and noise sources present in mass spectral data would not pose a serious concern[195]. In this paper, we confirm the effectiveness of FFNNs for biomolecule class identification in multi-omics analyses and its tolerance for handling experimental measurement errors, common in HMRP MS data, that can impede compound identification, elemental composition determination, and/or classification by exact mass measurement.

Specifically, we present an FFNN-based chemometric analysis tool for classification of small biomolecules (*e.g.*, lipids, polypeptides, nucleic acids, glycans) by utilizing their commonly acquired soft ionization MS data produced from TOF instrumentation. We show that a series of FFNNs can be trained using features extracted from the centroided isotopic envelopes (*i.e.*, the mass-to-charge (*m/z)* and relative isotopologue peak intensities/ratios) to classify individual sample components into a selection of biomolecular class targets.

Using these trained FFNNs, we show that a neural decision tree (NDT)[196] can be constructed to utilize MS data sets and sequentially classify analytes in a representative multi-class component sample containing polypeptides, metabolites, and lipids. Moreover, we show that classified lipids can be assigned to their respective subclasses and demonstrate how this technique can provide simple, qualitative results that can be interpreted by non-MS experts. In addition to classification of theoretically generated MS data, we confirm the validity of *In Silico* Fractionation (iSF) by using experimental data and successfully identifying its components. We also show that network classification operations can be rapid and comparable to time-of-flight (TOF) MS detection event time scales[197]. We provide an analysis and a discussion of the chemical basis for the iSF technique and its adaptability to various types of MS methodologies.

2.3 Materials and Methods

2.3.1 Neural Network Input Preprocessing

To generate theoretical *m/z* values and isotopic patterns for selected classes of biomolecules (*viz.*, lipids, glycans, non-lipid metabolites, polypeptides, DNA, and RNA), elemental compositions of a total of 36,973 molecules were acquired from four different sources as indicated in the following text. Elemental compositions for these 36,973 molecules were sourced from (1) the LIPID MAPS Structural Database (lipids, n = 7,473) [198], (2) the

Consortium of Functional Glycomics (glycans, n = 1,750), and (3) the Human Metabolome Database (organic, non-lipid metabolites, n = 7,454) [199]. Lastly, polypeptide (n = 7,465), deoxyribonucleic acid (DNA, n = 7,140), and ribonucleic acid (RNA, n = 5,691) elemental compositions were generated by (4) an in-house written python (CPython 3.6.2; Python Software Foundation, DE) script that pseudo-randomly generated heteropolymer sequences corresponding to expected ESI charge states based on biopolymer chain lengths [200] and converted them to elemental compositions (see Appendix A.1). For example, polypeptides in the +1 charge state result from a gaussian-like distribution of polymer chain lengths between about 5 and 17 residues as demonstrated by Xie et al. [200]. Therefore, the chain lengths for singly-charged, protonated polypeptides were generated by a gaussian probability function with bounds to match the experimentally observed distribution. The centroided isotopic distribution profiles for some protonated, deprotonated, and sodiated ESI adducts and charge states (respective to the class of biomolecule) were calculated in the enviPat 2.2 package [201] in R (ver. 3.4.3; R Foundation for Statistical Computing). The molecular adduct forms and charge states included in the training set were not inclusive of all commonly observed types; the training set was limited in this regard due to deteriorating network training performance as the inclusion of multiple forms presumably added extraneous dimensionality to the FFNN training set. Isotopic profiles were generated considering respective physical realities of ion production in ESI for each class as follows: polypeptides as

protonated ions ranging from $[M+H]^+$ to $[M+6H]^{6+}$, glycans as positively charged, sodiated ions $[M+Na]^+$ and $[M+2Na]^{2+}$ and as negatively charged, deprotonated ions ranging from $[M-H]^-$ to $[M-3H]^{3-}$, lipids as positively charged, protonated ions $[M+H]^+$, and oligonucleotides as negatively charged, deprotonated ions ranging from $[M-H]^-$ to $[M-6H]^{6-}$.

2.3.2 Neural Network Training

FFNNs were constructed with a custom script (see Appendix A.2) written in M (MATLAB R2018a; The MathWorks Inc., Natick, MA). Three FFNNs, denoted as PEPNET, LIPNET, and LIPSUBNET, were used in this study. PEPNET was a binary classifier to assign network inputs into the peptide (T_P) or non-peptide (T_{NP}) target class. Whereas LIPNET was a binary classifier to assign network inputs into the lipid (T_L) or non-lipid (T_{NL}) target class; for lipid subclass categorization, LIPSUBNET (an 8-target classifier) was used to assign previously classified T_L inputs into class targets corresponding to lipid subclasses. Designated lipid subclasses in LIPSUBNET included: fatty acyls (T_{FA}), glycerolipids (T_{GL}), glycerophospholipids (T_{GP}), polyketides (T_{PK}), prenol lipids (T_{PR}), saccharolipids (T_{SL}), sphingolipids (T_{SP}), and sterol lipids (T_{ST}).

The hidden layer architecture (i.e., number of hidden layers and number of nodes per hidden layer) for each type of FFNN was chosen such that classification accuracy for the input data set used for training would be maximized. For FFNNs with more than one hidden layer, the number of nodes

per layer were kept the same across layers as networks already exhibited high performance and further optimization was unnecessary. Training times were short enough that they were not considered in determining network architecture. 10 FFNNs were trained for each tested network architecture, and performance was evaluated by the mean percent error (percent of false positive and false negative classifications over all classifications). Network architectures were chosen for the minimization of mean percent error (for more information regarding network architecture optimization, see Appendix A.3 and A.4). When training the PEPNET and LIPNET binary classifiers, it became apparent that small changes in hidden layer sizes had notable effects on performance; therefore, hidden layer neurons were increased by 10 during optimization tests. The optimized hidden layer architecture of LIPNET and PEPNET was constituted by 2 layers with 30 perceptrons each. For the 8-target classifier (LIPSUBNET), network performance was relatively insensitive to small changes in hidden layer neurons; thus, hidden layer neurons were increased by the power function 2^n to evaluate a large range of hidden layer neurons. The optimized hidden layer architecture was constituted by 2 layers with 64 perceptrons each. Training bias due to uneven class representation was accounted for by using the "growth method" (Brantley et al. [202]) such that each class had equal numbers of representations in the dataset. The input data set was then randomly divided (*i.e.*, MATLAB function "dividerand") for network training as follows: 70% in the training set, 15% in the validation set, and 15% in the test set. Networks were

trained using scaled conjugate gradient backpropagation (*i.e.,* MATLAB function "trainscg") with an early stopping method [203] to avoid overfitting. Performance during training was monitored using cross-entropy as an error metric (MATLAB function "crossentropy"). Network inputs were 7 numbers (henceforth referred to as "input vectors") in the following order: (1) the exact mass monoisotopic *m/z* value to four decimal places (to be consistent with TOF data utilized in this demonstration) corresponding to the first isotopic peak or "A" (where "A" notation is based on the designation introduced by McLafferty et al. [186]), (2) the intensity of "A" normalized to the most abundant isotopologue (i.e., normalized to the highest peak within the isotopic envelope for the molecular ion designated as 100% relative abundance), (3) the relative intensity for the second isotopologue or A+1 [186] , (4) the relative intensity for the third isotopologue or A+2, (5) the ratio of A/A+1, (6) the ratio of A+1/A+2, and (7) the Kendrick mass defect [204] (KMD, relative to CH_2 integer mass scale) of the monoisotopic *m/z*. The output layers were limited to two (for PEPNET and LIPNET) or eight (for LIPSUBNET) nodes corresponding to the classification targets of each neural network. The PEPNET output layer nodes correspond to: 1) polypeptide and 2) non-polypeptide. The LIPNET output layer nodes correspond to: 1) lipid and 2) non-lipid. The LIPSUBNET output layer nodes correspond to: 1) fatty acyl, 2) glycerolipid, 3) glycerophospholipid, 4) polyketide, 5) prenol lipid, 6) saccharolipid, 7) sphingolipid, and 8) sterol lipid.

Supervised training target outputs were provided such that the presence of a class of biological molecules was indicated by a value of 1.0 and absence by a value of 0.0. The same network inputs were used for each FFNN, but the supervised training target outputs were changed depending on FFNN being trained (e.g., in PEPNET, polypeptide components had target outputs of 1.0 and lipid/metabolite/glycan/DNA components had target outputs of 0.0; the same components were used in LIPNET, but polypeptide component target outputs were changed to 0.0, and lipid component target outputs were changed to 1.0). Additionally, LIPSUBNET was trained with only lipid components. For each scored component (in either the training set or an experimental test set), the maximum predicted value from the set of values corresponding with each class predicted by the network is taken as the predicted class. Ten FFNNs were trained for each type of network (i.e., PEPNET, LIPNET, and LIPSUBNET) with the optimal hidden layer architectures, and the network with the minimum mean percent error across the training, validation, and test datasets was selected as the best performing network for further analyses. All networks were trained using a desktop computer (OptiPlex 7050 Tower; Dell Computer, Round Rock, TX) equipped with a 4-core processor (Intel® Core™ i5-7500) and 16 GB of RAM.

2.3.3 Mass Spectrometry Methods and Multi-Class Component Test Set Generation

To test the iSF approach on experimental data, a multi-class component test set was generated. *Firstly*, an LC-MS output matrix (rows and columns corresponding to LC scan numbers and *m/z* axis data points, respectively) from a rat brain tryptic protein digest analysis (collected in-house for *m/z* range 50 to 2000; see below for MS methodology) was added to LC-MS output matrix (also, 50 to 2000 *m/z* range) from a lipidomics LC-MS output. The lipidomics data was downloaded from the Chorus Project (Stratus Biosciences, Seattle, WA) mass spectrometry file sharing database [205]. Available lipidomics data from Chorus [205] included a limited range of LC elution time and hence, only scans corresponding to ~15 minutes (scan numbers up to ~900, at an acquisition rate of 1 Hz for both lipidomics and proteomics) was included for further analysis (e.g., Figure 5 shows analysis results for scan range of ~400 to 900). Matrix addition was performed by the MassLynx (V4.1, Waters, Milford, MA) "Combine all functions" tool. *Secondly*, LC eluting components were found in the combined data set and manually (on a random order) were selected for inclusion in the test set. The data features necessary for the neural network were exported from the MassLynx mass spectrum data viewer. *Thirdly,* only the eluting components that could be identified were include so that their true chemical classes would be known for the evaluation of FFNN. Components that were identified as peptides were sequenced and identified in ProteinLynx Global Server (PLGS, Waters Corp., Milford, MA); the LC retention time recorded by PLGS was confirmed for

each peptide. Components that were identified as lipids and metabolites were identified in Progenesis QI (Nonlinear Dynamics, Durham, NC) by exact mass matching (< 5 ppm mass measurement error) in the LIPID MAPS and Metlin databases, respectively. Because it was necessary to confirm to which group (lipidomics or proteomics data set) each component belonged, identified lipid components were confirmed to be absent in the original proteomics data set. Likewise, identified peptide components were confirmed to be absent in the original lipidomics data set.

The downloaded LC-MS lipidomics data set was acquired using a Waters Xevo-G2 QToF (Waters Corp., Milford, MA) in positive-ion mode [205]. To collect the rat brain protein tryptic digest data, ultra-performance liquid chromatography (UPLC)-ion mobility-enhanced MS^E ($HDMS^E$) on a Synapt G2-S HDMS (Waters Corp., Milford, MA) operating in positive-ion mode was used. Pettit et al. previously described the procedure for rat brain tissue sample preparation [206]. A 10 mg rat brain cortex tissue section was homogenized via ultrasonication probe in lysis solution (100 mM ammonium bicarbonate, 1% (w/v%) sodium dodecyl sulfate, 10 mM tris(2-carboxyethyl) phosphine hydrochloride, and 40 mM 2-chloroacetamide; all chemicals from Fisher Scientific, NH). Bottom-up proteomics samples were prepared according to the filter-aided sample preparation protocol [207], which involves sequential wash and high mass (> 3 kDa) filtration (MilliporeSigma, MA) that washes lipids and detergents from the sample. Samples were digested with sequencing-grade

trypsin (Promega, Madison, WI) overnight at 37°C. PLGS was employed for peptide identification of the tryptic protein digests. The lipidomics and proteomics MS data were both acquired at a mass resolution of ~22,000 (see Appendix A.5 for exemplary peptide and lipid isotopic envelopes from the combined dataset with mass resolution calculations).

HeLa digest peptides and a rat brain lipid extract were mixed and analyzed via UPLC-IM-MS on a Synapt G2-S HMDS operating in positive-ion mode. The HeLa digest peptides were purchased as a standard from Thermo Fisher Scientific (Waltham, MA). A 10 mg rat brain tissue section was homogenized via ultrasonication in ice-cold 0.1% ammonium acetate and prepared as described by Matyash et al. [208] The HeLa digest peptide mixture was reconstituted to a concentration of 60 ng/µL in 0.1% formic acid in water, and the lipid extract was reconstituted to a final volume of 100 µL in an isopropanol/acetonitrile/water (4:3:1, v/v/v) solution. Equivalent volumes of each mixture were mixed, and a volume of 5 µL was injected on column (150 ng of peptides and lipids extracted from 250 µg of rat brain tissue). The mixture was separated by reversed-phase chromatography on the NanoAquity UPLC using a Symmetry C18 trap column (5 µm, 180 µm x 20 mm, Waters Corp., Milford, MA) and a BEH130 C18 analytical column (1.8 µm, 100 µm x 100 mm, Waters Corp., Milford, MA). Analytes were separated with a binary solvent gradient with 10 mM ammonium formate and 0.1% formic acid in water (mobile phase A) and isopropanol/acetonitrile (90:10, v/v) with 10 mM ammonium formate and 0.1%

formic acid. The gradient ramped linearly from 5% to 99% B over 70 minutes with an isocratic hold at 99% B for 4 minutes at 0.4 µL/min. Capillary voltage was set at 2.7 kV, and the source temperature was set at 100 °C. Travelling wave ion mobility conditions were set at default settings. Analytes for classification were selected from three time periods of the chromatogram: peptides from 10-30 minutes, lipids from 55-80 minutes, and unknowns from 30-55 minutes. The time periods were selected based on an understanding peptide and lipid elution behavior in reversed-phase separations. The 10-30 minute period roughly corresponds to 13-40% organic solvent composition, which is used for reversed-phase peptide separations [206]. The 55-80 minute period roughly corresponds to 70-99% organic solvent composition, which is common for reversed-phase lipid separations [209]. Selected analytes were limited to those with monoisotopic *m/z* values between 500-1200 *m/z*. Before 10 minutes, only unretained components and background signal were observed, and, between 80-90 minutes, only strong polymer signal was detected, and therefore these periods were excluded.

2.4 Results and Discussion

The following sections describe the generation and implementation of the iSF workflow; a neural decision tree-based method is introduced that utilizes MS data for the classification of small molecules. The physical and MS principles that constitute the basis for iSF as well as feature selections for neural network training are discussed in detail. Additionally, examples of iSF analyses

of an experimental LC-MS datasets containing multiple biological classes of molecule is provided.

2.4.1 MS Isotopic Envelope Feature Selection

Classes of biomolecules, such as polypeptides and nucleic acids, consist of biopolymers that are composed of monomeric subunits and are limited in elemental diversity for a given polymer length by the number of possible monomers. Contrarily, biomolecules such as lipids are predominantly non-polymeric and are more structurally diverse. Figure 2.1 displays the histogram distributions for elemental compositions of nucleic acids (purple), glycans (green), polypeptides (blue), and lipids (red) as a function of their respective mass percent composition (w/w%) values for hydrogen (Fig. 2.1a), carbon (Fig. 2.1b), nitrogen (Fig. 2.1c), and oxygen (Fig. 2.1d). For example, lipids are hydrocarbon-rich and thus the mass percent compositions for carbon and hydrogen are in the ranges of ~35-90% (Fig 2.1b) and ~3-14% (Fig. 2.1a), respectively. Additionally, lipids are substituted with a wide variety of polar functional groups and hence they are composed of ~0-14% nitrogen (w/w%, Fig. 2.1c) and ~4-52% oxygen (w/w%, Fig. 2.1d), respectively.

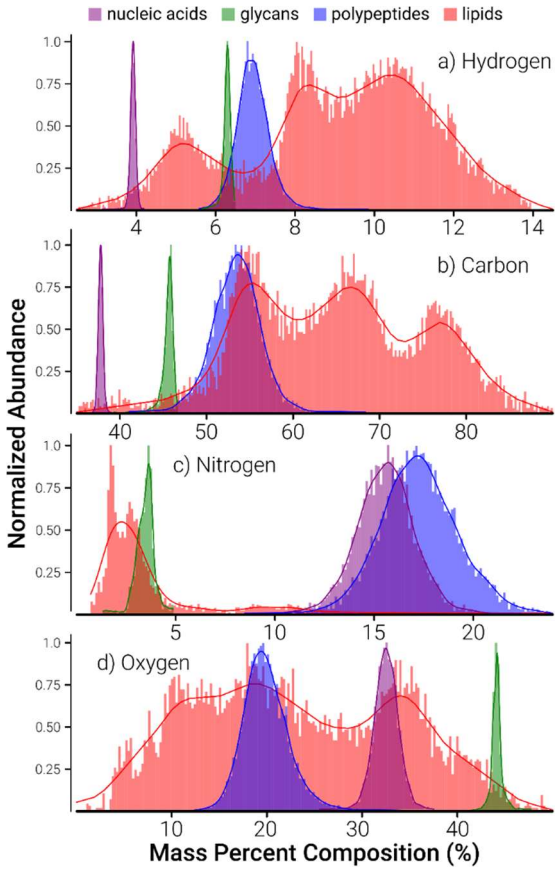


Figure 2.1. Histograms of nucleic acid (purple), glycan (green), polypeptide (blue), and lipid (red) compositions as a function of mass percent composition of (a) hydrogen, (b) carbon, (c) nitrogen, and (d) oxygen overlaid with a kernel density estimation line plot. The nucleic acid, polypeptide, and glycan class distributions are generally gaussian-like and overlap minimally, except for those for nitrogen mass percent composition in which there is significant overlap. Lipid species are generally distributed over a wide range of elemental mass compositions. Elemental composition of a significant number of the selected lipids lacked nitrogen (~37%), and hence the nitrogen histogram for lipids (which includes contributions from 63% of all lipids) does not display contributions from those species that contained no nitrogen (i.e., 37% of the lipids used to generate (c) contained 0% nitrogen).

Conversely, due to the constrained modes of elongation and incorporation of a limited number of possible monomers into their structures, polypeptides, nucleic acids, and glycans have narrower distributions of carbon (w/w%) and nitrogen (w/w%) contents; additionally, the distribution of oxygen composition in glycans is also very confined (green histogram in Fig. 2.1d). In fact, both nucleic acids and glycans have two distributions that are near completely

unconvolved with any other distribution (for nucleic acids, hydrogen (purple histogram in Fig. 2.1a) and carbon (purple histogram in Fig. 2.1b) and, for glycans, carbon (green histogram in Fig. 2.1b) and oxygen (green histogram in Fig. 2.1d)). The carbon w/w% of polypeptides range from ~41-62% (blue histogram in Fig. 2.1a), though slightly overlapping with that of glycans, this range is completely distinguished from the range of carbon w/w% for nucleic acids (~37-39%). Interestingly, an analytical tool for classification tasks in UHMRP MS data, called van Krevelen diagram, visualizes and groups detected compounds using molar ratios (e.g., the ratio of H/C and O/C atoms) derived from their calculated elemental compositions [10]. Given that elemental molar composition and mass composition are immediately related by elemental molar mass, classifications made by van Krevelen diagrams inherently utilize the described class distributions (Fig. 2.1). Given that van Krevelen diagram visualization requires ultrahigh mass resolving power FT-MS instruments (generally > 200,000 resolving power) to generate high confidence chemical compositions from exact mass measurements of monoisotopologue peaks, the technique is generally unviable for conventional HMRP instruments. However, the mass percent compositions of compounds also affect the MS isotopic pattern, which is readily resolved by conventional HMRP instrumentation and provides a wealth of class-specific information.

Another important intrinsic property of molecules that has often been utilized in mass spectrometry, for both small [186] and large molecule [210]

assignments, is the molecular isotopic pattern. Heavy isotopes of a given element uniquely contribute to heavy isotopologue peak intensities as molecular masses of analytes increase. For example, the second most abundant isotopes of carbon and nitrogen are ^{13}C (1.1% natural abundance (NA)) and ^{15}N (0.36% NA), respectively, which are ~1 Da heavier than their corresponding most abundant counterparts (^{12}C and ^{14}N). Hence, presence of either ^{13}C or ^{15}N increases the molecular mass of an ion by one atomic mass unit (*i.e.,* designated as A+1 elements that contribute to A+1 peak in a mass spectrum). Because of the larger NA contribution from ^{13}C (i.e., 1.1%) than ^{15}N (i.e., 0.36%) as well as greater mass proficiency (0.003355 for ^{13}C and 0.000109 for ^{15}N), contributions to average molecular weights are larger from ^{13}C isotopes than ^{15}N isotopes; moreover, biomolecules generally contain larger numbers of carbon atoms than nitrogen atoms and hence A+1 peaks (mostly contributions from ^{13}C) are often informally referred to as ^{13}C isotope peaks. It should be noted that, when present, several other isotopes can also contribute to A+1 peak (*e.g.*, ^{2}H, ^{17}O, *etc.*); UHMRP required to resolve these "fine structures" within the observed isotopic envelopes is beyond the reach of conventional mass spectrometers [137, 187, 211]. Therefore, observed A+1 peaks in mass spectra, that might be from a combination of various isotopes, are often unresolved. Similar to A+1 elements (*e.g.*, ^{13}C, ^{14}N, and ^{2}H), the second most abundant isotopes contribute to the observed relative abundance of A+2 peaks. For instance, oxygen and sulfur (*i.e.*, ^{18}O (0.21% NA) and ^{34}S (4.3% NA)) are ~2 Da heavier (*i.e.,* A+2 elements) than

their corresponding most abundant counterparts (^{16}O and ^{32}S) and thus contribute to A+2 peak. It should be noted that there are several other isotopes and numerous other combinations (*e.g.*, $^{2}H_2$, $^{13}C_2$, $^{15}N_2$, $^{2}H_1 + {}^{13}C_1$, $^{2}H_1 + {}^{15}N_1$, $^{13}C_1 + {}^{15}N_1$, *etc.*) that can also contribute to A+2 and other higher mass isotopologues (*e.g.*, A + 3, A + 4, …., A + n, where n is the number of observed peaks within an isotopic envelope for an analyte). Therefore, the relative contributions of each element to the mass of a compound affects isotopologue peak relative intensities in the mass spectrum. Figure 2.2 displays logarithmic plots of (A+1/A+2) peaks for nucleic acid, glycan, polypeptide, and lipid biomolecules as a function of their neutral monoisotopic molecular (or A) masses. Given the relative contributions of different heavy isotopes to A+1 and A+2, the ratio A+1/A+2 indirectly relates the quantities of carbon, nitrogen, hydrogen, and other A+1 elements to oxygen, sulfur, and other A+2 elements.

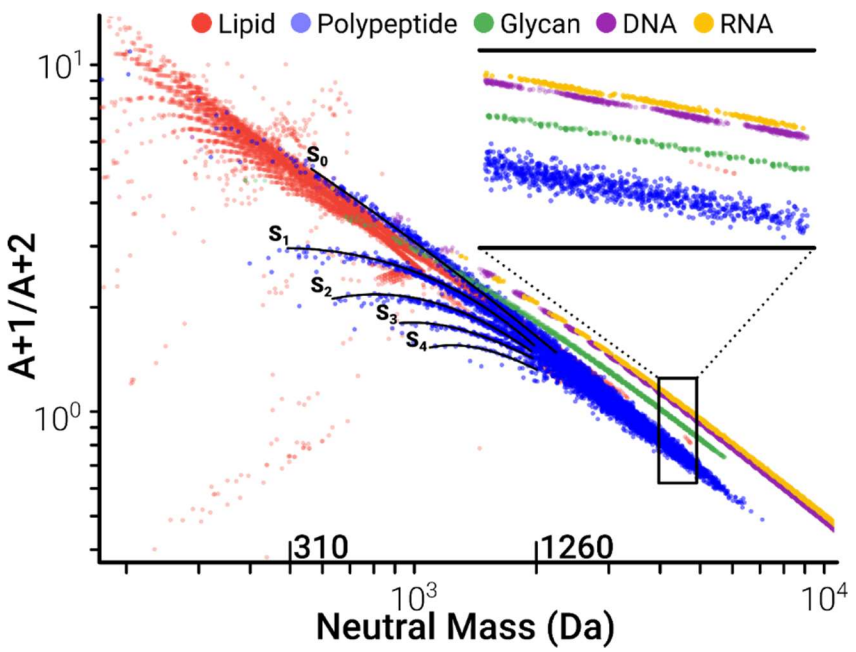


Figure 2.2. Log-log plots for theoretical MS isotopologue peak ratio, A+1/A+2, as a function of compound neutral mass (Da), show high degrees of separation between classes of biomolecules. Predominantly linear trends for each class show general correlations between the isotope ratios and neutral masses. Polypeptide sulfur content is also distinguished between the 310 to 1260 Da mass range as A+1/A+2 decreases with each sulfur atom inclusion (shown by blue trendlines labeled S_0-S_4). For the sake of figure clarity, the lowest and highest mass compounds are omitted.

The isotopic ratio plot (*i.e.*, A+1/A+2 as a function of neutral mass) presented in Figure 2.2 provides one example (of many possible ways) of how such visual displays can be used for biomolecule class differentiation; multidimensional raw data used to train FFNNs allow access to numerous other A+1/A+2 type plots (ratios of other isotopic peaks) and "hidden" relationships that are not easily discernable by using two-dimensional plots. For this discussion, the logarithmic plot of A+1/A+2 as a function of compounds' neutral mass (Figure 2.2) is helpful in so far as A+1/A+2 is influenced by both A+1 and A+2 elements. As expected, amplitude of the A+1/A+2 ratio decreases with increasing neutral mass up to certain values for different compounds- or a chemical class-dependent value.

The initial decrease is due to the generally increased presence of A+2 elements (*i.e.*, oxygen and sulfur) and the probability of larger molecules containing multiple A+1 element heavy isotopes (*e.g.*, $^{2}H_{2}$, $^{13}C_{2}$, $^{15}N_{2}$, $^{2}H_{1} + {}^{13}C_{1}$, $^{2}H_{1} + {}^{15}N_{1}$, $^{13}C_{1} + {}^{15}N_{1}$, *etc.*). Observed biomolecule class-dependent trends vary as a function of elemental mass composition. The top right inset in Figure 2.2 shows the expanded region from ~2000-2500 Da; although the parallel trends for different compound classes are close to each other, they are fully separated and correlate to the differences in mass percent compositions observed in Figure 2.1 that can be used for class differentiation. For instance, the ribonucleic acid (RNA, yellow) trendline is slightly higher than deoxyribonucleic acid (DNA, purple) because of the loss of CH_2 from ~25% of RNA residues (thymine exchanged with uracil). However, the magnitude of the RNA shift is reduced by the gain of oxygen at the ribose 2' position. From ~310-1260 Da (indicated number labels on the x-axis of Figure 2.2), the sulfur content of the polypeptides (Figure 2.2, blue) can be visually distinguished by the divergent trendlines (labeled S0-S4) due to the major contribution of ^{34}S (4.21% relative abundance) to A+2 [185]. The polypeptides along the S_0 trend contain no sulfur, and each descending trend (S_1-S_4) incorporates collections of polypeptides that contain an additional sulfur atom. Also, it should be noted that because of the polypeptide diversity (polymer growth possibilities to yield similar MW as compared to limited polymer growth possibilities for glycan, DNA, or RNA classes) and possibility of having various elemental compositions yielding similar polypeptide masses, the blue trendline

in the expanded region of Figure 2.2 is wider (in the y-axis, corresponding to (A+1/A+2) values) than the counterpart trendlines for glycans, DNAs, and RNAs. Such inter- and intra-class variations of isotopic patterns might be difficult to discern without the use of neural networks that often capitalize on "hidden" relationships for class discriminations.

The applicability of isotopic envelope features in their use by FFNNs can be limited by both instrumental and analyte-specific constraints. Firstly, the isotopic envelope was centroided and each peak integrated to reduce the complexity of the FFNN input layer and eliminate the effects of variance of resolution across MS instrumentation. The centroided monoisotopic *m/z* value (input 1) was chosen for the high correlation of *m/z* with isotopic ratios in organic compounds (Figure 2.1). However, with respect to macromolecules such as proteins that contain a large number of carbons, the monoisotopologue peak ($^{12}C_{all}$) relative intensity (input 2) is often very low (*i.e.*, the probability of having a large protein molecule with all of its carbons as ^{12}C) and often below the instrument detection limits. In our approach, the proper classification of biomolecules with FFNNs is dependent on successful identification and measurements of the monoisotopologue MS peak. Thus, application of the current approach to experimental data is limited to species with observable monoisotopologue peaks. Likewise, the measurement of the isotopologue peak relative intensity inputs, A (input 2), A+1 (input 3), and A+2 (input 4), and thus the calculated isotopic ratios, A/A+1 (input 5) and A+1/A+2 (input 6), are subject

to the same limitations of instrumental sensitivity and MRP. It could be argued that the calculated isotopic ratios (inputs 5 and 6) are unnecessary since the FFNN training could "discover" those relationships; however, the inclusion of these inputs improved FFNN training performance in all cases. As the fourth (and onwards) isotopologue peak(s) are often below instrument detection limits for low mass (and low abundance) analytes in ESI-MS experiments, these peak intensities were not included as inputs in our specific FFNN training sets. However, depending on the molecular weight (or m/z) ranges of interest, it is possible to utilize signals from other isotopologues (e.g., other relatively high abundance species) as inputs. In other words, two important criteria in considering a particular set of isotopologues to select for FFNN training sets is (a) relative abundance consideration (signal-to-noise consideration for MS detectability) and (b) computational cost (e.g., additional input may not necessarily improve the success rate sufficiently large but require unreasonably large computational times). The selection of three isotopologues in the presented study provided a good balance between the desired success rate and computational cost. The calculation of KMD (input 7) is also contingent on the detection and mass measurement accuracy of the monoisotopic *m/z*. The KMD input (7) was included for its ability to assist in MS ion classification challenges and for its contribution to network training performance improvements [18]. Within the context of small molecule analysis (*e.g.*, metabolomics, lipidomics,

and peptidomics), the above features are readily distinguished in most modern TOF-MS workflows.

2.4.2 FFNN Training and Performance

For PEPNET, the "non-polypeptide" portion of the supervised training set consisted of lipid, glycan, polar metabolite, and nucleic acid input vectors (i.e., one-dimensional data arrays containing network input values; N_{Total} = 23,816 post-growth method; N_{Train} = 16,673 (70%); N_{Test} = 3,543 (15%); $N_{Validation}$ = 3,600 (15%)). The "polypeptide" portion of the supervised training set consisted of polypeptide input vectors (N_{Total} = 23,816 post-growth method; N_{Train} = 16,669 (70%); N_{Test} = 3,602 (15%); $N_{Validation}$ = 3,545 (15%)). PEPNET had a consistent mean percent error of 7.2% for each training, validation, and test datasets; in other words, PEPNET incorrectly classified 7.2% of inputs submitted in the training set. For the supervised training test set, T_P and T_{NP} had true positive rates (TPR, percentage of actual positives classified as such) of 94.2% and 91.4%, respectively, and positive predictive values (PPV, percentage of actual positives in all predicted positives) of 91.7% and 93.9%, respectively. For LIPNET, the "non-lipid" portion of the supervised training set consisted of polypeptide, glycan, polar metabolite, and nucleic acid input vectors (N_{Total} = 23,809 post-growth method; N_{Train} = 16,712 (70%); N_{Test} = 3,552 (15%); $N_{Validation}$ = 3,545 (15%)). The "lipid" portion of the supervised training set consisted of lipid input vectors (N_{Total} = 23,809 post-growth method; N_{Train} = 16,620 (70%); N_{Test} = 3,591 (15%);

$N_{Validation}$ = 3,598 (15%)). The LIPNET training, validation, and test datasets had 12.1%, 12.6%, and 12.1% overall percent errors, respectively. T_L and T_{NL} had TPRs of 92.1% and 83.7%, respectively, and PPVs of 85.1% and 91.3%, respectively, for the supervised training test set. The training and validation confusion matrices [212] for PEPNET and LIPNET varied by ≤ 1%, suggesting that the networks were not overfit (for more information regarding PEPNET and LIPNET training performance metrics, see Appendix A.6 and A.7). PEPNET performed slightly worse than LIPNET presumably to the narrower variety of possible peptide chemical compositions as a function of mass relative to lipids, which exhibit great structural and compositional variety. LIPSUBNET was an 8-target classification network trained to classify previously-assigned T_L input vectors into T_{FA}, T_{GL}, T_{GP}, T_{PK}, T_{PR}, T_{SP}, T_{SL}, and T_{ST} target classes (*i.e.*, a comprehensive set of targets corresponding to all lipid subclasses). The LIPSUBNET training, validation, and test datasets had 7.9%, 8.1%, and 8.2% overall percent error, respectively. Respective TPR and PPV values for each target and predicted class is provided in a confusion matrix diagram in Appendix A.8.

Networks trained with multiple output nodes (e.g., LIPSUBNET) do not exhibit equal TPR and PPV values across each output node. Therefore, we utilized a parameter designated as “confidence threshold” to reduce variance in network performance characteristics (i.e., TPR and PPV) across network output nodes and improve the accuracy of the trained networks. Confidence threshold,

a user-defined scalar value from 0.0 to 1.0, constrained “confident” (unassociated with confidence interval) identifications to those that result from network scores higher than the confidence threshold value. Classifications with network scores less than the user-defined confidence threshold are removed from consideration in calculation of network performance characteristics. For example, with a defined confidence threshold of 0.7, a classification that resulted from a score of less than 0.7 would be excluded; conversely, a classification of that resulted from a score of 0.7 or higher would be included in calculation of network performance characteristics. Figure 2.3 displays the TPR of LIPSUBNET for all unique training set inputs as a function of confidence threshold. The TPR at a 0.0 confidence threshold is not uniform across class targets, ranging from 77.9% (prenol lipids, Figure 2.3, trace number 8 in blue-green) to 98.6% (saccharolipids, Figure 2.3, trace number 1 in light green).

The low performance of prenol lipid classification can be attributed to relatively high chemical structural similarity to the likes of sterol lipids (Figure 2.3, trace number 7 in orange), both of which share a common biosynthetic pathway [213], and, to a lesser extent, fatty acyls (Figure 2.3, trace number 6 in red) that are likewise mostly comprised of minimally branched hydrocarbon chains[213].

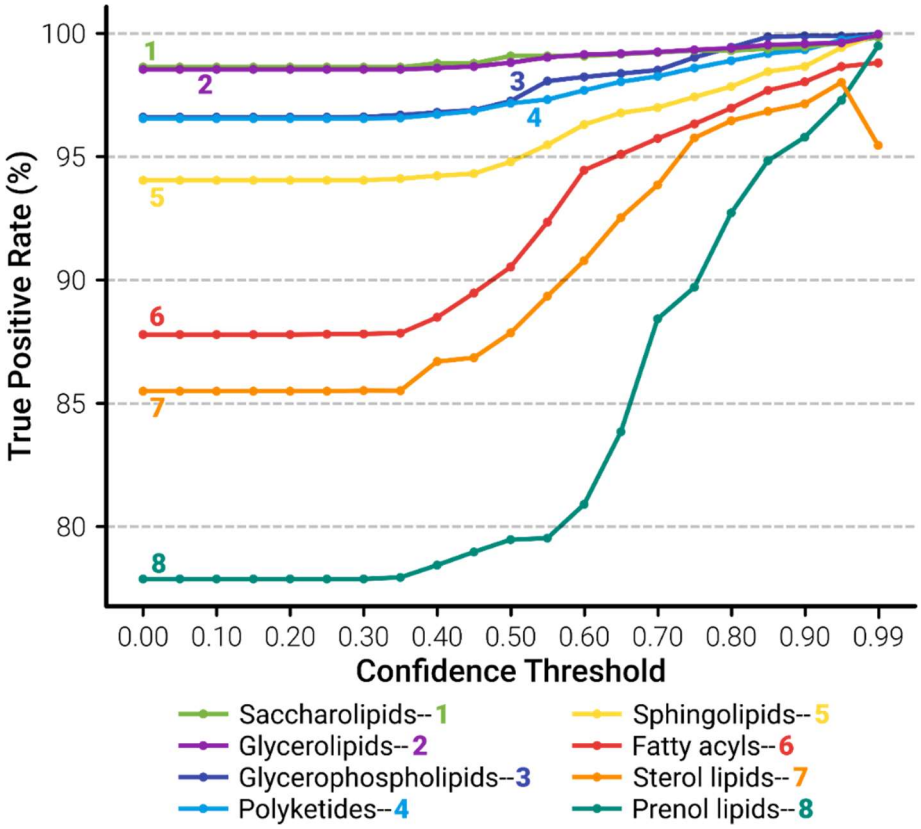


Figure 2.3. True positive rates for 8 trained networks (each aimed at a specific chemical class identification) for the computationally generated data at confidence thresholds (0 to 1.0) are shown. The true positive rate (TPR (%)) is defined as the percentage of actual positives that were classified as such. Increasing the confidence threshold value causes TPR to approach 100% for most classes. Sterol lipids (7, orange) exhibit a decrease in TPR in a rare case where more correct than incorrect classifications are removed by an increase in confidence threshold.

As confidence threshold is increased, the TPR generally increased for each class as more incorrect classifications with lower network output scores are removed than correct classifications (which resulted from generally higher network output scores). As shown in Figure 2.3, the lowest performing classes, prenol lipids (trace number 8 in turquoise), sterol lipids (trace number 7 in orange), and fatty acyls (trace number 6 in red), exhibit the largest improvements in TPR as a function of confidence threshold. By removing low scoring, incorrect predictions in these classes, the variance in performance between all classes is

reduced. For example, at confidence threshold 0.80, all classes exhibit TPRs of greater than or equal to 95%.

At high confidence thresholds (*i.e.*, > 0.95), it is possible for TPR to decrease; this is shown by the sterol lipid class that dropped by 2.56% TPR when the confidence threshold was increased from 0.95 to 0.99 (Figure 2.3, trace number 7 in orange). Decreases in TPR as a function of confidence threshold occur as more correct classifications than unconfident assignments are removed from the output. This behavior results from the networks' occasional proclivity to make confident, incorrect assignments primarily in the case of highly similar classes; for example, the sterol lipid class (orange, trace number 7 in Figure 2.3) exhibited decreases in accuracy between confidence threshold 0.95 and 0.99 in 8/10 trained networks due to high network output score (> 0.90), false positive classifications of prenol lipid inputs into the sterol lipid class.

Additionally, it is important to note that each class loses coverage (*i.e.*, the percentage of retained confident classifications of all classifications) as a function of confidence threshold as unconfident classifications are removed. As a rule, increased TPR as a consequence of higher confidence threshold selection results in some loss of coverage. The degree of coverage loss depends on the number of classifications removed; therefore, large increases in TPR result in large losses in coverage. The variance in class coverage is reported as estimated standard error of the mean (SEM % $= \frac{s}{\sqrt{n}} \times 100\%$). Based on the LIPSUBNET

results, fatty acyls, sterol lipids, and prenol lipids lost coverage (and gained TPR; Figure 2.3 traces 6, 7, and 8, respectively) at greater rates, causing high variance in coverage (76.6 ± 6.9%) between classes' TPR at a confidence threshold of 0.90. The relationships between class coverage and confidence threshold for each class in LIPSUBNET is displayed in Appendix A.9. The inverse relationship between TPR and coverage constitutes a "trade-off" that the user must manage depending on the user's specific needs for increased classification accuracy or higher coverage of all detected analytes. For general use of LIPSUBNET in this demonstration, the authors suggest use of a confidence threshold of 0.70 that balances TPR (96.3 ± 1.3%), total percent coverage (88.7 ± 3.9%), and uniformity of class representation.

