# Bioinformatics for quantitative, targeted lipidomics discovery

Justin Gareth Chitpin

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the M.Sc. degree in Biochemistry, with specialization in Bioinformatics

Department of Biochemistry, Microbiology, and Immunology
Faculty of Medicine
University of Ottawa

*I may not have gone where I intended to go, but I think I have ended up where I needed to be.*

—Douglas Adams, *The Long Dark Tea-Time of the Soul*

# Acknowledgements

I would like to thank my supervisors Drs. Steffany Bennett and Theodore Perkins for supporting me throughout my M.Sc studies. I never imagined, coming out of undergrad, that my project would lead me into the world of mass spectrometry and metabolism. Yet in hindsight, I couldn't be more thankful to be gifted a project where I could meet so many people and learn in such an awesome environment. I am very lucky to have both of you as my supervisors, and look forward to our continued adventures during my Ph.D. This thesis would not be possible without the present and past members of the Bennett lab. I couldn't ask for a more dynamic group of humorous, wise, and dedicated colleagues to work alongside. In particular, I would like to thank Graeme Taylor, Thao Nguyen, and Hongbin Xu for teaching me about the unlimited possibilities of mass spectrometry along the way. I acknowledge my sources of funding from the Ottawa Hospital Research Institute to the MATRIX CREATE training program which gave me the opportunity to attend academic and industry workshops around the country and internationally. I would also like to thank my thesis advisory committee members Drs. Lavallée-Adam and St-Pierre for their helpful comments, and Dr. Aris-Brosou for agreeing to evaluate my thesis. Finally, I would like to thank my mother, for I wouldn't be where I am without her.

# Abstract

Lipidomics uses high performance liquid chromatography coupled to tandem mass spectrometry via electrospray ionization to comprehensively identify and quantify all lipid species in a biological system. Divided into untargeted and targeted approaches, untargeted lipidomics is viewed as a discovery-driven method to possibly identify all lipids, while targeted lipidomics is considered an approach limited to quantifying known lipids. While there are many bioinformatics tools to facilitate lipid identification in untargeted lipidomics datasets, there are few tools to annotate lipid identities in targeted datasets. Thus, I present two novel bioinformatics tools to annotate and identify lipids from targeted lipidomics data. The first tool annotates lipid isotopes and in-source artifacts caused by electrospray ionization from targeted lipidomics. Using this tool, I highlight problems associated with the manual curation approach conventionally used to pick peaks from targeted lipidomics data. In Bayesian Annotations for Targeted Lipidomics (BATL), I present the first principled and accurate classifier for the identification of targeted lipidomics data. This program eliminates human subjectivity associated with peak identification. Combining these bioinformatics tools together, I conclude by presenting a fully unbiased, discovery-driven, targeted lipidomics pipeline, enabling the identification of all lipid isomers detected from targeted lipidomics data.

# List of abbreviations

| | |
|---|---|
| A | Single character code for relative area |
| AD | Alzheimer's disease |
| BATL | Bayesian annotations for targeted lipidomics |
| C8 | 8 carbon long hydrocarbon chain |
| C18 | 16 carbon long hydrocarbon chain |
| Cer() | Ceramide LIPID MAPS nomenclature |
| CID | Collision-induced dissociation |
| Constrained MAP | Constrained *Maximum a posteriori* |
| DDA | Data-dependent acquisition |
| DIA | Data-independent acquisition |
| DLB | Dementia with Lewy bodies |
| E | Trans unsaturation |
| EMS | Enhanced mass scan |
| EPI | Enhanced product ion scan |
| ESI | Electrospray ionization |
| F | Single character code for full width at half max |
| FA | Fatty acyl LIPID MAPS category |
| FDR | False discovery rate |
| GL | Glycerolipid LIPID MAPS category |
| Glc | Glucose |
| GlcCer() | Glucosylceramide LIPID MAPS nomenclature |
| GP | Glycerophospholipid LIPID MAPS category |
| GP01 | Glycerophosphocholine LIPID MAPS class |
| $H^+$ | Proton |
| H | Single character code for relative height |
| IDA | Information dependent acquisition |
| KS | Kolmogorov-Smirnov |
| L | Single character code for tailing factor |
| LC | Liquid chromatography |
| LIT | Linear ion trap |
| LPC() | Lysophosphocholine LIPID MAPS subclass |
| m/z | Mass to charge ratio |

| | |
|---|---|
| M | Molecular mass of an analyte |
| MAP | *Maximum a posteriori* |
| MCI | Mild cognitive impairment |
| MRM | Multiple reaction monitoring |
| MS | Mass spectrometry |
| MS-1 | Full scan |
| MS/MS | Tandem mass spectrometry |
| $MS^3$ | Multistage mass spectrometry to the third |
| MWBM | Maximum weighted bipartite matching |
| N | Single character code for retention time |
| Q/T | Qualifier/target |
| NL | Neutral loss |
| NP | Normal-phase |
| PD | Parkinson's disease |
| PDD | Parkinson's disease with dementia |
| PI | Potentially identified (peak) |
| PK | Polyketide LIPID MAPS category |
| PMI | *Post-mortem* interval |
| ppm | Parts per million |
| PR | Prenol LIPID MAPS category |
| R | Single character code for relative retention time |
| RT | Retention time |
| S | Single character code for subtracted retention time |
| SL | Saccharolipid LIPID MAPS category |
| *sn* | Stereospecific number |
| SP | Sphingolipid LIPID MAPS category |
| SP01 | Sphingoid base LIPID MAPS class |
| SP02 | Ceramide LIPID MAPS class |
| ST | Sterol LIPID MAPS category |
| Q1 | Quadrupole 1 (of a triple quadrupole) and/or parent ion m/z |
| Q1/Q3 | Parent and product ion transition |
| q2 | Quadrupole 2 (collision cell of a triple quadrupole) |

| | |
|---|---|
| Q3 | Quadrupole 3 (of a triple quadrupole) and/or product ion m/z |
| RP | Reverse-phase |
| SRM | Selected reaction monitoring |
| ToF | Time-of-flight |
| XIC | Extracted ion chromatogram |
| Y | Single character code for asymmetry factor |

# List of tables

1.1 Category definitions by LIPID MAPS Consortium . . . . . 5

2.1 Annotation depth per identified lipid artifact/isotope. . . . . 37

3.1 SRM and MRM peak features for naïve Bayes model. . . . . 48

4.1 Annotation depth per identified lipid artifact/isotope. . . . . 87

# List of figures


1.1 Lipid categories defined by the LIPID MAPS Consortium . . . . 3
1.2 General structure of sphingolipids . . . . . 4
1.3 General structure of glycerophospholipids . . . . . 4
1.4 Targeted lipidomics data acquisition modes . . . . . 17

2.1 Lipid artifacts and isotopes identified from the sphingolipid and glycerophospholipid datasets using the in-source lipid artifact and isotope annotation program. . . . . 38
2.2 Case study of a +2 isotope reported by the annotation tool . . . 39

3.1 Decision rules for peak assignments . . . . . 55
3.2 Identification errors using different decision rules . . . . . 61
3.3 MWBM is the best performing decision rule . . . . . 61
3.4 Subtracted retention time is the most discriminatory feature . . 63
3.5 Subtracted retention time accounts for systematic retention time shifts . . . . . 64
3.6 Relative retention time can induce systematic retention time shifts 66
3.7 Additional features significantly improve classifier identification accuracies . . . . . 66
3.8 Increasing model complexity improves cross validation identification accuracies by decreasing the proportion of unassigned peaks . . . . . 67
3.9 Best feature combinations identified from cross validation are validated by holdout datasets . . . . . 68
3.10 Unassignment rates are are similar between holdout and cross validation . . . . . 68
3.11 Minimum number of training data required for classifier . . . . 69
3.12 Classification of statistically independent testing datasets . . . . 73
3.13 Expanding model complexity correlates with increased unassignment rates . . . . . 74

3.14 Increasing model complexity improves classifier identification accuracies by correctly unassigning more peaks . . . . . 75
3.15 Identification accuracies with all eight features display equal or greater identification accuracies on cross validation . . . . . 76
3.16 Identification accuracies with all eight features features display equal or greater identification accuracies on testing datasets . . 77

4.1 Discovery dataset lipid artifacts and isotopes . . . . . 87
4.2 Discovery-versus-original dataset identification accuracies . . . . 92
4.3 Discovery dataset unassignment rates . . . . . 93
4.4 Discovery dataset unassignment accuracies . . . . . 93

5.1 BATL research contributions to the targeted lipidomics analysis pipeline . . . . . 103

# List of supplementary tables


A1 Description of sphingolipid training dataset . . . . . 117
A2 Description of glycerophospholipid training dataset . . . . . 118
A3 Description of sphingolipid blood testing dataset . . . . . 118
A4 Description of sphingolipid brain testing dataset . . . . . 119
A5 Description of glycerophospholipid blood testing dataset . . . . 119
A6 Description of glycerophospholipid brain testing dataset . . . . 119
A7 Complexity of sphingolipid training dataset . . . . . 120
A8 Complexity of sphingolipid holdout dataset . . . . . 120
A9 Complexity of sphingolipid blood testing dataset . . . . . 120
A10 Complexity of sphingolipid brain testing dataset . . . . . 121
A11 Complexity of glycerophospholipid training dataset . . . . . 121
A12 Complexity of glycerophospholipid holdout dataset . . . . . 121
A13 Complexity of glycerophospholipid blood testing dataset . . . . 122
A14 Complexity of glycerophospholipid brain testing dataset . . . . 122

B15 Sphingolipid and glycerophospholipid datasets. . . . . 124

D16 Sphingolipid blood discovery dataset. . . . . 136
D17 Sphingolipid brain discovery dataset. . . . . 137
D18 Glycerophospholipid blood discovery dataset. . . . . 138
D18 Glycerophospholipid blood discovery dataset, cont'd. . . . . 139
D19 Glycerophospholipid brain discovery dataset. . . . . 140
D19 Glycerophospholipid brain discovery dataset, cont'd. . . . . 141

# List of supplementary figures

C1 MWBM is the best performing decision rule . . . . . 126
C2 Retention time and subtracted retention time are the most . . . 127
C3 Additional features significantly improve classifier identification accuracies . . . . . 128
C4 Best feature combinations identified from cross validation are validated by holdout datasets . . . . . 129
C5 Increasing model complexity improves cross validation identification accuracies by decreasing the proportion of unassigned peaks . . . . . 130
C6 Classification of statistically independent testing datasets . . . . 131
C7 Expanding model complexity does not correlate with increased unassignment rates . . . . . 132
C8 Increasing model complexity improves classifier identification accuracies by correctly unassigning more peaks . . . . . 133
C9 Classifiers trained with all eight features display equal or greater identification accuracies on testing datasets . . . . . 134

D10 Discovery dataset relative areas of previously identified-versus-novel peaks . . . . . 143

# Table of contents


Acknowledgements . . . . . ii
Abstract . . . . . iii
List of abbreviations . . . . . iv
List of tables . . . . . vii
List of figures . . . . . viii
List of supplementary tables . . . . . x
List of supplementary figures . . . . . xi

**1 Introduction** **1**
1.1 Exploring lipid biology . . . . . 1
1.2 Lipid structure and nomenclature . . . . . 2
1.3 Technologies for high-throughput lipidomics . . . . . 5
1.3.1 Separation methods . . . . . 6
1.3.2 Ionization methods . . . . . 7
1.3.3 Mass analyzers . . . . . 8
1.3.4 Summary of mass spectrometry-based platforms for lipidomics . . . . . 9
1.4 Mass spectrometry data acquisition methods for lipidomics . 10
1.4.1 Untargeted lipidomics . . . . . 10
1.4.2 Targeted lipidomics . . . . . 12
1.5 Bioinformatics for lipidomics . . . . . 15
1.5.1 Overview of lipidome bioinformatics . . . . . 15
1.5.2 Spectral database matching algorithms . . . . . 18
1.5.3 *In silico* generation of tandem mass spectrometry spectra 19

1.5.4 LC-ESI-MS/MS peak detection and in-source ion annotation algorithms . . . 21
1.5.5 Retention time alignment algorithms . . . 23
1.5.6 Targeted lipid identification algorithms . . . 25
1.6 Addressing the paucity of targeted lipid identification algorithms 27
1.7 Hypothesis and objectives . . . 30

**2 In-source lipid artifact and isotope annotation tool 31**
2.1 Objective . . . 31
2.2 Statement of author contributions . . . 31
2.3 Summary . . . 32
2.4 Introduction . . . 32
2.5 Methods . . . 34
2.5.1 Algorithm . . . 34
2.5.2 Data . . . 35
2.6 Results . . . 36
2.7 Discussion . . . 37
2.8 Conclusion . . . 41

**3 Bayesian annotations for targeted lipidomics (BATL) 42**
3.1 Objective . . . 42
3.2 Statement of author contributions . . . 43
3.3 Summary . . . 43
3.4 Introduction . . . 44
3.5 Methods . . . 48
3.5.1 Algorithm . . . 48
3.5.2 Data . . . 57
3.6 Results . . . 58
3.6.1 Metrics . . . 58
3.6.2 Decision rule and feature selection . . . 58
3.6.3 Holdout validation of classifier . . . 62
3.6.4 Predictive performance of classifier . . . 69

3.6.5 Finalizing the trained classifier . . . . . 71
3.7 Discussion . . . . . 78
3.8 Conclusion . . . . . 80

**4 Discovery-driven targeted lipidomics 81**
4.1 Objective . . . . . 81
4.2 Statement of author contributions . . . . . 82
4.3 Summary . . . . . 82
4.4 Introduction . . . . . 83
4.5 Methods . . . . . 85
4.5.1 Data . . . . . 85
4.5.2 Data analysis pipeline using BATL . . . . . 85
4.6 Results . . . . . 86
4.6.1 Annotating lipid artifacts and inferring lipid families 86
4.6.2 Metrics . . . . . 88
4.6.3 Discovery-driven lipid identification . . . . . 90
4.7 Discussion . . . . . 91
4.8 Conclusion . . . . . 94

**5 Discussion 96**
5.1 In-source lipid artifact and isotope annotation . . . . . 97
5.2 Bayesian annotations for targeted lipidomics . . . . . 99
5.3 Discovery-driven targeted lipidomics . . . . . 101
5.4 Advancing targeted lipid identification and discovery . . . . 102

**References 104**

**Appendix: SRM/MRM data descriptions 117**

**Appendix: Feature normality 123**

**Appendix: Matrix-specific classifier results 125**

**Appendix: Discovery datasets 135**

# Chapter One

# Introduction

## 1.1 Exploring lipid biology

Lipids encompass a diverse group of hydrophobic and amphipathic molecules making up the intracellular and extracellular membranes of every known organism. Their functional roles are numerous and lipids are best known for self-assembling into water-tight bilayers and efficiently storing metabolic energy (Harayama and Riezman 2018; Marrink and Berendsen 1994; W. V. Allen 1976). Yet, lipids are more than structural and storage molecules that compartmentalize biochemical reactions and supply metabolic precursors within organelles. Technological advancements of the 21st century are enabling, for the first time, the comprehensive cataloguing and quantitation of hundreds to thousands of lipid species in a biological system (Han, Yang, and Gross 2012). This interdisciplinary field of lipidomics lies at the crossroads of lipid biochemistry, analytical chemistry, and computational biology. With emerging methods to study lipids with increasing granularity, the once simplistic view of lipids is being supplanted by discoveries of novel lipid species and lipid functions (Brown and R. C. Murphy 2009). How lipid composition alters cellular function has only begun to be explored (Shevchenko and Simons 2010). With structural, metabolic, and signalling roles, lipids represent an exciting avenue for understanding cell

regulation, disease, and potential therapeutics. Lipids are docking sites for proteins (Li et al. 2020; Ziegler et al. 2002; Hedger and Sansom 2016), can post-translationally modify proteins (Jiang et al. 2018), act as ligands that activate cellular receptors (Wymann and Schneiter 2008; Fernandis and Wenk 2007), and are themselves signalling second messengers (Whitehead et al. 2007; A. H. Merrill 1992), to name only a few roles. Their biological significance is underlined by the fact that hundreds of proteins are responsible for synthesizing and trafficking lipid species within mammalian cells. Understanding lipid localization, regulation, and organization is leading to breakthroughs in neurodegenerative diseases (Alecu and Bennett 2019; Bennett et al. 2013), cancer (Hajicek et al. 2019; Vasan et al. 2019), metabolic syndrome (Olzmann and Carvalho 2019; Back et al. 2019), and immunometabolic disorders (Hotamisligil 2017). Despite these breakthroughs, the most fundamental question that remains unsolved in lipidomics is determining the total number of biologically-occurring lipid species. This exact figure remains unknown, yet cataloguing all naturally-occurring lipid species has the potential to revolutionize biomedical science.

## 1.2 Lipid structure and nomenclature

Over twenty thousand unique lipid species have been empirically observed and curated, owing to their complex structural variations (Fahy, Subramaniam, Brown, et al. 2005; Sud et al. 2007). Unlike genes or proteins, lipids are synthesized from a variety of molecular precursors. These building blocks are combinatorially linked via multiple types of chemical bonds to produce distinct lipid species at the molecular level. As such, and unlike nucleic acids or proteins, lipids possess no sequential structure or motifs, and are generally classified by their physical structure instead of biological functions or sequence homologies (Fahy, Subramaniam, Brown, et al. 2005).

According to the LIPID MAPS Consortium (Fahy, Subramaniam, Brown, et al. 2005; Schmelzer et al. 2007; Fahy, Subramaniam, R. C. Murphy, et al.

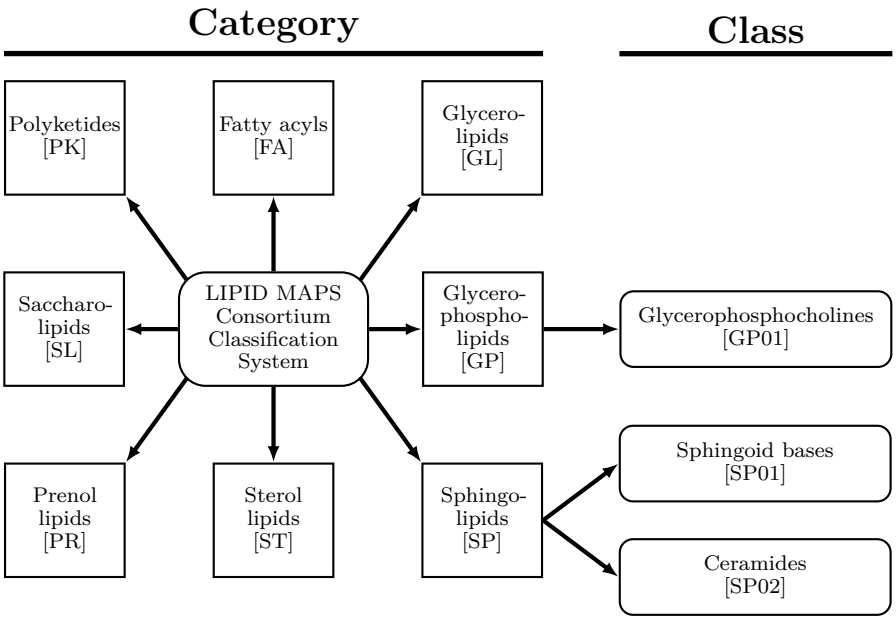



**Figure 1.1: Lipid categories defined by the LIPID MAPS Consortium.** Square brackets denote the classification acronyms for each category and class. Example classes are shown and do not reflect all sphingolipid or glycerophospholipid classes.

2009), lipids are classified into eight categories based on key structural characteristics (Figure 1.1 and Table 1.1). These categories are further divided into classes and subclasses based on the number and type of chemical modifications to these defining structural elements. For example, the presence of a sphingoid base defines the sphingolipid (SP) category. Subsequent modifications to the sphingoid base and the types of chemical bonds linking these molecules together characterize the many different SP classes and subclasses, which are also collectively referred to as families (R. C. Murphy and Axelsen 2011). Because of this structure-based classification system, lipid nomenclatures describe the structure of each lipid in detail.

The lipid nomenclature within each of the eight categories and their associated classes begin with an alphabetical shorthand which denotes the overarching structure of the lipid subclass. An example is "GlcCer()" which is a subclass of the neutral glycosphingolipid class in the sphingolipid cat-

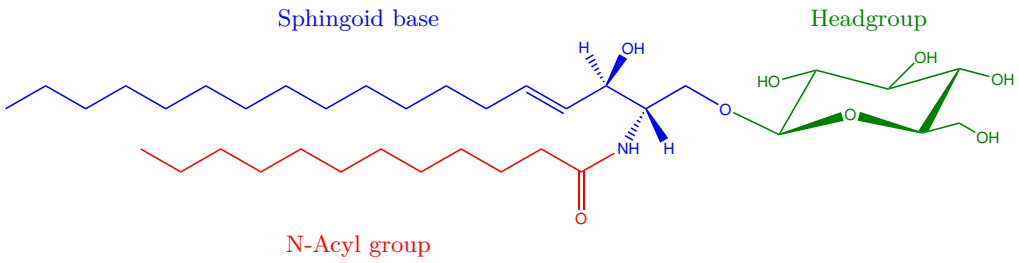



**Figure 1.2: General structure of sphingolipids.** Sphingolipids are defined by the sphingoid base, N-acyl group, and headgroup. The structure shown is GlcCer(d18:1(4E)/12:0).

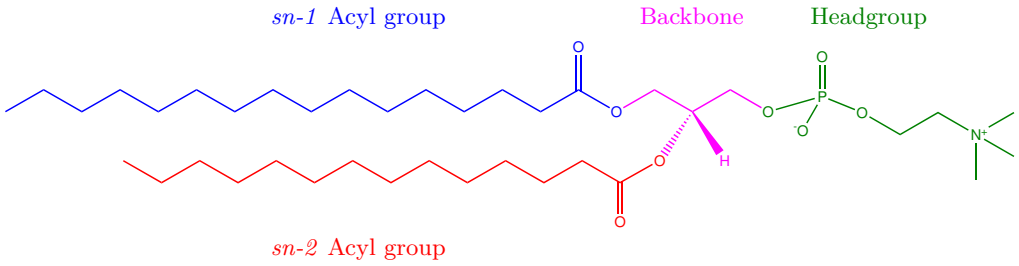



**Figure 1.3: General structure of glycerophospholipids.** Glycerophospholipids are defined by a backbone attached to an *sn-1* acyl group, *sn-2* acyl group, and a headgroup. The structure shown is PC(16:0/14:0).

egory. Glc indicates the presence of a glucose headgroup on a ceramide (Cer). Further structural insight includes knowledge of the total number of carbons and double bonds on both the ceramide backbone and the fatty N-acyl chain. This structural resolution is known as the positional isomer level (Koelmel et al. 2017). Thus, GlcCer(d30:1) refers to a glycosylated ceramide species with 30 carbons in both its respective sphingoid base and fatty acyl chain, and one degree of unsaturation either in the sphingoid base or the N-acyl chain. If the sphingoid base was 18 carbons long with 1 degrees of unsaturation (a sphingosine), the lipid name would be updated to GlcCer(d18:1/12:0) (Figure 1.2). An example in another category would be a glycerophospholipid wherein PC(16:0/14:0) is a member of the glycerophosphocholine class, denoted by "PC()". Similar to the example

**Table 1.1: Category definitions by LIPID MAPS Consortium.**

| Category | Defining structural characteristic |
|---|---|
| Fatty acyls [FA] | Repeating series of methylene groups |
| Glycerolipids [GL] | Glycerol backbone |
| Glycerophospholipids [GP] | Glycerol backbone and phosphate head |
| Sphingolipids [SP] | Sphingoid base backbone |
| Sterol lipids [ST] | Fused four-carbon rings and derivatives |
| Prenol lipids [PR] | Isopentenyl disphosphate and dimethylallyl diphosphate |
| Saccharolipids [SL] | Fatty acid linked directly to sugar backbone |
| Polyketides [PK] | Acetyl and propionyl subunits |

sphingolipid, the 16:0 and 14:0 refers to the hydrocarbon chain length and degrees of unsaturation of the fatty acyls linked at the *sn-1* and *sn-2* position, respectively (Figure 1.3).

To readdress how many biologically-occurring lipid species there exist, one must first identify each lipid present in biological life at the structural level. By enumerating all combinations of lipid building blocks, programs such as LipidBlast (Kind, Liu, et al. 2013) predict approximately 119,200 lipid species, although the structural identification is limited to the total number of carbons and double bonds for many of the complex glycolipids. Other tools such as VaLID (McDowell et al. 2014), which enumerate all possible double bond positions, predict up to 736,584 theoretical glycerophospholipid species alone. As lipid identities can also extend to account for *cis*/*trans* isomerism (Figure 1.2) and stereochemistry (Koelmel et al. 2017), this further expands the number of theoretical lipid species.

## 1.3 Technologies for high-throughput lipidomics

The following sections provide a general overview of mass spectrometry-based technologies employed in lipidomics.

### 1.3.1 Separation methods

Lipid extracts are complex mixtures containing hundreds to thousands of lipid analytes extracted from a given matrix (i.e. biological sample). Their abundances may span several orders of magnitude depending on the biological matrix profiled, treatment condition, or disease state (Quehenberger and Dennis 2011). This sample complexity is a challenge for high-throughput lipid identification and quantitation because lipid isomers and isobars are difficult to resolve by mass spectrometry methods alone. Lipid isobars share the same nominal mass and cannot be identified by low-resolution mass spectrometry. Conversely, lipid isomers share exact masses and cannot be resolved through either low or high-resolution mass spectrometry (discussed further in Section 1.3.3). Lipid extracts may also contain non-lipid molecules derived from the biological tissues, or matrices, during the lipid extraction process. These "matrix-specific effects" may competitively suppress or enhance lipid ionization, although the exact mechanisms causing this phenomenon remain unknown (Panuwet et al. 2016). To overcome these obstacles, lipid extracts are often separated via chromatography to improve lipid isomer identification and reduce matrix-specific effects (R. C. Murphy and Axelsen 2011).

The most common separation method in lipidomics is high performance liquid chromatography (LC). LC encompasses several techniques to separate analytes in a liquid solvent based on the desired chemical properties. Under high pressure, lipid extracts are mixed in a solvent system and eluted through a thin glass capillary, or analytical column, filled with chemically derivated silicon beads. The choice of solvent(s) and/or chemically-derivated beads defines the type of LC system. In lipidomics, reverse-phase liquid chromatography (RP-LC) and normal-phase (NP-LC) are used to separate complex mixtures of lipid species, including positional isomers belonging to the same family (Cajka and Fiehn 2016). Only RP-LC will be discussed here as data obtained from RP-LC were used to generate the bioinformatics solutions described in this thesis. In RP-LC, lipid extracts are eluted through

a reverse-phase gradient with an initially polar solvent that switches in composition to a non-polar solvent. The column is packed with a stationary phase of medium to long chain hydrocarbons derivated to the silicon beads. C8 and C18 hydrocarbon chains are typically used to profile sphingolipids and glycerophospholipids, respectively (Shaner et al. 2009; Whitehead et al. 2007). Using this mobile phase solvent system, lipids are separated in the column based on chemical hydrophobicity. The most hydrophilic lipids elute first as the hydrophobic species preferentially interact with the hydrocarbon chains packed in the column. As the solvent gradient becomes increasingly non-polar, the hydrophobic lipids detach from the silicon beads, eluting in order of increasing hydrophobicity. An optimized LC system thus produces clearly resolvable "peaks" generally consisting of a single lipid species. The elution time at the peak apex, where the maximal number of lipid analytes are exiting the column, is referred to as the "retention time." This information is highly informative for lipid identification and discussed further in Sections 1.5.4 and 1.5.6.

### 1.3.2 Ionization methods

After LC separation, lipid analytes must undergo ionization to be detected by the mass analyzer. This process is complex because the charged lipid analytes must also exist in a gas phase to fly through the components of the mass spectrometer. Several ionization methods have been applied to ionize lipids with mixed success (Blanksby and Mitchell 2010). Currently, electrospray ionization (ESI) remains the most common method in lipidomics to effectively ionize lipid species dissolved in a liquid solvent with minimal, premature fragmentation (Ho et al. 2003).

ESI uses high voltages to convert and ionize lipid species from a liquid to a gas phase. During this process, an electric field is applied to lipid extracts as they are pumped through a narrow capillary tip towards an opposing mass spectrometer inlet. This electric field is generated by a potential difference between the capillary and the inlet. When operated in positive ion

mode, the capillary functions as an anode or terminal where electrons are drawn from the lipid extract. Positively charged ions then repel against the capillary, concentrate, and accumulate at the tip of the capillary. As ions accumulate, Coulombe repulsion increases the charge density. When the drop reaches Rayleigh's limit, it breaks forming a cone, known as Taylor's cone, before being dispersed into a fine mist of positively charged droplets that radiate away from the tip of the capillary. The positive ions contained in these droplets are generally $H^+$ species from the acidified solvent system, although other metal cations present in the solvent may contribute to the positively charged droplets (Konermann et al. 2013). These droplets are then desolvated by a heated, inert gas, and the protonated, gaseous lipid species are directed into the inlet of the mass spectrometer where their mass over charge ratios (m/z) are detected (Ho et al. 2003).

### 1.3.3 Mass analyzers

Once ionized, charged analytes passing through the inlet of the mass spectrometer are scanned by the mass analyzer based on their m/z. Several types of mass analyzers exist and are characterized by their mass resolving power, mass accuracy, sensitivity, acquisition speed, and other functionalities (Kind, Tsugawa, et al. 2018). Mass resolving power and mass accuracy are related metrics that determine how well the mass analyzer can measure a known entity (Balogh 2004). The mass resolving power is defined as the ratio between the peak mass and the peak width required to separate that mass from a neighbouring one. For clarity, these "peaks" are not generated from LC separation, but represent a distribution of detected masses centered at the most intense scanned mass. Relatedly, mass accuracy measures how well the mass spectrometrist can calibrate their mass analyzer to measure a known entity. This metric is defined as the ratio between the measured mass error and the true mass, and is often expressed in parts per million (ppm). Thus, a mass analyzer, scanning a given lipid analyte, is not detecting a single mass but a distribution of masses that are centered on the true analyte mass.

With calibration, this distribution should be narrow such that analytes with different masses are not scanned together as one peak. The variance of this distribution is inversely proportional to the resolving power of the mass analyzer, while the mass accuracy is the error between the empirically measured and theoretical lipid mass. The other two metrics, sensitivity and acquisition speed, measure how effectively and rapidly the mass analyzer can scan incoming ions. A sensitive mass analyzer can measure extremely low concentrations of ions, and ones with high acquisition speeds can scan for more ions per second. No mass analyzer can excel at all four attributes at once, and the strengths and weaknesses of several common mass analyzers are discussed below.

There are three overarching types of mass analyzers. The quadrupole mass analyzer is commonly found in many types of mass spectrometers. While their mass resolving power and accuracy are low, these mass analyzers can select for ions with given m/z and fragment analytes via collision-induced dissociation (CID). Furthermore, some quadrupoles include a linear ion trap (LIT), which improves sensitivity by temporarily trapping ions with selected masses in the quadrupole before scanning them. The trapping of ions also enables information-dependent data acquisition (IDA) modes, which are discussed in Section 1.4.2. Time-of-flight (ToF) and orbital ion trap mass analyzers are found in high-resolution mass spectrometers. Compared to a quadrupole or LIT, ToF mass analyzers and orbital ion traps achieve superior mass accuracy and resolving power, but with no fragmentation functionality. Orbital ion traps achieve greater mass accuracies than ToF analyzers, although their tandem MS acquisition speeds are much lower and they also cannot select for ions (Kind, Tsugawa, et al. 2018).

### 1.3.4 Summary of mass spectrometry-based platforms for lipidomics

A mass spectrometry platform is the coupling of one or more mass analyzers via some ionization method and optional front-end LC separation. In

lipidomics, the simplest mass spectrometry platforms contain a single mass analyzer with no front end separation. These direct-infusion platforms contrast with chromatography-based mass spectrometry platforms which utilize LC to separate lipid extracts prior to data acquisition. More sophisticated tandem mass spectrometry platforms (MS/MS) are obtained by coupling multiple mass analyzers together. The combination of two or more mass analyzers enables unique data acquisition methods where ions are repeatedly scanned, selected, and/or fragmented. These data acquisition methods are discussed in Sections 1.4.1 and 1.4.2.

The majority of modern LC-ESI-MS/MS platforms for lipidomics consist of RP-LC, NP-LC, and/or variants coupled to a triple quadrupole mass spectrometer via ESI. A triple quadrupole is a series of three quadrupole mass analyzers, where the last quadrupole is sometimes replaced with a LIT, ToF, or orbital ion trap mass analyzer. The naming schemes of these tandem mass spectrometry configurations vary across manufacturers, but generally reflect the hyphenation of the mass analyzers present in the system (e.g. QqQ, QqTRAP, QqToF). The choice of mass spectrometry-based platform dictates what lipidomic data acquisition methods are feasible, and these methods are divided into untargeted and targeted approaches.

## 1.4 Mass spectrometry data acquisition methods for lipidomics

### 1.4.1 Untargeted lipidomics

Untargeted lipidomics is one of two major approaches used when conducting lipidomic studies. In literature, it is often described as an unbiased, discovery-oriented workflow to potentially detect and identify all lipid analytes in a sample, although many other definitions widely used in the literature are also discussed in Section 1.6. An unambiguous definition of untargeted lipidomics is the application of untargeted mass spectrometry

data acquisition modes to identify lipid analytes, providing semi-quantitative data where lipid isomers are quantified together and not individually. Admittedly, this definition becomes tenuous when describing workflows that target certain lipid classes during sample extraction and identify these species using untargeted mass spectrometry approaches. Nevertheless, untargeted lipidomics can be explained through the many untargeted mass spectrometry data acquisition modes.

The most common untargeted data acquisition method is full scan, or MS-1 scan, where all lipid analytes in a sample are detected by the mass analyzer. This acquisition method relies on high-resolution ToF and orbital ion trap-based mass spectrometers to identify lipid analytes based on the closest approximation of their exact mass. To resolve lipid isomers, untargeted tandem mass spectrometry data acquisition methods are used to consecutively scan parent lipid ions, fragment them, and scan the resulting product ions. This acquisition method is referred to as tandem MS or MS/MS because data is acquired from two or more mass analyzers instead of one. The first two mass analyzers are often quadrupoles to enable ion scanning, selection, and fragmentation. The first quadrupole (Q1) scans/selects for parent ions which are fragmented in the second quadrupole (q2). The q2 is written with a lower case "q" because this quadrupole functions solely as a collision cell with no capacity to scan or select for ions. The resulting product ions produced from q2 are scanned in the last quadrupole, LIT, ToF, or orbital ion trap. For simplicity, the last mass analyzer is abbreviated as Q3. Using these mass spectrometry platforms, tandem MS data acquisition modes are further subdivided into "data-dependent acquisition" (DDA) and "data-independent acquisition" (DIA). In DDA, only the most abundant parent ions scanned in Q1 are subjected to fragmentation in q2 and detection in Q3. DIA addresses the high lipid abundance bias in DDA by fragmenting all parent ions in Q1 within sequential m/z windows to obtain tandem MS spectra from both high and low abundance lipid species. A challenge of DIA, however, is reconstructing the link between observed tandem MS

spectra and the original parent ions detected prior to fragmentation. Using both DDA and DIA, the acquired tandem spectra can be matched to an empirical or *in silico* spectral database of lipid species for identification. Nevertheless, unambiguous identification of many lipid isomers, which share similar structure and therefore tandem MS spectra, is impossible. Another limitation of DDA and DIA is that lipid quantitation is still based on the parent ions scanned in Q1 because of the difficulty in quantitatively linking spectral data to the individual lipid isomers. Thus, untargeted lipidomics trades low abundance lipid isomer identification and quantitation for broader, albeit semi-quantitative, identification of lipids at the levels of total carbon number and degrees of unsaturation.

### 1.4.2 Targeted lipidomics

Targeted lipidomics represents the second major lipidomic approach used to quantifying select subclasses, or panels, of known lipid species. In contrast to untargeted lipidomics, targeted lipidomics is described in literature as a biased approach to quantify previously identified lipids (described in Section 1.6). In keeping with the proposed definition of untargeted lipidomics, targeted lipidomics can be defined as the application of targeted mass spectrometry data acquisition methods to identify lipid analytes, providing relative or absolute quantitation depending on the experimental design. Thus, targeted lipidomics can also be explained through the various targeted mass spectrometry data acquisition modes.

Nearly all targeted data acquisition modes employ a tandem mass spectrometer, usually with a triple quadrupole or triple quadrupole with the third quadrupole containing a LIT coupled to RP-LC-ESI, NP-LC-ESI, or chromatographic variants. Several acquisition methods are possible depending on whether quadrupoles Q1 and Q3 scan or select for precursor/product ions (Brodbelt 1997). A precursor ion scan involves scanning lipid analytes in Q1, fragmenting them in q2, and selecting for product ions with a specific m/z in Q3 (Figure 1.4a). As lipid subclasses are defined by a shared

chemical group, monitoring a specific product ion enables the detection of class-specific lipid species (Ekroos et al. 2002; Alfred H. Merrill et al. 2005). A product ion scan is the inverse of a precursor ion scan where lipid analytes are selected with given m/z in Q1, fragmented in q2, and scanned in Q3 (Figure 1.4b). This scan mode commonly arises when IDA modes are discussed below. A neutral loss scan detects parent ions that fragment by a loss of a neutrally charged fragment, regardless of the parent ion m/z (Figure 1.4c) (Brugger et al. 1997). This mode is thus useful for detecting diacylglycerols and triacylglycerols, as these lipids tend to produce neutrally charged fatty acyl groups (R. C. Murphy, James, et al. 2007). The most "targeted" data acquisition mode is selected reaction monitoring (SRM), where both precursor and product ions are selected with specific m/z in Q1 and Q3 (Figure 1.4e). A variant of SRM is multiple reaction monitoring (MRM), where multiple transitions are monitored per lipid species for improved lipid identification. In literature, the terms SRM and MRM are sometimes used interchangeably; monitoring multiple SRM transitions is sometimes referred to as an MRM method even if a single pair of parent/product ions is monitored per lipid species (Spickett and Pitt 2015; Djafi et al. 2013). Here, the term MRM will exclusively refer to targeted data acquisition modes where more than one transition is monitored per lipid species.

Using targeted lipidomic approaches, many lipid isomers can be resolved and quantified separately based on a unique product ion monitored in Q3. However, both parent and product ions of the lipid of interest must be known to monitor the correct transition corresponding to the single lipid target. One example is Cer(d14:1/22:1) and Cer(d16:1/20:1), which are ceramide isomers detected at the same protonated parent m/z of 564.5. To identify these lipid species using SRM, the mass of the d14:1 and d16:1 product ion backbones can be monitored in Q3, corresponding to 208.3 and 236.3 daltons, respectively, thereby quantifying both isomers separately. Some positional isomers are unidentifiable using mass spectrometry means alone, such as glycerophosphocholines PC(18:0/0:0) and PC(O-16:0/2:0).

These isomers have the same molecular formula with a protonated parent m/z of 523.3. When fragmented in q2, the phosphocholine head group common to both species is fragmented and no unique product ion is obtained capable of distinguishing between species. To resolve these positional isomers, targeted data acquisition modes are coupled to RP-LC, and the characteristic retention times of each isomer are used to discriminate between lipids with identical diagnostic fragments.

The benefit of SRM, and by extension MRM, mode is the capacity for relative and absolute lipid quantitation based on the product ion selected in Q3. Because orbital ion traps cannot select for ions, Q3 denotes either a quadrupole, LIT, or ToF. To account for lipid extraction, ionization and fragmentation efficiencies, internal standards, representing pure lipid species, are spiked into the samples at the time of lipid extraction. Internal standards sharing the same product ion as the detected SRM peaks are assumed to fragment and ionize similarly; relative quantitation is therefore achieved by normalizing the areas of SRM peaks to the corresponding internal standard sharing the same product ion. A limitation of relative normalization is that normalized abundances can only be compared between the same lipid species across conditions, as ionization and fragmentation efficiencies may differ between species. To overcome this problem, SRM peaks may also be absolutely quantified by generating a standard dilution curve using a synthetic lipid species to model the combined fragmentation and ionization efficiencies of the corresponding endogenous lipid (Whitehead et al. 2007).

Advanced targeted data acquisition modes are available if the quadrupole in Q3 is also a LIT. LITs enable the transient storage of ions in Q3, which is useful when detecting extremely low abundance lipids. For example, when running an enhanced mass scan (EMS), lipid analytes are only scanned in Q3, with no filtering or collision-induced dissociation in Q1 and q2, respectively. The Q3 LIT temporarily stores ions in the quadrupole before scanning them to increase detection sensitivity. The number of ions stored over a given duration are parameters set by the operator and can be tuned to

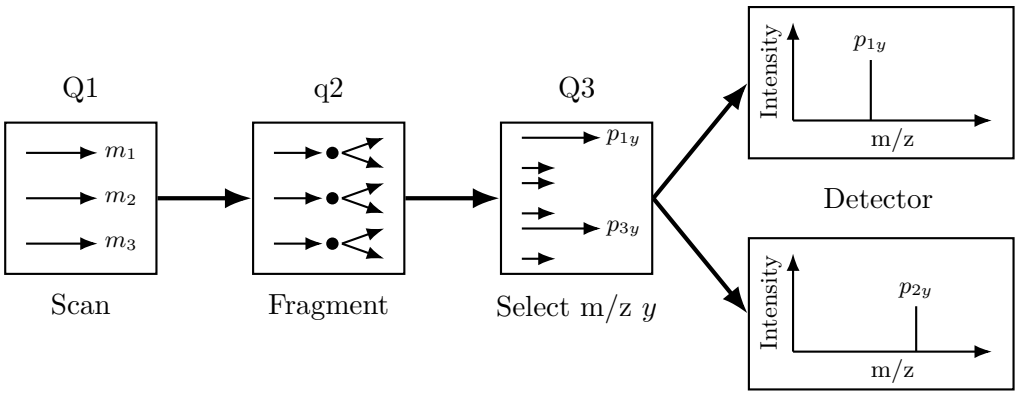
improve the detection of low abundance lipid species. The LIT also enables information-dependent acquisition (IDA), which share some similarities with the DDA methods used in untargeted lipidomics. IDA is a targeted mode that triggers when a set quantity of ions of a given m/z corralled in the LIT reaches a preset threshold from a precursor, product, or SRM "survey scan". The term "survey scan" denotes a data acquisition mode that triggers IDA when a user-defined ion threshold is met. LITs are also capable of fragmenting the stored ions. Because these traps are commonly used as the Q3 mass analyzer, the IDA fragmentation within these traps are referred to as enhanced product ion (EPI) spectra. Figure 1.4f demonstrates an example SRM survey scan coupled to an IDA-EPI acquisition. Once a critical threshold of SRM product ions are acquired, the trap releases all product ions and other ions passing from q2, producing targeted MS/MS/MS (also known as $MS^3$) data, which can provide enough structural information to unambiguously identify a lipid species.

