Computational Approaches Enabling Disparately Acquired Untargeted LC-MS Metabolomics Data Analysis

by

Hani Habra

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2022

Doctoral Committee:

Associate Professor Alla Karnovsky, Chair
Professor Veerabhadran Baladandayuthapani
Assistant Professor Charles Evans
Professor George Michailidis, UF Informatics Institute
Professor Alexey Nesvizhskii

Hani Habra

hhani@umich.edu

ORCID iD: 0000-0003-4838-0105

Dedication

I would like to dedicate this thesis to my mother, Randa Loutfi, my father Ghiyath Habra, my grandmother, Fahmia Hower-Loutfi, and to all those who supported and encouraged me throughout my education.

Acknowledgements

First and foremost, all praises are due onto Allah, the most glorified, the most high, for rescuing my parents from the dark, uncertain days of occupied Kuwait, for blessing me with the opportunity to pursue my dreams, and for being my strength in times of hardship and ease throughout this journey. I pray that the intellectual and experiential knowledge I've acquired during my studies will be useful and beneficial to the scientific community and mankind.

I would like to express my gratitude to my mentor and supervisor, Dr. Alla Karnovsky, for taking me as a student and for her continuous support, wisdom, and guidance. It has been a tremendous honor to work with Dr. Karnovsky's group, especially William Duren, Gayatri Iyer, Janis Wigginton, and other past and present lab members. I will always appreciate the encouragement, companionship, and insightful suggestions that have helped further our mission and enabled my growth as a researcher. I would like to thank Dr. Maureen Kachman, Dr. Charles Evans, and Dr. George Michailidis for answering my myriad questions in chemistry and statistics. I deeply appreciate them for taking the time to share their knowledge and assist in various projects.

I would like to thank the Department of Computational Medicine & Bioinformatics for accepting me to the master's and PhD program. The staff has been so welcoming, friendly, and helpful. I was truly blessed to be able to do my education here at the University of Michigan. A special thanks goes to Dr. Maureen Sartor and Dr. Raymond Cavalcante for guiding me towards pursuing graduate studies in Bioinformatics.

I have had the privilege of working with many scientists who contributed to the development and success of my projects. These include Dr. Veerabhadran Baladandayuthapani, Dr. Ivana Blaženović (Brightseed, CA), Dr. Clary Clish, Brady Anderson, Dr. Stephen Goutman, Dr. Eva Feldman, Dr. Charles Burant, and Dr. Jennifer Meijer (Dartmouth, NH). I greatly appreciate the help of all other scientists whom I have collaborated with and learned from throughout my graduate studies.

I would like to gratefully acknowledge my funding sources during my doctorate studies. The Advanced Proteome Informatics of Cancer Training Program (PICTP) headed by Dr. Alexey Nesvizhskii and funded by the National Cancer Institute generously supported me in the first two years of my doctoral training. In the latter years, the NIH-sponsored Michigan Compound Identification Development Core headed by Dr. Nesvizhskii and Dr. Charles Evans provided funds and extensive support.

Finally, I would like to express my love and gratitude to my family and friends. Thank you to my parents, Randa and Ghiyath, grandmother Fahmia, sister Reema, best friend Adam, cat Lulu, and all my aunts, uncles, cousins, roommates, and cherished companions that I have made throughout my life.

Hani Habra

April 21, 2022

Preface

This dissertation is an original work by Hani Habra, based on original ideas provided by my supervisor Dr. Alla Karnovsky. Primary research projects were led by Dr. Alla Karnovsky from the Department of Computational Medicine & Bioinformatics, who supervised my training, coordinated funding for research and educational activities, and contributed to all published projects described herein. Programmers William Duren and Janis Wigginton were responsible for the development of the *Binner*, *CorrelationCalculator* and *Filigree* software packages used in multiple research projects. With few noted exceptions, metabolomics experimentation was performed at the Michigan Regional Comprehensive Metabolomics Research Core (MRC2) at the University of Michigan in Ann Arbor, directed by Dr. Charles Burant and managed by Dr. Maureen Kachman. Dr. Kachman spearheaded the development of *Binner* alongside Dr. Karnovsky and provided her analytical chemistry expertise for multiple projects. The Broad Institute Metabolomics Platform directed by Dr. Clary Clish contributed two LC-MS metabolomics datasets used for software testing (see chapters 2 and 3), with Dr. Kevin Bullock responsible for conducting at least one of these experiments. Dr. Charles Evans generated experimental measurements for the study described in Chapter 4.2 and contributed significantly to all compound identification projects. Brady Anderson (Department of Chemistry) is principally responsible for formulating and carrying out the project outlined in Chapter 4.3, including experimental and computational identification tasks. Other Compound Identification Development Core (CIDC) members, including the West Coast Metabolomics Center, Georgia Technical Institute, and Pacific Northwest National Laboratories contributed datasets necessary

for the tasks outlined in Chapter 4.4. Dr. Stephen Goutman and Dr. Eva Feldman were the principal investigators of the ALS metabolomics study outlined in Section 5.2. Drs. Jennifer Meijer, Vasantha Padmanabhan, and Charles Burant coordinated the Michigan Mother Infant Pairs study and its associated metabolomics investigations; Dr. Meijer worked with us to formulate study directions, perform data analysis, and interpret our results. Dr. George Michailidis stared his statistical and computational expertise for every major metabolomics software development and data analysis project undertaken and Dr. Veerabhadran Baladandayuthapani advised me on modeling approaches for the *metabCombiner* project. My colleague Gayatri Iyer worked on the implementation of DNEA and assisted in its application for the ALS project. Finally, Dr. Alexander Raskind administered the virtual server used for computational projects and assisted in numerous technical queries over the years.

I was responsible for devising and testing multiple strategies for the *Binner* workflow steps, including data cleaning, clustering, and annotation, and researching other programs for carrying out the comparative performance analysis. I was also responsible for the conception, implementation, testing, and maintenance of *metabCombiner*. With few exceptions, I applied these computational tools in all studies described in this project, as well as testing of normalization strategies, assessments of feature overlaps, and statistical and bioinformatics methods, including partial correlation network construction.

Table Of Contents

List of Figures

List of Tables

List of Abbreviations

The following table is a list of abbreviations used in this document that stand for important terms in the study of metabolomics.

a)	**Da:**	Dalton or unified atomic mass
b)	**iDDA:**	iterative Data Dependent Analysis
c)	**DNEA:**	Differential Network Enrichment Analysis
d)	**ESI:**	Electrospray Ionization
e)	**EIC:**	Extracted Ion Chromatogram
f)	**FPA:**	Feature Pair Alignment
g)	**GC:**	Gas Chromatography
h)	**HILIC:**	Hydrophilic Interaction Chromatography
i)	**HPLC:**	High Performance Liquid Chromatography
j)	**HPU:**	High Priority Unknown
k)	**IS:**	Internal Standard
l)	**LC:**	Liquid Chromatography
m)	**m/z:**	mass-to-charge ratio
n)	**MMIP:**	Michigan Mother-Infant Pairs
o)	**MS:**	Mass Spectrometry
p)	**MSI:**	Metabolomics Standards Initiative
q)	**NIST:**	National Institute of Standards and Technologies
r)	**NMR:**	Nuclear Magnetic Spectroscopy
s)	**RPLC:**	Reversed Phase Liquid Chromatography
t)	**RT:**	Retention Time
u)	**Q:**	Relative Quantity
v)	**QTOF:**	Quadrupole Time-of-Flight

Abstract

Metabolomics is the systematic study of small molecule metabolites that are substrates, intermediates, and products of cellular metabolism. Metabolomics assays performed using liquid chromatography- mass spectrometry (LC-MS) typically detect thousands of analytes, or features characterized by mass-to-charge (m/z) ratio and retention time (RT). The objective of untargeted metabolomics is to detect, quantify, and identify as many compounds as possible and to relate their abundances with phenotypic outcomes. Metabolite identification is a major bottleneck in metabolite profiling studies, with only a small percentage of observed features unambiguously identified in a typical experiment. A substantial proportion of the detected features consists of in-source adducts, fragments, isotopologues, complexes, chemical and computational artifacts. Detecting and removing these redundancies is essential for improving statistical power in downstream analysis, as many metabolomics studies have limited sample sizes. LC-MS assays can be performed using a wide range of chromatographic conditions, instruments, and other analytical techniques. Differences in protocols between and within laboratories create numerous challenges for information transfer and meta-analysis, especially for unidentified compounds.

This dissertation is focused on developing computational methods for enabling disparately acquired LC-MS metabolomics data analysis and demonstrating their benefits in compound identification and biomedical investigations. First, I describe *Binner*, a standalone application for annotating in-source adducts, fragments, complexes, and isotopologues derived from a common metabolite, thus facilitating the reduction of feature tables to a parsimonious

expression of the detected metabolome. I highlight the unique capabilities of *Binner*, including its superior annotation performance compared to existing programs and its modules for facilitating the discovery of complex annotations. Second, I describe *metabCombiner*, a software package for aligning metabolomics measurements acquired under similar, but non-identical, conditions, concatenating their values to generate merged feature tables. *metabCombiner* uses a spline-based modeling approach to project across substantial gaps in retention times and a weighted similarity score to match features corresponding to identical analytes. This package forms the basis for expanded sample size analyses as well as information transfer between protocols, instruments, and laboratories. I detail multiple applications in which compound identification rates in plasma, urine, and other specimens are improved by coupling disparate LC-MS alignment to novel experimental and computational elucidation approaches. A framework consisting of alignment and normalization steps for the removal of intra-batch, inter-batch, and inter-experiment variation in retention times and acquired signal was developed and applied to metabolomics studies of ALS and pregnancy. Subsequent statistical and bioinformatics approaches using partial correlation networks performed on the aligned, normalized datasets illustrate the benefits of combining datasets, despite major differences in experimental conditions. Together, these computational methods address numerous data analysis challenges and unlock new opportunities in the metabolomics field as well as other fields that utilize LC-MS for high-throughput measurements.

Chapter 1

Introduction to Computational Metabolomics

1.1 General Introduction to Metabolomics

Metabolites are small (<1500 Da) molecules that constitute the reactants or byproducts of chemical reactions in biological systems. Metabolites exhibit a wide range of properties and belong to diverse biochemical classes, including lipids, nucleotides, carbohydrates, amino acids, short-length peptides, steroids, organic acids, and so forth.[1–3] Metabolomics is the systematic study of metabolites, or the "metabolome", with the goals of (i) providing (semi)quantitative information about each metabolite present in a sample; (ii) determining the structure, function, and interactions of each small compound present; and (iii) to infer the metabolic responses of living systems to physiological stimuli.[4,5] The metabolome consists of compounds endogenous to the studied organism and those derived from exogeneous sources, such as dietary, environmental, and microbiotic exposures.[6] The total size of the metabolome differs between species and is difficult to pinpoint, with estimates of 3800 metabolites in E. Coli[7] to over 100,000 entries currently listed in the Human Metabolome Database[8] and more than 200,000 for the plant kingdom.[9]

Advances in analytical chemistry, instrumentation technologies, hardware and informatics have enabled the large-scale detection and quantification of metabolites within cells, tissues, fluids, and organisms, establishing metabolomics as a powerful and widely used tool in biological and clinical investigations. Currently, there are two major analytical platforms used to

detect and quantify metabolites: Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS). NMR spectroscopy obtains structural information by measuring the interactions of an oscillating radiofrequency electromagnetic field with nuclei in an external magnetic field. NMR spectra consist of peaks whose positions are determined by chemical shifts, or differences in the intrinsic frequency of nuclear spinning (largely determined by the electronegativity of nearby atoms) relative to a reference compound.[4,10,11] NMR spectroscopy has numerous advantages in metabolomics research, such as ease of sample preparation, sample preservation and rapid analysis speed. Its cross-lab reproducibility makes this technique a gold standard in compound identification. The principal weakness of NMR is its lack of sensitivity as the technique is largely incapable of quantitation for compounds of extremely low concentration (~10^{-9} M or below); the relatively inferior coverage achieved by NMR makes it less suitable for performing high-throughput untargeted metabolomics.[10]

By contrast, mass spectrometry (MS) is a highly sensitive and versatile method that measures mass-to-charge (m/z) ratios of ionized molecules. A mass spectrometer consists of an ion source to convert compounds into ions; a mass analyzer to resolve these ions by time-of-flight or in an electromagnetic field; and a detector, to detect the ions and output signal corresponding to their m/z and abundance.[12] Mass spectrometry is usually coupled with chromatography or other techniques used to separate complex mixtures into their components, facilitating the analysis of different metabolite classes. The most common separation methods are the following: 1) Gas chromatography (GC), in which volatile molecules are vaporized and partition between a solid or liquid stationary phase coated onto and an inert carrier gas at elevated temperatures before introduction to the mass spectrometer[13] 2) Liquid chromatography (LC), in which compounds are typically separated by partitioning between a stationary phase

bonded to stationary particles in a column and a mobile phase consisting of liquid solvent. The extent of retention is determined by differential affinity for the solvent or the stationary phase.[14] 3) Capillary Electrophoresis, suited for analysis of charged compound where specimens travel at different through an electric-field stimulated capillary, separating compounds by charge & molecular size.[15] Of these separation techniques, liquid chromatography is the most versatile and achieves the greatest coverage, establishing it as the method of choice for untargeted metabolomics.[15] While other analytical approaches described here have uses and advantages in metabolite profiling, the work presented here focuses on analyzing LC-MS data.

Metabolomics experiments typically employ one of two approaches to determine the range and throughput of assayed metabolites: targeted and untargeted. Targeted metabolomics approaches are designed to measure a specified list of known metabolites or metabolite classes. Protocols for sample preparation, metabolite extraction and data acquisition are optimized to achieve quantitatively accurate measurements for a limited subset of the metabolome.[1,16] These experiments are typically hypothesis-driven, utilizing knowledge of biochemical pathways and metabolic enzymes to answer specific questions pertaining to phenotypic perturbations and physiological states. Recent examples include studies detecting decreased levels of pantothenic acid in brain tissues of Alzheimer's and Huntington's Disease patients[17] and finding a panel of serum bile acids whose levels reliably discriminate the presence or absence of nonalcoholic fatty liver disease in Type 2 Diabetes patients.[18] By contrast, the aim of untargeted metabolomics (alternatively, metabolite profiling) is to achieve a holistic view of the metabolome in a population of specimens by measuring as many metabolites as possible. The complexity of biological samples and the wide range of physicochemical properties of metabolites preclude a comprehensive read-out of the entire metabolome by any single analytical method, therefore

requiring multiple assays to achieve a global metabolic profile.[19] Untargeted metabolomics is an invaluable hypothesis-generating approach typically used to detect biomarkers and explore large-scale metabolic shifts. The breadth of coverage by untargeted metabolomics techniques comes at the expense of quantitation accuracy as experimental protocols are not optimized for the measurement of any specific metabolite class. Furthermore, untargeted metabolomics assays detect many unknown metabolites as well as contaminants and other signals not derived from the underlying samples, providing a central challenge to metabolomics data analysis. Untargeted metabolomics analyses are the central focus of the research presented in this document.

Untargeted metabolomics studies routinely produce large and complex datasets that require computers for storage, deconvolution, analysis, interpretation, and sharing of results. The growing capabilities of analytical techniques and instrumental technologies, coupled with the increasingly widespread use of metabolomics in academic, industrial, clinical, and other settings, has necessitated the development of computational resources to extract knowledge from complex data in an efficient manner. Computational metabolomics denotes the collective set of software, algorithms, databases, and other resources developed for the study of metabolomics. Three major challenges underlie the pursuit of computational solutions in metabolomics analyses. The first is detection and quantitation, consisting of methods to uncover metabolite-associated signals across analyzed samples and assign them a quantity that accurately reflects their comparative abundances across experimental samples. The second is annotation, consisting of all steps that assign chemical information, such as structural identity, molecular formulas, or biochemical class, to detected analytes. The third is interpretation, signifying the use of statistical and

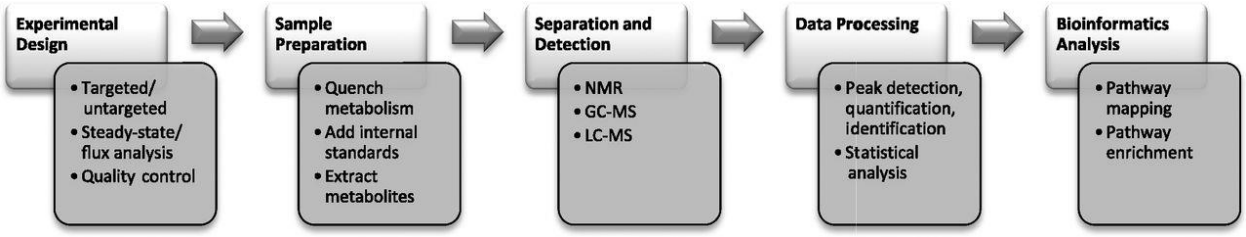


Figure 1.1 Metabolomics Workflow Brief overview of experimental and computational procedures routinely performed in metabolomics studies, adapted from Sas et al. (2015).[20]

informatics approaches to derive useful and meaningful information from metabolite identities and quantities. Subsequent sections discuss these individual problems in detail.

1.2 LC-MS Experimentation, Instrumentation & Data Acquisition

High performance liquid chromatography coupled with mass spectrometry (HPLC-MS) is an analytical technique of choice for metabolite profiling.[21] An understanding of LC-MS experimental design and instrumentation is a prerequisite to developing effective computational resources for metabolomics data analysis. Here, basic ideas about study design, sample collection and preparation, liquid chromatography and mass spectrometry are briefly described as they relate to untargeted metabolomics.

1.2.1 Experimental Design

The crucial first step to every metabolomics endeavor is to consider an appropriate study and experimental design. This begins with a specific aim or question that the researcher seeks to answer, ranging from characterizing the metabolome of specific specimens or phenotypes of interest to more complex queries involving a range of analytical or biological factors.[1] Depending on the study aims, the researcher must consider the appropriate population to obtain data from and ensure that enough samples can be obtained to draw appropriate conclusions. Next, the manner of sample collection, storage, and stabilization must be carefully planned. Fresh samples from clinical, animal, or cellular culture studies provide the most control over sample collection and storage, with procedures such as snap freezing utilized to cease enzymatic

activity and preserve the metabolic state to the greatest extent possible.[22] The dynamic, fluctuating levels of metabolites means that several factors must be considered to mitigate unintended sources of variation, such as diets, time of day at sample collection, and excessive variation in clinical covariates related to the outcome of interest. The choices of sample preparation and analytical technology are the next considerations, followed by the computational and statistical methodologies necessary to answer the underlying study question.

1.2.2 Sample Preparation

The goal of sample preparation in untargeted metabolomics studies is to maintain the original metabolite composition of samples as much as possible while converting the samples into a compatible medium for MS analysis.[23] Commonly applied steps include metabolite extraction, centrifugation, precipitation (e.g. to remove proteins), evaporation, reconstitution, or dilution, depending on the biological specimen of interest. Various metabolite extraction techniques are routinely applied, especially liquid-liquid extraction (LLE) by which metabolites are partitioned based on their relative solubilities in immiscible liquids, and solid phase extraction (SPE) where analytes are partitioned based on their affinity for a solid phase over a liquid.[24] Due to the challenges of surveying the chemically diverse metabolome, researchers may prioritize certain classes of metabolites based on physicochemical properties.[23]

It is standard practice in metabolomics studies to include quality controls in the experimental design. Most commonly, pooled aliquots of all study samples are inserted at intermittent points of experimental runs and measurements on negative control or "blank" samples derived from extraction solvents enable the assignment and elimination of experimental artifacts in post-processing steps. In addition, purified internal standard compounds are frequently mixed with biological samples before or after metabolite extraction to monitor and

remove systematic biases in measured intensities. Together, these procedures aid in the comprehensive detection of metabolites, improve the data quality, and facilitate biological interpretation of the results.

1.2.3 Liquid Chromatography

Chromatographic methods are designed to separate complex mixtures into their components. When coupled with mass spectrometry, this allows for improved detection sensitivity and data quality. High Performance Liquid Chromatography (HPLC) is an analytical technique for separating mixtures based on the differential affinity of molecules for liquid solvents versus a stationary phase. The stationary phase is a porous solid (e.g. polymers or silica) contained within a column which the solvent passes through. Compounds are introduced onto the column and partition between the stationary phase and solvent many times as they move through its length, eventually eluting (emerging) from the column outlet where they may be sensed by a detector. The output of HPLC is a chromatogram, which represents signal in relation to how long each compound was retained in the column before detection. The time taken for an analyte to be detected after its initial injection is the *retention time* (RT), a property that strongly depends on the compound's interactions with the chromatographic system composed of the liquid mobile phase and the stationary phase of the column. RTs are of central importance to LC-MS metabolomics as they play a crucial role in characterizing detected metabolites. Assigning, correcting for, and predicting chromatographic retention behavior is an important goal in computational metabolomics.

HPLC methods separate compounds based on certain physiochemical properties, including polarity, charge (Ion Exchange Chromatography (IEC)), or size (Size Exclusion Chromatography (SEC)).[25] LC approaches can be combined to attain a multidimensional

separation to increase the separation capacity.[26] However, one-dimensional separations by polarity are the most used for metabolomics, with two complementary approaches accounting for most applications: Reversed Phase Liquid Chromatography (RPLC) and Hydrophilic Interaction Chromatography (HILIC).[21,27] RPLC uses a stationary phase modified with nonpolar groups, such as aliphatic hydrocarbon chains (C18), and a mixture of aqueous and organic (e.g. acetonitrile, methanol) solvents. Most applications of RPLC utilize gradient elution, by which the liquid mobile phase gradually transitions from a predominantly aqueous to an organic solvent, which favors longer retention of metabolites with greater hydrophobicity. Non-polar & semi-polar substances are therefore well-resolved by RPLC, whereas highly polar and ionic analytes are poorly retained and often elute in the void volume. By contrast, HILIC, a variant of Normal Phase Liquid Chromatography, uses columns containing hydrophilic, anionic, cationic, or zwitterionic ligands for the improved separation hydrophilic compounds.[21] In HILIC, the solvent is initially mostly a non-polar organic solvent before gradually assuming a more polar composition, which results in the increasingly polar metabolites being retained longer and with enhanced resolution, at the expense of poor retention for nonpolar compounds. HILIC is commonly used as an orthogonal method to RPLC, though it suffers from reduced retention time reproducibility and requires more time for stabilization between runs, among other disadvantages that hinder its widespread adoption.[28,29] Most metabolomics studies discussed in this document use RPLC-MS, with some applications involving HILIC-MS.

Aside from polarity, numerous factors affect the retention of compounds on the column in liquid chromatography. These include the choice of solvents, the type of column, column dimensions, the age of the column, gradient slope, flow rate, and total chromatography time.[30] Additionally, minor imperfections or changes in chromatographic conditions between LC runs,

such as gradient delay, solvent mis-apportioning, changes in pressure and pH may introduce minor changes between LC runs.[30] In standard practice, experimental variables and analytical conditions are maintained and replicated to the greatest extent possible across all samples to obtain consistent retention time values. Importantly, measured retention times are specific to a metabolite's interactions with a particular chromatographic system and are not easily extrapolated to other systems. Retention indices, calculated linearly based on the elution patterns relative to internal standards of a particular class (such as nitroalkanes or alkyl ketones), are sometimes used as a dimensionless substitute for retention times, though their utility is uncommon for LC as compared to Gas Chromatography.[31–33]

1.2.4 Mass Spectrometry

Invented near the turn of the 20th century, mass spectrometry has become a primary means of chemical detection and characterization with wide-ranging academic and industrial applications.[34] Mass spectrometry can detect substances at extremely small concentrations and, when coupled with chromatographic separation, provides the most comprehensive metabolomic profile of any existing technique. Innovations and enhancements continue to improve the sensitivity, efficiency, versatility, and affordability of mass spectrometers. Important concepts are discussed as they relate to structure elucidation and quantitation in untargeted metabolomics.

Mass spectrometry measures the mass-to-charge ratios (*m/z*) of compounds, which requires that analytes be ionized before detection. The ion source is an essential component of the mass spectrometer where neutral metabolites preferentially form positively or negatively charged ions, depending on their physicochemical properties. Hard ionization refers to techniques that impart excess energy onto molecules, frequently fragmenting the analyte; chief among these is electron impact ionization (EI), where beams of electrons are emitted at a specific

voltage (typically 70eV) to cause the ejection of a single electron from gas-phase compounds, producing a radical cation (and resultant fragments) for mass analysis.[35] EI is the primary method in GC-MS and lacks compatibility with liquid chromatography applications. By contrast, soft ionization approaches impart small amounts of energy, minimizing fragmentation and allowing for accurate measurement of the mass of intact ions. Among these are Matrix-Assisted Laser Desorption Ionization (MALDI), whereby samples are entrapped with an energy-absorbent organic matrix and ionized by a laser beam; atmospheric pressure chemical ionization (APCI), where the sample is vaporized, then ionized in aerosol form via corona discharge using a needle electrode upon which a high voltage is applied;[36] and Electrospray Ionization (ESI), where mist-like electrically-charged droplets of a particular charge disperse and disintegrate, inducing charge upon the sample, then analytes are ionized and transferred to the gas phase either via evaporation from the surface of microdroplets (ion evaporation model)[37] or by successive Coulombic fissions due to the excessive build-up of similar charges (charged-residue model).[38] Of these methods, electrospray ionization is the most widely used ionization method for untargeted LC-MS metabolomics, largely due to its applicability to a wider range of metabolites. Most metabolites are singly charged ($z = 1$) when subjected to ESI, though some take on multiple charges. One notable disadvantage is that electrospray ionization is subject to matrix effects, particularly ion suppression whereby an analyte's ionization efficiency is reduced due to the presence of co-eluting compounds in the same matrix.[36,39] Moreover, certain compounds such as thermally stable, low-polarity metabolites are either inaccessible or better measured by APCI, which often serves as a complementary approach.[27,40] Nevertheless, in LC-MS metabolomics applications, ESI is the primary ionization method and all studies described herein use this technique.

Following ionization, compounds are propelled via a strong magnetic field towards the mass analyzer, which comprises the next component of the mass spectrometry workflow. A variety of commercially available instruments exist for high resolution mass analysis, with most based on Time of Flight (TOF) or Orbitrap technologies.[27] In TOF-MS, an electric field of a fixed potential accelerates ions through a flight tube and the time taken to traverse the tube is then mathematically related to the accurate mass of the ions.[41] Generally, ions of lower mass and higher charge travel faster through the flight tube. To facilitate the selection of low molecular weight compounds, a quadrupole is usually paired to TOF analysis and filters ions based on the trajectory of their oscillations between the rods such that only compounds of a specific m/z range will arrive at the detector. By contrast, Orbitraps consist of devices with inner and center electrodes, between which an electric field is generated such that ions oscillate harmonically around the central electrode. The resulting frequencies of these orbits are obtained through Fourier transformation, which are then converted to a mass spectrum.[42] Both QTOF and Orbitrap instruments are capable of high mass resolution, with QTOF achieving resolutions around 35000 and Orbitrap more than 200,000.[43] Orbitrap mass spectrometers typically have higher dynamic ranges, defined as the ratio of the most abundant ion to the least abundant detectable ion, while QTOF possess higher scan rates, facilitating their coupling with high-throughput chromatographic techniques.[27] Both QTOF & Orbitrap mass analyzers enable untargeted metabolomics analysis, and most laboratories use one or a combination of these instruments.

Depending on the application, an additional fragmentation step may be pursued in mass spectrometry analysis to reveal structural information about the detected compounds. The most widely used method is collision-induced dissociation (CID),[44] where selected ions are forced to collide with a neutral buffer gas, imparting internal energy that breaks molecular bonds and

forming product ions which can similarly be measured by m/z. In certain mass spectrometers, this process can be repeated multiple times when studying specific metabolites, with a single fragmentation denotated as MS/MS and multiple dissociations termed MS^n. The spectral output of these fragmentation workflows can be compared to spectra of known chemicals, and thus used to deduce the identity of the underlying compounds. More information on these workflows is contained in Section 1.4.

1.3 Metabolomics Preprocessing & Normalization

LC-MS assays generate files containing abundant and complex mass spectral information detected for each sample which provide the entry point to computational metabolomics data analyses. LC-MS preprocessing workflows extract quantitative values from raw signal, transforming raw spectral information in these files into a matrix of features, where each feature represents an analyte that is commonly detected across experimental files represented by an averaged m/z, retention time (RT), and per-sample signal abundance values. Various commercial solutions exist for preprocessing data from specific vendors; however, this discussion pertains to open-source data analysis software and methods, particularly the popular XCMS[45] and MZMine2[46] programs. The main steps in a preprocessing workflow typically consist of file conversion, peak detection and deconvolution, retention time correction, alignment, gap filling, imputation, filtering, and normalization.

1.3.1 File Conversion

Initially, mass spectral data is presented in a variety of proprietary formats depending on the instrument supplier, such as .d by Agilent (Santa Clara, CA) or .raw by Thermo Fisher Scientific (Waltham, MA). Though commercial software to extract information from these specific file types are widely used, most open-source software cannot operate on these intricate

files, motivating the development of various XML-based common formats to achieve platform-independence and ease of data exchange. Early efforts include the .mzData and .mzXML[47] formats, developed by the Human Proteomics Organization Protein Standards Initiative (HUPO PSI) and Systems Biology Institute (SBI) respectively, before the two philosophies were unified in 2006-2008 to create the .mzML format.[48] .mzML is designed to be simple, stable with some flexibility to encode new information, and is compatible with a wide range of open-source mass spectrometry software. To make use of these formats, files must first be converted from their native format to .mzML, .mzXML, or a related file type. The most prominent conversion tool is the MSConvert utility of Proteowizard,[49,50] which can be used to apply various filters and perform useful operations such as centroiding of profile data to reduce its complexity. Once converted, the files are compatible with open-source pre-processing software. Conversion therefore represents the first step in many computational metabolomics protocols.

1.3.2 Peak Detection

Mass spectral data are represented as a series of sequential scans ordered by retention time, with each scan containing a series of detected masses and their associated intensities. Peak detection algorithms identify robust signal arising from true metabolites, while excluding noisy signal from contaminants and chemicals of non-sample origin. A common first step is to determine "Regions of Interest (ROI)", or m/z traces in consecutive scans that meet thresholds for intensity and can thus be considered for modeling of peak-like behavior. For the original XCMS *matchedFilter* algorithm, the full m/z region is arbitrarily segmented into slices (by default 0.1 Da wide) and a second-derivative Gaussian peak model is fit through the highest signals within these slices.[45] The more widely used *centWave* method for high resolution MS data constructs ROIs by iteratively agglomerating nearby mass traces within set m/z distances

and updating the average mass of the ROI with each addition; then continuous wavelet transform (CWT) is applied to detect peaks from the signal, using the Mexican Hat wavelet with multiple scales to account for varying widths of the peaks.[51] In both methods, a signal to noise ratio eliminates low-lying peaks, where noise is estimated as the average signal in the m/z slice containing the signal. Similarly, MZmine2 determines extracted ion chromatograms (EIC) from detected mass traces, followed by one of multiple chromatogram deconvolution options, such as noise amplitude, Savitzky-Golay filter, and local minimum search.[46]

Despite setting the standard for LC-MS preprocessing, XCMS & MZMine2 peak-picking methods have been reported to generate many false positive peaks as part of their outputs.[52] Examples of false positives may include partial integrations of signal that do not encompass the full peak, the integration of noise and flat plateaus often originating from LC solvent ions. These signals inflate the number of metabolite peaks and without further post-processing or inspection, the problem is hidden from the user. Efforts have been made in recent years to improve CWT-based methods for reduced false positive peak detection. These include the Automated Data Analysis Pipeline (ADAP) implemented in MZMine2, which adds a coefficient/area filter parameter to limit non-Gaussian peak shapes.[53] Others have proposed deep learning methods for detecting and eliminating false positive peaks as a post-processing step.[54–56] Another simple method for eliminating background solvent ions is to remove peaks whose levels are comparable between negative control "blanks" and normal samples. Deconvoluting LC-MS signal remains a complex task with tradeoffs between identifying signal from true metabolites (including low abundance compounds) and reducing the presence of false positives in peak lists.

1.3.3 Retention Time Correction and Alignment

m/z	RT	samp1	samp2	samp2	samp3	samp4	samp5
278.9167	0.494	3799			28183	13478	
428.8893	0.495	4903	5248	13878	6716	4971	4011
360.9008	0.496	12322	11911	20954	19893	22657	11354
454.8415	0.497	8867	8723	8239	7571	7363	8036
162.9556	0.497		17943				15940

Table 1.1 Metabolomics Feature Table Each row represents an individual feature, with columns displaying the averaged mass-to-charge (m/z), retention time (RT), and individual sample values.

Most untargeted LC-MS metabolomics studies acquire data from multiple samples sequentially by the same instrument under nominally identical settings. After performing peak detection in each of the samples individually, the next step is to match peaks corresponding to the same metabolite across samples and assemble their values into a matrix, exemplified by **Table 1.1**. Each row of the table represents a "feature", or a consistently detected peak in m/z and retention time (RT) space. The weighted average of these values across the samples in which the analyte is detected is assigned as the feature's m/z and RT.

Simple alignment strategies, such as density-based grouping in XCMS or the join aligner in MZMine2, construct features by performing a matching of at most one peak per sample subject to constraints in m/z & RT distance. However, nonlinear shifts in RTs are often observed from sample to sample, due to slight variability in chromatographic mobile phase composition, column conditions, gradient, and temperature (47,48).[57,58] Accounting for these chromatographic perturbations is necessary to generate an accurate and comprehensive correspondence between peaks arising from the same analyte. Dozens of methods have been devised for simultaneous RT correction and peak matching, including independent software packages and implemented algorithms within LC-MS preprocessing software. A 2015 review of alignment approaches lists several commonalities, such as the use of monotonic time warping functions modeled after "landmark" or "anchor" features and/or representative samples, followed by direct peak

matching.[59] In XCMS, two methods have been implemented for RT correction: locally weighted smoothing spline (LOESS) of retention time shifts modelled after coarsely pre-defined peak groups;[45] and Ordered Bijective Interpolated Warping (Obiwarp), a derivative of dynamic time warping (DTW) which computes a warping function between chromatograms and a single reference chromatogram.[60] In MZMine2, the RANSAC aligner iteratively generates and updates a master list of features and their candidate alignment RTs, using LOESS to map between master list RTs and those of the next chromatogram, followed by joining.[46] These methods are designed for relatively small retention drifts between samples within the same assay. More significant shifts are observed between samples in separate batches or experiments requiring specialized methods. More details about alignment are contained in Chapter 3.

1.3.4 Gap-Filling and Imputation

Following alignment, the feature matrix typically contains missing abundance values. These may result from a failure to align peaks corresponding to the same metabolite or to detect peaks due to insufficient signal-to-noise ratios or poor peak-like behavior. Since many statistical methods require complete data, imputation of missing values is often necessary after feature alignment. It is common practice to filter the feature list to exclude features whose percentage of missingness across samples exceeds some threshold. By default, features not present in more than 50% of detected samples are excluded in XCMS during the peak grouping step.[45] The "80% rule" is commonly applied in studies, limiting analyses to features that have values for at least 80% of the samples.[61] After choosing which features to retain, a variety of imputation methods can be applied to eliminate missingness in metabolomics data.

One imputation approach with supported functionality in XCMS & MZMine2 is to fill missing feature values by the detected signal in the m/z & RT space, either indiscriminately or

with minor constraints on spectral behavior. This can often detect true peaks that may have been missed due to sub-optimal parameters in the peak-picking and alignment steps. Gap-filling approaches are admittedly controversial as they may be susceptible to incorporating noise or signal from other compounds distinct from the represented analyte.[62] Generalized approaches may be used for pre-processed data from pipelines lacking gap-filling functions. These range from simple approaches such as zero, minimum, half-minimum value imputation to machine-learning imputation strategies, such as Random Forest (RF) or Singular Vector Decomposition (SVD)-based imputation.[63,64] Comparative reviews have examined the strengths and weaknesses of imputation strategies for data that is "missing at random" (e.g. peaks missed due to sub-optimal preprocessing) or "not missing at random" (peaks not present or at insufficient abundance), with RF imputation cited as the most accurate.[65,66]

1.3.5 Normalization

Normalization is a post-processing step with the goal of reducing between-sample variability resulting from pre-analytical and analytical factors, such as sample quantity or volume differences, sample handling, instrument variation, sample run order, and batch effects, while retaining biological information. Some normalization methods are specific to the biological specimen, such as normalization by creatinine levels for urine metabolomics[67] or cell/ DNA/ protein count for cell culture studies.[68,69] Some methods employ one or multiple internal standard (IS) compounds for normalization, such as the NOMIS[70] & CCMN[71] methods. The main drawback of these methods is that the IS compounds selected may not sufficiently represent the diversity of the metabolome chemical space; moreover, compounds may be subject to matrix effects, introducing biases during normalization. Another strategy involves fitting nonlinear regression models of quality control sample metabolite values as a function of run order.[72,73]

More generalized location-scale methods include normalization by total signal or total "useful" signal, autoscaling or pareto scaling. A more detailed look at normalization and removal of batch effects can be found in Chapter 5.

1.4 Compound Identification

Annotation of LC-MS features is crucial to biological interpretation and represents a challenging area of ongoing metabolomics research. LC-MS provides a wealth of information about metabolites present in biological samples, including retention behavior (which is indicative of various physicochemical properties) and accurate masses of ionized small molecules. The isotopic distributions coupled to accurate mass can facilitate deduction of molecular formulas.[74] However, this knowledge is rarely sufficient to accurately deduce the structural identities of metabolites. Retention times are specific to the chromatography system measuring the analytes and cannot be informative by themselves to other systems. A given mass or molecular formula may correspond with numerous isomers with a variety of connectivities between the atoms. Furthermore, unlike nucleic acids and proteins, the metabolome encompasses a broad structural and physicochemical range that cannot be simplified to a polymeric sequence of known monomers[75]. As a result, most detected features lack unambiguous characterization in metabolite profiling studies,[76] underlying many challenges described throughout this work.

Table 1.2 Metabolite Identification Levels

Level	Name	Minimum Requirements
1	Identified compounds	Comparison of two or more orthogonal properties with an authentic chemical standard analysed under identical analytical conditions
2	Putatively annotated compounds	Based upon physicochemical properties or spectral similarity with public/commercial spectral libraries, without reference to authentic standards
3	Putatively characterized compound classes	Based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class
4	Unknown compounds	Although unidentified and unclassified, these metabolites can still be differentiated and quantified based upon spectral data

1.4.1 Identification Confidence Levels

In 2007, the Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI) listed four distinct levels of identification confidence for small compounds, as listed in **Table 1.2** below.[77,78] Identifications at level 1 (the highest confidence level) require matching two orthogonal measurements (such as retention time, mass spectrum, isotope pattern) to that of a pure standard for authentication. Laboratories typically maintain and grow libraries of compound standards and databases documenting their properties for ease of identifications in routine metabolite profiling experiments. However, authentic standards do not exist for every metabolite and obtaining every standard compound for experimentation is impractical. Therefore, computational strategies towards structural elucidation over the past decades have pursued ways of annotating features at levels 2 or 3, obtaining spectral similarity or property matches to entries in small compound databases or ascribing chemical class characteristics to unknown features. Two prominent classes of tools are briefly discussed here: LC-MS/MS spectral matching and RT prediction models.

1.4.2 LC-MS/MS Spectral Matching

For selected analytes of a given precursor mass and retention time, LC-CID-MS/MS reveals

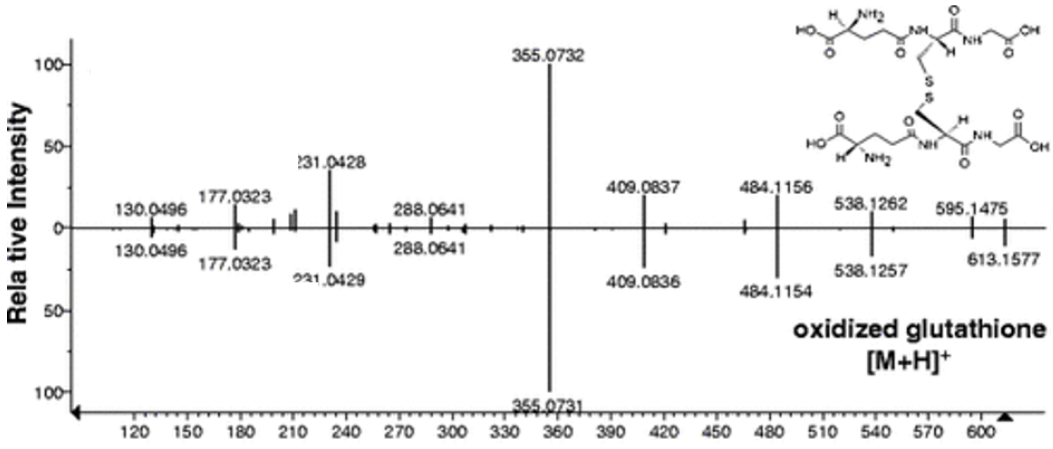


Figure 1.2 Example LC-MS/MS Spectrum Match Adapted from Dunn et al. (2013).[78]

favorable fragmentation reaction product masses and their relative abundances. Fragment masses can be useful for deducing molecular substructures, sometimes sufficient to determine the chemical class to which a particular compound belongs.[3] More importantly, tandem mass spectra are characteristic fingerprints of analytes that can be directly compared to experimentally derived MS/MS readouts collected in mass spectral databases. Library searching is a fast and accurate means of obtaining a set of compound annotations from MS/MS data, with numerous existing public or commercially available compound spectral databases such as the NIST, METLIN,[79] MoNA, and GNPS[80] databases. Tools for database search, including the prominent NIST MS Search and MSPepSearch utilities,[81] compare and score matches between library and observed spectra using dot product, hybrid, and other similarity measures. Hits from identity searches match the precursor ion of the collected spectra, while in-source hits ignore the precursor ion and score matches based on fragment ions alone. Hybrid searches combine direct peak matching and neutral-loss matching, where the mass difference between collected precursor ions and database entries is calculated before fragment ions are conditionally shifted by the same mass difference. Compound matches for each parent ion can be filtered based on composite scores and manually reviewed to confirm matches between library and empirically measured spectral peaks.

