Wrocław University of Science and Technology

FIELD OF SCIENCE: SCIENCE AND NATURAL SCIENCES

DISCIPLINE OF SCIENCE: CHEMICAL SCIENCES

DOCTORAL DISSERTATION

Bioinformatic analysis in targeted metabolomics, automating the process of translating raw NMR spectrum signals into qualitative and quantitative data for use in the analysis of disease states.

Mr. Łukasz Pruss, MSc.

Supervisor:
Prof. Piotr Młynarz, PhD.

Assistant supervisor:
Kaja Milanowska-Zabel, PhD.

Keywords: Bioinformatics, Metabolomics, 1D ^{1}H Nuclear magnetic resonance, Machine learning, Pipeline, Nextflow.

WROCŁAW 2024

Abstract

Metabolomics, a comprehensive study about metabolites, investigates dynamic changes in the concentrations of low-molecular-weight molecules (approx. 1-1.5 kDa) to provide insights into biological conditions influenced by various stressors. The identification of metabolites is commonly achieved using analytical techniques such as nuclear magnetic resonance (NMR) and mass spectrometry (MS), often coupled with separation techniques. Recently, numerous studies have explored the correlation between metabolite concentrations and conditions like diseases, genetic disorders, and cancer progression. Despite its significance in understanding complex biochemical processes, much of the existing metabolomics software remains inaccessible, unsupported, or behind paywalls, with most tools focusing on isolated steps of the analysis process rather than offering comprehensive, automated solutions.

This thesis addresses the growing need for robust and reproducible pipelines in metabolomics. To tackle these challenges, it introduces NASQQ, an open-source Nextflow pipeline specifically designed for the analysis of 1D proton NMR spectra. NASQQ integrates existing solutions with advanced machine learning models and pathway enrichment analysis to provide a robust and reproducible solution for metabolomics research. The implementation of the NASQQ pipeline includes a modular metabolomic workflow written in Nextflow framework, covering spectral processing and data analysis modules, including both univariate and multivariate approaches and pathway analysis.

In this work, the pipeline was evaluated using an open dataset on *Familial Dysautonomia*, involving the analysis of raw serum spectra from both patients and healthy relatives. This evaluation demonstrated the pipeline's effectiveness and applicability in disease studies. The application of NASQQ on *Familial Dysautonomia* samples highlighted significant findings from the spectral processing, univariate tests, machine learning assessments, and pathway enrichment analysis. By leveraging open-source bioinformatics tools, custom functions and machine learning, NASQQ offers an accessible end-to-end workflow that standardizes signal assignment methodologies, reduces operational confusion, and enhances reproducibility through automation and parallelization in a stable containerized environment. The pipeline effectively links raw spectral data with biological interpretation, while also paving the way for future improvements and expanded applications in metabolomics research.

keywords: Bioinformatics, Metabolomics, 1D ^{1}H nuclear magnetic resonance, Machine learning, Pipeline, Nextflow.

Streszczenie

Metabolomika, dziedzina zajmująca się badaniem metabolitów, analizuje dynamiczne zmiany w stężeniach molekuł o niskiej masie cząsteczkowej (ok. 1-1.5 kDa), aby dostarczyć wgląd w biologiczne warunki panujące pod wpływem różnych stresorów. Identyfikacja metabolitów jest zazwyczaj realizowana za pomocą technik analitycznych, takich jak magnetycznego rezonans jądrowy (NMR) i spektrometria mas (MS), często w połączeniu z technikami separacji. W ostatnich latach liczne badania skupiały się na korelacjach pomiędzy stężeniami metabolitów a stanami chorobowymi, zaburzeniami genetycznymi oraz progresją nowotworów. Pomimo swojego znaczenia w zrozumieniu złożonych procesów biochemicznych, wiele istniejącego oprogramowania do metabolomiki pozostaje niedostępne, niewspierane lub płatne, a większość narzędzi koncentruje się na pojedynczych etapach analizy, zamiast oferować kompleksowe, zautomatyzowane rozwiązania.

Niniejsza praca odpowiada na rosnącą potrzebę kompetentnych i powtarzalnych potoków przetwarzania danych (ang. *pipelines*) w metabolomice. Aby sprostać tym wyzwaniom, stworzono NASQQ, otwarty i ogólnodostępny potok przetwarzania danych, zaprojektowany do analizy jednowymiarowych protonowych widm NMR. NASQQ integruje istniejące rozwiązania z zaawansowanymi modelami uczenia maszynowego i analizą wzbogacenia ścieżek, aby zapewnić stabilne i powtarzalne rozwiązania dla badań metabolomicznych. Implementacja NASQQ obejmuje modułową metabolomiczną sekwencję zadań (ang. *workflow*) napisaną w języku Nextflow, obejmującą przetwarzanie widm i moduły analizy danych, w tym zarówno podejścia jednowymiarowe, jak i wielowymiarowe oraz analizę ścieżek biologicznych.

W ramach pracy potok przetwarzania danych został zwalidowany za pomocą otwartego zestawu danych obejmującego Dysautonomię rodzinną (ang. *Familial Dysautonomia*), obejmującego analizę surowych widm surowicy zarówno pacjentów z tą chorobą, jak i ich zdrowych krewnych. Ewaluacja wykazała skuteczność i przydatność NASQQ w badaniach nad chorobami. Co więcej, zastosowanie NASQQ do próbek Dysautonomii rodzinnej podkreśliło w testach wielowymiarowych oraz w analizie wzbogacenia ścieżek biologicznych istotne odkrycia w przetwarzanych widmach. Wykorzystując otwarte narzędzia bioinformatyczne, niestandardowe funkcje i uczenie maszynowe, NASQQ oferuje ogólnodostępny i skuteczny przepływ pracy, który standaryzuje metody przypisywania sygnałów, redukuje błędy operacyjne i zwiększa powtarzalność poprzez automatyzację i równoległe przetwarzanie w stabilnym środowisku opartym na kontenerach (ang. *containerized environment*). Potok przetwarzania danych skutecznie łączy surowe dane widmowe z interpretacją biologiczną, a także otwiera nowe perspektywy na przyszłe usprawnienia i rozszerzone zastosowania w badaniach metabolomicznych.

słowa kluczowe: Bioinformatyka, Metabolomika, 1D ^{1}H magnetyczny rezonans jądrowy, Uczenie maszynowe, Potok przetwarzania danych, Nextflow.

Acknowledgments

I would like to express my deep gratitude and appreciation to my supervisor, Prof. Piotr Młynarz, PhD. His invaluable intellectual support and experience were crucial at many stages of my doctorate. Thanks to him and his research team, many valuable materials, publications, and scientific collaborations were created. His valuable guidance helped steer my doctorate in the right scientific direction. I also thank my assistant supervisor, Kaja Milanowska-Zabel, PhD., for her help in shaping the study according to the research plan, identifying areas that needed refinement during the creation of this work, and for her support during the most challenging times, including emotional support.

I am also grateful to Ardigen SA for the opportunity to pursue my "Implementation Doctorate". I especially thank my colleagues at work for their scientific support, the opportunity to participate in scientific grants, and their support throughout the entire doctoral process. Being part of the Microbiome R&D team at Ardigen allowed me to grow professionally at an incredible pace thanks to working in such a diverse team with varied scientific and professional competencies. Special thanks are due to Janusz Homa, Łukasz Nowak, Jan Majta, Krzysztof Odrzywołek, Tomasz Jetka, Oskar Gniewek, Paweł Biernat, Michał Jakubczak, Konrad Zych, Błażej Szczerba and everyone else who has impact on the creation of this doctoral thesis.