## 1.5 Bioinformatics for lipidomics

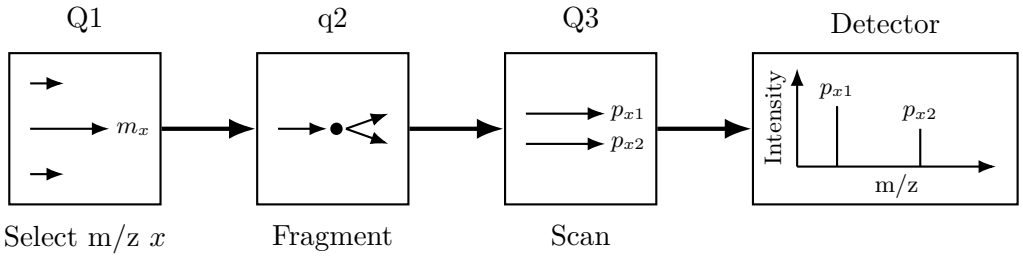
### 1.5.1 Overview of lipidome bioinformatics

The two main goals of lipidome bioinformatics are i) the accurate identification and quantitation of lipid species from different mass spectrometry technologies and data acquisition modes, and ii) the development of computational methods to analyze and interpret lipidomics data. A major challenge of lipid identification is accommodating the diversity of mass spectrometry platforms and data acquisition modes. This is especially important as untargeted and targeted lipidomics are complementary approaches, and their data acquisition modes offer strengths and weaknesses when identifying and quantifying lipid species. However, their methods of lipid identification are different. For untargeted lipidomics, numerous algorithms have been proposed to address the lipid identification problem for full scan, DDA, and DIA. Conversely, very few algorithms have been developed for targeted lipidomics

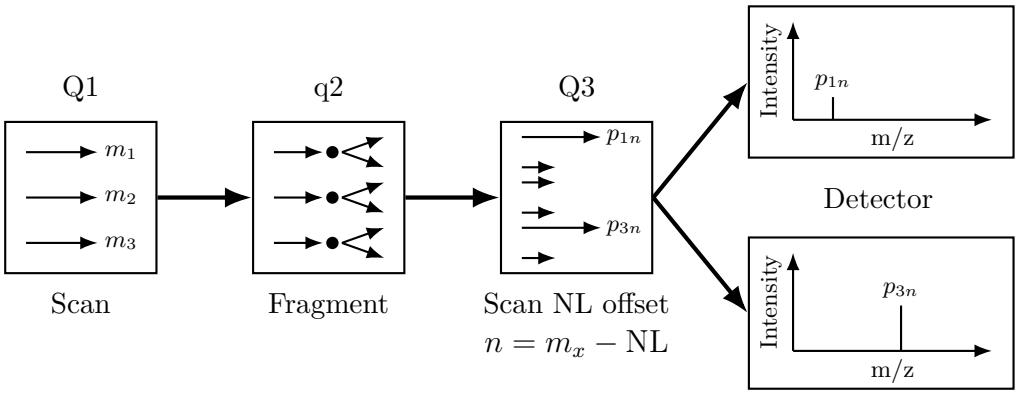
**(a)** Precursor ion scan



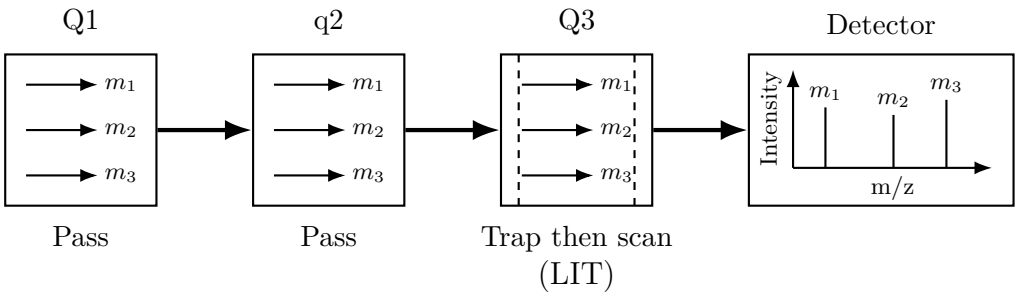

**(b)** Product ion scan



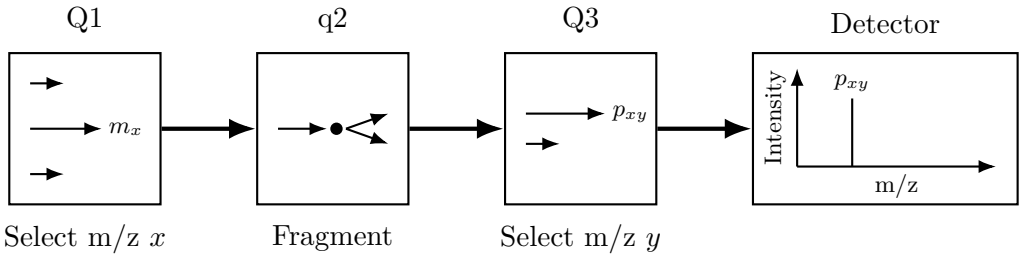

**(c)** Neutral loss scan




**(d)** Enhanced mass scan



**(e)** Selected reaction monitoring




**(f)** Selected reaction monitoring-information dependent acquisition-enhanced product ion scan

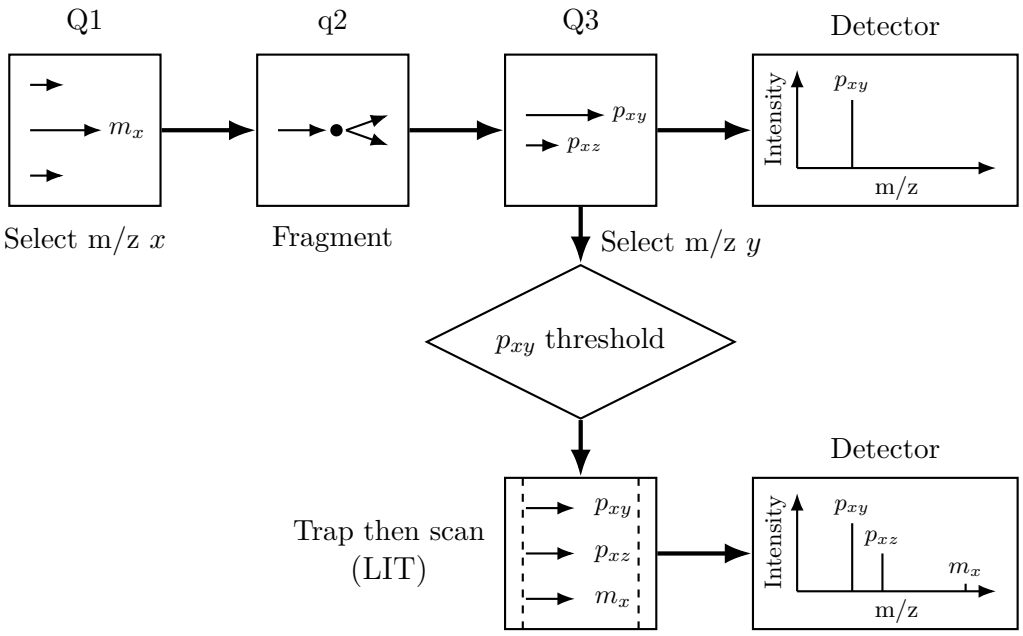



**Figure 1.4: Targeted lipidomics data acquisition modes.** $m_x$ denotes parent ion $x$; $p_{xy}$ and $p_{xz}$ denote product ions $y$ and $z$ fragmented from parent ion $x$; $p_{xn}$ denotes the neutral loss product ion $n$ fragmented from parent ion $x$. Short, unlabelled arrows in Q1 and Q3 are not scanned/selected during data acquisition. In this schematic, the detector is depicted measuring the quantity of ions passing from Q3. (a) Precursor ion scan. (b) Product ion scan. (c) Neutral loss scan. (d) Enhanced mass scan (requires a LIT). (e) Selected reaction monitoring. (f) Selected reaction monitoring-information dependent acquisition-enhanced product ion scan (requires a LIT).

data. The lack of targeted lipid identification algorithms is attributed to the incorrect assumption that an observed signal within a given transition and retention time window corresponds to a single lipid target (discussed in Section 1.5.5). Beyond the problem of identification, both lipidomic workflows are amenable to the same data analysis and interpretation tools. These include univariate and multivariate methods for differential lipidome analysis (MetaboAnalyst 4.0, Chong et al. (2018)), ontology enrichment analysis (ClassyFire, Djoumbou Feunang et al. (2016); ChemRICH, Barupal and Fiehn (2017)), and pathway analysis (BioPAN, Nguyen et al. (2017)), to name a few programs. These data analysis tools emphasize the potential of lipidomics in basic sciences and clinical applications (L. Zhang, Han, and Wang 2018). Yet, until methods for lipid identification are available and accurate across all data acquisition modes, any downstream data analysis may be prone to incorrect biological conclusions. The following sections review developments in lipid identification, highlighting the urgent need for targeted lipid identification tools.

### 1.5.2 Spectral database matching algorithms

The earliest lipidomic experiments employed direct-infusion mass spectrometry using both untargeted and targeted data acquisition modes (Ekroos et al. 2002; Han and Gross 2003). Lipids were manually identified based on matching MS and tandem MS spectra from full scan, DDA, and precursor ion scan modes. Subsequent direct-infusion experiments led to the creation of empirical, lipid spectra databases and spectral matching lipid identification programs (e.g. LipidBank, Watanabe, Yasugi, and Oshima (2000); LipidProfiler (now commercially branded LipidView), Houjou et al. (2005); FAAT, Leavell and Leary (2006); LIMSA, Haimi et al. (2006); LipidInspector, Schwudke et al. (2007); LipidQA, Song et al. (2007); LipidHome, Foster et al. (2013)). These programs annotated input MS and tandem MS spectra by matching them to a user-defined or internal database of known lipid species and computing spectral similarity scores to determine

the most appropriate lipid annotation. These scores were computed based on the spectral intensity similarities between input and reference spectra with matching m/z.

Without prior separation by liquid chromatography, MS and tandem MS spectral annotations from these programs were limited in structural identification to the level of total carbons and degrees of unsaturation. More importantly, the coverage and quality of assigned annotations were strongly dependent on the size of these spectral databases, type of ESI-MS/MS platform, and experimental settings used to acquire them. For example, tandem MS spectral intensities are influenced by the tissue matrix and mass spectrometer fragmentation parameters such as collision energy. The utility of these programs and their databases were potentially limited to data acquired from specific matrices and mass spectrometry platforms. These limitations thus motivated the simulation of tandem MS spectra for all theoretical lipid species.

### 1.5.3 *In silico* generation of tandem mass spectrometry spectra

In 2013, Kind, Liu, et al. developed LipidBlast, the first effort to simulate collision-induced dissociation, ESI-MS/MS spectra for 26 different lipid classes (Kind, Liu, et al. 2013). This innovative program exhaustively generated all theoretical lipid spectra based on lipid fragmentation patterns reported in literature and heuristic rules derived from experimentally observed spectra from over 500 lipid standards across different ESI-MS/MS platforms. LipidBlast addressed the main deficiency with empirical spectral databases at the time; spectra could now be generated where synthetic lipid standards did not exist or were too costly to synthesize. This program led to a surge of new *in silico* ESI-MS/MS small molecule prediction algorithms based on more advanced bond dissociation rules (Hill and Mortishire-Smith 2005), heuristic rules (e.g. MetFusion, Gerlich and Neumann (2013); MetFrag, Wolf et al. (2010) and Ruttkies et al. (2016); LipidFrag, Witting et al.

(2017); MS-DIAL, Tsugawa, Cajka, et al. (2015); Greazy, Kochen et al. (2016)), and machine learning approaches (CFM-ID, F. Allen, Pon, et al. (2014), F. Allen, Greiner, and Wishart (2015), and Djoumbou-Feunang et al. (2019)). Programs such as LipidXplorer (Herzog et al. 2011) and LDA (Hartler, Trötzmüller, et al. 2011; Hartler, Triebl, et al. 2017) avoided simulating spectral databases altogether by implementing lipid subclass-specific fragmentation rules to directly decipher unknown spectra "on the fly." In addition to simulating tandem MS spectra, these programs also explored different spectral database matching algorithms to better predict the most probable lipid identity for each unknown spectrum. Some scoring algorithms favourably weighed specific diagnostic ions that are known indicators of certain lipid subclasses (Kind, Liu, et al. 2013; Ruttkies et al. 2016). Other algorithms incorporated analyte citation counts to favourably score small molecules that have been reported in literature (Ruttkies et al. 2016; Djoumbou-Feunang et al. 2019; Blaženović, Kind, Torbašinović, et al. 2017), although this weight does bias these programs against discovering novel lipid species.

The *in silico* generation of tandem MS spectra has greatly increased the number of identified lipid species from untargeted lipidomics data. Data acquired from full scan, DDA, and DIA modes can be subjected to database-dependent or independent spectral matching algorithms. Some programs such as CFM-ID can also simulate spectra at various collision energies to accommodate data generated from ESI-MS/MS platforms with variable fragmentation parameters. Nevertheless, these approaches for lipid identification remain imperfect. Simulating tandem MS spectra is computationally challenging with processing times up to 150 minutes per spectrum, even with parallel computing (Ruttkies et al. 2016). Matrix-specific effects may alter the spectral intensities of detected lipid species, leading to misidentifications with simulated tandem MS databases. Finally, while tandem MS spectra can be simulated for all theoretical lipid species, untargeted lipidomic approaches cannot exploit this resource to identify low

abundance lipid isomers. This is the major weakness of untargeted lipidomics as low abundance lipid species may never trigger DDA and their DIA tandem spectra may be dominated by high abundance isomers. To overcome these problems, RP-LC was implemented by untargeted and targeted lipidomic approaches to separate lipid isomers, and NP-LC to separate classes, before ESI-MS/MS data acquisition.

### 1.5.4 LC-ESI-MS/MS peak detection and in-source ion annotation algorithms

The coupling of LC to ESI-MS/MS provides several benefits over direct-infusion approaches. LC separates lipid extracts into chromatographic peaks, reducing ion suppression and isobaric/isomeric overlap from spectra acquired from untargeted or targeted lipidomic approaches (Hu, Duan, and Han 2019). Integrating peak areas also enables the quantitation of peaks corresponding to individual lipid species using targeted lipidomic approaches or mixtures of co-eluting lipid isomers if using untargeted data acquisition modes. Accompanying these benefits, however, is the computational problem of "picking" chromatographic peaks for downstream identification and quantitation. Many peak detection algorithms have been developed to pick chromatographic peaks from LC-ESI-MS/MS data (e.g. MZmine, Katajamaa, Miettinen, and Orei (2006) and Pluskal et al. (2010); XCMS, C. A. Smith et al. (2006) and Benton et al. (2008); MRMDIFF, Tsugawa, Ohta, et al. (2014)), which are subsequently identified from their MS or tandem MS spectra. However, the number of picked peaks is generally an overrepresentation of the true number of unique lipid species. A single lipid species can be detected as multiple peaks with the same retention time across different m/z values. These additional peaks may reflect isotopes, adducts, or in-source artifacts. These additional peaks complicate downstream lipid identification because they can be confused for other monoisotopic lipid species based on similar m/z, tandem MS spectra, and/or retention times. Removing these peaks simplifies lipid identification and potentially reduces the number

of misidentified lipid species. Several programs have been developed to annotate and remove these redundant peaks (e.g. CAMERA, Kuhl et al. (2012); CliqueMS, Senan et al. (2019); LipidFinder, O'Connor et al. (2017); MS-FLO, Defelice et al. (2017)).

The fundamental principle of all isotope, adduct, and artifact annotation programs is that lipid monoisotopes, isotopes, adducts, and artifacts elute at identical retention times but with different m/z. Lipid isotopes and monoisotopes share the same hydrophobicity, while any lipid adducts or artifacts generated during ESI occur after RP-LC separation. Current programs exploit these principles by generating correlation matrices based on chromatographic similarities of detected peaks within or across samples. The most common approaches calculate pairwise Pearson correlation coefficients across peak intensities binned over time; chromatographically similar peaks, representing artifacts or isotopes, will yield greater Pearson similarities (Kuhl et al. 2012; Defelice et al. 2017). As peak shapes may not change linearly in intensity, cosine similarities have been shown outperform Pearson correlations (Senan et al. 2019). User-defined similarity thresholds are applied to identify candidate peak pairs representing possible isotopes, adducts, and artifacts. Annotations are then assigned to these peaks by computing the pairwise difference in m/z and searching a user-defined database of allowable m/z differences corresponding to isotopes, adducts, and lipid class-specific artifacts.

Currently, there are three limitations of isotope, adduct, and in-source lipid artifact annotations tools. The main assumption when assigning annotations is that the monoisotopic peak is picked in the dataset, otherwise annotations may be incorrectly assigned with respect to a non-monoisotopic peak. This problem rarely occurs, however, because the monoisotopic species is usually highly abundant. Another shortcoming of current tools is that some annotations may be assigned due to random chance, but the user has no means of assessing false discovery rates (FDRs) from similarity coefficients. Two lipid species may share similar hydrophobicities, co-elute, and have an

m/z difference that randomly matches a known isotope, adduct, or artifact. It is unclear whether some type of FDR could be estimated from these correlations, and the stringency of the correlation threshold parameters are left as user-specified parameters. Finally, these programs only work with untargeted lipidomics data formats, even though redundant peaks are also detected in targeted lipidomics data acquired via LC-ESI-MS/MS. This last limitation highlights the general lack of bioinformatics tools for targeted lipidomics SRM and MRM modes.

### 1.5.5 Retention time alignment algorithms

A common problem of LC-ESI-MS/MS lipidomic analyses is processing multi-dimensional data corresponding to chromatographic peaks linked to MS and tandem MS spectra. A single untargeted LC-ESI-MS/MS acquisition may yield hundreds of chromatographic peaks, each corresponding to hundreds of lipid species. Depending on the number of transitions monitored from a targeted lipidomics approach, hundreds of SRM peaks may be detected in a single panel. When acquiring data across multiple samples, peak retention times can shift linearly and non-linearly due to routine variations in LC and other experimental factors (R. Smith, Ventura, and Prince 2015). Manual inspection of peak retention times, exact mass, and/or tandem MS spectra is therefore unfeasible. Using retention time windows is impractical due to the number of closely eluting peaks and possible retention time shifts. Similarly, low abundance lipid species in certain conditions or matrices may not trigger DDA or produce tandem MS spectra with DIA for unambiguous identification. To address this problem, retention time alignment algorithms have been proposed to align chromatographic peaks across samples. This process simplifies lipid identification because, in theory, peaks sharing the same aligned retention time are assumed to be the same lipid species. Once all sample chromatograms have been aligned, all peaks can be simultaneously identified using a combination of exact mass, tandem MS spectra, and/or retention time.

Nearly 50 novel retention time alignment algorithms have been proposed to align chromatograms across samples and match peaks corresponding to the same lipid species (R. Smith, Ventura, and Prince 2015). These algorithms monotonically warp peak retention times to capture systematic retention time shifts until some objective function is minimized. Algorithms may vary in computational complexity, use of "anchor points" or other user-defined parameters to estimate retention time variation across samples. A crucial parameter for many algorithms is the choice of reference sample, which all other samples are aligned to. If chromatographic conditions vary across samples, the choice of reference may lead to different peak alignments. No formal study has comprehensively evaluated even the most well-cited retention time alignment algorithms across empirical or *in silico* benchmark datasets (Nielsen, Carstensen, and Smedsgaard 1998; Kassidas, Macgregor, and Taylor 1998; Eilers 2004; Listgarten et al. 2005; Prince and Marcotte 2006; C. A. Smith et al. 2006; Benton et al. 2008; Jaitly et al. 2006).

The main limitation of current retention time alignment algorithms is the use of monotonic functions to model retention time shifts. Some variations in LC affect all analytes, due to solvent pH, ambient temperature, column age, and platform stability. Other retention shifts are localized to a single analyte at a given time, and these local shifts may be attributed to fluctuations in column pressure, ionization efficiency, and matrix-specific ion suppression (R. Smith, Ventura, and Prince 2015). While systematic retention time shifts can be modelled using monotonic functions, component level retention time shifts cannot. The second fundamental assumption of all alignment algorithms is that the deviation in retention time across samples is equal to or lesser than the time between adjacent peaks (C. A. Smith et al. 2006; R. Smith, Ventura, and Prince 2015). This assumption may be violated due to a combination of systematic and component level variation, leading to peak misalignments and possible misidentifications across samples.

In untargeted lipidomics, the benefits of retention time alignment greatly

outweigh the possibilities of misaligning and misidentifying lipid species. Data analysis is greatly simplified by simultaneously annotating a single peak and extrapolating that identity to other sample peaks sharing the same aligned retention time. Retention time misalignments often occur when lipid isomers elute closely together and when the number of detected peaks varies across samples. If lipid identities and abundances are reported at the total carbon and double bond level, which is standard in untargeted lipidomics (Koelmel et al. 2017), some misaligned peaks may be end up correctly identified within the appropriate lipid class. In targeted lipidomics, however, where identifying and relatively/absolutely quantifying lipid isomers across biological conditions and matrices is desired, retention time misalignment may lead to erroneous biological conclusions. Coupled with the strong emphasis of manual peak picking and identification, very few retention time alignment algorithms have been developed specifically for targeted lipidomics data (e.g. MRMDIFF, Tsugawa, Cajka, et al. (2015); MRMAnalyzer, Cai, Weng, et al. (2015) and Cai and Zhu (2019)).

### 1.5.6 Targeted lipid identification algorithms

Targeted lipidomics SRM and MRM data consist of peaks monitored at discrete transitions and retention times. In a traditional, targeted lipidomics analysis, the expert operator constructs an SRM or MRM transitions to monitor each lipid of interest. The average retention times of these target lipids are generally known in advance from a prior untargeted or targeted lipidomic analysis of the same sample. Knowledge of target analyte retention times is useful because a single transition may contain dozens of detected peaks with different retention times. Post-data acquisition, expert operators manually interrogate extracted ion chromatograms (XICs) outputted by vendor-specific software packages to pick peaks with the closest retention time to the lipid targets. This process is time-consuming, especially when the number of detected peaks is fundamentally influenced by the sample treatment, disease state, and tissue matrix. Different analysts may disagree

on which peaks represent the lipid targets, or if their lipid targets are even present in the biological sample. Some degree of automated peak picking has been achieved by vendor-specific and open-source programs (e.g. MultiQuant, SCIEX, MRMAnalyzer (Cai, Weng, et al. 2015; Cai and Zhu 2019)), which pick peaks using a user-defined list of lipid species, monitored transitions, and retention time windows. MultiQuant selects the closest peak to the target retention time, while MRMAnalyzer initially aligns peak retention times, using lipid standards as anchor points, before assigning an identity to the most intense peak within a retention time window. Expert curation is still required to validate these peaks due to variations in LC, and the variable number of closely eluting peaks across biological samples. Nevertheless, manual curation is subjective, biased towards reporting the lipid targets expected by the analyst, and therefore prone to misidentification (Melnik et al. 2017).

No targeted lipid identification algorithms have been developed to identify SRM data collected across multiple platforms, where single transitions are monitored per lipid species. These algorithms are urgently required because retention time alignment algorithms commit strong assumptions about peak retention times, which are often violated due to a combination of systematic and component level variation. Only two programs for targeted lipid identification, MRMAnalyzer and MRMPROBS, are reported in literature. The program MRMAnalyzer claims to identify SRM peaks by aligning samples using several internal standards and assigning lipid identities using a 60 second retention time window to a user-defined reference library of known lipid species (Cai, Weng, et al. 2015). It is unclear how many lipid standards are required for effective alignment, especially since lipid standards are limited by commercial availability and price. When validating MRMAnalyzer against a mixture of 182 metabolite standards, Cai, Weng, et al. reported an identification rate of 83.5%, which highlights the limitation in using a retention time window to assign lipid identities to the most intense peaks. As this experiment was performed using standards, identification

rates may be much lower in experimental datasets profiling mixed-matrices with varying numbers of lipid isomers across biological conditions. The other program MRMPROBS (Tsugawa, Arita, et al. 2013) constructs a multivariate logistic regression using five MRM peak features from a user-defined reference library. When applied to SRM data, MRMPROBS can only construct the model using one of five features, namely retention time. Furthermore, MRMPROBS is designed such that the number of lipid identities in the library cannot exceed the number of peaks in the test sample. The names of the lipid identities must also match the putative lipid identity in the SRM method. MRMPROBS is thus incapable of identifying lipids from data acquired across biological conditions or SRM methods. With limitations using MRMAnalyzer and MRMPROBS, novel targeted lipid identification algorithms are vital to avoid biases associated with manual curation approaches.

## 1.6 Addressing the paucity of targeted lipid identification algorithms

It is difficult to understand why there are so few targeted lipid identification programs when peak identification remains such a challenge. A possible explanation may be attributed to incorrect perceptions of targeted lipidomic approaches. In the previous sections, untargeted and targeted lipidomics were defined by the data acquisition modes available to each approach. In doing so, the strengths and weaknesses of these approaches were contrasted, emphasizing the complementary use of both methods for lipid identification and quantitation. Untargeted lipidomics excels at profiling broad changes in the lipid composition of the most abundant species across all lipid families; targeted lipidomics excels at detecting and quantifying low and high abundance lipid isomers within a given family. However, different definitions of untargeted and targeted lipidomics are used in literature. Untargeted lipidomics is instead described as the approach for unbiased lipid identifica-

tion and discovery, while targeted lipidomics is described as the approach for lipid quantitation (Cajka and Fiehn 2014; Narváez-Rivas and Q. Zhang 2016; Bilgin et al. 2016; Collins et al. 2016; Contrepois et al. 2018; Gross 2017; Goracci et al. 2017; Forest et al. 2018; Lee and Yokomizo 2018; Blaženović, Kind, Sa, et al. 2019). These binary definitions are problematic because they overgeneralize the benefits of each approach, leading to incorrect perceptions of both untargeted and targeted lipidomics.

Targeted lipidomics is viewed as an approach to quantify already known lipid species. This belief is partly true as knowledge of the lipid target is required to monitor the appropriate transition to detect that species. Yet, Section 1.5.6 describes a host of problems associated with simultaneous peak picking/identification. Multiple peaks may be detected within a given transition, and misidentifications can occur when these "off-target" peaks are confused for the lipid of interest. Even if the analyst picks the correct peak corresponding to their lipid target, the other detected peaks, possibly representing genuine lipid species, are generally discarded. If there was a means to link lipid targets with their peak-specific features, parent and product ion transitions, and presence in discrete matrices, a targeted lipid identification classifier could unbiasedly "discover" all lipid identities within each monitored transition.

It is important to clarify the definition of lipid "discovery" because several meanings are used in literature, primarily describing untargeted lipidomics. To most analytical chemists, discovery is the unbiased detection and structural identification of possibly any and all lipid species in a single MS run. Bioanalytical chemists define discovery differently in the context of drug or biomarker discovery, where identified lipids are found to be differentially abundant across conditions. Another technological definition frames discovery as the use of high-resolution MS to identify lipids across all classes, instead of parking on a given product ion to identify lipids within a single class. The analytical chemistry definition of discovery is most prevalently used in literature; however, this definition overgeneralizes

the strengths of untargeted lipidomics, while misrepresenting the inherent weakness of untargeted mass spectrometry acquisition modes. Having discussed full scan, DDA, and DIA modes in Section 1.4.1, it is evident that untargeted lipidomics cannot meaningfully detect all lipid species in a sample. Some positional isomers cannot be resolved by high-resolution mass spectrometry, nor low abundance lipids whose spectra are dominated by high abundance species. These limitations of current untargeted mass spectrometry-based technologies demonstrate how untargeted lipidomics cannot discover all possible lipid signals. Furthermore, what lipids are detected by the mass spectrometer is fundamentally influenced by sample preparation, and the use of LC. Lipid extraction methods may concentrate certain lipid classes at the expense of others (Folch, Arsove, and Meath 1951; Bligh and Dyer 1959; Alfred H. Merrill et al. 2005; Reis et al. 2013), such that DDA or DIA cannot identify certain lipid species. LC may also alter which lipid isomers are irresolvable, or resolvable and thus identifiable by untargeted acquisition modes. A mass spectrometrist, optimizing their LC separation, will invariably cause some lipid isomers to co-elute, and others to separate. While the operator may be unaware of these outcomes, it can be argued nonetheless that they are biasing their platform to preferentially target certain species.

I assert that the term discovery should be re-interpreted as *the capacity to meaningfully identify all acquired signals*, constrained by the limits of currently available MS-based technologies and bioinformatics tools. This definition unifies similar views of untargeted and targeted lipidomics expressed by the Lipidomics Standards Initiative (LSI) Consortium (V. B. O'Donnell et al. 2020; V. O'Donnell B., R. Murphy C., and Watson 2014; Porta Siegel, Ekroos, and Ellis 2019; Liebisch, Ekroos, et al. 2017). Untargeted lipidomics is discovery-oriented because DDA and DIA can detect all high abundance lipids and identify them via spectral database matching algorithms. With appropriate tools for targeted lipid detection and identification, discovery-driven targeted lipidomics is achievable by identifying all low and high

abundance lipid species detected within discrete SRM and MRM transitions. To unlock this potential, it is imperative that novel bioinformatics tools be developed for targeted lipidomics to enable the discovery of these lipid species.

## 1.7 Hypothesis and objectives

The overall goal of my thesis is to address the dire paucity of bioinformatics methods for targeted lipidomics data. I hypothesize that accurate, robust, and simultaneous targeted lipid annotation, identification, and discovery are achievable through rigorous modelling of SRM features. Thus, I outline three objectives to address this hypothesis:

1. Develop a computational method to annotate in-source lipid artifacts and isotopes from targeted lipidomics data. (**Chapter 2**).
2. Develop a machine learning classifier to identify SRM peaks. (**Chapter 3**)
3. Use this classifier to explore a novel and unbiased, discovery-driven paradigm where all peaks, encompassing all detected signals, are picked for annotation and downstream targeted lipid identification. (**Chapter 4**)

# Chapter Two

# In-source lipid artifact and isotope annotation tool

## 2.1 Objective

Removing in-source lipid artifacts and isotopes to reduce data complexity remains an essential yet unaddressed problem in targeted lipidomics. My objective was to develop a computational tool to annotate artifactual and isotopic ions from LC-ESI-MS/MS data by exploiting the principles of LC coupled to mass spectrometry. I explored the utility of this program on a variety of datasets identified by expert operators, trained to avoid picking these peaks, and identified previously unknown lipid isotopes and artifacts selected in these datasets.

## 2.2 Statement of author contributions

Justin G. Chitpin, Steffany A. L. Bennett, and Theodore J. Perkins conceived the study. Justin G. Chitpin developed the algorithm, analyzed the data, prepared the figures, and wrote the chapter with Steffany A. L. Bennett and Theodore J. Perkins; mass spectrometry datasets were prepared and provided by Graeme P. Taylor, Hongbin Xu, Thao Nguyen, Irina Alec, and

Jayu Y. Gao. Thao Nguyen and Graeme P. Taylor further provided expert consultation and advice, and structurally validated matrix-specific lipid identities by SRM-IDA-EPI.

## 2.3 Summary

A challenge in all mass spectrometry-based technologies is that the number of detected signals is an overrepresentation of the true number of unique molecular species. In lipidomics, ESI is the main source generating artifactual ions, as lipid analytes may prematurely fragment "in-source" or dimerize prior to tandem MS. Isotopes, conversely, are generally caused by the abundance of naturally-occurring carbon-13 atoms that incorporate into lipid species, increasing their mass by an additional 1-4 daltons. When monitoring a list of SRM or MRM transitions, known as panels, it is imperative that these lipid artifacts and isotopes derived from other species are not confused for the targets of interest and picked for downstream identification. Current methods to detect lipid artifacts and isotopes are limited to data acquired from untargeted lipidomic approaches and are urgently required in targeted lipidomics to avoid lipid misidentification. Here, I present the first computational tool specifically designed for targeted lipidomics data to identify common sphingolipid and glycerophospholipid artifacts and isotopes.

## 2.4 Introduction

Lipidomics data acquisition is driven by liquid chromatography coupled to tandem mass spectrometry via electrospray ionization (LC-ESI-MS/MS). A unique challenge of any mass spectrometry-based assay is the detection and filtering of unwanted signals corresponding to artifactual analytes. In lipidomics, these artifacts are generally caused by premature lipid fragmentation during ionization, a necessary step to enable ion detection by the mass

analyzer. ESI is the "soft" ionization method of choice in lipidomics because it protonates/deprotonates analytes with little fragmentation (Gross 2017; Cajka and Fiehn 2016). Nevertheless, some lipid analytes contain functional groups that are especially prone to in-source fragmentation. Sphingolipids, for example, can be deglycosylated, dehydrated, or, in rare events, oxidatively dimerized or trimerized by the ESI source. With the exception of dimers and trimers, these in-source modifications always remove specific chemical groups from the monoisotopic sphingolipid. A single deglycosylation event would result in a loss of 162 daltons, while a dehydration event would result in a loss of 18 daltons. Dimers are the result of oxidation and/or are formed by the association of an ion $[M+H]^+$ with its neutral counterpart, forming $M + [M+H]^+$ of the same lipid analyte. These artifacts are often produced when a highly abundant lipid species elutes through the LC system, forming fusion products during ionization. A dimer, representing the fusion of the same two lipid species, would result in an artifactual lipid ion with twice the mass over charge as the single monoisotopic species. Moreover, multiple in-source fragmentation events may occur, and a +1 isotopic peak may lose both a sugar and hydroxyl group as water, equalling a total loss of 179 daltons.

Currently, computational methods to identify artifactual ions and isotopes are lacking for targeted lipidomics data. Several reasons may explain this paucity as in-source artifact and isotope annotation programs are common for untargeted lipidomics data. Firstly, targeted lipidomics data acquisitions methods are discrete and monitor specific parent and product ion pairs, unlike untargeted lipidomics data where signals are detected across the entire mass analyzer range. With fewer detected signals in targeted lipidomics data, there exist fewer possibilities of encountering in-source artifacts or isotopes. Monitoring both parent and product ions, as well, improves the selectivity for detecting the lipid species of interest, instead of background noise or other unwanted analytes. Secondly, targeted lipidomic analyses rely on expert operators to manually pick peaks corresponding to

their lipids of interest. Thus, filtering adducts is perceived as unnecessary if the analyst is only interested in targeting protonated or deprotonated lipid species, for example. The major assumption of this justification is that the analyst is skilled enough to avoid erroneously picking lipid artifacts and isotopes. Given the subjectivity of expert operator peak picking, this assumption is challenged in Chapter 3, where a novel computational tool is discussed for unbiased, targeted lipid identification. Nonetheless, it remains an empirical question whether expert operators unknowingly confuse artifactual or isotopic ions for their target lipid species.

To address the utility and urgency of identifying in-source lipid artifacts and isotopes, I present the first artifactual and isotopic ion detection program for targeted lipidomics data. I apply this tool to SRM datasets profiling sphingolipids (d18:1 backbones), glycerophosphocholines, and sphingomyelins, and show that 0.7-3.6% of all peaks picked by expert operators actually corresponded to lipid isotopes, deglycosylations, dehydrations, or dimers. While these figures may appear small, annotating these peaks improved the overall quality of SRM data and provided an empirical estimate of manual peak picking error.

## 2.5 Methods

### 2.5.1 Algorithm

Let $\mathbf{P} = \{P_1, P_2, \ldots, P_n\}$ represent an ordered list of peaks picked in a single sample monitoring the same product ion, and thus family of lipid species. Each peak is described by parent ion $q$ and retention time $r$. To assign annotations, all peaks in $\mathbf{P}$ must be paired to determine if the second peak in the pair is an artifact or isotope with respect to the first peak. An annotation is assigned if both peaks in the pair share the same retention time and whose difference in parent ion m/z, between the second and first peak, matches a known annotation described further below. For each pair of peaks $(P_i, P_j) : 1 \leq i \neq j \leq n$, which are indexed by $l$, let $r_i^l$

and $r_j^l$ denote the retention times of the $i^{th}$ and $j^{th}$ peak. Similarly, let $q_i^l$ and $q_j^l$ denote the parent ion m/z of the $i^{th}$ and $j^{th}$ peak. Finally, let $\mathbf{A} = (A_1, A_2, \ldots, A_m)$ represent a list of all possible peak artifacts, isotopes, and permissible combinations of artifacts and isotopes. Peak artifacts include a -1 dehydration (-18 m/z), -2 dehydration (-36 m/z), -1 deglycosylation (-162 m/z), -2 deglycosylation (-324 m/z), -3 deglycosylation (-486 m/z), +1 isotope (+1 m/z), ..., +4 isotope (+4 m/z), +2 dimer ($2 \times q_i^l$), and +3 dimer ($3 \times q_i^l$). The compound annotations contained in $\mathbf{A}$ are computed by enumerating all feasible subsets of artifacts and isotopes. Feasible compound annotations are determined by the product ion monitored. In our study, both artifact and isotope annotations are possible for sphingolipids with a d18:1 backbone because they are monitored at a product ion of 264.3 m/z. Only isotopes are allowed for glycerophosphocholines because their product ion is monitored at 184.1 m/z. Glycerophosphocholine lipids monitored at this product ion do not undergo the same modifications as sphingoid base lipids (i.e. dehydrations, deglycosylations, and dimers) because they contain no hydroxyl groups or sugars in their structure. The $k^{th}$ annotation in $\mathbf{A}$ is described by $d_k$ m/z. The tolerance parameter $\tau$ is specified according to the user's parent ion mass analyzer. Another tolerance parameter, T must also be specified according to the user's LC system. The retention time tolerance is defined as T significant figures after the decimal place. The method is described in Algorithm 1.

### 2.5.2 Data

SRM data were acquired from 440 and 499 samples profiling for sphingolipids and glycerophosphocholines, respectively. The exact biological conditions, number of samples/acquisitions, organisms, and matrices of each dataset are described in Appendix 5.4. Trained expert operators analyzed the raw MS data and manually picked peaks corresponding to their lipid targets of interest. All data were acquired using an Agilent 1290 Infinity II LC system coupled via ESI to a SCIEX QTRAP 5500 tandem mass spectrometer

**Algorithm 1:** In-source lipid artifact and isotope finder.

**Input:** Peaks $\mathbf{P}$, annotation list $\mathbf{A}$, retention time tolerance $T$, and machine tolerance $\tau$
**Output:** Annotations for peaks $\mathbf{P}$
**for** $1 \leq i \neq j \leq n$ **do**
  **if** $|r_i - r_j| < T$ **then**
    **for** *k in 1 .. m* **do**
      **if** $|q_j - q_i - d_k| < 2\tau$ **then**
        *Assign annotation* $A_k$ *to peak* $P_j$ *with respect to* $P_i$
      **end**
    **end**
  **end**
**end**

operating in positive ion mode as described by Granger et al. (2019) and Xu et al. (2019).

## 2.6 Results

The sphingolipid and glycerophosphocholine datasets, profiling blood and brain matrices, consisted of 53,427 and 165,273 peaks, respectively. Using the in-source lipid artifact and isotope annotation tool, 1,897 sphingolipid and 1,095 glycerophosphocholine artifacts and isotopes were reported across both datasets. For the sphingolipid dataset, 21 compound artifact/isotope annotations were assigned (e.g. double-dehydrated and deglycosylated lipid, dehydration of a +1 isotope species). Many peaks also shared the same retention times with multiple different peaks and were assigned more than one artifact or isotope annotation (Table 2.1). As there were too many unique compound annotations to graph, all annotations are presented individually in Figure 2.1. In total, 3.6% and 0.7% of all expert-picked sphingolipid and glycerophosphocholine peaks corresponded to possible off-target lipid artifacts and isotopes.

A shortcoming of the annotation algorithm is that artifacts and isotopes

**Table 2.1: Annotation depth per identified lipid artifact/isotope.**

| | # of peaks with one or more annotations | | | | |
|---|---|---|---|---|---|
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Sphingolipid | 1897 | 1781 | 371 | 29 | 1 |
| Glycerophospholipid | 1095 | 837 | 253 | 5 | 5 |

are identified using a retention time cutoff and machine tolerance cutoff. Therefore, no probabilistic scores were assigned to each peak annotation or compound annotation. This problem is ubiquitous across all in-source lipid artifact and isotope annotation programs (e.g. CAMERA, Kuhl et al. (2012); CliqueMS, Senan et al. (2019); LipidFinder, O'Connor et al. (2017); MS-FLO, Defelice et al. (2017)). A case study was conducted to validate the identities of several annotations by SRM-IDA-EPI. One example is highlighted in Figure 2.2, where a +2 sphingolipid isotope was reported at transition 652.9/264.3 with respect to the monoisotope detected at transition 650.6/264.3. Both peaks were detected with a retention time of 27.17 minutes (Figure 2.2a). The red peak indicates the monoisotope while the blue peak, which is much less intense than the red peak, indicates the +2 isotope. Only these two peaks were detected in the SRM method at 27.17 minutes (Figure 2.2b), and their identities, Cer(d18:1/24:0) and the corresponding +2 isotope, were structurally confirmed by IDA-EPI analysis.