Despite these advantages, mass spectral library searching is limited by several factors. Scoring algorithms can give inaccurate results for spectra that contain very few fragments. Experimentally obtained MS/MS spectra may contain chemical noise and other compounds, complicating spectral comparisons with library spectra measured from pure standards. Databases of experimentally derived MS/MS spectra are incomplete and cover only a small fraction of the metabolome. In recent years, in silico fragmentation tools have emerged to supplement experimentally derived MS/MS spectra with computationally predicted spectra of compounds

whose spectra are not present in databases.[82–85] The accurate prediction and validation of in-silico MS/MS spectra is the subject of much on-going research, employing techniques and concepts from computational chemistry and machine learning.

1.4.3 Retention Time Prediction

Retention time is an orthogonal measurement to the mass spectrum that is indicative of an analyte's physicochemical properties. Many studies over the past decade have sought to accurately predict compound retention times, with the goal of increasing annotation rates and minimizing false positive annotations. The predominant method for RT prediction is Quantitative Structure Retention Relationship (QSRR) modeling, a cheminformatics approach in which a series of molecular descriptors, or numeric values that represent the structural composition and properties of molecules, are calculated from analytes as features for machine learning model construction. Creek et al. (2011) first employed multiple linear regression to model the retention factor of 120 compounds in a HILIC system using six calculated properties, most prominently the log octanol-water partition coefficient (a measure of compound hydrophilicity).[86] A similar effort by Cao et al (2015) used Random Forest regression and calculated 346 molecular descriptors with the rcdk package, again finding log partition coefficient to be most predictive, along with other polarity and electrostatics descriptors.[87] Both studies applied their respective models to predict the expected retention times for thousands of compounds, showing that QSRR models can limit the pool of possible identities for a given m/z, making annotation more feasible. Some studies combine retention time prediction with mass spectrum or fragmentation property modeling.[88,89] Other notable studies include Aicheler et al. (2015) where support vector regression (SVR) is used to model lipidomics retention times;[90] Wen et al. (2018) modeled retention indices with Partial Least Squares combined with Genetic Algorithms;[91] the Retip R

package implements five different algorithms for QSRR model construction and is embedded in various mass spectrometry software packages.[92] An important consideration for QSRR modeling is the applicability domain, or the structural space of compounds used to design models. A large and diverse set of compounds is necessary to build robust models applicable to a wide range of possible metabolites.[93,94]

QSRR models have the limitation of being locked to predicting retention behavior within specific chromatography systems, with limited transferability to other systems. To be useful for other experimental set-ups, retention projection models are necessary to transfer retention times from one system to the other. To address this, an alternative retention time modelling approach derives predictions based on measured retention times in separate chromatography systems. The most prominent of these is the PredRet database which stores known compound retention times for multiple chromatography systems with routinely updated generalized additive model (GAM) fits between the shared compounds of similar systems to predict where new compounds may elute, along with prediction intervals.[95] More recently, the CALLC tool unites the QSRR & prediction across chromatography systems to achieve a generalized calibration for improved retention prediction over either modeling technique alone.[96] Overall, these approaches make great use of retention properties for compound annotation purposes, and when combined with MS/MS spectral matching, provide orthogonal information for narrowing down the list of potential identities for metabolomics features.

1.4.4 LC-MS Isotopes, Adducts and In-Source Fragments

When LC-MS preprocessing outputs report thousands of commonly detected <m/z, RT> features, it is important to recognize that these features do not all correspond to unique metabolites. During the electrospray ionization process, metabolites may give rise to multiple

features corresponding to adduct ions, in-source fragments, isotopologues and cluster ions.[97] The presence of these ion multiplicities creates multiple problems in metabolomics data analysis. They contribute a significant proportion of the overall feature count, exaggerating the total number of detected metabolites and complicating downstream statistical analysis. For biological interpretation, it is desirable to reduce data redundancy and obtain a parsimonious representation of the detected metabolome, i.e. through a single representative feature per metabolite. For metabolite identification workflows, "degenerate" features obfuscate the neutral mass of the underlying compound, and by extension its identity, unless the chemical transformation type is correctly assigned. Given these concerns, computationally assisted annotation of ion species is a prerequisite to compound identification as well as statistical analysis. A detailed description of algorithms for metabolomics feature annotation is contained in chapter 2.

1.5 Statistical & Bioinformatics Methods

For most metabolite profiling studies, the aim is to interrogate biological systems and correlate metabolite measurements with phenotype. The levels of metabolites are constantly changing in biological systems due to turnover from reactions, often mediated by specific enzymes, and the influence of environmental factors such as diet, temperature, and microorganisms.[98] As a result, metabolomics is thought to be the most direct signature of biochemical activity and the best indicator of many phenotypic perturbations.[1] Many statistical and bioinformatics approaches have been designed and adapted to extract patterns and contextualize information from metabolomics datasets. Metabolomics shares much in common with other high-throughput biological data types, such as transcriptomics and proteomics. Numerous approaches can be commonly applied to any of these data types, regardless of the biological entity, and many methods originally designed for these molecular analyses can be

similarly applied to metabolomics data.[99] Moreover, like genes and proteins, metabolites have been organized into functional biochemical pathways curated over many decades of biochemical research. Metabolomics is frequently integrated with genomics, transcriptomics, and proteomics to obtain a holistic systems-level view of biological phenomena.[100–103] However, several distinctions make metabolomics data more challenging to interpret. The first is the previously discussed problem of unknown and unidentified metabolites. Observations concerning unnamed compounds are difficult to contextualize and they prevent untargeted metabolomics datasets from being used to their full potential. Unlike transcripts and proteins, metabolites cannot be inferred from mapped genomic sequences and the true size of the metabolome is not fully understood.[104] Furthermore, various classes of metabolites, especially lipids, have not been unambiguously mapped to biochemical pathways as knowledge of substrate-enzyme relationships in biological systems remains incomplete.[105] Metabolomics assays are not yet capable of detecting the full metabolome, with many reaction intermediates missing or present at very low levels.[106] These challenges must be taken into consideration when designing and adapting statistical and bioinformatics approaches to metabolomics data analysis. This section briefly surveys the methods relevant to the research described in this thesis.

1.5.1 Statistical Approaches in Metabolomics

Depending on the study design and the factors of interest, a variety of statistical approaches can be employed to uncover patterns in metabolomics data, a few of which are discussed here. Classic analytical techniques commonly applied to metabolomics data include univariate, multivariate, supervised, and unsupervised approaches. Univariate analyses determine individual variables that have the strongest responses to the investigated conditions, such as differentially abundant metabolites. Student's t-tests & analysis of variance (ANOVA) are typical

univariate approaches for performing differential analysis. Generally, an approach to differential analysis is fitting linear regression models of the form,

$$Y_{m\times n}=X_{m\times p}\gamma_{p\times n}+W_{m\times q}\alpha_{q\times n}+e_{m\times n},$$

where Y represents the matrix of metabolite abundances, X represents observed factors of interest, W represents factors of corresponding to unwanted variation, γ and α are unobserved coefficients that weigh individual factors on the metabolite abundances, and e is an unobserved error matrix.[107] This method determines the influence of the observed factor on the metabolite outcome while accounting for potential confounding factors. Differential analyses yield a test statistic and associated p-value for each metabolite which may include many false positive results given the large number of tests. Multiple testing corrections like the Bonferroni correction or Benjamini-Hochberg correction are necessary to limit the number of false positives.[108]

Multivariate analysis expands on univariate approaches by examining dependencies and correlations between groups of metabolites. Principal Component Analysis (PCA) is an unsupervised multivariate approach used to explain sources of variation in a dataset through a projection onto "principal components" or a series of orthogonal vectors. PCA is often used as an exploratory step to determine primary sources of variability, such as experimental group, population factors such as gender, or batch effects. Likewise, cluster analysis techniques- most prominently hierarchical and k-means clustering- are unsupervised approaches to grouping observations or variables according to a predefined similarity metric, such as Euclidean distance between pairwise correlation vectors. On the other hand, Partial Least Squares (PLS) and its variant Partial Least Squares Discriminant Analysis (PLS-DA) are supervised approaches that seek to determine groups of metabolites associated with a response structure and infer which variables maximize the discrimination between experimental conditions.[99] Classification models,

such as logistic regression, support vector machines, and random forests are widely used in metabolomics to determine biomarkers of disease and other biological states.[109]

1.5.2 Bioinformatics Methods in Metabolomics

Bioinformatics methods facilitate interpretation of metabolomics information through knowledge-based or data-driven approaches. Central to these methods is functional enrichment testing, where sets of biochemically or functionally related metabolites that are overrepresented or "enriched" in their statistical association with some phenotypic trait. Frequently, these metabolite sets are represented as networks containing nodes representing metabolites and edges representing either enzyme-mediated reactions or a high degree of similarity. Known metabolic pathways are one such grouping, with each pathway serving a defined biological role as defined by prior research. Databases of curated pathways, including KEGG,[110,111] MetaCyc[112,113,114], HMDB,[8] and LipidMaps[115] are instrumental in storing and retrieving information about metabolite pathways, though pathway boundaries, the inclusion of certain metabolites in pathways, and the pathway designations are not fully consistent across databases as these are manually defined by researchers.[116] Pathway-based interpretations are the dominant paradigm in metabolomics research, with numerous tools for both enrichment analysis and visualization, such as IMPaLa,[117] MetaboAnalyst,[118] MPEA,[119] MetScape,[120] MetaMapp,[106] Paintomics,[121] 3Omics,[122] and MetExplore.[123] The main statistical methods for enrichment analysis are Fisher's Exact, Kolmogorov-Smirnov, or hypergeometric tests.[124]

The limitations of pathway-based metabolite groupings have led to alternative metabolite set definitions. A notable example is the Chemical Similarity Enrichment Analysis implemented in ChemRICH, where compounds are organized into non-overlapping sets based on their Tanimoto chemical similarity followed by Kolmogorov-Smirnov enrichment analysis.[105]

Grouping by chemical similarity facilitates the inclusion of metabolites that are poorly mapped by metabolic pathways, including important compound classes, resulting in greater utilization of metabolomics datasets than pathway-based methods. However, in its present state, ChemRICH is strictly limited to known, identified compounds with PubChem database identifiers and defined chemical structures.

A more generalized concept independent of biochemical domain knowledge is to construct data-driven metabolic networks using correlation similarities. Since metabolites involved in chemical reactions undergo concerted fluctuations due to factors such as changes in enzyme levels (i.e. due to different regulatory states), temperature and pH, high correlations may be an indicator of chemical relatedness.[98,99] The correlation network can then be clustered, generating a series of subnetworks or modules of metabolites that hypothetically serve a common biological function. A notable example of this methodology is the Weighted Gene Co-expression Network Analysis (WGCNA), which constructs networks by first relating variables by correlation raised to a power followed by module detection via hierarchical clustering.[125] WGCNA has been used in many contexts to determine modules of metabolites (as well as genes and proteins) associated with multiple traits.[126] The principal drawback of most correlation-based networks using Pearson's correlation is the inability to distinguish between direct and indirect associations between metabolites.[127] Two metabolites may be correlated with each other due to confounding factors as opposed to direct biochemical relationships, leading to dense networks of edges between directly and indirectly associated metabolites.

Partial correlations serve as an alternative similarity metric in which the relationship between two variables is conditioned against all remaining variables, thereby controlling for potential confounding relationships that may explain an association. Gaussian Graphical Models

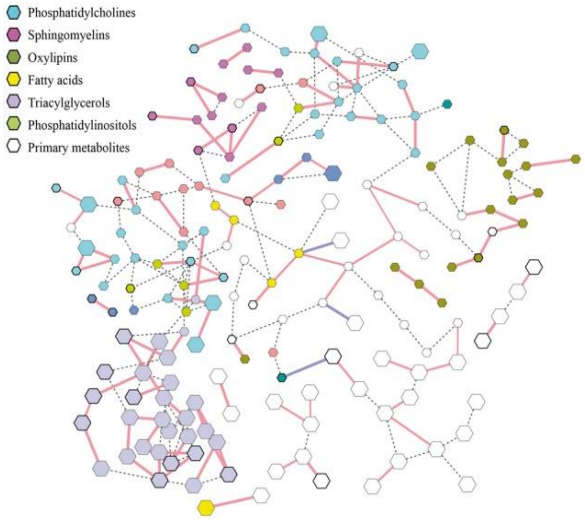


Figure 1.3 Example Partial Correlation Network Adapted from Basu et al. (2017).[131]

(GGM) are networks connecting functionally related metabolites with high partial correlations and have seen much usage in metabolomics studies.[127,128] The recently developed Differential Network Enrichment Analysis (DNEA) method extends the idea of functional enrichment testing to partial correlation networks to assess modules of individual metabolites whose abundance levels and pairwise partial correlations may be altered between two conditions (e.g. disease and control groups).[129,130] First, a partial correlation network is jointly estimated between metabolites for multiple conditions, followed by a consensus clustering to determine densely connected subnetworks, each a self-contained module of the detected metabolome. Then Network Gene Set Analysis (NetGSA) topology-based enrichment method identifies biologically relevant subnetworks that are altered significantly between conditions.[1] DNEA has been shown to generate biologically relevant metabolomics subnetworks associated with disease conditions without *a priori* knowledge of metabolic reactions.

A major limitation of partial correlations is the requirement of a high sample to variable count ratio, a property that is rarely met in the context of untargeted metabolomics experiments which typically contain thousands of features and at most several hundred samples. Certain modifications to partial correlation estimations have been proposed, such as debiased sparse

partial correlations (DSPC) to partially overcome this limitation.[131] Ultimately, applications of partial correlation networks have been limited to targeted metabolomics analyses, datasets with a manageable or reduced metabolite count or in large-scale metabolomics studies where an acceptable sample to metabolite count ratio can be attained.

1.6 Disparate LC-MS Metabolomics Data Analysis

Currently, there are two major repositories in which published metabolomics study data can be deposited: the National Metabolomics Data Repository, also known as Metabolomics Workbench,[132] and Metabolights.[133] These databases contain metabolomics data acquired from diverse species, specimens, disease states, technologies, assay types and contributing institutions. The meta-data for these studies vary in terms of information completeness, but may include both raw and processed data files, lists of identified compounds (with their respective m/z and RT values), and detailed experimental protocols listing instrumentation and parameters used to obtain the underlying data.

A brief look at these databases reveals a lack of standardization of untargeted LC-MS protocols across institutions, even for experiments performed on biologically similar specimens. For example, consider two untargeted RPLC-MS metabolomics studies of human plasma with Metabolomics Workbench identifiers ST000292 and ST000992 by the Southeast Center for Integrated Metabolomics (SECIM) and the Michigan Regional Comprehensive Metabolomics Resource Core (MRC2), respectively. Apart from study aims and design, there are differences in chromatography columns (ACE Excel 2 C18-PFP vs Waters Acquity HSS T3), column dimensions (100 x 2.1mm, 2um vs 50 x 2.1mm, 1.8um), total chromatography time (20 vs 30 minutes), organic solvents (acetonitrile vs methanol), gradients, MS instrument type (Orbitrap vs QTOF), and other important analytical parameters. Both institutions list identified metabolites

found in their respective data, with some commonalities as well as numerous non-overlapping compound annotations.

These two studies exemplify the current state of metabolomics in which laboratories independently develop their own standard operating procedures for metabolite profiling and generate data that cannot be easily harmonized due to m/z, RT, and metabolome coverage disparities.[30] Many cross-laboratory studies have examined the reproducibility and consistency of metabolomics assays across differences in instrumentation, protocols, and processing software.[134–140] To date, however, few tools or resources have been developed for the purpose of aligning and merging LC-MS metabolomics data acquired under disparate LC-MS conditions, defined as experiments in which major chromatographic, spectrometric, and other analytical parameters have been altered. The ability to align data from different laboratories, instruments, and protocols could unlock many opportunities for the metabolomics field, such as intersecting identified and unidentified compound measurements between assays, validating compound annotations, enabling reproducibility assessments, and performing expanded sample analyses. With respect to the latter goal, the integration of data from multiple sources presents an additional layer of technical variability in the form of inter-experimental effects that must be harmonized before further analysis. Addressing these gaps is a central goal of this dissertation.

1.7 Research Objectives

This chapter provides a brief overview of computational metabolomics, highlighting important methods and considerations for relative LC-MS metabolomics quantification, compound identification, and biological interpretation. Abundant tools and resources have been developed to further metabolomics research, though many daunting obstacles remain before untargeted metabolomics can be used to its full potential. Some key issues identified in this

chapter are: 1) high confidence compound identification is achieved for a small proportion of the detected metabolome, yielding a largely uncharacterized list of features; 2) inflated feature counts (due to the presence of redundant ion species) coupled to low sample sizes in untargeted metabolomics datasets, hindering the efficacy of statistical and bioinformatics approaches; 3) the limitation of computational workflows to metabolite profiling experiments performed within a single laboratory, instrument and protocol.

The research described herein was pursued with the following objectives: 1) develop new software and methods for the annotation and subsequent elimination of redundant ion species; 2) develop and implement a novel algorithm for the alignment of metabolomics features acquired from biologically similar specimens under disparate analytical and instrumental conditions; 3) design a comprehensive workflow for merging and normalizing disparately acquired metabolomics datasets, with applications in compound annotation and bioinformatics analysis.

To address objective #1, a standalone Java application called *Binner*, a tool that implements a workflow for the annotation of isotopologues, adducts, in-source fragments, and complexes for metabolomics studies has been developed. *Binner* takes pre-processed and aligned LC-MS metabolomics feature tables and generates a multi-tab report organizing features into retention time and correlation coefficient clusters, with associated pairwise correlation heatmaps and m/z differences displayed in separate tabs. Hypothesized annotations are provided for features combining traditional charge carrying adduct ions with neutral addition or loss groups. *Binner* provides additional resources for the determination of complex adducts and neutral losses that may be prevalent in the data but not defined in adduct rule tables. *Binner* facilitates investigations of metabolomics feature relationships and data reduction via the removal of degenerate features, maintaining features that represent unique analytes.

To address objective #2, the *metabCombiner* software package written in the R statistical programming language for feature matching and concatenation of non-identically acquired untargeted LC-MS metabolomics datasets was developed. The package workflow takes a pair of conventionally processed and aligned feature tables pertaining to biologically similar specimens in the same ionization mode as input and reports all possible matches (constrained by m/z distance) between features as output. *metabCombiner* uses a weighted similarity score metric to determine likely matches between features representing identical analytes in the corresponding tables and filtering unlikely and inferior matches. Multiple examples illustrating the performance of this package, along with applications for facilitating compound identification efforts.

To address objective #3, two population metabolomics studies are presented, each containing two experimental subsets acquired under disparate conditions that require data merging. The first study consists of plasma samples from amyotrophic lateral sclerosis (ALS) patients and healthy controls analyzed as two separate cohorts more than a year apart. The second study consists of women recruited during pregnancy, with maternal plasma collected at the first and third trimesters along with the umbilical cord plasma. Samples were analyzed in two subsets by untargeted LC-MS metabolomics more than three years apart with important chromatographic, instrumental, and experimental design factors altered between the assays. Workflows for the feature data concatenation, harmonization of measured spectral abundances, and statistical and bioinformatics analysis of the combined datasets are presented for each study. These serve as the blueprints for a new type of analysis that shifts the paradigm from the requirement of single-instrument, single-laboratory, common protocol studies towards the integration and analysis of disparately acquired LC-MS metabolomics datasets.

Chapter 2

Deep Annotation and Reduction of Untargeted LC-MS Metabolomics Data with *Binner*

2.1 Introduction

In LC-ESI-MS metabolomics applications, thousands of features are detected as m/z and retention time value pairs. A single metabolite can give rise to multiple mass spectral features corresponding to multiple ion species eluting at roughly the same retention time but with different m/z values. These include isotopologues, adducts, in-source fragments, complexes (e.g. homodimers and trimers), contaminant compounds of non-sample origin, and other artifacts. The ion species that appear and their relative abundances depend on the chemical properties of the metabolite as well as the sample matrix, solvents, mass spectrometer, and other properties.[78] An example mass spectrum is shown in **Figure 2.1** where the metabolite tryptophan generates at least seven peaks corresponding to different positively charged ionic species of distinct mass.

The most common ion types encountered in positive and negative ionization modes are protonated [M+H]+ and de-protonated [M-H]- ions, respectively. These are often considered to be the representative or "principal" metabolite ions as their m/z differs from the neutral mass by approximately one Dalton and they are generally found in high abundances for most metabolites in their respective ionization modes. Other cationic or anionic species, including sodium (Na^+), potassium (K^+), ammonium ($NH4^+$), and chlorine (Cl^-), may carry excess surface charge imparted by electrospray nanodroplets and form non-covalent interactions with metabolites.[141]

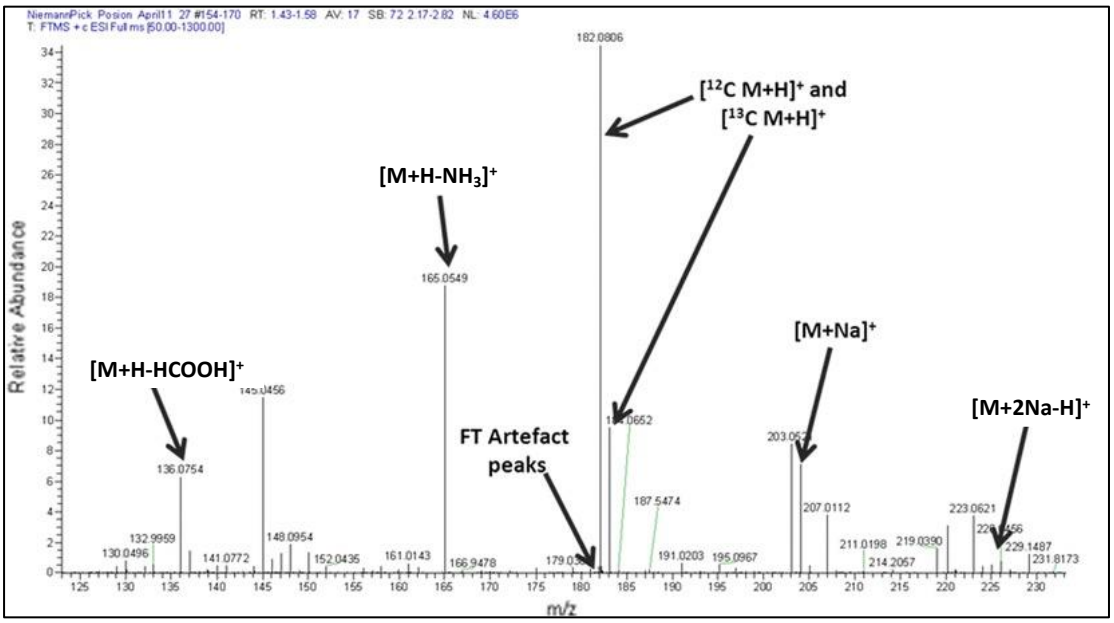


Figure 2.1 L-Tryptophan Mass Spectrum Detected peaks include the protonated form [M+H]+, a ^{13}C isotopologue, two sodium adducts [M+Na]+ & [M+2Na-H]+, two fragments [M+H-NH3]+ & [M+H-HCOOH]+, and a Fourier Transform artifact peak. Adapted from Dunn et al. (2013).

Solvent impurities, glassware, and high salt concentrations in sample matrices (such as plasma or urine) contribute to the presence of these ions.[142,143] In addition, the use of mobile phase additives for enhanced ESI efficiency, such as formic acid or ammonium acetate, contribute significantly to the formation of adducts in metabolomics, e.g. the formation of [M+HCOO$^-$] and [M+CH3COO$^-$] adduct ions in the negative ionization mode.[144]

Isotopologues, compounds with differing numbers of neutrons for at least one of the constituent atoms, are often reported as separate features with distinct m/z values. Some LC-MS preprocessing programs collapse isotopic envelopes automatically, while others such as XCMS report all isotopologues, requiring a post-processing step for their annotation and removal. The most common naturally occurring polyisotopic elements encountered in mass spectrometry-based metabolomics are hydrogen (H), carbon (C), oxygen (O), nitrogen (N), sulfur (S), and chlorine (Cl).[78] Of these, carbon is the most common with roughly 1.11% of natural carbon observed to be the heavier ^{13}C isotope compared to the lighter, more common ^{12}C variant.[145] The two isotopic variants have an observed difference of 1.00335 Da, which combined with the

observed proportions makes ^{13}C isotopologues relatively simple to detect. Accounting for charge, this difference becomes 1.00335/z, where differences between isotopologue m/z values enables the detection of the metabolite's charge state. Sulfur and Chlorine both have stable isotopes differing by close to 1.996 Da (^{32}S vs ^{34}S and ^{35}Cl vs ^{37}Cl), with the heavier isotopes having observed proportions of 4.77% and 24.2%, respectively.[146] Natural isotopes of Nitrogen, Hydrogen, and Oxygen all differ by close to one Dalton and their lesser isotopic forms do not feature as prominently in metabolomics data outside of isotope labeling experiments.

Despite being one of the softest ionization methods, some fragmentation does occur in ESI-MS as metabolite ions readily dissociate between the atmospheric pressure region of the ion source and the vacuum chamber of the mass spectrometer.[147] In source fragmentation patterns are characteristic of the structural make-up of the metabolite and bear many similarities to fragments observed when metabolites are subjected to low collision energy MS/MS.[148] Many prominent fragments correspond to losses of common neutral groups, such as that of H_2O, NH_3, and HCOOH, which can be predicted through the m/z distance between the parent and the fragment ions.[149] Other fragments are more difficult to identify from the data alone and require more advanced strategies, such as matching to knowledge bases of experimental MS/MS spectra.[148] In some cases, fragment ion features observed in metabolomics datasets can often be mistaken for distinct compounds, especially when the fragment ions mimic the m/z and chromatographic retention of common metabolites.[147]

A large proportion of metabolomics features may originate from solvents or contaminants of non-sample origin. A key example are salt complexes consisting of sodium, potassium, chloride, formate, acetate, or other ions that typically elute very early in the chromatogram in reversed phase liquid chromatography assays.[150] Since these clusters are mostly composed of

elements that contribute to high mass defect (defined by the decimal number after the nominal mass), one of the ways used to recognize them is to use filters based on a linear association between mass defect and measured m/z.[150] For background solvent ions detected as features, using negative control samples, robust peak-picking, or excising the early and late retention regions can help eliminate these signals which have little to no value for the analysis.

In many cases, the presence of adducts, fragments, and isotopologues can provide beneficial information for metabolite structure elucidation. Isotopic abundances can be used to deduce molecular formulas and their m/z distances are regularly used to determine charge states. Adduct formation with alkali metals (e.g. Na, K) enables the ionization of metabolites that cannot otherwise be protonated, allowing for more unique metabolites to be detected. Neutral losses point to sub-structures present in the molecule, and their masses may match compounds present in MS/MS fragmentation spectra. Together, adducts, fragments, and isotopologue features enable the triangulation of a common underlying neutral mass.

However, ion multiplicities pose many problems in metabolomics data analysis. Their presence in the dataset artificially increases the number of univariate statistical tests, leading to decreased power following multiple test corrections. In one study by Mahieu & Patti, some 25000 peaks detected in an untargeted metabolomics assay of an *Escherichia coli* extract could be systematically reduced to fewer than 1000 compounds.[151] Without accurate recognition of the ion species, the neutral mass of the compound giving rise to the features is obscured, potentially leading to misidentifications. Furthermore, the high spectral abundance correlation across samples for features arising from the same metabolite hampers searches for biologically meaningful correlations between distinct metabolites, which poses problems for building data-driven networks, such as demonstrated Figure 2.2 below.[131] Thus, annotating ion species in LC-

ESI-MS metabolomics data and performing feature reduction is a critically important task that requires a deep understanding of the complexities present in these datasets.

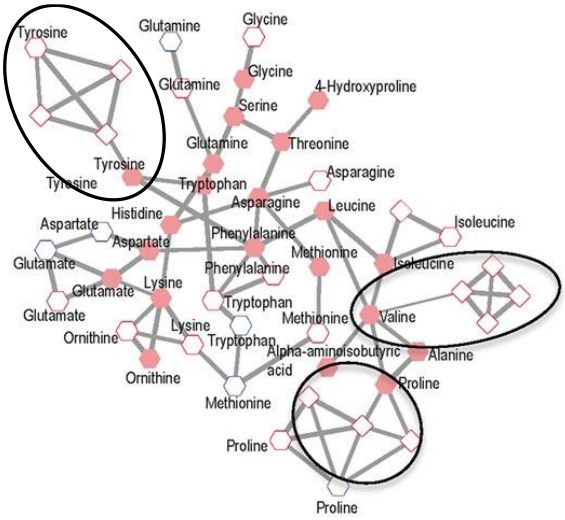


Figure 2.2 Partial Correlation Network with Degenerate Features Nodes enclosed in circles are adducts and fragments for the amino acids Proline, Valine, and Tyrosine. Adapted from Basu et al. (2018).[131]

2.2 Current Tools & Methods

A growing list of tools and strategies have been developed to perform feature annotation and reduction in untargeted LC-MS metabolomics datasets. Most tools follow a similar process in which features arising from the same compound are grouped together in an unsupervised manner, followed by annotation of features using mass relationship rules corresponding to known adducts and neutral losses. Related metabolite ion multiplicities are expected to co-elute, that is, their measured retention times should fall within a very small window. Depending on the software, chromatographic peak shape similarity or correlation of spectral abundances across experimental samples are used as a second similarity measure to detect features from the same metabolite. Major differences between the methods stem from how these tools determine feature groupings and how the peak annotation proceeds within these feature groups.

2.2.1 Pairwise Thresholding Methods

One of the earliest software tools developed for feature annotation is the *AStream* R package by Alonso et al. (2011).[152] The tool first identifies and eliminates samples harboring systematic differences in their metabolomic profile compared to other samples. Then pairwise correlations between all pairs of features are computed, and only feature pairs within retention time and correlation thresholds are retained for further analysis. Annotation of isotopologues follows based on the m/z differences of ^{13}C isotopes versus the more abundant ^{12}C. Finally, adduct and fragment annotations are annotated based on known m/z differences with respect to [M+H]+ and [M-H]- ions. AStream is a template for the simplest annotation workflow in which ion multiplicities are recognized based on simple thresholds of correlation, retention time and m/z difference tolerances. Isotopologues are typically recognized first to simplify the eventual workflow of adduct and fragment ion detection.

A similar workflow to AStream is implemented in the web-based Mass Spectral Feature List Optimizer (MS-FLO).[153] Along with isotopologue recognition, MS-FLO allows for a list of defined mass differences and has capabilities for outright removal of rows satisfying strict correlation cutoffs (if retention time and m/z difference tolerances are also met) or annotation flags if a soft correlation threshold is met. MS-FLO can also flag or remove duplicates and features deemed to be contaminants based on a list of m/z values. Unfortunately, MS-FLO cannot be used for the recognition of complexes (e.g. dimers) or multiply-charged species which cannot be detected by simple m/z distances. PUTMED-LCMS, a trio of workflows for LC-MS feature annotation implemented for Taverna Workbench, also uses simple tolerances for the detection of features arising from the same metabolite and additionally attempts to perform formula and metabolite library matches following data reduction.[154]

2.2.2 Unsupervised Clustering Methods

More sophisticated programs use clustering or generative algorithms for grouping similar metabolomics features. A case in point is the *MSClust* program, which implements subtractive fuzzy clustering, a technique that allows for multiple feature cluster memberships, each cluster defined by a feature "centrotype".[155] The authors of the *RAMClust* R package combined pairwise retention time distances and abundance correlations across samples into a single similarity metric.[156] RAMClust calculates this pairwise similarity between all features and performs a fast hierarchical clustering, separating features into clusters by cutting the resulting dendrogram at a specific height. The results of both MSClust and RAMClust are spectral clusters (assumed to belong to the same compound) that can be supplied to mass spectral search library tools for putative compound annotation. As published, these programs do not annotate ion multiplicities, though a later release of RAMClust makes use of an auxiliary R package, *InterpretMSSpectrum*.[157] The "findMain" function in InterpretMSSpectrum is a heuristic method that successively assigns neutral mass hypotheses to non-isotopologue "major" peaks whose measured intensities are above some specified threshold. The hypothesis that yields the highest weighted sum of explained intensity for the cluster is assigned by the program.

2.2.3 Chromatographic Peak Similarity Methods

One of the most prominent software tools for feature annotation is the *CAMERA* R package, a collection of annotation related methods for mass spectrometry data.[158] CAMERA groups ions by iteratively assigning each to the largest, most abundant overlapping peaks (*groupFWHM*). Subsequently, correlation-based grouping using chromatographic peak shape similarity within samples or Pearson correlation of spectral intensities across samples (or both) proceeds, with a graph connecting highly related features within the compound spectra (*groupCorr*). CAMERA applies one of two algorithms for graph separation to obtain the final set

of pseudospectra within which features are annotated. Isotopologues annotated within these spectra must conform to plausible abundance ratios applying the assumed ~1% natural abundance of ^{13}C (*findIsotopes*). Finally, CAMERA computes adduct, fragment, and multimer annotations based on a dynamic ruleset based on different ion types with pre-assigned score weights; the molecular mass hypothesis that maximizes the sum of scores determines the set of assigned annotations (*findAdducts*). Due to its flexible correlation grouping algorithm, CAMERA is one of the few tools that can handle annotation tasks for single and multiple sample experiments. Chromatographic peak shape correlations require access to the raw spectral data and is generally more computationally intensive to compute than spectral abundance correlations. CAMERA is therefore designed to be paired with XCMS package object inputs.

CAMERA is frequently cited as a benchmark for comparison to other annotation programs in terms of its functionality and performance. *CliqueMS* is a similar program that computes chromatographic peak shape similarities between peaks, using a cosine measure as opposed to Pearson correlations.[159] With a probabilistic generative model, CliqueMS identifies network cliques of related signals, assigning group labels until a customized likelihood function has been maximized. Subsequent adduct and fragment ion annotations are based on summing log plausibility scores, where default plausibility weights are based on observed ion type frequencies in the NIST MS library compound spectra. The performance of CliqueMS compares favorably with CAMERA for internal standard mixtures; however, it is designed for single sample analyses only as opposed to aligned multi-sample datasets.

2.2.4 Bayesian Probabilistic Methods

Bayesian probabilistic models have been explored in the context of feature annotation. Rogers et al. (2009) provided the basis for these efforts by showing that multiple sources of

information (especially isotopic distributions) can be used to generate probabilistic models assigning mass spectral peaks to molecular formulas.[160] *MetAssign* by Daly et al. (2014) is a software implementation that assigns related features to clusters centered around a common molecular formula, using a Bayesian Markov Chain Monte Carlo sampling approach that integrates m/z, RT, and the inter-peak dependency structure.[161] This method requires a list of formulas whose isotopic profiles and possible adduct m/z values are computed, with a prior probability distribution assigning to formulas based on m/z distance; a posterior probability is computed for cluster and formula assignments, giving some measure of confidence in the program-computed annotation. The Integrated Probabilistic Annotation (IPA) software package further modifies this formulation by inclusion of prior information, such as which adducts and fragments may potentially form for each metabolite (including the most abundant form), the RT range where the compound is most likely to be present, and biological connectivities.[162] One disadvantage is that methods utilizing Gibbs Sampling are more computationally intensive and slower to complete than typical annotation programs.

2.2.5 xMSannotator

A deterministic variant of these simultaneous adduct and compound assignment strategies is implemented in xMSannotator.[163] This software package annotates features through a six-step process which includes pairwise correlation calculations between abundance vectors, applying WGCNA for module detection of highly correlated features,[125] kernel density estimation for RT grouping, and subsequent matching to a database of compounds selected by the user (such as HMDB, KEGG, or LipidMaps). A score is assigned to compound and adduct matches based on the number of matched adducts and isotopes, correlations, sum of adduct weights (pre-assigned or user-defined), retention time range, and isotope distribution. As an added refinement,

biological network information may be incorporated to improve the scores of adduct and compound matches grouped in the same module as biochemically related molecules. Like with MetAssign and IPA, the computational burden of xMSannotator grows rapidly with the size of the compound database, the number of adduct and fragment annotations, and the feature count, leading to slow overall processing times compared to other methods.

2.2.6 mz.unity

The mz.unity R package contains a unique approach to enumerating all possible mass relationships in a list of metabolomics peaks.[164] Its foundation is an algorithm called mz.sum which enumerates peak relationships as the gain or loss of multiple charged formulas. mz.unity pursues many classes of relationships that are often not covered by most annotation methods, such as heteromers (complexes involving multiple metabolites), distal fragments, and complexes with background ions. It also provides graph-like structures to enable the visualization of mass relationships as a network. However, mz.unity does not attempt to group or cluster related features, leaving multiple potential interpretations for many of the assigned annotations that cannot be resolved without further information.

2.2.7 Credentialing Approaches

Credentialing is an experimental method for validating features derived from sample origin as opposed to chemical noise and contaminants.[165] This technique involves harvesting cellular cultures in different isotopically labeled media (e.g. ^{12}C and ^{13}C) and examining the resulting m/z spacing and intensity ratios. Isotopologue products must pass an expected ratio criterion to be deemed "credentialed", sample-derived metabolites. Mahieu et al. demonstrated its utility for optimizing preprocessing parameters using credentialed peaks as the primary criterion as opposed to total feature count.[165] The Peak Annotation and Verification Engine

(PAVE) method incorporates credentialing into its workflow.[166] The method requires four separate cultures (unlabeled, ^{13}C, ^{15}N, and $^{13}C + ^{15}N$) which it uses to estimate the carbon and nitrogen counts of labeled compounds. The atom counts serve as an additional criterion for adduct detection, which PAVE performs through common m/z-based relationship rules. While credentialing has its benefits, it is limited to sample types conducive to isotopic labeling. For a more generalizable experimental method, the authors of PAVE introduced the Buffer Modification Workflow (BMW) consisting of assays swapping mobile phase additives (e.g. $^{14}NH_3$ acetate and $^{15}NH_3$ formate) to credential metabolites.[167]

2.2.8 Miscellaneous Methods

Some recently developed tools consist of scripts or modular implementations within larger workflows. CROP (Correlation-based Reduction Of multiPlicities) is an R script in which features are clustered by pairwise feature intensity correlations as well as retention time.[168] Analytic Correlation Filtration,[169] a stand-alone tool as well as a module within the Workflow4Metabolomics,[170] uses thresholds of pairwise correlation (Pearson, Spearman, or partial) similarities, RT and m/z difference tolerances like other previously described tools. Both tools then choose a single representative feature among those that have been clustered or linked together, which is usually the feature with the highest mean abundance.

2.2.9 Limitations of Current Approaches

The proliferation of feature annotation and reduction methods reflects the significance of the problem of data redundancy and the need for automated approaches in routine untargeted metabolomics datasets. Still, many shortcomings of existing tools must be addressed. Some methods are limited to individual sample analyses, whereas most untargeted metabolomics datasets contain multiple samples with aligned spectral features. Visualizations of feature

relationships is another key aspect missing from all but a handful of programs. Metabolomics annotation programs frequently come with a packaged table of mass relationship rules, with no clear way of exploring if additional unannotated adducts or fragments may be present. Formation of in-source adducts and fragments may differ between LC-MS metabolomics protocols, and tailoring annotation rules to discover all possible multiplicities in specific LC-MS systems is an arduous process. Thus, many ion multiplicities are either left unannotated or are misannotated due to the unexpected feature relationships present in the data. Many methods have strict thresholds for correlation that are difficult to set objectively and hinder the discovery of true relationships if strict cutoffs are not met. Lastly, many software packages have time consuming processes and are difficult to operate for users without advanced computational training. Ideally, feature annotation should be fast, reproducible, informative, accurate, and easy to use to accommodate the routine nature of this task in metabolomics data analysis pipelines.