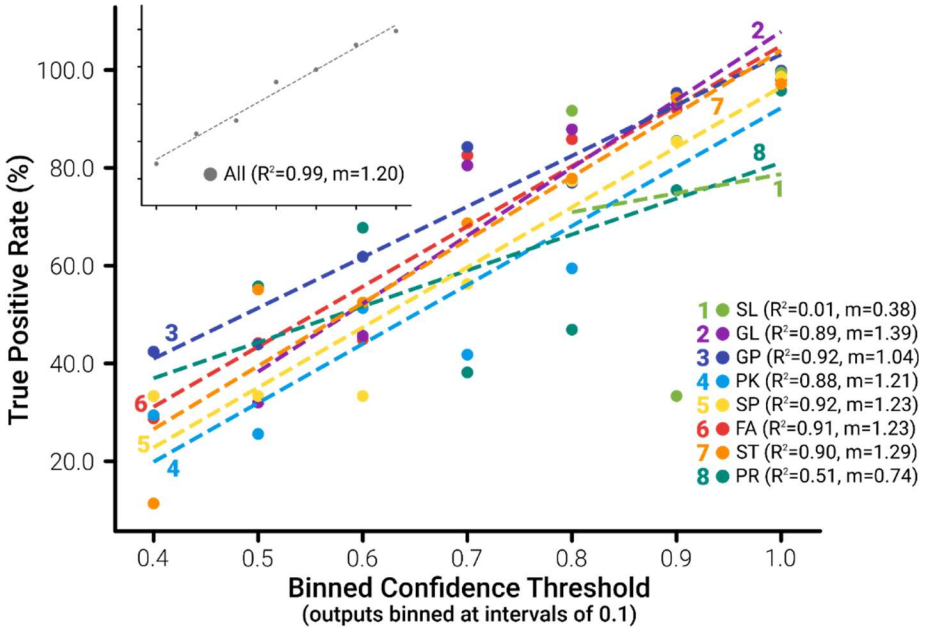


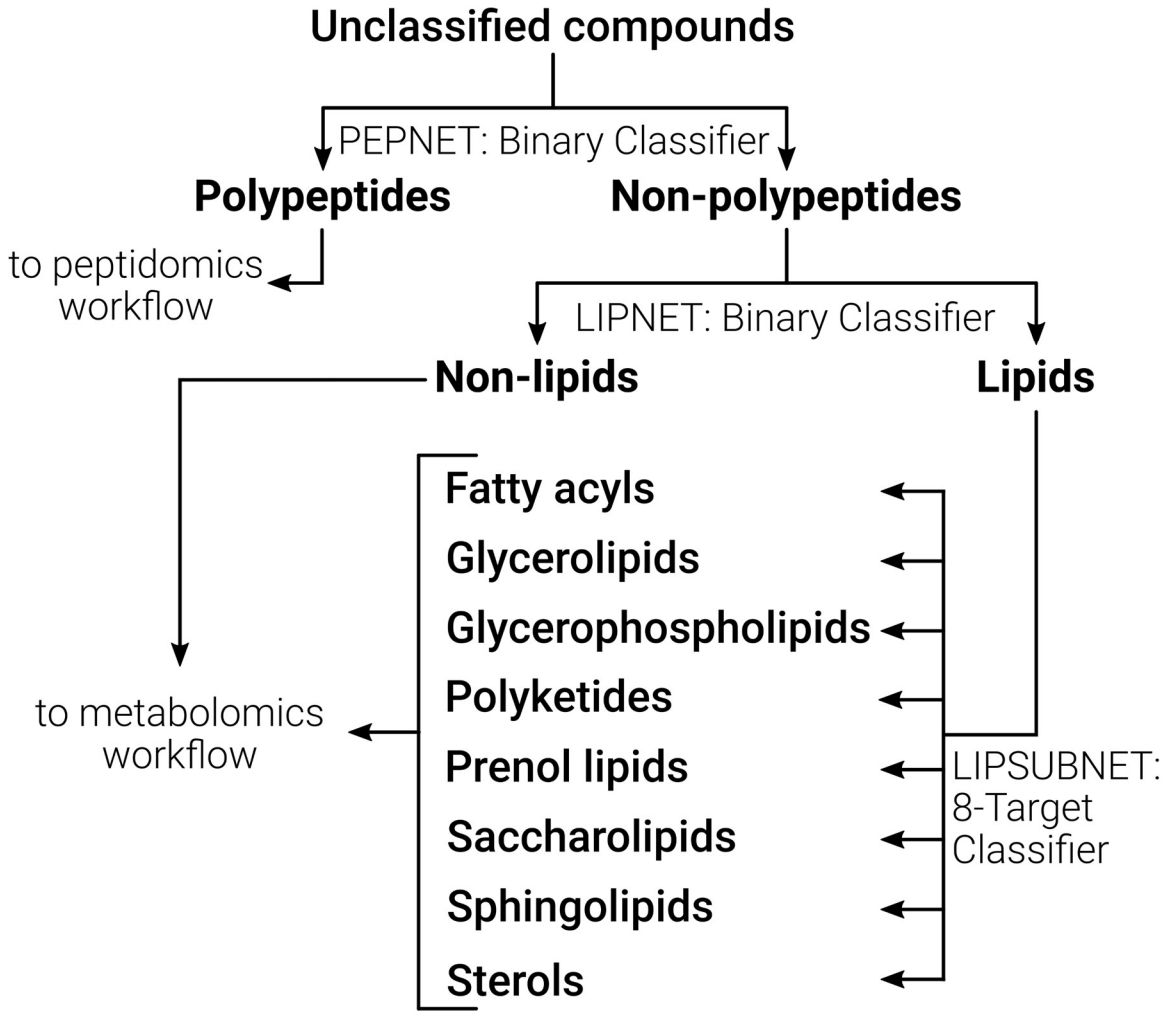
Figure 2.4. Visualization of the true positive rate (TPR) of LIPSUBNET network outputs binned at confidence threshold intervals of 0.1 (e.g., (0.1, 0.2], (0.2,0.3],...(0.9, 1.0], where each plotted point represents the TPR of the outputs which scored in the binned interval range; only bin intervals that contained at least 10 outputs were plotted. Subclasses are represented in the legend as follows: saccharolipids (SL, 1), glycerolipids (GL,2), glycerophospholipids (GP, 3), polyketides (PK, 4), sphingolipids (SP, 5), fatty acyls (FA, 6), sterol lipids (ST, 7), and prenol lipids (PR, 8). The "All" category (inset, with the identical x and y axis ranges of 0.4 to 1.0 binned intervals and 0 to 100% TPR) represents the combined lipid classification accuracy by LIPSUBNET.

The general positive trends for observed TPR values as a function of confidence threshold (after a confidence threshold of ~0.30; Figure 2.3) suggests a relationship between the magnitude of network output score and the probability that an actual positive is classified as such. To visualize the relationship between network output score and TPR without the influence of high scoring outputs at every threshold (as in Figure 2.3), the TPR of network output scores binned at confidence threshold intervals of 0.1 (*e.g.*, [0.0, 0.1], (0.1, 0.2], ..., (0.9, 1.0] bins for the 0.4 to 1.0 confidence threshold range) are displayed in Figure 2.4 (where only sufficiently populated bins with n ≥ 10 inputs

have been included). The raw network outputs of all lipid subclasses (x values in Figure 2.4) exhibit an appreciable linear correlation (positive m values from 0.38 to 1.39, Figure 2.4) with TPR; the mean R^2 value for all classes (excluding saccharolipids, SL with $R^2 = 0.01$, trace number 7 of Figure 2.4 in light green) is 0.85 ± 0.06 (± SEM). The exceptionally low R^2 of 0.01 and high TPR (of 98% as seen in Figure 2.3, trace 1 in light green) for SL indicate that LIPSUBNET may have been overfit for characterization of the SL target class. However, when all (binned) lipid outputs were considered together, a high linear correlation between accuracy and network output score was observed ($R^2 = 0.99$, Figure 2.4 inset). The slope (m) of the combined linear regression trend line (inset, Figure 2.4) was 1.20, suggesting a near 1:1 relationship between the magnitude of output score and the probability that a classification is correct; as such, the magnitude of raw network outputs may be useful for determining classification confidence levels.

2.4.3 In Silico Fractionation of Multi-Class Component Mixtures

The supervised training sets of PEPNET, LIPNET, and LIPSUBNET were constructed to demonstrate the *In Silico* Fractionation (iSF) approach. In this case, we present an application of the iSF approach that utilizes a NDT structure to parse a sample dataset containing lipid, polypeptide, and polar metabolite components (Scheme 2.1).



Scheme 2.1. Representative *In Silico Fractionation* neural decision tree workflow diagram of a biological sample dataset containing polypeptide, lipid, and polar metabolite components.

A visual representation of the iSF approach to an artificially combined, representative multi-class component experimental LC-MS data is presented in Figure 2.5. The analyzed LC-MS dataset (represented in Figure 2.5) is not intended to reflect realistic sample preparation or chromatographic conditions as all analytes were initially prepared, separated, and detected in ideal conditions. However, it was prepared to demonstrate a use case for iSF in which analyte signal could not be assumed to originate from a single class. It should also be noted that, with exception to enhancements to detection sensitivity, sample preparation and chromatographic conditions do not uniquely affect the spectral features of the MS isotopic envelope utilized by iSF.

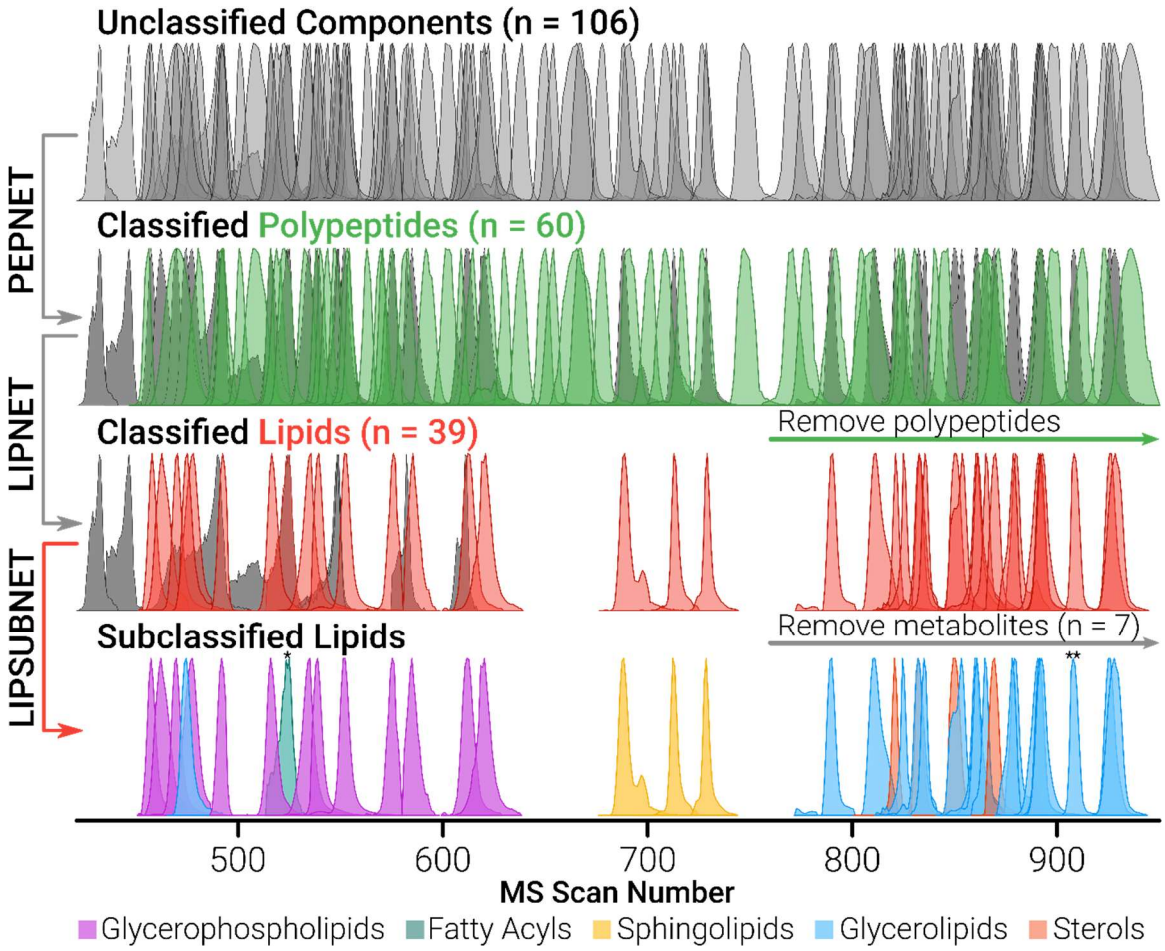


Figure 2.5. A visual representation of the *In Silico Fractionation* approach to a randomly selected portion of an LC-MS dataset that included a total of 106 unknown analytes. The original input data (106 gray LC peaks) in top row was created by adding a portion of an experimentally acquired LC-MS output from a rat brain tryptic digest analysis to lipidomics LC-MS data downloaded from Chorus) [43]. Unclassified components (gray LC peaks) were submitted to PEPNET which classified inputs as 60 polypeptides (T_P, LC peaks in second row from the top) or 46 non-polypeptides (T_{NP}, gray LC peaks). T_P inputs were removed, and the remaining T_{NP} inputs (46 species) were submitted to LIPNET for a second chemical class identification, which classified inputs as 39 lipids (T_L, red LC peaks in third row from the top) and 7 non-lipids (T_{NL}, gray LC peaks in the third row from the top). T_{NL} inputs were removed, and T_L inputs were submitted to LIPSUBNET, which classified T_L inputs into narrower lipid subclasses. PEPNET, LIPNET, and LIPSUBNET had true positive rates of 100%, 100%, and 95%, respectively. The * and ** symbols represent a prenol lipid component and a sterol component that were erroneously predicted as fatty acyl and glycerolipid, respectively.

Additionally, as Scheme 2.1 and Figure 2.5 suggest, the end goal of this work is to guide fractionation of the full MS dataset in state amenable for separate downstream analyses for classes of molecules in a multi-class mixture that have unique data processing requirements. However, the challenges associated with full integration of iSF into a multi-omics workflow will be addressed in a future study.

Raw network output scores for PEPNET, LIPNET, and LIPSUBNET related to Figure 2.5 are provided in Appendix A.10. Figure 2.5 demonstrates the application of iSF to the classification of multiple types of eluting compounds in a standard LC-ESI-MS analysis using experimentally gathered data. Each eluting compound is pictorially represented by its detected LC peak profile. Peaks that have yet to be positively classified (i.e., classified into a discrete category of biomolecule) were shown in gray and were assigned a color once positively classified. The iSF approach begins with generating an input vector for each of the unclassified compounds in a LC-MS dataset (Figure 2.5, top, gray). As the first step in Figure 2.5, PEPNET proceeds to classify network inputs (second row, Figure 2.5) as polypeptides (T_P, green) or non-polypeptides (T_{NP}, gray). In the dataset presented in Figure 2.5 (*i.e.*, the combined dataset discussed at the end of the Experimental section comprised of an eluting mixture of brain peptides, lipids, and polar metabolites detected by ESI-MS), polypeptides were classified by PEPNET with 100% TPR.

For the second step (row 3, Fig. 2.5) LIPNET classified the input vectors that PEPNET classified as T_{NP} (gray peaks in row 2 of Fig. 2.5) as either lipids (T_L, red) or non-lipids (T_{NL}, gray) as shown in row 3 of Figure 2.5 with 100% TPR. The vector inputs classified as T_L by LIPNET were then submitted to LIPSUBNET for classification into lipid subclasses (*i.e.*, subclasses listed by LipidMaps). LIPSUBNET classified 37 out of 39 inputs correctly (95% TPR for all class targets).

See Table 2.1 for statistical metrics for each class; FA, GL, GPL, PK, PR, SL, SP, and ST acronyms in Table 2.1 (column headers) are used for fatty acyl, glycerolipid, glycerophospholipid, polyketide, prenol lipid, saccharolipid, sphingolipid, and sterol lipid, respectively, where row headers TPR, TNR, FPR, and FNR stand for true positive rate, true negative rate, false positive rate, and false negative rate, respectively. One sterol lipid input vector was incorrectly classified as a glycerolipid with a network score of 0.597 and one prenol lipid input vector was incorrectly classified as a fatty acyl with a network score of 0.622. However, for both the sterol and prenol misclassifications, the true class targets received the second highest network scores of 0.323 and 0.217, respectively. If the previously suggested confidence threshold (of 0.70) is applied to the results of the iSF workflow reported herein, 5 classifications, including both incorrect classifications, are removed, resulting in 100% TPR with 87% (*i.e.*, 34 out of 39 input) coverage. Additionally, input vectors previously classified by PEPNET or LIPNET may be resubmitted to the other

(e.g., input vector classified as T_P by PEPNET can be reclassified by LIPNET) for secondary confirmation. All positively classified input vectors resubmitted to either PEPNET or LIPNET (respective to their previous classification) were classified with 100% TPR.

Table 2.1. LIPSUBNET statistical metrics for classification of experimental data

Metric	FA	GL	GPL	PK	PR	SL	SP	ST
TPR	-	94.7%	100.0%	-	0.0%	-	100.0%	80.0%
TNR	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FPR	2.6%	4.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FNR	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	20.0%
Overall	Accuracy	98.7%		TPR	94.9%		TNR	99.3%

To demonstrate an example of iSF's application to a real-world multi-class separation, PEPNET and LIPNET were used to analyze a separated mixture of trypsin digest peptides and lipids. The total ion LC chromatogram of the combined lipid and peptide separation is shown in Figure 2.6. The dotted lines mark the LC retention time boundaries between which, from left to right, peptides, unknowns, and lipids eluted. In this demonstration, the true classes of each of analyte signals were determined by their respective reverse-phase chromatographic elution times.

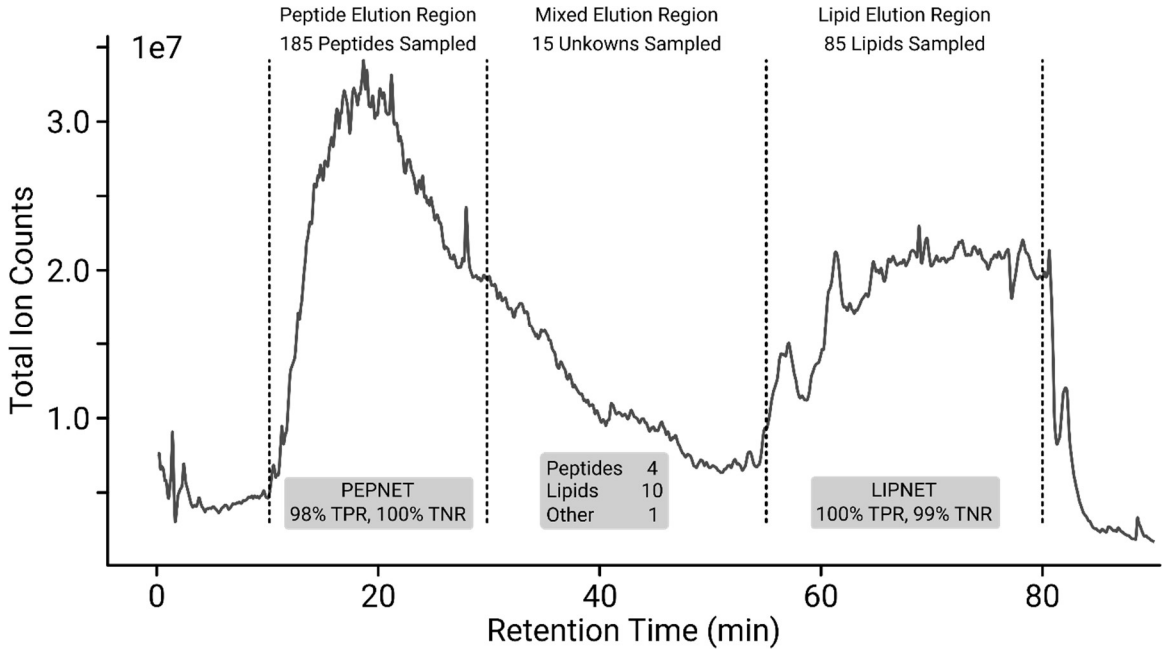


Figure 2.6. A visual representation of the *In Silico Fractionation* approach to a portion of a reversed-phase LC-MS dataset in which a mixture of HeLa digest peptides (total of 150 ng) and rat brain-extract lipids (from 250 μg rat brain) were separated and analyzed in the same acquisition. Known peptide data were extracted from between 10-30 minutes, lipid data from between 55-80 minutes, and unknowns from the middle region where both hydrophobic peptides and hydrophilic lipids were expected to elute. Regarding the classification of known components, PEPNET had a 98% TPR and 100% TNR, and LIPNET had a 100% TPR and 99% TNR. In the middle region, 4 unknowns were classified as peptides, 10 were classified as lipids, and 1 was classified as a non-lipid and non-peptide. No conflicting classifications were made for any unknown component. The period before 10 minutes was excluded as it contained only unretained components and background signal. The period after 80 minutes was excluded as it contained only a strong contaminate/polymer signal.

As indicated in Figure 2.6, LC eluting analytes between 10 and 30 minutes were determined to be peptides, and those that eluted between 55 and 80 minutes were determined to be lipids. The classification of each analyte was confirmed by independent LC-MS analyses of the peptide and lipid mixtures. Polypeptide analytes were classified by PEPNET as T_P with 98% TPR (181 out of 185 inputs). Of the 185 polypeptide components, 4 polypeptides were incorrectly classified by PEPNET as T_{NP}. Each of the 4 polypeptides were small (MW < 1000 Da), singly-

charged ions; however, it should be noted that most other small, singly-charged polypeptide components were classified correctly (12 out of 16 inputs). Lipid inputs were classified by LIPNET as T_L with 100% TPR (85 out of 85 inputs). Also, these lipid inputs submitted to PEPNET were classified as T_{NP} with 100% TNR. Peptide inputs submitted to LIPNET were classified as T_{NL} with 99% TNR (183 out of 185 inputs). The two polypeptide inputs that were misclassified by LIPNET as T_L were 2 of the 4 small, singly-charged polypeptide inputs misclassified by PEPNET. In the mixed elution region from 30 to 55 minutes, 15 unknown components were classified by both PEPNET and LIPNET. In this group of inputs, iSF classifications were self-consistent as 4 inputs were classified as T_P and T_{NL} (designated as "T_P/T_{NL}"), 10 inputs as T_L/T_{NP}, and 1 input that was classified as neither a peptide nor lipid (both T_{NP} and T_{NL} or T_{NP}/T_{NL}). No conflicting classifications (i.e., belonging to both T_P and T_L classifications) were made for any unknown input. Raw network output scores for PEPNET and LIPNET related to Figure 2.6 are provided in Appendix A.11.

2.4.4 Considerations for Future Applications

The generalization capabilities of FFNNs enabled PEPNET, LIPNET, and LIPSUBNET to accommodate non-trivial variance in the *m/z* domain. The supervised training sets for each class were constructed with elemental compositions corresponding to the commonly observed adduct ions in ESI (*e.g.*, $[M+H]^+$ for most polypeptides, $[M+Na]^+$ for glycans, $[M-H]^-$ for oligonucleotides,

etc.). However, there were a variety of detected lipid adduct forms (*e.g.*, $[M+NH_4]^+$, $[M+ ACN+H]^+$, $[M+ Na]^+$, *etc.*) that were successfully classified by LIPSUBNET and LIPNET, even though the training set of lipids for LIPSUBNET and LIPNET was restricted to protonated adduct lipid forms. Given the easy-to-obtain nature of the selected mass spectral vector inputs (*i.e.*, isotopologue peak relative intensity, exact mass, and KMD information) and the insensitivity of iSF to potential interferences from adduct ions, iSF could be applied to most broadband MS workflows in which analytes' molecular ion isotopic envelopes are preserved. Given its robustness, iSF can be applied as a pre-processing step in conceivably any concurrent multi-omics analysis (in which MS1 spectra are acquired) to provide crucial information about analyte class and guide future data processing. Such capabilities should be useful for characterization of complex biological systems such as bacterial differentiation in microbiome studies [214], class identification in biological MS imaging, and pictorial representation of biomarker panels (healthy controls vs disease states) in clinical studies. Although the presented work here has focused on positive-ion mode experiments and classification of intact molecular species, iSF can be applied to other complementary types of data. For example, in future contributions, we plan to evaluate the performance of iSF for utilizing data acquired under negative-ion mode and classification of fragments and other modified structures (such as metal adducts).

The neural network training period took several seconds to several minutes (for details, see Appendix Figures A.3 and A.4); however, once trained, each network classification was rapid (~85 μs of processor time). Thus, these operations can be performed at frequencies comparable to TOF data acquisition rates, which (depending on the measured *m/z* range) can range from ~10-100 kHz [105, 197, 215].

The presented approach is robust in dealing with instrumental noise and small variations in analytes' measured masses (data acquisition restrictions that are due to formation of various types of adducts, resolving power limits, and mass measurement errors). Additionally, our findings suggest that iSF can be applied successfully to MS data produced from low (L) MRP instrumentation. Versions of LIPNET, PEPNET, and LIPSUBNET were trained and tested with monoisotopic *m/z* and KMD values restricted to only two decimal places. The monoisotopic *m/z* and KMD values of the experimental test inputs from the artificially combined multi-omics separation were likewise restricted to two decimal places. These two types of networks (viz., networks trained using data sets with either (a) only two or (b) four decimal places for *m/z* values) had nearly identical performance characteristics both in application to computationally generated data and experimental MS data to LIPNET, PEPNET, and LIPSUBNET. The exciting implication of this finding is iSF's applicability to data from relatively LMRP instrumentation (with precision to only 2 decimal places) that exhibit mass measurement errors unsuitable for elemental composition

determination (~10 ppm error). Tabulated performance metrics for the networks trained with precision to 2 decimal places are shown in Appendix A.12.

Given that relative peak intensities and isotopic ratios are the primary dimensions of separation used by the neural network, the proposed approach is sensitive to convolution of analyte signals in the *m/z* domain (hence, peak capacity limits in the *m/z* dimension govern the level of sample complexity that can be tolerated). We recommend the use of pre-MS physical separations such as LC and/or IM to prevent potential peak convolutions in the *m/z* domain and increase isotopic envelope purity. IM profiles or trendlines for biomolecules have also shown class-dependent trends that may assist in classifications[183]; however, overlaps of *m/z*-mobility trends complicate classification of some groups of biomolecules[183]. It should be noted that, even with physical separation prior to MS injection, convolution of closely related chemicals across measurement domains is possible. In such instances of partial isotopic envelope convolutions of closely related lipids in the LC and *m/z* domains (for example, as reported in Figure 2.6) iSF is still able to correctly classify each convolved component. For instance, 11 of the 85 lipid inputs (all of which were correctly classified as T_L by LIPNET) were partially convoluted in the LC retention time and *m/z* domains. Of the 11 convolved lipids, 6 were partially convolved by closely related species that differed by a degree of unsaturation. In regions of full LC convolution (e.g. retention times at which two lipids that differ by one

degree of unsaturation are eluting; an example of which is shown in Appendix A.13), the third isotopologue peak of the singly-charged, unsaturated species (which eluted first in each case) was fully convolved with the monoisotopologue peak of the saturated species (which elutes second in each case). However, by sampling from the leading edge of the saturated species' LC peak and the trailing edge of unsaturated species' LC peak, sufficiently pure isotopic envelopes were obtained. The other 5 convolved lipid components were partially convolved in the *m/z* domain (i.e., each isotopologue peak convolved up to ~20% peak height; an example for which is shown in Appendix A.14), but extracted isotopic features were sufficiently pure for iSF to classify each component correctly.

2.5 Conclusions

In this study, we describe an *In Silico* Fractionation approach for classification of small biological compounds from MS data *via* isotopic ratio analysis using a neural decision tree. The FFNNs estimate the relationships that define and separate biomolecular classes (*e.g.*, lipids, glycans, polypeptides, *etc.*) based on their respective isotopic distribution patterns and, therefore, their elemental compositions. The FFNNs utilized to demonstrate iSF (PEPNET, LIPNET, and LIPSUBNET) were sensitive in their application to experimentally detected chemical components: PEPNET had a TPR of 100%, LIPNET had a TPR of 100%, and LIPSUBNET had a combined TPR of 95% (without confidence

threshold constraint). The specific demonstration presented here, constitutes one possible design and application of iSF; however, depending on the sample composition and class types present in the sample, the iSF workflow can be tailored for a wide variety of multi-class component mixtures.

2.6 Acknowledgements

The authors acknowledge Ian Anthony (Baylor University) for his assistance with code refactoring and manuscript review, and Dr. Christopher M. Kearney, Dr. S. M. Ashiqul Islam, and Reese Martin for their early conceptual input.

Funding: This work was supported by the National Science Foundation [grant numbers: IDBR-1455668 and CHE-1709526].

CHAPTER THREE

Chemical Classification for Improved Lipidomics Sample Annotation with In Silico Fractionation

3.1 Abstract

Chemical annotation is indispensable in untargeted lipidomics workflows. Conventionally, annotations in lipidomics data are often limited to identifications produced by instrumental methods (e.g., exact mass measurements and fragmentation spectral matching) and to chemical structures that are present in existing libraries or databases. Here we describe an application of *In Silico* Fractionation (iSF), a machine learning tool for classification of small biological molecules in mass spectrometry (MS) data, for subclassification of lipids in liquid chromatography (LC)-MS data. In this work, iSF uses an array of binary multi-layer feedforward neural networks to classify lipids using only relative abundance information for chemical isotopes. Here we show the performance of iSF classifiers in application to large lipidomic data sets and assess its performance by comparing the iSF results to LipiDex tandem MS identifications. Using this approach, we can accurately classify LC separated lipids into their respective sub-classes using data from single stage mass spectrometry alone. This approach does not require a priori chemical identification and hence is orthogonal to conventional and tandem MS assignment protocols.

3.2 Introduction

Lipid distributions and structures vary greatly with respect to organism, organ, tissue, and cell type[216-218], and their characterization is a vital step towards understanding of the chemistry involved in living organisms. Mass spectrometry (MS)-based lipidomics is a powerful approach for characterization of lipids and their prevalence in biological samples[219, 220]. For example, in early applications, direct injection electrospray ionization (ESI) coupled to triple quadrupole instruments was used to characterize and quantify lipids in cellular lipid extracts[221, 222]. To expand the analytical capabilities of direct injection MS, pre-MS in-line separation tools such as liquid chromatography (LC) and ion mobility (IM) are often coupled to MS to increase peak capacity [223] and improve lipidome characterization [224, 225]. Likewise, advanced MS acquisition modes such as tandem MS (MS/MS) are used to increase the accuracy and confidence in lipid identifications [226, 227]. However, lack of comprehensive of MS/MS spectral libraries might prevent full lipidome annotation and require computer-generated tandem mass spectral libraries to increase coverage [228]. Furthermore, the extent to which a sample can be accurately annotated with identifications made by MS/MS fragment spectral matching depends on the quality of MS/MS spectra utilized; for instance, LC coeluting compounds may require advanced deconvolution methods to ensure confident annotation [229].

In conventional tandem MS lipidomics studies, generally a subset of MS/MS analyte identifications can be highly confident; however, it is also possible that many of the MS/MS ion fragmentation pattern matches exhibit poorer qualities and yield less confident lipid identifications [226, 229]. Additionally, even with highly accurate measured masses (e.g., better than 1 parts-per-million (ppm) at mass resolving powers ($M/\Delta M_{50\%}$) greater than ~100,000), it is not often possible to differentiate potential isomers, isobars, or other confounding species in congested mass spectra. Hence, more complete and accurate sample characterization in LC-MS and -MS/MS lipidomics data should require annotation methods that provide relevant information, regarding lipid types and their functional groups, in cases that preclude definitive analyte identification.

Structure based classification are useful for lipid characterizations and have been used to address sample complexities in lipidomics studies. For example, more advanced lipidomics analysis workflows such as ClassyFire [230] and LIPID MAPS [198] utilize substructure or molecular fingerprint analyses [231-233] and structure-based chemical classification systems to enable specific structure-based class annotation. Alternatively, experimental methods have enabled annotation of metabolomics and lipidomics features through assignment of analyte class information in the absence of definitive identity assignment. For instance, the IM-MS community has thoroughly shown that analytes map to IM collision cross section (CCS) conformational space on a per-class basis [234] which

has assisted in identity assignment in untargeted acquisition methods [183, 209, 234]. Additionally, Kenrick mass defect (KMD) [19] has been shown to vary dependently with respect to analyte class; KMD analyses have been purposed towards classification of bulk lipid classes [14, 15, 23] and identification of cohorts of similar chemicals in MS imaging experiments [22]. In addition to these structure-based approaches, modern machine learning tools have been utilized to enhance lipidomics data analyses [14, 15, 235].

Machine learning approaches improve the classification and descriptive power of the MS measurements by elucidating hidden correlations between various measured data points. For examples, Dührkop *et al.* developed a two-step approach, Class Assignment aNd Ontology Prediction Using mass Spectrometry (CANOPUS), involving support vector machines and neural networks to classify analytes using MS/MS fragmentation spectra [235]. Additionally, McLean and colleagues employed a random forest machine learning method, Supervised Inference of Feature Taxonomy from Ensemble Randomization (SIFTER), and IM-MS data (i.e., monoisotopic *m/z*, KMD, and IM CCS) for analyte classification [14]. Neither CANOPUS nor SIFTER can be utilized using single-stage mass spectrometry data alone. For example, the CANOPUS method requires MS/MS data, and SIFTER uses ion mobility CCS and MS data as input. Herein, we present a method that predicts analyte classifications using only MS measurements. This approach avoids complexities associated with

variable quality of MS/MS measurements and does not require IM data for successful classification of lipids.

In this work, we use *In Silico* Fractionation (iSF), a feed-forward neural network (FFNN)-based approach, for the classification and annotation of lipid molecular ions in MS data [15]. Here, we evaluate the performance of iSF for specific lipid sub-classification in a large lipidomic dataset by comparing its results to LipiDex MS/MS findings. Overall, iSF FFNNs utilizes the measured features of the MS isotopic envelope for 1) supervised FFNN training using theoretical isotopic envelope calculations and 2) to classify analytes using experimentally acquired single-stage mass spectral data. Specifically, we show the capacity of iSF to accurately classify lipids in LC-MS analyses of complex samples into the eight main lipid classes as denoted by the LIPID MAPS Structure Database (LMSD). Additionally, we demonstrate the ability of iSF classification and annotation to refine lipid identification (especially those associated with low confidence MS/MS spectra) and to provide the basis for sample differentiation by iSF predicted class representation.

3.3 Materials and Methods

3.3.1 Training, Validation, and Test Dataset Generation

The iSF neural network training set was generated from the lipid elemental compositions in the LIPID MAPS Structure Database (LMSD, updated 2019-10-02). A custom Python (CPython 3.7.6; Python Software Foundation, DE) script was used to parse the LMSD in .sdf format and generate theoretical isotopic envelopes for each composition using the IsoSpecPy Python library. The neutral composition of each lipid was modified with five class-appropriate potential electrospray adducts (all resulting in singly charged molecular ions) for positive- and/or negative-ion modes, and the isotopic envelope was calculated for each molecular ion form. The adduct forms for each lipid class are detailed in Appendix B.1. The isotopic feature input for each adduct form of each lipid was collected in a "master" dataset (n = 200,805 structures) from which the training set for each type of FFNN (i.e., for each major class of lipid) was built. In general, the training set for each type of FFNN incorporated a pseudo-random selection of isotopic feature inputs in Python for the positive class targets (targets to be positively predicted) equal to the number of the least represented class (viz., saccharolipids, n = 6,580) from the master dataset; a pseudo-random selection of isotopic feature inputs for each of the other classes (all equal in size) were incorporated for the negative class targets (targets to be negatively predicted). The relative size of the negative inputs to positive inputs

was increased if FFNN overtraining was evident (e.g., if FFNN training and validation set performance metrics diverged substantially) [236]. Optimized ratios of negative-to-positive inputs were as follows for each class: 1.0 (fatty acyls, n_{Total} = 13,160), 1.5 (glycerolipids, n_{Total} = 16,450), 1.5 (glycerophospholipids, n_{Total} = 16,450), 1.0 (polyketides, n_{Total} = 13,160), 1.0 (prenol lipids, n_{Total} = 13,160), 1.0 (saccharolipids, n_{Total} = 13,160), 2.5 (sphingolipids, n_{Total} = 23,030), and 1.0 (sterol lipids, n_{Total} = 13,160). Supervised training target outputs were provided such that the presence of a class of biological molecules was indicated by a value of 1.0 and absence by a value of 0.0.

The neural network inputs for each molecular ion component were calculated from its respective isotopic envelope features and were structured as MATLAB (R2020a; The MathWorks Inc., Natick, MA) row vectors and are shown in Table 3.1. The monoisotopic *m/z* (input 1) was limited to 4 decimal places to be consistent with achievable mass resolving power of the Orbitrap employed ($m/\Delta m_{50\%}$ ~ 60,000) in this demonstration.

Table 3.1. FFNN input vector structure.

Position	Input	Position	Input
1	*m/z*	7	$A_{R.I.}/A+2_{R.I.}$
2	$A_{R.I.}$	8	KMD_{CH2}
3	$A+1_{R.I.}$	9	KMD_{S}
4	$A+2_{R.I.}$	10	KMD_{O}
5	$A_{R.I.}/A+1_{R.I.}$	11	KMD_{P}
6	$A+1_{R.I.}/A+2_{R.I.}$	12	KMD_{N}

The relative intensity (as subscripts R.I. in Table 3.1) for each isotopologue referenced by inputs 2-7 is denoted using the "A" notation introduced by McLafferty et al.[186] where each subsequent isotopologue proceeding the monoisotopologue is indicated with [A+N] nomenclature. Five Kendrick mass defect (KMD) values relative to CH_2, S, O, P, and N (inputs 8-12) were included as each provided a small improvement in training performance and were simple to calculate.

3.3.2 Neural Network Training

FFNNs were constructed and trained with the neural network package in M (MATLAB R2020a). Eight FFNNs were used in this study and they were denoted as FA-NET (for positive classification of fatty acyls), GL-NET (for positive classification of glycerolipids), GP-NET (for positive classification of glycerophospholipids), PK-NET (for positive classification of polyketides), PL-NET (for positive classification of prenol lipids), SL-NET (for positive classification of saccharolipids), SP-NET (for positive classification of sphingolipids), and ST-NET (for positive classification of sterol lipids). Each FFNN is a binary classifier trained to positively classify the class of lipids denoted in the name and negatively classify all other lipid classes.

The hidden layer architecture (i.e., number of hidden layers and number of nodes per hidden layer) was optimized by evaluating the mean percent error (i.e. percent of false-positive and false-negative classifications over all

classifications) of each architecture with 10 training samples of each architecture. FFNN architectures with up to 3 hidden layers and up to 100 nodes per layer (in 10 node increments) were tested; in every sample, the number of nodes per layer were held constant. Hidden layer architecture optimization data is shown in Appendix B.2. A network architecture of 2 hidden layers with 50 nodes in each layer was chosen for its low error and low training time. The input data set was then randomly divided (*i.e.,* MATLAB function "dividerand") for network training as follows: 70% in the training set, 15% in the validation set, and 15% in the test set. Networks were trained using scaled conjugate gradient (SCG) backpropagation (*i.e.,* MATLAB function "trainscg") and gradient descent with momentum and adaptive learning rate (GDX) backpropagation (i.e., MATLAB function "traingdx")—both with GPU acceleration. Seven networks were trained for each type of FFNN, for each training method (SCG and GDX), and for each training set (i.e., that included different ratios of negative-to-positive inputs). Additionally, an early stopping method [203] was used to avoid overfitting; FFNN training was completed after 4 validation set failures. The loss function used to measure FFNN training performance was cross-entropy (MATLAB function "cross-entropy"). Following training, the most accurate networks with sufficient generalization were selected for this demonstration; a generalized FFNN was indicated by insignificant performance differences between the training and test sets as well as positive and negative classification performances. All networks were trained using a desktop computer (Precision

Tower 7810; Dell Computer, Round Rock, TX) equipped with a 10-core/20-thread processor (Intel® Xeon™ E5-2640), a discrete graphics processing unit (NVIDIA GTX 1080), and 32 GB of RAM.

3.3.3 Lipidomics Data Sourcing and Feature Extraction

The lipidomics data used in this study was acquired and shared on the CHORUS mass spectrometry database by Coon et al. for an initial demonstration of LipiDex, an open-source software suit for LC-MS/MS lipidomics data analysis[226]. The sample preparation procedures and LC-MS/MS analysis parameters are comprehensively shown in the main text and supplemental material of the LipiDex publication[226] but are summarized here. Lipidomics datasets include three LC-MS/MS acquisitions for each positive- and negative-ion mode from 4 biological sources: human Hap1 cells, yeast (*S. cerevisiae*), mouse liver homogenate, and pooled human plasma. Hap1 cell, yeast, and mouse liver lipids were extracted by a chloroform/methanol liquid-liquid extraction and were reconstituted in ACN/IPA/H_2O (65:30:5, v/v/v). Human plasma lipids were extracted by methyl tert-butyl ether liquid-liquid extraction and was reconstituted in MeOH/Toluene (9:1, v/v). All acquisitions used a full MS resolution of 60,000 and an MS/MS resolution of 15,000. All mouse liver, human plasma, and yeast cell analyses used an MS/MS collision-induced dissociation (CID) stepped normalized collision energy (NCE) of 20 to 25, positive polarity Hap1 cell analyses used a stepped NCE of 20 to 25, and negative polarity Hap1

cell analyses used a stepped NCE of 20 to 30. Reversed-phase LC separations were conducted with a binary solvent gradient program; mobile phase A was 10 mM ammonium acetate in ACN/H2O (70:30, v/v) with 250 µL/L acetic acid, and mobile phase B was 10 mM ammonium acetate in IPA/ACN (90:10, v/v) with 250 µL/L acetic acid. Thermo Fisher Scientific *.raw files were downloaded from the Chorus Project (Stratus Biosciences, Seattle, WA) mass spectrometry file sharing database (https://chorusproject.org/pages/dashboard.html#/projects/all/1409/experiments)[205]. Additionally, unfiltered lipid identifications/peak lists were sourced from the LipiDex results files provided with the software in the Coon group's github repository (https://github.com/coongroup/LipiDex).

Several custom Python scripts were used to: 1) parse LipiDex results files for feature information including monoisotopic *m/z*, LC retention time, lipid identification/class, molecular ion charge, and MS/MS dot product score, and 2) extract MS isotopic envelope features (monoisotopic *m/z* and relative intensities of [A], [A+1], and [A+2] isotopologues) directly from Thermo Fisher Scientific *.raw files using Francois Allain's Thermo MSFileReader Python bindings (pymsfilereader, https://github.com/frallain/pymsfilereader). Additionally, isotopic features were extracted from five scans along the apex of the select ion chromatogram peak for each identified lipid feature to control for variable quality in isotopic envelopes over the duration of each LC elution event; each

identified lipid LC-MS feature is associated with five iSF inputs before further quality control filters were applied (described in section 2.4). Isotopic envelope features were used to calculate each of the iSF input vector values (as shown in Table 1).

3.3.4 iSF Input Testing and Results Filtering

All iSF inputs were submitted to each network, and results were filtered using several quality control (QC) parameters. Calculated QC parameters fall into two groups wherein they either 1) describe the quality of the LC-MS feature that produced the iSF input or 2) describe the confidence/validity of the LipiDex identification.