## 2.7 Discussion

Using the artifact and isotope annotation tool, approximately 0.7% and 3.6% of all expert-picked sphingolipid and glycerophosphocholine peaks were reported as off-target lipid artifacts and isotopes. An SRM-IDA-EPI case study was also provided to demonstrate the correctness of a +2 isotope assigned by the annotation tool. The intensity of the +2 isotope was also much weaker than the monoisotope and the theoretical isotopic distribution of Cer(d18:1/24:0) could have been computed to orthogonally validate the

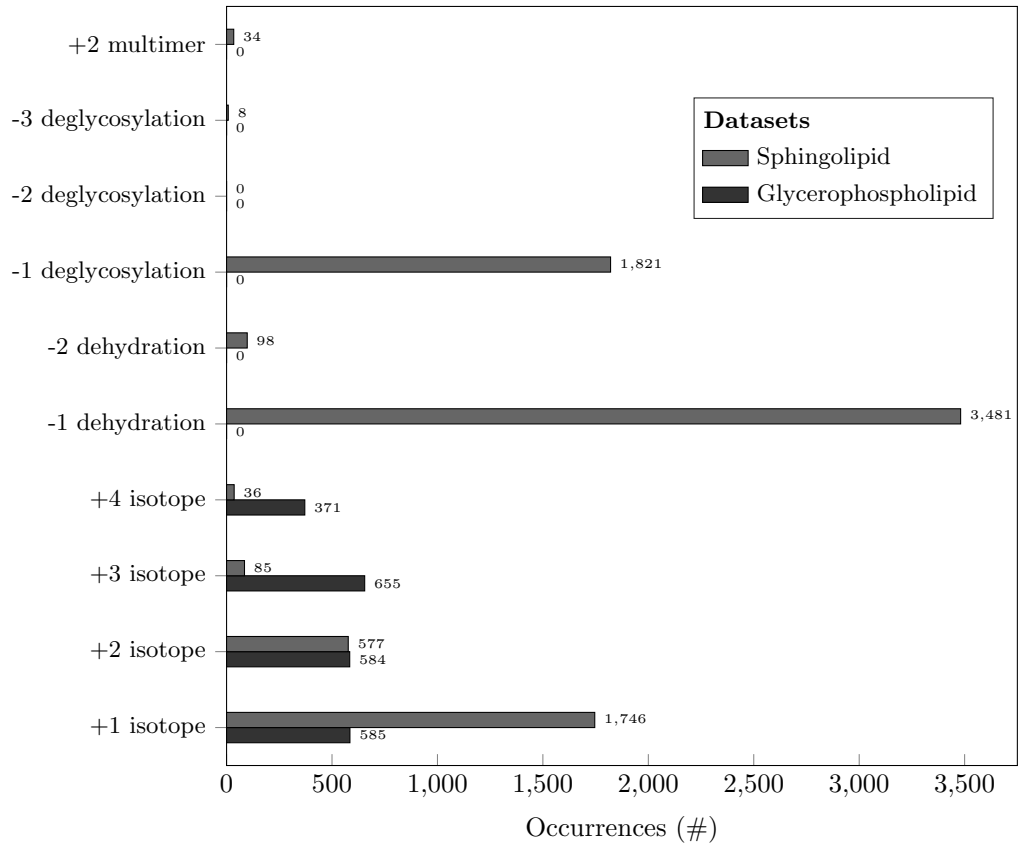
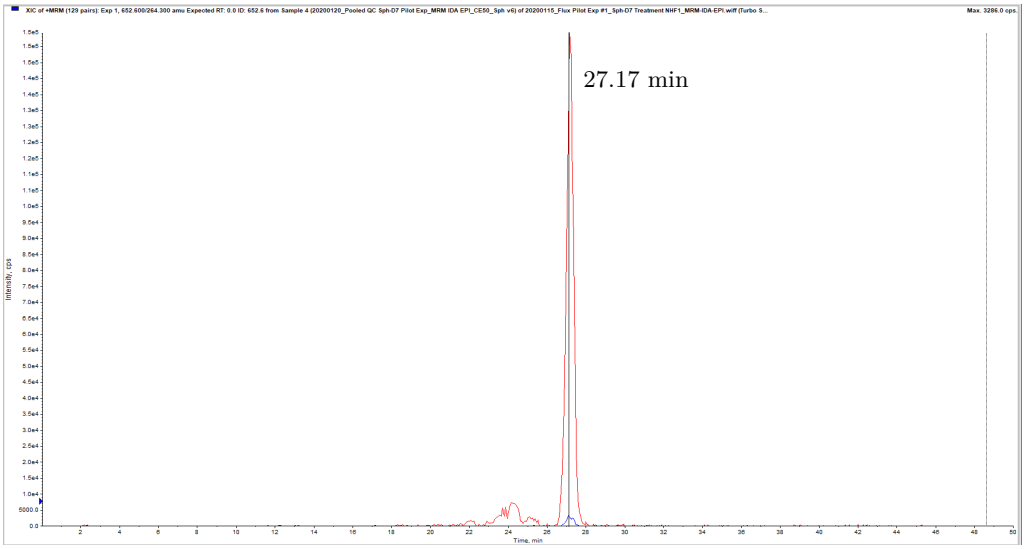



**Figure 2.1: Lipid artifacts and isotopes identified from the sphingolipid and glycerophospholipid datasets using the in-source lipid artifact and isotope annotation program.**

assigned peak annotations. However, theoretical isotope distributions require that data is not acquired from stable isotope labelling experiments, and flux experiments are common in lipidomics and metabolomics. While some of these annotations may correspond to false positives, analysts may interpret these assigned annotations at their own discretion. Operators may choose to ignore, filter, or experimentally validate these annotations, as demonstrated in the case study. Without the annotation tool, it would be highly inefficient to manually compute all permissible annotations and apply them to a list of SRM peaks. A total of 33 individual and compound annotations were computed for the sphingolipid dataset, and 4 isotope annotations for the

**(a)** XIC plot




**(b)** SRM plot

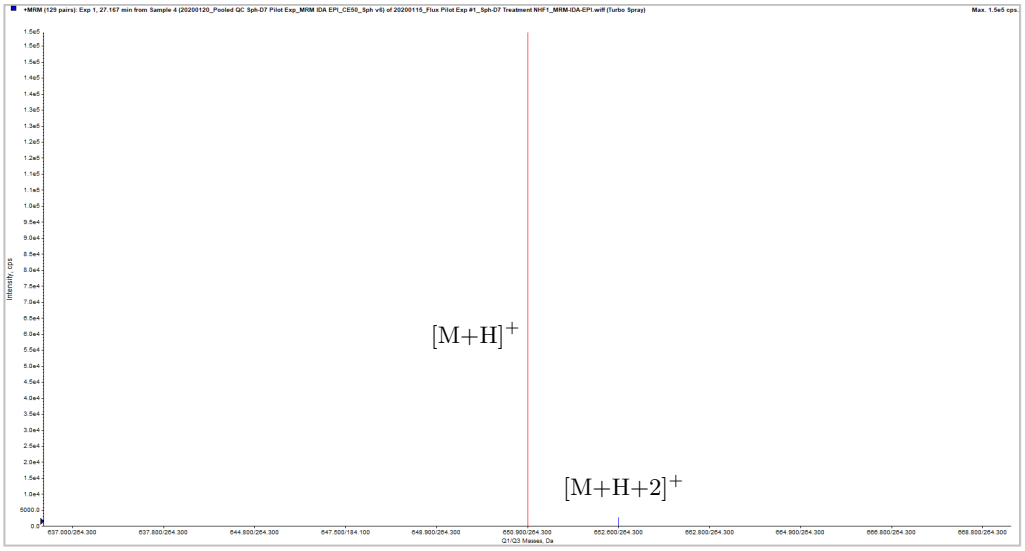



**Figure 2.2: Case study of a +2 isotope reported by the annotation tool and validated by SRM-IDA-EPI to be a +2 isotope of Cer(d18:1/24:0).** (a) XIC plot showing the overlay of the red and blue peaks which are both detected at 27.17 minutes (x-axis: retention time; y-axis: intensity). (b) SRM centroid spectra of the red and blue peaks detected at 27.17 minutes (x-axis: Q1/Q3 transition; y-axis: intensity). The red peak is the Cer(d18:1/24:0) monoisotope detected at 650.6/264.3; the less intense blue peak is the +2 isotope of the red peak.

glycerophosphocholine dataset. With a list of 100 or more SRM peaks per sample, manually finding all co-eluting peaks and determining if one of the 33 or 4 annotations are present is unfeasible for high-throughput targeted lipidomic analyses.

A strength of my approach is that it uses retention times, which are obtainable from text file list of peaks, to identify lipid artifacts and isotopes. In contrast with MS1 annotation programs for untargeted lipidomics, most software packages compare peak chromatography profiles stored in vendor-specific file formats. These programs bin peak intensities by retention time and compute either Pearson or cosine similarities, which are thresholded to identify lipid artifacts and isotopes (Kuhl et al. 2012; Uppal, Walker, and Jones 2017; Senan et al. 2019). A notable exception is MS-FLO (Defelice et al. 2017) which uses a similar retention time matching approach for artifact, isotope, and adduct annotation, although it works specifically for untargeted lipidomics data. Which approach is best remains unresolved in the greater lipidomics and metabolomics field. Senan et al. benchmarked their program, CliqueMS, against CAMERA, MS-FLO, and xMSannotator using a mixture of metabolite standards and found their cosine similarity algorithm to correctly annotate more artifacts, adducts, and isotopes (Senan et al. 2019). Their results on two biologically-relevant datasets acquired in positive and negative ion mode, however, were comparable to CAMERA. Between 6 and 20 artifacts were manually validated in these datasets and $\leq$ 75% of the artifacts were detected by either CliqueMS or CAMERA. Although CliqueMS correctly annotated 30-50% more peaks than CAMERA, the program also assigned twice as many annotations than CAMERA. These annotations may be incorrect and future evaluation of these MS-1 annotations programs on biologically-relevant datasets with a greater number of structurally validated annotations is required to compute both precision, recall, and false positive rates to determine the best performing program.

Future work remains to improve lipid artifact and isotope detection for targeted and untargeted lipidomics data. A major factor influencing

chromatographic similarity and retention times are the integration parameters used to detect peaks. Gaussian smoothing can dramatically alter the intensity profile of peak tails, but not retention times which are more robust against these effects. One limitation of my study was that expert operators may have changed peak integration settings while picking peaks to improve quantitation. While retention times are robust against Gaussian smoothing, some alterations nevertheless occur. Therefore the reported 3.6% and 0.7% off-target ions may represent a slightly conservative estimate of the true number of lipid artifacts and isotopes picked by the analysts. Furthermore, different programs (e.g. MultiQuant, SCIEX; MassHunter, Agilent; TraceFinder, ThermoFisher Scientific; MRMPROBS, Tsugawa, Arita, et al. (2013)) may utilize distinct peak detection algorithms and may disagree as to where a peak even starts and ends above the baseline noise. Thus, any differences in MS1 annotation tools using binned chromatographic similarity profiles or co-eluting peak apices could be confounded by the choice of vendor-specific or open-source peak detection programs. Subsequently, investigating whether these peak detection and integration parameters impact chromatographic similarity profiles and retention times may improve artifact and isotope detection in the future.

## 2.8 Conclusion

In this study, I present the first in-source lipid artifact and isotope annotation tool for targeted lipidomics data. This program is operable on any list of picked peaks exported as a text file, and is capable of annotating in-source sphingolipid artifacts and isotopes from any lipid category. Using this tool, I further show that manual peak picking, even by experienced operators, is susceptible to artifactual and isotopic errors.

# Chapter Three

# Bayesian annotations for targeted lipidomics (BATL)

## 3.1 Objective

My objective was to develop the first machine learning classifier capable of annotating lipid identities using the features inherent in targeted lipidomics data. To do so, I prioritized a naïve Bayes classifier, which is a family of probabilistic machine learning models. To develop the optimal naïve Bayes classifier for this purpose, my goals were to establish evidence-based classifier decision rules, identify ideal discriminative peak features, and establish criteria that define when not to assign a lipid identity to a peak. To facilitate software usage across multiple labs, my secondary objectives were to determine the minimum number of training data required for accurate lipid annotation and to establish whether accurate classification could be achieved with matrix-independent versus matrix-dependent datasets. The result of these investigations was *Bayesian Annotations for Targeted Lipidomics* or BATL, an R package that annotates lipid identities from datasets employing targeted lipidomic approaches.

## 3.2 Statement of author contributions

Justin G. Chitpin, Steffany A. L. Bennett, and Theodore J. Perkins conceived the study. Justin G. Chitpin developed the algorithm, analyzed the data, prepared the figures, and wrote the chapter with Steffany A. L. Bennett and Theodore J. Perkins; mass spectrometry datasets were prepared and provided by Graeme P. Taylor, Hongbin Xu, Thao Nguyen, Irina Alecu, and Jayu Y. Gao; Graeme P. Taylor further provided expert consultation and advice, and structurally validated matrix-specific lipid identities by SRM-IDA-EPI.

## 3.3 Summary

Targeted lipidomics excels in relative and absolute quantification through selective or multiple reaction monitoring (SRM and MRM) that reduces matrix complexity by monitoring pairs of parent and product ions. This approach contrasts with high-resolution mass accuracy approaches to identify lipid species by parking on a given product ion to reduce detection of competing ion signals. Targeted lipidomic approaches thus enable unparalleled sensitivity for both low and high abundance lipids within a given lipid family. While highly selective for detecting lipids of interest, SRM and MRM peaks are commonly and manually identified by expert operators under the incorrect assumption that an observed signal always corresponds to a single lipid target within a given retention time window. While a valid assumption for a single condition, lipid detection is fundamentally influenced by environmental changes. For example, in biologically relevant conditions, matrices may be mixed (e.g. blood and brain) due to the disease state. When multiple isomers are detected at a given transition or when the number of isomers changes between matrices or conditions, high-throughput lipid identification becomes problematic and manual curation unreliable. This expert-curation approach is not only prone to identification error but is biased towards reporting the *a priori* lipids of interest expected by expert

analysts. To address these biases and afford discovery within a targeted lipidomic context that analyzes lipid changes family by family, I developed the first targeted lipid identification classifier capable of high-throughput annotation of lipid species from SRM and MRM data across matrices and conditions. Using chemically and biologically diverse SRM datasets with structurally-validated identifies, I extensively validated my classifier on 182 independent datasets. Across all annotated peaks, I demonstrate an average accuracy of 92% for sphingolipid (d18:1 sphingoid backbones) and 79% for glycerophosphocholine datasets profiled across diverse organisms, matrices, and biological conditions.

## 3.4 Introduction

Lipidomics is an emerging "-omics" field aimed at discovering and characterizing all lipid species within a biological system (Brown and R. C. Murphy 2009). Advances in liquid chromatography coupled to tandem mass spectrometry via electrospray ionization (LC-ESI-MS/MS) are enabling the comprehensive, high-throughput detection of lipid species across numerous organisms, tissue matrices, and biological conditions (L. Zhang, Han, and Wang 2018). These lipidomic studies are generally divided into untargeted and targeted approaches depending on the pursued biological question. Untargeted approaches aim to identify the overarching changes in lipid composition of the most abundant species across families using the fewest number of MS analyses; identification is achieved by exploiting high-resolution mass accuracy and spectral assessment of analyte structure. Much interest in this workflow stems from the thousands of empirically observed and theoretically predicted lipid species with unknown biological functions (Kind, Liu, et al. 2013; Foster et al. 2013; Sud et al. 2007; Caffrey and Hogan 1992; Watanabe, Yasugi, and Oshima 2000; Kuo and Tseng 2018; Blanchard et al. 2013). Targeted approaches seek to identify and quantify the complete repertoire of metabolic changes in both high and

low abundance lipid species within a given family, requiring multiple MS analyses to encompass the entire lipidome. Differential targeted lipidome analyses of all species across samples and matrices, however, require an approach, whereby quantitation of individual lipid isomers (both low and high abundance) is achieved by monitoring diagnostic ions with discrete m/z values indicative of a given lipid family.

Currently, we lack a means of linking prior structural knowledge of the lipid target, its presence in discrete tissues, its parent and product ion pairs, and its peak-specific features across conditions to effectively annotate the appropriate lipid identity, regardless of expert or other bias. Neither SRM or MRM are sufficiently selective to monitor a single transition corresponding to a unique lipid species. Many lipid isomers share the same transition; thus, many peaks are detectable at a given SRM or MRM transition. Identification is achieved by comparing peaks retention times to target lipid retention times. When several peaks elute within a target lipid retention time window, the lipid identity is matched to the peak with the closest matching retention time. The assumption of this approach is that the closest detected peak within a retention time window corresponds to the lipid of interest "targeted" by the operator. State of the art programs implement this peak identification algorithm as the final data processing step after peak visualization and detection from extracted ion chromatograms (XCMS-MRM, Domingo-Almenara et al. (2018); MZmine2, Pluskal et al. (2010); MultiQuant, (SCIEX); peakPantherR, `https://github.com/phenomecentre/peakPantheR`; TarMet, `https://github.com/hcji/TarMet`; sRm, `https://github.com/wilsontom/sRm`). Because this approach is highly sensitive to peak retention time shifts across samples, the majority of SRM and MRM peaks are manually annotated (Bowden et al. 2018). Nevertheless, manual peak identification/annotation remains both a challenging task and a major bottleneck of targeted lipidomic studies.

Manual peak identification poses several challenges for high-throughput analysis. For one, peak retention times can shift considerably across samples

based, in part, on the complexity of their matrix and/or on routine variations in chromatography; one peak could be mistaken for another or even regarded as a novel lipid isomer altogether. Many retention time alignment algorithms have been proposed to address this correspondence problem (Nielsen, Carstensen, and Smedsgaard 1998; Eilers 2004; C. A. Smith et al. 2006; Tsugawa, Ohta, et al. 2014), yet all are founded on the assumption that retention time differences across samples are equal to or lesser than deviations between adjacent peaks (R. Smith, Ventura, and Prince 2015). This assumption may not be accurate as the chosen alignment parameters (i.e. bandwidth) or reference sample which all other samples are aligned to can impact peak alignment across MS runs. As such, peak identification is often based on raw retention time. Yet, given the same dataset, analysts may interpret different lipid species from the same peaks (Liebisch, Ejsing, and Ekroos 2015; Liebisch, Ekroos, et al. 2017; Koelmel et al. 2017). Differences may also arise when analysts consider features aside from retention time such as peak intensity, shape, or signal to noise ratio which may be affected by disease state or treatment condition. Thus, independent analyses of the same SRM/MSM dataset by different experts may misidentify and misquantify different lipid species, leading to conflicting biological conclusions (Melnik et al. 2017). An unbiased, high-throughput classifier is urgently needed to address this problem.

No program exists capable of high-throughput annotation of SRM or MRM data collected on multiple platforms, where single transitions are monitored per lipid species. While LipidView (SCIEX) contains targeted lipidomics data for over 25,000 analytes, it is a closed-source program restricted to SRM/MRM data acquired in proprietary format. LipidView also cannot annotate lipid molecular identities, as all annotations are reported by total carbon and total number of double bonds (SCIEX 2012). Non-proprietary software packages MRMPROBS (Tsugawa, Arita, et al. 2013) and MRMAnalyzer (Cai, Weng, et al. 2015; Cai and Zhu 2019) have sought to address the SRM/MRM identification problem for targeted

metabolomics and lipidomics data, yet face difference challenges that impede high-throughput applications. MRMPROBS accepts MRM data using a user-defined training library of known analytes and constructs a multivariate logistic regression model using five MRM features to assign lipid identities to peaks. These features are the peak intensity, retention time, shape and co-elution score, and qualifier/target (QT) ratio. The QT ratio is defined as the quotient between the most intense qualifier fragment and target fragment monitoring the same analyte. However, MRMPROBS requires multiple transitions per analyte as a peak monitored at a single transition can be matched to multiple lipid species. Thus, it is not possible for MRMPROBS to identify lipid isomers within a single transition from SRM data. Unlike MRMPROBS, MRMAnalyzer uses the OBI-WARP retention time alignment algorithm (Prince and Marcotte 2006) that aligns samples to reference standards. One major issue regarding the application of warping algorithms is the minimum number of standards required for successful alignment. The authors' validated their algorithm using a mixture of 182 metabolite standards which is uncommon in high-throughput lipidomics where pure lipid standards are limited by commercial availability and price. MRMAnalyzer achieves peak identification by direct match of aligned peaks where identities are assigned to the most intense peak within a very large, sixty second retention time window. Different peaks may also be misannotated due to variable lipid abundances within and across biological samples. Finally, the R package is not publicly available.

To address these challenges and provide a high-throughput annotation solution for targeted lipidomic approaches, I present a naïve Bayes classifier for targeted lipid identification and annotation. This classifier is vendor agnostic, thus applicable to SRM and MRM data acquired from any LC-ESI-MS/MS platform, and requires that only a single transition is monitored per target lipid species. Using datasets encompassing different organisms, matrices, and biologically divergent conditions, I demonstrate automated and accurate identification and annotation of sphingolipids, targeting species with d18:1

**Table 3.1: SRM and MRM peak features for naïve Bayes model.**

| Feature | Description |
|---|---|
| Q1/Q3 | Precursor and product ion m/z. |
| Retention time (RT) | Retention time. |
| Relative RT | Retention time divided by the internal standard retention time. |
| Subtracted RT | Retention time subtracted by the internal standard retention time. |
| Relative area | Peak area divided by the internal standard peak area. |
| Relative height | Peak height divided by the internal standard peak height. |
| Full width half max | Peak width at half the maximum height. |
| Asymmetry factor | The quotient between the back slope to center line width at 10% maximum height and front slope to center line width at 10% maximum height. |
| Tailing factor | The distance from the front to back slope of the peak at 5% maximum height divided by twice the distance from the center line of the peak to the front slope. |

backbones, and glycerophospholipids, targeting glycerophosphocholines.

## 3.5 Methods

### 3.5.1 Algorithm

Let $P_1, P_2, \ldots, P_m$ represent a list of SRM or MRM peaks within a sample with biological condition $C$. This condition $C$ could encompass a disease state, treatment condition, organism, or matrix. Peak $i$ is detected at parent ion $q_i^1$ and product ion $q_i^3$, and is described by feature vectors $F_i = (f_{i1}, f_{i2}, \ldots, f_{ik})$. All feature vectors have the same $k$ length, where a combination of up to eight features were explored describing peak intensity, shape, and retention time. Some features were normalized to a single

internal standard to account for LC-ESI-MS/MS variability across different acquisitions (details in Table 3.1). All features except subtracted retention time are exported by the MultiQuant software package by default.

The objective is to assign lipid identities to all peaks, including a possible "unassigned" identity if a peak cannot be confidently identified. The set of all lipid identities, also referred to as barcodes, is $I = \{B_1, B_2, \ldots, B_n, U\}$, where $U$ is the unassigned identity. The true identity of each peak is unknown but can be represented as variables $I_1, I_2, \ldots, I_m$ with lipid identities drawn from set $I$. Using these variables, two Bayesian approaches were formulated to address SRM/MRM peak identification. The individual identification method estimates the posterior probability of assigning barcodes to individual peaks, while the joint identification method computes the posterior probability of assigning barcodes to all $m$ peaks in a sample. While computing the posterior probabilities is shown to be nearly identical in both approaches, they differ in the decision rule used to assign barcodes to peaks.

**Individual lipid identification model**

The individual lipid identification method assigns lipid barcodes by considering the individual posterior probability of identifying peak $i$ belonging to condition $C$ from all candidate lipid identities monitored at the same $q_i^1 \pm \tau$ and $q_i^3 \pm \tau$ (where $\tau$ is the machine tolerance). Using Bayes's theorem, this posterior probability is written as

$$\Pr(I_i|F_i, C) = \frac{\Pr(I_i|C) \cdot \Pr(F_i|I_i, C)}{\Pr(F_i|C)}. \tag{3.1}$$

Under the naïve Bayes assumption that all input features are conditionally independent given the lipid identity:

$$\Pr(I_i|F_i, C) = \frac{\Pr(I_i|C) \cdot \prod_{j=1}^{k} \Pr(f_{ij}|I_i, C)}{\Pr(f_{i1}, f_{i2}, \ldots, f_{ik})}. \tag{3.2}$$

The naïve Bayes assumption of the 8 features in Table 3.1 is incorrect for many features, but greatly simplifies computing the likelihood term. Relative peak height and area, and retention time features are correlated,

for example. Other features are likely uncorrelated such as retention time and asymmetry or tailing factor. Estimating these likelihoods and priors is discussed in subsequent sections because the same methods are used for the individual and joint lipid identification models. The marginal likelihood can be omitted from Equation 3.2 because it is constant for every candidate barcode of peak $i$ and does not change the barcode ranks for a given peak. Thus, the final posterior probability of assigning a barcode is

$$\Pr(I_i|F_i, C) \propto \Pr(I_i|C) \cdot \prod_{j=1}^{k} \Pr(f_{ij}|I_i, C). \tag{3.3}$$

As previously stated, it is common for multiple lipid species to be detected at the same transition. According to the *maximum a posteriori* (MAP) decision rule, the most probable lipid assignment $\hat{y}$ is achieved by selecting the candidate barcode with the greatest posterior probability:

$$\hat{y} = \arg\max_{i \in I} \ \Pr(I_i|C) \cdot \prod_{j=1}^{k} \Pr(f_{ij}|I_i, C) \tag{3.4}$$

The main disadvantage with this individual lipid identification model using the MAP decision rule is that the same barcode can be assigned to two or more peaks within a sample. Experimentally, this is impossible as resolvable peaks must correspond to unique lipid species, otherwise they would elute at the same time. A modified MAP decision rule was implemented to overcome this problem. Constrained MAP (Algorithm 2) ranks candidate peak assignments, peaks paired with a single barcode, by order of posterior probability. Barcodes are then assigned to the peaks with the greatest posterior probability such that no peak is assigned more than one barcode and all barcodes, except the unassigned barcode $U$, are assigned no more than once per sample. This algorithm is greedy, however, because assigning the barcodes by ranked order of decreasing posterior probability is not guaranteed to maximize the posterior probability of all peak assignments. Assigning a peak to a barcode with the greatest posterior probability, for example, may result in assigning an unlikely barcode to another peak with

a very small posterior probability. This limitation is the main motivation for introducing the joint lipid identification model which maximizes the posterior probability of simultaneously identifying all peaks.

**Algorithm 2:** Constrained *maximum a posteriori*

**Input:** List of candidate peak assignments in a sample

**Output:** List of lipid identities assigned to each peak

Order candidate peak assignment list from greatest to smallest posterior probabilities

**for** *i in 1:length(assignments)* **do**

  **if** *Peak and identity not assigned* **then**

    Assign lipid identity to peak

  **else**

    Assign unassigned identity $U$ to peak

  **end**

**end**

### Joint lipid identification model

Instead of computing the posterior probability of observing an individual lipid barcode, lipid identification can be viewed as a joint posterior probability optimization problem. The objective is to maximize the posterior probability of identifying all peaks within a sample. This joint posterior probability is written as

$$\begin{aligned}&\Pr(I_1, I_2, \ldots, I_m | F_1, F_2, \ldots, F_m, C) \\ &\quad = \frac{\Pr(F_1, F_2, \ldots, F_m | I_1, I_2, \ldots, I_m, C) \cdot \Pr(I_1, I_2, \ldots, I_m | C)}{\Pr(F_1, F_2, \ldots, F_m | C)}.\end{aligned} \tag{3.5}$$

Unlike Equation 3.3, the marginal likelihood term cannot be ignored because it depends on the feature values of each peak. It will be shown later that this marginal term is actually constant with respect to the lipid identities $I_i$. For now, the denominator can be evaluated by summing over all possible joint lipid assignments using the law of total probability.

$$\begin{aligned}\Pr(F_1, F_2, \ldots, F_m|C) &= \sum_{I_1^{'}, I_2^{'}, \ldots, I_m^{'} \in I} \Pr(I_1^{'}, I_2^{'}, \ldots, I_m^{'}, F_1, F_2, \ldots, F_m|C) \\ &= \sum_{I_1^{'}, I_2^{'}, \ldots, I_m^{'} \in I} \Pr(F_1^{'}, F_2^{'}, \ldots, F_m^{'}|I_1^{'}, I_2^{'}, \ldots, I_m^{'}, C) \\ &\quad \cdot \Pr(I_1^{'}, I_2^{'}, \ldots, I_m^{'}|C)\end{aligned} \tag{3.6}$$

This equation is substituted for the marginal likelihood in the main equation.

$$= \frac{\Pr(F_1, F_2, \ldots, F_m|I_1, I_2, \ldots, I_m, C) \cdot \Pr(I_1, I_2, \ldots, I_m|C)}{\sum_{I_1^{'}, I_2^{'}, \ldots, I_m^{'} \in I} \Pr(F_1^{'}, F_2^{'}, \ldots, F_m^{'}|I_1^{'}, I_2^{'}, \ldots, I_m^{'}, C) \cdot \Pr(I_1^{'}, I_2^{'}, \ldots, I_m^{'}|C)} \tag{3.7}$$

The prior probability of observing all lipid barcodes given the condition is assumed to be independent. The prior probability term is simplified to

$$\Pr(I_1, I_2, \ldots, I_m|C) = \prod_{i=1}^{m} \Pr(I_i|C). \tag{3.8}$$

While the assumption of prior independence is not correct, it us unclear how prior dependencies could be effectively modelled. Some lipid species are matrix-specific, yet their prior probabilities could change over time due to disease progression or treatment response. Samples may inadvertently contain mixed matrices such as brain tissue samples containing traces of blood. Priors could be estimated by $C$ if it encodes the sample matrix, but coinciding lipids could still occur within $C$. Some lipid species may co-occur due to related biochemical pathways involved in their synthesis. Conversely, other lipids may never co-occur because the pathway upregulating the synthesis of one lipid may inhibit the production of another. Matrix-specific $C$'s are later explored and the main equation is updated to

$$= \frac{\Pr(F_1, F_2, \ldots, F_m|I_1, I_2, \ldots, I_m, C) \cdot \prod_{i=1}^{m} \Pr(I_i|C)}{\sum_{I_1^{'}, I_2^{'}, \ldots, I_m^{'} \in I} \Pr(F_1, F_2, \ldots, F_m|I_1^{'}, I_2^{'}, \ldots, I_m^{'}, C) \cdot \prod_{i=1}^{m} \Pr(I_i^{'}|C)}. \tag{3.9}$$

Several simplifying assumptions are made to estimate the likelihood. The first is that the feature vectors for different peaks are statistically independent.

$$\Pr(F_1, F_2, \ldots, F_m | I_1, I_2, \ldots, I_m, C) = \prod_{i=1}^{m} \Pr(F_i | I_1, I_2, \ldots, I_m, C) \quad (3.10)$$

This assumption is not necessarily true because systematic retention time shifts may delay peak retention times by a consistent value. Peak intensities may also be correlated due to differing lipid extraction, ionization, or fragmentation efficiencies. Normalizing features to those of an internal standard may reduce these "batch effects," although there is some measurement error when spiking these standards across many samples. Another simplifying assumption is that the features of peak $i$ only depend on the true identity of peak $i$ and the condition.

$$\Pr(F_i | I_1, I_2, \ldots, I_m, C) = \Pr(F_i | I_i, C) \quad (3.11)$$

The final naïve Bayes assumption is that input features are independent of each other given the lipid identity and condition. Once again, this is not necessarily true although it could tested by determining whether features are significantly correlated.

$$\Pr(F_i | I_i, C) = \prod_{j=1}^{k} \Pr(f_{ij} | I_i, C) \quad (3.12)$$

These final simplifying assumptions lead to the final equation

$$\begin{aligned} &\Pr(I_1, I_2, \ldots, I_m | F_1, F_2, \ldots, F_m, C) \\ &\quad = \frac{\prod_{i=1}^{m} \prod_{j=1}^{k} \Pr(f_{ij} | I_i, C) \cdot \Pr(I_i | C)}{\sum_{I'_1, I'_2, \ldots, I'_m \in I} \prod_{i=1}^{m} \prod_{j=1}^{k} \Pr(f_{ij} | I'_i, C) \cdot \Pr(I'_i | C)}. \end{aligned} \quad (3.13)$$

Estimating the marginal likelihood in the denominator is tricky because it is a sum over all candidate barcodes. It would be problematic if the same barcode was assigned to two or peaks within a sample because this is

experimentally impossible and the joint probability should evaluate to zero. Instead of computing the marginal likelihood overall all feasible barcodes, likelihood estimation is avoided altogether by log transforming Equation 3.13.

$$\begin{aligned}
&\log \Pr(I_1, I_2, \ldots, I_m | F_1, F_2, \ldots, F_m, C) \\
&\quad = \log \frac{\prod_{i=1}^{m} \prod_{j=1}^{k} \Pr(f_{ij}|I_i, C) \cdot \Pr(I_i|C)}{\sum_{I_1', I_2', \ldots, I_m' \in I} \prod_{i=1}^{m} \prod_{j=1}^{k} \Pr(f_{ij}|I_i', C) \cdot \Pr(I_i', C)} \\
&\quad = \sum_{i=1}^{m} \log \left( \prod_{j=1}^{k} \Pr(f_{ij}|I_i, C) * \Pr(I_i|C) \right) \\
&\qquad - \log \sum_{I_1', I_2', \ldots, I_m' \in I} \prod_{i=1}^{m} \prod_{j=1}^{k} \Pr(f_{ij}|I_i', C) * \Pr(I_i'|C)
\end{aligned} \tag{3.14}$$

The first term in Equation 3.14 represents the logarithmic sum of all unnormalized posterior probabilities of observing $I_1, I_2, \ldots, I_m$. Maximizing this joint probability is equivalent to assigning the most probable lipid barcodes to all $m$ peaks in a sample. Furthermore, the second term is constant with respect to $I_i$ because it depends on the data and not the assigned lipid barcodes which are being optimized.

The BATL program maximizes the log posterior probability of all peak identities by generating a weighted bipartite graph for all peaks assigned to the same candidate barcodes based on matching $q_i^1 \pm \tau$ and $q_i^3 \pm \tau$. The weights of each graph are the log posterior probabilities of the peak assignments. Maximizing the joint posterior probabilities of all peak assignments in each graph is equivalent to solving a maximum weighted bipartite matching (MWBM) problem. BATL implements the push-relabel algorithm (Goldberg and Tarjan 1988) to enforce the MWBM decision rule to assign barcodes to peaks. An example peak assignment graph of the MWBM versus MAP and constrained MAP decision rule is presented in Figure 3.1.

### Prior estimation

In both the individual and joint lipid identification models, the prior probability of observing a given barcode is empirically estimated from a labelled

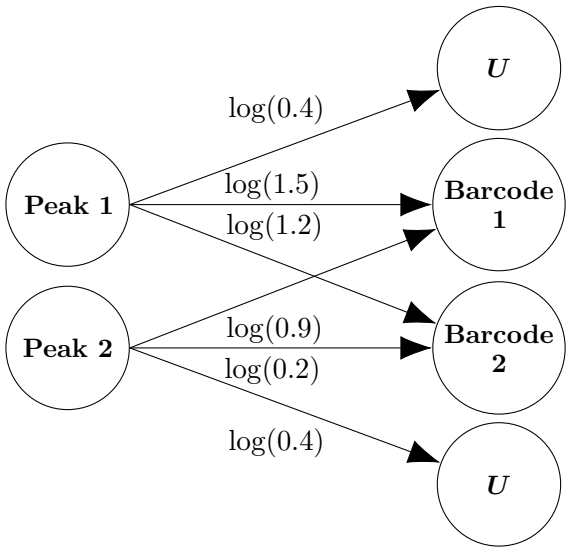



**Figure 3.1: An example illustrating peak assignments according to different decision rules.** MAP would assign both peak 1 and 2 to barcode 1. Constrained MAP would sequentially assign peak 1 to barcode 1 and peak 2 to $U$. MWBM would assign peak 1 to barcode 2 and peak 2 to barcode 1.

training dataset consisting of $l$ samples with pseudocount $\alpha$ set to 0 by default.

$$\Pr(I_i, C) = \frac{\sum_{i=1}^{l} \mathbb{I}(\Pr(I_i, C)) + \alpha}{l + \alpha} \tag{3.15}$$

**Likelihood estimation**

The likelihood of observing a feature value conditioned on the identity and condition is estimated using a normal or log-normal distribution. To determine which distribution best models the data, the mean and variance parameters of the normal and log-normal distribution are computed using the sample mean and variance of every lipid identity feature across all $l$ samples in the training dataset. To avoid taking the logarithm of a negative feature value, occurring whenever the internal standard elutes after a peak, a constant of 100 minutes was added to all subtracted retention times in the testing and training dataset. Addressing whether the normal/log-normal distribution is an appropriate estimator of the likelihoods is discussed in

Appendix 5.4.

For every identity feature, a Kolmogorov-Smirnov (KS) test is used to assess normality and log-normality at a $P < 0.05$ threshold. Let $P_{ij}$ represent the KS test $P$-value for normality of feature $j$ for lipid identity $I_i$. Let $P_{ij}^{\log}$ represent the KS test $P$-value for log-normality of feature $j$ for lipid identity $I_i$. The log-normal distribution is used for feature $j$ if the number of lipids for that feature is more log-normally than normally distributed according to the KS test. Otherwise, the normal distribution is used (Equation 3.16).

$$P(f_{ij}|I_i, C) = \begin{cases} \log \mathcal{N}(f_{ij}|\mu_{I_i}, \sigma_{I_i}, C), & \sum_{i=1}^{m} \mathbb{I}(P_{ij}^{\log} > 0.05) > \sum_{i=1}^{m} \mathbb{I}(P_{ij} > 0.05) \\ \mathcal{N}(f_{ij}|\mu_{I_i}, \sigma_{I_i}, C), & \text{otherwise.} \end{cases} \tag{3.16}$$

**Posterior probability of the unassigned identity** $U$

There are two general cases where the true identity of peak $i$ is not contained in the labelled training dataset. These picked peaks represent either background noise or novel lipid species that are absent from the set of barcodes $I$. If these peaks are detected at the same transition as known lipid identities in the training dataset, any peak assignment would be incorrect. To learn when not to assign a barcode, it is first assumed that barcode $U$ can be assigned to any peak and more than once per sample. It is also assumed that the posterior probability of an incorrect peak-barcode assignment is less than the smallest posterior probability of a correct peak-barcode assignment within that transition.

The posterior probabilities of $U$ are learned from the training dataset by estimating the minimum posterior probability cutoffs per transition by cross validation minus a small constant $c = 10^{-10}$. This approach thus assigns $U$ to a peak if all other candidate peak assignments seem implausible due to low posterior probabilities.

### 3.5.2 Data

Two sphingolipid and glycerophospholipid datasets consisting of 440 and 499 sample runs, respectively, were equally stratified by acquisition date into training and holdout datasets at random. The exact biological conditions, number of samples/acquisitions, organisms, and matrices of each dataset are described in Appendix 5.4. 338 and 541 transitions were monitored in the sphingolipid and glycerophosphocholine datasets, respectively. These datasets are a curated subset of the datasets used in Chapter 2, and all peaks were previously validated by SRM-IDA-EPI to provide ground truth labels. For secondary validation of the classifier, four statistically-independent testing datasets were selected, profiling sphingolipid and glycerophosphocholine blood and brain matrices; the lipid identities in these datasets were also validated by SRM-IDA-EPI. Both training/holdout and testing datasets were selected to represent a diversity of profiled organisms, tissues, and biological conditions; the exact biological conditions, number of samples/acquisitions and matrices are also described in Appendix 5.4. Each sample included two internal standards, Cer(d18:1/16:0-D31) and LPC(13:0/0:0). All data were acquired using Agilent 1290 Infinity II liquid chromatography system coupled via electrospray ionization to SCIEX QTRAP 5500 tandem mass spectrometer operating in positive ion mode as described by Granger et al. (2019) and Xu et al. (2019).

## 3.6 Results

### 3.6.1 Metrics

The following metrics were used to evaluate classifier performance.

$$\text{Identification accuracy} = \frac{\#\text{ correctly assigned peaks}}{\#\text{ assigned peaks}} \tag{3.17a}$$

$$\text{Unassignment rate} = \frac{\#\text{ unassigned peaks}}{\#\text{ detected peaks}} \tag{3.17b}$$

$$\text{Unassignment accuracy} = \frac{\#\text{ correctly unassigned peaks}}{\#\text{ unassigned peaks}} \tag{3.17c}$$

Of these three metrics, identification accuracy is the most important measure of classifier performance; however, because identification accuracy does not consider unassigned peaks, the unassignment rate and unassignment accuracy were also assessed. A correctly unassigned peak is defined as a peak that is not assigned a barcode by BATL and whose ground truth identity is not present in the training dataset. These three metrics were used to evaluate the calibration of BATL's posterior probability cutoffs.

### 3.6.2 Decision rule and feature selection

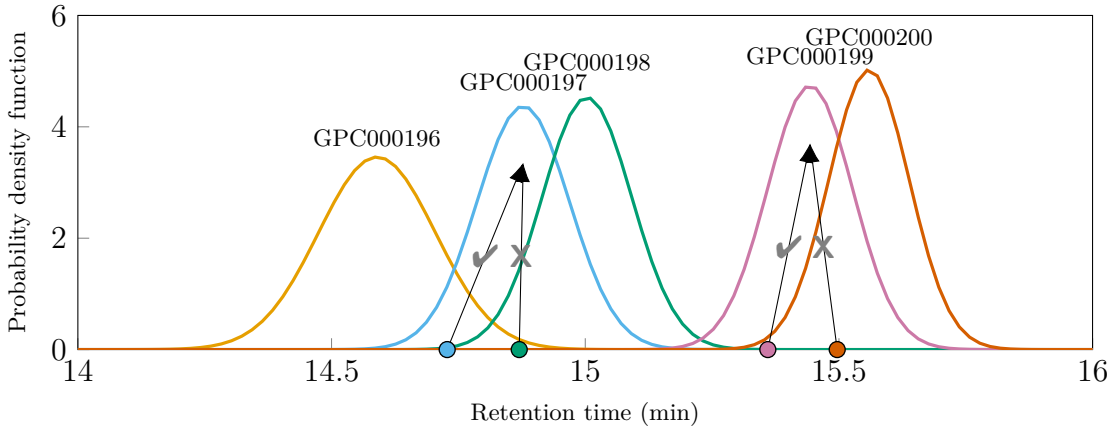
It is unclear which decision rule performs best nor which features are the most important for peak identification. Both aspects were simultaneously explored by generating models using every unique combination of features described in Table 3.1 with all decision rules. These models were evaluated by ten-fold cross validation of the sphingolipid and glycerophospholipid training datasets across blood and brain matrices. Identification accuracies for single feature models, such as retention time, showed that the MWBM decision rule performed equally or significantly better than either MAP or constrained MAP decision rules (Figure 3.3). This finding was consistent across all 255 feature combination models. Identification accuracies were generally poorer on the glycerophospholipid dataset where each peak could

match to up to eight lipid identities versus four in the sphingolipid dataset. Although constrained MAP was expected to outperform regular MAP by prohibiting duplicate assignments, worse identification accuracies were observed using constrained MAP, especially on the glycerophospholipid dataset (Figure 3.3b). All of these results were also consistent with matrix-dependent ten-fold cross validation (Appendix 5.4).

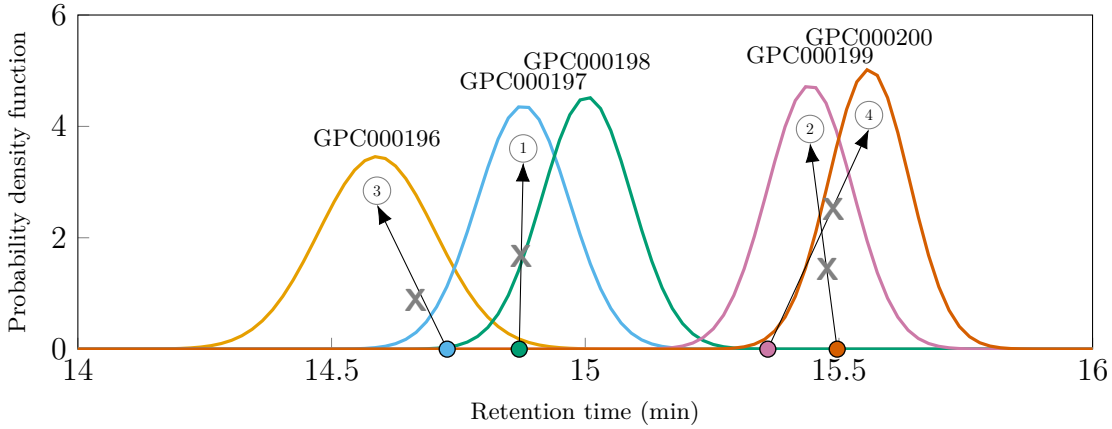
To understand why MWBM outperformed MAP and constrained MAP, identification errors were visualized using likelihood plots. An example plot is shown in Figure 3.2 where four glycerophospholipid peaks detected at transition 634.7/184.1 were matched to five candidate barcodes. It is evident that peak retention times were systematically shifted compared to the likelihoods of the true lipid identities. Due to the likelihood overlap between GPC000197/198 and GPC000199/200, MAP incorrectly assigned the same identity to two out of the four peaks (Figure 3.2a). Using constrained MAP, the first two peaks were incorrectly assigned, causing the subsequent misclassifications. When using the MWBM decision rule, all peaks were correctly assigned, even with these systematic feature shifts. These observations were consistent across all identification errors caused by the MAP and constrained MAP decision rule. Thus, all subsequent results were acquired using the joint lipid identification model employing the MWBM decision rule.