2.3 Methods

2.3.1 Overview of *Binner*

Binner is a standalone, platform-independent application for LC-MS metabolomics feature annotation written in Java.[171] The input for Binner is a table of aligned metabolomics features consistent with data generated by metabolomics pre-processing programs, where each row represents an individual analyte. The table must have columns containing the measured m/z, RT, and per-sample feature abundance values for at least three samples. A column of

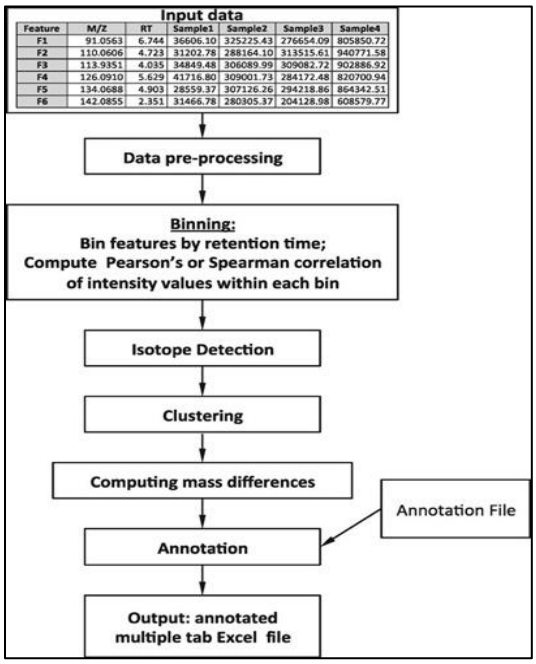


Figure 2.3 *Binner* Workflow Overview

character identifiers is also required for each feature as well. Additional columns in the input file may be included in the results. The main workflow for *Binner* is illustrated in **Figure 2.3**. After data cleaning and value transformation steps, features are organized into retention time bins, isotopologue groups and correlation clusters, before adduct and neutral loss fragment labels are assigned based on a user-defined annotation file. The output of *Binner* is a spreadsheet with multiple tabs consisting of summary analysis information, feature groupings, raw and adjusted abundance values, and separate sheets with the designated ion types. Each step is described in detail here.

2.3.2 Data Processing

Multiple processing steps are undertaken to aid subsequent analysis. First, missing feature abundance values must be appropriately handled as they may bias correlation estimates between features. As a preliminary step, features with abundance values missing above some proportion of samples (by default 30%) are filtered and thus excluded from further processing. Remaining missing values are imputed with the median abundance of the pertinent feature. Since correlation estimates are susceptible to outliers, these points are detected as values more than 'n' standard deviations from the mean (n = 4 by default) and treated similarly to missing values. As an optional, but recommended step, intensity values are log-transformed to correct for potential heteroscedasticity and obtain symmetric value distributions. Untransformed and transformed values are provided as output tabs in the output file, with detected outliers and their replacement values highlighted wherever they have been corrected.

2.3.3 Retention Time Binning

The central workflow step is the grouping of features by chromatographic RTs through a binning step. Binning segments the feature list into self-contained units or "bins", which greatly

simplifies subsequent clustering computations. Features are first sorted in RT order and a new bin is formed whenever the difference in RTs between consecutive analytes exceeds some user-defined gap parameter (by default 0.03 minutes). The gap parameter value should be chosen to exceed the expected difference in RT between features derived from a common metabolite, while simultaneously accounting for the density and separation of features. Pairwise Pearson or Spearman correlations are calculated between all pairs of features contained within the bins.

2.3.4 Isotopologue Detection

Isotopologue detection is an optional and highly recommended step for datasets where isotopes have not been pre-filtered. Since most isotopologue features in LC-MS metabolomics feature lists occur due to the presence of ^{13}C isotopes, the detection is performed based on the observed behavior of these isotopes. Binner searches for pairs of features within bins whose m/z values differ by 1.0033 / z Daltons (i.e.. 1.0033, 0.5016, and 0.3344 for singly, doubly, and triply charged compounds, respectively). Such features are considered as potential isotopologues. Additional criteria are imposed on these features, such as RT distance (0.1 minutes by default) and abundance correlation threshold (default 0.6). In addition, due to the lower abundance of ^{13}C relative to ^{12}C, isotopologues must decrease in median intensity with increasing mass, a condition met by all low mass compounds. The mass spacing of the detected isotopologues allows for the assignment of charge states to features at this stage, without which features are assumed to have charge state z = 1. Moreover, isotopologues apart from the parent monoisotopic mass feature are grouped together and temporarily removed from processing until the eventual output stage. This facilitates correlation clustering by size and complexity of bins. Isotopologues not detected at this stage may be detected as part of the downstream annotation process.

2.3.5 Correlation Clustering

Correlation clustering is a common problem in bioinformatics with numerous approaches for grouping together similar objects separately from those that are dissimilar. In this workflow, correlation clustering follows retention time binning to determine collections of tightly related features arising from a common parent metabolite. A key challenge is to determine the optimal number of clusters comprising each bin, accounting for the underlying correlation structure of the features. For each bin, Binner performs an average linkage hierarchical clustering using the Euclidean distances between pairwise correlation vectors, deriving the structure of pairwise dissimilarities between features. Then the hierarchical clustering dendrogram is cut at a height that produces the cluster number (k) that maximizes the average silhouette coefficient among all bin features. Silhouette coefficients are evaluated as the relative difference between intra-cluster and inter-cluster dissimilarities, with higher silhouette coefficient values (close to 1) reflecting high inter-cluster to intra-cluster dissimilarity ratios.[172]

An example of this binning and clustering formulation is shown in **Figure 2.4**. Three distinct bins form in the range (RT = 7.31 - 7.61 min) due to gaps between the retention times of consecutive ordered features exceeding 0.05 min. From there, Bin B, the largest of the three bins,

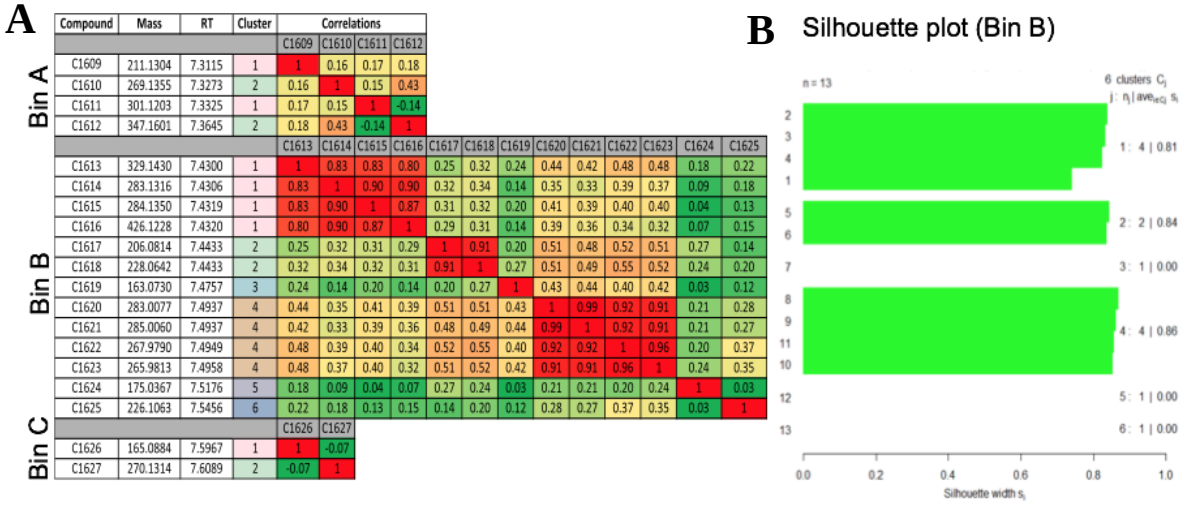


Figure 2.4 Example Bins and Correlation Clusters (A) Three bins displayed from RT = 7.31 to 7.61 min, separated by 0.05 min bin gaps. (B) Silhouette plot of the optimal cluster number (6) for Bin B.

contains six easily identifiable clusters characterized by highly correlated submatrices located along the matrix diagonal. The choice of six clusters maximizes the average silhouette coefficient, as illustrated by the silhouette value plot.

A few additional heuristics have been implemented to refine the clustering formulation described here. In the trivial case of small bins entirely composed of highly correlated, tightly eluting feature elements where no further clustering is necessary, a score is computed linking together average correlation, retention time range, and the total number of features (n) as

$$Score = \frac{(Corr_{avg})^2}{log_2(n)\sqrt{rt_{max} - rt_{min}}}$$

Higher scores imply greater cohesiveness of the bin, implying that it should not be clustered. A threshold value of 2 is set by default to prevent clustering of high-scoring bins. In many bins, the cluster number chosen by maximizing average silhouette value is often observed to select k = 2, leading to an under-clustered solution containing dissimilar elements. To address this, a weighted silhouette variant includes a multiplicative term to the silhouette coefficient that imposes a penalty to the within-cluster dissimilarity. Finally, as clusters are first determined purely through correlations, it is possible to have a wide RT spread among the cluster elements. To counter this, correlation clusters with excessive RT spreads are subdivided by a further clustering by pairwise RT distances (which proceeds similarly to correlation clustering), or alternatively, a re-binning of features occurs using the RT gap parameter. The resulting subclusters serve as the final, most refined grouping of metabolomics features prior to the annotation step.

2.3.6 Annotation Process

Binner requires a customized annotation file listing the possible "charge carrier" element(s) and neutral gain / loss groups that may be observed. Common charge carriers include +H, +Na,+ K, and combinations of these elements in the positive mode, as well as -H, +Cl, or

+COOH in the negative mode. Binner annotations may consist of these charge carriers alone (e.g. [M+H]+, [M+Na]+, [M-H]-, etc...) or a combination of charge carriers with neutral groups like H_2O or NaCOOH (e.g. [M+H-H_2O]+, [M+Na+NaCOOH]+). Dimers and trimers may also be considered in this formulation (e.g. [2M+H]+, [3M+2Na]2+), granting Binner a wide annotation search space. Unless isotopic evidence points to the existence of multiply charged (|z| > 1) species, features are assumed to be singly charged. Given a specific neutral mass M, total multiplicity n, and total charges of the associated charge carriers, z, an annotation is assigned to a feature if its m/z value falls within some error of the formula:

$$m/z \approx \frac{nM \pm mass(charge\ carriers) \pm mass(neutral)}{z}$$

The m/z search error parameter is set to 0.002Da by default. Increasing this value increases the total number of annotations, but typically at the expense of more frequent misannotations.

The central premise of the Binner annotation method is that the in-source formation of certain ion types are generally favored over others, leading to higher relative abundances. Protonated [M+H]+ and de-protonated [M-H]- forms are common and highly abundant in the positive and negative ionization modes, respectively, however it is possible for sodium [M+Na]+, chloride [M-Cl]-, or other forms to be the dominant form depending on the metabolite properties. It is also possible for protonated and deprotonated ions to be absent entirely, or for multiply charged ions to dominate. The idea is therefore to determine the most abundant ion form in the spectrum, denoted as the "principal ion", which can then be used to calculate the corresponding neutral mass that generates the most annotations within the clustered spectra.

In each cluster, *Binner* finds the feature with the highest median abundance across the experimental samples. This feature is iteratively assigned an adduct hypothesis corresponding to the most frequent ion types (M+H, M+Na, etc...), which are obtained from the charge carriers for

the given feature's charge state. For each hypothesis, an underlying neutral mass is calculated from the m/z value for the feature along with the mass and charge of the hypothesized adduct. *Binner* then uses each neutral mass to search for annotations that account for the reported m/z values of other features in the cluster. The hypothesis that optimizes the number of putatively annotated features determines the assignment of the principal ion, associated neutral mass, and related features annotated with respect to the neutral mass. These steps are repeated with remaining unannotated features in the cluster until the process has been attempted at least once for every feature. In the final output, features are categorized as principal ions, degenerate features, or unannotated features, with degenerates deemed removable for downstream analysis.

2.3.7 Evaluation Dataset

The main dataset used to evaluate Binner is a metabolite profiling study of plasma from 80 individuals in a population-based study of myocardial infarction in Costa Rican adults. Briefly, plasma samples were thawed on ice, deproteinized, and analyzed on an Agilent 1290 LC / 6530 qTOF MS system (Agilent Technologies, Inc., Santa Clara, CA USA) using the Waters Acquity HSST3 1.8 uL 2.1 x 100 mm column (Waters Corporation, Milford, MA). Each sample was analyzed twice, once in positive and once in negative ion mode. For both ion mode runs, mobile phase A was 100% water with 0.1% formic acid and mobile phase B was 100% methanol with 0.1% formic acid, with a gradient that proceeds as follows: 0-2 min 2% B, 2-20 minutes 2-75% B (linear), 20-22 min 75-98% B (linear), followed by a 7 minute reequilibration at starting conditions. The flow rate was 0.46 mL/min and the column temperature was 40°C. The injection volumes for positive and negative mode were 5 µL and 8 µL, respectively. Electrospray ionization is performed with an Agilent Jetstream ion source, with full-scan mass spectra acquired over the m/z range 50-1500 Da. Source parameters were: drying gas temperature

350°C, drying gas flow rate 10 L/min, nebulizer pressure 30 psig, sheath gas temp 350°C and flow 11 l/min, and capillary voltage 3500V, with internal reference mass correction.

Raw spectral files for all samples were converted to .mzXML format using the MSConvert tool of ProteoWizard 3.0. Positive and negative mode samples were processed using the XCMS R package Version 3.0.0. Peak picking was performed with the CentWave method, with peakwidth from 10-40s, ppm = 30, mzdiff = 0.01 and snthresh = 6. The obiwarp method was used to perform RT correction, with profStep = 0.5. Features were grouped across samples with the default density method, with bw = 5, mzwid = 0.033, and minfrac = 0.5. Missing spectral information was imputed with the fillPeaks() method. PCA revealed one sample to be an outlier and it was excluded from further processing, leaving 79 samples. Features were arranged in RT order and assigned generic compound labels (C1, C2, ...).

The resulting XCMS feature table was processed with *Binner*. Putative principal ions and other non-annotated ions were searched against the Michigan Regional Comprehensive Metabolomics Resource Core's (MRC^2) in house RT-MS/MS library using 0.005 Da mass accuracy, 0.01 Da MS/MS fragment mass accuracy and 0.1 min RT as matching criteria. MS/MS data were processed using the Find by Targeted MS/MS algorithm (Agilent) and library search (default weighting) with a score of at least 50 to be considered a match. Metabolites not in the MRC^2 library were searched against the NIST17 MS/MS library and required a reverse dot product score of 700 to be considered a match.

2.4 Results

2.4.1 Implementation and Output

Binner is implemented as a user-friendly Graphical User Interface containing five separate tabs ("Input", "Output Options", "Data Cleaning", "Feature Grouping", and

"Annotation") for loading data and adjusting parameters for the analysis. The output is an excel spreadsheet providing summary analysis information and multiple tabs containing organized and formatted feature summaries that facilitate further data exploration. **Figure 2.5** depicts an example output for two annotated clusters of closely eluting features. The first three columns contain the user-provided feature name, m/z and RT for every feature as read from the input file. The remaining columns show *Binner*-generated information, such as the retention time bin and correlation cluster to which the features belong, and pairwise abundance correlation coefficient matrices that are useful for validating feature relationships. The two clusters each contain seven features. In the first cluster, *Binner* annotated C908 and C909 as [M+Na] and [M+H], respectively, which have a correlation coefficient of 0.45. It is not uncommon for commonly occurring adduct pairs to have relatively low correlations that could easily be missed by a method that imposes a rigid correlation threshold. *Binner* was able to annotate these features as likely representing a single metabolite based on close RTs and the mass relationship with a common neutral mass (345.0022).

By contrast, the second cluster contains highly correlated features. C930 and C934 are identified as isotopologues of C931 and C935, respectively. The subsequent annotation process identified C931 and C925 as the [M+H] and [M+Na] adducts, respectively, and C935 as a loss of NH_3 fragment. These five features and isotopes belong to the feature group for kynurenine,

Feature	m/z	RT	Median Intensity	Isotopes	Annotations	Derivations	Bin	Cluster	Correlations						
									C899	C901	C902	C903	C908	C909	C911
C899	262.164217	2.000083	21775				5	2	1	0.36	0.15	0.44	0.37	0.44	0.36
C901	257.112558	2.014600	15892				5	2	0.36	1	0.54	0.34	0.49	0.24	0.35
C902	212.091170	2.025367	15459				5	2	0.15	0.54	1	0.27	0.28	0.26	0.17
C903	112.075556	2.030183	5239				5	2	0.44	0.34	0.27	1	0.30	0.28	0.23
C908	367.991464	2.065508	32740		[M151 + Na]	[345.0022 + 22.989221]	5	2	0.37	0.49	0.28	0.30	1	0.45	0.45
C909	346.009471	2.071500	146303		[M151 + H]	[345.0022 + 1.007276]	5	2	0.44	0.24	0.26	0.28	0.45	1	0.31
C911	347.016043	2.096658	63798				5	2	0.36	0.35	0.17	0.23	0.45	0.31	1
									C925	C926	C929	C930	C931	C934	C935
C925	231.073853	2.204233	14842		[M152 + Na]	[208.0848 + 22.989221]	5	3	1	0.77	0.88	0.80	0.85	0.85	0.88
C926	146.060892	2.204517	12056				5	3	0.77	1	0.82	0.84	0.75	0.82	0.79
C929	94.065463	2.208233	10370				5	3	0.88	0.82	1	0.87	0.93	0.90	0.96
C930	210.096310	2.208400	48170	[m125 + 1]			5	3	0.80	0.84	0.87	1	0.70	0.87	0.77
C931	209.092046	2.210325	391210	[m125]	[M152 + H]	[208.0848 + 1.007276]	5	3	0.85	0.75	0.93	0.70	1	0.80	0.99
C934	193.069473	2.212225	16750	[m126 + 1]			5	3	0.85	0.82	0.90	0.87	0.80	1	0.84
C935	192.065405	2.213025	128157	[m126]	[M152 + H - NH3]	[208.0848 + 1.007276 - 17.0254]	5	3	0.88	0.79	0.96	0.77	0.99	0.84	1

Figure 2.5 Example Annotated *Binner* Clusters

which was identified by accurate mass, RT and MS/MS compared against an authentic standard in our compound library. Two remaining unannotated features, C926 and C929, are highly correlated with the rest of the features within the cluster and elute at an identical RT. Referring to an in-house MS/MS library shows that these features matched fragment ions observed in low-energy CID fragmentation of kynurenine. This demonstrates the utility of correlation matrices in manually uncovering feature relationships that are not pre-specified or easily described, facilitating the discovery of additional degenerate features.

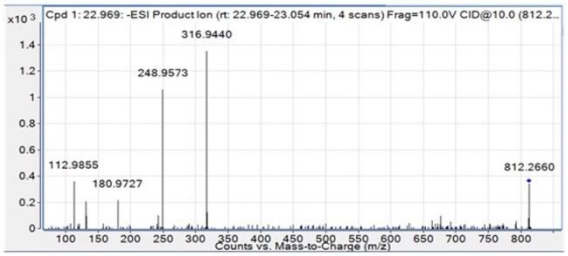
2.4.2 Mass Differences

In addition to annotations and correlation heatmaps, *Binner* generates pairwise mass difference matrices that can provide guidance for further discovery of feature relationships and new annotations within a cluster. An example is depicted in **Figure 2.6** of a cluster from the negative mode data containing LysoPC 16:0. The matrix of observed mass differences in panel C shows multiple instances of 67.987 and 57.958, which equate to the masses of NaCOOH and NaCl, respectively. There are linear combinations of these values, such as 135.974 (= 2 x NaCOOH), 193.933 (= 3 x NaCOOH), and 125.945 (= NaCOOH + NaCl), which when

A

Isotope	Annotations	Derivations
[m364]	[M423 + Cl-]	[495.3329 + 34.969402]
[m363]	[M423 + COOH-]	[495.3329 + 44.998201]
[m366]	[M423 + COOH- + NaCl]	[495.3329 + 44.998201 + 57.958622]
[m368]	[M423 + COOH- + NaCOOH]	[495.3329 + 44.998201 + 67.987424]
	[M423 + COOH- + NaCl + NaCOOH]	[495.3329 + 44.998201 + 125.946046]
[m369]	[M423 + COOH- + 2X NaCOOH]	[495.3329 + 44.998201 + 135.974848]
	[M423 + COOH- + 2X NaCOOH + NaCl]	[495.3329 + 44.998201 + 193.933448]
[m371]	[M423 - H + 3X NaCOOH + HCOOH]	[495.3329 - 1.007276 + 249.967751]
	[M423 + COOH- + 3X NaCOOH + NaCl]	[495.3329 + 44.998201 + 261.920872]
[m370]	[M423 + COOH- + 4X NaCOOH]	[495.3329 + 44.998201 + 271.949696]
	[M423 + COOH- + 4X NaCOOH + NaCl]	[495.3329 + 44.998201 + 329.908318]
[m373]	[M423 + COOH- + 5X NaCOOH]	[495.3329 + 44.998201 + 339.93712]

B



C

Isotopes	Annotations	Derivations	Mass Diffs							
			C2508	C2509	C2510	C2511	C2512	C2513	C2514	C2515
	[M423 + COOH- + 2X NaCOOH + NaCl]	[495.3329 + 44.998201 + 193.933448]	0.0000	57.9585	-10.0290	-135.9743	55.9535	-4.0115	123.9402	-78.0162
[m369]	[M423 + COOH- + 2X NaCOOH]	[495.3329 + 44.998201 + 135.974848]	-57.9585	0.0000	-67.9874	-193.9327	-2.0050	-61.9699	65.9817	-135.9747
[m371]	[M423 - H + 3X NaCOOH + HCOOH]	[495.3329 - 1.007276 + 249.967751]	10.0290	67.9874	0.0000	-125.9453	65.9824	6.0175	133.9691	-67.9873
	[M423 + COOH- + 4X NaCOOH + NaCl]	[495.3329 + 44.998201 + 329.908318]	135.9743	193.9327	125.9453	0.0000	191.9278	131.9628	259.9144	57.9580
[m369 + 2]			-55.9535	2.0050	-65.9824	-191.9278	0.0000	-59.9649	67.9867	-133.9697
	[M427 + COOH- + Na-H]	[671.2944 + 44.998201 + 21.9821]	4.0115	61.9699	-6.0175	-131.9628	59.9649	0.0000	127.9516	-74.0048
[m368 + 2]			-123.9402	-65.9817	-133.9691	-259.9144	-67.9867	-127.9516	0.0000	-201.9564
[m370]	[M423 + COOH- + 4X NaCOOH]	[495.3329 + 44.998201 + 271.949696]	78.0162	135.9747	67.9873	-57.9580	133.9697	74.0048	201.9564	0.0000

Figure 2.6 Complex Annotations of LysoPC 16:0 (A) Feature annotations of clustered ions derived from metabolite "M423" consisting of $COOH^-$, Cl^-, and Na^+ combinations. (B) MS/MS performed on m/z = 812.266, indicating spacings of 67.98 (NaCOOH), validating annotations. (C) Associated mass difference matrix.

considered with respect to the neutral mass of the LPC 16:0 principal ion enables the annotation of all features as complex adducts consisting of multiple Na, COOH, and Cl additions. Additional MS/MS experimental evidence was obtained for this group of features, with multiple fragments in the resulting spectra that corroborate the existence of sodium and formate cluster ions, as shown in panel B of **Figure 2.6**.

Binner also builds a histogram highlighting the most frequent mass differences observed throughout the dataset to enable the discovery of additional neutral mass gains and losses that may not have been considered or included in the annotation file. This concept is illustrated in **Figure 2.7** below for a cluster containing the Leucine metabolite detected in the positive ionization mode. Panel A shows the tabulated mass differences between features, with common mass differences highlighted in different colors, and Panel B shows part of the mass difference distribution computed for the full dataset. Column 3 in Panel A shows a mass difference between

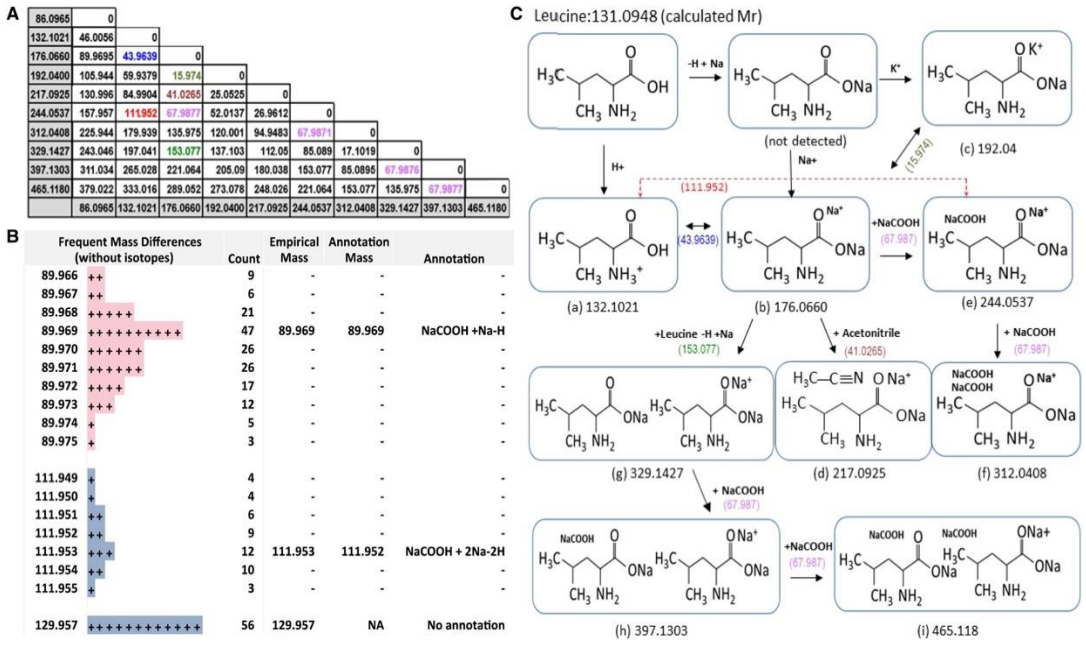


Figure 2.7 Deriving Complex Annotations from Mass Differences (A) matrix showing common pairwise m/z differences that can explain a cluster containing Leucine in HILIC-POS dataset. (B) Mass difference distribution output showing common pairwise differences (with known and unknown annotation explanations). (C) Flow chart showing the complex chemical additions of Leucine

feature 176.066 and Leucine M+H (132.1021) of 43.9639 that corresponds to +2Na-2H, and Column 4 shows a mass difference of 67.9877 between features 244.0537 and 176.066 that can be annotated as +NaCOOH. It can therefore be concluded that the frequently observed mass difference 111.952 shown in Panel B results from the combination of two masses: 43.9636 and 67.987. Based on these findings, the mass 111.952 was added to the annotation library as +2Na-2H+NaCOOH. In this way, *Binner* allows users to explicitly define series of combinations of smaller neutral gain/loss groups (complex adducts) that are observed in the data.

2.5 Evaluation

2.5.1 Criteria and Procedure

To evaluate *Binner*, the accuracy and thoroughness of its main annotation workflow were compared to those of three feature annotation programs: CAMERA[158], MS-FLO[153], and xMSannotator[163]. These tools were selected as they have a similar scope of operations to Binner in that they accept aligned spectral features as input and generate annotations as part of their automated processes. For this evaluation, 75 identified known compounds were detected in the positive ionization mode within the dataset described in Section 2.3.7. These 75 compounds are well distributed throughout the chromatogram, range in relative abundance in human plasma analyzed by RPLC-QTOF, and their principal ion annotations have been verified by accurate mass, retention time, and MS/MS matches to reference spectra. Each tool was equipped with a comparable, non-exhaustive set of adducts and neutral losses to search for. In addition, common parameter settings were used for each program, wherever applicable: annotation RT tolerances are set to 0.05 min, mass tolerances set to 0.005 Da (10 ppm for xMSannotator), and all correlation thresholds are set to 0.7.

The performance of each program was evaluated according to four different criteria. The first is the ability to correctly annotate the principal ions for the 75 known compounds, where principal ions are defined as the most abundant representative feature of each metabolite. The second is the total number of adducts, neutral losses, and multimers annotated by each program for the 75 compounds. For this, the union of these ion types grouped together with the identified principal ion across all four tools was used as the benchmark, with the restriction that only annotations explicitly labeled in annotation rule files may count towards this score. The third is the number of correctly assigned isotopologues, which was similarly benchmarked by the union of isotopologues found by all programs with respect to the 75 compounds and their multiple ion types. Isotopologues must be labeled exactly (i.e. M+3 isotopes labeled as M+1 do not count), have plausible relative abundances, and second (M+2) isotopes must have a mass difference between 1.995 to 2.011 to count for this analysis. Finally, the programs were evaluated in terms of how well they could group together features arising from a common metabolite within the same cluster. This criterion counts the total number of cluster groups encompassing the set of annotated adducts, in-source fragments, and multimers, penalizing instances in which these ion forms are placed into separate feature clusters. The features need not be annotated correctly to count for this final criterion.

2.5.2 Evaluation Results

The results of the evaluation are contained in **Table 2.1** and **Figure 2.8** below. *Binner* correctly annotated the principal ions for 64 out of the 75 selected compounds, compared to 51, 49, and 47 PIs for xMSannotator, MS-FLO, and CAMERA respectively. Only 32 principal ions were correctly annotated by all four annotation programs. *Binner* also outperformed the other programs in terms of accurate adduct/NL/multimer annotations, with 225 annotated consistently

with the correct principal ions compared to 184, 152, and 155 for CAMERA, MS-FLO, and xMSannotator, respectively. CAMERA and Binner are the closest in terms of their clustering performance, with a nearly identical number of total clusters containing the union of all ion types considered for this evaluation. One of the major advantages of Binner is that it takes relative abundances into consideration, particularly as most in source fragments and adducts tend to form at lower abundances in electrospray ionization in comparison to [M+H]+ and [M+Na]+ ions. On the other hand, CAMERA utilizes user-specified weights for each ion type and chooses the annotation set that maximizes the sum of weights without consideration of relative abundances, leading to the mis-annotation of many principal ion-fragment pairs.

The fixed mass difference approach used by MS-FLO fails to annotate any multimers or multiply charged adducts, which make up a large proportion of all potential annotations. While the number of annotations identified by MS-FLO could have been increased by explicitly defining certain fragment/parent ion pairs (e.g. M+H-H2O/M+H, mass diff = ~18.01), some neutral groups (such as NH3) can be either added or lost (e.g. M+H/M+NH3 and M+H/M+NH4) and these cannot be defined simultaneously within MS-FLO. The formula-based matching strategy employed by xMSannotator also had many limitations. The program assigns incorrect

Table 2.1. Summary of Annotation Evaluation Results

	Binner	**CAMERA**	**MS-FLO**	**xMSannotator**
No. of correctly annotated PIs	64	47	49	51
Total no. of adduct/NL /multimer annotations**	225	184	152	155
Total no. of isotope annotations**	201	220	157	86
Total no. of feature groups	87	86	N/A*	133

*MS-FLO not included in this comparison because it does not explicitly define feature groups.
**Isotopes and adducts only counted if consistent with the correct PI interpretation

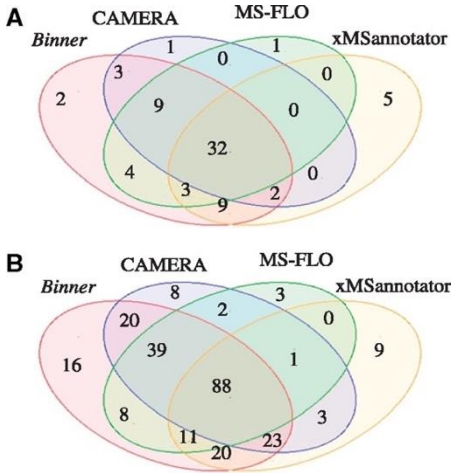


Figure 2.8 Annotation Results Venn Diagrams Overlap of annotations generated by different programs. (A) Principal Ion annotations; (B) Adduct and neutral loss fragment annotations

formulas to many features, mostly due to incidental mass matches. Further, although isotopes (including non-carbon isotopes) can be detected and used to enhance confidence in the assigned formulas, isotope detection in xMSannotator is largely limited to M+H as the charge carrier, which results in fewer annotations. One unique feature of xMSannotator is the use of database searching coupled with pathway information, which might provide an advantage in situations where feature annotations cannot be identified solely based on pairwise correlation and mass relationships. xMSannotator was the only program that could annotate metabolites with only one feature, which account for most principal ion annotations that Binner and other programs missed. Despite this advantage, there were much fewer confident assignments compared to the number of low confidence assignments.

Overall, this evaluation demonstrates that Binner has comparable clustering and isotopologue annotation performance with existing programs, and superior principal ion and adduct/fragment/ multimer annotation performance.

2.6 Limitations and Potential Improvements

There are several limitations to the Binner annotation workflow that are discussed here, along with suggestions for potential improvements to overcome these issues.

2.6.1 Large Bins

One of the most import problems with the current Binner formulation is dealing with feature-dense regions of the chromatogram that produce excessively large bins. In human plasma applications, it is common to see thousands of features associated with non-polar lipids elute very late (for RPLC-MS) or very early (HILIC-MS). In LC-MS metabolomics studies with inadequate separation between features, *Binner* typically amasses a single collection that encompasses 90-100% of the features, even with exceedingly small binning gap parameter

values. This has several negative implications for the analysis. The first is the excessive memory and computational time required to process large bins, with clustering analysis and output file writing serving as the principal bottlenecks. Depending on the computing environment, the issue may render *Binner* or the results files it produces unusable. Large bins can stretch for multiple minutes in RT, harboring analytes that elute much further apart than the plausible distance between ions derived from a common parent metabolite. Consequently, large clusters with wide RT spans form, despite post-processing steps designed to generate manageable sub-clusters. The visualizations of pairwise matrices are considerably less useful given their overwhelming size. Finally, the resulting complexity of the clustering problem in larger bins results in suboptimal groupings that erroneously separate features derived from a common parent or group together highly correlated features from non-identical analytes, leading to many mis-annotations.

In the current implementation, *Binner* processing halts whenever it encounters a bin containing more than 4000 features. With the growing ion detection capabilities of mass spectrometers, this failsafe proves insufficient to accommodate most metabolite profiling experiments. A solution to reduce bins to a manageable size is necessary for the continued utility of this tool. One potential solution is to set a maximum size value, K, such that whenever a bin exceeds this maximum size, the program excises the K consecutive features within the bin that harbors the smallest RT span. This process would then be repeated with remaining bins until every bin has less than or equal to K elements. The principal drawback of a bin size reduction approach is the potential to arbitrarily split related ion species into separate bins. Nonetheless, it is useful to explore how this proposed approach could alleviate the problems associated with excessive bin sizes and its effect on annotation accuracy.

2.6.2 Isotopologue Annotation

Isotopologue annotation is pursued separately from other ion species due to the ease with which most of these can be detected and to reduce the complexity of the clustering problem. However, while Binner accounts for the 1.003/z Da mass spacings of ^{13}C isotopologues, this workflow fails to annotate ^{34}S and ^{37}Cl isotopes. One reason is that the roughly 1.997 Da spacings fall outside the specified m/z window ([M+2] for ^{13}C have an expected mass spacing of 2.006/z) and, in many cases, the lack of an expected [M+1] in between the [M] & [M+2] isotopologue. In some cases, the assumption that higher mass isotopologue are less abundant may be violated, such as in compounds containing multiple chlorides (e.g. salt complexes). This partly accounts for the shortcomings of isotopologue detection performance compared to CAMERA in the evaluation. While it is possible to include a neutral mass difference of 1.997 in the annotation file, this will only work for suspected charge carriers (e.g. [M+Cl+isotope(+2)]-) and not with charge carrier/ neutral group combinations. It is therefore necessary to include an additional search for the mass difference of 1.997/z after accounting for ^{13}C isotopes in the isotopologue detection step, along with binomial distribution checks to facilitate annotation in cases for which the assumption of decreasing abundance is violated.

2.6.3 Need for Annotation Restrictions

Binner generates a list of potential annotations by combining one charge carrier with one neutral gain or loss mass, as specified in the annotation file. Dimer and trimer versions of each annotation is also included in this list. If, for example, the annotation file contains two charge carriers H & Na, and one neutral loss, H2O, the list of annotations consists of {[M+H], [2M+H], [3M+H], [M+Na], [2M+Na], [3M+Na], [M+H-H2O], [2M+H-H2O], [3M+H-H2O], [M+Na-H2O], [2M+Na-H2O], and [3M+Na-H2O]}. With more chemical groups in the annotation file, the search space becomes exceedingly large with a high chance of making spurious annotations

based on incidental mass matches. Complex multimer and neutral gain/ loss combinations are not expected to be as highly prevalent, but *Binner* arbitrarily searches for these with no possibility of restricting this behavior beyond removing chemical groups from the annotation file entirely, losing many desirable combinations. Under the default settings, all annotations - including less common and complex types- are weighted equally in the determination of principal ion adduct types, potentially resulting in mis-annotations. While the current annotation file format has the advantage of forming combinations without explicitly listing all possible adduct and fragment types, the annotation process could be greatly improved with greater customization control. The new alternative file format would lay each combination on separate lines with columns indicating the total mass, charge, and maximum multiplicity allowed. Thus, combinations that are likely to be spurious can be removed without the complete removal of chemical groups. Additional restrictions could be applied to condition certain annotations on the presence of a prerequisite ion type (e.g. [3M+Na+NaCOOH]+ requires [3M+Na]+) or to certain RT ranges.

2.6.4 Charge States

Given the relatively unrestricted nature of adduct & fragment annotation, it is appropriate that program behavior is restricted to annotations with charge of $z = \pm 1$, unless isotopic evidence demonstrates the existence of higher order charge states. However, many datasets have isotopologues removed prior to *Binner* analysis, removing the possibility of correctly annotating features with charge states greater than 1. In some cases, the charge state is pre-calculated and is available as a column in the input file, though this information cannot be accessed or used by *Binner* in its current form. An optional input column field for charge could overcome this issue, with associated annotations conforming to these charge states. In addition, the highest charge state currently searched by *Binner* is hard coded to 3, though higher charge states (+4, +5, ...)

may be observed. A separate maximum charge parameter could facilitate the detection of these additional isotopologues.

2.6.5 Semi-Supervised Annotation

Like most programs, the annotation process within *Binner* is largely unsupervised, with the only prior information introduced into this process coming from the file of likely chemical additions or losses. While this generalizes the method's applications and minimizes bias, user knowledge of known compounds and their respective ion types represents an opportunity for further refinement, as was shown previously.[162] Introducing limited supervision to the algorithm may hold many benefits to improve the annotation method. In many instances, incorrect annotations of known identified features can be averted through prioritizing pre-labeled annotations and using these to inform co-clustered features. Prior knowledge may be used to inform which annotations may be more likely or unlikely for certain metabolites or metabolite classes. These strategies may prove beneficial in a variety of cases, such as metabolites which tend to form single ions that cannot be accurately labeled through mass relationships or the labeling of lower-order fragment ions whose masses match up to peaks in low collision energy MS/MS spectra. The latter is exemplified by the two unlabeled Kynurenine fragment features in the second cluster previously shown in Figure 2.5.

2.7 Conclusion

Among the thousands of consistently detected signals in high throughput LC-MS metabolomics assays, a significant proportion correspond to redundant ions, artifacts, or chemical entities of non-sample origin. Their presence in data has many undesirable consequences, particularly if they are not accurately recognized. *Binner* can facilitate the determination of experiment-specific chemical additions and losses, whose formation may be

dependent on experimental factors, such as instruments, sample matrices, solvents, and mobile phase additives. Its automated workflow performs on par with or superior to existing methods, as evaluated in a complex human plasma metabolomics dataset. As acquisition of MS/MS data is widely considered an essential step in unknown compound identification for untargeted studies, using *Binner* to reduce data complexity can lead to a more thorough understanding of unmatched MS/MS spectra typically produced in comprehensive MS/MS analyses.

Chapter 3

metabCombiner: Alignment of Disparately Acquired Metabolomics Datasets

3.1 Introduction

Alignment of mass spectral peaks, characterized as commonly detected m/z and RT features, across a set of experimental samples is a crucial step in LC-MS data processing pipelines. The goal is to maximize the discovery of shared constituent analytes and assemble their respective abundance measurements into a unified table of spectral features to be used for further downstream analyses. Accurate and comprehensive alignment approaches must properly account for shifts in measured m/z and especially RTs due to changes in chromatographic conditions and sample matrices between runs. Depending on the application, analytical variations may be slight (within experimental batches), moderate (between batches), or significant (disparate experimental conditions). Each case is carefully examined here.

