Leveraging high-resolution mass spectrometry detection of stable isotopes for metabolomics

By

Javier David Gomez

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Biomolecular Engineering

May 10th, 2024

Nashville, Tennessee

Approved:

Jamey D. Young, Ph.D.

Paul E. Laibinis, Ph.D.

Marjan Rafat, Ph.D.

Renã A.S. Robinson, Ph.D.

To my three brothers, who I am very proud of.

ACKNOWLEDGEMENTS

As I stand in front of the fourth trusty computer, searching for the right words, I confront the feeling that each word I write is a signature admitting that I am incapable of truly expressing how thankful I am to every single person that, whether directly or indirectly, contributed to this journey. Even if my words fall short, I want to express my gratitude to those who guided me to this point.

First, I want to thank Jamey Young, my advisor, for accepting me in his lab and helping me become a better scientist and person. I can only describe Jamey embodies my vision of what an academic role model is. His expertise and intelligence were surpassed only by his compassion, kindness, and patience. If I were to go back to being a professor, I could only aspire to follow in his footsteps. His guidance and encouragement throughout this process were simply unvaluable. I would also like to thank the rest of the members of my committee: Paul Laibinis for his constant support and tough questions that always taught me something new and constantly gave me new perspectives; Marjan Rafat for her kindness and thought-provoking biological questions, keeping me focused on both the biological and the bioinformatic aspects of my research; and Renã Robinson for her insights on proteomics, always broadening the applicability horizons of my project.

To my collaborators, the people at the Donald Danforth Plant Science Center, Doug Allen, Brad Evans, and Shrikaar Kambhampati. This work would not have been done without your help, efforts, and insights. Thank you for your feedback that helped me improve my research and for being a reliable source of high-quality data that from the early stages of the project shaped the quality of my work.

I would like to express my gratitude to my past mentors for their unwavering support across countless stages of my life. To Oscar Alvarez, thank you for shaping me as a young scientist. Your

belief in me and the numerous opportunities you opened have meant a lot. Uniandes will forever hold a special place in my heart, filled with appreciation and respect. The decision to embark on this journey was influenced by everyone there. Professors Jorge Mario and Pradillosky, Cris, Mauro, and Andreita, you were with me from the very beginning, providing support throughout this entire process, and my gratitude for you will be eternal.

I'm grateful for finding a research group with incredible people. Thanks to Dr. Hasenour for his unwavering support, encouragement to reach new heights, take the next step, jump into challenges, and aim high. In addition to the insightful scientific conversations, thanks for the non-scientific discussions and for allowing me to be annoying. My respect and admiration will always follow him. Special thanks to Amy Zheng for her friendship, making this process more enjoyable with her joyous personality and her ability to create smiles regardless of the situation. Thanks to Deveena Banerjee for letting me be (as Irina called us) a “mean girl”, sharing laughs and gossip in the hallways. To everyone in the lab who made me feel welcome—thank you for being more than just lab mates: Irina, Sarah, Zach, Kevin, Bo, Tomasz, Mohsin, Baltazar, Rachel, Sara, Miles, Maggie, and Coston.

I want to thank all the amazing friends that I made here who will always be in my heart: Ben Hacker, Rabeb Layouni, Alya Nur Afiqah, Gerrit Gaillard, Felisha Baquera, Angie Pernell. Specially, to my Beloved Basement Big-Brained Best Buddies: Eden Paul and Allison Cordova, my support and love will always be with you, regardless of where you are. I have confidence that you will succeed in your Ph.D journey, and I look forward to call you “doctor” one day. Thanks for all the laughs, experiences, beers, tears, and talks we shared. You will always be in my heart, and I will always cherish those memories.

I want to thank Kyle Shadrick for supporting me and helping me in the toughest moments. It was a long journey, but thanks to you I was able to grow a lot and face the tough times with better resources. Also, I want to extend my deepest gratitude to Olivia Darrow for being my rock during the darkest moments. When the incompetence of other offices became glaringly evident, your unwavering support kept me alive and grounded. Your love and dedication to your job served as a beacon of hope, reminding me that despite the challenges, it's still possible to do the right thing with integrity, courage, and commitment.

I need to thank my best friend and brother, Nicolas Melo. His unconditional support and love held me together in the toughest moments. I could not have done this without having him by my side, and my gratitude and love will always be with him. Thank you for always listening and for always being there. I don't think words can fully express my gratitude and appreciation.

To my other brothers, Thibault Bignon and Daniel Forero, thanks for your support, example, and love. I admire both of you fondly, and I have learned so many things from you. You can always count on me, just as I have counted on you.

To two of my mentees and beloved friends, Kimia Koushesh and Diego Cadena, thank you for teaching me so much. I admire and respect you, and I have great expectations that so far you both have fulfilled. Your growth always served me as an inspiration and as motivation. I am so proud of you.

Lastly but not less importantly, I need to thank my family. Special thanks to my aunt Cristina and my uncle Pablo; you were always there to listen and to give me advice. Thank you for your presence and mentorship in my life that has shaped me into the person I am. Thanks to my grandmother Margarita for always listening to my rants and showing me compassion and love.

Thanks to my dads and my mom for all the effort and sacrifices that ultimately led me here. Thanks to my cousin Jero for giving me an excuse to take some breaks. To the rest of my family—my sister, my cousins, my uncles, my grandparents, and my dog—thanks for being the support net that brought me here.

In the words of Sir Isaac Newton, "if I have seen further, it is by standing on the shoulders of giants." I want to express my sincere thanks to every person who has been a part of my journey. Having incredible people in my life has been a privilege that I will always acknowledge and appreciate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
3-PGA	3-Phosphoglycerate
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CLI	Command Line Interface
CM	Correction Matrix
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
DHAP	Dihydroxyacetone phosphate
ESI	Electrospray Ionization
EIC	Extracted Ion Chromatogram
M+1	First isotopologue
GC-MS	Gas chromatography - mass spectrometry
G6P	Glucose-6-phosphate
GAP	Glyceraldehyde-3-phosphate
HF	High fat
HRMS	High resolution mass spectrometry
HPLC	High-performance liquid chromatography
IPM	Isotope purity matrix
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	Liquid chromatography - mass spectrometry
MID	Mass isotopomer distribution
MS	Mass spectrometry
m/z	Mass-to-charge ratio
Mc4r	Melanocortin 4 receptor
MFA	Metabolic flux analysis
Mo	Monoisotopic mass
NIST	National Institute of Standards and Technology
NMR	Nuclear magnetic resonance
PLS	Partial least squares
PEP	Phosphoenolpyruvate
PCA	Principal component analysis
RT	Retention time
SIM	Selective Ion Monitoring
SVM	Support vector machine
MS/MS	Tandem mass spectrometry
MS2	Tandem mass spectrometry
TBDMS	*tert*-Butyldimethylsilyl
TOF	Time-of-flight
TIC	Total Ion Chromatogram

TM	Tracer matrix
TCA	Tricarboxylic acid cycle

CHAPTER 1

INTRODUCTION

Over the last couple of decades, the use of stable isotopes in the research of metabolism has given rise to significant discoveries across several species. The premise of these studies is to feed a metabolically active tracer labeled with stable isotopes into a living organism and track the incorporation of the stable isotopes into metabolic byproducts as the tracer is metabolized [1], [2]. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) have emerged as the most useful measurement approaches given their capabilities to determine the mass shifts and the position of the incorporated isotopes [3], [4]. The use of stable isotopes has allowed researchers to map the connection between metabolites into clearer metabolic pathways and has served as a reliable method to calculate the metabolic fluxes within the pathways [5], [6], [7].

Despite the fact that bioinformatic tools are created and published at a high rate, they still cannot fully leverage all the technological advances in mass spectrometry. The high resolving power and accurate mass quantification achieved with high-resolution mass spectrometry (HRMS) opens the door to novel experiments that incorporate more than one tracer (e.g., $^{2}H/^{13}C$ or $^{13}C/^{15}N$) or combine untargeted methods (i.e., experiments where the identity of the objective metabolites is not known a priori) into typical metabolomics workflow [8]. It has to be noted that some experiments using dual tracers have been done in the past, but they have not leveraged the capabilities of HRMS [9], [10], [11], [12].

The overall theme of this dissertation is the development of bioinformatic tools that automate the data analysis process in stable isotope-resolved metabolomics experiments that use modern HRMS instruments as the main source of the data. This work builds upon previous in-house tools and

algorithms that were developed within the Young research group, optimizing and generalizing them for a wider variety of experimental designs and file formats. The overall objective of the work presented herein is to develop bioinformatic tools that enable the metabolomics community to fully leverage stable isotopes to profile metabolic network dynamics.

This research was accomplished by developing two Matlab-based tools: SUNDILE and PIRAMID. The former offers an untargeted dimension reduction approach to extract, assess, and identify the metabolites that exhibit detectable enrichment during a stable isotope labeling experiment. The latter offers a targeted approach to automate the tasks required for extracting the isotope enrichment information from vendor-agnostic mass spectrometry files. Overall, they serve as the basis for an efficient workflow in stable isotope-based metabolomics: first, untargeted studies are used to identify potential biomarkers and pathways that show differences in labeling between two or more conditions. Then, targeted studies are used to accurately quantify the enrichment patterns of specific metabolites, and subsequent studies use those results to infer biological information such as metabolic pathway activities determined by metabolic flux analysis (MFA) [13], [14].

The development of these tools also complements a previously existing project, the MFA Suite™, which is a set of bioinformatic tools designed to facilitate MFA. Previously the MFA Suite™ consisted of two post-data processing tools aimed to calculate cell-specific metabolic rates [15] and to model isotopic measurements for MFA [16]. With the addition of the tools that were developed during this dissertation research, the MFA Suite™ now offers support at every single step of the metabolomics workflow.

This dissertation is divided into the following chapters:

Chapter 2 delves into a comprehensive literature review, presenting the foundations that this work rests upon. First, background on MS measurement capabilities and limitations will be established, followed by the examination of MS applications to stable isotope labeling experiments in metabolomics. Finally, the advancements and gaps in the bioinformatic tools that support these experiments will be discussed.

Chapter 3 offers an exposition of the untargeted tool SUNDILE, unveiling its algorithms and potential applications. This tool converts a list of extracted peaks from an MS dataset where stable isotopes were used as a metabolic tracer and extracts the labeled metabolites and suggests their identity framing their enrichment into a metabolic pathway context.

Chapter 4 elucidates the functionalities and intricate architecture of the targeted tool PIRAMID. This tool takes mass spectrometry files in a vendor-agnostic format and a Matlab file with information of the target metabolites and extracts the information that is commonly used in stable isotope labeling-based metabolomics such as the peak intensity, the isotopic distribution, and the isotope enrichment.

Chapter 5 explores the application of the previously described tools to study the metabolism of soybean (*Glycine max*). The untargeted exploration and discovery of target pathways labeled with ^{13}C-glucose or ^{13}C/^{15}N-glutamine is discussed, followed by the targeted integration of specific metabolites that shed light onto the differences of the metabolism of two different cultivars with differing oil content.

Chapter 6 follows the same outline as Chapter 5 but focuses on the metabolism of mice *(Mus musculus*) with different diets and genotypes in the context of insulin resistance.

Chapter 7 draws together the threads of the developed bioinformatic tools and their applications to summarize the overall conclusions of the research. Building upon the insights presented throughout the dissertation, this chapter will discuss the implications and limitations of the research as well as the potential impact on future studies.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 Mass spectrometry

Mass spectrometry is a relatively old analytical technique that was developed in the early 20th century and has attracted special attention from researchers in life and health sciences in the last couple of decades [17], [18], [19], [20]. It operates based on the principle of the formation of gas-phase ions that are isolated using electric or magnetic fields based on their mass-to-charge ratio (*m/z*). In magnetic mass analyzers, lighter ions are deflected more than heavier ions, leading to a spatial separation across the detection path where their position (converted to a *m/z*) and intensity are recorded [21]. Quadrupole-based mass analyzers rely on the creation of an electric field induced by a radio frequency voltage in conjunction with a direct current offset voltage in a quadrupole geometry composed of rods or plates. By adjusting the voltages, the trajectory of the ions in the quadrupole can be manipulated, selectively allowing certain masses to have stable trajectories into the mass detector [22], [23], [24]. Finally, time-of-flight (TOF)-based mass analyzers accelerate ions through a chamber in an electric field of known strength. Hence, all ions in the sample are subject to the same kinetic energy, that is translated into different speeds depending on their masses. The time each ion takes to reach the mass analyzer is measured and is correlated with their mass-to-charge ratio [25].

Different compounds can yield similar ions which can confound the results of mass spectrometry when the sample is introduced without prior purification. This poses an issue given that a biological sample can have hundreds to thousands of compounds [26] and is why a separation step is often needed before the ionization step. Chromatography has emerged as a solution, separating

metabolites and compounds in a mobile phase based on their chemical affinity and interactions with a stationary phase. Following this logic, compounds with higher affinity will be retained by the stationary phase whereas compounds with lower affinity will elute faster. This creates a separation in time such that each compound can be analyzed with minimal interference from other compounds [27].

The coupling of chromatography with mass spectrometry has emerged as the leading analytical tool to perform metabolomics studies [28], [29]. Depending on the phase of the sample during the chromatographic separation, this technique is abbreviated to GC-MS for gas chromatography coupled with mass spectrometry and LC-MS for liquid chromatography coupled with mass spectrometry.

2.1.1 Mass spectrometry metadata

As discussed earlier, both GC-MS and LC-MS provide information on the compound's elution time in the chromatography step (also referred to as retention time or RT) and the distribution of different *m/z* values and intensities that were detected in the mass spectrometry step. This combination of descriptive variables creates a complex three-dimensional dataset that requires thorough exploration during data analysis. To achieve this, for each specific RT, a two-dimensional graph can be generated, illustrating the *m/z* values and their corresponding intensities. This is known as a *spectrogram* or *mass spectrum*. Analogously, for a specific *m/z* value a two-dimensional graph can be extracted, displaying the RTs at which the selected ion elutes and the corresponding intensities. As the elution of compounds progressively happens over time, the intensities take the form of peaks. This is known as a *chromatogram;* For a specific *m/z* value, the extracted data is usually called an *Extracted Ion Chromatogram* (EIC).

The utilization of spectrograms and chromatograms aids in the precise identification and quantification of compounds, significantly enhancing the understanding of the analytical results obtained from both GC-MS and LC-MS techniques. Depending on the applied energy during the ionization, each compound will have a unique fragmentation pattern and will yield different ions. The comparison of the resulting fragmentation patterns in the spectrograms to libraries of standards can elucidate the identity of a compound [30], [31]. However, co-eluting compounds can yield similar ion fragments which leaves the user with the task to solve the "puzzle" of determining which fragments correspond to each compound, a process known as *peak deconvolution*.

In certain cases, a minimum of energy is supplied in the form of chemical ionization, electrospray ionization, and photoionization to produce what is known as a "soft ionization" [32], [33], [34]. This process usually yields hydrogen adducts in the form pseudo-molecules regardless of the resulting charge of the ionization process [35]. In these cases, if the mass is resolved with enough accuracy, the formula of the metabolite can be inferred but the isomeric determination is left unresolved [36], [37].

2.1.2 Tandem mass spectrometry

Recent advances in MS technology have addressed some of the issues and limitations that were discussed earlier. Tandem MS allows the instrument to isolate ions in a specific *m/z* range before fragmentation, usually ensuring that the detected fragments correspond to a single precursor ion. Tandem MS relies on the use of multiple mass analyzers, a first analyzer (MS1) that gathers information of the unfragmented ions (known as precursor or parent ions) that are fragmented into smaller ions (known as product or daughter ions) to later be analyzed again by another mass analyzer (MS2). Therefore, tandem MS is known as MS/MS or MS^2. In theory, it is possible to

assemble a series of mass analyzers to re-fragment and analyze the product ions multiple times, which is also noted as MS^n [38], [39].

In a fully untargeted experiment using MS^2, the *m/z* values of the precursor ions derived from a biological sample are recorded, along with the *m/z* values of their corresponding product ions. While this technology effectively resolves the challenge of correlating each compound with its corresponding fragments, it also adds another dimension to the metadata that previously contained the information of the RT, the *m/z* values, and the ion intensities. This type of acquisition mode is known as data-independent acquisition (DIA) [40] because the *m/z* values of the product ions will be acquired for all precursor ions in a certain range of user-selected masses [41].

Analogously, after the *m/z* values and intensities of the precursor ions are analyzed, some instruments can fragment a certain number of ions depending on their abundance. Hence, the top *N* most abundant precursor ions are fragmented and analyzed. This method is known as data-dependent acquisition (DDA) [40].

2.1.3 High resolution mass spectrometry

In addition to using fragment ion spectra to elucidate compound identity, it is possible to infer chemical formula of compounds from accurate *m/z* values with the implementation of electromagnetic ion traps, ion cyclotron resonance, or time-of-flight mass analyzers. These techniques allow the instrument to resolve the *m/z* values with errors as low as 0.001 mass units for low molecular weight metabolites [42], [43].

It is possible to correlate an accurate mass to a chemical formula by iterating the elements and number of atoms that are typically found in organic compounds, and the possibilities can be sorted based on the distribution of naturally abundant isotopes [44], [45]. The disadvantage of this

technology is that as more accurate information on the *m/z* values is gathered, the size of the metadata increases significantly, requiring additional compression [46]. Consequently, various data processing algorithms that were developed for low-resolution MS are not compatible with HRMS.

2.2 Mass spectrometry-based metabolomics

2.2.1 Classic metabolomics

The use of MS to study the metabolism of different species can be done in two ways: the first one is known as metabolic profiling, where the metabolite concentrations are compared across different conditions. This is done in an untargeted manner to find metabolites that are associated with certain diseases or metabolic disorders [30], [47] and in a targeted manner to quantify the abundances of a defined list of metabolites [48], [49]. Most untargeted studies aim to determine which metabolites and pathways can be used as biomarkers of a given perturbation or metabolic state. For targeted studies, the accurate quantification of the abundance of metabolites (which is proportional to the peak intensity in MS) is needed. Therefore, targeted studies only require information on the intensity of the target peaks. However, metabolic profiling studies only represent a steady-state snapshot of metabolism, which is an inherently dynamic process. It is possible for a metabolite to remain at the same steady-state concentration while the flux of reactions producing and consuming the metabolite changes significantly [50], [51].

2.2.2 Stable isotope-assisted metabolomics

The second form of metabolic analysis performed with MS is known as metabolic flux analysis (MFA), which relies on the use of stable isotopes (e.g., ^{2}H, ^{13}C, ^{15}N) to track the dynamic incorporation of substrate atoms into metabolic products after the substrates are metabolized [30],

[52]. The metabolic fate of the labeled atoms can take numerous paths through the reaction network, leading to a distribution in the number of labeled atoms incorporated into downstream products. The multiple isomers that differ only in their isotopic composition with otherwise identical chemical structure are known as *isotopologues* [53] (e.g., [1-^{13}C]glutamine, [$^{15}N_2$]glutamine, [1,2-$^{13}C_2$]glutamine, and[U-$^{13}C_5$]glutamine are isotopologues of glutamine). On the other hand, groups of isotopologues of the same mass that cannot be separated by mass spectrometry alone are known as *mass isotopomers* [54], [55]. When measured at nominal mass resolution, mass isotopomers are denoted as M+*X*, where *X* is the nominal mass that is added to the monoisotopic mass. MFA requires precise measurements of the mass isotopomer distribution (MID), which emerges from the incorporation of labeled atoms following the metabolism of isotopically enriched tracers.

Recent technologies such as MS^2 and HRMS have found specific applications in metabolomics studies. While HRMS can lead to an exact formula, MS^2 is exceptionally valuable for metabolite identification, as the distinctive fragmentation patterns of the detected precursor ions can lead to the precise identification of the possible isomers when compared to libraries. Furthermore, it enables the filling of partial metabolomes by detecting intermediate compounds and their metabolic reactions, as the fragments themselves carry crucial information about the chemical properties of the metabolites. Hence, the placement of previously unknown or unconnected metabolites within pathways becomes feasible, even for metabolites involved in multiple pathways, overcoming a challenging aspect of metabolomics analysis [56], [57], [58].

As previously discussed, HRMS allows a more exact identification of compounds by comparing the exact *m/z* value and the isotopic distribution to theoretical simulated values. Furthermore, this technology also opens the door to a more complex type of study: stable isotope-based

metabolomics involving multiple tracers. In these studies, multiple stable isotopes are employed (e.g., ^{13}C and ^{15}N, ^{2}H and ^{13}C). HRMS plays a crucial role in resolving the mass defect between the resulting isotopologues, enabling simultaneous tracking of the fate of multiple labeled atoms [59], [60], [61].

The workflow of stable isotope-assisted mass spectrometry-based metabolomics starts with an experimental design where two metabolic conditions are set (e.g., wild-type vs. knock-out, different metabolic states, etc.), and then the organisms are given a stable isotope-labeled tracer that will enrich the metabolic pathways of interest. Data gathering follows, where samples are collected and analyzed using MS. Preprocessing is often needed to prepare the raw data: this step encompasses data transformation, peak alignment across samples, and metabolite normalization to improve the quality of the data and reduce unwanted variations. Next, untargeted metabolomics comes into play, by means of MS^2 acquired data and statistical tools, so that potential candidate biomarkers associated with specific conditions or treatments are identified. To obtain quantitative information on these candidate biomarkers, a targeted step follows, where the intensity or the enrichment of the metabolites is quantified in replicate samples to further validate their involvement and gain mechanistic insight to their role in the phenotype of interest. Lastly, researchers delve into advanced approaches like MFA and network analysis, enabling the assessment of dynamic metabolic pathways, metabolic fluxes, and intricate interrelationships between metabolites [14], [62], [63].

2.3 Metabolomics workflow

Following the acquisition of MS data, subsequent steps in the metabolomics workflow require the integration and use of bioinformatic tools to infer biological insights from the data. Typically the

metabolomics workflow consists of four steps: data preprocessing, untargeted analysis, metabolite identification, and targeted analysis (Figure 2-1) [13], [64].

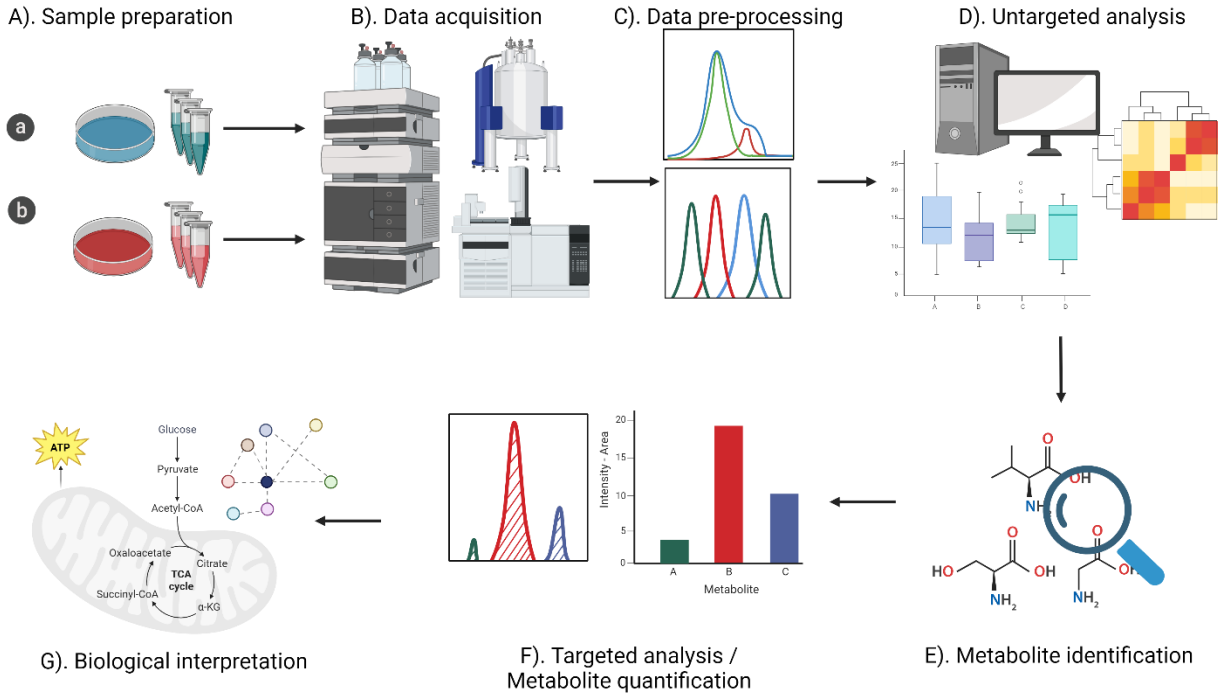


Figure 2-1. Typical metabolomics workflow. *A). Samples with differential treatments are prepared. B). Data acquisition by means of NMR or MS. C). Data preprocessing to reduce sample-specific errors. D). Untargeted analysis to determine features (i.e., combinations of retention time and m/z values) that indicate different metabolic phenotypes across samples. E). Metabolite identification. F). Metabolite quantification. G). Biological interpretation of the results.*

Some studies can adjust the workflow and omit certain steps if enough information has been previously gathered. Each step has a specific purpose and brings different types of information to the workflow. Usually, specialized software and algorithms are tailored to each step of the workflow. A review of these steps is presented below.

2.3.1 Preprocessing of MS data

While MS metadata itself might be conceptually similar across different instruments, the data formats used by distinct vendors differ significantly. This phenomenon becomes an issue as MS users encounter contrasting data structures, naming conventions, and metadata representations

across platforms and machines. To address this problem, standardization efforts have been pursued to set a common starting point for other bioinformatic tools used in later stages of the workflow [65], [66]. Currently, most MS datasets are converted into three vendor-agnostic formats: NetCDF [67] is typically used for low-resolution datasets that have limited dimensions in the metadata. When more dimensions are introduced (e.g., MS^2 or HRMS) the lack of compression makes the use of this format infeasible [68]. The mzML [69] and mzXML [70] formats have emerged as alternatives, where the former is considered a better choice due to the use of zlib compression [68]. The ProteoWizard [71] project, and specifically its msconvert tool, has emerged as the leading method to translate MS files from proprietary formats into standard formats.

Mass spectrometry data is prone to experimental errors that can affect the quantification and interpretation of the results. The possible sources of analytical errors and corrections that can be made during data preprocessing to minimize their significance are described in the following sections:

2.3.1.1 Baseline drift

The position of the signal baseline in mass spectrometry can shift between samples or over time within the same sample due to factors like temperature fluctuations, changes in solvents, or variations in mass analyzer conditions [72]. Often, this shift appears as a curve [73], impacting the accurate determination of peak area and height in the chromatogram. To correct for this baseline drift, data processing algorithms carry out linear and non-linear transformations on the raw data, fit the extracted chromatograms to "top-hat" or wavelet-like models, and implement subtraction techniques involving the use of quality control samples [74], [75].

2.3.1.2 Noise reduction

Noise in the resulting data can arise from multiple sources. Electronic noise, originating from components like amplifiers and detectors, introduces random fluctuations in the signal [76]. Variability in the ion source, influenced by factors such as temperature, pressure, and sample composition, can lead to fluctuations in the signal during ionization. Detectors, responsible for measuring ion intensity, introduce noise through factors like thermal noise (also known as Johnson noise) and dark current noise (i.e., noise that stems from a small electrical current in charge-coupled devices) [77]. Finally, unwanted signals from other compounds or environmental factors that contribute to the chemical noise within the sample matrix or instrument can also exacerbate the already existing noise [78]. Typical approaches to reduce noise through signal smoothing include the application of digital filters [79], [80] and integral transforms [81], [82].

2.3.1.3 Retention time drift correction

The performance of the chromatographic columns is not consistent throughout prolonged times and is affected by local changes in temperature or pressure, which leads to drifts in the retention times of the samples. To guarantee that the same compound elutes at the same retention time across multiple samples, the peaks appearing in the samples are aligned using a variety of algorithms that comprise warping techniques, graph-based methods, statistical techniques and deep learning methods [83], [84], [85], [86], [87].

2.3.1.4 Peak detection and extraction

As previously mentioned, the standardization of mass spectrometry data plays a crucial role in the data analysis pipeline. One of the most popular ways to achieve this standardization is to represent the mass spectrometry data as a list of all the peaks that includes their retention time (RT), *m/z*

value, and height or area. To achieve this goal, the peaks need to be detected and their properties (apex, width, height, area) need to be calculated. This is usually done using two methods: i) by scanning the data in the *m/z* and RT dimension annotating all the points above a certain threshold, which then are grouped into peaks by fitting them to wavelet or Gaussian models [88], [89], or ii) by processing the extracted ion chromatograms independently in the time domain using filters to find peak inflection points or by searching for areas above a threshold level calculated as the mean or median of the chromatogram signal [90], [91].

2.3.1.5 Data preprocessing software

Most of the aforementioned data preprocessing steps have been developed as stand-alone tools or applications. However, two packages stand out for their versatility and capability to implement all steps in a modular way: XCMS [90] and MZmine [92]. Their wide applicability has positioned them as the most powerful and widely used bioinformatic software in the metabolomics field [93]. XCMS is an R-based package that follows a standardized pipeline consisting of peak-picking, peak alignment, peak regrouping, and missing value imputation. The algorithms and the parameters used to accomplish this pipeline can be set by the user giving it an unmatched flexibility relative to other tools. There is also an online version of XCMS that offers limited capabilities but aids researchers who are not proficient in the use of command-line interface applications [94]. On the other hand, MZmine is a modular toolkit that offers peak detection, peak alignment, peak identification, visualization, normalization, isotope detection, and statistical analysis of datasets from GC-MS and LC-MS, even using data files in vendor-specific formats. The number of different algorithms that can be used in each one of these steps is also a significant advantage compared to other tools, and developers of both XCMS and MZmine constantly update their modules and algorithms to build a comprehensive toolkit that is not commonly found among other

tools. As an example, MZmine offers the feature of peak identification by comparison to online databases such as PubChem, KEGG, METLIN, HMDB, and ChemSpider.

2.3.2 Untargeted detection of candidate metabolites

Untargeted studies are usually the first step in the metabolomics pipeline and are aimed at comprehensively profiling and identifying possible relevant metabolites without prior knowledge or bias toward specific compounds [63]. This method involves the comparison of spectral peak areas or metabolite concentrations, between at least two experimental groups (i.e., control vs. test cases, treatment A vs. treatment B, etc.) [95]. The comparisons are usually done implementing univariate and multivariate statistical methods, also known as chemometrics [96], that are explained as follows:

2.3.2.1 Univariate analysis methods

Univariate methods analyze the calculated metabolomics features independently. A major disadvantage of these methods is that they do not take into account the presence of interactions between the metabolic features, and the large number of peaks that are found in a typical metabolomics experiment results in a need to correct for multiple tests [97]. In addition, the potential effect of confounding variables is not accounted for in these methods, increasing the rate of false positive and false negative results [98], [99]. On the other hand, their simplicity to use and interpret is one of the most notorious advantages. Parametric statistical tests such as t-test or ANOVA are the most common types of univariate analysis methods that are used in metabolomics [100]. Normality and homogeneity of the data are often tested using methods such as the Kolmogorov-Smirnov normality test and Barlett's homogeneity test, respectively. In cases where

the normality of the data is not confirmed, non-parametric tests such as Mann-Whitney U test or Kruskal-Wallis one-way analysis of variance are used [101].

2.3.2.2 Multivariate analysis and machine learning methods

In untargeted metabolomics studies, as large numbers of metabolites are simultaneously analyzed, it is expected that multiple metabolites are associated. To identify these relationship patterns between metabolites while taking into account all measured metabolic features, multivariate analyses are usually implemented. Most of these methods fall under the umbrella of machine learning methods and can be divided into three different groups:

2.3.2.2.1 Unsupervised methods

Unsupervised learning methods enable the identification of inherent groups or trends within data that are not pre-assigned to specific classes. The goal of these methods is to summarize, explore, and discover patterns within the data, requiring minimal prior assumptions and knowledge of it [102]. Typically, unsupervised learning serves as an initial step in data analysis, aiding in data visualization and uncovering potential issues with the experimental design by reducing the initial dimensions of the data to a subset of dimensions that share certain features in common. The most common unsupervised methods in metabolomics studies are:

i) Principal component analysis (PCA): This method substitutes a set of correlated variables with a significantly smaller number of uncorrelated variables, commonly known as principal components, that preserve the essential information present in the original dataset [103]. PCA presents a reduced and more interpretable set of variables that reflects the variance in the data without the redundancy of correlated information.

ii) Clustering: This method aims to identify groups in the original dataset by grouping parts of the data that share common features or trends in the same cluster. A variety of algorithms have been created for this purpose, leveraging different similarity indicators, such as distance between data dimensions and correlation coefficients. K-means and hierarchical clustering are the most popular clustering methods. The former uses a centroid-based algorithm to separate the data into multiple clusters while the latter builds a hierarchical structure in dendrograms [102].

iii) Self-organizing maps: Although this method is not very popular, it is a powerful tool to visualize high-dimensional data [104]. Similar to PCA, the idea behind self-organizing maps is to transform high-dimensional data into a lower- (usually two-) dimensional representation while preserving the inherent relationships among data points. This is done by creating networks that consist of nodes arranged in a grid, where each node represents a specific region in the input space. During the training steps, the self-organizing map learns to adapt the weights of the relationships between nodes, clustering similar data points close to each other on the map.

Unsupervised methods have been widely used in the context of metabolomics research. By means of PCA applied to the quantification of metabolites in *H. canadensis*-based dietary supplements, one study was able to train a method capable of detecting adulteration of botanical dietary supplements [105]. Using hierarchical clustering on the measured blood-metabolites of cancer patients, one study was able to divide the participants into a high-risk and a low-risk group based on their cancer progression and survival, suggesting possible metabolites that could be used as biomarkers to determine the risk stratification in head and neck cancer [106]. Similar studies have modeled the survival rate from chemotherapy based on subsets of metabolites in non-metastatic

breast cancer [107]. Different unsupervised clustering methods were used in metabolomics data coupled with proteomics data from *Chaetomium thermophilum* identifying 461 new key protein-metabolite interactions yielding a deeper understanding of the metabolism of the organism [108]. Finally, self-organizing maps have been used to visualize metabolic changes in breast cancer tissues [109] and to improve the clustering and elucidation of unexplored metabolic pathways using information from model organisms as a reference [110].

2.3.2.2.2 Supervised methods in metabolomics

Supervised methods, are typically used in biomarker discovery, classification, and prediction in datasets with response variables. The main goal of supervised methods is to determine the relationships between the response variables and predictors (also known as covariates) for accurate predictions. The term "supervised" implies the need for one or more response variables to guide model training. The response variables can be discrete (e.g., control vs disease groups) or continuous (e.g., metabolite concentration or gene or protein expression levels). In the first case, the algorithm solves a classification problem, whereas the second case solves a regression problem. This process is usually divided into two steps: first a training step, where algorithms fit a model to the training dataset, and a second testing step, where the accuracy of the predictions is evaluated using a separate testing dataset [111], [112].

Notable supervised methods include Partial Least Squares (PLS), that can be used both as a regression analysis or as a binary classifier (in its “discriminant analysis” form). As opposed to PCA, PLS does not aim to maximize the variance between the data, but the covariance between the response variable and the metabolomic data. Hence, the measured features (also known as loadings) of PLS components reflect the degree of contribution of a feature to the discrimination of the different experimental groups [113].

Another widely used method is the classification and regression through Support Vector Machines (SVM). The functioning of SVMs is based on the decomposition of the measured features in a plane to find an optimal hyperplane that separates the data points. In a binary classification scenario, this hyperplane is set to maximize a margin, which is defined as the distance between the hyperplane and the nearest data points from each class, which are called support vectors, as seen in Figure 2-2. The support vectors play a crucial role in defining the decision boundary. SVMs are effective in handling non-linear relationships using mathematical transformations of the data through kernel functions, which transform the input data into a higher-dimensional space where a hyperplane can effectively separate the classes.

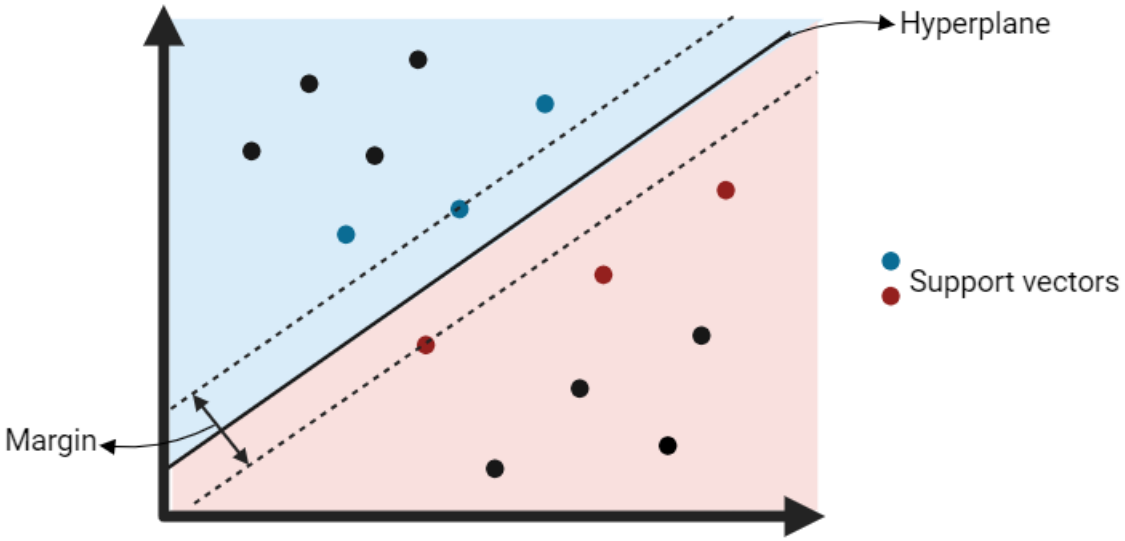


Figure 2-2. Graphical representation of a binary classification using support vector machines. *A hyperplane is created with the aim of separating the data into two groups. A margin is defined as the distance between the hyperplane and the closest data point in each section. The support vectors are defined as the nearest points to the margin.*

Recent studies have used these methods to elucidate biomarkers and determine their relationships with specific diseases. One study used variations of PLS on metabolites in urine to detect and predict two genetic inborn errors of metabolism, methylmalonic acidemia and isovaleric acidemia. The results correlated the concentrations of methylmalonic acid, methylcitric acid, and 3-OH-

propionic acid as potential metabolites that could be used to diagnose the aforementioned conditions [114]. Analogously, SVMs have also been used in the discovery of biomarkers to diagnose hepatocellular carcinoma based on the concentration of ten phospholipids in liver tissues [115] and to predict the survival rate of chronic obstructive pulmonary disease based on the presence or concentration of 79 plasma metabolites [116].

2.3.2.2.3 Deep learning methods in metabolomics

Finally, the use of deep learning algorithms has seen a rise in recent years thanks to the advanced capabilities of new computing systems [117]. Deep learning involves the use of neural networks with multiple layers (known as deep neural networks) to model complex systems that cannot be resolved with simpler methods. Each layer in the network transforms the input data and progressively extracts more abstract features, allowing the model to capture patterns and relationships within the data. Similar to supervised methods, deep learning requires a training step that involves adjusting the weights and biases of the network through backpropagation, learning from known training data to make predictions on new, unseen data.

Applications of deep learning in metabolomics have been seen in every step of the data analysis pipeline. A recent study proposed a model trained with neural networks to enhance the quality of the extracted peaks and reduce the number of false peaks in the preprocessing step of the workflow [118]. Another study created a model that is able to detect the presence of a pathogen, *L. monocytogenes,* based on the presence and concentration of certain metabolites analyzed with LC-MS [119]. Finally, deep learning models have been trained with data from three spectral databases, the Golm Metabolome Database, the Human Metabolome Database, and FiehnLib, to help in the identification of metabolites detected via GC-MS [120]. This is a process that is located between the untargeted and targeted steps within the metabolomics workflow.

2.3.2.3 Untargeted detection of labeled metabolites

In the context of stable isotope-resolved metabolomics, the vast majority of the peaks that are produced in a MS experiment will belong to metabolites that are not relevant to the experiment given their lack of isotope enrichment. Hence, the untargeted detection of labeled species becomes crucial in this type of study. The goal of an algorithm capable of detecting labeled metabolites in an untargeted manner is to receive MS datasets and i) group the peaks corresponding to the multiple isotopologues into their putative compounds and ii) detect labeling patterns by means of the previously mentioned methods.

To our knowledge, there is only one publicly available tool that serves this purpose: X13CMS [121] is an extension of XCMS with the main goal of detecting metabolites that have been enriched with a single isotope. This functionality is based on finding the statistically different enrichment patterns of the same compound across multiple samples, including an unlabeled sample. However, limitations within this software may constrain its applicability in metabolomics studies. Notably, X13CMS does not assign molecular formulae or structures to features, making it impossible to correct for the abundance of naturally occurring isotopes like ^{13}C. In studies employing non-^{13}C tracers that lack the resolution to resolve the mass defect between isotopes, this limitation introduces a potential source of error. Furthermore, the tool requires an unlabeled control sample for comparison, introducing biological variability between unlabeled samples as a source of error. Finally, X13CMS does not map results to metabolic pathways, limiting the analysis to a detection of labeled compounds [122].

2.3.3 Metabolite identification

The groups of *m/z* values and their respective retention times that are found in targeted studies are meaningless if they are not correlated to a metabolite identity. Hence, typically every untargeted analysis is immediately followed by an identification step. The identification of metabolites (in some studies referred to as metabolite annotation) can be done at different levels of confidence depending on the nature of the tools and data that are used in the process. To stratify these levels of confidence, a common consensus of "identification levels" has been embraced by metabolomics researchers. Figure 2-3 shows the hierarchy of the identification levels. Higher levels reflect poor and uncertain identifications, whereas lower levels reflect high-quality and specific identifications [63], [123]. Level 5 identification relies solely on the comparison of the exact mass of interest and can be done using only the *m/z* values of peaks detected in an HRMS experiment. A level 4 identification refines the guess to an unequivocal molecular formula using information of the isotopic information gathered from MS. A level 3 identification reflects a tentative candidate where a structure, or a chemical class, has been identified using MS^2 datasets. A level 2 identification portrays a probable structure after comparing the fragmentation fingerprints of the molecules against libraries of MS^2 data. Finally, a level 1 identification reflects a confirmed structure by a reference standard where the fragmentation spectra and the retention times in the experiment were compared and confirmed against the measurements of a pure standard [37].

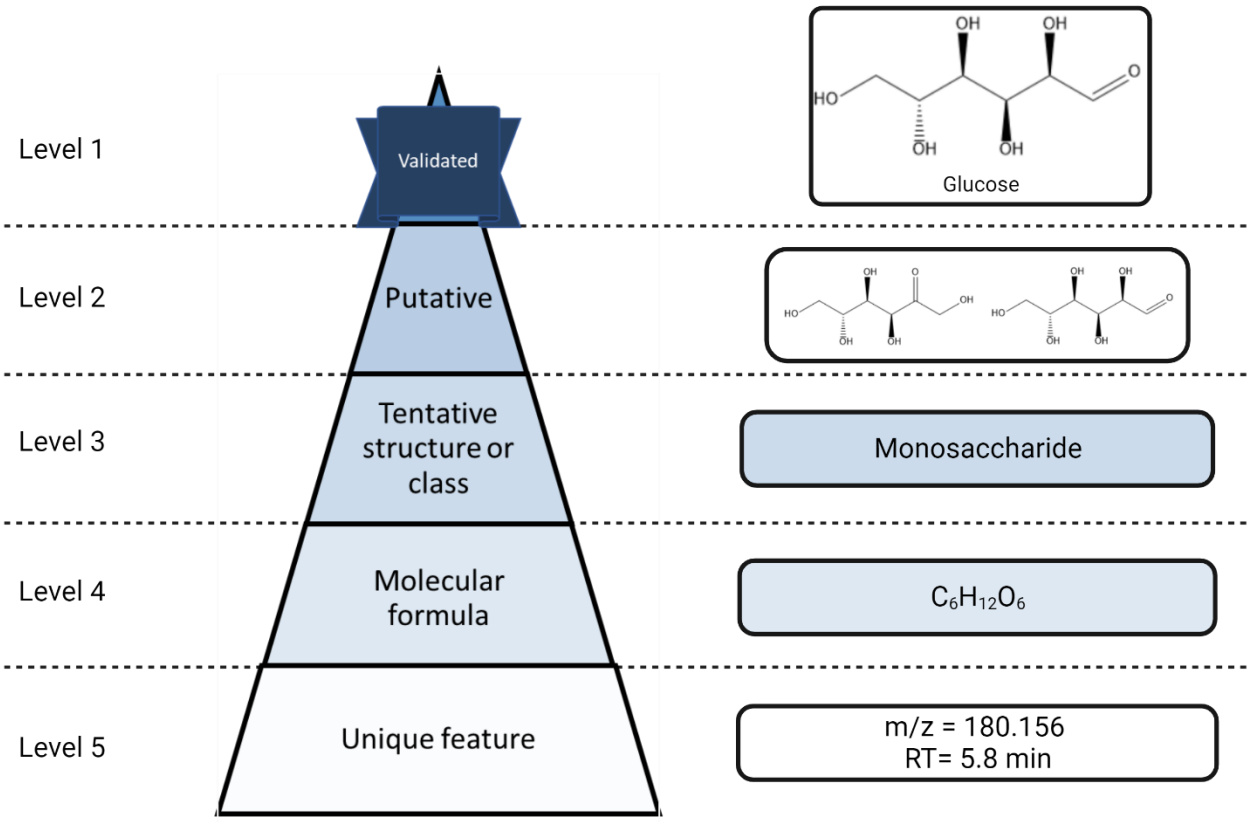


Figure 2-3. Identification confidence levels in MS datasets. *As lower levels are reached by means of the gathering of additional information, the identification becomes more accurate. Adapted from [37]*

While several bioinformatic tools have been developed to perform the comparisons needed to identify metabolites based on their MS spectra, three tools stand out for their advanced capabilities and modular workflows: CAMERA [124] is an R-based tool designed as a complement to XCMS that offers additional grouping algorithms based on the retention time, shape, and isotopic patterns of the peaks. All the isotopic patterns are based on a fixed mass which does not consider the mass defect of atoms other than carbon, limiting the analysis when using stable isotope-based datasets. The advantage of using positive and negative mode data simultaneously offers more accurate level 3 annotations. MS-DIAL [125] provides a compilation of multiple MS^2 spectral libraries to obtain level 1 identifications by spectral comparison. It also has built-in algorithms that pre-process the data to optimize the identification, yielding better results. For the peaks that do not have a spectral

match, a putative level 3 identification is given based on the parent ion and its isotopic patterns. Finally, SIRIUS [126] is an alternative that uses the isotopic patterns of multiple fragments from HRMS spectra to estimate the chemical formula of a compound aided by the database PubChem. Newer versions of this tool have been combined with other tools like CSI:FingerID [127] to support HRMS and MS^2 datasets in a more accurate manner.

2.3.4 Targeted metabolite quantification via chromatogram integration

To reduce the bias in the peak detection and pick picking algorithms, once a metabolite has been identified from untargeted studies it is subsequently quantified using targeted approaches. Measurements like the absolute area, the relative area compared to an internal standard, peak height, and relative abundance of its corresponding isotopologues are usually quantified with specialized software. However, some software tools rely on the manual and separate integration of each mass isotopomer, which has been demonstrated to yield quantitative errors if the parameters of the baseline correction or integration bounds are not consistent across all mass isotopomers of a single compound [128], [129]. For this purpose, proprietary tools are mostly used. Agilent MassHunter, Sciex MultiQuant, and ThermoFisher XCalibur are the most used options that appear in the metabolomics literature, despite not being optimized for stable isotope-tracking and quantification.

Some publicly available tools have arisen as viable options to perform targeted analysis in a semi-automated way: MAVEN [130] is one example that offers visualization and quantification of targeted LC-MS data; it was later refined and republished as a stand-alone tool, El-MAVEN [131] offering support for GC-MS, HRMS, and MS^2 data. El-MAVEN was integrated with Polly™, which is a biomedical data platform that allows the integration of MS data with other omics data. Finally, the Skyline project [132] was originally developed as an open-source stand-alone tool to

analyze MS-based proteomics datasets. Since then, the subsequent updates of the software have refined the algorithms to make it suitable for metabolomics analyses. Despite none of these tools being optimized for stable-isotope quantification, they are still being used as the most reliable tools to integrate the peaks of mass isotopomers, leaving the isotope enrichment and MID quantification as manual tasks for the user.

2.3.5 Natural abundance correction

The calculated MIDs need to be corrected to remove the impact of natural isotopic background when interpreting data from metabolic tracer experiments. Approximately 1% of the carbon atoms, 0.36% of the nitrogen atoms, and 0.2% of the oxygen atoms exist as naturally occurring heavy isotopes [133]. These values might not seem significant by themselves, but in large molecules that have multiple atoms, they are enough to skew the results of the measured MIDs. Approaches to correct the measured MIDs are based on the construction of a natural abundance correction matrix *CM* and the solution of the following linear algebra problem [134], [135] :

$$MID_u = MID_c\,[CM] \qquad \text{(Eq. 2-1)}$$

Where the subindex *u* stands for 'uncorrected' and *c* for 'corrected'.

The determination of the correction matrix relies on the nature of the utilized data. Recent investigations have provided insights into computing the correction matrix for MS^2 data [136] and experiments involving multiple labeled atoms in the tracers [137]. As a consequence, various software tools have been developed to correct for natural isotope abundance in specific MS data types [138], and certain peak integration tools enable the natural abundance correction of limited types of MS data [130], [131]. However, to our knowledge, there is currently no tool capable of

simultaneously quantifying and correcting the peaks from all the common types of MS acquisition modes used in stable isotope-based metabolomics (i.e., low-resolution, high-resolution, MS^2, and HRMS), adding yet another step of data processing to the metabolomics pipeline.

2.3.6 Association of metabolites to metabolic pathways

The information of the intensity and enrichment of metabolites offers valuable information by itself, but in order to extract relevant biological information it is imperative to frame these data in the context of the pathways the metabolites belong to. Recently, a handful of bioinformatic tools have been developed to integrate genomics and proteomics studies into the context of metabolic pathways, but the ones that are designed to use information based on metabolomics information are very scarce [139]. In the context of stable isotope-based metabolomics, one recent tool was used to leverage differential abundance of the measured compounds to elucidate active pathways. Mummichog [140] uses the relative position of the metabolite candidates in a reconstructed metabolic network to refine their identity, but most of the attention has fallen into the elucidation of active metabolic pathways given a list of identified metabolites. The use of this tool allowed researchers to determine the regulation mechanisms behind autophagy in virus-specific CD8 T-cells during acute lymphocytic choriomeningitis virus infection [141] and to determine associations between almost 1700 metabolites and the age, sex, and genotype of *Drosophila melanogaster*. Posterior multivariate analysis showed that individual metabolic profiles could be used to predict these traits with high accuracy [142].

2.3.7 Advanced analyses in metabolomics

While simple targeted metabolomics provides a snapshot of the current metabolic state of an organism, it is unable to capture the dynamic aspects of metabolism. To resolve the dynamic

changes and rates at which the metabolites are synthesized or consumed, more advanced analyses such as kinetic flux profiling or metabolic flux analysis become essential [143]. Both approaches entail the tracking of stable isotopes across multiple targeted measurements at different timepoints, enabling the observation of fluxes through metabolic pathways.

Kinetic flux profiling fits the behavior of the unlabeled isotopologue (M0) of each metabolite as a function of time to exponential decay models. The estimated fractional turnover rate is multiplied by the metabolite concentration, and the resulting value can be approximated to the value of the metabolic flux (i.e., the conversion of a metabolite through a reaction over time) [144].

On the other hand, metabolic flux analysis uses information on the labeled isotopologues of each metabolite, following the transitions of each atom or group of atoms in their corresponding reactions to determine the metabolic fluxes [145]. This is done by solving a set of stoichiometric equations based on the premise that in metabolic steady state the sum of the fluxes entering one metabolite is equal to the sum of fluxes that emanate from it [146]. The advantage of using additional information on multiple labeled isotopologues instead of a single unlabeled isotopologue is that it can clearly resolve branching points where a single metabolite serves as a precursor to multiple reactions.

An in-house developed tool, INCA [16], is designed to calculate metabolic fluxes in experiments involving stable isotopes, but it relies on accurate measurements of metabolite MIDs. INCA's algorithms are compatible with both MS^2 and HRMS datasets, offering insights into complex metabolic details by tracking multiple tracers and exploring new pathways and metabolites discovered through untargeted methods. Therefore, having reliable bioinformatics tools to assist researchers at earlier steps of the process becomes crucial.

2.4 Conclusions

Stable isotope-based metabolomics is a powerful field that enables the comprehensive understanding of fundamental biological processes and reactions, which could be used for pharmaceutical and synthetic biology applications. The data analysis workflow includes: i) a pre-processing step, ii) an untargeted step to find potential targets and metabolites of interest, iii) a metabolite identification step to determine the accurate identity of the detected biomarkers, iv) a targeted step where the identified markers are quantified, and vi) an advanced analysis to extract biological information from the results.

While several bioinformatic tools have been developed to aid in specific parts of the workflow, most of them are limited by the type of data that they can process and do not fully support the latest developments in MS technology such as MS^2 and HRMS. Hence, there is still a lack of comprehensive tools that are capable of fully leveraging these technologies and enabling the use of multiple isotopic tracers in a single experiment for the discovery of new biomarkers and active pathways, while simultaneously offering support for classic methods such as low-resolution MS.

This dissertation presents the development and application of two bioinformatics tools tailored for untargeted and targeted metabolomics analyses. The untargeted tool, SUNDILE, aided by supervised machine learning methods, detects metabolites enriched by a labeled tracer without constraints on the number or type of labeled atoms. Additionally, leveraging a KEGG-based network reconstruction, SUNDILE proposes identities for the detected labeled metabolites without the need to gather MS^2 data and extends its findings to a pathway-level analysis, suggesting new enriched pathways beyond the commonly explored canonical pathways. On the other hand, the targeted tool PIRAMID can accurately quantify chromatographic peak areas of target metabolites. PIRAMID calculates metabolite MIDs and enrichments, offering support for the most common

MS technologies employed in metabolomics experiments, including GC-MS, LC-MS, low-resolution MS, HRMS, single MS, and MS^2. Applications on plant and mammalian organisms are presented discussing the advantages and limitations of the developed software.

CHAPTER 3

UNTARGETED METABOLITE IDENTIFICATION AND PATHWAY ANALYSIS USING SUNDILE

Abstract

Untargeted metabolomics relies on the comparison of datasets under different experimental conditions followed by metabolite identification. The identification of unknown metabolite peaks in the spectral data is time-consuming, as it often requires the additional collection of MS^2 spectra, and in some cases the determination of the retention times and *m/z* values of known metabolites through the analysis of mixtures of metabolite standards. In the context of stable isotope-based metabolomics, most identified compounds do not exhibit isotope enrichment, which makes them irrelevant in the subsequent analysis of labeling dynamics and metabolic flux. This chapter presents SUNDILE, a software tool that detects compounds that exhibit stable isotope enrichment and provides a putative identity based on an *in silico* metabolic network reconstruction. SUNDILE processes a .csv list with the extracted peaks from mass spectrometry-based metabolomics experiments and provides a .xls file with a list of labeled compounds, their identities, and a score of metabolic pathway activity. Furthermore, SUNDILE can process data from stable isotope experiments that incorporate one or more isotopic labels. The capabilities and algorithms of the software are validated using two in-house datasets.

3.1 Introduction

Stable isotope labeling in metabolomics can elucidate dynamic changes in metabolism through the measurement of metabolic fluxes [147], [148], [52]. The latest research in this field focuses on specific pathways such as the urea and tricarboxylic acid (TCA) cycle, glycolysis, and the pentose phosphate pathway [149], leaving other pathways underexplored. Most of these studies require *a priori* knowledge of the isotopically enriched metabolites that will be analyzed through mass spectrometry (MS) [52], [150], [151]. As such, MS typically yields thousands of peaks, but only a narrow set of metabolites from targeted pathways are used to assess metabolic activity. The asymmetry between total measurement quantity and the limited number of measurements that are subjected to further analysis results, in part, from the impracticality of a global, untargeted analysis of all isotopically labeled metabolites detected by MS [152], [153], [154].

Untargeted metabolomics of isotope labeling experiments (ILEs) has grown in popularity to maximize the richness of MS datasets and discover novel pathway activities. These studies leverage recent advances in MS technology [155] that allow scientists to identify the mass of compounds with atoms enriched from single (e.g., ^{13}C) or dual (e.g., $^{13}C/^{15}N$) heavy isotopes with a high degree of accuracy [156], [157]. Most untargeted metabolomics of ILEs follow a similar workflow: organisms of interest are divided into two groups, one that will receive a stable isotopic label and one that will not (Fig. 3-1-A). Compounds are extracted and analyzed to detect mass differences of labeled compounds via MS or nuclear magnetic resonance (NMR) (Fig. 3-1-B). MS data are preprocessed to limit noise and bias by removing artifacts like baseline intensity of chromatograms, convolved peaks, or retention time drifts. The output from this preprocessing step is a list of features (i.e., duplets of *m/z* and retention time) and their respective properties (Fig. 3-1-C). The preprocessed data are then analyzed using specialized software that relies on statistical

comparisons between labeled and unlabeled samples. Newer software uses machine learning algorithms to facilitate these comparisons, which provide a list of flagged features (i.e., 'hits') that show a statistical difference between conditions (Fig. 3-1-D).

The next step in the process is considered one of the major analytical bottlenecks: *compound identification*. Starting from the *m/z* and retention time, the identities of the top hits are found using other techniques that carry multiple possible levels of confidence (i.e., techniques that use the comparison of fragment ion spectra to those of known standards yield higher confidence than techniques that only compare the monoisotopic mass to a theoretical value) [37], [63]; this step is particularly onerous since most identifications are based on comparisons against references, and the lack of comprehensive libraries leaves many identities unsolved (Fig. 3-1-E). This overall process can be implemented to test many biological hypotheses (e.g., the impact of gene knockout, drug treatment, etc.) (Fig. 3-1-F); however, the reliability of these outcomes will be heavily influenced by the accuracy of compound identification.

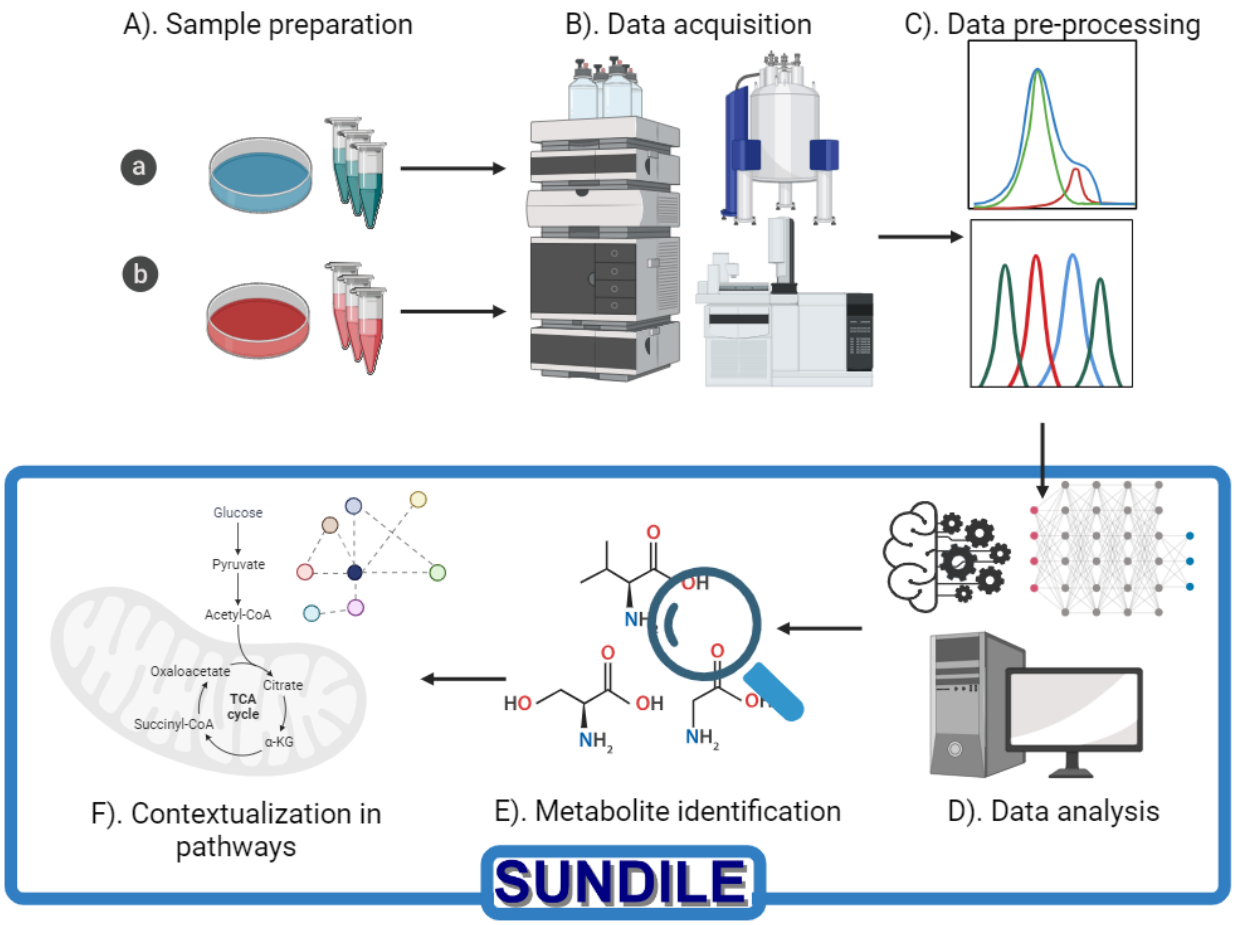


Figure 3-1. General untargeted metabolomics workflow. *A) Data from experimental groups is B) gathered using spectrometry-based analytical instruments (i.e., MS or NMR). C) The data must be preprocessed before it can be D) analyzed using specialized software that detects differences between groups. E) The compounds exhibiting differences are identified using their respective spectra and F) a biological interpretation of the experiment is built based on the detected hit compounds. We have developed new software, SUNDILE, that aids in data analysis and metabolite identification steps, which are major bottlenecks in the metabolomics workflow.*

A handful of bioinformatic tools have been developed to ease this workflow (Fig. 3-1), many of which specialize in a single step. Some tools are optimized to extract and annotate metabolites from MS datasets [124], [126], others to bin the preprocessed features into groups of isotopologues that account for their labeling [158], [159], and others focus on a broader scale, highlighting the active pathways in the experiment [149]. However, most have not been updated to support the latest technological advances (e.g., experiments with two or more isotope labels and high-resolution MS data).

With these limitations in mind, this chapter presents the development of a graphical user interface (GUI)-based software package (SUNDILE: Software for UNtargeted analysis of Dynamic Isotope Labeling Experiments) that improves the data analysis and identification steps in the untargeted ILE analysis workflow. SUNDILE receives an input of preprocessed and extracted features, including their intensities and retention times. SUNDILE then bins these features into compounds that exhibit a shift in mass isotopomer distribution (MID) due to isotopic labeling and provides a putative compound identity. Lastly, a pathway enrichment score is provided that describes the extent of compound labeling within annotated metabolic pathways. SUNDILE can process high-resolution MS datasets of metabolites enriched with one or more labeled atoms (e.g., due to ^{13}C and ^{15}N tracers administered in the same experiment) and where multiple sample time points are compared. The software was validated against previously reported results and tested using metabolomics data from a dual-isotope labeling experiment.

3.2 Methods

3.2.1 Soybean ILEs with dual-labeled [13C5,15N2]glutamine

Soybean (*Glycine max*) seeds from the cultivar PI 603338, *Glycine max* (L.) Merr., were used to generate this dataset. The seeds were planted in a 1-gallon pot containing Farard 4M soil and were grown under supplemental lighting in growth chambers ensuring a 14h/10h day/night photoperiod with temperatures ranging from 25°C−27°C/21°C−23°C and a humidity above 45%. Pods were harvested from each plant at multiple stages of reproductive development starting from the early seed-filling stage (R5) and placed on ice. Embryos were extracted from the pods, dissected from the seed coats, and transferred to sterile labeled culture medium. A modified Linsmalier and Soog medium [160], [161] with Gamborg's vitamins (Sigma) enriched with dual-labeled [$^{13}C_5$,$^{15}N_2$]glutamine was used as the culture medium. Culturing was performed under continuous

light at an intensity of 30–35 μE and a temperature of 26°C, and tissue was collected at multiple time points (0, 30, 60, 120, 240, 480, and 1200 minutes) following tracer administration. At the end of the labeling experiment, the seed metabolism was quenched by a rinse of the cotyledon surface with water prior to slicing and rapid freezing in liquid nitrogen and stored at -80°C until metabolite extraction.

Metabolites were extracted following a previously described protocol [162], and full-scan MS^1 data were acquired using a Thermo Scientific Q Exactive mass spectrometer at a resolution of 280,000 FWHM with a SeQuant ZIC HILIC column in positive and negative ionization modes analyzing an *m/z* scan range up to 1000 Da. The elution was performed at a flow rate of 0.25 mL/min with an increasing gradient of acetonitrile: 10 mM ammonium acetate (90:10, v/v) and 5 μM medronic acid, pH 9.0 (A) and 10 mM ammonium acetate in water, pH 9.0 (B). The HPLC eluent was introduced into an ESI source with the following conditions: ion spray voltage, 4.5 kV (ESI-); ion source temperature, 550°C; source gas 1, 45 psi; source gas 2, 40 psi; curtain gas, 35 psi; and entrance potential, 10. MS^2 data was acquired for the unlabeled sample (t=0) only to enable confirmation of compound identities by comparison to standard libraries.

These experiments were performed by collaborators at the Donald Danforth Plant Science Center. (St. Louis, MO).

3.2.2 T cell ILE with [^{13}C]glucose

C57BL/6, CD90.1 (Thy1.1$^+$), and OT-I transgenic mice (The Jackson Laboratory. Bar Harbor, ME) were bred and maintained under specific pathogen-free conditions at the Van Andel Institute in compliance with approved institutional protocols. Experiments were conducted using mice

within the age range of 8 and 12 weeks. The adoptive transfer and infection with *L. monocytogenes* was performed following previously described protocols [163], [164], [165].

T-cells were isolated from spleen and peripheral lymph nodes by negative selection (StemCell Technologies, Vancouver, BC) following a previously described protocol [166]. Cells were cultured in Iscove's Modified Delbecco's Medium lacking sodium pyruvate (IMDM) or Van Andel Institute-modified Iscove's Medium (VIM) supplemented with 10% dialyzed fetal bovine serum (FBS) (Wisent, St. Bruno, QC), penicillin-streptomycin (Invitrogen), and 2-mercaptoethanol (Sigma-Aldrich, St. Louis, MO). Glucose, L-glutamine, and a mixture of nutrients found at concentrations higher than 100 μM in mice (i.e., acetate,b-hydroxybutyrate [bOHB], citrate, lactate, and pyruvate) were added to cell culture media at previously described concentrations [165]. The activated cells were produced by stimulating naive CD8+ T cells ($1x10^6$ cells/mL) with plate-bound anti-CD3ε (clone 2C11) and anti-CD28 (clone 37.51) antibodies (eBioscience, San Diego, CA) and 50 U/mL IL2 (PeproTech, Rocky Hill, NJ) for 3 days.

The activated T cells were washed in IMDM or VIM containing 10% dialyzed FBS, and re-cultured ($2.5x10^6$ cells/well in 24-well plates) for 24 h containing 25mM [^{13}C]glucose in IMDM or 5mM [^{13}C]glucose in VIM (Cambridge Isotope Laboratories).

Metabolites were extracted by modified Bligh-Dyer extraction [167] by the addition of ice-cold methanol (A456, Fisher Scientific) and an equal volume of chloroform (A456, Fisher Scientific) directly to frozen cells. The sample was vortexed for 10 sec, incubated on ice for 30 min, and then 0.9 parts of LC-MS grade water (W6-4, Thermo Fisher Scientific) was added. The samples were vortexed and centrifuged at maximum speed to achieve phase separation. The top layer containing polar metabolites was aliquoted into a fresh tube and dried in a speedvac for LC-MS analysis. The bottom layer was not analyzed.

For LC-MS analysis, metabolite extracts were resuspended in 50 µL of 60% acetonitrile (A955, Fisher Scientific) and analyzed by high resolution accurate mass spectrometry using an ID-X Orbitrap mass spectrometer (Thermo Fisher Scientific) coupled to a Thermo Vanquish Horizon liquid chromatography system. An Acquity BEH Amide (1.7 µm, 2.1 mm x 150 mm) analytical column (#176001909, Waters, Eschborn, Germany) fitted with a pre-guard column (1.7 µm, 2.1mm x 5 mm; #186004799, Waters) using an elution gradient with a binary solvent system was used as a chromatographic separation method. Solvent A consisted of LC-MS grade water (W6-4, Fisher), and Solvent B was 90% LC-MS grade acetonitrile (A955, Fisher). For negative mode analysis, both mobile phases contained 10 mM ammonium acetate (A11450, Fisher Scientific), 0.1% (v/v) ammonium hydroxide, and 5 µM medronic acid (5191-4506, Agilent Technologies). For positive mode analysis, both mobile phases contained 10 mM ammonium formate (A11550, Fisher), and 0.1% (v/v) formic acid (A11710X1, Fisher). For both negative and positive mode analyses the 20-min analytical gradient at a flow rate of 400 µL/min was: 0–1.0 min ramp from 100% B to 90% B, 1.0–12.5 min from 90% B to 75% B, 12.5–19 min from 75% B to 60% B, and 19–20 min hold at 60% B. The H-ESI source was operated at spray voltage of 2500V for negative mode acquisition and 3500V for positive mode.

High resolution MS^1 data was collected with a 20-min full-scan method with *m/z* scan range using quadrupole isolation from 70 to 1000 *m/z*, mass resolution of 120,000 FWHM, RF lens at 35%, and standard automatic gain control (AGC). Unlabelled control samples were used for data dependent MS^2 acquisition for compound identification and annotation via the AquireX workflow (Thermo Scientific) using MS1 resolution at 60,000, MS2 resolution at 30,000, intensity threshold at $2.0x10^4$, and dynamic exclusion after one trigger for 10s.

These experiments were performed by collaborators at the Van Andel Institute (Grand Rapids, MI).

3.2.3 HRMS analysis of 13C-labeled mouse liver samples

Unlabeled and [$^{13}C_3$]propionate-labeled frozen liver samples from a previous study [168] were thawed and their metabolites were extracted following a biphasic methanol/water/chloroform extraction. The polar layer of the extract was isolated and air-dried before being reconstituted in a 9:1 water:acetonitrile mixture.

A Thermo Scientific Q Exactive mass spectrometer equipped with a Waters XBridge Amide HILIC column (2.1 × 100 mm, 3.5 µm) using mobile phase A containing 9:1 water:acetonitrile + 5mM ammonium formate and mobile phase B containing 9:1 acetonitrile:water with 5mM ammonium formate at a flowrate of 200 µL/min was operated in negative mode at a resolution of 120,000 FHWM was used to acquire the mass spectra in MS^1 full scan mode and in MS^2 DDA mode monitoring masses between 60 and 500 Da.

The MS raw files were converted to .mzml format using the ProteoWizard MSConvert tool. The extraction of peaks was done using XCMS using the following parameters obtained through the software IPO [169]:

method: CentWave, prefilter= [2,500], ppm=20, snthresh=2, peakwidth= [5,100], and noise=0. A retention time correction step was included using the obiwarp method.