Finally, I want to thank my parents and grandmother for their emotional support, for supporting my education, and for shaping me into the person I have become over the years. A special thank you goes to my wife, Margarita Pruss, for being by my side during the most challenging final stage of my doctorate, making this journey somewhat easier.

Łukasz Pruss

Wrocław, 2024

Table of Contents

Motivation and thesis outline

The motivation behind this thesis is to address the growing need for robust and reproducible pipelines in metabolomics, specifically targeting the niche in solutions for analyzing one-dimensional proton nuclear magnetic resonance (1D ^{1}H NMR) spectra. Metabolomics, a key discipline within systems biology, holds the promise of advancing our understanding of disease mechanisms through the detailed study of small molecules within biological systems. However, the complexity and variability inherent in metabolomic data necessitate advanced bioinformatics tools that can ensure accurate and consistent results. This thesis presents the **NASQQ** (**N**extflow **A**utomatization and **S**tandardization for **Q**ualitative and **Q**uantitative ^{1}H NMR Metabolomic) pipeline, specifically designed for 1D ^{1}H NMR analysis, and evaluates its performance and applicability in metabolomics research, with a focus on disease studies. The majority of methodology and results are part of the original scientific paper on the NASQQ pipeline by Pruss, L et al., 2024, which is currently under resubmission.

The main objective of this thesis is to perform automated bioinformatic analysis of metabolite-derived signals on blood serum spectra obtained by 1D ^{1}H NMR proton magnetic resonance. The side goals include:

- Using bioinformatics tools and machine learning methods to standardize signal assignment methodology and minimize confusion in assigning metabolites to the appropriate signals.
- Automation of the translation process of raw NMR spectra signals, parallelization of the whole process.
- Moving from metabolic profiles to qualitative-quantitative analysis in targeted metabolomics in a fully automated manner.

The thesis aims to explore the following hypotheses:

1) At each stage of the automated pipeline, there is a set of parameters that significantly affect the accuracy of the results obtained.
2) The use of machine learning methods will significantly improve the quality and accuracy of identifying significant metabolites in disease progression.
3) The time required for analysis using the NASQQ pipeline will be much less than manual curation of raw signals while maintaining similar quality.

The following section provides an overall description of all chapters included in the thesis, outlining the scope and content of the work. **Chapter I: Introduction** establishes the foundational knowledge necessary for understanding the rest of the thesis. It begins with an exploration of the fundamental principles of metabolomics within the broader context of systems biology, highlighting various approaches for investigating the metabolome. This chapter then explores the general concepts of nuclear magnetic resonance, with a focus on one-dimensional proton NMR. It further discusses the applications of metabolomics in disease research, detailing studies conducted by the research team as well as by external groups. Finally, the chapter introduces computational methods including bioinformatics, emphasizing its crucial role in metabolomics data analysis, and reviews relevant databases and state-of-the-art tools.

Next chapter, **Chapter II: Methodology - Evaluation of NASQQ pipeline on ^{1}H 1D NMR Spectra** details the methodology employed in the evaluation of the NASQQ pipeline. It provides a general overview of the metabolomic workflow, emphasizing the implementation of NextFlow and the use of a containerized computing environment. The spectral processing of raw 1D spectra and the identification of metabolites are meticulously described, followed by an explanation of the data analysis module, which encompassess both univariate tests and multivariate approaches, including machine learning models. The chapter concludes with a description of the biological interpretation of features derived from the data analysis module.

Third chapter, **Chapter III: Results - Application of NASQQ pipeline on *Familial Dysautonomia* serum samples** presents the results of applying the pipeline to an open dataset of raw spectra from serum samples of patients with *Familial Dysautonomia*. It covers the preparation of the raw spectra and corresponding metadata, the outcomes of the spectral processing modules, and the results from the univariate module tests. Additionally, it assesses the metabolites classified by machine learning models and discusses the intersection of metabolomic pathways based on the KEGG database.

The final chapter, **Chapter IV: Discussion - General conclusion and perspectives**, offers a comprehensive discussion on the general conclusions drawn from the usage of the NASQQ pipeline for metabolomic analysis. It addresses the limitations of the methodologies utilized and proposes future directions and perspectives for further development of the pipeline. The chapter aims to provide a balanced view of the pipeline's strengths and areas for improvement, setting the stage for continued advancements in the field of metabolomics.

The appendix includes supplementary materials that support the main content of the thesis, providing additional data, methodological details, and relevant information alongside full results that enhances the overall understanding of the research conducted.

★ ★ ★

The dissertation was created in collaboration with Ardigen S.A. as part of the "Implementation Doctorate" program (Agreement No. P/0180/434/2020).

Publications

The following scientific papers were produced as part of the doctoral work over the years 2020-2024:

1) Bogunia-Kubik K, Wojtowicz W, Swierkot J, Mielko KA, Qasem B, Wielińska J, Sokolik R, **Pruss Ł**, Młynarz P. Disease Differentiation and Monitoring of Anti-TNF Treatment in Rheumatoid Arthritis and Spondyloarthropathies. Int J Mol Sci. 2021 Jul 15;22(14):7389. doi: 10.3390/ijms22147389. PMID: 34299109.
2) Mielko KA, Jabłoński SJ, **Pruss Ł**, Milczewska J, Sands D, Łukaszewicz M, Młynarz P. Metabolomics Comparison of Drug-Resistant and Drug-Susceptible Pseudomonas aeruginosa Strain (Intra- and Extracellular Analysis). Int J Mol Sci. 2021 Oct 1;22(19):10820. doi: 10.3390/ijms221910820. PMID: 34639211.
3) Wojciechowski S, Majchrzak-Górecka M, Biernat P, Odrzywołek K, **Pruss Ł**, Zych K, Majta J, Milanowska-Zabel K. Machine learning on the road to unlocking microbiota's potential for boosting immune checkpoint therapy. Int J Med Microbiol. 2022 Oct;312(7):151560. doi: 10.1016/j.ijmm.2022.151560. PMID: 35238088.
4) Pudełko-Malik N, Drulis-Fajdasz D, **Pruss Ł**, Mielko-Niziałek KA, Rakus D, Gizak A, Młynarz P. A single dose of glycogen phosphorylase inhibitor improves cognitive functions of aged mice and affects the concentrations of metabolites in the brain. [**manuscript under review**]. Scientific reports. 2024.
5) **Pruss Ł**, Gniewek O, Jetka T, Wojtowicz W, Milanowska-Zabel K, Młynarz P. NASQQ: Nextflow automatization and standardization for 1H NMR metabolomics data preparation and analysis [**manuscript during resubmission**]. Oxford GigaScience, Technical note. 2024.