3.1.1 Conventional LC-MS Alignment

Most LC-MS metabolomics alignments are performed between samples acquired under replicated conditions, typically within one experimental batch. In single batch analyses, precise m/z and RT values may exhibit slight variation slightly from sample to sample. Typical methods for conventional LC-MS alignment consist of piecewise or spline-based warping functions modeling retention time shifts as a systematic function of chromatographic region; following correction, these methods perform a direct matching of two-dimensional peaks between raw chromatograms to obtain the resulting feature list. Numerous variations on this formula have

been implemented, including methods that use a reference sample as a basis for constructing or warp functions (50, 167),[60,173] while others use anchor points that are either randomly determined,[46] present in all samples,[45,174] or chosen by MS/MS matching.[175,176] Alignment steps in popular open-source metabolomics preprocessing software, such as XCMS, MZMine2, and OpenMS[177] typically use extracted ion chromatograms (EIC) constructed in chromatogram deconvolution steps as the alignment dimension, whereas many methods and standalone tools operate on the total ion chromatogram (TIC), reduced isotopic envelopes (RIE), or related peak (e.g. adduct) features.[59]

3.1.2 Inter-Batch LC-MS Alignment

Large scale metabolomics studies contain more samples than can be run in a single batch, requiring that they be divided into multiple analytical batches, which may be acquired over long time periods. Systematic and random variation in m/z and RT measurements are frequently observed between samples from different batches, posing significant challenges for data processing pipelines. A growing list of studies have highlighted shortcomings of existing conventional LC-MS alignment approaches in performing inter-batch alignment tasks. This has motivated the development of new tools for correcting mis-alignments due to errors in existing pre-processing methods or implementing novel alignment methods accounting for the batch membership of each sample. One example is the BatchCorr R package, which applies a misalignment correction to data pre-processed with XCMS.[178] BatchCorr requires features to be present in at least 80% of quality control samples for multiple batches, and subsequently features within user-defined m/z and RT windows between clusters are matched using a recursive sub-clustering algorithm. Another notable package is neighbor-wise compound-specific Graphical Time Warping (ncGTW), an XCMS plugin that devises RT warping functions in m/z slices as a function of sample run order, with the premise that adjacent samples are more alike than distant

samples.[179] Finally, the apLCMS 2.0 pre-processing package handles multiple sample batch inputs through a two-step process in which retention time drifts are first corrected within-batches, then between-batches with nonlinear kernel smoothing functions.[180] Subsequently, weak signals missed in peak picking stages are recovered based on the retention time corrections. With the growing sizes of experimental studies, the development of these methods is increasingly necessary to overcome chromatographic drift between distally acquired samples.

3.1.3 Disparate LC-MS Alignment

By contrast, fewer methods have previously been designed for aligning disparate LC-MS metabolite profiling data, defined as studies in which samples are analyzed under different instruments or experimental protocols, often by different laboratories, or with significant intervals between data acquisitions. Under these conditions, RTs for identical analytes may differ by multiple minutes between samples, far beyond the thresholds of conventional pre-processing tools and multi-batch correction upgrades. While some approaches to retention projection have been described previously (see Section 1.4.3), few have been incorporated into alignment algorithms due to the difficulty of matching between unknown features. One method was implemented in DIMEDR (Disparate Metabolomics Data Reassembler), which links multiple experimental datasets to create a unified matrix, employing a universal retention correction based on commonly detected "endogenous anchors".[181] Another method is PAIR-UP MS, which effectively bypasses RT comparisons by leveraging the correlation structure of biologically similar specimens to compare and match analytes of similar m/z in input datasets.[182] This method requires sufficient numbers of shared known identified metabolites and large sample sizes to work optimally. Finally, Mitra el al. (2014) reported a method for uncovering elution order distortions in LC-MS proteomics data generated by different laboratories using LOESS to model RTs of peaks with similar m/z and ranked abundances, followed by peak matching of peptides.[183]

3.1.4 Motivation

This work is primarily focused on computational methods for the third, most intricate case of alignment between non-identical LC-MS metabolomics assays. A versatile method that requires no prior knowledge of known compounds, no sample size minimums, pre-processing software independence and has the capability to bridge substantial RT shifts between samples collected from different assays would provide a powerful new approach for disparate LC-MS data analysis. This would provide three major benefits to computational metabolomics research. First, disparate LC-MS alignment facilitates the mapping of information between non-identically acquired metabolomics data, allowing for validation of annotated metabolites, determining intersecting entities lacking structural characterization, and providing putative identification hypotheses or other chemical characteristics (such as adduct or formula labels) whenever this is determined for at least one of the input datasets. Second, it enables reproducibility calculations, such as intra-class correlation coefficients, for both known and unknown metabolites, a common pursuit in multi-laboratory and multi-instrument studies. Third, aligning feature data generates merged sets of intersecting experimental measurements, with the expanded sample sizes providing increased statistical power to detect significant results that may otherwise be missed from the constituent datasets alone. Thus, alignment is the crucial first step to analysis of disparately acquired metabolomics datasets and its benefits extend to metabolomics quantitation, annotation, and interpretation, the three major goals of computational metabolomics.

3.2 Methods

3.2.1 *metabCombiner* Overview

metabCombiner is a software package written in the R statistical programming language. The methods and applications of this tool have been described previously.[184] In its simplest form,

metabCombiner aligns a pair of peak-picked and conventionally aligned feature tables generated by common LC-MS pre-processing programs, where each row represents a unique analyte. The datasets must be acquired in the same ionization mode, with no prior scaling or normalization that may distort their ranked abundance order. The datasets used as inputs must be acquired from biologically similar specimens with a strong expected overlap in their metabolic composition. Finally, chromatographic protocols used to acquire datasets must be similar enough that the elution order of compounds is comparable, if not identical. Post-processing steps for removal of features of non-sample origin (background ions, noise, processing artifacts) and isotopologues are desirable, but not required. Users may include feature identifiers, as well as adduct, fragment or formula labels for validation and parameter optimization purposes. Other dataset columns may be included as "extra" non-analyzed columns in the output table.

3.2.2 *metabCombiner* Workflow

The basic workflow for *metabCombiner* is depicted in **Figure 3.1**. For a pair of input datasets, one is designated as the projection (X) dataset, whose RTs will be mapped to the chromatogram of its complement, denoted as the reference (Y) dataset. In practice, the dataset with the shorter retention time range is usually designated as the reference since smaller absolute prediction errors are observed when mapping from a chromatogram with highly resolved peaks to one a less-resolved one than vice versa. The *metabCombiner* method is constructed around key observations for shared compounds in datasets that conform to the previously listed assumptions. The first is that m/z deviations for identical compounds are generally small, rarely

exceeding 5-10 mDa even if measured by different high-resolution mass analyzers, as long as proper instrument calibration is maintained throughout the analyses. The second is that while raw spectral abundances are generally not comparable between experiments, relative abundance (Q)-calculated as ranked median or mean intensity quantile values between 0 & 1 – serve as an additional dimension for comparison besides m/z & rt. Finally, a number of highly abundant common endogenous metabolites (e.g. creatinine and L-Carnitine in urine, saturated lipids and amino acids in plasma, etc...) are expected to be present in both datasets, due to the biological sample similarity requirement. These abundant compounds can be used to anchor a non-linear mapping between RTs. Each step of the workflow is described in detail in the following sections.

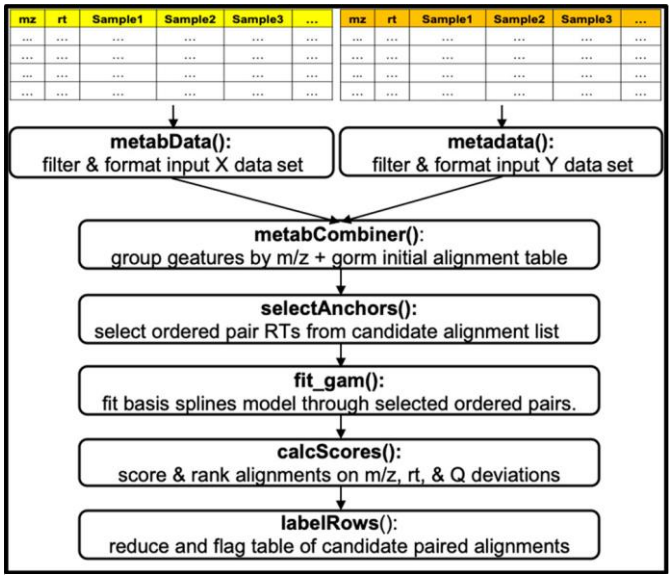


Figure 3.1 Basic *metabCombiner* Workflow

3.2.3 Package Objects and Terminology

There are two major object classes formulated in this package: *metabData* and *metabCombiner*. *metabData* objects are single dataset representation classes, containing a specially formatted representation of the data, as well as the names of analyzed samples and 'extra' columns, and feature statistics. *metabCombiner* objects are multi-dataset representation classes that serve as the main framework for executing the package workflow steps. *metabCombiner* classes contain a *combinedTable*, a data frame containing the set of possible feature pair alignments (FPAs), their associated per-sample abundances, and package-generated alignment information; *featdata*, a data frame closely linked to the *combinedTable* containing all

feature meta-data with special utility for alignment tasks involving more than two datasets; a *datasets* field storing identifiers of constituent datasets that are used to organize the meta-data, sample & extra column names; *samples* and *extra* lists merging the respective lists from input datasets (each with an associated unique dataset identifier); *anchors* and *model* fields for storing the results of retention time projection steps; *xy*, storing the character identifiers of the X and Y datasets, respectively; and *stats*, storing various object statistics.

3.2.4 Data Processing

Input feature tables are initially processed and formatted as separate *metabData* objects. First, the package checks for all required (m/z, RT, sample abundances) and optional (identifiers, adduct labels, and 'extra') feature meta-data. Subsequently, multiple filters are applied to reduce input feature lists. A RT range filter can limit the analysis to features between a start and end RT, eliminating the head and/or tail of the chromatogram which are often sparse and contain mostly background ions. Eliminating these chromatographic regions improves the quality of the RT fitting steps, at the expense of some features that cannot be accurately matched. The second filter eliminates features that are missing in more than a certain percentage (default: 50%) of analyzed samples, since relative abundance is difficult to estimate for overly missing features. Finally, pairs of features within specified m/z & RT tolerance values are discarded as duplicates, with one copy retained. Ranked relative abundance quantile (Q) values are then calculated from the mean or median abundances of the remaining features.

3.2.5 Feature Grouping by m/z

All features from a pair of input objects are pooled, sorted and binned in the m/z dimension. Distinct feature groups form whenever the difference between consecutive m/z values is less than a user-specified *binGap* argument (by default 5 mDa). Each group initially

contains *m* features from dataset X & *n* features from dataset Y ($m > 0$ & $n > 0$), with $m * n$ total FPAs. Direct feature matching occurs exclusively within these groups and no features from distinct groups may be aligned. The *metabCombiner* object and its associated *combinedTable* is constructed in this step, assembling all FPAs arranged by m/z groups in increasing order and datasets given identifiers to facilitate organization of information. Subsequent steps assess which FPAs correspond to matching metabolite entities.

3.2.6 Feature Pair Anchor Selection

A set of ordered pair features is required to construct map between dataset RTs. Confidently identified compounds shared between datasets would be useful for this purpose, however sufficient chromatographic coverage of known metabolites cannot be assumed in general. Therefore, the ordered pairs are selected among all possible FPAs using the process illustrated in **Figure 3.2A**. First, the most abundant feature (with the largest Q value) from the X dataset is selected and denoted as x_1. The most abundant Y dataset feature in the same m/z group containing x_1 is selected as the corresponding y-ordinate, y_1. Together, the RTs of x_1 & y_1 serve as the first anchor. All features within a small RT window (0.03 minutes by default) of x_1 and y_1 in their respective datasets are excluded from consideration as potential anchors. This selection approach is iterated for the remaining features, until all feature pairs have been either included or excluded as anchors, providing an initial list of ordered pairs (x_1, y_1), (x_2, y_2), ..., (x_n,y_n) which we denote as Set A). This process is repeated, this time choosing the most abundant features of dataset Y and their counterparts in dataset X, deriving a second anchor list, Set B. The final anchor set is obtained from the intersection of sets A & B and is expected to provide a rough outline of the nonlinear smooth curve between two sets of chromatographic retention times through which a spline may be fit. This largely unsupervised process may select a few

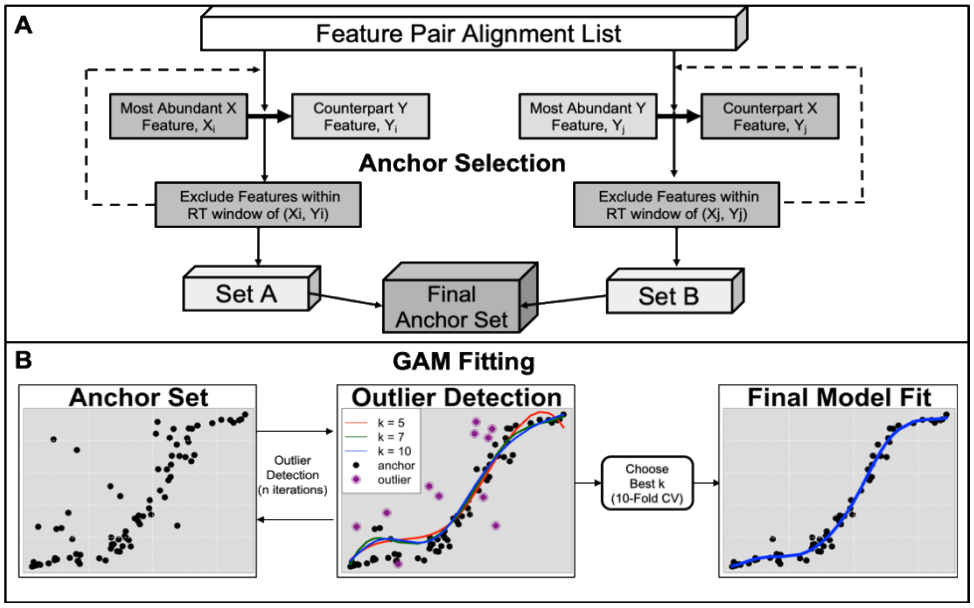


Figure 3.2 *metabCombiner* RT mapping procedure In (A), RT ordered pairs are selected from shared abundant (or identified) features, generating two lists that are subsequently intersected to obtain a final set

mismatching compounds that manifest as outliers from the outlined curve. Constraints can be placed on m/z, Q, and RT quantile differences of anchoring X & Y features to increase the robustness of anchor selection. Users may also incorporate features with shared identities as anchors, which the program selects first as ordered pairs before performing the unsupervised process. Incorporating prior knowledge improves and refines the RT mapping steps.

3.2.7 Spline-Based RT Mapping

Basis splines, implemented in the *mgcv* R package, is the main method for RT mapping in this workflow.[185,186] Basis splines are a type of generalized additive model (GAM), where a smooth curve is computed based on the sum of low-order polynomial basis functions joined at *k* control points. *k* determines the flexibility of the smooth curve and must be optimized from the underlying data. The RT mapping process is illustrated in **Figure 3.2B**. First, multiple GAMs with different values of *k* (e.g. 5, 7, 10, ...). are fit to the ordered pairs computed in the anchor selection step, modeling Y-ordinates as a function of the X-ordinates, i.e.

$$rt_Y \sim f(rt_X) + \epsilon$$

In each individual model, the function calculates the absolute value of the fitted vs observed residual for each ordered pair (rt_x, rt_y). Anchors with consistently high residuals, defined by default as twice the mean absolute model error in over half of the model fits (or alternatively twice the interquartile range plus the third quartile) are excluded as outliers. This is repeated for a specified number of iterations, removing anchors that deviate significantly from the outlined curve. Anchors selected by matching identities are a key exception as they are never filtered, even if high errors are observed. With the remaining points, the optimal *k* value is selected from among the provided options using ten-fold cross validation, minimizing mean absolute deviation. The final model is computed with this *k* value, which then maps RTs between datasets.

3.2.8 Feature Pair Similarity Scoring

Each feature may have one or multiple candidate matches in its counterpart dataset. To determine the most plausible FPAs, we assign to all pairs of grouped features F_x & F_y a similarity score between 0 and 1 according to the expression

$$\boldsymbol{S(F_x, F_y) = \exp(-A|mz_y - mz_x| - B\,\frac{|rt_y - f(rt_x)|}{range(rt_y)} - C\,|Q_y - Q_x|)}$$

where mz_x, rt_x, & Q_x are the respective m/z, RT, and Q values of feature F_x; mz_y, rt_y, & Q_y are the respective m/z, RT, and Q values of feature F_y ; *f* denotes the computed RT mapping function, with prediction errors normalized by the range of Y dataset RT values. *A, B, C* are positive weight parameters penalizing differences in feature m/z, rt, and Q, respectively. The choice of weight values should account for instrument mass accuracy, model fit, chromatographic range, and sample similarity. The most effective range of values used in testing for *A, B, C* are 50 - 120, 5 - 20, & 0 - 1, respectively. If the datasets contain a sufficiently representative set of shared identified compounds, the package method *evaluateParams*() finds the set of A, B, C weight values that optimize an objective function maximizing the positive difference between scores of

FPAs of true known matches from their respective misaligned pairings. A similarity score of 1 implies perfect concordance of m/z, RT prediction, and relative abundance of a pair of complementary features from the input datasets, implying a likely aligned compound match, whereas scores closer to 0 may be disregarded as misaligned compound pairs. Each feature's potential matches from the complementary dataset are ranked in reverse score order, with best matches (rankX = 1 & rankY = 1) displayed first.

3.2.9 Combined Table Reduction

An automated method reduces the *combinedTable* report by eliminating rows corresponding to misaligned pairs, as indicated by the similarity scores and ranked similarities computed from the previous step. Ideally, every feature should be aligned with at most one feature from the counterpart dataset, though in some cases multiple matches may need to be considered. The *labelRows()* package method facilitates this process by first placing thresholds on score, pairwise ranking, and retention time error, then flagging lower-ranked FPAs for review if the delta scores relative to the top-scoring FPA are below a user-defined threshold. Features failing to meet any of these thresholds are designated by the program as removable rows. *labelRows()* is useful for facilitating user inspection of all alignment results between a pair of datasets and visualizing all possibilities alongside the package row designations. The alternative *reduceTable()* method takes this approach further by automatically reducing the list of FPAs to contain only one-to-one compound matches. Conflicting FPAs are resolved by selection of rows lacking shared features (and subject to retention time order constraints) with the maximum sum of scores. This method is used more often in routine applications of *metabCombiner*.

3.2.10 Recovering Non-matched Features

By default, *metabCombiner* generates the intersection of features jointly detected between non-identically acquired metabolomics datasets. The new feature count of this intersected table is smaller than the inputs supplied to the program. There are three steps in the workflow in which features are removed and not included in the final *combinedTable*: 1) the filtering steps during the pre-processing of input feature lists in *metabData* construction; 2) during m/z grouping, where many features can be excluded due to the lack of signals with close m/z values in the complementary dataset, and 3) In the table reduction stage, where automated and/or manual removal of FPAs may leave features with no compound matches. In many applications, it is desirable to account for the union of input features, including those that are not part of the intersection. The *updateTables()* package method takes the *metabCombiner* object result of the previous steps and the original input X dataset and Y dataset objects used for its initial instruction as input and searches for features that are not present in the *combinedTable*. Features present in the X dataset but not in the program results are appended to the table as separate rows containing their respective m/z, RT, Q, id, adduct, sample abundance and extra column values, while their respective Y columns are filled in with missing values. A similar process is performed for missing Y dataset features, with missing values in X columns. This process accounts for the second and third criteria listed above since these features are part of package objects, whereas features filtered in the early analysis stages are eliminated altogether.

3.3 Methods Extensions: Multiple Disparate LC-MS Dataset Alignment

The package methodology presented in Section 3.2 defines the process of aligning metabolomics data in pairs of conventionally pre-processed LC-MS tables. Large-scale metabolomics studies or multi-laboratory investigations may require aligning more than two datasets. The *metabCombiner* approach can be extended to more than two datasets in a stepwise

manner, augmenting the results of an alignment task with additional single or multiple dataset objects. A straightforward extension of the multi-dataset alignment analysis approach called *batchCombine* is implemented for aligning batches of the same experiment.

3.3.1 Stepwise Disparate Multi-Dataset Alignment Workflow

As in the basic paired alignment workflow, multiple dataset alignment requires that all input tables be filtered and formatted as single dataset *metabData* objects. The main difference is that either the X, Y, or both datasets are themselves *metabCombiner* objects. Hence, the results of an alignment task can be supplied as input for further alignment with a new single dataset (*metabData*) or combined dataset (*metabCombiner*). **Figure 3.3** illustrates an example of this approach, starting with two datasets and merging the results with additional single datasets after performing the six *metabCombiner* alignment steps.

At the beginning of each metabCombiner alignment task, a *featdata* table is constructed and updated in each alignment task to contain all observed feature descriptors from the

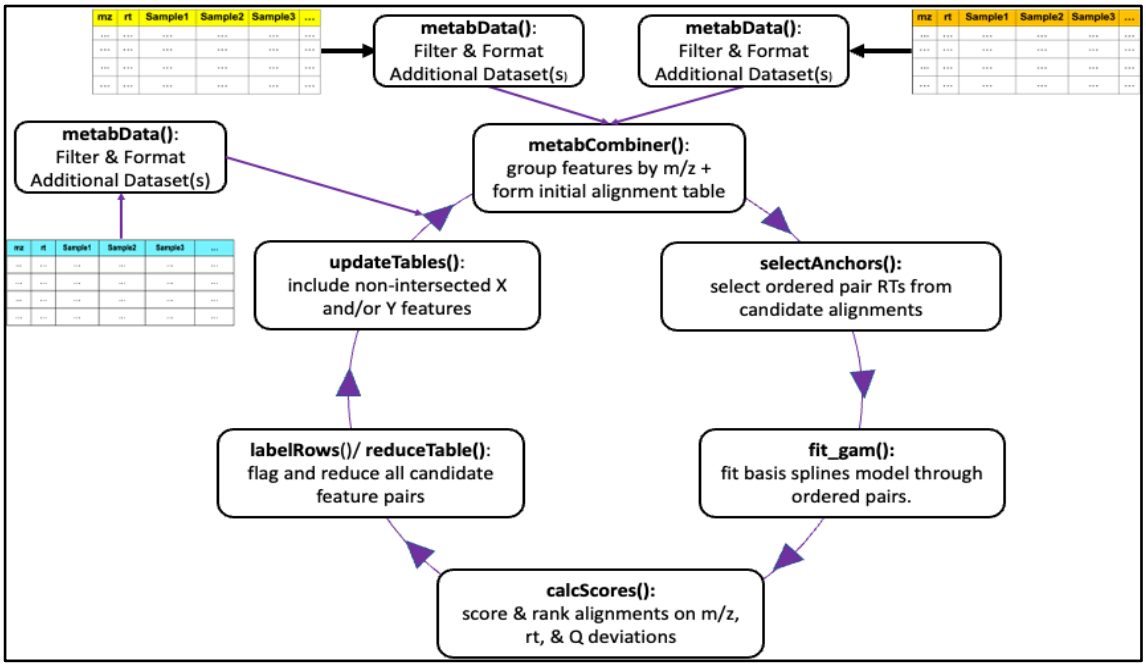


Figure 3.3 Stepwise Multi-Dataset Alignment Workflow *metabCombiner* steps with associated function names are the same as the basic workflow (Figure 3.1), but with the addition of *updateTables* for including non-intersected features and additional arrows implying the workflow's cyclic nature in which the result of one alignment analysis is supplied as input along with another single or multiple dataset object

constituent datasets that have been aligned together. The program draws the essential feature meta-data {m/z, RT, Q, id, adduct} from one selected dataset of a *metabCombiner* input (designated by its unique identifier) stored in the *featdata* table, and uses that to form the X or Y meta-data columns in the *combinedTable*. For quantitative descriptors, the mean value across all datasets can also be used to represent features as an alternative to selection from a single dataset, although averaging of RTs is not recommended for datasets acquired with major chromatographic differences. Once the new set of possible alignments are generated, the workflow steps proceed in the same manner as previously described.

A complication arises when accounting for features that are absent in the constituent dataset selected to represent the *metabCombiner* object. These features will contain missing quantitative descriptors {m/z, RT, and Q} essential to their alignment with a new feature list. One possible workaround is to have this information imputed with guidance from other constituent datasets for which the feature is present. Currently, only mean {m/z, RT, Q} imputation is implemented, which is not recommended for disparate LC-MS alignment applications. Future iterations of *metabCombiner* will explore appropriate imputation methods for RTs in non-identically acquired datasets.

3.3.2 *batchCombine*: Application to Multi-Batch Alignment

While originally designed for aligning non-identically acquired LC-MS metabolomics data, the process of feature-matching and data concatenation has useful applications in multi-batch LC-MS metabolomics experiments. As previously noted, retention behavior of compounds may shift between samples in large-scale experiments, despite efforts to closely replicate experimental variables from batch to batch. [178–180] These differences are especially pronounced between distally analyzed samples (in run order) compared to proximal samples. *batchCombine*

applies the *metabCombiner* process between batches of the same experiment, where each batch of samples is separately pre-processed using conventional tools such as XCMS or MZMine2.

The pre-processed batches should be arranged in acquisition order {1, 2, ..., n}, where n is the total number of batches, and formatted as *metabData* classes. The *batchCombine* workflow proceeds in a nearly identical manner to that previously shown in **Figure 3.3**. The first two batches are aligned in one *metabCombiner* cycle, and the results are aligned with the third batch, followed by the fourth, and so forth with a total of n-1 cycles. In each cycle, feature meta-data are drawn from the latter batch of the previous alignment step to be used for comparison with the next batch (i.e. batch 2 feature meta-data is aligned to batch 3, batch 3 is aligned to batch 4, and so forth) since retention behavior is more similar between proximal batches. m/z, RT, Q or a combination of these descriptors may be averaged prior to each alignment step, an option that could control for random shifts. Identical parameters for the six main *metabCombiner* functions are applied for all tasks, facilitating automation of batch alignment processes. It is expected that the overall chromatographic dimension (total run time and observed RTs) should be roughly similar between samples within a multi-batch experiment; therefore, imputing the average RT for features not detected within a specific batch is acceptable, unlike in disparate LC-MS alignment. *batchCombine* can determine either the union or the intersection of all features found across *n* input batch tables.

3.4 Evaluation

3.4.1 Evaluation with Plasma Metabolomics Datasets

Untargeted RPLC metabolomics data were acquired twice in the positive ionization mode for ten human plasma samples, five from a pooled plasma obtained from deidentified Red Cross (RC) donors and five from pooled plasma purchased from a commercial supplier for the NIH

Children's Health Exposure Analysis Resource (CHEAR) consortium. Samples and process blanks were analyzed using the same instrumentation and column, but with two different water-methanol gradient elution methods, with total run times of 30 and 20 minutes. Both datasets were processed with XCMS (39) and reduced by isotopologue and blank sample feature ratio filtering. A total of 137 common metabolites were identified according to Metabolomics Standards Initiative (MSI) criteria 1 or 2,[77] with 532 in-source adducts, fragments, and multimers of these metabolites annotated using a custom R script and *Binner*.[171]

For this pair of datasets, RT fitting and score-based matching of compounds was evaluated, using identified compounds as a benchmark. Known metabolites were partitioned into 50% training, 50% test sets. RT fitting was both semi-supervised (with all training set compounds included as anchors) or unsupervised (rt fitting without prior knowledge), with mean absolute deviation (MAD) of the fit calculated for the test set compounds. The *evaluateParams* package method was used to guide A, B, C weight value selection on the training set compounds. An "accurate match" is defined to be a best-scoring FPA (rankX = 1 & rankY = 1) between two identically annotated features with a score greater than 0.5. Feature matching accuracy is assessed "per-variant"- that is, weighing each adduct/fragment feature equally- and "per-compound" – summing the fractions of accurately matched adducts and fragments for each compound over the total number of test compounds. Different sample subsets (CHEAR or RC) analyzed in the 30-minute analysis are designated as dataset X, with the opposite set in the 20-minute analysis designated as dataset Y.

3.4.2 Exploration with Muscle Metabolomics Datasets

Muscle tissue from 10 sedentary and 10 exercised rats were analyzed in the Michigan Regional Comprehensive Metabolomics Resource Core ('*MiSE10*') as well as by the Broad

Institute ('*BrSE10*') in the negative ionization mode. Data were processed using XCMS and Progenesis QI (*nonlinear dynamics*), detecting 5335 & 8573 features respectively. There were greater differences in the experimental methods used in this case, including column type (Waters H3 TSS modified C18 vs unmodified C18), mobile phase solvents (methanol vs acetonitrile), mass analyzer (QTOF vs Orbitrap), and m/z scan range (50-1000 vs 70-850). Of the named identified compounds (200 in *MiSE10* & 80 in *BrSE10*), there were only 14 mostly non-polar overlapping compound identities. These compounds including their annotated adducts, are used to evaluate the efficacy of alignment under highly disparate conditions by *metabCombiner*.

3.4.3 Application of Multi-batch Alignment to ELEMENT Study

To test the capabilities of *batchCombine*, metabolomics data from the ELEMENT (Early Life Exposure to ENvironmental Toxicants) cohort were aligned to form a single cohesive batch-merged table.[187] A total of 402 individual plasma samples from adolescent subjects were partitioned into eight total batches and analyzed by LC-MS within the same laboratory, instrument, and protocol. Nine pooled QC samples were analyzed alongside fifty biological replicates in each batch. The raw spectral files of each batch were separately pre-processed with XCMS using identical peak-picking and alignment parameters, generating tables containing between 11100 and 13800 features for the eight batches.

3.5 Results

3.5.1 Program Output

The main output of *metabCombiner* is the combined table containing FPAs organized into separate m/z groups in order of increasing m/z. The signal abundance values of each feature are concatenated to form a combined table. An example m/z group from the plasma evaluation is shown in **Figure 3.4A**, consisting of 3 features from dataset X (30-minute analysis) and 3 from

A

	idx	idy	mzx	mzy	rtx	rty	rtProj	Qx	Qy	Group	RankX	RankY	Score
I	X7710	Y7385	426.3206	426.3205	28.145	15.316	15.299	0.950	0.937	2221	1	1	0.9706
II	X7753	Y7434	426.3205	426.3205	28.462	15.566	15.543	0.915	0.890	2221	1	1	0.9687
	X7753	Y7385	426.3205	426.3205	28.462	15.316	15.543	0.915	0.937	2221	2	2	0.8172
	X7710	Y7434	426.3206	426.3205	28.145	15.566	15.299	0.950	0.890	2221	2	2	0.7683
III	CHOLATE [M+NH4]+	CHOLATE [M+NH4]+	426.3219	426.3223	21.547	10.023	9.837	0.428	0.620	2221	1	1	0.7494
	X7710	CHOLATE [M+NH4]+	426.3206	426.3223	28.145	10.023	15.299	0.950	0.62	2221	3	2	0.0085
	X7753	CHOLATE [M+NH4]+	426.3205	426.3223	28.462	10.023	15.54	0.915	0.62	2221	3	3	0.007
	CHOLATE [M+NH4]+	Y7385	426.3219	426.3205	21.547	15.316	9.837	0.428	0.937	2221	2	3	0.0068
	CHOLATE [M+NH4]+	Y7434	426.3219	426.3205	21.547	15.566	9.837	0.428	0.890	2221	3	3	0.0056

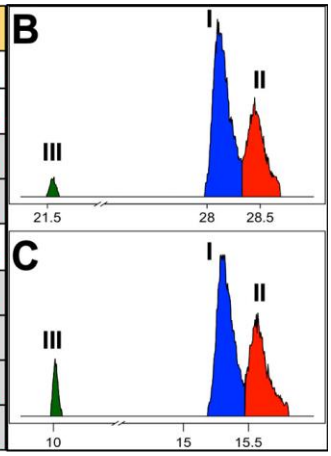


Figure 3.4 Example *metabCombiner* m/z Group (A) X and Y features associated meta-data, and alignment scores. Peaks of the same color are matched analytes. (B) X dataset EIC. (C) Y dataset EIC.

dataset Y (20-minute analysis), all within the m/z range 426.3205 - 426.3223. Two features each from the two complimentary datasets are unidentified isomers (X7710 & X7753, Y7385 & Y7434) while the third previously identified in both datasets as Cholate [M+NH4]+. A pairwise top match (rankX = 1 & rankY = 1) is assigned between [X7710, Y7385] and [X7785, Y7434], respectively, with alignment scores very close to 1. Alternative possible alignments for this pair of compounds are displayed as separate rows which can be quickly dismissed as misalignments. The alignment score of cholate [M+NH4]+ with itself is lower due to a higher retention time prediction error and a slight difference in the relative abundance of this compound between assays; nevertheless, it is correctly assigned the top-scoring FPA with its counterpart, and all other pairs score very poorly. Thus, three FPAs corresponding to three separate compounds remain in the final table, and six rows defined as misalignments are eliminated. A visual inspection of the peaks shown in **Figure 3.4B** & **Figure 3.4C** confirms this matching.

3.5.2 m/z Grouping

The size of the initial combined table is a function of input dataset feature counts, their degree of m/z overlap, and the *binGap* parameter in the m/z grouping step. In the plasma datasets, the number of FPAs is comparable to the initial dataset sizes and grows steadily with

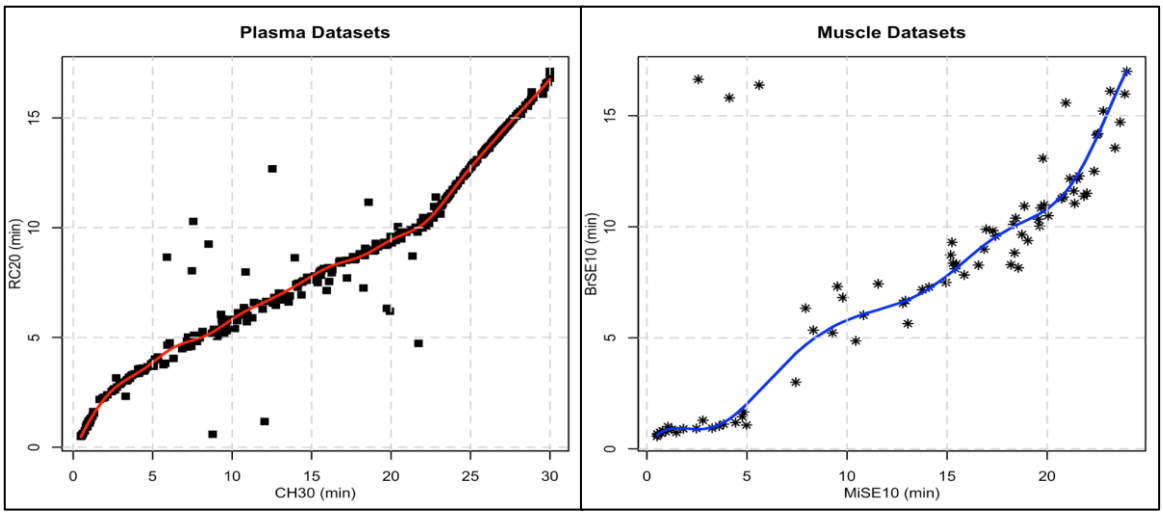


Figure 3.5 Example *metabCombiner* Model Fits Model points are selected from highly abundant feature pairs and (optionally) feature pairs with matching compound identities

increased gap values. On the other hand, a smaller FPA list between the muscle datasets, likely due to differences in m/z ranges surveyed by the respective analyses. In other surveyed datasets, the row count rapidly increases where there are high numbers of compounds in specific mass ranges (e.g. 100-200Da). While most matching known compounds display small (< 1mDa) m/z differences, larger errors (>5 mDa) may be observed in some cases due to instrumental and pre-processing software factors. The value of *binGap* reflects a tradeoff between compactness of the combined table and the ability to detect all true compound matches. *binGap* value is set to 5 mDa by default and can be altered as necessary.

3.5.3 Retention Time Mapping

The *plot* package method is useful for visualizing results of anchor selection and GAM-fitting. Plots for plasma and muscle datasets are displayed in **Figure 3.5**. In both cases, a moderate to high degree of fluctuation in the center of the chromatogram along the gradient slope is observed, indicating that these regions are generally more difficult to model accurately with low prediction errors. Moreover, there are differences in how well-represented each chromatographic region is in terms of ordered pair anchor selection. In the plasma case, all

regions (from polar to non-polar) are well-represented in anchor coverage whereas the muscle case contains noticeable gaps along the gradient. The plot serves as a useful tool for tuning parameters associated with model fitting and determining an appropriate RT penalty weight.

3.5.4 *metabCombiner* Evaluation with Plasma Datasets

CHEAR and Red Cross plasma aliquots were analyzed together in the same laboratory using two different RPLC protocols with 20 and 30-minute total chromatography run times. Of the 137 identified compounds common to both plasma datasets, all but three could be grouped by m/z using the default 5mDa *binGap* value. The principal ions of caffeine, glutamyl-phenylalanine, and creatine deviated by 0.006, 0.0088, and 0.02 daltons, respectively. When the *binGap* was increased to 0.0075, all metabolite ions could be successfully grouped except for creatine. Therefore 136 compounds were used for this analysis, with 68 each randomly partitioned into training and test sets. The choices of sample subset (CHEAR vs Red Cross) and the option to use known identity information affects the selection of anchors and the subsequent modeling & feature matching accuracy. The results of the evaluation are displayed in **Table 3.1**. The mean absolute error of each model is consistently around 0.06 min, with a slight advantage observed in semi-supervised models in which training set compounds are selected as anchors. In each model, more than 50 out of the 68 compounds could be predicted within 0.1 minutes. Prediction errors vs observed retention times for selected test set compounds are shown in **Figure 3.6**. Polar and very nonpolar metabolite RTs are mostly well-predicted, whereas

Table 3.1 Plasma Datasets Alignment Evaluation Results

Mode	X Dataset	Y Dataset	RT M.A.D	Accuracy (per Variant)	Weighted Accuracy (per Compound)
Semi-Supervised (including identities)	30 MIN CHEAR	20 MIN Red Cross	0.056	259/270	0.91
Unsupervised	30 MIN CHEAR	20 MIN Red Cross	0.066	254/270	0.88
Semi-Supervised (including identities)	30 MIN Red Cross	20 MIN CHEAR	0.054	254/270	0.88
Unsupervised	30 MIN Red Cross	20 MIN CHEAR	0.07	250/270	0.86

compounds of intermediate polarity were less predictable due to alterations in gradient slopes, with the highest retention time errors between 0.25-0.35 min. The inclusion of prior information provides a distinct advantage in predicting metabolite RTs, especially in sparse chromatographic regions. The fitted models were used to evaluate similarity scoring, using 270 annotated variants (adducts, in-source fragments, and multimers) of the test set metabolites as points of comparison. Score parameter arguments were chosen to be to be A = 100, B = 15, C = 0.3, as guided by *evaluateParams* on training set compounds. Most compounds accurately achieve the highest alignment score for all their adduct and fragment variants, with weighted per-compound average scores higher than 0.85 in all analyses. Four compounds scored at or below the threshold 0.5 level, mostly due to penalization of high m/z differences. In cases for which the correct alignment is not the top-ranked match, at least one feature may be more similar in m/z, Q, or fitted RT. The feature(s) may arise from closely eluting structural isomers, or peaks incorrectly divided due to pre-processing errors. No accurate FPA ranked poorer than 3rd best for the respective compounds, and the scores of all but one compound were within 0.2 from the top scoring FPA. Proceeding with table reduction, score, rankX, rankY, and delta score tolerance values were set to 0.5, 3, 2, 0.2 reducing the set of 14024 FPAs by 6765; further inspection reduces an additional 400, leaving roughly 6900 rows.

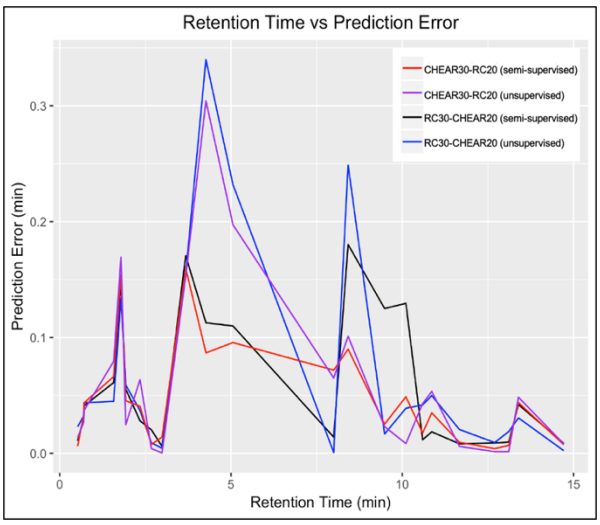


Figure 3.6 Retention Time vs Model Prediction Error

3.5.5 Alignment Analysis of Muscle Metabolomics Datasets

Experimental variables varied more for this pair of datasets compared to the plasma analysis. The protocols used to acquire *BrSE10* are optimized for the measuring metabolites of intermediate polarity (such as bile acids and free fatty acids), whereas *MiSE10* is acquired with a generalized metabolomics assay. This has several implications when aligning this pair of datasets. First, owing to different column types (Waters HSS T3 C18 for *MiSE10*, which has an embedded polar retention functionality, vs Waters BEH unmodified C18 for *BrSE10*), polar compounds were difficult to map and differentiate properly as they eluted very rapidly in *BrSE10*. Second, coverage of highly nonpolar metabolites differed, causing major distortions in initial model fitting attempts. To correct this, the late chromatographic portions where highly nonpolar compounds elute was excised by setting a maximum retention time of 24 & 17 min for *MiSE10* & *BrSE10*, respectively. These constraints remove 10-20% of the input features in each dataset. Third, numerous fatty acids observed to be present at high abundances in *BrSE10* were barely or not at all detectable in *MiSE10* samples, likely due to differences in sample extraction protocols between the two assays; therefore, quantile Q comparisons are less reliable in some cases. Finally, while there were few significant mass errors for shared compounds in preliminary analysis, the overall mean m/z for these two datasets differed by more than 200 Da (549.3 vs 316.1 in MiSE10 & BrSE10, respectively). With the binGap parameter kept to its default value of 5 mDa, this generated a small initial set of only 3247 possible FPAs, indicating a limited coverage overlap between the assays.