Using the resulting XCMS object, the software $X^{13}CMS$ was executed with the following parameters:

ppm = 3, massOfLabeledAtom = 12, noiseCutoff = 10000, alpha = 0.05.

Metabolites were identified at confidence level 1 using MS^2 data analyzed with MS-Dial by comparing against the library "*ESI (-)-MS/MS from standards+bio+in silico*" that is available on the software website using the default parameters.

3.2.4 Training of a machine learning regressor to determine the maximum atom number

Data from 111,375 metabolites from the PubChem Pharmacology and Biochemistry database [170] was curated to discard the compounds for which the molecular formula information was not available. The monoisotopic mass of the remaining compounds and the number of individual hydrogen, carbon, nitrogen, and oxygen atoms were calculated from their molecular formula. This data was discretized into 100 groups of molecular masses, and the maximum number of each of the atoms was assigned to each one of the resulting groups of masses.

The built-in MATLAB "Regression learner" application was used to compute the regressor models on the discretized data. A 5-fold cross-validation was implemented to avoid overfitting. Four groups of machine-learning methods were tested and the root-mean-squared error (RMSE) was used as a comparative parameter to choose the best method among the following:

i) **Linear regression:** Simple linear regression, linear with interaction terms, robust linear regression, and stepwise-solved linear model.

ii) **Regression trees:** Fine tree (leaf size=4), medium tree (leaf size=12), and coarse tree (leaf size=36).

iii) **Support Vector Machines (SVMs) using the following kernels:** Linear, quadratic, cubic, fine Gaussian (scales with squared number of predictors over 4), medium Gaussian (scales with squared number of predictors), and coarse Gaussian (scales with squared number of predictors times 4).

iv) **Ensemble of trees:** Boosted trees (LSBoost algorithm) and bagged (Bootstrap-aggregated).

v) **Gaussian Process Regression using the following kernels:** Squared exponential, matern 5/2, exponential, and rational quadratic.

3.2.5 Training of a machine learning classifier to assess the validity of a compound

Using the information of all the 18,815 metabolites that are reported in KEGG, similar data as the previous section was extracted (formula, number of hydrogen, carbon, nitrogen, and oxygen atoms). The theoretical MID of the compounds was calculated and the ratio between the first two isotopologues (M1/M0) was computed. From this data a supervised method was developed to predict if a compound is valid based on the M1/M0 ratio. As only positive data are available (i.e., data showing how a compound is supposed to appear), a one-class support vector machine classifier was used. This step was performed using the built-in MATLAB function "fitcsvm" with an automated kernel function.

3.2.6 Fitting of sigmoidal models and sigmoidal model selection

When applying sigmoidal models to fit the isotope enrichment data, the built-in MATLAB "*fmincon*" solver was applied to minimize the weighted sum of squared residuals (SSR) subject to user-specified constraint boundaries, using the inverse of the variance of the data as weights. Numerical calculations of the Jacobian were performed based on the partial derivatives of the SSR with respect to the fitted parameters. These derivatives were used to calculate a variance-covariance matrix, and its diagonal elements were extracted and used to calculate confidence intervals of the fitted parameters. To find the best sigmoidal model to be used in the program, the

Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were used to compare the models described by Equations 3.1a-f:

$$\mathrm{M0}(t,\mathbf{b}) = b_2 + (b_1 - b_2)\left(1 + \exp\left(\frac{t-b_3}{b_4}\right)\right)^{-1} \quad \text{Boltzmann model (Eq. 3.1a)}$$

$$\mathrm{M0}(\mathrm{t},\mathbf{b}) = 1 - (1 - \mathrm{b}_1)\left(1 + \exp\left(\frac{b_2-t}{b_3}\right)\right)^{-1} \quad \text{Modified Boltzmann model (Eq. 3-1b)}$$

$$\mathrm{M0}(\mathrm{t},\mathbf{b}) = \frac{\mathrm{b}_1}{b_1-b_2} b_3(\exp(-b_1 t) - \exp(-b_2 t)) + b_4 \exp(-b_2 t) \quad \text{Markov model (Eq. 3-1c)}$$

$$\mathrm{M0}(\mathrm{t},\mathbf{b}) = \exp(-b_1 t) \quad \text{Exponential decay model (Eq. 3-1d)}$$

$$\mathrm{M0}(\mathrm{t},\mathbf{b}) = \mathrm{b}_1 + (1 - \mathrm{b}_1)\exp\left(-exp\left(\frac{t-b_2}{b_3}\right)\right) \quad \text{Gompertz model (Eq. 3-1e)}$$

$$\mathrm{M0}(\mathrm{t},\mathbf{b}) = \mathrm{b}_1 b_2 \exp(b_3 t)\,(b_! + b_2(-1 + \exp(b_3 t)))^{-1} \quad \text{Verhulst model (Eq. 3-1f)}$$

In these equations, M0 is the fractional abundance of the monoisotopic mass isotopologue (i.e., the isotopologue with no tracer atoms incorporated), *t* is time, and **b** is a vector of the adjustable model parameters. The modified version of the Boltzmann model was proposed to reduce the number of parameters in the fit. It is based on the assumption that the sigmoidal function will converge to 1 as *t* approaches 0. This assumption is valid considering that the data at *t*=0 comes from unlabeled samples and has been corrected empirically to remove the contribution from natural isotopic background (see section 3.3.1).

The soybean and murine T cell datasets were processed to extract M0 values as a function of time for a total of 402 labeled compounds. These values were fit to the aforementioned models, and the AIC and BIC were calculated for each one of the metabolites. The selection of the best fitting model was based on the mean of these criteria across all the compounds analyzed.

3.3 SUNDILE Workflow

SUNDILE can be divided into two steps: i) a *binning* step and ii) an *identity suggestion* step. In brief, the *binning* step detects isotopologues that co-elute at the same retention time, bins them into groups associated with the same putative compound, then screens the compound groups to detect hits that indicate the incorporation of an isotopic label (i.e., a shift in the isotopologue distribution). The second step uses an *in silico* metabolic network reconstruction to suggest the identity of the labeled compounds, then assigns a pathway enrichment score based on the labeling of compounds within specific metabolic pathways.

3.3.1 Binning step

The algorithm starts with a list of preprocessed features (e.g., produced from commonly used software tools such as XCMS), a user-specified mass tolerance, and the supported atoms that were used as stable isotope tracers in the experiment (i.e., hydrogen, carbon, nitrogen, or oxygen). A backtracking algorithm is used to bin each of the features into an isotopologue series by scanning progressively through the preprocessed list. Sorted by increasing *m/z*, one feature at a time is assumed to be an ion of monoisotopic mass (i.e., an M0 isotopologue). The maximum number of heavy isotope substitutions is calculated from the monoisotopic mass based on a combination of different supervised machine learning algorithms trained using a PubChem compound database and classic methods. Decision tree regressors were used for the estimation of hydrogen substitutions, linear regressors were implemented for the estimation of carbon substitutions, a Gaussian process regression model was used for nitrogen substitutions, and a heuristic approach was utilized for oxygen substitutions. The rationale behind the selection of these methods is found in Section 3.4.2. The binning algorithm enumerates the possible heavy isotopologues and their molecular masses, finding all possible combinations of isotopes that could be present. These *m/z*

values are searched within the list of remaining unbinned features and assigned to a temporary compound group. The MID of the compound group, which is a vector of the relative intensities of each of the binned isotopologues, is calculated and corrected for natural isotope abundance using a correction matrix that assumes the distribution calculated for the unlabeled timepoint (t=0) is an accurate representation of the unenriched MID of the putative compound [171] as shown in Figure 3-2.

A)

MID Correction Matrix				
M0(0)	M1(0)	M2(0)	M3(0)	M4(0)
M0(1)	M1(1)	M2(1)	M3(1)	M4(1)
M0(2)	M1(2)	M2(2)	M3(2)	M4(2)
M0(3)	M1(3)	M2(3)	M3(3)	M4(3)

B)

Approximated Correction Matrix				
M0(E)	M1(E)	M2(E)	M3(E)	M4(E)
0	M0(E)	M1(E)	M2(E)	M3(E)
0	0	M0(E)	M1(E)	M2(E)
0	0	0	M0(E)	M1(E)

Figure 3-2. Calculation of the correction matrix based on experimental data from unlabeled samples. *A) Classic MID correction matrix. The elements of the correction matrix are expressed in the format Mi(j) where 'i' represents the number of heavy atoms (including naturally abundant isotopes, i=0 being the monoisotopic mass) and 'j' represents the number of labeled atoms incorporated into the molecule from the tracer. For instance, M1(2) represents the measurement of the M+1 isotopologue of a molecule containing 2 labeled atom replacements. B) Approximated correction matrix computed by shifting the elements of the experimentally determined unlabeled distribution ('E'=experimental). Typically, a classic correction matrix is calculated by filling the first row of the matrix with the simulated MID of an unlabeled molecule, followed by the consecutive filling of subsequent rows with the simulated MIDs of the same molecule replacing an increasing number of unlabeled atoms with labeled ones. The approach applied here approximates the MID of a labeled molecule as a shifted distribution of the experimentally determined values in the binning process.*

Typically, a classic correction matrix is calculated by filling the first row of the matrix with the simulated MID of an unlabeled molecule, followed by the consecutive filling of subsequent rows with the simulated MIDs of the same molecule replacing an increasing number of unlabeled atoms with labeled ones. The proposed approach shown in Fig. 3-2B approximates the MID of a labeled molecule as a shifted distribution using the empirically determined values derived from binning the isotopologues of unlabeled samples. This approximation neglects the so-called 'skew correction factor' [134] because the atoms that contribute to the higher mass isotopologues (M1, M2, etc.) cannot be inferred from the measurements at this stage of the workflow.

For a temporary compound group to be considered a valid, isotopically enriched compound it needs to fulfill the following conditions:

- The relative intensity of the monoisotopic mass (M0) must be greatest in the unlabeled samples; most small molecules of interest in metabolomics will fulfill this criterion, whereas some lipid and protein molecules may not. This condition is evaluated by performing a right-tailed T-test on the mass isotopomers of the different samples and replicates.

- M0 must decrease over time in samples collected following tracer administration. This condition is evaluated by a simple linear fit of the data. If sufficient datapoints ($n \geq 3$) are available to estimate the 95% confidence interval of the fitted slope, a compound group is considered to have a decreasing trend in M0 if the 95% confidence interval of the M0 slope includes only negative values. If insufficient replicates ($n < 3$) are available to perform proper statistical tests, this condition is fulfilled if the fitted slope is negative.

- The maximum mass difference between two consecutive isotopologues must be less than or equal to 3 Da; this condition avoids the incorrect binning of adducts.

- The ratio between the relative intensities of the first isotopologue (M1) and M0 must yield a passing score on a supervised support vector machine (see section 3.4.3) trained with the previously described data; this condition is necessary to avoid false positives in the binning of noisy peaks.

If the group of isotopologues is considered to represent a valid and labeled compound, the corresponding isotopologue features are removed from the list and the algorithm continues to analyze the next possible M0 species. When all features have been scanned, an adduct search is

conducted by comparing the *m/z* difference between each compound group and a list of possible adducts. The list of the possible adducts is chosen according to the acquisition mode, as previously described [172].

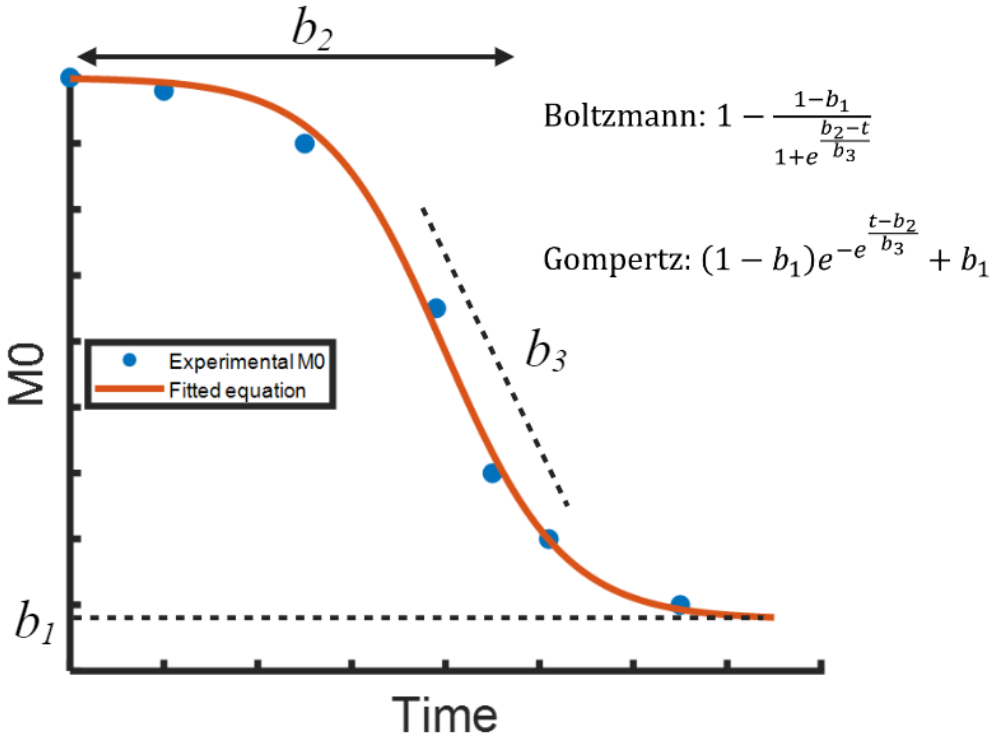


Figure 3-3. Logistic models used to extract labeling information from the empirically corrected values of M0 over time*. In both cases, the models yield information on the maximum enrichment (i.e., minimum M0 value, b_1) that is reached at steady state, the lag period prior to exponential labeling (b_2), and an exponential time constant (b_3) that describes the maximum rate of labeling.*

At this point, the algorithm will have generated a list of valid metabolite peaks (annotated by *m/z* and retention time) that exhibit significant isotope enrichment during the course of the ILE, as well as their corresponding isotopologues. The final step evaluates the dynamics of this enrichment. This is achieved by fitting the time course of the corrected M0 values to one of two possible user-selected logistic models, as depicted in Figure 3-3 (when at least 3 sample time points are available) or to a linear model if only 2 time points are available.

From the logistic models, the plateau value at which M0 stabilizes at t=∞ (i.e., the best-fit value of the b_1 parameter) reflects the maximum enrichment at steady state. Furthermore, the best-fit value of the b_2 parameter is indicative of the time delay preceding the onset of enrichment of a given metabolite. Finally, the best-fit value of the b_3 parameter represents the time constant of labeling during the period when the rate of change is maximal [173], [174]. In contrast, the slope of the linear model is the sole parameter that can be extracted is an indirect measurement of the rate of labeling. More information of these models is presented in section 3.2.6.

3.3.2 Metabolic network reconstruction

The underlying algorithms of the identity suggestion step require the *in silico* reconstruction of an organism-specific metabolic network using an undirected graph approach. This is accomplished by creating a binary $N \times N$ adjacency matrix (N = number of total compounds), where each element of the matrix has a value of 1 if the compounds represented by the corresponding row and column are connected via a metabolic reaction, or zero if they are unconnected. SUNDILE uses the KEGG API [175] as the source of reaction information for constructing an organism-specific adjacency matrix, although other databases could also be used for this purpose.

The algorithm starts by querying KEGG to generate a list of the metabolic pathways for a user-defined organism, and the metabolic reactions that comprise the pathways. One by one, each reaction is searched to find the compounds that participate in the reaction and to update the adjacency matrix for compounds that appear as substrates or products. For example, the reaction A → B would be represented by a value of 1 at positions (P_A, P_B) and (P_B, P_A) of the adjacency matrix, where P_i is the position corresponding to the i^{th} compound in the matrix. A priori, this approach assumes that the inverse reaction is always possible or that the directionality of the reactions is not considered in the creation of the matrix, which differs from directed graph

approaches used in other publications [176], [177]. The rationale for using an undirected graph is that it favors the discovery of new pathways, since the KEGG database often lacks comprehensive descriptions of reaction reversibility. Additionally, considering that organism-specific enzyme information is not employed, we aimed to keep the possibility of investigating new pathways unrestricted by the constraints imposed by the directionality of the reactions.

The adjacency matrix is transformed into a graph with active compounds as nodes and chemical reactions as paths, and the matrix is trimmed to remove rows and columns that have a degree equal to zero (i.e., corresponding to compounds that are unconnected in the network). Additional information of the compounds that remain in the adjacency matrix (such as their molecular mass, formula, and name) is retrieved from KEGG and stored locally as arrays in a .mat file. The local storing of this file avoids unnecessary duplication of this time-consuming step every time the algorithm is run. Finally, the names of the remaining compounds are compared to build another sparse matrix that has non-zero entries if two or more compounds are isomers of each other; this isomer matrix is used to avoid redundancies in the suggestions from the subsequent compound identification step.

3.3.3 Compound identity suggestion

Using the KEGG API [175], the algorithm finds the intersection of (i) the set of metabolic pathways that are connected to the tracer and (ii) the set of pathways associated with the specified organism, yielding a list of pathways that could be isotopically enriched during the ILE. Next, the compounds in these pathways are retrieved and their molecular masses are compared against the *m/z* values of the labeled peaks (i.e., 'hits') found by the binning algorithm; this yields a candidate list of potentially labeled metabolites. The network distances between all the potentially labeled metabolites and the tracer source node is calculated from the graph using the built-in MATLAB

function *"distances"* to compute the shortest path between a node pair (i.e., the path with the fewest reaction steps between the tracer and the measured metabolite). A function that scores the distance of each candidate metabolite to the point of tracer entry within the reconstructed network is used to prioritize compounds that are metabolically 'closer' to the source of labeling and, thus, more likely to have been labeled by the tracer. The score is computed as the value of the normal probability density function with a mean equal to zero and standard deviation customizable by the user (default = 3 reaction steps), evaluated using the shortest distance between each metabolite and the tracer source node. This standard deviation value was chosen based on the empirical data gathered during the early stages of algorithm development. The metabolite measurements in these test datasets were largely concentrated in pathways 'nearby' to the tracer node, so a higher value of the standard deviation may be appropriate for labeling studies of longer duration where the tracer atoms can reach more distal pathways. The algorithm picks the top 3 identity suggestions from KEGG with the highest scores (i.e., shortest network distances from the tracer source) and presents them to the user.

With the list of all possible suggestions, the algorithm estimates an overall activity score for each KEGG pathway as the average b_1 parameter value over all the metabolites within the pathway, divided by the number of metabolites detected in the pathway.

3.3.4 SUNDILE output

SUNDILE outputs a .xls file that contains the following information:

- A table with all the binned compounds and their respective isotopologues, including information on their molecular mass, retention time, number of labeled atoms, and adducts.

- A table with the monoisotopic mass and retention time of the compounds that were found to be isotopically enriched, the highest number of labeled atoms detected for each element traced, the corrected M0 abundance at each time point, and the identity suggestion of each compound.
- The active pathways in the experiment. This file provides the number of metabolite identities that were found in each pathway, the pathway coverage (calculated as the ratio between the number of found metabolites and the total number of metabolites in the pathway), and the enrichment score of the pathway.

3.4 Results

3.4.1 Software validation

SUNDILE's performance was evaluated across multiple performance tiers:

1) Software capabilities and ease-of-use attributes.
2) Feature reduction capacity.
3) Comparison of SUNDILE's binning results to X^{13}CMS.
4) Comparison of SUNDILE's compound suggestions to those identified with MS-DIAL.

3.4.1.1 Capability comparison

A recently published review compared the capabilities of several tools that quantify metabolite enrichment in ILEs [178]. For objectivity, we used the same rubrics to position our software among others in the review (Table 3-1). This comparison can be divided into two groups:

The first group contrasts the technical capabilities of each tool such as the programming language used to implement the tools, the use of statistical analysis to differentiate between labeled and unlabeled peaks, the capability to consider multiple time points and extract information of the labeling kinetics, and the ability to identify metabolites and correct MIDs for natural isotope abundance. None of the evaluated tools offered all the capabilities examined. It might be argued that SUNDILE offers all the capabilities in the list; however, it is limited in the sense that the suggested metabolite identities are based on the measured *m/z* of the base peak and the biological context of the experiment; thus, the identifications could be considered "Level 3" confidence level according to a previously published categorization [37].

The second group contrasts "ease-of-use" characteristics: availability of a GUI, ability to evaluate or display the quality of the peaks, whether the user needs to write code in the language used by the tool, ability to visualize the data in terms of enrichment or MID plots, and the output format. All the capabilities of SUNDILE are offered through a GUI, and no knowledge of MATLAB programming is required. One limitation of SUNDILE is that it does not offer a way to evaluate or visualize the chromatographic peaks, as the data have already been preprocessed and only the peak intensities are extracted. SUNDILE also outputs information that can be used to reconstruct MID or enrichment plots but does not offer that capability per se.

Feature	X[13]CMS	geoRge	HiResTEC	dynaMet	mzMatch-ISO	SUNDILE
Platform	R	R	R	Python	R	MATLAB
Statistical analysis	Y	Y	Y	N	Y	Y
Time course kinetics	N	Y	Y	Y	Y	Y
Metabolite identification	N	Y	N	Y	Y	Y
MID correction	Y	N	N	Y	N	Y
Ease of use						
GUI	N	N	N	Y	N	Y
Peak evaluation	Y	N	Y	Y	Y	N
Reliance on programming knowledge	Y	Y	Y	Y	Y	N
Automated data visualization	N	N	Y	Y	Y	N
Output format	.xslx, pdf	.xlsx	.xslx, pdf	Python table file	.tsv/.pdf	.xlsx

Table 3-1 Comparison between freely available software tools designed to detect isotopic labeling in an untargeted manner. The evaluation criteria and results were obtained from a previously published review [178], with the evaluation for SUNDILE added in the rightmost column. Unlike previous tools, SUNDILE offers all the features listed and offers a GUI that does not require previous knowledge of software platforms and programming languages.

3.4.1.2 Feature reduction capacity

MS datasets from both mouse and soybean labeling experiments were analyzed, quantifying the number of features that were retained at different steps in the workflow. The results of this analysis are presented in Fig. 3-4. The difference between the number of total features detected in each dataset can be attributed to differences in the study designs. In particular, the use of a dual-labeled tracer (i.e., [$^{13}C_5$,$^{15}N_2$]glutamine) in the soybean dataset requires a higher resolution to detect the mass defect between the carbon- and nitrogen-labeled isotopologues. Therefore, a higher number of peaks were detected in the dataset. In addition, the increase in the MS resolution also affected the IPO-optimized parameters that were used for peak extraction in the data-preprocessing step. Finally, the sample preparation steps were optimized to target a different range of metabolites in

each study. It is possible that the sample preparation used to produce the soybean dataset resulted in better metabolite recovery and thus a higher number of compound peaks.

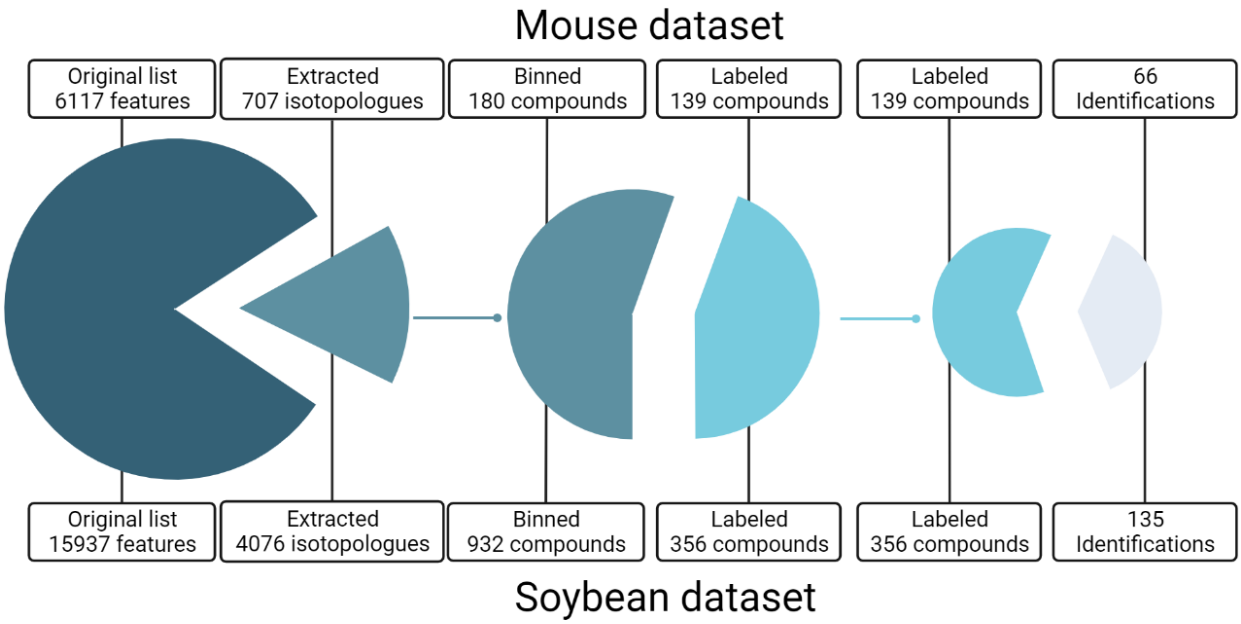


Figure 3-4. SUNDILE feature reduction performance. *The measurements in both analyzed datasets were reduced to less than 10% of the original number of features after the SUNDILE algorithm was used. 'Compounds' are groups of features that are considered to be isotopologues or adducts of the same metabolite.*

In the first part of the algorithm, after grouping the features into valid compounds (regardless of their isotope enrichment) ~80% of the features were discarded leaving around ~180 valid compounds in the mouse dataset and ~900 in the soybean dataset. After discarding those with insignificant labeling, ~140 labeled compounds (~540 features) and ~350 compounds (~1700 features) were kept in the mouse and soybean datasets, respectively. The algorithm suggested 66 compound identities in the mouse dataset and 135 in the soybean dataset. Overall, SUNDILE is capable of discarding ~90% of the features that are not of interest in an untargeted metabolomics experiment involving stable isotope labeling.

3.4.1.3 Comparison of binning results to X^{13}CMS

In the [$^{13}C_3$]propionate-labeled murine dataset, the number of labeled compounds detected by each software tool is presented in Fig. 3-5A. While X^{13}CMS found 112 labeled compounds, SUNDILE was able to find 181. Eighty-three of these compounds were found by both tools. Among the 29 compounds that X^{13}CMS found but SUNDILE excluded, 16 did not pass the compound validation filters and 13 were not detected due to differences in how mass tolerance is calculated by the two programs.

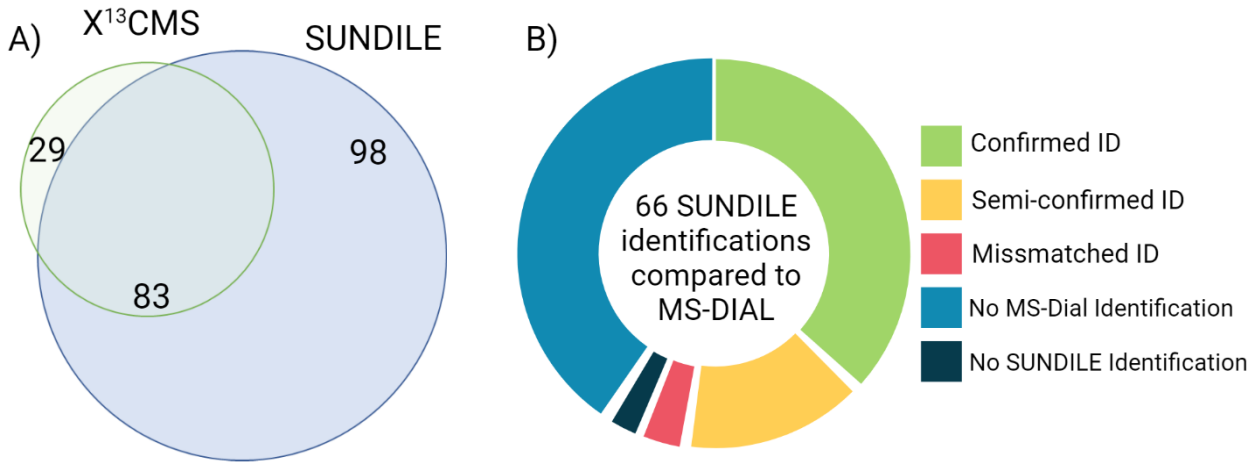


Figure 3-5. Binning and compound identification performance. *A). Comparison of the binning capabilities of SUNDILE versus X^{13}CMS. SUNDILE was able to find significantly more compounds that were not recovered by X^{13}CMS. B). Comparison of the ID suggestion capabilities of SUNDILE versus MS-DIAL. A vast majority of the compounds that were suggested by SUNDILE could be confirmed (Level 1 identification based on MS2 spectral matching) or semi-confirmed (Level 3 identification based on the isotopic pattern).*

The majority of compounds that were excluded by SUNDILE due to the validation filters show statistical differences between the labeled and unlabeled samples, but a physically unrealistic labeling trajectory was observed (i.e., the isotope enrichment decreased over time, rather than the expected increasing trend). Therefore, SUNDILE was able to exclude these false positives by evaluating the direction of labeling over time and removing compounds that do not exhibit the expected trend. On the other hand, the differences in the mass tolerances arise from the fact that X^{13}CMS calculates the parameter as a function of the two masses that are compared and the

isotopic mass difference (Eq. 3-2a), whereas SUNDILE only uses the monoisotopic mass of the measured peak (Eq. 3-2b).

$$\Delta m_b = m_b^*(1 \pm \epsilon) + \frac{m_b * \epsilon}{d\,(1 \pm \epsilon)} \quad \text{(Eq. 3-2a)}$$

$$\Delta m_b = m_b * \epsilon * 1 \times 10^{-6} \quad \text{(Eq. 3-2b)}$$

Where m_b is the *m/z* of the base peak, m_b* is the nearest whole number to m_b, ε is the ppm error, and *d* is the mass difference of the isotopic label.

3.4.1.4 Compound identification

As previously discussed, SUNDILE yielded 66 compound suggestions for the mouse T cell dataset. To confirm their validity, the results were compared against the output of MS-DIAL, a program designed to produce high-confidence "Level 1" identifications based on the acquired MS^2 spectra of the unlabeled samples. If there is not enough information to find a "Level 1" identification, the program will use the measured MS^1 isotopic pattern to suggest a lower confidence "Level 3" identification. The results of this comparison can be divided into five groups (Fig. 3-5B):

1. **Level-1 confirmed identities:** These are metabolites for which the MS-DIAL MS^2 analysis (Level-1 confidence) matched one of the three compound identities suggested by SUNDILE. Approximately 38% (25) of the metabolites found by SUNDILE fell into this category.

2. **Level-3 confirmed identities:** These are metabolites for which the MS-DIAL MS^2 analysis did not yield a match, but a suggestion of their identity was proposed by MS-DIAL

based on their isotopic patterns (Level-3 confidence). Approximately 14% (9) of the metabolites found by SUNDILE were in this group.

3. **Mismatched identities:** This group contains metabolites for which the identity suggestion provided by SUNDILE did not match the MS^2 identity found by MS-DIAL. Only 4.5% (3) of the SUNDILE results were in this group.

4. **No MS-DIAL-identified metabolites:** These are metabolites for which SUNDILE provided an identification, but no identity was found by MS-DIAL. This occurrence can be attributed to two primary factors: i) The MS^2 spectra of the metabolite suggested by SUNDILE was absent in the MS-DIAL library employed for the analysis, or ii) the *m/z* peaks associated with the precursor ions had comparatively lower intensities in relation to other metabolites within the sample, resulting in their exclusion from MS^2 analysis during the data-dependent LC-MS run. This category contains approximately 41% (29) of the metabolites identified by SUNDILE.

5. **No SUNDILE-identified metabolites:** This group refers to metabolites that were flagged by SUNDILE as labeled without presenting an identification, yet were matched to an identity by MS-DIAL. This occurrence is ascribed to compounds that were not matched to KEGG compounds during the analysis due to their absence in the reconstructed network. Such exclusion can result from a lack of association with the tracer or with the organism under investigation. Approximately 3% (2) of the total SUNDILE-proposed compound hits were found within this group.

3.4.2 Use of supervised machine learning algorithms to assess the number of atoms in a compound

During the binning step, it is necessary to estimate the maximum number of atoms that comprise a compound based solely on its measured monoisotopic *m/z* value. This value will determine the extent of isotopologues that are searched in the list of extracted masses. If this value is underestimated, it is possible to miss higher isotopologues that are enriched by the tracer, especially in experiments involving fully labeled tracers. On the other hand, if the number of atoms is overestimated, false positive isotopologues might be assigned to a compound, skewing the calculated MIDs and possibly misassigning a feature that belongs to another compound.

Heuristic formulas have been developed that enable metabolomics data analysis programs to estimate the elemental composition of a detected compound [47], [179], [180]. However, all existing approaches rely on the assumption that the compound is completely resolved from other isobaric compounds, and that its isotopic distribution is accurately measured by means of MS or NMR spectroscopy [181]. One of the most common approaches is to apply the "Seven Golden Rules" for heuristic filtering of molecular formulas obtained via high-resolution MS [182]. This approach starts by estimating the number of atoms belonging to the most common elements found in organic compounds (i.e., hydrogen, carbon, nitrogen, and oxygen) using linear models. Subsequently, these estimations are refined by comparing the measured isotopic distributions obtained by MS or NMR against the isotopic and heteroatom ratios calculated from libraries of known compounds.

However, the full application of these heuristics and subsequent refinements becomes unviable when isotopic information is lacking. For SUNDILE, estimating the number of atoms of each element in the chemical formula serves as a precursor for binning the isotopologues, and hence

precedes the determination of the compound's MID. Therefore, we tested whether accurate chemical formula predictions could be based on the compound's molecular mass alone. As proof of concept, we processed 111,375 compounds from the PubChem metabolite database to predict the maximum number of carbon (C), nitrogen (N), and hydrogen (H) atoms in a compound based on its molecular mass, mirroring the initial stages of the "Seven Golden Rules" approach mentioned earlier. The outcomes are presented in Figure 3-6. The results indicate that the accuracy of the heuristics degrades at higher *m/z* values.

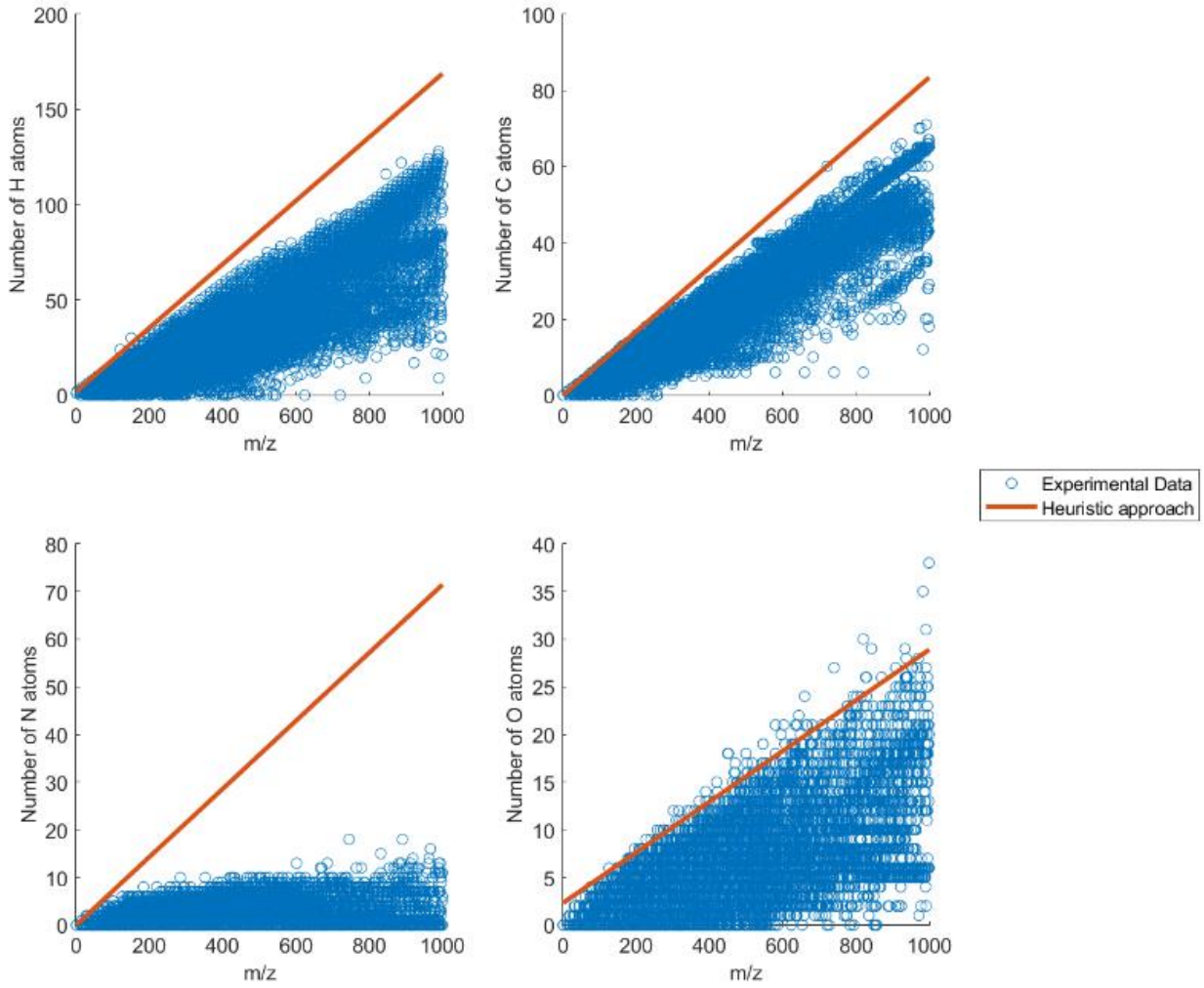


Figure 3-6. Atom number estimation as a function of the m/z value of a compound following commonly used heuristic rules. *These rules fail to estimate the maximum number of hydrogen, carbon, nitrogen, and oxygen atoms at m/z values higher than ~500.*

The heuristics are accurate over low ranges of *m/z* values (~0-200 Da). However, many metabolites have molecular masses outside of this range, which motivates the search for a more accurate approach.

As an alternative to the heuristic "Seven Golden Rules", different supervised machine learning algorithms were trained using the compound data in PubChem. The results of the regressor training for all possible models and datasets are presented in Table 3-2.

	RMSE			
Tested Model	**H**	**C**	**N**	**O**
Linear	3.254	1.6081	1.8656	**2.154**
Interactions Linear	3.254	1.6081	1.8656	**2.154**
Robust Linear	3.2563	1.6433	1.8655	2.1976
Stepwise Linear	3.254	1.6081	1.8656	**2.154**
Fine Tree	4.9674	2.6927	1.8608	2.5507
Medium Tree	9.7251	4.7915	2.0073	3.1455
Coarse Tree	18.872	10.186	2.4031	4.6558
Linear SVM	3.928	1.6846	1.8737	2.205
Quadratic SVM	3.1239	1.6804	1.7643	2.1961
Cubic SVM	3.2012	1.8237	**1.6729**	2.1896
Fine Gaussian SVM	4.2798	2.6974	1.7903	2.5669
Medium Gaussian SVM	3.4426	2.2052	1.7488	2.3162
Coarse Gaussian SVM	3.9333	1.7273	1.8871	2.3775
Boosted Trees	5.008	2.6712	1.8304	2.5913
Bagged Trees	4.1048	2.424	1.8247	2.4398
Squared Exponential GPR	2.7783	1.5767	1.7092	2.2092
Matern 5/2 GPR	**2.7252**	**1.5567**	1.735	2.3
Exponential GPR	3.1665	1.6774	1.7528	2.2208
Rational Quadratic GPR	2.7647	1.5712	1.735	2.2399

Table 3-2. Goodness of fit for the different regressors modeling the maximum number of different atoms of hydrogen, carbon, nitrogen, and oxygen. The root-mean-squared error (RMSE) was used as the estimator parameter to evaluate the models. Hydrogen and carbon are best fit by Gaussian progress regressors with a Matern 5/2 kernel, nitrogen is best fit by a support vector machine with a cubic kernel, and oxygen is best fit with multiple linear models. The lowest RMSE in each column is highlighted in bold.

The atoms of hydrogen, carbon, and nitrogen are best fit with non-linear kernels, whereas oxygen is best fit by a simple linear model without the need of a kernel. The results of the implementation of these regressor models on the test data are shown in Figure 3-7.

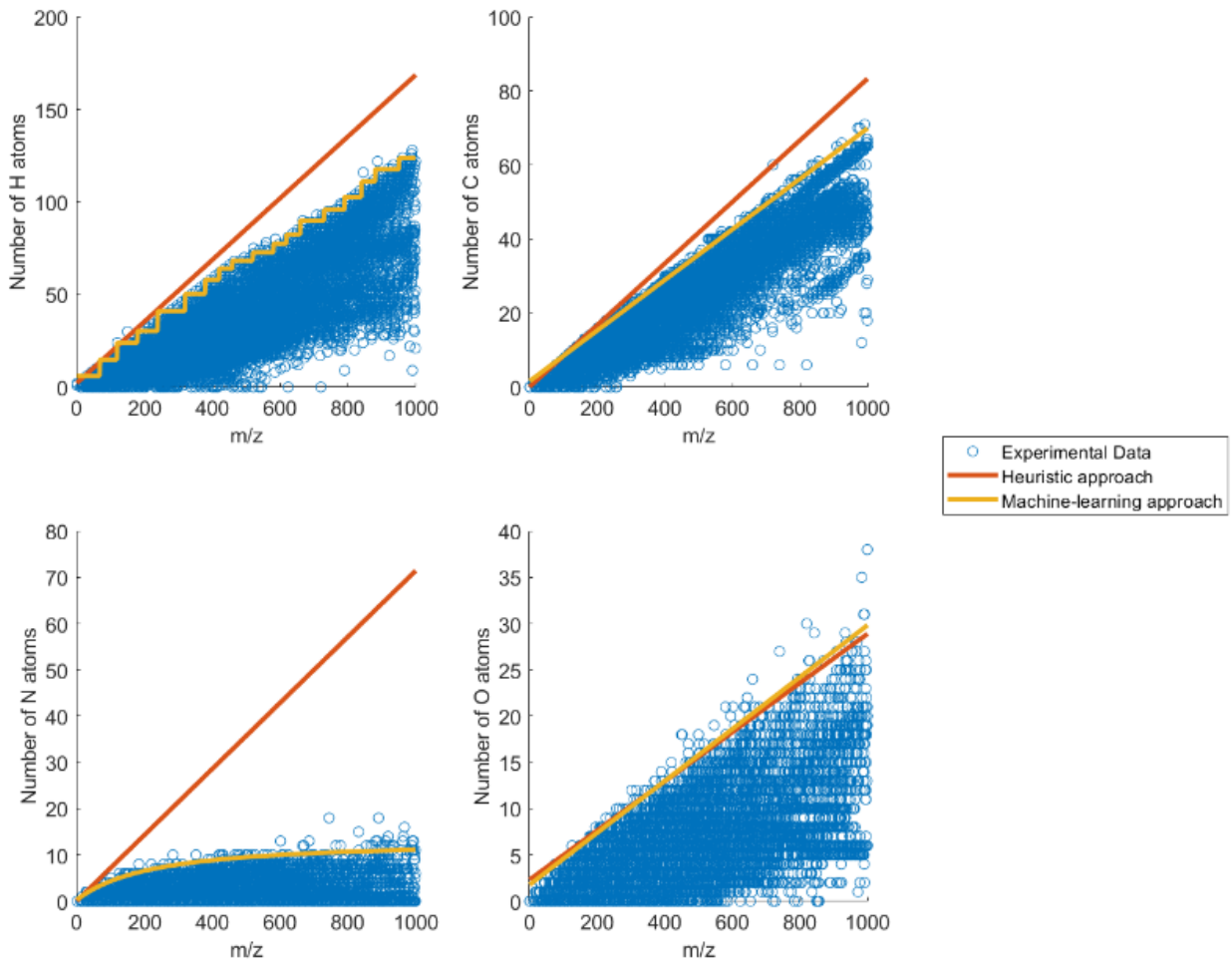


Figure 3-7. Application of the best models described in Table 3-2 to estimate the maximum number of atoms of each element as a function of the compound's molecular mass (m/z)*. The orange line shows estimates provided by the "Seven Golden Rules" heuristic compared to the machine learning estimates shown in yellow.*

Compared to the heuristic approach, the use of supervised machine learning methods outperformed the heuristic "Seven Golden Rules" approach. This difference is more noticeable at higher masses when regressing the maximum number of hydrogen and carbon atoms, given the previously discussed overestimates provided by the heuristic models. In the case of nitrogen atoms, the non-linear kernel provides an improved fit over all values of *m/z*. This is a significant improvement from the previous methods. Finally, in the case of oxygen atoms there is not a substantial difference in the implemented methods given that both approaches rely on simple linear methods. To avoid atom-dependent nuances in the programming of the algorithm that is implemented in SUNDILE,

the maximum oxygen atom linear regressor determined by the machine learning approach was implemented.

3.4.3 Use of supervised machine learning algorithms to validate a compound.

During the binning step (see section 3.3.1), a rule was applied to treat a set of isotopologues as a compound group if the monoisotopic mass (i.e., the M0 isotopologue) had the highest relative intensity in the unlabeled sample. However, the isotopic M1/M0 ratio can also yield information on the number of carbons and can be used to test the validity of a putative compound group. With increasing numbers of atoms in a molecule, the isotopic distributions become skewed such that isotopologues of higher mass exhibit increasing natural abundance. To take full advantage of the information available in the MS data, a supervised machine learning method was trained to recognize real compounds based on their measured values of *m/z* and M1/M0.

Given the nature of the training dataset (i.e., positive data from real compounds) and the lack of options to find negative data reflecting compounds that are "not real", a challenge emerges because most classifiers require both positive and negative data [183], [184]. However, support vector machines can be trained on "one-class" data, meaning that after being trained, they are able to recognize patterns that differ from the data that were used to train them.

A support vector machine model was trained using a normalized *m/z* value equal to the original *m/z* value divided over 1000. The classification loss computed through a 10-fold cross-validation yield was 0.0005, meaning that the model has a misclassification rate lower than 0.1%. The decision boundaries and the support vectors for the trained model are shown in Figure 3-8. The overlap of multiple support vectors creates multiple areas that are assigned scores depending on the density of support vectors. Consequently, an area with a high score represents a hyperplane

that is similar to the trained data. In contrast, an area with a lower score represents a region of parameter space that is different from the trained data. As only positive data were used in the training of the model, higher scores represent data that matches the M1/M0 ratio of real compounds. SUNDILE applies this SVM model to detect and exclude features that were incorrectly grouped as compounds, or that were grouped with peaks belonging to noise and not to real compounds, based on their measured M1/M0 ratios.

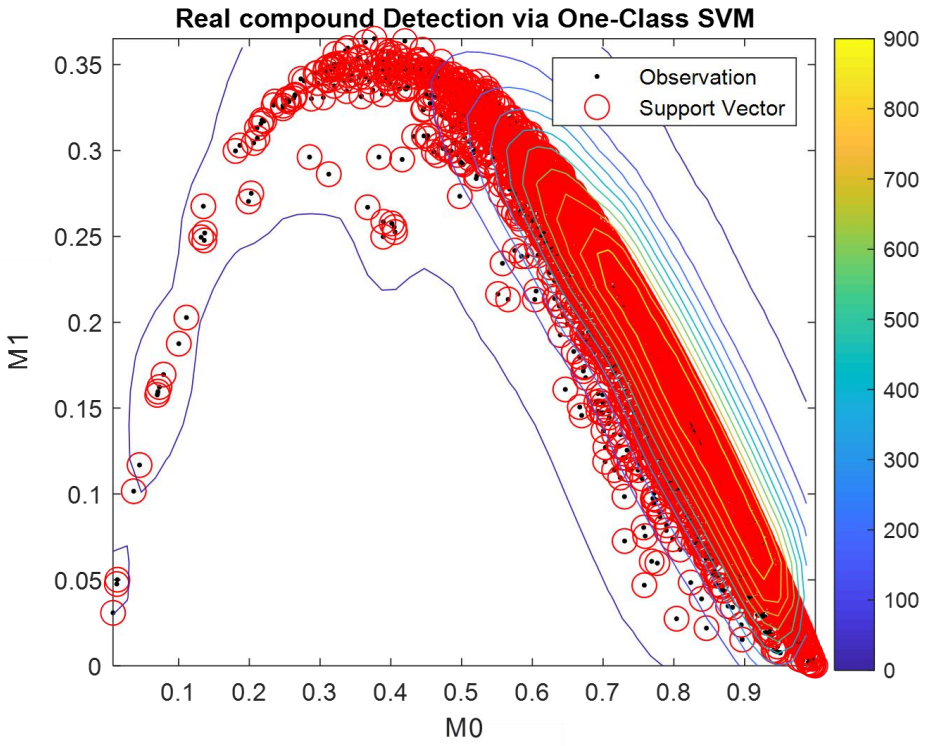


Figure 3-8. Results of the training of the one-class support vector machine to test the validity of compounds based on their M1/M0 ratio and m/z value . *The decision boundaries and their corresponding scores (colored lines), the data used to train the model (black dots), and the support vectors (red circles), are presented. Points that resemble the original data will receive higher scores. A region of feasibility is defined as an area represented with a positive score, and binned features that fall outside this area are considered false metabolites.*

There is a subgroup of support vectors that form low score areas far from the region where most of the data is found, creating hyperplanes that can yield misleading results. To better test the validity of the model, a minimum score is set as a constraint to determine if a group of

isotopologues can be considered a real compound or just the product of noise. Figure 3-9 shows a 2D plot where the scores assigned to pairs of *m/z* and M1/M0 can be found.

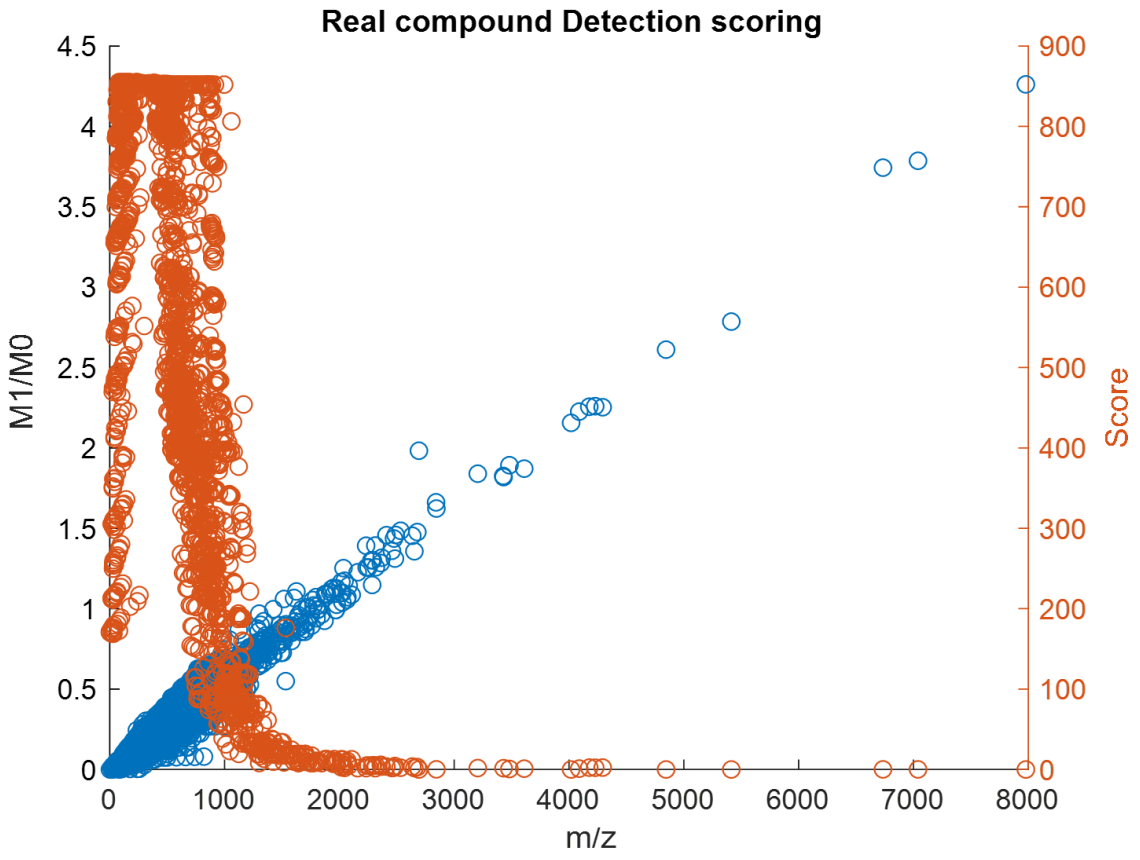


Figure 3-9. Scores (orange points) assigned to pairs of M1/M0 and m/z values (blue points) for each compound in the training dataset*. The SVM is not capable of assigning high scores to real compounds in the higher mass range, which might affect the algorithm if masses with values higher than 1000 Da are found. As most metabolites and lipids of interest are found below this threshold, the model is accepted.*

The interpretation of these results is that the SVM cannot assign high scores in a range of *m/z* values above 1000 Da, which is a defect of the model itself. The limitation that real compounds with masses higher than 1000 Da could be assigned low scores and discarded by the algorithm needs to be considered. This is acceptable for typical metabolomics and certain lipidomics studies, since most metabolites of interest have *m/z* values below that threshold. The data points in the training set with *m/z* values above 1000 Da were principally lipids and proteins that exceed the intended scope of the tool. Below this *m/z* threshold, 95% of the compounds were assigned a score

with a value at least higher than 100, and this value is used as the minimum value for confirmation purposes. In the main SUNDILE algorithm, after a temporary compound group has been binned and the corresponding MID has been calculated, the values of *m/z* and M1/M0 are input into the trained SVM model. If the predicted score is lower than 100, the temporary compound is discarded.

3.4.4 Use of logistic models to determine the labeling parameters of the compounds.

Historically, the turnover of metabolites and proteins has been quantified by fitting changes in isotope enrichment to an exponential decay model [144], [185], [186]. This approach is valid as long as the steepest decay occurs at t=0 and there is no lag phase in the labeling trajectory. In the specific case of untargeted metabolomics, some metabolites often exhibit a delayed response in the incorporation of the labeled tracer. This time delay can occur when there is spatial compartmentalization of certain pathways or when the measured metabolites are several steps downstream of the point where the tracer enters the metabolic network. Metabolites that are directly downstream or closely connected to the tracer tend to incorporate labeled atoms faster since their precursor substrates become labeled more quickly. Conversely, metabolites that are farther downstream or situated in parallel pathways might experience a delay in labeling (Fig. 3-10).

Understanding these dynamics is crucial for the accurate interpretation of ILEs. Therefore, we tested the use of several sigmoidal models that allow for a time delay in the labeling dynamics as alternatives to the simple exponential decay models that have been widely used in the past.

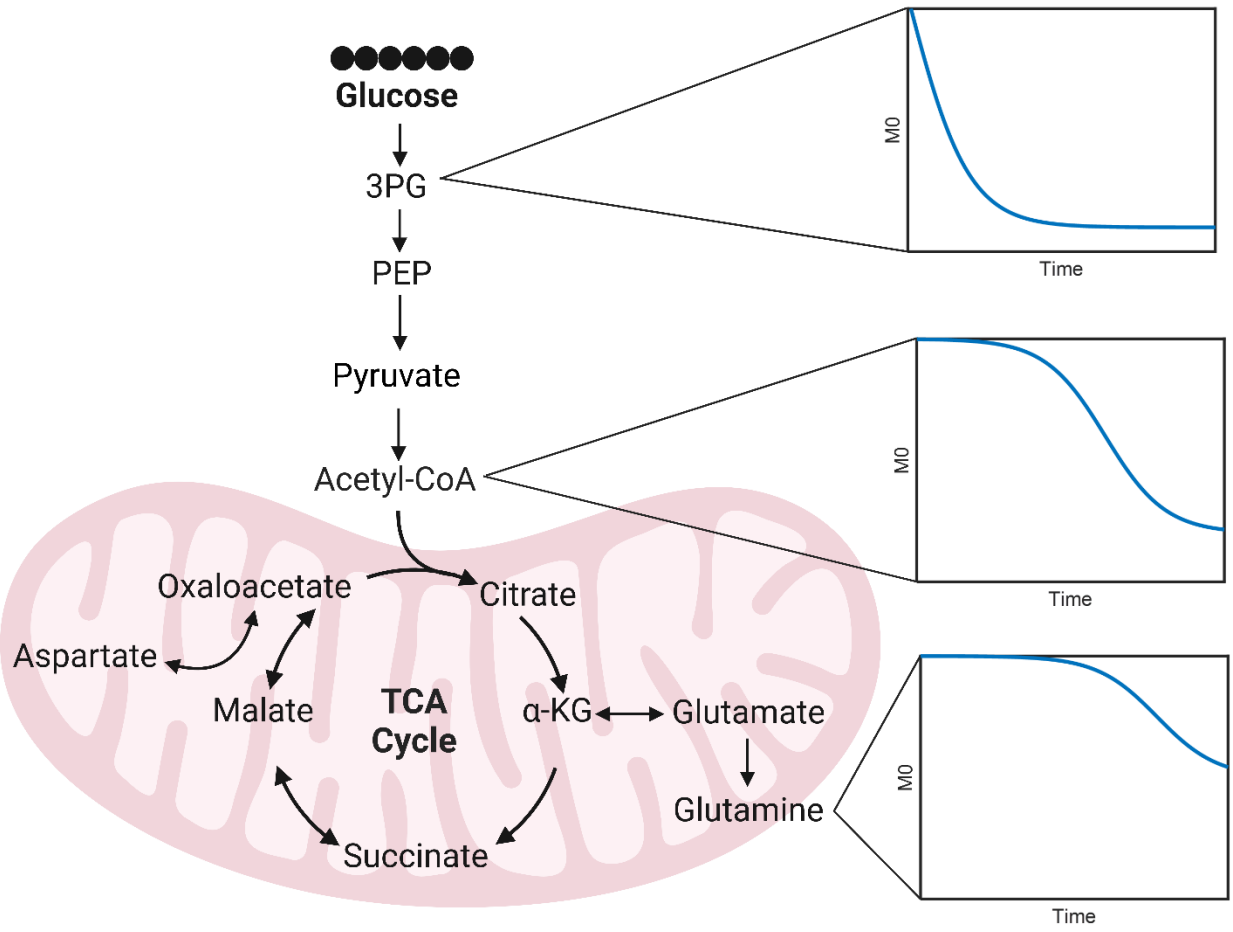


Figure 3-10. Hypothetical labeling dynamics of the glycolytic and TCA cycle pathways resulting from the introduction of a glucose tracer. *Metabolites immediately downstream of the tracer are rapidly labeled while metabolites that lie at a farther distance from the tracer take longer to exhibit detectable isotope enrichments.*

The models tested are described in section 3.2.6, and the results of the calculation of the AIC and BIC for the soybean and murine T cell datasets can be found in Figure 3-11.

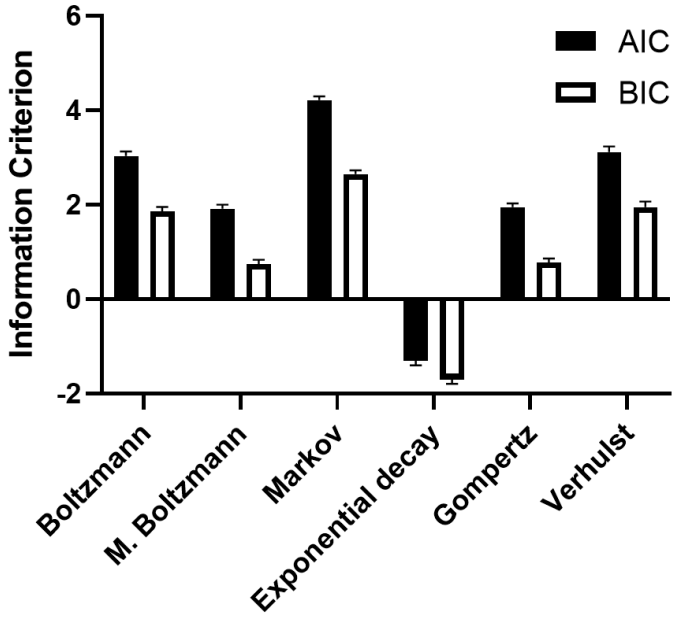


Figure 3-11. Comparison of the Akaike information criterion (AIC) and Bayesian information criterion (BIC) across multiple evaluated models. *A simple exponential decay model provides the lowest AIC and BIC, but the lower values are explained by the fact that this model only uses a single fitted parameter. The modified Boltzmann and Gompertz sigmoidal models show the best performance and were selected as the models implemented in the SUNDILE program.*

Normally, models with low AIC and BIC are preferred because they provide the best trade-off between goodness of fit and simplicity of the model. Among the evaluated models, the exponential decay exhibited notably low values for both the AIC and BIC. However, it is important to note that the performance of this model was influenced by the fact that it relied on only one adjustible parameter, which is rewarded by the criteria. Consequently, the selection of this model would not provide information regarding time delays in the labeling dynamics. In contrast, the modified Boltzmann and the Gompertz models emerged as the most promising alternatives that allow for time delays in the labeling dynamics. For this reason, they were selected as the models to be implemented in SUNDILE, leaving the user the final decision of which one to implement for a particular analysis.

An interesting caveat regarding the selected models is that they do not enforce a specific limit on the initial value of M0 at t=0. As discussed earlier, the samples collected at t=0 are assumed to be

unlabeled and corrected to remove isotopic background such that the M0 fractional abundance will be 1 when t=0. Consequently, the model parameters will be adjusted during the fitting process to give an M0 value that approaches 1 at this timepoint, although this is not enforced as a strict constraint, thus yielding some variability in the possible range of predicted values when t=0.

3.5 Discussion

3.5.1 Compound identification is facilitated using machine learning methods

The utilization of machine learning-based methods offers a significant enhancement in the accuracy and efficiency of estimating the elemental composition and the maximum number of atoms in a putative compound. The use of additional information such as the isotopic distribution to predict the molecular formula of a detected compound has been explored previously with heuristic approaches [182] and machine learning approaches [187]. However, no method by itself has been successful in identifying compounds based solely on raw spectral data from MS. The use of all available information is crucial in the compound identification process, and therefore the use of machine learning and deep learning methods can improve the identification process significantly.

Recent methods have added a retention time dimension into the analysis [188], [189], [190], [191], thereby reaching accuracies up to 70%. Other methods have used the fragmentation patterns of MS^2 data [192], [193], [194], but they are limited to a specific group of compounds (e.g., small molecules, peptides, and compounds with specific functional groups). While our implementation of machine learning models enhanced the estimation of the number of atoms in detected molecules, especially in the case of hydrogen and nitrogen, the models may not offer a substantial advantage in the metabolite identification process by themselves. However, our proposed approach integrates

multiple elements, including atom number prediction, isotopic patterns, and metabolic network reconstruction, in conjunction with analysis of stable isotope labeling dynamics. This combined methodology produces remarkable results when validated against the MS-DIAL program that provides Level-1 metabolite identifications based on MS^2 data, especially considering that exclusively MS^1 data is used in the SUNDILE workflow.

It is important to note that SUNDILE does not support the partial implementation of algorithms to merely suggest a molecular formula. Instead, it heavily depends on KEGG-based metabolic network reconstruction, and the identification of compounds beyond those contained in the database is currently unsupported. Compounds that are not present in the metabolic network reconstruction (e.g., pathways unrelated to the organism under study or disconnected from the tracer) may not appear as possible suggestions, potentially influencing the compound identification process. Even if an identity is not proposed, the user will still be given a list of labeled compounds, denoted by their monoisotopic *m/z* and retention time values. It is strongly recommended to use SUNDILE in tandem with MS^2-based identification tools such as MS-DIAL. This combined approach not only helps identify compounds that SUNDILE may not recognize but also serves to confirm the identities proposed by SUNDILE, thereby improving the overall level of confidence in the identification process.

3.5.2 Validation of compounds via machine learning-based methods

The problem of identifying groups of isotopologues in experiments involving stable isotopes has been addressed previously using different approaches [121], [158], [195], but to our knowledge, SUNDILE is the first software tool to include filters to reduce the rate of false positives in the binning process. In this approach, a crucial component is a supervised machine learning model that assesses the binned set of isotopologues based on their M1/M0 ratio. This method is highly

effective at minimizing false positives and ensuring the overall quality of the results. However, it's essential to acknowledge that this method's accuracy is primarily optimized for metabolites with *m/z* values below 1000 Da. This limitation has been previously recognized and attributed to the rapid increase of isomers with increasing *m/z* and the direct relationship between the mass resolution and *m/z* [196].

While it is true that most metabolites commonly analyzed in metabolomics studies fall below this 1000-Da threshold, SUNDILE users should be cautious when interpreting results for labeled compounds with higher masses. Furthermore, it's worth noting that the model's training data excluded instances where the compounds had a charge different than ±1. Again, it is important to emphasize that the vast majority of ionization methods typically used in metabolomics yield ions that fulfill this condition. Nevertheless, the algorithm was not validated against datasets containing multiply charged ions, which may introduce a potential source of error in such cases.

3.5.3 Use of sigmoid functions to estimate the rate of metabolite labeling

Another unique feature of our tool is that it uses sigmoid functions to fit the labeling dynamics of the detected metabolites. Sigmoidal models have been historically used to describe dose-response behaviors [197], growth of microorganisms [198], and the progressive labeling of proteins in proteomics studies [199], [200], but to our knowledge, these models have not been applied in the context of stable isotope-based metabolomics.

As discussed previously, the main advantage is the possibility to extract additional information from the fitted parameters such as time delay and the maximum degree of enrichment. The use of these parameters could enhance the analysis of untargeted stable-isotope-based metabolomics in

two ways: (i) by accounting for the distance between the tracer and tracee within the network and (ii) by enabling a more rigorous analysis of the labeling kinetics.

First, varying levels of enrichment are observed at different locations within metabolic networks. Initial data from our research group indicated that, in a small-scale network, the degree of enrichment displayed an inverse relationship with the distance between measured metabolites and the tracer source node [156]. We attempted to extrapolate these findings to predict the relative position of a compound in the metabolic network, but all efforts were unsuccessful due to the emergence of metabolic cold pools and other complexities that confound the analysis. Cold pools occur when the same metabolite exists in multiple compartmentalized or spatially segregated locations, and at least one of these pools is metabolically inactive or slow to label. Because untargeted metabolomics only measures the isotope enrichment of the total metabolite pool after it has been extracted from cells/tissues, there is no way to infer the impacts of pathway compartmentalization on the measurements.

Second, the analysis of labeling kinetics, achieved by fitting labeling trajectories to simple kinetic models, has enabled the approximation of metabolic fluxes with certain limitations through a technique known as kinetic flux profiling [144], [201]. This is a similar approach to the stable isotope-assisted protein turnover estimation methods that have been historically described [202], [203]. In essence, these methods rely on fitting labeling dynamics to simple exponential decay models, allowing for the extraction of the analyte turnover represented by the time constant in the model. Addressing the time delay before the labeling process commences has been approached through pulse-chase experiments that provide an isotopic tracer until a specific enrichment level is reached and then measure the rate of disappearance of metabolite labeling after the tracer is removed, often referred to as the "chase" period of the experiment [202]. This process becomes

tedious as the times required to reach the desired enrichment differ for each compound. The use of sigmoidal models allows the extraction of labeling time constants and time delay parameters directly from the "pulse" phase of the experiment, which describe the overall labeling dynamics of consecutive metabolite pools in series (with possible dilution from unlabeled input streams).

3.5.4 Use of SUNDILE in the context of stable isotope-based metabolomics

In its current form, SUNDILE facilitates the identification of labeled compounds and situates them within the framework of metabolic networks. These outcomes serve to assess the extent to which various metabolites and pathways are impacted by the introduction of a labeled tracer into a specific organism, as illustrated by the examples in Chapters 5 and 6. However, there remains a wealth of unexplored information within the outputs generated by SUNDILE. As an example, the labeling dynamics could be used in conjunction with MFA tools such as INCA [16] to perform detailed modeling of isotopologue patterns for precise metabolic flux determination. Furthermore, the extracted time delay and time constant parameters obtained from the sigmoidal models have potential to be used for *de-novo* reconstruction or refining of already existing metabolic networks.

In the context of stable isotope-based untargeted metabolomics, SUNDILE serves as a pivotal resource in the initial phases of hypothesis generation, particularly through the exploration of potentially labeled compounds and pathways. However, for a comprehensive and robust analysis, it is essential to incorporate SUNDILE alongside complementary compound identification software tools. These tools play a crucial role in confirming the identities suggested by SUNDILE and in identifying unknown compounds discovered by it.

To derive meaningful insights, the detected metabolites need to be further analyzed and filtered according to the specific objectives of the researchers. Subsequently, a targeted approach can be

employed with tools such as PIRAMID (Chapter 4) designed to quantify the degree of labeling in specific compounds. The quantification results of the selected metabolites can then be fed into software tools specializing in the reconstruction of metabolic networks or other advanced analyses like MFA. This integrated approach ensures a comprehensive exploration and analysis of the metabolomics data, ultimately leading to a deeper understanding of complex metabolic systems.

3.5.5 Future work and areas for improvement

The capabilities of SUNDILE have been extensively explored in the preceding sections, revealing its potential as a robust tool for metabolomics research. However, it's important to address the challenges and consider potential enhancements for future iterations of the software. One notable aspect is the vast amount of data presented to the user (i.e., the list of isotopologues, the summary of labeled compounds and their identities, the scoring of the compounds and pathways), which can be overwhelming. In Chapters 5 and 6, approaches to visualize and interpret the data (e.g., dendrograms and heatmaps) are used, but their implementation was performed manually. Incorporating these tasks into the core codebase of SUNDILE could streamline the data analysis process, but this would require the development of user-friendly graphical interfaces—a feature for future development.

SUNDILE's limitations were also highlighted. The performance of the SVM in detecting labeled compounds proved to be accurate for *m/z* values below 1000 Da, but further enhancements are needed to expand its applicability to higher mass ranges. While we presented reasons that this is not expected to be a limiting factor in metabolomics analyses, the exploration of additional dimensions of the data and the training of more robust machine learning models are proposed as future approaches to address this issue. This, in turn, may open doors for applying SUNDILE in other omics fields, such as proteomics and lipidomics.

Furthermore, the metabolic network reconstruction that is performed by SUNDILE is specific to a single organism. While this approach presents advantages to avoid misleading pathways, it also biases the results in specific cases where co-metabolism between multiple organisms is prevalent (Chapter 6). Changes in the algorithm will be needed to accommodate these types of cases. These changes can have the form of allowing the user to input families of organisms, or to include additional pathways that are known to be present for specific organisms. For instance, mammals do not have the enzymes required to synthesize essential amino acids, but labeled products or intermediates of these pathways may appear in circulation due to contributions from the endogenous metabolism of gut microbes.

In addition to expanding its scope to simultaneously include additional organisms, an important advancement for SUNDILE would be the incorporation of enzymatic information into its algorithms. Currently, certain pathways are extracted from KEGG for a specific organism, automatically encompassing all the metabolites associated with it. However, many organisms possess a limited set of enzymes that only catalyze a subset of reactions in the pathway. By integrating enzymatic information from the organism under analysis, SUNDILE could effectively filter out metabolites that cannot be synthesized or do not partake in the endogenous reactions of interest. This would significantly refine and narrow down the list of feasible metabolites in the analysis, but might also exclude valid metabolites due to missing or incorrect enzyme annotations. Notably, KEGG possesses this enzymatic data, making it relatively accessible for integration. However, it's important to acknowledge that the implementation of this information into the network reconstruction algorithm would be a time-intensive endeavor. Therefore, this feature is assigned for inclusion in future releases of the software, reflecting a commitment to continued improvement and refinement of SUNDILE.