The following QC parameters describe the quality of each LC-MS feature and estimate the isotopic purity (i.e., the potential for *m/z* convolution of isobaric peaks). The primary concern for an iSF application to lipidomics data was potential for convolution of the [A+2] isotopologue of one lipid component with the [A] isotopologue of a component that contains one less degree of unsaturation. The primary QC parameter controlling isotopic purity was the linear correlation (R^2) between isotopologue select ion chromatograms (SIC) to filter out potentially LC-MS convolved features. For each input, the SIC of each isotopologue were extracted, and each SIC was tested for linear correlation with the SIC of the proceeding peak; the term R^2_1 represents the linear correlation between $SIC_{[A]}$ and $SIC_{[A+1]}$, and R^2_2 represents the linear correlation between

$SIC_{[A+1]}$ to $SIC_{[A+2]}$. An additional concern was potential for feature extraction of unassociated peaks (i.e., non-isotopologues) in, for example, a case where the [A+2] isotopologue of a feature is below detection limits and a different peak is acquired during feature extraction. The difference between the $\Delta m/z_{[A+1]-[A]}$ and $\Delta m/z_{[A+2]-[A+1]}$ (termed $\Delta\Delta m/z$) was calculated to provide a metric with which to filter out such invalid iSF inputs. Given the theoretical mass difference of 1.00335 between ^{12}C and ^{13}C, the theoretical absolute value of $\Delta\Delta m/z$ approaches zero for organic molecules. Additionally, iSF inputs were filtered by the intensity of the monoisotopologue peak.

The following QC parameters describe the confidence and validity of each LipiDex identification. A QC parameter utilized in LipiDex identification filtering that was incorporated into this workflow was retention time (RT) median absolute deviation (MAD) for each subclass of lipid identified by LipiDex. The median and MAD was calculated for each subclass and identified LC-MS features were an outlier if its respective RT was outside the subclass median RT ± 3*MAD range. Additionally, iSF inputs were filtered by the LipiDex MS/MS Dot Product score of the lipid identification associated with the iSF input.

3.4 Results and Discussion

Lipid features identified by the LipiDex MS/MS analysis workflow from several different biological sources (Hap1 cells, liver, plasma, and yeast) and

from both positive- and negative-ion mode acquisitions were extracted from their respective Thermo .raw files. Classifications for each feature were then predicted by iSF FFNNs. The accuracy of iSF predictions were assessed by constraining inputs to those with high scoring LipiDex identifications. Additionally, the orthogonality of iSF in application to low confidence MS/MS identifications and as a means to describe and differentiate sample types were investigated.

3.4.1 In Silico Fractionation of Lipidomics LC-MS Data

In order to provide a performance baseline for each iSF network, iSF FFNNs were tested with inputs that were filtered by the abovementioned QC parameters to include only those produced by high quality LC-MS features and high confidence LipiDex identifications ($R^2_1 \geq 0.9$, $R^2_2 \geq 0.9$, $|\Delta\Delta m/z| \leq 0.01$, $Intensity_{[A]} \geq 1e5$, MS/MS Dot Product ≥ 700). Figure 3.1 displays the receiver operator characteristic (ROC) curve for each type of network which positively predicted lipid classes identified by LipiDex, which includes fatty acyls (Figure 3.1A), glycerolipids (Figure 3.1B), glycerophospholipids (Figure 3.1C), sphingolipids (Figure 3.1D), and sterol lipids (Figure 3.1E). A perfect classifier would have an ROC area under the curve (AUC) of 1.0 given that both true-positive rate (TPR) and false-positive rate (FPR) are ratios between 0 and 1. The FA-NET ROC curve (Figure 3.1A) was produced from 24 true-positive FA positive-ion mode inputs (solid green), had an AUC of 0.934, and had an optimal

cutoff value of 0.403 (TPR% = 87.5, FPR% = 11.5). The GL-NET ROC curve (Figure 3.1B) was produced from 6218 GL positive-ion mode inputs (solid green), had an AUC of 0.934, and had an optimal cutoff value of 0.322 (TPR% = 89.2, FPR% = 10.8).

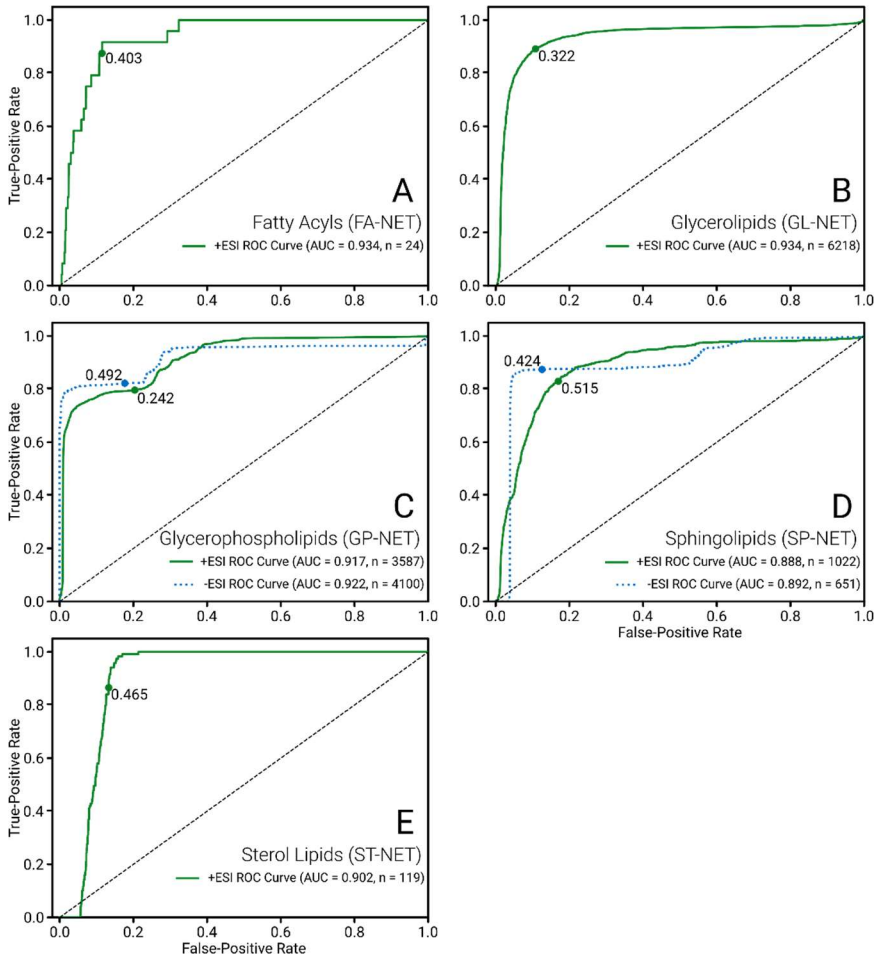


Figure 3.1. Receiver operator characteristic (ROC) curves for fatty acyls (A), glycerolipids (B), glycerophospholipids (C), sphingolipids (D), and sterol lipids (E) showing the performance of each iSF neural network for each class of lipid identified by LipiDex analysis. iSF inputs were constrained to those lipid features that produced high confidence LipiDex identifications (MS/MS Dot Product ≥ 700) and satisfied high isotopic purity standards ($R^2_1 \geq 0.9$, $R^2_2 \geq 0.9$, $|\Delta\Delta m/z| \leq 0.01$, $Intensity_A \geq 1e5$). Optimal score cut-offs are marked and labeled for each class and instrument polarity; positive instrument polarity is indicated by green solid lines and negative instrument polarity by dashed blue lines. The legend of each plot displays the instrument polarity (as + or -), area under the curve (AUC), and the number of true-positives (n) in the class. AUC is a general descriptor of classifier performance and is similar for those classes present in both positive- and negative-ion modes (i.e., glycerophospholipids and sphingolipids).

The GP-NET ROC curves (Figure 3.1C) were produced from 3587 GP positive-ion mode inputs (solid green) and 4100 GP negative-ion mode inputs (dotted blue). The GP positive-ion mode (Figure 3.1C, solid green) curve had an AUC of 0.917 with an optimal cut-off value of 0.242 (TPR% = 79.6, FPR% = 20.4). The GP negative-ion mode (Figure 3.1C, dotted blue) curve had an AUC of 0.922 with an optimal cut-off value of 0.492 (TPR% = 82.2, FPR% = 17.8). The SP-NET ROC curves (Figure 3.1D) were produced from 1022 SP positive-ion mode inputs (solid green) and 651 SP negative-ion mode inputs (dotted blue). The SP positive-ion mode (Figure 3.1D, solid green) curve had an AUC of 0.888 with an optimal cut-off value of 0.515 (TPR% = 83.0, FPR% = 17.0). The SP negative-ion mode (Figure 3.1D, dotted blue) curve had an AUC of 0.892 with an optimal cut-off value of 0.424 (TPR% = 87.4, FPR% = 12.6). The ST-NET ROC curve (Figure 3.1E) was produced from 119 ST positive-ion mode inputs (solid green), had an AUC of 0.902, and had an optimal cutoff value of 0.465 (TPR% = 86.6, FPR% = 13.3). As an example of a general application, iSF network performance characteristics applied to the same dataset with a cutoff of 0.5 are displayed in Table 3.2. As expected, application of a cutoff of 0.5 produced worse results than the optimized cutoff values for each class. Notably, the FA class inputs were predicted with a decreased TPR% of 45.83% in contrast with a TPR% of 87.5 with an optimized cutoff value of 0.403.

Table 3.2. iSF network performance metrics in application to inputs with high confidence LipiDex identifications (MS/MS Dot Product > 700), with high isotopic purity parameters ($R^2_1 > 0.9$, $R^2_2 > 0.9$, $|\Delta\Delta m/z| < 0.1$), and with a score cutoff at 0.5.

Metric	FA	GL	GP	PK	PL	SL	SP	ST
TPR% (n)	45.83 (11)	84.54 (5257)	79.96 (6417)	NA (0)	NA (0)	NA (0)	85.76 (1451)	80.67 (96)
TNR% (n)	96.92 (15560)	92.81 (9151)	91.33 (7355)	98.7 (15869)	91.67 (14739)	99.98 (16075)	84.29 (12126)	85.86 (13702)
FPR% (n)	3.08 (494)	7.19 (709)	8.67 (698)	1.3 (209)	8.33 (1339)	0.02 (3)	15.71 (2260)	14.14 (2257)
FNR% (n)	54.17 (13)	15.46 (961)	20.04 (1608)	NA (0)	NA (0)	NA (0)	14.24 (241)	19.33 (23)

In our initial demonstration of iSF, only positive-ion mode ESI-MS features were tested; however, the diversity of molecular adduct types produced in positive-ion mode suggested that slight changes in mass and composition would not significantly hinder classification [15]. In the present demonstration, comparison of iSF performance between positive- and negative-ion modes with glycerophospholipid and sphingolipid inputs showed general parity. As shown in Figure 3.1, iSF performed only slightly better in application to negative-ion mode inputs; the AUC for glycerophospholipids increased from 0.917 in positive-ion mode to 0.922 in negative-ion mode, and the AUC for sphingolipids increased from 0.888 in positive-ion mode to 0.892 in negative-ion mode. The shapes of the glycerophospholipid curves (Figure 3.1, bottom left) are more similar in shape relative to the shapes of the sphingolipid curves (Figure 3.1, bottom middle) because the subclass compositions (i.e., the relative portions of various lipid subclasses in the group of inputs) of positive- and negative-ion mode glycerophospholipids are more similar. For example, in both positive- and negative-ion mode, glycerophospholipid inputs that scored between 0.1-0.3

(roughly in the range where both curves "step-up"; Figure 3.1, bottom left), ~70% were lysophospholipids, which lack a fatty acyl chain compared to their diacyl and alkyl/acyl counterparts. Regarding the sphingolipid class, ceramides and hexosylceramides were the only sphingolipid subclasses detected in the negative-ion mode; sphingomyelins made up ~50% of sphingolipids identified in the positive-ion mode, and the rest were identified as ceramides and hexosylceramides.

3.4.2 Orthogonality of iSF Chemical Classification

Not all lipidomics identifications with LC-MS/MS are made with high confidence, and the confidence of such identifications depend on the quality of the MS/MS fragmentation spectra. Among the MS/MS identifications produced by LipiDex demonstration dataset, approximately 58% of identifications were produced with MS/MS dot product scores less than 700 (Appendix B.3), at which point LipiDex MS/MS identification performance begins to degrade[226]. Presumably, the probability for misassignment increases as the quality of the fragmentation spectra declines. However, classification by iSF should be unhindered even for low confidence MS/MS identifications provided that the quality of the MS isotopic pattern of the precursor ion is high (i.e., isotopic purity and S/N). We suggest that iSF classification could provide insight for such low confidence identifications.

Figure 3.2 displays the relationship between the LipiDex MS/MS Dot Product (binned by increments of 100) and the ROC AUC to show the level of "agreement" between iSF and LipiDex (i.e., the degree to which iSF predicted classifications match the classes determined by MS/MS identification).

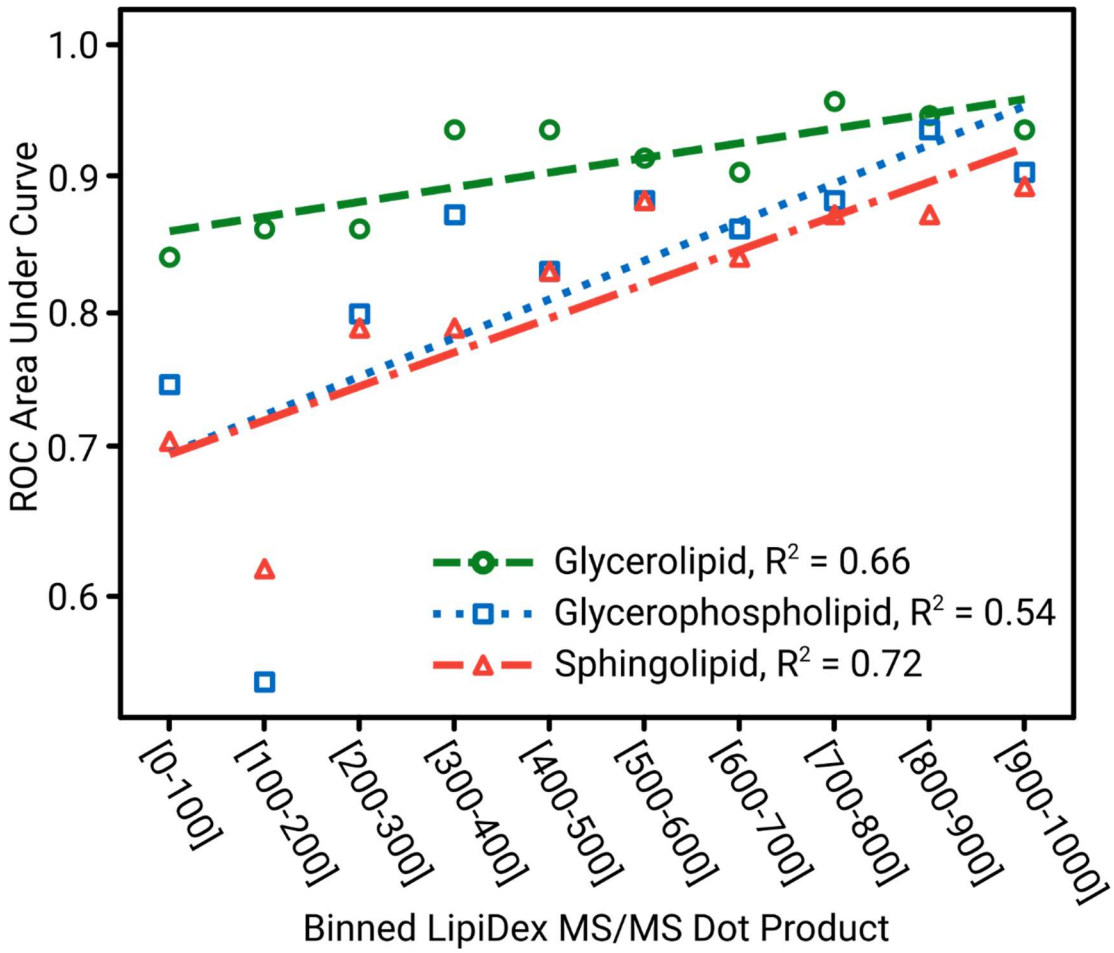


Figure 3.2. Scatter plot displaying receiver iSF neural network classifier operator characteristic (ROC) area under curve (AUC) as a function of LipiDex MS/MS Dot Product. iSF inputs were binned in increments of 100 dot product units. The fatty acyl and sterol lipid classes were omitted because these classes were not represented in most LipiDex MS/MS Dot Product bins. The glycerolipid (green circles), glycerophospholipid (blue squares), and sphingolipid (red triangles) classes are represented by at least 120 inputs per bin. Given that the true class of each lipid input is determined by the LipiDex identification and potential for LipiDex misidentification correlates inversely with Dot Product (as shown in Figure 3.2), ROC AUC is indicative of agreement between the iSF and LipiDex rather than true accuracy. For each class, the degree of "agreement" between the iSF predicted classifications and the expected classes as identified by LipiDex increases linearly (as confirmed by the R^2 score of the linear regression for each class) with respect to the LipiDex MS/MS Dot Product.

The MS/MS Dot Product score is a metric that describes the quality of MS/MS fragment spectral matching between experimental and reference spectra[226]. The LC-MS QC parameters of each iSF input included in Figure 3.2 were such that each LC-MS feature utilized was of high isotopic purity and sufficient intensity ($R^2_1 \geq 0.9$, $R^2_2 \geq 0.9$, $|\Delta\Delta m/z| \leq 0.01$, minimum intensity ≥ 1e5 counts). For each of the predicted lipid classes that had populated bins for each data point (i.e., glycerolipids (green circles), glycerophospholipids (blue squares), and sphingolipids (red triangles)), the ROC AUC was positively correlated with the LipiDex MS/MS Dot Product as shown for glycerolipids (green dashes, $R^2 = 0.66$), glycerophospholipids (blue dots, $R^2 = 0.54$), and sphingolipids (red dash-dots, $R^2 = 0.72$). Given that the QC parameters defining the quality of each included LC-MS feature, and therefore iSF input, were maintained, the correlation between iSF agreement and identification confidence cannot be primarily attributed to the degradation of iSF feature quality or classification performance; instead, it can be attributed to the increased propensity of lipid identification workflows to produce misidentifications associated with MS/MS spectra of low quality/confidence. Additionally, the ROC AUC for the 100-200 MS/MS Dot Product bin for glycerophospholipids and sphingolipids were notably lower than what the modeled linear relationship would suggest; this decrease in ROC AUC in the same bin was not observed for the glycerolipid class. Neither bin was underpopulated with 786 and 152 inputs for glycerophospholipids and sphingolipids, respectively, relative to the lowest number of inputs in any

glycerophospholipid and sphingolipid bin which was 357 and 124, respectively. Given the high QC constraints placed on the LC-MS features included in all bins, we suggest that the observed decrease in ROC AUC should be attributed to biases inherent to the LipiDex MS/MS identification workflow.

3.4.3 Differentiation Between Biological Sample Types

Distributions and biosynthesis of lipids vary significantly with respect to different biological sources[237-241]. Application of iSF to lipidomics LC-MS data provides a means to describe sample types by annotating species in terms of their respective class. The relative proportions of different identified lipid classes by a given LC-MS/MS workflow and sample preparation routine may then be utilized to discriminate between different sample types. Figure 3.3 displays the iSF-predicted class representation (%) of each lipid class identified by LipiDex analysis (i.e., glycerolipids, glycerophospholipids, sphingolipids, fatty acyls, and sterol lipids) for each biological source used to prepare lipidomics samples – Hap1 cells (red), liver tissue (green), plasma (yellow), and yeast (blue). Error bars are confidence intervals (confidence level = 95%) calculated for each set of measurement replicates (n = 3 for each positive- and negative-ion mode and each biological source; $n_{Total} = 24$).

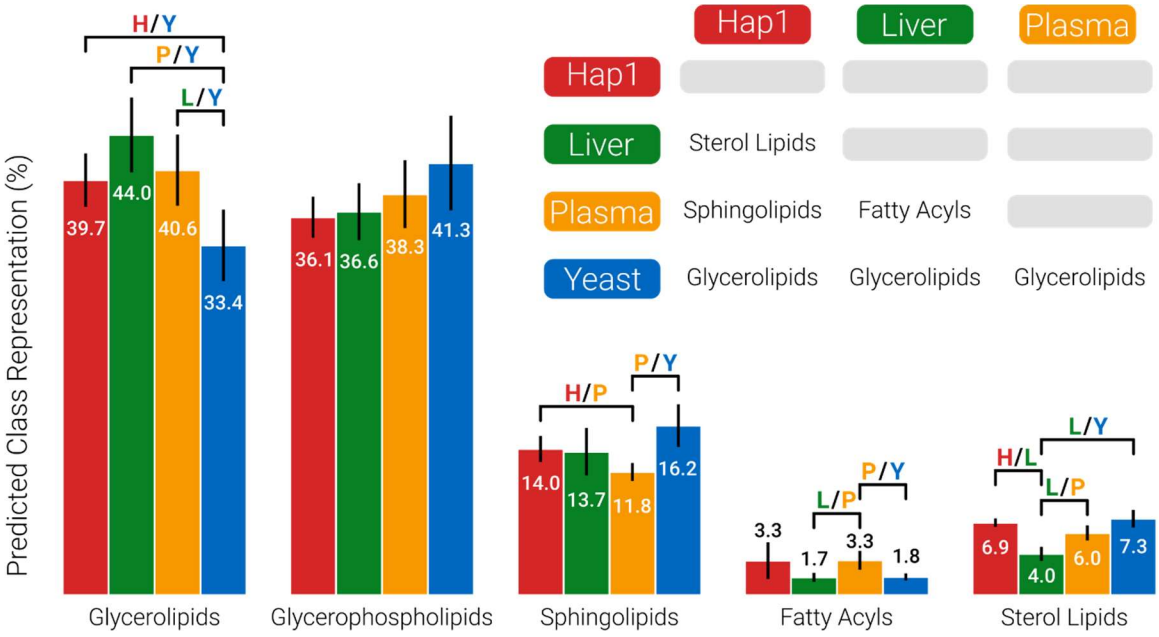


Figure 3.3. Bar charts displaying the predicted class representation for each of the sampled biological sources. Error bars are confidence intervals (confidence level = 95%) calculated for each set of measurement replicates (n = 3 for each polarity and each biological source; n_{Total} = 24). Confidence intervals were propagated when classifications from positive- and negative-ion mode from each biological source were accounted for. Each labeled bracket indicates two biological sample sources that are differentiated (defined by statistically significant difference by performing student t-test at the 95% confidence level) by the predicted lipid class representation and the label above each bracket denotes the two biological sources that were differentiated; biological sample source labels are H (Hap1), L (liver), P (plasma), and Y (yeast). The inset in the upper right depicts the primary discriminating lipid classes between each biological source.

Confidence intervals were propagated when classifications from positive- and negative-ion mode from each biological source were added; the class representation of each class for both positive and negative instrument polarities are displayed in Appendix B.4. Classes were considered discriminatory if tested positive for sign difference by the Student's *t*-test with respect to one biological source to another. By this approach, each biological source was differentiated by at least one lipid class, and the inset of Figure 3.3 displays the primary discriminating lipid classes between each biological source. Glycerolipids

differentiated Hap1 from yeast, plasma from yeast, and liver from yeast. Glycerophospholipids did not assist in differentiating sample types. Sphingolipids differentiated Hap1 from plasma and plasma from yeast. Fatty acyls differentiated liver from plasma and plasma from yeast. Sterol lipids differentiated Hap1 from liver, liver from plasma, and liver from yeast.

Sample annotation and description by iSF holds several distinct advantages. Though in this demonstration, all lipid features were previously identified by LipiDex MS/MS fragment spectral matching or exact mass assignment, iSF sample description does not require previous feature identification. Comparison of datasets/sample types for purposes of comparison and/or differentiation does require consistent experimental methodology (e.g., sample preparation, analyte separations, instrumental parameters, etc.). If all necessary analytical rigor is maintained, the quantitative results produced from iSF sample annotation are easily comparable, and qualitative conclusions about sample type are easy to draw. We expect that iSF sample annotation of lipidomics datasets may be useful for differentiating other clinically relevant biological sample types such as bacteria[242, 243] and cancer[244].

3.5 Conclusions

In this work, we demonstrated a means for sample annotation and description of complex lipidomics datasets via iSF classification. The accuracy of iSF prediction was characterized through assignment and testing against true-

positives given by high confidence LipiDex MS/MS fragmentation spectral matching. In general, each iSF network superseded 79% TPR for lipidomics data acquired with both positive- and negative-ion mode. Moreover, we demonstrated the importance of orthogonal means of class annotation and sample description in LC-MS and MS/MS workflows in which acceptance of unconfident assignments are likely to lead to erroneous annotations and sample metanalysis. Finally, we provided an initial demonstration of sample description and differentiation through comparison of iSF predicted class distributions in which each biological sample type was differentiated by at least one predicted class.

3.6 Acknowledgements

Funding: This work was supported by the National Science Foundation [grant numbers: IDBR-1455668 and CHE-1709526].

CHAPTER FOUR

Referenced Kendrick Mass Defect-Based Annotation and Filtering of Imaging MS Lipidomics Experiments

4.1 Abstract

Because of their diverse functionality in cells, lipids are of primary importance when characterizing molecular profiles in physiological and disease states. Imaging mass spectrometry (IMS) can provide the spatial distributions of lipid populations in tissues. Referenced Kendrick mass defect (RKMD) analysis is an effective mass spectrometry (MS) data analysis tool for classification and annotation of lipids. Herein, we extend the capabilities of RKMD analysis and demonstrate an integrated method for lipid annotation and chemical structure-based filtering for IMS datasets. Annotation of lipid features with lipid molecular class, radyl carbon chain length, and degree of unsaturation allows image reconstruction and visualization based on each chemical property. We show a proof-of-concept application of the method to a computationally generated IMS dataset and validate that the RKMD method is highly specific for lipid components in the presence of confounding background ions. Moreover, we demonstrate an application of the RKMD-based annotation and filtering to matrix-assisted laser desorption/ionization (MALDI) IMS lipidomic data from human kidney tissue analysis.

4.2 Introduction

Imaging mass spectrometry (IMS) provides valuable identity, abundance, and spatial distribution information for molecular components of complex biological tissues. Variety of IMS approaches have been used to explore molecular profiles of many biological systems and measure small metabolites,[245-247] lipids,[248-250] peptides,[60, 251, 252] glycans,[253, 254] and proteins.[255, 256] Among these molecular classes, lipids are essential for cell signaling, membrane composition, and metabolism[257-259] but are difficult to study by non-MS means such as immunostaining or transcriptomics. Matrix-assisted laser desorption/ionization (MALDI) IMS is a powerful tool to measure lipids at 10 μm spatial resolutions approaching the size of a mammalian cell[260]. In MALDI analyses, tissue sections between 5 and 20 μm are thaw mounted on conductive glass slides and uniformly covered with a chemical matrix that absorbs ultraviolet radiation and allows for ionization of lipids[60, 261, 262]. Ion intensities from mass spectra acquired from each laser position are correlated to produce spatially resolved ion images[26]. Because of the abundance and diversity of lipids, resultant IMS spectra can be congested[250]; some of the detected lipids can be isomeric and/or isobaric that are unresolvable by using high resolution MS instrumentation alone. Therefore, often ultrahigh mass resolving power instruments are used for isobar separation[130, 263] and other analysis dimensions such as ion mobility separation[250, 264], low energy CID[62], or chemical modification[265, 266] are utilized to assign double-bond position and

stereospecifically numbered (*sn*) position isomers. Given the direct biosynthetic relationships within families of lipids, methods that can identify lipids, link lipid families, and preserve their spatial distributions in tissues are essential for investigating lipid biochemistry.

KMD analysis is an approach that has been used to deduce families of chemically related compounds, such as lipids, using high resolution MS data in a variety of different fields of study[18, 20, 22]. In KMD analysis, the atomic mass unit reference is changed from ^{12}C to other groups, such as methylene (or CH_2, often using $^{12}C_1$ and $^{1}H_2$ isotopes for carbon and hydrogen atoms) or other units that repeat in polymer chain elongation. Thus, the Kendrick mass is the monoisotopic *m/z* value adjusted to the new reference and the resultant mass deficiency or defect, usually rounded to the nearest integer unit, can be used to discriminate molecular classes that contain varied mass deficiencies. Given that KM scale eliminates all CH_2 mass defect contributions, molecules such as lipids that differ by aliphatic chain length have the same KMD and those with differing degrees of unsaturation exhibit KMD differences of 0.01335 per unsaturation, which corresponds to the KMD of H_2. De Pauw et al. demonstrated a KMD-based IMS visualization tool that filtered MALDI-MS images based on lipid features clustered in KMD space[22]. Although molecular families could be grouped by untargeted clustering algorithms, analyte assignments were provided by exact mass matching, and molecular classes of clusters were inferred. As evidenced in this visualization tool, KMD analysis is well suited to lipidomics; however, the

more specialized referenced KMD (RKMD) approach can provide more direct information about lipid molecular families[23].

Lerno et al. demonstrated an adapted KMD analysis method, termed RKMD, that determined the class and degrees of unsaturation for lipidomics experiments[23]. In RKMD analysis, the reference KMD of a specified lipid headgroup is subtracted from the analyte KMD, and the difference is divided by 0.0134. Theoretically, if the resulting quotient is equal to integer value of zero or less, it is indicative of a positive classification for lipid molecular class with the specified head group. Moreover, the absolute value of the RKMD value is indicative of the degrees of unsaturation. However, mass measurement errors often preclude an error-free case, and thus RKMD values that predict correct chemical classes might not be an exact integer value. Additionally, the presence of confounding peaks in mass spectra (such as those from heavy isotopologues, MALDI matrix species, solvent clusters, and other molecular classes) present challenges for conventional RKMD analyses that lack controls to ensure specificity in lipid classification. Lerno et al. employed heuristic constraints that limited false-positive classifications but simultaneously narrowed the scope of application to lipids with less than or equal to six unsaturations, which comprise a narrower subset of lipids than might be detected during MS experiments. This provides an opportunity for method improvements to expand the coverage of the RKMD analysis to a wider subset of the lipidome.

Herein, we report a method for lipid feature annotation and class-based image filtering for lipidomics IMS data using an RKMD-based approach. We utilized both computationally generated and experimental MALDI-MS imaging datasets from human kidney tissue to assign lipid features via RKMD determination of lipid molecular classes, degrees of unsaturation, and numbers of radyl carbons. The latter is a novel extension of RKMD analysis that allows for increased method specificity and precision as well as lipid assignment. We show that class-specific spatial distributions of lipid populations can be used for automated image filtering and visualization of lipid descriptors such as molecular class, unsaturation, and radyl carbons. In previous approaches, spatial analyses depended on targeted identification of lipids by instrumental methods and user input to determine relationships in and between chemically related groups of lipids. In contrast, the presented method provides an integrated means for identification, annotation, and rapid visualization of related lipids in IMS datasets.

4.3 Materials and Methods

4.3.1 Sample Preparation

Human kidney tissue was surgically removed during a full nephrectomy and remnant tissue was processed for research purposes by the Cooperative Human Tissue Network at Vanderbilt University Medical Center. Remnant biospecimens were collected in compliance with the Cooperative Human Tissue

Network standard protocols and National Cancer Institute's Best Practices for the procurement of remnant surgical research material. The excised tissue was flash frozen over an isopentane, dry ice slurry, embedded in carboxymethylcellulose, and stored at −80 °C until use. Kidney tissue was cryosectioned to a 10 μm thickness, thaw mounted onto indium tin-oxide (ITO) coated glass slides (Delta Technologies, Loveland, CO, USA) for IMS analysis. Tissues were stored at -80 °C and returned to ~20 °C within a vacuum desiccator. IMS samples were coated with a 20 mg/mL solution of 1,5-diaminonapthalene dissolved in THF using an HTX TM M3 Sprayer (HTX Technologies, LLC, Chapel Hill, NC, USA) yielding a 1.67 mg/cm^2 coating (0.05 mL/hr, 5 passes, 40 °C spray nozzle). Tissue samples underwent IMS analysis immediately after matrix deposition.

4.3.2 MALDI timsTOF IMS

MALDI IMS was performed on a Bruker timsTOF pro MS system[260] (Bruker Daltonics, Bremen, Germany) in quadrupole-time of flight (Q-TOF) only analysis mode. The qTOF ion images were collected in positive-ion mode at 10 μm pixel size. The laser beam scan was set to 6 μm^2 and 200 laser (λ = 266 nm) shots per pixel at 10 kHz was used for laser desorption and 18.6% laser power (30% global attenuator and 62% local laser power). Mass spectrometry data were collected from *m/z* 50 – 2000 in centroid mode for lipid analysis. Lipids were identified using a combination of mass accuracy (≤ 3 ppm) and LIPIDMAPS[267, 268]

database searching. Only even chained lipids were considered because it is well known that mammalian systems do not generally produce odd-chain lipids, except in special circumstances[269, 270].

4.3.3 Computational Generation of IMS Data

Theoretical isotopic envelopes for lipids, MALDI matrix clusters, and peptide ions were calculated using the pyOpenMS (2.6.0) Python package to provide a proof-of-concept for and test the specificity and precision of the RKMD-based method. Peptides and MALDI matrix clusters were used to test the specificity of the method for lipids in the presence of confounding species. Lipid chemical formulas were acquired from the LIPIDMAPS structure database. Each lipid isotopic envelope was generated from the chemical formula of the protonated, singly-charged molecular ion and data for three isotopologues were calculated and used in subsequent analyses. The isotopic envelopes for MALDI matrix (M) cluster ions were calculated for monomeric $[M+H]^+$ and proton bound dimeric $[2M+H]^+$, trimeric $[3M+H]^+$, and tetrameric $[4M+H]^+$ ion clusters of 2,5-dihydroxybenzoic acid (DHB), α-cyano-4-hydroxycinnamic acid (CHCA), and DAN. In addition to inclusion of the isotopic envelopes for singly-charged protonated matrix molecule ions and clusters, isotopic envelopes for fragment ions resulting from common neutral losses (H_2O and CO_2 from DHB and CHCA and NH_3 from DAN) as well as sodium and potassium adducts were included as potential confounders; sodiated and potassiated cluster ions were generated

according to the rules described by Keller et al.[271] DHB, CHCA, and DAN are three common choices for MALDI matrix in positive mode lipidomics MALDI-IMS experiments and provide good ionization for a variety of lipid classes[272]. Peptide chemical formulas were converted from randomly generated peptide sequences with chain lengths between 1 and 25 amino acids. Each peptide isotopic envelope was synthetically generated from singly-charged and protonated species and included seven isotopologue peaks. Continuum mass spectra (i.e., with multiple sampled points over each peak) were generated by calculating the gaussian distribution of each isotopologue along an *m/z* axis from *m/z* 100-1500. The *m/z* centroid and relative isotopic abundance values were input for the mean and amplitude in the gaussian function. Given that resolving power is fixed across the mass range in TOF instrumentation[273], the gaussian sigma parameter was held constant across the *m/z* range and produced peaks with mass resolving powers ($m/\Delta m_{50\%}$) ranging from ~55,000 to 65,000.

To test the specificity and precision of the RKMD annotation method, five MS datasets consisting of theoretical centroid *m/z* values were generated for each type of molecular class tested, viz. protonated, sodiated, and potassiated lipids, MALDI matrix clusters, and peptides. Lipid components included protonated, sodiated, and potassiated ions of 500 lipids, including 62 phosphatidylcholine (PC), 60 phosphatidylethanolamine (PE), 61 phosphatidic acid (PA), 62 phosphatidylglycerol (PG), 59 diacylglycerol (DG), 22 sphingomyelin (SM), 32 triacylglycerol (TG), 40 ether-linked (O/P-) PC, 35 O/P-

PE, 33 O/P-PA, and 34 O/P-PG. Lipids were chosen pseudo-randomly from the LIPIDMAPS Structure Database (LMSD) and had even chains between 28 and 50 radyl carbons and between 0 and 9 unsaturations from the abovementioned 11 lipid classes. The peptide ion spectrum was generated from 1000 unique peptide sequences, resulting in 7000 total peaks. The MALDI matrix ion spectrum was generated from 1305 matrix cluster species, resulting in 3915 total peaks.

The computationally generated IMS dataset was based on a .PNG image depicting the letters "BU & VU", and each letter had a unique set of RGB color values. The RGB color values in the image were associated with collections of lipid ion isotopic envelopes related by lipid molecular class, degree of unsaturation, or number of radyl carbons. At each coordinate in an equivalently sized 2-dimensional array, a spectrum extending in the 3rd dimension was generated using the lipid species of the appropriate class. At "BU" coordinates were spectra containing MS isotopic envelopes from 86 O/P-PG lipids with 0-6 double bonds and 28-40 radyl carbons (even radyl carbon chains only). However, "&" coordinates included spectra containing MS peaks from 216 MS isotopic envelopes from lipids with 4 double bonds in the PC, PA, PG, and DG classes and with 28-40 radyl carbons (even radyl carbon chains only). Finally, "VU" coordinates included spectra that contained MS isotopic envelopes from 257 lipids with 34 radyl carbons in the PC, PA, PG, and DG classes and with 0-6 double bonds.

4.3.4 MALDI-IMS Data Preprocessing

Bruker MALDI-IMS data was converted to the imzML file format prior to peak picking and then to a native Python dictionary structure with custom Python (3.8.5, CPython, Python Foundation) scripts. To make the data amenable to numpy array operations and matplotlib image visualization in Python, data for unsampled coordinates between the maximum x and y image coordinate were filled with an empty spectrum. An internal, quadratic recalibration of the summed spectrum was performed using six common lipid features, $[PC(32:0)+H]^+$, $[PC(34:1)+H]^+$, $[PC(34:1)+Na]^+$, $[PC(36:1)+H]^+$, $[PC(34:1)+K]^+$, and $[PC(36:1)+Na]^+$, resulting in < 3 ppm error. Each peak in each MALDI-MS spectrum was aligned to the recalibrated summed spectrum.

4.3.5 RKMD-Based Lipid Annotation

Overall, we annotate lipids with their sum compositions using an RKMD-based workflow that uses mass spectrometry data to assign lipid sum compositions, namely, with headgroup, radyl carbon chain length, and unsaturation information. A representative schematic of the lipid annotation using this RKMD approach is included in Supporting Information (Appendix C.1). Firstly, the synthetically generated or experimentally acquired IMS dataset was input to the annotation workflow in a Python dictionary structure. On a per pixel basis, the centroid spectrum was read and aligned to the recalibrated summed spectrum. A recalibrated summed spectrum was used to bin *m/z* values

and enhance mass measurement accuracy by recalibrating the average *m/z* values in all acquired mass spectra. Once the mass spectrum was realigned, RKMD analysis was performed for each peak in the spectrum for the molecular class headgroup and adduct reference KMD. Headgroup elemental compositions used to calculate the reference KMD for each class and commonly observed adducts for each class are listed in Supporting Information (Appendix C.2 and C.3, respectively). Specifically, the reference KMD of the adducted headgroup of each lipid class was calculated and subtracted from the experimentally acquired KMD value. The resulting difference is then divided by 0.0134 (CH_2-based Kendrick mass defect of carbon) to produce the RKMD value[23].

For each calculated RKMD value, its distance from the closest integer value (δ) was determined. The features that produced an RKMD δ within a user-defined window (δ = 0.35 in this work) for RKMD values between 0 to -9 (corresponding to 0 and 9 unsaturations, respectively) were considered potential positive annotations for the class-of-interest; features that did not meet these criteria were excluded from further processing. The corresponding headgroup and unsaturation information were used to calculate the number of radyl carbons for each potential positive classification. Analogous to using δ acceptance windows, the distance from the calculated integer values indicating numbers of radyl carbons (ε) was used to exclude erroneous classifications. Peaks with radyl carbon ε values greater than 0.001 were excluded from

downstream processing steps as true positive identifications were found to have radyl carbon $\varepsilon \leq 0.001$. Positive integer results were considered unacceptable results not in agreement with physical reality[23]. For each potential annotation, *m/z*, lipid class, adduct, number of radyl carbons, degree of unsaturation, radyl carbon ε, RKMD δ, even number radyl carbons (as true for even numbers and false for odd numbers), and peak intensity were stored in a Python dictionary. This process is repeated for every molecular class headgroup and adduct of interest.

4.3.6 Image Filtering and Heuristic Constraints

Lipid distributions were then visualized based on lipid classes defined by similarities in lipid headgroup, degrees of unsaturation, and number of radyl carbons. As an example, a filter for the O/P-PE class is applied for RKMD assignments of *m/z* 790.5151, which include $[PC(O/P\text{-}35:5)+K]^+$, $[PE(O/P\text{-}38:5)+K]^+$, $[LPC(35:5)+K]^+$, and $[LPE(38:5)+K]^+$ (Appendix C.4). Firstly, assignments were rank ordered by ascending RKMD δ values for each peak at each pixel. To limit false positive identifications, several heuristic constraints were applied. Lower and upper limits were placed on the numbers of radyl carbons and degrees of unsaturation that were accepted for each lipid molecular class (Appendix C.5). Limits were based on commonly observed fatty acids[274] and radyl carbon chain lengths for each lipid molecular class in MALDI-IMS tissue analyses[249, 275]. Molecular class-specific degree of unsaturation limits were

necessary to limit false-positive identifications in which unrealistically high degrees of unsaturation were calculated as noted by Lerno et al.[23] Additionally, in this work, odd numbered of radyl carbon chains were excluded given that odd chain fatty acids are uncommon in human tissues[269, 270]. After excluding potential annotations by heuristic constraints, the top ranked assignment was compared to the filter criterion, and the RKMD δ was compared to an *m/z* value dependent error limit. If the filter criterion matched the assignment and the RKMD δ was below the error limit, the feature intensity was added to the pixel intensity of the filtered image. In this demonstration, the RKMD δ error limit was calculated by $r = \frac{p}{13415m^{-1}}$ where *r* is the error limit expressed in terms of RKMD δ, *p* is the error limit expressed in terms of ppm error, *m* is the *m/z* value of the feature, and 13415 is a constant that relates RKMD δ to ppm error to approximate a 2.5 ppm error threshold. The relationship between RKMD δ and ppm error is inversely related (Appendix C.6).