Abbreviations

16S rRNA: 16S Ribosomal Ribonucleic Acid;

AI: Artificial Intelligence;

ANOVA: Analysis of Variance;

AS: *Ankylosing Spondylitis*;

ASICS: Automatic Statistical Identification in Complex Spectra;

AWS: Amazon Web Services;

BMRB: BioMagResBank;

BMI: Body Mass Index;

CMD: Command prompt;

CNN: Convolutional Neural Network;

CPU: Central Processing Unit;

CRP: C-reactive Protein;

ChemFOnt: Chemical Functional Ontology;

DAG: Directed Acyclic Graph;

DAS28: Disease Activity Score-28;

Da: Dalton;

DL: Deep Learning;

DNA: Deoxyribonucleic Acid;

DSS: 4,4-Dimethyl-4-silapentane-1-sulfonic acid;

EDA: Exploratory Data Analysis;

EMBL-EBI: European Molecular Biology Laboratory - European Bioinformatics Institute;

FD: *Familial Dysautonomia*;

FAIR: Findable, Accessible, Interoperable, and Reusable;

FDR: False Discovery Rate;

FFT: Fast Fourier Transform;

FID: Free Induction Decay;

FT: Fourier Transform;

FT-ICR: Fourier Transform Ion Cyclotron Resonance;

FWER: Family-Wise Error Rate;

GC: Gas Chromatography;

GCG: glucagon gene;

GSEA: Gene Set Enrichment Analysis;

GUI: Graphical User Interface;

GMP: Guanosine Monophosphate;

HMDB: Human Metabolome Database;

HPC: High-Performance Computing;
HRMS: High-Resolution Mass Spectrometry;
Hz: Hertz;
ICT: Immune Checkpoint Therapy;
IMP: Inosine Monophosphate;
ID: Identifier;
ISS: International Space Station;
I/O: Input/Output;
IQR: Interquartile Range;
KEGG: Kyoto Encyclopedia of Genes and Genomes;
LASSO: Least Absolute Shrinkage and Selection Operator;
LC: Liquid Chromatography;
LOF: Local Outlier Factor;
ML: Machine Learning;
MS: Mass Spectrometry;
MS2: Tandem Mass Spectrometry;
MSI: Metabolomics Standards Initiative;
Mnova: Mestrelab NMR and MS Analysis;
MHz: Megahertz;
NAFLD: Non-Alcoholic Fatty Liver Disease;
NMR: Nuclear Magnetic Resonance;
NN: Neural Network;
PCA: Principal Component Analysis;
PDB: Protein Data Bank;
PLS-DA: Partial Least Squares Discriminant Analysis;
PQN: Probabilistic Quotient Normalization;
PsA: *Psoriatic Arthritis*;
PPM: Parts Per Million;
RAM: Random Access Memory;
RA: *Rheumatoid Arthritis*;
RDS: R Data Store;
RF: Radio Frequency;
RFE: Recursive Feature Elimination;
RMS: Root Mean Square;
RNA: Ribonucleic Acid;
ROC AUC: Receiver Operating Characteristic Area Under the Curve;
ROTS: Reproducibility-Optimized Test Statistic;

SNR: Signal-to-Noise Ratio;

SVG: Scalable Vector Graphics;

SVM: Support Vector Machine;

T: Tesla;

TCA cycle: Tricarboxylic Acid Cycle;

TMAO: Trimethylamine N-oxide;

TMS: Tetramethylsilane;

TSP: Trimethylsilyl Propionate;

UMP: Uridine Monophosphate;

UPLC: Ultra Performance Liquid Chromatography;

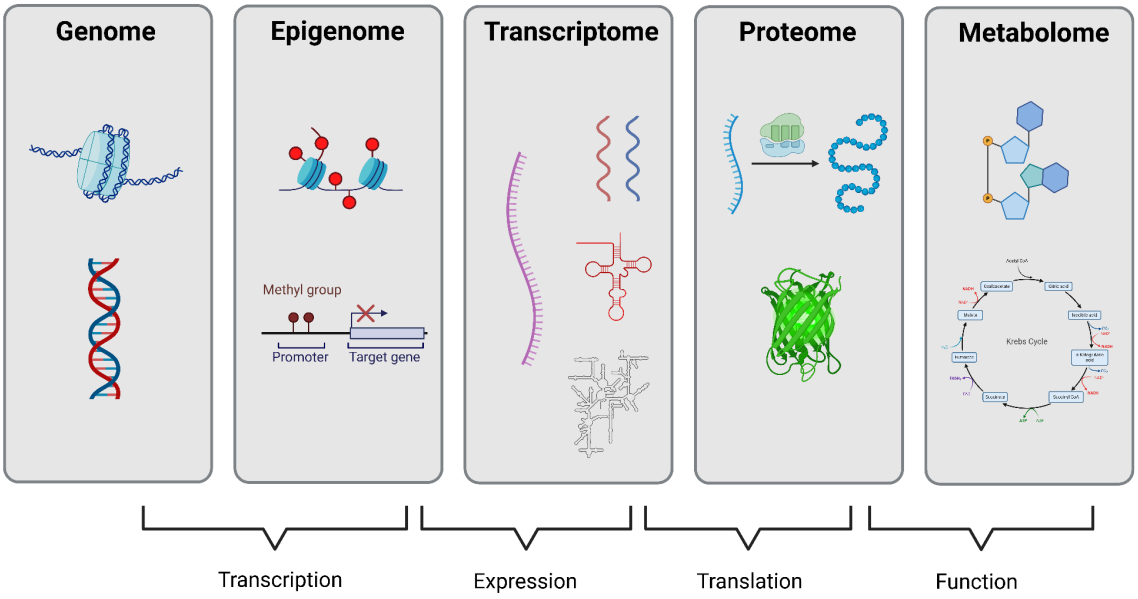
XML: Extensible Markup Language.

I. Introduction

I.1 Fundamental principles of metabolomics within systems biology

Metabolomics is the large-scale, comprehensive analysis of chemical processes that provides valuable insights into biological conditions by analyzing quantitative and qualitative alterations of metabolites [1]. Metabolites are small organic compounds typically ranging in molecular weight from 30 (formaldehyde, methanol, methylamine) to 1500 (somatostatin, aquacobalamin) daltons [Da] that are substrates, intermediate or final products of metabolism within biological systems. These molecules are omnipresent within cells, biofluids, tissues, organs or whole organisms, where they play crucial roles in cellular functions and physiological processes [2]. Metabolomics, alongside disciplines such as genomics, epigenomics, transcriptomics, and proteomics, is an integral part of "Systems Biology" – a holistic approach to understanding biological systems at the molecular level (see **Figure 1**). While each discipline focuses on distinct aspects of cellular function, they collectively contribute to a comprehensive understanding of biological processes [3].

Figure 1. Categorization of omics within the systems biology framework.



Source: Adapted from [3], created in Biorender.com.

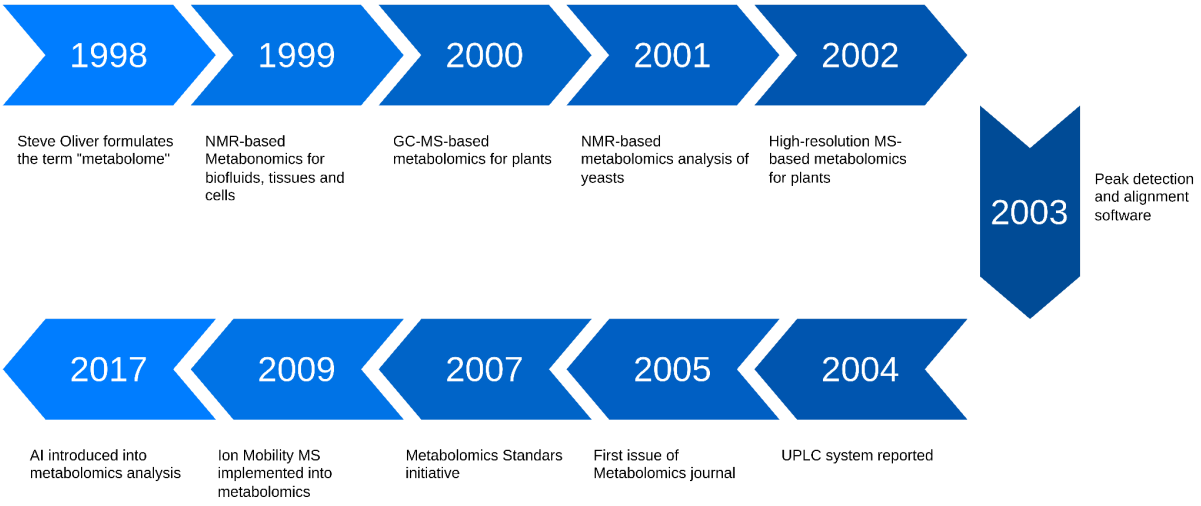
Genomics examines the complete set of genes within an organism, providing insights into DNA sequences and genetic variations [4-6]. Epigenomics focuses on the study of epigenetic modifications that influence gene expression without altering the underlying DNA sequence. These modifications, which include DNA methylation, histone modifications, and non-coding RNA regulation, play a significant role in regulating gene activity and cellular function. Epigenomics complements other omics disciplines by providing a deeper understanding of the regulatory