Lastly, as discussed in section 3.5.4, there is untapped potential in using the fitted parameters derived from sigmoidal models of the labeling dynamics. Leveraging this biological information to perform complex tasks, such as integrating with software for *de novo* metabolic network reconstruction or MFA is a promising avenue for future research. Employing unsupervised machine learning methods, like clustering, can facilitate the discovery of connections between the position of a metabolite within the metabolic network and the time delay before it exhibits a detectable level of enrichment from the tracer. Correlation analyses between the extracted time constants (as a proxy for metabolite turnover rate) and metabolic fluxes are recommended as a starting point for further exploration.

3.6 Conclusions

From the extracted list of MS peaks in an isotope-assisted metabolomics study, SUNDILE enables the user to (i) bin isotopologue peaks into compound groups, (ii) search the compound groups for metabolites that are labeled over time, and (iii) estimate the metabolite MIDs and labeling rate parameters from the raw MS data. It offers a level-3 identification of the 'hits' and frames them in the context of metabolic pathways, scoring the enrichment of the pathways based on the measured metabolites within them. The detection of metabolites is limited to compounds with *m/z* values below 1000 Da, and the identification of metabolites is limited to compounds for which there is enough information available in the KEGG database to obtain a compound match (i.e., molecular mass, formula, accurate metabolic pathway assignment, etc.).

The assessment of SUNDILE's performance included a comparison with other software tools designed for similar purposes, highlighting its unique capabilities in the detection and identification of labeled compounds. Remarkably, this performance was achieved using only MS^1 data, underscoring the tool's efficiency and accuracy. The utilization of machine learning methods

within SUNDILE’s workflow proved to be a significant improvement over previously described algorithms, demonstrating the power of advanced computational techniques in the field of metabolomics. SUNDILE's ability to outperform existing algorithms showcases its potential to greatly enhance the accuracy and efficiency of metabolite identification, marking a significant advancement in the domain of stable isotope-based untargeted metabolomics.

3.7 Acknowledgements

This work was funded by the National Institutes of Health Grant No. U01 CA235508.

CHAPTER 4

Portions of this chapter are adapted from Program for Integration and Rapid Analysis of Mass Isotopomer Distributions (PIRAMID) *and has been reproduced with the permission of the publisher and my co-authors.* [204]

PROGRAM FOR INTEGRATION AND RAPID ANALYSIS OF MASS ISOTOPOMER DISTRIBUTIONS (PIRAMID)

Abstract

Analysis of stable isotope labeling experiments requires accurate, efficient, and reproducible quantification of mass isotopomer distributions (MIDs), which is not a core feature of general-purpose metabolomics software tools that are normally optimized to quantify metabolite abundance. Here we present PIRAMID, a MATLAB-based tool that addresses this need by offering a user-friendly, graphical user interface (GUI)-driven program to automate the extraction of isotopic information from mass spectrometry (MS) data sets. This tool can extract the ion chromatograms from data files in the most common vendor-agnostic file formats, locate chromatographic peaks based on a targeted list of characteristic ions and retention times, and quantify MIDs for each target ion. Multiple metabolites can be analyzed in a consistent manner across a batch of data files (correcting for deviations in their retention times if needed), and the MIDs and integrated ion counts can be output in comma-separated value (CSV) format for easy import into spreadsheet programs. The MIDs can be optionally corrected for natural isotopic background based on the user-defined molecular formula of each target ion. PIRAMID includes several features that set it apart from other freely available tools for stable isotope-assisted metabolomics. In particular, it i) performs data extraction, peak processing, and

analysis/visualization steps within a single program; ii) supports a range of common file formats and both gas chromatography (GC) and liquid chromatography (LC) platforms; iii) does not place restrictions on the number/type of isotopes present (e.g., ^{13}C, ^{15}N, etc.); iv) does not require familiarity with scripting languages or command-line interfaces (CLIs); and v) provides support for tandem mass spectrometry (MS^2) and high-resolution mass spectrometry (HRMS) instruments.

4.1 Introduction

Recent advances in MS technologies have propelled the usage of stable isotopes in biological studies across multiple "omics" fields [205], [206]. However, the lack of efficient and reproducible data analysis workflows remains a major bottleneck for omics experiments [207]. Various software tools have been developed to automate the processing of MS data sets, but a minority of them are optimized to support high-throughput analysis of stable isotope labeling experiments. More specifically, isotope enrichments computed from raw MS data are sensitive to parameter values and require robust algorithms to minimize bias in the results. The use of consistent integration bounds and accurate baseline correction across all isotopologues is critical for accurate MID determination [128]. In cases of low-intensity peaks, it is imperative to recognize that inherent noise, stemming from experimental factors such as the MS detection of background signals, the presence of low-concentration peaks, fluctuations in the electronic components, thermal effects, or electromechanical variations, can introduce substantial measurement errors [129].

To our knowledge, there is no publicly available software package that has been optimized to (i) accurately extract the intensity of the isotopologues of multiple target metabolites from a batch of metabolomics data files, (ii) automate the entire analysis workflow from start to finish, and (iii) support a broad range of MS instruments and metabolomics experimental designs. One of the most common open-source programs used to process MS-based metabolomics data, Maven [130], has been used as a standalone tool and as a base to develop more extensive platforms [208] due to its broad capabilities to quantify MS data sets and to automatically detect and process the peak intensities of isotope-labeled metabolites. Nonetheless, it is not optimized for efficient processing of isotopologue data from large numbers of samples, and its file format support is limited. iMS2Flux [209] provides a standardized platform to automate the processing of isotope

enrichment data from experiments involving ^{13}C-labeled tracers. Despite its strengths, the tool is not able to process raw MS data files and therefore adds a file conversion and pre-processing step to the already time-consuming pipeline. TarMet [210] is a relatively recent GUI-based tool designed to quantify targeted MS peaks from a wide variety of vendor-agnostic data files. However, experiments using multiple isotopic labels are not supported, and the integration of multiple target peaks and/or data files must be done manually. DExSI [211] offers a GUI-based tool capable of annotating metabolites, determining mass and positional isotopomer abundance from multiple fragment ions, and correcting for their corresponding natural isotope abundance. However, it is limited to ^{13}C and ^{15}N labeling experiments and only offers support for GC-MS data. Some examples of command-line based tools are: AssayR [212], an R package designed to integrate the peak areas for all the isotopologues of a ^{13}C or ^{15}N labeled metabolite and export the data using bar plots, and mzMatch-ISO [159], another R package designed to annotate and quantitate isotope-labelled MS data, which serves as an extension to the package mzMatch.R [213]. A common issue with these R packages is that they only accept high-resolution LC-MS data and do not offer a graphical interface, which poses an additional knowledge barrier to metabolomics researchers and makes the data analysis pipeline more complex.

In this chapter, we introduce PIRAMID: a vendor-agnostic tool optimized to analyze data from targeted metabolomics experiments involving stable isotopes. In addition to automating the repetitive steps of peak selection and integration to determine MIDs of multiple target metabolites, PIRAMID can batch process a stack of data files and perform pairwise and time-series comparisons of MIDs all within a single GUI-driven workflow. The tool is designed to accept data from GC- or LC-MS instruments in the most common non-proprietary file formats (.cdf, mzml, and .mzxml). In addition, it can process data from either single MS or tandem MS^2 experiments,

acquired using either low- or high-resolution instruments. By automating the process of extracting, integrating, and analyzing high-throughput MS data in a user-friendly environment, we expect that PIRAMID will enable a growing number of researchers to incorporate stable isotopes into their metabolomics studies.

4.2 Methods

4.2.1 Low resolution sample preparation and analysis

An in-house mixture of 200 mM standards (Thermo Scientific™) containing polar metabolites was converted into methoxime-*tert*-butylsilyl derivatives (Table 4-1) by adding 50 µL of methoxyamine reagent and 70 µL of MTBSTFA + 1% TBDMCS (Thermo Fisher Scientific) and incubating at 70°C for 30 mins. The measurement of these metabolites was performed using an Agilent 7890A gas chromatograph equipped with two 15-meter DB-5ms capillary columns connected to an Agilent 5977B mass spectrometer operating under ionization by electron impact at 70 eV. An injection volume of 1 µl was introduced in split mode at an injection temperature of 270°C. The temperature of the MS source and quadrupole were held at 230 and 150°C, respectively. The detector was set to scan over the mass range 160–595 *m/z*.

Metabolite	Abbreviation	Retention Time	Formula
Pyruvate	Pyr	5.5	$C_6H_{12}O_3NSi$
Lactate	Lac	8.69	$C_{11}H_{25}O_3Si_2$
Alanine	Ala	9.5	$C_{11}H_{26}O_2NSi_2$
Glycine	Gly	9.9	$C_{10}H_{24}O_2NSi_2$
Norvaline	Nor	12.05	$C_{13}H_{30}NO_2Si_2$
Succinate	Suc	13.95	$C_{12}H_{25}O_4Si_2$
Serine	Ser	19.08	$C_{17}H_{40}NO_3Si_3$
Malate	Mal	21.78	$C_{17}H_{39}O_4Si_3$
Aspartate	Asp	22.68	$C_{18}H_{40}O_4NSi_3$
Glutamate	Glu	25.17	$C_{19}H_{42}O_4NSi_3$
Glutamine	Gln	33.3	$C_{19}H_{43}O_3N_2Si_3$
Citrate	Cit	33.34	$C_{20}H_{39}O_6Si_3$

Table 4-1. Metabolite standards in the low resolution dataset used for software validation. The abbreviations, retention times and molecular formulas are presented for each monitored ion.

4.2.2 High resolution sample preparation and analysis

An in-house equimolar mixture of standards composed of the Thermo Scientific™ Pierce™ Amino Acid standard H (Fisher Scientific, Fair Lawn, New Jersey, USA) with the addition of fumarate, succinate, malate, citrate, glucose, fructose, glucose-1-phosphate, and glucose-6-phosphate (HPLC grade, Sigma-Aldrich, St. Louis, Missouri, USA) (Table 4-2) at different dilutions (62.5, 125, and 250 µM) was analyzed by injecting 5 µL of sample to a Shimadzu HPLC linked to an AB Sciex QTRAP 6500 mass spectrometer operated with positive and negative full scan mode monitoring an *m/z* range between 70 and 800 Da at a resolution of 70,000 FWHM. The HPLC system was equipped with an Infinity Lab Poroshell 120 Z-HILIC column (2.7 µm, 100 × 2.1 mm; Agilent Technologies, Santa Clara, CA, USA). The metabolites were eluted with an increasing gradient of acetonitrile: 10 mM ammonium acetate (9:1 v/v) and 5 µM medronic acid, pH 9.0 (mobile phase A) and 10 mM ammonium acetate in water, pH 9.0 (mobile phase B) with a flow rate of 0.25 mL/min. The eluent was ionized under a 4.5 kV (ESI+) ion spray voltage; ion source temperature, 550°C; source gas 1, 45 psi; source gas 2, 40 psi; curtain gas, 35 psi; and entrance potential, 10.

Analogously, a mixture of glutamine standards with varying enrichment patterns (unlabeled, 1-^{13}C, 1,2-$^{13}C_2$, $^{15}N_2$, U-$^{13}C_5$) (Cambridge Isotope Laboratories, Tewksbury, MA, USA) was analyzed using a Thermo Scientific Q-Exactive mass spectrometer by injecting 10 µL of sample to a SeQuant ZIC HILIC column (2.1 × 100 mm, 3.5 µm) with mobile phase A containing 9:1 water:acetonitrile + 5mM ammonium formate and mobile phase B containing 9:1 acetonitrile:water with 5mM ammonium formate at a flowrate of 200 µL/min with the following gradient: 95% B for 2 min, 95 to 40% B over 16 min, 40% B held for 2 min, 40 to 95% B over 15 min, and 95% B held for 10 min (gradient length, 45 min). Full scan MS^1 data was acquired in negative mode between 80 and 500 Da at a resolution of 60,000. Source ionization parameters were as follows: spray voltage, 3.0 kV; transfer temperature, 280°C; S-lens level, 40; heater temperature, 325°C; sheath gas, 40; aux gas, 10; and sweep gas flow, 1.

The analyzed compounds and their respective formulae are presented in Table 4-1.

Metabolite	Abbreviation	Retention Time	Formula
Alanine	Ala	7.75	$C_3H_7NO_2$
Aspartate	Asp	7.75	$C_4H_7NO_4$
Glutamate	Glu	7.73	$C_5H_9NO_4$
Isoleucine	Iso	6.10	$C_6H_{13}NO_2$
Leucine	Leu	6.37	$C_6H_{13}NO_2$
Methionine	Met	6.14	$C_5H_{11}NO_2S$
Phenylalanine	Phe	5.25	$C_9H_{11}NO_2$
Proline	Pro	7.50	$C_5H_9NO_2$
Serine	Ser	8.05	$C_3H_7NO_3$
Threonine	Thr	7.45	$C_4H_9NO_3$
Tyrosine	Tyr	6.95	$C_9H_{11}NO_3$
Valine	Val	6.90	$C_5H_{11}NO_2$
Fumarate	Fum	7.35	$C_4H_4O_4$
Succinate	Suc	7.25	$C_4H_6O_4$
Malate	Mal	7.37	$C_4H_6O_5$
Citrate	Cit	6.20	$C_6H_8O_7$
Glucose	Glc	9.10	$C_6H_{12}O_6$
Fructose	Fru	7.77	$C_6H_{12}O_6$
Glucose-1-Phosphate	G1P	8.25	$C_6H_{13}O_9P$
Glucose-6-Phosphate	G6P	8.50	$C_6H_{13}O_9P$
Glutamine	Gln	4.90	$C_5H_{10}N_2O_3$

Table 4-2. Metabolite standards in the high resolution dataset used for software validation. The abbreviations, retention times and molecular formulas are presented for each monitored ion.

4.2.3 Methods for baseline estimation

The evaluation of different baselining algorithms was accomplished by comparing five baselining methods. This analysis was performed in both the low- and high-resolution datasets. The root mean squared errors across all samples were analyzed using a repeated measurements ANOVA at a 0.05 significance level. Each method is described further below.

4.2.3.1 Simple noise-dependent baseline estimation

This method is based on a workflow described in a previous study [128]. In this method, the baseline for each extracted chromatogram is determined by filtering all scans for which the differences between the intensity of the *i*-th scan and the intensities of the *i*+1, *i*+2, *i*+3, *i*+4, and *i*+5 scans fall below a threshold starting at the noise value. If no scans are found, this threshold is

doubled, and the process is repeated. When this step converges to a solution and potential baseline scans are found, the 40 baseline scans located nearest to the peak apex are selected and their intensity values are averaged, yielding the final baseline value.

4.2.3.2 Wavelet transform-based method

This method aims to separate the baseline signal from the rest of a chromatogram in two steps:

First, the original intensity data C^0 at scan n is decomposed into discrete approximations C^j and discrete details D^j by means of an orthogonal wavelet decomposition following Eqs. 4-1a and 4-1b.

$$\boldsymbol{C^j(n) = C^{j-1}(n) * \widetilde{H}} \quad \text{(Eq. 4-1a)}$$

$$\boldsymbol{D^j(n) = C^{j-1}(n) * \widetilde{G}} \quad \text{(Eq. 4-1b)}$$

where $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{G}}$ are respectively the low and high pass filters of the wavelet function, at a *j*-th resolution level. Among these discrete transformations, there will be one (or multiple) of the discrete approximations C^k, which resembles the drifting baseline. In the second step, the baseline approximations are subtracted from the original chromatogram following Eq. 4-1c.

$$\boldsymbol{bsln = C^0 - C^k} \quad \text{(Eq. 4-1c)}$$

4.2.3.3 Entropy based baseline estimation method

This method is based on a previous study [214] which proposed to use the statistical entropy as a metric to distinguish the baseline level. The algorithm follows the following steps. First, each extracted ion chromatogram is normalized to a value between zero and one. This is done by dividing each value in the chromatogram by the intensity of the highest peak within it. Second, the entropy value (*S*) over each extracted chromatogram is calculated following Eq. 4-2.

$$\boldsymbol{S^m = \sum_n I_n^m \log(I_n^m)} \quad \text{(Eq. 4-2)}$$

where I_n^m is the normalized intensity of a metabolite with *m/z* value of *m* at the *n*-th scan time. Third, the ion chromatograms for which the entropy value is lower than a chosen threshold (defined as the sum of the mean and standard deviation of all the entropy values for the other extracted chromatograms) are retained, while the others are discarded. The baseline value can then be calculated as the average difference between the mean of the retained chromatograms and the original chromatogram.

4.2.3.4 Baseline Estimation And Denoising Using Sparsity (BEADS) method

In this algorithm based on previously published research [215], the baseline is modeled as a low-pass signal and the series of chromatogram peaks is modeled as sparse data with sparse derivatives.

This approach models an *N*-point chromatogram as a vector following Eq. 4-3a:

$$\boldsymbol{y = x + f + w} \quad \text{(Eq. 4-3a)}$$

where the vector *x* consists of numerous peaks (modeled as a sparse-derivative signal), the vector *f* represents the baseline (modeled as a low-pass filter signal), and *w* is a stationary white Gaussian process. An estimated baseline $\hat{f}$ can be then computed by filtering the measured chromatogram and an estimate of the peaks $\hat{x}$ with a low-pass filter *L*, as follows:

$$\boldsymbol{\hat{f} = L(y - \hat{x})} \quad \text{(Eq. 4-3b)}$$

If the value of $\hat{s}$ can be approximated as a decomposition of high (*H*) and low (*L*)-pass filters as follows,

$$\boldsymbol{\hat{s} = \hat{f} + \hat{x} = L(y - \hat{x}) + \hat{x} = L(y) + H(\hat{x})} \quad \text{(Eq. 4-3c)}$$

then the estimate of $\hat{x}$ is calculated from the observed data y. This is done by optimizing the quadratic fidelity term:

$$\left\|y - \hat{f} - \hat{x}\right\|_2^2 = \|y - L(y) - H(\hat{x})\|_2^2 = \|H(y - \hat{x})\|_2^2 \quad \text{(Eq. 4-3d)}$$

4.2.3.5 Peak elbow-based method

This method is described in section 4.3.3.

4.2.4 Methods for chromatogram scaling

The evaluation of different scaling methods was performed by comparing five of the most frequently used scaling techniques. For all metabolites in the low- and high-resolution standard mixtures, the scaled intensity is calculated from the extracted chromatograms using the following methods (Eqs. 4-4) before the peak edge determination step.

4.2.4.1 Simple scaling

$$\widehat{I_j} = \left[I_j - mean(I_c)\right] * std(I_c)^{-1} \quad \text{(Eq. 4-4a)}$$

4.2.4.2 Pareto scaling

$$\widehat{I_j} = \left[I_j - mean(I_c)\right] * std(I_c)^{-\frac{1}{2}} \quad \text{(Eq. 4-4b)}$$

4.2.4.3 Range scaling

$$\widehat{I_j} = \left[I_j - mean(I_c)\right] * [max(I_c) - min(I_c)]^{-1} \quad \text{(Eq. 4-4c)}$$

4.2.4.4 Level scaling

$$\widehat{I_j} = \left[I_j - mean(I_c)\right] * mean(I_c)^{-1} \quad \text{(Eq. 4-4d)}$$

4.2.4.5 Quantile scaling

$$\widehat{I_j} = \left[I_j - Q2(I_c)\right] * [Q3(I_c) - Q1(I_c)]^{-1} \qquad \text{(Eq. 4-4e)}$$

where $\widehat{\boldsymbol{I_j}}$ is the scaled intensity of the *j*-th point, $\boldsymbol{I_j}$ is the measured intensity of the *j*-th point, $\boldsymbol{I_c}$ represents the intensity values in the extracted chromatogram, $\boldsymbol{std}$ represents the standard deviation, and $\boldsymbol{Q_x}$ represents the *X*-th quartile.

The root mean squared error across all uncorrected isotopologues and the computation time were used as benchmarks to determine the best method. The statistical significance of this comparison was assessed by repeated measurements ANOVA and a Tukey's honestly significant difference test. A 0.05 significance level was used throughout the analysis. The computation time was calculated by placing the "tic/toc" holders within the MATLAB code. All evaluations were performed on a computer equipped with a 4-core Intel® Core™ i7-8550U CPU @ 1.8 GHz, 1992 Mhz, and 16.0 GB of RAM.

4.2.5 Validation of the natural abundance correction algorithm

To assess the impact of the non-tracer matrix on the overall accuracy of the correction algorithm, an evaluation of the natural abundance correction for the combined isotopologues of carbon and nitrogen in eight metabolites was conducted. This evaluation was done using both PIRAMID's algorithm and AccuCor2, a software that implements the complete set of three matrices for natural abundance correction, at a resolution of 70,000 FWHM. The entire list of metabolites and their corresponding isotopologues and MIDs can be found in Table 4-3. The statistical significance of the difference of the root mean squared error calculated across all isotopologues of each metabolite was assessed via a paired t-test using a significance level of 0.05.

4.2.6 Software validation

The average error (measured vs. expected) in the acquired MIDs of the standards was calculated for each one of the concentrations in the high-resolution dataset and compared against two other quantification tools: Skyline and El-MAVEN. For each metabolite, the expected MID was computed theoretically based on the known natural isotope abundance of each element present in the molecular formula. In the case of labeled glutamine standards, this calculation was done accounting for the purity of the tracer according to the vendor.

The comparison of multiple tools in quantifying unlabeled metabolites was determined based on the root mean squared error across all corrected isotopologues. In the case of labeled glutamine, this comparison was performed for each possible isotopologue accounting for the combinations of labeled carbon and nitrogen atoms. A repeated measurements ANOVA was used to determine the statistical significance of the differences between the programs. Given the dispersity in the errors across multiple metabolites, a Geisser-Greenhouse correction was implemented to the test. To test the individual differences in the errors of the tools versus PIRAMID, a Tukey's honestly significant difference test was performed. A 0.05 significance level was used throughout the analysis.

Asparagine	C4H8N2O3
C0N0	9.50E-01
C1N0	4.16E-02
C2N0	6.82E-04
C3N0	4.97E-06
C4N0	1.36E-08
C0N1	6.99E-03
C1N1	3.06E-04
C2N1	5.01E-06
C3N1	3.65E-08
C4N1	9.99E-11
C0N2	1.28E-05
C1N2	5.62E-07
C2N2	9.22E-09
C3N2	6.72E-11
C4N2	1.84E-13
Aspartate	C4H7NO4
C0N0	9.54E-01
C1N0	4.17E-02
C2N0	6.84E-04
C3N0	4.99E-06
C4N0	1.36E-08
C0N1	3.51E-03
C1N1	1.53E-04
C2N1	2.52E-06
C3N1	1.83E-08
C4N1	5.01E-11
Threonine	C4H9NO3
C0N0	9.54E-01
C1N0	4.17E-02
C2N0	6.84E-04
C3N0	4.99E-06
C4N0	1.36E-08
C0N1	3.51E-03
C1N1	1.53E-04
C2N1	2.52E-06
C3N1	1.83E-08
C4N1	5.01E-11

Glutamine	C5H10N2O3
C0N0	9.40E-01
C1N0	5.14E-02
C2N0	1.12E-03
C3N0	1.23E-05
C4N0	6.72E-08
C5N0	1.47E-10
C0N1	6.91E-03
C1N1	3.78E-04
C2N1	8.26E-06
C3N1	9.04E-08
C4N1	4.94E-10
C5N1	1.08E-12
C0N2	1.27E-05
C1N2	6.95E-07
C2N2	1.52E-08
C3N2	1.66E-10
C4N2	9.08E-13
C5N2	1.99E-15
Glutamate	C5H9NO4
C0N0	9.44E-01
C1N0	5.16E-02
C2N0	1.13E-03
C3N0	1.23E-05
C4N0	6.74E-08
C5N0	1.47E-10
C0N1	3.47E-03
C1N1	1.90E-04
C2N1	4.15E-06
C3N1	4.54E-08
C4N1	2.48E-10
C5N1	5.42E-13

Lysine	C6H14N2O2
C0N0	9.30E-01
C1N0	6.10E-02
C2N0	1.67E-03
C3N0	2.43E-05
C4N0	1.99E-07
C5N0	8.72E-10
C6N0	1.59E-12
C0N1	6.84E-03
C1N1	4.49E-04
C2N1	1.23E-05
C3N1	1.79E-07
C4N1	1.47E-09
C5N1	6.41E-12
C6N1	1.17E-14
C0N2	1.26E-05
C1N2	8.25E-07
C2N2	2.25E-08
C3N2	3.29E-10
C4N2	2.70E-12
C5N2	1.18E-14
C6N2	2.15E-17
Glycine	C2H5NO2
C0N0	9.75E-01
C1N0	2.13E-02
C2N0	1.17E-04
C0N1	3.58E-03
C1N1	7.84E-05
C2N1	4.29E-07
Serine	C3H7NO3
C0N0	9.64E-01
C1N0	3.16E-02
C2N0	3.46E-04
C3N0	1.26E-06
C0N1	3.55E-03
C1N1	1.16E-04
C2N1	1.27E-06
C3N1	4.63E-09

Table 4-3. Experimentally determined MIDs of multiple amino acids used to assess the validity of the natural abundance correction algorithm in datasets containing dual labeled data.

4.3 PIRAMID workflow

The workflow of PIRAMID can be broken down into three steps: i) data extraction, ii) peak finding and integration, and iii) data analysis and output (Fig. 4-1). Each step is described in more detail below.

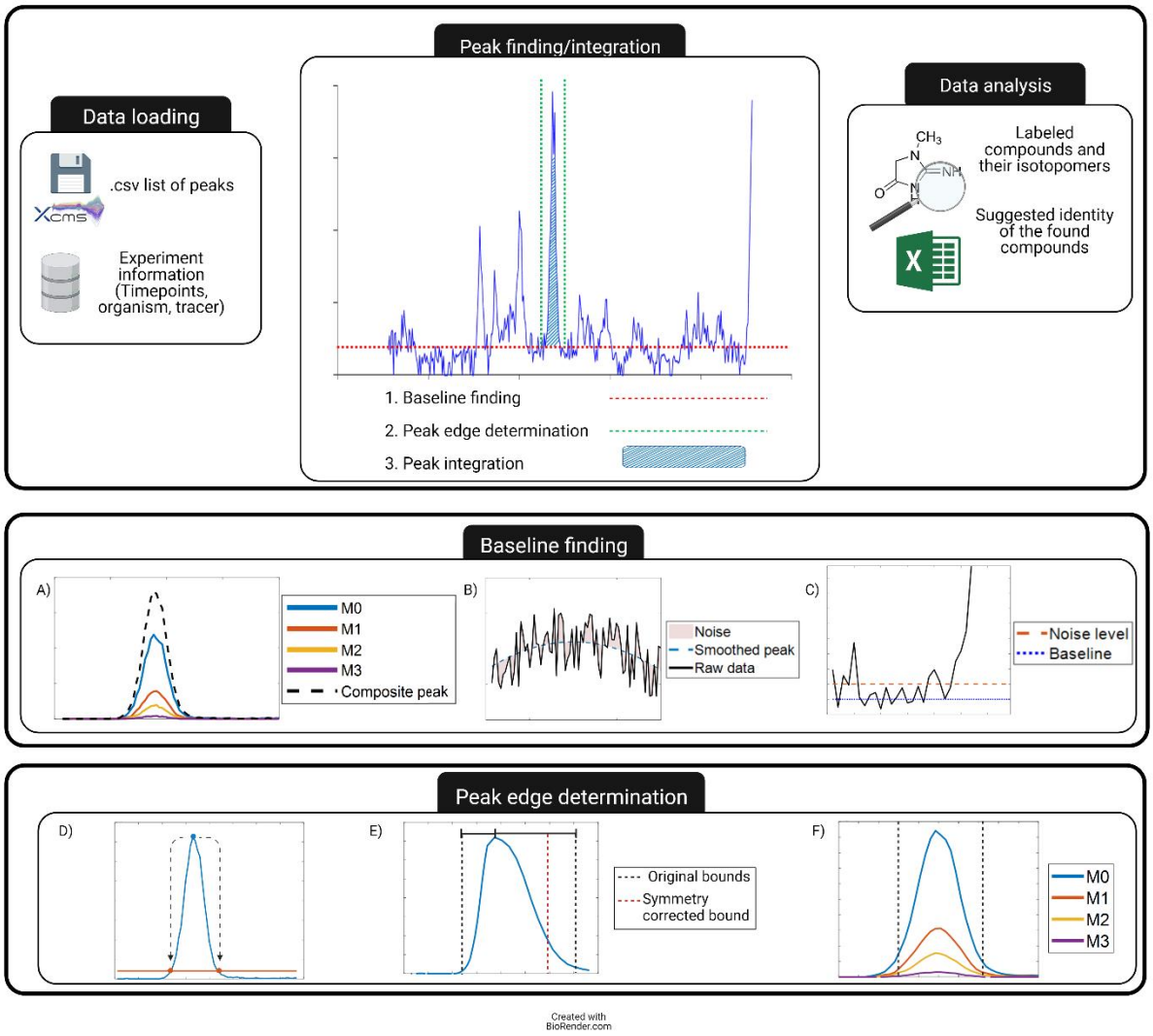


Figure 4-1. PIRAMID data processing workflow. *The workflow begins with a data extraction step that involves loading the raw MS files and a MATLAB .m method file containing information on the target peaks. The tool proceeds to find, process, and integrate the corresponding peaks in each file. Finally, the MIDs and total ion counts are evaluated, visualized, and exported.* ***A)*** *Creation of a composite peak comprising all isotopologues of a target ion.* ***B)*** *Noise level estimation.* ***C)*** *Baseline determination.* ***D)*** *Peak edge determination.* ***E)*** *Peak edge correction for asymmetric peaks.* ***F)*** *Consistent boundaries are applied to all isotopologues during peak integration to ensure accurate MIDs are obtained.*

4.3.1 Data extraction

PIRAMID requires two input files: (i) the raw MS data in .cdf, .mzml, or .mzxml format and (ii) a MATLAB .m method file containing information on each of the target compounds (e.g., their retention times (RTs), characteristic ion(s), and respective isotopologues to be quantified). Plugins are integrated into the main GUI to facilitate the creation and updating of the method file. Dual-labeled metabolites are entered by specifying the combinations of labeled atoms sorted by the first tracer atom, which is assumed to be carbon, followed by the second tracer atom (e.g., Glu-C0N0, Glu-C1N0, Glu-C2N0, Glu-C0N1, Glu-C1N1, Glu-C2N1, …).

Upon loading the files, the extracted ion chromatograms (EICs) associated with each target compound are extracted and loaded into a MATLAB structure containing the RT and intensity information of the monitored ions within a time window surrounding the expected RT. If multiple signals match the same *m/z* range (to within the *m/z* tolerance specified by the user), a composite EIC is created by summing the corresponding signals. In the specific case where deviations in the RT occur across multiple files, the extracted chromatograms can be aligned using a time-warping method based on the built-in MATLAB functions "*alignsignals*" and "*delayseq*".

4.3.2 Peak assignment and smoothing

For each of the target compounds in the method file, a peak is matched and assigned using a probabilistic method based on the compound's characteristic ion(s) and expected RT provided in the method file. First, the total intensity across all the characteristic ions of the target metabolite is summed and multiplied by a probability function that accounts for the proximity of the peak apex to its expected RT. If multiple characteristic ions are specified, a second probability function is multiplied that accounts for their expected relative intensities. The most likely peak location is

defined as the RT value at which the product of the probabilities is the greatest. The apex of the peak is found using the MATLAB built-in function "*findpeaks*". Next, a composite peak is created for each isotopic cluster, which contains all the isotopologues associated with a user-specified target ion of interest, by summing the intensities of the isotopologues to be quantified (Fig. 4-1A). Finally, the EICs of the composite peak and each individual isotopologue are smoothed using a Savitzky-Golay filter of user-specified order (second order is the default), and the noise level is estimated as the mean difference between the raw and smoothed signals (Fig. 4-1B).

4.3.3 Baseline calculation

The baseline is calculated as the mean of the intensities of the closest points to the elbows of the target peak for which their intensity values are lower than the noise level (Fig. 4-1C). The number of points used in this calculation is 2% of the total data points in the extracted chromatogram. The elbows of the peak are determined through the algorithm depicted in Fig. 4-2 and explained as follows:

1) The peak apex and the width at half-height are determined via the native function *findpeaks* (Fig. 4-2A).

2) From the peak apex, two lines are projected to the x-intercept at a value equal to 3 times the peak width in each direction (Fig. 4-2B).

3) From these lines, orthogonal projections to the signal are created. (Fig. 4-2C).

4) The length of the orthogonal projections to the signal is calculated. The maximum length intersects the point of highest inflection which is considered the elbow of the peak (Fig. 4-2D).

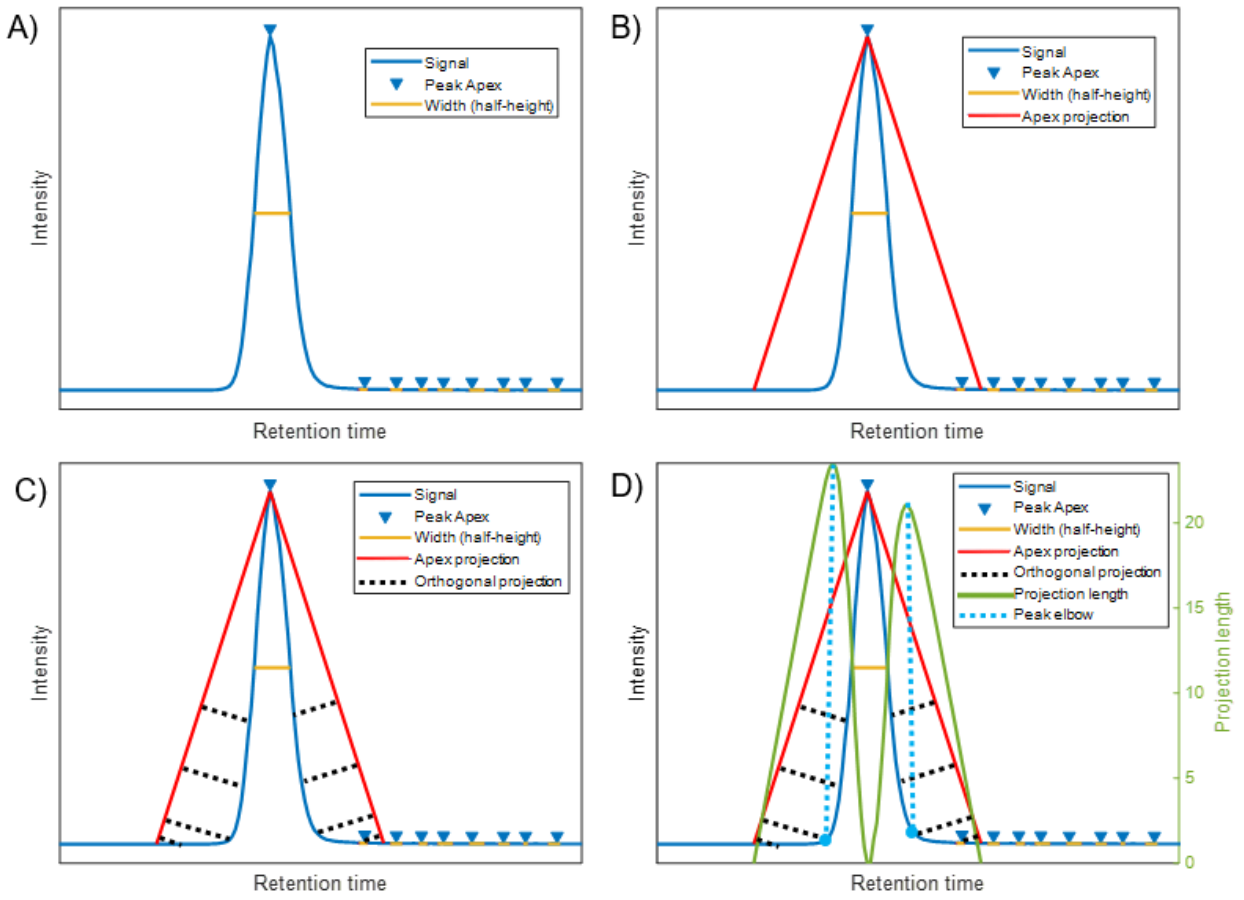


Figure 4-2. Elbow determination algorithm. *A) The peak apex and signal are determined. B) The apex is projected to the signal at a fixed distance. C) From the projection of the previous part, orthogonal projections are built to the signal. D) The length of the projections of the last step are calculated and the elbows are determined.*

In the case where not enough data points are available to use the described method accurately, the baseline will be set as the mean of intensities of the points for which their absolute derivative falls within the lowest decile of all points.

4.3.4 Peak edge determination

The edges of the peaks are determined by scanning away from the apex of the composite peak in both directions until a threshold signal (defined as the sum of the baseline and the average noise level) is found (Fig. 4-1D). To avoid pitfalls arising from common artifacts like peak tailing and shouldering, the edges of asymmetric peaks (defined as peaks with significantly different distances

between the apex and the upper and lower bounds) are corrected by adjusting the distance between the apex and the farthest edge on either side to be equidistant to the shortest edge (Fig. 4-1E).

4.3.5 Integration and MID determination

The integration of each EIC is computed as the sum of intensities between the peak edges minus the sum of the average noise level and value of the baseline multiplied by the number of data points between the peak edges. The MID of each target ion is calculated as the relative area abundance of its isotopologues. The same peak edges are applied to integrate all isotopologues of a given target ion (Fig. 4-1F), which ensures accurate quantification of low abundance isotopologues that may exhibit poor peak shape.

4.3.6 Theoretical MID calculation

Multiple algorithms require the estimation of the theoretical MID of a molecule in its natural state or with the incorporation of a certain number of labeled atoms. The isotopic distribution of any molecule can be calculated as the distribution of a random variable, *X*, which is the sum of the random variables corresponding to the isotopic distributions of all the elements, *E*, the molecule contains [216], [217]:

$$X = \sum X_E = X_H + X_C + X_N + X_O + X_s + \cdots \quad \text{(Eq. 4-5)}$$

The random variable X_E is described by the binomial random distribution of an element *E* according to Eq. 4-6:

$$p_{(k_I)} = \frac{N_E!}{k_I * \prod_{i=0}^{I-1} k_i!} * \prod_{i=0}^{I} p_i^{k_i} \quad \text{(Eq. 4-6)}$$

where N_E is the total number of atoms of the element *E*, and k_i is the fraction of each one of the possible isotopologues of *E* up to a maximum number of naturally occurring isotopes *I*, with the

corresponding probabilities p_i. The first term in Eq. 4-6 addresses the possible combinations of naturally occurring isotopes in a molecule of an element, while the second term addresses the probability of the natural occurrence of the isotopes in each atom. Overall, Equation 4-7 models the exact probability of isotope distributions for any atom. The general relationships of multinomial distributions hold:

$$N_E = \sum_{i=0}^{I} k_i \quad \text{(Eq. 4-7a)}$$

$$1 = \sum_{i=0}^{I} p_i \quad \text{(Eq. 4-7b)}$$

The probability distribution of a molecule containing atoms with different probability distributions is a convolution (in the case of the isotope distributions, a discrete convolution) of probability distributions of each random variable following Eq. 4-8:

$$P(X) = P(X_H) \otimes P(X_C) \otimes P(X_N) \otimes P(X_O) \otimes P(X_S) \otimes \ldots \quad \text{(Eq. 4-8)}$$

Given that Eq. 4-8 is not a closed form formula and that the number of possible atom configurations increases polynomially in large molecules, solving this problem isotope-by-isotope can become a resource-intensive task. To address this issue, a series of convolutions are performed on the mass isotopomer distribution vector (i.e., a vector containing the abundance distribution of each one of the naturally occurring isotopes) corresponding to each element present in the molecular formula. The mass isotopomer distribution vector for each atom is calculated based on a previously reported library [218], and the elements and corresponding isotopes taken into consideration for the algorithm are presented in Table 4-4.

Element	Stable isotopes	M0	M1	M2	M3	M4
Hydrogen	^{1}H, ^{2}H	0.9998	0.0002	-	-	-
Carbon	^{12}C, ^{13}C	0.9892	0.0108	-	-	-
Nitrogen	^{14}N, ^{15}N	0.9963	0.0037	-	-	-

Oxygen	^{16}O, ^{17}O, ^{18}O	0.9976	0.0004	0.0020	-	-
Fluorine	^{19}F	1	-	-	-	-
Silicon	^{28}Si, ^{29}Si, ^{30}Si	0.9222	0.0469	0.0309	-	-
Phosphorus	^{31}P	1	-	-	-	-
Sulfur	^{32}S, ^{33}S, ^{34}S, ^{36}S	0.9504	0.0075	0.0420	0	0.0002
Chlorine	^{35}Cl, ^{37}Cl	0.7577	0.0000	0.2423	-	-

Table 4-4. Elemental natural abundance of stable isotopes used in the software.

To make the process less computationally intensive, each convolution is calculated following the convolution theorem, as the inverse Fourier transform of the pointwise product of the Fourier transform of the convolved vectors [219] (Eq. 4-9)

$$g \otimes \mathrm{h} = \mathcal{F}^{-1}\big(\mathcal{F}(g) \cdot \mathcal{F}(h)\big) \qquad \text{(Eq. 4-9)}$$

where g and h are the vectors to be convolved.

For the algorithms that require the estimation of the MID of a molecule containing one or more labeled isotopes, the same method is followed except the mass isotopomer distribution vectors of the labeled atoms are replaced with a vector containing the relative abundance of each isotope according to the isotopic purity of the tracer (e.g., [0.01, 0.99] reflects a ^{13}C tracer with 99% purity, whereas [0, 0, 1] would reflect a pure ^{18}O isotope). The inclusion of the reported purity of the tracer in MID calculations has been discussed in previous studies [220], concluding that it increases the accuracy of the results compared to the cases where the tracer distribution is modeled as 100% pure.

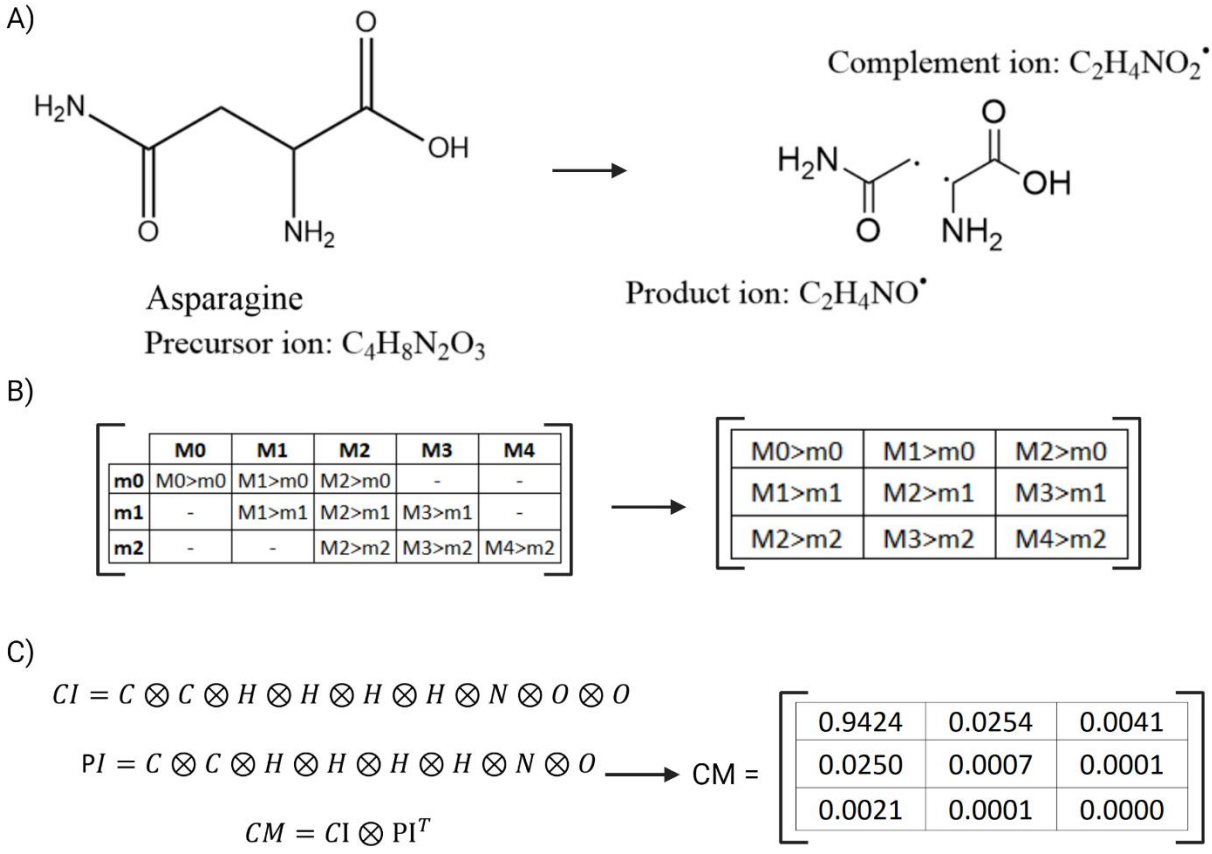
4.3.7 MID correction for natural isotope abundance

The MIDs can be corrected for natural isotope abundance based on the chemical formula of the ion using different methods depending on the type of data that is being analyzed. If low-resolution data is used, the problem in Eq. 2-1 is solved using a simple correction matrix [134] (Eq. 4-10).

$$CM = [MID_{UL}; MID_{L=1};\ MID_{L=2}\ ;\ \dots\ ] \qquad \text{(Eq. 4-10)}$$

where the rows of the correction matrix *CM* correspond to the row vector of the MIDs of the corrected molecule with increasing degrees of labeling (*L*). In the first row, the theoretical MID of the unlabeled molecule would be found. The second row would contain the theoretical MID of the molecule enriched with a single labeled atom, and the subsequent columns would proceed in a similar manner.

If MS^2 data is used, the natural abundance correction problem involves a 2D deconvolution rather than a simple matrix inversion. This problem is solved by means of a tandem correction matrix that considers the transitions from the precursor ion into the product ion [136]. The precursor ion is split into the product ion that is analyzed by the mass spectrometer and a complement ion that is not detected (Fig. 4-3A). The measured compact matrix is built based on the ions that are detected by MS2 taking into account only the feasible transitions (e.g., it is impossible for a precursor ion with 2 labeled atoms to yield a product ion with 3 labeled atoms) (Fig. 4-3B). The correction matrix CM is calculated as a series of convolutions between the vectors of the complement ion and the transposed vectors of the product ion (Fig. 4-3C).



	M0	M1	M2	M3	M4
m0	M0>m0	M1>m0	M2>m0	-	-
m1	-	M1>m1	M2>m1	M3>m1	-
m2	-	-	M2>m2	M3>m2	M4>m2

M0>m0	M1>m0	M2>m0
M1>m1	M2>m1	M3>m1
M2>m2	M3>m2	M4>m2

0.9424	0.0254	0.0041
0.0250	0.0007	0.0001
0.0021	0.0001	0.0000

Figure 4-3. Tandem MS natural abundance correction. *A) The target molecule is split into a complement ion and a product ion. B) A compact tandem MID matrix is built based on the feasible transitions between precursor and product ions. C) The correction matrix (CM) is calculated as the convolution between the complement ion (CI) and the transposed product ion (PI).*

Finally, if high-resolution data are used in conjunction with a dual-labeled tracer, the isotopologue distributions are only corrected for the contribution of the potentially labeled atoms given the complexity of accounting for the natural abundance of all the unlabeled atoms in the molecule. This process was accomplished using an adaptation of a previously described approach [137]. In this case, it is assumed that the MS resolution can distinguish isotopologues based on the mass defect between the isotopes of different atoms (i.e., the mass shift between ^{12}C and ^{13}C is different from the mass shift between ^{14}N and ^{15}N). Hence, the measured vectors and the correction matrix need to account for the combinations of each labeled atom. Following this logic, the isotopologues

are resolved into specific combinations of isotopes. An example with carbon and nitrogen is presented in Eqs. 4-11.

$$M + 2 \rightarrow M_{C2N0} + M_{C1N1} + M_{C0N2} \quad \text{(Eq. 4-11A)}$$

$$M + 3 \rightarrow M_{C3N0} + M_{C2N1} + M_{C1N2} + M_{C0N3} \quad \text{(Eq. 4-11B)}$$

The correction matrix needs to be adapted accordingly. For this specific algorithm, the correction matrix (*CM*) is calculated as follows:

$$CM = TM \times IPM \quad \text{(Eq. 4-12)}$$

where *TM* is the tracer matrix, which contains information of the atoms that can be labeled, and *IPM* is the isotopic purity matrix, which contains information of the purity of the tracers. Theoretically, a non-tracer matrix accounting for the natural abundance of the remaining atoms should be included. However, the calculation of this matrix depends on the equipment resolution and the combinatorial probabilities of the contribution of multiple elements, which made the algorithm significantly more complicated and slower. As shown in section 4.4.3, the differences between the corrected data with versus without inclusion of the non-tracer matrix were negligible, so the decision was made to neglect it.

Using the example of a $^{13}C/^{15}N$ labeling experiment, the tracer matrix is a square matrix with $(m + 1) \times (n + 1)$ rows and columns, where *m* represents the number of carbon atoms and *n* the number of nitrogen atoms present in the measured molecule. The tracer matrix is built so that it is composed of $(n + 1) \times (n + 1)$ blocks where each block is an $(m + 1) \times (m + 1)$ submatrix. In general, the matrix in the *i*-th row and *j*-th column of the block matrix and the *f*-th row and the *g*-th column of the submatrix is filled as follows:

$$TM[(i-1)\times(m+1)+f,(j-1)\times(m+1)+g] = {}_{m+1-g}^{f-g}P_C \times {}_{m+1-j}^{i-j}P_E \quad \text{(Eq. 4-13)}$$

where ${}_{y}^{x}P_E$ represents the probability of having *x* isotopes out of *y* natural atoms for the element *E*. Similarly, the *IPM* is a square matrix composed of $(n+1)\times(n+1)$ blocks where each block is an $(m+1)\times(m+1)$ submatrix that is filled by computing the probability of finding a non-labeled atom in the pool of expected labeled atoms due to the impurities in the tracer. It follows the equation:

$$IPM[(i-1)\times(m+1)+f,(j-1)\times(m+1)+g] = {}_{g-1}^{g-f}P_{IPC} \times {}_{j-1}^{j-i}P_{IPE} \quad \text{(Eq. 4-14)}$$

Where ${}_{y}^{x}P_{IPE}$ represents the probability of having *x* unlabeled atoms out of *y* atoms of *E* in the tracer.

4.3.8 Post-processing calculations

Using the information of the corrected data, PIRAMID calculates the atom percent enrichment (APE) of each target analyte, defined as $\frac{\sum_{i=0}^{N} i\times Mi}{N}$, where M_i represents the relative abundance of the mass isotopomer with *i* heavy atoms and N represents the maximum number of atoms than can be enriched by the isotope tracer. If the timepoint data information is available, the program will calculate the root mean-square error between the unlabeled samples (*t*=0) and the theoretically predicted MID based on the target ion formula.

4.3.9 Data analysis and output.

Optionally, a statistical analysis of up to 10 combinations of sample time points and experimental groups can be performed. Once the comparisons are specified, A two-tailed t-test is performed to

compare the intensities of each selected isotopologue and the APE of the metabolite. The user can also choose to normalize the total ion counts of target analytes to an internal standard peak in order to compare relative metabolite abundances across different samples. Plots of the data time courses or group comparisons can be generated interactively. The results are exportable as two .xls files: one with the integrated MIDs, APEs and ion abundances for each sample and, if sufficient information is provided by the user, another containing the summarized results of each timepoint/experimental group and the statistical comparison of the selected pairs.

4.4 Results

4.4.1 Optimization of baselining algorithm

The root mean squared error of all analyzed metabolites in both the low (-L) and high (-H) resolution datasets is presented in Table 4-5. The ANOVA test yielded an F-statistic equal to 1.84 and a p-value of 0.125, meaning that there were no statistically significant differences between the methods. However, there are some caveats that require further exploration. Some algorithms were not able to fully process the data from specific extracted chromatograms. For instance, the wavelet transform-based algorithm could not process the peaks from Pyr-L. The simple noise-dependent baseline estimation method was not able to process the peaks from Ser-H and Val-H. Finally, the entropy-based algorithm was unable to process the peaks of Iso-H, Ser-H, and Suc-H. The explanation of these differences can be attributed to the variable quality of the chromatograms. Figure 4-4 shows the EICs of the monoisotopic mass isotopologue and the first two carbon isotopologues of the aforementioned metabolites.

	Root-mean-squared error				
Metabolite	**Simple noise-dependent**	**Wavelet transform**	**BEADS**	**Entropy**	**Peak elbow**
Ala-L	0.0311	0.004	0.0276	0.0019	0.002
Asp-L	0.0037	0.0013	0.0074	0.0028	0.0005
Cit-L	0.0014	0.0015	0.0021	0.0019	0.0015
Gln-L	0.0043	0.0044	0.0062	0.0084	0.0042
Glu-L	0.0007	0.0006	0.0016	0.001	0.0008
Gly-L	0.003	0.003	0.0008	0.0069	0.0025
Lac-L	0.0129	0.0004	0.0008	0.0011	0.0045
Mal-L	0.0029	0.0047	0.0111	0.008	0.0029
Nor-L	0.0005	0.0006	0.0026	0.0044	0.0003
Pyr-L	0.0418	**0.1126**	0.0444	0.0024	0.0319
Ser-L	0.0017	0.0018	0.0031	0.0111	0.0017
Suc-L	0.0004	0.0004	0.0007	0.0008	0.0004
Ala-H	0.0044	0.0044	0.0044	0.0044	0.0042
Asp-H	0.0014	0.0015	0.0015	0.002	0.0016
Cit-H	0.0025	0.0587	0.0554	0.0039	0.0187
Frc-H	0.0352	0.0539	0.0538	0.0434	0.0539
G1P-H	0.0355	0.0355	0.0355	0.0355	0.0355
G6P-H	0.0355	0.0355	0.0355	0.0355	0.0355
Glc-H	0.0013	0.0013	0.0013	0.0014	0.0012
Glu-H	0.0011	0.0015	0.0015	0.0019	0.0015
Iso-H	0.0327	0.0327	0.0327	**0.378**	0.0327
Leu-H	0.0007	0.0007	0.0007	0.0021	0.0006
Mal-H	0.0012	0.0012	0.0011	0.001	0.0011
Met-H	0.0011	0.0016	0.0012	0.0032	0.0024
Phe-H	0.0008	0.0007	0.0008	0.0037	0.0007
Pro-H	0.0009	0.001	0.001	0.0017	0.0009
Ser-H	**0.378**	0.0332	0.0332	**0.378**	0.0332
Suc-H	0.0334	0.0319	0.0318	**0.4472**	0.0314
Thr-H	0.0023	0.0037	0.0034	0.002	0.0029
Tyr-H	0.0009	0.0008	0.0008	0.0013	0.001
Val-H	**0.4472**	0.0345	0.0345	0.0345	0.0314

Table 4-5. Calculated root-mean-squared errors for each of the tested baselining algorithms based on the difference between the expected natural abundance and the measured MIDs for the unlabeled metabolite mixture in the low (-L) and high (-H) resolution datasets. Specific metabolites such as Pyr-L, Ser-H, or Val-H could not be correctly processed by some of the methods, yielding high errors (highlighted in bold).

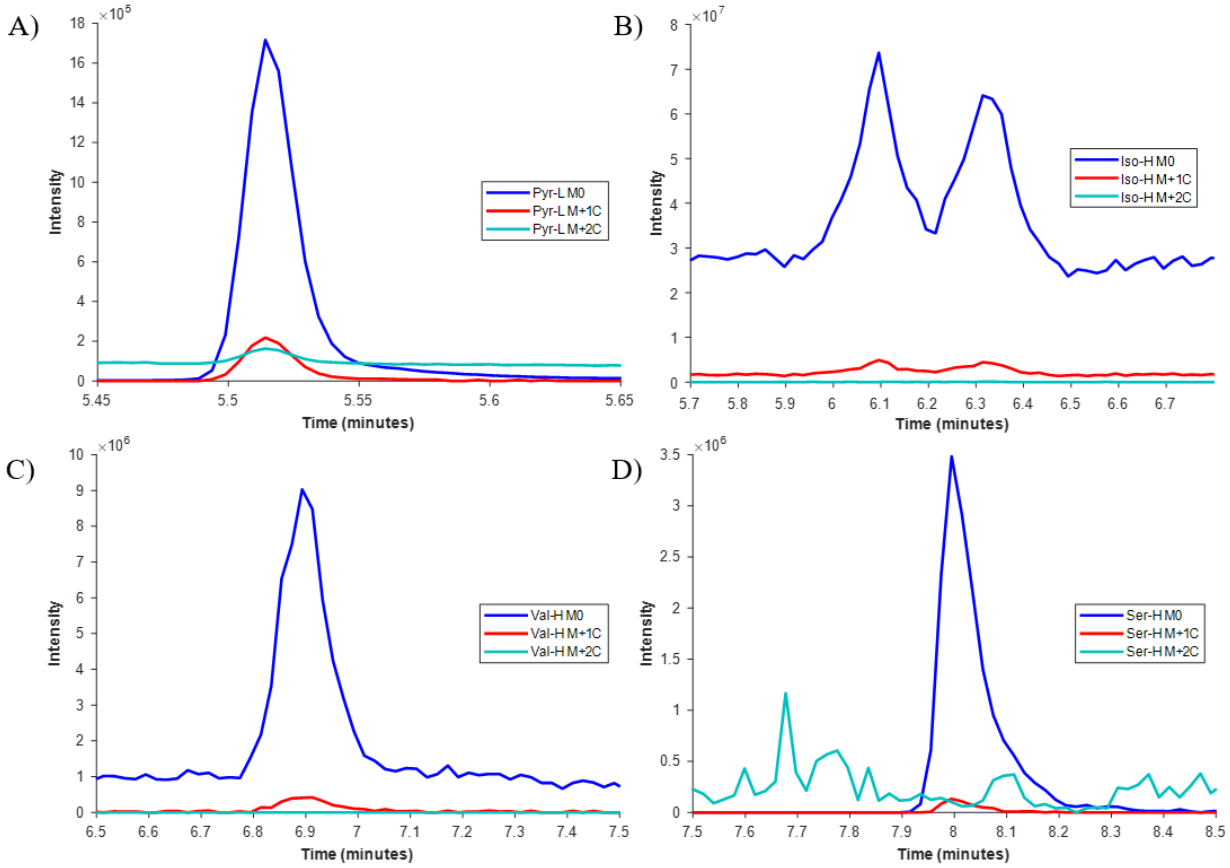


Figure 4-4. Extracted ion chromatograms of metabolites that could not be processed by certain methods. *All metabolites show noisy chromatograms or elevated baselines which interferes with the methods. A) Pyruvate – Low resolution. B) Isoleucine – High resolution. C) Valine – High resolution. D) Serine - High resolution.*

In the wavelet transform-based algorithm, the methodology is optimized for high-resolution datasets that contain multiple data points. To execute the wavelet transform, it is necessary to rescale the data so that the total number of data points becomes a power of 2, as the dataset is halved at each decomposition step. In the case of low-resolution MS, data is sampled at a lower rate, leading to a reduced number of data points. This reduction in data points has a direct impact on the wavelet transform, as it diminishes the number of data points available for the decomposition process. For isotopologues that exhibit asymmetric peak shapes, like Pyr-L M+2C, the resulting number of data points may not be sufficient to accurately perform the decomposition. This limitation can introduce errors into the results of the wavelet transform-based method.

Regarding the simple noise-dependent approach, the algorithm has a tendency to fail if the baseline is not reached within the first few iterations of the algorithm. As the threshold value increases in proportion to 2^N, it grows rapidly with each subsequent iteration, encompassing even the points belonging to the peak region as potential baseline points and leading to an overestimation of the baseline. In the specific case of Iso-H M0, the algorithm fails to converge to a threshold value as the final iteration considers all data points within the extracted chromatogram as potential baseline points, resulting in an error in the analysis.

Finally, in the case of the entropy-based method, the algorithm is optimized to include several ions in the calculation of the entropy threshold. It also estimates a single baseline value for the entire chromatogram, which was found to be inaccurate in cases where certain isotopologues clearly show a different baseline from others.

Once these less robust methodologies are discarded due to the aforementioned limitations, the only remaining options are the BEADS method and the peak elbow-based method. A t-test between the results of these two algorithms resulted in a p-value of 0.047, meaning that the difference between their performances is barely significant from a statistical standpoint. As the peak elbow-based method yielded a lower average error across metabolites analyzed in both low- and high-resolution, it was selected as the baselining algorithm implemented in PIRAMID.

4.4.2 Determination of a chromatogram scaling algorithm

Recent studies have shown the effects of chromatogram scaling, in some cases also known as normalization, in the inference of biological information from mass spectrometry-based experiments [221], [222], [223]. To assess whether an extra scaling step would enhance the accuracy of metabolite integration, we employed five of the most commonly used scaling methods

in the integration of both the low and high resolution datasets and compared the resulting relative errors to the results obtained without scaling, as depicted in Figure 4-5A. However, the determination of whether an additional step should be incorporated into the overall data processing workflow should not be solely guided by accuracy considerations. The introduction of an extra step inevitably extends computation time and, in some instances, can complicate the data, potentially influencing overall performance. Hence, we also quantified the additional computation time required for each of the tested scaling methods, as shown in Figure 4-5B.

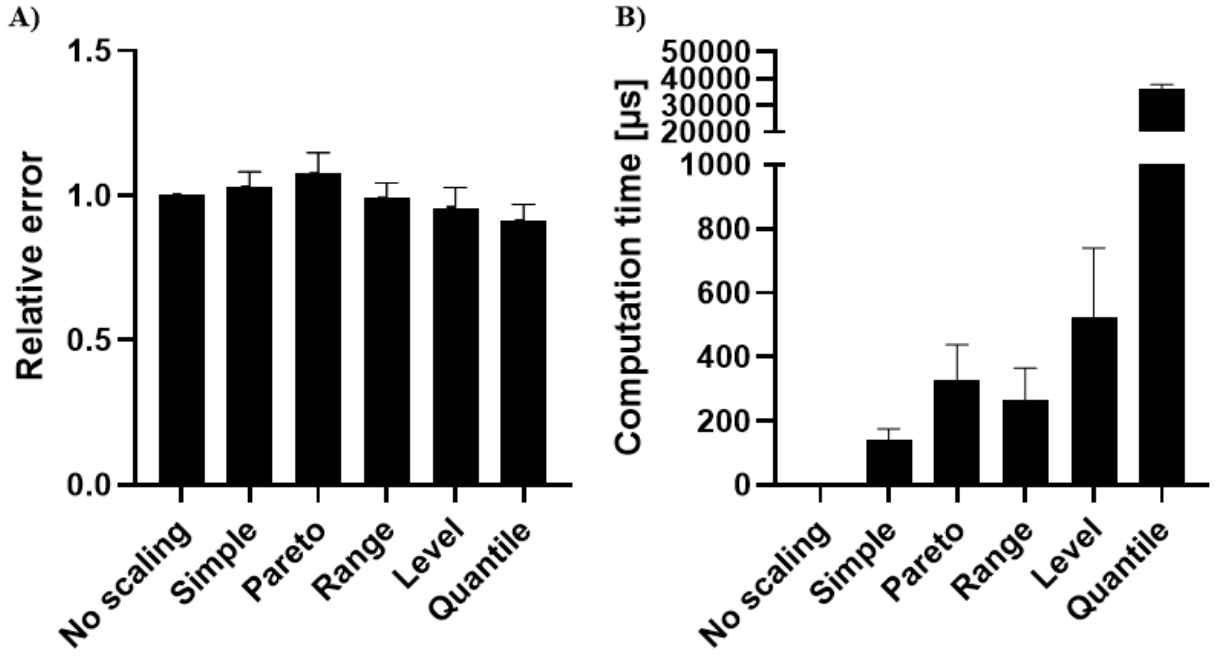


Figure 4-5. Performance comparison evaluating the implementation of different scaling methods. *A) Relative error compared to the results obtained without scaling the data before integration. B) Average additional computation time per metabolite required to run the implemented method. The implementation of a scaling step barely reduced the errors in the integration but took proportionally longer times to compute.*

No apparent improvement in the accuracy of the integration was found in any of the evaluated cases. The results of the statistical significance analysis on the relative error are presented in Table 4-6.

	ANOVA	Simple vs NS	Pareto vs NS	Range vs NS	Level vs NS	Quant vs NS
F-test statistic	0.959	0.001	0.243	0	0.083	0.343
p-value	0.446	0.999	0.941	1	0.995	0.884

Table 4-6. Statistical analysis of the differences of root mean squared errors due to implementing multiple scaling algorithms prior to metabolite integration. No statistically significant differences were found between the conditions.

No statistically significant differences were found between the various implemented scaling algorithms and the 'no scaling' control. In all cases, there was an increase in the computation time due to scaling, and the data could not be processed by the quantile scaling algorithm in one case. Given that no significant improvement in the accuracy of the integration was found, it was decided to not scale the data before the analysis to avoid cases where the implementation of the scaling algorithm interferes with the normal function of PIRAMID.

4.4.3 Validation of the natural abundance correction algorithm

The conventional algorithm employed for correcting dual-labeled data for isotopic natural abundance creates a correction matrix that is a function of three matrices: a labeled atom matrix, a non-labeled atom matrix, and a tracer purity matrix [137]. As discussed in section 4.3.7, PIRAMID's approach for correcting natural abundance does not incorporate the non-labeled atom matrix. The calculated root mean squared errors for each of the analyzed metabolites are presented in Figure 4-6.

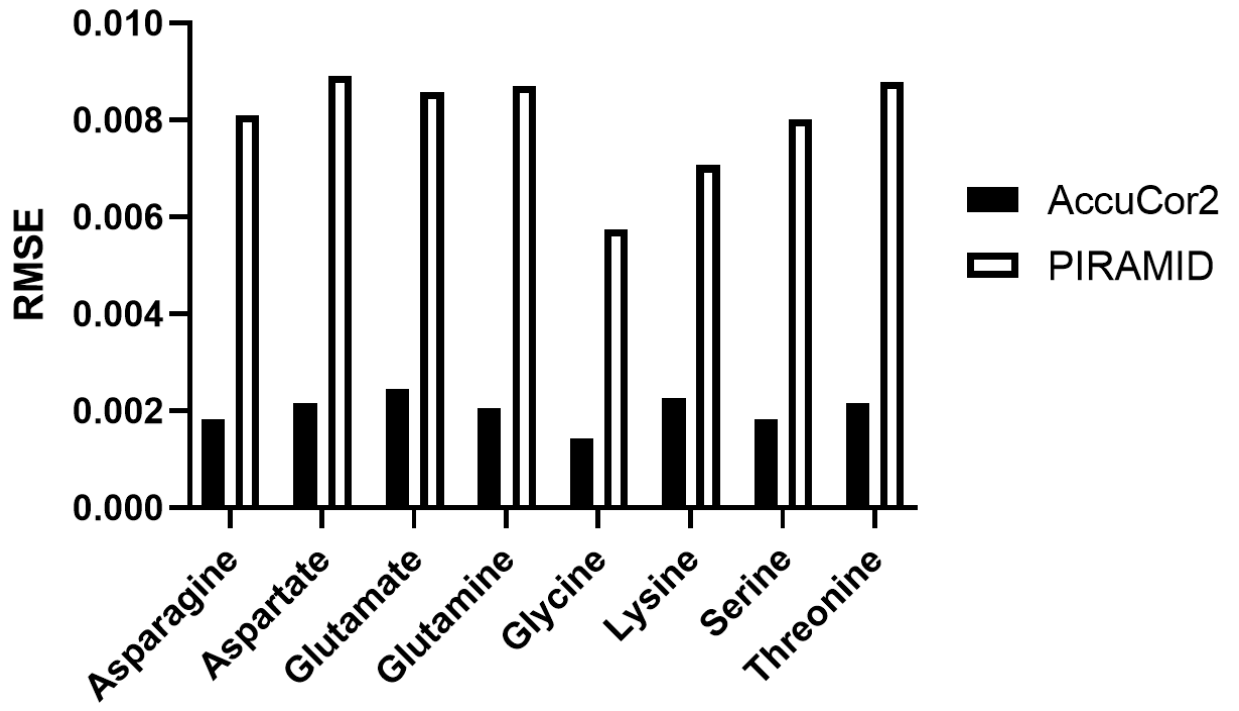


Figure 4-6. Comparison of the errors associated with the incorporation of a non-labeled atom matrix (AccuCor2) and its omission (PIRAMID). *The implementation of the non-labeled atom matrix reduces the errors significantly. However, compared to the inherent errors found in HRMS datasets, the errors found from omission of the non-labeled atom matrix are negligible.*

Despite employing theoretical MIDs in the calculations, both tools exhibit some degree of error in the results. The errors in AccuCor2 can be attributed to the inherent challenge of precisely distinguishing isotopologues from the natural abundance of non-tracer atoms, such as hydrogen and oxygen. Notably, when the non-labeled atom matrix is omitted from the natural abundance correction, the errors are markedly higher than when it is included in the calculation. In our practical experience, the acceptable range of errors in LC-MS and HRMS measurements typically varies around 1-3%, while GC-MS tends to yield errors around 0.5-1%. It's noteworthy that all the errors in our approximations fall below the 1% threshold.

Our software has the limitation of requiring carbon to be one of the labeled atoms used as a tracer. It is important to note that most stable isotope experiments use carbon labeling in combination with other isotopes. Hence, we believe this limitation should not significantly hinder potential

users from leveraging PIRAMID in their studies. Additionally, our previous efforts to fully integrate the non-labeled atom matrix into the natural abundance correction process have proven to be unsuccessful. As a result, we warn users that there may be an increased margin of error in the correction process. The comprehensive implementation of an algorithm capable of correcting the natural abundance of MIDs containing any combination of atoms, irrespective of carbon, and incorporating the non-labeled atom matrix remains a subject for future research.

4.4.4 Software validation

The calculated root mean squared errors across all isotopologues of each metabolite in the high-resolution dataset are presented in Table 4-7. Across all metabolites tested, PIRAMID demonstrated consistently lower errors compared to El-Maven and Skyline, two widely used programs for metabolomics data analysis. However, the differences between the programs were not statistically significant, as shown in Table 4-8.