4.4 Results and Discussion

4.4.1 RKMD-Based Annotation and Filtering of Computationally Generated IMS Data

Data processing and analysis methods in IMS have advanced significantly in recent years to accelerate analysis of large data volumes. Imaging data analysis in MS is often conducted manually by selective visualization of *m/z* values or by unsupervised data reduction (e.g., principal component analysis[276])

and/or segmentation approaches[277] that group pixels/spectra by similarity[278]. However, manual analysis can be time consuming, and unsupervised analyses do not describe the relationships between pixel groups or may produce uninterpretable results[277]. Biologically relevant conclusions are therefore dependent on accurate annotation of molecular species in biomarker discovery workflows. Chemical class annotation is useful to analyze global trends in data, and one attractive option for lipid class annotation is the RKMD method[23]. However, in conventional RKMD, there is a potential for false-positive classifications from confounding ions because the only criterion for chemical classification is an acceptance window for RKMD values (determined by mass measurement error). To address this drawback and adapt RKMD for imaging applications, we implemented an additional data curation criterion to reduce potential incorrect results and expanded the RKMD approach's analytical capabilities from classification to full sum composition lipid annotation.

In this approach, we increase the specificity and precision of RKMD-based annotation via exclusion of true-negative peaks based on the distance from calculated radyl carbon integer values. The specificity and precision of the RKMD-based annotation method are demonstrated in application to a computationally generated complex dataset containing lipids and potentially confounding species, including MALDI matrix cluster and peptide ions (Figure 4.1). Confounding species were included to evaluate the performance of the approach for their effective exclusion.

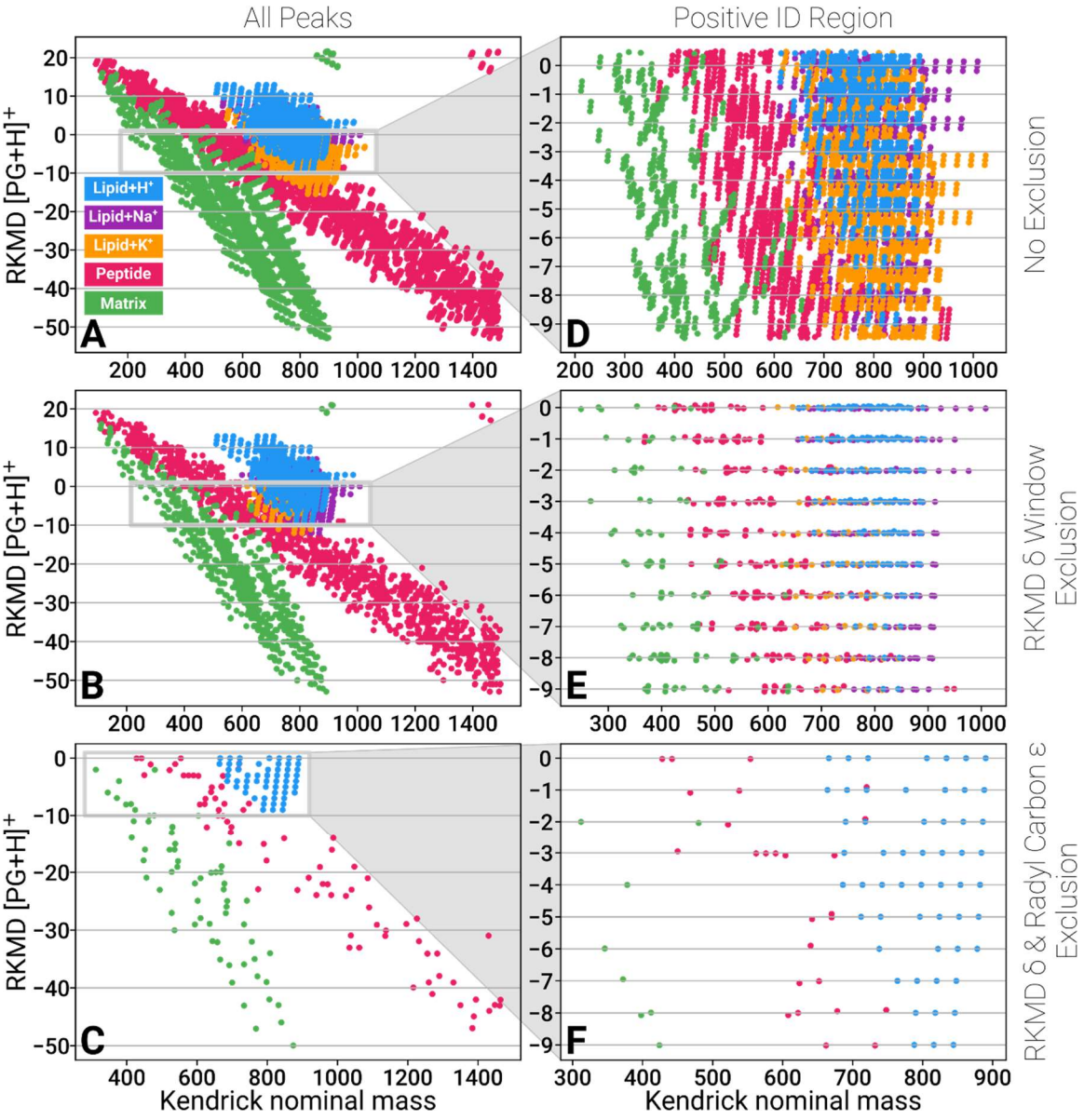


Figure 4.1. Computationally generated $[PG+H]^+$ RKMD plots A, B, and C with their respective zoom regions in D, E, and F demonstrate the utility of using data curation parameters, RKMD δ and radyl carbon ε exclusion windows, to enhance specificity and precision of the RKMD-based annotation method in the presence of protonated lipids (blue, 500 w/ three isotopes), sodiated lipids (purple, 500 w/ three isotopes), potassiated lipids (orange, 500 w/ three isotopes), peptides (pink, 1000 w/ seven isotopes), and MALDI matrix clusters (green, 1305 w/ three isotopes). The top row plots (A, D) include all datapoints, Plots in the second row (B, E) include datapoints with $\delta \leq 0.1$, and plots in the third row (C, F) include datapoints with $\delta \leq 0.1$ and $\varepsilon \leq 0.001$.

MALDI matrix cluster ions are often observed as background ions in MALDI experiments[279]. Likewise, peptides are potential confounders in tissue IMS lipidomics; although they are not often detected concurrently with lipids in tissue IMS experiments except in single cell[251] and small metabolite[246] analyses.

The computationally generated ions were subjected to RKMD analysis and results as are displayed as plots of RKMD as a function of KNM (Figure 4.1 A-F). Each row of plots was subjected to a different level of data curation. In this work, [PG+H]+ was chosen as a reference lipid headgroup as it exhibited the lowest specificity of all classes included in the dataset. In Figure 4.1, MALDI matrix cluster ions are displayed in green, peptide ions are displayed in pink, and lipids are shown in three different colors of blue (protonated), purple (sodiated), and orange (potassiated). Plots labeled "All Peaks" (Figure 4.1A-C) contain the entire dataset whereas plots labeled "Positive ID Region" (Figure 4.1D-F) show the relevant regions for RKMD classifications. Prior to any data curation, 1112 peptide, 498 MALDI matrix cluster, 2596 true-negative lipid MS datapoints were observed in the zoomed positive ID region (Figure 4.1D).

For comparison, lipid, matrix cluster, and peptide data were first curated by RKMD δ exclusion only with a window of 0.1 or ~1.9 ppm mass error (Figure 4.1B & E). This case reflects a conventional application of RKMD wherein retained datapoints would indicate positively classified species for the specified headgroup. By imposing the RKMD δ exclusion value of 0.1, 20.1% (1410) of the 7000 peptide datapoints in the total space were retained (Figure 4.1B), and 20.6%

(229) of the 1112 peptide datapoints in the positive ID region were retained (pink, Figure 4.1E). Similarly, 20.4% (799) of the 3915 MALDI matrix cluster datapoints in the total space were retained (Figure 4.1B), and 19.5% (97) of the 498 matrix cluster datapoints in the positive ID region were retained (green, Figure 4.1E). Of the retained lipid datapoints, 55 corresponded to $[PG+H]^+$ monoisotopic peaks, and 358 corresponded to heavy isotopologues and/or peaks from other lipid molecular classes. In the total RKMD space (Figure 4.1B), the specificity for correct exclusion of non-$[PG+H]^+$ monoisotopologues was 83.3% (ratio of true-negative indications to all negatives), and the precision (ratio of true-positive indications to all positive indications) imparted by RKMD δ exclusion was 2.1%. In the positive ID region, these numbers for the specificity and precision improved to 89.9% and 11.4%, respectively (Figure 4.1E). Although a significant portion of the confounders were excluded by utilizing an RKMD δ window (Figure 4.1E), this conventional approach lacks the desired level of specificity and precision for confident lipid annotation.

To demonstrate the enhancement provided in the presented RKMD-based annotation workflow, the number of radyl carbons were calculated for each feature assuming a $[PG+H]^+$ headgroup, and data were curated by a radyl carbon ε exclusion window of 0.001 in addition to an RKMD δ exclusion window of 0.1 (Figures 4.1C & F). Application of radyl carbon ε exclusion with a window of 0.001 significantly decreased the number of retained peptide datapoints from 1410 (peptides retained by RKMD δ exclusion only) to only 79 (1.1% of 7000 total)

and matrix cluster peaks from 773 to 64 (1.6% of 3915 total) in the total space (Figure 1C). In the RKMD positive ID region (Figure 4.1F), radyl carbon ε exclusion decreased retained peptide datapoints from 229 to 26 (2.3% of 1112) and MALDI matrix cluster datapoints from 97 to 8 (1.6% of 498). All potential lipid false-positives were eliminated, leaving only the 55 peaks corresponding to the $^{12}C_{all}$ isotopologues of $[PG+H]^+$ components (Figure 4.1C & F). This corresponds to a true-positive rate of 100% for positive identification of all $[PG+H]^+$ lipids. In the positive ID region (Figure 4.1F), specificity was increased to 98.8% (from 89.9% for RKMD δ exclusion only) and precision to 78.6% (from 11.4% with only RKMD δ). The exclusion of a most matrix cluster ions and peptides (98.4% and 97.7%, respectively) suggest that the method is robust in excluding non-lipid components (Figure 4.1).

Moreover, the observed enhancement in this new approach enabled assignment of highly unsaturated lipids with greater confidence, relative to conventional RKMD δ windowing exclusion approach via elimination of false-positive lipid assignments from both heavy isotopologue peaks and monoisotopic peaks of other classes. For instance, when solving for $[PC+H]^+$ RKMD values, $[PA(34:1)+K]^+$, $[PA(38:4)+K]^+$, and $[PA(36:6)+K]^+$ monoisotopic peaks produce low RKMD δ values at integers -7, -10, and -12, respectively (purple diamonds, Appendix C.1, node III), and therefore reduce confidence in highly unsaturated PC assignments in conventional RKMD analyses. However, solving for each $[PA+K]^+$ components radyl carbon chain length (assuming a

[PC+H]$^+$ headgroup) produces large radyl carbon ε values exceeding the threshold of 0.001 used in this work (purple diamonds, Appendix C.1, node IV), restoring confidence to highly unsaturated [PC+H]$^+$ assignments.

As a proof-of-concept, the RKMD-based annotation and filtering method was applied to a computationally generated IMS dataset comprised of theoretical MS peaks (Figure 4.2**)**. The total ion current (TIC) image of the dataset (Figure 4.2B) displays contributions from 559 protonated lipid MS peaks spatially arranged to display the text "BU & VU". The summed mass spectrum of all included lipid components is displayed on top of the total ion current (TIC) image and is notably complex with regions of severe congestion (Figure 4.2A). RKMD-based annotation correctly assigned each lipid, and the filtering method reconstructed each select class image using RKMD δ and radyl carbon ε exclusion windows of 0.1 and 0.001, respectively. Specifically, reconstructed images for 86 ether-linked phosphatidylglycerol lipids at "BU" coordinates (Figure 4.2C), 216 lipids with four degrees of unsaturation from PC, PA, PG, and DG chemical classes at "&" coordinates (Figure 4.2D), and 257 lipid features with 34 radyl carbons from PC, PA, PG, and DG chemical classes at "VU" coordinates (Figure 4.2E) are shown in Figure 4.2C-E. Each SC image represents a filtering mode that utilizes a different criterion, namely lipid chemical class (Figure 4.2C), degrees of unsaturation (Figure 4.2D), and number of radyl carbons (Figure 4.2E).

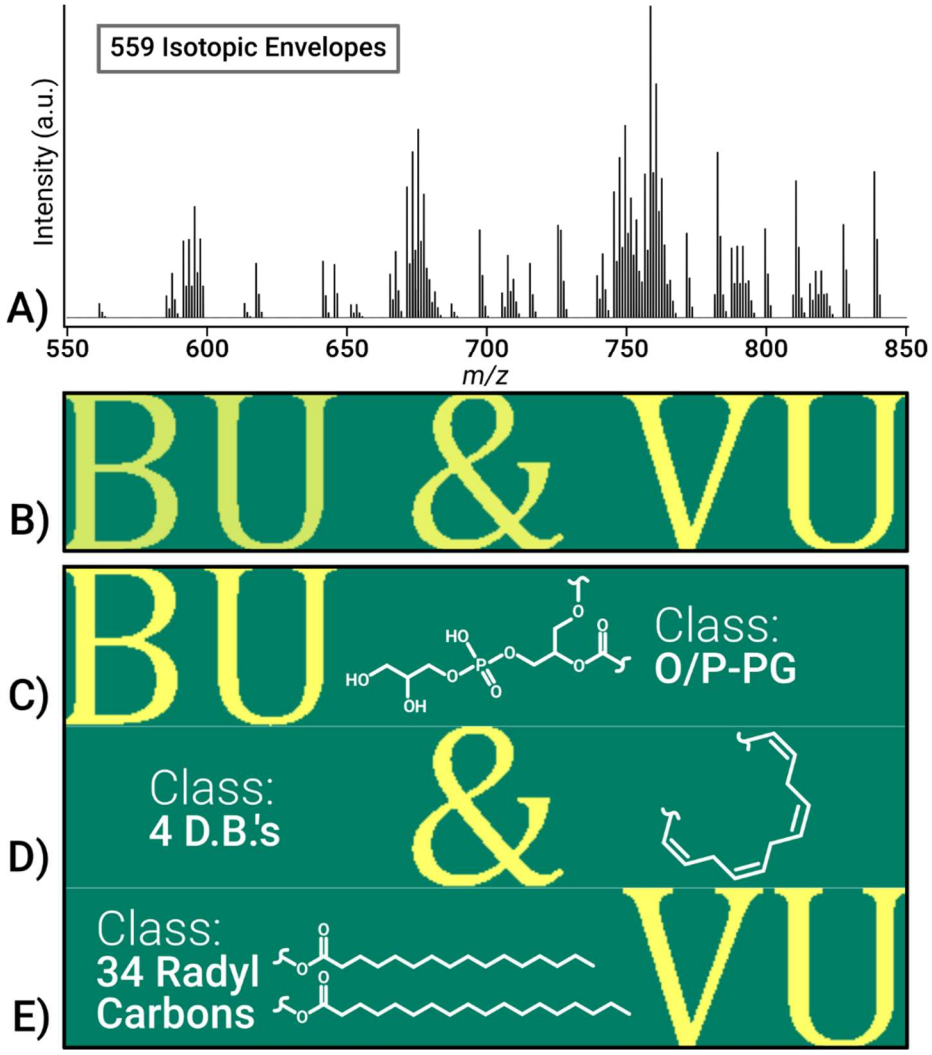


Figure 4.2. A computationally generated summed mass spectrum (A) for an MS dataset from 37044 pixels that included 559 lipids with pseudo-randomized relative abundances was used to generate the total ion image show in (B) and a series of RKMD-based filtered mass spectrometry images (C-E). Total ion current (TIC) image (84 x 441 pixels) in B depicts the summed intensity for each coordinate. The selected class (SC) images were filtered based on molecular class, degrees of unsaturation, and radyl carbon chain length, respectively. The molecular class image (C) was filtered for ether-linked phosphatidylglycerol (O/P-PG) lipids; the dataset included 86 O/P-PG lipids at the "BU" coordinates. The degree of unsaturation image (D) was filtered for lipids containing 4 unsaturations; this dataset included 216 lipids containing PC, PA, PG, or DG headgroups and all contained four double bonds at the "&" coordinates. The radyl carbon chain length image (E) was filtered for lipids with 34 radyl carbons; the dataset included 257 lipids with 34 radyl carbons from the same four molecular classes.

The RKMD-based reconstructed images demonstrate that this class-based filtering approach can be used to ascertain the character and localization of related groups of lipids in IMS data (Figure 4.2). To evaluate the utility of the RKMD-based lipid annotation for tissue image reconstruction, we analyzed MALDI-IMS data from a human kidney tissue and highlighted the advantages of a class-based approach for spatial tissue characterization.

4.4.2 MALDI-IMS of Kidney Tissue Lipids

The presented workflow was applied to a MALDI tissue imaging analysis of human kidney lipids. Lipids are of primary importance to the healthy functioning of kidney tissues and characterization of renal disease[280, 281]. MALDI-MS has enabled detailed interrogations into the spatial distribution and composition of different lipids in human kidney tissues that have provided key insights into physiological and disease mechanisms[282-284]. The imaged kidney section (220,000 pixels) contains portions of the medulla and cortex. Subsections of these regions are visible at varying degrees in all class-based images, such as medullary rays, tubules, collecting ducts, vasculature, and glomeruli (Figure 4.3). A composite of all saturated lipids and monounsaturated PC lipids are shown in Figure 4.3 (top and bottom, respectively). To confirm the accuracy of the RKMD-based annotation method, 44 *m/z* values that resulted in identifications made by a combination of mass accuracy and LIPIDMAPS database searching were submitted to RKMD-based annotation.

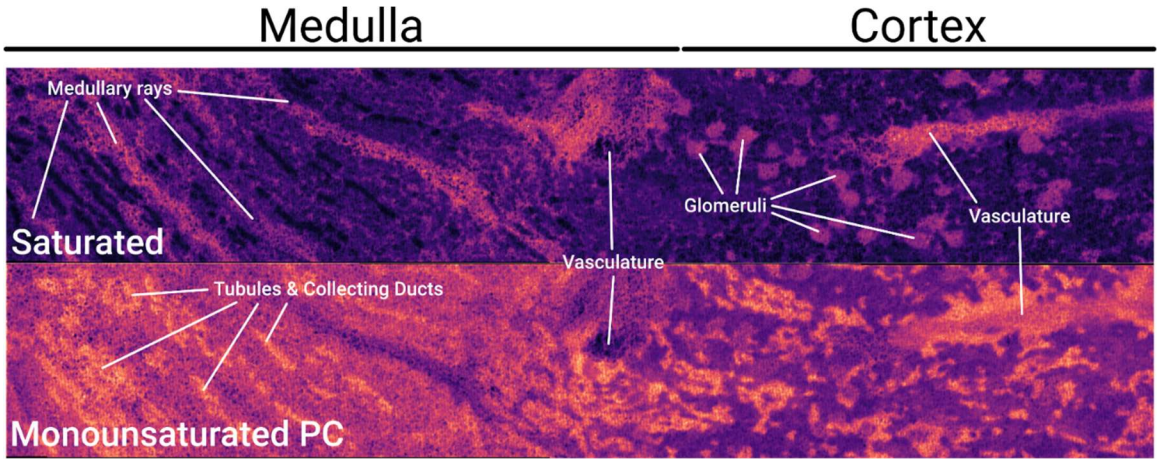


Figure 4.3. Labeled renal tissue structures spanning parts of the medulla and cortex region in class composite images depicting saturated (top) and monounsaturated PC (bottom).

The RKMD-based method produced equivalent assignments in each case after application of the heuristic constraints used in the presented image filtering workflow (Appendix C.7).

Observed lipids from MALDI-IMS of kidney tissue from several different molecular classes were detected and assigned by the RKMD workflow. At the highest level, resultant images are composites of all assigned lipid components that are grouped by molecular class, unsaturation, and radyl carbon chain length (Appendix C.8). Although the high-level class composite images may be useful to evaluate broad differences in lipid class distributions in tissue, localization of related lipids can vary significantly with respect to other characteristics, such as localization of a lipid class with varying radyl carbon chain length or degree of unsaturation. Some lipid isomers can even have differing spatial distributions in tissues; however, these differences cannot be visualized without an orthogonal dimension of separation such as ion

mobility[250]. However, in the interest of preserving spatial information, all molecular metadata for each component was retained such that subclass images of more specific groupings of lipids could be easily reconstructed and compared to evaluate localization of lipid classes with finer differences.

For example, lipid distributions corresponding to saturated and monounsaturated PC and PE and mono- and diunsaturated SM components were evaluated (Figure 4.4). Images of saturated PC (Figure 4.4A) and PE (Figure 4.4B) and monounsaturated SM (Figure 4.4C) components are highly colocalized in the kidney tissue, showing high abundance within the glomeruli, vasculature, and medullary rays (Figure 4.3). Previously, SM lipids have been characterized throughout the renal cortex and medulla; moreover, studies have characterized localization of monounsaturated SM to glomeruli in healthy rat subjects[285] and in diabetic mouse subjects in response to a high fat diet[282]. In contrast, images of monounsaturated PC and PE (Figure 4.4D & E) show different spatial distributions. Although some low signal may be observed from the glomeruli in the cortex, signal arises primarily from the surrounding structures. In the medulla, monounsaturated PC and PE (Figure 4.4D & E) are colocalized to the renal tubules and their collecting ducts; in the cortex, monounsaturated PC and PE highlight elements of the renal cortical labyrinth that surround the glomeruli. Concerning the vasculature, saturated (Figure 4.4A & B) and unsaturated PC and PE (Figure 4.4D & E) are negatively correlated.

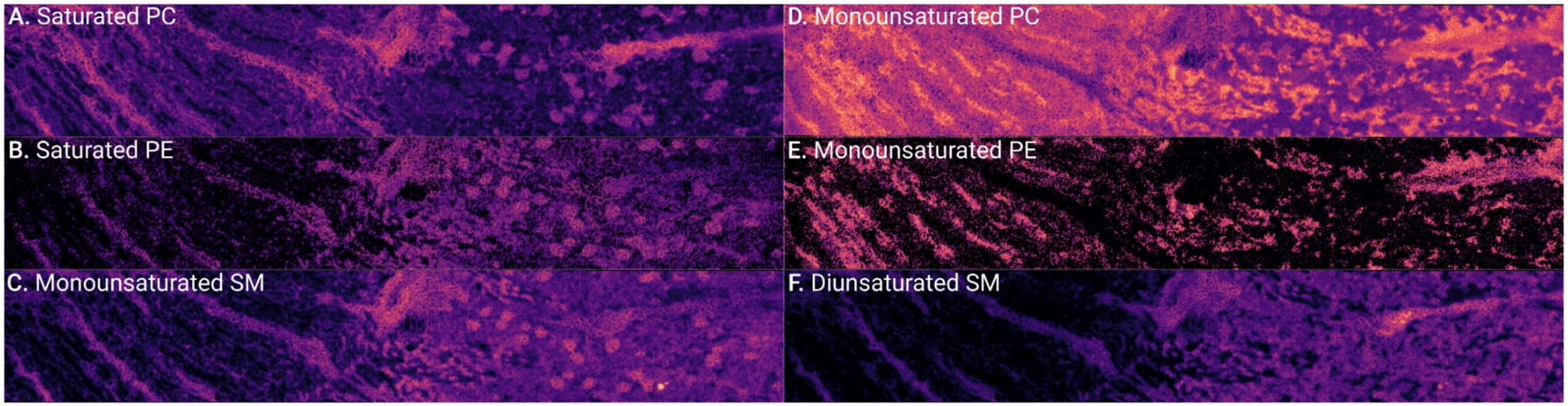


Figure 4.4. RKMD-based filtering applied to a MALDI-IMS dataset from a human kidney section with medulla and cortex visible in all images: A) saturated PC, B) saturated PE, C) monounsaturated SM, D) monounsaturated PC, E) monounsaturated PE, and F) diunsaturated SM. Images generated by RKMD-based filtering can be used to rapidly determine lipid trends among functional regions. For instance, glomeruli possess higher levels of saturated PC and PE lipids (A-B) and monounsaturated SM (C), and elements of the vasculature and medulla are represented very differently by saturated PC/PE (A and B) and SM (C and F) compared to monounsaturated PC/PE (D and E).

Saturated PC and PE colocalized to structures above the artery, and unsaturated PC and PE have higher abundance in the surrounding tissues. In contrast, diunsaturated SM (Figure 4.4F) maintains localization to the medullary rays and vasculature relative to saturated PE, PC, and monounsaturated SM; however, this lipid class is completely absent from glomeruli.

In the interest of further characterizing the behavior of SM localization, we grouped and displayed SM subclasses by radyl carbon chain lengths (Figure 4.5). Each image is a composite of at least three SM components (except for SM with 38 radyl carbons (Figure 4.5C) which has two). The 38 radyl carbon SM composite image was included to show continuity in the progression of increasing chain lengths in the SM class. Each image shows conservation of some features including the medullary rays, vasculature, and tubules. For example, 34 radyl carbon SM (Figure 4.5A) and 42 radyl carbon SM (Figure 4.5E) are uniquely colocalized to the glomeruli. Although 42 radyl carbon SM components produced less signal in these regions (as compared to Figure 4.5A for 34 radyl carbon SM), they contrasted with 36, 38, and 40 radyl carbon SM subclasses (Figure 4.5B-D), which are completely absent in the glomeruli. Based on the localization of monounsaturated SM to the glomeruli (Figure 4.4C), we presumed and confirmed that a major component was SM(34:1), which was characterized previously as an important mediator for ATP production in glomeruli[282].

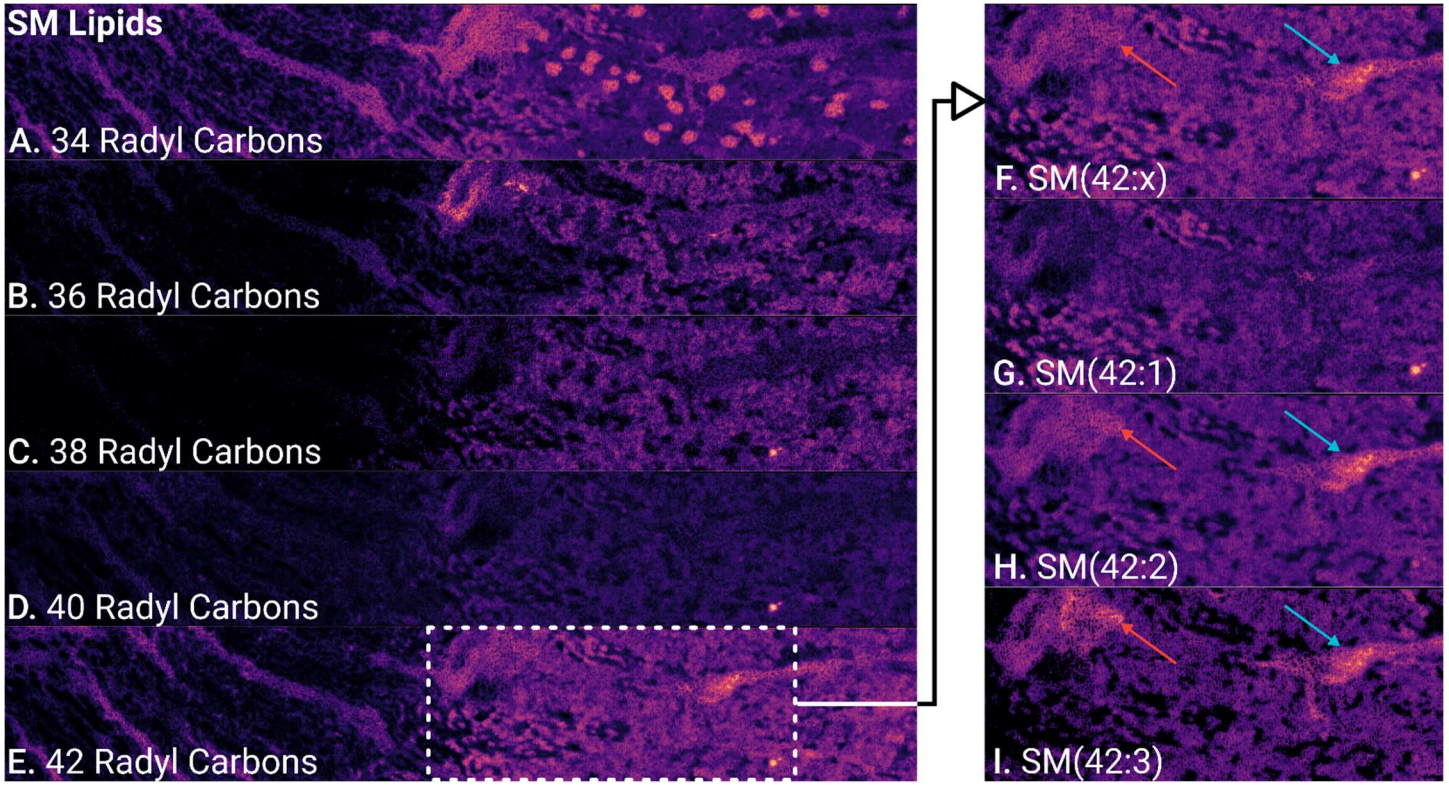


Figure 4.5. RKMD-based filtering applied to a MALDI-IMS dataset from a human kidney section with 34, 36, 38, 40, and 42 radyl carbons in A-E followed by an enlarged region in the cortex for SM with 42 radyl carbons (F) with further classifications by the degree of unsaturation for 1(G), 2 (H), and 3 (I) unsaturations. Red and blue arrows indicate vascular structures represented in the 42 radyl carbon SM composite (F) that are colocalized to SM(42:2) (H) and SM(42:3) (I) and absent in SM(42:1). Images were used to further characterize SM localization in the glomeruli and reducing class composite images enables localization of discrete sum compositions and attribution of observed morphological features to more specific groups of lipids.

The RKMD-based image filtering approach can be applied to any group or subgroup of lipids with increasing specificity down to individual lipid sum compositions, reducing composite images to observe colocalization of very closely related lipids. Of course, lipid sum composition does not describe *sn* position or double bond position/geometry but is the greatest level of specificity provided by single-stage MS measurement and the RKMD-based method. For example, the 42 radyl carbon SM composite class image (Figure 4.5E & F) was reduced to the spatial distributions of the three contributing sum compositions, SM(42:1), SM(42:2), and SM(42:3), were visualized in an enlarged region of the cortex (Figure 4.5G-I). The red and blue arrows (Figure 4.5F, H, & I) indicate vascular structures observed in the 42 radyl carbon SM composite image (Figure 4.5F) that were observed with higher intensity relative to the surrounding tissue in the SM(42:2) and SM(42:3) (Figure 4.5H & I, respectively); this is in contrast to the SM(42:1) image (Figure 4.5G) in which signal intensities for the indicated areas are similar to the surrounding tissue.

The RKMD-based annotation and image filtering approach provides the framework for an intuitive and data-driven approach for spatial analysis of lipids. High-level class composite images should allow investigators to make broad inferences about their data that inform subsequent interrogations with increasing levels of specificity.

4.5 Conclusions

This work has demonstrated a method for RKMD-based lipid annotation and IMS image filtering. The enhanced specificity and precision of the annotation method was shown through calculation of radyl carbon chain length and dataset curation by exclusion of features with distances from radyl carbon integer values, ε, larger than a window defined in this work as 0.001. When applied to peptide, MALDI matrix cluster, and lipid MS features, the specificity and precision were broadly enhanced by radyl carbon ε exclusion when compared to conventional exclusion only by RKMD δ, or distance from RKMD integer values. A proof-of-concept application to a computationally generated IMS dataset showed the outputs of the method which were filtered and reconstructed images that use RKMD calculated molecular class, degree of unsaturation, and radyl carbon chain length as criteria.

Finally, we applied the method to MALDI-IMS lipidomic data from human kidney tissue section that spanned the cortex and medulla regions. The filtering method was used to visualize spatial distribution of subgroups of PC, PE, and SM lipids. Colocalization of saturated PC and PE and monounsaturated SM components was observed throughout the tissue, namely in glomeruli, medullary rays, and vasculature. However, the addition of one unsaturation to each molecular class reduced the previously observed correlations between PC/PE and SM. Of particular note was the colocalization of SM to cortical glomeruli. To evaluate the extent of SM localization to glomeruli, we visualized

distributions of SM components with varying chain lengths noting unique colocalization of SM with 34 and 42 radyl carbons with glomerular structures. Finally, we unpacked the 42 radyl carbon SM composite image to visualize each sum composition component. Building on this work, future studies may utilize this workflow to intuitively analyze spatial distributions of lipid classes within and between samples to enhance analysis of lipidomics IMS datasets.

4.6 Acknowledgements

This work was supported by the National Science Foundation [grant number: 1709526]. The authors would like to thank Jamie L. Allen, Maya Brewer, and Prof. Mark de Caestecker for assistance with tissue processing. Support was provided by the NIH Common Fund and National Institute of Diabetes and Digestive and Kidney Diseases (U54DK120058 awarded to J.M.S. and R.M.C.), NIH National Institute of Allergy and Infectious Disease (R01 AI138581 awarded to J.M.S.), and the National Science Foundation Major Research Instrument Program (CBET–1828299 awarded to J.M.S. and R.M.C.). E.K.N. is supported by a National Institute of Environmental Health Sciences training grant (T32ES007028). The Cooperative Human Tissue Network is supported by the NIH National Cancer Institute (5 UM1 CA183727-08).

CHAPTER FIVE

Conclusion

5.1 Dissertation Overview

The work presented in this dissertation focused on providing chemical classification and annotation tools that could enhance metabolomics and lipidomics analysis workflows for characterization of biological samples. After presenting mass spectrometry and existing data analysis tools in Chapter one, subsequent chapters introduced novel chemical class annotation tools for interpretation of mass spectrometry data. More specifically, Chapter one provided a brief history of the mass spectrometry and scope of the research presented in this dissertation followed by a general overview of the methods and concepts that might be necessary for better understanding of the presented research. The reviewed methods and concepts pertained to components of instrument configurations used to gather and analyze data that served to validate the chemical classification tools presented.

Chapter two introduced a supervised feedforward neural network (FFNN) tool for chemical classification or *In Silico* Fractionation (iSF) of biomolecules. In this approach, iSF FFNNs utilized direct MS measurement inputs including *m/z* values, ion isotope relative abundances, isotope ratios, and KMD values to classify features. Classifiers were arranged in a neural decision tree structure

that sequentially classified inputs with increasing levels of specificity. In this demonstration, inputs that were positively predicted by the lipid classifier, LIPNET, were submitted to the lipid subclassifier, LIPSUBNET, to predict lipid subclasses. In addition, a performance enhancing measure, termed "confidence thresholding", was implemented to improve the accuracy of iSF classifiers by limiting classification to those with an output exceeding the threshold value. The accuracy and robustness of iSF classifiers were demonstrated using two experimental datasets in which peptides, lipids, and metabolites were successfully classified.

Chapter three demonstrated an application of iSF classifiers to a large lipidomics LC-MS/MS dataset. An array of iSF binary classifiers (i.e., one for each lipid subclass) were used to subclassify lipid features from lipid extracts from biological samples that were previously characterized and identified by LC-MS/MS and a fragmentation spectral matching workflow, LipiDex. The orthogonality of iSF to the LipiDex workflow was demonstrated by comparing the agreement of the iSF classifiers predicted classes with LipiDex-determined classes as a function of LipiDex match score. When the LC-MS quality control parameters were controlled to maintain high quality iSF inputs, it was shown that iSF agreement increased as a function of Lipidex MS/MS fragment spectral match score (i.e., the score that describes the similarity between experimental and reference library MS/MS spectra), up to a certain match score that the LipiDex authors describe as "confident". These results demonstrate that iSF

classification could provide useful information in cases of low scoring, and therefore ambiguous, identification results produced by, for instance, MS/MS fragment spectral matching.

Chapter four introduced an enhanced referenced Kendrick mass defect (RKMD) method for sum composition annotation of lipids in MS imaging datasets and class-based image filtering. Previously, RKMD was used as a chemical classification tool in high resolving power MS data that could also yield information on lipid unsaturation. However, in the presented research in Chapter four, this information was used to calculate radyl carbon chain length as well. Calculation of radyl carbon chain length using the information derived from RKMD analysis enabled lipid sum composition determination, significant improvements in automated data curation, and image filtering based on chain length. The RKMD-based annotation and image filtering workflow was applied to a MALDI-MS imaging analysis of human kidney tissue lipids. Spatial distributions of sphingomyelins, phosphatidylcholines, and phosphatidylethanolamines were visualized based on the degrees of unsaturation to show differential localization in kidney substructures. Kidney structures colocalized to sphingomyelin distributions led to visualization of sphingomyelin distributions based on radyl carbon chain length.

5.2 Future Directions

5.2.1 iSF Application to Whole X-ome LC-MS Separations

As demonstrated in Chapters two and three, iSF FFNNs can be trained to accommodate a wide diversity of biological chemical classes. Although lipidomics and metabolomics are important fields for application of orthogonal chemical classifiers, the field of whole x-omics requires flexible and comprehensive annotation strategies and hence, could benefit greatly from methods such as iSF. In whole x-omics research, investigators seek to unify analyses of various molecular classes in biological samples to interrogate inter-class relationships and increase throughput. However, annotation workflows increase in their complexity with greater number of classes accommodated in an analysis as investigators lose the ability to use *a priori* assumptions to guide analysis. Regarding MS data analysis, *a priori* assumptions about analyte class inform: 1) allowable elements and numbers of atoms in EC determination, 2) MS database usage for mass accuracy and MS/MS fragmentation spectral matching, and 3) peptide and nucleotide sequencing workflows. In such cases where identification workflows are severely impaired, orthogonal chemical classification methods such as iSF could yield information on general sample composition and suggest strategies for identification on a feature-by-feature basis.

Recently, a proprietary sample preparation and LC-MS method, Omni-MS, was presented at ASMS 2019 and 2020 that provides “one-shot” analysis of peptides, lipids, metabolites, metals, and nucleotides. Omni-MS is being used at Dalton Bioanalytics to perform quantitative analyses of various molecules in plasma. The presenting authors were kind to supply a dataset used for Omni-MS method validation. In general, the method involves sequential digestion of peptides and nucleotides and followed by extraction of the solution containing all soluble molecular components. Samples were then separated by UPLC with a triphasic solvent gradient using water, ACN, and a buffer constituted of volatile salts. The resulting LC-MS acquisitions produced molecular signatures from peptides, nucleotides, metabolites, lipids, and metals. Preliminary iSF FFNNs were trained to classify each molecular class except for metals. Application to iSF to these datasets showed promising results with strong classifications from each chemical class that correlated well with class predictions based on LC retention and/or manual inspection of LC-MS features.

Although a trained user can often discriminate between different classes of molecules in LC-MS data (i.e., by charge state, LC retention, isotope ratios, etc.), data volumes produced in such acquisitions make such expectations untenable. Additionally, an automated implementation of chemical class prediction informs downstream processing steps and provides a means for sample description that is mostly independent of user bias. Of the other available orthogonal chemical classification methods[13, 14], iSF is the only viable

available option for pure LC-MS data (i.e., iSF is not reliant on IM or MS/MS fragmentation data).

5.2.2 Comparison of Internal Spatial Class Correlations between Lipidomics MS Images to Differentiate Sample Conditions

As demonstrated in Chapter four, the RKMD-based lipid annotation and image filtering workflow can effectively describe lipids in terms of their headgroup, degrees of unsaturation, and radyl carbon chain length (i.e., their sum composition) and visualize lipid distributions based on each selected feature. The presented RKMD-based workflow provided an intuitive means for comparing the colocalization of lipid populations within samples; however, it can also support means for automated comparison of lipid spatial distributions between samples and conditions.

In conventional MS imaging analyses, between condition/sample comparisons are difficult because tissue sections for different samples/conditions are unique in their size, shape, and distribution of internal structures due to differences in tissue morphology between equivalent sections from different animals. Therefore, comparing similarities or differences in localization of different molecules is not trivial; investigators must often evaluate each *m/z* specific image and infer whether molecules are likely to be differentially localized and then quantitatively evaluate select regions-of-interest that contain the localized features. Given the difficulty of direct comparisons, we suggest that the RKMD-based annotation and filtering workflow can facilitate

and automate indirect comparisons between samples by comparing internal spatial correlations ($class_A$ vs. $class_B$ within a sample where A and B are different lipid classes) between different samples with different conditions ($correlation_{AB1}$ vs. $correlation_{AB2}$ for hypothetical conditions 1 and 2). Given equivalent sample preparation (e.g., uniform matrix deposition in MALDI, same MALDI matrix, etc.), and sampling conditions (e.g., laser fluence, wavelength, spot size, etc.), internal correlations between molecules in two equivalent tissue sections (e.g., two control samples) should be similar. Our hypothesis is that changes in localization of different molecules due to changing experimental conditions will result in significant changes in internal spatial correlations. The degree of similarity between samples of the same condition will need to be assessed to provide a baseline for comparisons between conditions.