The eight peak features for SRM classification are broadly divided into three categories based on peak retention time, intensity and shape. Identification accuracies of models trained with all single features were compared to the model constructed with retention time because that is the most commonly used feature in manual peak identification. Unsurprisingly, models trained with the retention time features correctly identified more peaks than the intensity or shape features, which yielded the worst identification accuracies. Subtracted retention time performed significantly better than regular retention time on the sphingolipid dataset (Figure 3.4a), with no significant change on the glycerophospholipid dataset (Figure 3.4b). Visual-

**(a)** MAP




**(b)** Constrained MAP




**(c)** MWBM

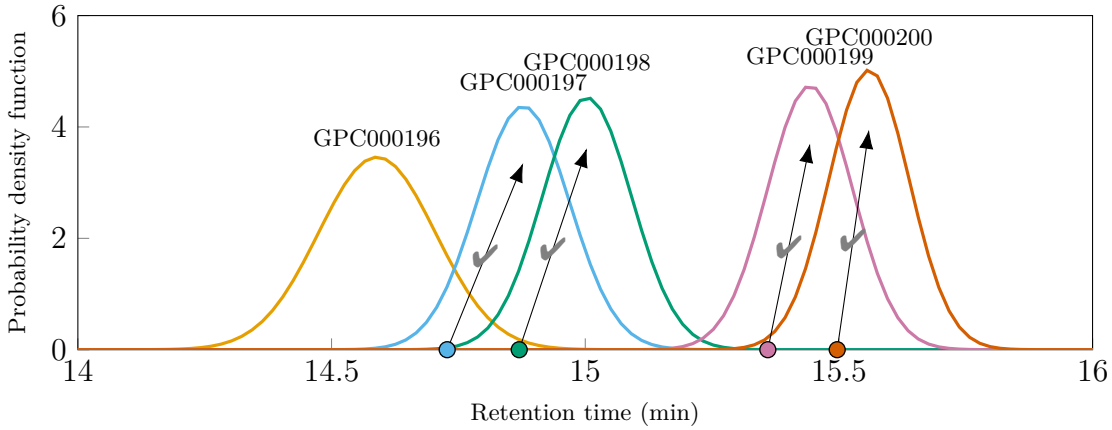


**Figure 3.2: MWBM decision rule outperforms MAP and constrained MAP.** Glycerophospholipid data from a single cross validation sample is shown for transition 634.6/184.1. Point markers correspond to peaks and colours denote the ground truth lipid identities. Checkmarks denote correct assignments; crosses denote incorrect assignments; $U$ denotes peaks that were not assigned a lipid identity. (a). MAP identification errors are often caused by assigning the same barcode to multiple peaks. (b). Initial misidentifications using constrained MAP cause successive errors. The first and second incorrect assignments cause the third and fourth incorrect assignments. (c). The MWBM decision rule correctly assigns all peaks.

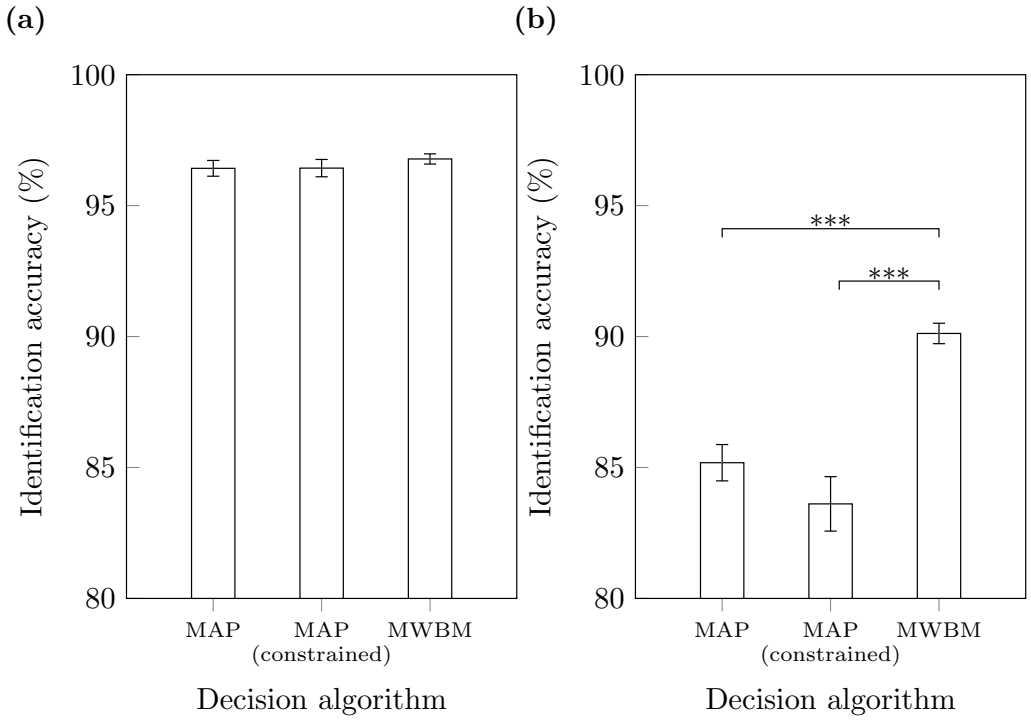



**Figure 3.3: MWBM is the best performing decision rule.** Ten-fold cross validation identification accuracies of naïve Bayes classifier using the retention time feature ($^{*}Q < 0.05$, $^{**}Q < 0.01$, $^{***} < 0.001$, pairwise $t$-tests adjusted with the Benjamini-Hochberg method). 95% confidence intervals are shown. (a) Sphingolipid dataset. (b) Glycerophospholipid dataset.

izing the sphingolipid likelihood plots confirmed that subtracted retention time effectively normalized against samples exhibiting systematic retention time shifts (Figure 3.5b). Without retention time normalization, samples that eluted quicker or slower than the Gaussian mean were more likely to be misclassified or not assigned a barcode (Figure 3.5a). In contrast with subtracted retention time, relative retention time warped peak retention times by dividing against the internal standard retention time. This normalization was shown to induce non-linear forward or backward shifting in the absence of systematic retention time shifts (Figure 3.6a). One explanation for this event is local shifting of the peaks and corresponding internal stand in opposing directions.

To determine whether a combination of features could improve targeted lipid identification, identification accuracies were plotted for models generated with the best combinations of $N$ feature for $N = (1, \ldots, 8)$ (Figure 3.7). Including three or more features increased sphingolipid and glycerophospholipid identification accuracies from 97.9% to 99.3% and 90.1% to 94.9%, respectively. The two most discriminatory features for both datasets included retention time and intensity features, indicating that peak area and height were not condition-specific. Beyond these two feature types, there was no discernible pattern in feature combinations that maximized identification accuracies. Unassignment rates were checked to ensure that increasing model complexity was not associated with improvements in identification accuracy (Figure 3.8). In fact, increasing model complexity significantly decreased unassignment rates across both datasets to less than 2%, partly explaining gains in lipid identification due to well-calibrated posteriors.

### 3.6.3 Holdout validation of classifier

Briefly, the optimal feature combinations maximizing identification accuracies were validated by holdout datasets described in Section 3.5.2. For this analysis, the entire holdout datasets were classified by BATL trained using the entire training dataset. Visually, similar identification accuracies were

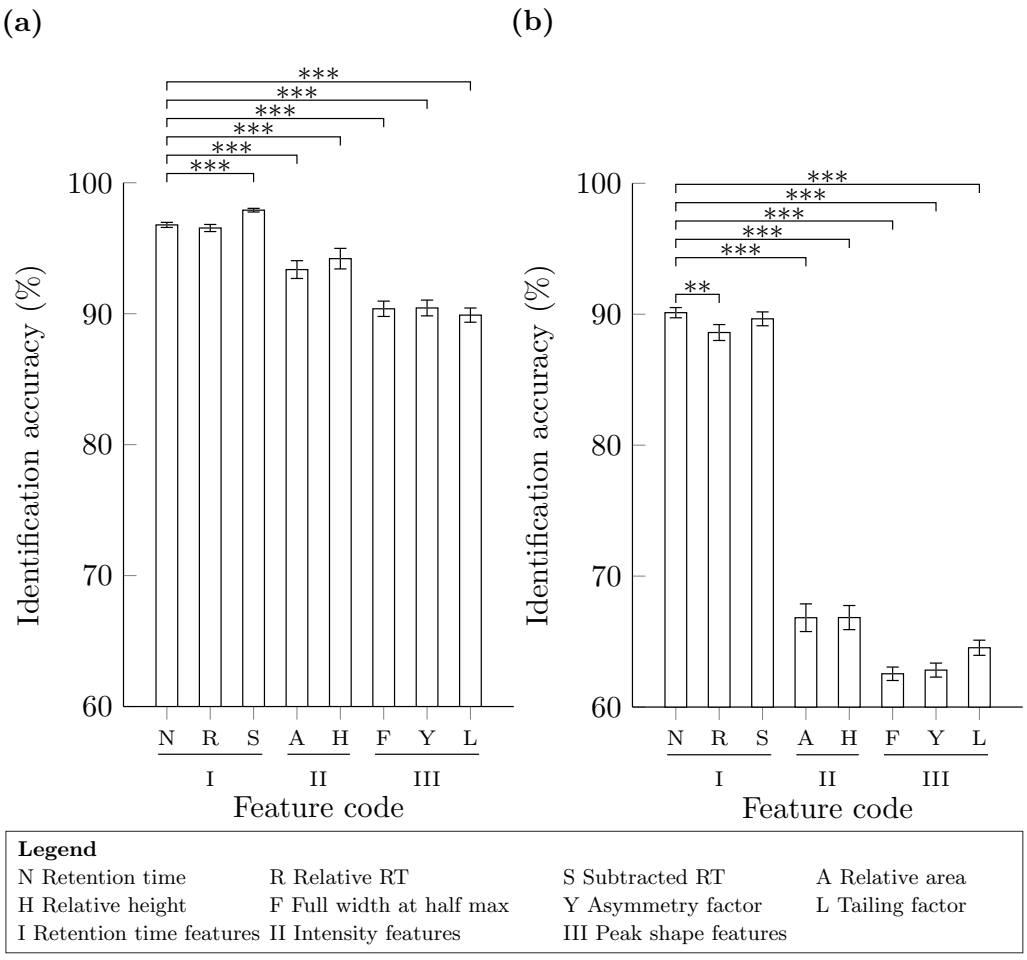



**Figure 3.4: Subtracted retention time is the most discriminatory MRM feature.** Ten-fold cross validation identification accuracies of naïve Bayes classifier trained on single features ($^{*}Q < 0.05$, $^{**}Q < 0.01$, $^{***} < 0.001$, pairwise $t$-tests against the retention time feature adjusted with the Benjamini-Hochberg method). 95% confidence intervals are shown. I indicates hydrophobicity features; II indicates intensity features; III indicates peak shape features. (a) Sphingolipid dataset. (b) Glycerophospholipid dataset.

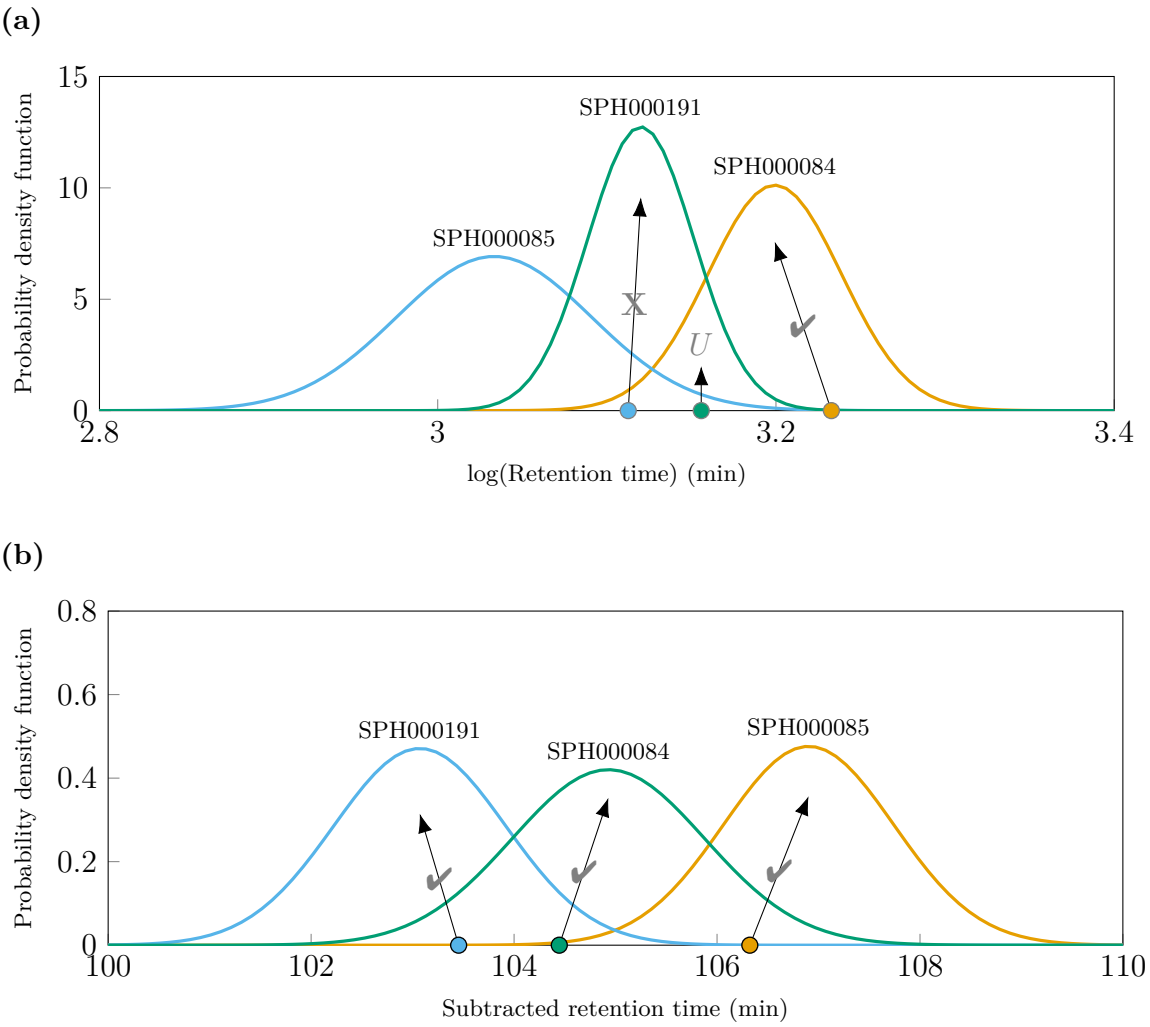
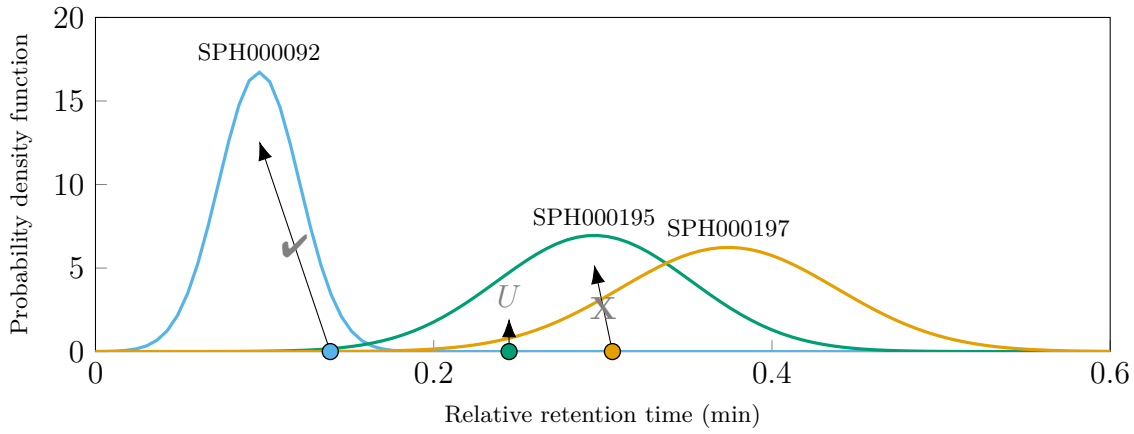
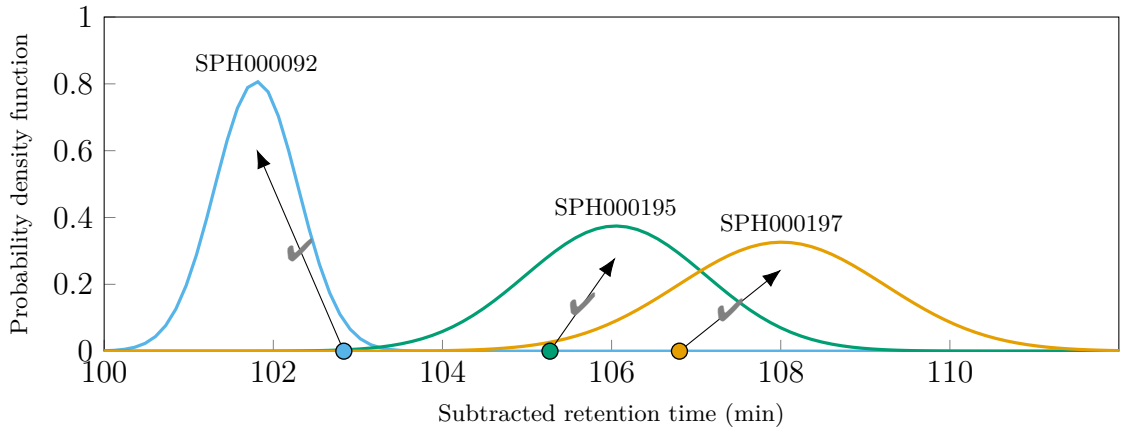



**Figure 3.5: Subtracted retention time accounts for systematic retention time shifts.** Point markers correspond to peaks and colours denote the ground truth lipid identities. Checkmarks denote correct assignments; crosses denote incorrect assignments; $U$ denotes peaks that were not assigned a lipid identity. (a) The retention times of the peaks are shifted forward, causing the blue peak to be misclassified to barcode SPH000191 and the green peak to be unassigned. (b) Correct classification of all three peaks with subtracted retention time.

**(a)**


SPH000092
SPH000195
SPH000197
U
X
Probability density function
Relative retention time (min)
0
5
10
15
20
0.2
0.4
0.6

**(b)**


SPH000092
SPH000195
SPH000197
Probability density function
Subtracted retention time (min)
0
0.2
0.4
0.6
0.8
1
100
102
104
106
108
110

**(c)**

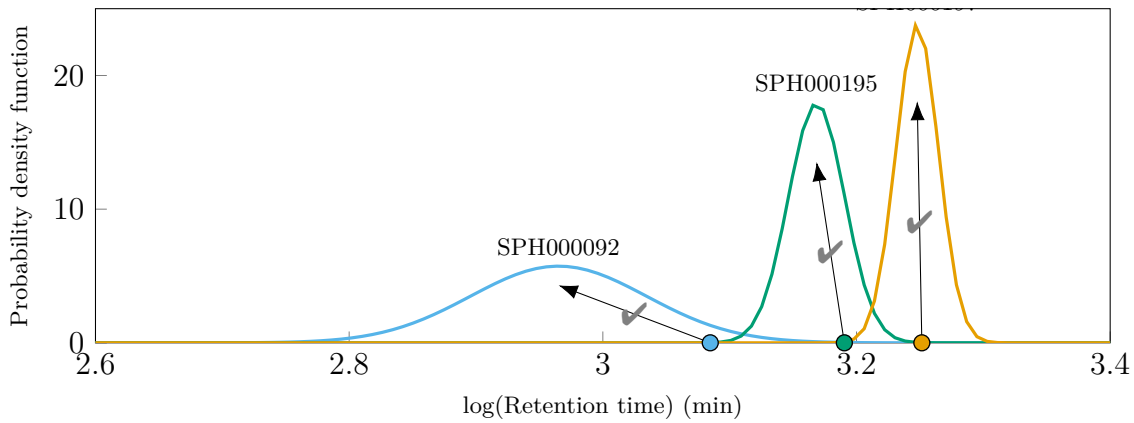

SPH000195
SPH000092
Probability density function
log(Retention time) (min)
0
10
20
2.6
2.8
3
3.2
3.4

**Figure 3.6: Relative retention time can induce systematic retention time shifts.** Point markers correspond to peaks and colours denote the ground truth lipid identities. Checkmarks denote correct assignments; crosses denote incorrect assignments; $U$ denotes peaks that were not assigned a lipid identity. (a) Relative retention time normalization causes a retention time shift. Incorrect classification of the green peak and the red peak is unassigned. (b) Correct classification of all three peaks with subtracted retention time. (c) Correct classification of all three peaks with regular retention time.

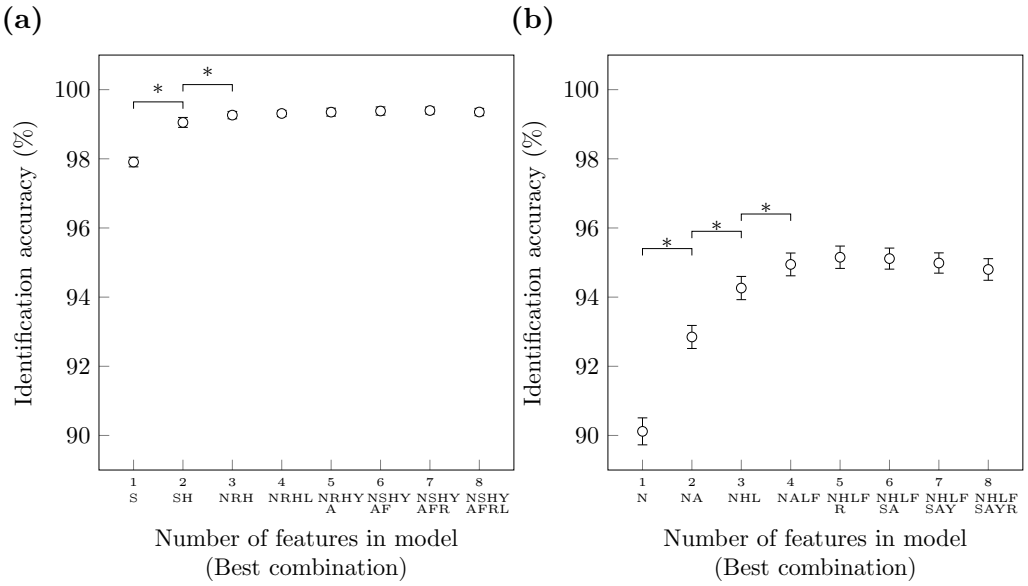



**Figure 3.7: Additional features significantly improve classifier identification accuracies.** Significant testing between the best combination of $N$ and $N+1$ features for $N = (1, \ldots, 7)$ ($t$-test, $^*P < 0.05$). Feature codes are defined in Figure 3.4. 95% confidence intervals are shown. (a) Sphingolipid dataset. (b) Glycerophospholipid dataset.

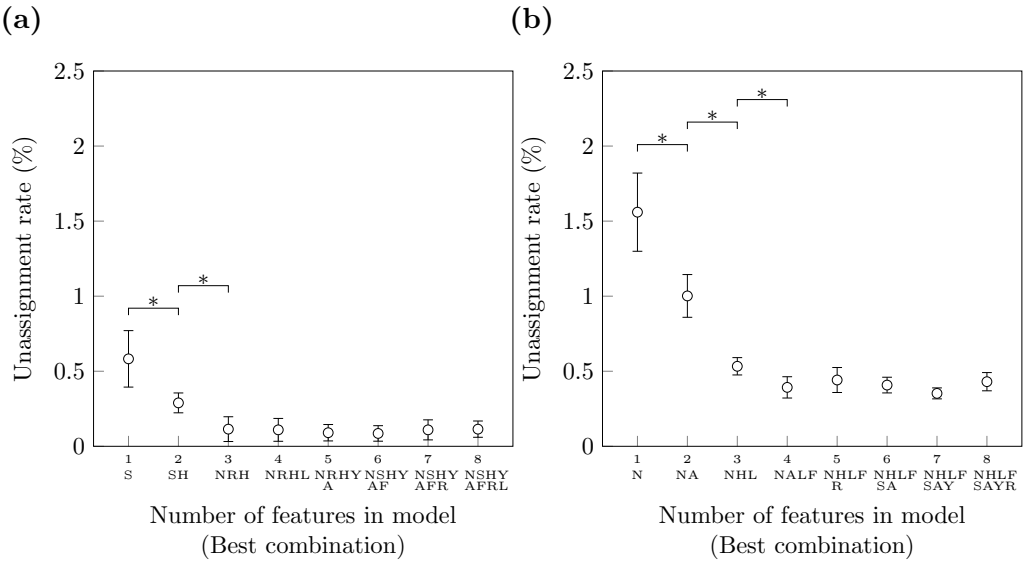



**Figure 3.8: Increasing model complexity improves cross validation identification accuracies by decreasing the proportion of unassigned peaks.** Significance testing only performed on significant feature combinations identified in Figure 3.7 ($t$-test, $^*P < 0.05$). Feature codes are defined in Figure 3.4. (a) Sphingolipid cross validation dataset. (b) Glycerophospholipid cross validation dataset.

observed for the sphingolipid dataset (Figure 3.9a), yet identification accuracies were consistently greater in the glycerophospholipid dataset by nearly 2% when one feature was used (Figure 3.9b). Unassignment accuracies were also similar between holdout and cross validation datasets with less than 1% unassigned peaks across all models (Figure 3.10). These results indicated appropriately fit models trained with every optimal combination of features.

Potentially, the reduction in training dataset size due to cross validation affected performance. To determine whether additional training samples improved identification accuracies, holdout datasets were re-evaluated with varying proportions of training data re-sampled ten times at 10% intervals without replacement (Figure 3.11). A significant difference in identification accuracy was observed in the sphingolipid holdout dataset from 40% to 30% training data used, although accuracies changed by less than 0.2%. Even though identification accuracies dropped by 0.9% between 100% and 10%

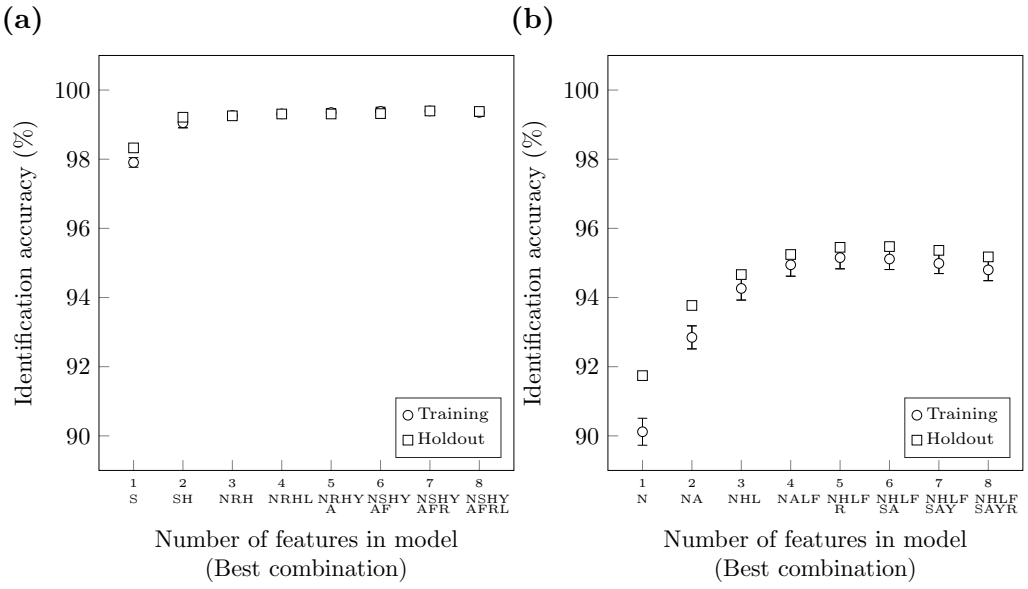



**Figure 3.9: Best feature combinations identified from cross validation are validated by holdout datasets.** 95% confidence intervals are shown for training only. Feature codes are defined in Figure 3.4. (a) Sphingolipid dataset. (b) Glycerophospholipid dataset.

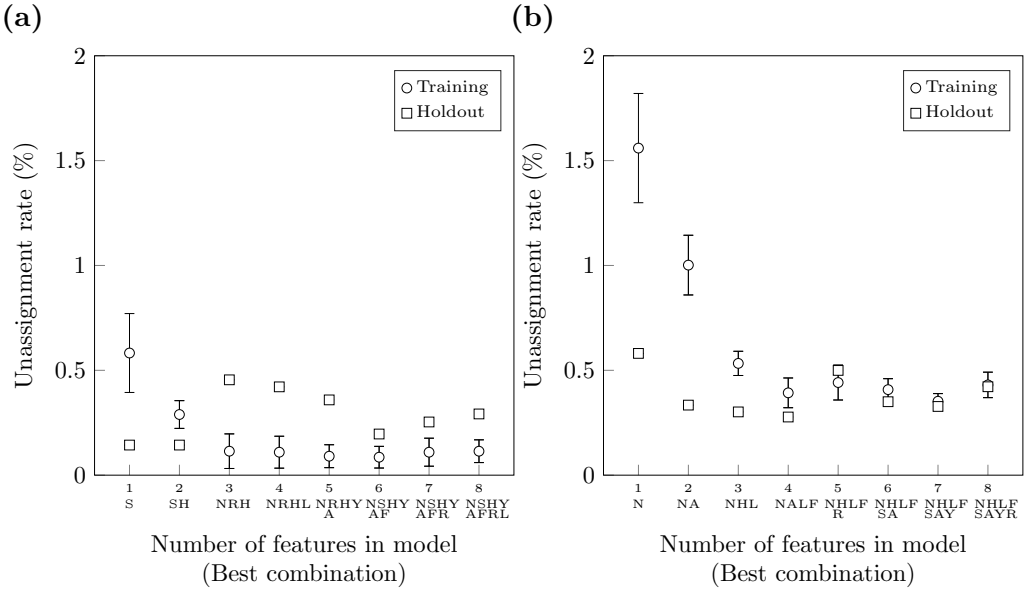



**Figure 3.10: Unassignment rates are are similar between holdout and cross validation datasets.** 95% confidence intervals are shown for training only. Feature codes are defined in Figure 3.4. (a) Sphingolipid dataset. (b) Glycerophospholipid dataset.

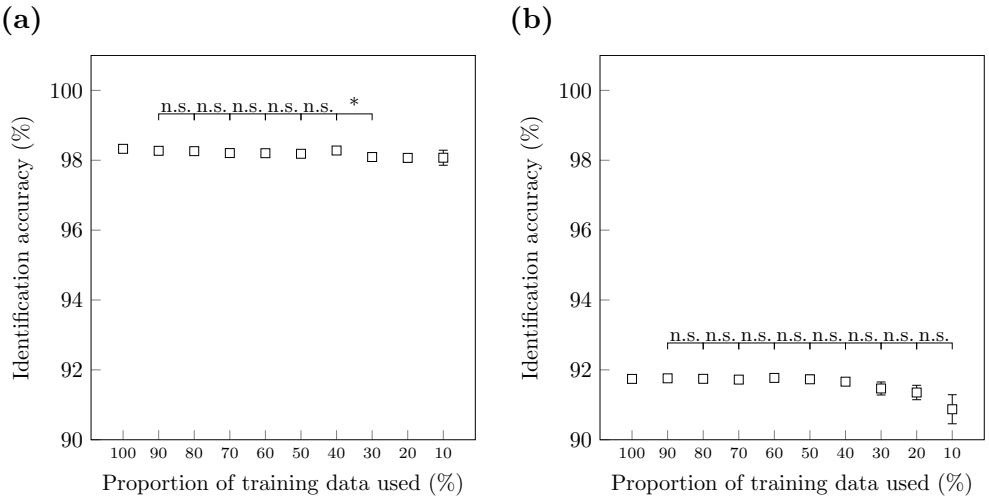



**Figure 3.11: Identification acuracies plateau when only 10% of the training data is used.** Training data was resampled ten times for each proportion. Top-down model selection was used to determine if training with 10% fewer samples significantly changed identification accuracies ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. (a) Sphingolipid dataset trained with subtracted retention time. Every 10% increment reflects 22 samples. (b) Glycerophospholipid dataset trained with retention time. Every 10% increment reflects 23 samples.

glycerophospholipid training data, no significant differences were observed in 10% training data increments. While these findings only partially explain the increased identification accuracies in the holdout dataset, they highlight that BATL is trainable on as few as 10% of the data, or 22-23 samples, with diminishing performance returns as further samples are added.

### 3.6.4 Predictive performance of classifier

Four statistically-independent sphingolipid and glycerophospholipid datasets, composed of samples derived from blood and brain matrices, were each classified ten times using the same training dataset partitions from ten-fold cross validation. Each model was trained using the optimal feature combinations identified in Figure 3.7. Following the holdout results, testing dataset identification accuracies positively correlated with increased model complexity.

Identification accuracies significantly increased in the sphingolipid blood dataset from 92.9% to 94.8% with the addition of the relative height feature. The use of all eight features-versus-one increased identification accuracies, with significant testing reported in subsequent sections. Compared to the cross validation and holdout analyses, sphingolipid and glycerophospholipid testing dataset identification accuracies decreased by 5-10% and 15-25%, respectively. Several factors may explain the increased classification error, especially on the glycerophospholipid testing datasets. The lipid-specific feature distributions of the testing dataset peaks may have changed due to possible batch effects from sample preparation and data acquisition. These batch effects are well-documented in mass spectrometry-based assays where time of day, reagent lot, and mass spectrometer calibration may alter lipid abundances and retention times. Different tissue dissection and lipid extraction operators, for example, may contribute to highly correlated lipid abundances within studies. The exact quantity of tissue procured may also depend on the tissue density which could correlate within subjects belonging to the same cohort. Any of these factors may alter peak retention times, intensities, and shapes in subsequent testing datasets, leading to poorly calibrated posterior probabilities and subsequent lipid misclassification.

Following identification accuracies, testing dataset unassignment rates were 10-15% greater than cross validation datasets (Figure 3.13). This finding was partly expected as some lipid species were only present in the testing dataset. Curiously, additional features increased unassignment rates by 10-15% in all but the sphingolipid brain dataset. The inclusion of relative and subtracted retention time, increased unassignment rates in the glycerophospholipid brain dataset, possibility indicating a batch effect propagated by retention time normalization. Unassignment accuracies also varied depending on the testing dataset and model complexity (Figure 3.14). The sphingolipid blood dataset unassignment accuracies, in particular, varied greatly between 65-97%, and the addition of relative and subtracted retention time in the glycerophospholipid brain dataset decreased unassign-

ment accuracies by 10%. These findings suggested that posterior probability cutoffs were sensitive to model complexity, and perhaps too stringent, as identification accuracies remained equal to or significantly greater than single feature models with diminishing unassignment rates and accuracies.

### 3.6.5 Finalizing the trained classifier

Previously, models for all combinations of features were explicitly computed and the best combination of $N$ features using the MWBM decision rule were determined by cross validation. These combinations of features were applied to the testing datasets, where increased model complexity correlated with increased identification accuracies in some datasets, but not others. Additional model features altered unassignment rates and accuracies in some testing datasets too, but these effects were not always correlated with changes in identification accuracies. The use of two-versus-one features in the glycerophospholipid brain dataset, for example, significantly improved identification accuracy with no significant changes in unassignment rates or accuracies. These findings were also consistent in matrix-dependent analyses, where increasing model complexity positively correlated with improved identification accuracies without reduced unassignment accuracies in the sphingolipid and glycerophospholipid brain datasets (Appendix 5.4). As identification accuracies remained the same or significantly increased with model complexity in matrix-agnostic and matrix-specific analyses, these results suggested that the use of all eight features would yield the best model for all future SRM peak identifications. Models with all eight features were compared to the least complex models with the greatest identification accuracy reported from the cross validation datasets. No significant differences in identification accuracy were observed on either sphingolipid or glycerophospholipid cross validation datasets (Figure 3.15). To check for model overfitting, these models were compared on the testing datasets with similar identification accuracies (3.15). Thus, an end user can expect identification accuracies as low as 73% on extraordinarily divergent

biological conditions and as high as 95% otherwise.

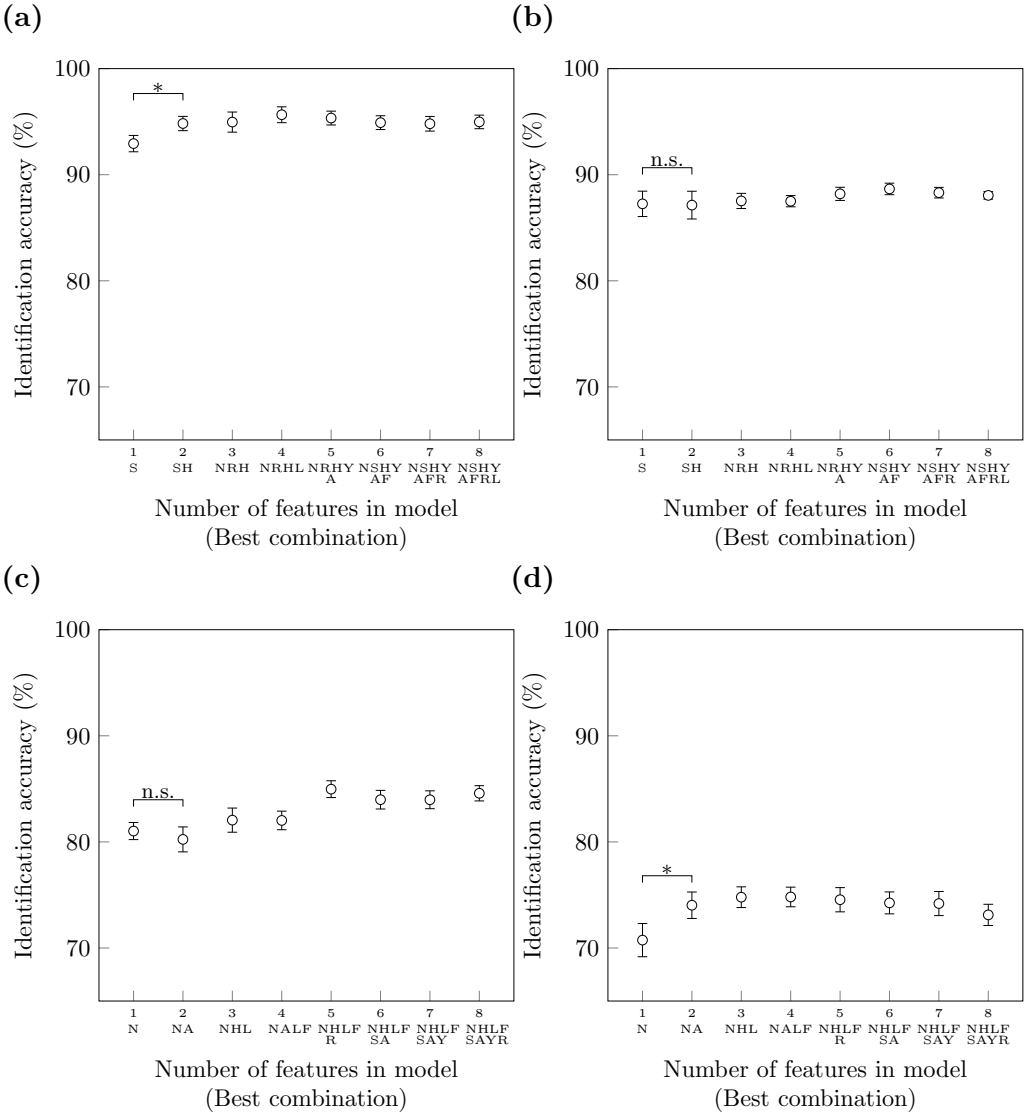



**Figure 3.12: Classification of statistically independent testing datasets.** Significance testing only performed on significant feature combinations identified in Figure 3.7 ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure 3.4. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.

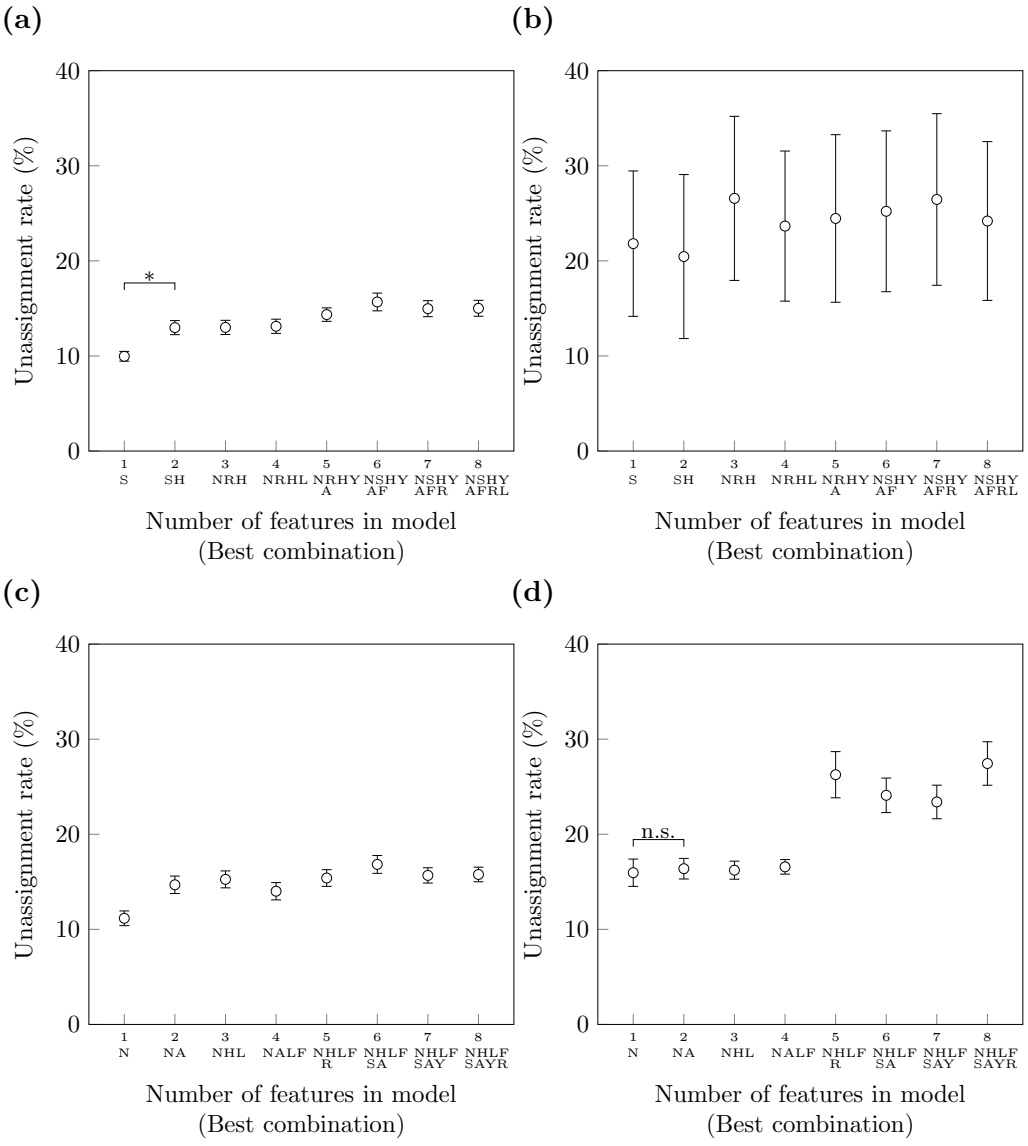



**Figure 3.13: Expanding model complexity correlates with increased unassignment rates.** Significance testing only performed on significant feature combinations identified in Figure 3.12 ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure 3.4. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.