	El-MAVEN			Skyline			PIRAMID		
	62.5 µM	**125 µM**	**250 µM**	**62.5 µM**	**125 µM**	**250 µM**	**62.5 µM**	**125 µM**	**250 µM**
Ala	0.168%	0.247%	0.100%	0.321%	0.267%	0.343%	0.256%	0.216%	0.273%
Asp	0.051%	0.068%	0.047%	0.114%	0.047%	0.270%	0.049%	0.025%	0.084%
Cit	0.858%	0.460%	0.474%	0.510%	0.452%	1.866%	0.483%	0.496%	0.999%
Fru	0.590%	0.549%	0.403%	0.706%	3.435%	2.312%	0.706%	0.364%	0.967%
Fum	0.618%	0.274%	0.313%	0.665%	0.443%	0.355%	0.354%	0.366%	0.392%
G1P	0.097%	0.089%	0.162%	0.161%	0.148%	0.183%	0.116%	0.165%	0.140%
G6P	0.091%	0.085%	0.157%	0.104%	0.108%	0.098%	0.039%	0.099%	0.048%
Glc	0.445%	0.660%	0.401%	0.841%	0.760%	0.907%	0.799%	0.477%	0.885%
Glu	0.043%	0.066%	0.170%	0.654%	0.111%	0.618%	0.105%	0.024%	0.040%
Iso	0.009%	0.020%	0.053%	0.111%	0.181%	0.063%	0.022%	0.044%	0.016%
Leu	0.009%	0.020%	0.053%	0.149%	0.135%	0.105%	0.028%	0.069%	0.017%
Mal	0.301%	0.334%	0.294%	0.438%	0.385%	0.442%	0.282%	0.273%	0.272%
Met	0.071%	0.105%	0.045%	0.241%	0.202%	0.651%	0.147%	0.057%	0.211%
Phe	0.024%	0.023%	0.021%	0.115%	0.107%	0.117%	0.026%	0.020%	0.023%
Pro	0.073%	0.070%	0.144%	0.133%	0.138%	0.116%	0.044%	0.085%	0.062%
Ser	9.439%	6.365%	2.664%	16.260%	3.704%	18.488%	2.743%	5.128%	0.459%
Suc	0.263%	0.305%	0.251%	0.614%	0.637%	0.454%	0.324%	0.309%	0.442%
Thr	0.396%	0.342%	0.107%	3.254%	3.359%	2.358%	0.959%	0.306%	0.692%
Tyr	0.024%	0.019%	0.022%	0.103%	0.117%	0.081%	0.014%	0.013%	0.035%
Val	0.060%	0.050%	0.013%	0.085%	0.124%	0.153%	0.023%	0.025%	0.069%
Avg-Measured	0.681%	0.508%	0.295%	1.279%	0.743%	1.499%	0.376%	0.428%	0.306%
Avg Program	0.495%			1.174%			0.370%		

Table 4-7. Calculated errors for each of the programs tested based on the difference between the expected natural abundance and the measured MIDs for the unlabeled metabolite mixture. The data analysis using PIRAMID consistently produced a smaller error relative to theory.

	62.5 µM	**125 µM**	**250 µM**
F-test statistic	0.7175	0.3461	1.6972
ANOVA p-value	0.4923	0.7088	0.1923
PIRAMID vs El-MAVEN p-value	0.9163	0.9780	0.9998
PIRAMID vs Skyline p-value	0.4712	0.7047	0.2595

Table 4-8. Results of the ANOVA and Tukey's honestly significant difference tests applied to the calculated errors of all metabolites in the mixture of standards. At all concentrations, the differences in the root mean squared errors are not statistically significant at a confidence level of 95%

In the analysis of metabolite enrichment, PIRAMID consistently yielded lower calculated errors. However, it is important to emphasize that these differences in accuracy, while consistent, did not achieve statistical significance. Furthermore, most of the examined metabolites displayed errors below the acceptable threshold of <1% (Fig. 4-7A). However, three specific metabolites (i.e., fructose, serine, and threonine) stood out for contributing the highest errors (Fig. 4-7B). Across all

three software applications tested, these three metabolites consistently surpassed a 1% error threshold.

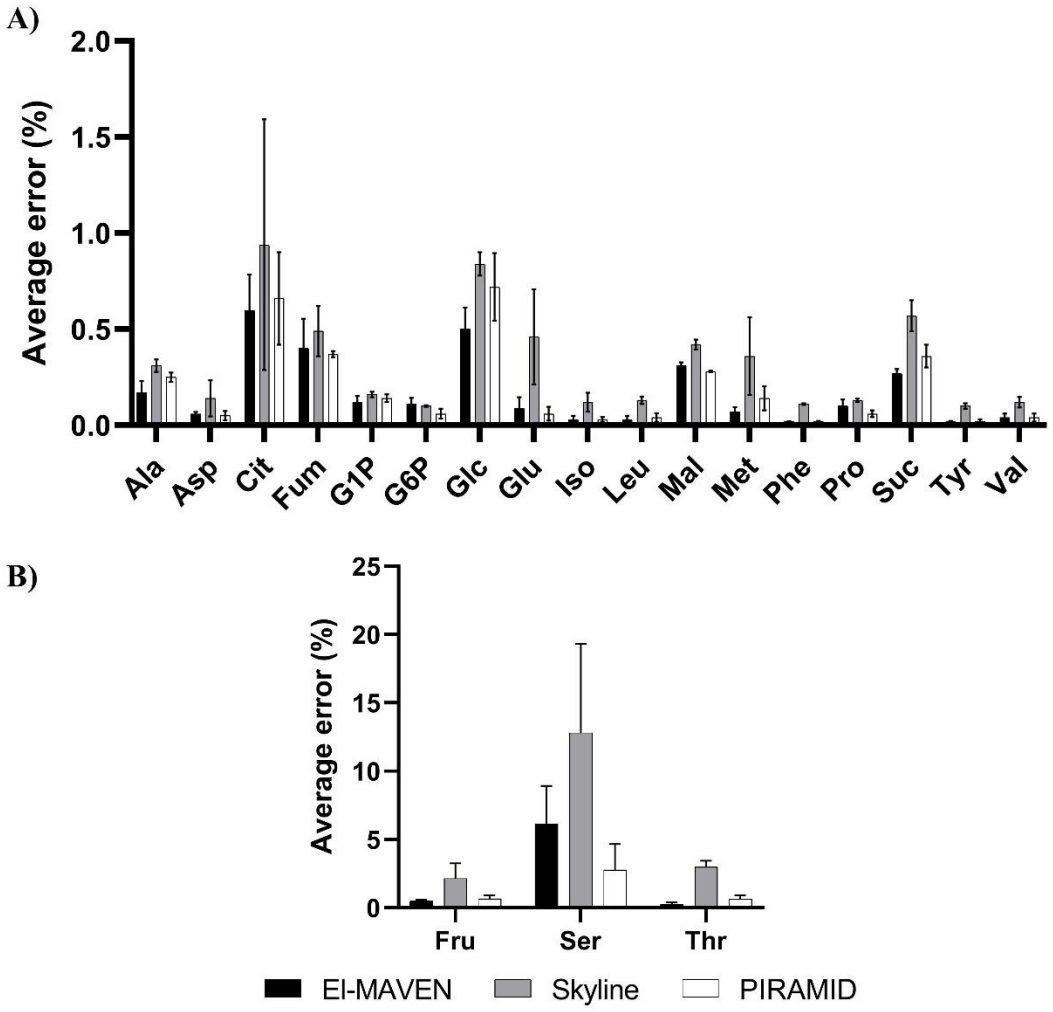


Figure 4-6. Calculated errors of the metabolites in the standard mixture across all concentrations. *A) Metabolites with errors lower than 1%. Heuristically, for metabolomics purposes this error is considered acceptable. B) Metabolites with errors higher than 1%. Serine stands out as the metabolite with the highest error across all platforms. Fructose and threonine show high errors when using Skyline, but acceptable errors using El-MAVEN and PIRAMID.*

The 250 µM concentration standard mixture produced the lowest errors in both El-MAVEN and PIRAMID, which can be attributed to the direct correlation between intensity and peak height. This relationship often results in higher-quality peaks for metabolites present at higher concentrations, leading to a more accurate measurement of MIDs. However, an exception was

observed in Skyline, primarily due to an outlier in the case of serine, which significantly skewed the average error. The nature of this outlier will be explored in the next section.

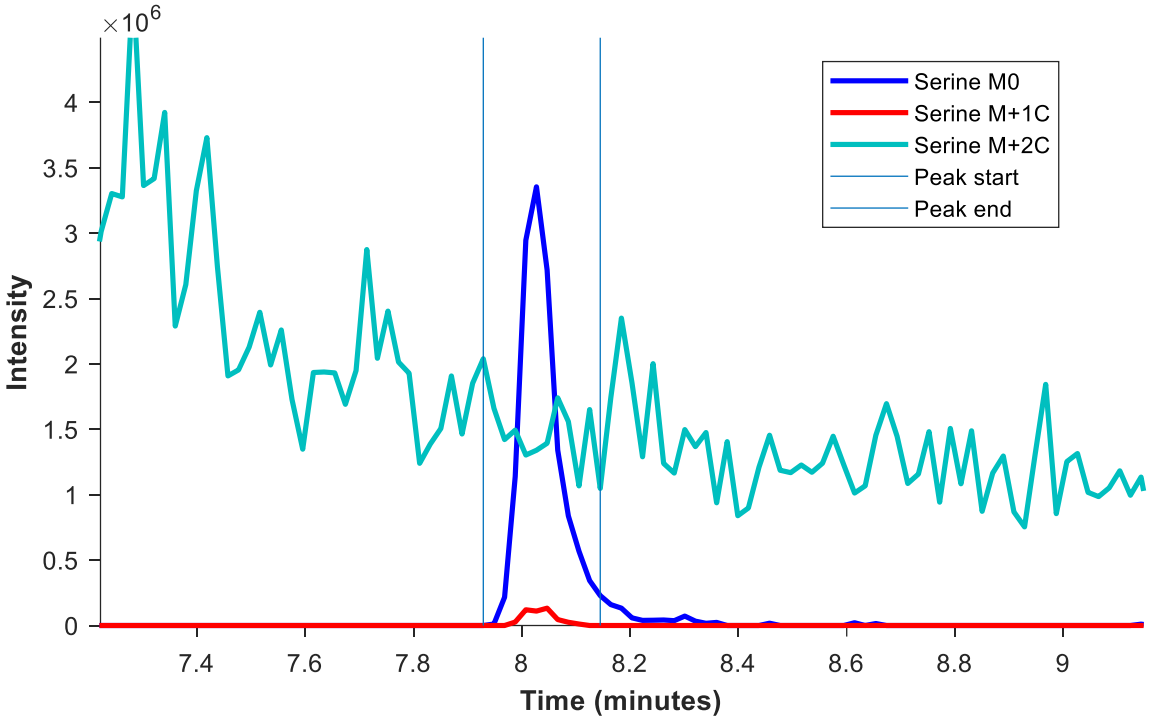


Figure 4-7. **Intensity of detected serine isotopologues.** The M+2C isotopologue shows a high baseline and noise possibly from the tailing of a higher intensity peak that elutes at a similar retention time. PIRAMID outperforms the other tools by detecting these types of defects and adjusting the baseline accordingly.

Upon closer inspection, the baselining algorithm contributed substantially to errors in the metabolites with the lowest accuracies. In the specific case of serine, which is the compound that shows the highest error of all metabolites, the M+2C isotopologue presents an unusually high baseline caused by the tailing of a peak that elutes before it (Fig. 4-8). Skyline and El-MAVEN miscalculate the baseline of the M+2C peak, yielding a higher integrated intensity that skews the MID. In certain cases, these errors can exhibit values reaching as high as 10% (Table 4-5). PIRAMID stands out as the only tool that allows the user to manually modify the value of the baseline in cases where the algorithm yields an erroneous value.

Table 4-9 provides the calculated errors for the glutamine standard mixture. Overall, PIRAMID provided errors in the same order of magnitude as the other two tools. Once again, the differences were not large enough to be statistically significant, as shown in Table 4-10.

	Maven					Skyline					PIRAMID				
Isotopologue	**Unlabeled**	**[1-^{13}C]**	**[1,2-$^{13}C_2$]**	**[$^{15}N_2$]**	**[U-$^{13}C_5$]**	**Unlabeled**	**[1-^{13}C]**	**[1,2-$^{13}C_2$]**	**[$^{15}N_2$]**	**[U-$^{13}C_5$]**	**Unlabeled**	**[1-^{13}C]**	**[1,2-$^{13}C_2$]**	**[$^{15}N_2$]**	**[U-$^{13}C_5$]**
C0N0	0.14%	0.44%	0.06%	0.04%	0.07%	0.88%	0.36%	0.06%	0.06%	0.07%	0.06%	0.42%	0.06%	0.03%	0.06%
C1N0	0.15%	0.44%	0.50%	0.03%	0.01%	0.20%	2.19%	0.56%	0.02%	0.01%	0.13%	0.26%	0.52%	0.02%	0.00%
C2N0	0.04%	0.13%	0.03%	0.06%	0.07%	0.00%	0.11%	0.27%	0.09%	0.09%	0.01%	0.04%	0.07%	0.00%	0.00%
C3N0	0.00%	0.03%	0.18%	0.00%	0.10%	0.00%	0.03%	0.05%	0.00%	0.11%	0.00%	0.03%	0.11%	0.00%	0.11%
C4N0	0.00%	0.01%	0.01%	0.00%	0.67%	0.00%	0.00%	0.01%	0.00%	1.08%	0.00%	0.01%	0.02%	0.00%	0.79%
C5N0	0.01%	0.58%	0.00%	0.00%	0.72%	0.00%	0.55%	0.00%	0.00%	0.63%	0.00%	0.57%	0.00%	0.00%	0.92%
C0N1	0.06%	0.17%	0.02%	1.33%	0.02%	0.06%	0.03%	0.02%	1.42%	0.01%	0.08%	0.06%	0.00%	1.39%	0.03%
C1N1	0.01%	0.00%	0.24%	0.10%	0.01%	0.02%	0.04%	0.00%	0.10%	0.00%	0.01%	0.00%	0.12%	0.10%	0.00%
C2N1	0.00%	0.01%	0.02%	0.00%	0.00%	0.00%	0.00%	0.07%	0.00%	0.00%	0.00%	0.01%	0.06%	0.00%	0.00%
C3N1	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.01%	0.01%	0.02%	0.00%	0.02%	0.02%	0.00%	0.00%
C4N1	0.00%	0.00%	0.00%	0.00%	0.36%	0.00%	0.00%	0.00%	0.00%	0.04%	0.00%	0.00%	0.00%	0.00%	0.23%
C5N1	0.00%	0.00%	0.00%	0.00%	0.70%	0.71%	1.82%	0.16%	0.00%	0.26%	0.00%	0.00%	0.18%	0.31%	0.70%
C0N2	0.01%	0.01%	0.02%	1.02%	0.02%	0.01%	0.00%	0.00%	1.67%	0.03%	0.00%	0.00%	0.01%	0.75%	0.08%
C1N2	0.00%	0.00%	0.00%	0.39%	0.00%	0.00%	0.00%	0.00%	0.36%	0.00%	0.00%	0.00%	0.00%	0.26%	0.00%
C2N2	0.00%	0.00%	0.00%	0.11%	0.00%	0.02%	0.00%	0.01%	0.07%	0.01%	0.00%	0.00%	0.00%	0.06%	0.00%
C3N2	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.01%	0.02%	0.02%	0.26%	0.00%	0.00%	0.00%	0.02%	0.04%
C4N2	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%
C5N2	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.02%	0.01%	0.08%	0.06%	0.00%	0.00%	0.00%	0.18%	0.00%
Avg-Measured	0.02%	0.10%	0.06%	0.17%	0.15%	0.11%	0.29%	0.07%	0.22%	0.15%	0.02%	0.08%	0.07%	0.17%	0.17%
Avg Program	0.10%					0.17%					0.10%				

Table 4-9. Calculated errors for each tested program based on the difference between the expected versus the measured isotopologue distribution for the glutamine standard mixture. The data analysis using PIRAMID consistently produced a smaller error relative to theory.

	Unlabeled	**[1-^{13}C]**	**[1,2-$^{13}C_2$]**	**[$^{15}N_2$]**	**[U-$^{13}C_5$]**
F-test statistic	1.9737	1.4701	0.0181	0.0691	0.0147
ANOVA p-value	0.1494	0.2394	0.9821	0.9332	0.9853
PIRAMID vs El-MAVEN p-value	0.992	0.9847	0.9957	0.9999	0.9916
PIRAMID vs Skyline p-value	0.1872	0.2715	0.9944	0.9472	0.9851

Table 4-10. Results of the ANOVA and Tukey's honestly significant difference tests applied to the isotopologues of all labeled glutamine standards. For all standards tested, the differences in the root mean squared errors are not statistically significant at a confidence level of 95%

The [$^{15}N_2$]glutamine standard showed the highest errors across all standards, specifically in the C0N1 isotopologue. This occurrence could be explained by impurities in the tracer stemming from glutamine being partially labeled with nitrogen in higher proportions than what was reported by the manufacturer. This phenomenon highlights the importance of accounting for the tracer impurity in the calculations of natural abundance and correction of MIDs.

A comparison via a linearity analysis of the performance of PIRAMID versus El-MAVEN and Skyline is presented in Figures 4-9 and 4-10, respectively. This analysis provides a visual representation of the similarities between the errors of PIRAMID and the other selected tools. The isotopologues above the 1:1 line represent values for which PIRAMID yielded comparatively higher errors, while isotopologues below the 1:1 line represent isotopologues where PIRAMID provided lower errors. The results show a high correlation for nearly all isotopologues. The measurement of the M+5C isotopologue in the [1-^{13}C]glutamine standard shows a deviation between PIRAMID and the other tools. In this case, PIRAMID matches the expected value precisely while the other tools overestimate the expected value. However, once again the errors fall below 1% and are within the expected range of accuracy for isotopologue measurements.

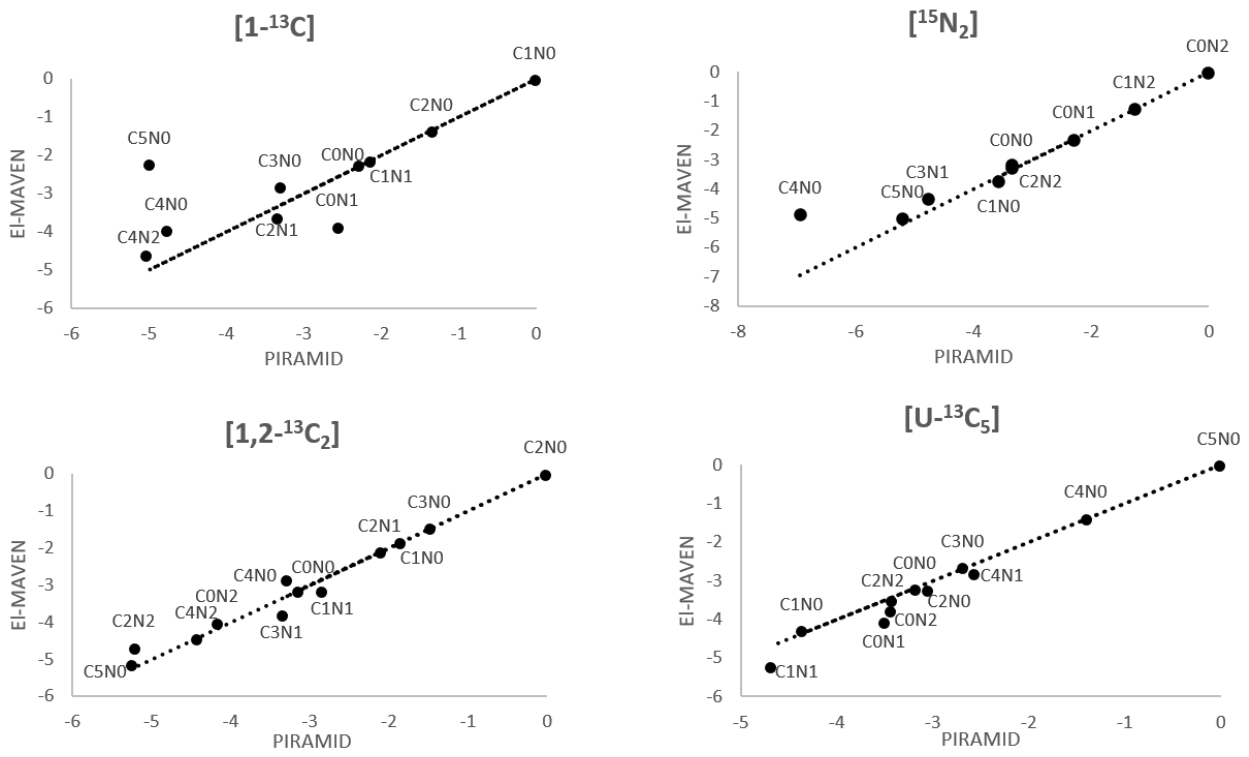


Figure 4-8. Linearity analysis of the labeled standards analyzed using PIRAMID versus EI-MAVEN. *The axes labels indicate the relative abundance of each isotopologue on a log_{10} scale. Isotopologues that were not detected by any of the tools and showed a zero value in their intensity are not presented. The dotted line represents 1:1 agreement between the programs.*

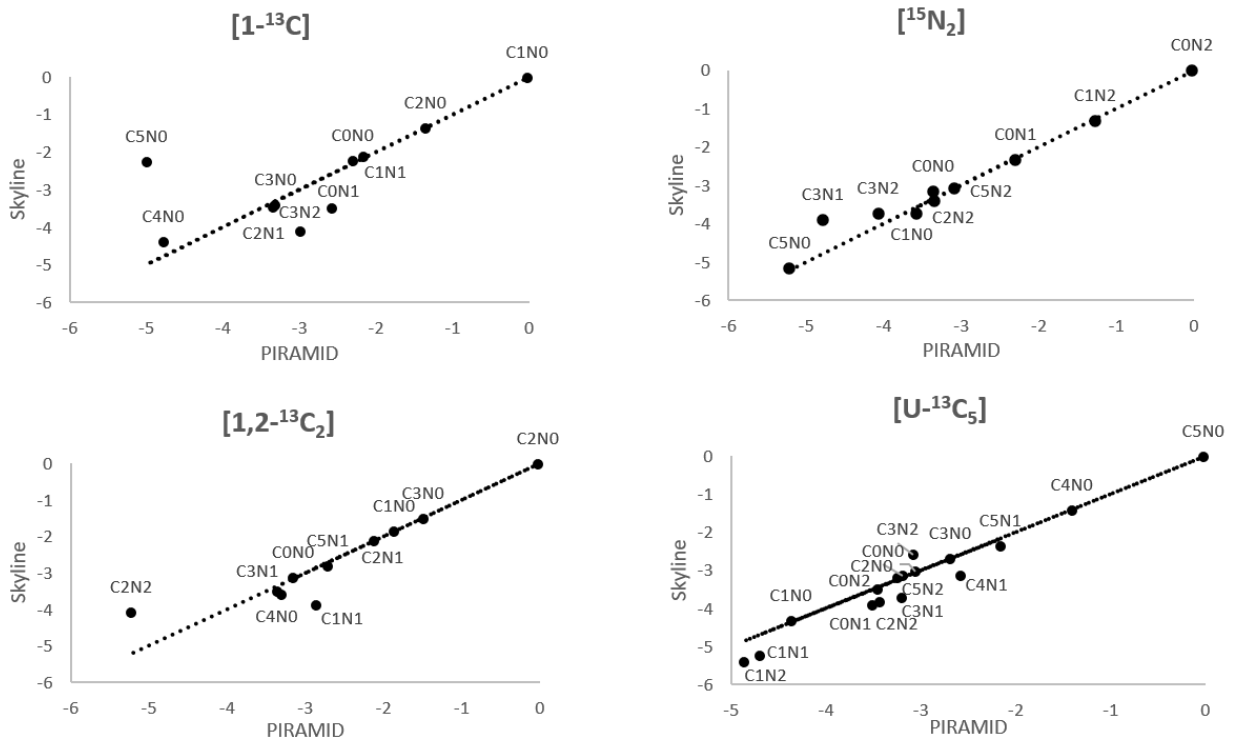


Figure 4-9. Linearity analysis of the labeled standards analyzed using PIRAMID versus Skyline. *The axes labels indicate the relative abundance of each isotopologue on a log_{10} scale. Isotopologues that were not detected by any of the tools and showed a zero value in their intensity are not presented. The dotted line represents 1:1 agreement between the programs.*

In general, it cannot be concluded from the limited examples tested that PIRAMID is any more or less accurate than the other tools. A major advantage of PIRAMID is the number of additional features that are useful in data analysis and that account for the most common artifacts and challenges in the process of analyzing MS data from isotope-based metabolomics experiments. A table comparing these features against other popular free tools that are typically used for metabolomics data processing is presented in Table 4-11. The tools that were considered for the comparison are iMS2Flux [209], eMZed3 [224], TarMet [210], DExSI [211], AssayR [212], mzMatch-ISO [159], El-MAVEN [208], and Skyline [132].

Features	iMS2Flux	eMZed3	TarMet	DExSI	AssayR	mzMatch-ISO	El-MAVEN	Skyline	PIRAMID
Load .cdf files	-	-	✓	✓	-	-	-	-	✓
Load .mzml/.mzxml files	-	✓	✓	-	✓	-	✓	✓	✓
Warp retention times	-	✓	✓	-	-	-	-	-	✓
Supports ^{2}H	✓	-	✓	✓	✓	-	✓	✓	✓
Supports ^{13}C	✓	-	✓	✓	✓	✓	✓	✓	✓
Supports ^{15}N	✓	-	✓	✓	✓	-	✓	✓	✓
Supports ^{18}O	✓	-	✓	-	-	-	-	-	✓
Supports ^{34}S	✓	-	-	-	-	-	✓		✓
Supports MS2	-	✓	✓	-	-	✓	✓	✓	✓
Supports HRMS	✓	✓	✓	-	✓	✓	✓	✓	✓
Supports dual-label experiments	-	✓	-	-	-	-	-	✓	✓
Consistent integration bounds for isotopic clusters	-	-	-	-	-	-	-	-	✓
MID calculation	✓	✓	✓	✓	✓	-	-	-	✓
Percent enrichment calculation	-	-	-	✓	-	-	-	-	✓
Natural abundance correction	✓	-	-	✓	-	-	-	-	✓
File/group comparisons	-	✓	-	✓	-	✓	✓	✓	✓
Visualize MIDs in real time before exporting	-	✓	✓	✓	✓	✓	-	-	✓
GUI-based	-	-	✓	✓	-	-	✓	✓	✓

Table 4-11. Feature comparison across multiple freely available programs used to analyze metabolomics experiments involving stable isotopes. PIRAMID offers a wide range of capabilities focused on simplifying the data analysis pipeline. The use of fixed integration bounds for all the isotopologues within an isotopic cluster improves the accuracy of MID quantification. The ability to correct for natural isotope abundance in experiments involving dual-labeled tracers or MS^2 acquisition provides added versatility in processing stable isotope data.

PIRAMID distinguishes itself as a versatile software tool with comprehensive support for various mass spectrometry data types. In addition to Tarmet, it stands out as the sole software capable of handling data originating from both GC-MS and LC-MS. This flexibility is reflected in its ability to process the respective vendor-agnostic formats .cdf and .mzml/.mzxml. Furthermore, PIRAMID supports data collected in both high-resolution MS (HRMS) and tandem MS (MS^2) modes. PIRAMID offers support for a wide range of commonly used tracer atoms, a feature shared only with iMS2Flux. Notably, PIRAMID surpasses iMS2Flux by providing compatibility with dual-labeled experiments. Regarding post-analysis capabilities, PIRAMID provides valuable features

like correcting MIDs for natural isotope abundance, calculating atom percent enrichment, and facilitating simple statistical comparisons across experimental groups.

Arguably, one of PIRAMID's most significant attributes, setting it apart from other software tools, is its capacity to integrate peaks consistently within a defined retention time window across all isotopologues of a compound. This approach has been proven to significantly enhance the precision of results, making PIRAMID a promising tool for isotopic analysis in a broad spectrum of research applications.

4.5 Discussion

4.5.1 PIRAMID in the context of stable isotope-based metabolomics

The field of stable isotope-based metabolomics has witnessed significant advancements in recent years, with the development of various analytical techniques aimed at enhancing the accuracy and efficiency of data analysis. However, the inclusion of stable isotopes in metabolomics experiments is hindered by the lack of publicly available software tools that implement these techniques [225].

Broadly, metabolomics studies start with an untargeted analysis aimed at discovering biomarkers and metabolites of interest, which is then followed by a targeted analysis that quantifies selected metabolites. When isotope tracers are administered, the MS measurements are often analyzed using kinetic flux profiling or metabolic flux analysis to assess metabolite turnover or flux through specific pathways [13], [14], [226]. PIRAMID’s purpose is to perform the targeted extraction and quantification of metabolite labeling that results from such experiments.

Data analysis is a significant bottleneck in targeted metabolomics experiments, especially in cases where large datasets are produced in high throughput [227]. In some cases, the available programs can only process one file at a time, requiring the user to manually integrate each file [228] . The

need for software capable of consecutively processing several files with no user input often arises in large studies with many samples. PIRAMID offers an efficient software solution to automate the processing of such large-scale datasets, which is optimized for accurate quantification of stable isotope enrichment.

4.5.2 Strengths of PIRAMID

The tool developed in this chapter, PIRAMID, exhibits notable strengths that make it an asset in the field of stable isotope-based metabolomics. These strengths contribute to its versatility and utility in handling complex mass spectrometry data, as well as enhancing the accuracy of MID calculations. One of the significant strengths of PIRAMID is its ability to accommodate the most commonly used types of mass spectrometry instruments in the field, namely GC-MS and LC-MS, without the need for a pre-processing or peak-extraction step. This feature is important for researchers who combine the use of both types of mass spectrometry for the detection of different metabolites (i.e., GC-MS typically yields more consistent and accurate results but not all metabolites can be derivatized and analyzed in the gas phase). The algorithms of PIRAMID have been optimized to handle challenging data scenarios in the extracted chromatograms, such as noisy chromatograms, elevated baselines, asymmetric peak shapes, and tailing behaviors. This robustness ensures that the software can effectively extract and analyze isotopic data from various types of biological samples, where such complexities are often encountered due to the simultaneous extraction of several metabolites and the co-elution of interfering compounds.

A unique feature of PIRAMID is its use of a consistent retention time window throughout the integration of all isotopologues of a compound. This uniform approach significantly enhances the accuracy of MID calculations. By maintaining a consistent window, PIRAMID minimizes the potential for errors caused by variations in integration bounds, making it a reliable tool for precise

MID determination. PIRAMID offers the advantage of correcting for the natural abundance of isotopes without the need for additional tools or external resources. This built-in capability streamlines the data processing pipeline, reducing the complexity of isotopic analysis. Furthermore, it stands as the only software capable of correcting several types of MS, including low resolution-MS, MS^2, and HRMS with single- or dual- labeled tracers. Finally, PIRAMID simplifies the data analysis process by providing users with direct calculations of the most important experimental outputs, including MIDs and APEs. Moreover, it organizes the data in a manner that eases further downstream analysis.

4.5.3 Limitations of PIRAMID

It is essential to acknowledge the limitations of PIRAMID, which may impact its applicability in certain experimental scenarios. These limitations should be taken into account when considering the use of the software in metabolomics research. One of the limitations of PIRAMID is its dependence on specific data formats. The software does not accept data directly from vendor files, requiring the conversion of data into .cdf or .mzml/.mzxml format for processing. In the case of the latter, it's important to note that z-lib compression is not supported, potentially resulting in larger dataset sizes. In the specific context of dual-labeled data analyzed with HRMS, PIRAMID's natural abundance correction algorithm has certain limitations. This correction can only be performed if one of the labeled atoms is carbon, though we have limited experience of studies involving dual labeling where ^{13}C is not included. Furthermore, the correction is applied exclusively to the labeled atoms. This constraint may pose challenges in cases where the analyzed metabolites have multiple non-tracer atoms, and their natural abundance interferes with measurements of the isotopologue distribution of the labeled atoms. Finally, PIRAMID's dependency on MATLAB for its execution may limit its accessibility to a wider user base.

MATLAB is proprietary, commercial software, which means that users need to have an appropriate license to run PIRAMID. This reliance on a commercial software package may restrict the reach of PIRAMID, particularly for researchers who do not have access to MATLAB due to cost constraints or licensing limitations.

4.5.4 Comparison of PIRAMID to other software

Despite its limitations that provide opportunities for further development, PIRAMID currently stands out as a unique tool that offers a wider variety of features than other tools that have been developed for the purpose of extracting and quantifying MIDs. While a statistical analysis did not show significant differences between the accuracy of PIRAMID compared to two other widely used tools, El-MAVEN and Skyline, PIRAMID implements novel features that provide specific advantages over other software tools. As mentioned previously, PIRAMID can process data from several different MS acquisition modes, with the extent of data formats only matched by one (Tarmet) out of the eight other tools that were evaluated. Such versatility ensures that PIRAMID can effectively accommodate diverse experimental setups, making it a versatile choice for researchers working across different MS platforms. Another distinguishing feature of PIRAMID is its capability to perform time warping, effectively correcting for retention time drifts that can occasionally occur during chromatographic separation. This feature, coupled with the use of consistent integration windows across all isotopologues, significantly improves the accuracy of MID extraction and integration. The combination of these features was unmatched by the tools used in the comparison. Furthermore, only two of the evaluated tools offered the feature of natural abundance correction, while the rest of the tools require the use of separate tools for natural abundance correction. This reliance on external tools can introduce complexity and additional preprocessing steps into the workflow, potentially hindering efficiency. PIRAMID's intrinsic

natural abundance correction capability eliminates the need for such additional steps, streamlining the process and simplifying data analysis. Finally, a notable benefit of PIRAMID is its user-friendly GUI. This interface is particularly advantageous for researchers who may not be proficient in coding or scripting. The GUI provides an intuitive and accessible means for users to interact with the software, making it more inclusive and user-friendly for a broader research audience.

4.5.5 Future work

While PIRAMID positions itself in the metabolomics field as a strong and competitive tool, there are several areas where future research and development efforts are warranted to further enhance the software's capabilities and address certain limitations. One of the challenges encountered during the development of PIRAMID is the absence of a universal, one-size-fits-all solution for the baseline estimation algorithm. Although an algorithm was selected to minimize errors in challenging cases, the quest for an optimal baseline algorithm remains a topic for future research. While PIRAMID supports a wide variety of chromatography-coupled MS datasets, it currently does not extend its capabilities to non-chromatographic MS data types. Notably, mass spectrometry imaging techniques, such as Matrix-Assisted Laser Desorption/Ionization (MALDI), employ a different data format where the retention time dimension is substituted by two spatial dimensions. In the future, it would be invaluable to broaden PIRAMID's compatibility to encompass these non-chromatographic experiments. This expansion would render PIRAMID even more versatile, accommodating an even wider spectrum of research methodologies.

As previously mentioned, the natural abundance correction method in PIRAMID has certain limitations in the context of dual-isotope tracer data. To address these limitations, future work will focus on the implementation of the non-tracer atom matrix in the calculation. Specifically, adjustments to the software's GUI are needed to ensure that all necessary additional information is

gathered from the user. Finally, there is an opportunity to enhance the interoperability of PIRAMID with other software tools used in the field of metabolomics. The ability to directly import the results of other software, such as SUNDILE, to create lists of metabolites for integration within PIRAMID could significantly enhance the capabilities of both tools. Additionally, PIRAMID's results could be imported into other downstream analysis software like INCA [16], reducing analysis time and offering a more streamlined workflow for researchers.

4.6 Conclusions

PIRAMID is a user-friendly tool that automates the extraction, integration, and analysis of MS data sets obtained from stable isotope labeling experiments. PIRAMID can be used for many purposes including: i) automatic peak finding and integration of MIDs of multiple target metabolites from a batch of MS data files; ii) quantification and normalization of integrated ion counts of the target metabolites; iii) generation of statistical comparisons and interactive plots to identify significant differences between two or more experimental conditions or sample time points; and iv) output the data into a .xls format that can be readily imported by spreadsheet programs or other tools.

PIRAMID is at least as accurate as other bioinformatics tools that are used to integrate target metabolites in MS experiments, and its underlying algorithms are optimized to process data in the context of stable isotope labeling, correcting for the abundance of naturally occurring isotopes and calculating the enrichment and the mass isotopomer distribution of the metabolites. It also enables the analysis of advanced MS acquisition modes such as MS^2 and HRMS, facilitating the quantification of isotope enrichment in metabolites that are labeled with more than one tracer atom (e.g., ^{15}N and ^{13}C). PIRAMID is expected to accelerate the analysis of high-throughput omics experiments, thus eliminating one of the major bottlenecks in the field.

4.7 Acknowledgments

This work was funded by the National Institutes of Health Grant No. U01 CA235508.

CHAPTER 5

ELUCIDATING SOYBEAN METABOLISM USING STABLE ISOTOPE-BASED METABOLOMICS

Abstract

Soybean stands as a pivotal contributor to the US economy mainly due to its high contents of protein and oil. The composition of soybean seeds plays a decisive role in determining overall crop value and utility, motivating research on the metabolic optimization of both protein and oil yields. While previous studies have examined the factors influencing seed development and composition, certain biochemical pathways, particularly those occurring outside the primary metabolic routes, remain unexplored. Using substrates labeled with stable isotopes and an untargeted metabolomics approach, we show that entry of glucose into the tricarboxylic acid (TCA) cycle is constrained during seed development, and glucose carbon is instead used to fuel the synthesis of essential amino acids and complex molecules such as anthraquinones and isoflavonoids. In contrast, glutamine emerges as a primary contributor to the TCA cycle, serving as a key precursor for pyruvate production from malate via the malic enzyme, which, in turn, serves as a precursor for fatty acid biosynthesis. Furthermore, our investigation reveals a proportional relationship between malic enzyme expression and the final seed oil content, unveiling a potential regulatory mechanism that controls lipid biosynthesis. These findings collectively enhance our understanding of soybean metabolism, providing valuable insights into optimizing seed composition for enhanced economic and agricultural benefits.

5.1 Introduction

Soybean (*Glycine max*) is considered one of the most important oilseed crops worldwide [229]. It is of particular importance in the economy of the United States, being the country with the most soybean exports around the globe [230]. The market value of this seed crop depends on its protein and oil contents [231]. However, seed composition is a highly variable trait: protein content is typically between 35% and 40% [232], whereas oil content can vary from 6.5% to 28.7% [233]. Several studies have focused on genetic engineering to produce optimized strains with high content of both oil and protein [234], [235], [236]. However, the optimization efforts to increase oil content have typically come at the expense of lowering protein content, and vice versa [237], [238], [239].

Primary pathways in central carbon metabolism are responsible for the biosynthesis and accumulation of lipids, proteins, and carbohydrates. The biochemistry of these pathways is common among different plant species and soybean strains, but the metabolic activity (i.e., 'flux') through these pathways depends on the genetic and environmental context of each plant [240], [241]. In seeds, nutrients are supplied to the developing embryo from the seed coat, which contains sugars and amino acids derived from the maternal plant [242], [243], [244]. Therefore, the final seed composition is a consequence of the availability of these nutrients and their metabolic fate within the embryo [245], [246]. Thus, understanding the differences in activity between protein- and oil-yielding metabolic pathways is crucial for designing strategies to engineer seeds and tailor their compositions.

Metabolic flux analysis (MFA) based on isotope labeling experiments (ILEs) is the preferred approach to elucidate pathway activities and determine the flow of nutrients through specific metabolic routes [247], [248]. However, isotope labeling measurements used for MFA are often

limited to a targeted list of compounds within core metabolic pathways, which leaves many other pathways unexplored.

Prior MFA studies of oilseed crops have reported that ~10% of hexose flux entering rapeseed embryos is directed into the oxidative pentose phosphate (OPP) pathway, which provides up to 22% of the reductant (i.e., NADPH) required for fatty acid biosynthesis [249]. This model, however, ignored protein-producing pathways and left the synthesized amino acids as a sink of carbon. Another study found that 40% of mitochondrial pyruvate is produced *in situ* by malic enzyme instead of being imported from the cytosol, and nearly all citrate produced in the mitochondria is exported to provide cytosolic acetyl-CoA for fatty acid elongation rather than entering the oxidative reactions of the TCA cycle [250]. This study was based on a model where PEP stems from 3-phosphoglycerate (3PGA) but ignores other branching pathways such as the gluconeogenesis and glycerate-serine pathways that will also be labeled by 3PGA [251]. In sunflower embryos, MFA revealed that the flux of malate into oil biosynthesis was low (contributing <9% of carbon to fatty acids) and that futile cycling consumed <20% of the total ATP produced [252], suggesting a high efficiency in the use of carbon allocated to produce other unknown compounds that are ignored in the study.

These studies primarily focused on the precursors of fatty acid biosynthesis, providing valuable insights. However, they overlooked additional branching pathways that could offer crucial information about protein biosynthesis. In the context of soybean, incorporating these pathways is essential for a comprehensive understanding of seed metabolism. Their inclusion would facilitate genetic manipulation and optimization strategies to enhance oil and protein content of soybean crops.

In the case of soybean embryos, multiple studies have shed light on the pathways involved in the production (or utilization) of fatty acids and proteins during seed development. One study determined that a substantial amount of carbon is directed through the oxidative pentose phosphate pathway and the gluconeogenic pathway, from triose-phosphate to fructose-6-phosphate, in a gluconeogenetic route, and in the opposite direction, to pyruvate within the plastid. Moreover, the flux through the glyoxylate shunt, an anaplerotic pathway bypassing certain oxidative steps in the TCA cycle and metabolizing acetate units from lipid breakdown during germination, was found to be negligible. This suggests a metabolic state favoring the production and storage of fatty acids [253]. In another study using labeled glutamine as a substrate, approximately 41% of the total carbon from glutamine was transformed into glutamate-derived amino acids in proteins. Additionally, 10% of the carbon in fatty acids was labeled, suggesting a notable production of pyruvate from the TCA cycle, specifically via the conversion of malate to pyruvate. The measured enrichment of malate suggested that approximately 25% of fatty acid carbon was derived from this pool. Comparative analysis with other oilseeds revealed that in soybean embryos, amino acids contributed a higher proportion of carbon to fatty acid biosynthesis [254].

Overall, fluxes connecting the pathways of glucose and glutamine metabolism to the TCA cycle have been calculated in developing oilseed embryos leading to the general consensus that malate provides relatively low amounts of carbon for fatty acid biosynthesis, compared to the carbon contribution that is acquired from glycolytic routes [250], [252], [254]. Nevertheless, previous research has not delved into the roles of glucose and glutamine in pathways beyond core metabolism. This gap underscores the importance of clarifying how carbon is allocated in these secondary pathways.

This chapter seeks to unravel the dynamics of carbon flow beyond the commonly explored core pathways, shedding light on secondary pathways and metabolites. The aim is to expand our understanding of the fatty acid and protein biosynthesis pathways in soybeans. To do so, this chapter focuses on uncovering the metabolic differences between two high-protein soybean cultivars with varying levels of oil content. By tracing the fate of the two major carbon sources feeding the TCA cycle (i.e., glucose and glutamine), an untargeted isotope-based metabolomics study was conducted using the analysis tools described in Chapters 3 and 4. This not only enabled a comprehensive analysis of nutrient metabolism in developing soybean embryos, but also served as a relevant testbed for software validation. Differences in isotope enrichment from glucose and glutamine tracers were identified, revealing unique metabolic signatures of the two cultivars examined. A metabolic fate map was developed by quantifying tracer contributions to metabolites both inside and outside the core pathways of energy and lipid metabolism. Our results serve as a template for improving knowledge of nutrient metabolism in soybean and other crops that can be leveraged to enhance the production of oil and/or modify the composition of seeds through genetic engineering.

5.2 Methods

Note: The experimental measurements described in Sections 5.2.1, 5.2.2, 5.2.3, and 5.2.5 were performed by Dr. Shrikaar Kambhampati, Dr. Stewart Morley, Dr. Doug Allen, and Dr. Bradley Evans of the Donald Danforth Plant Science Center in the context of a scientific collaboration.

5.2.1 Plant growth, tissue collection, and culture of soybean embryos with isotopic labels

Two soybean cultivars, PI603338 [*Glycine max* (L.) Merr., Cha se dou B] and PI587778 [*Glycine max* (L.) Merr., Jing huang 18], were obtained from the USDA-ARS Germplasm Resources

Information Network (GRIN). PI 603338 has a reported protein/oil content of 50.2/20% and was designated as the high oil (H) cultivar. PI 587778 has a reported protein/oil content of 57.4/9% and was designated as the low oil (L) cultivar. Plants were grown under greenhouse conditions, described previously [255], under a 14/10-hour day/night photoperiod. Developing seeds at peak oil and protein filling stage, R5, were collected for further analysis. Seeds were excised from pods collected at the appropriate stage; seed coats were removed; and the cotyledons were used for culturing in sterile 24-well plates with stable isotopes. The culture medium consisted of 300 µL of modified Linsmaier and Skoog medium [160], [161] with Gamborg's vitamins (Sigma) and 5 mM MES buffer adjusted to pH 5.8. Parallel labeling experiments were performed with either 200 mM [U-$^{13}C_6$]glucose (for ^{13}C labeling) or 20 mM [$^{13}C_5$, $^{15}N_2$]glutamine (for $^{13}C/^{15}N$ labeling). Culturing was performed under 30-35 µE continuous light at 26°C, and 4 independent replicates were collected at 30, 60, 120, 240, 480 and 1200 min to enable time-course assessments of isotope labeling, similar to a previous study [162]. Uncultured cotyledons, collected after excision from pods, were used as unlabeled controls.

5.2.2 Metabolite extraction

Metabolite extracts were obtained from seed embryos using a biphasic extraction method as described previously [162], with slight modifications. Briefly, 20-30 mg of ground lyophilized leaf tissue was collected in Eppendorf tubes, and metabolites were extracted using 700 µL of 3:1 (v/v) methanol:chloroform (-20ºC) with 0.3 µM [U-$^{13}C_{12}$]sucrose as an internal standard. Samples were vortexed vigorously and incubated at 4ºC on a rotary shaker for two hours, after which 300 µL of ddH_2O was added. Samples were centrifuged at 14,000 rpm for 10 min to achieve phase separation. The upper aqueous fraction was collected into a new centrifuge tube and dried using a speed vacuum centrifuge (Labconco®, Kansas City, USA). One hundred µL of methanol was added to

the remaining organic fraction, and the mixture was centrifuged further at 14,000 rpm for 5 min to pellet the debris. The supernatant was transferred into a fresh centrifuge tube and dried similarly to the aqueous fraction. Dried samples were stored in -80ºC until further use.

5.2.3 Metabolomics data acquisition

Two different chromatographic methods were used for untargeted metabolomics, a hydrophilic interaction chromatography (HILIC) method and a reverse-phase (C18) method, to obtain wide coverage of different compound classes. Aqueous-phase extracts, re-suspended in 1:1 methanol: H_2O, were used for HILIC and C18 chromatography. A Dionex UltiMate 3000 ultra-high performance liquid chromatography (UHPLC) system interfaced to a benchtop Q-Exactive Orbitrap mass spectrometer was used for all untargeted liquid chromatography-mass spectrometry (LC-MS) analysis. HILIC separation was achieved using an Agilent zic-pHILIC (100 x 2.1 x 3 µm) column with mobile phases 10 mM ammonium bicarbonate in ddH2O (solvent A) and 10 mM ammonium bicarbonate in 95:5 (v/v) acetonitrile:ddH_2O (solvent B) and a flow rate of 250 µL/min. The following gradient was used for HILIC: 2 min hold at 100% B, 3 min linear gradient from 100% to 85% B, 16 min linear gradient from 85% to 50% B, 17 min linear gradient from 50% to 30% B, 18 min hold at 30% B, 20 min linear ramp back to 100% B and equilibration for up to 30 min. Reverse-phase chromatography was performed using an Acquity HSS T3 (100 x 2.1 x 1.8 µm) column with mobile phases 0.1% formic acid in ddH_2O (solvent A) and 0.1% formic acid in acetonitrile (solvent B) and a flow rate of 250 µL/min. The gradient conditions used for the C18 method are as follows: 3 min hold at 2% B, 13 min linear gradient from 0% to 100% B, 16 min hold at 100% B, 19 min linear gradient back to 2% from 100% B and equilibration up to 30 min.

Data for untargeted metabolomics using both chromatographic methods were acquired over a mass range of 70-1000 *m/z* by full-scan MS at 70,000 resolution in both positive and negative ionization

modes. The automatic gain control (AGC) and maximum injection time (IT) were 5 x 10^5 and 100 ms respectively. The heated electrospray ionization (HESI) source was operated with sheath gas, 15 arbitrary units; auxiliary gas, 5 arbitrary units; capillary temperature, 250ºC; auxiliary gas heater temperature, 50ºC; and S-lens RF level, 50. The spray voltage was 4.2 and 3.9 kV in positive and negative modes, respectively. One unlabeled control sample from each group was used for a 'top 12' data-dependent acquisition experiment in both ionization modes to generate MS^2 datasets for compound identification. These experiments involved full-scan MS at 70,000 resolution, AGC target of 5 x 10^5, and maximum IT of 100 ms interlaced with MS^2 scans at 17,500 resolution, AGC target of 5 x 10^4, maximum IT of 50, 2.0 *m/z* isolation window, stepped collision energy of 15, 25 and 35 eV, intensity threshold 1 x 10^4, and 15 sec dynamic exclusion.

5.2.4 Metabolomics data processing and analysis

Raw data files in Thermo .RAW format obtained in the profile mode were first centroided by conversion into .mzml format using ProteoWizard [71] with peak picking filter applied. Mass spectral features were detected, and a pre-processed data table was created using the XCMS program [90]. XCMS parameters for data pre-processing were optimized using IPO [169]. Raw data files (.RAW format) as well as the pre-processed data (.csv format) are publicly available at the National Metabolomics Data Repository (NMDR) [181]. The extracted list of features was analyzed with SUNDILE (see Chapter 3) using a resolution of 20 ppm and a retention time window of 0.1 minutes for feature binning, only scanning for the possible ^{13}C isotopologues. Metabolites were identified using MS^2 data analyzed using MS-DIAL [125] with default parameters, matching the ion fragmentation patterns against the library "ESI(-)-MS/MS from standards+bio+in silico", which is available on the software website. In order to rigorously quantify isotope enrichments of selected compounds, targeted integrations were performed for some of the identified metabolites

using PIRAMID (see Chapter 4). Isotopologues were binned using a mass tolerance of 10 ppm; the peaks were smoothed using a second-order Savitzky-Golay filter with a frame length of 11 data points; and the isotopic purity of the labeled atoms was assumed to be 99%.

5.2.5 Malic enzyme measurements

$NADP^+$-dependent malic enzyme activity was determined using a previously described method [256] with some modifications. Briefly, ground homogeneous lyophilized seed tissue was resuspended by bead milling (Retsch MM 400 Mixer Mill, Retsch, Haan, Germany) in extraction buffer consisting of 100 mM Tris-HCl pH 8, 0.5M urea, 5 mM $MgCl_2$, 2 mM EDTA, 10% v/v glycerol, 10 mM β-mercaptoethanol, and 2% polyvinylpolypyrrolidone (w/v). Urea was added to aid solubilization and clarification of dense protein extracts from soy seed tissues. Malic enzyme standard (Cat. M5257-25UN, Sigma-Aldrich, St. Louis, USA) was analyzed either with or without 0.5 M urea to determine any positive or negative effects from the additional buffer component. No noticeable effect was observed on malic enzyme standard activity due to urea. Approximately 10 mg of lyophilized tissue was resuspended in 0.5 mL of extraction buffer and clarified by centrifugation at 16,000g for 15 minutes. Clarified extract was further diluted until total protein concentration was ≤2 mg/mL as determined by a Bradford protein assay. Crude extract was assessed for enzyme activity immediately using a 96-well microplate reader and recording the change over time in absorbance at 340 nm corresponding to NAD(P)H production. To prepare the assay, crude extract was added to an assay buffer consisting of 50 mM Tris pH 8, 10 mM $MgCl_2$, and 0.5 mM $NADP^+$ or NAD^+. Malate solution buffered to pH 8 was added to a concentration of 10 mM to begin the reaction. Simultaneous background readings were collected by substituting H_2O for malate. Assays were performed at 25°C in a total volume of 100 µL per well. Total units, U, were determined according to Beer's law with inputs of the 340 nm absorbance change per

minute, the reaction volume (converted to liters), path length, b, and millimolar extinction coefficient, ε = 6.22 according to the following equations (including a- multiplication factor of 1000 for Units of enzyme):

$$U = \frac{\Delta A340_{1min} * V * 1000}{\varepsilon b} \quad \text{(Eq. 5-1a)}$$

$$b = \frac{Sample(OD_{1000} - OD_{900})}{Cuvette(OD_{1000} - OD_{900})} \quad \text{(Eq. 5-1b)}$$

A Bradford assay was used to quantify total soluble protein in crude extracts to determine U/mg protein.

5.3 Results

5.3.1 Untargeted pathway analysis of isotope enrichment in soybean

Two cultivars with different oil yield levels (i.e., high (H) and low (L)) were labeled with [U-$^{13}C_6$]glucose (Glc) or [$^{13}C_5$, $^{15}N_2$]glutamine (Gln), resulting in four possible experimental conditions that were measured over 6 timepoints. The data acquired in negative ionization mode was processed using SUNDILE to identify ^{13}C-labeled metabolites and to uncover differences in isotope labeling dynamics between the experimental conditions. Though information on the ^{15}N-labeled isotopologues was available, the study exclusively focused on the ^{13}C labeling. Several pathways that exhibited differences between the four datasets are presented in Fig. 5-1. The entire list of pathways and their respective enrichment scores can be found in Table 5-A1.

Some core metabolic pathways were found to have significant differences in labeling across the four datasets: TCA cycle, amino acid biosynthesis, and pentose phosphate pathway were differentially enriched. Labeling from [U-$^{13}C_6$]glucose was diverted away from the TCA cycle and into biosynthetic pathways leading to nucleotides and phosphoenolpyruvate (PEP)-derived

metabolites such as aromatic amino acids and betalains. On the other hand, [$^{13}C_5$,$^{15}N_2$]glutamine provided substantial enrichment of TCA cycle intermediates, pyruvate-derived amino acids, and pathways that involve acetyl-CoA, such as fatty acid biosynthesis, propanoate metabolism, lysine degradation, and lysine biosynthesis.

The enrichment scores of these pathways were consistent to previous research on glucose and glutamine metabolism in soybean. In one study, differences in the flux maps were analyzed based on the C:N ratios of the supplied substrates [245]. The study revealed that under low C:N conditions, the pentose phosphate pathway exhibited higher carbon allocation from sugars than from glutamine, and amino acids in the aspartate and glutamate families showed the opposite trend [245]. Similarly, our study identified high scores in the "D-amino acid metabolism," "alanine, aspartate, and glutamate metabolism," and "arginine biosynthesis" pathways originating from glutamine but not from glucose.

Other studies determined the metabolic fluxes involved in the production of different amino acids when fully labeled glutamine was employed as a tracer. In these investigations, PEP-derived amino acids, such as phenylalanine, tyrosine, and tryptophan, showed negligible enrichments when glutamine was used as a tracer [254], [256], consistent with our findings. One of these studies also reported a higher carbon allocation to amino acids and fatty acids from pyruvate, rather than the conversion of pyruvate into citrate, enriching the TCA cycle [254]. Our results similarly indicate low enrichment in the TCA cycle and high enrichments in the “Phenylalanine, tyrosine, and tryptophan biosynthesis” pathways when glucose is used as a tracer, suggesting the above-mentioned diversion of carbon.

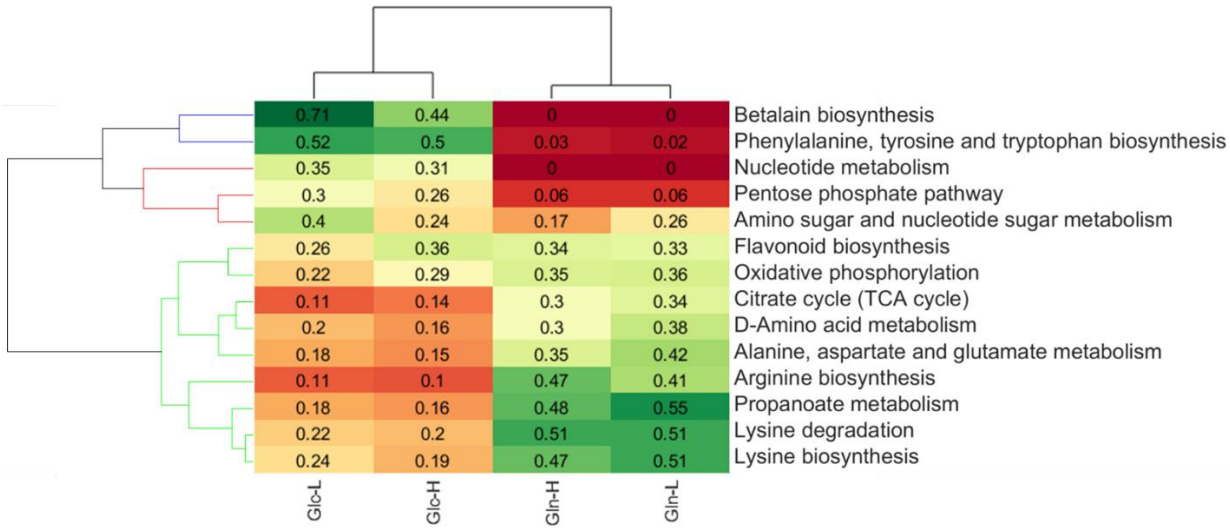


Figure 5-1. Hierarchical clustering of the pathway enrichment scores comparing the different experimental groups. *The score serves as an indicator of the average isotope enrichment of pathway metabolites. The TCA cycle was more enriched using glutamine as a tracer than when using glucose. Analogously, pathways branching from glycolysis (e.g., biosynthesis of aromatic amino acids and betalains) before entering the TCA cycle were highly enriched when using glucose as a tracer. 'H' denotes the high oil cultivar and 'L' denotes the low oil cultivar. Glc-H and Glc-L were labeled with [$^{13}C_6$]glucose; Gln-H and Gln-L were labeled with [$^{13}C_5$,$^{2}N_2$]glutamine.*

5.3.2 Targeted analysis of the highlighted pathways

To further explore the isotope enrichments in the pathways identified by SUNDILE, a targeted study was performed in two steps. First, core pathways were explored by quantifying isotope enrichments in central metabolites using PIRAMID. The metabolites that were found in the first part of the targeted study were glucose-6-phosphate (G6P) in the glycolysis pathway, malate and fumarate in the TCA cycle, and phenylalanine, tryptophan, tyrosine, and aspartate in the amino acid biosynthesis pathway. In addition, glutamate was found to be adjacent to the TCA cycle connecting glutamine to the pathway (Fig. 5-2).

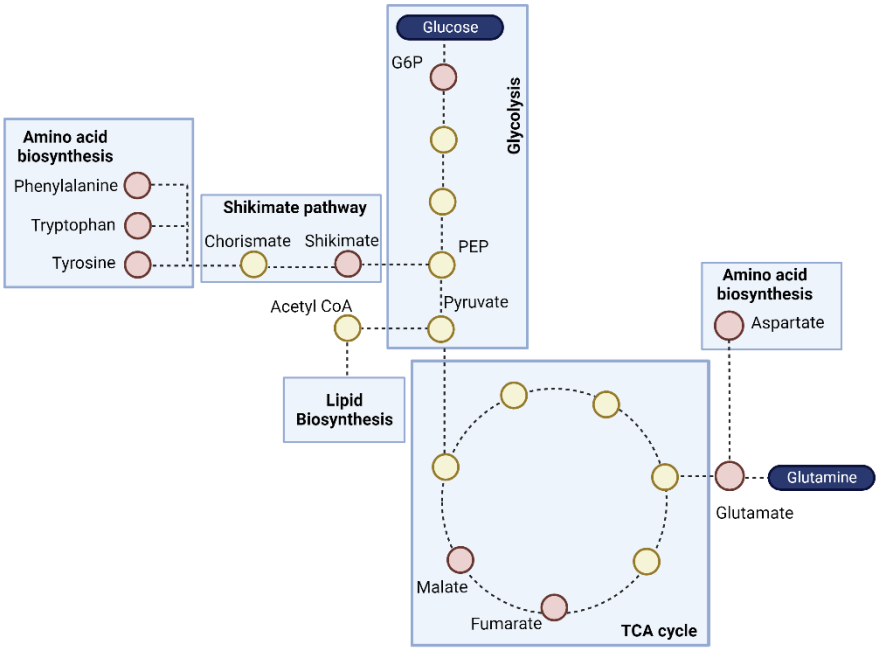


Figure 5-2. Diagram of central metabolism and location of labeled metabolites in the core pathways. *The enrichment of these metabolites was quantified.*

The analysis confirmed that [U-$^{13}C_6$]glucose enriched several intermediates in glycolysis, pentose phosphate pathway, and aromatic amino acid biosynthesis: G6P, shikimate, phenylalanine, tryptophan, and tyrosine. On the other hand, TCA cycle intermediates and amino acids in the aspartate and glutamate families were more highly enriched by [$^{13}C_5$,$^{15}N_2$]glutamine (Fig. 5-3).

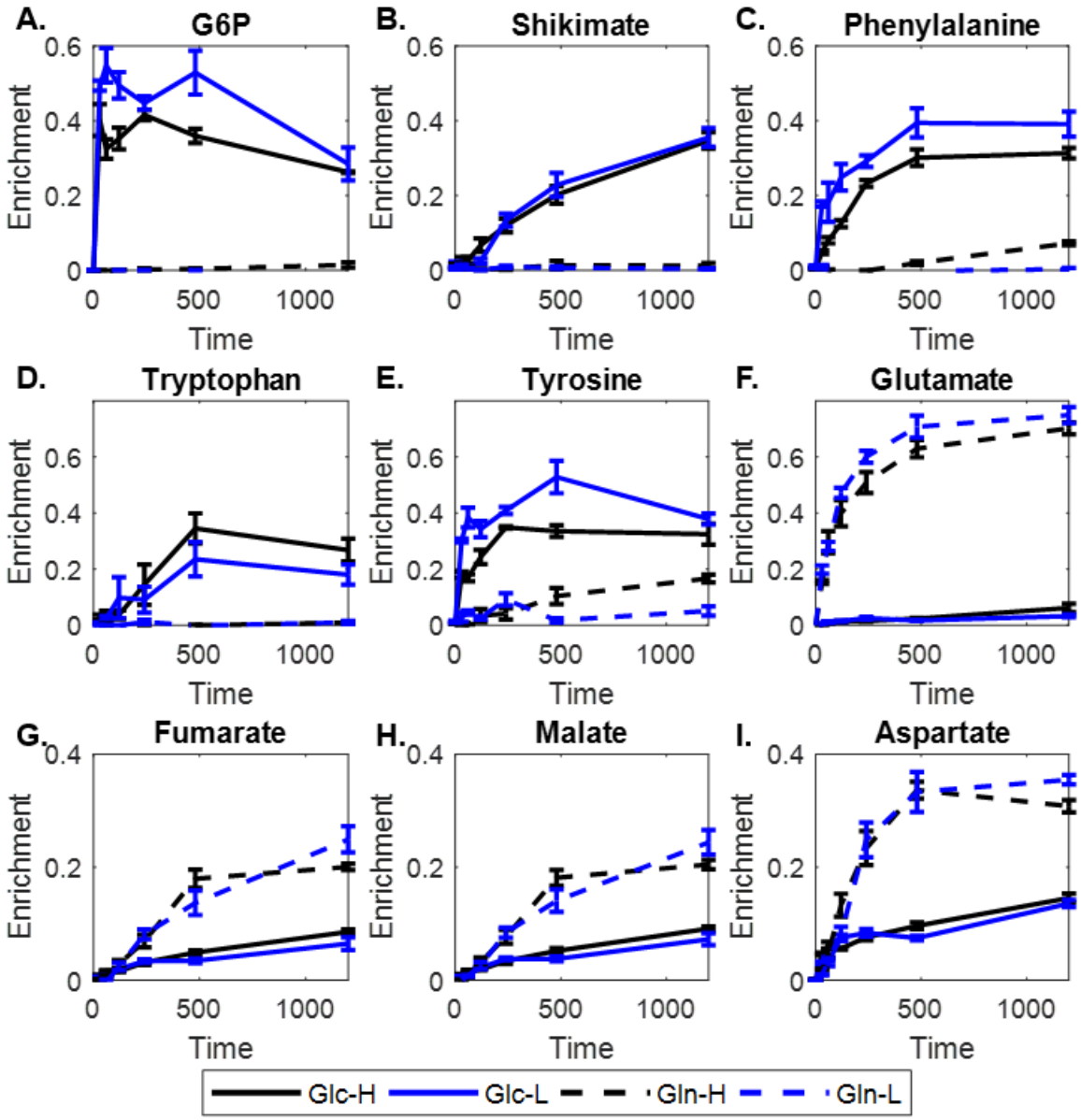


Figure 5-3. Results of the targeted analysis of labeled compounds within core metabolic pathways. *Metabolites in glycolysis and aromatic amino acid biosynthesis (A-E) exhibit high enrichment with [$^{13}C_6$]glucose but low enrichment with [$^{13}C_5$,$^{2}N_2$]glutamine. On the other hand, metabolites derived from the TCA cycle (F-I) exhibit the opposite trend.*

The results reveal a cluster of metabolites, including G6P, shikimate, phenylalanine, and tryptophan, exhibiting substantial enrichments when glucose is used as a substrate, with negligible enrichments when glutamine serves as the substrate. Statistical analysis using a two-tailed t-test indicated no significant difference in enrichment between high- and low-oil conditions for each metabolite in this cluster, regardless of the labeled substrate, at a 95% confidence level (p-value =

0.076). Conversely, another subgroup of metabolites showed the opposite trend, with higher enrichments in the glutamine tracer experiment. This subgroup includes glutamate, fumarate, malate, and aspartate. Once again, no statistically significant difference in enrichment of was observed for these metabolites between the high versus low oil cultivars.

Lastly, PIRAMID was applied to quantify the isotope enrichments of metabolites in nucleotides, nucleotide-sugars, flavonoids, and other metabolites outside of central metabolism that were detected by untargeted analysis of ^{13}C enrichment with SUNDILE (Fig. 5-4). Nucleotides and nucleotide sugars exhibited significantly higher enrichment when glucose was utilized as tracer. Notably, nucleotides displayed no enrichment when glutamine served as tracer, whereas nucleotide sugars showed low enrichment. Furthermore, a statistically significant difference was observed between low- and high-oil cultivars, with the low oil cultivar demonstrating higher enrichment in these pathways. In the isoflavonoid biosynthesis pathway, three metabolites displayed mixed behaviors: daidzin consistently exhibited high enrichment when glucose was used as tracer, regardless of the cultivar; rubiadin (originally marked as daidzein by SUNDILE) showed slightly higher enrichments in the high oil cultivar; and emodin (originally marked as genistein by SUNDILE) exhibited slower labeling dynamics, culminating in higher enrichment when glucose was used as the labeled substrate.

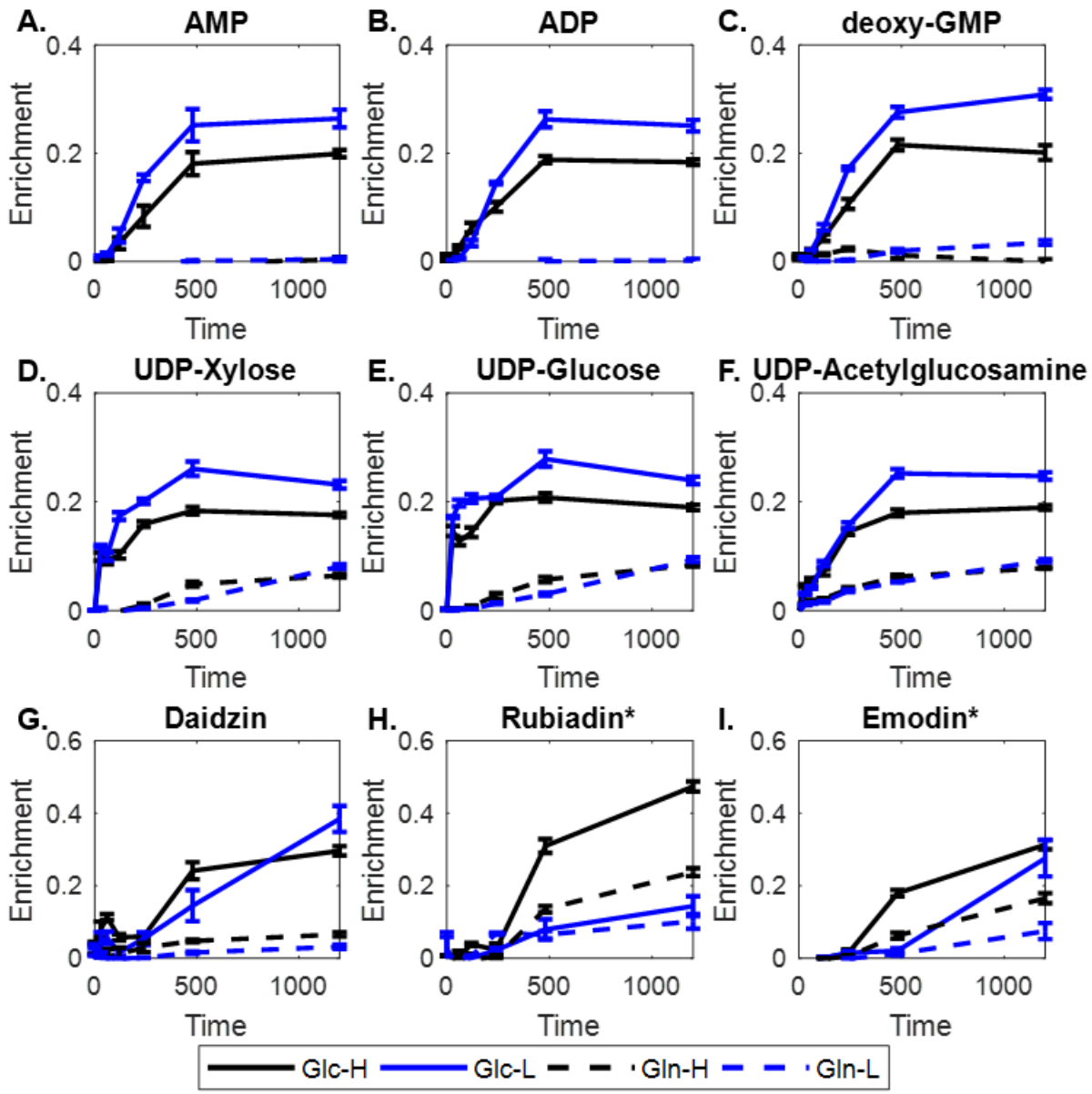


Figure 5-4. Targeted analysis of labeled metabolites outside of core metabolic pathways. *Nucleotides (A-C), nucleotide sugars (D-F), and chorismate derivatives (G-I) show different labeling patterns depending on the tracer applied (glucose (Glc) vs. glutamine (Gln)) and the oil content of the cultivar (H vs. L). These pathways indicate rerouting of carbon into nucleotide biosynthesis and away from flavanoids in the low oil cultivar..*Rubiadin and emodin were initially mislabeled by SUNDILE as daidzein and genistein, which belong to the isoflavonoid biosynthesis pathway.*

Finally, as a proof of concept that it is plausible for carbon to be diverted from TCA cycle into pyruvate (and eventually fatty acids or amino acids), the expression of malic enzyme (the enzyme required to catalyze the conversion of malate to pyruvate) was measured in both cultivars at two stages of seed maturation (i.e., R5 when the seed begins to form, and R5.5 when the seed is halfway

matured into a full seed) (Fig. 5-5). Malic enzyme activity was detected under both conditions but was significantly elevated in the high oil cultivar (p-values= 0.0048 and 0.0052 for R5 and R5.5, respectively).