Preliminary data analysis has shown that the similarity of spatial distributions of lipids can be assessed by determining their structural similarity index[286]. Using Python scripting, the intensity values of select class and *m/z* images from the human kidney lipidomics MALDI-IMS dataset were scaled between 0 and 1 and then compared by calculating the structural similarity index (from the scikit-image Python package) for each pair of images. Visual inspection of pairs of images of high structural similarity index scores confirmed that the visualized lipids were highly colocalized. We suggest application of the RKMD-based method to multiple control and disease condition samples to validate the approach and assess changes in correlation

between equivalent samples and between conditions. The RKMD-based method is particularly suited for this type of analysis as it facilitates visualization and analysis of lipid classes with increasing specificity. We suggest that observed changes in internal spatial correlations, resulting from comparisons between high-level classes, will indicate general differences between conditions and direct analysis towards comparisons of more specific groups (and/or individual lipid sum compositions). Ultimately, the goal is to determine which classes and individual species are most affected by the condition(s).

APPENDICES

APPENDIX A

Using Isotopic Envelopes and Neural Decision Tree-based *In Silico Fractionation* for Biomolecule Classification

A.1 Python Script for Generation of Pseudo-random Polypeptide and Nucleic Acid Sequences and Elemental Compositions

```
import os, xlwt
import pandas as pd
import numpy as np
import random as rn
import datetime as dt

def parse_csv(fileName):
    with open(fileName) as f:
        parsed = pd.read_csv(f)
    parsed.head()
    return parsed

def build_dict(sequence, symbolType, atoms):
    dictionary = {}
    for i, symbol in enumerate(sequence[symbolType]):
        dictionary[symbol] = {atom: sequence[atom][i] for atom in atoms}
    return dictionary

def build_filename(baseName, length, average, std):
    time = dt.datetime.now().strftime("%Y-%m-%d_%H-%M-%S")
    return f'{time}_{length}__AVG{average}_STDEV{std}_{baseName}.xlsx'

def generate_chemform(sequence, dictionary, atoms):
    atomNumbers = {atom: 0 for atom in atoms}
    for monomer in sequence:
        molecules = dictionary.get(str(monomer))
        for atom in atoms:
            atomNumbers[atom] = int(atomNumbers[atom] + molecules.get(atom))
    chemForm = ""
    for atom in atoms:
        if atomNumbers[atom] != 0:
            chemForm = chemForm + atom + str(atomNumbers[atom])
    return chemForm

def generate_sequence(length, df, symbolType, monoRange, ptm=False):
    sequence = [df[symbolType][rn.randint(*monoRange)] for _ in range(length)]
    if ptm:
        sequence[-1] = ptm
    return sequence
```

```
def generate_form(length, avgLen, atoms, df, d, symbolType, monoRange, std,
ptm=False):
    formList = []
    for _ in range(length):
        lng = int(np.round(np.random.normal(loc=avgLen, scale=std, size=1),
0))
        sequence = generate_sequence(lng, df, symbolType, monoRange)
        formList.append(generate_chemform(sequence, d, atoms))
    return formList

def save_form(formList, header, fileName, avg, std, index=False):
    dataframe = pd.DataFrame(formList)
    dataframe.columns = header
    fileName = build_filename(fileName, len(formList), avg, std)
    dataframe.to_excel(fileName, index=index)

directory = os.path.dirname(os.path.abspath(__file__))
atName = 'AminoTable.csv'
ntName = 'NucleicAcidTable.csv'
dfAmino = parse_csv(atName).dropna(axis=0,how='all').reset_index(drop=True)
dfNucleic = parse_csv(ntName)
peptideDict = build_dict(dfAmino, 'Long Symbol', ['C', 'H', 'N', 'O', 'S'])
nucleicDict = build_dict(dfNucleic, 'Short Symbol', ['C', 'H', 'N', 'O', 'P'])
```

Listing A.1. Code for generating random peptide and nucleic sequences and elemental compositions.

The above script builds a dictionary containing the elemental compositions of each amino acid and nucleotide monomer. Polymer sequences are then generated pseudo-randomly. The mean length of the sequences to be generated is provided, and the length of the output sequences follow a normal distribution around the provided mean and with a provided standard deviation. Using the nucleic acid or amino acid dictionary, the sequence is used to generate an elemental composition. The list of elemental compositions is then saved to an excel spreadsheet.

```
function [performance, time, percentErrors] = NN_Train(Hidden_Nodes,
Train_Set, Targ_Set)

    x = Train_Set';
    t = Targ_Set';

    trainFcn = 'trainscg';

    hiddenLayerSize = Hidden_Nodes;
    net = patternnet(hiddenLayerSize, trainFcn);

    net.input.processFcns = {'removeconstantrows','mapminmax'};

    net.divideFcn = 'dividerand';
    net.divideMode = 'sample';
    net.divideParam.trainRatio = 70/100;
    net.divideParam.valRatio = 15/100;
    net.divideParam.testRatio = 15/100;

    net.performFcn = 'crossentropy';
    net.trainParam.max_fail = 6;
    net.trainParam.epochs = 1000;

    net.plotFcns = {'plotperform','plottrainstate','ploterrhist', ...
        'plotconfusion', 'plotroc'};

    [net,tr] = train(net,x,t);

    y = net(x);
  e = gsubtract(t,y);
    performance = perform(net,t,y)
    tind = vec2ind(t);
    yind = vec2ind(y);
    percentErrors = sum(tind ~= yind)/numel(tind);

    trainTargets = t .* tr.trainMask{1};
    valTargets = t .* tr.valMask{1};
    testTargets = t .* tr.testMask{1};
    trainPerformance = perform(net,trainTargets,y)
    valPerformance = perform(net,valTargets,y)
    testPerformance = perform(net,testTargets,y)

    if (true)
        genFunction(net,'NNetFcn','MatrixOnly','yes');
        y = myNeuralNetworkFunction(x);
    end
end
```

Listing A.2. MATLAB code for training feedforward neural networks.

The above script takes in the number of hidden neurons, a matrix for the training set of values, and a matrix for the training targets. We utilized the "trainscg" function (i.e., scaled conjugate gradient backpropogation) for all trained networks. Matrices were processed by removing rows with constant values and mapping minimum and maximum row values to [-1,1]. Training, validation, and test sets were selected from the provided training data at random. Divisions were as follows: training set with 70%, validation set with 15%, and test set with 15%. The "crossentropy" function was used to monitor network performance. The early stopping method was employed when a max of 6 validation checks were failed. The MATLAB functions for the trained networks, PEPNET, LIPNET, and LIPSUBNET, are available at: www.github.com/luketrichardson/in-silico-Fractionation.

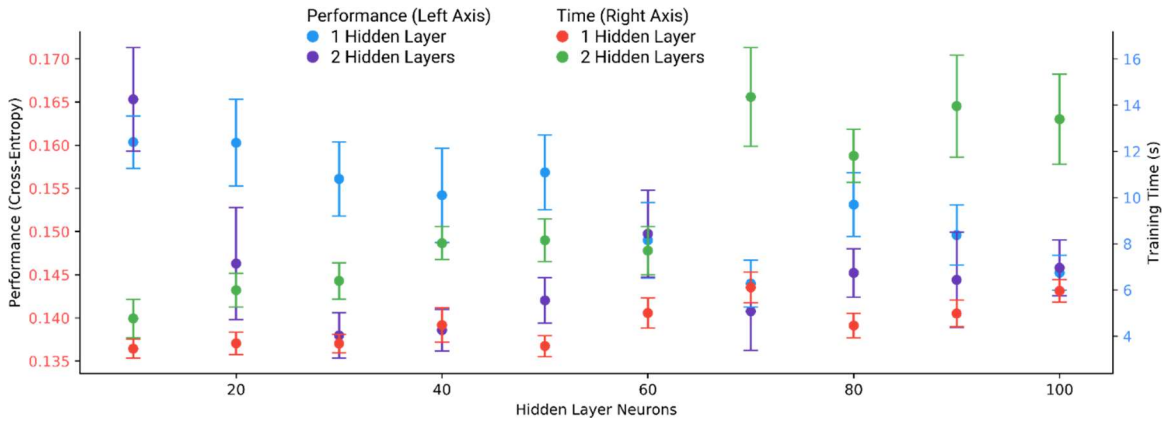


Figure A.1. Performance (left axis) and training time (right axis) for one and two hidden layers as a function of hidden layer neurons for the binary classifier architecture.

The plot in Figure A.1 displays the mean performance (left axis) and training time (right axis) for neural network training. A two hidden layer network with 30 neurons each was selected for the binary classifier (PEPNET and LIPNET) architecture for its relatively consistent and high performance (lowest mean cross-entropy) within an acceptable training timeframe.

A.4 Multi-target Classifier Neural Network Training Characteristics

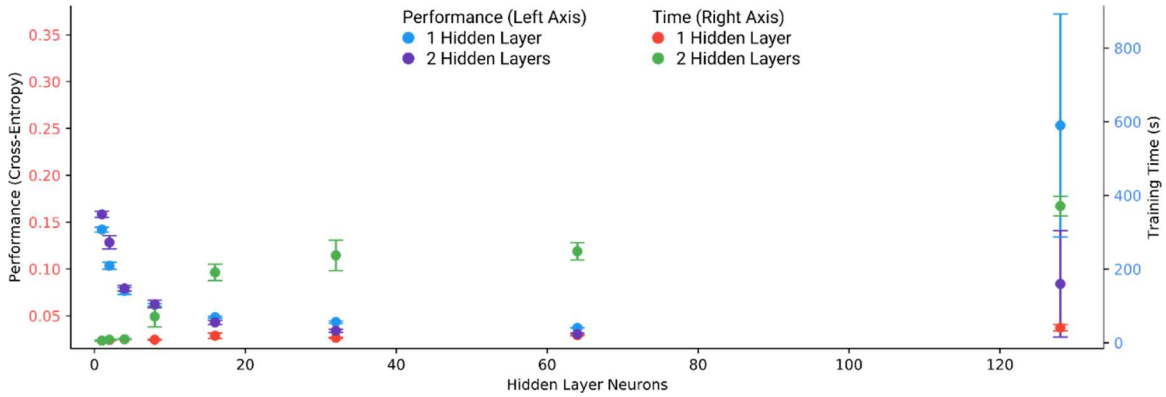


Figure A.2. Performance (left axis) and training time (right axis) for one and two hidden layers as a function of hidden layer neurons for the multi-target classifier architecture.

The plot in Figure A.2 displays the mean performance (left axis) and training time (right axis) for neural network training. A two hidden layer network with 64 neurons each was selected for the poly-target classifier (LIPSUBNET) architecture for its relatively consistent and high performance (lowest mean cross-entropy) within an acceptable training timeframe.

A.5 Mass Resolution Evaluation for Peptide and Lipid MS Peaks

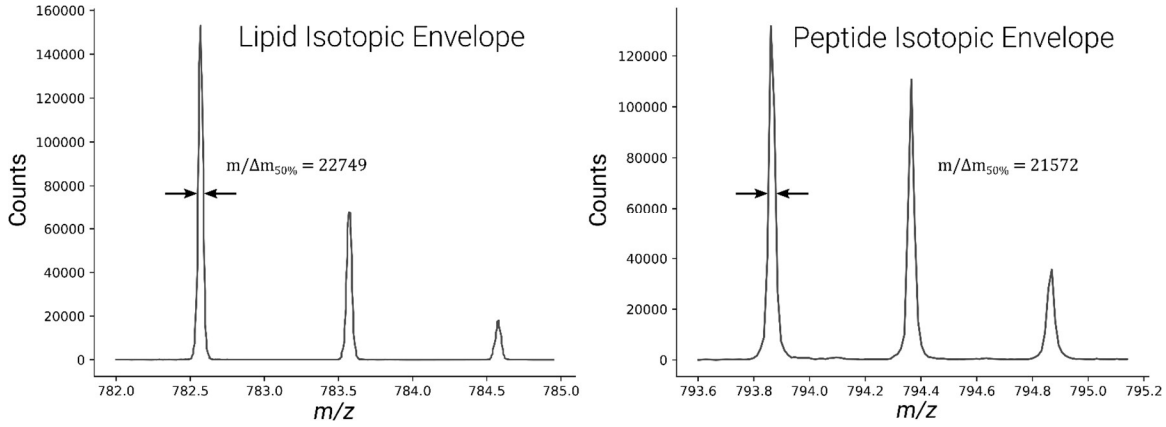


Figure A.3. Evaluation of mass resolution of peptide and lipid MS peaks.

The mass spectra in Figure A.3 display exemplary isotopic envelopes from the lipid (left) and polypeptide (right) classes. Components with similar *m/z* values were chosen to show the mass resolution of the experimentally acquired spectra utilized in this demonstration. Mass resolution was calculated as the ratio of the monoisotopic *m/z* to the monoisotopic peak width at half of maximum intensity ($m/\Delta m_{50\%}$). The mass resolution of the measured lipid and polypeptide analytes were 22749 and 21572, respectively.

A.6 PEPNET Binary Classifier Confusion Matrix

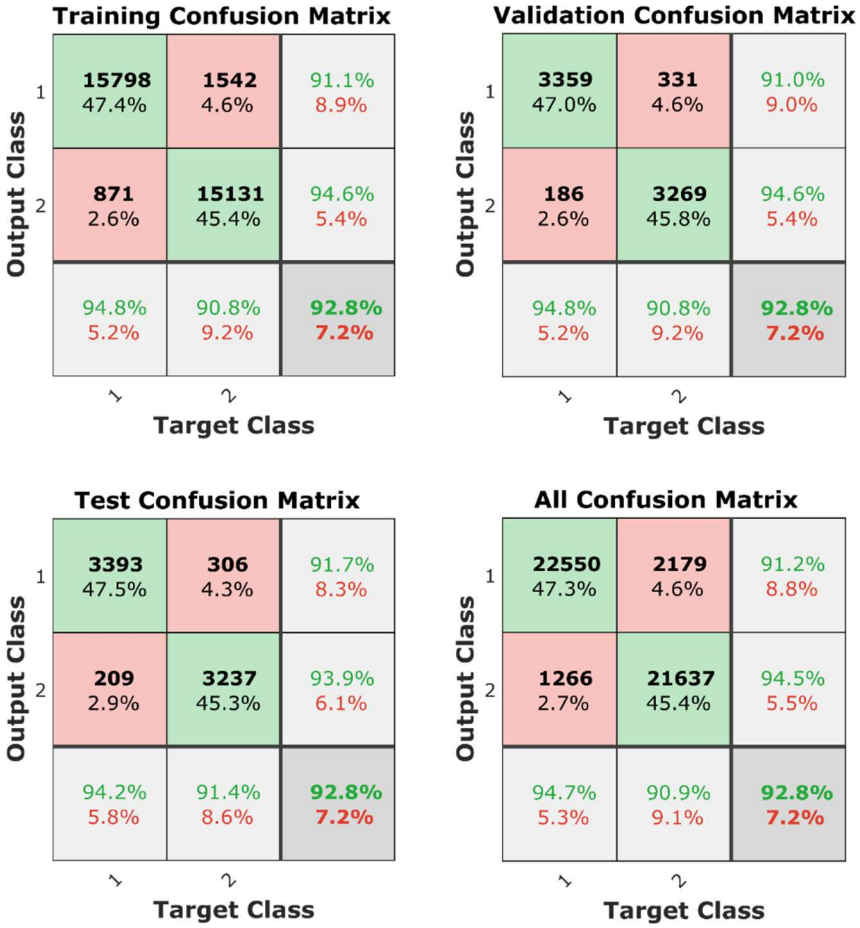


Figure A.4. Confusion matrix of PEPNET binary classifier.

The diagram in Figure A.4 displays the neural network confusion matrix for PEPNET. The “Target Class” columns correspond to the true class and “Output Class” rows correspond to the predicted class. The green-shaded cells cor-respond to correct classifications, and the red-shaded cells correspond to incorrect classifications. The number of classifications (top) and percentage of the total number of classification (bottom) are shown in each cell. The far-right column (excluding the cell in the bottom-right corner) displays the positive

predictive values (top, green) and false discovery rates (red, bottom) for each class. The bottom-most row (excluding the cell in the bottom-right corner) displays the true positive rates (top, green) and false negative rates (bottom, red) for each class. The cell in the bottom-right corner displays the overall accuracy (top, green) and error (bottom, red). Classes 1 and 2 are polypeptides and non-polypeptides, respectively.

A.7 LIPNET Binary Classifier Confusion Matrix

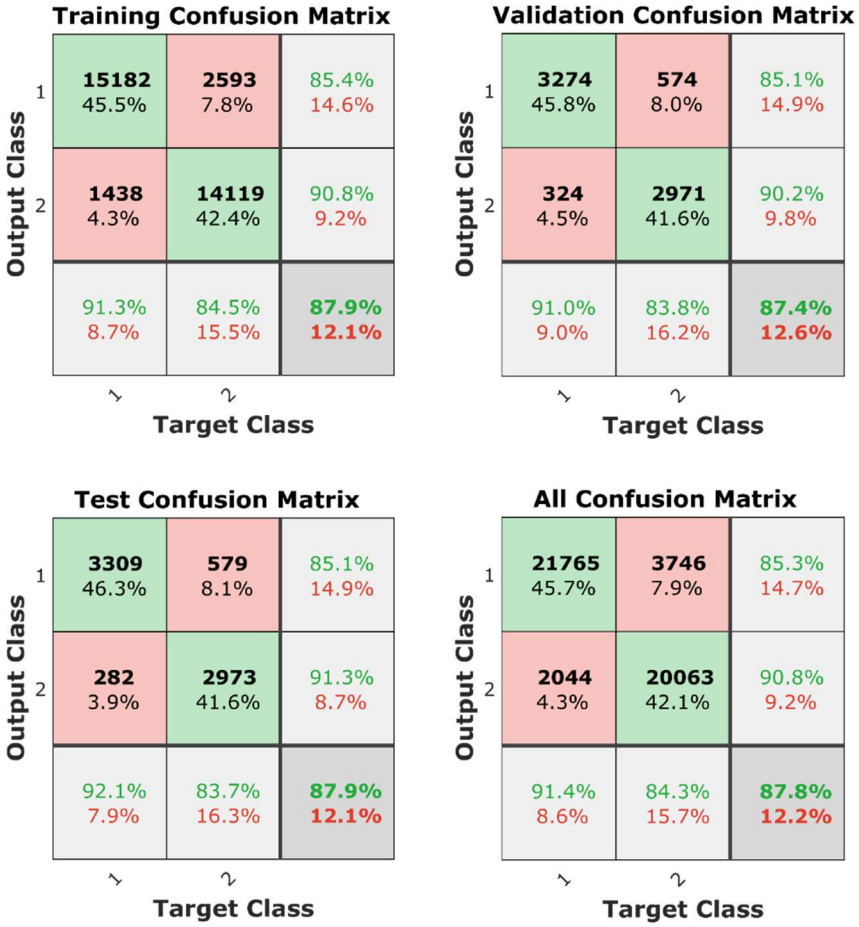


Figure A.5. Confusion matrix of LIPNET binary classifier.

The diagram in Figure A.5 displays the neural network confusion matrix for LIPNET. The “Target Class” columns correspond to the true class and “Output Class” rows correspond to the predicted class. The green-shaded cells correspond to correct classifications, and the red-shaded cells correspond to incorrect classifications. The number of classifications (top) and percentage of the total number of classification (bottom) are shown in each cell. The far-right column (excluding the cell in the bottom-right corner) displays the positive

predictive values (top, green) and false discovery rates (red, bottom) for each class. The bottom-most row (excluding the cell in the bottom-right corner) displays the true positive rates (top, green) and false negative rates (bottom, red) for each class. Classes 1 and 2 are lipids and non-lipids, respectively.

A.8 LIPSUBNET Multitarget Classifier Confusion Matrix

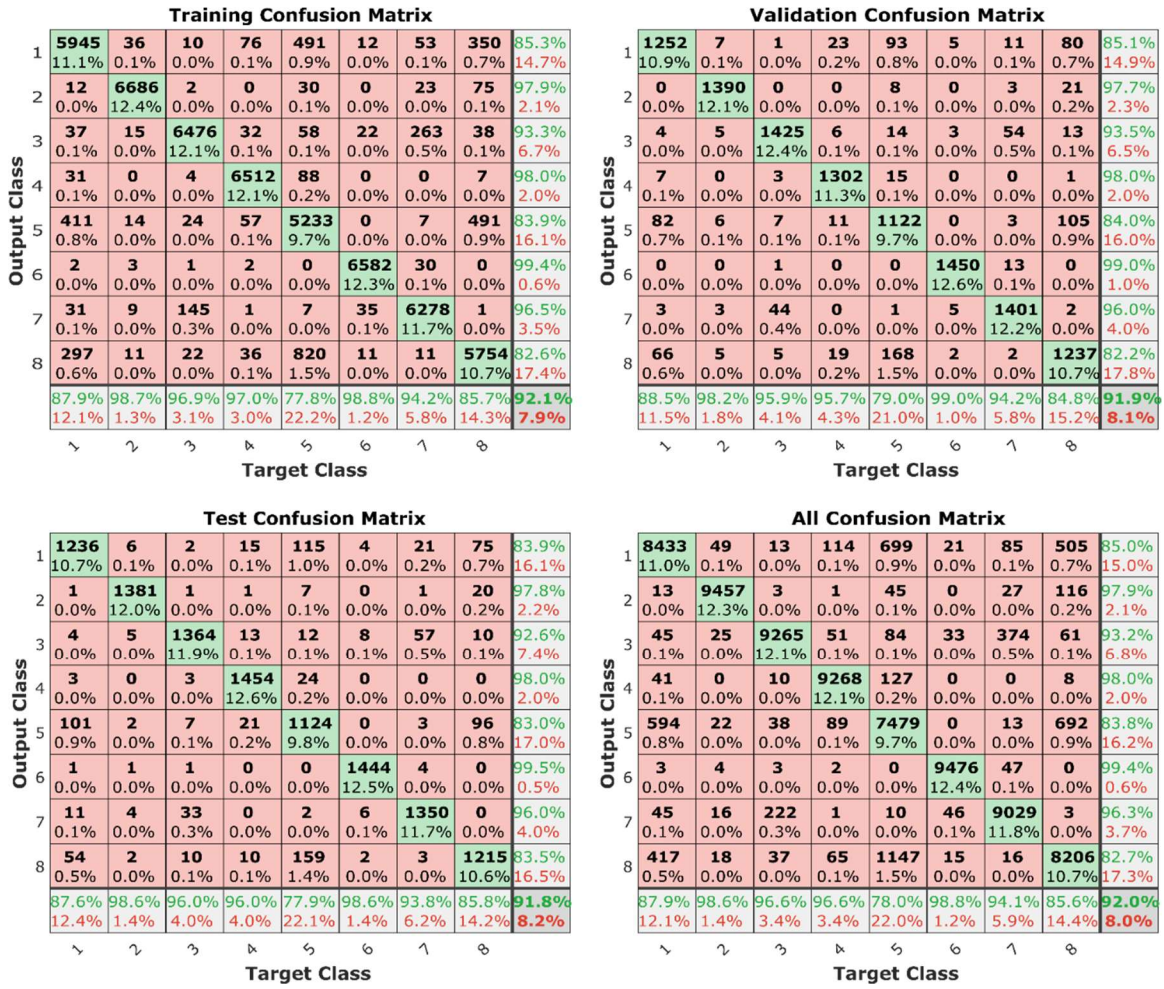


Figure A.6. Confusion matrix of LIPSUBNET multi-target classifier.

The diagram in Figure A.6 displays the neural network confusion matrix for LIPSUBNET. The “Target Class” columns correspond to the true class and “Output Class” rows correspond to the predicted class. The green-shaded cells correspond to correct classifications, and the red-shaded cells correspond to incorrect classifications. The number of classifications (top) and percentage of the total number of classification (bottom) are shown in each cell. The far-right

column (excluding the cell in the bottom-right corner) displays the positive predictive values (top, green) and false discovery rates (red, bottom) for each class. The bottom-most row (excluding the cell in the bottom-right corner) displays the true positive rates (top, green) and false negative rates (bottom, red) for each class. Classes 1-7 are as follows in order: (1) fatty acyls, (2) glycerolipids, (3) glycerophospho-lipids, (4) polyketides, (5) prenol lipids, (6) saccharolipids, (7) sphingolipids, and (8) sterol lipids.

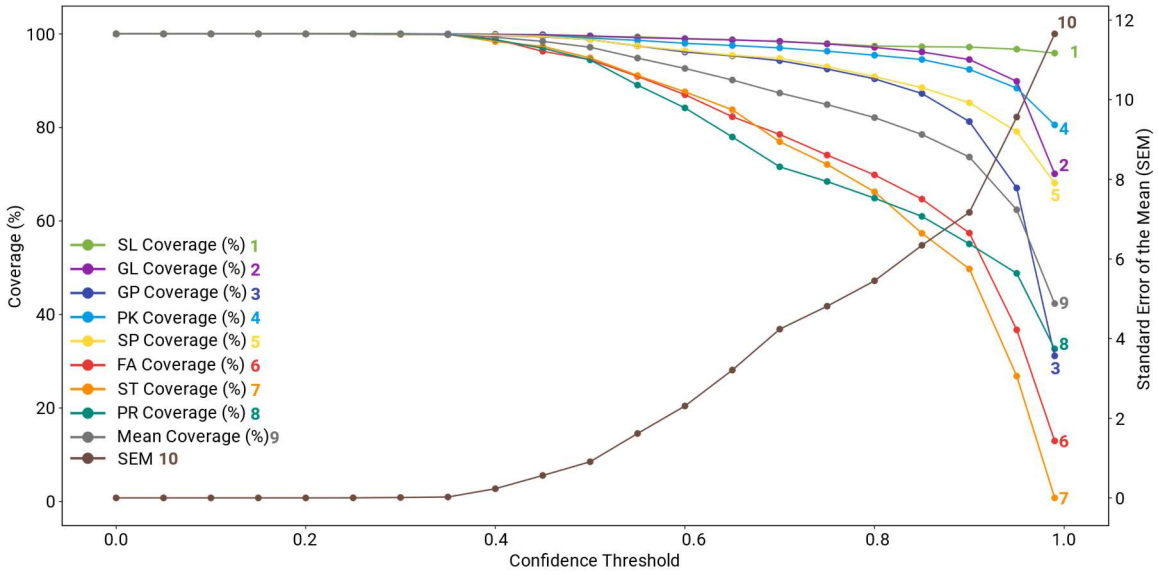


Figure A.7. Line plot show lipid subclass coverage (left axis) and the coverage standard error of the mean (SEM, right axis) between subclasses as a function of Confidence Threshold.

The plot in Figure A.7 displays network coverage percent as a function of coverage threshold for the theoretical training set of LIBSUBNET. The coverage of each class (as well as the mean coverage) is plotted on the left axis, and standard error of the mean coverage (gold) is plotted on the right axis to represent the degree of variance in class representation. Coverage does not decrease until a confidence threshold of 0.25 after which each class begins to lose coverage at different rates. Typically, classes that lose the most coverage gain the most with regard to TPR, which results in higher total accuracy for all classes (~99%) at the cost of equal class representation in a sample. Subclasses are represented in the legend as follows: saccharolipids (SL, 1), glycerolipids

(GL, 2), glycerophospholipids (GP, 3), polyketides (PK, 4), sphingolipids (SP, 5), fatty acyls (FA, 6), sterol lipids (ST, 7), and prenol lipids (PR, 8). The percent coverage for each subclass and mean percent coverage (Mean, 9) is plotted against the left axis, and the standard error of the mean (SEM, 10) is plotted against the right axis.

A.10 PEPNET, LIPNET, and LIPSUBNET Score Outputs for the Multi-Class Sample

Table A.1. FFNN output scores for PEPNET, LIPNET, and LIPSUBNET.

PEPNET		LIPNET		LIPSUBNET							
P	NP	L	NL	FA	GL	GP	PK	PR	SL	SP	ST
.008	.992	.992	.008	.000	.000	.988	.001	.000	.000	.010	.000
.006	.994	.993	.007	.000	.000	.991	.001	.000	.000	.008	.000
.008	.992	.991	.009	.002	.001	.974	.001	.000	.001	.020	.000
.006	.994	.991	.009	.000	.000	.952	.001	.000	.000	.047	.000
.011	.989	.985	.015	.001	.999	.000	.000	.000	.000	.000	.000
.007	.993	.993	.007	.000	.001	.986	.001	.000	.001	.011	.000
.009	.991	.992	.008	.000	.006	.985	.001	.000	.001	.005	.002
.005	.995	.992	.008	.000	.000	.947	.001	.000	.000	.052	.000
.008	.992	.992	.008	.002	.009	.958	.001	.000	.006	.023	.001
.006	.994	.993	.007	.000	.000	.970	.001	.000	.001	.028	.000
.009	.991	.989	.011	.019	.036	.799	.002	.000	.064	.078	.001
.006	.994	.994	.006	.000	.001	.975	.001	.000	.003	.019	.000
.004	.996	.993	.007	.000	.000	.848	.001	.000	.000	.151	.000
.004	.996	.993	.007	.000	.000	.941	.001	.000	.000	.058	.000
.003	.997	.996	.004	.000	.000	.158	.000	.000	.000	.842	.000
.004	.996	.994	.006	.001	.000	.407	.002	.000	.001	.590	.000
.003	.997	.994	.006	.000	.000	.367	.002	.000	.000	.631	.000
.041	.959	.994	.006	.000	.988	.000	.000	.012	.000	.000	.000
.036	.964	.995	.005	.000	.996	.000	.000	.001	.000	.000	.003
.008	.992	.997	.003	.001	.078	.000	.000	.036	.000	.000	.886
.027	.973	.991	.009	.001	.762	.000	.000	.000	.012	.000	.225

Table A.1. FFNN output scores for PEPNET, LIPNET, and LIPSUBNET.

P	NP	L	NL	FA	GL	GP	PK	PR	SL	SP	ST
.009	.991	.997	.003	.000	.459	.000	.000	.006	.000	.000	.534
.011	.989	.997	.003	.001	.855	.012	.000	.000	.000	.042	.090
.026	.974	.995	.005	.000	.981	.000	.000	.000	.000	.000	.019
.011	.989	.996	.004	.001	.145	.000	.000	.003	.000	.000	.852
.010	.990	.997	.003	.001	.702	.038	.000	.001	.000	.154	.104
.013	.987	.996	.004	.000	.994	.000	.000	.000	.000	.001	.005
.015	.985	.996	.004	.000	.983	.001	.000	.000	.000	.004	.012
.103	.897	.960	.040	.000	.000	.000	.000	.000	.000	.000	.000
.010	.990	.994	.006	.011	.377	.000	.000	.000	.000	.001	.610
.018	.982	.996	.004	.000	.995	.000	.000	.000	.000	.000	.005
.021	.979	.996	.004	.000	.992	.000	.000	.000	.000	.001	.007
.011	.989	.996	.004	.000	.996	.000	.000	.000	.000	.003	.001
.013	.987	.996	.004	.000	.997	.000	.000	.000	.000	.002	.001
.017	.983	.996	.004	.000	.997	.000	.000	.000	.000	.002	.002
.009	.991	.994	.006	.050	.597	.000	.000	.000	.000	.030	.323
.015	.985	.996	.004	.000	.997	.000	.000	.000	.000	.002	.000
.017	.983	.996	.004	.000	.992	.001	.000	.000	.000	.006	.001
.915	.085	.005	.995								
.994	.006	.001	.999								
.876	.124	.015	.985								
.866	.134	.017	.983								
.735	.265	.121	.879								
.881	.119	.031	.969								

Table A.1. FFNN output scores for PEPNET, LIPNET, and LIPSUBNET.

P	NP	L	NL	FA	GL	GP	PK	PR	SL	SP	ST
.993	.007	.000	.000								
.953	.047	.027	.973								
.988	.012	.001	.999								
.987	.013	.002	.998								
.855	.145	.138	.862								
.851	.149	.132	.868								
.669	.331	.253	.747								
.939	.061	.019	.981								
.989	.011	.000	.000								
.526	.474	.045	.955								
.736	.264	.201	.799								
.988	.012	.001	.999								
.980	.020	.003	.997								
.895	.105	.034	.966								
.959	.041	.000	.000								
.989	.011	.001	.999								
.955	.045	.018	.982								
.637	.363	.032	.968								
.872	.128	.032	.968								
.991	.009	.001	.999								
.711	.289	.095	.905								
.898	.102	.053	.947								
.997	.003	.002	.998								

Table A.1. FFNN output scores for PEPNET, LIPNET, and LIPSUBNET.

P	NP	L	NL	FA	GL	GP	PK	PR	SL	SP	ST
.966	.034	.008	.992								
.931	.069	.006	.994								
.832	.168	.055	.945								
.903	.097	.034	.966								
.932	.068	.016	.984								
.987	.013	.000	.000								
.989	.011	.000	.000								
.843	.157	.012	.988								
.883	.117	.015	.985								
.989	.011	.001	.999								
.988	.012	.007	.993								
.987	.013	.000	.000								
.988	.012	.000	.000								
.640	.360	.285	.715								
.896	.104	.016	.984								
.968	.032	.016	.984								
.978	.022	.001	.999								
.989	.011	.000	.000								
.930	.070	.016	.984								
.793	.207	.020	.980								
.955	.045	.008	.992								
.944	.056	.018	.982								
.987	.013	.011	.989								

Table A.1. FFNN output scores for PEPNET, LIPNET, and LIPSUBNET.

P	NP	L	NL	FA	GL	GP	PK	PR	SL	SP	ST
.936	.064	.018	.982								
.825	.175	.023	.977								
.980	.020	.000	.000								
.942	.058	.014	.986								
.644	.356	.019	.981								
.953	.047	.010	.990								
.001	.999	.251	.749								
.061	.939	.161	.839								
.009	.991	.702	.298	.622	.000	.000	.001	.217	.000	.000	.160
.050	.950	.164	.836								
.011	.989	.184	.816								
.064	.936	.166	.834								
.011	.989	.184	.816								
.009	.991	.184	.816								
.015	.985	.188	.812								
.007	.993	.185	.815								

Neural network output scores for PEPNET, LIPNET, and LIPSUBNET corresponding to Figure 2.5 are displayed in Table A.1. Green cell shading indicates a correct component class prediction. Red cell shading indicates an incorrect component class prediction. For incorrectly predicted components, yellow cell shading indicates the true class. T_P and T_{NP} denote the peptide and non-peptide class targets for PEPNET, respectively. T_L and T_{NL} denote the lipid

and non-lipid class targets for LIPNET, respectively. Target classes in LIPSUBNET (*i.e.*, lipid subclasses) are represented in the legend as follows: fatty acyls (T_{FA}), glycerolipids (T_{GL}), glycerophospholipids (T_{GP}), polyketides (T_{PK}), prenol lipids (T_{PR}), saccharolipids (T_{SL}), sphingolipids (T_{SP}), and sterol lipids (T_{ST}).

A.11 PEPNET and LIPNET Score Outputs for the Lipid and Peptide Sample

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

PEPNET		LIPNET		
P	NP	L	NL	LC Region
0.967	0.033	0.008	0.992	10-30 minutes
0.949	0.051	0.044	0.956	Peptides
0.977	0.023	0.035	0.965	
0.988	0.012	0.011	0.989	
0.990	0.010	0.010	0.990	
0.973	0.027	0.000	1.000	
0.905	0.095	0.000	1.000	
0.908	0.092	0.044	0.956	
0.939	0.061	0.007	0.993	
0.736	0.264	0.000	1.000	
0.923	0.077	0.000	1.000	
0.946	0.054	0.004	0.996	
0.938	0.062	0.333	0.667	
0.916	0.084	0.006	0.994	
0.985	0.015	0.001	0.999	
0.990	0.010	0.004	0.996	
0.919	0.081	0.016	0.984	
0.959	0.041	0.000	1.000	
0.989	0.011	0.001	0.999	
0.988	0.012	0.015	0.985	
0.787	0.213	0.003	0.997	
0.993	0.007	0.001	0.999	
0.993	0.007	0.001	0.999	
0.973	0.027	0.018	0.982	
0.957	0.043	0.000	1.000	
0.948	0.052	0.005	0.995	
0.974	0.026	0.004	0.996	
0.958	0.042	0.009	0.991	
0.874	0.126	0.006	0.994	
0.798	0.202	0.000	1.000	
0.849	0.151	0.006	0.994	
0.966	0.034	0.000	1.000	
0.969	0.031	0.027	0.973	
0.944	0.056	0.018	0.982	
0.934	0.066	0.009	0.991	

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

P	NP	L	NL	LC Region
0.963	0.037	0.010	0.990	10-30 minutes
0.957	0.043	0.009	0.991	Peptides
0.985	0.015	0.005	0.995	
0.936	0.064	0.001	0.999	
0.890	0.110	0.013	0.987	
0.970	0.030	0.000	1.000	
0.842	0.158	0.009	0.991	
0.938	0.062	0.008	0.992	
0.908	0.092	0.020	0.980	
0.989	0.011	0.000	1.000	
0.993	0.007	0.001	0.999	
0.932	0.068	0.000	1.000	
0.511	0.489	0.004	0.996	
0.939	0.061	0.001	0.999	
0.929	0.071	0.021	0.979	
0.991	0.009	0.012	0.988	
0.989	0.011	0.041	0.959	
0.972	0.028	0.000	1.000	
0.967	0.033	0.000	1.000	
0.983	0.017	0.012	0.988	
0.990	0.010	0.001	0.999	
0.973	0.027	0.000	1.000	
0.939	0.061	0.009	0.991	
0.958	0.042	0.029	0.971	
0.915	0.085	0.000	1.000	
0.951	0.049	0.004	0.996	
0.939	0.061	0.032	0.968	
0.982	0.018	0.003	0.997	
0.927	0.073	0.044	0.956	
0.933	0.067	0.002	0.998	
0.948	0.052	0.000	1.000	
0.990	0.010	0.008	0.992	
0.990	0.010	0.009	0.991	
0.920	0.080	0.016	0.984	
0.920	0.080	0.079	0.921	
0.987	0.013	0.005	0.995	
0.838	0.162	0.008	0.992	
0.964	0.036	0.006	0.994	
0.612	0.388	0.016	0.984	

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

P	NP	L	NL	LC Region
0.960	0.040	0.073	0.927	10-30 minutes
0.553	0.447	0.439	0.561	Peptides
0.950	0.050	0.000	1.000	
0.842	0.158	0.003	0.997	
0.876	0.124	0.047	0.953	
0.945	0.055	0.036	0.964	
0.963	0.037	0.026	0.974	
0.882	0.118	0.009	0.991	
0.886	0.114	0.006	0.994	
0.963	0.037	0.000	1.000	
0.955	0.045	0.004	0.996	
0.953	0.047	0.010	0.990	
0.951	0.049	0.000	1.000	
0.964	0.036	0.000	1.000	
0.946	0.054	0.009	0.991	
0.905	0.095	0.000	1.000	
0.815	0.185	0.003	0.997	
0.955	0.045	0.029	0.971	
0.954	0.046	0.021	0.979	
0.946	0.054	0.022	0.978	
0.901	0.099	0.000	1.000	
0.979	0.021	0.167	0.833	
0.937	0.063	0.434	0.566	
0.867	0.133	0.260	0.740	
0.944	0.056	0.009	0.991	
0.944	0.056	0.000	1.000	
0.882	0.118	0.002	0.998	
0.931	0.069	0.003	0.997	
0.955	0.045	0.002	0.998	
0.960	0.040	0.375	0.625	
0.890	0.110	0.016	0.984	
0.959	0.041	0.006	0.994	
0.931	0.069	0.031	0.969	
0.945	0.055	0.006	0.994	
0.947	0.053	0.007	0.993	
0.890	0.110	0.006	0.994	
0.704	0.296	0.000	1.000	
0.932	0.068	0.021	0.979	
0.941	0.059	0.020	0.980	

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

P	NP	L	NL	LC Region
0.940	0.060	0.013	0.987	10-30 minutes Peptides
0.874	0.126	0.015	0.985	
0.979	0.021	0.004	0.996	
0.822	0.178	0.017	0.983	
0.970	0.030	0.000	1.000	
0.312	0.688	0.000	1.000	
0.916	0.084	0.032	0.968	
0.830	0.170	0.000	1.000	
0.857	0.143	0.009	0.991	
0.659	0.341	0.154	0.846	
0.964	0.036	0.011	0.989	
0.477	0.523	0.503	0.497	
0.932	0.068	0.033	0.967	
0.934	0.066	0.000	1.000	
0.934	0.066	0.005	0.995	
0.880	0.120	0.013	0.987	
0.948	0.052	0.000	1.000	
0.893	0.107	0.005	0.995	
0.885	0.115	0.059	0.941	
0.951	0.049	0.005	0.995	
0.633	0.367	0.014	0.986	
0.295	0.705	0.033	0.967	
0.831	0.169	0.000	1.000	
0.942	0.058	0.414	0.586	
0.903	0.097	0.010	0.990	
0.945	0.055	0.000	1.000	
0.950	0.050	0.007	0.993	
0.849	0.151	0.045	0.955	
0.932	0.068	0.002	0.998	
0.810	0.190	0.337	0.663	
0.580	0.420	0.008	0.992	
0.896	0.104	0.000	1.000	
0.845	0.155	0.000	1.000	
0.942	0.058	0.035	0.965	
0.945	0.055	0.000	1.000	
0.933	0.067	0.011	0.989	
0.936	0.064	0.003	0.997	
0.842	0.158	0.000	1.000	
0.932	0.068	0.000	1.000	

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

P	NP	L	NL	LC Region
0.933	0.067	0.026	0.974	10-30 minutes Peptides
0.934	0.066	0.027	0.973	
0.942	0.058	0.011	0.989	
0.610	0.390	0.000	1.000	
0.810	0.190	0.050	0.950	
0.925	0.075	0.000	1.000	
0.590	0.410	0.014	0.986	
0.880	0.120	0.029	0.971	
0.474	0.526	0.840	0.160	
0.933	0.067	0.026	0.974	
0.930	0.070	0.005	0.995	
0.929	0.071	0.004	0.996	
0.741	0.259	0.023	0.977	
0.748	0.252	0.090	0.910	
0.994	0.006	0.000	1.000	
0.832	0.168	0.259	0.741	
0.866	0.134	0.008	0.992	
0.912	0.088	0.019	0.981	
0.929	0.071	0.016	0.984	
0.721	0.279	0.009	0.991	
0.890	0.110	0.030	0.970	
0.938	0.062	0.008	0.992	
0.917	0.083	0.000	1.000	
0.924	0.076	0.014	0.986	
0.649	0.351	0.147	0.853	
0.602	0.398	0.025	0.975	
0.890	0.110	0.000	1.000	
0.881	0.119	0.006	0.994	
0.790	0.210	0.000	1.000	
0.912	0.088	0.005	0.995	
0.918	0.082	0.042	0.958	
0.902	0.098	0.204	0.796	
0.793	0.207	0.008	0.992	
0.010	0.990	0.992	0.008	55-80 minutes Lipids
0.012	0.988	0.996	0.004	
0.010	0.990	0.993	0.007	
0.350	0.650	0.909	0.091	
0.006	0.994	0.978	0.022	
0.005	0.995	0.989	0.011	

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

P	NP	L	NL	LC Region
0.026	0.974	0.969	0.031	55-80 minutes Lipids
0.029	0.971	0.992	0.008	
0.013	0.987	0.985	0.015	
0.015	0.985	0.886	0.114	
0.010	0.990	0.990	0.010	
0.010	0.990	0.991	0.009	
0.320	0.680	0.793	0.207	
0.297	0.703	0.856	0.144	
0.009	0.991	0.990	0.010	
0.019	0.981	0.986	0.014	
0.005	0.995	0.987	0.013	
0.483	0.517	0.996	0.004	
0.005	0.995	0.925	0.075	
0.005	0.995	0.991	0.009	
0.237	0.763	0.991	0.009	
0.003	0.997	0.956	0.044	
0.000	1.000	0.992	0.008	
0.136	0.864	0.883	0.117	
0.060	0.940	0.948	0.052	
0.146	0.854	0.987	0.013	
0.010	0.990	0.981	0.019	
0.007	0.993	0.993	0.007	
0.019	0.981	0.970	0.030	
0.007	0.993	0.971	0.029	
0.008	0.992	0.988	0.012	
0.006	0.994	0.992	0.008	
0.245	0.755	0.996	0.004	
0.468	0.532	0.997	0.003	
0.279	0.721	0.994	0.006	
0.003	0.997	0.995	0.005	
0.026	0.974	0.851	0.149	
0.011	0.989	0.982	0.018	
0.012	0.988	0.988	0.012	
0.018	0.982	0.978	0.022	
0.047	0.953	0.877	0.123	
0.009	0.991	0.996	0.004	
0.005	0.995	0.994	0.006	
0.005	0.995	0.995	0.005	
0.005	0.995	0.802	0.198	

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

P	NP	L	NL	LC Region
0.006	0.994	0.991	0.009	55-80 minutes
0.006	0.994	0.996	0.004	Lipids
0.006	0.994	0.983	0.017	
0.007	0.993	0.956	0.044	
0.014	0.986	0.990	0.010	
0.019	0.981	0.979	0.021	
0.017	0.983	0.994	0.006	
0.020	0.980	0.992	0.008	
0.040	0.960	0.956	0.044	
0.014	0.986	0.976	0.024	
0.004	0.996	0.893	0.107	
0.036	0.964	0.967	0.033	
0.110	0.890	0.998	0.002	
0.008	0.992	0.990	0.010	
0.008	0.992	0.938	0.062	
0.021	0.979	0.983	0.017	
0.004	0.996	0.981	0.019	
0.012	0.988	0.888	0.112	
0.011	0.989	0.990	0.010	
0.005	0.995	0.995	0.005	
0.004	0.996	0.969	0.031	
0.364	0.636	0.990	0.010	
0.145	0.855	0.678	0.322	
0.116	0.884	0.996	0.004	
0.098	0.902	0.963	0.037	
0.012	0.988	0.995	0.005	
0.012	0.988	0.986	0.014	
0.011	0.989	0.995	0.005	
0.032	0.968	0.997	0.003	
0.095	0.905	0.931	0.069	
0.017	0.983	0.993	0.007	
0.005	0.995	0.986	0.014	
0.012	0.988	0.957	0.043	
0.012	0.988	0.987	0.013	
0.497	0.503	0.994	0.006	
0.017	0.983	0.815	0.185	
0.006	0.994	0.990	0.010	
0.010	0.990	0.995	0.005	
0.005	0.995	0.987	0.013	

Table A.2. FFNN output scores for PEPNET and LIPNET in the Lipid and Peptide Separation.