mechanisms that affect gene expression and cellular function [7-8]. Transcriptomics studies gene expression patterns by focusing on the transcriptome, encompassing all RNA molecules transcribed from the genome [9]. Proteomics explores the entire complement of proteins expressed within a cell or organism, shedding light on protein structure, function, and interactions [10]. Metabolomics offers insights into the dynamic interaction of small molecules (1-1.5 kDa) within biological systems. The complete set of metabolites inside a biological system and their interactions at any given time point are known as the metabolome. The metabolome represents the complete set of metabolites in a biological cell, tissue, organ, or organism, which are the substrates, intermediate substances or products of cellular processes. By quantifying and characterizing metabolite profiles or metabolomes, the study has a chance to unravel the metabolic pathways, regulatory networks, and metabolic signatures underlying cellular processes and responses to internal and external stimuli, ultimately contributing to our understanding of health, disease, and environmental interactions [11-12].

I.1.1 Approaches for investigating the metabolome

Over the years, the field of metabolomics has undergone significant evolution, driven by advancements in analytical chemistry and computational biology (demonstrated on **Figure 2**) [13]. In 2000, shortly after the conceptualization of metabolomics, Fiehn et al. [14] demonstrated gas chromatography (GC) and mass spectrometry (MS) for metabolomic analysis in plants. In the early days of metabolomics, basic analytical techniques such as gas chromatography and mass spectrometry were primarily relied upon to characterize metabolite composition in biological samples. These techniques laid the foundation for metabolomic studies by offering initial insights into the broad spectrum of metabolites presented in biological systems. This approach enabled the analysis of volatile and thermally stable metabolites, making it well-suited for the study of small molecules like amino acids, organic acids, fatty acids etc. [15-16]. Similarly, liquid chromatography coupled with mass spectrometry (LC-MS) has arisen as a powerful analyzing system in metabolomics, offering high sensitivity and selectivity for detecting and quantifying metabolites. LC-MS combines the separation capabilities of liquid chromatography with the detection and quantification capabilities of mass spectrometry, allowing for the analysis of a wide range of metabolites, including polar and non-polar compounds. This technique has been used in the identification and quantification of metabolites in complex biological samples such as blood, urine, and tissue extracts [17-18].

Figure 2. Timeline of significant events in the field of metabolomics.



Source: Based on [13], created in Draw.io.

However, as metabolomics evolved, so did the need for more sophisticated devices and computational tools to address the complexity of biological samples. In 2004, Plumb et al. [19] presented a significant breakthrough with the introduction of ultra-performance liquid chromatography (UPLC), enhancing separation technology for metabolomics. The development of high-resolution mass spectrometry (HRMS) platforms, such as Fourier transform ion cyclotron resonance (FT-ICR) and Orbitrap mass analyzers, revolutionized metabolomic analyses by offering enhanced mass resolution, accuracy, and dynamic range [20]. Time-of-flight (TOF) and quadrupole-based mass analyzers allow identification of metabolites based on their mass spectra, contributing to the characterization of metabolic pathways and metabolic signatures associated with various physiological conditions [21].

Concurrently, nuclear magnetic resonance (NMR) spectroscopy has also played a significant role in metabolomic analyses by offering a non-destructive and highly reproducible method for studying metabolite structure and dynamics in unmodified biological samples. In contrast to MS, NMR provides quantitative results with minimal sample preparation without the need of separation methods like liquid or gas chromatography. Furthermore, NMR analysis preserves samples without causing any damage during the process of acquiring spectrum. Samples analysis can be accomplished in just a few minutes, which allows immediate confirmation of the structure of a small molecule compound. These capabilities, coupled with the ability to automatically collect data from numerous samples, have made NMR widely accessible across various public institutions, including colleges, universities, and private chemical or pharmaceutical companies [22-23]. In 1998, Steve Oliver and colleagues coined the term “Metabolome” setting the stage for advancements in metabolomics [24]. Shortly thereafter, Nicholson et al. introduced NMR-based metabonomics for the analysis of biofluids,

tissues, and cells [25]. Building on this progress, Raamsdonk et al. demonstrated the application of NMR-based metabolomics specifically in yeast in 2001 [26]. Consequently, NMR has become an essential tool for diverse research programs. Early NMR studies provided foundational insights into metabolite composition and metabolic pathways, laying the groundwork for further advancements in the field. The combination of HRMS and NMR spectroscopy has significantly expanded the analytical capabilities of metabolomics, allowing to obtain complementary information about metabolite identities, concentrations, and interactions within biological systems [27-29].

Nowadays, the field encompasses a wide range of methodologies and techniques developed over time to measure and investigate metabolites. Each method comes with its own set of advantages and limitations, making the selection of an appropriate approach a mandatory step in study design. Metabolomic studies typically adopt either a targeted or untargeted approach, each guiding decisions related to experimental design, sample preparation, and analytical techniques. Untargeted methods aim to comprehensively analyze samples, detecting and quantifying as many metabolites as possible without prior assumptions. In contrast, targeted approaches focus on specific metabolite groups to investigate treatment effects or genetic modifications. Within these general frameworks, a problem-dependent choice from a multitude of techniques is made, with MS and NMR spectroscopy still emerging as primary tools on the global stage, owing to their versatility and robustness in metabolite analysis [30-32].