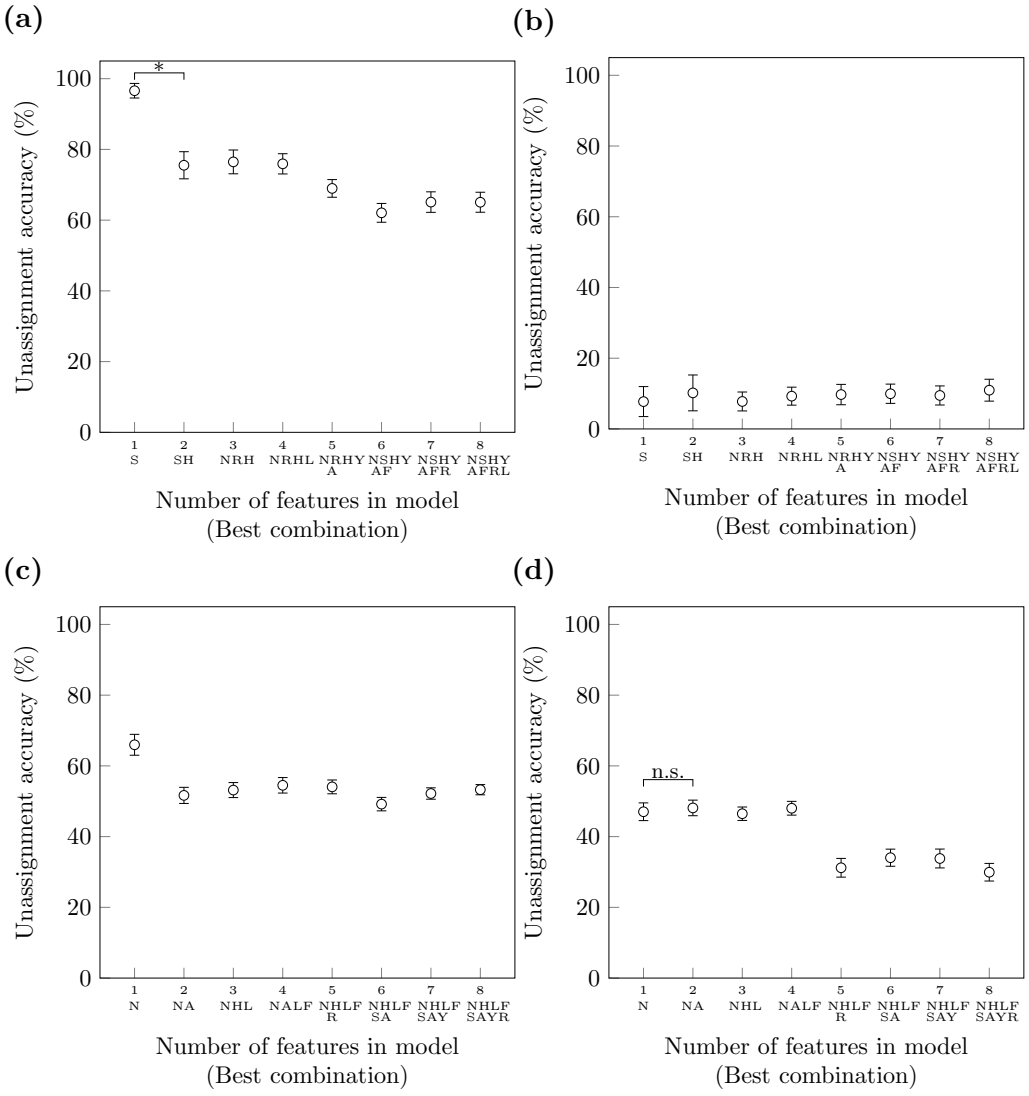



**Figure 3.14: Increasing model complexity improves classifier identification accuracies by correctly unassigning more peaks.** Significance testing only performed on significant feature combinations identified in Figure 3.12 ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure 3.4. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.

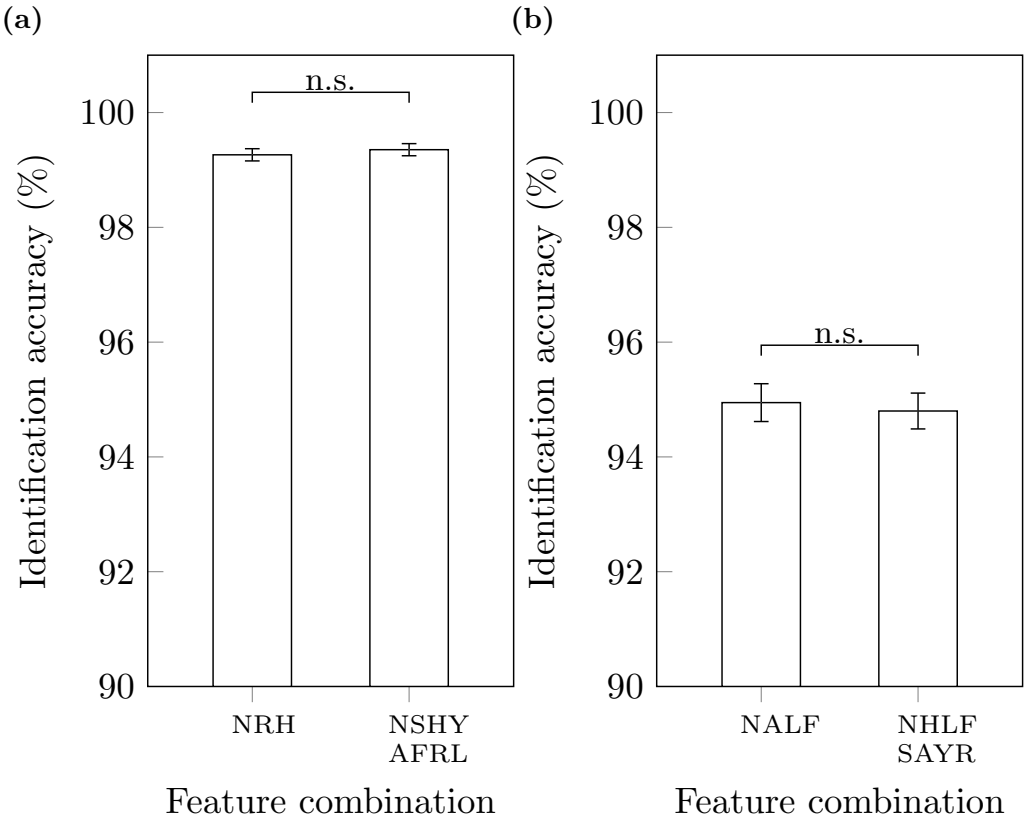



**Figure 3.15: Classifiers trained with all eight features display equivalent identification accuracies on cross validation datasets than the optimal, least complex model identified from cross validation.** 95% confidence intervals are shown ($t$-test, $^*P < 0.05$). Feature codes are defined in Figure 3.4. (a) Sphingolipid dataset. (b) Glycerophospholipid dataset.

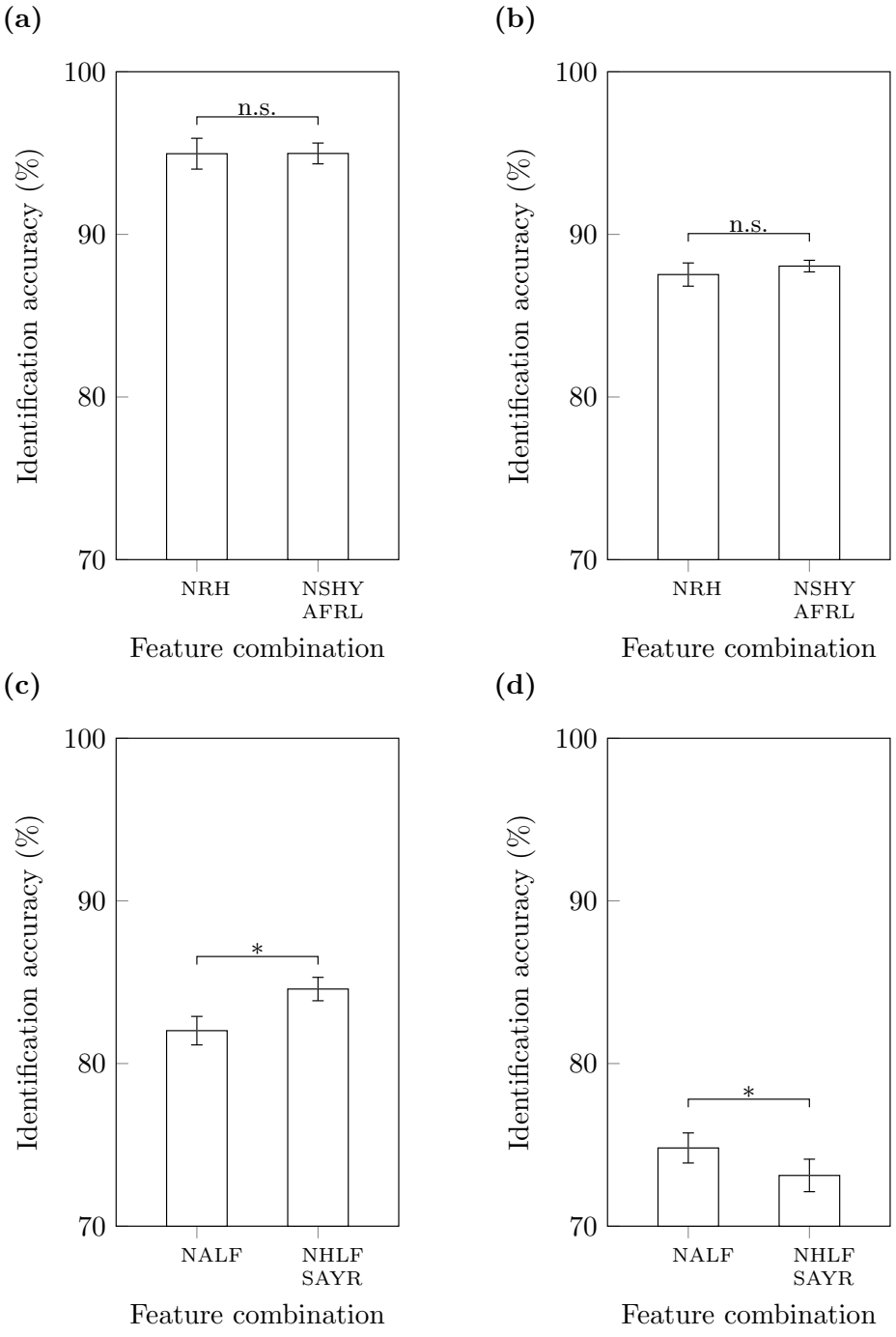



**Figure 3.16: Classifiers trained with all eight features display similar identification accuracies on testing datasets than the optimal, least complex model identified from cross validation.** 95% confidence intervals are shown ($t$-test, $^*P < 0.05$). Feature codes are defined in Figure 3.4. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.

## 3.7 Discussion

In this study, I introduced the first naïve Bayes classifier, BATL, for automated and accurate targeted lipid identification. By framing lipid identification as a joint posterior probability optimization problem, the MWBM decision was proven to be an optimal method for peak identification under certain assumption of feature and lipid independence. Using a variety of structurally-validated datasets, optimal lipid identification was empirically demonstrated on the cross validation datasets using the MWBM decision rule. As the joint lipid identification model was computationally straightforward, the most discriminative combinations of features were identified that maximized identification accuracy. Notably, it was shown that subtracted retention time, instead of regular or relative retention time, is the most important feature for SRM classification. Subtracted retention time effectively corrects against run-to-run LC-ESI-MS/MS variation in the presence of systematic retention time shifts. It was further shown that models using all eight features performed comparably to those generated with the least complex combination of features that maximized identification accuracy, and these findings were validated on holdout and testing datasets.

In the joint lipid identification model, several assumptions were proposed to simplify computing the posterior probabilities of peak-barcode assignments. The main assumptions included statistical independence of peak features, Gaussian likelihoods, independence of features conditioned on the lipid identity (naïve Bayes assumption), and independence of lipid priors. Future investigations remain to determine which assumptions contribute most to misidentification, and whether more sophisticated modelling could improve classification. By classifying variants of the same dataset with simulated "batch effects," the number of misidentifications could be correlated with the magnitude of systematic feature shifts. Artificially skewing these datasets may also establish whether the Gaussian distribution is a poor likelihood estimator versus non-parametric approaches, such as kernel density estimation. Modelling feature dependencies using Bayesian networks

may also improve classification, although it remains unclear if the training data and computation required to do so would significantly improve identification accuracies. Finally, although matrix-specific prior probabilities were investigated in Appendix 5.4, detailed analyses are required to determine if improvements in identification accuracy were due, in part, to reduced batch effects from the smaller training datasets.

A major problem for any targeted lipid identification software is the heterogeneity of sample pre-processing and data acquisition methods. Each laboratory may use different lipid extraction protocols, LC-ESI-MS/MS platforms, and machine parameters that are optimized for their own SRM and MRM panels. Thus, any laboratory interested in using BATL must generate and manually validate their own training datasets to construct the probabilistic model and learn the feature likelihoods and transition-specific posterior probability cutoffs. This obstacle is currently unavoidable in the lipidomics field until mass spectrometry platforms and protocols are standardized to enable cross-laboratory data comparisons (Liebisch, Ekroos, et al. 2017). To overcome this problem, BATL was designed to be vendor-agnostic and compatible with SRM/MRM peak lists (in text format) acquired from any LC-ESI-MS/MS platform. Training BATL with either $k$-fold or leave-one-out cross validation takes minutes, given the simplicity of the naïve Bayes model and the implementation of parallel computing to decrease training times. Note that BATL is applicable to other targeted lipidomics acquisition modes such as precursor ion scan, product ion scan, and neutral loss scan (data not shown). BATL also uses a naïve Bayes model which rapidly learns on as few as 22-23 samples, and has been shown to compete against more sophisticated and computationally expensive classifiers in other applications (Domingo-Almenara et al. 2018). Unlike complex classifiers, learning the posterior probability cutoffs for BATL is also feasible on consumer-grade hardware, and scalable with increasing training samples due extensive code vectorization. Once trained, BATL takes on the order of seconds to compute and assign lipid identities that

maximize the joint posterior probability, effectively eliminating the biases associated with manual lipid identification.

To assess real-world classifier performance, testing datasets profiling sphingolipid and glycerophospholipid datasets were chosen across different organisms, matrices, and biological conditions. These datasets were classified using matrix-agnostic training datasets with identification accuracies between 73-95% or greater using all eight features. These results are compelling because only the 17% and 36% of the sphingolipid and glycerophospholipid training datasets, respectively, consisted of samples profiling brain matrices. Nevertheless, identification accuracies of these brain matrix datasets ranged from 73-85%, even though the majority of training samples were derived from blood matrices. This finding suggests that BATL can overcome matrix-specific effects and accurately identify lipids across matrices. Excitingly, training BATL on all available biological matrices may also enable the unbiased identification of novel lipid biomarkers across mixed matrices and complex biological conditions.

## 3.8 Conclusion

I present the first open-source and platform-independent R package for targeted lipidomics identification. BATL is the first algorithm to rigorously address the SRM/MRM high-throughput identification problem, addressing the need to streamline, standardize, and improve the quality of targeted lipid identification between laboratories and across platforms.

# Chapter Four

# Discovery-driven targeted lipidomics

## 4.1 Objective

Targeted lipidomic approaches are often viewed as biased because data acquisitions are limited to the operator's lipids of interest, dictated by the transitions monitored in Q1 and Q3. My final thesis objective was to establish and assess an unbiased, discovery-driven targeted lipidomics workflow, thus enabling the identification of novel lipid species from SRM data. To accomplish this goal, I picked every peak from two samples randomly selected from each testing dataset described in Chapter 3. These peaks encompassed both known lipid identities, background noise, and potentially novel, undiscovered lipid species not "targeted" by other analyzers. Using my in-source lipid artifact annotation tool (Objective 1), I filtered out redundant lipid artifacts and isotopes, and subjected the remaining peaks to identification using my trained classifier BATL (Objective 2). Using this methodology, I demonstrate the potential of BATL for the unbiased analysis of targeted lipidomics data that facilitates the discovery of unknown lipid species across disease states, treatment conditions, and biological matrices.

## 4.2 Statement of author contributions

Justin G. Chitpin, Steffany A. L. Bennett, and Theodore J. Perkins conceived the study. Justin G. Chitpin developed the algorithms, analyzed the data, prepared the figures, and wrote the chapter with Steffany A. L. Bennett and Theodore J. Perkins; mass spectrometry datasets were prepared by Justin G. Chitpin and provided by Graeme P. Taylor, Hongbin Xu, Thao Nguyen, Irina Alecu, and Jayu Y. Gao; Graeme P. Taylor further provided expert consultation and advice, and structurally validated matrix-specific lipid identities by SRM-IDA-EPI.

## 4.3 Summary

To enable discovery-driven, targeted lipidomics, I generated four targeted lipidomic "discovery datasets," where all SRM peaks were picked encompassing known lipids, background noise, and possibly novel lipid species. These datasets were analyzed using my in-source lipid artifact and isotope annotation tool to reduce data redundancy, and my targeted lipidomics classifier BATL to assign lipid annotations to all peaks. Combining these programs, I present an unbiased, targeted lipidomics pipeline for the simultaneous identification and discovery of novel lipid species across mixed matrices, organisms, and conditions. When only evaluating BATL annotations that have been previously validated by SRM-IDA-EPI, I report conservative identification accuracies between 33.4-58.0%. However, when considering the validity of novel lipid species annotated by BATL, I report potential identification accuracies between 68.3-87.5%, corresponding to a 37.6-81.4% increase in discovered lipid species from sphingolipid (d18:1 sphingoid backbones) and glycerophosphocpholine discovery datasets.

## 4.4 Introduction

The primary goal of lipidomics is to identify and quantify the totality of lipid species within a biological system. Lipidome estimates vary within and across matrices (Sampaio et al. 2011; Quehenberger and Dennis 2011; Bozek et al. 2015; Khrameeva et al. 2018), and understanding the total number of biologically-occurring lipid species is both a fundamental scientific question, and necessary for the advancement of functional and clinical lipidomics (Gross et al. 2005; L. Zhang, Han, and Wang 2018). Currently, global lipid profiling studies are exclusively conducted using untargeted lipidomic approaches, which are considered "discovery-driven" in contrast to targeted lipidomic approaches. As discussed in Section 1.6, the term "discovery" is poorly defined in literature (Contrepois et al. 2018; Blaženović, Kind, Sa, et al. 2019). Discovery may be interpreted by analytical chemists as the identification of any analyte present in a single MS run. Bioanalytical chemists define discovery in the context of drug or biomarker discovery where differentially abundant lipids are identified across conditions. Other fields may define lipid discovery slightly differently, although the analytical chemistry definition is the most prevalent in literature. Yet, untargeted mass spectrometry data acquisition modes used by untargeted lipidomic approaches cannot meaningfully detect and identify all lipid species in a single run. Full scan, DDA, and DIA are biased towards identifying high abundance lipids, and the many experimental parameters chosen by the MS operator influence what lipid species are more likely to be detected. I thus asserted that "discovery" should be reinterpreted as *the capacity to meaningfully identify all acquired signals*, constrained by the limits of currently available MS-based technologies and bioinformatics tools. Using this definition, targeted lipidomics is currently a biased approach to quantify known lipid species because no bioinformatics algorithms exist to identify low and high abundance lipids within a given family at the positional isomer level from targeted lipidomics data. Conversely, untargeted lipidomic approaches are discovery-driven because many computational algorithms have been

developed to exploit this data to identify all high abundance lipids across many different lipid families.

Until presently, no computational tools have been developed specifically for the annotation of in-source artifacts and isotopes from targeted lipidomics data, nor the unbiased identification of SRM or MRM peaks. With the development of the in-source lipid artifact annotation tool (Chapter 2) and BATL (Chapter 3), unbiased, targeted lipidomic analyses are now possible by picking peaks, annotating possible artifacts and isotopes, and assigning lipid identities using BATL. To remove any biases associated with expert operators picking certain peaks, but not others, we propose a discovery-driven, targeted lipidomics analysis, where all detected peaks in a given dataset are selected for downstream identification. By detecting all SRM signals, this analysis enables the potential identification of all lipid isomers in samples, where novel lipids can be flagged as unassigned for future structural identification.

To demonstrate that novel lipid species can be "discovered" in targeted lipidomic datasets, I developed a proof-of-concept R pipeline that integrated my in-source lipid artifact annotation tool and targeted lipid identification classifier BATL. The testing datasets for this study consisted of sphingolipid and glycerophosphocholine samples containing every detectable peak corresponding to genuine lipid species or background noise. Using this pipeline, I demonstrate the robustness of BATL to effectively annotate and identify peaks with lipid identities previously validated by SRM-IDA-EPI. Peaks representing background noise or possibly undiscovered lipid species absent from the training dataset were designated as unlabelled peaks, enabling users to interrogate their identities in the future. While identification accuracies were lower in datasets containing all detected peaks, as opposed to datasets from expert operator-picked peaks, hundreds of novel lipid species were reported by BATL and thousands flagged as unassigned, where they could be prioritized for structural interrogation by SRM-IDA-EPI. Thus, I highlight the potential of this targeted lipidomics analysis for the unbiased

discovery of lipid species across mixed matrices, treatment conditions, and disease states.

## 4.5 Methods

### 4.5.1 Data

Two biological samples were randomly selected from each of the four sphingolipid and glycerophosphocholine testing datasets described in Chapter 3. These datasets profiled unique treatment conditions within blood and brain matrices, and the conditions are described in Appendix 5.4. The terms "original dataset" and "discovery dataset" are used to refer to the representative, testing dataset samples containing only known lipid species validated by SRM-IDA-EPI and those containing all detected peaks, respectively. All discovery dataset peaks were picked in MultiQuant (SCIEX) meeting a signal to noise ratio criterion equal or greater than ten with Gaussian smoothing and peak splitting parameters set to zero and one, respectively. These peak integration settings were specifically chosen to objectively detect all lipid signals and prevent any peak-picking biases attributed to peak integration settings.

### 4.5.2 Data analysis pipeline using BATL

All analyzed peaks in the four discovery datasets were exported as Qsession output files using MultiQuant. Discovery dataset peaks representing known lipid species previously validated by SRM-IDA-EPI were labelled with the appropriate lipid identities from the corresponding, original testing dataset samples using a 15 second retention time window. This step was necessary because some original dataset peaks were manually integrated with different Gaussian smoothing and other peak picking parameters which slightly altered their retention times. The retention time window thus enabled the correct matching of peaks in the discovery dataset to the

lipid barcodes in the original dataset. The terms "previously identified peaks" and "novel peaks" are used to refer to peaks with validated lipid identities and those with unknown identities, respectively. Lipid artifacts and isotopes were detected using the in-source lipid artifact annotation tool integrated into BATL, using a machine tolerance of 0.5 m/z units and retention time matching to two significant figures after the decimal place. All in-source artifacts and lipids were filtered out of the discovery datasets and the remaining peaks were classified using BATL. BATL was trained with all eight peak features previously described in Table 3.1, using the same matrix-independent sphingolipid and glycerophosphocholine training dataset cross-validation folds from Chapter 3. The training dataset matrices and disease state/treatment conditions are described in Appendix 5.4. The machine tolerance for peak-barcode matching was set to 0.5 m/z; the prior pseudocount was set to zero and the MWBM decision rule employed. All statistical analyses were performed using R (version 3.6.3).

## 4.6 Results

### 4.6.1 Annotating lipid artifacts and inferring lipid families

The sphingolipid blood and brain discovery datasets comprised 1,402 and 2,534 previously identified and novel peaks. A total of 5,473 and 7,781 peaks were present in the glycerophosphocholine blood and brain discovery datasets. Using the in-source lipid artifact and isotope annotation program integrated into BATL, 1,438 artifactual/isotopic peaks were detected across all four discovery datasets, representing 7-9% of the starting peaks in each dataset. These annotations included isotopes and sphingolipid-specific dehydrations, deglycosylations, and dimers. In addition to single artifact/isotope annotations, several dozen unique combinations of lipid annotations were identified (e.g. +1 isotope and -1 dehydration of a mono-isotopic peak with

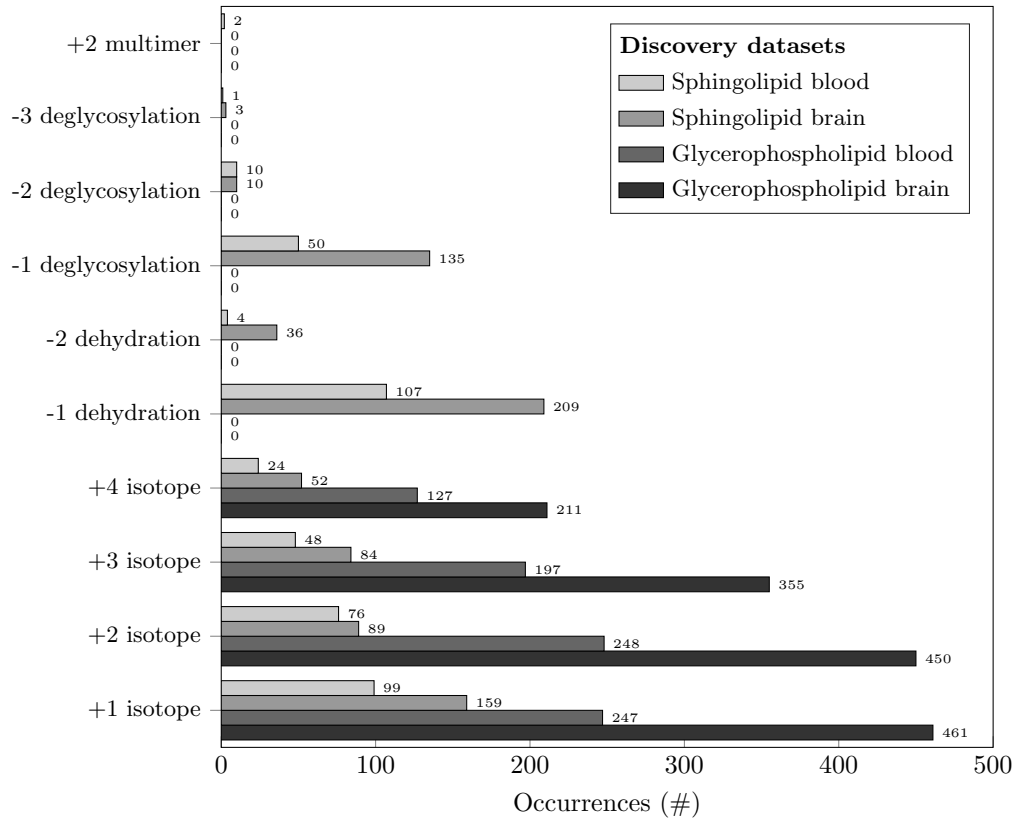



**Figure 4.1: Lipid artifacts and isotopes identified from all discovery datasets using the in-source lipid artifact annotation program.**

**Table 4.1: Annotation depth per identified lipid artifact/isotope.**

| | # of peaks with one or more annotations | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Discovery datasets | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Sphingolipid blood | 124 | 92 | 37 | 15 | 8 | 3 | 2 | 2 | 1 | 1 |
| Sphingolipid brain | 234 | 170 | 50 | 16 | 7 | 3 | 2 | 1 | - | - |
| Glycerophospholipid blood | 397 | 281 | 90 | 36 | 11 | 4 | - | - | - | - |
| Glycerophospholipid brain | 683 | 504 | 184 | 60 | 26 | 12 | 8 | - | - | - |

a given parent ion). For visual clarity, all types of annotations are presented in Figure 4.1, and the number of peaks assigned multiple annotations are presented in Table 4.1.

Due to the number of discovery dataset peaks, there was concern that

some assigned lipid annotations were false positives. This limitation was discussed in Chapter 2 as annotations are assigned based on machine and retention time tolerance cutoffs. Thus, it was possible that two or more peaks corresponding to different molecular identities randomly eluted at the same retention time with parent ions that coincidentally matched a particular annotation or compound annotation. One method considered to validate these annotations was SRM-IDA-EPI; however, it was unfeasible in this study to analyze all IDA-EPI spectra for the 1,438 artifactual and isotopic peaks. Some assigned annotations indicated the presence of lipid identities absent from the original datasets. Several peaks were assigned triple deglycosylations in the sphingolipid brain dataset, suggesting the presence of globotriaosylceramides. These lipids are characterized by three sugars linked to the sphingosine backbone. Globotriaosylceramides were absent from the training dataset, highlighting the possible utility of the annotation tool for database-independent lipid predictions.

### 4.6.2 Metrics

There are seven possible cases when classifying previously identified and novel peaks.

*A*. # previously identified peak correctly barcoded

*B*. # previously identified peak incorrectly barcoded

*C*. # previously identified peak correctly not barcoded

*D*. # previously identified peak incorrectly not barcoded

*E*. # novel peak (incorrectly) barcoded

*F*. # novel peak (correctly) not barcoded

*G*. # novel peak barcoded and barcode absent in testing sample

Using the six possible cases above, the following metrics were used to evaluate classifier performance against the discovery datasets.

$$\text{Conservative identification accuracy} = \frac{A}{A+B+E} \tag{4.1a}$$

$$\text{Potential identification accuracy} = \frac{A+G}{A+B+E} \tag{4.1b}$$

$$\text{Unassignment rate} = \frac{C+D+F}{A+B+C+D+E+F} \tag{4.1c}$$

$$\text{Unassignment accuracy} = \frac{C+F}{C+D+F} \tag{4.1d}$$

It is unclear how many novel peaks across all discovery datasets correspond to background noise or genuine lipid species which may be present or absent in the training dataset. If all novel peaks without SRM-IDA-EPI validation are assumed to be noise, the conservative identification accuracy is defined as the number of correct peak-barcode matches, verified by SRM-IDA-EPI, divided by the total number of assigned peaks. This metric is conservative because it penalizes any novel peaks that may be assigned to genuine lipid species. Thus, the conservative identification accuracy serves as a lower bound of identification accuracy. An upper bound of identification accuracy is represented by the potential identification accuracy. This metric accounts not only for correct peak-barcode matches validated by SRM-IDA-EPI, but also novel peaks assigned to lipids that have never been reported in these samples. The assumption of this metric is that novel peaks representing lipid species absent from the training datasets are not incorrectly assigned any lipid identities due to well-trained posterior probability cutoffs. When applied to the original datasets, both the conservative and potential identification accuracies are equivalent because these samples contain no novel peaks. Lastly, the unassignment rate and unassignment accuracy metrics measure the calibration of BATL's posterior probability cutoffs. In the context of the discovery datasets, all four metrics evaluate the robustness of BATL to accurately assign lipid identities, while not assigning annotations to peaks

representing noise or novel lipids absent from the training datasets.

### 4.6.3 Discovery-driven lipid identification

All peaks with in-source lipid artifact and isotope annotations were filtered out from the four sphingolipid and glycerophosphocholine discovery datasets. Of the remaining previously identified and novel peaks, the majority were novel (87.6-94.6%). These discovery datasets were classified ten times with BATL using the same ten cross validation partitions from Chapter 3. The number of previously identified and novel peaks detected across each transition are presented in Appendix 5.4. Conservative identification accuracies were significantly lower in the discovery datasets compared to the original ones, decreasing between 34.2-47.8% (Figure 4.2). An analysis of the classification errors revealed that most incorrect peak assignments were not caused by novel peaks falsely matching to previously identified peak barcodes. Rather, the majority of classification errors were caused by novel peaks that matched to training dataset barcodes not previously detected in these samples. Relative area histograms of previously identified-versus-novel peaks indicated that these peak assignments were possibly genuine as a small proportion of novel lipid species showed similar intensities as previously identified peaks (Appendix 5.4). A notable example where many new lipid species were identified is the glycerophospholipid blood dataset, where potential identification accuracies were not statistically different from original dataset identification accuracies. Although approximately 79 previously identified peaks were misclassified or unassigned in this discovery dataset, approximately 182 novel peaks were discovered instead. Similar findings were also observed in the glycerophosphocholine brain dataset where 206 novel peaks, which were absent in the original dataset, were identified using this discovery-driven lipid identification approach. While the potential identification accuracies of the other datasets remained significantly lower than the original datasets, the accuracies varied by only 10% or less. As the identities of these novel peaks assigned by BATL have not been

structurally validated by SRM-IDA-EPI, the true identification accuracies lie somewhere between the conservative and potential estimates across all datasets. Given the large proportion of novel peaks in all discovery datasets, the unassignment rates for BATL ranged between 86.6-93.0% (Figure 4.3). This was expected as not only were there many novel peaks likely corresponding to background noise, but also fewer training dataset barcodes per transition. More importantly, the unassignment accuracies of the discovery datasets were greater than 97% (Figure 4.4), indicating that BATL was effectively matching previously identified peaks to their correct barcodes and resistant to classifying peaks corresponding to background noise. As both unassignment rates and accuracies depend strongly on the number of novel peaks present in the discovery datasets, no statistics were calculated between original-versus-discovery datasets, or between all four discovery datasets.

## 4.7 Discussion

In this study, I developed and tested the first discovery-driven approach for targeted lipid identification. This methodology overcomes the two main biases associated with targeted lipidomics which are biased peak picking and biased peak identification. Traditionally, peaks are simultaneously picked and assigned lipid identities by expert operators. As discussed in Chapter 3, this approach is prone to identifying lipid species expected by the expert operator, while neglecting other detected peaks which may correspond to genuine lipid species. Due to the magnitude of detected peaks within each discovery dataset, removing in-source lipid artifacts and isotopes was essential in reducing data complexity. These artifacts could have potentially impeded downstream classification with BATL, and annotating these peaks yielded valuable structural information regarding possible peak identities, that could be confirmed by IDA-EPI.

Lipid identification of the discovery datasets proved challenging for

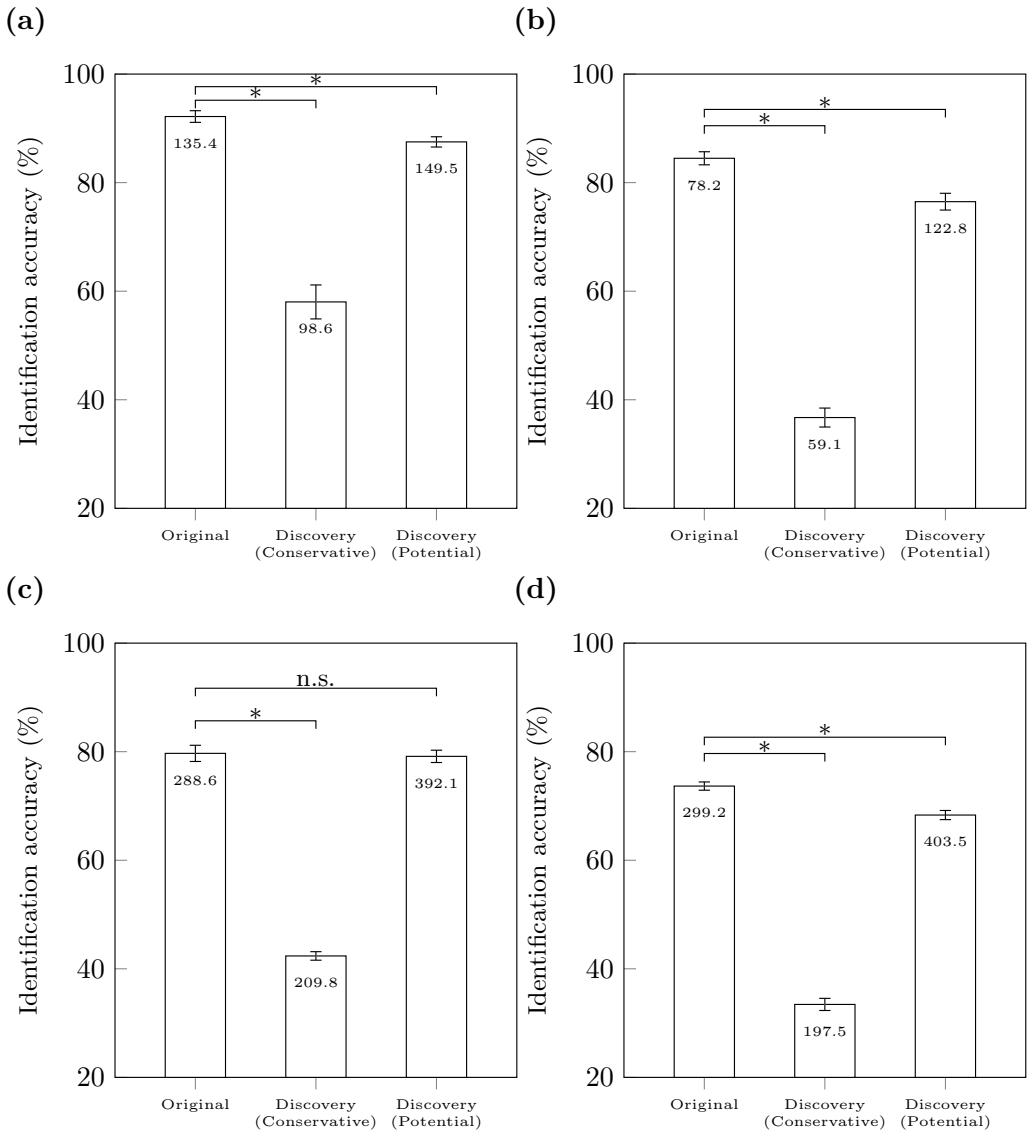



**Figure 4.2: Discovery-versus-original dataset identification accuracies.** Significance testing performed between original dataset and discovery dataset conservative/potential identification accuracies ($t$-test, $^*P < 0.05$). Labels inside bars indicate the average number of identified peaks. 95% confidence intervals shown. (a) Sphingolipid blood discovery dataset. (b) Sphingolipid brain discovery dataset. (c) Glycerophosphocholine blood discovery dataset. (d) Glycerophosphocholine brain discovery dataset.

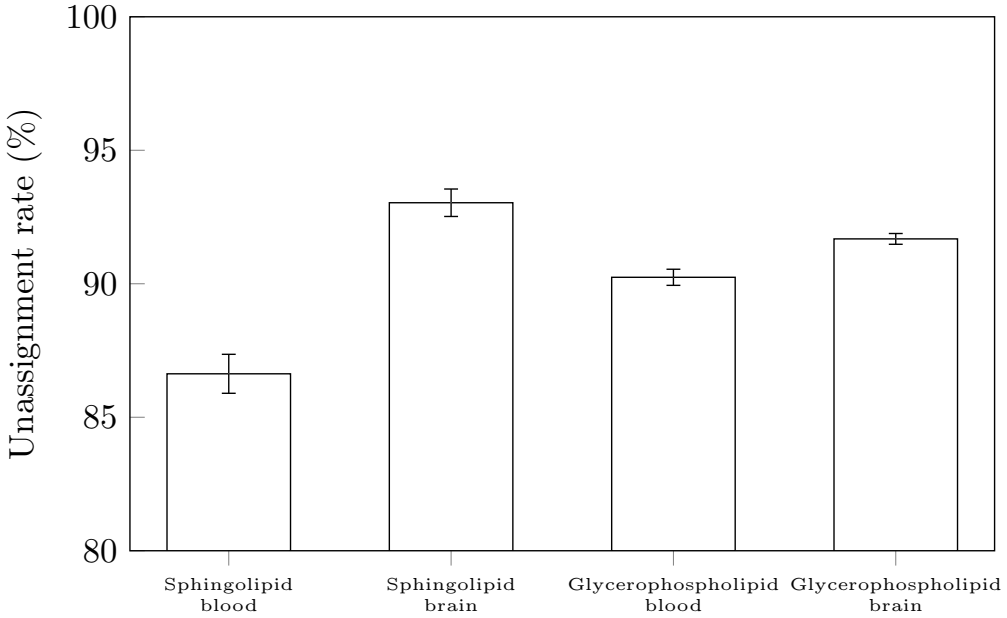



**Figure 4.3: BATL unassignment rates for sphingolipid and glycerophospholipid discovery datasets.** 95% confidence intervals shown.

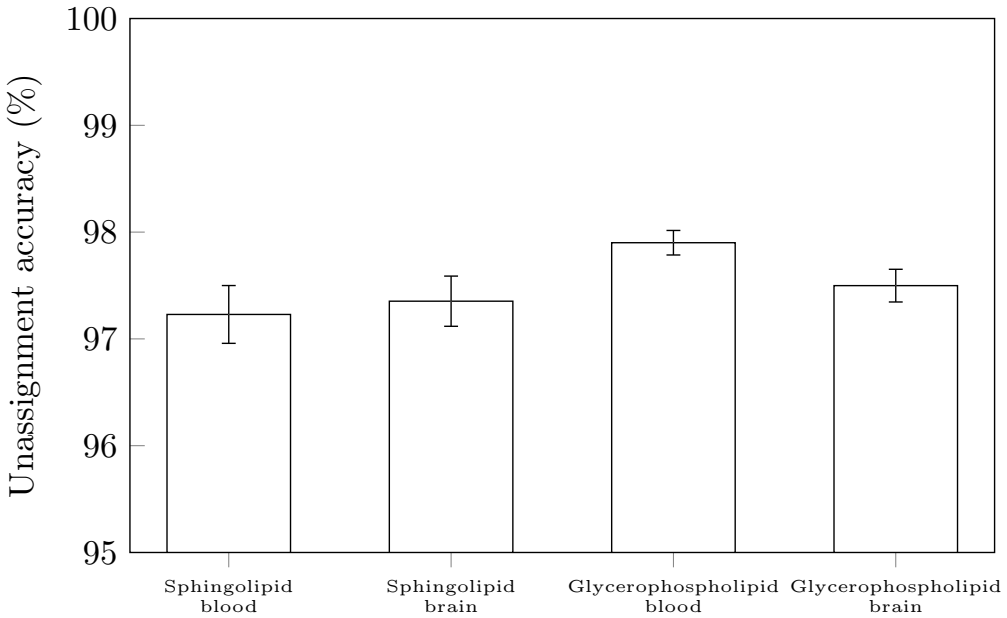



**Figure 4.4: BATL unassignment accuracies for sphingolipid and glycerophospholipid discovery datasets.** 95% confidence intervals shown.

BATL, given the magnitude of previously identified and novel peaks. These peaks represented both genuine lipid species and background noise, unlike the original datasets which only contained validated peaks chosen by expert operators. The addition of novel peaks increased classification difficulty as there were more possible peak-barcode assignments, and batch effects, such as retention time shifts, could lead to incorrect peak assignments. For this reason, the identification accuracies decreased significantly when classifying the discovery datasets. However, this metric was very also stringent because it only considered structurally validated peaks as correct peak-barcode assignments. Despite the strong assumptions calculating the potential identification accuracies, these findings indicated that dozens to hundreds of peaks representing bona fide lipid species were discarded from the expert operator analyses of the original datasets.