P	NP	L	NL	LC Region
0.009	0.991	0.989	0.011	
0.005	0.995	0.883	0.117	30-55 minutes
0.006	0.994	0.877	0.123	Unknowns
0.009	0.991	0.905	0.095	
0.106	0.894	0.776	0.224	
0.015	0.985	0.946	0.054	
0.233	0.767	0.101	0.899	
0.942	0.058	0.020	0.980	
0.851	0.149	0.029	0.971	
0.973	0.027	0.004	0.996	
0.660	0.340	0.000	1.000	
0.016	0.984	0.942	0.058	
0.025	0.975	0.917	0.083	
0.011	0.989	0.955	0.045	
0.012	0.988	0.947	0.053	
0.009	0.991	0.952	0.048	

Neural network output scores for PEPNET and LIPNET corresponding to Figure 2.6 are displayed in Table A.2. Green cell shading indicates a correct component class prediction (both true positives and true negatives). Red cell shading indicates an incorrect component class prediction (both false positives and false negatives). For incorrectly predicted components, yellow cell shading indicates the true class. TP and TNP denote the peptide and non-peptide class targets for PEPNET, respectively. TL and TNL denote the lipid and non-lipid class targets for LIPNET, respectively.

A.12 Neural Network Performance Characteristics with Simulated Low Mass Resolution Data

Table A.3. Neural network performance metrics for PEPNET, LIPNET, and LIPSUBNET with simulated low mass resolution MS data.

	Performance Metric		
Network	Accuracy	TPR	TNR
PEPNET	100.0%	100.0%	100.0%
LIPNET	98.1%	100.0%	98.5%
LIPSUBNET	99.4%	97.4%	99.6%

Neural network performance metrics for PEPNET, LIPNET, and LIPSUBNET trained with *m/z* and KMD values restricted to two decimal places applied to the artificially combined, experimental test set used in Figure 5. The results for each network are very similar the networks trained with *m/z* and KMD values restricted to four decimal places. PEPNET exhibited perfect accuracy. LIPNET maintained 100% TPR, but the TNR fell to 98.5% (compared to 100% for the four-decimal network). LIPSUBNET exhibited a slight increase in performance with 97.4% TPR (compared to 94.9% for the four-decimal network) and 99.6% TNR (compared to the 99.3% for the four-decimal network).

A.13 Partial LC and Full m/z Convolution of Lipids that Differ by One Degree of Unsaturation

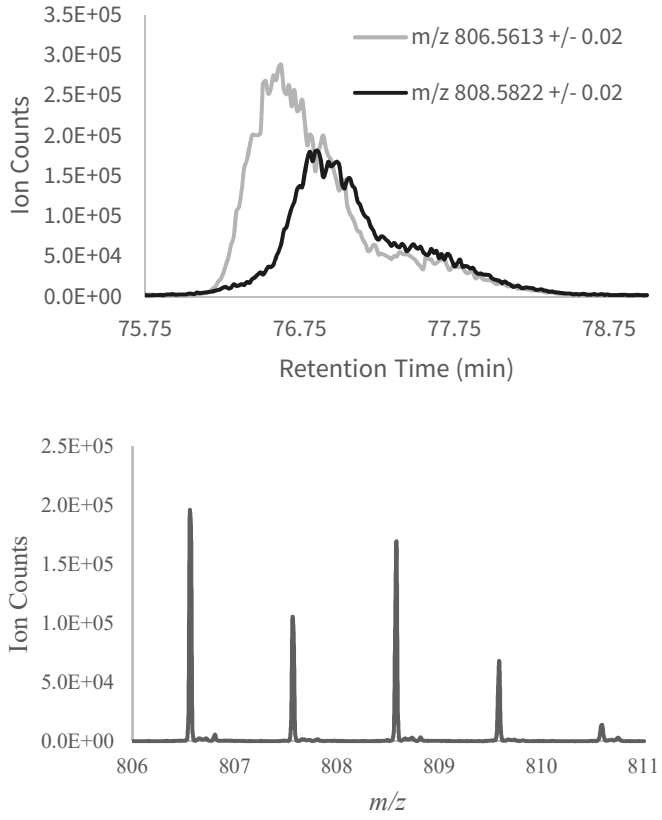


Figure A.8. Select ion chromatograms (top) and mass spectrum (bottom) of two partially convolved lipids that differ by one degree of saturation.

The plots display select ion chromatograms (left) and an averaged mass spectrum (right) of two partially LC-MS convoluted lipids that differ by one degree of unsaturation separated in the combined analysis of peptides and lipids. The first isotopologue of the saturated lipid (808.5822 m/z) becomes fully convolved with the third isotopologue of the unsaturated lipid (806.5613 m/z) shortly after the unsaturated lipid begins to elute.

A.14 Partial m/z Convolution of Lipids

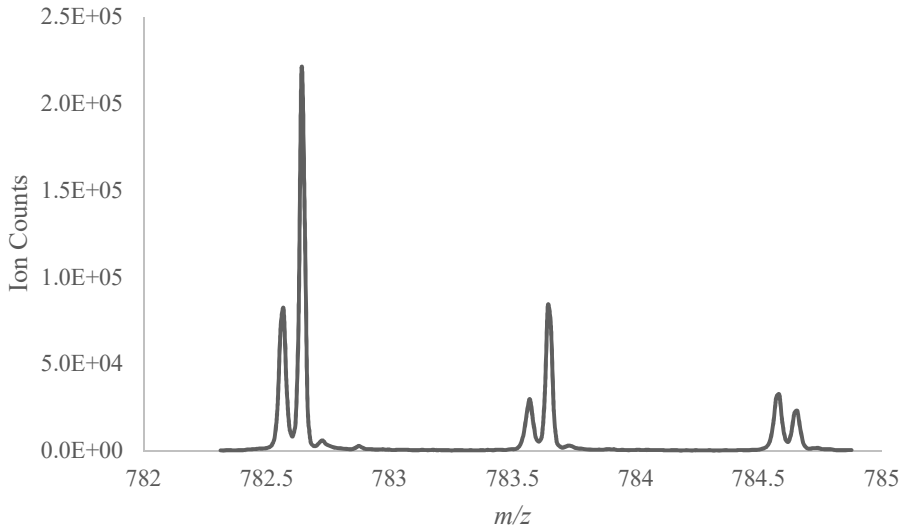


Figure A.9. Mass spectrum of two partially *m/z* convolved lipids.

The plot in Figure A.9 displays an averaged mass spectrum of two partially convolved lipids in the *m/z* domain separated in the combined analysis of peptides and lipids. Each isotopologue peak of both lipids is partially convolved with the neighboring isotopologue of the other lipid up to ~20% peak height.

APPENDIX B

Chemical Classification for Improved Lipidomics Sample Annotation with In Silico Fractionation

B.1 Table of Adducts that Modified Lipid Elemental Compositions in the iSF Training Set

Table B.1. Table displaying positive and negative polarity adducts commonly generated in ESI-MS experiments that were used to modify lipid elemental compositions of each class for iSF training set generation.

FA	GL	GP	PK	PL	SL	SP	ST
H	H	H	H	H	K	H	H
-H	NH_4	-H	NH_4	NH_4	NH_4	HCO_2	NH_4
NH_4	Na	$-CH_3$	Na	Na	Na	Na	Na
Na	K	Na	K	-H	-H	-H	-H
$-H_3O$	C_2H_4N	HCO_2	C_2H_4N	$-H_3O$	$-H_3O$	$C_2H_3O_2$	$-H_3O$

FA (fatty acyls), GL (glycerolipids), GP (glycerophospholipids), PK (polyketides), PL (prenol lipids), SL (saccharolipids), SP (sphingolipids), ST (sterol lipids)

This table displays the iSF training set adduct modifications for each lipid class, and each adduct was chosen for each class because it was commonly observed in positive and negative polarity LC-ESI-MS experiments.

B.2 *iSF Feedforward Neural Network Training Time as a Function of Performance*

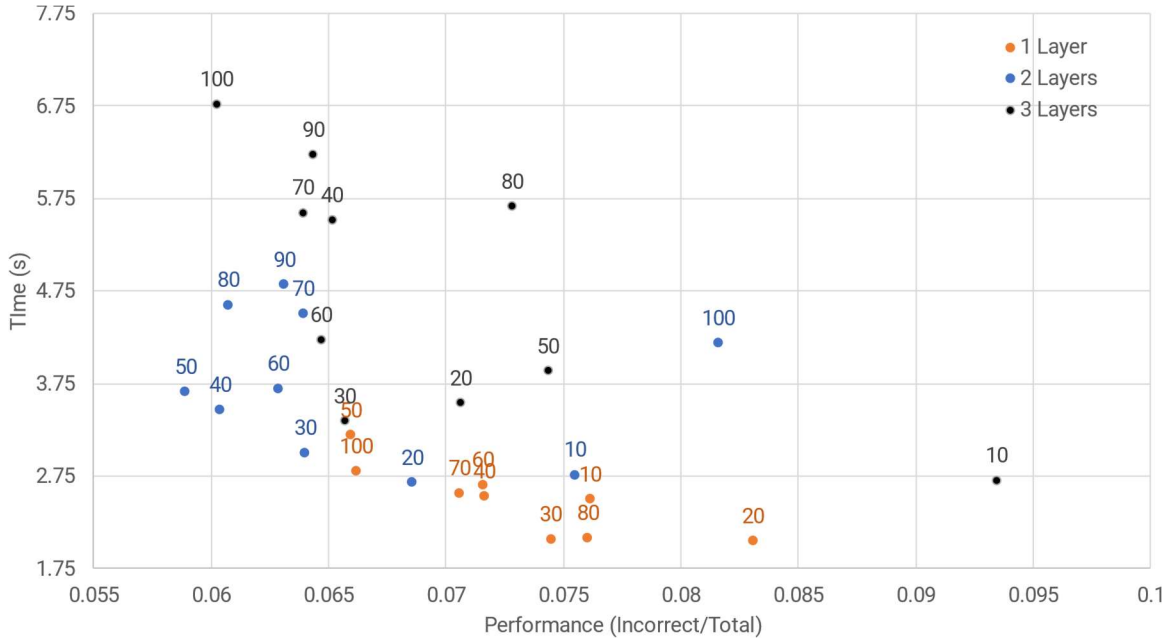


Figure B.1. Scatter plot displaying FFNN training time as a function of performance (incorrect predictions/total predictions).

To optimize the FFNN architecture, ten FFNNs were trained with each architecture (number of layers and hidden nodes per layer), and the mean training time and performance of each architecture was determined. Based on the relatively short, required training times, a configuration of two hidden layers with 50 nodes each was chosen for its slightly higher performance. The datum for 1 layer with 90 nodes had very poor performance (off the plot to the right) was not displayed for sake of visual clarity.

B.3 Proportion of Confident and Unconfident LipiDex MS/MS Lipid Assignments

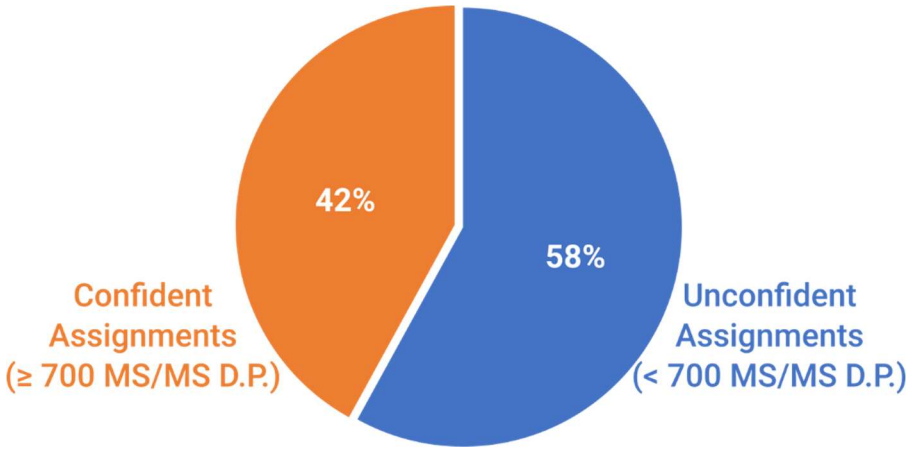


Figure B.2. Pie chart depicting the proportions of confident (orange) and unconfident (blue) assignments as defined by a MS/MS dot product score threshold of 700.

The plot displays the relative proportions of unconfident and confident assignments produced by the LipiDex MS/MS fragmentation spectral matching workflow. As show, over half of all assignments (58%) produced are unconfident and more likely to produce erroneous assignments.

B.4 Representation of each Lipid Class in the LipiDex Validation Positive and Negative Polarity LC-MS Experiments

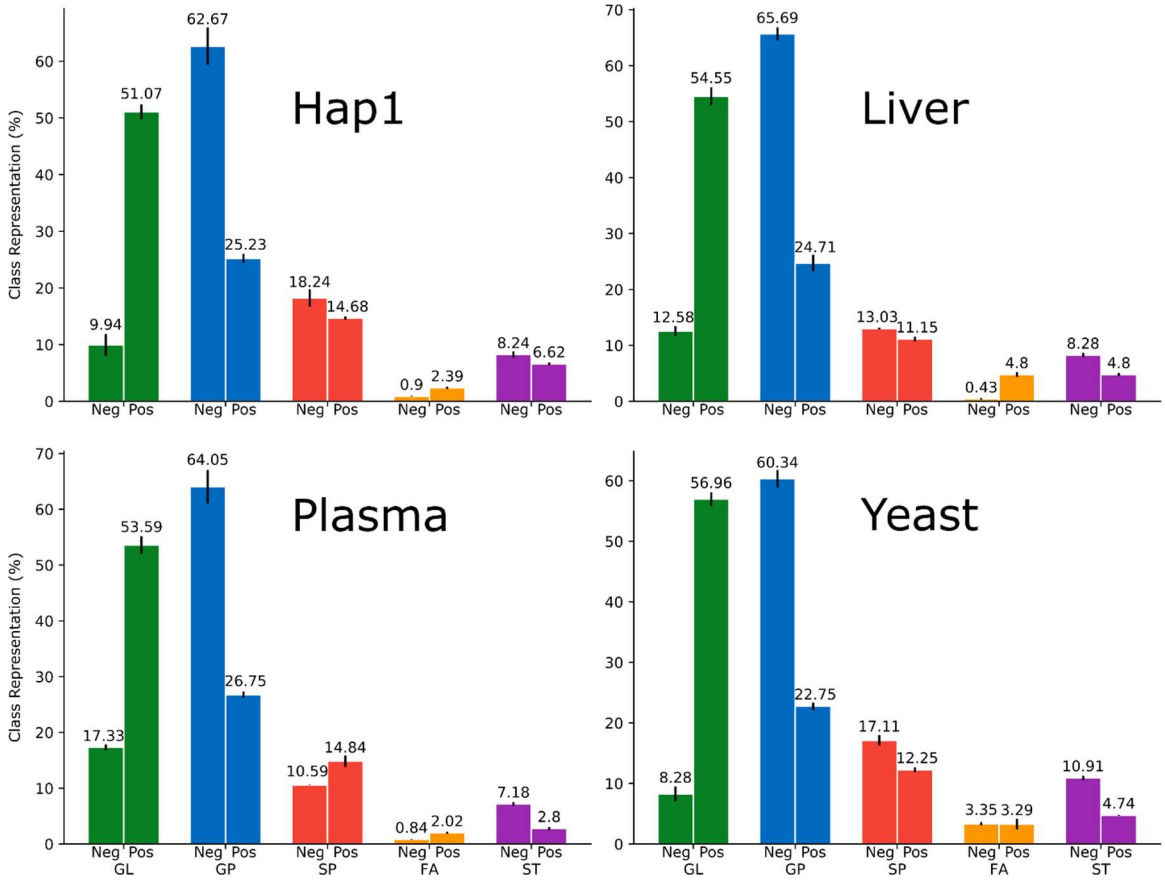


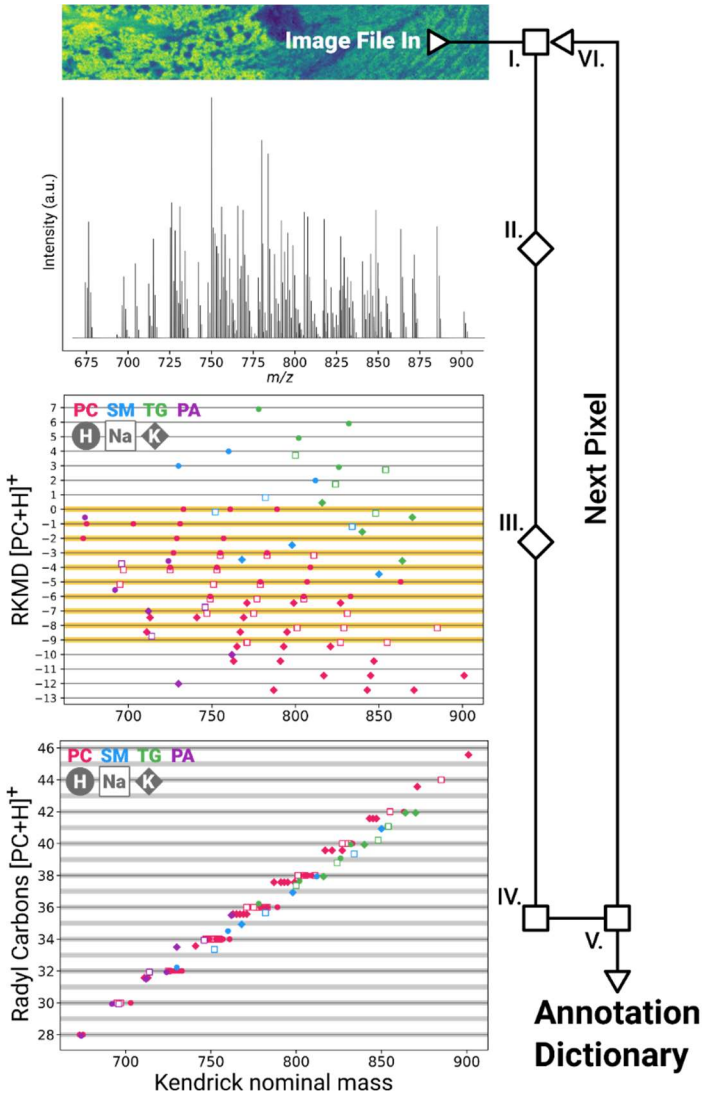
Figure B.3. Bar charts depicting the iSF predicted lipid class representation for both positive and negative polarity LC-MS experiments for each biological sample type.

The bar plots show the representation of each class present in each set of measurement replicates (n = 3) as predicted by iSF. Error bars represent the relative standard deviation. As shown, each class is consistently represented across each set of measurement replicates for each biological sample type.

APPENDIX C

Referenced Kendrick Mass Defect-Based Annotation and Filtering of Imaging MS Lipidomics Experiments

C.1 Representative Schematic of the RKMD-Based Annotation Workflow



Scheme C.1. A representative schematic of the RKMD-based annotation workflow.

A representative schematic of the RKMD-based lipid annotation workflow for assigning lipid classes, degree of unsaturation, and the number of radyl carbons. Input data utilizes the externally calibrated centroided image data saved in Python dictionary structure. At node I, data from the image is loaded, and the main processing loop begins operating on mass spectra at each image coordinates. The example mass spectrum shown in Fig. S1 (node II) contained 279 peaks (corresponding to 31 lipids) and was computationally generated by using the elemental compositions of singly-charged protonated [M+H]+, sodiated [M+Na]+, and potassiated [M+K]+ lipids; hence, the masses in this example did not require calibration. The relative abundance for each species (between m/z range ~670-910) was randomized (from zero to one hundred percent using Python's numpy package); pyOpenMS (2.6.0) Python package was used to calculate the theoretical isotopic patterns, and the first three isotopologues were included to generate the mass spectrum shown at node II. These 31 lipids included 21 phosphatidylcholine (PC), 3 sphingomyelin (SM), 4 triacylglycerol (TG), and 3 phosphatidic acid (PA) lipids, which were selected to show examples of different types of RKMD results (as discussed below). At node II, the centroided peaks are read for each pixel and aligned to the recalibrated averaged spectrum constructed from all mass spectra. At node III, the RKMD values calculated for [PC+H]+ for the theoretical 12Call peaks of several PC (pink), SM (blue), TG (green), and PA (purple) lipids with H+ (circle), Na+ (square), and K+ (diamond) adducts are displayed as a function of Kendrick

nominal mass (KNM). Mass spectrometry peaks are retained with RKMD distances (δ) within a user-defined window, covering RMKD values between 0 to -9 (thick horizontal yellow lines). Positive integer results, such as those for SM (blue) and TG (green) observed in the RKMD plot at node III, were considered unacceptable results not in agreement with physical reality as non-zero values that correspond to non-zero degrees of unsaturation result in negative integer values. At node IV, the number of radyl carbons of each feature is calculated and data points with radyl carbon distances (ε) within a user-defined window (thick horizontal gray lines) are retained. At node V, lipid peak assignment data for the molecular class and associated adducts are saved in a Python dictionary structure and, at node VI, the main processing loop proceeds to the next pixel until the last pixel is processed. The procedure is repeated for each lipid class and adduct combination.

C.2 Lipid Headgroup Elemental Compositions Used to Calculate RKMD

Table C.1. Elemental compositions of lipid headgroups for each lipid class used calculate RKMD.

Class	Headgroup elemental composition
TG	$C_6H_8O_6$
DG	$C_5H_8O_5$
MG	$C_4H_8O_4$
FA	$C_2H_4O_2$
SM	$C_9H_{21}N_2O_6P$
CER	$C_4H_9NO_3$
CERP	$C_4H_{10}NO_6P$
HCER	$C_{10}H_{19}NO_8$
PE	$C_7H_{14}NO_8P$
PC	$C_{10}H_{20}NO_8P$
PA	$C_5H_9O_8P$
PG	$C_8H_{15}O_{10}P$
O/P-PE	$C_6H_{14}NO_7P$
O/P-PC	$C_9H_{20}NO_7P$
O/P-PA	$C_4H_9O_7P$
O/P-PG	$C_7H_{15}O_9P$
LPE	$C_6H_{14}NO_7P$
LPC	$C_9H_{20}NO_7P$
LPA	$C_4H_9O_7P$
LPG	$C_7H_{15}O_9P$

This table displays the elemental compositions of each lipid headgroup that was used to calculate RKMD for each lipid class. Abbreviations are as follows: triacylglycerol (TG), diacylglycerol (DG), monoacylglycerol (MG), fatty acyl (FA), sphingomyelin (SM), ceramide (CER), ceramide-1-phosphate (CERP), hexosylceramides (HCER), phosphatidylethanolamine (PE), phosphatidylcholine (PC), phosphatidic acid (PA), phosphatidylglycerol (PG), ether-linked (O/P-), and lyso- (L).

C.3 Charged Adducts that Modified Reference Lipid Headgroup Elemental Compositions

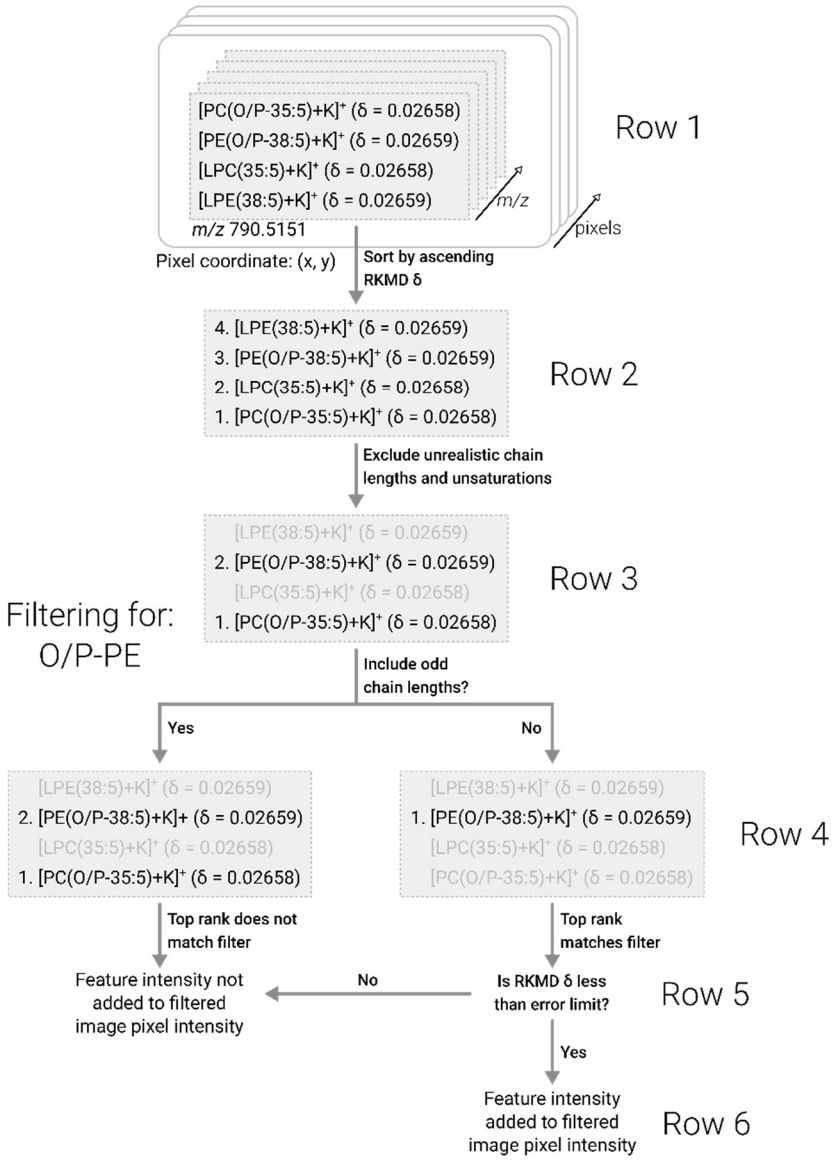
Table C.2. Adducts that were used to modify reference lipid headgroup elemental compositions to calculate RKMD for each lipid class.

TG	DG	MG	FA	SM	CER	CERP	HCER	PE	PC
Na^+	Na^+	Na^+	H^+	H^+	H^+	H^+	H^+	H^+	H^+
	$[H\text{-}H_2O]^+$	$[H\text{-}H_2O]^+$	Na^+	Na^+	Na^+	Na^+	Na^+	Na^+	Na^+
			K^+	K^+	K^+	K^+	K^+	K^+	K^+
					$[H\text{-}H_2O]^+$	$[H\text{-}H_2O]^+$	$[H\text{-}H_2O]^+$		

PA	PG	O/P-PE	O/P-PC	O/P-PA	O/P-PG	LPE	LPC	LPA	LPG
H^+	H^+	H^+	H^+	H^+	H^+	H^+	H^+	H^+	H^+
Na^+	Na^+	Na^+	Na^+	Na^+	Na^+	Na^+	Na^+	Na^+	Na^+
K^+	K^+	K^+	K^+	K^+	K^+	K^+	K^+	K^+	K^+

This table displays the charged adducts that were used to modify the reference lipid headgroup elemental compositions that were used to calculate RKMD for each lipid class. Abbreviations are as follows: triacylglycerol (TG), diacylglycerol (DG), monoacylglycerol (MG), fatty acyl (FA), sphingomyelin (SM), ceramide (CER), ceramide-1-phosphate (CERP), hexosylceramides (HCER), phosphatidylethanolamine (PE), phosphatidylcholine (PC), phosphatidic acid (PA), phosphatidylglycerol (PG), ether-linked (O/P-), and lyso- (L).

C.4 Representative Schematic of the Class-Based Image Filtering Workflow



Scheme C.2. A representative schematic of the class-based image filtering workflow.

An example schematic of the image filtering workflow for m/z 790.5151 where its RKMD assignments are filtered for the ether-linked PE (O/P-PE) class. Annotations are sorted and rank-ordered by ascending RKMD distance (δ, row 2)

from the nearest RKMD integer value and filtered by three heuristic constraints: 1) limiting numbers of radyl carbons, 2) limiting degrees of unsaturation, 3) and including odd chain lengths. In the third row, the calculated number of radyl carbons of the LPE and LPC annotations, 35 and 38, respectively, were larger than the upper limit for lysoglycerophospholipids and were, therefore, removed from consideration (text in grey, row 3). In the fourth row on the right, the exclusion of assignments with odd numbers of radyl carbons filters out the $[PC(O/P\text{-}35:5)+K]^+$ assignment. In the fourth row on the left, odd numbers of radyl carbons are included, and the $[PC(O/P\text{-}35:5)+K]^+$ assignment is retained in the top rank. If the top ranked assignment for the m/z 790.5151 matches the filter criteria (i.e., class) and the RKMD δ is less than an *m/z* dependent error limit, its peak intensity is added to the filtered image pixel intensity. In the fifth row on the left, the top ranked annotation, $[PC(O/P\text{-}35:5)+K]^+$, does not match the filter criterion, and, therefore, the feature intensity is not added to the image pixel intensity. In the fifth row on the right, the top ranked annotation, $[PE(O/P\text{-}38:5)+K]^+$, matches the filter criterion, and, therefore, the feature intensity is added to the image pixel intensity (row 6).

C.5 Heuristic Limits on Radyl Carbon Chain Lengths and Unsaturations for Filtering RKMD Annotations

Table C.3. Heuristic limits on radyl carbon chain lengths and unsaturations for filtering RKMD annotations.

Class	Minimum radyl carbons	Maximum radyl carbons	Maximum unsaturations
TG	48	70	9
DG	26	47	9
MG	12	25	5
FA	0	25	5
SM	26	47	3
CER	26	47	3
CERP	26	47	3
HCER	26	47	3
PE	26	47	9
PC	26	47	9
PA	26	47	9
PG	26	47	9
O/P-PE	26	47	9
O/P-PC	26	47	9
O/P-PA	26	47	9
O/P-PG	26	47	9
LPE	0	25	5
LPC	0	25	5
LPA	0	25	5
LPG	0	25	5

In order to limit false-positive assignments that result from RKMD indications of unrealistic degrees of unsaturation and radyl carbon chain length, we implemented the above heuristic limits based on commonly observed lipid species.

C.6 Relationship Between RKMD δ and PPM Mass Measurement Error

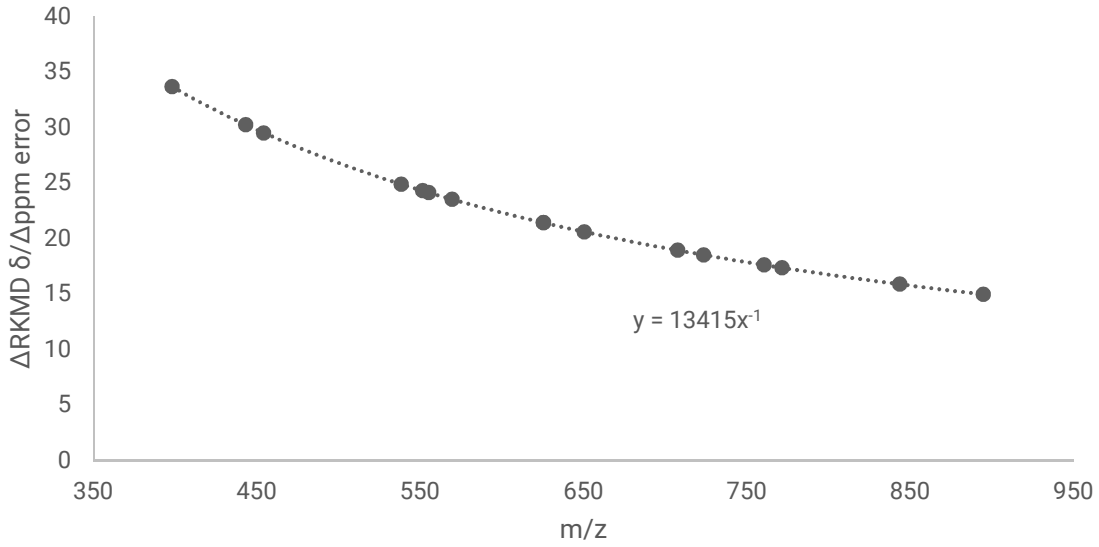


Figure C.1. Line plot displaying the inverse power function that relates ppm mass measurement error to RKMD δ as a function of *m/z*.

To calculate the datapoints in the above plot, the monoisotope *m/z* of 16 lipids from several lipid molecular classes were adjusted by ppm error factors between 0 and 1 by increments of 0.1, and the RKMD δ was calculated for each error adjusted *m/z* value. For each m/z, RKMD δ was plotted as a function of ppm error. Linear regression determined the slope which corresponds to the ΔRKMD δ/Δppm error. The ΔRKMD δ/Δppm error factor was plotted as a function of *m/z*, and a power regression determined an inverse power relationship. The root mean squared error of the fit was 2.49×10^{-4}.

Table C.4. LIPIDMAPS and RKMD-based lipid assignments from kidney tissue MALDI-IMS analysis.

		Ranked RKMD Assignments					
m/z	LIPIDMAPS	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
666.4815	CERP(36:2)+Na	CERP(36:2)+Na* (0.13368)	CERP(38:5)+H (0.31286)	None	None	None	None
697.478	PA(34:1)+Na	PA(34:1)+Na* (0.00937)	PA(36:4)+H (0.1698)	TG(38:4)+K (0.17839)	None	None	None
703.5781	SM(34:1)+H	SM(34:1)+H* (0.24233)	None	None	None	None	None
721.4783	PA(36:3)+Na	PA(36:3)+Na* (0.03184)	PA(38:6)+H (0.14734)	TG(40:6)+K (0.15593)	None	None	None
723.4945	PA(36:2)+Na	PA(36:2)+Na* (0.07279)	PA(38:5)+H (0.10639)	TG(40:5)+K (0.11498)	None	None	None
725.5586	SM(34:1)+Na	SM(34:1)+Na* (0.13467)	None	None	None	None	None
731.607	SM(36:1)+H	SM(36:1)+H* (0.06342)	None	None	None	None	None
732.5547	PC(32:1)+H	PC(32:1)+H* (0.06863)	PE(35:1)+H (0.06864)	None	None	None	None
734.5697	PC(32:0)+H	PC(32:0)+H* (0.02012)	PE(35:0)+H (0.02014)	None	None	None	None
737.4533	PA(36:3)+K	PA(36:3)+K* (0.11106)	LPA(39:9)+Na (0.12269)	None	None	None	None
739.4669	PA(36:2)+K	PA(36:2)+K* (0.0418)	None	None	None	None	None
741.5312	SM(34:1)+K	SM(34:1)+K* (0.03499)	None	None	None	None	None

Table C.4. LIPIDMAPS and RKMD-based lipid assignments from kidney tissue MALDI-IMS analysis.

		Ranked RKMD Assignments					
m/z	LIPIDMAPS	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
745.4753	PA(38:5)+Na	PA(38:5)+Na* (0.19169)	PG(O/P-32:1)+K (0.20323)	LPG(32:1)+K (0.20323)	None	None	None
747.4931	PA(38:4)+Na	PA(38:4)+Na* (0.03147)	PG(O/P-32:0)+K (0.04301)	LPG(32:0)+K (0.04301)	PA(40:7)+H (0.21065)	TG(42:7)+K (0.21924)	None
749.5076	PA(38:3)+Na	PA(38:3)+Na* (0.11725)	PA(40:6)+H (0.29643)	TG(42:6)+K (0.30501)	None	None	None
756.5526	PC(32:0)+Na	PE(37:3)+H (0.0878)	PC(34:3)+H (0.08781)	PC(32:0)+Na* (0.09137)	PE(35:0)+Na (0.09138)	None	None
758.57	PC(34:2)+H	PC(34:2)+H* (0.04259)	PE(37:2)+H (0.0426)	None	None	None	None
760.5842	PC(34:1)+H	PE(37:1)+H (0.06554)	PC(34:1)+H* (0.06555)	None	None	None	None
763.4679	PA(38:4)+K	PA(38:4)+K* (0.03284)	None	None	None	None	None
768.5888	PC(O-34:1)+Na	PC(O/P-34:1)+Na* (0.0776)	LPC(34:1)+Na (0.0776)	PE(O/P-37:1)+Na (0.07761)	LPE(37:1)+Na (0.07761)	PE(O/P-39:4)+H (0.10156)	LPE(39:4)+H (0.10156)
772.5241	PE(O-38:6)+Na	PE(O/P-38:6)+Na* (0.07867)	LPE(38:6)+Na (0.07867)	PC(O/P-35:6)+Na (0.07868)	LPC(35:6)+Na (0.07868)	PE(35:0)+K (0.0903)	PC(32:0)+K (0.09031)
772.5842	PE(38:2)+H	PE(38:2)+H* (0.06549)	PC(35:2)+H (0.0655)	None	None	None	None
774.5977	PE(38:1)+H	PE(38:1)+H* (0.22581)	PC(35:1)+H (0.22582)	None	None	None	None
782.5688	PC(36:4)+H	PE(39:4)+H (0.04675)	PC(36:4)+H* (0.04676)	PC(34:1)+Na (0.13242)	PE(37:1)+Na (0.13243)	None	None
784.5845	PC(36:3)+H	PE(39:3)+H (0.04307)	PC(36:3)+H* (0.04308)	PC(34:0)+Na (0.13609)	PE(37:0)+Na (0.13611)	None	None

Table C.4. LIPIDMAPS and RKMD-based lipid assignments from kidney tissue MALDI-IMS analysis.