Parameters for anchor selection and GAM-fitting were optimized using a grid search of potential values, using the mean absolute RT deviation for fourteen shared identified compounds as the error metric. The final model mapped five compounds accurately to within 0.1 minutes; two additional compounds were predicted within 0.25 minutes; five had errors between 0.4-0.6

minutes; two (cholate & glycocholate) could not be predicted by any of the models to within 1.25 minutes. Given these factors, scoring parameters were chosen as A = 100, B = 7, C = 0.2. Of the shared known compounds, cholate (0.35) & glycocholate (0.39) score the lowest due to the high RT fitting errors. Ten out of fourteen compounds accurately achieved the highest alignment score in all their respective adduct forms; the remaining metabolites have one misaligned variant each and only one variant ranked worse than the 5th best. On this basis, FPAs with scores below 0.35 and ranking worse than 5th were removed, eliminating 1765 alignments; further inspection of conflicting alignments removed an additional 400-450 rows, reducing to under 1000 of the original 3247 FPAs.

3.5.6 Multi-batch Alignment of ELEMENT study with *batchCombine*

Initial processing with *metabCombiner* removes less than 20 features from each batch each due to missingness and duplicate feature filtering. For *batchCombine* parameters, a common m/z binGap value of 0.0075 for initial feature grouping; anchor selection and RT fitting parameters were kept close to the default values. m/z, RT, and Q score weight values were set to 70, 35, and 0.8, respectively, and table reduction parameters were kept at default values. All quantitative descriptors were averaged after each step and the union option enabled, imputing missing feature information with the mean value of features when absent.

In total, 23701 rows appear in the final table, of which a little over one-fourth (6493) are present in all eight batches and those present in only batch account for 5453 rows. **Table 3.2** summarizes the frequency with which features are present in batches across the aligned dataset. Sparsely detected features appear to have unique m/z values or lacking a close RT counterpart in the other pre-processed batches. For most features, RTs were highly consistent from batch to

Table 3.2 ELEMENT Study Batch-Aligned Feature Results

Batch Presence	1 batch	2 batches	3 batches	4 batches	5 batches	6 batches	7 batches	8 batches
# Features	5453	2914	2659	2288	1692	1072	1129	6493

batch, with more than 92% of the features having a standard deviation of batch RT measurements (listed in the constructed *featdata*) of less than 0.05 minutes. Higher retention time drifting occurred later in the chromatogram and for some sparse feature matches. Two notable examples were closely examined and depicted in **Figure 3.7**. **Figure 3.7A** shows three isomers with m/z value = 332.332, all three of which have been detected in all 8 batches between RT = 19.2 and 20.5 minutes. There is a notable shift in the RTs of all three isomers in batch 4, which could lead to misalignments between compounds in traditional correspondence algorithms. **Figure 3.7B** shows the drifting of peaks with m/z = 650.6453 from 29.49 to 28.92 min, a differential of over half a minute from the first to the eighth batch. The gradual drift from between adjacent batches validates the *batchCombine* approach of arranging and concatenating batches in sequential order. The compound characterizing this feature has an $-H_2O$ loss fragment with an identical drift pattern, providing additional confidence in this feature matching. The feature meta-data that is reported alongside the array of spectral abundances is useful for inspecting alignments between consistently measured peaks, peaks with plausible drifting patterns, or spuriously matched signal across batches.

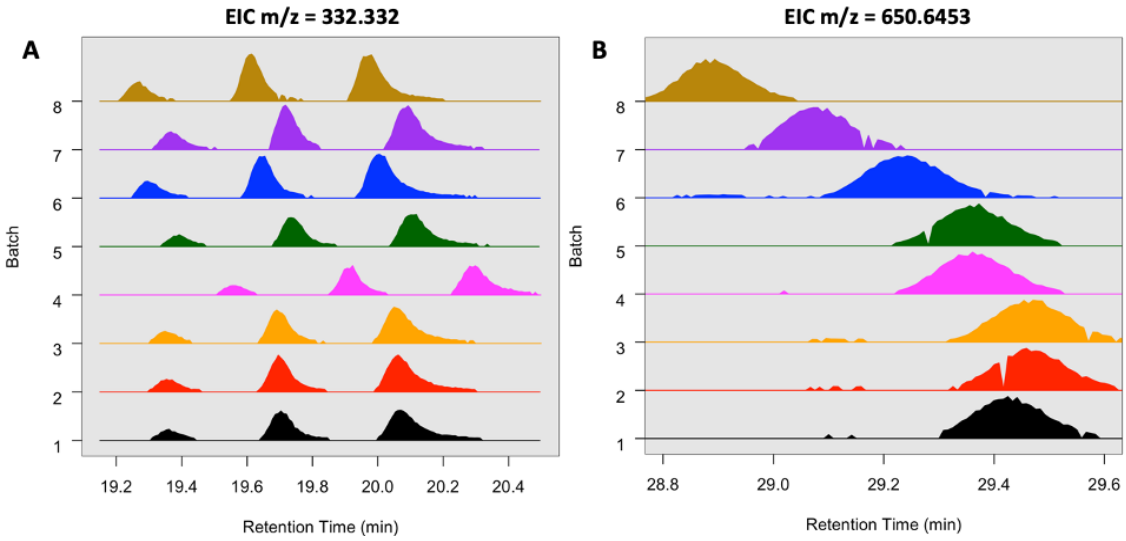


Figure 3.7 Retention Time Drifts Between Batches (A) Batch 4 exhibits a major shift for three isomer peaks (m/z = 332.332); (B) Gradual drifting in RT observed from batch 1 to 8, depicted for this analyte (m/z = 650.645)

3.6 Limitations

metabCombiner is one of the first computational methods specifically designed for disparate LC-MS data alignment. There are some limitations to the methodology that are addressed in this section.

3.6.1 Use of Pre-Processed Feature Tables

Unlike most LC-MS feature alignment tools that operate on raw spectral files for each individual sample, *metabCombiner* uses traditionally peak-picked and aligned metabolomics data as input. This simplifies the process by generating tables of similarly acquired experimental samples, where each aligned feature is represented by a single m/z and RT value, as opposed to separate m/z and RT values for every peak across all similar and disparately acquired samples. Furthermore, this allows for identically acquired spectral data to be peak-picked and aligned using any method, as opposed to being designed for a single tool, such as XCMS. However, it is important to acknowledge that important spectral information could be lost in translation from raw MS data to the required tabular format. This method assumes that spectral patterns and spectral deconvolution performance is similar between all samples in single datasets that are initially supplied as input, which may not hold for all detected compounds. Different preprocessing software tools were used in this study to determine potential sources of error resulting from dataset generation. Similar errors to those reported previously, such as incomplete or multi-peak integration, low signal-to-noise peaks, and non-extraction of true peaks. Such errors complicate the one-to-one alignment of features, as illustrated in **Figure 3.8**. Factors affecting the accurate estimation of m/z, central RT, peak area calculation as well as overall quantity of features have important ramifications for this analysis. Therefore, the preprocessing

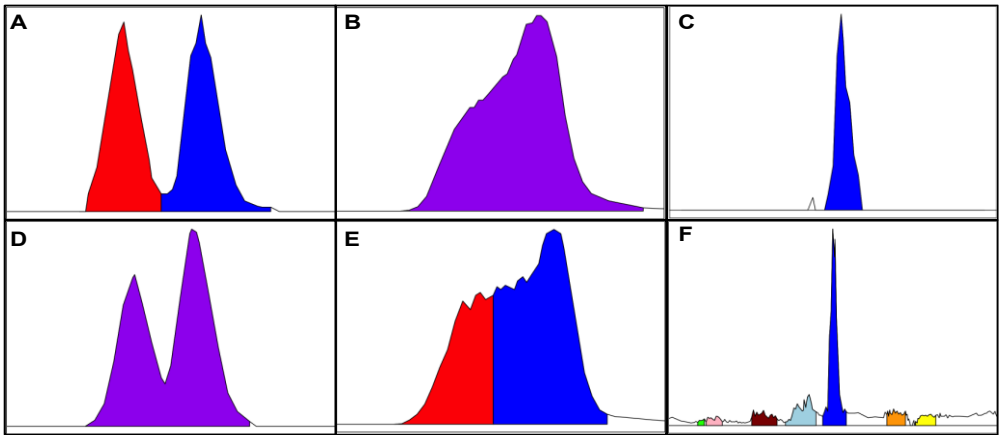


Figure 3.8 Common Peak Detection Errors. The top row (A-C) displays correctly integrated peaks, whereas the second row (D-F) display errors. In (A), two isomers are integrated as separate peaks, but in (D) they are fused as one feature. In (B), a wide peak is integrated as one feature, but in (E) it is split into two features. In (F), one peak is compared to multiple low signal/noise features, with only the abundant peak representing the true compound match. These errors complicate 1-1 matching between features.

method and choice of parameters should be considered carefully for each dataset. Future improvements in disparate LC-MS alignment methods may draw information from the raw spectral dimension to ascertain the validity of assigned spectral matches.

3.6.2 Union of Disparately Acquired Features

As previously discussed, imputation methods for missing features within a constituent dataset chosen as representative are limited to estimating the mean of {m/z, RT, and Q} values of the feature where they may exist in other datasets. This option is sensible for typical multi-batch alignment tasks, but not for disparate data alignment analyses. Consequently, the tool is not yet capable of generating an accurate union of features since signal missing from the dataset representative cannot be brought forth in the given *metabCombiner* cycle. Instead, one "primary" dataset can be selected as the X or Y feature list in all n-1 alignment tasks with the other datasets serving as "target" sets to be aligned with it. The only way to overcome this limitation under the current framework is to impute RT values for the features absent in the chosen representative dataset using information from datasets in which the feature is present, using RT projection models. The alternative is to explore a framework for simultaneous multi-dataset alignment as opposed to stepwise paired alignments.

3.6.3 Stepwise vs Simultaneous Alignment

metabCombiner was originally designed for aligning two LC-MS metabolomics feature tables and generating their intersection. While it was expanded to facilitate the alignment of more than two datasets, the workflow still proceeds as a stepwise pairwise merging task. Apart from the *batchCombine* functionality for multi-batch alignment tasks, simultaneous one-step alignment of three or more datasets has not been implemented. The primary reason stems from the need to carefully refine parameters in alignment tasks, especially those for anchor selection, RT modeling, and pairwise feature similarity scoring. Haphazardly applying identical or default parameters in every application often leads to suboptimal results. Future developments must focus on automating the process to a greater extent to allow for a hands-off approach compatible with non-expert users.

3.6.4 Gap Filling for Missing Feature Abundances

Related to the previous limitations, *metabCombiner* lacks a dedicated capability to re-extract signal that may have been missed in the initial peak-picking phases, such as performed by fillPeaks() in XCMS or Gap Filling in MZMine2. In theory, computed RT projection models can estimate the RT location of missed peaks, which can then be useful for extracting signal present at the <m/z, RT> coordinates in raw spectral LC-MS files, though this could inevitably lead to complications if the absolute RT fitting errors of the models are high. For now, features entirely absent from input feature tables are left with missing signal abundance values.

3.6.5 The Use of Relative Abundance

metabCombiner uses relative abundance (Q) in a unique manner compared to other alignment tools. It is first utilized when selecting ordered pair anchors for RT mapping and then it is incorporated alongside RT and m/z in pairwise alignment scoring. While useful for

contrasting high and low-abundance compounds, disparities in relative abundance may occur due to experimental factors, such as differences in sample preparation and in-source ionization. In many cases, formation and relative abundances of in-source adducts and fragments may differ for the same compounds between datasets. *metabCombiner* can be configured to give relative abundance a less prominent role than RT and m/z by setting the specific weight parameter in the similarity scoring step to 0.

3.6.6 RT Projection Error

metabCombiner is not the first tool to use a GAM for the purposes of RT-mapping between chromatograms. GAMs have distinct advantages over local regression (LOESS) and regression tree ensemble approaches. Their simplicity, versatility, and robustness to overfitting have been noted.[95] Setting the default "family" argument to *scat* (scaled t-family for heavy-tailed data) helps to eliminate the influence of outlier points, which may cause other overfitting in other approaches. One limitation of the workflow is that the RT mapping approach does not yet provide prediction intervals and only point estimates are used to weigh pairwise alignment scores. Prediction intervals may provide great utility due to the non-uniformity of RT mapping errors throughout the chromatogram. The influence of chromatographic variables was examined in testing, such as gradients, column types and dimensions, and mobile phase solvents, yet numerous other variables have yet to be fully explored. In general, datasets acquired from HILIC methods were more difficult to align than those by RPLC, a difficulty shared with previous studies attempting to predict compound retention in HILIC assays.[188]

3.6.7 Incorporating Additional Information for Feature Alignment

metabCombiner is designed for flexibility to accommodate as many LC-MS alignment cases as possible. Only m/z, RT, and sample abundances, all expected components of

conventionally pre-processed feature tables, are required to use the program. Feature identifiers can enhance the process by improving RT modeling steps and adduct labels can be compared between assigned feature matches, though neither is required for the overall procedure. In difficult alignment cases, e.g. between groups of isomers or features at the score threshold borderline (i.e. with relatively high m/z, Q, or predicted RT deviations), these measures are often insufficient to confidently and accurately determine feature correspondence. Additional parameters could increase confidence or point to alternative hypotheses in ambiguous cases that are not resolved by the five required and optional descriptors. MS/MS fragmentation, chromatographic peak shapes,[189] or isotopic envelope[190] information may serve as useful bases of comparison wherever such data exists.

3.7 Conclusion

LC-MS metabolomics has long been constrained by the requirement of replicated protocols and the incompatibility of data acquired under disparate analytical conditions. *metabCombiner* is a computational method for comprehensively mapping features from distinct untargeted metabolomics experiments and generating aligned datasets in an automated manner. This provides opportunities to build bridges between previously incomparable metabolomics data and increase the utility of studies beyond their initial uses. *metabCombiner* is a versatile approach with wide applicability to a variety of metabolomics datasets, without requiring prior knowledge of shared metabolite coverage. This tool has numerous applications, such as facilitating inter-laboratory comparisons, reproducibility assessments, collaborative compound annotation efforts, and generating expanded datasets suitable for meta-analysis. While it is designed for metabolomics data, the methods may be adapted to other untargeted LC-MS analyses of complex mixtures, provided that input datasets meet the core assumptions.

Chapter 4

Applications of Disparate LC-MS Alignment to Compound Identification

4.1 Introduction

Assigning unambiguous compound identities to detected LC-MS features is one of the most significant bottlenecks in computational metabolomics data analysis, with only a small portion readily identified in most studies.[191–193] Improving compound identification rates would facilitate a more complete mapping of data to biochemical pathways and gaining deeper biological insight. Currently, no single method can comprehensively detect and identify all metabolites in a complex mixture. A combination of chromatographic techniques (RPLC & HILIC), compound class-specific approaches (most notably lipidomics), and ionization modes are required to survey the diverse chemical space occupied by the metabolome.[194] As previously described in Section 1.4, numerous tools and algorithms have been developed to facilitate and expand upon structure elucidation capabilities in metabolomics studies, such as compound databases, mass spectral database search tools, in silico fragmentation modeling approaches, and RT prediction models.[94] Purified standards are required for fully authenticated compound identities, which is generally not feasible to obtain for most compounds; therefore, identifications assigned by computational tools are generally at levels 2 ("probable") or 3 ("possible"), according to the Metabolomics Standards Initiative Compound Identification Working Group.[195]

As an alignment approach designed for disparate LC-MS data, *metabCombiner* falls into the category of RT modeling tools. Specifically, it belongs to the class consisting of RT projection models in which information is mapped between similar but non-identical LC systems, estimating where identified compounds may fall in a different separation space. By itself, *metabCombiner* cannot be used to infer the structural characteristics of unknown compounds, perform spectral library matching or generate de novo identifications of novel metabolites. Instead, its primary utility in compound identification is information sharing and establishing consensus annotations between confidently matched metabolomics features. *metabCombiner* can be paired to traditional and novel identification methods to map compounds beyond the limited chromatographic spaces in which they are assigned.

In this chapter, three separate applications of disparate LC-MS alignment using metabCombiner are described in furtherance of the goal of increasing identification rates in untargeted LC-MS metabolomics studies. First, data from a published large-scale identification study of human urine samples acquired using HILIC-MS were aligned to a data acquisition of stock urine samples from a different laboratory, and resulting annotations were compared. Second, an expanded gradient LC-MS analysis of plasma designed to achieve substantially higher identification rates is mapped to conventional length LC-MS data of the same samples. Finally, *metabCombiner* was used to align datasets from a multi-institutional untargeted lipidomics study, where known and unknown lipids have been detected and annotated across multiple specimens using separate protocols from each laboratory.

4.2 Alignment of Urine Mass Spectral Features Analyzed by HILIC-MS

4.2.1 Background

Human urine is among the most widely studied biofluids as it provides a rich source of biomarkers reflecting excretory processes influence by various physiological conditions and

stimuli. The urinary metabolome has been associated with health-related conditions such as urological and non-urological cancers, asthma, hepatitis, liver cirrhosis, metabolic syndrome, and a wide range of toxicities, exposures, treatments, and substance uses.[196] The National Metabolomics Data Repository[132] and Metabolights[133] repository contain dozens of human urine metabolomics studies. Comprehensive characterization of the full urinary metabolome was the subject of numerous endeavors, using library searches,[194] literature-mining,[197] molecular networking,[198] and recurrent cross-laboratory mass spectral signature curation.[199]

In one study, Blaženović et al attempted the comprehensive characterization of all mass spectra in a dataset consisting of 43 human urine samples from interstitial cystitis patients.[194] Biphasic extraction performed on separated urinary metabolomes into their lipid and polar fractions, and separately analyzed with two chromatographic methods: charged-surface hybrid chromatography (CSH) for lipidomics and hydrophilic interaction chromatography (HILIC) for profiling polar metabolites. Of the thousands of signals extracted and aligned, roughly 42% triggered MS/MS fragmentation spectra in Data Dependent Acquisition (DDA) mode. Using over 1000 authentic compound standards, various mass spectral fragmentation and retention time libraries, and tools such as NIST Hybrid Mass Spectral Search,[200] CSI FingerID for in-silico spectral prediction,[201] and ClassyFire for chemical ontology assignment,[3] features were annotated at MSI levels 1, 2, 3, or 4. The overwhelming majority were classified at level 3 or 4, but more than 500 could be identified at more confident MSI levels 1 (1-2%) and 2 (10-15%), higher annotation counts than can be found in most routine LC-MS metabolite profiling studies.

To take advantage of these efforts performed on a specific sample set within one laboratory, untargeted HILIC-MS metabolomics measurements were acquired from stock human urine samples under disparate conditions and were aligned to the published study data using

metabCombiner. Pre-processing and simple compound identification efforts were performed on the new dataset. Confident alignments from the two datasets provide information about consistently (and inconsistently) named compounds as well as hypothesized annotations for features named in one dataset but not its complement.

4.2.2 Experimental Methods

Data from 3 pooled replicates of healthy human urine obtained from BioIVT (Westbury, NY) and NIST Standard Reference Material SRM3673 (together called 'B3N3') were aligned to the study data published by Blaženović et al consisting of 43 samples from interstitial cystitis patients (hereafter referred to as '*IC43*'). *B3N3* samples were thawed and extracted using a biphasic aqueous / methanol tert-butyl ether solvent system, exactly as previously described for *IC43*; only the aqueous layer was used for subsequent HILIC-MS analysis. Samples were analyzed using an Agilent 6545 LC-qTOF mass spectrometer (as opposed to a Thermo Q-Exactive LC-MS as in *IC43*). The chromatographic approach used for *B3N3*, including gradient length and mobile phase composition, were replicated from the *IC43* methods, but with a shorter column (Waters Xbridge Amide 1.7 um, 2.1 mm ID, 100 mm in length as opposed to the 150 mm length columns used for *IC43*), inducing major RT shifts.

For MS1 analyses, MS source conditions were as follows: Agilent Dual Jetstream ESI, positive ion mode, source gas temp 275 C, drying gas 12 L/min, nebulizer 45psi, sheath gas temp 325 C, sheath gas flow 12 L/min, capillary voltage 4000, MS scan range 50-1200 Da, 2 spectra/sec, reference mass correction enabled. For MS2 analysis in *B3N3*, all parameters were the same as *IC43* except the MS2 scan range was 25-1200 Da with a rate of 2 spectra/sec. The isolation width was narrow, collision energy was 20, 3 precursor ions were allowed per cycle with active exclusion enabled after 2 spectra for 0.5 minutes. Four runs of iterative MS/MS

(rolling-precursor ion exclusion between replicate LC injections of a sample) were used with a mass error tolerance of 20ppm and RT tolerance of +/- 0.5 min.

4.2.3 Data Pre-Processing & Feature Identification

Raw MS1 mass spectral files from *B3N3* and *IC43* (the latter downloaded from Metabolomics Workbench Study ID ST001122) were both processed by MZMine2 using the ADAP pipeline,[53] extracting 10624 & 22313 features respectively. MS2 data for *B3N3* were loaded into Masshunter Qualitative Workflows (Agilent) and features were detected using the "Find by Auto MS/MS" tool and resulting MS/MS spectra was exported in MGF format. These data were simultaneously searched against the NIST 2017 tandem MS library and the MoNA LC-MS/MS positive mode library (http://massbank.us, downloaded 12/2019), using the NIST MSPepSearch software tool. Resulting MS/MS hits were considered "identified" (MSI level 2) if the NIST score was >650, the dot product score was >750, and visual review confirmed that spectra were good matches with multiple well-aligned fragment ions.

For both datasets, custom R scripts were deployed for the purposes of matching compound information to pre-processed MS1 features. For *B3N3*, the associated m/z and RT values were searched with tolerances of 0.007 Da and 0.25 min. In addition to the main adduct form, a list of adduct and fragment variants of these metabolites were searched using their associated mass-based rules, including [M]+, [M+H]+, [M+Na]+, [M+2Na-H]+, [M+NH4]+, [M+K]+, [2M+H]+, [2M+Na]+, [M+H-H2O]+, [M+H-NH3]+ and [M+H-HCOOH]+. Compound and adduct annotations were examined for consistency and validity (small RT & m/z deviations from with respect to other variants). This process yielded 123 assigned compound identities with at least one annotated variant, all at MSI level 2. A similar process was employed for *IC43*, using identities drawn from Table S2 of the published manuscript which were labeled

as either mzrt or MS2 matches. An in-house spectral library (*uclib*) was also used in this search. For feature identity assignments, mzrt named matches were prioritized, followed by MS2, and lastly uclib, with 101, 199, and 144 respective compound matches.

In the initial data processing steps of *metabCombiner*, RT ranges were set to 0.5 to 10.5 min (*B3N3*) and 0.5 to 11 min (*IC43*), excising sparse head and tail regions, and reducing by 213 and 268 features, respectively. The default missingness threshold 50% reduced *B3N3* by an additional 88 features and by over 11000 features in IC43. The final feature count was 10320 and 11014 for *B3N3* and *IC43*.

4.2.4 Results and Discussion

A preliminary survey of the datasets revealed subpar mass accuracy, particularly for the *B3N3* dataset, due to errors related to instrumental or pre-processing software factors. Therefore, a wider m/z grouping *binGap* tolerance value of 0.01 was used, despite a substantial enlargement to 95898 rows in the initial FPA table (compared to 31248 and 55224 rows for binGap values of 0.005 and 0.0075 Da, respectively). This large initial table size is attributable to the density of detected features in the low m/z range (especially between 100-200Da). *metabCombiner* analysis for these datasets was conducted in two stages. First, an unsupervised analysis was conducted without relying on named features and assessing the validity of alignments. This was followed by a semi-supervised analysis using consistently named compound identities to obtain a more accurate RT mapping and aligned table reduction.

The anchor selection step produced 66 ordered pairs for mapping between RTs in both datasets, using *B3N3* as dataset X and *IC43* as dataset Y. Scoring parameter values were set to A = 60, B = 8, C = 0.3. Forty-one consistently named compounds between both datasets achieved the highest-scoring alignments among their respective groups. Three more best-scoring

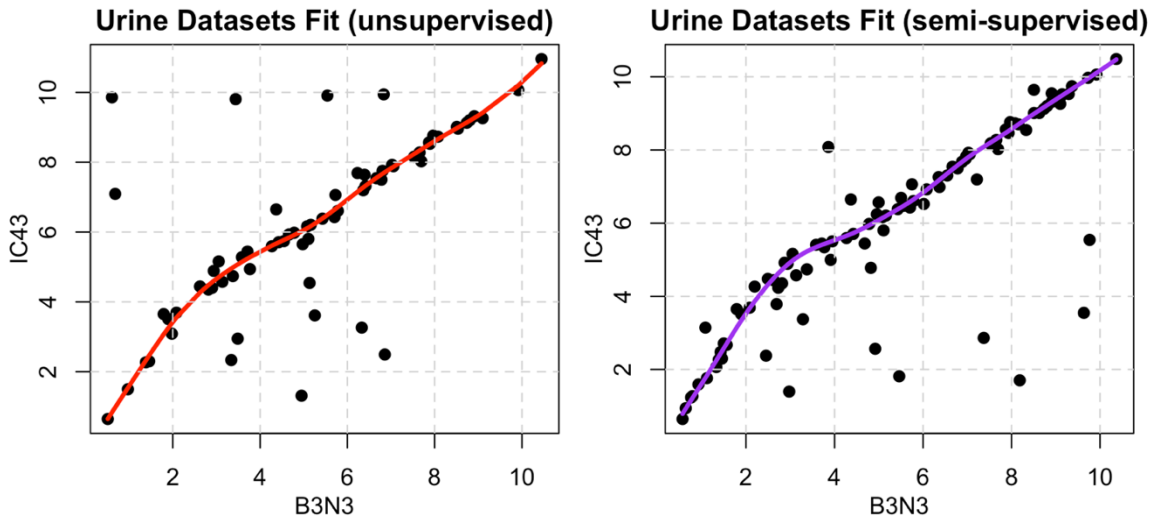


Figure 4.1 Unsupervised vs Semi-Supervised RT Fits No prior information was used in the unsupervised fit (left), whereas features with identity agreement were incorporated as anchors in the semi-supervised analysis (right). IDs refine the RT mapping, especially in sparsely-anchored chromatographic regions.

alignments were between features named as positional isomers, e.g. 4- and 3-hydroxypyridine, which were counted among the consistent set. The computed GAM mapped RTs accurately for most compounds, with 28 and 33 fitted within 0.1 and 0.2 min (1-2% error) of the observed RT, respectively. One named compound, Ornithine, eluted 0.67 min later than predicted as its elution order changed considerably between the two datasets. Seven identically named features did not score highly, mainly due to excessive RT fit deviations. In fourteen cases, high-scoring alignments were observed between features with mismatched identities; six of these could be resolved through manual review of MS/MS or correcting adduct annotations of these features. Many assigned features had no probable match in the counterpart dataset (particularly among drug-related metabolites) and a few others could not be definitively assigned as a match due to low scores or the presence of conflicting feature(s). On the other hand, aligning *B3N3* to the well-annotated *IC43* provides moderate-to-high scoring alignments to 167 distinct features that were named in *IC43* but not *B3N3*. These alignments provide a list of putative identities which can be subsequently verified with authentic standards.

In the second stage, semi-supervised alignment analysis was performed with the aid of the 44 consistently named metabolites. With this adjustment, 98 ordered pairs were selected for anchoring the updated RT mapping. A visual of the two model fits is shown in **Figure 4.1**. The greatest differences in the model-predicted RTs are observed in the early to middle chromatographic regions. Score parameters, as guided by the *evaluateParams* method, were similar to those used in the first stage, with only *B* changed to 7. The table was then reduced from 95898 to 3265 FPA rows, or roughly 3% of the original table size.

This analysis demonstrates that large-scale compound identification results within data generated by other laboratories are transferrable and usable by other laboratories for generating many putative annotations, even for similar but non-identical specimens. This generates identification hypotheses that laboratories can test without excessive duplication of efforts as well as focusing deeper efforts on mutually detected unknowns. This application demonstrates that the use of shared known identified features as anchors can be useful for increasing the modeling and feature matching performance of *metabCombiner*, leading to further discovery of accurate named compound matches.

4.3 Modifying Chromatography Conditions for Improved Unknown Feature Identification in Untargeted Metabolomics

4.3.1 Background

Compound identification efforts for unknown features usually focus on acquiring MS/MS data that can be searched against experimental and *in silico* spectral databases. Due to the incompleteness of spectral libraries, many high-quality MS/MS spectra cannot be matched until new compound standards are acquired. In addition, features may be difficult to identify because MS/MS spectra they produce contain few unique product ions[202] or due their low abundances where minor fragments may fall below noise thresholds. Multiple experimental approaches can

be employed to increase LC-MS/MS coverage and, by extension, compound identification rates. One such approach is to improve the peak capacity, or the number of separate peaks within a retention window, by using longer columns and/or extended gradients.[203] Another approach is to perform consecutive data-directed LC–MS/MS runs with rolling precursor ion exclusion, which reduces the collection of redundant spectra and allows for annotation of more low-abundance features. Finally, studies show that increasing sample concentrations or injection volumes improve identification rates slightly, even with degraded chromatographic performance.[204]

This section outlines an application of *metabCombiner* exploring how modified chromatographic and MS/MS variables coupled to disparate LC-MS alignment can improve compound identification rates.[205] The first goal of this study was to explore the potential for improving feature identification rates in untargeted metabolomics substantially by altering chromatographic parameters, such as gradient length and mass loading, and employing rolling precursor ion exclusion. The second goal was to determine a procedure for categorizing features as "high-priority unknowns" (HPUs), or spectra that have a higher likelihood of being identifiable to follow-up structural elucidation analyses. Finally, this study aimed to map putatively identified features and HPUs identified in modified chromatographic conditions data to features detected within conventional length LC–MS runs, transferring the beneficial properties of conditions optimized for compound identification to routine, high-throughput untargeted metabolomics analyses. Modified chromatographic conditions yielded a substantial improvement of confident compound identifications over conventional LC-MS/MS analysis in a human plasma sample, based on data from automated spectral searching aided by manual review.

4.3.2 Experimental Methods

Pooled human plasma was obtained from the Red Cross of Michigan and metabolites extracted using a 1:1:1:1 methanol:acetonitrile:acetone:ethanol solution.[206] Depending on the concentration and volume required, one to several aliquots were pooled for analysis. Concentrations of extract samples injected for LC–MS analysis are reported relative to undiluted human plasma (e.g., 0.4× plasma signifies a reconstituted extract with 40% concentration of pre-extraction plasma, assuming full recovery of metabolites)

LC–MS was performed on an Agilent (Santa Clara, CA) 1290 Infinity II LC system coupled to an Agilent 6545 QTOF mass spectrometer. RPLC separations were performed using a Waters (Milford, MA) Acquity UPLC HSS T3 column (2.1 × 100 mm; 1.8 μm). The flow rate was set to 0.4 mL/min and mobile phases consisted of (A) water with 0.1% formic acid and (B) methanol with 0.1% formic acid. HILIC separations were performed on a Waters Acquity UPLC BEH amide column (2.1 × 100 mm 1.7 μm). The flow rate was set to 0.3 mL/min, and mobile phases consisted of (A) 95:5 acetonitrile/water with 0.125% v/v formic acid and 10 mM ammonium formate (with a 10-minute sonication to ensure that the ammonium formate thoroughly dissolves) and (B) 95:5 water/acetonitrile with 0.125% v/v formic acid and 10 mM ammonium formate. Both columns were maintained at 55 °C for separations. Different gradients and total chromatography times were compared for this analysis. In RPLC-MS and HILIC-MS methods, 21-22 minutes represents a conventional length LC-MS analysis.

Instrument settings for positive mode ESI were: Sheath gas flow rate, 11 L/min; drying gas, 8 L/min; drying gas temperature, 320 °C; nebulizer, 35 psi; capillary voltage, 3500 V; nozzle voltage, 1000 V; fragmentor, 175 V; skimmer, 65 V; Octupole 1 RF Vpp, 750 V; collision energy, 20; iterative MS/MS mass error tolerance, ± 20 ppm; iterative MS/MS retention time (RT) exclusion tolerance, ± 0.5 min (0.1 min also evaluated); spectrum data type, centroid.

Data-dependent MS/MS parameters were: mass range, 25–1200 *m/z*; rate, 2 spectra/s; max precursor ions per cycle, 3; absolute precursor threshold, 5000 counts; relative precursor threshold, 0.001%; active exclusion enabled after 2 spectra and released after 0.5 min; isolation width, narrow (~1.3 *m/z*).

4.3.3 Compound Identification Methods

Briefly, LC–MS/MS data were loaded into an interface for the input and output of the NIST *MSPepSearch* (www.chemdata.nist.gov), searching against NIST20, LipidBlast,[207] Metlin,[79] and Massbank of North America (MONA, massbank.us) spectral libraries and results were reviewed. Spectral search hits were ranked by score values and visualized with head-to-tail plots, and pattern recognition entropy (PRE) and total intensity were calculated for all spectra.[208] Compound matches must meet multiple criteria (mass accuracy within 20 ppm, at least two fragments, EICs must resemble metabolite peaks and not background ions) and must not be degenerate feature per *Binner*[171] analysis to be considered for this workflow. Compound class annotations (MSI level 3) were assigned to features with search results that demonstrate good fragmentation alignment without acceptable precursor ion agreement, as determined in manual review. Remaining features were considered MSI level 4 (unknowns).

After reviewing MS2 spectra, a subset of the annotated and unknown compounds (MSI3 and MSI4) was classified as High Priority Unknowns through a linear discriminant analysis- k-nearest neighbors (LDA-KNN) method. Identified metabolites (MSI2) are distinguished from annotated (MSI3) or unknown (MSI4) compounds, using Z-transformed RT, precursor *m/z*, PRE, and total intensity values as predictors. Then, unknown spectral features surrounded by neighbors in 4D space consisting of more identified MSI2 than unidentified spectra for a majority of iterations were designated as HPUs.

4.3.4 LC-MS Pre-processing & Disparate Alignment Methods

Metabolite identifications and HPUs uncovered by the previous steps within modified (longer run, higher sample loading) RPLC and HILIC conditions were mapped to extracted features in conventional (i.e., ~20-minute run time) LC-MS metabolomics data. First, MS1 feature detection was performed with MZmine2 v. 2.42 for modified and conventional LC-MS conditions.[46] The targeted peak detection module was used to extract features in modified LC-MS conditions data (consisting of one RPLC and four HILIC samples), based on the m/z and RT values of identified compounds and HPUs. MSI2 identifications HPU features from the RPLC and HILIC experiments were filtered such that only the feature with the highest database score/match factor for a given compound identification was maintained for *metabCombiner* alignment. Untargeted data analysis modules, consisting of chromatogram deconvolution, isotopic peaks grouping, join alignment and gap filling steps, were performed for the conventional LC-MS data (four RPLC and five HILIC samples). For the conventional conditions datasets, features with an average sample to average blank peak area ratio less than two were removed, leaving total feature counts of 35,878 (RPLC) and 10,715 (HILIC).

The resulting feature lists were then aligned using *metabCombiner*. In both analyses, features were grouped by m/z with a binning gap (binGap) value of 0.01 Da. Alignment of all feature pairs was scored from 0 (poor alignment) to 1 (excellent alignment) based on differences in *m/z*, RT fitting error, and relative abundance, with specific weights 60, 10, and 0.1 respectively. Scores above 0.75 were classified as high-confidence alignments, while scores below 0.5 were rejected. Alignment scores between 0.5 and 0.75 were classified as moderate-confidence matches, for which manual validation is recommended.

4.3.5 Compound Identification Results

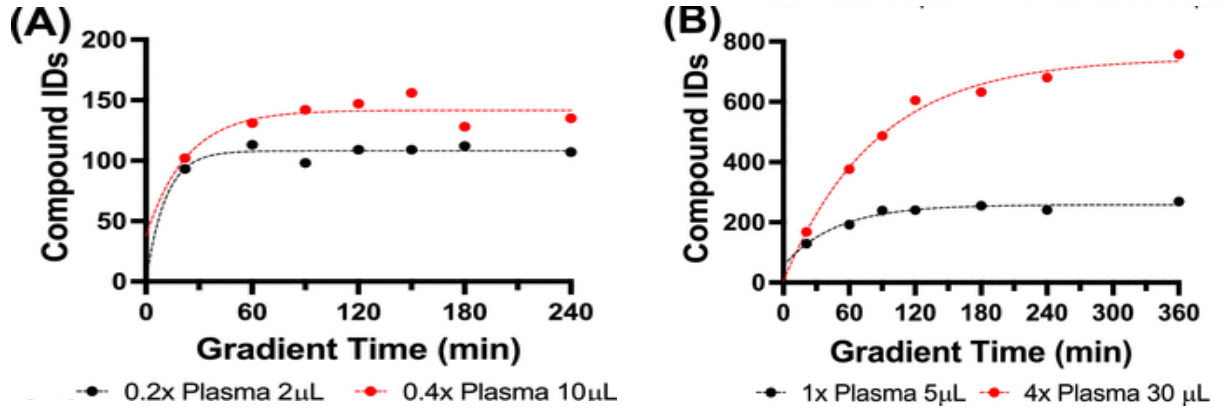


Figure 4.2 Effect of Total Gradient Time on Compound Identification (A) HILIC-MS; (B) RPLC-MS

Identifications were evaluated based upon library match score values and spectral characteristic thresholds from a summary of manually reviewed features. Generally, features assigned at MSI3 and MSI4 had lower MS/MS database search scores, total intensity, and/or spectral entropy than MSI2 identifications. Thresholds and criteria selected for NIST MS/MS database search score, intensity, and entropy removed more low-quality spectra compared to thresholds based on spectral agreement alone,[94,199] while retaining valid identifications.

The impact of sample loading, gradient length, and iterative acquisition on compound identification performance was assessed by the number of unique MSI2 identifications from LC–MS/MS features searched against each of the libraries. Total chromatography times ranged from 21 to 360 min, with total uncovered compound identities displayed in **Figure 4.2**. For HILIC, the modified gradient length and sample loadings selected for improved compound ID was 120 min (10 µL of 0.4x plasma), above which no substantial increase in unique, identified (MSI2) features was observed. Though slight improvements were observed with longer runs, 180 min runs was selected for RPLC (30 µL of 4× plasma) to complete the iterative MS/MS acquisitions in a reasonable 24 hour timeframe.

Increasing the sample injection volume alone resulted in a plateau in the number of unique identifications, whereas increasing sample concentration slightly increases the compound

ID count. At low sample concentrations and sample volumes, low-abundance compounds are difficult to detect and often produce too few fragments to be confidently identified by spectral database searches. Increases in identification rates with higher column loading can be attributed to improvements in MS/MS spectral data quality for low abundance features and acquisition of spectra for compounds whose levels were too low to trigger MS/MS acquisition under conventional conditions. There were tradeoffs observed between sample loading and chromatographic resolution, with detector saturation observed for excessive sample loads. Nevertheless, the results show that increasing sample loading improves compound identification rates, especially when coupled to longer LC-MS analyses.

In the HILIC data set, 449 unique compounds were assigned at MSI level 2 or better to 30.4% of collected MS/MS features, verified by manual review. RPLC data were assessed in a semi-automated fashion, with limited manual review of a subset of spectra. 1885 unique identifications were made with the modified RPLC conditions, corresponding to 9.5% of collected MS/MS features. Equivalent semi-automated analyses of both conventional HILIC and RPLC conditions were also performed. In total, 2052 unique compounds were identified (MSI2) by the two modified methods, compared to 214 which could be identified using a single conventional LC–MS/MS run on both HILIC and RPLC. LDA-KNN analysis, using k = 7 as the most accurate with respect to feature classification, selected 576 HILIC and 749 RPLC features in the modified conditions datasets as HPUs, or spectra likely to be identifiable by MS/MS matching when the underlying compound is present in a database. Identified spectra in the kNN space were typically more abundant than unidentified ones and had midrange entropy values; HPUs were thus found to be abundant and produced multiple key fragment ions in common with nearby identified features (in *m/z* and RT).

4.3.6 Disparate LC-MS Alignment Results

Increasing sample loading and gradient duration achieves superior data-dependent MS/MS spectral quality and compound identification performance. However, it is impractical to employ these conditions for every sample in routine metabolomics studies. Therefore, identifications made using modified LC-MS conditions were mapped to unidentified features in data with shorter, conventional length LC-MS assays with lower sample loading, using *metabCombiner*. Spectral peaks were extracted for the list of MSI2 compounds (1885 RPLC & 449 HILIC) and HPUs (749 RPLC & 576 HILIC) detected under modified conditions and aligned to data from conventional ~20-min LC–MS methods.