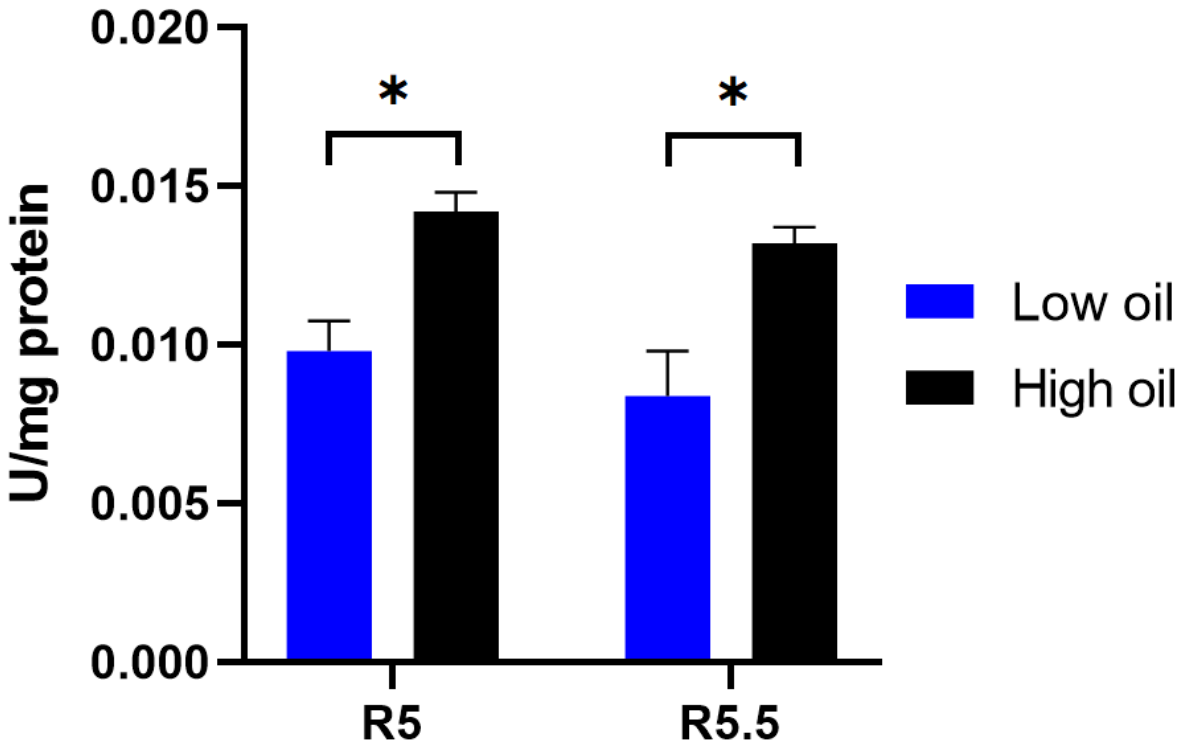


Figure 5-5. Malic enzyme activity measurements at different stages of seed development. *(R5=Beginning seed, R5.5=Half maturity into full seed). The measurements were normalized to total protein content assessed by a Bradford total protein assay. High oil producing cultivars consistently showed higher malic enzyme activity. The lines marked with * represent groups that are statistically different according to a t-test with a confidence level of 95%.*

5.4 Discussion

The untargeted analysis of ^{13}C-labeled metabolites indicates that entry of glucose into the TCA cycle is inhibited during seed development, and glucose carbon is instead diverted toward the synthesis of essential amino acids and complex molecules such as anthraquinones and isoflavonoids. The first step of glucose metabolism is its phosphorylation to G6P. As expected, G6P was highly enriched by glucose but not by glutamine (Fig. 5-3A). However, the presence of

an unlabeled pool of G6P hints at contributions from other endogenous sources of glucose (e.g., starch). Shikimate reaches similar final enrichments as G6P, but the labeling dynamics were slower. This behavior is expected since shikimate is more distant from glucose in the metabolic network, requiring more time before the labeled carbons reach the shikimate pathway (Fig. 5-3B). This pathway siphons flux away from glycolysis at the PEP branch point and from erythrose-4-phosphate in the pentose phosphate pathway [257]. The observed enrichment patterns suggest a metabolic state where glucose is preferentially used for biosynthesis of aromatic amino acids and isoflavanoids at this developmental stage.

Three aromatic amino acids—phenylalanine, tryptophan, and tyrosine—show higher enrichments in the glucose-labeled samples (Fig. 5-3C-E), which is consistent with the high enrichment of shikimate given that all three aromatic amino acids are derived from shikimate and share chorismate as a common precursor (Fig. 5-2) [258], [259]. Of the three, tryptophan is the only metabolite that exhibits increased enrichment in the high oil cultivar, suggesting that there is a shift in the carbon routing after chorismate that correlates with oil content. Seed embryos from the high oil cultivar direct flux from chorismate into anthranilate to produce tryptophan, whereas embryos of the low oil cultivar upregulate flux from chorismate into prephenate leading to synthesis of tyrosine and phenylalanine [260]. External tryptophan administration to growing seeds has been found to be a promoter of fatty acid accumulation in other oilseeds such as canola plants (*Brassica napus L.*) [261]. The hypothesized metabolic mechanism behind this phenomenon is through the synthesis of a class of plant hormones called auxins that are derived from tryptophan [262]. The enhanced growth and fatty acid accumulation stimulated by different types of auxins has been reported in other species such as microalgae [263], safflower [264], and in soybeans [265].

On the other hand, when glutamine was used as tracer, low enrichments in the glycolytic pathway indicate that TCA cycle intermediates were not used as a carbon source for gluconeogenesis. Aspartate equilibrates with oxaloacetate, which acts as an intermediate metabolite in the gluconeogenic pathway from the TCA cycle [266]. Despite evidence of enrichment in aspartate when glutamine was used as tracer, the lack of enrichment in G6P suggests that flux from the TCA cycle does not divert into glucose production. Furthermore, gluconeogenesis and glycolysis are enzymatically regulated to be active under opposing conditions [266], [267]. Considering the unlikelihood of simultaneous operation of both directions in the same pathway (i.e., glycolysis and gluconeogenesis) and the evidence of a metabolic state favoring glycolysis, we infer that gluconeogenic pathways are not active in soybeans at this developmental stage. This finding is consistent with studies reporting that gluconeogenesis becomes important only in later stages of plant development, after seed germination [266], [268], [269].

Analogous to aspartate, glutamate was highly enriched by glutamine but not by glucose (Fig. 5-3F), which reflects the direct conversion of glutamine to glutamate via glutamate synthase and the low carbon flux from glucose into the TCA cycle. Aspartate (Fig. 5-3I) was highly enriched by glutamine in both cultivars, significantly more than the enrichment achieved from glucose. Aspartate biosynthesis from oxaloacetate occurs in the cytoplasm [270], the mitochondria [271], and in chloroplasts [272]. However, the enrichments of other four-carbon intermediates in the TCA cycle (e.g., fumarate and malate) were lower than aspartate (Fig. 5-3G-H), indicating that aspartate biosynthesis occurs in a compartment where TCA cycle intermediates are preferentially labeled by glutamine. This observation aligns with prior research that has elucidated the compartmentalization of these pathways in soybeans. Specifically, while the TCA cycle takes place in the mitochondria [273], some of its intermediates are exported into other organelles where

their enrichment is diluted when exposed to unlabeled pools. For instance, malate and pyruvate are exported to plastids for fatty acid biosynthesis with acetyl-CoA as a intermediate metabolite [274], [275]. On the other hand, oxaloacetate is exported to the peroxisome to produce precursors of amino acids such as glycine via glyoxylate [276], [277]. The compartmentalization of these pathways can be seen in Fig. 5-6.

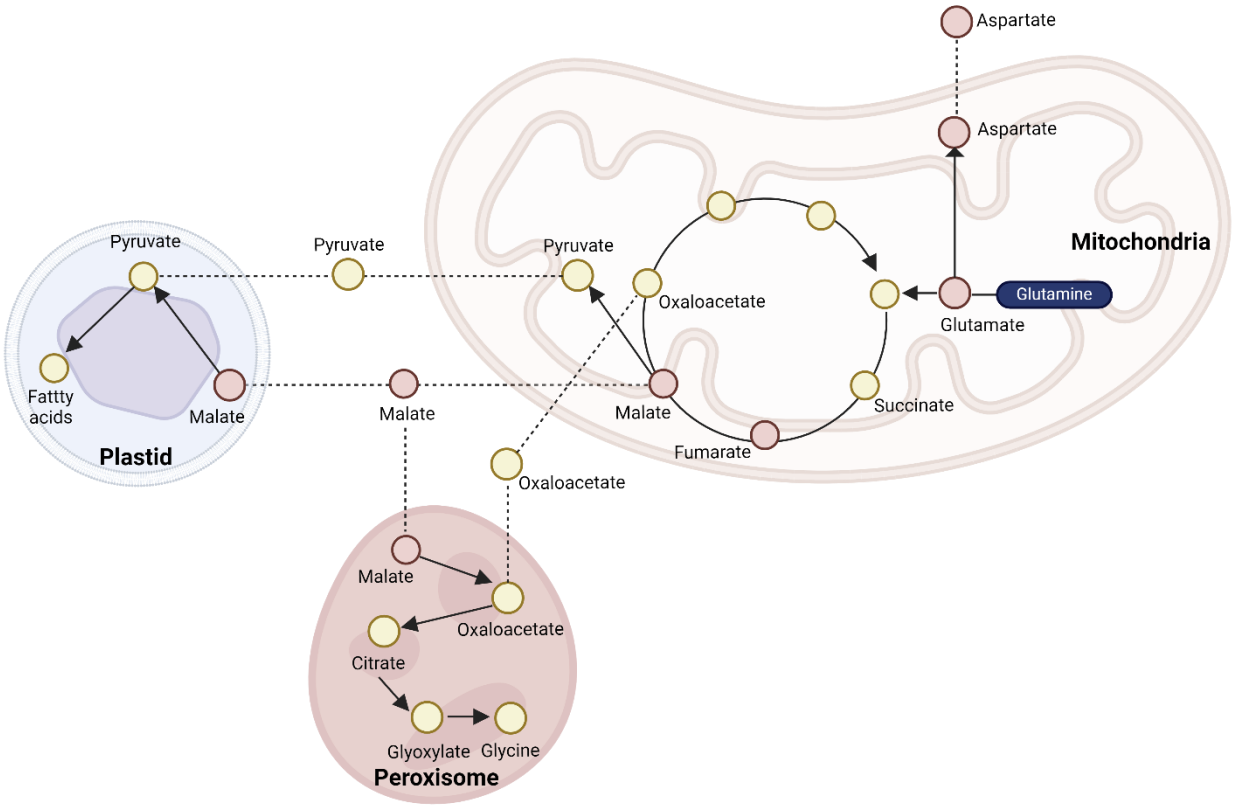


Figure 5-6. Compartmentalization of the TCA cycle, fatty acid biosynthesis, and amino acid biosynthesis pathways*. While the TCA cycle occurs in the mitochondria, some intermediate metabolites are exported into other organelles, diluting their enrichment. The higher enrichment of aspartate can be explained by analyzing the local enrichment of its precursors in the mitochondria.*

In general, nucleotides and nucleotide sugars showed a higher enrichment when using glucose as the tracer (Fig. 5-4A-F), which is logical since they are produced from the pentose phosphate pathway. However, they also consistently exhibited lower enrichments in the high oil cultivar, suggesting a rerouting of carbon away from nucleotide synthesis that could potentially be used to

produce oil. Furthermore, differences in the flavonoid and anthraquinone biosynthesis pathways were linked to three specific metabolites: daidzin, rubiadin, and emodin (Fig. 5-4G-I). However, SUNDILE initially annotated rubiadin and emodin as the isoflavanoids daidzein and genistein, respectively. Based on closer examination of their labeling trajectories, it was concluded that the identities suggested by SUNDILE for these two compounds were inconsistent with their labeling trajectories since they were more enriched than the isoflavanoid precursor phenylalanine and were also unexpectedly enriched by labeled glutamine (Fig. 5-7).

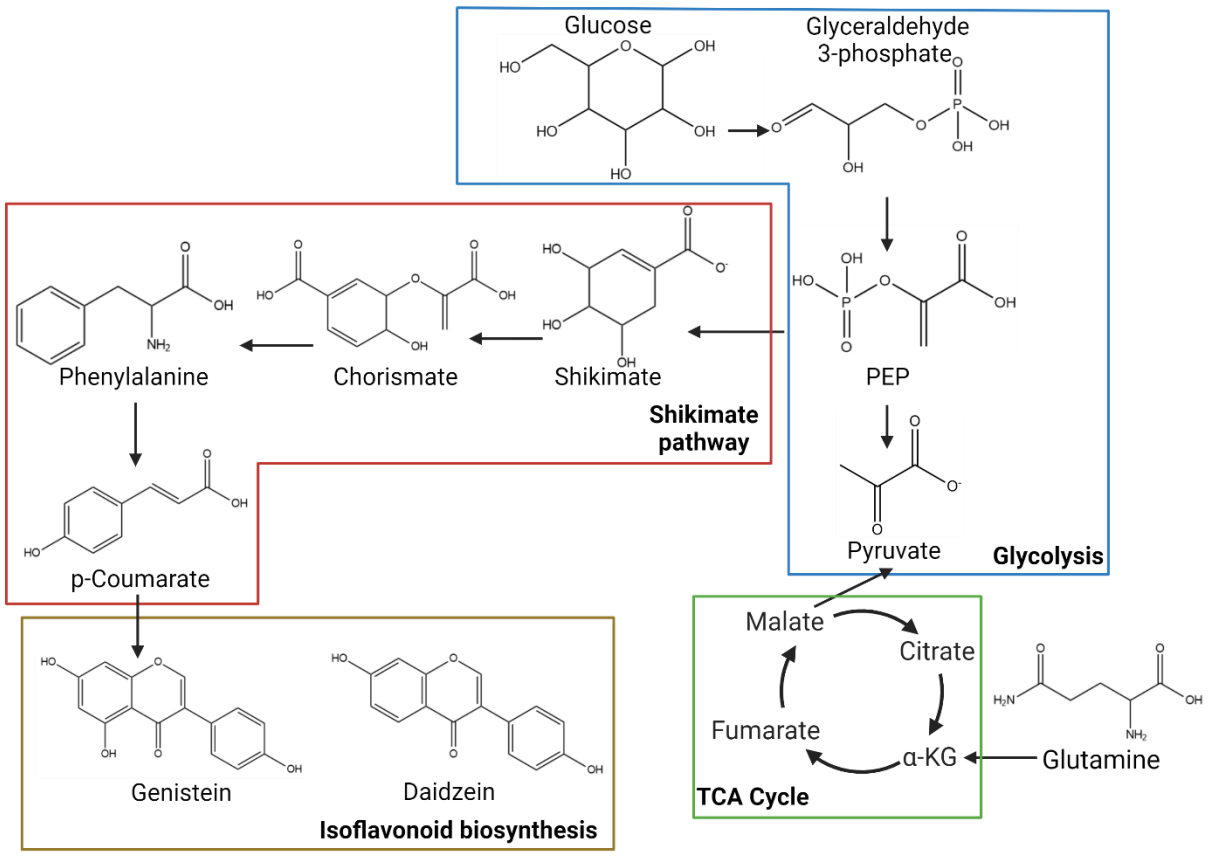


Figure 5-7. Isoflavonoid biosynthesis pathway. Isoflavonoids are synthesized from the shikimate pathway that has PEP as a precursor. *Given the low enrichments of phenylalanine and shikimate from glutamine labeling, it is unlikely to find enrichment of isoflavonoids, suggesting that the identities of genistein and daidzein proposed by SUNDILE were erroneous.*

Given the impossibility of the identities suggested by SUNDILE, additional information was required to determine the correct identities of these compounds. MS^2 analysis of peaks

corresponding to these compounds suggested that their identities were rubiadin and emodin instead, which are isomers of daidzen and genistein, respectively (Fig. 5-8).

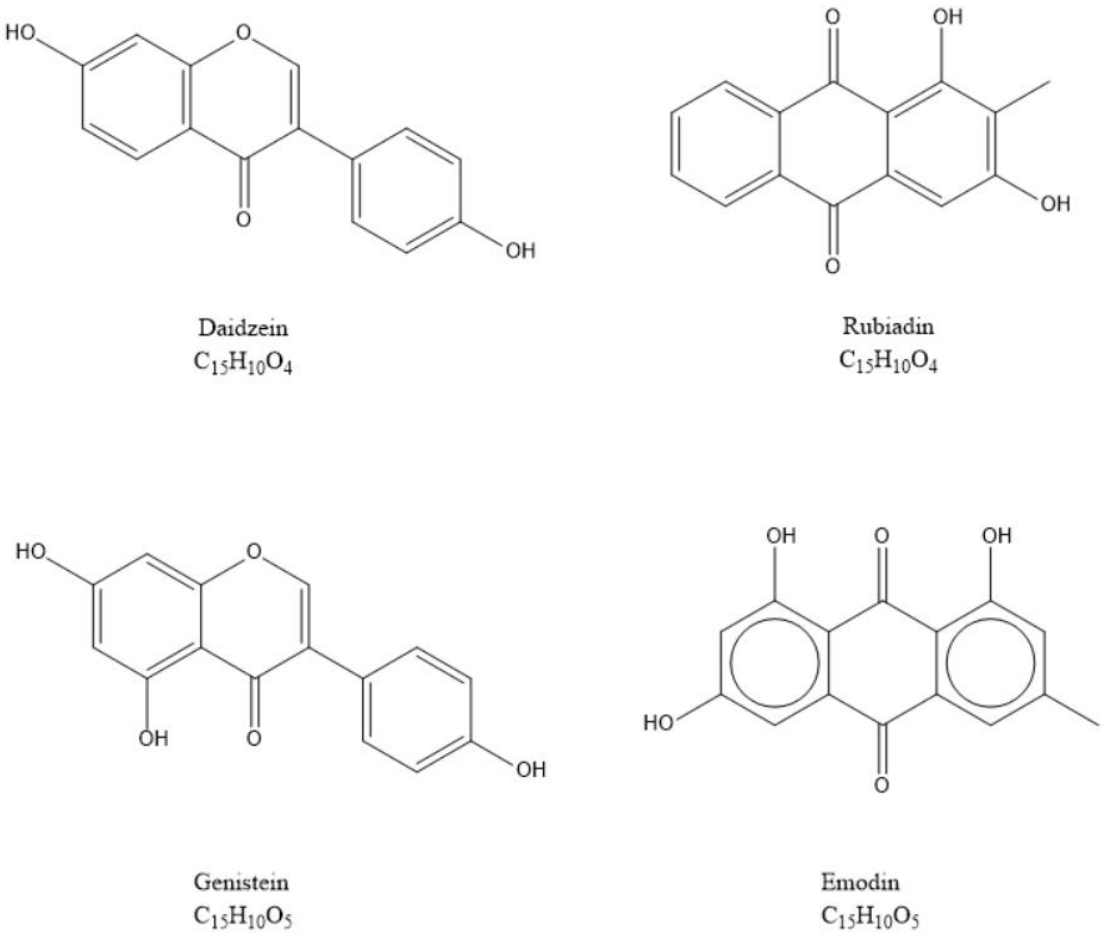


Figure 5-8. Molecular comparison of the compound identities proposed by SUNDILE (daidzein and genistein) against the suggestions from MS^2 analysis of these same peaks (rubiadin and emodin). *Each pair of suggestions have the same molecular formula. While SUNDILE suggested compounds from the isoflavonoid family, the MS^2 analysis pointed to compounds in the anthraquinone family.*

This classification proved to be more logical given that anthraquinones (the class rubiadin and emodin belong to) are derived from PEP, acetyl-CoA, and TCA cycle intermediates, thus explaining their enrichment from labeled glutamine [278] (Fig. 5-9). The misclassification by SUNDILE can be attributed to the fact that anthraquinones are defined in KEGG as compounds, but their biosynthesis pathways are not available. Consequently, the program was not able to link these metabolites with the labeled tracer, hence removing them from the list of possibilities.

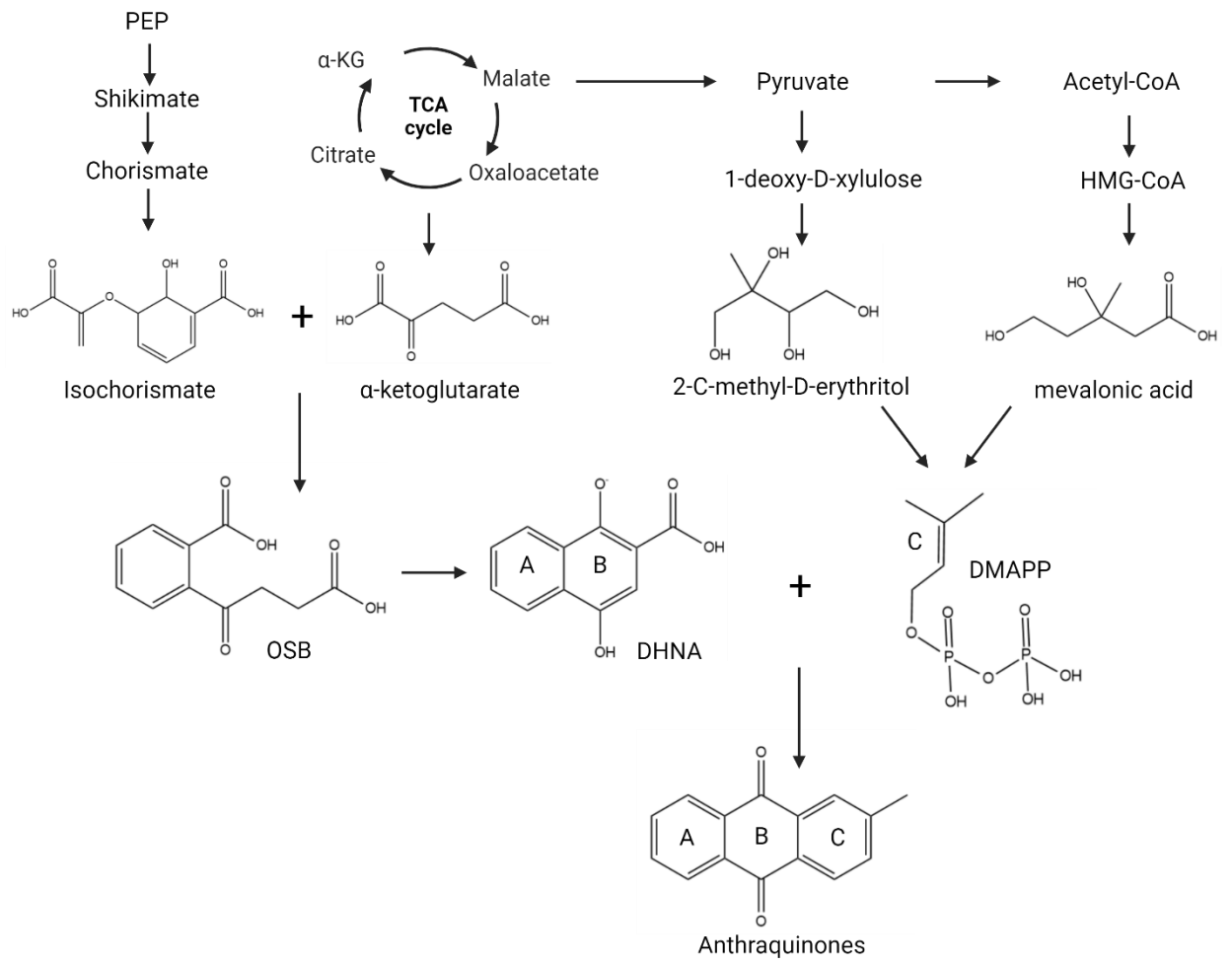


Figure 5-9. Anthraquinone biosynthesis. Anthraquinones are synthesized from a backbone that is a product of both PEP and alpha-ketoglutarate. *The remaining part of the molecule comes from pyruvate derivatives. The labeling found in these compounds using glutamine as a tracer can be explained by the enrichment coming from α-ketoglutarate in the TCA cycle and from the derivatives of pyruvate that are enriched from malate. On the other hand, the labeling found using glucose as a tracer can be explained by the enrichment of the shikimate pathway coming from PEP. OSB=o-succinylbenzoic acid. DHNA=1,4-dihydroxy-2-naphtolic acid. DMAPP=3,3-dimethylallyl diphosphate.*

For this hypothesis to be considered valid, the labeling of anthraquinones from glutamine must have been derived from either α-ketoglutarate (whose enrichment can be inferred from other TCA cycle metabolites) or acetyl-CoA. Unfortunately, acetyl-CoA was not detected in the targeted analysis, so information of acetyl-CoA labeling must be deduced from other metabolites. One possibility for acetyl-CoA to become labeled from glutamine involves carbon routing from the

TCA cycle into pyruvate via malic enzyme [279]. High malic enzyme activity was observed in both cultivars (Fig. 5-5), confirming its potential for contributing to acetyl-CoA production from glutamine. In addition, the high oil cultivar showed a significantly higher expression of malic enzyme in both developmental stages. Although acetyl-CoA could also be labeled from glutamine after entering the TCA cycle through the α-ketoglutarate node and converted into citrate via isocitrate dehydrogenase, ultimately leading to acetyl-CoA by ATP-citrate lyase, enzyme measurements for these steps were not obtained, making conclusions regarding this alternative pathway impossible to be made.As previously mentioned, approximately a quarter of the carbon in soybean fatty acids originates from pyruvate synthesized from malate [254]. The difference in the levels of malic enzyme expression between the low- and high-oil cultivars provides a feasible strategy for optimizing oil production in soybeans by focusing on the manipulation of malic enzyme. This approach has been previously tested by expressing *Arabidopsis* malic enzyme alleles in homozygous transgenic soybean plants, resulting in an increase in fatty acid content of up to 2% [256]. In other organisms, the overexpression of different isoforms of malic enzyme increased the fatty acid content by 30% in *Mucor circinelloides* [280] and by ~40% in *Phaeodactylum tricornutum* [281].

5.5 Conclusions

Labeling of soybean embryos with ^{13}C-glucose resulted in low enrichment of TCA cycle intermediates, and glucose carbon was instead diverted into the production of aromatic compounds and nucleotides. On the other hand, glutamine carbon directly enriched the TCA cycle and was converted into pyruvate by means of malic enzyme activity; labeled carbon was further routed into the production of lipids and amino acids derived from pyruvate, aspartate, and glutamate (Fig. 5-10).

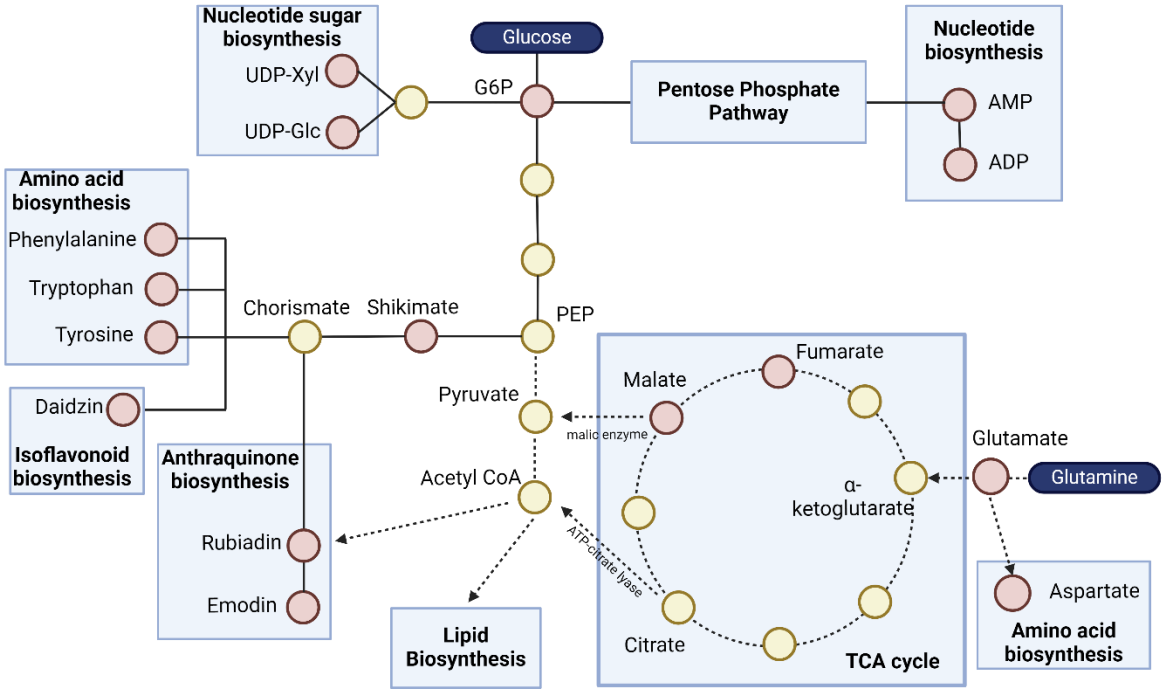


Figure 5-10. Metabolic map depicting the carbon routing when [U-$^{13}C_6$]glucose or [$^{13}C_5$,$^{2}N_2$]glutamine were used as tracers. *The metabolites in red were integrated in a targeted manner, the dotted lines represent the carbon routes traced by glutamine and the solid lines represent the carbon routes traced by glucose.*

Measurements of malic enzyme activity at different developmental stages further support the role of this enzyme in generating pyruvate from TCA cycle intermediates, with increasing activity correlated to higher oil content (Fig. 5-5). This is consistent with previous research that applied targeted studies designed to directly assess malic enzyme flux [245], [254], [256]. Furthermore, the low oil cultivar consistently exhibited higher enrichment in nucleotides and nucleotide sugars, suggesting that glucose carbon could be re-routed from these pathways to improve oil synthesis. It has been previously reported that oil accumulation competes with synthesis of other major biomass components (protein and nucleotides) in soybean [282], [283] and other plant species such as *Oliveria decumbens* [284], *Cuminum cyminum L* [285], and *Salvia officinalis* [286].

5.6 Acknowledgements

We thank Dr. Shrikaar Kambhampati, Dr. Doug Allen, and Dr. Bradley Evans in the Donald Danforth Plant Science Center for their support in the raw data gathering and their input in our scientific discussions.

This work was funded by the National Institutes of Health Grant No. U01 CA235508.

5.7 Appendix

Pathway	Glc-H	Glc-L	Gln-H	Gln-H
Glycolysis / Gluconeogenesis	0.337	0.170	0.206	0.317
Citrate cycle (TCA cycle)	0.138	0.111	0.300	0.340
Pentose phosphate pathway	0.262	0.302	0.062	0.057
Pentose and glucuronate interconversions	0.230	0.331	0.228	0.358
Fructose and mannose metabolism	0.064	0.080	0.663	0.663
Galactose metabolism	0.323	0.263	0.069	0.050
Ascorbate and aldarate metabolism	0.206	0.189	0.143	0.169
Ubiquinone and other terpenoid-quinone biosynthesis	0.268	0.300	0.452	0.377
Oxidative phosphorylation	0.292	0.222	0.353	0.355
Photosynthesis	0.331	0.504	0.150	0.109
Arginine biosynthesis	0.097	0.110	0.468	0.406
Purine metabolism	0.326	0.350	0.000	0.000
Caffeine metabolism	0.047	0.027	0.000	0.007
Pyrimidine metabolism	0.108	0.056	0.301	0.456
Alanine, aspartate and glutamate metabolism	0.153	0.180	0.355	0.418
Glycine, serine and threonine metabolism	0.190	0.273	0.216	0.356
Monobactam biosynthesis	0.210	0.281	0.167	0.360
Cysteine and methionine metabolism	0.210	0.273	0.230	0.310
Valine, leucine and isoleucine degradation	0.259	0.245	0.606	0.613
Valine, leucine and isoleucine biosynthesis	0.212	0.318	0.204	0.265
Lysine biosynthesis	0.193	0.241	0.472	0.509
Lysine degradation	0.197	0.218	0.507	0.510
Arginine and proline metabolism	0.126	0.112	0.284	0.439
Histidine metabolism	0.164	0.147	0.539	0.580
Tyrosine metabolism	0.228	0.243	0.340	0.471
Phenylalanine metabolism	0.341	0.397	0.289	0.341
Tryptophan metabolism	0.610	0.432	0.000	0.000
Phenylalanine, tyrosine and tryptophan biosynthesis	0.503	0.525	0.029	0.020
beta-Alanine metabolism	0.031	0.038	0.160	0.295
Taurine and hypotaurine metabolism	0.225	0.169	0.428	0.495
Phosphonate and phosphinate metabolism	0.266	0.367	0.117	0.058
Selenocompound metabolism	0.350	0.317	0.131	0.231

Cyanoamino acid metabolism	0.409	0.394	0.312	0.367
D-Amino acid metabolism	0.162	0.200	0.296	0.377
Glutathione metabolism	0.195	0.104	0.286	0.370
Starch and sucrose metabolism	0.270	0.261	0.032	0.057
Amino sugar and nucleotide sugar metabolism	0.453	0.443	0.320	0.373
Glycerolipid metabolism	0.298	0.349	0.098	0.057
Inositol phosphate metabolism	0.173	0.160	0.062	0.057
Pyruvate metabolism	0.122	0.072	0.336	0.406
Glyoxylate and dicarboxylate metabolism	0.131	0.127	0.287	0.383
Propanoate metabolism	0.160	0.184	0.478	0.551
Butanoate metabolism	0.136	0.153	0.426	0.479
C5-Branched dibasic acid metabolism	0.105	0.164	0.368	0.481
Carbon fixation in photosynthetic organisms	0.308	0.440	0.145	0.231
Thiamine metabolism	0.413	0.577	0.254	0.360
Riboflavin metabolism	0.210	0.458	0.458	0.000
Vitamin B6 metabolism	0.255	0.577	0.637	0.597
Nicotinate and nicotinamide metabolism	0.183	0.204	0.300	0.302
Pantothenate and CoA biosynthesis	0.222	0.504	0.155	0.207
Biotin metabolism	0.221	0.300	0.000	0.000
Lipoic acid metabolism	0.202	0.209	0.445	0.617
Porphyrin metabolism	0.079	0.217	0.387	0.659
Terpenoid backbone biosynthesis	0.000	0.000	0.117	0.006
Carotenoid biosynthesis	0.010	0.000	0.127	0.000
Zeatin biosynthesis	0.258	0.419	0.188	0.032
Nitrogen metabolism	0.166	0.060	0.673	0.723
Sulfur metabolism	0.247	0.222	0.654	0.681
Phenylpropanoid biosynthesis	0.421	0.451	0.477	0.548
Flavonoid biosynthesis	0.360	0.260	0.335	0.331
Anthocyanin biosynthesis	0.000	0.754	0.000	0.000
Isoflavonoid biosynthesis	0.304	0.216	0.263	0.230
Flavone and flavonol biosynthesis	0.167	0.236	0.078	0.242
Stilbenoid, diarylheptanoid and gingerol biosynthesis	0.434	0.449	0.389	0.679
Isoquinoline alkaloid biosynthesis	0.441	0.206	0.261	0.062
Tropane, piperidine and pyridine alkaloid biosynthesis	0.349	0.420	0.045	0.018
Betalain biosynthesis	0.435	0.711	0.000	0.000
Glucosinolate biosynthesis	0.413	0.401	0.045	0.018
Aminoacyl-tRNA biosynthesis	0.286	0.278	0.267	0.333
Metabolic pathways	0.216	0.246	0.264	0.301

Biosynthesis of secondary metabolites	0.215	0.238	0.237	0.222
Carbon metabolism	0.230	0.249	0.298	0.365
2-Oxocarboxylic acid metabolism	0.227	0.276	0.244	0.261
Biosynthesis of amino acids	0.217	0.259	0.224	0.235
Nucleotide metabolism	0.314	0.351	0.000	0.000
Biosynthesis of cofactors	0.275	0.340	0.246	0.259
Biosynthesis of nucleotide sugars	0.446	0.491	0.320	0.373
ABC transporters	0.167	0.192	0.117	0.230
Plant hormone signal transduction	0.008	0.000	0.000	0.032
Sulfur relay system	0.258	0.189	0.131	0.231

Table 5-A1. Enrichment scores of common pathways found using different tracers (glucose=Glc, glutamine=Gln) and soybean cultivars (H=high oil, L=low oil) .

CHAPTER 6

COMPREHENSIVE ISOTOPE-BASED METABOLOMICS OF MAMMALIAN METABOLISM UNDER OBESOGENIC CONDITIONS

Abstract

Obesity and diabetes are complex metabolic disorders with far-reaching health implications. To gain deeper insights into the underlying mechanisms and pathways involved, *in vivo* experiments involving stable isotope tracers have been historically used. However, these approaches have typically been performed in a targeted manner, which leaves many metabolites and pathways unexplored. This chapter presents two pilot studies combining untargeted and targeted metabolomics analysis, aided by the bioinformatic tools introduced in previous chapters. One study focused on fasting mice that received an infusion of [$^{13}C_3$]propionate, and another study focused on mice under a hyperinsulinemic/euglycemic clamp that received an infusion of [$^{13}C_6$]glucose. Important insights for future experiments using our developed workflow include variations in pathway or metabolite enrichment when comparing untargeted and targeted results between the negative and positive acquisition modes, and deviations in the calculated enrichments of the nucleotide biosynthesis pathway between the untargeted and targeted steps of the workflow. The latter are attributed to the high molecular masses of pathway intermediates, which are reflected in higher tolerances and the misassignment of peaks into the corresponding isotopologues. Furthermore, variation in pentose phosphate pathway activity under obese conditions was predicted by SUNDILE and confirmed by PIRAMID. Finally, unexpected enrichments in metabolites that are not typically synthesized by mammals (i.e., aminoisobutyric acid, threonate, and propionate) were detected. This phenomenon could be explained by the reversibility of certain reactions or the production of these metabolites by gut microbiota. In addition to the experimental

and analytical results, the chapter outlines potential areas of improvement in SUNDILE algorithms and proposes a roadmap for future work. The combined findings and insights from this research are poised to shape future investigations that apply stable isotopes to assess metabolic responses to obesity, diabetes, and related conditions, which offers promising avenues for basic research with potential application to clinical practice.

6.1 Introduction

Obesity has grown as a public health threat over the last few decades [287], [288], with increasing body-mass index (BMI) trends observed in children, adolescents, and adults across the globe [289], [290]. Obesity is associated with premature mortality and the development of the most prevalent diseases in modern society such as type 2 diabetes, hypertension, cardiovascular disease, viral infections, and certain cancers [291], [292], [293], [294].

One variable which must be tightly controlled in obesity research is chronic and acute nutrient availability. Several long-term factors such as age, underlying medical conditions, and genetics play a crucial role in the metabolic response to obesity [295], [296], [297]. Similarly, short-term variables such as stress, sleep, and diet will also be reflected in the metabolic state of an organism [298], [299], [300]. These variables are routinely manipulated to test their impact on metabolic control. Furthermore, specialized diets have been developed to more closely replicate human diseases in rodent models [301], [302], [303], [304]. Needless to say, *in vivo* mammalian metabolism is often studied under different conditions with targeted manipulation of metabolic states.

Metabolic research in obesogenic conditions has been performed across species including zebrafish, nematodes, flies, primates, and rodents; the latter being the most used model given their cost effectiveness, multiparity, physiological resemblance to humans, and genetic malleability [305], [306]. Several metabolomics approaches have been developed to study different metabolic conditions in mice [307], [308]. However, most of them are based on comparison of the static abundance of metabolites in body fluids or tissues [309] or whole-body assessment of bioenergetics by measuring dynamic changes in oxygen consumption or carbon dioxide production in the live animal [310]. These approaches lack the ability to resolve changes in metabolic pathway

flux (i.e., the primary indicator of metabolic activity), which can only be achieved through the use of radioactive or stable isotopes to trace the dynamic conversion of substrates into products [311].

In stable isotope-based *in vivo* metabolomics, isotopes are administered to animals in solid [312] or liquid [313] diets, or by direct infusion into the organism [314]. Tracers administered in the diet rely on the process of absorption in the gastrointestinal system, which leads to variable and unpredictable delays in the appearance and disappearance of the tracer in circulation. On the other hand, an intravenous infusion exposes relevant tissues to a well-defined quantity of the tracer that can be maintained constant in the circulation [315]. To perform advanced analyses like ^{13}C metabolic flux analysis (MFA), the system needs to reach isotopic and metabolic steady state (i.e., the isotope enrichment and concentrations of metabolites remain constant over time) [316], [317]. Therefore, direct infusion is the preferred method of tracer administration for rigorous studies of steady-state metabolism.

Within the framework of obesity and diabetes research, fasting conditions are often used to limit the impact of dietary nutrients and normalize basal metabolic activity. The choice of which labeled nutrient(s) to administer (e.g., alanine, glycerol, lactate, propionate, glucose, etc.) must be tailored to match the metabolic state of interest and to address relevant physiological questions [318], [319], [320], [321], [322]. While some questions only require data from fasting subjects, an organism's metabolic state may need to be manipulated to test certain hypotheses. A salient example of this type of manipulation is the hyperinsulinemic-euglycemic clamp, which is considered the "gold-standard" assay for assessing insulin sensitivity [323]. In these studies, insulin is infused at a constant rate to elevate circulating insulin levels and glucose is infused at a variable rate to maintain euglycemia. Radioactive and stable isotope tracers are often administered

to assess plasma glucose kinetics. Thus, the glucose infusion rate as well as rates of glucose production and disappearance are indicators of insulin sensitivity [324], [325].

As discussed previously, infused tracers circulate throughout the body and enrich a variety of tissues and metabolic pathways. However, the vast majority of *in vivo* stable isotope studies collect metabolomics data in a targeted manner [326], leaving a wide range of metabolic pathways unexplored. To address this issue, we analyzed two murine isotope-based metabolomics datasets obtained under different physiological conditions. Specifically, mice were 1) fasted and infused with a cocktail of ^{13}C and ^{2}H isotopes or 2) subjected to a hyperinsulinemic-euglycemic clamp that included the infusion of [$^{13}C_6$]glucose. Tissues from mice in both cohorts were analyzed in an untargeted manner to detect metabolites from core and peripheral metabolic pathways, followed by a targeted study to validate and quantify the extent of biomarker enrichment. By employing this two-pronged strategy, we were able to substantiate the utility of our bioinformatic tools described in Chapters 3 and 4 using an *in vivo* model system. Based on these results, we also propose alternative strategies to the experimental design which may further leverage unique capabilities of the tools.

6.2 Methods

All the datasets that had been tested in the prior development of SUNDILE and PIRAMID were data from simpler *ex vivo* or *in vitro* experiments. Before applying the workflow to a large-scale *in vivo* experiment, we performed a preliminary test on samples from a limited number of mice (n=2) to see if the implementation of our tools was feasible and yielded logical results. The design of this experiment is described in section 6.2.1. After finding success in this limited dataset, we proceeded to implement the workflow in an experiment with a full experimental design. This

experiment is described in section 6.2.2. All procedures were performed with approval from the Vanderbilt Animal Care and Use Committee.

6.2.1 In *vivo* procedures used to study fasted mice

Selected samples from a previous study [168] were re-analyzed using high-resolution mass spectrometry (HRMS). Briefly, male mice (n=2) presenting hyperphagia due to knockout of the melanocortin-4 receptor (*Mc4r*) gene that regulates feeding behavior and satiety [327], [328] were given ad libitum access to food and water and maintained on a 12/12h light-dark cycle at a constant temperature (23°C) and humidity. After weaning, mice were provided a standard chow diet (5L0D, 29% protein, 58% carbohydrates, 13% fat by caloric contribution; LabDiet, St. Louis, MO) for 28 weeks. Indwelling two-part catheters were surgically implanted in the jugular vein for infusing and in the carotid artery for sampling ~1 week before the study. Overnight-fasted mice received an intravenous primed (440 μmol/kg) continuous (4.4 μmol/kg/min) infusion of [6,6-$^{2}H_2$]glucose and a bolus of $^{2}H_2O$ to enrich body water to 4.5% (abbreviated collectively as ^{2}H). Two hours later, mice received a primed (1.1 mmol/kg) continuous (0.055 mmol/kg/min) infusion of sodium [$^{13}C_3$]propionate (Cambridge Isotope Laboratories, Tewksbury, MA). The isotope infusions took place during a ~4.5-h time window. At the end of the study, mice were euthanized by cervical dislocation, and liver tissue was rapidly excised and freeze-clamped in liquid nitrogen. The total fast duration was ~20 h at the time of sample collection. All samples were stored at −80°C.

6.2.2 *In vivo* procedures in the hyperinsulinemic-euglycemic clamp

Male diet-induced obese (DIO) and control C57Bl/6J (Jackson laboratories) mice (n=4 per group) were given ad libitum access to diets and water and maintained on a 12/12h light-dark cycle at a constant temperature (23°C) and humidity. DIO mice were placed on a high-fat (HF) diet

(Research Diets D12492, New Brunswick, NJ) at 6 weeks of age while control mice (n=4) were provided a control diet. The mice were maintained on these diets until they were 23 weeks old. Indwelling catheters were surgically implanted in the jugular vein for infusing and in the carotid artery for sampling ~1 week before the study. Mice were fasted for 5 hours before the experiment to allow the gut to empty. A blood sample was taken directly prior to the start of the clamp to measure the initial glucose level (euglycemic) and other basal metabolic parameters. The insulin clamp was conducted at a continuous 4 mU/kg/min infusion rate of insulin, and a variable infusion of D50 (i.e., a 50% dextrose solution commonly used to treat hypoglycemia) to maintain euglycemia; the D50 solution was enriched up to ~60% with [$^{13}C_6$]glucose. Euglycemia (~150 mg/dl) was maintained during clamps by measuring blood glucose every 10 minutes and adjusting the [$^{13}C_6$]glucose infusion rate accordingly. The clamp was maintained for a ~2h period. At the end of the study, mice were euthanized by cervical dislocation, and kidney and liver tissue were rapidly excised and freeze-clamped in liquid nitrogen. All samples were stored at −80°C.

The unlabeled control group for this study was taken from unlabeled kidney and liver tissues that were collected from wild-type C57Bl/6J mice that did not receive any type of infusion or surgical procedure. These samples were used for metabolite identification through comparison to reference libraries and to assess the expected retention time and background enrichment of the detected metabolites.

6.2.3 Metabolite extraction

Tissues were homogenized at 4ºC using a bead blender and metabolites were extracted from tissues using a biphasic extraction using a 1:1:1 methanol/water/chloroform mixture. The samples were maintained on ice during the extraction. The polar layer of the extract was isolated and air-dried

before being reconstituted in a 9:1 water:acetonitrile mixture. The non-polar layer of the extract was discarded.

6.2.4 MS data acquisition

A Thermo Scientific Q-Exactive mass spectrometer equipped with a Millipore SeQuant ZIC HILIC column (2.1 × 100 mm, 3.5 μm) was used to analyze the metabolite extract samples. Mobile phase A (9:1 water:acetonitrile + 5mM ammonium formate) and mobile phase B (9:1 acetonitrile:water with 5mM ammonium formate) were pumped through the column at a total flowrate of 200 μL/min using the following gradient: 95% B for 2 min, 95 to 40% B over 16 min, 40% B held for 2 min, 40 to 95% B over 15 min, and 95% B held for 10 min (gradient length, 45 min). Full scan MS^1 and MS^2 DDA data were acquired in negative mode between 80 and 500 Da at a resolution of 60,000. Source ionization parameters were: spray voltage, 3.0 kV; transfer temperature, 280°C; S-lens level, 40; heater temperature, 325°C; sheath gas, 40; aux gas, 10; and sweep gas flow, 1.

6.2.5 Untargeted data analysis

The raw MS data files were converted to .mzml format using the ProteoWizard MSConvert tool to process the data from positive and negative ionization modes into separate files. The extraction of peaks was performed with XCMS [90] using the following parameters: method=CentWave, prefilter=[2,500], ppm=20, snthresh=2, peakwidth=[5,100], and noise=0. A retention time correction step was included using the 'obiwarp' method. Labeled pathways were detected using SUNDILE (Chapter 3) with a resolution of 20 ppm and a retention time window of 0.1 minutes, monitoring only for ^{13}C isotopologues. The list of pathways was filtered to only include the pathways directly associated with murine metabolism [329]. Pathways associated with genetic or

environmental information processing, cellular processes, organismal systems, human diseases, and drug development were excluded from the analysis.

Pathway scores were normalized to the enrichment of plasma glucose, which was calculated using PIRAMID (Chapter 4). The normalized pathway scores were analyzed using agglomerative hierarchical clustering in two dimensions (pathways × sample type, i.e., tissue of origin and diet treatment) using the Euclidean distance between the scores. The cutoff of the optimal number of clusters (5) was determined by the maximization of the Calinski-Harabasz score, optimizing the ratio of the variances between clusters.

Identification of metabolites was performed using MS-DIAL to compare the acquired MS^2 data against the library "*ESI(-)-MS/MS from standards+bio+in silico*" for negative mode and "*ESI(-)-MS/MS from standards+bio+in silico"* for positive mode data. Both libraries are publicly available on the software website. The default software parameters were used throughout the analysis.

6.2.6 Targeted data analysis

A list of target metabolites was generated by matching the metabolites identified by MS-DIAL with the list of labeled metabolites detected by SUNDILE. The entire list of target metabolites was searched and integrated using our custom software tool PIRAMID (Chapter 4). Extracted features were assigned to isotopologues of the target compounds using a mass tolerance of 10 ppm; the peaks were smoothed using a second-order Savitzky-Golay filter with a frame length of 11 data points; and the isotopic purity of the tracers was assumed to be 99%.

6.3 Results

6.3.1 Evaluating the efficacy of untargeted and targeted approaches *in vivo*: A case study using the Mc4r$^{-/-}$ dataset

Liver tissue samples previously collected from genetically obese *Mc4r*$^{-/-}$ mice infused with [$^{13}C_3$]propionate [168] were re-analyzed by HRMS using a Q-Exactive mass spectrometer operated in both positive and negative ionization modes. Prior to the study, mice were fasted overnight to deplete liver glycogen and shift liver metabolism to a gluconeogenic state. Untargeted analysis of metabolic pathway activity was followed by compound identification and targeted analysis of metabolite ^{13}C enrichments as described in the subsections below.

6.3.1.1 Untargeted detection of active pathways using SUNDILE

SUNDILE was applied to detect active metabolic pathways and identify isotopically labeled metabolites in these pathways. The list of active pathways detected by SUNDILE is presented in Table 6-A1, and the list of labeled metabolites identified by SUNDILE is provided in Table 6-A2. A heatmap of the pathway enrichment scores (i.e., an indicator of isotopic enrichment of the metabolites within each pathway) is shown in Fig. 6-1.

Pathways associated with sugar metabolism consistently exhibited high scores. Conversely, pathways linked to amino acid metabolism and TCA cycle associated pathways displayed consistently low scores across the samples. This behavior contradicted the biological expectations of the experiment: since propionate was used as a metabolic tracer, it was expected that the highest enrichments would be found in propionate and metabolites immediately downstream from it, leading to a high score in the “propanoate metabolism” pathway. However, this pathway showed

a relatively low activity score, specifically compared to pathways related to the metabolism of sugars and nucleotide sugars (Fig. 6-1).

ActiveMap	MapScore
Biosynthesis of nucleotide sugars	0.2742
Amino sugar and nucleotide sugar metabolism	0.2730
Starch and sucrose metabolism	0.2686
Galactose metabolism	0.2601
Pentose phosphate pathway	0.2461
Fructose and mannose metabolism	0.2461
Glyoxylate and dicarboxylate metabolism	0.2107
Valine, leucine and isoleucine biosynthesis	0.2101
Glycolysis / Gluconeogenesis	0.2097
2-Oxocarboxylic acid metabolism	0.1877
Valine, leucine and isoleucine degradation	0.1823
Tyrosine metabolism	0.1814
Carbon metabolism	0.1808
Citrate cycle (TCA cycle)	0.1640
Butanoate metabolism	0.1590
Glutathione metabolism	0.1540
Pyruvate metabolism	0.1509
Biosynthesis of amino acids	0.1451
Cysteine and methionine metabolism	0.1387
ABC transporters	0.1382
Phenylalanine, tyrosine and tryptophan biosynthesis	0.1342
Propanoate metabolism	0.1212
Glycerophospholipid metabolism	0.1205
Ascorbate and aldarate metabolism	0.1170
Arginine biosynthesis	0.1109
Alanine, aspartate and glutamate metabolism	0.1021
Pyrimidine metabolism	0.0619
Histidine metabolism	0.0615
D-Amino acid metabolism	0.0614
Taurine and hypotaurine metabolism	0.0497
beta-Alanine metabolism	0.0282

Figure 6-1. Heatmap of pruned pathways based on their enrichment score. *Metabolic pathways related to sugar metabolism showed higher scores whereas pathways related to amino acids showed low scores. These findings contradict the biological expectations given that [$^{13}C_3$]propionate was used as a tracer, which should be reflected in higher enrichments in the "propanoate metabolism" pathway..*

The discrepancies between the expected and the observed values in the pathway scores can be explained by analyzing the enrichment of the individual metabolites that comprise each pathway. Based on untargeted analysis with SUNDILE, all the metabolites in the six highest scoring pathways were attributable to seven mass spectral features. Given the nature of the biochemical

reactions in these pathways, a single *m/z* value can be attributed to several different metabolites (e.g., glucose/dextrose, mannose, and fructose all share the same molecular mass and can be found within the same pathways). These associations are shown in Figure 6-2.

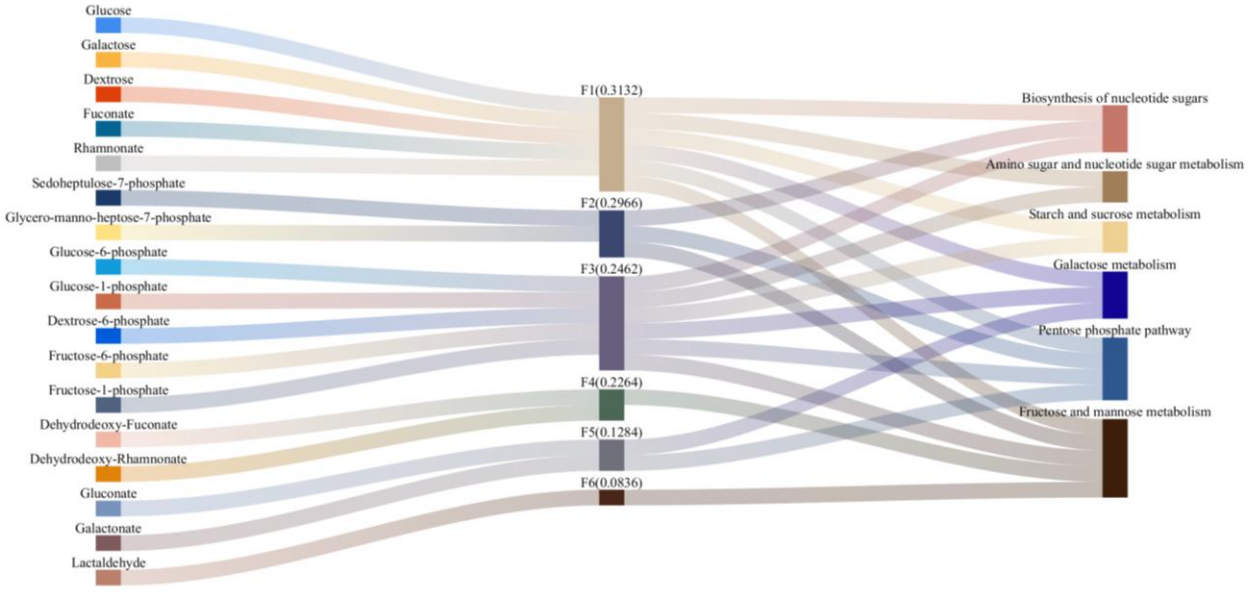


Figure 6-2. Associations among metabolites, scored features (F), and pathways. *The Sankey plot shows the relationships between 17 metabolite identities proposed by SUNDILE, which map to 7 independent features, and the six most highly scored pathways. The enrichment score of each feature is shown in parentheses.*

Analysis of MS^2 data with MS-DIAL identified F1, F3, F4, and F5 as glucose, glucose-6-phosphate, 3-hydroxy-3-methylglutarate, and gluconate, respectively. To assess the accuracy of the score estimates, a targeted integration of these features was performed with PIRAMID, and the equivalent pathway enrichment scores were calculated using the same method that SUNDILE uses to calculate the scores based on the individual metabolite enrichments. The results are shown in Figure 6-3.

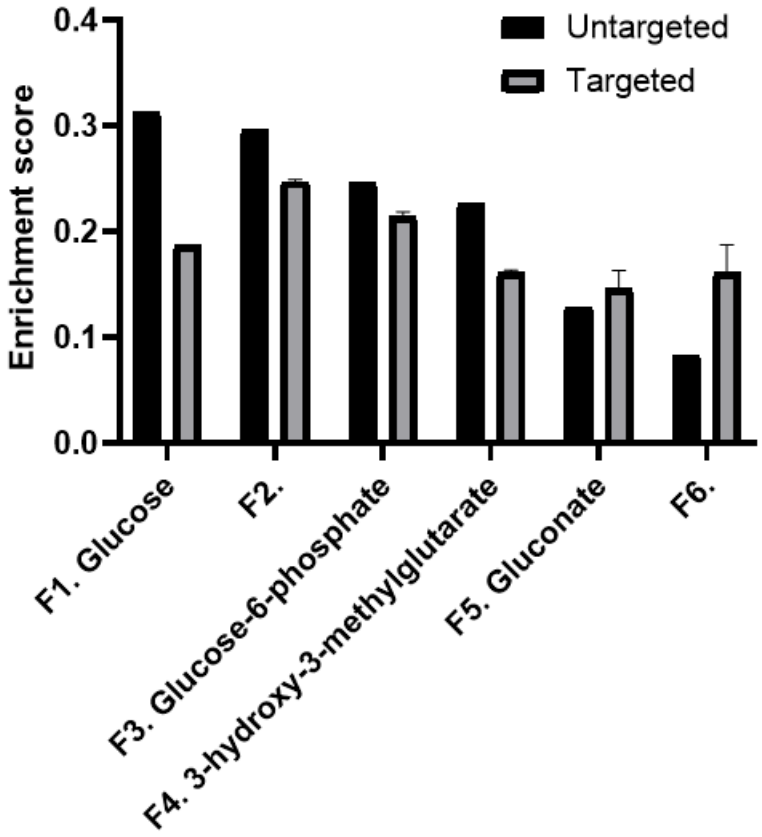


Figure 6-3. Enrichment score comparison across the features found in the highest scoring pathways*. There were significant differences in the scores determined by the untargeted and targeted analysis for the six most common features. Features F2 and F6 were not identified by the MS^2 analysis.*

6.3.1.2 Targeted pathway and compound labeling quantification using PIRAMID

Based on untargeted analysis with SUNDILE, notably higher scores were obtained in four out of the six cases re-examined with PIRAMID. Specifically for features F1 and F3, which are present in all the top-scoring pathways, the untargeted metabolite enrichment scores exhibited an overestimation of the enrichment in comparison to the actual targeted scores. These inflated metabolite enrichment scores subsequently had a cascading effect on the pathway scores. Furthermore, 3-hydroxy-3-methylglutarate, which is an acetyl-CoA derivative belonging to pathways immediately adjacent to the TCA cycle [330], was mislabeled by SUNDILE as a sugar associated to the "Fructose and mannose metabolism" pathway due to the lack of associations

between the compound and any pathway in KEGG. As acetyl-CoA is only a few reactions away from propionate, it is expected that the enrichment of its derivatives would be relatively higher than others that lie further from the tracer. The misidentification of 3-hydroxy-3-methylglutarate by SUNDILE, exacerbated by the overestimation of its enrichment as shown in Figure 6-3, led to an elevated activity score for the "Fructose and mannose metabolism" pathway.

The nature of the overestimation of these values is not attributed to the enrichment quantification by SUNDILE, but instead it stemmed from the peak extraction step in the pre-processing phase, which is conducted before the application of SUNDILE. Manual examination of the peaks showed wide, high intensity peaks for glucose and all of its mass isotopomers, as shown in Figure 6-4. Although the peaks appeared noisy, they were free from interference or obvious shape defects.

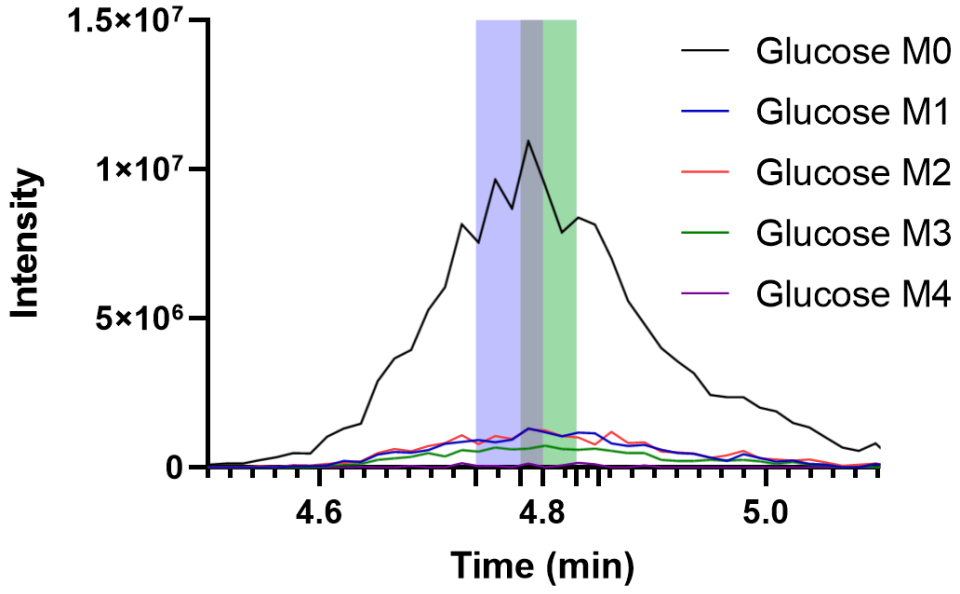


Figure 6-4. Glucose mass isotopomers and their corresponding extraction window in the pre-processing step using XCMS before the untargeted analysis. *All mass isotopomers were quantified using different extraction windows, which skewed the intensities and could have biased the analysis of isotope enrichment scores.*

The reported peak apex for all mass isotopomers was consistently observed at 4.8 ± 0.01 minutes. However, the edges of the ion chromatograms calculated by XCMS exhibited irregular extraction windows, leading to a distortion in the reported intensity values subsequently used in the

untargeted analysis. This phenomenon can be attributed to the inherent noise within the peaks, which hampered the ability of XCMS to accurately identify the edges during the peak-picking step. A similar trend can be observed in other metabolites in Table 6-A2, where a single compound is extracted as two different peaks with the same *m/z* value and similar retention times. Multiple efforts were made to address this issue by modifying the parameters in the XCMS peak processing step such as adjusting the peak-picking algorithm (using the MatchedFilter or CentWave methods), the peak width parameter, and the signal/noise acceptable threshold. However, all of these changes resulted in unfavorable consequences, adversely impacting other metabolites to a greater extent.

The exploration of other actively labeled pathways was performed based on the detection of metabolites that were subsequently identified by MS^2 analysis. Nevertheless, the metabolite matches within these pathways often shared common precursors with other metabolic routes. For instance, pyruvate serves as an intermediate metabolite linking both the glycolysis and the TCA cycle pathways, and it is also a precursor to a diverse array of amino acids that interact with multiple downstream pathways. Targeted analysis of the isotopically enriched metabolites in these pathways is summarized in Figure 6-5, and a reconstruction of the metabolic network based on these results is presented in Figure 6-6.

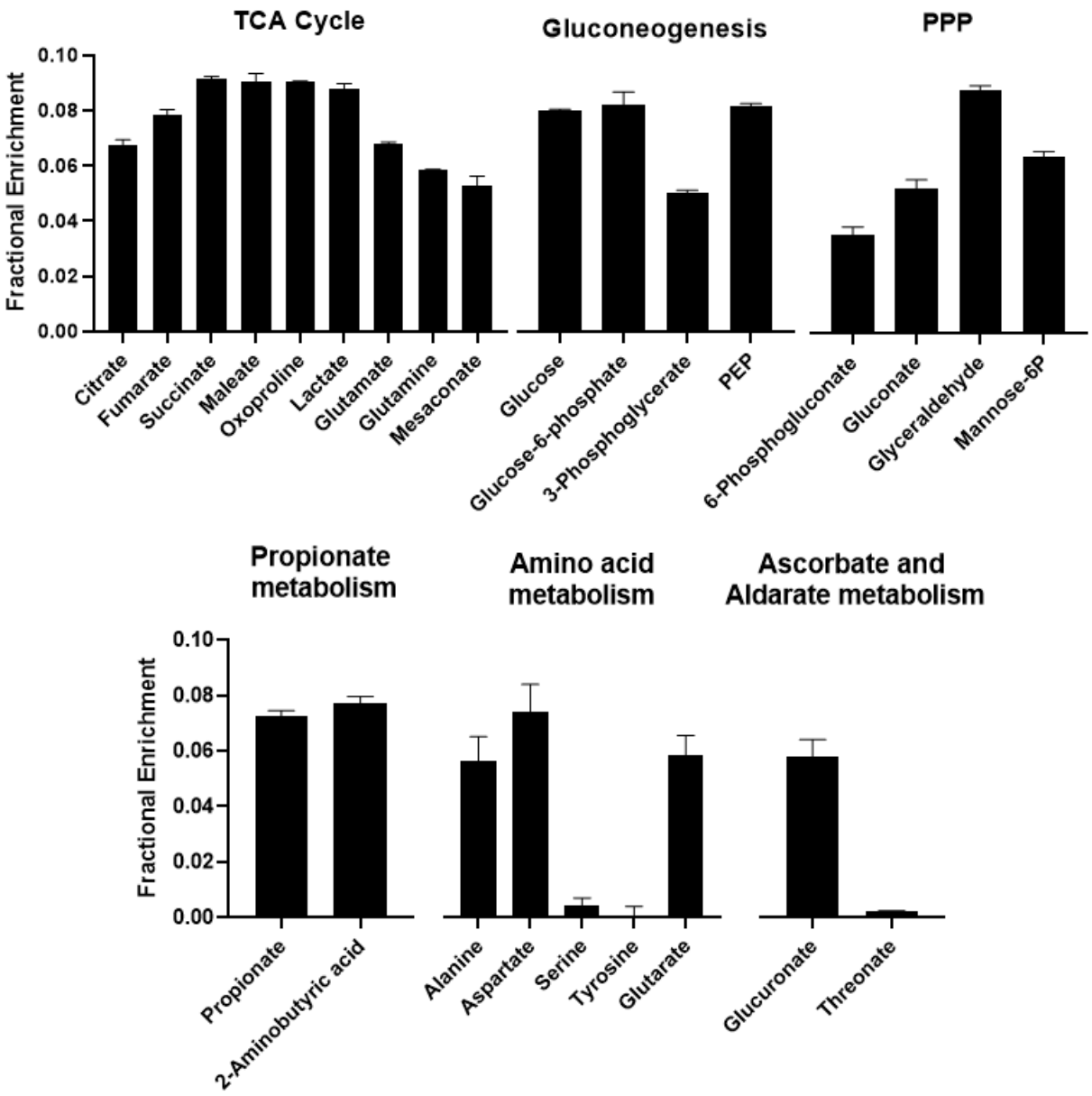


Figure 6-5. Fractional enrichment of the MS^2-confirmed metabolites in the fasted Mc4r$^{-/-}$ mouse dataset. *The upper row shows metabolites in the classic TCA cycle, glycolytic/gluconeogenic, and pentose phosphate pathways. Analogously, the bottom row shows some adjacent pathways and the labeled metabolites that were found within them. The metabolites comprising the "amino acid metabolism" and "ascorbate and aldarate metabolism" pathways show variable levels of enrichment. PEP=phosphoenolpyruvate.*

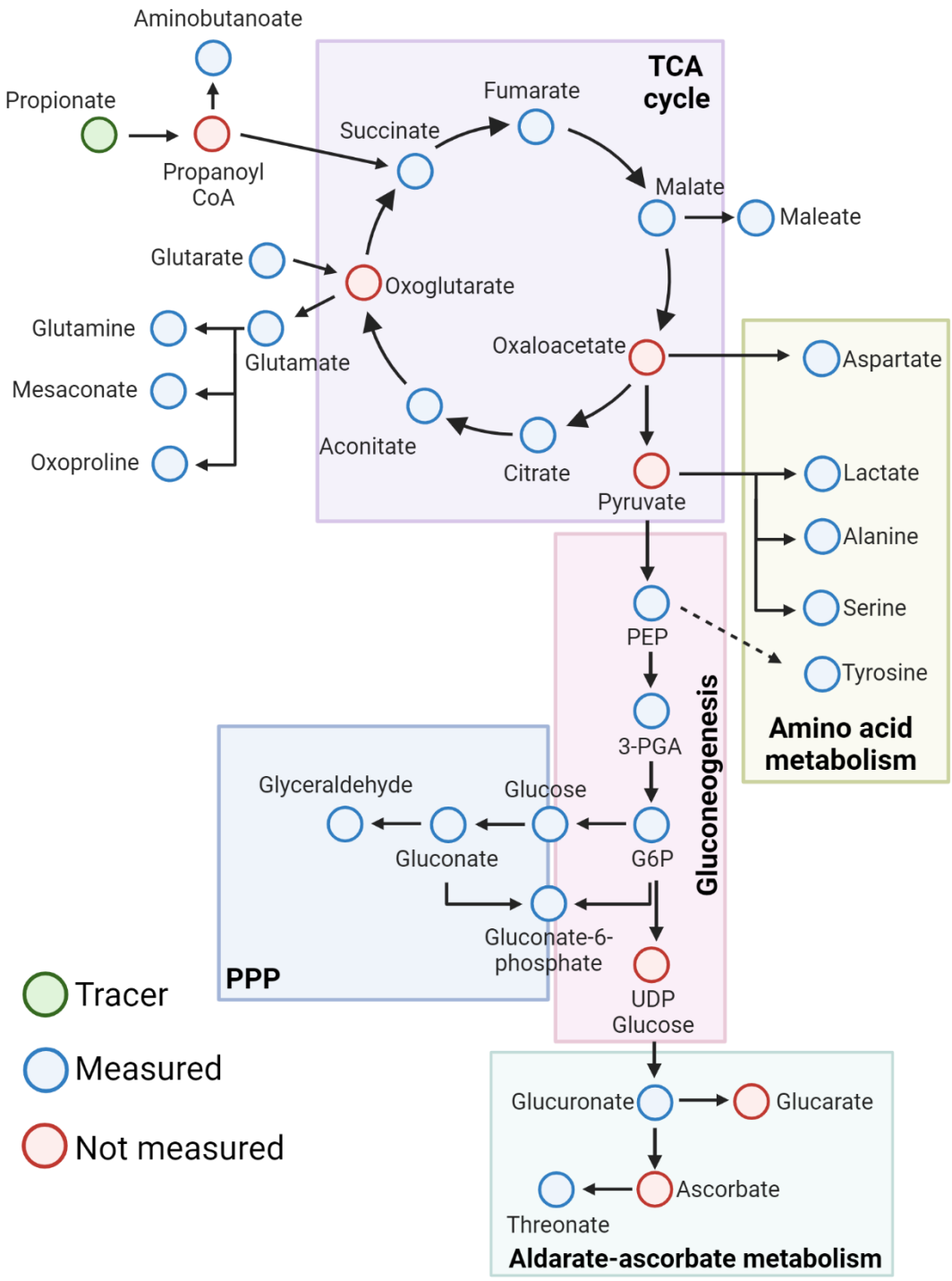


Figure 6-6. Reconstructed metabolic network based on labeled metabolites detected in the livers of overnight-fasted Mc4r$^{-/-}$ mice. *The gluconeogenic pathway stemming from the TCA cycle and branching into the pentose phosphate pathway, amino acid metabolism, and the aldarate and ascorbate metabolism pathways can be mapped into a single interconnected network. The solid lines represent chemical reactions that are reported in mammals. The dotted line represents a reaction that is not active in mammals, explaining the lack of enrichment in tyrosine. PEP=phosphoenolpyruvate, 3-PGA=3-phosphoglycerate, G6P=glucose-6-phosphate, UDP=uridine diphosphate. PPP=pentose phosphate pathway.*

6.3.1.3 Analysis of isotope enrichment in canonical pathways of glucose and energy metabolism

Metabolites within the TCA cycle—namely citrate, fumarate, and succinate—exhibited a comparable degree of enrichment. However, subtle distinctions in their enrichment levels, which reflect a lack of isotopic equilibration among these intermediates, can be ascribed to cellular/subcellular compartmentalization or potential isotope dilutions stemming from catabolism of unlabeled carbon sources. The direct conversion of propionate into succinate explains why succinate has the highest enrichment in the TCA cycle. Furthermore, citrate enrichment can be diluted due to unlabeled carbon entering from fatty acid catabolism or amino acid degradation pathways, which might contribute to its lower enrichment.