I.1.2 General concepts of nuclear magnetic resonance

Nuclear magnetic resonance spectroscopy is a spectroscopic technique that utilizes as its core the unique behavior of various atomic nuclei, such as ^{1}H, ^{13}C, ^{31}P, ^{15}N. Fundamentally, NMR spectroscopy relies on the nuclear resonance, wherein atomic nuclei reorient themselves in response to an external magnetic field, resonating at characteristic frequencies [33]. This phenomenon operates based on the principle of the Larmor equation, which describes the resonance frequency (ω in Hz or MHz) of a nucleus with a magnetic moment (γ) in an external magnetic field (B_0 in T):

$$\omega = \gamma B_0 \tag{1.1}$$

where:

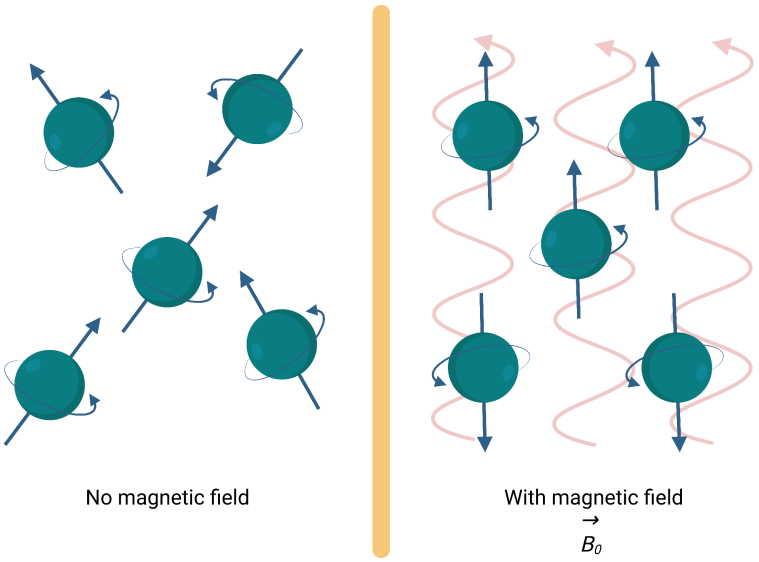
- ω is the resonance frequency,
- γ is the gyromagnetic ratio (specific to each nucleus),
- B_0 is the strength of the external magnetic field.

This resonance behavior originates from the property of atomic nuclei known as spin [34]. Each nucleus possesses an native angular momentum or spin (I), which is quantized in discrete units. The associated magnetic moment ($\vec{\mu}$) of a nucleus is collinear with its spin moment ($\vec{I}$) and is given by:

$$\vec{\mu} = \gamma \hbar \vec{I} \quad (1.2)$$

Here, γ represents the gyromagnetic ratio of the nucleus, expressed in rad T^{-1} s^{-1}, and $\hbar$ is Planck's constant ($\hbar = 1.054 \times 10^{-34} J \cdot s$) divided by 2π. All atomic nuclei characterized by a non-zero spin moment ($I \neq 0$) are observable by NMR. When there is no external magnetic field present, the nuclear spins within the ensemble are randomly oriented, where each nucleus behaves like a bar magnet. However, once an external magnetic field is applied $\vec{B}_0 > 0$, the nuclei respond by aligning themselves (see **Figure 3**).

Figure 3. Alignment of nuclear spins in the presence of a magnetic field.



Source: Based on [34], created in Biorender.com.

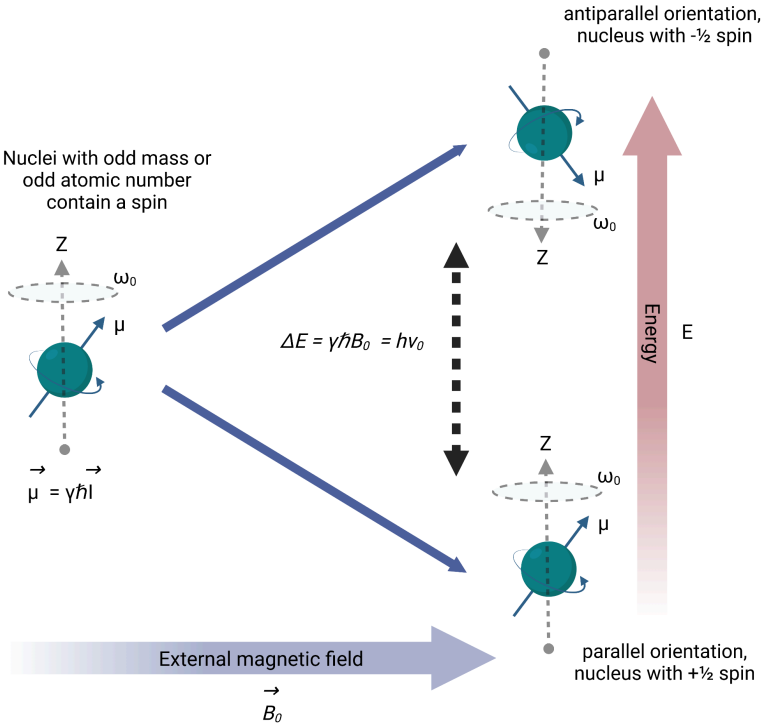
For a given nucleus, the spins can exist in two states: + ½ and − ½, corresponding to two potential orientations of the magnetic moment ($\vec{\mu}$). For instance, the proton of ^{1}H has a spin $I = ½$ and is the most recorded nucleus in NMR due to its high gyromagnetic ratio (42.576 MHz/T) and natural abundance of more than 99.98 %. Consequently, ^{1}H NMR provides a fundamental framework for analyzing the structures and dynamics of molecules across various chemical and biological systems. The widespread presence of hydrogen in known molecules and the highly distinctive nature of NMR spectra for individual compounds and functional groups further enhances its utility in scientific investigations, particularly those of organic compounds [35-36].

By manipulating magnetic fields and applying radiofrequency (RF) pulses and flipping the nuclei from its + ½ position to the − ½, NMR spectroscopy generates complex signals known as free induction decays (FID), which contain information about the molecular composition and structure [37]. The frequency is determined by the equation:

$$\Delta E = h\nu_0 \quad (1.3)$$

where ΔE represents the energy difference between two spin states, h is Planck's constant and ν_0 is resonant frequency corresponding to the energy gap between two spin states induced by B_0. More detailed information is included in **Figure 4**.

Figure 4. Spin orientation and energy transitions in NMR.



Source: Adapted from @steve_woodley NMR principle template, created in Biorender.com.

In NMR spectroscopy, the initial magnetization (M_0) of the sample plays a significant role in determining the outcome of pulse sequences. Before any RF pulses are applied, the sample's nuclear spins are aligned along the direction of the external magnetic field (B_0), resulting in an equilibrium magnetization represented by M_0. Precise control and manipulation of M_0 are essential for designing effective pulse sequences that influence nuclear spins during NMR experiments. Following this, each spin experiences precession at its unique Larmor frequency around the z-axis, thereby inducing a signal in the receiver coil. However, motions within the solution that lead to time-varying magnetic fields result in spin relaxation, causing the received FID signal to gradually diminish [38-39].

The chemical environment of a nucleus influences its resonance frequency through the shielding effect. Shielding of the nucleus alters the resonance frequency, leading to different chemical shifts for nuclei in distinct environments. Chemical shifts are reported relative to a reference signal, often e.g. tetramethylsilane (TMS), in parts per million (ppm), providing valuable structural

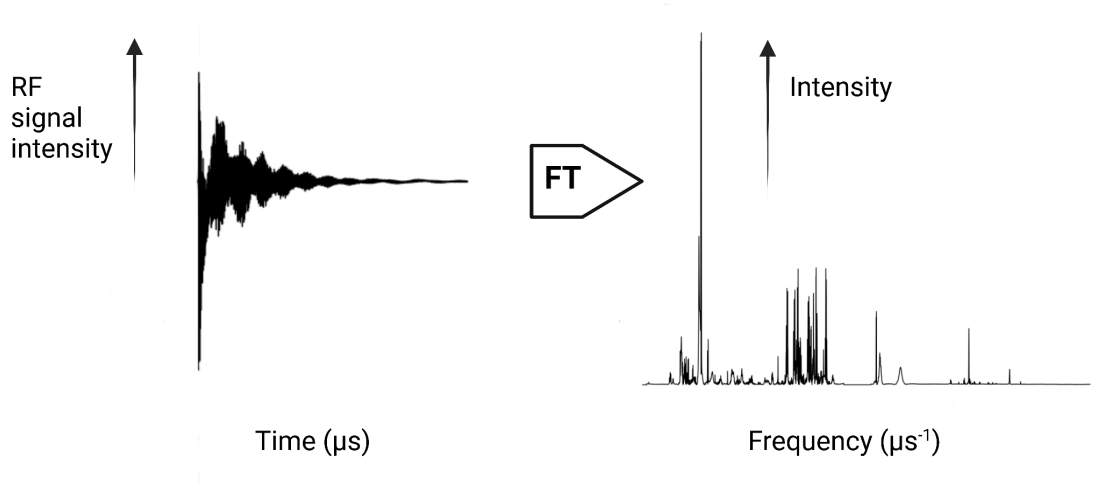
information. The dispersion of chemical shift values varies among nuclei, with ^{1}H signals exhibiting a small dispersion compared to other nuclei. Understanding this information is crucial for interpreting NMR spectra and revealing molecular structures [41-42].