Peak-picking with *MZMine2* was more challenging in the modified LC-MS conditions for multiple reasons. The raw mass spectral files were significantly larger than conventional LC-MS raw files, demanding more computational resources and time to process using untargeted analysis modules, particularly for the ADAP chromatogram deconvolution step.[53] The degraded chromatographic performance led to challenges with assigning accurate RT values for MS1 peaks, particularly for elongated peaks and peaks with low signal-to-noise ratios. Using targeted peak detection for the modified conditions is significantly faster than using the untargeted detection modules and it enables more assignments of identified compounds to MS1 peaks, though this comes with several tradeoffs. Features are detected by integrating signal located at the supplied m/z and RT values, with minimal peak quality assurance. Many extracted signals exhibit poor peak-like characteristics, including signal resembling background or noise. Targeted detection only uncovers features associated with identified metabolites or HPUs, leaving out thousands of peaks not associated with these compounds. Consequently, percentile-based relative quantitation (Q) values will be less informative within *metabCombiner* and alternative alignment

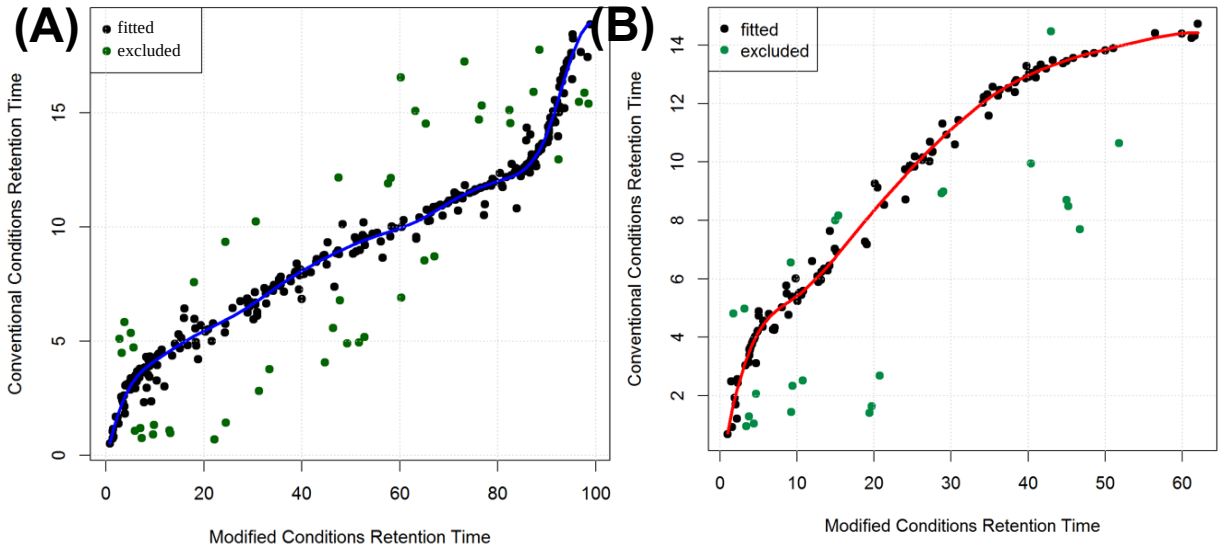


Figure 4.3 Modified and Conventional Conditions LC-MS RT Mapping (A) RPLC (B) HILIC

hypotheses involving these undetected features will be missed the process. Finally, this targeted detection process must achieve sufficient coverage to span the breadth of the chromatogram as a prerequisite to disparate LC-MS alignment.

Chromatographic mapping steps in *metabCombiner* projected the highly separated modified conditions retention space onto the narrower conventional conditions RTs. To improve RT mapping performance, identifications from the modified separation feature lists were capped at 99 and 64 min for RPLC & HILIC data, respectively, since highly retained compounds could not be effectively mapped to features under conventional conditions without spline fit distortions in the tail chromatographic region. Conventional conditions feature lists were capped at 19 and 15 minutes for RPLC & HILIC, respectively, and features eluting before 0.5 min were removed. Images of the mappings for the RPLC and HILIC experiments are shown in **Figure 4.3**.

Using the weighted scoring metric which considers *m/z*, RT alignment, and relative abundance differences, *metabCombiner* determines the most likely feature matches between the long and short length LC-MS conditions. **Table 4.1** summarizes the number of identified compounds mapped to the shorter method for the RPLC and HILIC experiments.

Table 4.1 Modified to Conventional LC-MS Conditions Alignment Results						
	HILIC			RPLC		
Description	Total	Known	HPU	Total	Known	HPU
Initial List of Unique IDs	1025	449	576	2634	1885	749
Unique IDs not assigned to any MS1 peak	25	9	16	32	21	11
Unique IDs Filtered by RT	2	2	0	296	196	100
Unique IDs filtered by Missingness (HILIC only)	11	3	8	N/A	N/A	N/A
Unique IDs Present after initial filters	987	435	552	2306	1668	638
Unique IDs with no m/z matches	101	37	64	135	93	42
Unique IDs with rejected alignments	190	55	135	434	331	103
Unique IDs with moderate-confidence alignments	202	91	111	543	379	164
Unique IDs with at least one high-scoring match	494	252	242	1194	865	329

High-to-moderate-confidence alignments were achieved for 68.0% of the newly identified MSI2 features (calculated as HILIC high and moderate-confidence features + RPLC high and moderate-confidence features / total feature count). Many HPU features from the HILIC and RPLC data sets were also successfully mapped to the high-throughput MS1 runs (63.8 and 65.8%), allowing for quantification of these compounds using conventional LC-MS conditions while orthogonal identification approaches are considered. Identified compounds and HPUs that could not be accurately aligned to a feature in the conventional length LC-MS dataset either lacked any features within close m/z proximity, had all alignments rejected due to weighted scores lower than 0.5 and/or pairwise score ranks higher than 3, falling into filtered chromatographic regions, missing from two or three samples (HILIC only), or the compound could not be detected by MZMine2 at the given m/z and RT within MS1 data.

Figure 4.4 below shows an example pair of EICs from the conventional and modified RPLC-MS conditions. *metabCombiner* accurately matched two peaks eluting at 11.7 and 76.5 minutes, respectively, which were eventually confirmed to be 2-acetyl-1-alkyl-sn-glycero-3-phosphocholine through manual spectral review. This showcases how exhaustively curated identities can be extrapolated between datasets with as much as five-to-ten-fold differences in

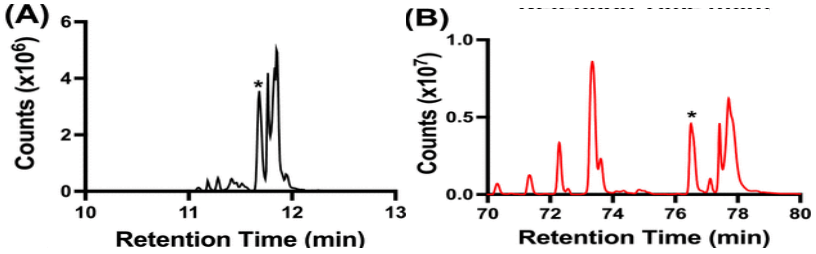


Figure 4.4 Example Conventional vs Modified Conditions EIC Comparison m/z ~ 546.35 for (A) conventional & (B) modified conditions.

total liquid chromatography lengths. However, it is notable that the effects of increased sample loading or reduced ionization suppression from better separation resulted in intensity shifts between modified and conventional conditions, which is demonstrated by the leftmost peaks (11 - 11.5 minutes in conventional LC, 70-75 minutes in modified LC). This suggests further weaknesses in relative abundance comparisons as differences in LC variables may result in substantial changes in the quantitation of various metabolite species.

4.3.7 Discussion

This study demonstrates the potential to substantially improve identification rates in routine metabolomics studies using altered chromatographic protocols coupled to disparate LC-MS alignment. As this study shows, major differences exist between LC parameters favoring quantification versus compound identification. Quantification requires moderate-to-fast run times and adequate specimen amounts to permit large scale biological sample analysis without saturating the detector, whereas greater sample amounts, longer total chromatography times, and iterative acquisition yielded a roughly nine-fold increase in identifications over conventional conditions. Disparate LC-MS alignment is essential for translating the benefits of optimized LC-MS conditions to routine metabolite profiling experiments, with the majority of identified compounds matching unknown features in conventional conditions data with moderate to high confidence. This generates hundreds of new annotation hypotheses for subsequent validation in follow-up experimentation.

4.4 Inter-laboratory Study of Unknown Lipids in Untargeted Lipidomics Data

4.4.1 Background

Lipids are among the most complex and structurally diverse compound classes found in organisms. According to some estimates, the number of distinct lipid structures in mammalian cells may number in the hundreds of thousands,[209] more than the endogenous metabolome (consisting of amino acids, carbohydrates, and nucleic acids) combined.[210] The Lipid Maps classification system divides lipids into eight categories, each with their own subclassifications: fatty acyls, fatty acyls, glycerolipids, sphingolipids glycerophospholipids, saccharolipids; and sterol lipids and prenol lipids.[211,212] Lipids are crucial components of cellular membranes and play essential roles in signaling and energy storage, among other critical functions.[213] Concentrations of lipids are wide-ranging and highly dynamic, changing with physiological, pathological, and environmental conditions. Given their significance, efforts are underway to comprehensively map the lipidome, or all lipids that may be present in biological specimens.

Multiple software tools, databases and other resources have been developed for assisting lipidomics identification. LipidMaps contains a structure database with 47454 known lipid entities in total (as of 04/22), as well as tools for aiding mass spectrometry-focused identification and statistical analysis.[115] The LipidBlast *in-silico* mass spectral library contains computer-generated fragmentation spectra for over 100,000 lipids, spanning dozens of classes and belonging to mammalian, plant, bacterial and other organisms.[207] Software packages such as Lipid Annotator,[214] LipidMS,[215] Greazy,[216] Lipidex,[217] and LipidHunter[218] apply algorithms for confident identification in experimental LC-MS/MS spectral data. Several inter-laboratory and multi-instrument investigations have been reported for harmonizing lipidomics datasets or determining the consistency of lipidomics measurements. For example, Bowden et al. (2017)

details a study among 31 laboratories of NIST SRM 1950, a plasma reference material, identifying 1527 unique lipids in total across all laboratories, with measurement estimates and uncertainties assigned for a subset of commonly detected lipids.[219] Cajka et al. (2017) investigated the quantitative performance of nine mass spectrometers in measuring human plasma lipids[220] and Spanier et al. (2021) compared lipid profiles of *C. elegans* in four separate laboratories.[221] The focus of these investigations was lipid quantitation, with an exclusive focus on shared known compounds and with identical samples, sample preparation procedures, chromatography conditions, and data processing steps. No inter-laboratory collaborative identification efforts for unknown and unidentified lipids have been reported to date.

This section describes the results of an inter-laboratory study conducted by four NIH Compound Identification Development Cores (CIDC), assessing the reproducibility of lipidomics measurements in multiple specimens (plasma, liver, brain, heart, and muscle). The primary goals of this consortium-wide investigation are to determine the extent of detected compound overlap between different laboratories, as well as consensus annotations and common unidentified features to prioritize for in-depth lipid identification efforts. Accomplishing these tasks required the disparate multi-dataset alignment functionality of *metabCombiner* to compare datasets that were acquired using distinct lipidomics protocols employed by each institution. After aligning the datasets, the extent of feature overlap and annotation agreement was assessed between the four institutional participants.

4.4.2 Methods

Hereafter, the four institutions taking part in this inter-laboratory study and their respective datasets are designated as "I", "II", "III", and "IV". Each received NIST1950 reference human plasma along with three additional pooled plasma samples representing diverse

populations; human skeletal muscle extract prepared from a bulk pool of frozen tissues from deidentified human donors; bovine heart and liver total lipid extracts, porcine brain total lipid extracts, and UltimateSsplash ONE lipid internal standard mixture from Avanti Polar Lipids (Alabaster, AL). Plasma and muscle samples prepared by liquid-liquid extraction and tissue total lipid extracts (heart, liver, brain) were mixed with the internal standard compounds and transferred to autosampler vials.

Each laboratory performed untargeted LC-MS/MS analyses on all samples, once using a "common method" protocol and once using individualized "in-house" standard operating procedures that each laboratory typically performs for untargeted lipidomics. The analysis described here focuses on the in-house data in the positive and negative ionization modes, which harbor greater between-dataset disparities than the common pipeline data. Laboratories employed different plasma lipid extraction methods, mass spectrometers, liquid chromatographic columns, instruments, and gradient elution methods to profile lipids from the five specimen types. In addition, each laboratory separately performed data pre-processing and compound identification workflows, with a focus on MS1 peaks that yielded MS/MS spectra.

The four preprocessed tables per ionization mode were aligned with *metabCombiner*, using the stepwise multi-dataset alignment framework. Sample measurement columns corresponding to one chosen specimen type (plasma) were used for relative quantitation (Q) comparisons between datasets, with the remaining sample columns designated as "extra" in the initial data processing step. Each table was filtered based on RT constraints, percent missingness, and duplicate m/z & RT criteria. Datasets *I* and *II* were aligned and merged first, then the result of this process is aligned to dataset *III*, followed by *IV*. In each cycle, 'primary' dataset *I* was the designated "X" data counterpart and its numerical feature descriptors (m/z, RT, Q) were selected

for pairwise comparisons with features of 'target' datasets *II*, *III*, and *IV*. At the start of each alignment cycle, the m/z grouping step generated lists of potential feature matches using a common binGap value of 0.0075 Da. Different parameter values for anchor selection, RT projection model fitting, feature pair scoring, and table reduction were used in each cycle, with guidance from the fit images. Features lacking complementary matches in pairwise alignments were recovered at the end of every *metabCombiner* workflow cycle and joined with the remaining matched features. Only features present in *I* could be aligned across multiple datasets under this formulation; features absent in *I* and detected in one of the target datasets are displayed in the results as unmatched entities

Once aligned, the extent of overlap between the four tables is assessed for all features and the subset of matching compound identities. Full aligned feature overlap counts are based on analytes detected in primary dataset *I*. Features present in the target datasets but not *I* are counted as found in one dataset only. Identities were considered equivalent if the lipid class, number of carbons, and double bonds matched between tables (e.g. TG 56:2), although multiple isomers may be detected in the datasets with the same annotation. Intermediate results generated after every alignment step were inspected to determine the degree of overlap between dataset pairs.

4.4.3 Results

Initial and processed feature counts for all four datasets in both ionization modes are displayed in **Table 4.2**. Initial table sizes range from 3819 to 29900 features in the positive mode and 2262 to 14865 in the negative mode. Dataset *I* contains a long chromatographic tail region between 14 and 20 minutes consisting of features that couldn't be mapped to other target dataset features, thus a RT filter was set with a cutoff of 14 minutes to obtain better RT projection results. The missingness filter excised a considerable fraction of datasets *I* and *IV*, and at most

Table 4.2 Inter-laboratory Lipidomics Study Initial and Processed Feature Counts						
Mode	**Dataset**	**Initial Feature Count**	**RT Filter**	**Missingness Filter**	**Duplicate Feature Filter**	**Final Feature Count**
POS	**I**	15004	320	4088	211	10385
	II	3819	0	0	10	3809
	III	27289	0	2278	106	24905
	IV	29900	17	10508	6	19369
NEG	**I**	10962	1224	2774	161	6799
	II	2262	0	0	15	2247
	III	7643	0	315	20	7308
	IV	14865	155	5774	1	8935

hundreds of duplicates were detected and removed. The remaining counts were 3809 to 24905 and 2247 to 8935 for positive and negative mode datasets, respectively. Pre-processing software and associated parameters used by each laboratory is a significant determinant of resulting table sizes, with open-source software, such as MZMine2 and XCMS, generating higher initial feature counts than vendor software (Agilent Profinder).

Chromatographic RT mapping images computed by *metabCombiner* are illustrated in **Figure 4.5** for both ionization modes. Three stepwise alignments were performed between primary dataset *I* and targets *II* (**Figure 4.5 A** and **D**), *III* (**Figure 4.5 B** and **E**), and *IV* (**Figure 4.5 C and F**). Generally, the early portions of the gradient are well-mapped based on the locations of the anchors with respect to the computed curves, whereas ordered pair selection and spline fitting appears to be less informative in the later chromatogram. Inspection of features in the later chromatogram of dataset *I* reveals mostly poorly resolved and excessively wide peaks resembling background ions, which is the likely cause of these observations.

Table 4.3 summarizes the total count of features found in one, two, three, or all four datasets in this analysis. Surprisingly, only 470 positive mode and 186 negative mode features were discovered in common across all four lipidomics assays. In the pairwise alignment, a total of 1208, 2090, and 2215 features were found in common between *I* & *II*, *I* & *III*, and I & IV, respectively, in the positive mode; in the negative mode, the totals come to 603, 1219, and 1537

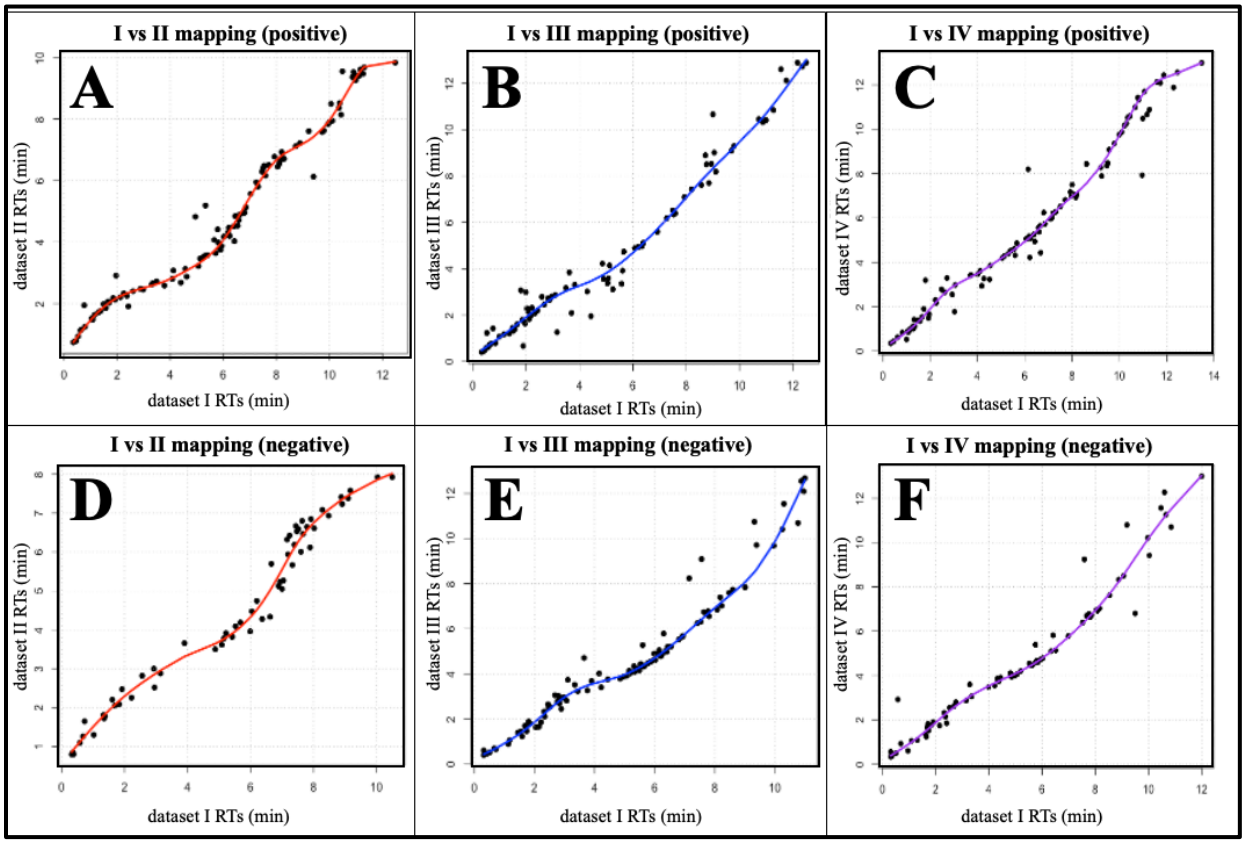


Figure 4.5 RT Projection Model Fits for Inter-laboratory Lipidomics Study Pairwise mappings in both ionization modes for I vs II (A & D), I vs III (B & E), and I vs IV (C & F).

matched features. A closer inspection reveals that 60% or more features in each of the target datasets lacked a counterpart within close m/z proximity to primary dataset *I*, which is a major contributor to the lower than anticipated intersection rate. Differences in experimental and pre-processing parameters, imbalances in the separation of isomers and high m/z or RT fitting errors for some compounds may also contribute to these findings.

Comparisons of annotations for aligned features revealed mostly commonalities with very few mismatched identities. **Figure 4.6** shows Venn diagrams of shared and unique lipid identifications for positive and negative mode features. The list of 55 aligned positive mode features with identity agreement across all datasets includes 16 Phosphatidylcholines (PC), 5 triacylglycerides (TAG), 12 sphingomyelins (SM), and 6 phosphoethanolamines (PE) in the positive mode; 11 out of 16 identically named aligned features in the negative mode were PEs.

Table 4.3 Inter-laboratory Lipidomics Study Aligned Feature Summary								
	Positive				Negative			
Detected across 4 Datasets	470				186			
Detected in I, II, & III	148				39			
Detected in I, II, and IV	335				140			
Detected in I, III, and IV	507				503			
Detected in I and II	255				238			
Detected in I and III	965				491			
Detected in I & IV	903				708			
Detected in I only	6802				4494			
	Total	**II**	**III**	**IV**	**Total**	**II**	**III**	**IV**
Detected Outside I	42570	2601	22815	17154	15131	1644	6089	7398

The four laboratories had varying annotation rates, with laboratory II assigning at higher rates (per total feature count) and IV identifying the fewest. Those found in only one target dataset include features with no alignments to dataset *I* features as well as features that are mapped to other unidentified features. Based on the alignment results, at least 159 positive mode and 70 negative mode unannotated dataset *I* features can be putatively assigned using identifications shared by two or more target datasets outside of *I*. In terms of mismatched identities, most were annotated as the same lipid class but with differing carbon numbers (negative mode) or they conflicted in assigning PC and PE lipids of similar mass (positive mode). For many features whose identities are mismatched between pairs of laboratory datasets, a consensus annotation could be imposed if two or three of the other laboratories came to the same conclusion.

4.4.4 Discussion

This study demonstrates that differences in analytical and computational parameters between laboratories drive substantial disparities in coverage of detected lipids. At most 5-10% of all primary dataset features could be found across the three target datasets, with most analytes

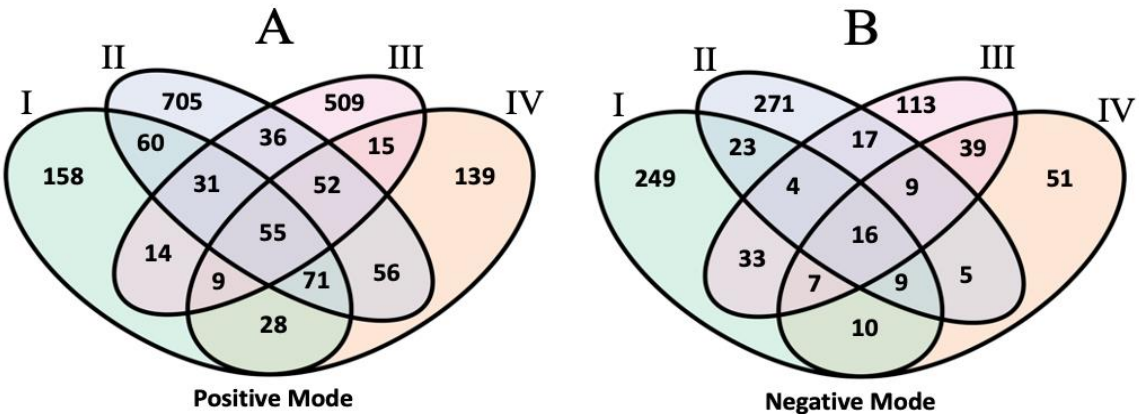


Figure 4.6 Venn Diagrams of Identified Lipids Intersections imply the numbers of features that are both aligned and named identically (class and double bond definition) in (A) positive and (B) negative mode datasets.

lacking any corresponding matches. Whether these non-aligned features are derived from unique lipids, non-biological artifacts or shared lipids that failed to meet *metabCombiner* cut-offs requires further investigation. Nevertheless, coupling multiple structure elucidation methods performed by multiple institutions and disparate LC-MS alignment with *metabCombiner* is a powerful strategy to improving identification rates in lipidomics studies. Merging inter-laboratory study data derives plentiful information, including the generation of hundreds of consensus lipid annotations, hypothesized identifications for unidentified features in both ionization modes, and a short list of aligned unknown features for further investigation.

4.5 Conclusion

Structure elucidation is the most significant obstacle to obtaining a full understanding of large-scale metabolic changes in organisms. Differences in experimental protocols between institutions and the relatively poor reproducibility of LC-MS magnify this problem by forming a critical barrier to information transferability between acquired datasets. In this chapter, the benefits of disparate LC-MS alignment are explored in three studies where computational and experimental approaches, such as MS/MS library searches, authentic compound standards, *in silico* databases, LC gradient and sample loading optimizations, and inter-laboratory

comparisons were applied with the aim of increasing metabolite identification rates in diverse specimens. In all three scenarios, *metabCombiner* mapped hundreds of identifications to unidentified features in a new dataset, with minimal duplication of exhaustive, time-consuming efforts. While comprehensive identification of all detectable metabolites remains a daunting challenge for the field, incorporating alignment for the discovery of common analytes acquired under non-identical conditions can efficiently determine "known unknowns" and prioritize true unknowns for additional follow-up experimental investigation.

Chapter 5

Applications of Disparate LC-MS Alignment to Bioinformatics Analysis

5.1 Introduction

For most metabolomics studies, the ultimate objective is to discover metabolites that distinguish samples from two or more experimental groups or populations, define novel diagnostic and prognostic biomarkers of specific phenotypes, and elucidate mechanisms underlying complex physiological events. Like other high-throughput -omics assays, metabolite profiling generates datasets containing thousands of measured variables, and therefore require sufficient sample sizes to perform well-powered statistical analyses.[222] For large-scale experiments involving hundreds or thousands of samples, the data must be acquired in multiple batches, preferably within a short period of time to minimize differences in analytical conditions between batch runs. In some cases, long time intervals may elapse between batch analyses or protocols may be altered between experimental subsets analyzed within one or multiple laboratories. Non-biological systematic and random variations are routinely observed in signal acquisition and chromatographic RT measurements in large-scale metabolomics studies;[178] this unwanted variability is compounded when the data are acquired under non-identical conditions. The consequences for measured RTs were discussed previously in section 3.1.

Overcoming signal intensity variation in the form of intra-batch, inter-batch, and inter-experiment effects is critical for obtaining accurate results from statistical and bioinformatics analyses applied to LC-MS metabolomics datasets. Signal drift is observed within LC-MS

batches due to changes in instrumental response throughout the course of a batch analysis resulting from the build-up of residues that cannot be fully removed during the detection process.[223] Additionally, inter-batch and inter-experiment effects may be caused by temperature changes, reagents lots, experimental operators, the condition of the LC column, differences in batch preparation, sample handling, and other latent environmental or technical factors.[224]

Common strategies for managing sample to sample variation within a batch typically involve nonlinear modeling of signal drift within quality control (QC) samples as a function of the order in which samples are analyzed. Quality controls have similar matrix compositions to biological samples as they are usually obtained by pooling aliquots from all subjects, and theoretically their feature abundance measurements should be the same regardless of run order or batch.[178] Modeling approaches used for these analyses include LOESS,[72] support vector regression,[73] and Random Forests,[225] after which experimental sample intensities are normalized to interpolated values predicted by the model. Inter-batch effects correction would then normalize to the median or mean QC sample values of each batch, bringing the values to a common scale. The principal drawback of these methods is that overfitting may occur if QCs do not sufficiently represent subject samples,[226] which may be observed for many metabolites.

Well-known batch effects removal methods designed for other high-throughput omics data have been adapted to metabolomics with reasonable results. Location-scale methods, such as centering, scaling, and quantile normalization, assume that metabolite abundances in different batches follow a similar distribution. The popular ComBat method for batch effects removal based on Empirical Bayes is a standard approach that has been used for many multi-batch metabolomics studies.[227] These methods are not recommended for use in multi-batch metabolomics studies in which sample groups are not evenly distributed between the batches.

Since they are not specifically designed for the intricacies of metabolomics data, such as variation due to sample order, they should be combined with methods for intra-batch corrections. Another class of methods uses matrix factorization approaches, such as Independent Component Analysis (ICA)[228] or singular vector decomposition.[229] A notable method called WaveICA performs Discrete Wavelet Transforms (DWT) to decompose sample metabolomics values (arranged by run order) into different frequencies based on biological and technical variability, followed by Independent Component Analysis to eliminate components associated with batch effects, then inverse DWT to return to the original scale.[223] This method is attuned to the specific design of metabolomics experiments and performs superior to most methods for batch normalization; one drawback is that the method assumes lower-frequency signal between samples is due to temporal drifts, whereas some biological variation may be represented (i.e. if metabolite sample groups are analyzed in a non-random order) and thus lost if the low frequencies are removed. Other emerging methods, such as TIGER[230] and NormAE,[231] use neural network architectures for batch effects removal. Batch effect removal approaches are typically designed for and applied to multi-batch experiments where practitioners attempt to replicate conditions as much as possible between batches. Since few tools have been previously developed for aligning disparately acquired data, problems associated with untargeted LC-MS metabolomics datasets merged from multiple experiments have not been fully explored.

This chapter describes two studies for which experimental sample subsets were obtained under non-identical conditions. In the first study, the plasma metabolome of volunteers with amyotrophic lateral sclerosis (ALS) was compared to healthy controls to identify significant metabolic dysregulation associated with the disease. Second, maternal plasma metabolite levels were measured at the first and third trimesters of gestation, as well as those of umbilical cord

blood. Data analysis for these studies outlines a procedure for extracting biological insights from the multiple merged experimental datasets, where strategies for normalization to overcome significant non-biological variation were addressed, followed by differential and partial correlation network analyses.

5.2 Metabolomics Identifies Dysregulation in Amyotrophic Lateral Sclerosis Cohorts

5.2.1 Background

Amyotrophic lateral sclerosis (ALS) is a fatal, progressive neurodegenerative disease,[232] with no known effective cure or treatments. Its pathogenesis is complex and influenced by genetic,[233] epigenetic,[234] and environmental[235,236] factors. Metabolomics provides a readout of metabolic dysregulation caused by internal and external factors, is useful for investigating complex diseases arising in ALS and other diseases. For example, studies of oxidized metabolites, such as nitric oxide, its toxic metabolite, peroxynitrite,[237] and oxidized lipids[238] in ALS patients identified oxidative stress as a disease characteristic. Multiple studies have employed untargeted metabolomics to identify metabolic differences between ALS versus control subjects, uncovering dysregulated pathways,[239–241] such as lipid,[240,242,243] amino acid,[240,244–246] and polyamine[244] metabolism. However, most previous studies were limited in sample size or metabolite count.

This section describes a study employing a commercial untargeted metabolomics platform to yield insights into ALS mechanisms.[247] Plasma metabolite profiles were obtained roughly one year apart from two independent cohorts consisting of ALS and healthy control subjects, with the goal of identifying consistent disease biomarkers and mechanisms between the cohorts. Measurements of metabolites common to both datasets were combined to perform differential network enrichment analysis (DNEA) with sufficient statistical power, uncovering

data-driven modules of correlated metabolites associated with the disease. Multiple data filtering and normalization steps were necessary to remove non-biological variation arising from latent experimental factors and confounding variables. This study serves to illustrate the benefits and challenges of analyzing metabolomics datasets merged from separate cohorts.

5.2.2 Methods

ALS patients older than 18 years were recruited at the University of Michigan Pranger ALS Clinic.[236,248] Control participants were also recruited through the University of Michigan Institute for Clinical and Health Research. Participants' age, sex, height, and weight were recorded, and ALS features were obtained from medical records. Participants' plasma was obtained through peripheral venipuncture, centrifuged at 2000*g* for 10 min at 4ºC, aliquoted into cryovials, and stored at -80 ºC. Metabolomics profiling of plasma samples was performed by Metabolon (Durham, North Carolina), using multiple semi-untargeted LC-MS/MS assays.[249] Metabolites were extracted and analyzed by RPLC, in both negative and positive ion modes, and HILIC. Metabolites were identified by Metabolon by comparing retention time/index, m/z, and fragmentation spectra with authentic standards. Importantly, only named metabolites are provided in data reports, with no m/z, RT, or unknown compound information that could be used for disparate LC-MS alignment analysis with *metabCombiner*.

For this analysis, only controls and case subjects whose diagnoses are listed as ALS or ALS/FTD (frontotemporal dementia) are included for analysis. This retains 196 subjects from cohort 1 (71 controls, 125 cases) and 328 subjects from cohort 2 (104 controls, 224 cases). For ALS patients, only the plasma sample from the initial visit is included, excluding additional measurements for subjects who participated in follow-up visits. Alignment between datasets is achieved trivially by matching features from the two cohort datasets based on matching

compound identifiers, as defined by Metabolon. A total of 954 compounds were found in common between the datasets, 101 of which were removed as drug-related metabolites not of interest to this workflow. Metabolites missing in over >20% of samples in one or both cohort datasets were excluded from further analysis, leaving a total of 640 remaining metabolites. Remaining missing abundance values were imputed with the cohort-specific minimum for the respective metabolites, followed by log-transformation of all quantities.

Inter-cohort batch effects were assessed using principal components analysis (PCA). For each dataset separately, metabolites were linearly adjusted for age, sex, and BMI. Fifteen missing BMI values were imputed using linear models based on the top 13 and 23 metabolites from cohort 1 and cohort 2, respectively, which correlated most strongly with BMI. After adjustment, metabolomics values from the two cohorts were separately Z-scaled (mean centered and scaled by the standard deviation) to obtain $N(0,1)$ distributions for each metabolite. The adjusted and normalized datasets from the two original and replication cohorts were then re-assembled into the merged dataset.

DNEA was employed to identify metabolite subnetworks that differentiate ALS from control samples.[129,130] Due to an imbalance in the number of samples in ALS versus control groups, a subsampling procedure coupled with PCN estimation to obtain robust network edges was used, as previously described.[130] The network was then clustered into densely-connected metabolite subnetworks, followed by enrichment analysis using NetGSA[250] which takes into account differential metabolite abundances and differences in network structure between cases and controls. The results of DNEA consist of subnetworks and their respective p-values and false discovery rate (FDR)-adjusted q-values, corresponding to the significance of subnetwork changes between ALS versus control groups.

5.2.3 Results

Figure 5.1 illustrates plots of the combined dataset projected onto the first two principal components before and after normalization steps. Panel A shows two homogeneous clouds consisting of sample observations from each cohort separately, indicating substantial inter-experiment variability. This technical variation is due to latent experimental factors that were intentionally or unintentionally changed in the interval between data acquisitions. Without additional normalization steps, non-biological batch effects would obscure differences in metabolite levels between sample groups. Following covariate adjustment and Z-scaling, PCA on the transformed dataset appears to show that the cohort-specific differences are successfully resolved, with samples from both cohorts mixed in a non-uniform manner.

Differential Network Enrichment Analysis of the merged and normalized multi-cohort dataset identified metabolite subnetworks that differentiate ALS from control samples. A total of 15 different metabolite subnetworks, 9 of which were significantly enriched at the 0.01 FDR cutoff. Xenobiotics-related pathways, such as “benzoate metabolism”, “food component/plant metabolism”, and “xanthine metabolism”, contributed the greatest number of metabolites to the

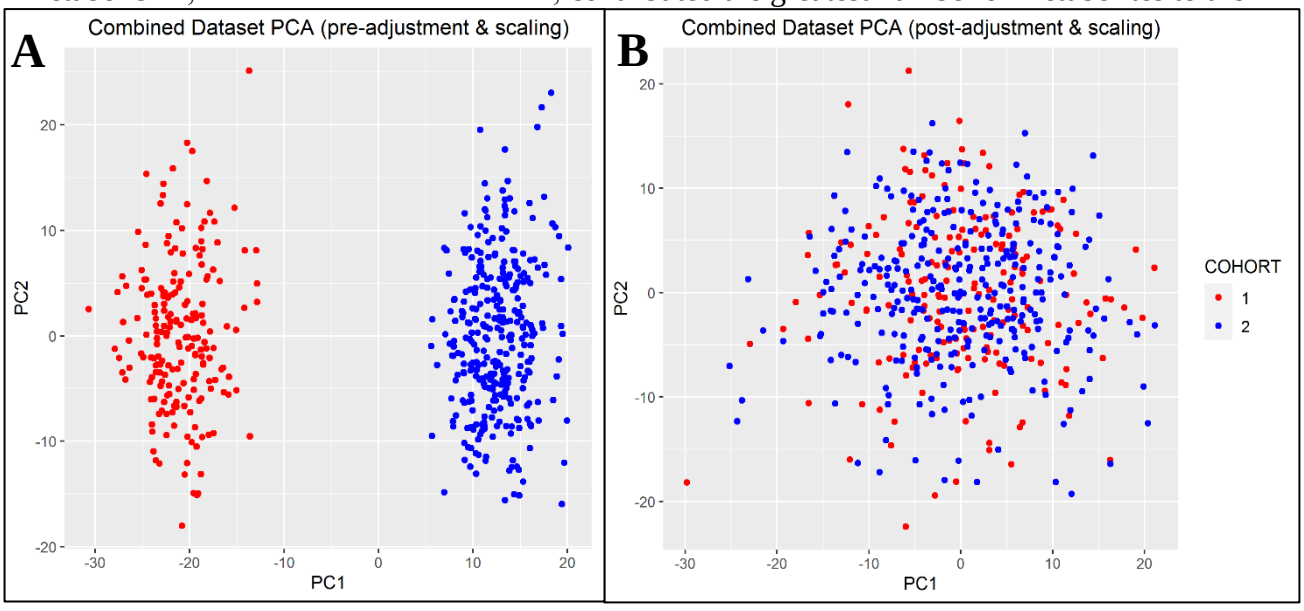


Figure 5.1 Combined ALS Dataset PCA Plots (A) Pre-normalization (B) Post-Normalization

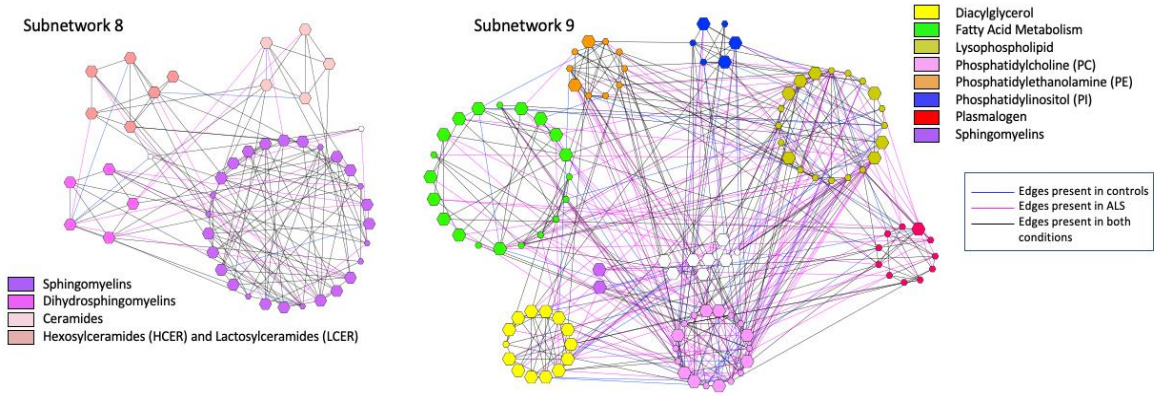


Figure 5.2 Overview of Lipid Subnetworks Nodes are colored according to the sub-pathways, as shown in the figure legends. Node sizes correspond to the directional changes of metabolites in ALS compared to controls.

two most significant subnetworks. Energy-related metabolites, including TCA cycle, amino acid metabolism, and purine/pyrimidine metabolism, featured in subnetwork 7. It also features creatine and its breakdown product creatinine, the two most significantly altered metabolites between cases and controls, with creatine elevated in ALS subjects at the expense of creatinine. Subnetworks S8 and S9, depicted in **Figure 5.2**, encompass multiple complex lipid species and reveal important sources of metabolic dysregulation, including impairment in sphingomyelin metabolism; increases in ceramides and glucosylceramides; higher fatty acid levels; and decreases in various phospholipids, including phosphoethanolamines and phosphotidylcholines, both of which have roles in signaling, membrane formation, and mitochondrial function. These highlight the continued need for research into the complex role played by lipids in ALS pathogenesis and progression. Other subnetworks are described in Goutman et al (2022).[247]

5.2.4 Discussion

A workflow for the merging, normalization, and meta-analysis of two ALS and control patient cohorts is performed for extracting deeper biological insights from the combined sample set over individual cohort datasets. The framework involves simple data correction approaches applied to the two datasets individually, followed by Z-transformation and re-assembly, to obtain

a normalized dataset amenable to statistical approaches. The increased sample size and reduction of metabolite count provided a manageable sample to metabolite ratio, amenable to DNEA. The results of DNEA show substantial changes in lipid metabolism, xenobiotic, primary metabolism, and creatine/creatinine along with other large-scale metabolic changes associated with ALS. Limitations of this analysis include the inability to assess the effect of sample order or the m/z and RTs of unknown compounds as this information was not available; a noted sex imbalance between the two cohorts, featuring disproportionately more female controls in the second cohort; and non-fasting ALS individuals may have introduced confounding factors that cannot be accounted for. Overall, this study shows that merging metabolomics datasets could shed additional light on clinical applications, provided that appropriate measures are taken to overcome sources of technical variability from inter-cohort effects.