From an expert operator perspective, analyzing the discovery datasets mimicked the peak picking of an inexperienced analyst. Newly trained analysts often struggle to pick peaks while accounting for natural retention time drifts between samples, and matrix-specific effects that may alter peak intensities. For example, they may expect a peak to elute at a given retention time and incorrectly assume that a peak eluting at that exact time point is their lipid of interest, when it was simply background noise. Inexperienced operators are also prone to picking additional peaks which represent either genuine lipid species, noise, or artifact ions and isotopes. Still, it would be rare for an analyst to pick every detectable peak in their dataset, and the high unassignment rates and accuracies suggest BATL is capable of classifying SRM/MRM data collected by analysts with diverse peak-picking expertise.

## 4.8 Conclusion

Using the bioinformatics tools developed in previous objectives, I present the first unbiased targeted lipid identification pipeline. This workflow addresses

the major biases associated with traditional targeted lipidomics studies and enables the discovery of new lipid species across mixed matrices and biologically-divergent conditions. I highlight that this analysis is not intended to compete with untargeted lipidomic approaches but to rather complement them for improved lipid identification and discovery. Furthermore, this analysis can be used to mine pre-existing SRM and MRM datasets for novel lipid species with foreseeable increases in BATL identification accuracies as more lipid species are identified and incorporated into training datasets. In the future, this discovery-driven concept may become a standard workflow for targeted lipidomic analyses, pushing the boundaries of lipidomics in the pursuit of lipid discovery.

# Chapter Five

# Discussion

In this thesis, I challenged the common assumption that targeted lipidomic analyses can only quantify previously known lipid species. This misconception has led to the paucity of targeted lipid identification tools, in contrast with the numerous identification methods for untargeted lipidomics. I assert that both approaches are complementary with respect to lipid identification and that targeted lipidomics excels at identifying low and high abundance lipid isomers within a given family. I thus developed two bioinformatics tools for targeted lipid annotation and identification, demonstrating how these technologies could be exploited to enable discovery-driven, targeted lipidomics. These two bioinformatics tools are the in-source lipid artifact and isotope annotation tool (Chapter 2) and targeted lipidomics classifier BATL (Chapter 3). Using the artifact and isotope annotation tool, I demonstrated that manual peak picking is prone to error; expert operators are trained to avoid picking redundant peaks, yet up to 3.6% of SRM peaks corresponded to unwanted lipid artifacts and isotopes. Excluding these redundant peaks, I described a host of "real-world" problems associated with manual peak picking that could lead to lipid misidentification. For example, peak retention times may shift within samples and across acquisitions, and the number of lipid isomers detected at a given transition may vary according to the profiled matrix and biological condition. Current

approaches fail to model this information, naïvely aligning peak retention times and assigning lipid identities to the closest or most intense peak within a given retention time window. To address these challenges, BATL models peak features as Gaussian distributions, and identifies lipid species by maximizing the joint probabilities of all peak's identities under this model. Using BATL, a fully discovery-oriented targeted identification pipeline was proposed, whereby all detected peaks within discrete SRM transitions were subjected for lipid identification. Overall, these bioinformatics tools are the first to rigorously address the targeted lipid identification problem, enabling accurate, automatic, and discovery-driven targeted lipidomics.

## 5.1 In-source lipid artifact and isotope annotation

Avoiding artifactual and isotopic analytes is a problem of all MS-based assays. However, programs to detect these ions have only been designed to handle untargeted, rather than targeted, lipidomics data. This bioinformatics gap was addressed in Chapter 2, where an annotation tool was developed to detect artifactual and isotopic ions from targeted lipidomics approaches using LC-ESI-MS/MS. This tool was capable of identifying sphingolipid-specific artifacts, in addition to isotopes which are present in all biological analytes. When applied to empirical SRM datasets, even peaks picked by expert operators were found to represent artifacts and isotopes. This finding highlights one of many challenges associated with manual peak identification; no analyst can effectively monitor and avoid all possible combinations of detectable artifacts and isotopes within their dataset. Filtering these redundant peaks not only reduces data complexity, but potentially decreases the probability of downstream lipid misidentification.

Several challenges remain to improve the detection and annotation of artifactual and isotopic ions. In the future, the artifact list could be expanded to handle in-source modifications from all lipid categories and classes, not just

sphingolipids. One possible solution would be implementing a user-defined list of artifacts and adducts, enabling this program to handle any type of SRM/MRM data. Other limitations of my tool are shared across all in-source lipid annotation programs reported in literature. The major assumption of all programs is that the monoisotopic peak is picked, otherwise artifacts and isotopes will remain undetected. Furthermore, no annotation tool provides probabilistic scores that could be used to estimate the possibility of falsely identifying a lipid artifact. Chromatography similarity coefficients, m/z, and retention time thresholds are always user-defined parameters, and it is unclear how many false annotations are assigned at different thresholds. FDR analysis may provide a solution to this problem, although it remains unclear if meaningful FDRs could be estimated from similarity coefficients or other metrics of peak co-elution. An alternative approach is incorporating MS and peak detection parameters to estimate the variance of peak retention times. In SRM or MRM mode, the mass analyzer is rapidly cycling through each transition with a sequence of repeating dwell and pause times. Analytes are selected during the dwell time, while no signals are detected during the pause time when the mass analyzer is switching to the next transition. The sum of all dwell and pause times across all transitions is known as the cycle time. A cycle time of 0.033 minutes was programmed in the SRM method used to generate all empirical sphingolipid datasets. While this delay may not affect isotopes detected at neighbouring transitions, a hypothetical monoisotope and artifact detected at the start and end of the transition list, respectively, would vary in retention time by 0.033 minutes. This measurement error is greater than the retention time threshold used in Chapter 2, indicating that some deglycosylated lipid species could have been missed by the annotation tool. Different combinations of dwell and pause times may also alter the shapes of chromatographic peaks, and these shapes are further modified with peak detection parameters such as Gaussian smoothing or peak splitting. Taken together, these parameters may explain variations in peak retention times, which could be modelled to better estimate retention time cutoffs

and identify lipid artifacts.

## 5.2 Bayesian annotations for targeted lipidomics

SRM and MRM peak identification is a complex problem where lipid-specific features are influenced by chromatography, date of acquisition, type of mass spectrometer, biological condition and presence in discrete tissue matrices. Previous efforts have attempted to identify SRM and MRM peaks, although these approaches frame peak identification as a retention time alignment problem. To address these challenges, I developed BATL, the first classifier for SRM peak identification. Unlike retention time alignment or manual curation strategies, BATL is a Gaussian naïve Bayes classifier that models lipid-specific peak features across biological conditions. A noteworthy feature of BATL is the use of posterior probability cutoffs to learn when not to assign a lipid identity to a peak. This feature is absent from alignment approaches, which assign lipid identities to the closest or most intense peak within an arbitrary retention time window. Using three performance metrics, BATL was extensively validated using cross-validation, holdout, and statistically-independent testing datasets across diverse biological conditions and matrices.

BATL has the future potential to standardize targeted lipid identification across multiple LC-ESI-MS/MS platforms. As a simple probabilistic classifier, very little data is required for training. Using only 10% of the training dataset, or 22-23 samples, yielded well-calibrated posterior probability cutoffs and identification accuracies during cross-validation. This finding is important because lipid-specific peak features are strongly dependent on the type of LC-ESI-MS/MS platform. While laboratories using BATL must provide their own training data specific to their platform, I show that very few samples (less than 24) are required for effective classification. Furthermore, BATL has been coded to accept any continuous feature. For example,

lists of picked peaks may output other features such as total peak width, start time at 5% peak height, width at 5% peak height, or area/height. These features, although strongly correlated with the ones used in Chapter 3, could be explored in future work.

Refining the BATL algorithm for improved identification accuracy is an area of immediate future interest. The current BATL algorithm makes several assumptions when modelling targeted lipidomics data. In Chapter 3, the prior probabilities of observing all lipid identities were assumed to be independent given the condition. Biologically, this assumption is incorrect because some lipid isomers are only present in certain tissue matrices. Similarly, the disease state or treatment condition may alter lipid remodelling pathways, suggesting dependent lipid priors. A limited investigation of matrix-specific priors was performed in Appendix 5.4, but few comparisons were drawn between matrix-agnostic results in Chapter 3. Prior dependencies could be quantified by investigating SRM-IDA-EPI spectra for each SRM transition under a given biological condition/matrix. The frequency of lipid isomer co-occurrences could be estimated by structurally identifying all lipid species from each SRM-IDA-EPI spectrum. This manual analysis does require expert knowledge of lipid fragmentation pathways to decipher all lipid identities. This future work thus highlights another crucial future direction which is developing spectral database matching algorithms for targeted lipidomics data. Other final assumptions of BATL are Gaussian distributed features and the independency of features given the lipid identity. Using kernel density estimation instead of Gaussian distributions to estimate the likelihoods may improve identification accuracies. Future work is also required to quantify the possible dependencies between peak retention time, intensity, and shape, especially across samples experiencing batch effects. However, it remains unclear whether modelling these features using a more complex Bayesian network would substantially increase identification accuracies.

Although BATL was tested and validated with SRM datasets, the pro-

gram is conceptually designed to work with nearly any type of targeted lipidomics data. Precursor and product ion scans both detect lipid species by scanning and selecting pairs of parent and product ions. Similarly, neutral loss scans detect lipid species with a given parent ion and neutral loss fragment. Data acquired from these targeted acquisition modes consist of picked peaks detected at a parent ion and product ion/neutral loss. With appropriate lipid barcodes, BATL could be trained and applied to annotate peaks from any of these scanning modes. While familiarity with the R language is required to effectively use BATL, development of a front-end graphical user interface and/or cloud-based solution would allow end-users of all backgrounds to incorporate BATL into their targeted lipidomic analyses.

## 5.3 Discovery-driven targeted lipidomics

Combining the bioinformatics tools from previous sections, I developed the first discovery-driven, targeted lipidomics approach to lipid identification. This innovative method overcomes the major biases associated with targeted lipidomics data by detecting and identifying all possible lipid isomers within discrete SRM transitions. Compared to a traditional, targeted lipidomics approach, analysts are not biased towards reporting the lipids of interest at the expense of other detected signals. When coupled with untargeted lipidomic approaches, discovery-driven targeted lipidomics is an invaluable method to potentially identify all low and high abundance lipid isomers with a family across matrices and biological conditions. Identification of new lipid isomers using this approach may facilitate lipid biomarker discoveries and uncover mechanistic insight driving lipid remodelling pathways and lipid flux under disease conditions.

To become a standard data analysis workflow, discovery-driven targeted lipidomics must overcome several limitations of my study. While not explicitly mentioned in Chapter 4, all 17,190 picked peaks in the discovery datasets were manually picked from XICs visualized in MultiQuant. This

painstaking process of selecting peaks satisfying the signal to noise ratio (S/N) criteria took nearly a month to process eight SRM testing samples. This technical problem could be fixed with a software update for MultiQuant and other vendor-specific peak picking programs, to output a list of all detected peaks meeting user-defined criteria. Additionally, BATL must demonstrate greater identification accuracies in testing datasets to ensure reliable lipid annotations. Algorithmic changes to BATL were proposed in previous sections, and further performance increases could be achieved by expanding the training dataset to encompass more diverse lipid species across different conditions and matrices.

## 5.4 Advancing targeted lipid identification and discovery

The research contributions of this thesis lie in the demonstration that manual peak identification is prone to error and how bioinformatics tools can, for the very first time, enable automated and accurate targeted lipid annotation, identification, and discovery. BATL positions itself as the first SRM/MRM classifier for targeted lipidomics data. Having discussed the challenges of peak identification, I anticipate my work will help to eliminate misconceptions of both targeted and untargeted lipidomics, reinforcing the need for both approaches to fully interrogate the lipidome. I also anticipate that BATL will encourage future peak identification algorithms for targeted lipidomics, including much needed spectral database matching algorithms for IDA-EPI spectra. Using BATL, laboratories conducting targeted lipidomic experiments can quickly and efficiently analyze their datasets without subjective human biases. As the field of lipidomics continues to mature, bioinformatics tools such as BATL will help to push the boundaries of targeted lipidomics, uncovering the biological breakthroughs hidden within our membranes.

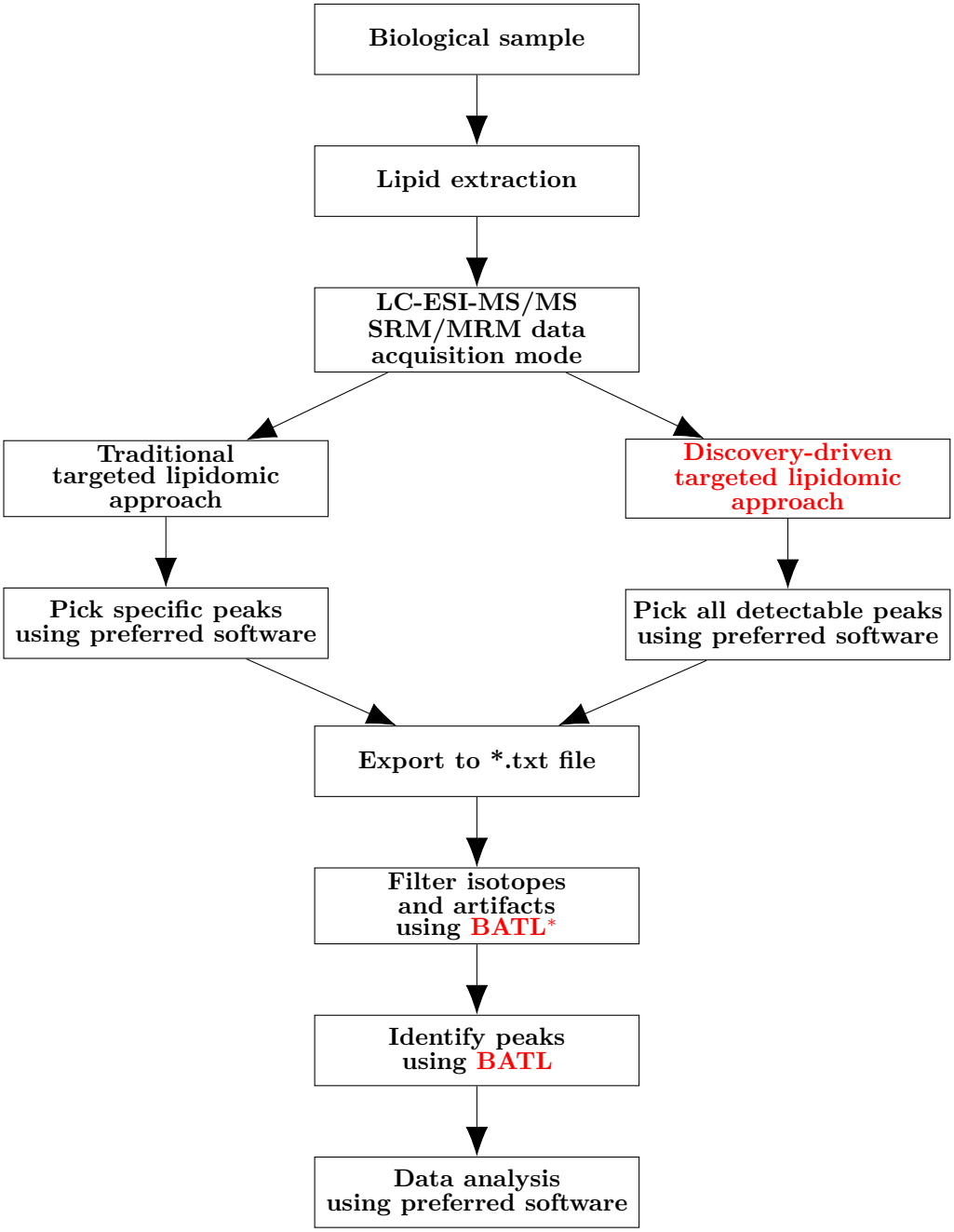



**Figure 5.1: BATL research contributions to the targeted lipidomics analysis pipeline.** *The in-source lipid artifact and isotope annotation tool is built into BATL.

# References


Alecu, Irina and Steffany A. L. Bennett (2019). "Dysregulated lipid metabolism and its role in $\alpha$-synucleinopathy in Parkinson's disease". In: *Front. Neurosci.* 13, p. 328.

Allen, Felicity, Russ Greiner, and David Wishart (2015). "Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification". In: *Metabolomics* 11.1, pp. 98–110.

Allen, Felicity, Allison Pon, et al. (2014). "CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra". In: *Nucleic Acids Res.* 42.Web Server issue, W94–W99.

Allen, W. V. (1976). "Biochemical aspects of lipid storage and utilization in animals". In: *Am. Zool.* 16.4, pp. 631–647.

Back, M. et al. (2019). "Inflammation and its resolution in atherosclerosis: mediators and therapeutic opportunities". In: *Nat. Rev. Cardiol.* 16.7, pp. 389–406.

Balogh, Michael P. (2004). "Debating resolution and mass accuracy in mass spectrometry". In: *Spectroscopy* 19.10, p. 34.

Barupal, Dinesh and Oliver Fiehn (2017). "Chemical similarity enrichment analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets". In: *Sci. Rep.* 7.1, p. 14567.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing". In: *J R STAT SOC B* 57.1, pp. 289–300.

Bennett, Steffany A. L. et al. (2013). "Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer's disease". In: *Front. Physiol.* 4, p. 168.

Benton, H. et al. (2008). "XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization". In: *Anal. Chem.* 80.16, pp. 6382–6389.

Bilgin, Mesut et al. (2016). "Lipid discovery by combinatorial screening and untargeted LC-MS/MS". In: *Sci. Rep.* 6.1, p. 27920.

Blanchard, Alexandre P. et al. (2013). "Visualization and phospholipid identification (VaLID): Online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass". In: *Bioinformatics* 29.2, pp. 284–285.

Blanksby, Stephen J. and Todd W. Mitchell (2010). "Advances in mass spectrometry for lipidomics". In: *Annu. Rev. Anal. Chem.* 3.1, pp. 433–465.

Blaženović, Ivana, Tobias Kind, Michael R. Sa, et al. (2019). "Structure annotation of all mass spectra in untargeted metabolomics". In: *Anal. Chem.* 91.3, pp. 2155–2162.

Blaženović, Ivana, Tobias Kind, Hrvoje Torbašinović, et al. (2017). "Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: Database boosting is needed to achieve 93% accuracy". In: *Journal of Cheminformatics* 9.1, p. 32.

Bligh, E. G. and W. J. Dyer (1959). "A rapid method of total lipid extraction and purification". In: *Can. J. Biochem.* 37.8, pp. 911–917.

Bowden, John et al. (2018). "NIST lipidomics workflow questionnaire: An assessment of community-wide methodologies and perspectives". In: *Metabolomics* 14.5, pp. 1–11.

Bozek, Katarzyna et al. (2015). "Organization and evolution of brain lipidome revealed by large-scale analysis of human, chimpanzee, macaque, and mouse tissues". In: *Neuron* 85.4, pp. 695–702.

Brodbelt, Jennifer (1997). "Mass spectrometry: Principles and applications". In: *J. Am. Soc. Mass Spectrom.* 8.11, pp. 1193–1194.

Brown, H. Alex and Robert C. Murphy (2009). "Working towards an exegesis for lipids in biology". In: *Nat. Chem. Biol.* 5.9, pp. 602–606.

Brugger, B. et al. (1997). "Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem mass spectrometry". In: *Proc. Natl. Acad. Sci. U. S. A.* 94.6, pp. 2339–2344.

Caffrey, Martin and Jacqueline Hogan (1992). "LIPIDAT: A database of lipid phase transition temperatures and enthalpy changes. DMPC data subset analysis". In: *Chem. Phys. Lipids* 61.1, pp. 1–109.

Cai, Yuping, Kai Weng, et al. (2015). "An integrated targeted metabolomic platform for high-throughput metabolite profiling and automated data processing". In: *Metabolomics* 11.6, pp. 1575–1586.

Cai, Yuping and Zheng-Jiang Zhu (2019). "A high-throughput targeted metabolomics workflow for the detection of 200 polar metabolites in central carbon metabolism". In: *Microbial metabolomics: methods and protocols*. New York, NY: Springer, pp. 263–274.

Cajka, Tomas and Oliver Fiehn (2014). "Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry". In: *Trends Anal. Chem.* 61.

— (2016). "Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics". In: *Anal. Chem.* 88.1, pp. 524–545.

Chong, Jasmine et al. (2018). "MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis". In: *Nucleic Acids Res.* 46.W1, W486–W494.

Collins, James R. et al. (2016). "LOBSTAHS: An adduct-based lipidomics strategy for discovery and identification of oxidative stress biomarkers". In: *Anal. Chem.* 88.14, pp. 7154–7162.

Contrepois, Kévin et al. (2018). "Cross-platform comparison of untargeted and targeted lipidomics approaches on aging mouse plasma". In: *Sci. Rep.* 8.1, p. 17747.

Defelice, Brian C. et al. (2017). "Mass spectral feature list optimizer (MS-FLO): A tool to minimize false positive peak reports in untargeted liquid chromatography-mass spectroscopy (LC-MS) data processing". In: *Anal. Chem.* 89.6, pp. 3250–3255.

Djafi, Nabila et al. (2013). "Multiple reaction monitoring mass spectrometry is a powerful tool to study glycerolipid composition in plants with different level of desaturase activity". In: *Plant Signal. Behav.* 8.5, e24118.

Djoumbou Feunang, Yannick et al. (2016). "ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy". In: *J. Cheminform.* 8, p. 61.

Djoumbou-Feunang, Yannick et al. (2019). "CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification". In: *Metabolites* 9.4.

Domingo-Almenara, Xavier et al. (2018). "XCMS-MRM and METLIN-MRM: A cloud library and public resource for targeted analysis of small molecules". In: *Nat. Methods* 15.9, pp. 681, +.

Eilers, Paul H. C. (2004). "Parametric time warping". In: *Anal. Chem* 76.2, pp. 404–411.

Ekroos, Kim et al. (2002). "Quantitative profiling of phospholipids by multiple precursor ion scanning of a hybrid quadrupole time-of-flight mass spectrometer". In: *Anal. Chem.* 74.5, pp. 941–949.

Fahy, Eoin, Shankar Subramaniam, H. Alex Brown, et al. (2005). "A comprehensive classification system for lipids". In: *J. Lipid Res.* 46.5, pp. 839–862.

Fahy, Eoin, Shankar Subramaniam, Robert C. Murphy, et al. (2009). "Update of the LIPID MAPS comprehensive classification system for lipids". In: *J. Lipid Res.* 50.Supplement, S9–S14.

Fernandis, Aaron Z. and Markus R. Wenk (2007). "Membrane lipids as signaling molecules". In: *Curr. Opin. Cell Biol.* 18.2, pp. 121–128.

Folch, J., S. Arsove, and J. A. Meath (1951). "Isolation of brain strandin, a new type of large molecule tissue component". In: *J. Biol. Chem.* 191.2, pp. 819–831.

Forest, Anik et al. (2018). "Comprehensive and reproducible untargeted lipidomic workflow using LC-QTOF validated for human plasma analysis". In: *J. Proteome Res.* 17.11, pp. 3657–3670.

Foster, Joseph M. et al. (2013). "LipidHome: A database of theoretical lipids optimized for high throughput mass spectrometry lipidomics". In: *PLoS ONE* 8.5, e61951.

Gerlich, Michael and Steffen Neumann (2013). "MetFusion: Integration of compound identification strategies". In: *J. Mass Spectrom.* 48.3, pp. 291–298.

Goldberg, Andrew and Robert Tarjan (1988). "A new approach to the maximum-flow problem". In: *J. ACM* 35.4, pp. 921–940.

Goracci, Laura et al. (2017). "Lipostar, a comprehensive platform-neutral cheminformatics tool for lipidomics". In: *Anal. Chem.* 89.11, pp. 6257–6264.

Granger, Matthew W. et al. (2019). "Distinct disruptions in Land's cycle remodeling of glycerophosphocholines in murine cortex mark symptomatic onset and progression in two Alzheimer's disease mouse models". In: *J. Neurochem.* 149.4, pp. 499–517.

Gross, Richard W. (2017). "The evolution of lipidomics through space and time". In: *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1862.8, pp. 731–739.

Gross, Richard W. et al. (2005). "Functional lipidomics: The roles of specialized lipids and lipid–protein interactions in modulating neuronal function". In: *Prostaglandins Other Lipid Mediat.* 77.1-4, pp. 52–64.

Haimi, Perttu et al. (2006). "Software tools for analysis of mass spectrometric lipidome data". In: *Anal. Chem.* 78.24, pp. 8324–8331.

Hajicek, Nicole et al. (2019). "Structural basis for the activation of PLC-$\gamma$ isozymes by phosphorylation and cancer-associated mutations". In: *eLife* 8.

Han, Xianlin and Richard W. Gross (2003). "Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: A bridge to lipidomics". In: *J. Lipid. Res.* 44.6, pp. 1071–1079.

Han, Xianlin, Kui Yang, and Richard W. Gross (2012). "Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses". In: *Mass Spectrom. Rev.* 31.1, pp. 134–178.

Harayama, Takeshi and Howard Riezman (2018). "Understanding the diversity of membrane lipid composition". In: *Nat. Rev. Mol. Cell Biol.* 19.5, pp. 281–296.

Hartler, Jürgen, Alexander Triebl, et al. (2017). "Deciphering lipid structures based on platform-independent decision rules". In: *Nat. Methods* 14.12, pp. 1171–1174.

Hartler, Jürgen, Martin Trötzmüller, et al. (2011). "Lipid Data Analyzer: Unattended identification and quantitation of lipids in LC-MS data". In: *Bioinformatics* 27.4, pp. 572–577.

Hedger, George and Mark S. P. Sansom (2016). "Lipid interaction sites on channels, transporters and receptors: Recent insights from molecular dynamics simulations". In: *Biochim. Biophys. Acta Biomembr.* 1858.10, pp. 2390–2400.

Herzog, Ronny et al. (2011). "A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language". In: *Genome Biol.* 12.1, R8–R8.

Hill, Alastair W. and Russell J. Mortishire-Smith (2005). "Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach". In: *Rapid Commun. Mass Spectrom.* 19.21, pp. 3111–3118.

Ho, C. S. et al. (2003). "Electrospray ionisation mass spectrometry: Principles and clinical applications". In: *Clin. Biochem. Rev.* 24.1, pp. 3–12.

Hotamisligil, Gökhan S. (2017). "Inflammation, metaflammation and immunometabolic disorders". In: *Nature* 542.7640, pp. 177–185.

Houjou, Toshiaki et al. (2005). "A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry". In: *Rapid Commun. Mass Spectrom.* 19.5, pp. 654–666.

Hu, Changfeng, Qiao Duan, and Xianlin Han (2019). "Strategies to improve/eliminate the limitations in shotgun lipidomics". In: *Proteomics*, e1900070.

Jaitly, Navdeep et al. (2006). "Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline". In: *Anal. Chem.* 78.21, pp. 7397–7409.

Jiang, Hong et al. (2018). "Protein lipidation: Occurrence, mechanisms, biological functions, and enabling technologies". In: *Chem. Rev.* 118.3, pp. 919–988.

Kassidas, Athanassios, John F. Macgregor, and Paul A. Taylor (1998). "Synchronization of batch trajectories using dynamic time warping". In: *Am Inst Chem Eng J* 44.4, pp. 864–875.

Katajamaa, Mikko, Jarkko Miettinen, and Matej Orei (2006). "MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data". In: *Bioinformatics* 22.5, pp. 634–636.

Khrameeva, Ekaterina et al. (2018). "Lipidome evolution in mammalian tissues". In: *Mol. Biol. Evol.* 35.8, pp. 1947–1957.

Kind, Tobias, Kwang-Hyeon Liu, et al. (2013). "LipidBlast in silico tandem mass spectrometry database for lipid identification". In: *Nat. Methods* 10.8, pp. 755–758.

Kind, Tobias, Hiroshi Tsugawa, et al. (2018). "Identification of small molecules using accurate mass MS/MS search". In: *Mass Spectrom. Rev.* 37.4, pp. 513–532.

Kochen, Michael A. et al. (2016). "Greazy: Open-source software for automated phospholipid tandem mass spectrometry identification". In: *Anal. Chem.* 88.11, pp. 5733–5741.

Koelmel, Jeremy P. et al. (2017). "Common cases of improper lipid annotation using high-resolution tandem mass spectrometry data and corresponding limitations in biological interpretation". In: *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1862.8, pp. 766–770.

Konermann, Lars et al. (2013). "Unraveling the mechanism of electrospray ionization". In: *Anal. Chem.* 85.1, pp. 2–9.

Kuhl, Carsten et al. (2012). "CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets". In: *Anal. Chem.* 84.1, pp. 283–289.

Kuo, Tien-Chueh and Yufeng Jane Tseng (2018). "LipidPedia: A comprehensive lipid knowledgebase." In: *Bioinformatics* 34.17, pp. 2982–2987.

Leavell, Michael D. and Julie A. Leary (2006). "Fatty acid analysis tool (FAAT): An FT-ICR MS lipid analysis algorithm". In: *Anal. Chem.* 78.15, pp. 5497–5503.

Lee, Hyeon-Cheol and Takehiko Yokomizo (2018). "Applications of mass spectrometry-based targeted and non-targeted lipidomics". In: *Biochem. Biophys. Res. Commun.* 504.3, pp. 576–581.

Li, Peiqi et al. (2020). "Cryo-EM reconstruction of a VPS13 fragment reveals a long groove to channel lipids between membranes". In: *J. Cell Biol.* 219.5.

Liebisch, Gerhard, Christer S. Ejsing, and Kim Ekroos (2015). "Identification and annotation of lipid species in metabolomics studies need improvement". In: *Clin. Chem.* 61.12, pp. 1542–1544.

Liebisch, Gerhard, Kim Ekroos, et al. (2017). "Reporting of lipidomics data should be standardized". In: *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1862.8, pp. 747–751.

Listgarten, Jennifer et al. (Jan. 2005). "Multiple alignment of continuous time series". In: *Adv Neural Inf Process Syst* 17, pp. 817–824.

Marrink, Siewert-Jan and Herman J. C. Berendsen (1994). "Simulation of water transport through a lipid membrane". In: *J. Phys. Chem.* 98.15, pp. 4155–4168.

McDowell, Graeme S. V. et al. (2014). "Predicting glycerophosphoinositol identities in lipidomic datasets using VaLID (visualization and phospholipid identification)–an online bioinformatic search engine". In: *Biomed Res. Int.* 2014, p. 8.

Melnik, Alexey V. et al. (2017). "Coupling targeted and untargeted mass spectrometry for metabolome-microbiome-wide association studies of human fecal samples". In: *Anal. Chem.* 89.14, pp. 7549–7559.

Merrill, A. H. (1992). "Ceramide: A new lipid second "messenger"?" In: *Nutr. Rev.* 50.3, pp. 78–80.

Merrill, Alfred H. et al. (2005). "Sphingolipidomics: High-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry". In: *Methods* 36.2, pp. 207–224.

Murphy, Robert C. and Paul H. Axelsen (2011). "Mass spectrometric analysis of long-chain lipids". In: *Mass Spectrom. Rev.* 30.4, pp. 579–599.

Murphy, Robert C., Patrick F. James, et al. (2007). "Detection of the abundance of diacylglycerol and triacylglycerol molecular species in cells using neutral loss mass spectrometry". In: *Anal. Biochem.* 366.1, pp. 59–70.

Narváez-Rivas, Mónica and Qibin Zhang (2016). "Comprehensive untargeted lipidomic analysis using core–shell C30 particle column and high field orbitrap mass spectrometer". In: *J. Chromatogr. A* 1440, pp. 123–134.

Nguyen, An et al. (2017). "Using lipidomics analysis to determine signalling and metabolic changes in cells". In: *Curr. Opin. Biotechnol.* 43, pp. 96–103.

Nielsen, Niels-Peter Vest, Jens Michael Carstensen, and Jørn Smedsgaard (1998). "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping". In: *J. Chromatogr. A* 805.1, pp. 17–35.

O'Connor, Anne et al. (2017). "LipidFinder: A computational workflow for discovery of lipids identifies eicosanoid-phosphoinositides in platelets". In: *JCI insight* 2.7, e91634.

O'Donnell, Valerie B. et al. (2020). "Lipidomics: Current state of the art in a fast moving field". In: *Wiley Interdiscip. Rev. Sys. Biol. Med.* 12.1, e1466.

O'Donnell B., Valerie, Robert Murphy C., and Steve Watson P. (2014). "Platelet lipidomics: Modern day perspective on lipid discovery and characterization in platelets". In: *Circ. Res.* 114.7, pp. 1185–1203.

Olzmann, James A. and Pedro Carvalho (2019). "Dynamics and functions of lipid droplets". In: *Nat. Rev. Mol. Cell Biol.* 20.3, pp. 137–155.

Panuwet, Parinya et al. (2016). "Biological matrix effects in quantitative tandem mass spectrometry-based analytical methods: Advancing biomonitoring". In: *Crit. Rev. Anal. Chem.* 46.2, pp. 93–105.

Pluskal, Tomás et al. (2010). "MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data". In: *BMC Bioinformatics* 11.1, p. 395.

Porta Siegel, Tiffany, Kim Ekroos, and Shane R. Ellis (2019). "Reshaping lipid biochemistry by pushing barriers in structural lipidomics". In: *Angew. Chem. Int. Ed.* 58.20, pp. 6492–6501.

Prince, John T. and Edward M. Marcotte (2006). "Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping". In: *Anal. Chem.* 78.17, pp. 6140–6152.

Quehenberger, Oswald and Edward A. Dennis (2011). "The human plasma lipidome". In: *N. Engl. J. Med.* 365.19, pp. 1812–1823.

Reis, Ana et al. (2013). "A comparison of five lipid extraction solvent systems for lipidomic studies of human LDL". In: *J. Lipid Res.* 54.7, pp. 1812–1824.

Ruttkies, Christoph et al. (2016). "MetFrag relaunched: Incorporating strategies beyond in silico fragmentation". In: *J. Cheminform.* 8.1, pp. 1–16.

Sampaio, Julio L. et al. (2011). "Membrane lipidome of an epithelial cell line". In: *Proc. Natl. Acad. Sci. U. S. A.* 108.5, pp. 1903–1907.

Schmelzer, Kara et al. (2007). "The LIPID MAPS initiative in lipidomics". In: *Meth. Enzymol.* 432, pp. 171–183.

Schwudke, Dominik et al. (2007). "Shotgun lipidomics by tandem mass spectrometry under data-dependent acquisition control". In: *Meth. Enzymol.* 433, pp. 175–91.

SCIEX (2012). *LipidView Software User Guide*. Version 1.2. SCIEX. 70 pp.

Senan, Oriol et al. (2019). "CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network". In: *Bioinformatics* 35.20, pp. 4089–4097.

Shaner, Rebecca L. et al. (2009). "Quantitative analysis of sphingolipids for lipidomics using triple quadrupole and quadrupole linear ion trap mass spectrometers". In: *J. Lipid Res.* 50.8, pp. 1692–1707.

Shevchenko, Andrej and Kai Simons (2010). "Lipidomics: Coming to grips with lipid diversity". In: *Nat. Rev. Mol. Cell Biol.* 11.8, pp. 593–598.

Smith, Colin A. et al. (2006). "XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification". In: *Anal. Chem.* 78.3, pp. 779–787.

Smith, Rob, Dan Ventura, and John T. Prince (2015). "LC-MS alignment in theory and practice: A comprehensive algorithmic review". In: *Brief. Bioinform.* 16.1, pp. 104–117.

Song, Haowei et al. (2007). "Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching". In: *J. Am. Soc. Mass Spectrom.* 18.10, pp. 1848–1858.

Spickett, Corinne M. and Andrew R. Pitt (2015). "Oxidative lipidomics coming of age: Advances in analysis of oxidized phospholipids in physiology and pathology". In: *Antioxid. Redox Signal.* 22.18, pp. 1646–1666.

Sud, Manish et al. (2007). "LIPID MAPS online tools for lipid research". In: *Nucleic Acids Res.* 35.Web Server issue, W606–W612.

Tsugawa, Hiroshi, Masanori Arita, et al. (2013). "MRMPROBS: A data assessment and metabolite identification tool for large-scale multiple reaction monitoring based widely targeted metabolomics". In: *Anal. Chem.* 85.10, pp. 5191–5199.

Tsugawa, Hiroshi, Tomas Cajka, et al. (2015). "MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis". In: *Nat. Methods* 12.6, pp. 523–526.

Tsugawa, Hiroshi, Erika Ohta, et al. (2014). "MRM-DIFF: Data processing strategy for differential analysis in large scale MRM-based lipidomics studies." In: *Front. Genet.* 5, pp. 471–471.

Uppal, Karan, Douglas I. Walker, and Dean P. Jones (2017). "xMSannotator: An R package for network-based annotation of high-resolution metabolomics data". In: *Anal. Chem.* 89.2, pp. 1063–1067.

Vasan, Neil et al. (2019). "Double mutations in cis increase oncogenicity and sensitivity to PI3K$\alpha$ inhibitors". In: *Science* 366.6466, pp. 714–723.

Watanabe, Kiyohiro, Etsuko Yasugi, and Mieko Oshima (2000). "How to search the glycolipid data in "LIPIDBANK for Web" the newly developed lipid database in Japan". In: *Trends Glycosci. Glycotechnol.* 12.65, pp. 175–184.

Whitehead, Shawn N. et al. (2007). "Identification and quantitation of changes in the platelet activating factor family of glycerophospholipids over the course of neuronal differentiation by high-performance liquid chromatography electrospray ionization tandem mass spectrometry". In: *Anal. Chem.* 79.22, pp. 8539–8548.

Witting, Michael et al. (2017). "LipidFrag: Improving reliability of in silico fragmentation of lipids and application to the Caenorhabditis elegans lipidome". In: *PLoS One* 12.3, e0172311.

Wolf, Sebastian et al. (2010). "In silico fragmentation for computer assisted identification of metabolite mass spectra". In: *BMC Bioinformatics* 11.1, p. 148.

Wymann, Matthias P. and Roger Schneiter (2008). "Lipid signalling in disease". In: *Nat. Rev. Mol. Cell Biol.* 9.2, pp. 162–16276.

Xu, Hongbin et al. (2019). "DMS as an orthogonal separation to LC/ESI/MS/MS for quantifying isomeric cerebrosides in plasma and cerebrospinal fluid". In: *J. Lipid Res.* 60.1, pp. 200–211.

Zhang, Linlin, Xianlin Han, and Xiangdong Wang (2018). "Is the clinical lipidomics a potential goldmine?" In: *Cell Biol. Toxicol.* 34.6, pp. 421–423.

Ziegler, Wolfgang H. et al. (2002). "A lipid-regulated docking site on vinculin for protein kinase C". In: *J. Biol. Chem.* 277.9, pp. 7396–7404.

# Appendix: SRM/MRM data descriptions

## Dataset descriptions

**Table A1: Sphingolipid training dataset.**

| Acquisition # | Sample runs | Organism | Matrix | [1]Condition |
|---|---|---|---|---|
| 1 | 192 | Human | Blood (plasma) | AD, MCI, DLB, PD, PDD, control |
| 2 | 127 | Human | Blood (plasma) | DLB, control |
| 3 | 89 | Mouse | Brain (hippocampus) | PD model |
| 4 | 32 | Mouse | Brain (temporal cortex) | PD model |

[1] AD: Alzheimer's disease. MCI: mild cognitive impairment. DLB: Dementia with Lewy bodies. PD: Parkinson's disease. PDD: Parkinson's disease with dementia.

**Table A2: Glycerophospholipid training dataset.**

| Acquisition # | Sample run | Organism | Matrix | [1]Condition |
|---|---|---|---|---|
| 1 | 192 | Human | Blood (plasma) | AD, MCI, DLB, PD, PDD, control |
| 2 | 127 | Human | Blood (plasma) | DLB, control |
| 3 | 38 | Mouse | Brain (hippocampus) | Normal chow |
| 4 | 38 | Mouse | Brain (temporal cortex) | Normal chow |
| 5 | 25 | Mouse | Brain (hippocampus) | Omega 3 enriched chow |
| 6 | 25 | Mouse | Brain (temporal cortex) | Omega 3 enriched chow |
| 7 | 27 | Mouse | Brain (hippocampus) | Omega 6/9 enriched chow |
| 8 | 27 | Mouse | Brain (temporal cortex) | Omega 6/9 enriched chow |

[1] AD: Alzheimer's disease. MCI: mild cognitive impairment. DLB: Dementia with Lewy bodies. PD: Parkinson's disease. PDD: Parkinson's disease with dementia.

**Table A3: Sphingolipid blood testing dataset.**

| Acquisition # | Sample runs | Organism | Matrix | [1]Condition |
|---|---|---|---|---|
| 1 | 47 | Human | Blood (serum) | MCI |

[1] MCI: mild cognitive impairment.

**Table A4: Sphingolipid brain testing dataset.**

| Acquisition # | Sample runs | Organism | Matrix | [1]Condition |
|---|---|---|---|---|
| 1 | 12 | Mouse | Brain (temporal cortex) | Normal chow |

[1] Samples are serially diluted.

**Table A5: Glycerophospholipid blood testing dataset.**

| Acquisition # | Sample runs | Organism | Matrix | Condition |
|---|---|---|---|---|
| 1 | 23 | Mouse | Blood (plasma) | Normal chow |
| 2 | 24 | Mouse | Blood (plasma) | Omega 3 enriched chow |
| 3 | 18 | Mouse Mouse | Blood (plasma) | Omega 6/9 enriched chow |

**Table A6: Glycerophospholipid brain testing dataset.**

| Acquisition # | Sample runs | Organism | Matrix | [1]Condition |
|---|---|---|---|---|
| 1 | 30 | Mouse | Brain (hippocampus) | 0-24 hours PMI |
| 2 | 28 | Mouse | Brain (temporal cortex) | 0-24 hours PMI |

[1] PMI: *post-mortem* interval.

# Dataset isomer/isobar complexity

Guaranteed assignments are defined as the number of peaks that could only match the correct lipid identity in the training dataset. None of the candidate assignments include the "no assignment" label unique to each peak.