The Fourier transformation (FT) of the free induction decay data obtained from nuclear magnetic resonance spectroscopy enables the conversion of time-domain signals into frequency-domain spectra. This transformation enables the identification and analysis of various molecular properties, including chemical shifts [43]. Mathematically, the Fourier transform is represented by the equation:

$$F(\nu) = \int_{-\infty}^{\infty} f(t) e^{-i2\pi\upsilon t} dt \quad (1.4)$$

where $F(\nu)$ is the frequency-domain spectrum, $f(t)$ is the time-domain signal (FID), υ is the frequency, i represents the imaginary unit and dt is the time interval between data points in the FID. This transformation allows analysis of the spectral components of the signal in terms of their frequencies rather than their amplitudes over time. The Fourier transform aids in determining molecular structures and dynamics by revealing resonance peaks corresponding to specific atomic nuclei and their interactions (see **Figure 5**) [44].

Figure 5. Converting time-domain FID to frequency-domain spectrum using Fourier transform (FT).



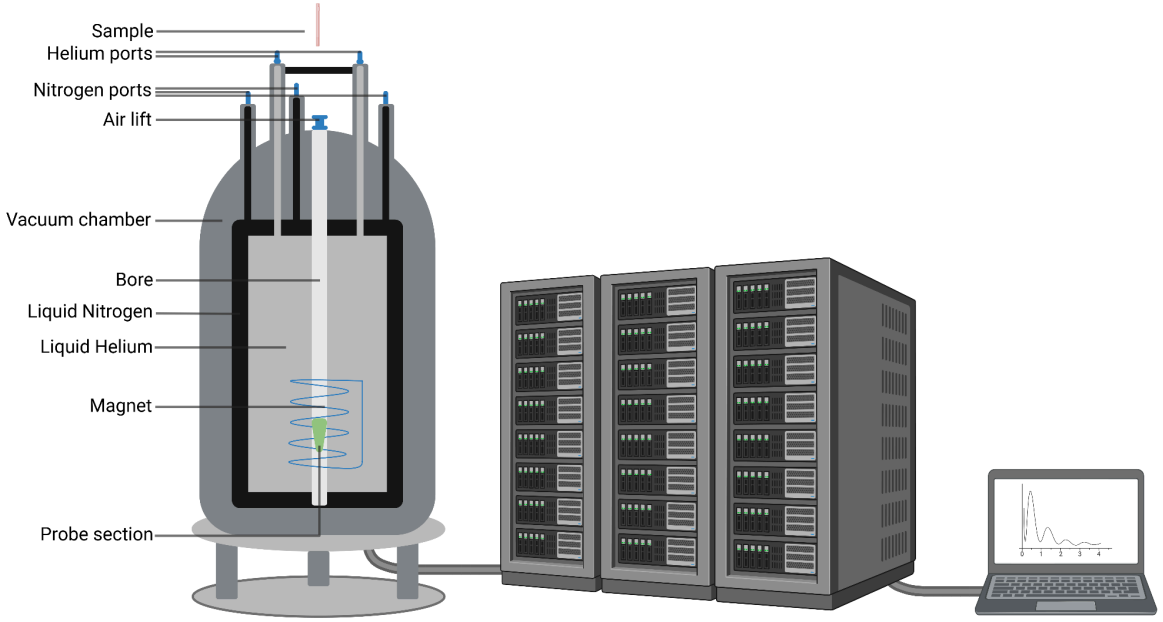
Source: Created in GIMP, own elaboration.

I.1.3 One-dimensional ^{1}H nuclear magnetic resonance

The construction of an NMR device involves integrating diverse components to enable precise manipulation and analysis of atomic nuclei within a sample (see **Figure 6**) [45-46]. At its core lies a powerful magnet that generates a strong and uniform magnetic field that is essential for NMR experiments. Surrounding the magnet is a vacuum chamber that maintains the required environment,

including liquid nitrogen and liquid helium for cooling purposes. The sample is positioned within the bore, held by a sample lift mechanism that uses air to move the sample. Supporting this setup is a magnet controller, featuring sophisticated electronic configurations and controls to regulate magnetic field intensity and stability. Helium and nitrogen ports allow for the controlled introduction of cooling agents. The device includes an RF transmitter calibrated to emit radiofrequency pulses targeted at the sample, inducing transitions between nuclear spin states and emitting NMR signals. These signals are captured by a sensitive detector located in the probe section, equipped with specialized coils and electronics. Finally, advanced data acquisition and processing units, including manufacturer software algorithms and computing resources, analyze and interpret the acquired NMR data.

Figure 6. Schematic diagram of an NMR spectrometer build.



Source: Based on [46], created in Biorender.com.

In an NMR experiment, one (1D) or a series (2D, 3D) of free induction decay curves are detected and recorded [47]. The process of one-dimensional NMR spectroscopy typically involves three phases: sample preparation, pulse (application of radiofrequency pulses to manipulate magnetization), and signal detection (followed by further data analysis). During the pulse phase, a specific pulse program is applied to manipulate the spin system and achieve a desired state, causing precession at each spin's unique Larmor frequency. However, spin relaxation occurs due to motions in the solution, gradually diminishing the received FID signal. To address this, NMR multi-pulse sequences like CMPG1D have been developed, incorporating RF pulses, timed delays, and magnetic field gradients for systematic manipulation of nuclear spins. To enhance the signal-to-noise ratio (SNR), experiments are repeated multiple times, and the data are summed before undergoing Fourier

transformation to produce the final 1D spectrum. Following preparation and detection, the measurable signal (FID) is acquired over several seconds, accumulating over repeated scans to improve SNR. [48]

Analyzing peak positions and intensities in the resulting spectrum provides insights into the chemical composition, structure, and dynamics of molecules in the sample, aiding in understanding molecular properties and interactions. Analysis of such data involves techniques like peak integration, spectral fitting and where applicable, deconvolution. Spectral alignment and peak integration ensure accurate comparison and integration of data across samples. Deconvolution helps distinguish and quantify individual components from complex mixtures, enabling detailed characterization and understanding of molecular structures and interactions. However, due to its complexity, deconvolution is often considered an auxiliary approach in spectroscopic studies. Despite challenges in metabolite identification due to sample complexity, advances in spectral databases and computational tools have improved capabilities. These tools extract valuable information from NMR spectra, revealing chemical shifts, molecular connectivity, and quantitative data about metabolites [49]. Innovative computational methods, such as pattern recognition algorithms [50-51], machine learning approaches [52], and metabolic pathway analysis tools [53], are also employed for metabolite identification and characterization. These methods utilize spectral features, chemical properties, and biological context to accurately annotate metabolites and infer metabolic pathways. By combining multiple analytical techniques and data integration strategies, it is possible to enhance metabolite identification and gain deeper insights into metabolic processes and disease mechanisms [54].