Certain metabolites can serve as effective proxies for estimating the enrichment of their upstream precursors. For instance, maleate can be used as a proxy for assessing the enrichment of malate, while lactate can offer insights into the labeling of pyruvate, given the rapid equilibration observed among these metabolites [331], [332]. Additionally, the activity of certain enzymes can be inferred from the enrichment patterns of pathway intermediates. This is the case of the metabolites that stem from pyruvate, lactate, alanine, and serine. The comparatively high enrichment of lactate presumably reflects a higher activity of lactate dehydrogenase, the enzyme responsible for conversion of pyruvate to lactate [333]. These results are in accordance with previous research that has reported rapid lactate kinetics that yield a quick isotopic equilibration between pyruvate and lactate [334]. This logic can also be extrapolated to serine, which shows a significantly lower enrichment than its upstream precursor 3-phosphoglycerate (3-PGA), leading to the hypothesis

that the enzyme responsible for this reaction (phosphoserine aminotransferase) exhibits low activity.

The gluconeogenic pathway generally exhibits equilibration in isotope enrichment, with the notable exception of 3-PGA, which displays a significantly lower level of enrichment. This deviation can be ascribed to potential dilution effects and/or compartmentalization of the 3-PGA pool (Fig. 6-5). Similarly, the pentose phosphate pathway demonstrates a relatively high degree of enrichment at its entry point, glyceraldehyde, but exhibits lower enrichment levels in downstream metabolites, suggesting entry of unlabeled carbon at subsequent steps in the pathway. Noteworthy, glyceraldehyde is not typically considered part of the pentose phosphate pathway, but in certain organisms, glyceraldehyde serves as a bridge that connects some sugar phosphates and glycerate-2-phosphate [335]. As SUNDILE's algorithm does not use organism-specific information in the construction of pathways, this glyceraldehyde was included in this pathway.

A few specific amino acids exhibited relatively high enrichment scores in the study. Notably, 2-aminobutyric acid, aspartate, and alanine demonstrated high enrichments. While aspartate and alanine are produced directly from the central carbon metabolites oxaloacetate and pyruvate, respectively, 2-aminobutyric acid likely derives from microbial origins. It is worth noting that serine and tyrosine, both conditionally essential amino acids, displayed low enrichment levels. This phenomenon was in line with expectations that serine biosynthesis occurs predominantly in the kidneys [336] and that animals do not express the complete pathway to synthesize tyrosine, which is normally only produced de novo by bacteria, other microorganisms, and plants [258].

6.3.1.4 Analysis of isotope enrichment in non-canonical metabolic pathways

One intriguing finding was the relatively high enrichment of glutarate. In mammals, glutarate is typically produced as a byproduct of the catabolism of essential amino acids [337], [338]. Partial reversibility in these catabolic reactions is a possible explanation for glutarate labeling, but no evidence to support this hypothesis has been documented. Alternative metabolic routes leading to glutarate via the TCA cycle intermediate α-ketoglutarate have been documented in prokaryotes [339], [340]. This finding raises the possibility of metabolite exchange between mouse liver and gut microbes. To explore the possibility of misidentification by the MS^2 analysis, we investigated other potential metabolite candidates for the labeled compound. Aceto-lactic acid, dimethyl malonate, and hydroxy-2-oxopentanoaic acid emerged as candidates; however, no known routes within mammalian metabolism could account for the formation of these metabolites, which supports the intriguing notion of metabolite exchange with gut microbes as a potential mechanism at play in this context.

The “ascorbate and aldarate metabolism” pathway exhibited a pattern of variable enrichment and a disconnection between its intermediate steps. As shown in Figure 6-6, the first metabolite connecting this pathway with gluconeogenesis (glucuronate) displayed a relatively high enrichment. However, its metabolic byproduct, threonate, exhibited no enrichment. This discrepancy might be explained by the diversion of flux into alternative byproducts, such as glucarate.

Despite the lack of measurements of metabolites to corroborate all the proposed hypotheses, this study provided valuable insights into previously unexplored pathways. For example, substantial enrichments were observed among several glutamate-derived metabolites including oxoproline, glutamine, and mesaconate Overall, the reconstruction of this network and the measurement of the enrichments of the metabolites within it opens the door for the resolution of fluxes outside of core

metabolism via model-based MFA. Encouraged by these findings, we sought to further test and replicate the workflow in the context of a hyperinsulinemic-euglycemic clamp that reflects metabolism in the fed state.

6.3.2 Application of the developed untargeted and targeted tools in an *in vivo* hyperinsulinemic-euglycemic clamp reveals metabolic impacts of diet-induced obesity

Liver and kidney tissue samples were obtained from control and DIO mice infused with [$^{13}C_6$]glucose during the course of a hyperinsulinemic-euglycemic clamp and subsequently analyzed by HRMS in positive and negative acquisition modes. Prior to study, the mice were fasted for five hours to place the liver into a postabsorptive metabolic state. Untargeted analysis of metabolic pathway activity was followed by compound identification and targeted analysis of metabolite ^{13}C enrichments as described in the subsections below.

6.3.2.1 Untargeted metabolic activity-based clustering of liver and kidney datasets using SUNDILE

SUNDILE was first applied to detect active metabolic pathways and identify isotopically labeled metabolites in liver and kidney samples. Following data processing with SUNDILE, a hierarchical clustering analysis (HCA) was conducted based on the enrichment scores of metabolic pathways to compare the effects of diet and tissue type across all samples. The outcomes of this analysis are presented in Figure 6-7, with emphasis on the first 5 clusters of the pathway analysis. The individual scores for each tissue, diet, and acquisition mode are presented in Table 6-A3. Additionally, a table with the number of metabolites that were found on each pathway is presented in Table 6-A4.

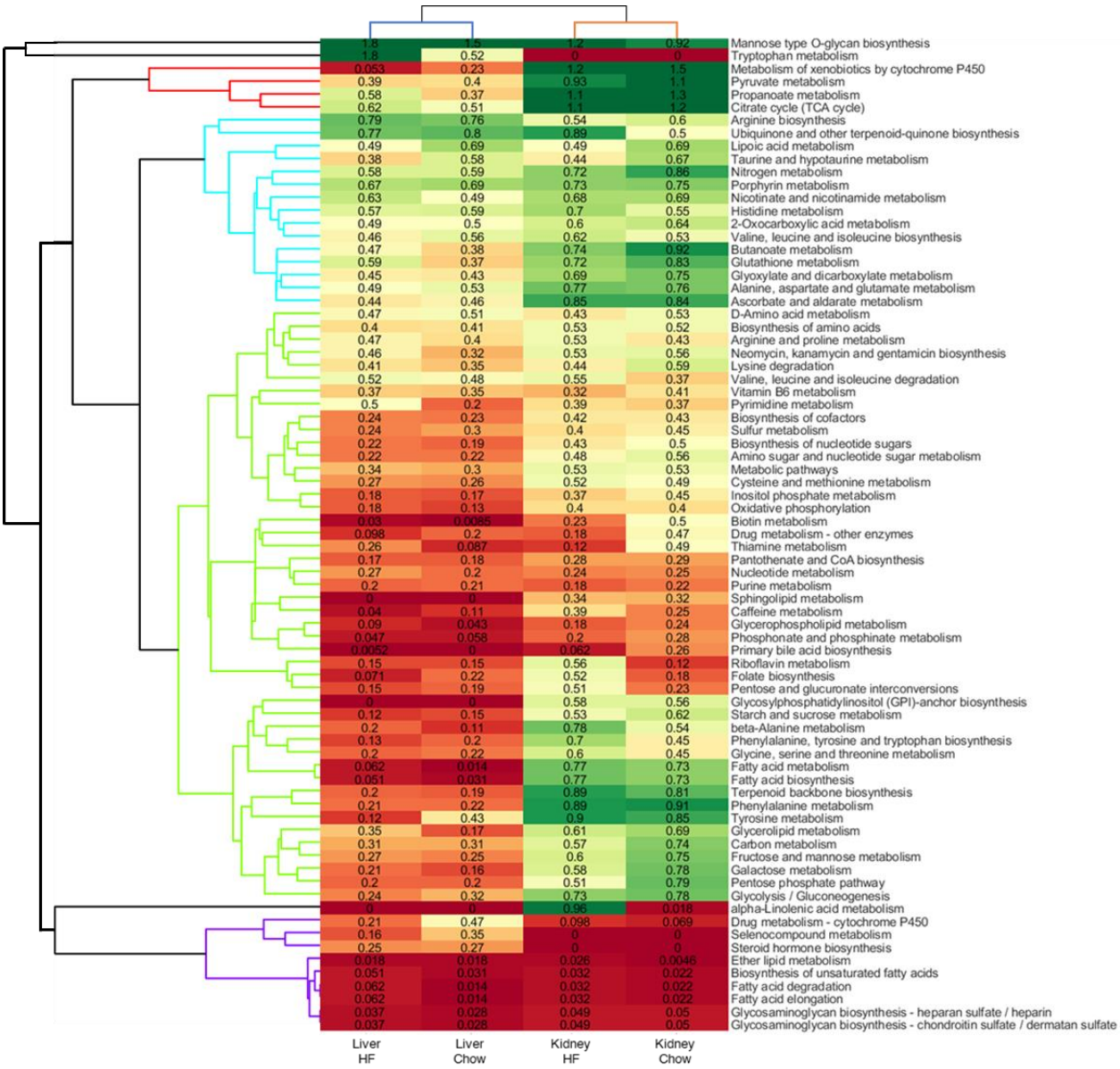


Figure 6-7. Hierarchical clustering analysis of the enrichment scores calculated from the untargeted analysis of kidney and liver samples collected from chow and high-fat fed mice. *The first five clusters are separated in the pathway dimension, indicated by colored branches on the left of the clustergram. The optimal number of clusters was determined by the maximization of the ratio of variances between clusters.*

HCA indicates that the pathway activities are predominantly influenced by the physiological characteristics of the tissue type more so than the animal's diet. This observation aligns with prior research findings that have consistently demonstrated that, under similar disease conditions, an organ's tissue-specific physiology dictates its metabolic activity [341]. On the other hand, the basal clades (represented by the upper black branches in Figure 6-7) give rise to a smaller cluster,

composed by the “mannose type-O glycan biosynthesis” and “tryptophan metabolism” pathways, that is characterized by exceptionally high scores observed in the liver. These elevated scores can be attributed to individual metabolites with remarkably high scores, resembling the phenomenon previously described in the fasted *Mc4r*$^{-/-}$ mouse dataset. However, in this particular instance, the “mannose type-O glycan biosynthesis” pathway score is determined by a single highly enriched metabolite, identified as cytidine triphosphate. The origin of this outlier data point can be attributed to poor extraction of the low intensity peaks in the preprocessing step, similar to the case shown in Figure 6-4. When the relative abundance of the extracted peaks grouped by SUNDILE is compared against the targeted integration results obtained from PIRAMID, significant discrepancies occur in the first two mass isotopomers, as seen in Figure 6-8. As the enrichment calculation in SUNDILE is based on the M0 abundance, the quantification of this metabolite results in a significant overestimation of enrichment that biases the results. In turn, when only this single metabolite is found in a pathway, the enrichment score of the entire pathway is equal to the enrichment score of the metabolite itself, opening the door for biased activity estimates.

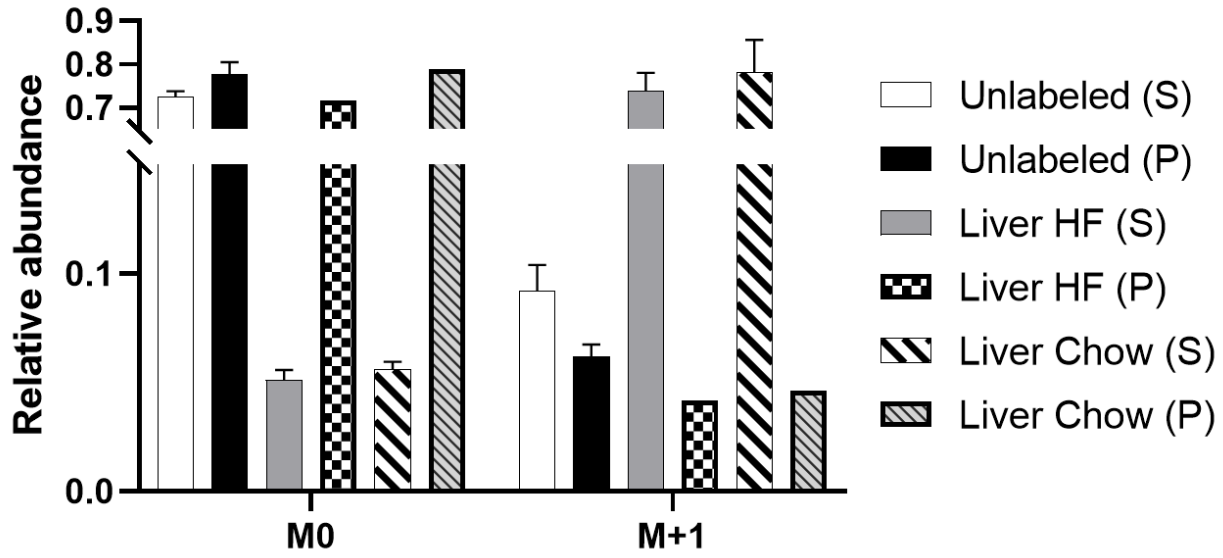


Figure 6-8. Comparison of the relative abundance of the two first isotopologues of the feature assigned to cytidine triphosphate using SUNDILE (S) and PIRAMID (P). *In all labeled samples, the extraction of the peaks by XCMS resulted in significantly lower values for the M0 isotopologue, which overestimates the enrichment score of the metabolite and its associated pathways in SUNDILE.*

Within the subsequent set of clades, the purple cluster in Figure 6-7 contains pathways linked to polysaccharides, fatty acids, and hormones. What unites the pathways within this cluster is the consistently low activity observed across both diets and both organs. This outcome aligns with expectations, as it can be attributed to the protracted turnover rates of metabolites within these pathways and the multistep polymerization processes involved in the synthesis of these macromolecules.

The red cluster contains pathways closely associated with the TCA cycle. For instance, the propanoate metabolism pathway consists of multiple coenzyme-A derivatives capable of conversion into precursors that can enter the TCA cycle, such as succinyl-CoA, and acetyl-CoA. Similarly, the pyruvate metabolism pathway encompasses the diverse fates of pyruvate, which can serve as both a precursor and a product of the TCA cycle. This cluster demonstrates notably high activity in the kidney, whereas it was assigned relatively lower scores in the liver.

The cyan cluster includes several pathways linked to the urea cycle and to glycolytic intermediates. Within this cluster, pathways associated with the biosynthesis of amino acids arising from the urea cycle (e.g., arginine, proline, and ornithine) and from a 3-PGA (serine, cysteine, and glycine) are featured. Additionally, byproducts originating from the glutamate/glutamine module, which ultimately feed into nitrogen-based pathways such as "nicotinate and nicotinamide metabolism," "nitrogen metabolism," and "glutathione metabolism" are enriched, despite the absence of a nitrogen-based tracer in the experiment. Across this cluster, relatively high enrichment scores can be found in the kidney with slightly lower scores in the liver.

Finally, the green cluster is characterized by consistently low scores observed across the liver samples. Notably, this cluster can be further subdivided into two distinct sub-clusters, as determined by HCA. The first sub-cluster exhibits moderate to low scores in the kidney samples and encompasses various metabolites that are not absolutely required for the survival of the organism, pathways related to the biosynthesis of several amino acids derived from pyruvate or the TCA cycle are also included within this cluster. The second sub-cluster, in contrast, exhibits relatively high scores in the kidney samples and is associated with the metabolism of sugars. This sub-cluster encompasses pathways such as glycolysis/gluconeogenesis and the pentose phosphate pathway.

6.3.2.2 Analysis of ^{13}C enrichment in target metabolites with PIRAMID reveals sources of variation between positive and negative acquisition modes

Intuitively, the pathways associated with sugar metabolism were expected to have the highest scores in the hyperinsulinemic-euglycemic clamp given that the metabolic tracer was [$^{13}C_6$]glucose. Despite relatively high scores in glucose metabolism of the kidney, unexpectedly low scores were found in the liver. Furthermore, higher scores in downstream pathways (such as

the TCA cycle) were also inconsistent with the expectations of the experiment. To understand the origin of these inconsistencies, a targeted analysis of metabolite ^{13}C enrichments in selected pathways was performed.

The calculation of the pathway scores is based on the mean of non-zero values obtained from both positive and negative acquisition modes. However, it is worth noting that in several instances, the scores for the same pathway can significantly differ between these acquisition modes, potentially introducing score distortions. Additionally, as illustrated in Table 6-A4, the number of identified metabolites within each pathway can vary across acquisition modes, diets, and organs. Given the possible existence of "cold pools" (i.e., metabolites that exhibit low enrichment due to compartmentalization or metabolic channeling [51], [342], [343]), pathway scores may vary markedly depending on which identified metabolites were included in the scoring function. Figure 6-9 shows some examples of pathways exhibiting divergent scores under different acquisition modes, highlighting the need for further interpretation and consideration of these variations in pathway scores.

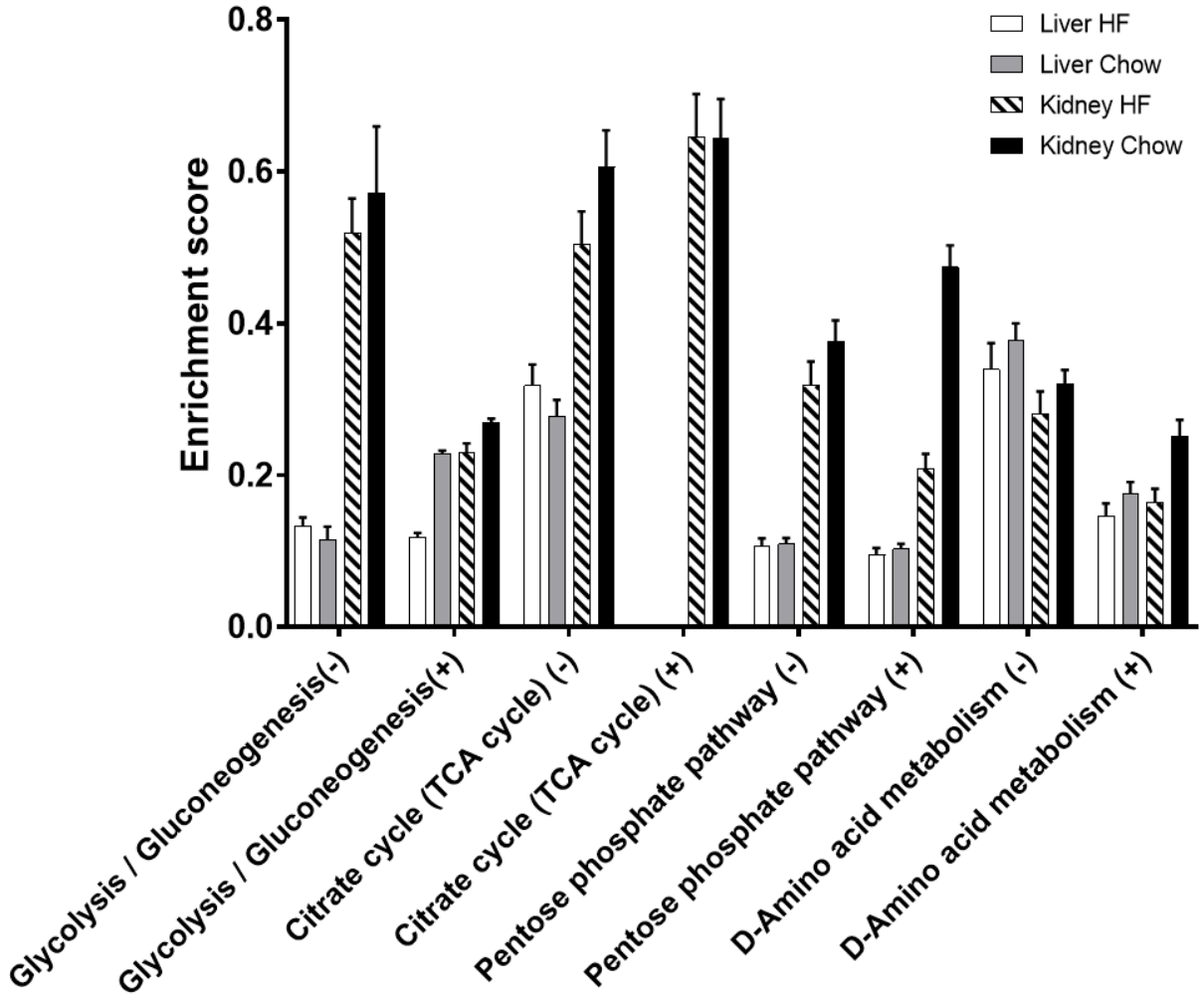


Figure 6-9. Enrichment score of a selected group of pathways calculated from the untargeted analysis in positive (+) and negative (-) acquisition modes. *Certain conditions show significant differences between the acquisition modes, which biases the overall score of the pathway.*

Arguably, one of the pathways in the experiment that was anticipated to be most enriched is the glycolytic/gluconeogenic pathway, since fasting conditions stimulate active gluconeogenesis in both liver and kidney. However, SUNDILE determined different enrichment scores for this pathway depending on whether positive or negative mode datasets were analyzed. Specifically, while the Liver HF scores exhibited consistency between acquisition modes, the Liver Chow scores and kidney scores from both diets exhibited distinct results without any clear correlation between scores based on positive versus negative mode data. This striking divergence is not

exclusive to the glycolytic/gluconeogenic pathway but is similarly observed across numerous other pathways. The underlying cause of this phenomenon can be attributed to the peak extraction discrepancies discussed in the previous section. Furthermore, this effect is also perceptible in the fractional ^{13}C enrichments of selected target metabolites determined by PIRAMID, albeit to a lesser degree, as depicted in Figure 6-10.

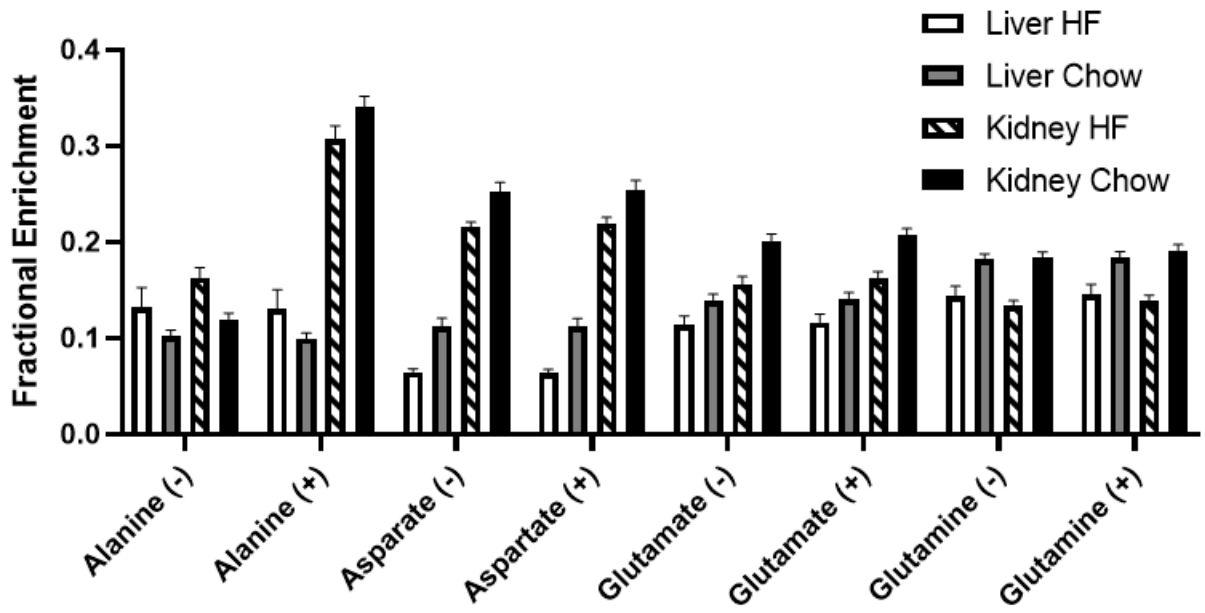


Figure 6-10. Fractional ^{13}C enrichment of a selected group of metabolites in the amino acid biosynthesis pathway calculated with PIRAMID using positive (+) and negative (-) acquisition mode datasets. *Alanine shows significant differences between the acquisition modes, which biases the overall enrichment score of the pathway.*

In the targeted integration of metabolites within the D-amino acid biosynthesis pathway, the fractional enrichments of most metabolites exhibit a consistent pattern between the positive and negative acquisition modes. However, a notable exception arises with alanine, where significant differences were observed in the kidney samples. These differences were attributed to interferences from other ions biasing the measurements of the MID of alanine.

6.3.2.3 Analysis of isotope enrichment in selected central metabolic pathways

With the aforementioned considerations in mind, a targeted analysis of selected pathways that showed differences in activity between the experimental conditions was performed. The results of the targeted integration of the MS^2-confirmed metabolites in three relevant pathways (i.e., the pentose phosphate pathway, urea cycle, and nucleotide metabolism) are shown in Figure 6-11.

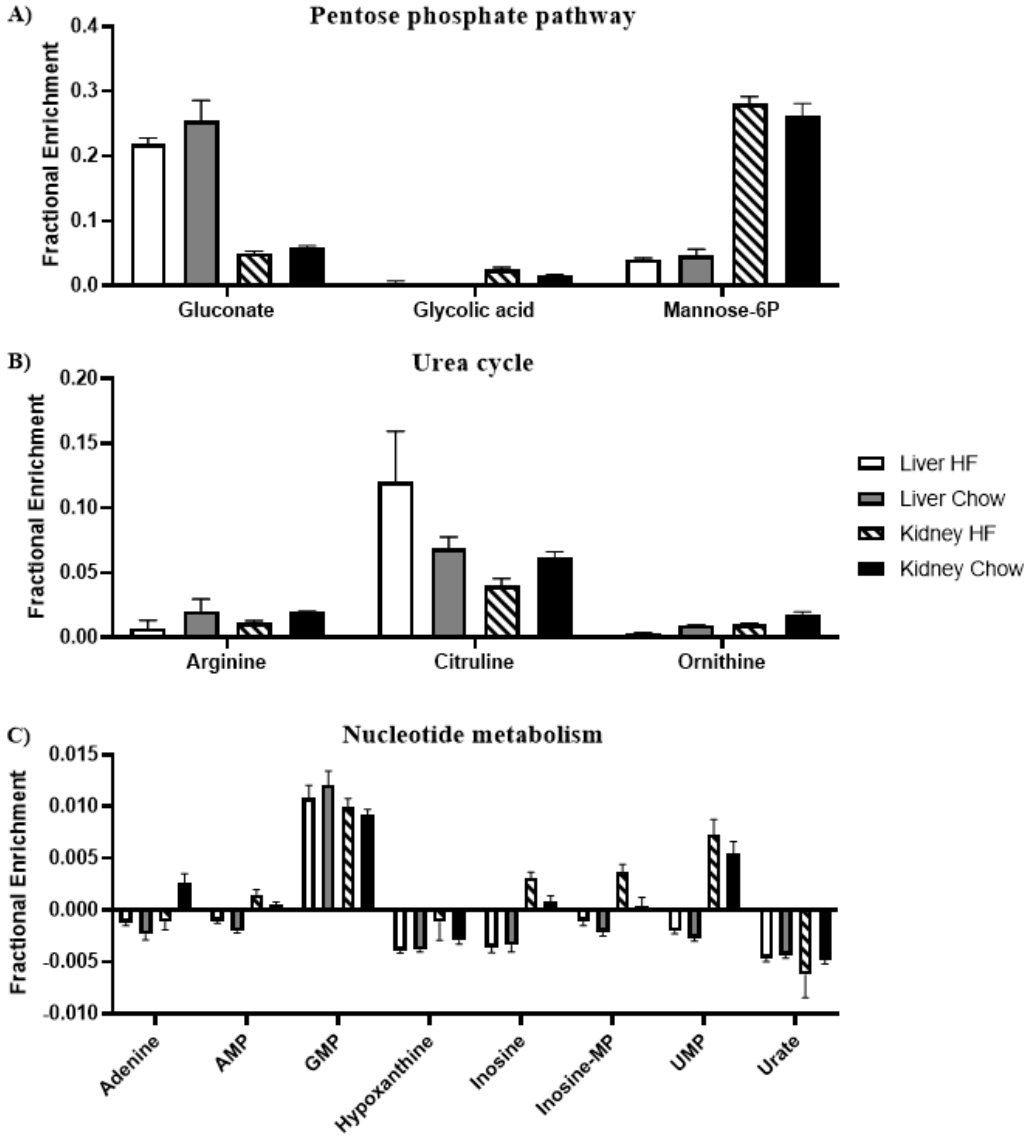


Figure 6-11. Fractional enrichment of the confirmed metabolites in multiple pathways: *A) Pentose phosphate pathway, B) Urea cycle, C) Nucleotide metabolism.*

6.3.2.3.1 Pentose phosphate pathway

The pentose phosphate pathway (Fig. 6-11A) exhibits contrasting behaviors among its metabolites. For instance, while gluconate demonstrates high enrichments within the liver of mice on both diets, it low gluconate enrichments were measured in the kidney. The role of this metabolite in the glycolytic and pentose phosphate pathway is not completely understood. Gluconate has been identified as an alternative route to convert glucose to 6-phosphogluconate and into the pentose phosphate pathway [344] following the reactions shown in Figure 6-12.

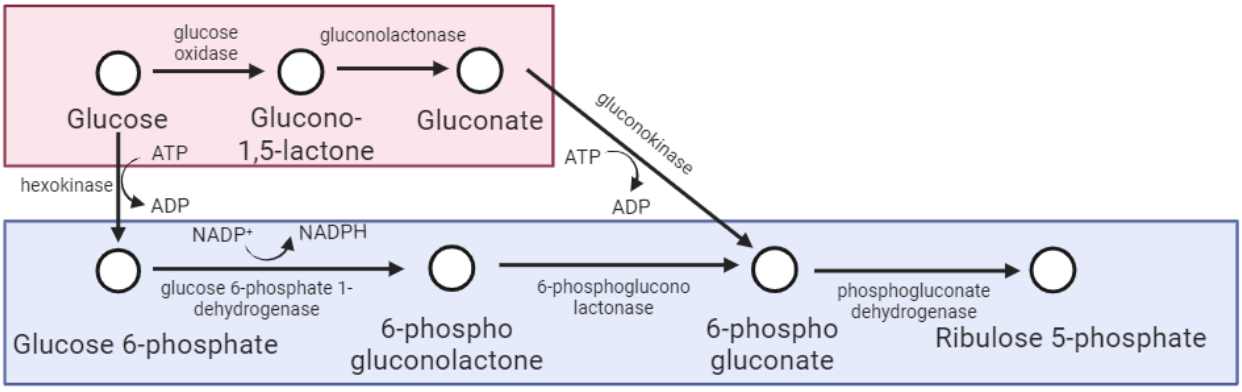


Figure 6-12. Alternative routes for the conversion of glucose into ribulose 5-phosphate*. Red: Alternative route using gluconate as an intermediate. Blue: Canonical route of the pentose phosphate pathway.*

Both routes involve the consumption of energy through the conversion of ATP to ADP, but the canonical route additionally yields NADPH, crucial for *de novo* lipogenesis. The liver is recognized as a major site of *de novo* lipogenesis [345], and there exists a correlation between elevated glucose availability and increased lipid production [346]. It is possible that both routes are active in the liver. Hence, the higher enrichment of gluconate in the liver could be an indicator of high activity of the pentose phosphate pathway which provides useful cofactors for *de novo* lipogenesis.

In contrast, mannose-6-phosphate exhibits the opposite pattern, with significantly lower enrichments in the liver compared to the kidney. Previous research has indicated that glycogen accumulation is highly upregulated in the liver of dogs undergoing a hyperinsulinemic-euglycemic clamp [347]. This upregulation of glycogen synthesis might divert the labeling from the ^{13}C-glucose tracer away from competing reactions at the G6P node (Fig. 6-13). This result is expected given that it has been reported that glycogen accumulation in the kidney is typically negligible except in prolonged hyperglycemic conditions [348], [349].

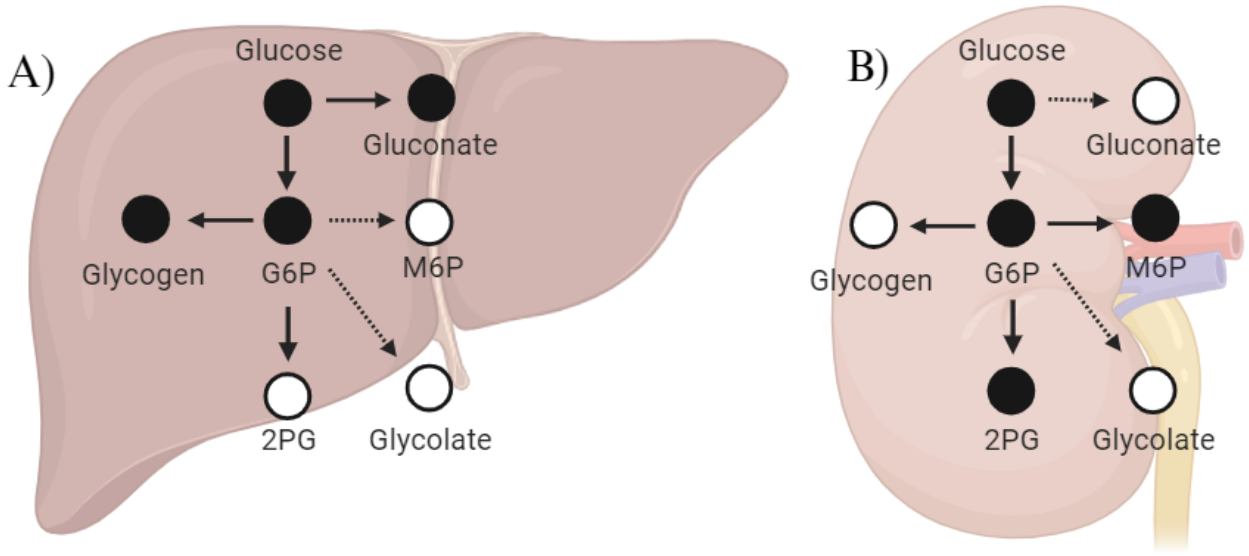


Figure 6-13. Proposed pathway explaining the differences between the enrichments of the labeled (filled) metabolites within the pentose phosphate pathway in different organs. *A) Liver. B) Kidney. The enrichment of the pentose phosphate pathway shows different trends between the liver and the kidney at the G6P node. High glycogen accumulation has only been observed in the kidney under prolonged hyperglycemia. Hence glycogen synthesis is not expected to be elevated in the kidney.*

The observed differences in pathway activation between each organ may reflect distinct physiological traits associated with each tissue type. At the cellular level, normal renal physiology heavily relies on the proper functioning of lysosomes. Lysosomes play a critical role in the selective catabolism of low molecular weight proteins that pass through the glomerular filter into the lumen of the nephrons. These proteins are subsequently endocytosed by the cells of the proximal convoluted tubules and degraded within lysosomes [350], [351]. The degradation of

these proteins requires hydrolases that must be tagged with mannose-6-phosphate before being transported to the endosomal/lysosomal compartment [352]. Following this logic, it is expected that renal cells exhibit an upregulation of pathways related to mannose-6-phosphate metabolism, which does not appear to be impacted by diet.

Clinically, the principal biomarkers used to predict the progression of diabetic kidney disease are albuminuria and the glomerular filtration rate. However, not all cases of diabetic kidney disease are accompanied by albuminuria [353], [354], creating a need to find other biomarkers of disease progression [355]. Given that the effects of metabolic diseases such as diabetes in the proximal tubules has shown higher glucose absorption under hyperglycemic conditions [356], our results suggest that M6P in the kidney could be used as a possible biomarker linking a metabolic alteration in the kidney to hyperglycemia under obesogenic conditions in the future.

6.3.2.3.2 Urea cycle

The metabolites of the urea cycle, as illustrated in Figure 6-11B, exhibited an inconsistent labeling pattern, with only citrulline displaying a moderate level of enrichment across all conditions. The variability in the enrichment within the urea cycle arises due to the carbon atom transitions within the pathway, as depicted in Figure 6-14A.

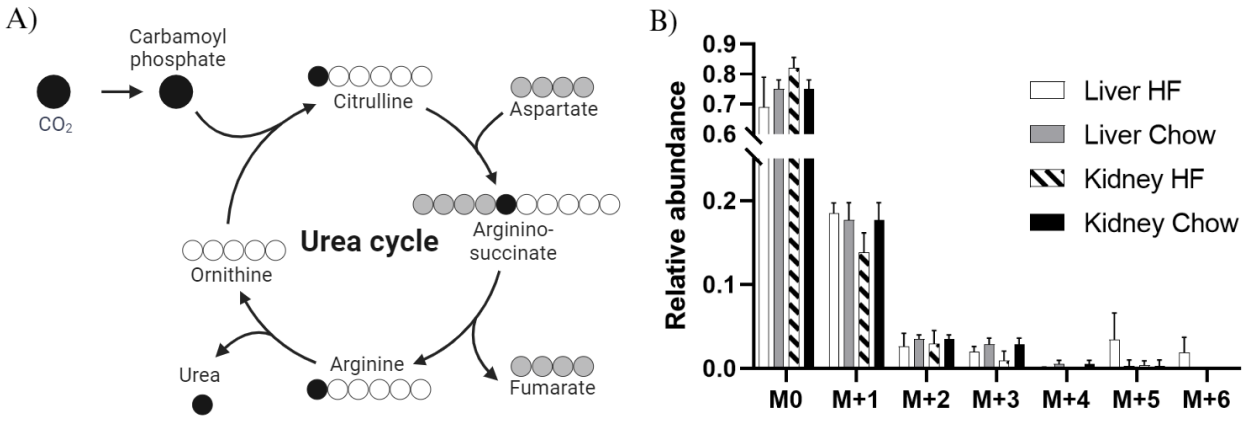


Figure 6-14. Urea metabolism and targeted integration of citrulline, corrected for natural isotope abundance*. A) Carbon transitions in the urea cycle. B) MID of citrulline showing most of the enrichment comes from the M+1 mass isotopomer. Black carbons come from glucose oxidation and from the oxidative decarboxylation of carboxylic acids in the TCA cycle. The resulting CO_2 from these reactions enters the cycle and exits in the biosynthesis of urea. Gray carbons coming from aspartate leave in the form of fumarate and do not add to the labeling of arginine. White carbons are constantly cycled through the urea cycle and are not expected to be labeled in this experiment.*

Notably, the intermediate metabolic steps between citrulline and arginine do not introduce any modifications to the existing enriched carbon atoms originating from citrulline. Hence, it is expected that arginine would exhibit a similar enrichment as citrulline. The absence of enrichment in arginine could be attributed to its potentially higher pool size that would slow down its labeling relative to citrulline. To confirm this hypothesis in a quantitative manner, the sum of the areas of all mass isotopomers (i.e., the total counts) normalized to the area of a standard with known concentration would need to be compared. Unfortunately, this study did not use such a standard. A qualitative approximation can be obtained by analyzing the total counts of both metabolites across all experimental conditions as depicted in Figure 6-15, showing that the total counts of arginine were one order of magnitude above their citrulline counterparts, partially confirming the proposed hypothesis.

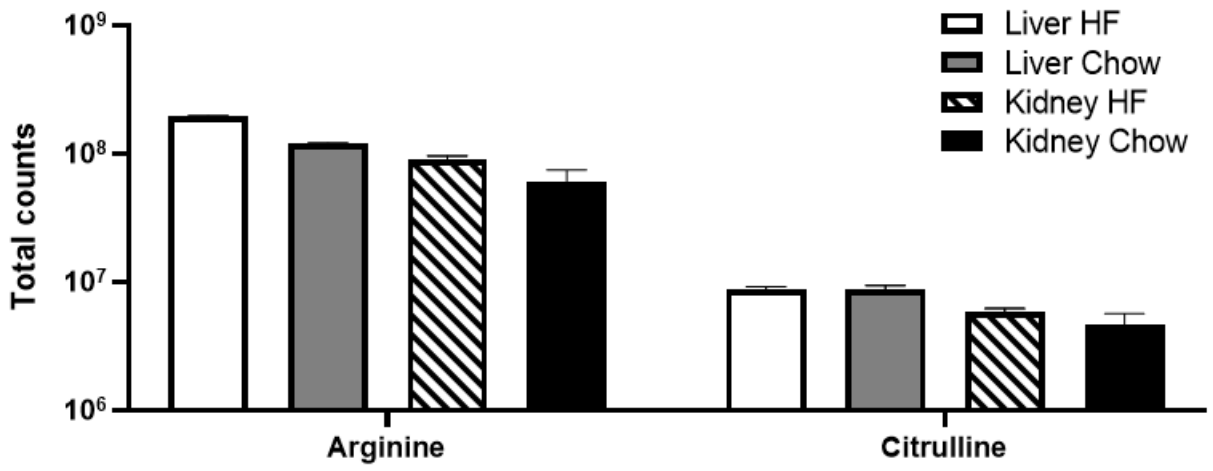


Figure 6-15. Total counts of arginine and citrulline. *For all experimental conditions, arginine exhibits significantly higher counts than citrulline.*

A closer examination of the mass isotopomer distribution (MID) of citrulline, as depicted in Figure 6-14B, reveals that the enrichment predominantly comes from the M+1 mass isotopomer. This observation suggests that the contribution of carbamoyl phosphate is responsible for the observed labeling pattern. Furthermore, the MID of citrulline shows some interesting differences between the analyzed conditions: The M+1 enrichment in the liver was similar under both diets but there was a decrease in the HF condition in the kidney. While ureagenesis is primarily localized to the liver given that it is the only organ that expresses all the enzymes required for the operation of the urea cycle [357], some partial reactions of this pathway have also been documented in the kidney [358], [359]. Diet-dependent differences in the citrulline response were found only in the kidney where high fat diet fed mice showed lower M+1 abundance than their chow diet fed counterparts. This phenomenon might be explained by the overall lower uptake of labeled glucose in the HF group, leading to a lower enrichment of the CO_2 pool that stems from glycolytic and TCA cycle oxidative routes. Additionally, the higher fat content in the HF group introduces unlabeled CO_2 when fatty acids are oxidized, thereby diluting the available CO_2 pool for urea synthesis.

Arginine has been used as a biomarker to assess kidney disease [360], and its administration to diabetic rats has been studied as a therapeutic tool to prevent kidney-disease related conditions such as glomerular hyperfiltration and proteinuria [361]. Given the close relationship between arginine and citrulline, and the physiological differences found in the kidney citrulline labeling patterns across both diets, kidney citrulline could be used as a biomarker of metabolic dysregulation in obese mouse models of kidney disease.

6.3.2.3.3 Nucleotide metabolism

Several pathways related to nucleotides and purines were identified with low labeling in the untargeted HCA as shown in Figure 6-7. Additionally, a targeted analysis was conducted across several metabolites within these pathways, as illustrated in Figure 6-11C. Notably, the targeted analysis revealed no significant enrichment in these metabolites. The difference between the low predicted enrichments in the untargeted analysis and the total lack of enrichment found in the targeted analysis can be explained by the overestimation of the enrichment of certain metabolites in this pathway by SUNDILE. As previously discussed, some metabolites exhibit an overestimation in their enrichment scores due to an inaccurate extraction of their areas by XCMS. A case in point is cytidine triphosphate, illustrated in Figure 6-8, which displayed an exceptionally high enrichment score in the untargeted study that was not corroborated in the targeted study. This phenomenon is relatively common among nucleotides given their relative low intensities and similar masses and retention times that causes errors in the peak extraction step by XCMS. Since these metabolites contribute to the calculation of the "nucleotide metabolism" pathway enrichment score, the overall pathway score is inflated as well. This phenomenon underscores the critical importance of precise peak extraction during the data preprocessing step. Ensuring accurate peak identification and assignment is essential for portraying biological results in a more reliable and

meaningful manner, especially when dealing with metabolites of higher molecular weights and complex isotopologue distributions.

6.3.2.4 Unexpected appearance of ^{13}C in pathways outside of central carbon metabolism

Finally, the targeted integration of certain pathways revealed unexpected isotope enrichment in a few specific metabolites, as shown in Figure 6-16.

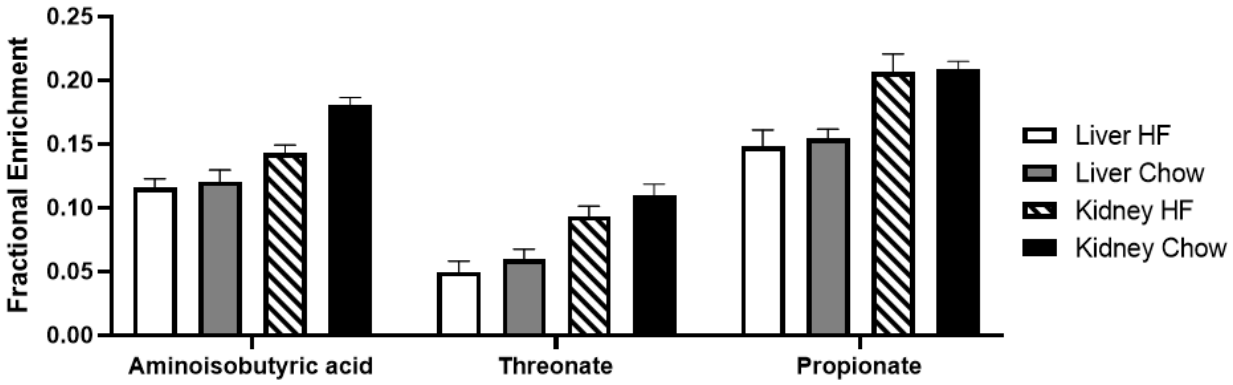


Figure 6-16. Targeted integration of other metabolites of interest. *Secondary metabolites with high enrichment can be correlated to disease models or to previously unexplored metabolic routes.*

Aminobutyric acid is typically produced as a byproduct of thymine [362] and valine [363]. Its relatively high enrichment in the context of this study raises interesting questions, particularly because pyrimidines were not shown to exhibit significant enrichment, and because valine is an essential amino acid that should not be labeled by mammalian metabolism. The possibility of co-metabolism with valine-producing bacteria could be hypothesized as recent research has found key bacteria in the clostridiaceae family that are both present in mammalian intestines [364], [365] and have the ability to biosynthesize valine [366].

Moreover, a correlation between aminobutyric acid concentrations and glucose regulation has been documented. There exists an inverse correlation between aminobutyric acid concentration in plasma and metabolic risk factors. In healthy mice, the metabolite has been found to increase the

expression of brown adipocyte-specific genes in white adipose tissue and the oxidation of fatty acids in hepatocytes, improving glucose homeostasis through the production of leptin [367], [368]. In genetically obese *ob/ob* mice (which exhibit loss of leptin production) treated with high doses of aminobutyric acid, hepatic inflammation was reduced but the increase in fatty acid oxidation did not exhibit a dose-dependent trend, indicating that there was an insensitivity to the protective effects of aminobutyric acid in this model [369].

The analysis presented in Figure 6-16 shows a reduction in the enrichment of aminobutyric acid in the kidneys of high-fat fed mice. Given the established correlation between the presence of this metabolite and the metabolic state of mice under obesogenic conditions, it could be proposed as a potential biomarker that might be useful to gain deeper insights into glucose and fatty acid regulation *in vivo*. Following this logic, reduced enrichment of aminobutyric acid in kidney tissue might be a predictor of poor glucose regulation and can give a snapshot of the ability to oxidize fatty acids.

Threonate was also substantially enriched in both the liver and kidney of mice infused with ^{13}C-glucose. Threonate is a byproduct of the spontaneous breakdown of ascorbate, also known as vitamin C [370]. In certain animals, ascorbate is produced from the metabolism of UDP-glucose, a nucleotide sugar. While direct measurements of the enrichment of this precursor were not obtained, the study identified multiple glycolytic precursors that showed significant levels of enrichment. Additionally, the untargeted analysis showed moderate scores in the “nucleotide sugar metabolism” pathway. These factors suggest that the pathway responsible for the synthesis of ascorbate is likely active in the tissues examined.

Ascorbate has been identified as an inhibitor of visceral obesity and nonalcoholic fatty liver disease in diet-induced obese mice [371]. Moreover, it has been suggested to play a role in the regulation

of glucose metabolism and leptin secretion in isolated adipocytes [372]. However, the precise metabolic mechanisms underpinning these effects are yet to be elucidated. While certain mammals like primates lack the gene that encodes L-gulonolactone oxidase, the enzyme responsible for the last step of ascorbate synthesis from glucose [373], the fate of carbon in precursor metabolites within this pathway still remains an interesting target for further investigation. Consequently, studies of the "aldarate/ascorbate metabolism" pathway could complement analyses of the pentose phosphate and the glycolysis pathways in the context of obesity and diabetes research.

Finally, propionate enrichments in the range of 15-20% were detected in our study. Propionate is recognized as a major microbial fermentation metabolite, and its synthesis stems from glucose-derived pyruvate. This process involves the transformation of pyruvate into propionyl-CoA through microbial routes: i) pyruvate→lactate→lactoyl-CoA→propionyl-CoA and ii) pyruvate→TCA cycle→succinate→succinyl-CoA→propionyl-CoA [374]. The interplay between host metabolism and gut microbiota is a dynamic and intricate relationship. Metabolites are produced by gut microbes and subsequently absorbed from the gut lumen, resulting in associations between the microbiome and plasma metabolites. Simultaneously, metabolites synthesized by the host influence microbial growth and the integrity of the gut barrier, which in turn impacts further metabolite absorption [375], [376]. The extent of co-metabolism between mammals and their gut bacteria is significant, with estimates suggesting that microbiomes are involved in the metabolism of at least 71% of fecal metabolites and 15% of blood metabolites [375]. The high propionate enrichments observed during the hyperinsulinemic/euglycemic clamp reveal a rapid equilibration of propionate and its precursor metabolites between the blood and the gut.

While the enzymes required for the biosynthesis of propionate via succinyl-CoA have been documented in the liver of mice [377], and its direct production in mammalian tissues is

theoretically possible, no evidence has been presented to support this hypothesis. The similar enrichments between succinate and propionate as seen in Figure 6-17 hint at the possibility of the biosynthesis of propionate from succinate.

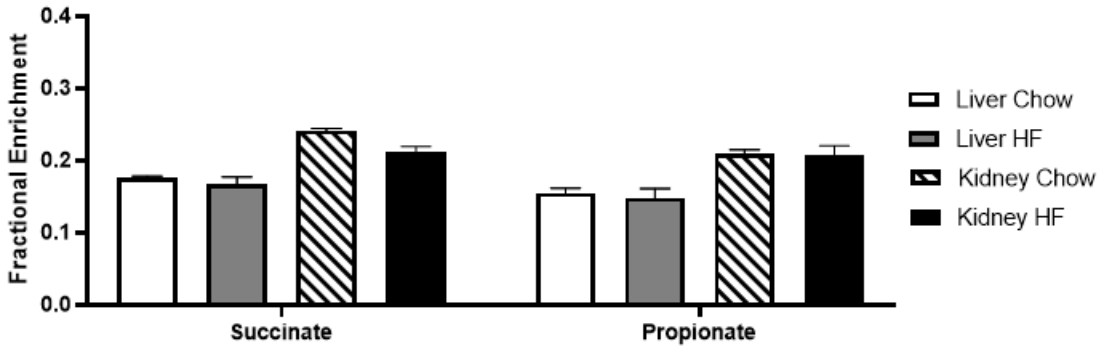


Figure 6-17. Fractional enrichment of succinate and propionate. *The similar patterns exhibited in both metabolites and the previously reported existence of the enzymes required to synthesize propionate from succinate via succinyl-CoA suggest that the biosynthesis of propionate could be feasible in the liver of mice.*

6.4 Discussion

6.4.1 Use of the developed bioinformatics tools in the context of stable isotope-based *in vivo* metabolomics

The use of stable isotope tracers, and specifically ^{13}C, has proven to be a potent tool for quantifying intracellular metabolic fluxes and enhancing our understanding of the physiological state of an organism [378]. A vast majority of the research employing this methodology is conducted in a targeted manner, wherein specific predetermined pathways and metabolites are the primary focus [379], and only a handful of studies have successfully gleaned significant biological insights from fully untargeted approaches [380]. Typically, these untargeted workflows provide at best an approximation of metabolic turnover rates using the labeling kinetics as a proxy for relative fluxes [381] or as a filtering step to distinguish metabolically active compounds from background features in a qualitative manner [382], [383], [384].

With the aid of the previously developed tools SUNDILE and PIRAMID, we propose a new two-step workflow that combines an untargeted approach followed by a targeted approach to quantitatively assess the isotope labeling of pathways and metabolites that are highlighted in the first step. While good concordance was found between the results of both steps of the process (Fig. 6-3), there are some caveats that need to be considered before implementing the workflow.

First, accurate peak extraction in the preprocessing step is crucial for achieving reliable results using SUNDILE. An incorrect extraction of the isotopologue peaks of a single metabolite can result in a snowball effect where the enrichment score of an entire pathway can be skewed (Figure 6-2). Abundant isotopologue peaks will be extracted by XCMS using the same, or similar, retention time windows, but noisy or low-intensity peaks will result in uneven extractions that can affect the calculated enrichments and pathway activity score, thus biasing the analysis (Fig. 6-4).

Previous researchers have addressed the limitations of the popular peak extraction algorithms [385]. While some automatic optimization tools have been developed to reduce the number of inconsistencies and improve the extraction [169], it remains widely acknowledged that achieving a flawless extraction is an exceedingly challenging endeavor. In the future, it would be beneficial to implement additional algorithms within SUNDILE capable of evaluating the quality of the peaks (or the extraction process) or adjusting the extracted areas based on the extraction window to homogenize the input parameters of the untargeted study and reduce false positives. Arguably, the simultaneous peak extraction of all the samples could potentially reduce the errors associated with the XCMS peak extraction process, but further exploration of this hypothesis is needed.

Moreover, significant progress has been made in the development of robust machine learning and deep learning-based peak picking and extraction algorithms over recent years, although these advancements have often been presented as separate codes and programs [386], [387], [388],

[389]. It is imperative that XCMS algorithms are updated to incorporate these more accurate and recently developed methods, ensuring that the metabolomics community benefits from the latest and most reliable tools for peak extraction and data analysis.

To a lesser extent, experimental errors can also introduce variations that impact the analysis. Inherent factors associated with mass spectrometry acquisition may exhibit minor inconsistencies capable of affecting the results. For instance, the polarity setting in mass spectrometry yielded differing outcomes in the calculation of the MIDs for one specific metabolite, alanine (Figures 6-9 & 6-10). Prior studies have proposed strategies to mitigate instrumental errors in metabolomics experiments [49]. However, it is essential to emphasize the importance of tailoring the analysis of the results to the specific experimental setup. In this context, a manual filtering step that excludes possibly skewed data based on factors such as polarity, chromatographic column, and retention times is strongly recommended for future experiments.

Finally, biological characteristics can introduce bias into the results and add complexity to the extraction of biological insights from the data, posing a challenge in identifying a theory that effectively explains all the results. Notably, the presence of large and/or metabolically inactive pools can delay the labeling patterns for certain intermediate metabolites, as exemplified by the citrulline-arginine dynamics within the urea cycle (depicted in Figure 6-14). Addressing this challenge entails leveraging metabolite pool sizes as a valuable source of information for interpreting the results. Although the enrichments were normalized to plasma glucose levels to facilitate a more meaningful comparison, it is crucial to acknowledge that the inherent differences in the tissues introduce significant challenges that render direct comparisons of a given metabolite in different tissues infeasible. Different tissues may differ in their rates of tracer absorption, metabolite pool sizes, and active routes within the same pathway, making it exceedingly complex

to equate enrichments across different tissues. Given these inherent variations, it is suggested for future researchers to use an exogenous internal standard that can be added at a known concentration before the data acquisition.

6.4.2 Biological findings of ^{13}C tracing in mice under obesogenic conditions

Previous research has highlighted the reactions catalyzed by glucose-6-phosphate dehydrogenase as a rate-limited step into the pentose phosphate pathway [390]. Our findings indicate evidence of significant enrichment in an alternative route through gluconate, offering a supplementary carbon shunting pathway into the pentose phosphate pathway from glucose. This suggests the potential to overcome kinetic limitations of glucose-6-phosphate dehydrogenase. Other research has established a direct correlation between this enzyme's activity and the severity of renal hypertrophy in diabetes [391]. However, the absence of measurements of additional metabolites within the pentose phosphate pathway limits further exploration beyond the G6P node in our study. Moreover, while prior studies have reported variations in the activities of reactions leading to the pentose phosphate pathway in the livers of high-fat diet-fed mice [392], we did not find significant differences between the diet conditions in any of the organs. This effect could be attributed to the similar glucose infusion rates supplied to both mouse cohorts, which could hint at better than expected insulin sensitivity in the high-fat fed mice used in our study.

One interesting discovery is the emergence of the "ascorbate/aldarate metabolism" pathway in both studies. In the hyperinsulinemic-euglycemic clamp, notable differences were found between the kidney and the liver. Previous investigations utilizing untargeted approaches have identified significant concentration differences of threonate in genomic- [393] and diet-based studies [394] that did not involve the use of stable isotope tracers. In addition, a prior study reported significantly different pool sizes of threonate between two mouse models of obesity after oral administration of

[^{13}C]glucose. However, no differences in the enrichment in this metabolite were reported given that the enrichment analysis was done only as a targeted analysis [312]. Compared to the results of these studies, our findings highlight the importance of implementing an untargeted analysis to select the important metabolites of interest.

Furthermore, the study shed light on aminobutyric acid and propionate, suggesting a significant degree of co-metabolism with microbial organisms, potentially originating from the gut microbiota of the mice. These findings carry implications for elucidating the flux of labeled carbon in metabolic flux analysis models by considering both source and sink metabolites associated with microbial metabolism. The consideration of the interplay between host and microbiome in metabolic processes could more broadly leverage MFA to assess these cross-species interactions and play a pivotal role in understanding their role in various physiological and pathological conditions related to obesity and diabetes.

The lack of substantial isotope labeling in pathways outside of central metabolism (e.g., glycolytic pathway, pentose phosphate pathway, amino acid biosynthesis, TCA cycle, etc.), while disappointing, is not entirely unexpected. Both labeling experiments were conducted over a relatively short time span. We hypothesize that metabolites further from the entry point of the tracer, such as nucleotides, fatty acids, and macromolecules, would likely require longer infusion times to achieve significant labeling. However, extending the infusion period is challenging due to the practical constraints of the experiment. In previous *in vivo* experiments with mice, infusion periods have been extended to as long as 24 hours [81], but it's important to note that the costs of such experiments increase linearly with the duration of experimentation. There have been alternative approaches employed in various contexts that are not specific to ^{13}C labeling, such as the use of implantable osmotic pumps to deliver prolonged and steady release of medications [82]

or injectable thermoresponsive hydrogels that enable the slow and sustained release of drugs triggered by ultrasound for durations of up to one week [83]. These approaches may provide avenues to deliver stable isotopes over longer periods for future studies aimed at investigating distal metabolic pathways outside of central metabolism.

Collectively, our findings provide valuable insights into the implications of untargeted metabolomics studies in *in vivo* mammalian models. Our proposed workflow paves the way for more thoughtful experiment design in future research, allowing researchers to leverage isotopic tracers to identify which pathways are affected and place them in a biological context. Importantly, this study demonstrates the feasibility of employing an untargeted-to-targeted workflow to study *in vivo* mouse models, thereby expanding the scope of analyzed metabolic pathways and compounds that are typically used in targeted analysis.

By combining untargeted and targeted approaches, researchers can gain a more comprehensive understanding of the metabolic landscape in complex biological systems. This approach not only enhances the depth of metabolic insights by gathering information that can be linked to enzyme activity and metabolite fluxes, but also holds the potential to uncover novel biomarkers and pathways of interest: The use of our untargeted tool can highlight untapped metabolites that show differential enrichments under different metabolic conditions. The posterior exploration of these metabolites in targeted studies and the implementation of advanced analyses such as MFA can elucidate changes in the activity of the reactions underlying the molecular mechanisms of diseases and metabolic conditions.

6.5 Conclusions

In Chapter 5, we demonstrated the feasibility of implementing a workflow that integrates the use of our developed bioinformatic tools, SUNDILE and PIRAMID, within the context of plant seed metabolism. Using this model, the acquisition of multiple timepoints and extended labeling periods is feasible. However, when it comes to *in vivo* mammalian studies, these advantages are not readily available. Gathering organs for analysis requires sacrificing the animal, and usually only comparisons between unlabeled and labeled samples collected at a single time point are possible. Furthermore, the infusion methods in live animal models necessitate significant interventions that render prolonged infusions a challenging and expensive undertaking.

Additionally, the amount of tracer required for *in vivo* experiments increases significantly compared to the amounts needed for tissue culture experiment, which raises the overall cost of the experiment when added to the costs of the breeding and maintenance of the animals and the cost of the experimental procedures. We aimed to confirm that our proposed workflow was applicable to these more complex *in vivo* models. To achieve this, we tested the pipeline in two experimental setups: one involved the direct infusion of ^{13}C-propionate into long-term fasted mice, while the other used a hyperinsulinemic-euglycemic clamp with a direct infusion of ^{13}C-glucose into short-term fasted mice.

Our implementation of this workflow revealed several caveats that can be categorized into three groups: experimental, preprocessing, and software considerations:

6.5.1 Experimental considerations:

- The addition of an internal standard at a known concentration is crucial for estimating metabolite pool sizes and explaining possible dilutions of labeling due to "cold pools." This

internal standard should be a metabolite that doesn't interfere with chromatographic elution or spectrometric detection of other endogenous metabolites. It's preferable if the internal standard cannot be synthesized or metabolized by gut microbiota to avoid interference due to microbial co-metabolic exchange with the host.

- The experimental labeling time should align with the research question's scope and expectations. In the tested cases, the labeling time was not sufficient to reach distant pathways like nucleotide or fatty acid biosynthesis. Alternative labeling strategies to address this limitation were proposed.

- Manual analysis and filtering are required at various stages of the workflow. A deep understanding of the physiological implications of the results is essential for selecting the appropriate pathways and metabolites to integrate, as well as making sense of the untargeted and targeted analysis results.

6.5.2 Preprocessing considerations:

- Some metabolites may exhibit different ionization patterns in positive and negative acquisition modes. To account for these variations, it is advisable to analyze the results of each mode separately. While most metabolites show similar results in both modes, addressing exceptions due to discrepancies between modes is crucial to avoid potential biases.

- Metabolites that are only found in specific conditions or tissues can introduce variability into the analysis. To standardize comparisons across different samples, it is suggested to perform the peak extraction for all samples simultaneously, rather than extracting similar

conditions (e.g., grouped by diet or by organ) separately. This approach helps maintain consistency in the preprocessing step.

- To enhance the precision of peak picking and extraction in untargeted metabolomics studies, it's crucial to incorporate advanced peak picking algorithms, such as machine learning and deep learning-based methods, into existing tools like XCMS. These advanced algorithms should strive to ensure consistent retention time windows across various isotopologues, reducing variability in peak extraction. Additionally, efforts should be directed towards simultaneously integrating information from multiple samples to identify and align related isotopologues or fragments across different mass spectrometry datasets during the extraction step.

6.5.3 Software considerations and future work:

- As mentioned, the uneven extraction windows in the preprocessing algorithms can introduce bias into the results, and this issue could potentially be rectified by a supplementary step: normalizing the extracted peak areas to their respective retention time windows. However, this approach will require future experimentation and algorithm refinement to validate its effectiveness.

- Another challenge arises when binning *m/z* features into compounds, particularly in the context of high molecular mass metabolites where multiple isotopologues are assigned and explored. The ability to resolve different *m/z* values decreases at higher molecular weights, making SUNDILE more susceptible to mismatches in isotopologue assignment. This, in turn, can lead to inaccuracies in estimating enrichments, highlighting the need for greater precision in isotopologue assignment.

- A single metabolite with inaccurately calculated enrichment can significantly skew pathway-level scores, leading to misleading results. To address this issue, implementing an outlier detection algorithm is recommended when computing pathway scores based on individual metabolite enrichments. Additionally, to minimize the impact of pathways enriched due to a single metabolite, a threshold could be set by defining a minimum number of labeled metabolites within a pathway for it to be considered a potential target. This approach helps reduce the list of pathways and eases the process of manually detecting potential errors.

- Considering the potential for co-metabolism between the host and its microbiota, the current algorithm that restricts the list of pathways to a single organism may exclude potential metabolite matches from outside the host metabolic network. Therefore, future efforts to adapt the algorithm to incorporate multiple organisms or include microbial pathways are encouraged.

- Several pathways in KEGG exhibit redundant information, which can complicate the interpretation of results. For instance, portions of the TCA cycle may be duplicated in adjacent pathways, potentially skewing pathway scores if TCA cycle metabolites are labeled. Additionally, a considerable number of non-metabolic pathways are explored, creating an extensive list that may dilute the most relevant information. Therefore, it is suggested that the list of possible pathways should be filtered based on the specific needs of the researcher to enhance the relevance of the results.

- Finally, after identifying a list of relevant pathways, the proposed workflow currently employs manual hierarchical clustering analysis to present users with information about

pathway enrichment and comparisons to other experimental conditions. Automating this step within SUNDILE for future experiments is recommended to improve efficiency and user-friendliness.

Although the primary focus of this study was not purely biological, several noteworthy findings emerged that have the potential to shape future research in the fields of obesity and diabetes.

6.5.4 Biological findings

- In the pentose phosphate pathway, two organ-specific observations highlight the physiological distinctions between the liver and kidneys. The liver exhibited a relatively high enrichment in gluconate, suggesting an alternative route to channel carbon into the pentose phosphate pathway. This pathway, in turn, produces co-factors crucial for de novo lipogenesis, a process known to occur prominently in the liver. Conversely, M6P showed higher enrichment in the kidneys than in the liver. This discrepancy is primarily attributed to the local lysosome-assisted protein degradation process, wherein M6P is required to tag proteins for degradation. Both processes originate from the glucose/G6P node. Understanding the differential activity in the reactions of this pathway provides insights into the local activity of the metabolic networks, facilitating the resolution of corresponding fluxes through metabolic flux analysis.

- In the urea cycle pathway, citrulline exhibited a notably higher enrichment compared to other metabolites in the pathway. While carbon tracing in the pathway reactions would suggest similar enrichments between citrulline and arginine, an examination of the arginine pool size explained this phenomenon. The observed difference is attributed to a dilution effect on the labeled carbon originating from citrulline within the arginine pool.

- Despite appearing as a pathway with high activity scores in the untargeted analysis, no significant enrichment was detected in the “nucleotide metabolism” pathway. This discrepancy was attributed to inaccuracies in the peak extraction step during data preprocessing.

- Two compounds, aminobutyric acid and propionate, which have not been reported to be biosynthesized in mammals, exhibited labeling. The presence of labeling in these compounds could be linked to co-metabolism between mice and their gut bacteria. However, the unusually high enrichments in both compounds are not typical for co-metabolic processes that lack rapid exchange between metabolites in the bloodstream or organs and the gastrointestinal system. Additionally, the possibility of a non-reported reaction for propionate production, suggested by its similar enrichment to succinate, was proposed, but further research is necessary to confirm this hypothesis.

- Enrichment in threonate hinted at carbon allocation into the "ascorbate and aldarate metabolism" pathway. While it is acknowledged that certain reactions within this pathway are not possible in humans, the exploration of this pathway and its metabolites, such as UDP-glucose and UDP-glucuronate, remains an intriguing subject for future research due to their roles in obesity and diabetes.

6.6 Appendix

Metabolic Pathway	Number of Matches	Map Score
Chemical carcinogenesis - DNA adducts	1	0.391467
Chemical carcinogenesis - reactive oxygen species	1	0.391467
Insulin signaling pathway	1	0.313225
Type II diabetes mellitus	1	0.313225
Non-alcoholic fatty liver disease	1	0.313225
AGE-RAGE signaling pathway in diabetic complications	1	0.313225
Metabolism of xenobiotics by cytochrome P450	2	0.28318
Neomycin, kanamycin and gentamicin biosynthesis	2	0.279731
Insulin secretion	2	0.279731
Prolactin signaling pathway	2	0.279731
Insulin resistance	2	0.279731
Biosynthesis of nucleotide sugars	6	0.274189
Amino sugar and nucleotide sugar metabolism	5	0.273032
Starch and sucrose metabolism	3	0.268566
Galactose metabolism	6	0.260092
Inositol phosphate metabolism	1	0.246237
Thyroid hormone synthesis	1	0.246237
Pentose phosphate pathway	5	0.24614
Fructose and mannose metabolism	8	0.246073
Pathways in cancer	2	0.244755
Renal cell carcinoma	2	0.244755
Carbohydrate digestion and absorption	4	0.239072
Diabetic cardiomyopathy	3	0.238071
Mineral absorption	3	0.22603
Glycerolipid metabolism	2	0.212544
Glycine, serine and threonine metabolism	1	0.212346
Glyoxylate and dicarboxylate metabolism	4	0.210704
Valine, leucine and isoleucine biosynthesis	3	0.210097
Glycolysis / Gluconeogenesis	9	0.209692
Terpenoid backbone biosynthesis	1	0.207601
AMPK signaling pathway	3	0.195919
Glucagon signaling pathway	7	0.193969
2-Oxocarboxylic acid metabolism	5	0.187721
Valine, leucine and isoleucine degradation	2	0.182295
Oxidative phosphorylation	1	0.18142
Tyrosine metabolism	2	0.18142
Carbon metabolism	11	0.180834
Choline metabolism in cancer	1	0.17885
HIF-1 signaling pathway	2	0.17683
Bile secretion	2	0.17683

Proximal tubule bicarbonate reclamation	4	0.170971
FoxO signaling pathway	2	0.170761
Citrate cycle (TCA cycle)	5	0.163991
Butanoate metabolism	5	0.158993
Glutathione metabolism	1	0.154029
Pyruvate metabolism	6	0.150877
Taste transduction	5	0.149569
Central carbon metabolism in cancer	8	0.146554
Biosynthesis of amino acids	10	0.145134
Metabolic pathways	90	0.145034
Nicotinate and nicotinamide metabolism	5	0.139289
Cysteine and methionine metabolism	3	0.138744
ABC transporters	4	0.138241
Pentose and glucuronate interconversions	8	0.135242
Phenylalanine, tyrosine and tryptophan biosynthesis	2	0.134186
Propanoate metabolism	6	0.121189
Phenylalanine metabolism	2	0.120703
Glycerophospholipid metabolism	3	0.120452
Ascorbate and aldarate metabolism	8	0.117008
Arginine biosynthesis	2	0.110928
Alanine, aspartate and glutamate metabolism	4	0.102102
Biosynthesis of cofactors	12	0.094236
Pyrimidine metabolism	6	0.061899
Histidine metabolism	3	0.061533
D-Amino acid metabolism	5	0.061399
Selenocompound metabolism	1	0.051641
Pantothenate and CoA biosynthesis	1	0.051641
Taurine and hypotaurine metabolism	3	0.049662
Sulfur relay system	2	0.049118
Protein digestion and absorption	4	0.049033
Folate biosynthesis	1	0.046595
Phototransduction	1	0.046595
Lysine degradation	1	0.040435
Lipoic acid metabolism	1	0.040435
GABAergic synapse	1	0.040435
Purine metabolism	4	0.037445
cGMP-PKG signaling pathway	2	0.037445
Olfactory transduction	2	0.037445
Antifolate resistance	2	0.037445
Nucleotide metabolism	4	0.031479
Aminoacyl-tRNA biosynthesis	2	0.030445
cAMP signaling pathway	1	0.028296
PI3K-Akt signaling pathway	1	0.028296

mTOR signaling pathway	1	0.028296
Regulation of lipolysis in adipocytes	1	0.028296
Parathyroid hormone synthesis, secretion and action	1	0.028296
Renin secretion	1	0.028296
Aldosterone synthesis and secretion	1	0.028296
Cortisol synthesis and secretion	1	0.028296
Longevity regulating pathway	1	0.028296
Parkinson disease	1	0.028296
Pathways of neurodegeneration - multiple diseases	1	0.028296
Morphine addiction	1	0.028296
Cushing syndrome	1	0.028296
beta-Alanine metabolism	3	0.02816
Neuroactive ligand-receptor interaction	2	0.027875
Tryptophan metabolism	1	0.023589
Phosphonate and phosphinate metabolism	2	0.022135
Vitamin B6 metabolism	1	0.018974
Ether lipid metabolism	1	0.003655
Arachidonic acid metabolism	5	-0.02629
Ovarian steroidogenesis	3	-0.02629
Vascular smooth muscle contraction	4	-0.02629
Serotonergic synapse	3	-0.02629
Inflammatory mediator regulation of TRP channels	2	-0.02629

Table 6-A1. List of pathways that were potentially labeled in the fasted Mc4r$^{-/-}$ mouse dataset, including the number of metabolites that were found within each pathway, and their corresponding enrichment scores.