5.3 Alignment and Analysis of a Disparately Acquired Multi-Batch Metabolomics Study of Maternal Pregnancy Samples

5.3.1 Background

Major metabolic changes occur throughout pregnancy to enable mothers adequate nutrients to support infant development, affecting the long-term health of the newborn.[251] In recent years, profiling of small molecular weight compounds in a biological sample by metabolomics has been used to obtain an objective measurement of the metabolic environment to which the developing fetus is exposed. Metabolomics has been applied in multiple developmental studies to identify changes in nutrient availability across pregnancy.[252–255] For example, a targeted metabolomics study showed increases in long-chain fatty acids (FFA) and long-chain acylcarnitines (AC) in maternal plasma from the first trimester to term,[254] reflecting increases in lipolysis in late-gestation to fuel fetal growth. Maternal metabolite levels, placental transfer, and metabolite interactions can affect the relative levels of the umbilical cord blood

(CB) metabolome.[256,257] Using a lipidomics platform, changes in 573 lipid species, including phospholipid, ceramide (CER), cholesteryl ester (CE), and triglyceride (TG) levels, were recorded for 106 mother-infant pairs throughout pregnancy.[251] DNEA revealed fluctuating correlations among lipid groups and compound class-specific associations with infant birthweight at different time points.[130] An untargeted metabolomics study was performed on the first trimester maternal plasma (M1), delivery maternal plasma (M3), and umbilical cord blood (CB) for the same mother-infant pairs as the lipidomics. A subset of these samples was analyzed in 2016, whereas the remaining data was acquired by the same laboratory in 2019, but with a different chromatography system, mass spectrometer, and experimental protocol, introducing many challenges to joint analysis of the full sample set. Here, a framework is presented for enabling disparately acquired multi-batch data analysis which consists of alignment steps to assemble the two experimental subsets into a single table and normalization steps to correct for significant inter-experiment and inter-batch variation. Subsequent analysis of the aligned and normalized data demonstrated enhanced statistical power to uncover significant changes in maternal plasma metabolome between the first trimester (M1) and third trimester (M3), and between maternal plasma (M3) and infant umbilical cord plasma (CB).

5.3.2 Experimental Methods

Pregnant women were recruited at the first prenatal appointment to the Michigan Mother Infant Pairs (MMIP) cohort. Eligibility criteria for MMIP includes age between 18 and 42 years old, had a spontaneously conceived singleton pregnancy, and intended to deliver at the University of Michigan Hospital. A subset of mother-infant dyads was selected for untargeted metabolomics measures. The initial study visits occurred at 8-14 weeks gestation (M1), where participants provided blood samples. Women were recontacted prior to delivery and maternal

blood samples (M3) and umbilical cord blood samples (CB) were collected at delivery. Characteristics of the study population can be found in Table 1 of the previously published work on the lipidomics analysis.[1] Mothers were on average 32.1 years at baseline with an average BMI of 25.8 kg/m^2. The average birth weight was 3.51 kilograms with 51 male and 55 female infants.

Two different experimental protocols were performed in this study on distinct sample sets, referred to as *exp616* & *exp946,* with most important parameter differences listed in **Table 5.1**. *ex616* is composed of M1, M3, and CB timepoint samples form 56 mother-infant pairs, divided into three batches, whereas *ex946* is partitioned into 2 batches consisting of data from 48 individuals. Sample preparation procedures were largely the same for *ex616* & *ex946*. 100 µL of each plasma sample was combined with 400µL Methanol: Acetonitrile: Acetone (1:1:1) extraction solvent containing 20 µM of internal standards (L-15N-Anthranalic acid, L-$^{15}N_2$-Tryptophan, Gibberelic acid, and L-Epibrassinolide) in micro-centrifuge tubes. Samples were vortexed for 5 minutes and centrifuged for 10 min at 15000 rpm. 250 µL supernatant was transferred to another clean vial, dried under N2 stream, and reconstituted with addition of 100 µL Methanol: H2O (2:98) containing Zeatin. QCs were prepared differently between experiments, with batch-specific pools employed for ex616 versus master pooled aliquots of all samples in *ex946*. Importantly, sample run order was not fully randomized across *ex616* as M1

Table 5.1 Major Experimental Parameter Differences between MMIP Subsets

Experiment	**ex616**				**ex946**		
Year of Acquisition	2016				2019		
Number of Subjects (M1, M3, CB samples)	56				48		
Batch Counts	3				2		
Pooled QCs per Batch	10				14		
Pooled QC type	batch-specific pools				full-experiment pools		
Gradient (min)	0-2	2-20	20-22	22-30	1-16	16-20	20
%solvent B	2%	2 to 75%	75 to 98%	98%	0-99%	99%	1%
Mass Spectrometer	Agilent QTOF 6530				Agilent QTOF 6545		
Ion-source gas temp	325				350		
nebulizer pressure	45 psi				30 psi		
sheath gas temperature	400ºC				350ºC		
sheath gas flow	12L/min				10L/min		
capillary voltage	4000V				3000V		

samples were run as a group towards the beginning or the end of each batch, separately from M3 and CB samples; by contrast, sample order arrangement was appropriately randomized in *ex946*.

For both experiments, a 1290 Infinity Binary LC system from Agilent is used for LC separation together with a Water Acquity UPLC HSS T3 1.8 µm x 100mm column. Mobile phase A was 100% water with 0.1% formic acid and mobile phase B was 100% methanol with 0.1% formic acid. The gradients and total chromatography times differ substantially between experiments, as listed in Table 1. An upgraded, more sensitive QTOF mass spectrometer was used for *ex946,* and several spectrometric parameters were altered, but in both experiments full-scan mass spectra were acquired over the range 50-1000 m/z with an acquisition rate 2 spectra/s and internal mass correction. Iterative Data Dependent (iDDA) MS/MS analysis was performed on pooled plasma in *ex946* only. iDDA captures MS/MS in stepwise fashion, with precursor ions excluded from MS/MS acquisition at the same RT during subsequent replicate runs. For our untargeted platforms, we collect 8 rounds of iDDA at 3 different collision energies (10, 20, and 40 eV). Analysis of iDDA spectra using *msPepSearch* and NIST20 spectral library was performed to provide MSI Level II & III identifications for statistically significant features.

5.3.3 Data Analysis Methods

The main computational workflow for this study is shown in **Figure 5.3**. Disparate LC-MS metabolomics data analysis must first correct for variations in retention times and acquired signal abundance values. The alignment steps merge the batch feature lists of each experiment into a single cohesive table. The normalization procedure consisted of data filtering, imputation, batch effects removal and scaling within each experimental set before re-assembly into the final matrix. Data from both experiments were first pre-processed using Agilent Profinder software creating separate feature tables for each batch in both ionization modes. Adducts, in-source

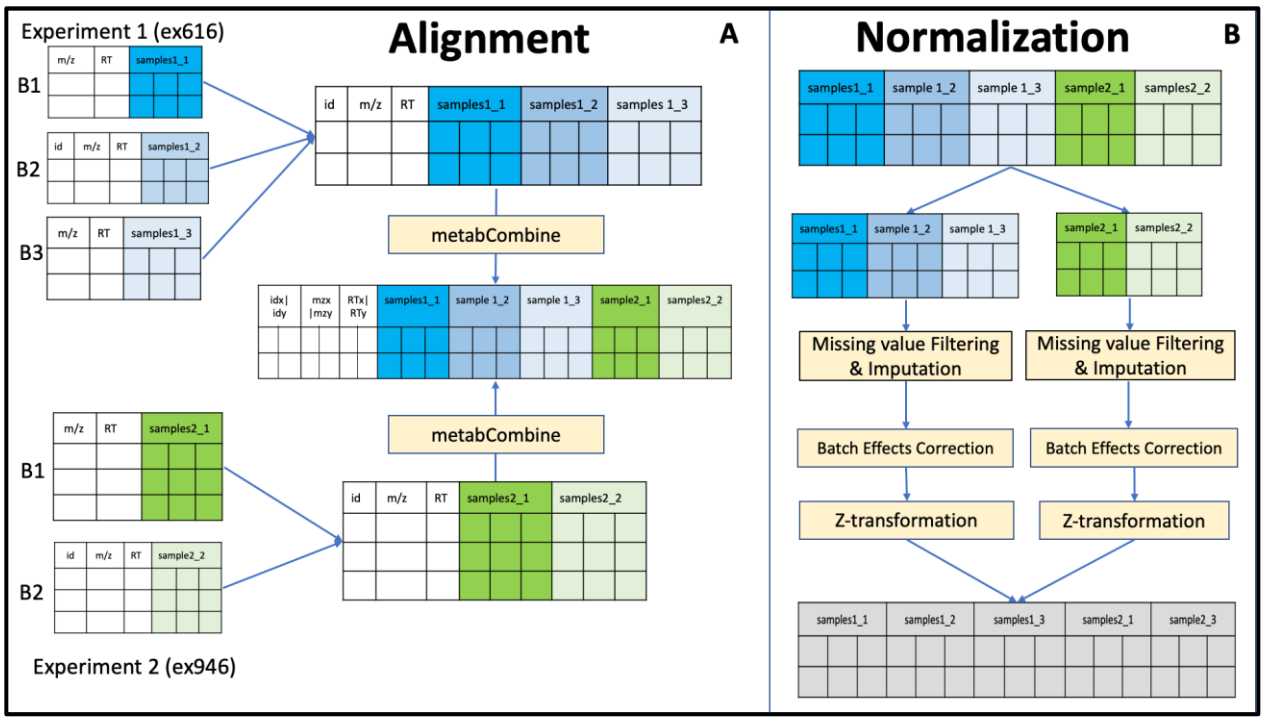


Figure 5.3 MMIP Study Analytical Workflow (A) alignment steps for merging batch and experimental data, (B) normalization steps for removing technical variation including within-experiment missing value handling, batch effects correction, Z-transformation, and re-assembly.

fragments, and multimer labels were annotated on a representative batch using *Binner* 1.1.0,[171] and known metabolites were annotated using in-house m/z and RT libraries obtained from authentic standards run on either chromatography system.

Alignment of metabolomics features between batches and experiments was performed using the *metabCombiner* R package.[184] Within *metabCombiner*, the quality control (QC) samples of each batch were selected for relative quantitation comparisons, with normal M1, M3, and CB samples designated as "extra" columns. The program filters features missing in over 50% of each set of batch QC samples before analysis. In each experiment separately, feature tables corresponding to each batch were aligned in an iterative and stepwise manner to construct batch-merged tables. Subsequently, features from the two non-identically acquired experimental tables were overlapped to construct a single table of sample measurements consisting of features detected in all batches from the two experiments. Any features missing from over 50% of all

experimental samples from either *ex616* or *ex946* were eliminated. For the remaining features, the data were log-transformed and missing data imputation applied to feature quantities from the two experiments separately, using random-forest based imputation implemented in the *MissForest* R package.[63]

To address the significant technical variation brought on by changes in instrumentation and other analytical factors between the experiments, separate approaches for batch effects correction were applied to the experimental subsets, attuned to their specific study designs. For *ex616*, which lacks uniform QC samples and contains non-randomness in its sample run order assignment, the QC-RLSC method corrects for intra-batch effects by normalizing to LOESS curves of feature values from batch QC samples for each individual batch[72]; then ComBat implemented in the *sva* R package is applied for the removal of inter-batch effects.[258] On the other hand, intra- & inter-batch effects are simultaneously handled in *ex946* with the WaveICA method.[223] Subsequently, abundance values in both experiments were Z-transformed, followed by re-assembly based on the previously determined feature matches. To determine the efficacy of this approach, Principal Components Analyses (PCA) and Principal Components Partial R-Square (PC-PR2)[259] analyses were performed. PC-PR2 method combines PCA with multivariate linear regressions to express a measure of the proportion of variability explained by variables, which are chosen to be experiment, batch, and sample type variables in this application.[259]

Following normalization and re-assembly of aligned datasets, annotations from pre-alignment steps were harmonized to obtain a list of known and unknown compounds for analysis. Metabolites were named when annotated in *ex616* or *ex946*. Annotated adducts, fragments, and features derived from internal standard compounds were removed at this stage.

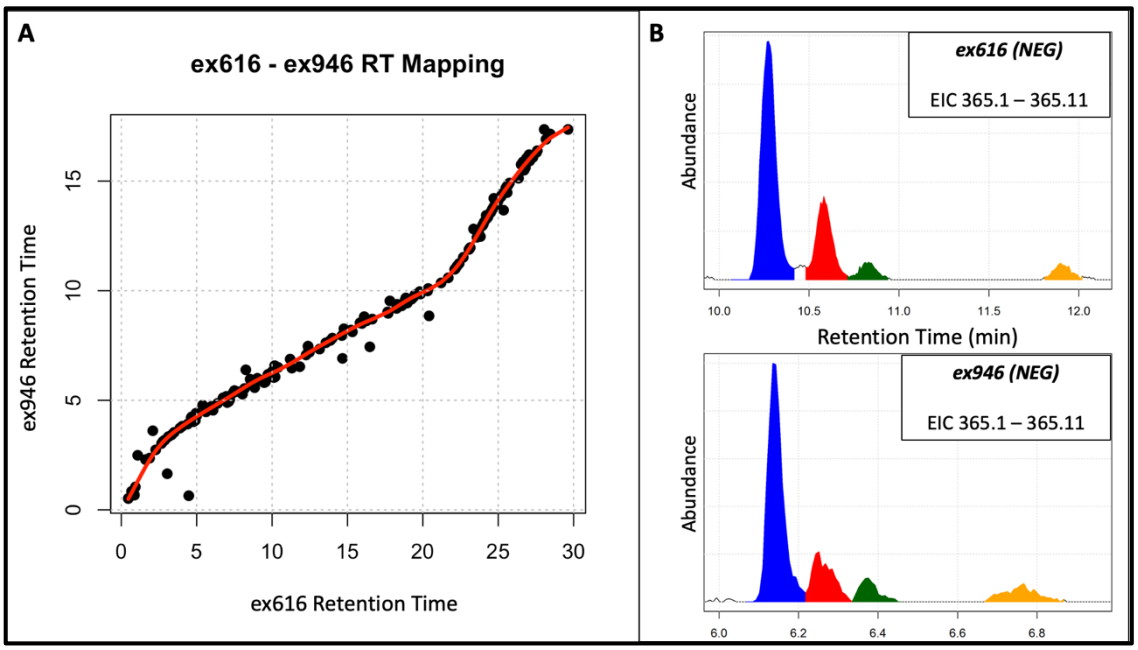


Figure 5.4 MMIP Study RT Mapping and Feature Matching (A) Plotted spline fit generated by *metabCombiner* mapping ex616 (30-minute total chromatography) RTs to ex946 (20-minute). (B) Selected EIC for the two experiments in m/z range 365.1-365.11 (negative mode), with matching colors for identical compounds, as assigned by *metabCombiner*.

Differential analyses were conducted to compare the metabolite levels of maternal plasma at the first and third trimesters (M1 & M3) and the umbilical cord blood (CB). Paired t-test statistics, p-values, and Bonferroni-adjusted p-values were obtained for the merged experimental data as well as *ex616* & *ex946* separately to determine the consistency of the differential features. Thresholds for nominal and statistical significance were set to 0.05 for both unadjusted and adjusted p-values. Partial correlation networks were generated with the Correlation Calculator program.[131] Due to sample size constraints, only annotated compound from aligned negative and positive mode datasets together were included. The resulting networks were visualized in Cytoscape[260] using Metscape.[120] Network nodes represent metabolites, while edges represent significant ($p_{adj} < 0.05$) partial correlations.

5.3.4 Alignment and Normalization Results

Experimental alignment consisted of within-batch pre-processing, merging between batches of the same experiment, and disparate alignment between experiments using

Table 5.2 MMIP Feature Counts from Alignment and Filtering Steps										
Ionization Mode	Positive					Negative				
Experiment	**ex616**			**ex946**		**ex616**			**ex946**	
Batch number	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**	**1**	**2**
Batch feature counts	8254	7474	7149	16216	13391	4830	4905	5135	9549	8773
Batch-merged feature counts	3971			10466		2616			5643	
Experiment-merged feature counts	2343					1583				
Filtered by Missingness	342					202				
Degenerate Features	512					327				
Annotated and Unannotated Features	1489					1054				
Annotated Features	199					129				

metabCombiner. RTs of *ex616* (30-minute run) were mapped to *ex946* (20-minute run) by selecting m/z and abundance quantile (Q)-matched ordered pair anchors through which basis splines curves are fit. RT mapping is shown in **Figure 5.4A**, alongside a visual confirmation of matched features for an example pair of EICs from both experiments (**Figure 5.4B**). Features were reduced in positive and negative ionization mode, beginning with the signal extraction and alignment within each batch to the final set of intersected analytes found across all experimental batches (**Table 5.2**). Metabolite alignment steps consisting of initial processing, inter-batch, and inter-experimental merging derived 2343 positive and 1583 negative mode features in common. These lists were subsequently reduced to 1489 and 1054, respectively, after applying missing value and degenerate feature filters.

Figure 5.5 illustrates the pre- and post-normalization of features projected onto the first two principal components, colored by experimental batch (**Figure 5.5A, 5.5C**) or sample type (**Figure 5.5B, 5.5D**). PC-PR2 bar plots represent the proportion of overall metabolomics variability explained by experiment, batch, and sample type, along with the total variation explained by these covariates (R^2) (**Figure 5.5E-F**). In the pre-normalized data, inter-experimental effects constitute the most significant source of variability, followed by sample type and within-experiment batch effects (**Figure 5.5E**). In the post-normalization data,

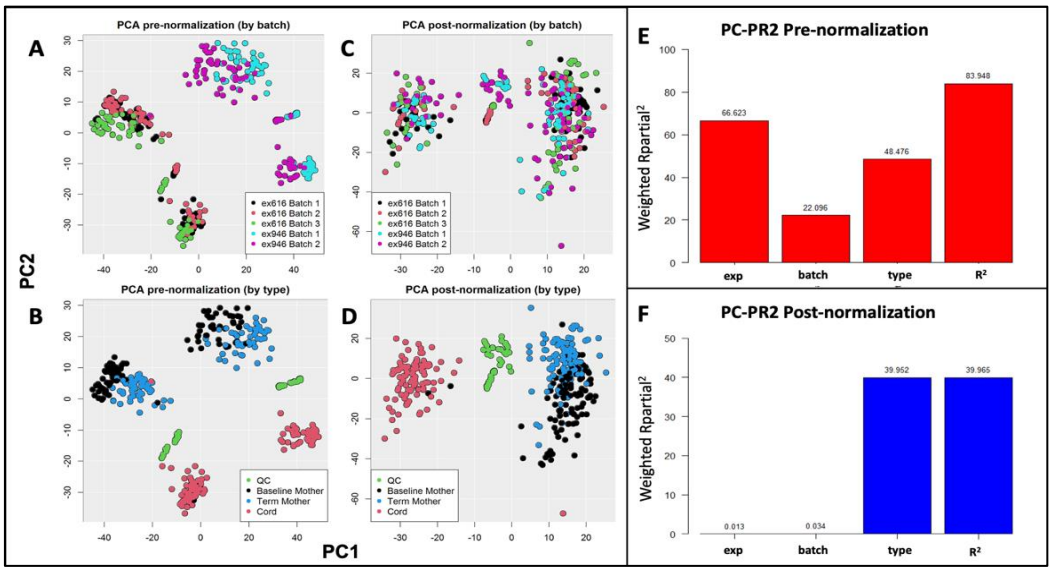


Figure 5.5 MMIP Study Pre- & Post-Normalization Plots (A-B) Pre-normalized aligned negative mode dataset projected onto the first two principal components, colored by experimental batch and sample type. (C-D) PC plot of the post-normalized dataset colored by (c) experimental batch and (D) sample type (E-F) PC-PR2 plots generated (E) before and (F) after normalization.

metabolomics variability arising from experiment and batch effects is eliminated, whereas variation due to sample types is largely preserved (**Figure 5.5F**). After normalization, separation is observed between the maternal (M1 and M3) and the CB metabolite levels (**Figure 5.5D**), indicating substantial metabolic differences between mother and infants. These visuals highlight the success of the normalization procedure in significantly reducing the influences of non-biological study variables, while retaining biological differences.

5.3.5 Bioinformatics Analysis Results

Using the aligned and normalized data, differential metabolites were identified between M1 and M3, which represents the change in the metabolome during gestation, and M3 and CB, representing the transfer of metabolites to support fetal development in late gestation and maternal-child metabolic differences. Comparing M1 and M3 samples across both experiments, 32% of aligned positive mode features and 47% of aligned negative mode features changed significantly (adjusted p-value < 0.05). Overall, 72% of the significant features increased from M1 to M3, demonstrating an increase of metabolite availability later in gestation to support fetal

growth.[261] Of the 968 significantly differential features in both ionization modes, 158 were annotated compounds. In a similar comparison M3 and CB timepoints, 64% of aligned positive mode features and 73% of aligned negative mode features exhibited significant changes (adjusted p-value < 0.05). Overall, 58% of the significant features were higher in CB compared to M3. Only 215 of the 1718 differential features were annotated compounds.

Differential analysis results were compared for the merged metabolomics dataset with those obtained from experiment *ex616* and *ex946* separately. There was an increase in the number of significant features in both M3 vs M1 and CB vs M3 for the merged metabolomics dataset vs. the experiments separately (**Table 5.3A**). This illustrates the advantage of increased sample sizes obtained by merging experimental data, even when acquired under disparate conditions. Among significant differential features found in common between *ex616* and *ex946*, the majority changed in the same direction, with 99-100% consistency in M1 vs. M3 and 96-97% consistency in CB vs M3 (**Table 5.3B**).

To further analyze observed metabolite changes, a partial correlation network of all known compounds was constructed (**Figure 5.6**) to enable visualization and biological interpretation of the data. Long chain free fatty acids (FFA) and lipid species containing long chain fatty acids, including diacylglycerides (DG), ceramides (CER), and sphingomyelins (SM)

A

Comparison	Ionization Mode	Number of significant features (% of total features)		
		ex616	ex946	combined dataset
M3 vs M1	Positive	403 (27%)	236 (16%)	471 (32%)
	Negative	464 (44%)	316 (30%)	497 (47%)
CB vs M3	Positive	922 (62%)	714 (48%)	947 (64%)
	Negative	728 (69%)	616 (58%)	771 (73%)

B

Comparison	Ionization Mode	Number of common significant features in ex616 and ex946	Number of significant features (% of total significant features)	
			Consistent Direction	Inconsistent Direction
M3 vs M1	Positive	163	163 (100%)	0 (0%)
	Negative	250	248 (99%)	2 (1%)
CB vs M3	Positive	568	545 (96%)	23 (4%)
	Negative	493	480 (97%)	13 (3%)

Table 5.3 MMIP Differential Analysis Summary (A) number of significantly differential features (Bonferroni adjusted p_{adj}< 0.05) for both timepoint comparisons (M3 vs M1 & CB vs M3) and ionization modes for the combined dataset and subsets *ex616* & *ex946*. (B) Comparison of differential results between experimental subsets.

largely increased from M1 to M3. The increase of lipid species containing long chain fatty acids from M1 to M3 is consistent with previous analyses by using a lipidomics platform,[251] suggesting the mobilization of FFA to support fetal brain development late in gestation.[262] On the other hand, 41% of phosphatidylcholine (PC) and phosphatidylethanolamine (PE) lipids decreased between M1 and M3, most of them polyunsaturated with very-long chains. Many medium and long chain acylcarnitines increased between M1 and M3 (AC 10:3, 11:1. 12:1, 14:0, 14:1, 16:2), consistent with a targeted analysis conducted for a subset of these women[254], indicating increased reliance on fat metabolism for energy later in gestation.[263] Decreases in essential amino acids were observed across pregnancy, including tryptophan and BCAA metabolites (ketoleucine, AC 4:0, and AC 5:1), consistent with other cohorts,[253] as protein anabolism is favored during pregnancy.[264] Cortisol and 11-deoxycortisol increased between M1 and M3, probably due to collection of M3 samples during parturition.[251] MS/MS library searches applied to the top 50 most significantly differential unnamed features between M1 and M3 in both ionization modes, using NIST Hybrid Search[200] and the NIST20 library. Top matches for

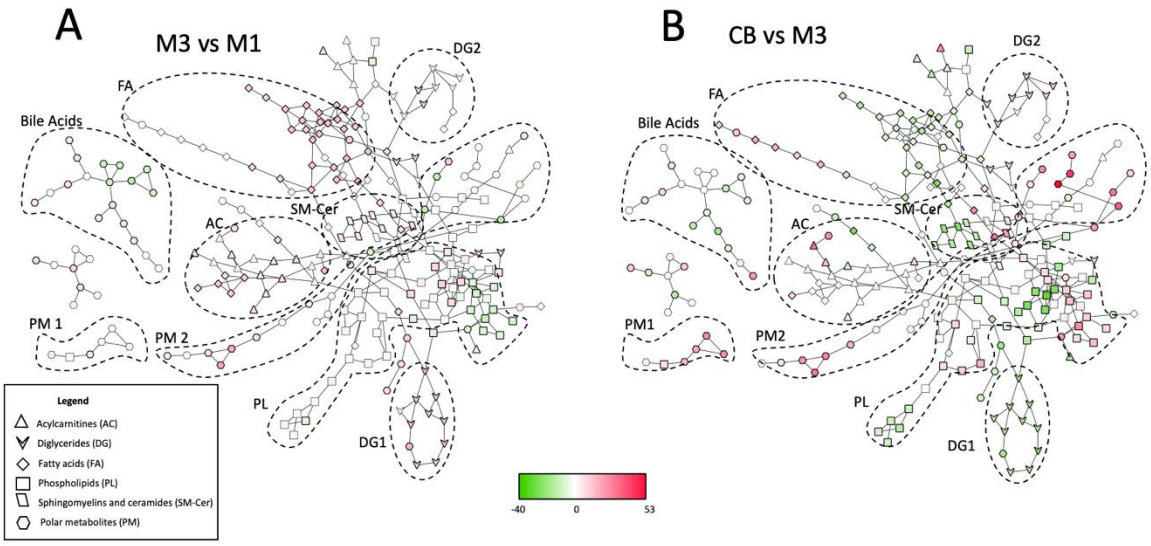


Figure 5.6 Partial Correlation Network Constructed from MMIP Metabolomics Data Nodes represent metabolites and edges represent computed partial correlations. Significant metabolites (q-value <0.05) have bold borders and node colors are based on t-statistics (A. M3 vs. M1. B. CB vs M3). The dotted lines outline subnetworks that include metabolites from different chemical classes.

most of these searches consisted of steroids and their associated sulfidated and glucuronidated modifications. The top feature (adjusted p = 3.3e-52) in the negative ionization mode, (m/z = 463.1967 Da, RT = 6.64 min in *ex946*) returns a match to estriol-16-glucoronide, a glucosiduronic acid metabolite of estriol that has been previously isolated in the urine and amniotic fluid of mothers during pregnancy.[265,266] Other matches include epiandrosterone sulfate, testosterone glucoronide, 7α,17α-dimethyl-5β-androstane-3α,17β-diol glucuronide, and 5α-pregnane-3α,17α-diol-20-one 3-sulfate. Increases in steroids were previously demonstrated during pregnancy to promote fetal and placenta development and ultimately parturition.[267]

Limited trends were observed from phospholipids, lysophospholipids, and diacylglycerols, apart from several long-chain monounsaturated and polyunsaturated PCs and PEs being elevated in CB. Cord blood levels of amino acids were consistently higher than M3, including branched chain amino acid metabolites, which is consistent with previous studies[268] and with high levels of branched-chain aminotranferases in human placenta tissue.[269]. Methionine, a major contributor to one carbon metabolism and DNA methylation, is higher in CB. Fluctuations in bile acids were observed with lower levels in CB of glycochenodeoxycholate (primary bile acid), deoxycholate, glycodeoxycholate, glycoursodeoxycholate, ursodeoxycholate (secondary bile acids) and higher levels of CB of glycocholate alpha-muricholate (primary bile acids), and taurocholate and taurodeoxycholate (secondary bile acids). Hormonal differences were observed with higher levels of androsterone sulfate, 11-deoxycortisol, and cortisol in M3 and higher levels of dehydroepiandrosterone sulfate (DHEA-S), pregnenolone sulfate, 11 beta-hydroxyandrost-4-ene-3-17-dione, and 17 alpha-20-alpha-dihydroxypregn-4-en-3-one in CB.

5.3.6 Discussion

This study presents a framework for merging LC-MS metabolomics data and overcoming substantial technical variation due to changes in protocols, instrumentation, and analytical conditions. Strategies are outlined for the alignment of known and unknown features and normalization to remove inter-batch and inter-experiment variation between disparately acquired datasets. Multiple analytical factors differed between, *ex616* and *ex946*, two metabolite profiling experiments staged 3 years apart. First, a more sensitive QTOF instrument derived substantially more extracted features in *ex946*, hence most extracted signals inevitably lacked one-to-one matches within the smaller *ex616* table. Second, the shorter total chromatography time in *ex946* caused gaps of up to 13 minutes in measured RTs for identical compounds between experiments. Alignment of these extracted feature matrices required mapping between chromatographic RTs or bypassing RT comparisons altogether.[182] LC-MS metabolomics meta-analysis studies[270–272] and methods[273,274] typically require replicated experimental conditions or limit their scope to shared known compounds, whereas *metabCombiner* overcomes this barrier and enables the joint analysis of LC-MS data of known and unknown compounds.

To normalize the expanded dataset, conventional intra-batch and inter-batch effects correction approaches were applied to two untargeted metabolomics experimental sets acquired by the same laboratory through different protocols, followed by Z-transformation and re-assembly of the matched features. This framework requires a similarly balanced study design between experimental sets as Z-transform normalization assumes similar abundance distributions.[223] The composition of M1, M3, CB, and quality control samples is proportionally similar between the experiments, though differences in how QC samples were prepared (per-batch pooled plasma in *ex616* vs all *ex946* sample pooled plasma) led to the use of alternative batch effects correction approaches. Recently, a linear mixed model approach was presented for

reducing inter-study variation, demonstrating higher ICC values for shared quantified sample measurements among eight pooled targeted metabolomics studies.[275] Successfully adapting normalization approaches to untargeted metabolomics data pooled from disparate sources may require the consideration of additional factors beyond those in targeted metabolomics.

Differences in the maternal plasma metabolome (M3 vs. M1) were assessed between the maternal and infant plasma (CB vs M3). Given the analytical differences between *ex616* and *ex946*, it was important to assess whether merging would enhance differential signal between sample groups or dilute it. The results show that the number of significant results increased with sample count, and that 96-100% significant changes shared between the two experiments occurred in the same direction (increasing or decreasing). Changes that occur in inconsistent directions could be indicative of inaccurate alignment between non-identical analytes, improper normalization between or within batches, or quantitation inconsistencies arising in the experimental or computational pre-processing stages. Statistical analysis of the aligned dataset revealed significant metabolic alterations between M1, M3, and CB, including coordinated changes observed between the first and third trimesters of gestation as well as infant cord blood consistent with previous studies.

5.4 Conclusion

Pooling experimental data between study batches or distinct studies is an emerging area of metabolomics research with the goal of increasing sample sizes and, by extension, statistical power. The major obstacles to data merging are the high inter-experiment variations in chromatographic RTs and signal acquisition. In the two examples described in this chapter, a common workflow consisting of alignment, missing value handling, within-experiment batch effects removal, and Z-transformation steps was applied to overcome these obstacles. In the ALS

study, data merging was achieved through straightforward metabolite ID matching, whereas merging the mother-infant pairs experimental data required the disparate LC-MS alignment capabilities of *metabCombiner*. The normalization steps performed in each study reduced inter-experiment variation in measured signal intensities, enabling analyses of the full sample set using statistical and bioinformatics techniques. These studies have similar objectives to differentiate between the metabolomes of population groups or timepoints, and in both cases increases in significant findings were reported over analyses of individual subsets. These studies benefit from having similarly proportional sample group distributions among the merged experiments, a requirement for location-scale batch effects removal techniques. It remains to be seen whether this framework can enhance metabolomics studies with alternative objectives, such as determining statistical associations between metabolite levels and numeric phenotypic outcomes. In summary, these studies establish the viability and utility of disparately acquired metabolomics data analyses, which have the potential to unlock many opportunities for the field of metabolomics that would otherwise be inaccessible through the traditional requirement of replicated experimental conditions.

Chapter 6

General Conclusions and Future Perspectives

6.1 General Conclusions

This dissertation has focused on three specific challenges in the field of computational metabolomics - 1) the annotation and reduction of redundant ion species present in pre-processed data as distinct features, 2) the alignment of metabolomics measurements corresponding to identical compounds measured under non-identical chromatographic and instrumental conditions, and 3) the harmonization of merged pooled disparately acquired multi-experimental datasets. As previously highlighted, tackling these issues would significantly reduce data redundancy, improve compound identification rates, minimize technical variation in compound measurements, and generate expanded feature tables with increased sample sizes for increased statistical power. To address the first two computational challenges, two new programs were developed, namely *Binner* and *metabCombiner*. The third challenge was addressed using a workflow of alignment and normalization steps designed for merged experiments based on study-specific properties. This chapter briefly summarizes the novel features of these new tools and the main findings of their applications in compound identification and bioinformatics.

6.1.1 Deep Annotation and Reduction of Untargeted Metabolomics Data with *Binner*

The standalone application *Binner* was developed to address the problem of data redundancy characterized by the multiplicity of detected ions derived from a common metabolite, as described in chapter 2. *Binner* takes pre-processed metabolomics features and

performs retention time binning, pairwise correlation hierarchical clustering, and annotation of isotopologues, adducts, fragments, and complexes based on mass relationships centered around principal ions. The program computes a distribution of common pairwise mass differences among binned features to facilitate the discovery of frequent mass relationships unexplained by user-supplied chemical addition or loss groups. *Binner* outperformed three existing tools in terms of adduct and fragment annotation accuracy, with comparable isotopologue annotation and metabolite ion grouping performance to CAMERA. With *Binner*, metabolomics datasets can be substantially reduced to the set of features that represent unique metabolites, eliminating potential false positive hits in statistical tests. Simultaneously, *Binner* calculates the underlying neutral masses, narrowing the list of possible metabolite identifications. Finally, annotated chemical modifications can be useful for validating correspondences between features in disparate LC-MS alignment results.

6.1.2 *metabCombiner*: Alignment of Disparately Acquired Metabolomics Datasets

metabCombiner is a software package implementing a method for aligning metabolomics features detected and measured in biologically similar specimens by non-identical LC-MS assays. The workflow steps are individual dataset filtering and formatting, grouping possible complementary feature matches by m/z proximity, selection of feature pair anchors based on high relative abundance or shared identities, retention time projection through spline-fitting, pairwise similarity scoring penalizing m/z, projected RT, and abundance quantile (Q) differences, and reduction to one-to-one feature matches. The method can be extended to multiple experimental results tables, including batches of a single experiment and inter-laboratory studies, concatenating abundance values of matched features found in each constituent dataset to generate a unified table with an augmented sample size. *metabCombiner* serves as the

cornerstone for disparate LC-MS metabolomics data analysis, making it possible to pool data from multiple experiments or studies for meta-analysis and enabling feature information transfer for known and unknown compounds.

6.1.3 Applications of Disparate LC-MS Metabolomics Data Analysis

Five different applications of disparate LC-MS data alignment were explored in this dissertation, the first three aimed to improve metabolite annotation rates and the latter two to enhance statistical power in biomedical studies. In each of the first three applications, experimental and computational approaches, such as MS/MS library searching, the use of authentic standards, *in silico* database searching, elongated LC gradients, increased sample loading, and in-house laboratory methods were applied to one or multiple experimental datasets, exhaustively characterizing as many features as possible. Well-characterized target dataset(s) were then aligned with data generated in the reference laboratory using *metabCombiner*, transferring hundreds of putative compound identities with considerably less time and effort. In the latter two applications, a common framework for correcting inter-batch and inter-experimental variation in RTs and acquired signal intensities was applied in two studies where subsets of samples were separately analyzed at least one year apart. The datasets were merged, normalized, and subjected to univariate differential and partial correlation network analyses, demonstrating more statistically significant results than with separate experimental subsets alone. These applications demonstrate the considerable potential of disparate LC-MS alignment and data analysis to augment chemical and biological knowledge, with many untapped opportunities from the vast expanse of publicly available metabolomics data.

6.2 Future Perspectives

The methods described in the preceding chapters for feature table reduction, dataset merging, and inter-experimental normalization represent important contributions to the field of computational metabolomics. They were designed to be versatile, accounting for specific properties observed in a wide range of LC-MS metabolomics datasets, and with numerous options for obtaining the most accurate results depending on data-specific properties and user preferences. However, there is considerable room for improvement. Here are some ideas that could be explored to improve the quality and performance of these methods.

6.2.1 *Binner*

One potential area of improvement for Binner is in the RT binning step, which currently cannot operate on tightly packed chromatographic regions without reducing the bin gap to a value near zero or manually breaking up the feature table into separate files for analysis. A heuristic method for accommodating excessively large bins is needed for the tool to handle the challenge of increasingly sensitive mass spectrometers and datasets with insufficient separation between features. Another area of improvement is the annotation method, which lacks any restrictions in terms combining neutral mass multiplicity, charge carriers, and neutral gain or loss groups, leading to many questionable annotations based on coincidental mass relationships. A procedure for annotating M+2 isotopologues without an M+1 prerequisite (e.g. ^{37}Cl or ^{34}S isotopes) should be added to the tool. Addition of prior information, such as compound identities, ion charge states, fragmentation information or RT range restrictions can prove useful for further improvements of annotation accuracy and reducing inaccurate annotations.

6.2.2 *metabCombiner*

Two drawbacks in *metabCombiner* in its current implementation are its detachment from the raw signal and the dependence of its feature matching on specific weights for the three parameters (m/z, RT, Q distances). While taking pre-processed tables rather than raw files simplifies the process by allowing conventional pre-processing to handle within-experiment alignment, the program cannot discern the quality of extracted peaks, often leading to problems in determining one-to-one alignments between incongruent spectra. Currently, no uncertainty estimates are provided for calculated scores or predicted RTs, and consequently thresholds for accepting feature pair alignments are difficult to set for balancing true and false matches. The current *metabCombiner* workflow supports the alignment of two datasets at a time, which precludes correspondence determinations for features absent from the chosen projection (X) or reference (Y) datasets when applying stepwise alignment procedure. Overcoming this requires an appropriate imputation procedure for missing features or otherwise re-designing the workflow to handle multiple dataset alignments simultaneously as opposed to dataset pairs.

6.2.3 Applications of Disparately Aligned LC-MS Metabolomics Data Analysis

For metabolite identification studies, similarity scores calculated from m/z, RT, and Q differences cannot be considered sufficient evidence of a confident compound identity. Incorporation of additional information into the algorithm, such as fragmentation spectra, ion mobility collisional cross sections (CCS), isotopic envelopes, or chromatographic shapes may be useful for improving confidence in assigned alignments. It is important to perform experimental validation of compound annotations, even on a limited scale, to assess the efficacy of information transfer.

The biomedical studies discussed in this work were similar in that they sought to differentiate metabolite levels between groups of samples, and their respective experimental

subsets were analyzed within one institution with appropriately balanced sample group proportions. Metabolomics applications involving data from multiple studies, acquired in separate institutions, using distinct experimental or study designs, measured within subjects drawn from dissimilar populations, or planned for statistical investigations other than differential analysis may require a different strategy to the ones used in this work. Linear mixed modeling approaches may provide an alternative framework for studies in which location-scale methods cannot be applied. Research into computational methods for harmonizing untargeted metabolomics data pooled from multiple experiments or studies is still in its infancy and must evolve with applications of disparate LC-MS data analysis.

6.3 Final Words

Metabolomics is a unique branch in the study of high throughput molecular phenotyping with its own opportunities and challenges distinct from the more established genomics and proteomics fields. As analytical technologies continue to improve their molecular detection capabilities and databases cataloguing known and theoretical metabolites grow alongside public data repositories, the field is well-positioned to enrich the collective functional understanding of living organisms. The way forward for metabolomics must consist of efforts to close the gap between the detected and known metabolome, to discern unique sample-derived metabolites from experimental and computational artifacts, and to harmonize data from incongruent metabolite profiling assays. This will help generate reproducible conclusions and maximize the use of available datasets, tapping into the field's unrealized potential. The work described in this dissertation provides important contributions towards achieving this goal, with the hope that they will be find wide-ranging applications in metabolomics and related scientific disciplines.

References

1. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* **13**, 263–269 (2012).
2. Tan, S. Z., Begley, P., Mullard, G., Hollywood, K. A. & Bishop, P. N. Introduction to metabolomics and its applications in ophthalmology. *Eye* **30**, 773–783 (2016).
3. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform* **8**, 61 (2016).
4. Nicholson, J. K., Lindon, J. C. & Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**, 1181–1189 (1999).
5. Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb Mol Case Stud* **1**, a000588 (2015).
6. Walker, D. I. *et al.* The metabolome: A key measure for exposome research in epidemiology. *Curr Epidemiol Rep* **6**, 93–103 (2019).
7. Sajed, T. *et al.* ECMDB 2.0: A richer resource for understanding the biochemistry of *E. coli*. *Nucleic Acids Res* **44**, D495–D501 (2016).
8. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research* **46**, D608–D617 (2018).
9. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends in Plant Science* **24**, 83–98 (2019).