Since its inception in the mid-20th century, 1D NMR and its applications have undergone significant evolution, expanding over the years to accommodate a wide range of research needs [55]. The emphasis of the work in this thesis is primarily on one-dimensional solutions, while leaving room for potential integration of multidimensional input handling in the future.

I.2 Metabolomics applications in disease research

Recent evidence increasingly supports the significance of specific metabolites in various diseases, spanning civilization diseases, genetic neurological disorders, and cancer [56-58]. Metabolites detected and analyzed in biological samples act as biomarkers, offering valuable insights into disease onset, progression, and treatment. Analyzing shifts in metabolite concentrations holds promise for assessing biomarkers of disease susceptibility, refining diagnostic methodologies, evaluating treatment efficacy, and optimizing pharmaceutical dosing regimens. This emphasizes metabolomics as a powerful tool for biomarker discovery and disease diagnosis, crucial in disease research [59].

I.2.1 Importance of metabolomics shown in own research

In the following section, a description of the research conducted by our team is provided, as documented in publications [64, 67]. These studies highlight our exploration into autoimmune diseases and antibiotic resistance through advanced metabolomics approaches. Autoimmune diseases encompass a diverse group of conditions where the body's immune system mistakenly targets its tissues, leading to chronic inflammation and tissue damage. This dysregulation can result from a complex interplay of genetic predisposition and environmental triggers [60]. For instance, rheumatoid arthritis, characterized by joint inflammation and destruction, is influenced by both genetic factors and environmental exposures such as smoking [61]. Similarly, systemic lupus erythematosus involves the immune system attacking various organs and tissues, with genetic susceptibility and environmental factors like infections contributing to its development [62]. Multiple sclerosis, an autoimmune disorder affecting the central nervous system, is thought to arise from a combination of genetic predisposition and environmental triggers [63]. Understanding aforementioned factors is essential for advancing disease research, aiding in the development of precise therapies, and enhancing patient outcomes. In study [64], metabolomics analyses were conducted to investigate alterations in low-molecular-weight compounds across patients diagnosed with rheumatoid arthritis (RA), ankylosing spondylitis (AS), and psoriatic arthritis (PsA). Metabolomics analyses were conducted at three intervals: before treatment initiation and at 3 and 6 months following the administration of a biologic agent. This approach enabled the detailed tracking of biochemical changes over time, providing insights into the metabolic pathways affected by these diseases and their treatments. Tracking parameters like DAS28 (Disease Activity Score using 28 joint counts, assessing disease activity in rheumatoid arthritis based on joint tenderness and swelling) and CRP (C-reactive protein, indicating inflammation) in RA patients showed notable improvement after 6 months, with partial improvement noted at 3 months. Analysis of AS-tracking parameters showed improvements after 3 months, however, PsA patients showed unsatisfactory disease activity parameters even after 6 months. The study identified several metabolite alterations: RA showed increases in alanine, tryptophan, tyrosine, glutamine, leucine, isoleucine, and keto acids like 2-oxoisocaproate and 3-methyl-2-oxovalerate, alongside a rise in citrate. AS patients exhibited elevated levels of glutamine, tryptophan, histidine, leucine, citrate, and formic acid, while isobutyrate and acetone levels decreased. In PsA, notable changes included a decrease in acetate and ethanol levels.The study identified several metabolite alterations: RA showed increases in alanine, tryptophan, tyrosine, glutamine, leucine, isoleucine, and keto acids like 2-oxoisocaproate and 3-methyl-2-oxovalerate, alongside a rise in citrate. AS patients exhibited elevated levels of glutamine, tryptophan, histidine, leucine, citrate, and formic acid, while isobutyrate and acetone levels decreased. In PsA, notable changes included a decrease in acetate and ethanol levels. Study has demonstrated a link between DAS28 and histidine, among other amino acids, indicating treatment effectiveness,

as lower inflammation is associated with higher levels of specific amino acids. Furthermore, associations between TLR4 gene polymorphisms and RA progression, as well as response to anti-TNF therapy, suggest a role for the microbiome in the development of rheumatic diseases. In summary, the study emphasizes the significance of metabolomics in uncovering potential biomarkers and understanding the biochemical mechanisms of autoimmune diseases. It also highlights the difficulties in treating PsA and proposes that targeted therapies, such as microbiome modulation, could enhance treatment outcomes for rheumatic diseases.

Metabolomics emerges its value in enhancing our understanding of antibiotic resistance mechanisms of microbiota [65]. Leveraging NMR spectroscopy, key metabolites and pathways implicated in antibiotic resistance can be pinpointed, laying the groundwork for targeted interventions and improvements in disease management strategies [66]. For instance, in a recent study [67], the metabolic profiles of two strains of *Pseudomonas aeruginosa*, encompassing both antibiotic-susceptible and antibiotic-resistant phenotypes, were analyzed. Pseudomonas is a genus of bacteria known for its adaptability to diverse environments, ranging from soil and water to the human body. Certain species, such as *Pseudomonas aeruginosa*, are notorious pathogens capable of causing a wide range of infections, particularly in immunocompromised individuals [68]. Investigation yielded direct molecular insights into *P. aeruginosa*'s response to antibiotics, highlighting metabolic disparities between resistant and susceptible strains. Notably, these differences appear to correlate with the activation of antibiotic resistance mechanisms. Comparative analysis of intracellular and extracellular metabolite profiles unveiled distinct metabolic signatures between drug-resistant and drug-susceptible *Pseudomonas aeruginosa* strains, particularly within the intracellular amino acid pool. The study identified significant alterations, including decreased pyruvate and increased lactate levels in the resistant strain, suggesting a more intense reduction reaction of pyruvate. Additionally, branched-chain amino acids (valine, leucine, and isoleucine) were upregulated, while isocitrate levels were lower, and succinate levels were higher in the resistant strain. Most intracellular amino acids were elevated in the resistant strain, except for glycine, glutamate, and tyrosine, which were also higher in this strain. Moreover, UMP levels were notably lower in the resistant strain, which is significant as UMP is involved in virulence and biofilm formation. The concentration of intracellular free amino acids reflects a balance between various processes, including protein synthesis, environmental uptake, biosynthesis, and degradation. This comprehensive understanding of metabolic dynamics holds promise in guiding the selection of optimal therapeutic approaches and identifying potential targets for future drug development, thus advancing our arsenal against antibiotic-resistant pathogens. These metabolomics-driven findings have significant implications for both current treatment strategies and the development of novel therapeutics to combat antibiotic resistance, ultimately advancing our arsenal against antibiotic-resistant pathogens.