**Table A7: Sphingolipid training dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 13222 | 6351 |
| 2 | 11403 | 418 |
| 3 | 4590 | 56 |
| 4 | 1023 | 0 |

**Table A8: Sphingolipid holdout dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 13069 | 6200 |
| 2 | 11307 | 455 |
| 3 | 4545 | 80 |
| 4 | 1008 | 0 |

**Table A9: Sphingolipid blood testing dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 28664 | 28664 |
| 2 | 18831 | 1558 |
| 3 | 5640 | 235 |
| 4 | 0 | 0 |

**Table A10: Sphingolipid brain testing dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 4548 | 4546 |
| 2 | 4869 | 678 |
| 3 | 3168 | 0 |
| 4 | 72 | 0 |

**Table A11: Glycerophospholipid training dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 9492 | 5155 |
| 2 | 16880 | 481 |
| 3 | 39002 | 0 |
| 4 | 52940 | 56 |
| 5 | 45202 | 0 |
| 6 | 25033 | 0 |
| 7 | 7490 | 0 |
| 8 | 5831 | 0 |

**Table A12: Glycerophospholipid holdout dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 9474 | 5091 |
| 2 | 16963 | 505 |
| 3 | 38838 | 0 |
| 4 | 53361 | 53 |
| 5 | 44937 | 0 |
| 6 | 25760 | 0 |
| 7 | 7483 | 0 |
| 8 | 5810 | 0 |

**Table A13: Glycerophospholipid blood testing dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 24290 | 24290 |
| 2 | 36220 | 593 |
| 3 | 94408 | 101 |
| 4 | 137097 | 57 |
| 5 | 137676 | 0 |
| 6 | 31790 | 0 |
| 7 | 24570 | 0 |
| 8 | 2730 | 0 |

**Table A14: Glycerophospholipid brain testing dataset.**

| No. candidate assignments | Detected peaks | Guaranteed assignments |
|---|---|---|
| 1 | 23286 | 19716 |
| 2 | 35786 | 1316 |
| 3 | 115970 | 517 |
| 4 | 142480 | 0 |
| 5 | 129492 | 0 |
| 6 | 45170 | 0 |
| 7 | 37660 | 0 |
| 8 | 4340 | 0 |

# Appendix: Feature normality

## Assessing whether barcode features were Gaussian distributed

The normal or log-normal distribution was used to estimate the likelihood of observing each barcode feature in the training dataset. To determine which distribution better modelled each feature, a KS test was used to assess whether a log transformation increased the proportion of barcodes that failed to reject the Gaussian, null hypothesis at $P < 0.05$. However, this statistical test could not be used to evaluate whether the normal or log-normal distribution itself was an appropriate model for each barcode feature. This was because the normal and log-normal distribution parameters were estimated directly from the training dataset to establish the null statistical mode. When the same data is used to construct and test the null hypothesis of normally-distributed barcode features, any $P$-values obtained are likely incorrect. To assess whether the normal and log-normal distribution were appropriate models for each feature, the following procedure was used to simulate an empirical distribution of KS test statistics:

For every barcode feature represented as a vector of size $n$:

1. Estimate the sample mean and variance.
2. Draw 1000 samples of size $n$ from a Gaussian distribution with parameters $\mu$ and $\sigma^2$ estimated from the original data.

3. For each sample, compute the KS test $D$ statistic with Gaussian parameters $\mu$ and $\sigma^2$ estimated from that sample.

4. Estimate a non-parametric distribution of KS test $D$ statistics using the LOGSPLINE R package.

5. Compute the KS test $D$ statistic using the original feature vector with Gaussian parameters estimated from the same data (Step 1).

6. Compute the $P$-value using the KS test $D$ statistic computed from Step 5 against the empirical distribution of $D$ statistics in Step 4.

All $j$ barcode feature $P$-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (Benjamini and Hochberg (1995)). The proportion of normally distributed barcodes for a given feature is

$$\frac{\sum_{i=1}^{j} I(Q_i >= \alpha)}{j} \tag{1}$$

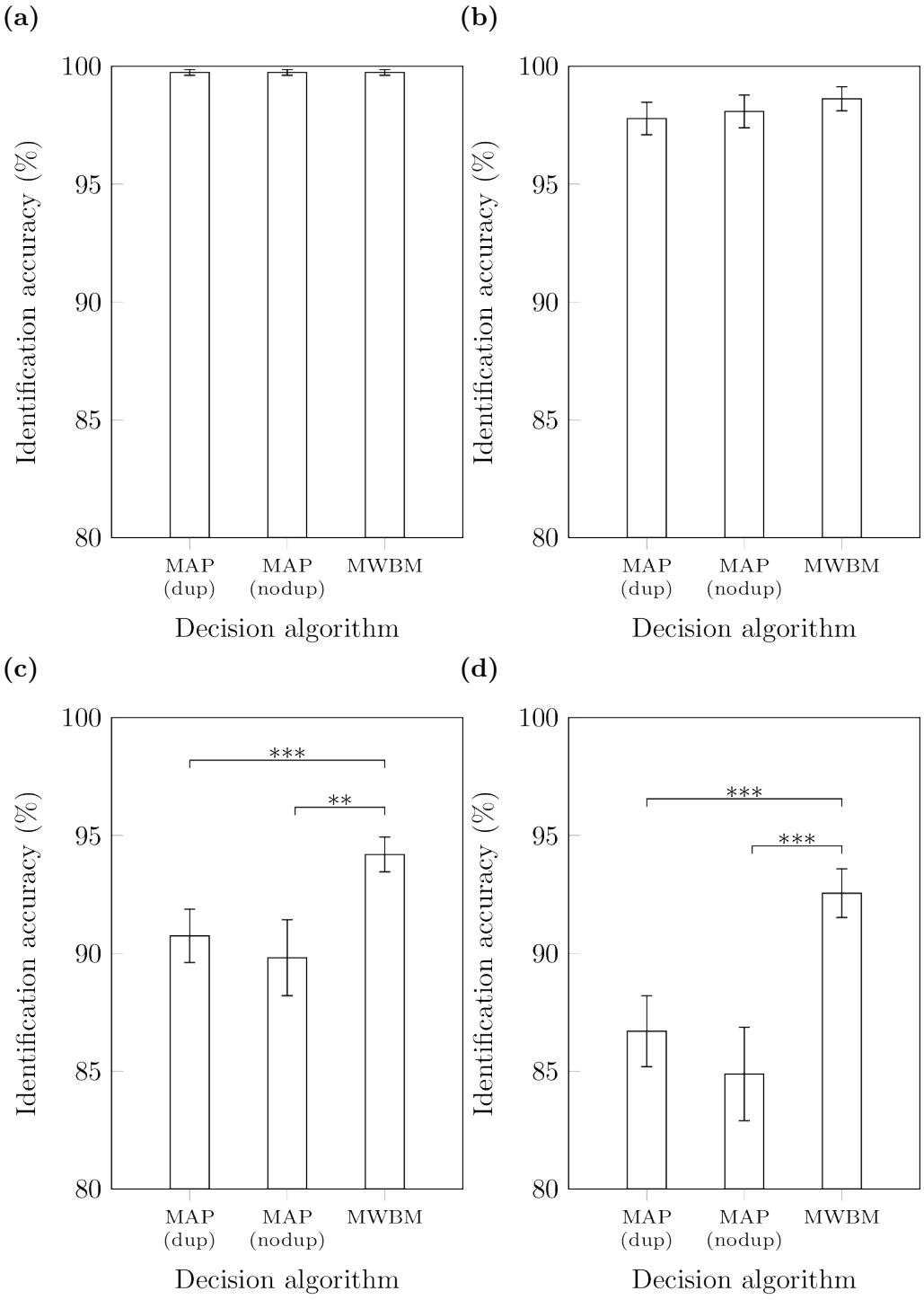
where $\alpha$ was set to 0.05. This procedure was also performed with log-transformed barcode features to estimate the proportion of those that were log-normally distributed. The results are summarized in Table B15.

**Table B15: Sphingolipid and glycerophospholipid datasets.**

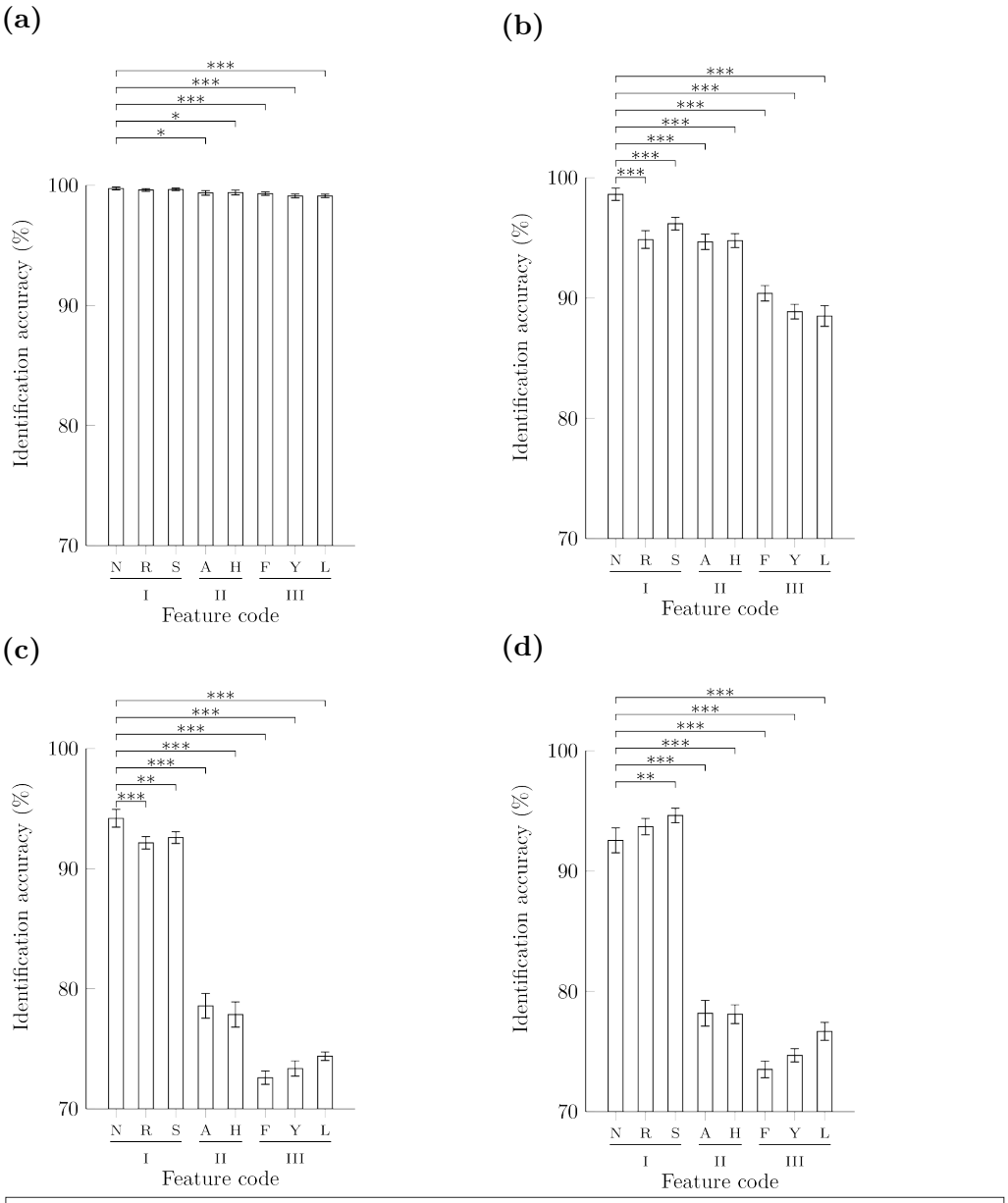
| | Barcode normality[1] (%) | | | |
|---|---|---|---|---|
| | Sphingolipid | | Glycerophospholipid | |
| Feature | Linear | Log | Linear | Log |
| Retention time | **6.88** | 5.62 | **5.85** | 5.85 |
| Area | 1.88 | **25.62** | 2.82 | **18.95** |
| Height | 1.25 | **29.38** | 2.82 | **18.95** |
| Full width at half maximum | **20.62** | 19.38 | 19.76 | **25.6** |
| Tailing factor | 16.25 | **33.12** | 13.1 | **29.84** |
| Asymmetry factor | 18.75 | **50** | 16.33 | **32.66** |
| Relative retention time | **1.88** | 1.25 | **2.42** | 2.22 |
| Subtracted retention time | **1.88** | 0.62 | **3.02** | 1.61 |

[1] Combined training and testing datasets.

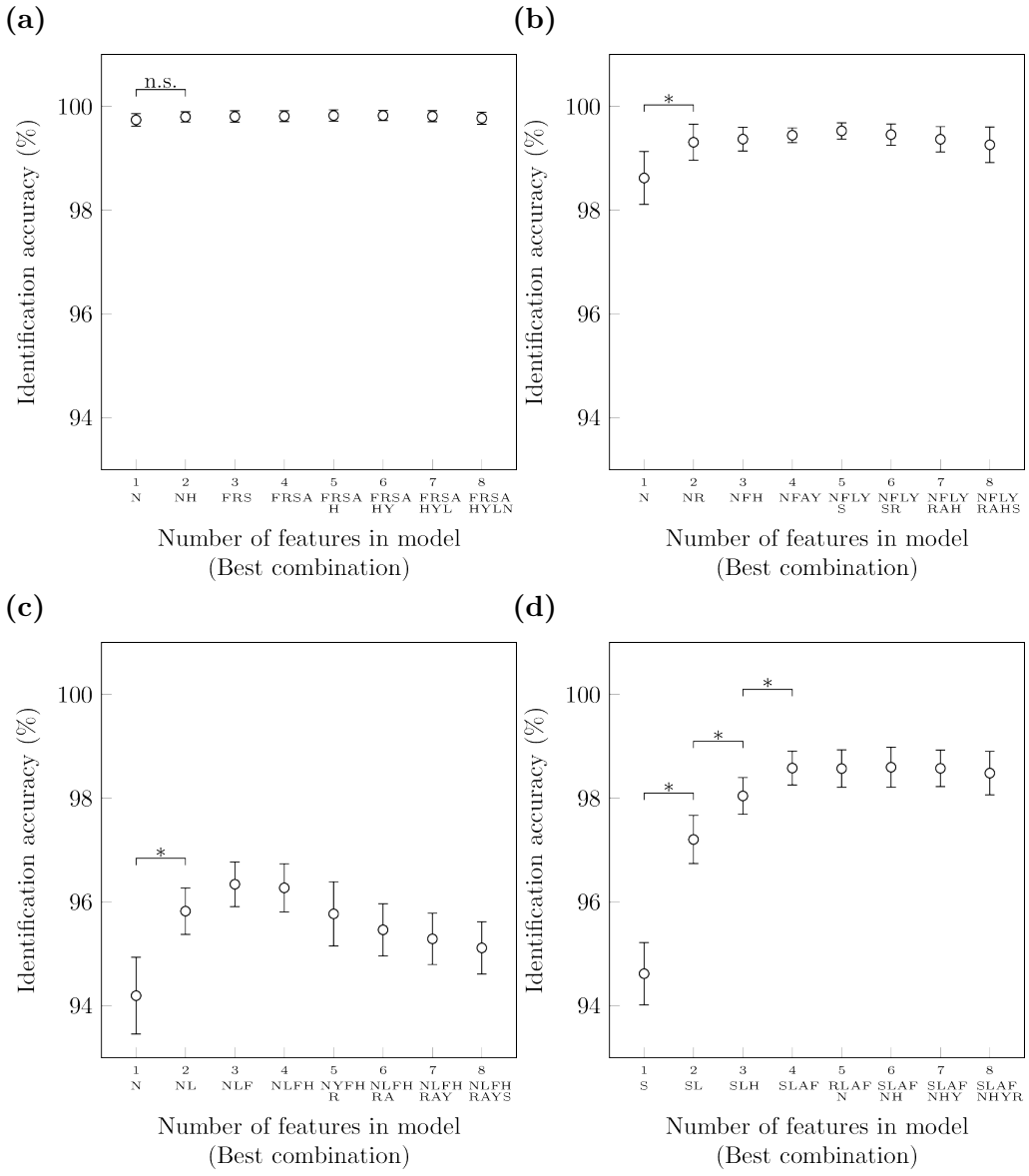
# Appendix: Matrix-specific classifier results



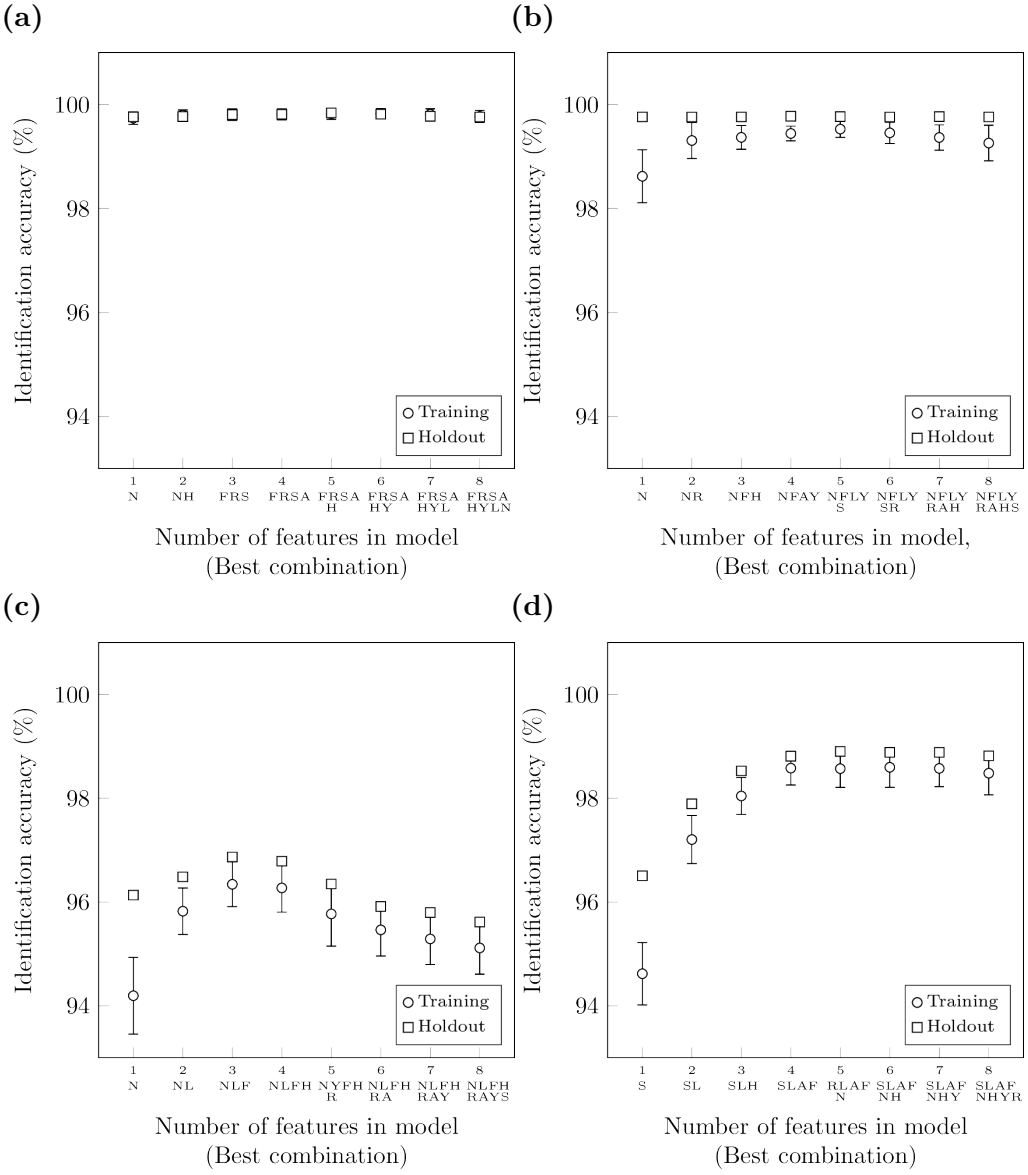

**Supplementary Figure C1: MWBM is the best performing decision rule.** Ten-fold cross validation identification accuracies of naïve Bayes classifier using the retention time feature ($^{*}Q < 0.05$, $^{**}Q < 0.01$, $^{***} < 0.001$, pairwise $t$-tests adjusted with the Benjamini-Hochberg method). 95% confidence intervals are shown. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.



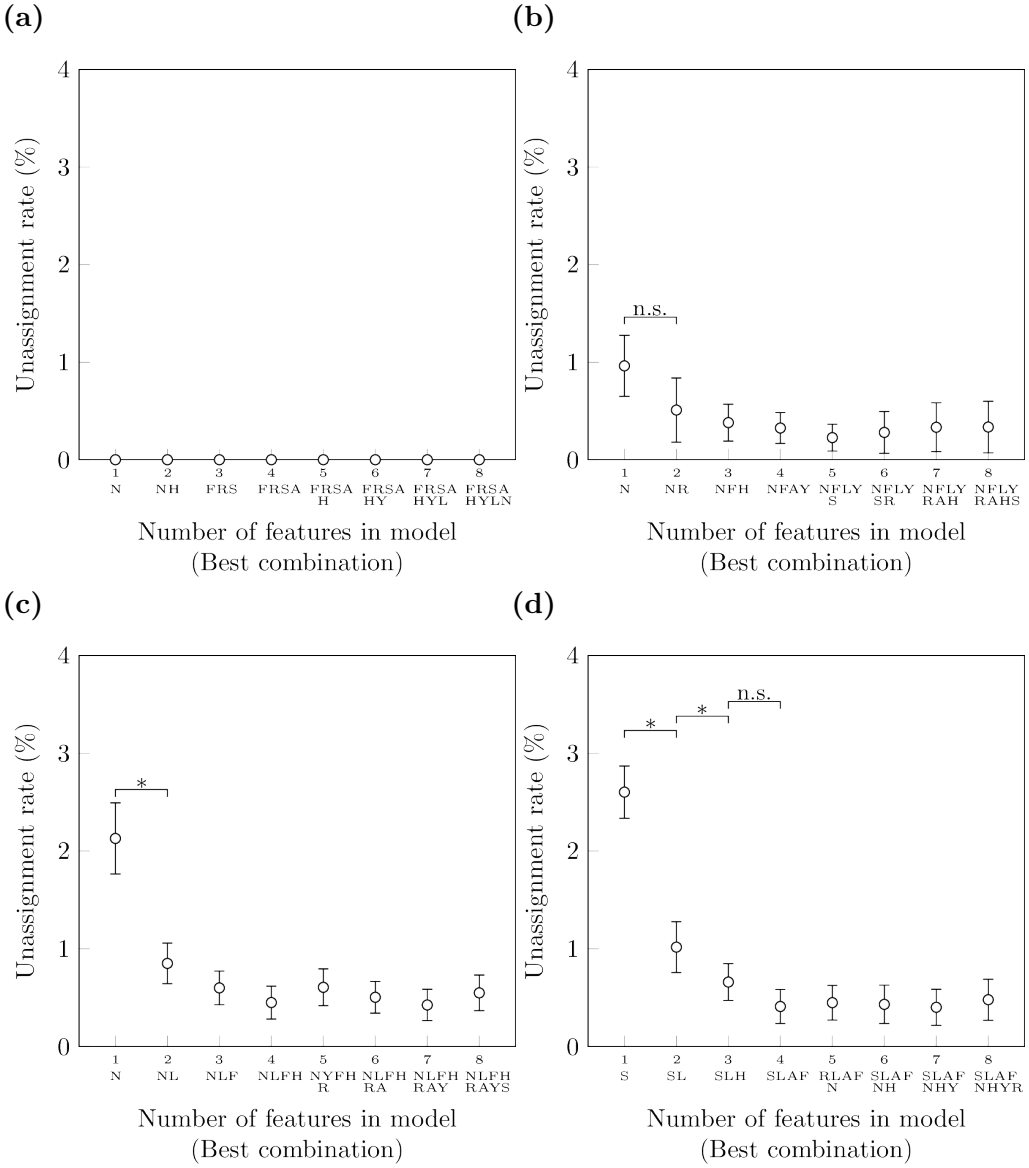

**Supplementary Figure C2: Retention time and subtracted retention time are the most discriminatory MRM features.** Ten-fold cross validation identification accuracies of naïve Bayes classifier trained on single features ($^{*}Q < 0.05$, $^{**}Q < 0.01$, $^{***} < 0.001$, pairwise $t$-tests against the retention time feature, adjusted with the Benjamini-Hochberg method). 95% confidence intervals are shown. I indicates retention time features; II indicates intensity features; III indicates peak shape features. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.



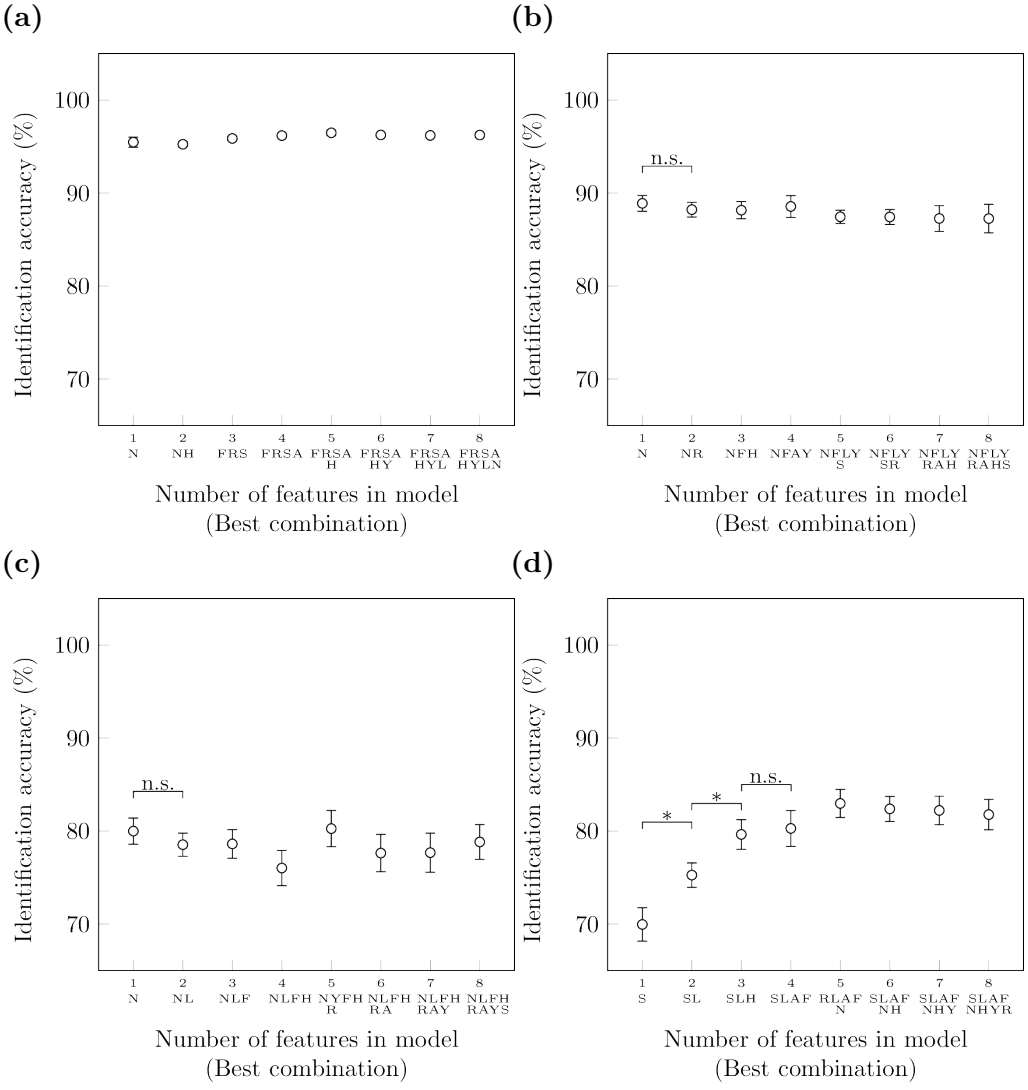

**Supplementary Figure C3: Additional features significantly improve classifier identification accuracies.** Significance testing between the best combination of $N$ and $N+1$ feature combinations for $N = (1, \ldots, 7)$ ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure C2. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.



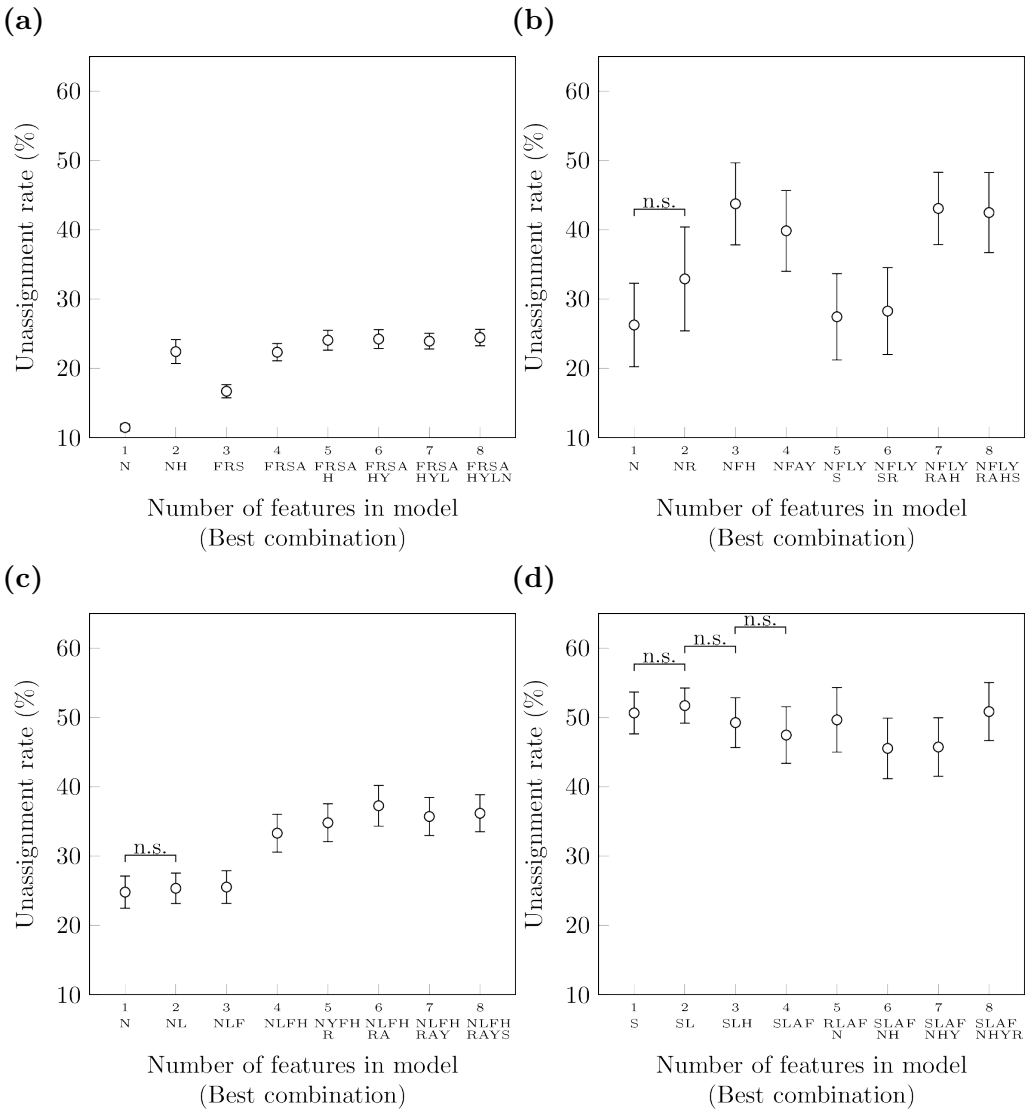

**Supplementary Figure C4: Best feature combinations identified from cross validation are validated by holdout datasets.** 95% confidence intervals are shown. Feature codes are defined in Figure C2. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.



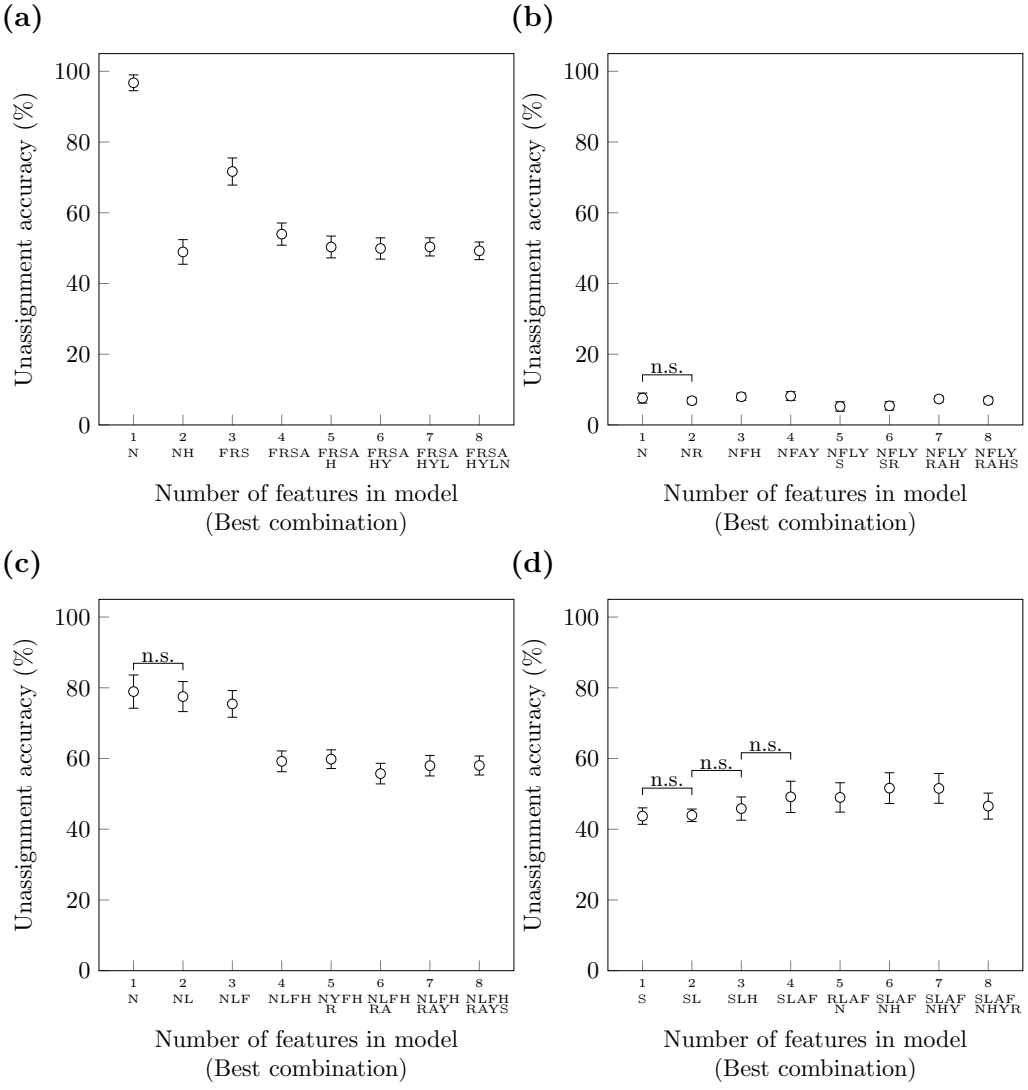

**Supplementary Figure C5: Increasing model complexity improves cross validation identification accuracies by decreasing the proportion of unassigned peaks.** Significance testing only performed on significant feature combinations identified in Figure C3 ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure C2. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.



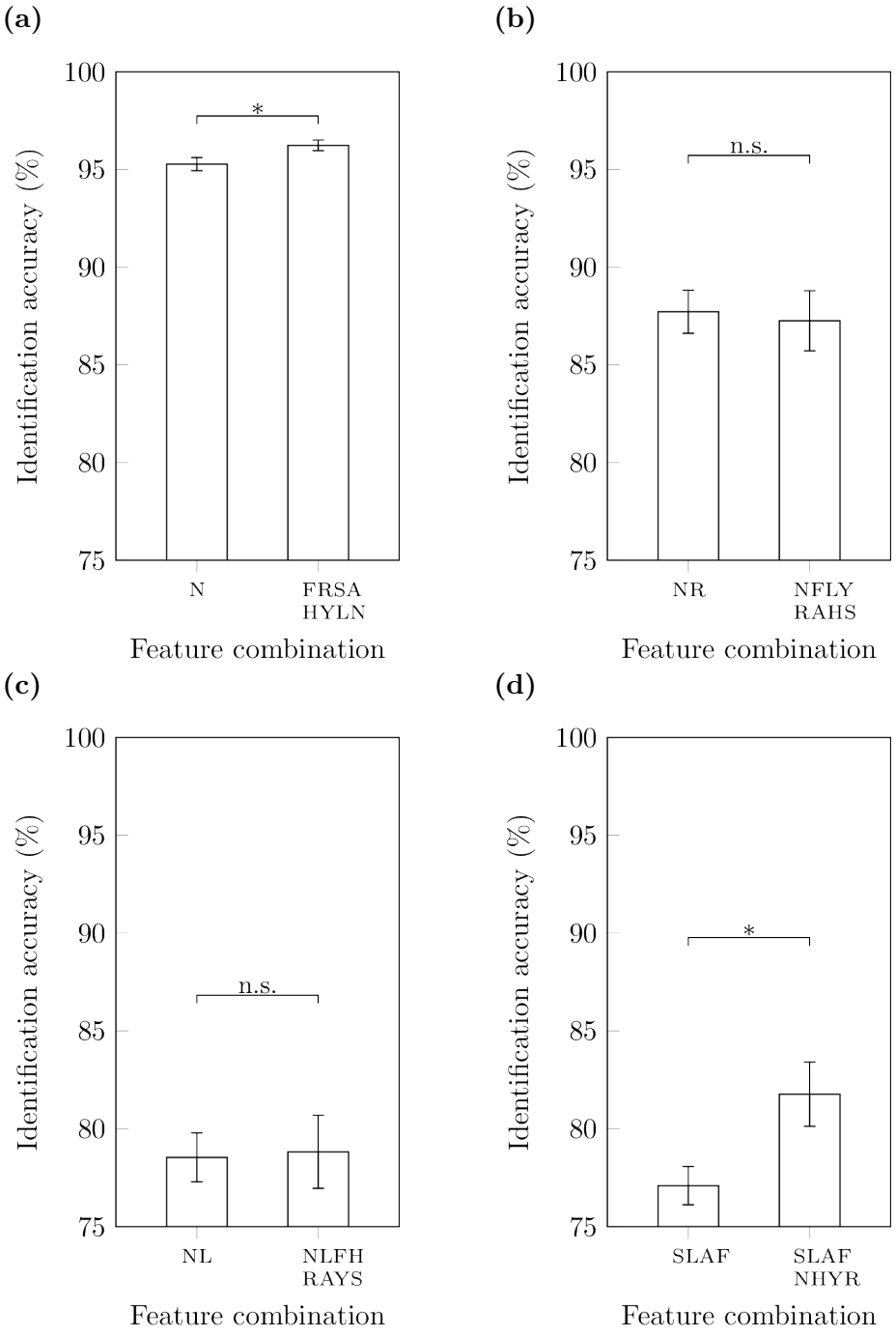

**Supplementary Figure C6: Classification of statistically independent testing datasets.** Significance testing only performed on significant feature combinations identified in Figure C3 ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure C2. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.




**Supplementary Figure C7: Expanding model complexity does not correlate with increased unassignment rates.** Significance testing only performed on significant feature combinations identified in Figure C6 ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure C2. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.




**Supplementary Figure C8: Increasing model complexity improves classifier identification accuracies by correctly unassigning more peaks.** Significance testing only performed on significant feature combinations identified in Figure C6 ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure C2. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.




**Supplementary Figure C9: Classifiers trained with all eight features display equal or greater identification accuracies on testing datasets than the optimal, least complex model identified from cross validation.** ($t$-test, $^*P < 0.05$). 95% confidence intervals are shown. Feature codes are defined in Figure C2. (a) Sphingolipid blood matrix dataset. (b) Sphingolipid brain matrix dataset. (c) Glycerophospholipid blood matrix dataset. (d) Glycerophospholipid brain matrix dataset.