10. Emwas, A.-H. M., Salek, R. M., Griffin, J. L. & Merzaban, J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics* **9**, 1048–1072 (2013).
11. Olivo, H. F. A Complete Introduction to Modern NMR Spectroscopy By Roger S. Macomber. Wiley Interscience, New York, NY. 1998. xvii + 382 pp. 21.5 × 28 cm. ISBN 0-471-15736-8. $49.95. *J. Med. Chem.* **41**, 3758–3758 (1998).
12. Zhou, B., Xiao, J. F., Tuli, L. & Ressom, H. W. LC-MS-based metabolomics. *Mol. BioSyst.* **8**, 470–481 (2012).
13. Fiehn, O. Metabolomics by Gas Chromatography–Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Current Protocols in Molecular Biology* **114**, (2016).
14. Gika, H. G., Wilson, I. D. & Theodoridis, G. A. LC–MS-based holistic metabolic profiling. Problems, limitations, advantages, and future perspectives. *Journal of Chromatography B* **966**, 1–6 (2014).
15. Theodoridis, G. A., Gika, H. G. & Wilson, I. D. *Metabolic profiling: methods and protocols*. (2018).
16. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted Metabolomics. *Current Protocols in Molecular Biology* **98**, (2012).
17. Scholefield, M. *et al.* Substantively Lowered Levels of Pantothenic Acid (Vitamin B5) in Several Regions of the Human Brain in Parkinson's Disease Dementia. *Metabolites* **11**, 569 (2021).

18. Wu, T. *et al.* Serum Bile Acid Profiles Improve Clinical Prediction of Nonalcoholic Fatty Liver in T2DM patients. *J Proteome Res* **20**, 3814–3825 (2021).
19. Wolfender, J.-L., Nuzillard, J.-M., van der Hooft, J. J. J., Renault, J.-H. & Bertrand, S. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography-High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics. *Anal Chem* **91**, 704–742 (2019).
20. Sas, K. M., Karnovsky, A., Michailidis, G. & Pennathur, S. Metabolomics and diabetes: analytical and computational approaches. *Diabetes* **64**, 718–732 (2015).
21. Patti, G. J. Separation strategies for untargeted metabolomics. *J Sep Sci* **34**, 3460–3469 (2011).
22. Barnes, S. *et al.* Training in metabolomics research. I. Designing the experiment, collecting and extracting samples and generating metabolomics data: Design and execution of a metabolomics experiment. *J. Mass Spectrom.* **51**, 461–475 (2016).
23. Vuckovic, D. Sample Preparation in Global Metabolomics of Biological Fluids and Tissues. in *Proteomic and Metabolomic Approaches to Biomarker Discovery* 51–75 (Elsevier, 2013). doi:10.1016/B978-0-12-394446-7.00004-2.
24. Worsfold, P. J., Townshend, A. & Poole, C. F. *Encyclopedia of analytical science*. (Elsevier, 2010).
25. Bird, I. M. High performance liquid chromatography: principles and clinical applications. *BMJ* **299**, 783–787 (1989).

26. Lv, W., Shi, X., Wang, S. & Xu, G. Multidimensional liquid chromatography-mass spectrometry for metabolomic and lipidomic analyses. *TrAC Trends in Analytical Chemistry* **120**, 115302 (2019).

27. Perez de Souza, L., Alseekh, S., Scossa, F. & Fernie, A. R. Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research. *Nat Methods* **18**, 733–746 (2021).

28. Wang, L., Wei, W., Xia, Z., Jie, X. & Xia, Z. Z. Recent advances in materials for stationary phases of mixed-mode high-performance liquid chromatography. *TrAC Trends in Analytical Chemistry* **80**, 495–506 (2016).

29. Berthelette, K. D. *et al.* Evaluating MISER chromatography as a tool for characterizing HILIC column equilibration. *J Chromatogr A* **1619**, 460931 (2020).

30. Abate-Pella, D. *et al.* Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods. *Journal of Chromatography A* **1412**, 43–51 (2015).

31. Baker, J. K. & Ma, C.-Y. Retention index scale for liquid—liquid chromatography. *Journal of Chromatography A* **169**, 107–115 (1979).

32. Hill, D. W., Kelley, T. R., Langner, K. J. & Miller, K. W. Determination of mycotoxins by gradient high-performance liquid chromatography using an alkylphenone retention index system. *Anal. Chem.* **56**, 2576–2579 (1984).

33. Bogusz, M. & Aderjan, R. Corrected Retention Indices in HPLC: Their Use for the Identification of Acidic and Neutral Drugs. *Journal of Analytical Toxicology* **12**, 67–72 (1988).

34. Griffiths, J. A Brief History of Mass Spectrometry. *Anal. Chem.* **80**, 5678–5683 (2008).

35. Dunn, W. B. Chapter two - Mass Spectrometry in Systems Biology: An Introduction. in *Methods in Enzymology* (eds. Jameson, D., Verma, M. & Westerhoff, H. V.) vol. 500 15–35 (Academic Press, 2011).

36. Clarke, W. Chapter 1 - Mass spectrometry in the clinical laboratory: determining the need and avoiding pitfalls. in *Mass Spectrometry for the Clinical Laboratory* (eds. Nair, H. & Clarke, W.) 1–15 (Academic Press, 2017). doi:10.1016/B978-0-12-800871-3.00001-8.

37. Iribarne, J. V. On the evaporation of small ions from charged droplets. *J. Chem. Phys.* **64**, 2287 (1976).

38. Felitsyn, N., Peschke, M. & Kebarle, P. Origin and number of charges observed on multiply-protonated native proteins produced by ESI. *International Journal of Mass Spectrometry* **219**, 39–62 (2002).

39. Zhou, W., Yang, S. & Wang, P. G. Matrix effects and application of matrix effect factor. *Bioanalysis* **9**, 1839–1844 (2017).

40. Nordström, A., Want, E., Northen, T., Lehtiö, J. & Siuzdak, G. Multiple Ionization Mass Spectrometry Strategy Used To Reveal the Complexity of Metabolomics. *Anal. Chem.* **80**, 421–429 (2008).

41. Stewart, D. *et al.* Chapter 4 - Omics Technologies Used in Systems Biology. in *Systems Biology in Toxicology and Environmental Health* (ed. Fry, R. C.) 57–83 (Academic Press, 2015). doi:10.1016/B978-0-12-801564-3.00004-3.

42. Zubarev, R. A. & Makarov, A. Orbitrap Mass Spectrometry. *Anal. Chem.* **85**, 5288–5296 (2013).

43. Fenaille, F., Barbier Saint-Hilaire, P., Rousseau, K. & Junot, C. Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: Where do we stand? *J Chromatogr A* **1526**, 1–12 (2017).

44. Johnson, A. R. & Carlson, E. E. Collision-Induced Dissociation Mass Spectrometry: A Powerful Tool for Natural Product Structure Elucidation. *Anal. Chem.* **87**, 10668–10678 (2015).

45. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779–787 (2006).

46. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).

47. Pedrioli, P. G. A. *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **22**, 1459–1466 (2004).

48. Deutsch, E. W. Mass spectrometer output file format mzML. *Methods Mol Biol* **604**, 319–331 (2010).

49. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).

50. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918–920 (2012).

51. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504 (2008).

52. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal. Chem.* **89**, 8689–8695 (2017).

53. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal Chem* **89**, 8696–8703 (2017).

54. Kantz, E. D., Tiwari, S., Watrous, J. D., Cheng, S. & Jain, M. Deep Neural Networks for Classification of LC-MS Spectral Peaks. *Anal. Chem.* **91**, 12407–12413 (2019).

55. Melnikov, A. D., Tsentalovich, Y. P. & Yanshole, V. V. Deep Learning for the Precise Peak Detection in High-Resolution LC–MS Data. *Anal. Chem.* **92**, 588–592 (2020).

56. Guo, J. *et al.* EVA: Evaluation of Metabolic Feature Fidelity Using a Deep Learning Model Trained With Over 25000 Extracted Ion Chromatograms. *Anal. Chem.* **93**, 12181–12186 (2021).

57. Yang, J. *et al.* Strategy for metabonomics research based on high-performance liquid chromatography and liquid chromatography coupled with tandem mass spectrometry. *Journal of Chromatography A* **1084**, 214–221 (2005).

58. Lange, E., Tautenhahn, R., Neumann, S. & Gröpl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **9**, 375 (2008).

59. Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in Bioinformatics* **16**, 104–117 (2015).

60. Prince, J. T. & Marcotte, E. M. Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal. Chem.* **78**, 6140–6152 (2006).

61. Smilde, A. K., van der Werf, M. J., Bijlsma, S., van der Werff-van der Vat, B. J. C. & Jellema, R. H. Fusion of Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **77**, 6729–6736 (2005).

62. Müller, E., Huber, C. E., Brack, W., Krauss, M. & Schulze, T. Symbolic Aggregate Approximation Improves Gap Filling in High-Resolution Mass Spectrometry Data Processing. *Anal. Chem.* **92**, 10425–10432 (2020).

63. Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).

64. Hastie, T. *et al.* Imputing Missing Data for Gene Expression Arrays. *Technical report, Stanford Statistics Department* **1**, (2001).

65. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci Rep* **8**, 663 (2018).

66. Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J. & Hanhineva, K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics* **20**, 492 (2019).

67. Tang, K., Toh, Q. & Teo, B. Normalisation of urinary biomarkers to creatinine for clinical practice and research – when and why. *smedj* **56**, 7–10 (2015).

68. Silva, L. P. *et al.* Measurement of DNA Concentration as a Normalization Strategy for Metabolomic Data from Adherent Cell Lines. *Anal. Chem.* **85**, 9536–9542 (2013).

69. Cao, B. *et al.* GC–TOFMS analysis of metabolites in adherent MDCK cells and a novel strategy for identifying intracellular metabolic markers for use as cell amount indicators in data normalization. *Anal Bioanal Chem* **400**, 2983–2993 (2011).

70. Sysi-Aho, M., Katajamaa, M., Yetukuri, L. & Orešič, M. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics* **8**, 93 (2007).

71. Redestig, H. *et al.* Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data. *Anal. Chem.* **81**, 7974–7980 (2009).

72. Dunn, W. B. *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* **6**, 1060–1083 (2011).

73. Kuligowski, J., Sánchez-Illana, Á., Sanjuán-Herráez, D., Vento, M. & Quintás, G. Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC). *Analyst* **140**, 7810–7817 (2015).

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

75. Wishart, D. S. Advances in metabolite identification. *Bioanalysis* **3**, 1769–1782 (2011).

76. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nat Rev Chem* **1**, 0054 (2017).

77. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).

78. Dunn, W. B. *et al.* Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9**, 44–66 (2013).

79. Guijas, C. *et al.* METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).

80. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**, 828–837 (2016).

81. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).

82. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).

83. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **8**, 3 (2016).

84. Böcker, S. Searching molecular structure databases using tandem MS data: are we there yet? *Current Opinion in Chemical Biology* **36**, 1–6 (2017).

85. Wang, Y., Kora, G., Bowen, B. P. & Pan, C. MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics. *Anal. Chem.* **86**, 9496–9503 (2014).

86. Creek, D. J. *et al.* Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Anal. Chem.* **83**, 8703–8710 (2011).

87. Cao, M. *et al.* Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **11**, 696–706 (2015).

88. Broeckling, C. D. *et al.* Enabling Efficient and Confident Annotation of LC−MS Metabolomics Data through MS1 Spectrum and Time Prediction. *Anal. Chem.* **88**, 9226–9234 (2016).

89. Hall, L. M. *et al.* Development of Ecom $_{50}$ and Retention Index Models for Nontargeted Metabolomics: Identification of 1,3-Dicyclohexylurea in Human Serum by HPLC/Mass Spectrometry. *J. Chem. Inf. Model.* **52**, 1222–1237 (2012).

90. Aicheler, F. *et al.* Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches. *Anal. Chem.* **87**, 7698–7704 (2015).

91. Wen, Y. *et al.* Retention Index Prediction Using Quantitative Structure–Retention Relationships for Improving Structure Identification in Nontargeted Metabolomics. *Anal. Chem.* **90**, 9434–9440 (2018).

92. Bonini, P., Kind, T., Tsugawa, H., Barupal, D. K. & Fiehn, O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal. Chem.* **92**, 7515–7522 (2020).

93. Sahigara, F. *et al.* Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **17**, 4791–4810 (2012).

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

95. Stanstrup, J., Neumann, S. & Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal. Chem.* **87**, 9421–9428 (2015).

96. Bouwmeester, R., Martens, L. & Degroeve, S. Generalized Calibration Across Liquid Chromatography Setups for Generic Prediction of Small-Molecule Retention Times. *Anal. Chem.* **92**, 6571–6578 (2020).

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

98. Camacho, D., de la Fuente, A. & Mendes, P. The origin of correlations in metabolomics data. *Metabolomics* **1**, 53–63 (2005).

99. Bartel, J., Krumsiek, J. & Theis, F. J. STATISTICAL METHODS FOR THE ANALYSIS OF HIGH-THROUGHPUT METABOLOMICS DATA. *Computational and Structural Biotechnology Journal* **4**, e201301009 (2013).

100. Wanichthanarak, K., Fahrmann, J. F. & Grapov, D. Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomark**Insights* **10s4**, BMI.S29511 (2015).

101. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 83 (2017).

102. Huang, S., Chaudhary, K. & Garmire, L. X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* **8**, 84 (2017).

103. Buescher, J. M. & Driggers, E. M. Integration of omics: more than the sum of its parts. *Cancer Metab* **4**, 4 (2016).

104. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the Metabolome: Size, Flux, and Regulation. *Molecular Cell* **58**, 699–706 (2015).

105. Barupal, D. K. & Fiehn, O. Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Sci Rep* **7**, 14567 (2017).

106. Barupal, D. K. *et al.* MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics* **13**, 99 (2012).

107. De Livera, A. M., Olshansky, M. & Speed, T. P. Statistical Analysis of Metabolomics Data. in *Metabolomics Tools for Natural Product Discovery* (eds. Roessner, U. & Dias, D. A.) vol. 1055 291–307 (Humana Press, 2013).

108. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

109. Xi, B., Gu, H., Baniasadi, H. & Raftery, D. Statistical Analysis and Modeling of Mass Spectrometry-Based Metabolomics Data. in *Mass Spectrometry in Metabolomics* (ed. Raftery, D.) vol. 1198 333–353 (Springer New York, 2014).

110. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).

111. Okuda, S. *et al.* KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Research* **36**, W423–W426 (2008).

112. Caspi, R. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research* **34**, D511–D516 (2006).

113. Karp, P. D. The EcoCyc and MetaCyc databases. *Nucleic Acids Research* **28**, 56–59 (2000).

114. Karp, P. D. The MetaCyc Database. *Nucleic Acids Research* **30**, 59–61 (2002).

115. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Research* **35**, D527–D532 (2007).

116. Barupal, D. K., Fan, S. & Fiehn, O. Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. *Current Opinion in Biotechnology* **54**, 1–9 (2018).

117. Kamburov, A., Cavill, R., Ebbels, T. M. D., Herwig, R. & Keun, H. C. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* **27**, 2917–2918 (2011).

118. Xia, J. & Wishart, D. S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc* **6**, 743–760 (2011).

119. Kankainen, M., Gopalacharyulu, P., Holm, L. & Orešič, M. MPEA—metabolite pathway enrichment analysis. *Bioinformatics* **27**, 1878–1879 (2011).

120. Gao, J. *et al.* Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics* **26**, 971–973 (2010).

121. Garcia-Alcalde, F., Garcia-Lopez, F., Dopazo, J. & Conesa, A. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics* **27**, 137–139 (2011).

122. Kuo, T.-C., Tian, T.-F. & Tseng, Y. J. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst Biol* **7**, 64 (2013).

123. Cottret, L. *et al.* MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Research* **38**, W132–W137 (2010).

124. Marco-Ramell, A. *et al.* Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics* **19**, 1 (2018).

125. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

126. Pei, G., Chen, L. & Zhang, W. WGCNA Application to Proteomic and Metabolomic Data Analysis. in *Methods in Enzymology* vol. 585 135–158 (Elsevier, 2017).

127. Krumsiek, J., Suhre, K., Illig, T., Adamski, J. & Theis, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol* **5**, 21 (2011).

128. Zuo, Y., Yu, G., Tadesse, M. G. & Ressom, H. W. Biological network inference using low order partial correlation. *Methods* **69**, 266–273 (2014).

129. Ma, J. *et al.* Differential network enrichment analysis reveals novel lipid pathways in chronic kidney disease. *Bioinformatics* **35**, 3441–3452 (2019).

130. Iyer, G. R. *et al.* Application of Differential Network Enrichment Analysis for Deciphering Metabolic Alterations. *Metabolites* **10**, 479 (2020).

131. Basu, S. *et al.* Sparse network modeling and Metscape-based visualization methods for the analysis of large-scale metabolomics data. *Bioinformatics* btx012 (2017) doi:10.1093/bioinformatics/btx012.

132. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* **44**, D463–D470 (2016).

133. Haug, K. *et al.* MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research* **41**, D781–D786 (2013).

134. Benton, H. P. *et al.* Intra- and Interlaboratory Reproducibility of Ultra Performance Liquid Chromatography–Time-of-Flight Mass Spectrometry for Urinary Metabolic Profiling. *Anal. Chem.* **84**, 2424–2432 (2012).

135. Telu, K. H., Yan, X., Wallace, W. E., Stein, S. E. & Simón-Manso, Y. Analysis of human plasma metabolites across different liquid chromatography/mass spectrometry platforms: Cross-platform transferable chemical signatures: Analysis of human plasma metabolites across different LC/MS platforms. *Rapid Commun. Mass Spectrom.* **30**, 581–593 (2016).

136. Gika, H. G. *et al.* Does the Mass Spectrometer Define the Marker? A Comparison of Global Metabolite Profiling Data Generated Simultaneously via UPLC-MS on Two Different Mass Spectrometers. *Anal. Chem.* **82**, 8226–8234 (2010).

137. Martin, J.-C. *et al.* Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study. *Metabolomics* **11**, 807–821 (2015).

138. Djekic, D., Pinto, R., Vorkas, P. A. & Henein, M. Y. Replication of LC–MS untargeted lipidomics results in patients with calcific coronary disease: An interlaboratory reproducibility study. *International Journal of Cardiology* **222**, 1042–1048 (2016).

139. Gürdeniz, G., Kristensen, M., Skov, T. & Dragsted, L. O. The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. *Metabolites* **2**, 77–99 (2012).

140. Hao, L. *et al.* Comparative Evaluation of MS-based Metabolomics Software and Its Application to Preclinical Alzheimer's Disease. *Sci Rep* **8**, 9291 (2018).

141. Enke, C. G. A Predictive Model for Matrix and Analyte Effects in Electrospray Ionization of Singly-Charged Ionic Analytes. *Anal. Chem.* **69**, 4885–4893 (1997).

142. Brown, M. *et al.* Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst* **134**, 1322 (2009).

143. Kruve, A. & Kaupmees, K. Adduct Formation in ESI/MS by Mobile Phase Additives. *J. Am. Soc. Mass Spectrom.* **28**, 887–894 (2017).

144. Creydt, M. & Fischer, M. Plant Metabolomics: Maximizing Metabolome Coverage by Optimizing Mobile Phase Additives for Nontargeted Mass Spectrometry in Positive and Negative Electrospray Ionization Mode. *Anal. Chem.* **89**, 10474–10486 (2017).

145. Chokkathukalam, A., Kim, D.-H., Barrett, M. P., Breitling, R. & Creek, D. J. Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. *Bioanalysis* **6**, 511–524 (2014).

146. Rosman, K. J. R. & Taylor, P. D. P. Isotopic Compositions of the Elements 1997. *Journal of Physical and Chemical Reference Data* **27**, 1275–1287 (1998).

147. Xu, Y.-F., Lu, W. & Rabinowitz, J. D. Avoiding Misannotation of In-Source Fragmentation Products as Cellular Metabolites in Liquid Chromatography–Mass Spectrometry-Based Metabolomics. *Anal. Chem.* **87**, 2273–2281 (2015).

148. Domingo-Almenara, X. *et al.* Autonomous METLIN-Guided In-source Fragment Annotation for Untargeted Metabolomics. *Anal. Chem.* **91**, 3246–3253 (2019).

149. Domingo-Almenara, X., Montenegro-Burke, J. R., Benton, H. P. & Siuzdak, G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Anal. Chem.* **90**, 480–489 (2018).

150. McMillan, A., Renaud, J. B., Gloor, G. B., Reid, G. & Sumarah, M. W. Post-acquisition filtering of salt cluster artefacts for LC-MS based human metabolomic studies. *J Cheminform* **8**, 44 (2016).

151. Mahieu, N. G. & Patti, G. J. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Anal. Chem.* **89**, 10397–10406 (2017).

152. Alonso, A. *et al.* AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics* **27**, 1339–1340 (2011).

153. DeFelice, B. C. *et al.* 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. *Anal. Chem.* **89**, 3250–3255 (2017).

154. Brown, M. *et al.* Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics* **27**, 1108–1112 (2011).

155. Tikunov, Y. M., Laptenok, S., Hall, R. D., Bovy, A. & de Vos, R. C. H. MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics* **8**, 714–718 (2012).

156. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **86**, 6812–6817 (2014).

157. Jaeger, C., Méret, M., Schmitt, C. A. & Lisec, J. Compound annotation in liquid chromatography/high-resolution mass spectrometry based metabolomics: robust adduct ion determination as a prerequisite to structure prediction in electrospray ionization mass spectra. *Rapid Commun Mass Spectrom* **31**, 1261–1266 (2017).

158. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).

159. Senan, O. *et al.* CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics* **35**, 4089–4097 (2019).

160. Rogers, S., Scheltema, R. A., Girolami, M. & Breitling, R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* **25**, 512–518 (2009).

161. Daly, R. *et al.* MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **30**, 2764–2771 (2014).

162. Del Carratore, F. *et al.* Integrated Probabilistic Annotation: A Bayesian-Based Annotation Method for Metabolomic Profiles Integrating Biochemical Connections, Isotope Patterns, and Adduct Relationships. *Anal. Chem.* **91**, 12799–12807 (2019).

163. Uppal, K., Walker, D. I. & Jones, D. P. xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal. Chem.* **89**, 1063–1067 (2017).

164. Mahieu, N. G., Spalding, J. L., Gelman, S. J. & Patti, G. J. Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. *Anal. Chem.* **88**, 9037–9046 (2016).

165. Mahieu, N. G., Huang, X., Chen, Y.-J. & Patti, G. J. Credentialing Features: A Platform to Benchmark and Optimize Untargeted Metabolomic Methods. *Anal. Chem.* **86**, 9583–9589 (2014).

166. Wang, L. *et al.* Peak Annotation and Verification Engine for Untargeted LC–MS Metabolomics. *Anal. Chem.* **91**, 1838–1846 (2019).

167. Lu, W. *et al.* Improved Annotation of Untargeted Metabolomics Data through Buffer Modifications That Shift Adduct Mass and Intensity. *Anal. Chem.* **92**, 11573–11581 (2020).

168. Kouřil, Š., de Sousa, J., Václavík, J., Friedecký, D. & Adam, T. CROP: correlation-based reduction of feature multiplicities in untargeted metabolomic data. *Bioinformatics* **36**, 2941–2942 (2020).

169. Monnerie *et al.* Analytic Correlation Filtration: A New Tool to Reduce Analytical Complexity of Metabolomic Datasets. *Metabolites* **9**, 250 (2019).

170. Giacomoni, F. *et al.* Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **31**, 1493–1495 (2015).

171. Kachman, M. *et al.* Deep annotation of untargeted LC-MS metabolomics data with *Binner*. *Bioinformatics* **36**, 1801–1806 (2020).

172. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

173. Nielsen, N.-P. V., Carstensen, J. M. & Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **805**, 17–35 (1998).

174. Zhang, X., Asara, J. M., Adamec, J., Ouzzani, M. & Elmagarmid, A. K. Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics* **21**, 4054–4059 (2005).

175. Jaitly, N. *et al.* Robust Algorithm for Alignment of Liquid Chromatography−Mass Spectrometry Analyses in an Accurate Mass and Time Tag Data Analysis Pipeline. *Anal. Chem.* **78**, 7397–7409 (2006).

176. Li, L. *et al.* An alignment algorithm for LC-MS-based metabolomics dataset assisted by MS/MS information. *Analytica Chimica Acta* **990**, 96–102 (2017).

177. Lange, E. *et al.* A geometric approach for the alignment of liquid chromatography—mass spectrometry data. *Bioinformatics* **23**, i273–i281 (2007).

178. Brunius, C., Shi, L. & Landberg, R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* **12**, 173 (2016).

179. Wu, C.-T. *et al.* Targeted realignment of LC-MS profiles by neighbor-wise compound-specific graphical time warping with misalignment detection. *Bioinformatics* **36**, 2862–2871 (2020).

180. Liu, Q. *et al.* Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. *Scientific Reports* **10**, 13856 (2020).

181. Mak, T. D., Goudarzi, M., Laiakis, E. C. & Stein, S. E. Disparate Metabolomics Data Reassembler: A Novel Algorithm for Agglomerating Incongruent LC-MS Metabolomics Datasets. *Anal. Chem.* **92**, 5231–5239 (2020).

182. Hsu, Y.-H. H. *et al.* PAIRUP-MS: Pathway analysis and imputation to relate unknowns in profiles from mass spectrometry-based metabolite data. *PLoS Comput Biol* **15**, e1006734 (2019).

183. Mitra, V. *et al.* Inversion of peak elution order prevents uniform time alignment of complex liquid-chromatography coupled to mass spectrometry datasets. *Journal of Chromatography A* **1373**, 61–72 (2014).

184. Habra, H. *et al. metabCombiner* : Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets. *Anal. Chem.* **93**, 5028–5036 (2021).

185. Eilers, P. H. C. & Marx, B. D. Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11**, (1996).

186. Wood, S. N. P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Stat Comput* **27**, 985–989 (2017).

187. Perng, W. *et al.* Early Life Exposure in Mexico to ENvironmental Toxicants (ELEMENT) Project. *BMJ Open* **9**, e030427–e030427 (2019).

188. Wang, N. & Boswell, P. G. Accurate prediction of retention in hydrophilic interaction chromatography by back calculation of high pressure liquid chromatography gradient profiles. *Journal of Chromatography A* **1520**, 75–82 (2017).

189. Kadjo, A. F., Dasgupta, P. K. & Srinivasan, K. Shape-Based Peak Identity Confirmation in Liquid Chromatography. *Anal. Chem.* **93**, 3848–3856 (2021).

190. Gutierrez, M., Handy, K. & Smith, R. Quantitative Evaluation of Algorithms for Isotopic Envelope Extraction via Extracted Ion Chromatogram Clustering. *J. Proteome Res.* **17**, 3774–3779 (2018).

191. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci USA* **112**, 12549–12550 (2015).

192. Uppal, K. *et al.* Computational Metabolomics: A Framework for the Million Metabolome. *Chem. Res. Toxicol.* **29**, 1956–1975 (2016).

193. Peisl, B. Y. L., Schymanski, E. L. & Wilmes, P. Dark matter in host-microbiome metabolomics: Tackling the unknowns–A review. *Analytica Chimica Acta* **1037**, 13–27 (2018).

194. Blaženović, I. *et al.* Structure Annotation of All Mass Spectra in Untargeted Metabolomics. *Anal. Chem.* **91**, 2155–2162 (2019).

195. Viant, M. R., Kurland, I. J., Jones, M. R. & Dunn, W. B. How close are we to complete annotation of metabolomes? *Current Opinion in Chemical Biology* **36**, 64–69 (2017).

196. Zhang, T. & Watson, D. G. A short review of applications of liquid chromatography mass spectrometry based metabolomics techniques to the analysis of human urine. *Analyst* **140**, 2907–2915 (2015).

197. Bouatra, S. *et al.* The Human Urine Metabolome. *PLoS ONE* **8**, e73076 (2013).

198. Neto, F. C. & Raftery, D. Expanding Urinary Metabolite Annotation through Integrated Mass Spectral Similarity Networking. *Anal. Chem.* **93**, 12001–12010 (2021).

199. Simón-Manso, Y. *et al.* Mass Spectrometry Fingerprints of Small-Molecule Metabolites in Biofluids: Building a Spectral Library of Recurrent Spectra for Urine Analysis. *Anal. Chem.* **91**, 12021–12029 (2019).

200. Cooper, B. T. *et al.* Hybrid Search: A Method for Identifying Metabolites Absent from Tandem Mass Spectrometry Libraries. *Anal. Chem.* **91**, 13924–13932 (2019).

201. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12580–12585 (2015).

202. Kind, T. *et al.* Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom Rev* **37**, 513–532 (2018).

203. Neue, U. D. Theory of peak capacity in gradient elution. *J Chromatogr A* **1079**, 153–161 (2005).

204. Lenz, E. M. *et al.* HPLC-NMR with severe column overloading: fast-track metabolite identification in urine and bile samples from rat and dog treated with [14C]-ZD6126. *J Pharm Biomed Anal* **43**, 1065–1077 (2007).

205. Anderson, B. G., Raskind, A., Habra, H., Kennedy, R. T. & Evans, C. R. Modifying Chromatography Conditions for Improved Unknown Feature Identification in Untargeted Metabolomics. *Anal. Chem.* **93**, 15840–15849 (2021).

206. Bruce, S. J. *et al.* Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. *Anal Chem* **81**, 3285–3296 (2009).

207. Kind, T. *et al.* LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat Methods* **10**, 755–758 (2013).

208. Chatterjee, S. *et al.* Using pattern recognition entropy to select mass chromatograms to prepare total ion current chromatograms from raw liquid chromatography-mass spectrometry data. *J Chromatogr A* **1558**, 21–28 (2018).

209. Dennis, E. A. Lipidomics joins the omics evolution. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 2089–2090 (2009).

210. Quehenberger, O. & Dennis, E. A. The Human Plasma Lipidome. *N Engl J Med* **365**, 1812–1823 (2011).

211. Fahy, E. *et al.* A comprehensive classification system for lipids. *J Lipid Res* **46**, 839–861 (2005).

212. Fahy, E. *et al.* Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res* **50 Suppl**, S9-14 (2009).

213. Yang, K. & Han, X. Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. *Trends in Biochemical Sciences* **41**, 954–969 (2016).

214. Koelmel, J. P. *et al.* Lipid Annotator: Towards Accurate Annotation in Non-Targeted Liquid Chromatography High-Resolution Tandem Mass Spectrometry (LC-HRMS/MS) Lipidomics Using A Rapid and User-Friendly Software. *Metabolites* **10**, E101 (2020).

215. Alcoriza-Balaguer, M. I. *et al.* LipidMS: An R Package for Lipid Annotation in Untargeted Liquid Chromatography-Data Independent Acquisition-Mass Spectrometry Lipidomics. *Anal Chem* **91**, 836–845 (2019).

216. Kochen, M. A. *et al.* Greazy: Open-Source Software for Automated Phospholipid Tandem Mass Spectrometry Identification. *Anal Chem* **88**, 5733–5741 (2016).

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

218. Ni, Z., Angelidou, G., Lange, M., Hoffmann, R. & Fedorova, M. LipidHunter Identifies Phospholipids by High-Throughput Processing of LC-MS and Shotgun Lipidomics Datasets. *Anal Chem* **89**, 8800–8807 (2017).

219. Bowden, J. A. *et al.* Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950–Metabolites in Frozen Human Plasma. *Journal of Lipid Research* **58**, 2275–2288 (2017).

220. Cajka, T., Smilowitz, J. T. & Fiehn, O. Validating Quantitative Untargeted Lipidomics Across Nine Liquid Chromatography-High-Resolution Mass Spectrometry Platforms. *Anal Chem* **89**, 12360–12368 (2017).

221. Spanier, B. *et al.* Comparison of lipidome profiles of Caenorhabditis elegans-results from an inter-laboratory ring trial. *Metabolomics* **17**, 25 (2021).

222. Nyamundanda, G., Gormley, I. C., Fan, Y., Gallagher, W. M. & Brennan, L. MetSizeR: selecting the optimal sample size for metabolomic studies using an analysis based approach. *BMC Bioinformatics* **14**, 338 (2013).

223. Deng, K. *et al.* WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. *Anal Chim Acta* **1061**, 60–69 (2019).

224. De Livera, A. M. *et al.* Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem* **87**, 3606–3615 (2015).

225. Fan, S. *et al.* Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal. Chem.* **91**, 3590–3596 (2019).

226. Shen, X. *et al.* Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* **12**, 89 (2016).

227. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

228. Renard, E., Branders, S. & Absil, P.-A. Independent Component Analysis to Remove Batch Effects from Merged Microarray Datasets. in *Algorithms in Bioinformatics* (eds. Frith, M. & Storm Pedersen, C. N.) vol. 9838 281–292 (Springer International Publishing, 2016).

229. Renard, E. & Absil, P.-A. Comparison of location-scale and matrix factorization batch effect removal methods on gene expression datasets. in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1530–1537 (IEEE, 2017). doi:10.1109/BIBM.2017.8217888.

230. Han, S. *et al.* TIGER: technical variation elimination for metabolomics data using ensemble learning architecture. *Briefings in Bioinformatics* **23**, bbab535 (2022).

231. Rong, Z. *et al.* NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **92**, 5082–5090 (2020).

232. Goutman, S. A. Diagnosis and Clinical Management of Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders: *CONTINUUM: Lifelong Learning in Neurology* **23**, 1332–1359 (2017).

233. Goutman, S. A., Chen, K. S., Paez-Colasante, X. & Feldman, E. L. Emerging understanding of the genotype–phenotype relationship in amyotrophic lateral sclerosis. in *Handbook of Clinical Neurology* vol. 148 603–623 (Elsevier, 2018).

234. Paez-Colasante, X., Figueroa-Romero, C., Sakowski, S. A., Goutman, S. A. & Feldman, E. L. Amyotrophic lateral sclerosis: mechanisms and therapeutics in the epigenomic era. *Nat Rev Neurol* **11**, 266–279 (2015).

235. Goutman, S. A. *et al.* High plasma concentrations of organic pollutants negatively impact survival in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **90**, 907–912 (2019).

236. Su, F.-C. *et al.* Association of Environmental Toxins With Amyotrophic Lateral Sclerosis. *JAMA Neurol* **73**, 803 (2016).

237. Cassina, P. *et al.* Peroxynitrite triggers a phenotypic transformation in spinal cord astrocytes that induces motor neuron apoptosis: Astrocytes Induce Motor Neuron Death. *J. Neurosci. Res.* **67**, 21–29 (2002).

238. Dodge, J. C., Yu, J., Sardi, S. P. & Shihabuddin, L. S. Sterol auto-oxidation adversely affects human motor neuron viability and is a neuropathological feature of amyotrophic lateral sclerosis. *Sci Rep* **11**, 803 (2021).

239. Rozen, S. *et al.* Metabolomic analysis and signatures in motor neuron disease. *Metabolomics* **1**, 101–108 (2005).

240. Lawton, K. A. *et al.* Plasma metabolomic biomarker panel to distinguish patients with amyotrophic lateral sclerosis from disease mimics. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* **15**, 362–370 (2014).

241. Bjornevik, K. *et al.* Prediagnostic plasma metabolomics and the risk of amyotrophic lateral sclerosis. *Neurology* 10.1212/WNL.0000000000007401 (2019) doi:10.1212/WNL.0000000000007401.

242. Blasco, H. *et al.* Lipidomics Reveals Cerebrospinal-Fluid Signatures of ALS. *Sci Rep* **7**, 17652 (2017).

243. Chaves-Filho, A. B. *et al.* Alterations in lipid metabolism of spinal cord linked to amyotrophic lateral sclerosis. *Sci Rep* **9**, 11642 (2019).

244. Patin, F. *et al.* Omics to Explore Amyotrophic Lateral Sclerosis Evolution: the Central Role of Arginine and Proline Metabolism. *Mol Neurobiol* **54**, 5361–5374 (2017).

245. Wuolikainen, A. *et al.* Multi-platform mass spectrometry analysis of the CSF and plasma metabolomes of rigorously matched amyotrophic lateral sclerosis, Parkinson's disease and control subjects. *Mol. BioSyst.* **12**, 1287–1298 (2016).

246. Blasco, H. *et al.* Biomarkers in amyotrophic lateral sclerosis: combining metabolomic and clinical parameters to define disease progression. *Eur J Neurol* **23**, 346–353 (2016).

247. Goutman, S. A. *et al.* Metabolomics identifies shared lipid pathways in independent amyotrophic lateral sclerosis cohorts. *Brain* awac025 (2022) doi:10.1093/brain/awac025.

248. Goutman, S. A. *et al.* Untargeted metabolomics yields insight into ALS disease mechanisms. *J Neurol Neurosurg Psychiatry* **91**, 1329–1338 (2020).

249. DeHaven, C. D., Evans, A. M., Dai, H. & Lawton, K. A. Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *J Cheminform* **2**, 9 (2010).

250. Ma, J., Shojaie, A. & Michailidis, G. Network-based pathway enrichment analysis with incomplete network information. *Bioinformatics* **32**, 3165–3174 (2016).

251. LaBarre, J. L. *et al.* Maternal lipid levels across pregnancy impact the umbilical cord blood lipidome and infant birth weight. *Sci Rep* **10**, 14209 (2020).

252. Barker, D. J. P. The origins of the developmental origins theory. *J Intern Med* **261**, 412–417 (2007).

253. Lindsay, K. L. *et al.* Longitudinal Metabolomic Profiling of Amino Acids and Lipids across Healthy Pregnancy. *PLoS One* **10**, e0145794 (2015).

254. Marchlewicz, E. H. *et al.* Lipid metabolism is associated with developmental epigenetic programming. *Sci Rep* **6**, 34857 (2016).

255. Luan, H. *et al.* Pregnancy-Induced Metabolic Phenotype Variations in Maternal Plasma. *J. Proteome Res.* **13**, 1527–1536 (2014).

256. Desert, R., Canlet, C., Costet, N., Cordier, S. & Bonvallot, N. Impact of maternal obesity on the metabolic profiles of pregnant women and their offspring at birth. *Metabolomics* **11**, 1896–1907 (2015).

257. Lowe, W. L. *et al.* Maternal BMI and Glycemia Impact the Fetal Metabolome. *Diabetes Care* **40**, 902–910 (2017).

258. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

259. Fages, A. *et al.* Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. *Metabolomics* **10**, 1074–1083 (2014).

260. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

261. Pinto, J. *et al.* Following healthy pregnancy by NMR metabolomics of plasma and correlation to urine. *J Proteome Res* **14**, 1263–1274 (2015).

262. Innis, S. M. Dietary (n-3) fatty acids and brain development. *J Nutr* **137**, 855–859 (2007).

263. Herrera, E. & Amusquivar, E. Lipid metabolism in the fetus and the newborn. *Diabetes Metab Res Rev* **16**, 202–210 (2000).

264. Kalhan, S. C. Protein metabolism in pregnancy. *Am J Clin Nutr* **71**, 1249S–55S (2000).

265. Sugar, J. *et al.* Estriol-3-Glucuronide and Estriol-16-Glucuronide in Amniotic Fluid during Normal Pregnancy*. *The Journal of Clinical Endocrinology & Metabolism* **50**, 137–143 (1980).

266. Yang, Y. J., Lee, J., Choi, M. H. & Chung, B. C. Direct determination of estriol 3- and 16-glucuronides in pregnancy urine by column-switching liquid chromatography with electrospray tandem mass spectrometry. *Biomed. Chromatogr.* **17**, 219–225 (2003).

267. Banker, M. *et al.* Association of Maternal-Neonatal Steroids With Early Pregnancy Endocrine Disrupting Chemicals and Pregnancy Outcomes. *The Journal of Clinical Endocrinology & Metabolism* **106**, 665–687 (2021).

268. van den Akker, C. H. P. *et al.* Human fetal amino acid metabolism at term gestation. *Am J Clin Nutr* **89**, 153–160 (2009).

269. Battaglia, F. C. & Regnault, T. R. Placental transport and metabolism of amino acids. *Placenta* **22**, 145–161 (2001).

270. Yang, X. *et al.* Metabolomics study and meta-analysis on the association between maternal pesticide exposome and birth outcomes. *Environmental Research* **182**, 109087 (2020).

271. Pang, Z., Zhou, G., Chong, J. & Xia, J. Comprehensive Meta-Analysis of COVID-19 Global Metabolomics Datasets. *Metabolites* **11**, 44 (2021).

272. Roointan, A., Gheisari, Y., Hudkins, K. L. & Gholaminejad, A. Non-invasive metabolic biomarkers for early diagnosis of diabetic nephropathy: Meta-analysis of profiling metabolomics studies. *Nutrition, Metabolism and Cardiovascular Diseases* **31**, 2253–2272 (2021).

273. Tautenhahn, R. *et al.* metaXCMS: Second-Order Analysis of Untargeted Metabolomics Data. *Anal. Chem.* **83**, 696–700 (2011).

274. Llambrich, M., Correig, E., Gumà, J., Brezmes, J. & Cumeras, R. Amanida: an R package for meta-analysis of metabolomics non-integral data. *Bioinformatics* **38**, 583–585 (2022).

275. Viallon, V. *et al.* A New Pipeline for the Normalization and Pooling of Metabolomics Data. *Metabolites* **11**, 631 (2021).