RT [min]	Mo [Da]	Enrichment score	Suggested ID(Score)
4.818	161.046	0.125	(2R,3S)-2,3-Dimethylmalate(0.37741)-(R)-2-Ethylmalate(0.19377)-2-Dehydro-3-deoxy-L-rhamnonate(0.19377)
18.025	161.047	0.328	(2R,3S)-2,3-Dimethylmalate(0.37741)-(R)-2-Ethylmalate(0.19377)-2-Dehydro-3-deoxy-L-rhamnonate(0.19377)
6.350	133.014	0.386	(R)-Malate(0.49561)-(S)-Malate(0.48895)-3-Dehydro-L-threonate(0.015434)
6.071	133.014	0.230	(R)-Malate(0.49561)-(S)-Malate(0.48895)-3-Dehydro-L-threonate(0.015434)
4.847	101.025	0.212	(S)-Methylmalonate semialdehyde(0.41418)-2-Oxobutanoate(0.40155)-Acetoacetate(0.091911)
4.155	128.035	0.154	1-Pyrroline-4-hydroxy-2-carboxylate(0.46581)-5-Oxoproline(0.44603)-5-Oxo-D-proline(0.047167)
6.303	264.952	0.155	2,3-Bisphospho-D-glycerate(0.50386)-3-Phospho-D-glyceroyl phosphate(0.49614)
1.068	319.228	-0.026	20-HETE(0.2)-19(S)-HETE(0.2)-5,6-EET(0.2)
4.713	215.033	0.208	2-C-Methyl-D-erythritol 4-phosphate(1)
5.747	111.009	0.190	2-Furoate(1)
4.832	113.025	0.060	2-Hydroxy-2,4-pentadienoate(1)
3.991	124.991	0.055	2-Hydroxyethanesulfonate(1)
3.634	124.991	0.059	2-Hydroxyethanesulfonate(1)
4.152	144.030	0.135	2-Oxoglutaramate(0.50112)-4-Oxoglutaramate(0.49888)
4.396	145.014	0.040	2-Oxoglutarate(0.79139)-Dehydro-D-arabinono-1,4-lactone(0.20861)
3.019	196.029	0.019	3-Hydroxy-2-methylpyridine-4,5-dicarboxylate(0.98201)-Clavulanate-9-aldehyde(0.017986)
3.365	187.042	0.589	3-Hydroxy-2-naphthoate(0.64819)-1-Hydroxy-2-naphthoate(0.35181)
6.100	166.975	0.022	3-Phosphonopyruvate(0.46551)-Phosphoenolpyruvate(0.46536)-Hydroxypyruvaldehyde phosphate(0.069135)
5.687	351.057	0.190	4-(4-Deoxy-alpha-D-gluc-4-enuronosyl)-D-galacturonate(0.51273)-Arbutin 6-phosphate(0.48727)
5.099	158.989	0.147	5-Chloro-1,2,4-trihydroxybenzene(1)
5.626	346.056	0.028	AMP(0.65812)-3'-AMP(0.27056)-dGMP(0.071319)
4.713	285.083	0.391	Benzo[a]pyrene-7,8-dihydrodiol(0.95956)-(-)-Vestitone(0.04044)-(S)-DNPA(5.8976e-62)
4.088	187.073	0.136	cis-2,3-Dihydro-2,3-dihydroxybiphenyl(1)
4.638	187.073	0.131	cis-2,3-Dihydro-2,3-dihydroxybiphenyl(1)
4.788	173.009	0.268	cis-Aconitate(0.95082)-Dehydroascorbate(0.049183)
4.921	147.030	0.209	D-erythro-3-Methylmalate(0.20323)-(R)-2-Methylmalate(0.19961)-2-Dehydro-3-deoxy-L-arabinonate(0.19933)
5.024	195.051	0.214	D-Galactonate(0.40016)-D-Altronate(0.20027)-D-Mannonate(0.20005)
5.399	195.051	0.043	D-Galactonate(0.40016)-D-Altronate(0.20027)-D-Mannonate(0.20005)
6.024	259.022	0.246	D-Glucose 1-phosphate(0.31536)-D-Glucose 6-phosphate(0.17116)-alpha-D-Glucose 6-phosphate(0.17116)

5.065	127.051	0.175	gamma-Amino-gamma-cyanobutanoate(0.70316)-Naphthalene(0.14842)-(R)-5,6-Dihydrothymine(0.14842)
5.962	362.051	0.047	GMP(0.6548)-Guanosine 3'-phosphate(0.1726)-Precursor Z(0.1726)
5.488	102.056	0.152	L-3-Aminoisobutanoate(0.44334)-(S)-2-Aminobutanoate(0.42573)-4-Aminobutanoate(0.064545)
5.046	88.041	0.052	L-Alanine(0.80181)-beta-Alanine(0.10972)-Sarcosine(0.05955)
4.850	165.041	0.011	L-Arabinonate(0.57727)-D-Xylonate(0.28865)-7-Methylxanthine(0.06704)
5.767	254.991	0.288	L-Ascorbate 6-phosphate(1)
5.846	305.069	0.357	Leucocyanidin(1)
5.569	154.062	0.009	L-Histidine(1)
4.803	179.056	0.313	L-Rhamnonate(0.31536)-D-Glucose(0.17116)-D-Galactose(0.17116)
5.504	115.004	0.156	Maleic acid(0.50575)-Fumarate(0.49425)
4.980	115.004	0.207	Maleic acid(0.50575)-Fumarate(0.49425)
4.705	269.088	0.304	Medicarpin(1)-Dihydroxycarbazepine(2.0989e-59)
4.215	73.030	0.084	Propanoate(0.48676)-(S)-Lactaldehyde(0.12831)-(R)-Lactaldehyde(0.12831)
5.778	323.029	0.047	Pseudouridine 5'-phosphate(0.6548)-UMP(0.1726)-3'-UMP(0.1726)
2.884	166.018	0.024	Quinolinate(1)-Thioguanine(1.2973e-62)
18.684	121.030	0.072	Salicylaldehyde(0.46575)-Benzoate(0.45143)-4-Hydroxybenzaldehyde(0.082823)
5.994	289.033	0.297	Sedoheptulose 7-phosphate(0.60339)-D-glycero-beta-D-manno-Heptose 7-phosphate(0.32749)-D-glycero-beta-D-manno-Heptose 1-phosphate(0.069124)
5.435	214.049	0.004	sn-Glycero-3-phosphoethanolamine(1)
5.916	171.007	0.179	Toluene-4-sulfonate(0.33714)-sn-Glycerol 1-phosphate(0.33377)-sn-Glycerol 3-phosphate(0.32909)
0.056	125.061	0.001	Toluene-cis-dihydrodiol(1)
5.244	403.001	0.004	UDP(1)

Table 6-A2. List of labeled metabolites for which an identity was suggested in the fasted Mc4r$^{-/-}$ dataset.

	Liver HF		Liver Chow		Kidney HF		Kidney Chow	
Pathways	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**
Glycolysis / Gluconeogenesis	0.13	0.12	0.11	0.23	0.52	0.23	0.57	0.27
Citrate cycle (TCA cycle)	0.32	0.00	0.28	0.00	0.50	0.65	0.61	0.64
Pentose phosphate pathway	0.11	0.10	0.11	0.10	0.32	0.21	0.38	0.47
Pentose and glucuronate interconversions	0.15	0.01	0.16	0.04	0.40	0.12	0.34	-0.09
Fructose and mannose metabolism	0.20	0.07	0.17	0.10	0.39	0.22	0.48	0.33
Galactose metabolism	0.21	0.01	0.13	0.04	0.44	0.16	0.47	0.37
Ascorbate and aldarate metabolism	0.22	0.00	0.25	0.00	0.36	0.51	0.34	0.57
Fatty acid biosynthesis	0.03	0.00	0.02	0.00	0.39	0.00	0.39	0.00
Fatty acid elongation	0.03	0.00	0.01	0.00	0.02	0.00	0.01	0.00
Fatty acid degradation	0.03	0.00	0.01	0.00	0.02	0.00	0.01	0.00
Primary bile acid biosynthesis	0.00	0.00	0.00	0.00	0.02	0.04	0.00	0.14
Ubiquinone and other terpenoid-quinone biosynthesis	0.00	0.40	0.11	0.75	0.00	0.45	0.36	0.18
Steroid hormone biosynthesis	0.00	0.13	0.00	0.15	0.00	0.00	0.00	0.00
Oxidative phosphorylation	0.17	0.01	0.13	0.01	0.41	0.00	0.42	0.01
Arginine biosynthesis	0.39	0.41	0.47	0.35	0.35	0.21	0.41	0.25
Purine metabolism	0.09	0.11	0.10	0.13	0.12	0.07	0.13	0.10
Caffeine metabolism	0.02	0.00	0.06	0.00	0.20	0.00	0.13	0.00
Pyrimidine metabolism	0.23	0.28	0.16	0.05	0.21	0.19	0.22	0.19
Alanine, aspartate and glutamate metabolism	0.28	0.23	0.31	0.27	0.42	0.38	0.46	0.36
Glycine, serine and threonine metabolism	0.09	0.11	0.09	0.15	0.40	0.22	0.32	0.16
Cysteine and methionine metabolism	0.15	0.13	0.18	0.10	0.37	0.16	0.29	0.24
Valine, leucine and isoleucine degradation	0.31	0.22	0.23	0.28	0.39	0.17	0.38	0.02
Valine, leucine and isoleucine biosynthesis	0.25	0.22	0.32	0.28	0.46	0.17	0.39	0.18
Lysine degradation	0.32	0.10	0.26	0.12	0.27	0.19	0.35	0.29
Arginine and proline metabolism	0.30	0.18	0.26	0.17	0.33	0.21	0.25	0.21
Histidine metabolism	0.21	0.37	0.40	0.24	0.37	0.35	0.43	0.16
Tyrosine metabolism	0.12	0.01	0.23	0.00	0.46	0.00	0.46	0.00
Phenylalanine metabolism	0.20	0.01	0.16	0.08	0.46	0.00	0.49	0.00
Tryptophan metabolism	0.92	0.00	0.28	0.00	0.00	0.00	0.00	0.00
Phenylalanine, tyrosine and tryptophan biosynthesis	0.08	0.06	0.11	0.10	0.36	0.00	0.24	0.00
beta-Alanine metabolism	0.16	0.04	0.06	0.05	0.40	0.00	0.57	0.01
Taurine and hypotaurine metabolism	0.19	0.20	0.40	0.23	0.30	0.16	0.48	0.25
Phosphonate and phosphinate metabolism	0.02	0.00	0.03	0.00	0.02	0.18	0.01	0.28
Selenocompound metabolism	0.00	0.08	0.00	0.19	0.00	0.00	0.00	0.00
D-Amino acid metabolism	0.34	0.15	0.38	0.18	0.28	0.16	0.32	0.25
Glutathione metabolism	0.41	0.20	0.25	0.15	0.45	0.29	0.49	0.40
Starch and sucrose metabolism	0.10	0.03	0.10	0.06	0.48	0.07	0.50	0.16
Mannose type O-glycan biosynthesis	0.91	0.00	0.80	0.00	0.62	0.00	0.50	0.00
Amino sugar and nucleotide sugar metabolism	0.10	0.12	0.10	0.14	0.43	0.07	0.44	0.16
Neomycin, kanamycin and gentamicin biosynthesis	0.16	0.31	0.21	0.13	0.44	0.11	0.38	0.22

Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	0.03	0.01	0.02	0.01	0.05	0.00	0.04	0.01
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	0.03	0.01	0.02	0.01	0.05	0.00	0.04	0.01
Glycerolipid metabolism	0.28	0.08	0.11	0.07	0.43	0.19	0.53	0.22
Inositol phosphate metabolism	0.12	0.06	0.08	0.10	0.31	0.07	0.32	0.16
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.30
Glycerophospholipid metabolism	0.07	0.02	0.02	0.03	0.08	0.10	0.15	0.11
Ether lipid metabolism	0.00	0.02	0.00	0.02	0.00	0.01	0.00	0.00
alpha-Linolenic acid metabolism	0.00	0.00	0.00	0.00	0.49	0.00	0.01	0.00
Sphingolipid metabolism	0.00	0.00	0.00	0.00	0.16	0.18	0.15	0.19
Pyruvate metabolism	0.33	0.07	0.25	0.18	0.51	0.44	0.65	0.48
Glyoxylate and dicarboxylate metabolism	0.26	0.20	0.28	0.18	0.42	0.29	0.46	0.35
Propanoate metabolism	0.30	0.00	0.20	0.00	0.55	0.00	0.71	0.00
Butanoate metabolism	0.26	0.23	0.28	0.13	0.50	0.26	0.55	0.45
Thiamine metabolism	0.13	0.00	0.05	0.00	0.00	0.06	0.00	0.26
Riboflavin metabolism	0.02	0.13	0.01	0.15	0.02	0.56	0.01	0.12
Vitamin B6 metabolism	0.16	0.21	0.13	0.25	0.13	0.20	0.19	0.25
Nicotinate and nicotinamide metabolism	0.34	0.31	0.31	0.22	0.55	0.15	0.53	0.21
Pantothenate and CoA biosynthesis	0.06	0.12	0.05	0.14	0.20	0.08	0.30	0.02
Biotin metabolism	0.02	0.02	0.00	0.00	0.02	0.21	0.26	0.28
Lipoic acid metabolism	0.25	0.00	0.37	0.00	0.00	0.25	0.00	0.37
Folate biosynthesis	0.07	0.00	0.05	0.19	0.27	0.00	0.10	0.00
Porphyrin metabolism	0.36	0.33	0.47	0.27	0.59	0.17	0.45	0.35
Terpenoid backbone biosynthesis	0.07	0.13	0.05	0.15	0.35	0.56	0.32	0.55
Nitrogen metabolism	0.33	0.27	0.38	0.26	0.43	0.31	0.51	0.42
Sulfur metabolism	0.13	0.12	0.23	0.09	0.35	0.06	0.37	0.11
Metabolism of xenobiotics by cytochrome P450	0.00	0.03	0.00	0.12	0.00	0.60	0.00	0.80
Drug metabolism - cytochrome P450	0.18	0.04	0.47	0.04	0.00	0.05	0.07	0.00
Drug metabolism - other enzymes	0.05	0.00	0.11	0.00	0.09	0.00	0.25	0.00
Biosynthesis of unsaturated fatty acids	0.03	0.00	0.02	0.00	0.02	0.00	0.01	0.00
Metabolic pathways	0.19	0.16	0.17	0.15	0.32	0.23	0.31	0.26
Carbon metabolism	0.17	0.14	0.18	0.15	0.41	0.18	0.44	0.36
2-Oxocarboxylic acid metabolism	0.33	0.17	0.37	0.16	0.36	0.25	0.34	0.35
Fatty acid metabolism	0.03	0.00	0.01	0.00	0.39	0.00	0.39	0.00
Biosynthesis of amino acids	0.19	0.22	0.25	0.20	0.29	0.26	0.29	0.26
Nucleotide metabolism	0.18	0.09	0.12	0.10	0.13	0.11	0.13	0.14
Biosynthesis of cofactors	0.16	0.09	0.14	0.11	0.27	0.16	0.30	0.17
Biosynthesis of nucleotide sugars	0.09	0.14	0.09	0.12	0.34	0.10	0.34	0.20

Table 6-A3. List of scores of the potentially labeled pathways in the hyperinsulinemic-euglycemic clamp dataset.

	Liver HF		Liver Chow		Kindey HF		Kindey Chow	
Pathways	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**
Glycolysis / Gluconeogenesis	12	7	12	8	7	7	7	7
Citrate cycle (TCA cycle)	5	0	5	0	4	2	4	2
Pentose phosphate pathway	16	5	15	5	12	3	11	5
Pentose and glucuronate interconversions	16	1	16	1	13	2	13	5
Fructose and mannose metabolism	12	3	11	3	9	4	9	3
Galactose metabolism	13	3	13	3	12	4	11	5
Ascorbate and aldarate metabolism	15	0	15	0	12	2	12	2
Fatty acid biosynthesis	2	0	2	0	2	0	2	0
Fatty acid elongation	1	0	1	0	1	0	1	0
Fatty acid degradation	1	0	1	0	1	0	1	0
Primary bile acid biosynthesis	1	0	0	0	1	2	0	2
Ubiquinone and other terpenoid-quinone biosynthesis	0	1	1	1	0	3	1	3
Steroid hormone biosynthesis	0	1	0	1	0	0	0	0
Oxidative phosphorylation	2	1	2	1	3	1	3	1
Arginine biosynthesis	5	4	6	4	7	5	7	5
Purine metabolism	18	14	16	14	16	18	14	19
Caffeine metabolism	5	0	10	0	5	0	8	0
Pyrimidine metabolism	11	7	11	5	14	6	15	7
Alanine, aspartate and glutamate metabolism	6	6	6	6	10	3	10	4
Glycine, serine and threonine metabolism	3	6	3	5	6	9	8	9
Cysteine and methionine metabolism	6	7	6	8	4	6	6	6
Valine, leucine and isoleucine degradation	1	1	1	1	3	1	3	1
Valine, leucine and isoleucine biosynthesis	3	1	3	1	4	2	4	2
Lysine degradation	3	5	5	5	8	8	8	8
Arginine and proline metabolism	4	5	5	6	10	10	11	13
Histidine metabolism	3	4	2	3	3	3	3	3
Tyrosine metabolism	3	1	1	0	6	0	4	0
Phenylalanine metabolism	2	3	3	3	3	0	3	0
Tryptophan metabolism	1	0	3	0	0	0	0	0
Phenylalanine, tyrosine and tryptophan biosynthesis	3	3	2	3	2	0	3	0
beta-Alanine metabolism	1	2	2	2	4	0	3	2
Taurine and hypotaurine metabolism	3	2	2	2	3	3	2	3
Phosphonate and phosphinate metabolism	6	0	4	0	3	2	1	2
Selenocompound metabolism	0	1	0	1	0	0	0	0
D-Amino acid metabolism	9	14	10	14	16	20	16	20
Glutathione metabolism	3	3	3	3	3	5	3	5
Starch and sucrose metabolism	7	5	7	5	7	4	7	4
Mannose type O-glycan biosynthesis	1	0	1	0	1	0	1	0
Amino sugar and nucleotide sugar metabolism	8	5	8	5	8	2	8	2
Neomycin, kanamycin and gentamicin biosynthesis	4	4	4	3	5	3	6	4
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	1	1	1	1	1	1	1	1

Glycosaminoglycan biosynthesis - heparan sulfate / heparin	1	1	1	1	1	1	1	1
Glycerolipid metabolism	7	4	8	4	4	4	3	4
Inositol phosphate metabolism	5	2	5	2	3	1	3	1
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	0	0	0	0	0	1	0	1
Glycerophospholipid metabolism	2	7	5	6	2	7	1	8
Ether lipid metabolism	1	2	1	2	0	1	0	2
alpha-Linolenic acid metabolism	0	0	0	0	1	0	1	0
Sphingolipid metabolism	0	0	0	0	1	2	1	2
Pyruvate metabolism	5	1	5	1	6	1	6	1
Glyoxylate and dicarboxylate metabolism	11	4	10	4	12	8	11	8
Propanoate metabolism	4	0	4	0	6	0	6	0
Butanoate metabolism	6	3	6	3	7	1	7	1
Thiamine metabolism	1	0	1	0	0	1	0	1
Riboflavin metabolism	1	1	1	1	1	1	1	2
Vitamin B6 metabolism	3	1	3	1	2	2	2	2
Nicotinate and nicotinamide metabolism	5	7	5	4	6	3	5	3
Pantothenate and CoA biosynthesis	2	2	3	2	5	2	4	3
Biotin metabolism	1	1	0	1	1	1	1	1
Lipoic acid metabolism	1	0	1	0	0	2	0	2
Folate biosynthesis	4	3	4	3	3	0	3	0
Porphyrin metabolism	2	1	2	1	4	3	4	3
Terpenoid backbone biosynthesis	3	2	3	2	1	2	1	2
Nitrogen metabolism	2	2	2	2	2	2	2	2
Sulfur metabolism	5	3	3	3	7	5	5	5
Metabolism of xenobiotics by cytochrome P450	0	2	0	2	0	1	0	1
Drug metabolism - cytochrome P450	3	5	4	5	0	2	1	1
Drug metabolism - other enzymes	1	0	1	0	1	0	1	0
Biosynthesis of unsaturated fatty acids	2	0	2	0	1	0	1	0
Metabolic pathways	161	101	165	92	178	108	176	121
Carbon metabolism	27	11	26	11	25	11	24	13
2-Oxocarboxylic acid metabolism	8	7	9	8	11	7	12	9
Fatty acid metabolism	1	0	1	0	2	0	2	0
Biosynthesis of amino acids	20	16	22	17	24	17	25	19
Nucleotide metabolism	16	12	17	12	15	17	15	16
Biosynthesis of cofactors	22	15	20	16	20	17	17	18
Biosynthesis of nucleotide sugars	11	6	11	7	9	4	9	3

Table 6-A4. Number of potentially labeled metabolites detected in each pathway based on the hyperinsulinemic-euglycemic clamp dataset.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

This dissertation describes the development of two bioinformatics tools in the context of a comprehensive stable isotope-based metabolomics workflow. The application of these tools to two biological case studies is discussed, revealing new enriched pathways and metabolites that would not be considered in typical targeted metabolomics studies.

Chapter 3 describes the development of the first tool, SUNDILE, which is an untargeted software designed to detect labeled compounds in stable isotope-tracing experiments. This tool receives a preprocessed list of peaks extracted from a mass spectrometry file and, aided by machine learning algorithms, groups and identifies compounds that exhibit isotope enrichment. Subsequently, by employing a metabolic network reconstruction derived from the online database KEGG, SUNDILE generates a list of level-3 identities for the labeled compounds that were detected. Finally, it evaluates the enrichment of metabolites and contextualizes their enrichment scores through a pathway-level analysis, providing the user with a list of labeled pathways for further investigation. More broadly, this program has the potential to uncover untapped biomarkers and metabolites of interest in metabolomics studies.

Chapter 4 introduces the second tool, PIRAMID, which is a targeted software designed to quantify the abundance of isotopologues and isotopic enrichment of a given compound. This tool offers support for GC-MS and LC-MS files in non-vendor file formats, and its algorithms are optimized to handle data from various MS techniques, including single and tandem MS, low- and high-resolution MS, as well as single- and multiple-labeled tracers. Additionally, it has the capability to

correct the calculated distribution of relative isotopologue intensities for the natural abundance of their isotopes. PIRAMID was compared to other freely available programs that are widely used in metabolomics studies, finding comparable accuracies while also offering a significant number of useful features that make it a versatile and user-friendly alternative for automating the targeted quantification of metabolites. Overall, this tool proves to be a robust tool for the reliable quantification of metabolites across multiple files, facilitating targeted analyses in the metabolomics workflow.

Chapter 5 shows the practical application of a workflow that implements both tools in researching the metabolism of two cultivars of soybean seed embryos with distinct oil yield levels. These cultivars were separately enriched with [U-$^{13}C_6$]glucose and [$^{13}C_6$,$^{15}N_2$]glutamine. The findings elucidate the fate of carbon depending on the source: glutamine was fed directly into the TCA cycle, eventually being transformed to pyruvate from malate via malic enzyme. Analogously, glucose exhibited limited entry into the TCA cycle, with its carbon flux redirected at the PEP node. However, the shikimate pathway displayed significant enrichment, impacting amino acids, isoflavonoids, and anthraquinones. Regarding the inter-cultivar differences, the expression of malic enzyme was found to be directly proportional to the oil yield. Furthermore, the cultivar presenting the lowest oil yield showed higher enrichments in nucleotides and nucleotide sugars, suggesting potential targets for optimizing soybean oil production through bioengineering the entry points of these pathways.

Chapter 6 discusses the practical application of the developed tools into *in vivo* experiments involving mice under obesogenic conditions. To establish the feasibility of this implementation, the algorithms underwent initial testing on a small dataset from genetically modified mice under fasting conditions. While several limitations and considerations for future implementations were

identified, the study revealed evidence of enrichment in compounds in the previously unexplored "ascorbate and aldarate metabolism" pathway. In addition, expected behaviors in the glycolysis/gluconeogenesis and pentose phosphate pathway were validated. Motivated by these findings, the tools were tested on a more complex experiment, a hyperinsulinemic-euglycemic clamp on mice fed diets with different fat contents, focusing on metabolites from liver and kidney tissue. This study offered insights into additional experimental considerations for implementing the proposed workflow in *in vivo* experiments. More importantly, the results exhibited differences in the enrichment of gluconate in the livers of all mice, suggesting an alternative route for funneling carbon into the pentose phosphate pathway, responding to the liver's physiological feature of synthesizing fatty acids. Furthermore, the presence of enrichment in metabolites that are derived from essential amino acids, aminobutyric acid and propionate, was observed. This finding could be attributed to the presence of co-metabolism with gut bacteria following rapid exchanges of labeled metabolites or could unveil previously unknown routes for the biosynthesis of these metabolites in mammals. This hypothesis is supported by the similar enrichment found in propionate, a metabolite typically produced by gut bacteria, and succinate, a metabolite in the TCA cycle. Finally, this study also revealed the presence of enrichment in metabolites in the "ascorbate and aldarate metabolism" pathway, suggesting that the metabolites within it could be a potential target in future studies aimed at the understanding of metabolism under obesogenic conditions.

7.2 Future directions

7.2.1 Bioinformatics

While SUNDILE is able to predict the activity score of metabolites and pathway, the analysis of these scores requires manual curation of the results. Further updates on the software should aim to implement unsupervised methods to present the information to the user in a more comprehensive

way. Furthermore, the underlying supervised algorithms of SUNDILE were trained and developed based on databases of metabolites with masses below 1000 Da. While this value is acceptable for metabolomics purposes, the refinement and re-training of these algorithms along with the implementation of new online databases will be necessary if the reach of the tool is expanded to proteomics or lipidomics studies in the future. Finally, the current algorithms in SUNDILE do not account for all the available enzymatic information in KEGG. While relying solely on the metabolic reactions in a pathway has advantages such as the inclusion of reactions that have not been reported for certain organisms, it also opens the door for errors when organisms do not express certain pathway enzymes that lead to false positives in the studies. Ideally, the use of this feature should be a decision of the user who can discern the results accordingly. Hence, the implementation of enzyme-level information into the network reconstruction algorithms is left for future updates of SUNDILE.

Concerning PIRAMID, the implementation of an optimized baselining algorithm was accomplished through the comparison of multiple algorithms, minimizing the resulting error. However, there is no one-size-fits-all solution that can avoid all errors stemming from miscalculation of the baseline. Hence, further investigation is needed into the development and integration of more precise algorithms or better ways to present the potential errors in the calculation to users, allowing for manual corrections. Furthermore, as recent studies have focused on the use of a relatively new MS technique, MALDI, to gather spatial information of the metabolites in a sample, enhancing PIRAMID's support to encompass non-chromatographic types of MS data would be valuable in the future. While, in theory, the processing algorithms of PIRAMID should be compatible with this type of data, the actual implementation of algorithms to process these types of files, ultimately extracting position versus intensity data remains a task for

future updates. Lastly, the incorporation of an algorithm capable of exporting the results from PIRAMID into MFA-specialized software such as INCA, would significantly reduce the time investment for metabolism researchers. This feature is also recommended for future implementation.

7.2.2 Biological

While the results of Chapter 5 showed enrichments in the isoflavonoid and anthraquinone biosynthesis pathways, the mechanisms of the carbon rerouting in each condition and the possible influence of these compounds on the oil yield are still unknown. Further research will be required to elucidate these mechanisms and dependencies. Furthermore, potential bioengineering targets were proposed to optimize the oil yield. The overexpression of malic enzyme has yielded encouraging results in other organisms [280], [281], but the results have yet to be extrapolated to soybeans. In addition, metabolic routes leading to the synthesis of nucleotides and nucleotide sugars could also be diverted to funnel more carbon into the glycolysis pathways that ultimately ends in the production of fatty acid precursors.

The results in Chapter 6 show a difference in the enrichment of pathways leading to the pentose phosphate pathway, but no measurements of flux were calculated. Future studies should use the information of the surrounding metabolites to perform MFA on this pathway to confirm this hypothesis. Furthermore, the presence of carbon exchange between succinate and propionate was suggested given the enrichment data of these metabolites. To test this hypothesis, enzymatic information in the reactions between these metabolites should be gathered and the determination of the exact conditions where this phenomenon is present is also necessary. In addition, acknowledging that humans are unable to fully perform the reactions in the “aldarate and ascorbate metabolism” pathway, further targeted analyses focusing on this pathway should be conveyed to

test the extent of carbon funneling into it and its relevance in the context of obesogenic conditions. Finally, it was recognized as a limitation of this study that the labeling process using [$^{13}C_6$]glucose only occurred for a relative short period of time, which could restrict the labeled carbons from reaching into pathways that are further from the entry point. As the feasibility of the implementation of these tools was established, future studies would benefit from the use of prolonged labeling conditions such as implanted pumps and orally administrated tracers.

7.3 Contribution

Overall, these studies highlight the versatility and applicability of the developed tools in the field of metabolomics studies involving stable isotope tracers. While previous tools have focused on specific subsets of mass spectrometry-based metabolomics, the developed tools in this dissertation are optimized to leverage the use of stable isotopes, offering simultaneous support for the most common types of mass spectrometry techniques. They open the door to the analysis of experiments involving multiple tracers, or dual-labeled tracers, significantly expanding the reach of the studies. Furthermore, they allow the discovery of new metabolites that are enriched in the experiments, and their contextualization in a pathway-level analysis. A greater understanding of the impacted pathways and metabolites in a study involving stable isotopes can aid in the reconstruction of metabolic networks and in the elucidation of metabolic fluxes by providing additional information in the models that are used in MFA studies. Ultimately, this process leads to a better understanding of the metabolism, which can have positive impacts in biomedical and agricultural research.

REFERENCES

[1] A. Chokkathukalam, D. H. Kim, M. P. Barrett, R. Breitling, and D. J. Creek, "Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks," *Bioanalysis*, vol. 6, no. 4, p. 511, Feb. 2014, doi: 10.4155/BIO.13.348.

[2] C. Jang, L. Chen, and J. D. Rabinowitz, "Metabolomics and Isotope Tracing," *Cell*, vol. 173, no. 4, pp. 822–837, 2018, doi: 10.1016/j.cell.2018.03.055.

[3] D. Yu, L. Zhou, X. Liu, and G. Xu, "Stable isotope-resolved metabolomics based on mass spectrometry: Methods and their applications," *TrAC Trends in Analytical Chemistry*, vol. 160, p. 116985, Mar. 2023, doi: 10.1016/J.TRAC.2023.116985.

[4] A. N. Lane and T. W. M. Fan, "NMR-Based Stable Isotope Resolved Metabolomics in Systems Biochemistry," *Archives of biochemistry and biophysics*, vol. 628, p. 123, Aug. 2017, doi: 10.1016/J.ABB.2017.02.009.

[5] Y. Wang, F. E. Wondisford, C. Song, T. Zhang, and X. Su, "Metabolic Flux Analysis—Linking Isotope Labeling and Metabolic Fluxes," *Metabolites*, vol. 10, no. 11, pp. 1–21, Nov. 2020, doi: 10.3390/METABO10110447.

[6] J. A. Reisz and A. D'Alessandro, "Measurement of metabolic fluxes using stable isotope tracers in whole animals and human patients," *Current Opinion in Clinical Nutrition and Metabolic Care*, vol. 20, no. 5, pp. 366–374, 2017, doi: 10.1097/MCO.0000000000000393.

[7] C. Balcells, C. Foguet, J. Tarragó-Celada, P. de Atauri, S. Marin, and M. Cascante, "Tracing metabolic fluxes using mass spectrometry: Stable isotope-resolved metabolomics in health and disease," *TrAC Trends in Analytical Chemistry*, vol. 120, p. 115371, Nov. 2019, doi: 10.1016/J.TRAC.2018.12.025.

[8] L. Perez de Souza, S. Alseekh, F. Scossa, and A. R. Fernie, "Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research," *Nature Methods 2021 18:7*, vol. 18, no. 7, pp. 733–746, May 2021, doi: 10.1038/s41592-021-01116-4.

[9] J. G. Jones, M. A. Solomon, S. M. Cole, A. D. Sherry, and C. R. Malloy, "An integrated 2H and 13C NMR study of gluconeogenesis and TCA cycle flux in humans," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 281, no. 4 44-4, 2001, doi: 10.1152/AJPENDO.2001.281.4.E848/ASSET/IMAGES/LARGE/H11010544007.JPEG.

[10] L. M. Blank, R. R. Desphande, A. Schmid, and H. Hayen, "Analysis of carbon and nitrogen co-metabolism in yeast by ultrahigh-resolution mass spectrometry applying 13C- and 15N-labeled substrates simultaneously," *Analytical and bioanalytical chemistry*, vol. 403, no. 8, pp. 2291–2305, Jun. 2012, doi: 10.1007/S00216-012-6009-4.

[11] C. M. Hasenour *et al.*, "Mass spectrometry-based microassay of (2)H and (13)C plasma glucose labeling to quantify liver metabolic fluxes in vivo," *American journal of physiology. Endocrinology and metabolism*, vol. 309, no. 2, pp. E191–E203, Jul. 2015, doi: 10.1152/AJPENDO.00003.2015.

[12] C. M. Hasenour *et al.*, "Liver AMP-Activated Protein Kinase Is Unnecessary for Gluconeogenesis but Protects Energy State during Nutrient Deprivation," *PLoS ONE*, vol. 12, no. 1, Jan. 2017, doi: 10.1371/JOURNAL.PONE.0170382.

[13] E. Gorrochategui, J. Jaumot, S. Lacorte, and R. Tauler, "Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow," *TrAC*

Trends in Analytical Chemistry, vol. 82, pp. 425–442, Sep. 2016, doi: 10.1016/J.TRAC.2016.07.004.

[14] C. P. Long and M. R. Antoniewicz, "High-resolution 13C metabolic flux analysis," *Nature Protocols 2019 14:10*, vol. 14, no. 10, pp. 2856–2877, Aug. 2019, doi: 10.1038/s41596-019-0204-0.

[15] T. A. Murphy and J. D. Young, "ETA: Robust software for determination of cell specific rates from extracellular time courses," *Biotechnology and Bioengineering*, vol. 110, no. 6, pp. 1748–1758, Jun. 2013, doi: 10.1002/BIT.24836.

[16] J. D. Young, "INCA: A computational platform for isotopically non-stationary metabolic flux analysis," *Bioinformatics*, vol. 30, no. 9, pp. 1333–1335, 2014, doi: 10.1093/bioinformatics/btu015.

[17] D. I. Papac and Z. Shahrokh, "Mass spectrometry innovations in drug discovery and development," *Pharmaceutical Research*, vol. 18, no. 2, pp. 131–145, 2001, doi: 10.1023/A:1011049231231/METRICS.

[18] Z. Meng, T. A. Simmons-Willis, and P. A. Limbach, "The use of mass spectrometry in genomics," *Biomolecular Engineering*, vol. 21, no. 1, pp. 1–13, 2004, doi: 10.1016/j.bioeng.2003.08.001.

[19] W. C. S. Cho, "Proteomics Technologies and Challenges," *Genomics, Proteomics & Bioinformatics*, vol. 5, no. 2, pp. 77–85, Jan. 2007, doi: 10.1016/S1672-0229(07)60018-7.

[20] F. Meissner, J. Geddes-McAlister, M. Mann, and M. Bantscheff, "The emerging role of mass spectrometry-based proteomics in drug discovery," *Nature Reviews Drug Discovery 2022 21:9*, vol. 21, no. 9, pp. 637–654, Mar. 2022, doi: 10.1038/s41573-022-00409-3.

[21] C. G. Herbert and R. A. W. Johnstone, *Mass Spectrometry Basics*. CRC Press, 2002. [Online]. Available: https://books.google.com/books?id=983ZkPwXXNYC

[22] I. V. Chernushevich, A. V. Loboda, and B. A. Thomson, "An introduction to quadrupole–time-of-flight mass spectrometry," *Journal of Mass Spectrometry*, vol. 36, no. 8, pp. 849–865, 2001, doi: 10.1002/jms.207.

[23] S. U. a. H. Syed, S. Maher, and S. Taylor, "Quadrupole mass filter operation under the influence of magnetic field," *Journal of Mass Spectrometry*, vol. 48, no. 12, pp. 1325–1339, 2013, doi: 10.1002/jms.3293.

[24] S. Maher, S. U. Syed, D. M. Hughes, J. R. Gibson, and S. Taylor, "Mapping the Stability Diagram of a Quadrupole Mass Spectrometer with a Static Transverse Magnetic Field Applied," *J. Am. Soc. Mass Spectrom.*, vol. 24, no. 8, pp. 1307–1314, Aug. 2013, doi: 10.1007/s13361-013-0654-5.

[25] W. C. Wiley and I. H. McLaren, "Time-of-Flight Mass Spectrometer with Improved Resolution," *Review of Scientific Instruments*, vol. 26, no. 12, pp. 1150–1157, Dec. 2004, doi: 10.1063/1.1715212.

[26] K. Saito and F. Matsuda, "Metabolomics for Functional Genomics, Systems Biology, and Biotechnology," *https://doi.org/10.1146/annurev.arplant.043008.092035*, vol. 61, pp. 463–489, May 2010, doi: 10.1146/ANNUREV.ARPLANT.043008.092035.

[27] H. M. McNair, J. M. Miller, and N. H. Snow, *Basic Gas Chromatography*. Wiley, 2019. [Online]. Available: https://books.google.com/books?id=VimjDwAAQBAJ

[28] T. O. Metz *et al.*, "Future of liquid chromatography–mass spectrometry in metabolic profiling and metabolomic studies for biomarker discovery," *http://dx.doi.org/10.2217/17520363.1.1.159*, vol. 1, no. 1, pp. 159–185, May 2007, doi: 10.2217/17520363.1.1.159.

[29] H. G. Gika, G. A. Theodoridis, R. S. Plumb, and I. D. Wilson, “Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics,” *Journal of Pharmaceutical and Biomedical Analysis*, vol. 87, pp. 12–25, Jan. 2014, doi: 10.1016/J.JPBA.2013.06.032.
[30] K. Dettmer, P. A. Aronov, and B. D. Hammock, “Mass spectrometry-based metabolomics,” *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 51–78, Jan. 2007, doi: 10.1002/MAS.20108.
[31] B. P. Bowen and T. R. Northen, “Dealing with the unknown: Metabolomics and metabolite atlases,” *Journal of the American Society for Mass Spectrometry*, vol. 21, no. 9, pp. 1471–1476, Sep. 2010, doi: 10.1016/J.JASMS.2010.04.003/METRICS.
[32] M. Anbar and W. H. Aberth, “Field Ionization Mass Spectrometry: A New Tool for the Analytical Chemist,” *Analytical Chemistry*, vol. 46, no. 1, pp. 59A-64A, Jan. 1974, doi: 10.1021/AC60337A712/ASSET/AC60337A712.FP.PNG_V03.
[33] N. Washida, H. Akimoto, H. Takagi, and M. Okuda, “Gas Chromatography/Photoionization Mass Spectrometry.,” *Analytical Chemistry*, vol. 50, no. 7, pp. 910–915, 1978, doi: 10.1021/AC50029A023/ASSET/AC50029A023.FP.PNG_V03.
[34] J. H. Gross, *Mass Spectrometry: A Textbook*. Springer Berlin Heidelberg, 2006. [Online]. Available: https://books.google.com/books?id=Wr_qCAAAQBAJ
[35] I. Horman and H. Traitler, “Pseudo-molecular ions in ion trap detector electron impact mass spectra: practical consequences,” *Analytical Chemistry*, vol. 61, no. 17, pp. 1983–1984, 1989.
[36] M. Zhu, H. Zhang, and W. G. Humphreys, “Drug Metabolite Profiling and Identification by High-resolution Mass Spectrometry,” *J Biol Chem*, vol. 286, no. 29, pp. 25419–25425, Jul. 2011, doi: 10.1074/jbc.R110.200055.
[37] E. L. Schymanski *et al.*, “Identifying small molecules via high resolution mass spectrometry: Communicating confidence,” *Environmental Science and Technology*, vol. 48, no. 4, pp. 2097–2098, Feb. 2014, doi: 10.1021/ES5002105/ASSET/IMAGES/LARGE/ES-2014-002105_0001.JPEG.
[38] A. K. Shukla and J. H. Futrell, “Tandem mass spectrometry: dissociation of ions by collisional activation,” *Journal of Mass Spectrometry J. Mass Spectrom*, vol. 35, pp. 1069–1090, 2000, doi: 10.1002/1096-9888.
[39] M. L. Vestal and J. M. Campbell, “Tandem Time-of-Flight Mass Spectrometry,” vol. 402, pp. 79–108, Jan. 2005, doi: 10.1016/S0076-6879(05)02003-3.
[40] J. Guo and T. Huan, “Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics,” *Analytical Chemistry*, vol. 92, no. 12, pp. 8072–8080, Jun. 2020, doi: 10.1021/ACS.ANALCHEM.9B05135/SUPPL_FILE/AC9B05135_SI_001.PDF.
[41] A. Doerr, “DIA mass spectrometry,” *Nature Methods 2015 12:1*, vol. 12, no. 1, pp. 35–35, Dec. 2014, doi: 10.1038/nmeth.3234.
[42] A. H. B. Wu, R. Gerona, P. Armenian, D. French, M. Petrie, and K. L. Lynch, “Role of liquid chromatographyhigh-resolution mass spectrometry (LC-HR/MS) in clinical toxicology,” *Clinical Toxicology*, vol. 50, no. 8, pp. 733–742, Sep. 2012, doi: 10.3109/15563650.2012.713108.
[43] Y. H. Lai and Y. S. Wang, “Advances in high-resolution mass spectrometry techniques for analysis of high mass-to-charge ions,” *Mass Spectrometry Reviews*, p. e21790, 2022, doi: 10.1002/MAS.21790.

[44] H. Yamamoto and J. A. McCloskey, “Calculations of Isotopic Distribution in Molecules Extensively Labeled with Heavy Isotopes,” *Analytical Chemistry*, vol. 49, no. 2, pp. 281–283, Feb. 1977, doi: 10.1021/AC50010A025/ASSET/AC50010A025.FP.PNG_V03.
[45] L. Patiny and A. Borel, “ChemCalc: A building block for tomorrow’s chemical infrastructure,” *Journal of Chemical Information and Modeling*, vol. 53, no. 5, pp. 1223–1228, 2013, doi: 10.1021/ci300563h.
[46] E. N. Fung, M. Jemal, and A. F. Aubry, “High-resolution MS in regulated bioanalysis: where are we now and where do we go from here?,” *https://doi.org/10.4155/bio.13.81*, vol. 5, no. 10, pp. 1277–1284, May 2013, doi: 10.4155/BIO.13.81.
[47] S. Alseekh and A. R. Fernie, “Metabolomics 20 years on: what have we learned and what hurdles remain?,” *The Plant Journal*, vol. 94, no. 6, pp. 933–942, Jun. 2018, doi: 10.1111/TPJ.13950.
[48] M. Piraud *et al.*, “ESI-MS/MS analysis of underivatised amino acids: a new tool for the diagnosis of inherited disorders of amino acid metabolism. Fragmentation study of 79 molecules of biological interest in positive and negative ionisation mode,” *Rapid Communications in Mass Spectrometry*, vol. 17, no. 12, pp. 1297–1311, Jun. 2003, doi: 10.1002/RCM.1054.
[49] Z. Lei, D. V. Huhman, and L. W. Sumner, “Mass spectrometry strategies in metabolomics,” *Journal of Biological Chemistry*, vol. 286, no. 29, pp. 25435–25442, Jul. 2011, doi: 10.1074/jbc.R111.238691.
[50] K. Nöh and W. Wiechert, “The benefits of being transient: Isotope-based metabolic flux analysis at the short time scale,” *Applied Microbiology and Biotechnology*, vol. 91, no. 5, pp. 1247–1265, Sep. 2011, doi: 10.1007/S00253-011-3390-4.
[51] Y. E. Cheah and J. D. Young, “Isotopically nonstationary metabolic flux analysis (INST-MFA): putting theory into practice,” *Current Opinion in Biotechnology*, vol. 54, pp. 80–87, Dec. 2018, doi: 10.1016/J.COPBIO.2018.02.013.
[52] M. R. Antoniewicz, “A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications,” *Metabolic Engineering*, vol. 63, pp. 2–12, Jan. 2021, doi: 10.1016/J.YMBEN.2020.11.002.
[53] IUPAC, *Gold book - Compendium of chemical terminology*. 2014.
[54] M. K. Hellerstein and R. A. Neese, “Mass isotopomer distribution analysis at eight years: Theoretical, analytic, and experimental considerations,” *American Journal of Physiology - Endocrinology and Metabolism*, vol. 276, no. 6 39-6, 1999, doi: 10.1152/AJPENDO.1999.276.6.E1146.
[55] D. Weindl, A. Wegner, and K. Hiller, “Non-targeted Tracer Fate Detection,” *Methods in Enzymology*, vol. 561, pp. 277–302, Jan. 2015, doi: 10.1016/BS.MIE.2015.04.003.
[56] J. F. Xiao, B. Zhou, and H. W. Ressom, “Metabolite identification and quantitation in LC-MS/MS-based metabolomics,” *TrAC Trends in Analytical Chemistry*, vol. 32, pp. 1–14, Feb. 2012, doi: 10.1016/J.TRAC.2011.08.009.
[57] Y. Sawada and M. Yokota Hirai, “Integrated LC-MS/MS system for plant metabolomics,” *Computational and Structural Biotechnology Journal*, vol. 4, no. 5, p. e201301011, Jan. 2013, doi: 10.5936/CSBJ.201301011.
[58] L. Cui, H. Lu, and Y. H. Lee, “Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases,” *Mass Spectrometry Reviews*, vol. 37, no. 6, pp. 772–792, Nov. 2018, doi: 10.1002/MAS.21562.

[59] T. W. M. Fan, P. K. Lorkiewicz, K. Sellers, H. N. B. Moseley, R. M. Higashi, and A. N. Lane, "Stable isotope-resolved metabolomics and applications for drug development," *Pharmacology & Therapeutics*, vol. 133, no. 3, pp. 366–391, Mar. 2012, doi: 10.1016/J.PHARMTHERA.2011.12.007.

[60] Y. Yang, T. W. M. Fan, A. N. Lane, and R. M. Higashi, "Chloroformate derivatization for tracing the fate of Amino acids in cells and tissues by multiple stable isotope resolved metabolomics (mSIRM)," *Analytica Chimica Acta*, vol. 976, pp. 63–73, Jul. 2017, doi: 10.1016/J.ACA.2017.04.014.

[61] A. N. Lane, R. M. Higashi, and T. W. M. Fan, "NMR and MS-based Stable Isotope-Resolved Metabolomics and applications in cancer metabolism," *TrAC Trends in Analytical Chemistry*, vol. 120, p. 115322, Nov. 2019, doi: 10.1016/J.TRAC.2018.11.020.

[62] T. Cajka and O. Fiehn, "Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics," *Analytical Chemistry*, vol. 88, no. 1, pp. 524–545, Jan. 2016, doi: 10.1021/ACS.ANALCHEM.5B04491/ASSET/IMAGES/LARGE/AC-2015-04491V_0008.JPEG.

[63] A. C. Schrimpe-Rutledge, S. G. Codreanu, S. D. Sherrod, and J. A. McLean, "Untargeted Metabolomics Strategies—Challenges and Emerging Directions," *Journal of the American Society for Mass Spectrometry*, vol. 27, no. 12, pp. 1897–1905, Dec. 2016, doi: 10.1007/S13361-016-1469-Y/ASSET/IMAGES/LARGE/JS8B05178_0004.JPEG.

[64] T. Züllig, M. Zandl-Lang, M. Trötzmüller, J. Hartler, B. Plecko, and H. C. Köfeler, "A Metabolomics Workflow for Analyzing Complex Biological Samples Using a Combined Method of Untargeted and Target-List Based Approaches," *Metabolites*, vol. 10, no. 9, Art. no. 9, Sep. 2020, doi: 10.3390/metabo10090342.

[65] S. D. Patterson, "Data analysis—the Achilles heel of proteomics," *Nature Biotechnology 2003 21:3*, vol. 21, no. 3, pp. 221–222, Mar. 2003, doi: 10.1038/nbt0303-221.

[66] M. Askenazi, J. R. Parikh, and J. A. Marto, "mzAPI: a new strategy for efficiently sharing mass spectrometry data," *Nature methods*, vol. 6, no. 4, p. 240, 2009, doi: 10.1038/NMETH0409-240.

[67] R. Rew and G. Davis, "NetCDF: An Interface for Scientific Data Access," *IEEE Computer Graphics and Applications*, vol. 10, no. 4, pp. 76–82, 1990, doi: 10.1109/38.56302.

[68] L. Martens *et al.*, "mzML—a Community Standard for Mass Spectrometry Data," *Molecular & Cellular Proteomics : MCP*, vol. 10, no. 1, Jan. 2011, doi: 10.1074/MCP.R110.000133.

[69] M. Turewicz and E. W. Deutsch, "Spectra, chromatograms, Metadata: mzML-the standard data format for mass spectrometer output," *Methods in molecular biology (Clifton, N.J.)*, vol. 696, pp. 179–203, 2011, doi: 10.1007/978-1-60761-987-1_11.

[70] P. G. A. Pedrioli *et al.*, "A common open representation of mass spectrometry data and its application to proteomics research," *Nature biotechnology*, vol. 22, no. 11, pp. 1459–1466, 2004, doi: 10.1038/NBT1031.

[71] M. C. Chambers *et al.*, "A Cross-platform Toolkit for Mass Spectrometry and Proteomics," *Nature biotechnology*, vol. 30, no. 10, p. 918, Oct. 2012, doi: 10.1038/NBT.2377.

[72] N. Dyson and R. M. Smith, *Chromatographic Integration Methods*, 2nd ed., vol. 3. in RSC Chromatography Monographs, vol. 3. Cambridge: Royal Society of Chemistry, 1998.

[73] F. Chau and A. Kai-man Leung, “Chapter 9 - Application of Wavelet Transform in Processing Chromatographic Data,” in *Data Handling in Science and Technology*, vol. 22, B. Walczak, Ed., in Wavelets in Chemistry, vol. 22. , Elsevier, 2000, pp. 205–223. doi: 10.1016/S0922-3487(00)80034-9.

[74] J. H. Christensen, J. Mortensen, A. B. Hansen, and O. Andersen, “Chromatographic preprocessing of GC–MS data for analysis of complex chemical mixtures,” *Journal of Chromatography A*, vol. 1062, no. 1, pp. 113–123, Jan. 2005, doi: 10.1016/j.chroma.2004.11.037.

[75] Z. Wang, M. Zhang, and P. de B. Harrington, “Comparison of Three Algorithms for the Baseline Correction of Hyphenated Data Objects,” *Anal. Chem.*, vol. 86, no. 18, pp. 9050–9057, Sep. 2014, doi: 10.1021/ac501658k.

[76] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, and F. Suits, “A noise model for mass spectrometry based proteomics,” *Bioinformatics*, vol. 24, no. 8, pp. 1070–1077, Apr. 2008, doi: 10.1093/bioinformatics/btn078.

[77] S. Medhe, “Mass Spectrometry: Detectors Review,” *Chemical and Biomolecular Engineering*, vol. 3, no. 4, pp. 51–58, 2018, doi: 10.11648/j.cbe.20180304.11.

[78] A. N. Krutchinsky and B. T. Chait, “On the nature of the chemical noise in MALDI mass spectra,” *Journal of the American Society for Mass Spectrometry*, vol. 13, no. 2, pp. 129–134, Feb. 2002, doi: 10.1016/S1044-0305(01)00336-1.

[79] J. Ding, J. Shi, G. G. Poirier, and F.-X. Wu, “A novel approach to denoising ion trap tandem mass spectra,” *Proteome Sci*, vol. 7, no. 1, p. 9, Mar. 2009, doi: 10.1186/1477-5956-7-9.

[80] J. Li, W. Gao, H. Wu, S. Shi, J. Yu, and K. Tang, “Application of zero-phase digital filtering for effective denoising of field asymmetric waveform ion mobility spectrometry signal,” *Rapid Communications in Mass Spectrometry*, vol. 36, no. 1, p. e9211, 2022, doi: 10.1002/rcm.9211.

[81] E. Mostacci, C. Truntzer, H. Cardot, and P. Ducoroy, “Multivariate denoising methods combining wavelets and principal component analysis for mass spectrometry data,” *PROTEOMICS*, vol. 10, no. 14, pp. 2564–2572, 2010, doi: 10.1002/pmic.200900185.

[82] L. Chiron, M. A. van Agthoven, B. Kieffer, C. Rolando, and M.-A. Delsuc, “Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 4, pp. 1385–1390, Jan. 2014, doi: 10.1073/pnas.1306700111.

[83] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, “Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping,” *Journal of Chromatography A*, vol. 805, no. 1, pp. 17–35, May 1998, doi: 10.1016/S0021-9673(98)00021-1.

[84] W. Yu, Z. He, J. Liu, and H. Zhao, “Improving Mass Spectrometry Peak Detection Using Multiple Peak Alignment Results,” *J. Proteome Res.*, vol. 7, no. 1, pp. 123–129, Jan. 2008, doi: 10.1021/pr070370n.

[85] A. Antoniadis, J. Bigot, and S. Lambert-Lacroix, “Peaks detection and alignment for mass spectrometry data,” *Journal de la société française de statistique*, vol. 151, no. 1, pp. 17–37, 2010.

[86] J. Wang and H. Lam, “Graph-based peak alignment algorithms for multiple liquid chromatography-mass spectrometry datasets,” *Bioinformatics*, vol. 29, no. 19, pp. 2469–2476, Oct. 2013, doi: 10.1093/bioinformatics/btt435.
[87] M. Li and X. R. Wang, “Peak alignment of gas chromatography–mass spectrometry data with deep learning,” *Journal of Chromatography A*, vol. 1604, p. 460476, Oct. 2019, doi: 10.1016/j.chroma.2019.460476.
[88] W. Wang *et al.*, “Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards,” *Analytical Chemistry*, vol. 75, no. 18, pp. 4818–4826, 2003, doi: 10.1021/ac026468x.
[89] M. Katajamaa and M. Orešič, “Processing methods for differential analysis of LC/MS profile data,” *BMC Bioinformatics*, vol. 6, 2005, doi: 10.1186/1471-2105-6-179.
[90] C. A. Smith, E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak, “XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification,” *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787, 2006, doi: 10.1021/ac051437y.
[91] D. Radulovic *et al.*, “Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry*,” *Molecular & Cellular Proteomics*, vol. 3, no. 10, pp. 984–997, Oct. 2004, doi: 10.1074/mcp.M400061-MCP200.
[92] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič, “MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data,” *BMC Bioinformatics*, vol. 11, 2010, doi: 10.1186/1471-2105-11-395.
[93] H. Tian, B. Li, and G. Shui, “Untargeted LC–MS Data Preprocessing in Metabolomics,” *J. Anal. Test.*, vol. 1, no. 3, pp. 187–192, Jul. 2017, doi: 10.1007/s41664-017-0030-8.
[94] R. Tautenhahn, G. J. Patti, D. Rinehart, and G. Siuzdak, “XCMS online: A web-based platform to process untargeted metabolomic data,” *Analytical Chemistry*, vol. 84, no. 11, pp. 5035–5039, 2012, doi: 10.1021/ac300698c.
[95] I. Gertsman and B. A. Barshop, “Promises and Pitfalls of Untargeted Metabolomics,” *J Inherit Metab Dis*, vol. 41, no. 3, pp. 355–366, May 2018, doi: 10.1007/s10545-017-0130-7.
[96] R. Madsen, T. Lundstedt, and J. Trygg, “Chemometrics in metabolomics—A review in human disease diagnosis,” *Analytica Chimica Acta*, vol. 659, no. 1, pp. 23–33, Feb. 2010, doi: 10.1016/j.aca.2009.11.042.
[97] D. I. Broadhurst and D. B. Kell, “Statistical strategies for avoiding false discoveries in metabolomics and related experiments,” *Metabolomics*, vol. 2, no. 4, pp. 171–196, Dec. 2006, doi: 10.1007/s11306-006-0037-z.
[98] L. G. Rasmussen, F. Savorani, T. M. Larsen, L. O. Dragsted, A. Astrup, and S. B. Engelsen, “Standardization of factors that influence human urine metabolomics,” *Metabolomics*, vol. 7, no. 1, pp. 71–83, Mar. 2011, doi: 10.1007/s11306-010-0234-7.
[99] M. K. Townsend *et al.*, “Reproducibility of Metabolomic Profiles among Men and Women in 2 Large Cohort Studies,” *Clinical Chemistry*, vol. 59, no. 11, pp. 1657–1667, Nov. 2013, doi: 10.1373/clinchem.2012.199133.
[100] E. Saccenti, H. C. J. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. W. B. Hendriks, “Reflections on univariate and multivariate analysis of metabolomics data,” *Metabolomics*, vol. 10, no. 3, pp. 361–374, Jun. 2014, doi: 10.1007/s11306-013-0598-6.

[101] A. Alonso, S. Marsal, and A. Julià, "Analytical Methods in Untargeted Metabolomics: State of the Art in 2015," *Frontiers in Bioengineering and Biotechnology*, vol. 3, 2015, Accessed: Nov. 24, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2015.00023
[102] "Computational and statistical analysis of metabolomics data | Metabolomics." Accessed: Nov. 24, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s11306-015-0823-6
[103] I. T. Jolliffe, Ed., "Principal Component Analysis for Special Types of Data," in *Principal Component Analysis*, New York, NY: Springer New York, 2002, pp. 338–372. doi: 10.1007/0-387-22440-8_13.
[104] T. Kohonen, "Learning Vector Quantisation and the Self Organising Map," in *Theory and Applications of Neural Networks*, J. G. Taylor and C. L. T. Mannion, Eds., London: Springer London, 1992, pp. 235–242.
[105] E. D. Wallace, D. A. Todd, J. M. Harnly, N. B. Cech, and J. J. Kellogg, "Identification of adulteration in botanical samples with untargeted metabolomics," *Anal Bioanal Chem*, vol. 412, no. 18, pp. 4273–4286, Jul. 2020, doi: 10.1007/s00216-020-02678-6.
[106] R. C. Eldridge *et al.*, "Unsupervised Hierarchical Clustering of Head and Neck Cancer Patients by Pre-Treatment Plasma Metabolomics Creates Prognostic Metabolic Subtypes," *Cancers*, vol. 15, no. 12, Art. no. 12, Jan. 2023, doi: 10.3390/cancers15123184.
[107] C. Bailleux *et al.*, "Survival analysis of patient groups defined by unsupervised machine learning clustering methods based on patient metabolomic data.," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 5136–5143, Jan. 2023, doi: 10.1016/j.csbj.2023.10.033.
[108] Y. Li *et al.*, "Coupling proteomics and metabolomics for the unsupervised identification of protein–metabolite interactions in Chaetomium thermophilum," *PLOS ONE*, vol. 16, no. 7, p. e0254429, Jul. 2021, doi: 10.1371/journal.pone.0254429.
[109] O. Beckonert, J. Monnerjahn, U. Bonk, and D. Leibfritz, "Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps," *NMR in Biomedicine*, vol. 16, no. 1, pp. 1–11, 2003, doi: 10.1002/nbm.797.
[110] D. H. Milone, G. Stegmayer, M. López, L. Kamenetzky, and F. Carrari, "Improving clustering with metabolic pathway data," *BMC Bioinformatics*, vol. 15, no. 1, p. 101, Apr. 2014, doi: 10.1186/1471-2105-15-101.
[111] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," in *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, NLD: IOS Press, Jun. 2007, pp. 3–24.
[112] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," *Behavior Therapy*, vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/j.beth.2020.05.002.
[113] J. Xia, D. I. Broadhurst, M. Wilson, and D. S. Wishart, "Translational biomarker discovery in clinical metabolomics: an introductory tutorial," *Metabolomics*, vol. 9, no. 2, pp. 280–299, Apr. 2013, doi: 10.1007/s11306-012-0482-9.
[114] Q. Yang, S.-S. Lin, J.-T. Yang, L.-J. Tang, and R.-Q. Yu, "Detection of inborn errors of metabolism utilizing GC-MS urinary metabolomics coupled with a modified orthogonal partial least squares discriminant analysis," *Talanta*, vol. 165, pp. 545–552, Apr. 2017, doi: 10.1016/j.talanta.2017.01.018.

[115] E. B. Evangelista *et al.*, “Phospholipids are A Potentially Important Source of Tissue Biomarkers for Hepatocellular Carcinoma: Results of a Pilot Study Involving Targeted Metabolomics,” *Diagnostics*, vol. 9, no. 4, Art. no. 4, Dec. 2019, doi: 10.3390/diagnostics9040167.
[116] V. Pinto-Plata *et al.*, “Plasma metabolomics and clinical predictors of survival differences in COPD patients,” *Respir Res*, vol. 20, no. 1, p. 219, Oct. 2019, doi: 10.1186/s12931-019-1167-y.
[117] I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” *SN Comput Sci*, vol. 2, no. 6, p. 420, 2021, doi: 10.1007/s42979-021-00815-1.
[118] Y. Gloaguen, J. A. Kirwan, and D. Beule, “Deep Learning-Assisted Peak Curation for Large-Scale LC-MS Metabolomics,” *Analytical Chemistry*, vol. 94, no. 12, pp. 4930–4937, Mar. 2022, doi: 10.1021/ACS.ANALCHEM.1C02220/ASSET/IMAGES/LARGE/AC1C02220_0005.JPEG.
[119] Y. Feng *et al.*, “Novel method for rapid identification of Listeria monocytogenes based on metabolomics and deep learning,” *Food Control*, vol. 139, p. 109042, Sep. 2022, doi: 10.1016/j.foodcont.2022.109042.
[120] D. D. Matyushin, A. Yu. Sholokhova, and A. K. Buryak, “Deep Learning Driven GC-MS Library Search and Its Application for Metabolomics,” *Anal. Chem.*, vol. 92, no. 17, pp. 11818–11825, Sep. 2020, doi: 10.1021/acs.analchem.0c02082.
[121] X. Huang, Y. J. Chen, K. Cho, I. Nikolskiy, P. A. Crawford, and G. J. Patti, “X13CMS: Global tracking of isotopic labels in untargeted metabolomics,” *Analytical Chemistry*, vol. 86, no. 3, pp. 1632–1639, 2014, doi: 10.1021/ac403384n.
[122] E. M. Llufrio, K. Cho, and G. J. Patti, “Systems-level analysis of isotopic labeling in untargeted metabolomic data by X13CMS,” *Nat Protoc*, vol. 14, no. 7, Art. no. 7, Jul. 2019, doi: 10.1038/s41596-019-0167-1.
[123] N. A. Reisdorph, S. Walmsley, and R. Reisdorph, “A perspective and framework for developing sample type specific databases for LC/MS-based clinical metabolomics,” *Metabolites*, vol. 10, no. 1, Jan. 2020, doi: 10.3390/METABO10010008.
[124] C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, and S. Neumann, “CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets,” *Analytical Chemistry*, vol. 84, no. 1, pp. 283–289, 2012, doi: 10.1021/ac202450g.
[125] H. Tsugawa *et al.*, “MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis,” *Nature Methods*, vol. 12, no. 6, pp. 523–526, 2015, doi: 10.1038/nmeth.3393.
[126] K. Dührkop *et al.*, “SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information,” *Nature Methods*, vol. 16, no. 4, pp. 299–302, 2019, doi: 10.1038/s41592-019-0344-8.
[127] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker, “Searching molecular structure databases with tandem mass spectra using CSI:FingerID,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 41, pp. 12580–12585, Oct. 2015, doi: 10.1073/PNAS.1509788112/SUPPL_FILE/PNAS.201509788SI.PDF.

[128] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos, “Accurate assessment of amino acid mass isotopomer distributions for metabolic flux analysis,” *Analytical Chemistry*, vol. 79, no. 19, pp. 7554–7559, 2007, doi: 10.1021/ac0708893.
[129] L. Krämer, C. Jäger, J. P. Trezzi, D. M. Jacobs, and K. Hiller, “Quantification of Stable Isotope Traces Close to Natural Enrichment in Human Plasma Metabolites Using Gas Chromatography-Mass Spectrometry,” *Metabolites 2018, Vol. 8, Page 15*, vol. 8, no. 1, p. 15, Feb. 2018, doi: 10.3390/METABO8010015.
[130] M. F. Clasquin, E. Melamud, and J. D. Rabinowitz, “LC-MS data processing with MAVEN: A metabolomic analysis and visualization engine,” *Current Protocols in Bioinformatics*, vol. 37, no. 1, 2012, doi: 10.1002/0471250953.bi1411s37.
[131] S. Agrawal *et al.*, “EL-MAVEN: A fast, robust, and user-friendly mass spectrometry data processing engine for metabolomics,” *Methods in Molecular Biology*, vol. 1978, 2019, doi: 10.1007/978-1-4939-9236-2_19.
[132] B. MacLean *et al.*, “Skyline: An open source document editor for creating and analyzing targeted proteomics experiments,” *Bioinformatics*, vol. 26, no. 7, pp. 966–968, 2010, doi: 10.1093/bioinformatics/btq054.
[133] R. D. Vocke, “Errata: Atomic weights of the elements: (Technical Report): (Pure Appl. Chem., (1999), 71(8), (1593–1607) 10.1351/pac199971081593),” *Pure and Applied Chemistry*, vol. 71, no. 9, p. 1808, Sep. 1999, doi: 10.1351/PAC199971091808/MACHINEREADABLECITATION/RIS.
[134] C. A. Fernandez, C. Des Rosiers, S. F. Previs, F. David, and H. Brunengraber, “Correction of 13C mass isotopomer distributions for natural stable isotope abundance,” *Journal of Mass Spectrometry*, vol. 31, no. 3, pp. 255–262, 1996, doi: 10.1002/(SICI)1096-9888(199603)31:3<255::AID-JMS290>3.0.CO;2-3.
[135] W. A. Van Winden, C. Wittmann, E. Heinzle, and J. J. Heijnen, “Correcting mass isotopomer distributions for naturally occurring isotopes,” *Biotechnology and Bioengineering*, vol. 80, no. 4, pp. 477–479, Nov. 2002, doi: 10.1002/BIT.10393.
[136] J. Choi and M. R. Antoniewicz, “Tandem mass spectrometry for 13C metabolic flux analysis: Methods and algorithms based on EMU framework,” *Frontiers in Microbiology*, vol. 10, no. JAN, 2019, doi: 10.3389/fmicb.2019.00031.
[137] Y. Wang, L. R. Parsons, and X. Su, “AccuCor2: isotope natural abundance correction for dual-isotope tracer experiments,” *Laboratory Investigation*, vol. 101, no. 10, pp. 1403–1410, 2021, doi: 10.1038/s41374-021-00631-4.
[138] R. Raaisa, S. Lathwal, V. Chubukov, R. G. Kibbey, and A. K. Jha, “>Corna - An Open Source Python Tool For Natural Abundance Correction In Isotope Tracer Experiments,” *bioRxiv*, p. 2020.09.19.304741, Sep. 2020, doi: 10.1101/2020.09.19.304741.
[139] B. H. Banimfreg, A. Shamayleh, and H. Alshraideh, “Survey for Computer-Aided Tools and Databases in Metabolomics,” *Metabolites 2022, Vol. 12, Page 1002*, vol. 12, no. 10, p. 1002, Oct. 2022, doi: 10.3390/METABO12101002.
[140] S. Li *et al.*, “Predicting Network Activity from High Throughput Metabolomics,” *PLoS Computational Biology*, vol. 9, no. 7, 2013, doi: 10.1371/journal.pcbi.1003123.
[141] X. Xu *et al.*, “Autophagy is essential for effector CD8 T cell survival and memory formation,” *Nat Immunol*, vol. 15, no. 12, pp. 1152–1161, Dec. 2014, doi: 10.1038/ni.3025.
[142] J. M. Hoffman, Q. A. Soltow, S. Li, A. Sidik, D. P. Jones, and D. E. L. Promislow, “Effects of age, sex, and genotype on high-sensitivity metabolomic profiles in the fruit fly,

Drosophila melanogaster," *Aging Cell*, vol. 13, no. 4, pp. 596–604, 2014, doi: 10.1111/acel.12215.
[143] G. Basler, A. R. Fernie, and Z. Nikoloski, "Advances in metabolic flux analysis toward genome-scale profiling of higher organisms," *Bioscience Reports*, vol. 38, no. 6, p. BSR20170224, Nov. 2018, doi: 10.1042/BSR20170224.
[144] J. Yuan, B. D. Bennett, and J. D. Rabinowitz, "Kinetic flux profiling for quantitation of cellular metabolic fluxes," *Nature protocols*, vol. 3, no. 8, p. 1328, 2008, doi: 10.1038/NPROT.2008.131.
[145] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos, "Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions," *Metabolic Engineering*, vol. 9, no. 1, pp. 68–86, 2007, doi: 10.1016/j.ymben.2006.09.001.
[146] H. Shimizu, "Metabolic engineering — Integrating methodologies of molecular breeding and bioprocess systems engineering," *Journal of Bioscience and Bioengineering*, vol. 94, no. 6, pp. 563–573, Dec. 2002, doi: 10.1016/S1389-1723(02)80196-7.
[147] N. Zamboni, "13C metabolic flux analysis in complex systems," *Current Opinion in Biotechnology*, vol. 22, no. 1, pp. 103–108, Feb. 2011, doi: 10.1016/J.COPBIO.2010.08.009.
[148] S. B. Crown and M. R. Antoniewicz, "Parallel labeling experiments and metabolic flux analysis: Past, present and future methodologies," *Metabolic Engineering*, vol. 16, no. 1, pp. 21–32, Mar. 2013, doi: 10.1016/J.YMBEN.2012.11.010.
[149] A. Karnovsky and S. Li, "Pathway Analysis for Targeted and Untargeted Metabolomics," *Methods in Molecular Biology*, vol. 2104, pp. 387–400, 2020, doi: 10.1007/978-1-0716-0239-3_19/FIGURES/3.
[150] Y. Wang, S. Hui, F. E. Wondisford, and X. Su, "Utilizing tandem mass spectrometry for metabolic flux analysis," *Laboratory Investigation 2020 101:4*, vol. 101, no. 4, pp. 423–429, Sep. 2020, doi: 10.1038/s41374-020-00488-z.
[151] L. Liang, F. Sun, H. Wang, and Z. Hu, "Metabolomics, metabolic flux analysis and cancer pharmacology," *Pharmacology & Therapeutics*, vol. 224, p. 107827, Aug. 2021, doi: 10.1016/J.PHARMTHERA.2021.107827.
[152] M. Wagner, D. Naik, and A. Pothen, "Protocols for disease classification from mass spectrometry data," *PROTEOMICS*, vol. 3, no. 9, pp. 1692–1698, Sep. 2003, doi: 10.1002/PMIC.200300519.
[153] P. Hernandez, M. Müller, and R. D. Appel, "Automated protein identification by tandem mass spectrometry: Issues and strategies," *Mass Spectrometry Reviews*, vol. 25, no. 2, pp. 235–254, Mar. 2006, doi: 10.1002/MAS.20068.
[154] K. Guo and L. Li, "High-performance isotope labeling for profiling carboxylic acid-containing metabolites in biofluids by mass spectrometry," *Analytical Chemistry*, vol. 82, no. 21, pp. 8789–8793, Nov. 2010, doi: 10.1021/AC102146G/SUPPL_FILE/AC102146G_SI_006.XLS.
[155] S. Eliuk and A. Makarov, "Evolution of Orbitrap Mass Spectrometry Instrumentation," *https://doi.org/10.1146/annurev-anchem-071114-040325*, vol. 8, pp. 61–80, Jul. 2015, doi: 10.1146/ANNUREV-ANCHEM-071114-040325.
[156] D. K. Allen and J. D. Young, "Tracing metabolic flux through time and space with isotope labeling experiments," *Current Opinion in Biotechnology*, vol. 64, pp. 92–100, Aug. 2020, doi: 10.1016/J.COPBIO.2019.11.003.