# Appendix: Discovery datasets

**Table D16: Sphingolipid blood discovery dataset.**

| | Original | | Filtered | |
|---|---|---|---|---|
| Transition | [1]PI | Novel | PI | Novel |
| 371.9 / 264.3 | 0 | 2 | 0 | 2 |
| 380.9 / 264.3 | 0 | 4 | 0 | 4 |
| 385.8 / 264.3 | 0 | 5 | 0 | 5 |
| 392.8 / 264.3 | 0 | 7 | 0 | 7 |
| 394.8 / 264.3 | 0 | 3 | 0 | 2 |
| 398.8 / 264.3 | 0 | 7 | 0 | 7 |
| 399.9 / 264.3 | 0 | 13 | 0 | 13 |
| 401.9 / 264.3 | 0 | 3 | 0 | 3 |
| 407.8 / 264.3 | 0 | 15 | 0 | 14 |
| 408.9 / 264.3 | 0 | 6 | 0 | 3 |
| 520.9 / 264.3 | 0 | 6 | 0 | 6 |
| 538.9 / 264.3 | 1 | 1 | 1 | 1 |
| 564.6 / 264.3 | 0 | 7 | 0 | 7 |
| 566.7 / 264.3 | 2 | 4 | 2 | 4 |
| 568.1 / 264.3 | 2 | 4 | 2 | 3 |
| 580.6 / 264.3 | 1 | 0 | 1 | 0 |
| 588.8 / 264.3 | 2 | 11 | 2 | 11 |
| 594.7 / 264.3 | 2 | 2 | 2 | 2 |
| 604.9 / 264.3 | 2 | 12 | 2 | 10 |
| 608.8 / 264.3 | 2 | 3 | 2 | 3 |
| 618.7 / 264.3 | 2 | 7 | 0 | 7 |
| 620.8 / 264.3 | 2 | 7 | 2 | 5 |
| 622.9 / 264.3 | 2 | 10 | 2 | 9 |
| 632.9 / 264.3 | 0 | 14 | 0 | 10 |
| 634.7 / 264.3 | 2 | 11 | 1 | 7 |
| 637.0 / 264.3 | 2 | 8 | 2 | 8 |
| 637.8 / 264.3 | 2 | 2 | 2 | 2 |
| 644.8 / 264.3 | 0 | 9 | 0 | 8 |
| 647.0 / 264.3 | 2 | 8 | 2 | 7 |
| 648.9 / 264.3 | 2 | 14 | 2 | 13 |
| 650.9 / 264.3 | 2 | 33 | 1 | 30 |
| 661.8 / 264.3 | 2 | 13 | 2 | 13 |
| 662.8 / 264.3 | 2 | 8 | 2 | 4 |
| 664.9 / 264.3 | 4 | 2 | 4 | 2 |
| 666.8 / 264.3 | 2 | 10 | 1 | 10 |
| 667.2 / 264.3 | 0 | 13 | 0 | 12 |
| 668.8 / 264.3 | 0 | 7 | 0 | 5 |
| 672.7 / 264.3 | 2 | 4 | 2 | 4 |
| 675.9 / 264.3 | 2 | 9 | 2 | 9 |
| 676.8 / 264.3 | 2 | 8 | 2 | 8 |
| 678.8 / 264.3 | 2 | 9 | 2 | 8 |
| 682.7 / 264.3 | 0 | 16 | 0 | 14 |
| 689.6 / 264.3 | 0 | 10 | 0 | 9 |
| 698.9 / 264.3 | 0 | 7 | 0 | 6 |
| 700.9 / 264.3 | 2 | 4 | 2 | 4 |
| 703.9 / 264.3 | 2 | 4 | 2 | 4 |
| 717.5 / 264.3 | 2 | 8 | 1 | 7 |
| 726.7 / 264.3 | 2 | 4 | 2 | 4 |
| 729.0 / 264.3 | 0 | 6 | 0 | 5 |
| 729.9 / 264.3 | 2 | 5 | 1 | 4 |
| 731.8 / 264.3 | 2 | 3 | 1 | 3 |
| 744.7 / 264.3 | 0 | 5 | 0 | 5 |
| 750.1 / 264.3 | 2 | 6 | 2 | 6 |
| 757.1 / 264.3 | 2 | 3 | 2 | 3 |
| 758.1 / 264.3 | 0 | 2 | 0 | 1 |
| 759.6 / 264.3 | 2 | 7 | 2 | 7 |
| 772.7 / 264.3 | 2 | 12 | 2 | 11 |
| 780.7 / 264.3 | 2 | 8 | 2 | 8 |
| 782.8 / 264.3 | 2 | 15 | 2 | 13 |
| 784.9 / 264.3 | 1 | 14 | 1 | 13 |
| 787.6 / 264.3 | 2 | 13 | 2 | 13 |
| 796.7 / 264.3 | 2 | 12 | 2 | 10 |
| 798.7 / 264.3 | 2 | 13 | 2 | 12 |
| 800.7 / 264.3 | 2 | 12 | 2 | 12 |
| 801.8 / 264.3 | 0 | 10 | 0 | 9 |
| 806.8 / 264.3 | 0 | 6 | 0 | 6 |
| 808.7 / 264.3 | 2 | 9 | 2 | 5 |
| 810.8 / 264.3 | 2 | 13 | 2 | 12 |
| 812.8 / 264.3 | 2 | 19 | 2 | 18 |
| 813.9 / 264.3 | 2 | 21 | 2 | 20 |
| 814.7 / 264.3 | 2 | 15 | 2 | 9 |
| 815.6 / 264.3 | 2 | 17 | 2 | 12 |
| 819.9 / 264.3 | 2 | 9 | 2 | 8 |
| 824.8 / 264.3 | 0 | 17 | 0 | 16 |
| 826.8 / 264.3 | 2 | 9 | 2 | 9 |
| 827.1 / 264.3 | 0 | 9 | 0 | 9 |
| 828.8 / 264.3 | 2 | 18 | 2 | 15 |
| 834.7 / 264.3 | 2 | 6 | 2 | 6 |
| 838.8 / 264.3 | 0 | 7 | 0 | 7 |
| 840.8 / 264.3 | 0 | 10 | 0 | 10 |
| 845.9 / 264.3 | 2 | 10 | 2 | 9 |
| 848.0 / 264.3 | 2 | 11 | 2 | 11 |
| 850.0 / 264.3 | 2 | 15 | 2 | 15 |
| 860.0 / 264.3 | 2 | 11 | 2 | 11 |
| 862.0 / 264.3 | 2 | 20 | 2 | 20 |
| 862.9 / 264.3 | 2 | 15 | 1 | 12 |
| 870.0 / 264.3 | 2 | 4 | 2 | 4 |
| 872.0 / 264.3 | 2 | 9 | 2 | 6 |
| 874.0 / 264.3 | 2 | 19 | 2 | 18 |
| 876.0 / 264.3 | 2 | 17 | 2 | 16 |
| 878.1 / 264.3 | 2 | 14 | 2 | 14 |
| 884.0 / 264.3 | 2 | 6 | 2 | 6 |
| 885.2 / 264.3 | 2 | 9 | 2 | 7 |
| 887.9 / 264.3 | 1 | 26 | 1 | 26 |
| 890.0 / 264.3 | 1 | 17 | 1 | 16 |
| 896.0 / 264.3 | 2 | 7 | 2 | 7 |
| 898.0 / 264.3 | 2 | 14 | 2 | 14 |
| 899.0 / 264.3 | 0 | 24 | 0 | 21 |
| 900.0 / 264.3 | 2 | 23 | 2 | 21 |
| 901.0 / 264.3 | 0 | 35 | 0 | 31 |
| 902.0 / 264.3 | 2 | 21 | 2 | 19 |
| 904.0 / 264.3 | 2 | 11 | 2 | 11 |
| 911.1 / 264.3 | 2 | 14 | 2 | 14 |
| 913.1 / 264.3 | 2 | 12 | 2 | 10 |
| 921.9 / 264.3 | 0 | 8 | 0 | 8 |
| 923.8 / 264.3 | 0 | 16 | 0 | 16 |
| 925.8 / 264.3 | 0 | 17 | 0 | 17 |
| 927.8 / 264.3 | 0 | 20 | 0 | 19 |
| 930.0 / 264.3 | 0 | 14 | 0 | 14 |
| 947.0 / 264.3 | 2 | 8 | 2 | 7 |
| 971.1 / 264.3 | 1 | 6 | 1 | 6 |
| 973.1 / 264.3 | 2 | 10 | 2 | 9 |
| 974.9 / 264.3 | 0 | 15 | 0 | 10 |
| 975.1 / 264.3 | 2 | 9 | 0 | 4 |
| 990.1 / 264.3 | 2 | 5 | 2 | 5 |
| 991.9 / 264.3 | 2 | 7 | 2 | 6 |
| 1025.0 / 264.3 | 2 | 6 | 1 | 6 |
| 1041.5 / 264.3 | 2 | 4 | 2 | 4 |
| 1109.0 / 264.3 | 1 | 2 | 1 | 2 |
| 1135.1 / 264.3 | 1 | 2 | 1 | 2 |
| 1151.9 / 264.3 | 2 | 5 | 2 | 5 |
| 1154.0 / 264.3 | 2 | 9 | 2 | 8 |
| 1170.7 / 264.3 | 1 | 3 | 1 | 3 |

[1]Previously identified peak

**Table D17: Sphingolipid brain discovery dataset.**

| | Original | | Filtered | |
|---|---|---|---|---|
| Transition | [1]PI | Novel | PI | Novel |
| 346.6 / 264.3 | 2 | 6 | 2 | 6 |
| 371.9 / 264.3 | 0 | 1 | 0 | 1 |
| 385.8 / 264.3 | 1 | 0 | 1 | 0 |
| 398.8 / 264.3 | 0 | 2 | 0 | 2 |
| 399.9 / 264.3 | 0 | 5 | 0 | 5 |
| 408.9 / 264.3 | 0 | 3 | 0 | 3 |
| 510.7 / 264.3 | 2 | 1 | 2 | 1 |
| 520.9 / 264.3 | 0 | 14 | 0 | 13 |
| 538.9 / 264.3 | 2 | 5 | 2 | 5 |
| 564.6 / 264.3 | 2 | 22 | 1 | 21 |
| 566.7 / 264.3 | 2 | 38 | 2 | 35 |
| 568.1 / 264.3 | 2 | 10 | 2 | 8 |
| 580.6 / 264.3 | 2 | 8 | 2 | 8 |
| 588.8 / 264.3 | 2 | 24 | 2 | 24 |
| 594.7 / 264.3 | 2 | 26 | 2 | 23 |
| 604.9 / 264.3 | 2 | 10 | 2 | 10 |
| 608.8 / 264.3 | 2 | 10 | 2 | 9 |
| 618.7 / 264.3 | 2 | 28 | 2 | 24 |
| 620.8 / 264.3 | 2 | 28 | 1 | 26 |
| 622.9 / 264.3 | 2 | 12 | 2 | 9 |
| 632.9 / 264.3 | 2 | 26 | 0 | 20 |
| 634.7 / 264.3 | 2 | 30 | 2 | 25 |
| 637.0 / 264.3 | 2 | 13 | 2 | 10 |
| 637.8 / 264.3 | 0 | 15 | 0 | 11 |
| 644.8 / 264.3 | 2 | 16 | 1 | 14 |
| 647.0 / 264.3 | 2 | 15 | 1 | 11 |
| 648.9 / 264.3 | 2 | 24 | 1 | 19 |
| 650.9 / 264.3 | 0 | 23 | 0 | 17 |
| 661.8 / 264.3 | 0 | 16 | 0 | 15 |
| 662.8 / 264.3 | 2 | 9 | 2 | 4 |
| 664.9 / 264.3 | 2 | 24 | 2 | 23 |
| 666.8 / 264.3 | 1 | 29 | 0 | 26 |
| 667.2 / 264.3 | 0 | 15 | 0 | 15 |
| 668.8 / 264.3 | 0 | 35 | 0 | 30 |
| 672.7 / 264.3 | 0 | 9 | 0 | 8 |
| 675.9 / 264.3 | 0 | 10 | 0 | 8 |
| 676.8 / 264.3 | 0 | 15 | 0 | 13 |
| 678.8 / 264.3 | 0 | 12 | 0 | 11 |
| 682.7 / 264.3 | 2 | 17 | 1 | 14 |
| 689.6 / 264.3 | 2 | 9 | 2 | 8 |
| 698.9 / 264.3 | 0 | 8 | 0 | 8 |
| 700.9 / 264.3 | 2 | 22 | 2 | 21 |
| 703.9 / 264.3 | 2 | 8 | 2 | 8 |
| 717.5 / 264.3 | 0 | 3 | 0 | 3 |
| 726.7 / 264.3 | 2 | 23 | 0 | 22 |
| 729.0 / 264.3 | 2 | 22 | 2 | 21 |
| 729.9 / 264.3 | 0 | 26 | 0 | 22 |
| 731.8 / 264.3 | 2 | 26 | 2 | 22 |
| 744.7 / 264.3 | 2 | 22 | 2 | 22 |
| 750.1 / 264.3 | 0 | 6 | 0 | 6 |
| 757.1 / 264.3 | 2 | 5 | 2 | 5 |
| 758.1 / 264.3 | 0 | 14 | 0 | 13 |
| 759.6 / 264.3 | 2 | 10 | 2 | 10 |
| 770.7 / 264.3 | 2 | 28 | 2 | 27 |
| 772.7 / 264.3 | 2 | 30 | 2 | 28 |
| 780.7 / 264.3 | 0 | 31 | 0 | 26 |
| 782.8 / 264.3 | 2 | 38 | 1 | 31 |
| 784.9 / 264.3 | 2 | 38 | 2 | 37 |
| 787.6 / 264.3 | 2 | 26 | 2 | 24 |
| 796.7 / 264.3 | 2 | 38 | 2 | 34 |
| 798.7 / 264.3 | 4 | 36 | 4 | 33 |
| 800.7 / 264.3 | 2 | 38 | 2 | 34 |
| 801.8 / 264.3 | 0 | 40 | 0 | 36 |
| 806.8 / 264.3 | 0 | 33 | 0 | 32 |
| 808.7 / 264.3 | 2 | 38 | 2 | 34 |
| 810.8 / 264.3 | 2 | 38 | 2 | 35 |
| 812.8 / 264.3 | 4 | 36 | 4 | 33 |
| 813.9 / 264.3 | 0 | 40 | 0 | 37 |
| 814.7 / 264.3 | 2 | 38 | 2 | 33 |
| 815.6 / 264.3 | 0 | 26 | 0 | 20 |
| 819.9 / 264.3 | 0 | 13 | 0 | 13 |
| 824.8 / 264.3 | 2 | 24 | 2 | 22 |
| 826.8 / 264.3 | 4 | 44 | 4 | 42 |
| 827.1 / 264.3 | 0 | 29 | 0 | 22 |
| 828.8 / 264.3 | 2 | 38 | 1 | 33 |
| 834.7 / 264.3 | 0 | 32 | 0 | 31 |
| 838.8 / 264.3 | 2 | 32 | 2 | 32 |
| 840.8 / 264.3 | 4 | 26 | 4 | 25 |
| 845.9 / 264.3 | 2 | 26 | 2 | 26 |
| 848.0 / 264.3 | 0 | 12 | 0 | 12 |
| 850.0 / 264.3 | 0 | 16 | 0 | 13 |
| 860.0 / 264.3 | 2 | 17 | 2 | 15 |
| 862.0 / 264.3 | 0 | 17 | 0 | 16 |
| 862.9 / 264.3 | 2 | 26 | 2 | 25 |
| 870.0 / 264.3 | 2 | 14 | 2 | 12 |
| 872.0 / 264.3 | 0 | 21 | 0 | 18 |
| 874.0 / 264.3 | 0 | 21 | 0 | 20 |
| 876.0 / 264.3 | 0 | 22 | 0 | 21 |
| 878.1 / 264.3 | 0 | 23 | 0 | 21 |
| 884.0 / 264.3 | 0 | 16 | 0 | 15 |
| 885.2 / 264.3 | 0 | 13 | 0 | 12 |
| 887.9 / 264.3 | 0 | 28 | 0 | 25 |
| 890.0 / 264.3 | 2 | 31 | 2 | 30 |
| 896.0 / 264.3 | 0 | 19 | 0 | 19 |
| 898.0 / 264.3 | 0 | 26 | 0 | 25 |
| 899.0 / 264.3 | 0 | 17 | 0 | 13 |
| 900.0 / 264.3 | 0 | 21 | 0 | 18 |
| 901.0 / 264.3 | 0 | 17 | 0 | 15 |
| 902.0 / 264.3 | 0 | 21 | 0 | 20 |
| 904.0 / 264.3 | 0 | 16 | 0 | 12 |
| 911.1 / 264.3 | 2 | 11 | 2 | 11 |
| 913.1 / 264.3 | 2 | 14 | 2 | 12 |
| 921.9 / 264.3 | 2 | 19 | 2 | 19 |
| 923.8 / 264.3 | 0 | 17 | 0 | 17 |
| 925.8 / 264.3 | 2 | 31 | 2 | 30 |
| 927.8 / 264.3 | 2 | 22 | 2 | 22 |
| 930.0 / 264.3 | 0 | 19 | 0 | 17 |
| 947.0 / 264.3 | 2 | 13 | 2 | 13 |
| 971.1 / 264.3 | 0 | 10 | 0 | 9 |
| 973.1 / 264.3 | 2 | 11 | 2 | 11 |
| 974.9 / 264.3 | 0 | 25 | 0 | 23 |
| 975.1 / 264.3 | 0 | 18 | 0 | 14 |
| 990.1 / 264.3 | 0 | 9 | 0 | 8 |
| 991.9 / 264.3 | 0 | 21 | 0 | 19 |
| 1025.0 / 264.3 | 4 | 17 | 4 | 16 |
| 1041.5 / 264.3 | 2 | 13 | 2 | 13 |
| 1109.0 / 264.3 | 0 | 15 | 0 | 15 |
| 1135.1 / 264.3 | 0 | 4 | 0 | 4 |
| 1151.9 / 264.3 | 0 | 4 | 0 | 4 |
| 1154.0 / 264.3 | 0 | 9 | 0 | 9 |
| 1170.7 / 264.3 | 0 | 8 | 0 | 8 |
| 1176.0 / 264.3 | 0 | 4 | 0 | 4 |
| 1193.0 / 264.3 | 0 | 2 | 0 | 2 |

[1]Previously identified peak

**Table D18: Glycerophospholipid blood discovery dataset.**

| | Original | | Filtered | |
|---|---|---|---|---|
| Transition | [1]PI | Novel | PI | Novel |
| 454.3 / 184.1 | 2 | 0 | 2 | 0 |
| 468.3 / 184.1 | 4 | 6 | 4 | 6 |
| 480.4 / 184.1 | 2 | 1 | 2 | 1 |
| 482.4 / 184.1 | 6 | 4 | 6 | 4 |
| 486.8 / 184.1 | 2 | 1 | 2 | 1 |
| 494.3 / 184.1 | 4 | 11 | 4 | 11 |
| 496.3 / 184.1 | 4 | 7 | 4 | 7 |
| 506.5 / 184.1 | 4 | 3 | 4 | 3 |
| 508.3 / 184.1 | 6 | 2 | 6 | 2 |
| 514.5 / 184.1 | 2 | 7 | 2 | 7 |
| 518.4 / 184.1 | 4 | 9 | 4 | 9 |
| 520.3 / 184.1 | 4 | 9 | 4 | 9 |
| 524.4 / 184.1 | 4 | 10 | 4 | 9 |
| 528.3 / 184.1 | 2 | 11 | 2 | 11 |
| 529.4 / 184.1 | 2 | 7 | 2 | 4 |
| 538.3 / 184.1 | 4 | 11 | 4 | 11 |
| 540.4 / 184.1 | 2 | 10 | 2 | 9 |
| 542.3 / 184.1 | 4 | 9 | 4 | 9 |
| 544.4 / 184.1 | 4 | 6 | 4 | 6 |
| 546.3 / 184.1 | 6 | 7 | 6 | 3 |
| 548.4 / 184.1 | 4 | 6 | 4 | 3 |
| 550.4 / 184.1 | 4 | 7 | 4 | 7 |
| 552.4 / 184.1 | 4 | 13 | 4 | 9 |
| 556.4 / 184.1 | 2 | 5 | 2 | 5 |
| 564.5 / 184.1 | 0 | 12 | 0 | 12 |
| 568.4 / 184.1 | 4 | 6 | 4 | 6 |
| 570.3 / 184.1 | 8 | 7 | 8 | 4 |
| 572.3 / 184.1 | 6 | 1 | 5 | 1 |
| 578.3 / 184.1 | 6 | 15 | 6 | 15 |
| 580.4 / 184.1 | 6 | 6 | 6 | 5 |
| 584.4 / 184.1 | 2 | 13 | 2 | 13 |
| 592.5 / 184.1 | 2 | 20 | 2 | 20 |
| 594.3 / 184.1 | 4 | 4 | 3 | 4 |
| 595.5 / 184.1 | 2 | 14 | 2 | 8 |
| 596.4 / 184.1 | 4 | 7 | 3 | 6 |
| 598.4 / 184.1 | 2 | 5 | 2 | 3 |
| 606.4 / 184.1 | 4 | 7 | 4 | 7 |
| 608.5 / 184.1 | 8 | 12 | 8 | 12 |
| 610.3 / 184.1 | 2 | 14 | 2 | 14 |
| 614.5 / 184.1 | 6 | 6 | 5 | 6 |
| 619.5 / 184.1 | 4 | 17 | 4 | 17 |
| 620.4 / 184.1 | 4 | 10 | 2 | 7 |
| 622.4 / 184.1 | 6 | 4 | 6 | 4 |
| 624.4 / 184.1 | 2 | 13 | 2 | 12 |
| 632.6 / 184.1 | 2 | 7 | 2 | 7 |
| 634.6 / 184.1 | 10 | 5 | 9 | 5 |
| 636.6 / 184.1 | 2 | 1 | 2 | 0 |
| 647.6 / 184.1 | 0 | 9 | 0 | 9 |
| 661.7 / 184.1 | 0 | 17 | 0 | 17 |
| 673.7 / 184.1 | 2 | 5 | 2 | 5 |
| 675.6 / 184.1 | 2 | 18 | 2 | 18 |
| 677.7 / 184.1 | 0 | 17 | 0 | 15 |
| 689.6 / 184.1 | 2 | 36 | 2 | 36 |
| 692.6 / 184.1 | 0 | 25 | 0 | 25 |
| 701.7 / 184.1 | 2 | 28 | 2 | 28 |
| 703.5 / 184.1 | 2 | 63 | 1 | 61 |
| 705.7 / 184.1 | 2 | 48 | 2 | 42 |
| 706.6 / 184.1 | 2 | 44 | 2 | 38 |
| 715.8 / 184.1 | 0 | 26 | 0 | 26 |
| 717.7 / 184.1 | 4 | 49 | 4 | 49 |
| 718.8 / 184.1 | 2 | 39 | 2 | 38 |
| 719.7 / 184.1 | 2 | 38 | 2 | 37 |
| 720.7 / 184.1 | 4 | 42 | 4 | 40 |
| 727.7 / 184.1 | 2 | 24 | 2 | 24 |
| 729.7 / 184.1 | 2 | 28 | 2 | 25 |
| 730.7 / 184.1 | 2 | 33 | 2 | 31 |
| 731.7 / 184.1 | 2 | 29 | 2 | 28 |
| 732.7 / 184.1 | 2 | 32 | 2 | 28 |
| 733.8 / 184.1 | 0 | 41 | 0 | 34 |
| 734.7 / 184.1 | 2 | 32 | 2 | 26 |
| 738.7 / 184.1 | 2 | 37 | 2 | 36 |
| 739.8 / 184.1 | 2 | 41 | 2 | 41 |
| 742.7 / 184.1 | 4 | 70 | 4 | 69 |
| 744.7 / 184.1 | 4 | 52 | 4 | 50 |
| 746.6 / 184.1 | 2 | 48 | 2 | 46 |
| 748.7 / 184.1 | 0 | 31 | 0 | 31 |
| 752.9 / 184.1 | 4 | 36 | 4 | 36 |
| 753.6 / 184.1 | 2 | 56 | 2 | 53 |
| 754.5 / 184.1 | 8 | 49 | 7 | 46 |
| 756.6 / 184.1 | 10 | 40 | 10 | 37 |
| 757.7 / 184.1 | 0 | 56 | 0 | 47 |
| 758.6 / 184.1 | 2 | 73 | 2 | 63 |
| 759.8 / 184.1 | 2 | 71 | 2 | 63 |
| 760.6 / 184.1 | 2 | 56 | 1 | 44 |
| 762.7 / 184.1 | 2 | 44 | 2 | 40 |
| 764.7 / 184.1 | 0 | 35 | 0 | 33 |
| 766.7 / 184.1 | 0 | 42 | 0 | 42 |
| 768.6 / 184.1 | 4 | 49 | 4 | 47 |
| 769.3 / 184.1 | 2 | 61 | 1 | 58 |
| 770.6 / 184.1 | 2 | 35 | 2 | 31 |
| 772.6 / 184.1 | 6 | 60 | 5 | 57 |
| 773.8 / 184.1 | 0 | 40 | 0 | 37 |
| 774.6 / 184.1 | 6 | 51 | 6 | 48 |
| 775.9 / 184.1 | 2 | 42 | 2 | 39 |
| 776.7 / 184.1 | 0 | 37 | 0 | 30 |
| 777.8 / 184.1 | 2 | 34 | 2 | 32 |
| 778.6 / 184.1 | 0 | 37 | 0 | 32 |
| 780.6 / 184.1 | 8 | 36 | 8 | 35 |
| 782.4 / 184.1 | 6 | 45 | 6 | 43 |
| 784.6 / 184.1 | 4 | 71 | 4 | 62 |
| 785.8 / 184.1 | 4 | 70 | 4 | 62 |
| 787.8 / 184.1 | 2 | 65 | 2 | 59 |
| 788.7 / 184.1 | 2 | 57 | 2 | 53 |
| 790.6 / 184.1 | 2 | 26 | 2 | 22 |

[1]Previously identified peak

**Table D18: Glycerophospholipid blood discovery dataset, cont'd.**

| | Original | | Filtered | |
|---|---|---|---|---|
| Transition | PI | Novel | PI | Novel |
| 791.3 / 184.1 | 2 | 25 | 2 | 16 |
| 792.7 / 184.1 | 1 | 31 | 1 | 24 |
| 794.6 / 184.1 | 2 | 55 | 2 | 49 |
| 795.5 / 184.1 | 2 | 45 | 1 | 45 |
| 796.6 / 184.1 | 2 | 56 | 2 | 55 |
| 798.7 / 184.1 | 2 | 69 | 2 | 64 |
| 799.8 / 184.1 | 6 | 65 | 6 | 61 |
| 800.7 / 184.1 | 6 | 62 | 6 | 56 |
| 801.8 / 184.1 | 0 | 58 | 0 | 54 |
| 803.8 / 184.1 | 2 | 31 | 2 | 29 |
| 804.4 / 184.1 | 2 | 32 | 2 | 29 |
| 806.6 / 184.1 | 4 | 30 | 4 | 29 |
| 808.6 / 184.1 | 4 | 51 | 4 | 46 |
| 810.7 / 184.1 | 4 | 64 | 4 | 62 |
| 811.8 / 184.1 | 2 | 72 | 2 | 64 |
| 812.7 / 184.1 | 4 | 72 | 4 | 62 |
| 813.8 / 184.1 | 2 | 69 | 2 | 62 |
| 814.6 / 184.1 | 2 | 75 | 2 | 61 |
| 815.6 / 184.1 | 1 | 62 | 1 | 48 |
| 816.6 / 184.1 | 4 | 35 | 4 | 31 |
| 819.4 / 184.1 | 2 | 16 | 1 | 15 |
| 820.8 / 184.1 | 0 | 53 | 0 | 51 |
| 823.8 / 184.1 | 0 | 56 | 0 | 51 |
| 827.9 / 184.1 | 2 | 45 | 2 | 44 |
| 830.6 / 184.1 | 2 | 37 | 2 | 33 |
| 831.8 / 184.1 | 0 | 43 | 0 | 40 |
| 834.7 / 184.1 | 7 | 29 | 7 | 25 |
| 836.7 / 184.1 | 4 | 42 | 3 | 42 |
| 838.7 / 184.1 | 4 | 51 | 4 | 48 |
| 839.6 / 184.1 | 0 | 58 | 0 | 47 |
| 842.7 / 184.1 | 6 | 26 | 6 | 23 |
| 844.8 / 184.1 | 2 | 29 | 2 | 28 |
| 846.6 / 184.1 | 0 | 42 | 0 | 40 |
| 848.8 / 184.1 | 0 | 49 | 0 | 47 |
| 850.7 / 184.1 | 0 | 45 | 0 | 42 |
| 854.6 / 184.1 | 0 | 25 | 0 | 24 |
| 857.7 / 184.1 | 2 | 44 | 2 | 44 |
| 858.5 / 184.1 | 0 | 31 | 0 | 29 |
| 859.7 / 184.1 | 2 | 39 | 2 | 37 |
| 860.7 / 184.1 | 0 | 47 | 0 | 46 |
| 862.7 / 184.1 | 0 | 40 | 0 | 37 |
| 864.7 / 184.1 | 1 | 36 | 1 | 34 |
| 874.7 / 184.1 | 2 | 28 | 2 | 28 |
| 876.7 / 184.1 | 0 | 50 | 0 | 49 |
| 878.8 / 184.1 | 0 | 39 | 0 | 38 |
| 881.7 / 184.1 | 2 | 51 | 2 | 49 |
| 885.8 / 184.1 | 2 | 55 | 2 | 53 |
| 887.7 / 184.1 | 0 | 42 | 0 | 42 |
| 902.7 / 184.1 | 0 | 37 | 0 | 37 |
| 904.7 / 184.1 | 0 | 32 | 0 | 30 |
| 909.8 / 184.1 | 2 | 39 | 2 | 39 |

**Table D19: Glycerophospholipid brain discovery dataset.**

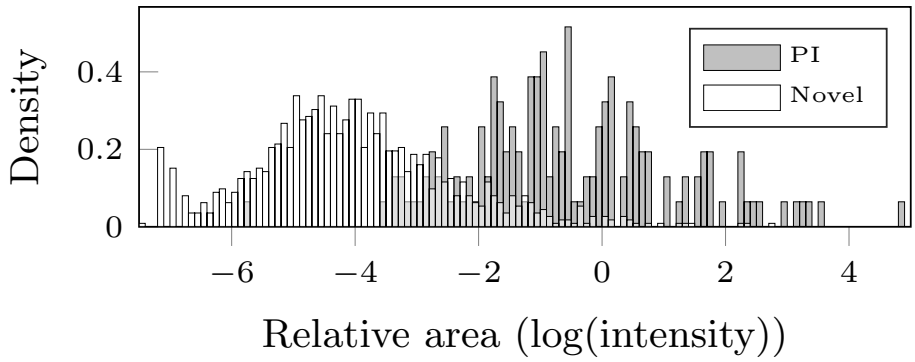
| | Original | | Filtered | |
|---|---|---|---|---|
| Transition | [1]PI | Novel | PI | Novel |
| 454.3 / 184.1 | 2 | 0 | 2 | 0 |
| 468.3 / 184.1 | 4 | 5 | 4 | 5 |
| 480.4 / 184.1 | 2 | 3 | 2 | 3 |
| 482.4 / 184.1 | 6 | 5 | 6 | 5 |
| 486.8 / 184.1 | 2 | 0 | 2 | 0 |
| 494.3 / 184.1 | 4 | 6 | 4 | 6 |
| 496.3 / 184.1 | 4 | 5 | 4 | 5 |
| 506.5 / 184.1 | 2 | 4 | 2 | 4 |
| 508.3 / 184.1 | 10 | 4 | 9 | 4 |
| 510.4 / 184.1 | 8 | 3 | 8 | 3 |
| 512.4 / 184.1 | 4 | 9 | 4 | 6 |
| 514.5 / 184.1 | 2 | 2 | 2 | 2 |
| 518.4 / 184.1 | 4 | 2 | 4 | 2 |
| 520.3 / 184.1 | 4 | 4 | 4 | 4 |
| 522.4 / 184.1 | 4 | 8 | 4 | 8 |
| 524.4 / 184.1 | 4 | 5 | 4 | 2 |
| 528.3 / 184.1 | 2 | 11 | 2 | 10 |
| 529.4 / 184.1 | 2 | 7 | 0 | 4 |
| 536.4 / 184.1 | 6 | 3 | 6 | 3 |
| 538.3 / 184.1 | 6 | 6 | 6 | 6 |
| 540.4 / 184.1 | 2 | 1 | 2 | 1 |
| 542.3 / 184.1 | 2 | 8 | 2 | 8 |
| 544.4 / 184.1 | 2 | 13 | 2 | 13 |
| 546.3 / 184.1 | 6 | 5 | 6 | 3 |
| 548.4 / 184.1 | 4 | 5 | 4 | 1 |
| 550.4 / 184.1 | 5 | 4 | 5 | 4 |
| 552.4 / 184.1 | 4 | 12 | 4 | 8 |
| 556.4 / 184.1 | 2 | 10 | 2 | 10 |
| 564.5 / 184.1 | 2 | 8 | 2 | 8 |
| 566.4 / 184.1 | 8 | 8 | 8 | 8 |
| 568.4 / 184.1 | 2 | 13 | 2 | 12 |
| 570.3 / 184.1 | 6 | 9 | 6 | 6 |
| 572.3 / 184.1 | 4 | 4 | 4 | 2 |
| 580.4 / 184.1 | 6 | 6 | 6 | 6 |
| 584.4 / 184.1 | 0 | 8 | 0 | 8 |
| 592.5 / 184.1 | 2 | 17 | 2 | 17 |
| 594.3 / 184.1 | 4 | 7 | 4 | 7 |
| 595.5 / 184.1 | 0 | 17 | 0 | 13 |
| 596.4 / 184.1 | 4 | 10 | 4 | 8 |
| 598.4 / 184.1 | 4 | 4 | 4 | 2 |
| 606.4 / 184.1 | 6 | 14 | 6 | 14 |
| 610.3 / 184.1 | 2 | 13 | 2 | 12 |
| 614.5 / 184.1 | 4 | 7 | 4 | 6 |
| 619.5 / 184.1 | 2 | 3 | 2 | 3 |
| 620.4 / 184.1 | 4 | 9 | 4 | 9 |
| 622.4 / 184.1 | 4 | 4 | 4 | 3 |
| 624.4 / 184.1 | 4 | 8 | 3 | 4 |
| 632.6 / 184.1 | 2 | 14 | 2 | 14 |
| 634.6 / 184.1 | 6 | 11 | 6 | 11 |
| 636.6 / 184.1 | 2 | 3 | 2 | 2 |
| 647.6 / 184.1 | 0 | 6 | 0 | 6 |
| 661.7 / 184.1 | 0 | 8 | 0 | 8 |
| 673.7 / 184.1 | 0 | 7 | 0 | 7 |
| 675.6 / 184.1 | 2 | 8 | 2 | 8 |
| 677.7 / 184.1 | 0 | 21 | 0 | 20 |
| 689.6 / 184.1 | 0 | 21 | 0 | 21 |
| 692.6 / 184.1 | 2 | 37 | 2 | 37 |
| 701.7 / 184.1 | 2 | 6 | 2 | 6 |
| 703.5 / 184.1 | 2 | 45 | 2 | 44 |
| 705.7 / 184.1 | 4 | 66 | 4 | 63 |
| 706.6 / 184.1 | 2 | 62 | 2 | 55 |
| 715.8 / 184.1 | 0 | 31 | 0 | 31 |
| 719.7 / 184.1 | 0 | 68 | 0 | 68 |
| 720.7 / 184.1 | 6 | 70 | 5 | 67 |
| 727.7 / 184.1 | 4 | 17 | 4 | 17 |
| 729.7 / 184.1 | 4 | 57 | 4 | 57 |
| 730.7 / 184.1 | 4 | 55 | 4 | 51 |
| 731.7 / 184.1 | 2 | 74 | 2 | 69 |
| 732.7 / 184.1 | 2 | 68 | 1 | 62 |
| 733.8 / 184.1 | 0 | 80 | 0 | 73 |
| 734.7 / 184.1 | 2 | 78 | 2 | 59 |
| 738.7 / 184.1 | 2 | 53 | 2 | 50 |
| 739.8 / 184.1 | 4 | 37 | 4 | 36 |
| 742.7 / 184.1 | 2 | 38 | 2 | 33 |
| 744.7 / 184.1 | 6 | 59 | 6 | 59 |
| 746.6 / 184.1 | 3 | 71 | 3 | 66 |
| 748.7 / 184.1 | 2 | 72 | 2 | 67 |
| 752.9 / 184.1 | 4 | 38 | 4 | 38 |
| 753.6 / 184.1 | 2 | 64 | 2 | 61 |
| 754.5 / 184.1 | 4 | 50 | 4 | 49 |
| 756.6 / 184.1 | 6 | 74 | 5 | 70 |
| 757.7 / 184.1 | 2 | 71 | 2 | 59 |
| 758.6 / 184.1 | 4 | 71 | 3 | 62 |
| 759.8 / 184.1 | 2 | 68 | 2 | 61 |
| 760.6 / 184.1 | 2 | 77 | 2 | 66 |
| 762.7 / 184.1 | 2 | 77 | 2 | 67 |
| 764.7 / 184.1 | 2 | 76 | 2 | 71 |
| 766.7 / 184.1 | 2 | 78 | 2 | 75 |
| 768.6 / 184.1 | 4 | 72 | 4 | 67 |
| 769.3 / 184.1 | 2 | 78 | 1 | 66 |
| 770.6 / 184.1 | 8 | 71 | 7 | 64 |
| 772.6 / 184.1 | 11 | 68 | 8 | 60 |
| 773.8 / 184.1 | 2 | 78 | 1 | 71 |
| 774.6 / 184.1 | 4 | 76 | 4 | 64 |
| 775.9 / 184.1 | 0 | 73 | 0 | 64 |
| 776.7 / 184.1 | 2 | 54 | 2 | 46 |
| 777.8 / 184.1 | 2 | 65 | 2 | 57 |
| 778.6 / 184.1 | 2 | 76 | 1 | 61 |
| 780.6 / 184.1 | 4 | 71 | 4 | 65 |
| 784.6 / 184.1 | 6 | 74 | 6 | 72 |
| 785.8 / 184.1 | 0 | 79 | 0 | 72 |
| 786.7 / 184.1 | 4 | 75 | 3 | 61 |
| 787.8 / 184.1 | 0 | 78 | 0 | 65 |
| 788.7 / 184.1 | 2 | 78 | 2 | 61 |
| 790.6 / 184.1 | 2 | 78 | 2 | 60 |
| 791.3 / 184.1 | 2 | 59 | 1 | 52 |

[1]Previously identified peak

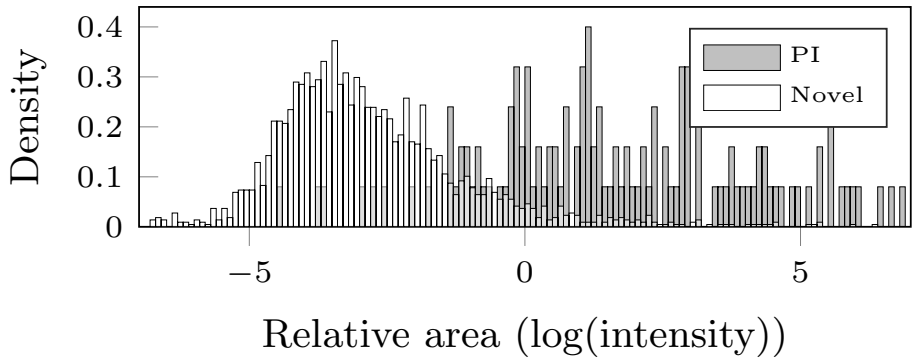
**Table D19: Glycerophospholipid brain discovery dataset, cont'd.**

| | Original | | Filtered | |
|---|---|---|---|---|
| Transition | PI | Novel | PI | Novel |
| 792.7 / 184.1 | 6 | 68 | 6 | 62 |
| 794.6 / 184.1 | 6 | 72 | 6 | 62 |
| 795.5 / 184.1 | 2 | 69 | 2 | 61 |
| 796.6 / 184.1 | 6 | 65 | 6 | 56 |
| 798.7 / 184.1 | 1 | 79 | 1 | 75 |
| 799.8 / 184.1 | 6 | 71 | 5 | 61 |
| 800.7 / 184.1 | 4 | 72 | 4 | 67 |
| 801.8 / 184.1 | 2 | 73 | 2 | 72 |
| 803.8 / 184.1 | 4 | 55 | 3 | 49 |
| 804.4 / 184.1 | 8 | 55 | 4 | 48 |
| 806.6 / 184.1 | 6 | 73 | 6 | 70 |
| 808.6 / 184.1 | 4 | 74 | 4 | 62 |
| 810.7 / 184.1 | 6 | 73 | 6 | 63 |
| 811.8 / 184.1 | 0 | 80 | 0 | 65 |
| 812.7 / 184.1 | 4 | 76 | 4 | 68 |
| 813.8 / 184.1 | 4 | 66 | 4 | 64 |
| 814.6 / 184.1 | 6 | 55 | 2 | 46 |
| 815.6 / 184.1 | 2 | 64 | 0 | 53 |
| 816.6 / 184.1 | 4 | 66 | 4 | 58 |
| 818.7 / 184.1 | 6 | 51 | 6 | 48 |
| 819.4 / 184.1 | 2 | 74 | 1 | 64 |
| 820.8 / 184.1 | 6 | 74 | 5 | 67 |
| 822.6 / 184.1 | 8 | 71 | 7 | 66 |
| 823.8 / 184.1 | 2 | 78 | 2 | 69 |
| 824.8 / 184.1 | 2 | 78 | 2 | 72 |
| 827.9 / 184.1 | 4 | 48 | 4 | 47 |
| 830.6 / 184.1 | 2 | 61 | 2 | 60 |
| 831.8 / 184.1 | 2 | 53 | 2 | 45 |
| 832.7 / 184.1 | 4 | 72 | 4 | 67 |
| 834.7 / 184.1 | 6 | 74 | 5 | 67 |
| 836.7 / 184.1 | 4 | 75 | 4 | 69 |
| 838.7 / 184.1 | 8 | 64 | 8 | 55 |
| 839.6 / 184.1 | 0 | 66 | 0 | 54 |
| 842.7 / 184.1 | 8 | 67 | 7 | 67 |
| 844.8 / 184.1 | 2 | 70 | 2 | 66 |
| 846.6 / 184.1 | 2 | 68 | 2 | 64 |
| 848.8 / 184.1 | 3 | 69 | 2 | 64 |
| 850.7 / 184.1 | 2 | 76 | 2 | 72 |
| 854.6 / 184.1 | 2 | 41 | 2 | 40 |
| 857.7 / 184.1 | 0 | 60 | 0 | 58 |
| 858.5 / 184.1 | 2 | 62 | 2 | 61 |
| 859.7 / 184.1 | 2 | 58 | 2 | 53 |
| 860.7 / 184.1 | 2 | 67 | 1 | 63 |
| 862.7 / 184.1 | 2 | 77 | 2 | 74 |
| 864.7 / 184.1 | 4 | 76 | 3 | 74 |
| 874.7 / 184.1 | 8 | 51 | 8 | 51 |
| 876.7 / 184.1 | 0 | 80 | 0 | 74 |
| 878.8 / 184.1 | 2 | 73 | 2 | 70 |
| 881.7 / 184.1 | 2 | 72 | 2 | 69 |
| 885.8 / 184.1 | 2 | 56 | 2 | 54 |
| 887.7 / 184.1 | 2 | 72 | 2 | 69 |
| 902.7 / 184.1 | 0 | 80 | 0 | 80 |
| 904.7 / 184.1 | 0 | 78 | 0 | 74 |
| 909.8 / 184.1 | 2 | 55 | 2 | 55 |

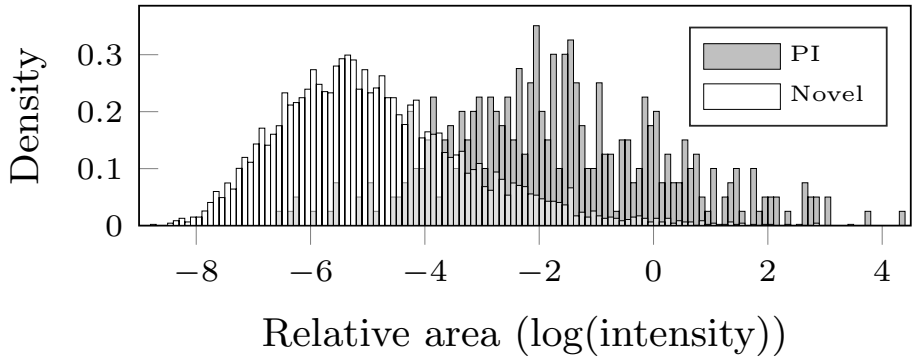
**(a)**


Density
0.4
0.2
0
PI
Novel
−6
−4
−2
0
2
4
Relative area (log(intensity))

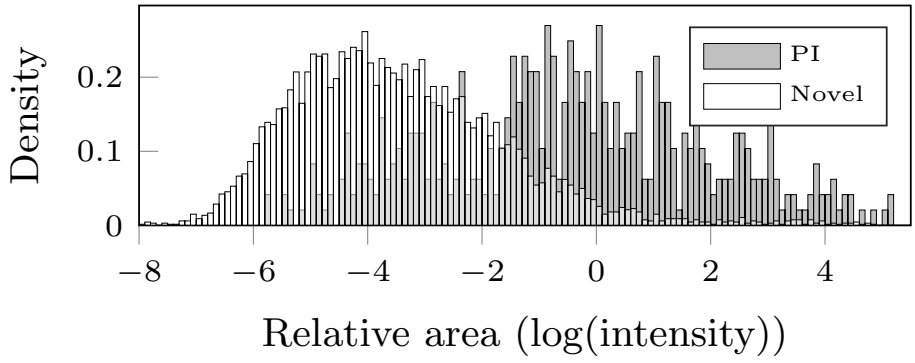
**(b)**


Density
0.4
0.3
0.2
0.1
0
PI
Novel
−5
0
5
Relative area (log(intensity))

**(c)**


Density
0.3
0.2
0.1
0
PI
Novel
−8
−6
−4
−2
0
2
4
Relative area (log(intensity))

**(d)**


Density
0.2
0.1
0
PI
Novel
−8
−6
−4
−2
0
2
4
Relative area (log(intensity))

**Supplementary Figure D10: Discovery dataset relative areas of previously identified-versus-novel peaks.** PI denotes previously identified peak. (a) Sphingolipid blood discovery dataset. (b) Sphingolipid brain discovery dataset. (c) Glycerophosphocholine blood discovery dataset. (d) Glycerophosphocholine brain discovery dataset.