		Ranked RKMD Assignments					
m/z	LIPIDMAPS	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
786.5998	PC(36:2)+H	PE(39:2)+H (0.06921)	PC(36:2)+H* (0.06923)	None	None	None	None
788.6134	PC(36:1)+H	PE(39:1)+H (0.22208)	PC(36:1)+H* (0.22209)	None	None	None	None
790.5151	PE(P-38:5)+K	PC(O/P-35:5)+K (0.02658)	LPC(35:5)+K (0.02658)	PE(O/P-38:5)+K* (0.02659)	LPE(38:5)+K (0.02659)	None	None
796.5261	PC(34:2)+K	PC(34:2)+K* (0.05888)	PE(37:2)+K (0.05889)	PC(O/P-37:8)+Na (0.07051)	LPC(37:8)+Na (0.07051)	PE(O/P-40:8)+Na (0.07052)	LPE(40:8)+Na (0.07052)
798.5404	PE(O-40:7)+Na	PE(O/P-40:7)+Na* (0.03016)	LPE(40:7)+Na (0.03016)	PC(O/P-37:7)+Na (0.03018)	LPC(37:7)+Na (0.03018)	PE(37:1)+K (0.04179)	PC(34:1)+K (0.04181)
806.5687	PC(38:6)+H	PE(41:6)+H (0.0541)	PC(38:6)+H* (0.05411)	PC(36:3)+Na (0.12507)	PE(39:3)+Na (0.12508)	None	None
808.5847	PC(38:5)+H	PE(41:5)+H (0.02806)	PC(38:5)+H* (0.02807)	PC(36:2)+Na (0.15111)	PE(39:2)+Na (0.15112)	None	None
810.6008	PC(38:4)+H	PC(38:4)+H* (0.00542)	PE(41:4)+H (0.00543)	PC(36:1)+Na (0.1846)	PE(39:1)+Na (0.18461)	None	None
813.686	SM(42:2)+H	SM(42:2)+H* (0.11938)	None	None	None	None	None
815.7006	SM(42:1)+H	SM(42:1)+H* (0.04106)	None	None	None	None	None
818.5448	PE(O-40:5)+K	PE(O/P-40:5)+K* (0.09268)	LPE(40:5)+K (0.09268)	PC(O/P-37:5)+K (0.09269)	LPC(37:5)+K (0.09269)	None	None

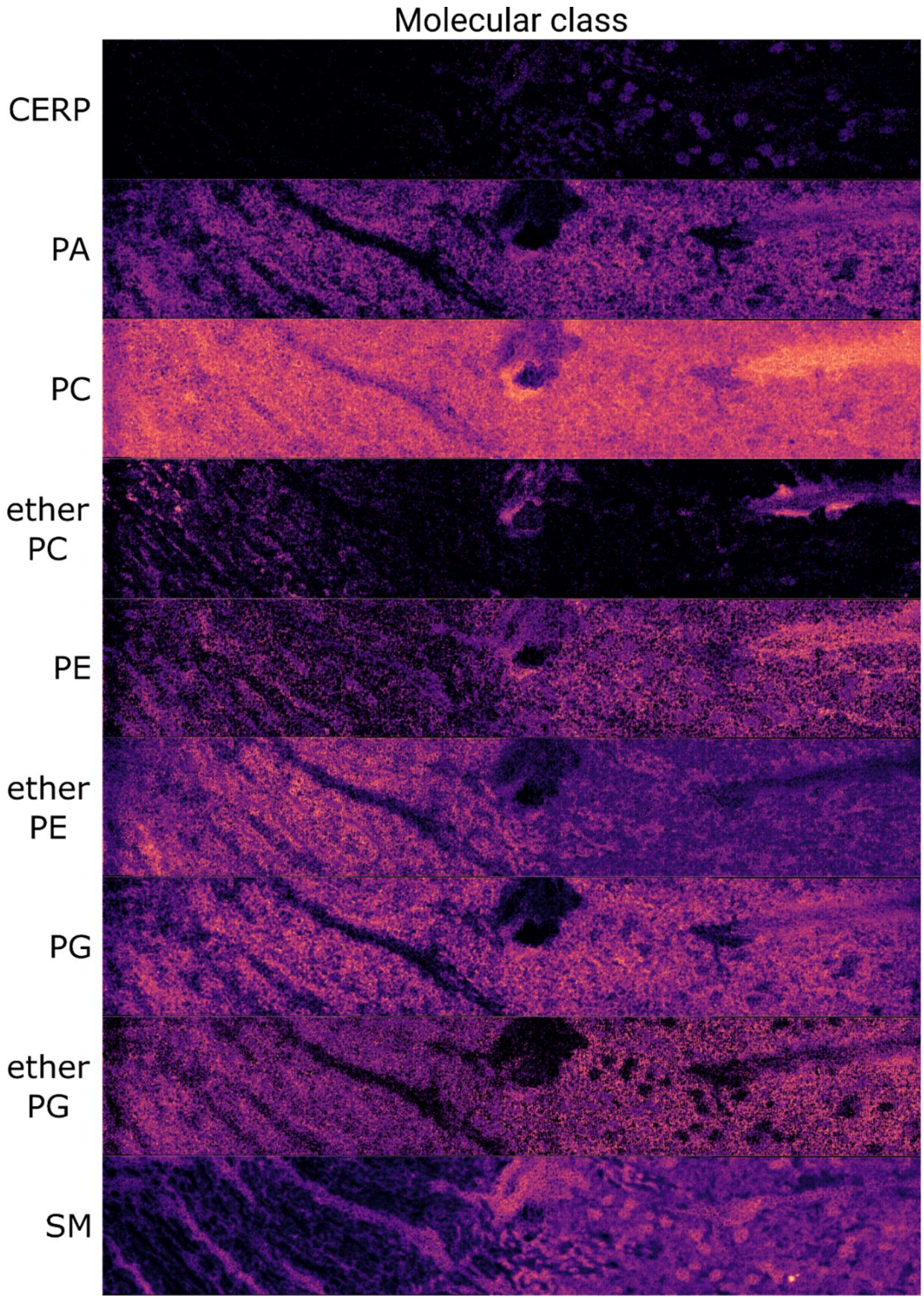
Table C.4. LIPIDMAPS and RKMD-based lipid assignments from kidney tissue MALDI-IMS analysis.

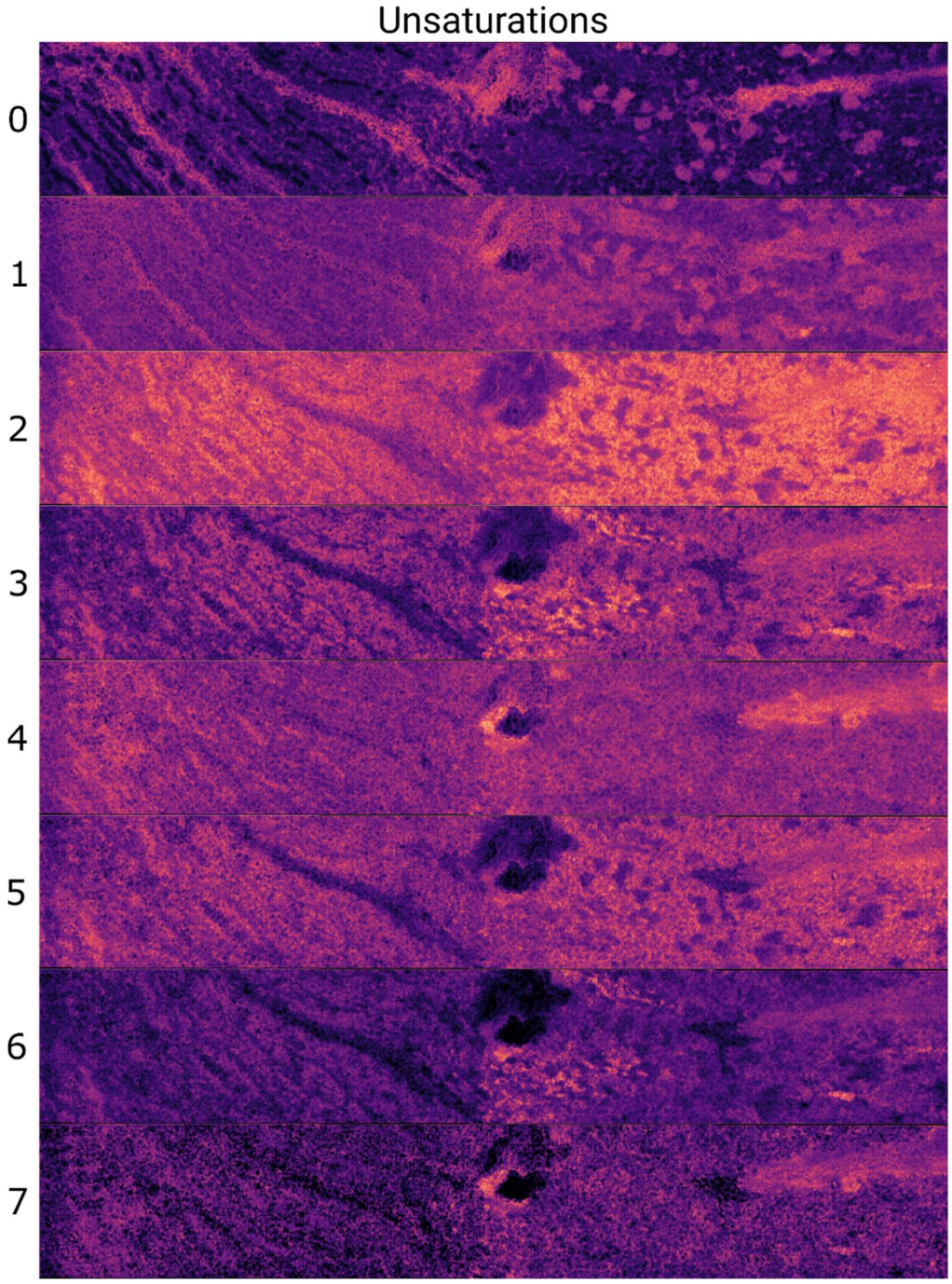
		Ranked RKMD Assignments					
m/z	LIPIDMAPS	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
822.5417	PC(36:3)+K	PC(36:3)+K* (0.0552)	PE(39:3)+K (0.05522)	PC(O/P-39:9)+Na (0.06683)	LPC(39:9)+Na (0.06683)	PE(O/P-42:9)+Na (0.06685)	LPE(42:9)+Na (0.06685)
824.5578	PC(36:2)+K	PC(36:2)+K* (0.0887)	PE(39:2)+K (0.08871)	PC(O/P-39:8)+Na (0.10032)	LPC(39:8)+Na (0.10032)	PE(O/P-42:8)+Na (0.10034)	LPE(42:8)+Na (0.10034)
826.5708	PE(O-42:7)+Na	PE(O/P-42:7)+Na* (0.09725)	LPE(42:7)+Na (0.09725)	PC(O/P-39:7)+Na (0.09727)	LPC(39:7)+Na (0.09727)	PE(39:1)+K (0.10888)	PC(36:1)+K (0.1089)
830.5633	PC(38:5)+Na	PE(41:5)+Na (0.27735)	PC(38:5)+Na* (0.27736)	None	None	None	None
832.5837	PC(38:4)+Na	PC(38:4)+Na* (0.07667)	PE(41:4)+Na (0.07668)	PE(43:7)+H (0.1025)	PC(40:7)+H (0.10251)	None	None
834.6001	PC(40:6)+H	PE(43:6)+H (0.04664)	PC(40:6)+H* (0.04666)	PC(38:3)+Na (0.13252)	PE(41:3)+Na (0.13254)	None	None
835.6693	SM (42:2)+Na	SM(42:2)+Na* (0.22045)	None	None	None	None	None
848.5602	PC(38:4)+K	PC(38:4)+K* (0.2677)	PE(41:4)+K (0.26772)	None	None	None	None

Kidney tissue lipids detected by MALDI-MS were identified by mass accuracies resulting in < 3 ppm mass error and LIPIDMAPS database searching, resulting in 44 lipid assignments. The m/z values that produced LIPIDMAPS identifications were submitted for RKMD-based annotation and are displayed by ascending RKMD δ. Final RKMD

assignments (marked with asterisk) were made through application of the same heuristics used in the RKMD image filtering workflow and were in full agreement with assignments based on mass accuracy, biological intuition, and prior experimentation. In most cases, the top ranked RKMD assignment matches the LIPIDMAPS assignment. All cases in which the top ranked RKMD assignment does not match the LIPIDMAPS assignment are produced from PC and PE species, particularly for ether linkage isomers. The even chain PC and odd chain PE assignments were structural isomers for twenty-seven m/z values and, therefore, could not be discriminated by mass accuracy alone. Additionally, differences in RKMD δ values of ~3E-5 between two putative assignments were not sufficient for identification. Similarly, the rank order of O/P- and lysoglycerophospholipids with the same headgroup was insignificant given their identical headgroup KMD values; for instance, m/z 769.5888, 772.5241, 790.5151 in Table C.4 produced O/P-PC and LPC (as well as O/P-PE and LPE) assignments with identical RKMD δ values. The final assignments in these cases (bolded) were made excluding odd chain radyl carbon chains and considering realistic radyl carbon chain lengths for each class.

C.8 High Level Lipid Class Images Produced by the RKMD-based Annotation and Image Filtering Method




Unsaturations
0
1
2
3
4
5
6
7

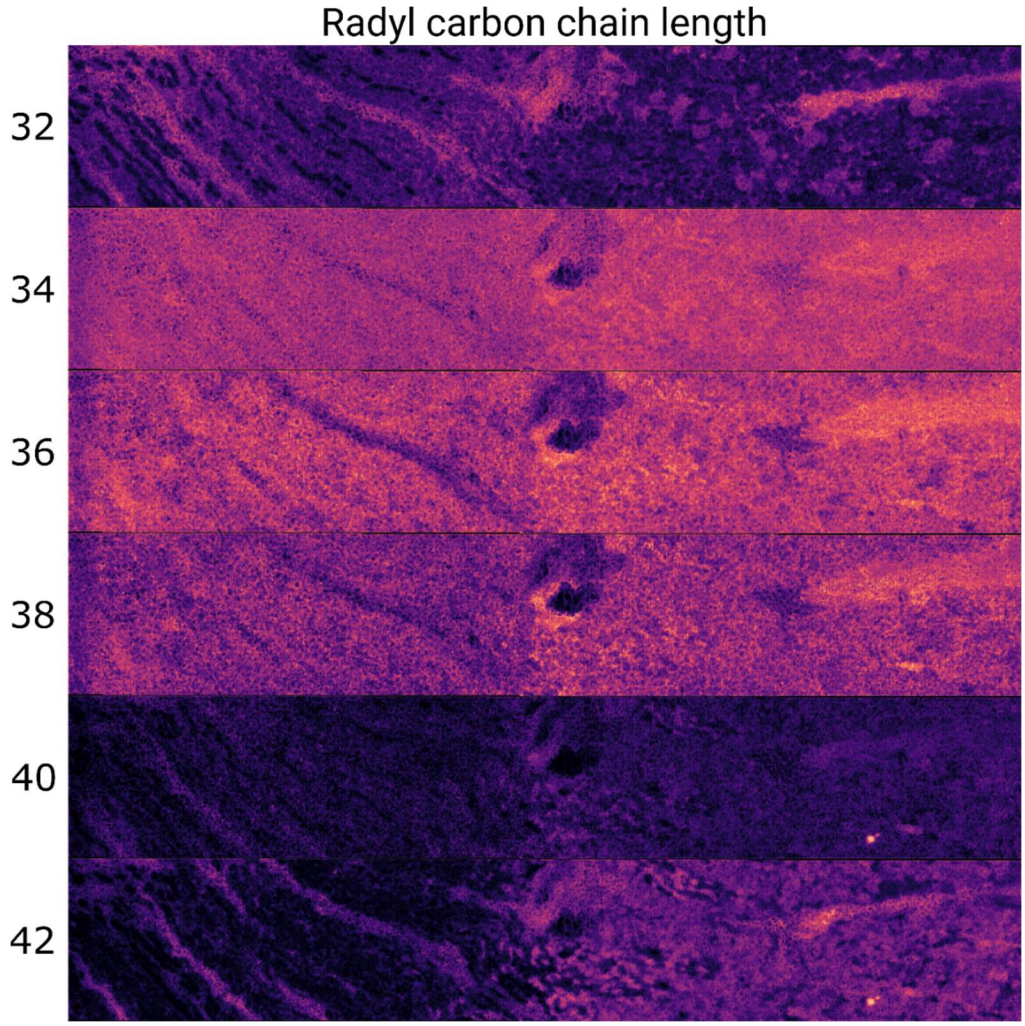


Figure C.2. High level class images from the MALDI-IMS analysis of human kidney tissue.

High level class images include those composite images that are reconstructed from contributions of broad groups of lipids that are related by their headgroup, degree of unsaturation, or number of radyl carbons. These images can be used to assess broad differences in localization between classes of lipids.

BIBLIOGRAPHY

1. Beger, R., A Review of Applications of Metabolomics in Cancer. *Metabolites* **2013,** *3* (3), 552-574.

2. Madsen, R.; Lundstedt, T.; Trygg, J., Chemometrics in metabolomics—A review in human disease diagnosis. *Analytica Chimica Acta* **2010,** *659* (1-2), 23-33.

3. Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O., Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018,** *8* (2), 31.

4. Griffiths, W. J.; Wang, Y., Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chemical Society Reviews* **2009,** *38* (7), 1882.

5. Blaženović, I.; Kind, T.; Sa, M. R.; Ji, J.; Vaniya, A.; Wancewicz, B.; Roberts, B. S.; Torbašinović, H.; Lee, T.; Mehta, S. S.; Showalter, M. R.; Song, H.; Kwok, J.; Jahn, D.; Kim, J.; Fiehn, O., Structure Annotation of All Mass Spectra in Untargeted Metabolomics. *Analytical Chemistry* **2019,** *91* (3), 2155-2162.

6. Kachman, M.; Habra, H.; Duren, W.; Wigginton, J.; Sajjakulnukit, P.; Michailidis, G.; Burant, C.; Karnovsky, A., Deep annotation of untargeted LC-MS metabolomics data with Binner. *Bioinformatics* **2020,** *36* (6), 1801-1806.

7. Liebisch, G.; Vizcaíno, J. A.; Köfeler, H.; Trötzmüller, M.; Griffiths, W. J.; Schmitz, G.; Spener, F.; Wakelam, M. J. O., Shorthand notation for lipid structures derived from mass spectrometry. *Journal of Lipid Research* **2013,** *54* (6), 1523-1530.

8. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016,** *8* (1).

9. Hastings, J.; Adams, N.; Ennis, M.; Hull, D.; Steinbeck, C., Chemical ontologies: what are they, what are they for and what are the challenges. *Journal of Cheminformatics* **2011,** *3* (S1), O4.

10. Kim, S.; Kramer, R. W.; Hatcher, P. G., Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. *Analytical Chemistry* **2003,** *75* (20), 5336-5344.

11. Minor, E. C.; Steinbring, C. J.; Longnecker, K.; Kujawinski, E. B., Characterization of dissolved organic matter in Lake Superior and its watershed using ultrahigh resolution mass spectrometry. *Organic Geochemistry* **2012,** *43*, 1-11.

12. Pereira, T. M. C.; Vanini, G.; Oliveira, E. C. S.; Cardoso, F. M. R.; Fleming, F. P.; Neto, A. C.; Lacerda, V.; Castro, E. V. R.; Vaz, B. G.; Romão, W., An evaluation of the aromaticity of asphaltenes using atmospheric pressure photoionization Fourier transform ion cyclotron resonance mass spectrometry – APPI(±)FT-ICR MS. *Fuel* **2014,** *118*, 348-357.

13. Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S., Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* **2021,** *39* (4), 462-471.

14. Picache, J. A.; May, J. C.; McLean, J. A., Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility-Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization. *Analytical Chemistry* **2020,** *92* (15), 10759-10767.

15. Richardson, L. T.; Brantley, M. R.; Solouki, T., Using isotopic envelopes and neural decision tree-based in silico fractionation for biomolecule classification. *Analytica Chimica Acta* **2020,** *1112*, 34-45.

16. Zhang, H.; Zhang, Y.; Shi, Q.; Ren, S.; Yu, J.; Ji, F.; Luo, W.; Yang, M., Characterization of low molecular weight dissolved natural organic matter along the treatment trait of a waterworks using Fourier transform ion cyclotron resonance mass spectrometry. *Water Research* **2012,** *46* (16), 5197-5204.

17. Wang, C.-F.; Fan, X.; Zhang, F.; Wang, S.-Z.; Zhao, Y.-P.; Zhao, X.-Y.; Zhao, W.; Zhu, T.-G.; Lu, J.-L.; Wei, X.-Y., Characterization of humic acids extracted from a lignite and interpretation for the mass spectra. *RSC Advances* **2017,** *7* (33), 20677-20684.

18. Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K., Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra. *Analytical Chemistry* **2001,** *73* (19), 4676-4681.

19. Kendrick, E., A Mass Scale Based on CH2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. *Analytical Chemistry* **1963,** *35* (13), 2146-2154.

20. Spiegel, M. T.; Anthony, I. G. M.; Brantley, M. R.; Hassell, A.; Farmer, P. J.; Solouki, T., Reactivities of Aromatic Protons in Crude Oil Fractions toward Br2 Tagging for Structural Characterization by Nuclear Magnetic Resonance and Electron Paramagnetic Resonance Spectroscopy and Mass Spectrometry. *Energy & Fuels* **2018,** *32* (10), 10549-10555.

21. Korf, A.; Vosse, C.; Schmid, R.; Helmer, P. O.; Jeck, V.; Hayen, H., Three-dimensional Kendrick mass plots as a tool for graphical lipid identification. *Rapid Communications in Mass Spectrometry* **2018,** *32* (12), 981-991.

22. Kune, C.; McCann, A.; Raphaël, L. R.; Arias, A. A.; Tiquet, M.; Van Kruining, D.; Martinez, P. M.; Ongena, M.; Eppe, G.; Quinton, L.; Far, J.; De Pauw, E., Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging. *Analytical Chemistry* **2019,** *91* (20), 13112-13118.

23. Lerno, L. A.; German, J. B.; Lebrilla, C. B., Method for the Identification of Lipid Classes Based on Referenced Kendrick Mass Analysis. *Analytical Chemistry* **2010,** *82* (10), 4236-4245.

24. Gross, J. H., *Mass spectrometry: a textbook*. Springer Science & Business Media: 2006.

25. Skoog, D. A.; Holler, F. J.; Nieman, T. A., *Principles of instrumental analysis*. 5th ed. ed.; Saunders College Pub. ; Harcourt Brace College Publishers: Philadelphia; Orlando, Fla., 1998.

26. El-Aneed, A.; Cohen, A.; Banoub, J., Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. *Applied Spectroscopy Reviews* **2009,** *44* (3), 210-230.

27. Herbert, C. G.; Johnstone, R. A., *Mass spectrometry basics*. CRC press: 2002.

28. Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S., Unraveling the Mechanism of Electrospray Ionization. *Analytical Chemistry* **2013,** *85* (1), 2-9.

29. Kebarle, P.; Verkerk, U. H., Electrospray: From ions in solution to ions in the gas phase, what we know now. *Mass Spectrometry Reviews* **2009,** *28* (6), 898-917.

30. Andrade, F. J.; Shelley, J. T.; Wetzel, W. C.; Webb, M. R.; Gamez, G.; Ray, S. J.; Hieftje, G. M., Atmospheric Pressure Chemical Ionization Source. 1. Ionization of Compounds in the Gas Phase. *Analytical Chemistry* **2008,** *80* (8), 2646-2653.

31. Robb, D. B.; Covey, T. R.; Bruins, A. P., Atmospheric Pressure Photoionization: An Ionization Method for Liquid Chromatography−Mass Spectrometry. *Analytical Chemistry* **2000,** *72* (15), 3653-3659.

32. Huang, M.-Z.; Yuan, C.-H.; Cheng, S.-C.; Cho, Y.-T.; Shiea, J., Ambient Ionization Mass Spectrometry. *Annual Review of Analytical Chemistry* **2010,** *3* (1), 43-65.

33. Kruve, A.; Kaupmees, K., Adduct Formation in ESI/MS by Mobile Phase Additives. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (5), 887-894.

34. Zhu, J.; Cole, R. B., Formation and decompositions of chloride adduct ions, [M + Cl]−, in negative ion electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2000,** *11* (11), 932-941.

35. Hsu, F.-F.; Bohrer, A.; Turk, J., Formation of lithiated adducts of glycerophosphocholine lipids facilitates their identification by electrospray ionization tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry* **1998,** *9* (5), 516-526.

36. Byrdwell, W. C., Atmospheric pressure chemical ionization mass spectrometry for analysis of lipids. *Lipids* **2001,** *36* (4), 327-346.

37. Hanold, K. A.; Fischer, S. M.; Cormia, P. H.; Miller, C. E.; Syage, J. A., Atmospheric Pressure Photoionization. 1. General Properties for LC/MS. *Analytical Chemistry* **2004,** *76* (10), 2842-2851.

38. Houck, M. M.; Siegel, J. A., Chapter 5 - Light and Matter. In *Fundamentals of Forensic Science (Third Edition)*, Houck, M. M.; Siegel, J. A., Eds. Academic Press: San Diego, 2015; pp 93-119.

39. Knochenmuss, R., Ion formation mechanisms in UV-MALDI. *Analyst* **2006,** *131* (9), 966-986.

40. Niessen, W. M. A., *Liquid Chromatography-Mass Spectrometry*. CRC Press: 2006.

41. Ahadi, E.; Konermann, L., Ejection of Solvated Ions from Electrosprayed Methanol/Water Nanodroplets Studied by Molecular Dynamics Simulations. *Journal of the American Chemical Society* **2011,** *133* (24), 9354-9363.

42. Daub, C. D.; Cann, N. M., How Are Completely Desolvated Ions Produced in Electrospray Ionization: Insights from Molecular Dynamics Simulations. *Analytical Chemistry* **2011,** *83* (22), 8372-8376.

43. Nguyen, S.; Fenn, J. B., Gas-phase ions of solute species from charged droplets of solutions. *Proc Natl Acad Sci U S A* **2007,** *104* (4), 1111-1117.

44. Smith, J. N.; Flagan, R. C.; Beauchamp, J. L., Droplet Evaporation and Discharge Dynamics in Electrospray Ionization. *The Journal of Physical Chemistry A* **2002,** *106* (42), 9957-9967.

45. Fenn, J. B., Electrospray Wings for Molecular Elephants (Nobel Lecture). *Angewandte Chemie International Edition* **2003,** *42* (33), 3871-3894.

46. Cech, N. B.; Enke, C. G., Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews* **2001,** *20* (6), 362-387.

47. Hogan, C. J.; Carroll, J. A.; Rohrs, H. W.; Biswas, P.; Gross, M. L., Combined Charged Residue-Field Emission Model of Macromolecular Electrospray Ionization. *Analytical Chemistry* **2009,** *81* (1), 369-377.

48. Dole, M.; Hines, R. L.; Mack, L. L.; Mobley, R. C.; Ferguson, L. D.; Alice, M. B., Gas Phase Macroions. *Macromolecules* **1968,** *1* (1), 96-97.

49. Iavarone, A. T.; Williams, E. R., Mechanism of Charging and Supercharging Molecules in Electrospray Ionization. *Journal of the American Chemical Society* **2003,** *125* (8), 2319-2327.

50. Kaltashov, I. A.; Mohimen, A., Estimates of Protein Surface Areas in Solution by Electrospray Ionization Mass Spectrometry. *Analytical Chemistry* **2005,** *77* (16), 5370-5379.

51. Fernandez de la Mora, J., Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Analytica Chimica Acta* **2000,** *406* (1), 93-104.

52. Heck, A. J. R.; van den Heuvel, R. H. H., Investigation of intact protein complexes by mass spectrometry. *Mass Spectrometry Reviews* **2004,** *23* (5), 368-389.

53. Verkerk, U. H.; Kebarle, P., Ion-Ion and Ion-Molecule Reactions at the Surface of Proteins Produced by Nanospray. Information on the Number of Acidic Residues and Control of the Number of Ionized Acidic and Basic Residues. *Journal of the American Society for Mass Spectrometry* **2005,** *16* (8), 1325-1341.

54. Patriksson, A.; Marklund, E.; van der Spoel, D., Protein Structures under Electrospray Conditions. *Biochemistry* **2007,** *46* (4), 933-945.

55. Fersht, A.; Fersht, U. A.; Freeman, W. H.; Company, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman: 1999.

56. Ahadi, E.; Konermann, L., Modeling the Behavior of Coarse-Grained Polymer Chains in Charged Water Droplets: Implications for the Mechanism of Electrospray Ionization. *The Journal of Physical Chemistry B* **2012,** *116* (1), 104-112.

57. Chung, J. K.; Consta, S., Release Mechanisms of Poly(ethylene glycol) Macroions from Aqueous Charged Nanodroplets. *The Journal of Physical Chemistry B* **2012,** *116* (19), 5777-5785.

58. Maier, S. K.; Hahne, H.; Gholami, A. M.; Balluff, B.; Meding, S.; Schoene, C.; Walch, A. K.; Kuster, B., Comprehensive Identification of Proteins from MALDI Imaging*. *Molecular & Cellular Proteomics* **2013,** *12* (10), 2901-2910.

59. Yang, J.; Caprioli, R. M., Matrix Sublimation/Recrystallization for Imaging Proteins by Mass Spectrometry at High Spatial Resolution. *Analytical Chemistry* **2011,** *83* (14), 5728-5734.

60. Caprioli, R. M.; Farmer, T. B.; Gile, J., Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Analytical Chemistry* **1997,** *69* (23), 4751-4760.

61. Jackson, S. N.; Ugarov, M.; Egan, T.; Post, J. D.; Langlais, D.; Albert Schultz, J.; Woods, A. S., MALDI-ion mobility-TOFMS imaging of lipids in rat brain tissue. *Journal of Mass Spectrometry* **2007,** *42* (8), 1093-1098.

62. Murphy, R. C.; Hankin, J. A.; Barkley, R. M., Imaging of lipid species by MALDI mass spectrometry. *Journal of Lipid Research* **2009,** *50*, S317-S322.

63. Zemski Berry, K. A.; Hankin, J. A.; Barkley, R. M.; Spraggins, J. M.; Caprioli, R. M.; Murphy, R. C., MALDI Imaging of Lipid Biochemistry in Tissues by Mass Spectrometry. *Chemical Reviews* **2011,** *111* (10), 6491-6512.

64. Fujimura, Y.; Miura, D., MALDI Mass Spectrometry Imaging for Visualizing In Situ Metabolism of Endogenous Metabolites and Dietary Phytochemicals. *Metabolites* **2014,** *4* (2), 319-346.

65. Yang, H.; Ji, W.; Guan, M.; Li, S.; Zhang, Y.; Zhao, Z.; Mao, L., Organic washes of tissue sections for comprehensive analysis of small molecule metabolites by MALDI MS imaging of rat brain following status epilepticus. *Metabolomics* **2018,** *14* (4), 50.

66. Bae, Y. J.; Kim, M. S., A Thermal Mechanism of Ion Formation in MALDI. *Annual Review of Analytical Chemistry* **2015,** *8* (1), 41-60.

67. Zenobi, R.; Knochenmuss, R., Ion formation in MALDI mass spectrometry. *Mass Spectrometry Reviews* **1998,** *17* (5), 337-366.

68. Jaskolla, T. W.; Karas, M., Compelling evidence for Lucky Survivor and gas phase protonation: the unified MALDI analyte protonation mechanism. *Journal of the American Society for Mass Spectrometry* **2011,** *22* (6), 976-988.

69. Karas, M.; Glückmann, M.; Schäfer, J., Ionization in matrix-assisted laser desorption/ionization: singly charged molecular ions are the lucky survivors. *Journal of Mass Spectrometry* **2000,** *35* (1), 1-12.

70. Karas, M.; Krüger, R., Ion Formation in MALDI: The Cluster Ionization Mechanism. *Chemical Reviews* **2003,** *103* (2), 427-440.

71. Knochenmuss, R.; Zenobi, R., MALDI Ionization: The Role of In-Plume Processes. *Chemical Reviews* **2003,** *103* (2), 441-452.

72. Dreisewerd, K., The Desorption Process in MALDI. *Chemical Reviews* **2003,** *103* (2), 395-426.

73. Chevrier, M. R.; Cotter, R. J.; Roepstorff, P., A matrix-assisted laser desorption time-of-flight mass spectrometer based on a 600ps, 1.2 mJ nitrogen laser. *Rapid Communications in Mass Spectrometry* **1991,** *5* (12), 611-617.

74. Alves, S.; Fournier, F.; Afonso, C.; Wind, F.; Tabet, J.-C., Gas-Phase Ionization/Desolvation Processes and Their Effect on Protein Charge State Distribution under Matrix-Assisted Laser Desorption/Ionization Conditions. *European Journal of Mass Spectrometry* **2006,** *12* (6), 369-383.

75. Menzel, C.; Dreisewerd, K.; Berkenkamp, S.; Hillenkamp, F., The role of the laser pulse duration in infrared matrix-assisted laser desorption/ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2002,** *13* (8), 975-984.

76. Gimon, M. E.; Preston, L. M.; Solouki, T.; White, M. A.; Russell, D. H., Are proton transfer reactions of excited states involved in UV laser desorption ionization? *Organic Mass Spectrometry* **1992,** *27* (7), 827-830.

77. Bencsura, A.; Navale, V.; Sadeghi, M.; Vertes, A., Matrix–Guest Energy Transfer in Matrix-assisted Laser Desorption. *Rapid Communications in Mass Spectrometry* **1997,** *11* (6), 679-682.

78. Bencsura, A.; Vertes, A., Dynamics of hydrogen bonding and energy transfer in matrix-assisted laser desorption. *Chemical Physics Letters* **1995,** *247* (1), 142-148.

79. Vertes, A.; Irinyi, G.; Gijbels, R., Hydrodynamic model of matrix-assisted laser desorption mass spectrometry. *Analytical Chemistry* **1993,** *65* (17), 2389-2393.

80. Vertes, A.; Balazs, L.; Gijbels, R., Matrix-assisted laser desorption of peptides in transmission geometry. *Rapid Communications in Mass Spectrometry* **1990,** *4* (7), 263-266.

81. Knochenmuss, R., A quantitative model of ultraviolet matrix-assisted laser desorption/ionization. *Journal of Mass Spectrometry* **2002,** *37* (8), 867-877.

82. Hillenkamp, F.; Peter-Katalinic, J., *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications*. Wiley: 2013.

83. Kinsel, G. R.; Knochenmuss, R.; Setz, P.; Land, C. M.; Goh, S.-K.; Archibong, E. F.; Hardesty, J. H.; Marynick, D. S., Ionization energy reductions in small 2,5-dihydroxybenzoic acid–proline clusters. *Journal of Mass Spectrometry* **2002,** *37* (11), 1131-1140.

84. Lin, Q.; Knochenmuss, R., Two-photon ionization thresholds of matrix-assisted laser desorption/ionization matrix clusters. *Rapid Communications in Mass Spectrometry* **2001,** *15* (16), 1422-1426.

85. Mukamel, S.; Abramavicius, D., Many-Body Approaches for Simulating Coherent Nonlinear Spectroscopies of Electronic and Vibrational Excitons. *Chemical Reviews* **2004,** *104* (4), 2073-2098.

86. Knochenmuss, R.; Zhigilei, L. V., Molecular Dynamics Model of Ultraviolet Matrix-Assisted Laser Desorption/Ionization Including Ionization Processes. *The Journal of Physical Chemistry B* **2005,** *109* (48), 22947-22957.

87. Karas, M.; Bachmann, D.; Hillenkamp, F., Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry* **1985,** *57* (14), 2935-2939.

88. Kim, S. K.; Li, S.; Bernstein, E. R., Excited state intermolecular proton transfer in isolated clusters: 1-naphthol/ammonia and water. *The Journal of Chemical Physics* **1991,** *95* (5), 3119-3128.

89. Zhang, W.; Niu, S.; Chait, B. T., Exploring infrared wavelength matrix-assisted laser desorption/ionization of proteins with delayed-extraction time-of-flight mass spectrometry. *Journal of the American Society for Mass Spectrometry* **1998,** *9* (9), 879-884.

90. Chen, X.; Carroll, J. A.; Beavis, R. C., Near-ultraviolet-induced matrix-assisted laser desorption/ionization as a function of wavelength. *Journal of the American Society for Mass Spectrometry* **1998,** *9* (9), 885-891.

91. Liao, P.-C.; Allison, J., Enhanced detection of peptides in matrix-assisted laser desorption/ionization mass spectrometry through the use of charge-localized derivatives. *Journal of Mass Spectrometry* **1995,** *30* (3), 511-512.

92. Krüger, R.; Pfenninger, A.; Fournier, I.; Glückmann, M.; Karas, M., Analyte Incorporation and Ionization in Matrix-Assisted Laser Desorption/Ionization Visualized by pH Indicator Molecular Probes. *Analytical Chemistry* **2001,** *73* (24), 5812-5821.

93. Zhu, Y. F.; Lee, K. L.; Tang, K.; Allman, S. L.; Taranenko, N. I.; Chen, C. H., Revisit of MALDI for small proteins. *Rapid Communications in Mass Spectrometry* **1995,** *9* (13), 1315-1320.

94. Fournier, I.; Brunot, A.; Tabet, J. C.; Bolbach, G., Delayed extraction experiments using a repulsing potential before ion extraction: evidence of non-covalent clusters as ion precursor in UV matrix-assisted laser desorption/ionization. Part II—Dynamic effects with α-cyano-4-hydroxycinnamic acid matrix. *Journal of Mass Spectrometry* **2005,** *40* (1), 50-59.

95. Fournier, I.; Marinach, C.; Tabet, J. C.; Bolbach, G., Irradiation effects in MALDI, ablation, ion production, and surface modifications. Part II: 2,5-dihydroxybenzoic acid monocrystals. *Journal of the American Society for Mass Spectrometry* **2003,** *14* (8), 893-899.

96. Cooks, R. G.; Glish, G. L.; Mc Luckey, S. A.; Kaiser, R. E., Ion trap mass spectrometry. *Chemical and Engineering News; (United States)* **1991**, Medium: X; Size: Pages: 26.

97. Harris, D. C., *Quantitative Chemical Analysis*. 10th ed.; W. H. Freeman: 2020.

98. Barrow, M. P.; Burkitt, W. I.; Derrick, P. J., Principles of Fourier transform ion cyclotron resonance mass spectrometry and its application in structural biology. *Analyst* **2005,** *130* (1), 18-28.

99. Coles, J. N.; Guilhaus, M., Resolution limitations from detector pulse width and jitter in a linear orthogonal-acceleration time-of-flight mass spectrometer. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (8), 772-778.

100. Brown, R. S.; Lennon, J. J., Mass Resolution Improvement by Incorporation of Pulsed Ion Extraction in a Matrix-Assisted Laser Desorption/Ionization Linear Time-of-Flight Mass Spectrometer. *Analytical Chemistry* **1995,** *67* (13), 1998-2003.

101. Mamyrin, B. A.; Karataev, V. I.; Shmikk, D. V.; Zagulin, V. A., The mass-reflectron A new nonmagnetic time-of-flight high resolution mass-spectrometer. *Zhurnal Ehksperimental'noj i Teoreticheskoj Fiziki* **1973,** *64* (1), 82-89.

102. Macfarlane, R. D.; Torgerson, D. F., Californium-252 plasma desorption mass spectroscopy. *Science* **1976,** *191* (4230), 920.

103. Yefchak, G. E.; Schultz, G.; Allison, J.; Enke, C.; Holland, J., Beam deflection for temporal encoding in time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.* **1990,** *1*, 440-7.

104. Dawson, J. H. J.; Guilhaus, M., Orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Communications in Mass Spectrometry* **1989,** *3* (5), 155-159.

105. Guilhaus, M.; Selby, D.; Mlynski, V., Orthogonal acceleration time-of-flight mass spectrometry. *Mass Spectrometry Reviews* **2000,** *19* (2), 65-107.

106. Pringle, S. D.; Giles, K.; Wildgoose, J. L.; Williams, J. P.; Slade, S. E.; Thalassinos, K.; Bateman, R. H.; Bowers, M. T.; Scrivens, J. H., An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *International Journal of Mass Spectrometry* **2007,** *261* (1), 1-12.

107. Ladislas Wiza, J., Microchannel plate detectors. *Nuclear Instruments and Methods* **1979,** *162* (1), 587-601.

108. Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A., An introduction to quadrupole–time-of-flight mass spectrometry. *Journal of Mass Spectrometry* **2001,** *36* (8), 849-865.

109. Knight, R. D., Storage of ions from laser-produced plasmas. *Applied Physics Letters* **1981,** *38* (4), 221-223.

110. Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R., The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* **2005,** *40* (4), 430-443.

111. Gillig, K. J.; Bluhm, B. K.; Russell, D. H., Ion motion in a Fourier transform ion cyclotron resonance wire ion guide cell. *International Journal of Mass Spectrometry and Ion Processes* **1996,** *157-158*, 129-147.

112. Solouki, T.; Gillig, K. J.; Russell, D. H., Detection of High-Mass Biomolecules in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: Theoretical and Experimental Investigations. *Analytical Chemistry* **1994,** *66* (9), 1583-1587.

113. Makarov, A., Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry* **2000,** *72* (6), 1156-1162.

114. Perry, R. H.; Cooks, R. G.; Noll, R. J., Orbitrap mass spectrometry: Instrumentation, ion motion and applications. *Mass Spectrometry Reviews* **2008,** *27* (6), 661-699.

115. Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S., Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *Analytical Chemistry* **2006,** *78* (7), 2113-2120.

116. Senko, M. W.; Canterbury, J. D.; Guan, S.; Marshall, A. G., A High-performance Modular Data System for Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **1996,** *10* (14), 1839-1844.

117. Lange, O.; Damoc, E.; Wieghaus, A.; Makarov, A., Enhanced Fourier transform for Orbitrap mass spectrometry. *International Journal of Mass Spectrometry* **2014,** *369*, 16-22.