[157] Q. Sun, T. W. M. Fan, A. N. Lane, and R. M. Higashi, “Applications of chromatography-ultra high-resolution MS for stable isotope-resolved metabolomics (SIRM) reconstruction of metabolic networks,” *TrAC - Trends in Analytical Chemistry*, vol. 123, 2020, doi: 10.1016/j.trac.2019.115676.
[158] J. Capellades *et al.*, “GeoRge: A Computational Tool to Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics,” *Analytical Chemistry*, vol. 88, no. 1, pp. 621–628, 2016, doi: 10.1021/acs.analchem.5b03628.
[159] A. Chokkathukalam, A. Jankevics, D. J. Creek, F. Achcar, M. P. Barrett, and R. Breitling, “MzMatch-ISO: An R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data,” *Bioinformatics*, vol. 29, no. 2, pp. 281–283, 2013, doi: 10.1093/bioinformatics/bts674.
[160] J. F. Thompson, J. T. Madison, and A. maria E. Muenster, “In vitro Culture of Immature Cotyledons of Soya Bean (Glycine max L. Merr.),” *Annals of Botany*, vol. 41, no. 1, pp. 29–39, Jan. 1977, doi: 10.1093/OXFORDJOURNALS.AOB.A085281.
[161] F. C. Hsu and R. L. Obendorf, “Compositional analysis of in vitro matured soybean seeds,” *Plant Science Letters*, vol. 27, no. 2, pp. 129–135, Oct. 1982, doi: 10.1016/0304-4211(82)90141-9.
[162] S. Kambhampati *et al.*, “Temporal changes in metabolism late in seed development affect biomass composition,” *Plant Physiology*, vol. 186, no. 2, pp. 874–890, Jun. 2021, doi: 10.1093/PLPHYS/KIAB116.
[163] J. Blagih *et al.*, “The Energy Sensor AMPK Regulates T Cell Metabolic Adaptation and Effector Responses In Vivo,” *Immunity*, vol. 42, no. 1, pp. 41–54, Jan. 2015, doi: 10.1016/j.immuni.2014.12.030.
[164] E. H. Ma *et al.*, “Serine Is an Essential Metabolite for Effector T Cell Expansion,” *Cell Metabolism*, vol. 25, no. 2, pp. 345–357, Feb. 2017, doi: 10.1016/j.cmet.2016.12.011.
[165] I. Kaymak *et al.*, “Carbon source availability drives nutrient utilization in CD8+ T cells,” *Cell Metab*, vol. 34, no. 9, pp. 1298-1311.e6, Sep. 2022, doi: 10.1016/j.cmet.2022.07.012.
[166] E. H. Ma *et al.*, “Metabolic Profiling Using Stable Isotope Tracing Reveals Distinct Patterns of Glucose Utilization by Physiologically Activated CD8+ T Cells,” *Immunity*, vol. 51, no. 5, pp. 856-870.e5, Nov. 2019, doi: 10.1016/j.immuni.2019.09.003.
[167] E. G. Bligh and W. J. Dyer, “A rapid method of total lipid extraction and purification,” *Can. J. Biochem. Physiol.*, vol. 37, no. 8, pp. 911–917, Aug. 1959, doi: 10.1139/o59-099.
[168] C. M. Hasenour *et al.*, “Vitamin E does not prevent Western diet-induced NASH progression and increases metabolic flux dysregulation in mice,” *Journal of Lipid Research*, vol. 61, no. 5, pp. 707–721, May 2020, doi: 10.1194/JLR.RA119000183.
[169] G. Libiseller *et al.*, “IPO: A tool for automated optimization of XCMS parameters,” *BMC Bioinformatics*, vol. 16, no. 1, 2015, doi: 10.1186/s12859-015-0562-8.
[170] S. Kim, P. A. Thiessen, T. Cheng, B. Yu, and E. E. Bolton, “An update on PUG-REST: RESTful interface for programmatic access to PubChem,” *Nucleic Acids Research*, vol. 46, no. Web Server issue, p. W563, Jul. 2018, doi: 10.1093/NAR/GKY294.
[171] Klaus Biemann, *Mass Spectrometry: Organic Chemical Applications*, 1st ed. McGraw-Hill, 1962.
[172] N. Huang, M. M. Siegel, G. H. Kruppa, and F. H. Laukien, “Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data,” *Journal*

of the American Society for Mass Spectrometry, vol. 10, no. 11, pp. 1166–1173, 1999, doi: 10.1016/S1044-0305(99)00089-6.
[173] M. K. Hellerstein and R. A. Neese, "Mass isotopomer distribution analysis: a technique for measuring biosynthesis and turnover of polymers," *https://doi.org/10.1152/ajpendo.1992.263.5.E988*, vol. 263, no. 5 26-5, 1992, doi: 10.1152/AJPENDO.1992.263.5.E988.
[174] C. Papageorgopoulos, K. Caldwell, C. Shackleton, H. Schweingrubber, and M. K. Hellerstein, "Measuring Protein Synthesis by Mass Isotopomer Distribution Analysis (MIDA)," *Analytical Biochemistry*, vol. 267, no. 1, pp. 1–16, Feb. 1999, doi: 10.1006/ABIO.1998.2958.
[175] S. Kawashima, T. Katayama, Y. Sato, and M. Kanehisa, "KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System," *Genome Informatics*, vol. 14, pp. 673–674, 2003, doi: 10.11234/GI1990.14.673.
[176] N. Cocco, M. Llabrés, M. Reyes-Prieto, and M. Simeoni, "MetNet: A two-level approach to reconstructing and comparing metabolic networks," *PLoS ONE*, vol. 16, no. 2, Feb. 2021, doi: 10.1371/JOURNAL.PONE.0246962.
[177] J. M. Posma, S. L. Robinette, E. Holmes, and J. K. Nicholson, "MetaboNetworks, an interactive Matlab-based toolbox for creating, customizing and exploring sub-networks from KEGG," *Bioinformatics*, 2014, doi: 10.1093/bioinformatics/btt612.
[178] M. C. Dange *et al.*, "Evaluation of freely available software tools for untargeted quantification of 13C isotopic enrichment in cellular metabolome from HR-LC/MS data," *Metabolic Engineering Communications*, vol. 10, p. e00120, Jun. 2020, doi: 10.1016/j.mec.2019.e00120.
[179] W. B. Dunn *et al.*, "Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics," *Metabolomics*, vol. 9, no. SUPPL.1, pp. 44–66, May 2013, doi: 10.1007/S11306-012-0434-4/FIGURES/7.
[180] F. Giacomoni *et al.*, "Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics," *Bioinformatics*, vol. 31, no. 9, pp. 1493–1495, May 2015, doi: 10.1093/BIOINFORMATICS/BTU813.
[181] M. P. Balogh, "Debating resolution and mass accuracy," *LC-GC North America*, vol. 22, no. 2, pp. 118–126, Feb. 2004.
[182] T. Kind and O. Fiehn, "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry," *BMC Bioinformatics*, vol. 8, 2007, doi: 10.1186/1471-2105-8-105.
[183] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121–134, Oct. 2016, doi: 10.1016/J.PATCOG.2016.03.028.
[184] K. L. Li, H. K. Huang, S. F. Tian, and W. Xu, "Improving one-class SVM for anomaly detection," *International Conference on Machine Learning and Cybernetics*, vol. 5, pp. 3077–3081, 2003, doi: 10.1109/ICMLC.2003.1260106.
[185] R. Busch *et al.*, "Measurement of protein turnover rates by heavy water labeling of nonessential amino acids," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1760, no. 5, pp. 730–744, May 2006, doi: 10.1016/J.BBAGEN.2005.12.023.
[186] X. Shi, B. Xi, P. Jasbi, C. Turner, Y. Jin, and H. Gu, "Comprehensive Isotopic Targeted Mass Spectrometry: Reliable Metabolic Flux Analysis with Broad Coverage," *Analytical Chemistry*, vol. 92, no. 17, pp. 11728–11738, Sep. 2020, doi:

10.1021/ACS.ANALCHEM.0C01767/ASSET/IMAGES/LARGE/AC0C01767_0006.JPEG.
[187] K. Dührkop, F. Hufsky, and S. Böcker, “Molecular Formula Identification Using Isotope Pattern Analysis and Calculation of Fragmentation Trees,” *Mass Spectrom (Tokyo)*, vol. 3, no. Spec Iss 2, p. S0037, 2014, doi: 10.5702/massspectrometry.S0037.
[188] P. Bonini, T. Kind, H. Tsugawa, D. K. Barupal, and O. Fiehn, “Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics,” *Analytical Chemistry*, vol. 92, no. 11, pp. 7515–7522, Jun. 2020, doi: 10.1021/ACS.ANALCHEM.9B05765/ASSET/IMAGES/LARGE/AC9B05765_0002.JPEG.
[189] S. H. Giese, L. R. Sinn, F. Wegner, and J. Rappsilber, “Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry,” *Nat Commun*, vol. 12, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41467-021-23441-0.
[190] E. S. Fedorova, D. D. Matyushin, I. V. Plyushchenko, A. N. Stavrianidi, and A. K. Buryak, “Deep learning for retention time prediction in reversed-phase liquid chromatography,” *J Chromatogr A*, vol. 1664, p. 462792, Feb. 2022, doi: 10.1016/j.chroma.2021.462792.
[191] X. Domingo-Almenara *et al.*, “The METLIN small molecule dataset for machine learning-based retention time prediction,” *Nat Commun*, vol. 10, no. 1, Art. no. 1, Dec. 2019, doi: 10.1038/s41467-019-13680-7.
[192] R. J. Arnold, N. Jayasankar, D. Aggarwal, H. Tang, and P. Radivojac, “A machine learning approach to predicting peptide fragmentation spectra,” in *Biocomputing 2006*, WORLD SCIENTIFIC, 2005, pp. 219–230. doi: 10.1142/9789812701626_0021.
[193] Y. Li, M. Kuhn, A.-C. Gavin, and P. Bork, “Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features,” *Bioinformatics*, vol. 36, no. 4, pp. 1213–1218, Feb. 2020, doi: 10.1093/bioinformatics/btz736.
[194] C. Zhou, L. D. Bowler, and J. Feng, “A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data,” *BMC Bioinformatics*, vol. 9, no. 1, p. 325, Jul. 2008, doi: 10.1186/1471-2105-9-325.
[195] T. U. H. Baumeister *et al.*, “DeltaMS: a tool to track isotopologues in GC- and LC-MS data,” *Metabolomics*, vol. 14, no. 4, p. 41, Feb. 2018, doi: 10.1007/s11306-018-1336-x.
[196] S. Böcker, M. C. Letzel, Z. Lipták, and A. Pervukhin, “Decomposing Metabolomic Isotope Patterns,” in *Algorithms in Bioinformatics*, P. Bücher and B. M. E. Moret, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 12–23. doi: 10.1007/11851561_2.
[197] M. L. Steinhauser *et al.*, “Multi-isotope imaging mass spectrometry quantifies stem cell division and metabolism,” *Nature*, vol. 481, no. 7382, Art. no. 7382, Jan. 2012, doi: 10.1038/nature10734.
[198] F. Weber, T. Zaliznyak, V. P. Edgcomb, and G. T. Taylor, “Using Stable Isotope Probing and Raman Microspectroscopy To Measure Growth Rates of Heterotrophic Bacteria,” *Applied and Environmental Microbiology*, vol. 87, no. 22, pp. e01460-21, Oct. 2021, doi: 10.1128/AEM.01460-21.
[199] P. Navarro *et al.*, “General Statistical Framework for Quantitative Proteomics by Stable Isotope Labeling,” *J. Proteome Res.*, vol. 13, no. 3, pp. 1234–1247, Mar. 2014, doi: 10.1021/pr4006958.

[200] M. Taubert *et al.*, “Time resolved protein-based stable isotope probing (Protein-SIP) analysis allows quantification of induced proteins in substrate shift experiments,” *PROTEOMICS*, vol. 11, no. 11, pp. 2265–2274, 2011, doi: 10.1002/pmic.201000788.
[201] R. Heise *et al.*, “Flux profiling of photosynthetic carbon metabolism in intact plants,” *Nat Protoc*, vol. 9, no. 8, Art. no. 8, Aug. 2014, doi: 10.1038/nprot.2014.115.
[202] J. M. Pratt *et al.*, “Dynamics of Protein Turnover, a Missing Dimension in Proteomics *,” *Molecular & Cellular Proteomics*, vol. 1, no. 8, pp. 579–591, Aug. 2002, doi: 10.1074/mcp.M200046-MCP200.
[203] M. Alevra, S. Mandad, T. Ischebeck, H. Urlaub, S. O. Rizzoli, and E. F. Fornasiero, “A mass spectrometry workflow for measuring protein turnover rates in vivo,” *Nat Protoc*, vol. 14, no. 12, Art. no. 12, Dec. 2019, doi: 10.1038/s41596-019-0222-y.
[204] J. D. Gomez *et al.*, “Program for Integration and Rapid Analysis of Mass Isotopomer Distributions (PIRAMID),” *Bioinformatics*, vol. 39, no. 11, p. btad661, Nov. 2023, doi: 10.1093/bioinformatics/btad661.
[205] X. Han, A. Aslanian, and J. R. Yates, “Mass spectrometry for proteomics,” *Current Opinion in Chemical Biology*, vol. 12, no. 5, pp. 483–490, 2008, doi: 10.1016/j.cbpa.2008.07.024.
[206] W. D. Lehmann, “A timeline of stable isotopes and mass spectrometry in the life sciences,” *Mass Spectrometry Reviews*, vol. 36, no. 1, pp. 58–85, 2017, doi: 10.1002/mas.21497.
[207] S. Cappadona, P. R. Baker, P. R. Cutillas, A. J. R. Heck, and B. Van Breukelen, “Current challenges in software solutions for mass spectrometry-based quantitative proteomics,” *Amino Acids*, vol. 43, no. 3, pp. 1087–1108, 2012, doi: 10.1007/s00726-012-1289-8.
[208] S. Agrawal *et al.*, “El-MAVEN: A Fast, Robust, and User-Friendly Mass Spectrometry Data Processing Engine for Metabolomics,” *Methods in molecular biology (Clifton, N.J.)*, vol. 1978, pp. 301–321, 2019, doi: 10.1007/978-1-4939-9236-2_19.
[209] C. H. Poskar, J. Huege, C. Krach, M. Franke, Y. Shachar-Hill, and B. H. Junker, “IMS2Flux - a high-throughput processing tool for stable isotope labeled mass spectrometric data used for metabolic flux analysis,” *BMC Bioinformatics*, vol. 13, no. 1, 2012, doi: 10.1186/1471-2105-13-295.
[210] H. Ji, Z. Zhang, and H. Lu, “TarMet: a reactive GUI tool for efficient and confident quantification of MS based targeted metabolic and stable isotope tracer analysis,” *Metabolomics*, vol. 14, no. 5, 2018, doi: 10.1007/s11306-018-1363-7.
[211] M. J. Dagley and M. J. McConville, “DExSI: A new tool for the rapid quantitation of 13 C-labelled metabolites detected by GC-MS,” *Bioinformatics*, vol. 34, no. 11, pp. 1957–1958, 2018, doi: 10.1093/bioinformatics/bty025.
[212] J. Wills, J. Edwards-Hicks, and A. J. Finch, “AssayR: A Simple Mass Spectrometry Software Tool for Targeted Metabolic and Stable Isotope Tracer Analyses,” *Analytical Chemistry*, vol. 89, no. 18, pp. 9616–9619, 2017, doi: 10.1021/acs.analchem.7b02401.
[213] R. A. Scheltema, A. Jankevics, R. C. Jansen, M. A. Swertz, and R. Breitling, “PeakML/mzMatch: A file format, Java library, R library, and tool-chain for mass spectrometry data analysis,” *Analytical Chemistry*, vol. 83, no. 7, pp. 2786–2793, 2011, doi: 10.1021/ac2000994.
[214] S. Krishnan *et al.*, “Instrument and process independent binning and baseline correction methods for liquid chromatography–high resolution-mass spectrometry deconvolution,” *Analytica Chimica Acta*, vol. 740, pp. 12–19, Aug. 2012, doi: 10.1016/j.aca.2012.06.014.

[215] X. Ning, I. W. Selesnick, and L. Duval, “Chromatogram baseline estimation and denoising using sparsity (BEADS),” *Chemometrics and Intelligent Laboratory Systems*, vol. 139, pp. 156–167, Dec. 2014, doi: 10.1016/j.chemolab.2014.09.014.
[216] W. -N P. Lee, L. O. Byerley, E. A. Bergner, and J. Edmond, “Mass isotopomer analysis: Theoretical and practical considerations,” *Biological Mass Spectrometry*, vol. 20, no. 8, pp. 451–458, Aug. 1991, doi: 10.1002/BMS.1200200804.
[217] R. G. Sadygov, “Poisson Model To Generate Isotope Distribution for Biomolecules,” *Journal of proteome research*, vol. 17, no. 1, p. 751, Jan. 2018, doi: 10.1021/ACS.JPROTEOME.7B00807.
[218] K. J. R. Rosman and P. D. P. Taylor, “Isotopic compositions of the elements 1997 (Technical Report),” *Pure and Applied Chemistry*, vol. 70, no. 1, pp. 217–235, Jan. 1998, doi: 10.1351/pac199870010217.
[219] C. D. McGillem and G. R. Cooper, *Continuous and Discrete Signal and System Analysis*. in Oxford series in electrical and computer engineering. Oxford University Press, 1991. [Online]. Available: https://books.google.com/books?id=6BVGAQAAIAAJ
[220] P. Heinrich *et al.*, “Correcting for natural isotope abundance and tracer impurity in MS-, MS/MS- and high-resolution-multiple-tracer-data from stable isotope labeling experiments with IsoCorrectoR,” *Scientific Reports*, vol. 8, no. 1, p. 17910, Dec. 2018, doi: 10.1038/S41598-018-36293-4.
[221] M. J. Noonan, H. V. Tinnesand, and C. D. Buesching, “Normalizing Gas-Chromatography-Mass Spectrometry Data: Method Choice can Alter Biological Inference,” *Bioessays*, vol. 40, no. 6, p. e1700210, Jun. 2018, doi: 10.1002/bies.201700210.
[222] P. Filzmoser and B. Walczak, “What can go wrong at the data normalization step for identification of biomarkers?,” *Journal of Chromatography A*, vol. 1362, pp. 194–205, Oct. 2014, doi: 10.1016/j.chroma.2014.08.050.
[223] J. Chen *et al.*, “Influences of Normalization Method on Biomarker Discovery in Gas Chromatography–Mass Spectrometry-Based Untargeted Metabolomics: What Should Be Considered?,” *Anal. Chem.*, vol. 89, no. 10, pp. 5342–5348, May 2017, doi: 10.1021/acs.analchem.6b05152.
[224] P. Kiefer, U. Schmitt, and J. A. Vorholt, “eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows,” *Bioinformatics*, vol. 29, no. 7, pp. 963–964, Apr. 2013, doi: 10.1093/BIOINFORMATICS/BTT080.
[225] R. Spicer, R. M. Salek, P. Moreno, D. Cañueto, and C. Steinbeck, “Navigating freely-available software tools for metabolomics analysis,” *Metabolomics*, vol. 13, no. 9, p. 106, 2017, doi: 10.1007/s11306-017-1242-7.
[226] N. G. Mahieu, J. L. Genenbacher, and G. J. Patti, “A Roadmap for the XCMS Family of Software Solutions in Metabolomics,” *Curr Opin Chem Biol*, vol. 30, pp. 87–93, Feb. 2016, doi: 10.1016/j.cbpa.2015.11.009.
[227] J. Zhou and Y. Yin, “Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry,” *Analyst*, vol. 141, no. 23, pp. 6362–6373, 2016, doi: 10.1039/C6AN01753C.
[228] W. M. B. Edmands, D. K. Barupal, and A. Scalbert, “MetMSLine: an automated and fully integrated pipeline for rapid processing of high-resolution LC–MS metabolomic datasets,” *Bioinformatics*, vol. 31, no. 5, pp. 788–790, Mar. 2015, doi: 10.1093/bioinformatics/btu705.

[229] S. Savary, L. Willocquet, S. J. Pethybridge, P. Esker, N. McRoberts, and A. Nelson, “The global burden of pathogens and pests on major food crops,” *Nature Ecology & Evolution 2019 3:3*, vol. 3, no. 3, pp. 430–439, Feb. 2019, doi: 10.1038/s41559-018-0793-y.
[230] G. L. Hartman, E. D. West, and T. K. Herman, “Crops that feed the World 2. Soybean-worldwide production, use, and constraints caused by pathogens and pests,” *Food Security*, vol. 3, no. 1, pp. 5–17, Mar. 2011, doi: 10.1007/S12571-010-0108-X/FIGURES/10.
[231] E. R. Cober, S. R. Cianzio, V. R. Pantalone, and I. Rajcan, “Soybean,” *Oil Crops*, pp. 57–90, 2009, doi: 10.1007/978-0-387-77594-4_3.
[232] L. Li *et al.*, “A systems biology approach toward understanding seed composition in soybean,” *BMC Genomics*, vol. 16, no. 3, pp. 1–18, Jan. 2015, doi: 10.1186/1471-2164-16-S3-S9/FIGURES/7.
[233] R. J. Weselake *et al.*, “Increasing the flow of carbon into seed oil,” *Biotechnology Advances*, vol. 27, no. 6, pp. 866–878, Nov. 2009, doi: 10.1016/J.BIOTECHADV.2009.07.001.
[234] T. E. Clemente and E. B. Cahoon, “Soybean Oil: Genetic Approaches for Modification of Functionality and Total Content,” *Plant Physiology*, vol. 151, no. 3, pp. 1030–1040, Nov. 2009, doi: 10.1104/PP.109.146282.
[235] V. Kumar *et al.*, “Omics advances and integrative approaches for the simultaneous improvement of seed oil and protein content in soybean (Glycine max L.),” *https://doi.org/10.1080/07352689.2021.1954778*, vol. 40, no. 5, pp. 398–421, 2021, doi: 10.1080/07352689.2021.1954778.
[236] H. Zhang *et al.*, “Selection of GmSWEET39 for oil and protein improvement in soybean,” *PLOS Genetics*, vol. 16, no. 11, p. e1009114, Nov. 2020, doi: 10.1371/JOURNAL.PGEN.1009114.
[237] Y. Assefa *et al.*, “Spatial Characterization of Soybean Yield and Quality (Amino Acids, Oil, and Protein) for United States,” *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–11, Oct. 2018, doi: 10.1038/s41598-018-32895-0.
[238] S. Lee *et al.*, “Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV,” *Theoretical and Applied Genetics*, vol. 132, no. 6, pp. 1639–1659, Jun. 2019, doi: 10.1007/S00122-019-03304-5/FIGURES/7.
[239] J. R. Wilcox and R. M. Shibles, “Interrelationships among Seed Quality Attributes in Soybean,” *Crop Science*, vol. 41, no. 1, pp. 11–14, Jan. 2001, doi: 10.2135/CROPSCI2001.41111X.
[240] D. K. Allen, P. D. Bates, and H. Tjellström, “Tracking the metabolic pulse of plant lipid production with isotopic labeling and flux analyses: Past, present and future,” *Progress in Lipid Research*, vol. 58, pp. 97–120, Apr. 2015, doi: 10.1016/J.PLIPRES.2015.02.002.
[241] P. D. Bates and J. Browse, “The significance of different diacylgycerol synthesis pathways on plant oil composition and bioengineering,” *Frontiers in Plant Science*, vol. 3, no. JUL, p. 29493, Jul. 2012, doi: 10.3389/FPLS.2012.00147/BIBTEX.
[242] C. Hernández-Sebastià, F. Marsolais, C. Saravitz, D. Israel, R. E. Dewey, and S. C. Huber, “Free amino acid profiles suggest a possible role for asparagine in the control of storage-product accumulation in developing seeds of low- and high-protein soybean lines,” *Journal of Experimental Botany*, vol. 56, no. 417, pp. 1951–1963, Jul. 2005, doi: 10.1093/JXB/ERI191.

[243] T. Maekawa, M. Maekawa-Yoshikawa, N. Takeda, H. Imaizumi-Anraku, Y. Murooka, and M. Hayashi, “Gibberellin controls the nodulation signaling pathway in Lotus japonicus,” *Plant Journal*, vol. 58, no. 2, pp. 183–194, Apr. 2009, doi: 10.1111/J.1365-313X.2008.03774.X.
[244] A. Pipolo, … T. S.-A. of A., and undefined 2004, “Protein and oil concentration of soybean seed cultured in vitro using nutrient solutions of differing glutamine concentration,” *Annals of Applied Biology*, vol. 144, no. 2, pp. 223–227, Apr. 2004, doi: 10.1111/j.1744-7348.2004.tb00337.x.
[245] D. K. Allen and J. D. Young, “Carbon and Nitrogen Provisions Alter the Metabolic Flux in Developing Soybean Embryos,” *Plant Physiology*, vol. 161, no. 3, pp. 1458–1475, Feb. 2013, doi: 10.1104/PP.112.203299.
[246] Q. Truong, K. Koch, J. M. Yoon, J. D. Everard, and J. V. Shanks, “Influence of carbon to nitrogen ratios on soybean somatic embryo (cv. Jack) growth and composition,” *Journal of Experimental Botany*, vol. 64, no. 10, pp. 2985–2995, Jul. 2013, doi: 10.1093/JXB/ERT138.
[247] M. Dieuaide-Noubhani, A. P. Alonso, D. Rolin, W. Eisenreich, and P. Raymond, “Metabolic flux analysis: recent advances in carbon metabolism in plants.,” *EXS*, vol. 97, pp. 213–243, 2007, doi: 10.1007/978-3-7643-7439-6_10/COVER.
[248] R. G. Ratcliffe and Y. Shachar-Hill, “Measuring multiple fluxes through plant metabolic networks,” *The Plant Journal*, vol. 45, no. 4, pp. 490–511, Feb. 2006, doi: 10.1111/J.1365-313X.2005.02649.X.
[249] J. Schwender, J. B. Ohlrogge, and Y. Shachar-Hill, “A Flux Model of Glycolysis and the Oxidative Pentosephosphate Pathway in Developing Brassica napus Embryos,” *Journal of Biological Chemistry*, vol. 278, no. 32, pp. 29442–29453, Aug. 2003, doi: 10.1074/JBC.M303432200.
[250] J. Schwender, Y. Shachar-Hill, and J. B. Ohlrogge, “Mitochondrial Metabolism in Developing Embryos of Brassica napus,” *Journal of Biological Chemistry*, vol. 281, no. 45, pp. 34040–34047, Nov. 2006, doi: 10.1074/JBC.M606266200.
[251] A. U. Igamberdiev and L. A. Kleczkowski, “The Glycerate and Phosphorylated Pathways of Serine Synthesis in Plants: The Branches of Plant Glycolysis Linking Carbon and Nitrogen Metabolism,” *Frontiers in Plant Science*, vol. 9, 2018, Accessed: Nov. 09, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2018.00318
[252] A. P. Alonso, F. D. Goffman, J. B. Ohlrogge, and Y. Shachar-Hill, “Carbon conversion efficiency and central metabolic fluxes in developing sunflower (Helianthus annuus L.) embryos,” *The Plant Journal*, vol. 52, no. 2, pp. 296–308, Oct. 2007, doi: 10.1111/J.1365-313X.2007.03235.X.
[253] G. Sriram *et al.*, “Quantification of Compartmented Metabolic Fluxes in Developing Soybean Embryos by Employing Biosynthetically Directed Fractional 13C Labeling, Two-Dimensional [13C, 1H] Nuclear Magnetic Resonance, and Comprehensive Isotopomer Balancing,” *Plant Physiology*, vol. 136, no. 2, pp. 3043–3057, Oct. 2004, doi: 10.1104/PP.104.050625.
[254] D. K. Allen, J. B. Ohlrogge, and Y. Shachar-Hill, “The role of light in soybean seed filling metabolism,” *The Plant Journal*, vol. 58, no. 2, pp. 220–234, Apr. 2009, doi: 10.1111/J.1365-313X.2008.03771.X.
[255] S. Kambhampati *et al.*, “On the Inverse Correlation of Protein and Oil: Examining the Effects of Altered Central Carbon Metabolism on Seed Composition Using Soybean Fast

Neutron Mutants," *Metabolites 2020, Vol. 10, Page 18*, vol. 10, no. 1, p. 18, Dec. 2019, doi: 10.3390/METABO10010018.
[256] S. A. Morley *et al.*, "Expression of malic enzyme reveals subcellular carbon partitioning for storage reserve production in soybeans," *New Phytologist*, Mar. 2023, doi: 10.1111/NPH.18835.
[257] R. Yokoyama, M. V. V. de Oliveira, B. Kleven, and H. A. Maeda, "The entry reaction of the plant shikimate pathway is subjected to highly complex metabolite-mediated regulation," *The Plant Cell*, vol. 33, no. 3, p. 671, Mar. 2021, doi: 10.1093/PLCELL/KOAA042.
[258] A. Parthasarathy, P. J. Cross, R. C. J. Dobson, L. E. Adams, M. A. Savka, and A. O. Hudson, "A Three-Ring circus: Metabolism of the three proteogenic aromatic amino acids and their role in the health of plants and animals," *Frontiers in Molecular Biosciences*, vol. 5, no. APR, p. 342220, Apr. 2018, doi: 10.3389/FMOLB.2018.00029/BIBTEX.
[259] T. Wiggins, S. Kumar, S. R. Markar, S. Antonowicz, and G. B. Hanna, "Tyrosine, phenylalanine, and tryptophan in gastroesophageal malignancy: A systematic review," *Cancer Epidemiology Biomarkers and Prevention*, vol. 24, no. 1, pp. 32–38, Jan. 2015, doi: 10.1158/1055-9965.EPI-14-0980/67949/AM/TYROSINE-PHENYLALANINE-AND-TRYPTOPHAN-IN-GASTRO.
[260] R. Jeannotte, "METABOLIC PATHWAYS | Nitrogen Metabolism," *Encyclopedia of Food Microbiology: Second Edition*, pp. 544–560, Jan. 2014, doi: 10.1016/B978-0-12-384730-0.00199-3.
[261] M. G. Dawood and M. S. Sadak, "Physiological response of canola plants (Brassica napus L.) to tryptophan or benzyladenine," *Lucrari stiintifice*, vol. 50, no. 9, pp. 198–207, 2007.
[262] Y. Zhao, "Auxin biosynthesis: a simple two-step pathway converts tryptophan to indole-3-acetic acid in plants," *Mol Plant*, vol. 5, no. 2, pp. 334–338, Mar. 2012, doi: 10.1093/mp/ssr104.
[263] G.-H. Dao, G.-X. Wu, X.-X. Wang, L.-L. Zhuang, T.-Y. Zhang, and H.-Y. Hu, "Enhanced growth and fatty acid accumulation of microalgae Scenedesmus sp. LX1 by two types of auxin," *Bioresource Technology*, vol. 247, pp. 561–567, Jan. 2018, doi: 10.1016/j.biortech.2017.09.079.
[264] D. Pashang, W. Weisany, and F. G.-K. Ghajar, "Changes in the Fatty Acid and Morphophysiological Traits of Safflower (Carthamus tinctorius L.) Cultivars as Response to Auxin Under Water-Deficit Stress," *J Soil Sci Plant Nutr*, vol. 21, no. 3, pp. 2164–2177, Sep. 2021, doi: 10.1007/s42729-021-00512-1.
[265] W. Liu, D. F. Hildebrand, and G. B. Collins, "Auxin-regulated changes of fatty acid content and composition in soybean zygotic embryo cotyledons," *Plant Science*, vol. 106, no. 1, pp. 31–42, Mar. 1995, doi: 10.1016/0168-9452(95)04067-5.
[266] R. P. Walker, Z. H. Chen, and F. Famiani, "Gluconeogenesis in Plants: A Key Interface between Organic Acid/Amino Acid/Lipid and Sugar Metabolism," *Molecules*, vol. 26, no. 17, Sep. 2021, doi: 10.3390/MOLECULES26175129.
[267] Y. Xiong, Q.-Y. Lei, S. Zhao, and K.-L. Guan, "Regulation of Glycolysis and Gluconeogenesis by Acetylation of PKM and PEPCK," *Cold Spring Harb Symp Quant Biol*, vol. 76, pp. 285–289, 2011, doi: 10.1101/sqb.2011.76.010942.
[268] P. J. Eastmond *et al.*, "Arabidopsis uses two gluconeogenic gateways for organic acids to fuel seedling establishment," *Nature Communications*, vol. 6, Apr. 2015, doi: 10.1038/NCOMMS7659.

[269] E. L. Rylott, A. D. Gilday, and I. A. Graham, “The Gluconeogenic Enzyme Phosphoenolpyruvate Carboxykinase in Arabidopsis Is Essential for Seedling Establishment,” *Plant Physiology*, vol. 131, no. 4, pp. 1834–1842, Apr. 2003, doi: 10.1104/PP.102.019174.
[270] A. D’Aniello, G. Fisher, N. Migliaccio, G. Cammisa, E. D’Aniello, and P. Spinelli, “Amino acids and transaminases activity in ventricular CSF and in brain of normal and Alzheimer patients,” *Neuroscience Letters*, vol. 388, no. 1, pp. 49–53, Nov. 2005, doi: 10.1016/J.NEULET.2005.06.030.
[271] S. L. Zhou, R. E. Gordon, M. Bradbury, D. Stump, C. L. Kiang, and P. D. Berk, “Ethanol up-regulates fatty acid uptake and plasma membrane expression and export of mitochondrial aspartate aminotransferase in HepG2 cells,” *Hepatology*, vol. 27, no. 4, pp. 1064–1074, Apr. 1998, doi: 10.1002/HEP.510270423.
[272] S. E. Wilkie, J. M. Roper, A. G. Smith, and M. J. Warren, “Isolation, characterisation and expression of a cDNA clone encoding plastid aspartate aminotransferase from Arabidopsis thaliana,” *Plant Molecular Biology*, vol. 27, no. 6, pp. 1227–1233, Mar. 1995, doi: 10.1007/BF00020897/METRICS.
[273] J. H. Bryce and D. A. Day, “Tricarboxylic Acid Cycle Activity in Mitochondria from Soybean Nodules and Cotyledons,” *Journal of Experimental Botany*, vol. 41, no. 8, pp. 961–967, Aug. 1990, doi: 10.1093/jxb/41.8.961.
[274] J. B. Ohlrogge, D. N. Kuhn, and P. K. Stumpf, “Subcellular localization of acyl carrier protein in leaf protoplasts of Spinacia oleracea.,” *Proceedings of the National Academy of Sciences*, vol. 76, no. 3, pp. 1194–1198, Mar. 1979, doi: 10.1073/pnas.76.3.1194.
[275] S. Rawsthorne, “Carbon flux and fatty acid synthesis in plants,” *Progress in Lipid Research*, vol. 41, no. 2, pp. 182–196, Mar. 2002, doi: 10.1016/S0163-7827(01)00023-6.
[276] T. B. Moreira *et al.*, “A Genome-Scale Metabolic Model of Soybean (Glycine max) Highlights Metabolic Fluxes in Seedlings1[OPEN],” *Plant Physiol*, vol. 180, no. 4, pp. 1912–1929, Aug. 2019, doi: 10.1104/pp.19.00122.
[277] J. E. Lunn, “Compartmentation in plant metabolism,” *Journal of Experimental Botany*, vol. 58, no. 1, pp. 35–47, Jan. 2007, doi: 10.1093/jxb/erl134.
[278] Y. S. Han, R. Van Der Heijden, and R. Verpoorte, “Biosynthesis of anthraquinones in cell cultures of the Rubiaceae,” *Plant Cell, Tissue and Organ Culture*, vol. 67, no. 3, pp. 201–220, 2001, doi: 10.1023/A:1012758922713/METRICS.
[279] X. Sun, G. Han, Z. Meng, L. Lin, and N. Sui, “Roles of malic enzymes in plant development and stress responses,” *Plant Signaling & Behavior*, vol. 14, no. 10, Oct. 2019, doi: 10.1080/15592324.2019.1644596.
[280] G. Hao *et al.*, “Role of Malic Enzyme during Fatty Acid Synthesis in the Oleaginous Fungus Mortierella alpina,” *Appl Environ Microbiol*, vol. 80, no. 9, pp. 2672–2678, May 2014, doi: 10.1128/AEM.00140-14.
[281] B.-H. Zhu, R.-H. Zhang, N.-N. Lv, G.-P. Yang, Y.-S. Wang, and K.-H. Pan, “The Role of Malic Enzyme on Promoting Total Lipid and Fatty Acid Production in Phaeodactylum tricornutum,” *Frontiers in Plant Science*, vol. 9, 2018, Accessed: Nov. 13, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2018.00826
[282] T. McClure, J. C. Cocuron, V. Osmark, L. K. McHale, and A. P. Alonso, “Impact of Environment on the Biomass Composition of Soybean (Glycine max) seeds,” *Journal of Agricultural and Food Chemistry*, vol. 65, no. 32, pp. 6753–6761, Aug. 2017, doi: 10.1021/ACS.JAFC.7B01457/ASSET/IMAGES/LARGE/JF-2017-01457N_0004.JPEG.

[283] J. L. Rotundo, L. Borrás, J. De Bruin, and P. Pedersen, “Physiological strategies for seed number determination in soybean: Biomass accumulation, partitioning and seed set efficiency,” *Field Crops Research*, vol. 135, pp. 58–66, Aug. 2012, doi: 10.1016/J.FCR.2012.06.012.
[284] H. Esmaeili, A. Karami, and F. Maggi, “Essential oil composition, total phenolic and flavonoids contents, and antioxidant activity of Oliveria decumbens Vent. (Apiaceae) at different phenological stages,” *Journal of Cleaner Production*, vol. 198, pp. 91–95, Oct. 2018, doi: 10.1016/J.JCLEPRO.2018.07.029.
[285] S. Alinian, J. Razmjoo, and H. Zeinali, “Flavonoids, anthocynins, phenolics and essential oil produced in cumin (Cuminum cyminum L.) accessions under different irrigation regimes,” *Industrial Crops and Products*, vol. 81, pp. 49–55, Mar. 2016, doi: 10.1016/J.INDCROP.2015.11.040.
[286] M. Ghorbanpour, “Major essential oil constituents, total phenolics and flavonoids content and antioxidant activity of Salvia officinalis plant in response to nano-titanium dioxide,” *Indian Journal of Plant Physiology*, vol. 20, no. 3, pp. 249–256, Sep. 2015, doi: 10.1007/S40502-015-0170-7/TABLES/4.
[287] C. Bouchard, “Genetics of Obesity: What We Have Learned Over Decades of Research,” *Obesity*, vol. 29, no. 5, pp. 802–820, May 2021, doi: 10.1002/OBY.23116.
[288] A. Khan, N. Choudhury, S. Uddin, L. Hossain, and L. A. Baur, “Longitudinal trends in global obesity research and collaboration: a review using bibliometric metadata,” *Obesity Reviews*, vol. 17, no. 4, pp. 377–385, Apr. 2016, doi: 10.1111/OBR.12372.
[289] J. Bentham *et al.*, “Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults,” *The Lancet*, vol. 390, no. 10113, pp. 2627–2642, Dec. 2017, doi: 10.1016/S0140-6736(17)32129-3.
[290] M. Di Cesare *et al.*, “Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants,” *Lancet (London, England)*, vol. 387, no. 10026, pp. 1377–1396, Apr. 2016, doi: 10.1016/S0140-6736(16)30054-X.
[291] I. Bhattacharya, C. Ghayor, A. Pérez Dominguez, and F. E. Weber, “From Influenza Virus to Novel Corona Virus (SARS-CoV-2)–The Contribution of Obesity,” *Frontiers in Endocrinology*, vol. 11, Oct. 2020, doi: 10.3389/fendo.2020.556962.
[292] A. Must, J. Spadano, E. Coakley, A. Field, G. Colditz, and W. Dietz, “The Disease Burden Associated With Overweight and Obesity,” *JAMA*, vol. 282, no. 16, p. 1523, Oct. 1999, doi: 10.1001/jama.282.16.1523.
[293] C. M. Petrilli *et al.*, “Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study,” *BMJ*, p. m1966, May 2020, doi: 10.1136/bmj.m1966.
[294] X. Zhao *et al.*, “Obesity Increases the Severity and Mortality of Influenza and COVID-19: A Systematic Review and Meta-Analysis,” *Frontiers in Endocrinology*, vol. 11, Dec. 2020, doi: 10.3389/fendo.2020.595109.
[295] G. Kastenmüller, J. Raffler, C. Gieger, and K. Suhre, “Genetics of human metabolism: an update,” *Human Molecular Genetics*, vol. 24, no. R1, pp. R93–R101, Oct. 2015, doi: 10.1093/hmg/ddv263.
[296] A. Keys, H. L. Taylor, and F. Grande, “Basal metabolism and age of adult man,” *Metabolism*, vol. 22, no. 4, pp. 579–587, Apr. 1973, doi: 10.1016/0026-0495(73)90071-1.

[297] T. L. Wall, “Genetic associations of alcohol and aldehyde dehydrogenase with alcohol dependence and their mechanisms of action,” *Therapeutic Drug Monitoring*, vol. 27, no. 6, pp. 700–703, Dec. 2005, doi: 10.1097/01.ftd.0000179840.78762.33.
[298] A. V. Nedeltcheva and F. A. J. L. Scheer, “Metabolic effects of sleep disruption, links to obesity and diabetes,” *Current Opinion in Endocrinology, Diabetes & Obesity*, vol. 21, no. 4, pp. 293–298, Aug. 2014, doi: 10.1097/MED.0000000000000082.
[299] C. Rabasa and S. L. Dickson, “Impact of stress on metabolism and energy balance,” *Current Opinion in Behavioral Sciences*, vol. 9, pp. 71–77, Jun. 2016, doi: 10.1016/j.cobeha.2016.01.011.
[300] S. Sharma and M. Kavuru, “Sleep and Metabolism: An Overview,” *International Journal of Endocrinology*, vol. 2010, pp. 1–12, 2010, doi: 10.1155/2010/270832.
[301] E. C. Lien and M. G. Vander Heiden, “A framework for examining how diet impacts tumour metabolism,” *Nature Reviews Cancer*, vol. 19, no. 11, pp. 651–661, Nov. 2019, doi: 10.1038/s41568-019-0198-5.
[302] P. Louis, K. P. Scott, S. H. Duncan, and H. J. Flint, “Understanding the effects of diet on bacterial metabolism in the large intestine,” *Journal of Applied Microbiology*, vol. 102, no. 5, pp. 1197–1208, May 2007, doi: 10.1111/j.1365-2672.2007.03322.x.
[303] H. N. MUNRO, “General Aspects of the Regulation of Protein Metabolism by Diet and by Hormones,” in *Mammalian Protein Metabolism*, Elsevier, 1964, pp. 381–481. doi: 10.1016/B978-1-4832-3209-6.50017-7.
[304] W. D. van Marken Lichtenbelt, R. P. Mensink, and K. R. Westerterp, “The effect of fat composition of the diet on energy metabolism,” *Zeitschrift für Ernährungswissenschaft*, vol. 36, no. 4, pp. 303–305, Dec. 1997, doi: 10.1007/BF01617803.
[305] M. Kleinert *et al.*, “Animal models of obesity and diabetes mellitus,” *Nature Reviews Endocrinology 2018 14:3*, vol. 14, no. 3, pp. 140–162, Jan. 2018, doi: 10.1038/nrendo.2017.161.
[306] T. Martins *et al.*, “Murine Models of Obesity,” *Obesities 2022, Vol. 2, Pages 127-147*, vol. 2, no. 2, pp. 127–147, Mar. 2022, doi: 10.3390/OBESITIES2020012.
[307] J. L. Griffin, “Understanding mouse models of disease through metabolomics,” *Current Opinion in Chemical Biology*, vol. 10, no. 4, pp. 309–315, Aug. 2006, doi: 10.1016/j.cbpa.2006.06.027.
[308] N. Vinayavekhin, E. A. Homan, and A. Saghatelian, “Exploring Disease through Metabolomics,” *ACS Chemical Biology*, vol. 5, no. 1, pp. 91–103, Jan. 2010, doi: 10.1021/cb900271r.
[309] W. W. Chen, E. Freinkman, T. Wang, K. Birsoy, and D. M. Sabatini, “Absolute Quantification of Matrix Metabolites Reveals the Dynamics of Mitochondrial Metabolism,” *Cell*, vol. 166, no. 5, pp. 1324-1337.e11, Aug. 2016, doi: 10.1016/j.cell.2016.07.040.
[310] W. I. Sivitz, “Techniques to investigate bioenergetics of mitochondria,” *Neuromethods*, vol. 123, pp. 67–94, 2017, doi: 10.1007/978-1-4939-6890-9_4/FIGURES/8.
[311] A. Srivastava, G. M. Kowalski, D. L. Callahan, P. J. Meikle, and D. J. Creek, “Strategies for extending metabolomics studies with stable isotope labelling and fluxomics,” *Metabolites*, vol. 6, no. 4, 2016, doi: 10.3390/metabo6040032.
[312] H. C. Williams *et al.*, “Oral Gavage Delivery of Stable Isotope Tracer for In Vivo Metabolomics,” *Metabolites*, vol. 10, no. 12, pp. 1–18, Dec. 2020, doi: 10.3390/METABO10120501.

[313] R. C. Sun *et al.*, "Noninvasive liquid diet delivery of stable isotopes into mouse models for deep metabolic network tracing," *Nature Communications 2017 8:1*, vol. 8, no. 1, pp. 1–10, Nov. 2017, doi: 10.1038/s41467-017-01518-z.
[314] D. Broekaert and S. M. Fendt, "Measuring in vivo tissue metabolism using 13C glucose infusions in mice," *Methods in Molecular Biology*, vol. 1862, pp. 67–82, 2019, doi: 10.1007/978-1-4939-8769-6_5/FIGURES/6.
[315] C. T. Hensley *et al.*, "Metabolic Heterogeneity in Human Lung Tumors," *Cell*, vol. 164, no. 4, pp. 681–694, Feb. 2016, doi: 10.1016/J.CELL.2015.12.034.
[316] M. R. Antoniewicz, "Methods and advances in metabolic flux analysis: a mini-review," *Journal of Industrial Microbiology and Biotechnology*, vol. 42, no. 3, pp. 317–325, Mar. 2015, doi: 10.1007/S10295-015-1585-X.
[317] Z. Dai and J. W. Locasale, "Understanding metabolism with flux analysis: from theory to application," *Metabolic engineering*, vol. 43, no. Pt B, p. 94, Sep. 2017, doi: 10.1016/J.YMBEN.2016.09.005.
[318] E. B. Marliss, T. T. Aoki, R. H. Unger, J. S. Soeldner, and G. F. Cahill, "Glucagon levels and metabolic effects in fasting man," *The Journal of Clinical Investigation*, vol. 49, no. 12, pp. 2256–2270, Dec. 1970, doi: 10.1172/JCI106445.
[319] G. F. Cahill, "Physiology of Insulin In Man: The Banting Memorial Lecture 1971," *Diabetes*, vol. 20, no. 12, pp. 785–799, Dec. 1971, doi: 10.2337/diab.20.12.785.
[320] H. Stingl *et al.*, "Changes in hepatic glycogen cycling during a glucose load in healthy humans," *Diabetologia*, vol. 49, no. 2, pp. 360–368, Feb. 2006, doi: 10.1007/s00125-005-0099-x.
[321] K. S. Ray and P. R. Singhania, "Glycemic and insulinemic responses to carbohydrate rich whole foods," *Journal of Food Science and Technology*, vol. 51, no. 2, p. 347, Feb. 2014, doi: 10.1007/S13197-011-0497-7.
[322] S. T. Chung, S. K. Chacko, A. L. Sunehag, and M. W. Haymond, "Measurements of Gluconeogenesis and Glycogenolysis: A Methodological Review," *Diabetes*, vol. 64, no. 12, pp. 3996–4010, Dec. 2015, doi: 10.2337/db15-0640.
[323] J. E. Ayala, D. P. Bracy, O. P. McGuinness, and D. H. Wasserman, "Considerations in the Design of Hyperinsulinemic-Euglycemic Clamps in the Conscious Mouse," *Diabetes*, vol. 55, no. 2, pp. 390–397, Feb. 2006, doi: 10.2337/DIABETES.55.02.06.DB05-0686.
[324] J. J. Robert, "Methods for the measurement of insulin resistance. Hyperinsulinemic euglycemic clamp.," *Presse Medicale (Paris, France : 1983)*, vol. 24, no. 15, pp. 730–734, Apr. 1995.
[325] Y. Zhang, L. Xu, X. Liu, and Y. Wang, "Evaluation of insulin sensitivity by hyperinsulinemic-euglycemic clamps using stable isotope-labeled glucose," *Cell Discovery 2018 4:1*, vol. 4, no. 1, pp. 1–4, Apr. 2018, doi: 10.1038/s41421-018-0016-3.
[326] R. Zhang, B. Chen, H. Zhang, L. Tu, and T. Luan, "Stable isotope-based metabolic flux analysis: A robust tool for revealing toxicity pathways of emerging contaminants," *TrAC Trends in Analytical Chemistry*, vol. 159, p. 116909, Feb. 2023, doi: 10.1016/J.TRAC.2022.116909.
[327] G. Baldini and K. D. Phelan, "The melanocortin pathway and control of appetite-progress and therapeutic implications," *J Endocrinol*, vol. 241, no. 1, pp. R1–R33, Apr. 2019, doi: 10.1530/JOE-18-0596.

[328] L. M. de Souza Cordeiro *et al.*, “Hypothalamic MC4R regulates glucose homeostasis through adrenaline-mediated control of glucose reabsorption via renal GLUT2 in mice,” *Diabetologia*, vol. 64, no. 1, pp. 181–194, Jan. 2021, doi: 10.1007/s00125-020-05289-z.
[329] M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.
[330] F. C. Baker and D. A. Schooley, “Biosynthesis of 3-hydroxy-3-methylglutaryl-CoA, 3-hydroxy-3-ethylglutaryl-CoA, mevalonate and homomevalonate by insect corpus allatum and mammalian hepatic tissues,” *Biochim Biophys Acta*, vol. 664, no. 2, pp. 356–372, May 1981, doi: 10.1016/0005-2760(81)90058-8.
[331] R. M. Bock and R. A. Alberty, “Studies of the Enzyme Fumarase. I. Kinetics and Equilibrium,” *Journal of the American Chemical Society*, vol. 75, no. 8, pp. 1921–1925, 1953, doi: 10.1021/ja01104a043.
[332] M. J. van der Werf, W. J. van den Tweel, and S. Hartmans, “Purification and Characterization of Maleate Hydratase from Pseudomonas pseudoalcaligenes.,” *Appl Environ Microbiol*, vol. 59, no. 9, pp. 2823–2829, Sep. 1993, doi: 10.1128/aem.59.9.2823-2829.1993.
[333] I. M. Washington and G. Van Hoosier, “Chapter 3 - Clinical Biochemistry and Hematology,” in *The Laboratory Rabbit, Guinea Pig, Hamster, and Other Rodents*, M. A. Suckow, K. A. Stevens, and R. P. Wilson, Eds., in American College of Laboratory Animal Medicine. , Boston: Academic Press, 2012, pp. 57–116. doi: 10.1016/B978-0-12-380920-9.00003-1.
[334] R. R. Wolfe, F. Jahoor, and H. Miyoshi, “Evaluation of the isotopic equilibration between lactate and pyruvate,” *American Journal of Physiology-Endocrinology and Metabolism*, vol. 254, no. 4, pp. E532–E535, Apr. 1988, doi: 10.1152/ajpendo.1988.254.4.E532.
[335] T. J. G. Ettema, H. Ahmed, A. C. M. Geerling, J. van der Oost, and B. Siebers, “The non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) of Sulfolobus solfataricus: a key-enzyme of the semi-phosphorylative branch of the Entner–Doudoroff pathway,” *Extremophiles*, vol. 12, no. 1, pp. 75–88, Jan. 2008, doi: 10.1007/s00792-007-0082-1.
[336] M. Holeček, “Serine Metabolism in Health and Disease and as a Conditionally Essential Amino Acid,” *Nutrients*, vol. 14, no. 9, May 2022, doi: 10.3390/NU14091987.
[337] A. Besrat, C. E. Polan, and L. M. Henderson, “Mammalian Metabolism of Glutaric Acid,” *Journal of Biological Chemistry*, vol. 244, no. 6, pp. 1461–1467, Mar. 1969, doi: 10.1016/S0021-9258(18)91782-5.
[338] E. Minogue *et al.*, “Glutarate regulates T cell metabolism and anti-tumour immunity,” *Nature Metabolism 2023*, pp. 1–18, Aug. 2023, doi: 10.1038/s42255-023-00855-2.
[339] J. Wang, Y. Wu, X. Sun, Q. Yuan, and Y. Yan, “De Novo Biosynthesis of Glutarate via α-Keto Acid Carbon Chain Extension and Decarboxylation Pathway in Escherichia coli,” *ACS Synthetic Biology*, vol. 6, no. 10, pp. 1922–1930, Oct. 2017, doi: 10.1021/ACSSYNBIO.7B00136/ASSET/IMAGES/LARGE/SB-2017-001367_0005.JPEG.
[340] J. L. Yu, X. X. Xia, J. J. Zhong, and Z. G. Qian, “Enhanced production of C5 dicarboxylic acids by aerobic-anaerobic shift in fermentation of engineered Escherichia coli,” *Process Biochemistry*, vol. 62, pp. 53–58, Nov. 2017, doi: 10.1016/J.PROCBIO.2017.09.001.

[341] I. Elia, R. Schmieder, S. Christen, and S. M. Fendt, “Organ-specific cancer metabolism and its potential for therapy,” *Handbook of Experimental Pharmacology*, vol. 233, pp. 321–353, Mar. 2016, doi: 10.1007/164_2015_10/TABLES/1.
[342] A. G. McAtee, L. J. Jazmin, and J. D. Young, “Application of isotope labeling experiments and 13C flux analysis to enable rational pathway engineering,” *Current Opinion in Biotechnology*, vol. 36, pp. 50–56, Dec. 2015, doi: 10.1016/j.copbio.2015.08.004.
[343] S. Kambhampati *et al.*, “SIMPEL: using stable isotopes to elucidate dynamics of context specific metabolism,” *Commun Biol*, vol. 7, no. 1, pp. 1–11, Feb. 2024, doi: 10.1038/s42003-024-05844-z.
[344] Ó. Rolfsson, G. Paglia, M. Magnusdóttir, B. Ø. Palsson, and I. Thiele, “Inferring the metabolism of human orphan metabolites from their metabolic network context affirms human gluconokinase activity,” *Biochemical Journal*, vol. 449, no. 2, pp. 427–435, Dec. 2012, doi: 10.1042/BJ20120980.
[345] G. Solinas, J. Borén, and A. G. Dulloo, “De novo lipogenesis in metabolic homeostasis: More friend than foe?,” *Molecular Metabolism*, vol. 4, no. 5, pp. 367–377, May 2015, doi: 10.1016/j.molmet.2015.03.004.
[346] Z. Song, A. M. Xiaoli, and F. Yang, “Regulation and Metabolic Significance of De Novo Lipogenesis in Adipose Tissues,” *Nutrients*, vol. 10, no. 10, p. 1383, Sep. 2018, doi: 10.3390/nu10101383.
[347] M. C. Moore *et al.*, “Morning Hyperinsulinemia Primes the Liver for Glucose Uptake and Glycogen Storage Later in the Day,” *Diabetes*, vol. 67, no. 7, pp. 1237–1245, Jul. 2018, doi: 10.2337/DB17-0979.
[348] A. Lal, J. L. Parai, and C. M. Milroy, “Liver Pathology in First Presentation Diabetic Ketoacidosis at Autopsy,” *Acad Forensic Pathol*, vol. 6, no. 2, pp. 271–280, Jun. 2016, doi: 10.23907/2016.028.
[349] M. A. Sullivan and J. M. Forbes, “Glucose and glycogen in the diabetic kidney: Heroes or villains?,” *EBioMedicine*, vol. 47, pp. 590–597, Aug. 2019, doi: 10.1016/j.ebiom.2019.07.067.
[350] E. I. Christensen and H. Birn, “Megalin and cubilin: multifunctional endocytic receptors,” *Nature reviews. Molecular cell biology*, vol. 3, no. 4, pp. 258–268, 2002, doi: 10.1038/NRM778.
[351] A. Saito, H. Sato, N. Iino, and T. Takeda, “Molecular Mechanisms of Receptor-Mediated Endocytosis in the Renal Proximal Tubular Epithelium,” *Journal of Biomedicine and Biotechnology*, vol. 2010, 2010, doi: 10.1155/2010/403272.
[352] M. F. Coutinho, M. J. Prata, and S. Alves, “Mannose-6-phosphate pathway: A review on its role in lysosomal function and dysfunction,” *Molecular Genetics and Metabolism*, vol. 105, no. 4, pp. 542–550, Apr. 2012, doi: 10.1016/j.ymgme.2011.12.012.
[353] M. E. Molitch *et al.*, “Development and Progression of Renal Insufficiency With and Without Albuminuria in Adults With Type 1 Diabetes in the Diabetes Control and Complications Trial and the Epidemiology of Diabetes Interventions and Complications Study,” *Diabetes Care*, vol. 33, no. 7, pp. 1536–1543, Jul. 2010, doi: 10.2337/dc09-1098.
[354] M. C. Thomas *et al.*, “Nonalbuminuric Renal Impairment in Type 2 Diabetic Patients and in the General Population (National Evaluation of the Frequency of Renal Impairment cO-existing with NIDDM [NEFRON] 11),” *Diabetes Care*, vol. 32, no. 8, pp. 1497–1502, Aug. 2009, doi: 10.2337/dc08-2186.

[355] H. C. Looker, M. Mauer, and R. G. Nelson, “Role of Kidney Biopsies for Biomarker Discovery in Diabetic Kidney Disease,” *Advances in chronic kidney disease*, vol. 25, no. 2, p. 192, Mar. 2018, doi: 10.1053/j.ackd.2017.11.004.
[356] V. Vallon, “The proximal tubule in the pathophysiology of the diabetic kidney,” *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 300, no. 5, pp. R1009–R1022, May 2011, doi: 10.1152/ajpregu.00809.2010.
[357] A. Ginguay and L. A. Cynober, “Amino Acids | Amino Acid Metabolism,” *Encyclopedia of Biological Chemistry: Third Edition*, vol. 1, pp. 2–9, Jan. 2021, doi: 10.1016/B978-0-12-819460-7.00059-1.
[358] S. Ratner, “Enzymes of arginine and urea synthesis,” *Advances in enzymology and related areas of molecular biology*, vol. 39, pp. 1–90, Jan. 1973, doi: 10.1002/9780470122846.CH1.
[359] H. Kato, I. Oyamada, M. Mizutani-funahashi, and H. Nakagawa, “New radioisotopic assays of argininosuccinate synthetase and argininosuccinase,” *Journal of biochemistry*, vol. 79, no. 5, pp. 945–953, 1976, doi: 10.1093/OXFORDJOURNALS.JBCHEM.A131162.
[360] S. Klahr and J. Morrissey, “L-arginine as a therapeutic tool in kidney disease,” *Seminars in Nephrology*, vol. 24, no. 4, pp. 389–394, Jul. 2004, doi: 10.1016/J.SEMNEPHROL.2004.04.010.
[361] A. A. Reyes, I. E. Karl, J. Kissane, and S. Klahr, “L-arginine administration prevents glomerular hyperfiltration and decreases proteinuria in diabetic rats,” *Journal of the American Society of Nephrology*, vol. 4, no. 4, pp. 1039–1045, Oct. 1993, doi: 10.1681/ASN.V441039.
[362] R. Note *et al.*, “Mitochondrial and metabolic effects of nucleoside reverse transcriptase inhibitors (NRTIs) in mice receiving one of five single- and three dual-NRTI treatments,” *Antimicrobial agents and chemotherapy*, vol. 47, no. 11, pp. 3384–3392, Nov. 2003, doi: 10.1128/AAC.47.11.3384-3392.2003.
[363] I. Audzeyenka, M. Szrejder, D. Rogacka, S. Angielski, M. A. Saleem, and A. Piwkowska, “β-Aminoisobutyric acid (L-BAIBA) is a novel regulator of mitochondrial biogenesis and respiratory function in human podocytes,” *Sci Rep*, vol. 13, p. 766, Jan. 2023, doi: 10.1038/s41598-023-27914-8.
[364] G. Attwood, D. Li, D. Pacheco, and M. Tavendale, “Production of indolic compounds by rumen bacteria isolated from grazing ruminants,” *Journal of Applied Microbiology*, vol. 100, no. 6, pp. 1261–1271, 2006, doi: 10.1111/j.1365-2672.2006.02896.x.
[365] J. Avbersek *et al.*, “Diversity of Clostridium difficile in pigs and other animals in Slovenia,” *Anaerobe*, vol. 15, no. 6, pp. 252–255, Dec. 2009, doi: 10.1016/j.anaerobe.2009.07.004.
[366] N. Ma and X. Ma, “Dietary Amino Acids and the Gut-Microbiome-Immune Axis: Physiological Metabolism and Therapeutic Prospects,” *Comprehensive Reviews in Food Science and Food Safety*, vol. 18, no. 1, pp. 221–242, 2019, doi: 10.1111/1541-4337.12401.
[367] K. Begriche, J. Massart, and B. Fromenty, “Effects of β-aminoisobutyric acid on leptin production and lipid homeostasis: mechanisms and possible relevance for the prevention of obesity,” *Fundamental & Clinical Pharmacology*, vol. 24, no. 3, pp. 269–282, Jun. 2010, doi: 10.1111/J.1472-8206.2009.00765.X.

[368] L. D. Roberts *et al.*, “β-Aminoisobutyric Acid Induces Browning of White Fat and Hepatic β-oxidation and is Inversely Correlated with Cardiometabolic Risk Factors,” *Cell metabolism*, vol. 19, no. 1, p. 96, Jan. 2014, doi: 10.1016/J.CMET.2013.12.003.
[369] C. Maisonneuve *et al.*, “Effects of Zidovudine, Stavudine and β-Aminoisobutyric Acid on Lipid Homeostasis in Mice: Possible Role in Human Fat Wasting,” *https://doi.org/10.1177/135965350400900513*, vol. 9, no. 5, pp. 801–810, Jul. 2004, doi: 10.1177/135965350400900513.
[370] C. L. Linster and E. Van Schaftingen, “Vitamin C: Biosynthesis, recycling and degradation in mammals,” *The FEBS Journal*, vol. 274, no. 1, pp. 1–22, Jan. 2007, doi: 10.1111/J.1742-4658.2006.05607.X.
[371] H. Lee, J. Ahn, S. S. Shin, and M. Yoon, “Ascorbic acid inhibits visceral obesity and nonalcoholic fatty liver disease by activating peroxisome proliferator-activated receptor α in high-fat-diet-fed C57BL/6J mice,” *International Journal of Obesity 2018 43:8*, vol. 43, no. 8, pp. 1620–1630, Oct. 2018, doi: 10.1038/s41366-018-0212-0.
[372] D. F. Garcia-Diaz, P. Lopez-Legarrea, P. Quintero, and J. A. Martinez, “Vitamin C in the Treatment and/or Prevention of Obesity,” *Journal of Nutritional Science and Vitaminology*, vol. 60, no. 6, pp. 367–379, 2014, doi: 10.3177/JNSV.60.367.
[373] M. Y. Lachapelle and G. Drouin, “Inactivation dates of the human and guinea pig vitamin C genes,” *Genetica*, vol. 139, no. 2, pp. 199–207, Feb. 2011, doi: 10.1007/s10709-010-9537-x.
[374] E. Hosseini, C. Grootaert, W. Verstraete, and T. Van de Wiele, “Propionate as a health-promoting microbial metabolite in the human gut,” *Nutrition reviews*, vol. 69, no. 5, pp. 245–258, May 2011, doi: 10.1111/J.1753-4887.2011.00388.X.
[375] A. Visconti *et al.*, “Interplay between the human gut microbiome and host metabolism,” *Nature Communications 2019 10:1*, vol. 10, no. 1, pp. 1–10, Oct. 2019, doi: 10.1038/s41467-019-12476-z.
[376] X. Zheng *et al.*, “The footprints of gut microbial-mammalian co-metabolism,” *Journal of Proteome Research*, vol. 10, no. 12, pp. 5512–5522, Dec. 2011, doi: 10.1021/PR2007945/SUPPL_FILE/PR2007945_SI_001.PDF.
[377] C. L. Linster *et al.*, “Ethylmalonyl-CoA Decarboxylase, a New Enzyme Involved in Metabolite Proofreading,” *J Biol Chem*, vol. 286, no. 50, pp. 42992–43003, Dec. 2011, doi: 10.1074/jbc.M111.281527.
[378] W. Wiechert, “13C metabolic flux analysis,” *Metabolic Engineering*, vol. 3, no. 3, pp. 195–206, 2001, doi: 10.1006/mben.2001.0187.
[379] D. Weindl *et al.*, “Bridging the gap between non-targeted stable isotope labeling and metabolic flux analysis,” *Cancer & Metabolism 2016 4:1*, vol. 4, no. 1, pp. 1–14, Apr. 2016, doi: 10.1186/S40170-016-0150-Z.
[380] D. Gaglio *et al.*, “Oncogenic K-Ras decouples glucose and glutamine metabolism to support cancer cell growth,” *Molecular Systems Biology*, vol. 7, no. 1, p. 523, Jan. 2011, doi: 10.1038/MSB.2011.56.
[381] D. J. Creek, A. Chokkathukalam, A. Jankevics, K. E. V. Burgess, R. Breitling, and M. P. Barrett, “Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation,” *Analytical Chemistry*, vol. 84, no. 20, pp. 8442–8447, Oct. 2012, doi: 10.1021/AC3018795/SUPPL_FILE/AC3018795_SI_003.ZIP.

[382] C. Bueschl *et al.*, “A novel stable isotope labelling assisted workflow for improved untargeted LC–HRMS based metabolomics research,” *Metabolomics*, vol. 10, no. 4, pp. 754–769, Aug. 2014, doi: 10.1007/s11306-013-0611-0.
[383] D. Weindl, A. Wegner, C. Jäger, and K. Hiller, “Isotopologue ratio normalization for non-targeted metabolomics,” *Journal of Chromatography A*, vol. 1389, pp. 112–119, Apr. 2015, doi: 10.1016/j.chroma.2015.02.025.
[384] S. Yang, J. C. Hoggard, M. E. Lidstrom, and R. E. Synovec, “Comprehensive discovery of 13C labeled metabolites in the bacterium Methylobacterium extorquens AM1 using gas chromatography–mass spectrometry,” *Journal of Chromatography A*, vol. 1317, pp. 175–185, Nov. 2013, doi: 10.1016/j.chroma.2013.08.059.
[385] O. E. Albóniga, O. González, R. M. Alonso, Y. Xu, and R. Goodacre, “Optimization of XCMS parameters for LC–MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results,” *Metabolomics*, vol. 16, no. 1, pp. 1–12, Jan. 2020, doi: 10.1007/S11306-020-1636-9/FIGURES/3.
[386] J. Behrmann, C. Etmann, T. Boskamp, R. Casadonte, J. Kriegsmann, and P. Maaß, “Deep learning for tumor classification in imaging mass spectrometry,” *Bioinformatics*, vol. 34, no. 7, pp. 1215–1223, Apr. 2018, doi: 10.1093/BIOINFORMATICS/BTX724.
[387] C. Bueschl *et al.*, “PeakBot: machine-learning-based chromatographic peak picking,” *Bioinformatics*, vol. 38, no. 13, pp. 3422–3428, Jun. 2022, doi: 10.1093/BIOINFORMATICS/BTAC344.
[388] P. Opgenorth *et al.*, “Machine Learning Applications for Mass Spectrometry-Based Metabolomics,” *Metabolites 2020, Vol. 10, Page 243*, vol. 10, no. 6, p. 243, Jun. 2020, doi: 10.3390/METABO10060243.
[389] L. L. Xu, A. Young, A. Zhou, and H. L. Röst, “Machine Learning in Mass Spectrometric Analysis of DIA Data,” *PROTEOMICS*, vol. 20, no. 21–22, p. 1900352, Nov. 2020, doi: 10.1002/PMIC.201900352.
[390] H. A. Krebs and L. V. Eggleston, “The regulation of the pentose phosphate cycle in rat liver,” *Advances in Enzyme Regulation*, vol. 12, pp. 421–434, Jan. 1974, doi: 10.1016/0065-2571(74)90025-9.
[391] K. A. Steer, M. Sochor, and P. McLean, “Renal Hypertrophy in Experimental Diabetes: Changes in Pentose Phosphate Pathway Activity,” *Diabetes*, vol. 34, no. 5, pp. 485–490, May 1985, doi: 10.2337/diab.34.5.485.
[392] E. Akbay *et al.*, “Effects of rosiglitazone treatment on the pentose phosphate pathway and glutathione-dependent enzymes in liver and kidney of rats fed a high-fat diet,” *Current Therapeutic Research*, vol. 65, no. 1, pp. 79–89, Jan. 2004, doi: 10.1016/S0011-393X(04)90007-0.
[393] P. Spegel, A. Chawade, S. Nielsen, P. Kjellbom, and M. Rützler, “Deletion of glycerol channel aquaporin-9 (Aqp9) impairs long-term blood glucose control in C57BL/6 leptin receptor–deficient (db/db) obese mice,” *Physiological Reports*, vol. 3, no. 9, p. e12538, 2015, doi: 10.14814/phy2.12538.
[394] Y. Gu *et al.*, “Very Low Carbohydrate Diet Significantly Alters the Serum Metabolic Profiles in Obese Subjects,” *J. Proteome Res.*, vol. 12, no. 12, pp. 5801–5811, Dec. 2013, doi: 10.1021/pr4008199.