118. Forcisi, S.; Moritz, F.; Kanawati, B.; Tziotis, D.; Lehmann, R.; Schmitt-Kopplin, P., Liquid chromatography–mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal of Chromatography A* **2013,** *1292*, 51-65.

119. Scigelova, M.; Makarov, A., Advances in bioanalytical LC–MS using the Orbitrap™ mass analyzer. *Bioanalysis* **2009,** *1* (4), 741-754.

120. Kruve, A.; Rebane, R.; Kipper, K.; Oldekop, M.-L.; Evard, H.; Herodes, K.; Ravio, P.; Leito, I., Tutorial review on validation of liquid chromatography–mass spectrometry methods: Part I. *Analytica Chimica Acta* **2015,** *870*, 29-44.

121. Cajka, T.; Fiehn, O., Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Analytical Chemistry* **2016,** *88* (1), 524-545.

122. Cao, D.; Hao, Z.; Hu, M.; Geng, F.; Rao, Z.; Niu, H.; Shi, Y.; Cai, Y.; Zhou, Y.; Liu, J.; Kang, Y., A feasible strategy to improve confident elemental composition determination of compounds in complex organic mixture such as natural organic matter by FTICR-MS without internal calibration. *Science of The Total Environment* **2021,** *751*, 142255.

123. McNaught, A. D.; Wilkinson, A.; Pure, I. U. o.; Chemistry, A., *IUPAC Compendium of Chemical Terminology*. IUPAC: 2003.

124. Kind, T.; Fiehn, O., Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews* **2010,** *2* (1), 23-60.

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

126. Fievre, A.; Solouki, T.; Marshall, A. G.; Cooper, W. T., High-Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry of Humic and Fulvic Acids by Laser Desorption/Ionization and Electrospray Ionization. *Energy & Fuels* **1997,** *11* (3), 554-560.

127. Höring, M.; Ejsing, C. S.; Krautbauer, S.; Ertl, V. M.; Burkhardt, R.; Liebisch, G., Accurate quantification of lipid species affected by isobaric overlap in Fourier-transform mass spectrometry. *Journal of Lipid Research* **2021,** *62*, 100050.

128. Gorshkov, M. V.; Fornelli, L.; Tsybin, Y. O., Observation of ion coalescence in Orbitrap Fourier transform mass spectrometry. *Rapid Communications in Mass Spectrometry* **2012,** *26* (15), 1711-1717.

129. Wang, M.; Huang, Y.; Han, X., Accurate mass searching of individual lipid species candidates from high-resolution mass spectra for shotgun lipidomics. *Rapid Communications in Mass Spectrometry* **2014,** *28* (20), 2201-2210.

130. Bowman, A. P.; Blakney, G. T.; Hendrickson, C. L.; Ellis, S. R.; Heeren, R. M. A.; Smith, D. F., Ultra-High Mass Resolving Power, Mass Accuracy, and Dynamic Range MALDI Mass Spectrometry Imaging by 21-T FT-ICR MS. *Analytical Chemistry* **2020,** *92* (4), 3133-3142.

131. Eliuk, S.; Makarov, A., Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annual Review of Analytical Chemistry* **2015,** *8* (1), 61-80.

132. Scigelova, M.; Hornshaw, M.; Giannakopulos, A.; Makarov, A., Fourier Transform Mass Spectrometry. *Molecular & Cellular Proteomics* **2011,** *10* (7), M111.009431.

133. Miura, D.; Tsuji, Y.; Takahashi, K.; Wariishi, H.; Saito, K., A Strategy for the Determination of the Elemental Composition by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Based on Isotopic Peak Ratios. *Analytical Chemistry* **2010,** *82* (13), 5887-5891.

134. Nagao, T.; Yukihira, D.; Fujimura, Y.; Saito, K.; Takahashi, K.; Miura, D.; Wariishi, H., Power of isotopic fine structure for unambiguous determination of metabolite elemental compositions: In silico evaluation and metabolomic application. *Analytica Chimica Acta* **2014,** *813*, 70-76.

135. Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B., Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst* **2009,** *134* (7), 1322-1332.

136. Böcker, S.; Letzel, M. C.; Lipták, Z.; Pervukhin, A., SIRIUS: decomposing isotope patterns for metabolite identification†. *Bioinformatics* **2009,** *25* (2), 218-224.

137. Nikolaev, E. N.; Jertz, R.; Grigoryev, A.; Baykut, G., Fine Structure in Isotopic Peak Distributions Measured Using a Dynamically Harmonized Fourier Transform Ion Cyclotron Resonance Cell at 7 T. *Analytical Chemistry* **2012,** *84* (5), 2275-2283.

138. Thompson, C. J.; Witt, M.; Forcisi, S.; Moritz, F.; Kessler, N.; Laukien, F. H.; Schmitt-Kopplin, P., An Enhanced Isotopic Fine Structure Method for Exact Mass Analysis in Discovery Metabolomics: FIA-CASI-FTMS. *Journal of the American Society for Mass Spectrometry* **2020,** *31* (10), 2025-2034.

139. Thurman, E. M.; Ferrer, I., The isotopic mass defect: a tool for limiting molecular formulas by accurate mass. *Analytical and Bioanalytical Chemistry* **2010,** *397* (7), 2807-2816.

140. Xu, Y.; Heilier, J.-F. O.; Madalinski, G.; Genin, E.; Ezan, E.; Tabet, J.-C.; Junot, C., Evaluation of Accurate Mass and Relative Isotopic Abundance Measurements in the LTQ-Orbitrap Mass Spectrometer for Further Metabolomics Database Building. *Analytical Chemistry* **2010,** *82* (13), 5490-5501.

141. Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R., Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **2013,** *9* (S1), 44-66.

142. Xiao, J. F.; Zhou, B.; Ressom, H. W., Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *TrAC Trends in Analytical Chemistry* **2012,** *32*, 1-14.

143. Yan, Z.; Yan, R., Improved Data-Dependent Acquisition for Untargeted Metabolomics Using Gas-Phase Fractionation with Staggered Mass Range. *Analytical Chemistry* **2015,** *87* (5), 2861-2868.

144. Guo, J.; Huan, T., Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Analytical Chemistry* **2020,** *92* (12), 8072-8080.

145. Zhang, C.; Zuo, T.; Wang, X.; Wang, H.; Hu, Y.; Li, Z.; Li, W.; Jia, L.; Qian, Y.; Yang, W.; Yu, H., Integration of Data-Dependent Acquisition (DDA) and Data-Independent High-Definition MSE (HDMSE) for the Comprehensive Profiling and Characterization of Multicomponents from Panax japonicus by UHPLC/IM-QTOF-MS. *Molecules* **2019,** *24* (15), 2708.

146. Jones, O. A. H., Illuminating the dark metabolome to advance the molecular characterisation of biological systems. *Metabolomics* **2018,** *14* (8).

147. Wu, Z.; Rodgers, R. P.; Marshall, A. G., Two- and Three-Dimensional van Krevelen Diagrams: A Graphical Analysis Complementary to the Kendrick Mass Plot for Sorting Elemental Compositions of Complex Organic Mixtures Based on Ultrahigh-Resolution Broadband Fourier Transform Ion Cyclotron Resonanc. *Analytical Chemistry* **2004,** *76* (9), 2511-2516.

148. Rivas-Ubach, A.; Liu, Y.; Bianchi, T. S.; Tolić, N.; Jansson, C.; Paša-Tolić, L., Moving beyond the van Krevelen Diagram: A New Stoichiometric Approach for Compound Classification in Organisms. *Analytical Chemistry* **2018,** *90* (10), 6152-6160.

149. Dey, A., Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies* **2016,** *7* (3), 1174-1179.

150. Kotsiantis, S. B.; Zaharakis, I. D.; Pintelas, P. E., Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review* **2006,** *26* (3), 159-190.

151. Singh, A.; Thakur, N.; Sharma, A. In *A review of supervised machine learning algorithms*, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 16-18 March 2016; 2016; pp 1310-1315.

152. Bishop, C. M.; Bishop, P. N. C. C. M.; Hinton, G.; Press, O. U., *Neural Networks for Pattern Recognition*. Clarendon Press: 1995.

153. Sarah, G.; A., R. J.; Travis, N.; C., H. K.; Angelo, D. A., Characterization of rapid extraction protocols for high-throughput metabolomics. *Rapid Communications in Mass Spectrometry* **2017,** *31* (17), 1445-1452.

154. Travis, N.; C., H. K.; Angelo, D. A., A three-minute method for high-throughput quantitative metabolomics and quantitative tracing experiments of central carbon and nitrogen pathways. *Rapid Communications in Mass Spectrometry* **2017,** *31* (8), 663-673.

155. Guder, J. C.; Schramm, T.; Sander, T.; Link, H., Time-Optimized Isotope Ratio LC–MS/MS for High-Throughput Quantification of Primary Metabolites. *Analytical Chemistry* **2017,** *89* (3), 1624-1631.

156. Lubin, A.; Geerinckx, S.; Bajic, S.; Cabooter, D.; Augustijns, P.; Cuyckens, F.; Vreeken, R. J., Enhanced performance for the analysis of prostaglandins and thromboxanes by liquid chromatography-tandem mass spectrometry using a new atmospheric pressure ionization source. *Journal of Chromatography A* **2016,** *1440*, 260-265.

157. Lubin, A.; De Vries, R.; Cabooter, D.; Augustijns, P.; Cuyckens, F., An atmospheric pressure ionization source using a high voltage target compared to electrospray ionization for the LC/MS analysis of pharmaceutical compounds. *Journal of Pharmaceutical and Biomedical Analysis* **2017,** *142*, 225-231.

158. Ai, W.; Nie, H.; Song, S.; Liu, X.; Bai, Y.; Liu, H., A Versatile Integrated Ambient Ionization Source Platform. *Journal of The American Society for Mass Spectrometry* **2018,** *29* (7), 1408-1415.

159. Zhang, W.; Li, F.; Nie, L., Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology* **2010,** *156* (2), 287-301.

160. Ahn, S.-Y.; Jamshidi, N.; Mo, M. L.; Wu, W.; Eraly, S. A.; Dnyanmote, A.; Bush, K. T.; Gallegos, T. F.; Sweet, D. H.; Palsson, B.; Nigam, S. K., Linkage of organic anion transporter-1 to metabolic pathways through integrated omics-driven network and functional analysis. *Journal of Biological Chemistry* **2011**.

161. Gong, Y.; Cao, R.; Ding, G.; Hong, S.; Zhou, W.; Lu, W.; Damle, M.; Fang, B.; Wang, C. C.; Qian, J.; Lie, N.; Lanzillotta, C.; Rabinowitz, J. D.; Sun, Z., Integrated omics approaches to characterize a nuclear receptor corepressor-associated histone deacetylase in mouse skeletal muscle. *Molecular and Cellular Endocrinology* **2018,** *471*, 22-32.

162. Szulejko, J. E.; Luo, Z.; Solouki, T., Simultaneous determination of analyte concentrations, gas-phase basicities, and proton transfer kinetics using gas chromatography/Fourier transform ion cyclotron resonance mass spectrometry (GC/FT-ICR MS). *International Journal of Mass Spectrometry* **2006,** *257* (1), 16-26.

163. Ruddy, B. M.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G., Positive Ion Electrospray Ionization Suppression in Petroleum and Complex Mixtures. *Energy & Fuels* **2018,** *32* (3), 2901-2907.

164. Cohen, S. L.; Chait, B. T., Influence of Matrix Solution Conditions on the MALDI-MS Analysis of Peptides and Proteins. *Analytical Chemistry* **1996,** *68* (1), 31-37.

165. Fuchs, B.; Schiller, J., Recent Developments of Useful MALDI Matrices for the Mass Spectrometric Characterization of Apolar Compounds. *Current Organic Chemistry* **2009,** *13* (16), 1664-1681.

166. Steven, R. T.; Bunch, J., Repeat MALDI MS imaging of a single tissue section using multiple matrices and tissue washes. *Analytical and Bioanalytical Chemistry* **2013,** *405* (14), 4719-4728.

167. Schwartz, A. J.; Williams, K. L.; Hieftje, G. M.; Shelley, J. T., Atmospheric-pressure solution-cathode glow discharge: A versatile ion source for atomic and molecular mass spectrometry. *Analytica Chimica Acta* **2017,** *950*, 119-128.

168. Gross, J. H., Direct analysis in real time—a critical review on DART-MS. *Analytical and Bioanalytical Chemistry* **2014,** *406* (1), 63-80.

169. Dong, X.; Cheng, J.; Li, J.; Wang, Y., Graphene as a Novel Matrix for the Analysis of Small Molecules by MALDI-TOF MS. *Analytical Chemistry* **2010,** *82* (14), 6208-6214.

170. Cheng, S.-C.; Jhang, S.-S.; Huang, M.-Z.; Shiea, J., Simultaneous Detection of Polar and Nonpolar Compounds by Ambient Mass Spectrometry with a Dual Electrospray and Atmospheric Pressure Chemical Ionization Source. *Analytical Chemistry* **2015,** *87* (3), 1743-1748.

171. Lang, L. M.; Dalsgaard, P. W.; Linnet, K., Quantitative analysis of cortisol and 6β-hydroxycortisol in urine by fully automated SPE and ultra-performance LC coupled with electrospray and atmospheric pressure chemical ionization (ESCi)-TOF-MS. *Journal of Separation Science* **2013,** *36* (2), 246-251.

172. Nyadong, L.; Galhena, A. S.; Fernández, F. M., Desorption Electrospray/Metastable-Induced Ionization: A Flexible Multimode Ambient Ion Generation Technique. *Analytical Chemistry* **2009,** *81* (18), 7788-7794.

173. Vaikkinen, A.; Shrestha, B.; Nazarian, J.; Kostiainen, R.; Vertes, A.; Kauppila, T. J., Simultaneous Detection of Nonpolar and Polar Compounds by Heat-Assisted Laser Ablation Electrospray Ionization Mass Spectrometry. *Analytical Chemistry* **2013,** *85* (1), 177-184.

174. Mirabelli, M. F.; Zenobi, R., Solid-Phase Microextraction Coupled to Capillary Atmospheric Pressure Photoionization-Mass Spectrometry for Direct Analysis of Polar and Nonpolar Compounds. *Analytical Chemistry* **2018,** *90* (8), 5015-5022.

175. Chen, C.; Weng, L.; Chen, K.; Sheu, F.; Lin, C., Symmetric Atmospheric Plasma Source Integrated With Electrospray Ionization for Ambient Mass Spectrometry Detections. *IEEE Transactions on Plasma Science* **2019,** *47* (2), 1114-1120.

176. Nie, W.; Yan, L.; Lee, Y. H.; Guha, C.; Kurland, I. J.; Lu, H., Advanced mass spectrometry-based multi-omics technologies for exploring the pathogenesis of hepatocellular carcinoma. *Mass Spectrometry Reviews* **2016,** *35* (3), 331-349.

177. Bock, C.; Farlik, M.; Sheffield, N. C., Multi-Omics of Single Cells: Strategies and Applications. *Trends in Biotechnology* **2016,** *34* (8), 605-608.

178. Ishii, N.; Tomita, M., Multi-Omics Data-Driven Systems Biology of E. coli. In *Systems Biology and Biotechnology of Escherichia coli*, Lee, S. Y., Ed. Springer Netherlands: Dordrecht, 2009; pp 41-57.

179. Hasin, Y.; Seldin, M.; Lusis, A., Multi-omics approaches to disease. *Genome Biology* **2017,** *18* (1), 83.

180. Siddiqui, G.; Srivastava, A.; Russell, A. S.; Creek, D. J., Multi-omics Based Identification of Specific Biochemical Changes Associated With PfKelch13-Mutant Artemisinin-Resistant Plasmodium falciparum. *The Journal of Infectious Diseases* **2017,** *215* (9), 1435-1444.

181. Kaplan, K.; Jackson, S.; Dwivedi, P.; Davidson, W. S.; Yang, Q.; Tso, P.; Siems, W.; Woods, A.; Hill, H. H., Monitoring dynamic changes in lymph metabolome of fasting and fed rats by matrix-assisted laser desorption/ionization-ion mobility mass spectrometry (MALDI-IMMS). *International Journal for Ion Mobility Spectrometry* **2013,** *16* (3), 177-184.

182. Woods, A. S.; Ugarov, M.; Egan, T.; Koomen, J.; Gillig, K. J.; Fuhrer, K.; Gonin, M.; Schultz, J. A., Lipid/Peptide/Nucleotide Separation with MALDI-Ion Mobility-TOF MS. *Analytical Chemistry* **2004,** *76* (8), 2187-2195.

183. Fenn, L. S.; Kliman, M.; Mahsut, A.; Zhao, S. R.; McLean, J. A., Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples. *Analytical and Bioanalytical Chemistry* **2009,** *394* (1), 235-244.

184. Quach, A.; Lomenick, B.; Yoon, A. J.; Cohn, W.; Whitelegge, J. P.; Faull, K. F. In *Omni-MS: a method for concurrent LC-MS analysis of electrolytes, small molecules, lipids, proteins, nucleic acids, and polysaccharides*, American Society for Mass Spectrometry Conference, , Atlanta, GA, Atlanta, GA, 2019.

185. Solouki, T.; Emmett, M. R.; Guan, S.; Marshall, A. G., Detection, Number, and Sequence Location of Sulfur-Containing Amino Acids and Disulfide Bridges in Peptides by Ultrahigh-Resolution MALDI FTICR Mass Spectrometry. *Analytical Chemistry* **1997,** *69* (6), 1163-1168.

186. McLafferty, F. W., *Interpretation of mass spectra*. 3rd ed.; University Science Books: Mill Valley, California, 1980.

187. Shi, S. D.-H.; Hendrickson, C. L.; Marshall, A. G., Counting individual sulfur atoms in a protein by ultrahighresolution Fourier transform ion cyclotron resonance mass spectrometry: Experimental resolution of isotopic fine structure in proteins. *Proceedings of the National Academy of Sciences* **1998,** *95* (20), 11532-11537.

188. Kim, S.; Rodgers, R. P.; Marshall, A. G., Truly "exact" mass: Elemental composition can be determined uniquely from molecular mass measurement at ~0.1mDa accuracy for molecules up to ~500Da. *International Journal of Mass Spectrometry* **2006,** *251* (2), 260-265.

189. Kujawinski, E. B.; Behn, M. D., Automated Analysis of Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectra of Natural Organic Matter. *Analytical Chemistry* **2006,** *78* (13), 4363-4373.

190. Erve, J. C. L.; Gu, M.; Wang, Y.; DeMaio, W.; Talaat, R. E., Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination. *Journal of the American Society for Mass Spectrometry* **2009,** *20* (11), 2058-2069.

191. Yergey, J. A., A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics* **1983,** *52* (2), 337-349.

192. Bishop, C. M., *Neural Networks for Pattern Recognition*. Oxford University Press, Inc.: 1995; p 482.

193. Rosenblatt, F., The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **1958,** *65* (6), 386-408.

194. Bebis, G.; Georgiopoulos, M., Feed-forward neural networks. *IEEE Potentials* **1994,** *13* (4), 27-31.

195. Giles, C. L.; Maxwell, T., Learning, invariance, and generalization in high-order neural networks. *Appl. Opt.* **1987,** *26* (23), 4972-4978.

196. Balestriero, R., Neural Decision Trees. *arXiv:1702.07360 [cs, stat]* **2017**.

197. Mühlberger, F.; Saraji-Bozorgzad, M.; Gonin, M.; Fuhrer, K.; Zimmermann, R., Compact Ultrafast Orthogonal Acceleration Time-of-Flight Mass Spectrometer for On-Line Gas Analysis by Electron Impact Ionization and Soft Single Photon Ionization Using an Electron Beam Pumped Rare Gas Excimer Lamp as VUV-Light Source. *Analytical Chemistry* **2007,** *79* (21), 8118-8124.

198. Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H., Jr.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S., LMSD: LIPID MAPS structure database. *Nucleic Acids Res* **2007,** *35* (Database issue), D527-D532.

199. Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A., HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research* **2013,** *41* (D1), D801-D807.

200. Liu, H.; Zhang, J.; Sun, H.; Xu, C.; Zhu, Y.; Xie, H., The Prediction of Peptide Charge States for Electrospray Ionization in Mass Spectrometry. *Procedia Environmental Sciences* **2011,** *8*, 483-491.

201. Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H., Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees. *Analytical Chemistry* **2015,** *87* (11), 5738-5744.

202. Brantley, M.; Solouki, T., Rapid and Non-Targeted Detection of Chemical Substructures using Feedforward Neural Networks. **Under Internal Review**.

203. Prechelt, L., Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks* **1998,** *11* (4), 761-767.

204. Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K., Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra. *Analytical Chemistry* **2001,** *73* (19), 4676-4681.

205. Fernández, F. M., TBI Lipidomics Dataset. 11/03/17 ed.; Chorus Project: 2017.

206. Pettit, M. E.; Donnarumma, F.; Murray, K. K.; Solouki, T., Infrared laser ablation sampling coupled with data independent high resolution UPLC-IM-MS/MS for tissue analysis. *Analytica Chimica Acta* **2018,** *1034*, 102-109.

207. Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nature methods* **2009,** *6* (5), 359.

208. Matyash, V.; Liebisch, G.; Kurzchalia, T. V.; Shevchenko, A.; Schwudke, D., Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of lipid research* **2008,** *49* (5), 1137-1146.

209. Paglia, G.; Angel, P.; Williams, J. P.; Richardson, K.; Olivos, H. J.; Thompson, J. W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A.; Plumb, R. S.; Grant, D. F.; Palsson, B. O.; Langridge, J.; Geromanos, S.; Astarita, G., Ion Mobility-Derived Collision Cross Section As an Additional Measure for Lipid Fingerprinting and Identification. *Analytical Chemistry* **2015,** *87* (2), 1137-1144.

210. Senko, M. W.; Beu, S. C.; McLaffertycor, F. W., Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry* **1995,** *6* (4), 229-233.

211. Popov, I. A.; Nagornov, K.; N.Vladimirov, G.; Kostyukevich, Y. I.; Nikolaev, E. N., Twelve Million Resolving Power on 4.7 T Fourier Transform Ion Cyclotron Resonance Instrument with Dynamically Harmonized Cell—Observation of Fine Structure in Peptide Mass Spectra. *Journal of The American Society for Mass Spectrometry* **2014,** *25* (5), 790-799.

212. Stehman, S. V., Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment* **1997,** *62* (1), 77-89.

213. Fahy, E.; Cotter, D.; Sud, M.; Subramaniam, S., Lipid classification, structures and tools. *Biochimica et biophysica acta* **2011,** *1811* (11), 637-647.

214. Szulejko, J.; Hall, S.; Jackson, M.; Solouki, T., Differentiation Between Pure Cultures of Streptococcus pyogenes and Pseudomonas aeruginosa by FT-ICR-MS Volatile Analysis. *The Open Spectroscopy Journal* **2009,** *3*.

215. Verentchikov, A. N.; Ens, W.; Standing, K. G., Reflecting time-of-flight mass spectrometer with an electrospray ion source and orthogonal extraction. *Analytical Chemistry* **1994,** *66* (1), 126-133.

216. Casares, D.; Escribá, P. V.; Rosselló, C. A., Membrane Lipid Composition: Effect on Membrane and Organelle Structure, Function and Compartmentalization and Therapeutic Avenues. *Int J Mol Sci* **2019,** *20* (9), 2167.

217. Orešič, M.; Hänninen, V. A.; Vidal-Puig, A., Lipidomics: a new window to biomedical frontiers. *Trends in Biotechnology* **2008,** *26* (12), 647-652.

218. Fahy, E.; Cotter, D.; Sud, M.; Subramaniam, S., Lipid classification, structures and tools. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2011,** *1811* (11), 637-647.

219. Li, L.; Han, J.; Wang, Z.; Liu, J. a.; Wei, J.; Xiong, S.; Zhao, Z., Mass spectrometry methodology in lipid analysis. *Int J Mol Sci* **2014,** *15* (6), 10492-10507.

220. Lee, H.-C.; Yokomizo, T., Applications of mass spectrometry-based targeted and non-targeted lipidomics. *Biochemical and Biophysical Research Communications* **2018,** *504* (3), 576-581.

221. Brügger, B.; Erben, G.; Sandhoff, R.; Wieland, F. T.; Lehmann, W. D., Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem mass spectrometry. *Proceedings of the National Academy of Sciences* **1997,** *94* (6), 2339-2344.

222. Han, X.; Gross, R. W., Electrospray ionization mass spectroscopic analysis of human erythrocyte plasma membrane phospholipids. *Proceedings of the National Academy of Sciences* **1994,** *91* (22), 10635-10639.

223. Ruotolo, B. T.; Gillig, K. J.; Stone, E. G.; Russell, D. H., Peak capacity of ion mobility mass spectrometry. *Journal of Chromatography B* **2002,** *782* (1-2), 385-392.

224. Cajka, T.; Fiehn, O., Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. *Trends Analyt Chem* **2014,** *61*, 192-206.

225. Shvartsburg, A. A.; Isaac, G.; Leveque, N.; Smith, R. D.; Metz, T. O., Separation and classification of lipids using differential ion mobility spectrometry. *J Am Soc Mass Spectrom* **2011,** *22* (7), 1146-1155.

226. Hutchins, P. D.; Russell, J. D.; Coon, J. J., LipiDex: An Integrated Software Package for High-Confidence Lipid Identification. *Cell Systems* **2018,** *6* (5), 621-625.e5.

227. Berry, K. A. Z.; Barkley, R. M.; Berry, J. J.; Hankin, J. A.; Hoyes, E.; Brown, J. M.; Murphy, R. C., Tandem Mass Spectrometry in Combination with Product Ion Mobility for the Identification of Phospholipids. *Analytical Chemistry* **2017,** *89* (1), 916-921.

228. Kind, T.; Liu, K.-H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O., LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nature Methods* **2013,** *10* (8), 755-758.

229. Koelmel, J. P.; Li, X.; Stow, S. M.; Sartain, M. J.; Murali, A.; Kemperman, R.; Tsugawa, H.; Takahashi, M.; Vasiliou, V.; Bowden, J. A.; Yost, R. A.; Garrett, T. J.; Kitagawa, N., Lipid Annotator: Towards Accurate Annotation in Non-Targeted Liquid Chromatography High-Resolution Tandem Mass Spectrometry (LC-HRMS/MS) Lipidomics Using A Rapid and User-Friendly Software. *Metabolites* **2020,** *10* (3), 101.

230. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016,** *8* (1), 61.

231. van der Hooft, J. J. J.; Wandy, J.; Young, F.; Padmanabhan, S.; Gerasimidis, K.; Burgess, K. E. V.; Barrett, M. P.; Rogers, S., Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Analytical Chemistry* **2017,** *89* (14), 7569-7577.

232. Haslam, R. P.; Feussner, I., Green light for lipid fingerprinting. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2017,** *1862* (8), 782-785.

233. Subramaniam, S.; Fahy, E.; Gupta, S.; Sud, M.; Byrnes, R. W.; Cotter, D.; Dinasarapu, A. R.; Maurya, M. R., Bioinformatics and systems biology of the lipidome. *Chem Rev* **2011,** *111* (10), 6452-6490.

234. May, J. C.; Morris, C. B.; McLean, J. A., Ion Mobility Collision Cross Section Compendium. *Analytical chemistry* **2017,** *89* (2), 1032-1044.

235. Dührkop, K.; Nothias, L. F.; Fleischauer, M.; Ludwig, M.; Hoffmann, M. A.; Rousu, J.; Dorrestein, P. C.; Böcker, S., Classes for the masses: Systematic classification of unknowns using fragmentation spectra. *bioRxiv* **2020**, 2020.04.17.046672.

236. Tetko, I. V.; Livingstone, D. J.; Luik, A. I., Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences* **1995,** *35* (5), 826-833.

237. Wang, X.; Nijman, R.; Camuzeaux, S.; Sands, C.; Jackson, H.; Kaforou, M.; Emonts, M.; Herberg, J. A.; Maconochie, I.; Carrol, E. D.; Paulus, S. C.; Zenz, W.; Van Der Flier, M.; De Groot, R.; Martinon-Torres, F.; Schlapbach, L. J.; Pollard, A. J.; Fink, C.; Kuijpers, T. T.; Anderson, S.; Lewis, M. R.; Levin, M.; McClure, M., Plasma lipid profiles discriminate bacterial from viral infection in febrile children. *Scientific Reports* **2019,** *9* (1).

238. Muhamadali, H.; Weaver, D.; Subaihi, A.; Almasoud, N.; Trivedi, D. K.; Ellis, D. I.; Linton, D.; Goodacre, R., Chicken, beams, and Campylobacter: rapid differentiation of foodborne bacteria via vibrational spectroscopy and MALDI-mass spectrometry. *The Analyst* **2016,** *141* (1), 111-122.

239. Dortet, L.; Potron, A.; Bonnin, R. A.; Plesiat, P.; Naas, T.; Filloux, A.; Larrouy-Maumus, G., Rapid detection of colistin resistance in Acinetobacter baumannii using MALDI-TOF-based lipidomics on intact bacteria. *Scientific Reports* **2018,** *8* (1).

240. Liu, X.; Zhang, M.; Cheng, X.; Liu, X.; Sun, H.; Guo, Z.; Li, J.; Tang, X.; Wang, Z.; Sun, W.; Zhang, Y.; Ji, Z., LC-MS-Based Plasma Metabolomics and Lipidomics Analyses for Differential Diagnosis of Bladder Cancer and Renal Cell Carcinoma. *Frontiers in Oncology* **2020,** *10* (717).

241. Gaiser, R. A.; Pessia, A.; Ateeb, Z.; Davanian, H.; Fernández Moro, C.; Alkharaan, H.; Healy, K.; Ghazi, S.; Arnelo, U.; Valente, R.; Velagapudi, V.; Sällberg Chen, M.; Del Chiaro, M., Integrated targeted metabolomic and lipidomic analysis: A novel approach to classifying early cystic precursors to invasive pancreatic cancer. *Scientific Reports* **2019,** *9* (1).

242. Layre, E.; Moody, D. B., Lipidomic profiling of model organisms and the world's major pathogens. *Biochimie* **2013,** *95* (1), 109-115.

243. Appala, K.; Bimpeh, K.; Freeman, C.; Hines, K. M., Recent applications of mass spectrometry in bacterial lipidomics. *Analytical and Bioanalytical Chemistry* **2020,** *412* (24), 5935-5943.

244. Eiriksson, F. F.; Nøhr, M. K.; Costa, M.; Bödvarsdottir, S. K.; Ögmundsdottir, H. M.; Thorsteinsdottir, M., Lipidomic study of cell lines reveals differences between breast cancer subtypes. *PLoS One* **2020,** *15* (4), e0231289-e0231289.

245. Li, B.; Neumann, E. K.; Ge, J.; Gao, W.; Yang, H.; Li, P.; Sweedler, J. V., Interrogation of spatial metabolome of Ginkgo biloba with high-resolution matrix-assisted laser desorption/ionization and laser desorption/ionization mass spectrometry imaging. *Plant, Cell & Environment* **2018,** *41* (11), 2693-2703.

246. Neumann, E. K.; Migas, L. G.; Allen, J. L.; Caprioli, R. M.; Van de Plas, R.; Spraggins, J. M., Spatial Metabolomics of the Human Kidney using MALDI Trapped Ion Mobility Imaging Mass Spectrometry. *Analytical Chemistry* **2020**.

247. Zhou, D.; Guo, S.; Zhang, M.; Liu, Y.; Chen, T.; Li, Z., Mass spectrometry imaging of small molecules in biological tissues using graphene oxide as a matrix. *Analytica Chimica Acta* **2017,** *962*, 52-59.

248. Neumann, E. K.; Ellis, J. F.; Triplett, A. E.; Rubakhin, S. S.; Sweedler, J. V., Lipid Analysis of 30 000 Individual Rodent Cerebellar Cells Using High-Resolution Mass Spectrometry. *Analytical Chemistry* **2019,** *91* (12), 7871-7878.

249. Neumann, E. K.; Comi, T. J.; Rubakhin, S. S.; Sweedler, J. V., Lipid Heterogeneity between Astrocytes and Neurons Revealed by Single-Cell MALDI-MS Combined with Immunocytochemical Classification. *Angewandte Chemie International Edition* **2019,** *58* (18), 5910-5914.

250. Djambazova, K. V.; Klein, D. R.; Migas, L. G.; Neumann, E. K.; Rivera, E. S.; Van de Plas, R.; Caprioli, R. M.; Spraggins, J. M., Resolving the Complexity of Spatial Lipidomics Using MALDI TIMS Imaging Mass Spectrometry. *Analytical Chemistry* **2020,** *92* (19), 13290-13297.

251. Do, T. D.; Ellis, J. F.; Neumann, E. K.; Comi, T. J.; Tillmaand, E. G.; Lenhart, A. E.; Rubakhin, S. S.; Sweedler, J. V., Optically Guided Single Cell Mass Spectrometry of Rat Dorsal Root Ganglia to Profile Lipids, Peptides and Proteins. *ChemPhysChem* **2018,** *19* (10), 1180-1191.

252. Taban, I. M.; Altelaar, A. F. M.; van der Burgt, Y. E. M.; McDonnell, L. A.; Heeren, R. M. A.; Fuchser, J.; Baykut, G., Imaging of peptides in the rat brain using MALDI-FTICR mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2007,** *18* (1), 145-151.

253. Drake, R. R.; Powers, T. W.; Norris-Caneda, K.; Mehta, A. S.; Angel, P. M., In situ imaging of N-glycans by MALDI imaging mass spectrometry of fresh or formalin-fixed paraffin-embedded tissue. *Current protocols in protein science* **2018,** *94* (1), e68.

254. Gustafsson, O. J. R.; Briggs, M. T.; Condina, M. R.; Winderbaum, L. J.; Pelzing, M.; McColl, S. R.; Everest-Dass, A. V.; Packer, N. H.; Hoffmann, P., MALDI imaging mass spectrometry of N-linked glycans on formalin-fixed paraffin-embedded murine kidney. *Analytical and Bioanalytical Chemistry* **2015,** *407* (8), 2127-2139.

255. McDonnell, L. A.; Corthals, G. L.; Willems, S. M.; van Remoortere, A.; van Zeijl, R. J.; Deelder, A. M., Peptide and protein imaging mass spectrometry in cancer research. *Journal of Proteomics* **2010,** *73* (10), 1921-1944.

256. Goodwin, R. J. A.; Pennington, S. R.; Pitt, A. R., Protein and peptides in pictures: Imaging with MALDI mass spectrometry. *PROTEOMICS* **2008,** *8* (18), 3785-3800.

257. Yèagle, P. L., Lipid regulation of cell membrane structure and function. *The FASEB Journal* **1989,** *3* (7), 1833-1842.

258. Dennis, E. A., Introduction to Thematic Review Series: Phospholipases: Central Role in Lipid Signaling and Disease. *Journal of Lipid Research* **2015,** *56* (7), 1245-1247.

259. Walther, T. C.; Jr., R. V. F., Lipid Droplets and Cellular Lipid Metabolism. *Annual Review of Biochemistry* **2012,** *81* (1), 687-714.

260. Spraggins, J. M.; Djambazova, K. V.; Rivera, E. S.; Migas, L. G.; Neumann, E. K.; Fuetterer, A.; Suetering, J.; Goedecke, N.; Ly, A.; Van de Plas, R.; Caprioli, R. M., High-Performance Molecular Imaging with MALDI Trapped Ion-Mobility Time-of-Flight (timsTOF) Mass Spectrometry. *Analytical Chemistry* **2019,** *91* (22), 14552-14560.

261. Eriksson, C.; Masaki, N.; Yao, I.; Hayasaka, T.; Setou, M., MALDI Imaging Mass Spectrometry—A Mini Review of Methods and Recent Developments. *Mass Spectrometry* **2013,** *2* (Special_Issue), S0022-S0022.

262. Korte, A. R.; Lee, Y. J., MALDI-MS analysis and imaging of small molecule metabolites with 1,5-diaminonaphthalene (DAN). *Journal of Mass Spectrometry* **2014,** *49* (8), 737-741.

263. Kompauer, M.; Heiles, S.; Spengler, B., Atmospheric pressure MALDI mass spectrometry imaging of tissues and cells at 1.4-μm lateral resolution. *Nature Methods* **2017,** *14* (1), 90-96.

264. Jeanne Dit Fouque, K.; Ramirez, C. E.; Lewis, R. L.; Koelmel, J. P.; Garrett, T. J.; Yost, R. A.; Fernandez-Lima, F., Effective Liquid Chromatography–Trapped Ion Mobility Spectrometry–Mass Spectrometry Separation of Isomeric Lipid Species. *Analytical Chemistry* **2019,** *91* (8), 5021-5027.

265. Tang, F.; Guo, C.; Ma, X.; Zhang, J.; Su, Y.; Tian, R.; Shi, R.; Xia, Y.; Wang, X.; Ouyang, Z., Rapid In Situ Profiling of Lipid C═C Location Isomers in Tissue Using Ambient Mass Spectrometry with Photochemical Reactions. *Analytical Chemistry* **2018,** *90* (9), 5612-5619.

266. Zhang, W.; Shang, B.; Ouyang, Z.; Xia, Y., Enhanced Phospholipid Isomer Analysis by Online Photochemical Derivatization and RPLC-MS. *Analytical Chemistry* **2020,** *92* (9), 6719-6726.

267. Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S., LIPID MAPS online tools for lipid research. *Nucleic Acids Research* **2007,** *35* (suppl_2), W606-W612.

268. Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H., Jr; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S., LMSD: LIPID MAPS structure database. *Nucleic Acids Research* **2006,** *35* (suppl_1), D527-D532.

269. Hodson, L.; Skeaff, C. M.; Fielding, B. A., Fatty acid composition of adipose tissue and blood in humans and its use as a biomarker of dietary intake. *Progress in Lipid Research* **2008,** *47* (5), 348-380.

270. Khaw, K.-T.; Friesen, M. D.; Riboli, E.; Luben, R.; Wareham, N., Plasma Phospholipid Fatty Acid Concentration and Incident Coronary Heart Disease in Men and Women: The EPIC-Norfolk Prospective Study. *PLOS Medicine* **2012,** *9* (7), e1001255.

271. Keller, B. O.; Li, L., Discerning matrix-cluster peaks in matrix-assisted laser desorption/ionization time-of-flight mass spectra of dilute peptide mixtures. *Journal of the American Society for Mass Spectrometry* **2000,** *11* (1), 88-93.

272. Leopold, J.; Popkova, Y.; Engel, K. M.; Schiller, J., Recent Developments of Useful MALDI Matrices for the Mass Spectrometric Characterization of Lipids. *Biomolecules* **2018,** *8* (4), 173.

273. Mamyrin, B. A., Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry* **2001,** *206* (3), 251-266.

274. Řezanka, T.; Sigler, K., Odd-numbered very-long-chain fatty acids from the microbial, animal and plant kingdoms. *Progress in Lipid Research* **2009,** *48* (3), 206-238.

275. Berry, K. A. Z.; Hankin, J. A.; Barkley, R. M.; Spraggins, J. M.; Caprioli, R. M.; Murphy, R. C., MALDI imaging of lipid biochemistry in tissues by mass spectrometry. *Chem Rev* **2011,** *111* (10), 6491-6512.

276. Klerk, L. A.; Broersen, A.; Fletcher, I. W.; Van Liere, R.; Heeren, R. M. A., Extended data analysis strategies for high resolution imaging MS: New methods to deal with extremely large image hyperspectral datasets. *International Journal of Mass Spectrometry* **2007,** *260* (2-3), 222-236.

277. Alexandrov, T., MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics* **2012,** *13* (S16).

278. Buchberger, A. R.; Delaney, K.; Johnson, J.; Li, L., Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Analytical Chemistry* **2018,** *90* (1), 240-265.

279. Calvano, C. D.; Monopoli, A.; Cataldi, T. R. I.; Palmisano, F., MALDI matrices for low molecular weight compounds: an endless story? *Analytical and Bioanalytical Chemistry* **2018,** *410* (17), 4015-4038.

280. Wahl, P.; Ducasa, G. M.; Fornoni, A., Systemic and renal lipids in kidney disease development and progression. *American Journal of Physiology-Renal Physiology* **2016,** *310* (6), F433-F445.

281. Harrison-Bernard, L. M., Sphingolipids, new kids on the block, promoting glomerular fibrosis in the diabetic kidney. *Am J Physiol Renal Physiol* **2015,** *309* (8), F685-F686.

282. Miyamoto, S.; Hsu, C.-C.; Hamm, G.; Darshi, M.; Diamond-Stanic, M.; Declèves, A.-E.; Slater, L.; Pennathur, S.; Stauber, J.; Dorrestein, P. C.; Sharma, K., Mass Spectrometry Imaging Reveals Elevated Glomerular ATP/AMP in Diabetes/obesity and Identifies Sphingomyelin as a Possible Mediator. *EBioMedicine* **2016,** *7*, 121-134.

283. Ruh, H.; Salonikios, T.; Fuchser, J.; Schwartz, M.; Sticht, C.; Hochheim, C.; Wirnitzer, B.; Gretz, N.; Hopf, C., MALDI imaging MS reveals candidate lipid markers of polycystic kidney disease[S]. *Journal of Lipid Research* **2013,** *54* (10), 2785-2794.

284. Abbas, I.; Noun, M.; Touboul, D.; Sahali, D.; Brunelle, A.; Ollero, M., Kidney Lipidomics by Mass Spectrometry Imaging: A Focus on the Glomerulus. *International Journal of Molecular Sciences* **2019,** *20* (7), 1623.

285. Muller, L.; Kailas, A.; Jackson, S. N.; Roux, A.; Barbacci, D. C.; Albert Schultz, J.; Balaban, C. D.; Woods, A. S., Lipid imaging within the normal rat kidney using silver nanoparticles by matrix-assisted laser desorption/ionization mass spectrometry. *Kidney International* **2015,** *88* (1), 186-192.

286. Zhou, W.; Bovik, A. C., A universal image quality index. *IEEE Signal Processing Letters* **2002,** *9* (3), 81-84.