I.2.2 Other examples of metabolomics importance

In functional genomics, metabolomics serves as an invaluable resource for identifying phenotypic changes resulting from genomic alterations [69]. *Familial dysautonomia* (FD) stands as a rare genetic neurologic disorder characterized by impaired neuronal development and progressive degeneration affecting both the peripheral and central nervous systems. The condition is monogenic, with the majority of patients harboring an identical point mutation in the elongator acetyltransferase complex subunit 1 (ELP1) gene, rendering it a relatively straightforward genetic landscape for the identification of modifiable factors influencing its pathology [70-71]. As demonstrated in the FD studies [72-73], gastrointestinal symptoms and metabolic deficits are prevalent among FD patients, reinforcing the notion that the gut microbiome and metabolome undergo alterations and dysfunction compared to those of healthy individuals. Investigation revealed significant discrepancies in gut microbiome composition, as evidenced by 16S rRNA gene sequencing of stool samples, alongside distinct profiles in stool and serum metabolomes determined through NMR analysis, when comparing a cohort of FD patients with their healthy cohabitating relatives. In terms of identified metabolites, FD patients exhibited significantly lower levels of xanthine and methanol in their serum, alongside elevated levels of urea. Additionally, stool samples from FD patients showed significantly elevated choline levels, which correlated with lower microbiome alpha diversity and decreased richness. These findings align with clinical reports suggesting metabolic energy deficits in FD patients and highlight additional metabolic pathways implicated in the disease's phenotypic expression. Overall, the study provides evidence linking ELP1 gene mutation in FD with profound metabolic alterations, offering insights into potential therapeutic targets for mitigating neurodegeneration in FD patients. Additionally, analysis demonstrated that key discoveries made in human subjects are also replicated in a neuron-specific Elp1-deficient mouse model.

Metabolomics has found intriguing application in research investigating multi-omics samples from rodent models subjected to gravity conditions. In a recent study [74], liver tissues from mice transported to the International Space Station (ISS) underwent histological, transcriptomic, and proteomic analyses. Pathway analysis was subsequently performed on the data from these animals to uncover molecular changes elucidated by omics techniques. The findings revealed significant alterations in lipid localization, as well as lipid and fatty acid metabolism and processing, irrespective of strain or flight conditions. The analysis revealed fluctuations in pathways, with notable similarities to metabolic changes observed in NAFLD (nonalcoholic fatty liver disease). Specifically, disruptions in lipid metabolism were evident, resembling early stages of NAFLD characterized by insulin resistance and altered fatty acid metabolism, leading to liver lipid accumulation. Proteomic analysis highlighted inhibition of apolipoproteins, potentially increasing NAFLD risk, along with dysregulation of glucagon (GCG) and insulin pathways. Activation of glucose metabolism pathways and upregulation of cell cycle pathways further suggested disturbances in lipid metabolism associated

with NAFLD. Disrupted circadian rhythm in microgravity environments may exacerbate NAFLD progression, potentially through increased lipid deposition in the liver. Findings suggest persistent activation of NAFLD-related pathways in space, indicating gradual lipid deposition in the liver and increased risk of NASH and irreversible liver fibrosis. Metabolomics alongside other omics has proven valuable in exploring the impact of gravity conditions. Nevertheless, further studies are necessary to understand the underlying mechanisms and assess risks to astronauts during longer-duration spaceflight missions.

The examples provided demonstrate that metabolomics alone, as a component of systems biology, offers numerous advantages in disease research. However, when integrated into a more comprehensive approach like multi-omics, its potential becomes even more pronounced. Multiomics methodologies aim to simultaneously analyze diverse molecular components, enabling a more comprehensive grasp of biological processes, disease mechanisms, and complex phenotypes. Nevertheless, implementing such approaches necessitates not only meticulous experimental design, appropriate equipment, and skilled personnel but also relies on advanced computational methods and substantial computing resources, particularly vital in deciphering multi-omics samples [75]. Bioinformatics approaches are crucial for processing and analyzing the vast amount of data generated in metabolomics studies. Bioinformatics tools and algorithms assist in identifying metabolites, analyzing pathways, and integrating metabolomic data with other omics datasets like genomics and proteomics. This integration enables a comprehensive understanding of disease mechanisms and facilitates the discovery of biomarkers.

I.3 Computational methods in metabolomics

Within metabolomics research, several computational methods are employed to dissect and interpret the complex metabolic data. Bioinformatics plays a crucial role in processing and analyzing large-scale metabolomic datasets to unravel metabolic pathways and identify biomarkers associated with physiological states or diseases [76]. Bioinformatics emerged as a field at the intersection of biology and computer science, aiming to analyze and interpret biological data using computational tools and techniques [77]. The roots of bioinformatics can be traced back to the latter half of the 20th century with the development of early algorithms for sequence alignment and protein structure prediction. These pioneering efforts laid the foundation for the modern era of bioinformatics, characterized by the explosion of biological data generated by high-throughput technologies, which have become increasingly available since the turn of the 21st century and continue to evolve [78]. Artificial Intelligence (AI) is now integral to this evolution, providing a variety of approaches for pattern recognition, predictive modeling, and data integration. These technologies enhance the ability to derive meaningful insights from complex datasets, further advancing the field of metabolomics.

I.3.1 The role of bioinformatics in metabolomics data analysis

Bioinformatics methods encompass a wide range of computational techniques and analytical approaches used to extract meaningful insights from biological data including those metabolomic-derived. To obtain reliable results, comprehensive software is needed to allow for precise analysis of raw spectra [79]. According to the sources [80-81] where an extensive collection of current metabolomics tools and databases were presented, among the most commonly used programming languages in metabolomics, Python, R, Ruby, and Matlab clearly dominate with BASH scripting being essential for automating data processing pipelines and integrating various bioinformatics tools. Additionally, authors mentioned that scripting languages are increasingly combined with "low-level" languages such as C, C++, or Rust. In addition to software and programmed methods, bioinformatics has embraced the adoption of automated workflows known as pipelines. These pipelines have gained popularity in recent years due to their ability to streamline and standardize complex data analysis tasks in bioinformatics. By automating the sequential execution of multiple software tools and algorithms, pipelines facilitate the efficient processing, analysis, and interpretation of large-scale biological datasets. They offer a systematic approach to reproducibly perform bioinformatics analyses, saving time and minimizing errors. Moreover, pipelines often include built-in documentation and version control mechanisms, enhancing transparency and reproducibility in bioinformatics research [82]. Luigi [83], SnakeMake [84], and Nextflow [85] are examples of workflow management systems used in bioinformatics. Luigi is a Python-based workflow management system developed by Spotify, designed to handle complex pipelines in a simple, scalable, and maintainable way. SnakeMake, another Python-based workflow management system, focuses on enabling reproducible and scalable data analysis workflows. Nextflow, on the other hand, is a data-driven computational workflow framework written in Groovy that enables scalable and reproducible scientific workflows using software containers such as Docker [86] or Singularity [87]. Supporting containerization, enabling users to encapsulate software dependencies and ensure reproducibility across different computing environments. It also features built-in support for task parallelization, data parallelization, and fault tolerance, making it suitable for large-scale data analysis on local workstations, high-performance computing (HPC) clusters, and cloud platforms. Additionally, Nextflow provides extensive logging, monitoring, and visualization capabilities for workflow management and optimization. Overall, while all three workflow management systems serve the purpose of orchestrating and automating bioinformatics pipelines, their differences in syntax, scalability, and community support make them suitable for different use cases and preferences within the bioinformatics community [88-89]. Despite the evident potential, there remains a minority of publicly available metabolomics software that supports NMR spectra analysis, while focusing more on various variants of mass spectrometry equipment (e.g. GC/ LC-MS).

Moreover, authors develop their tools mainly focusing on individual steps in metabolomic analyses, indicating a clear lack of comprehensive and fully automated pipelines.

I.3.2 Metabolomics data interpretation: preprocessing

Metabolomics data, from the initial phase of data acquisition to the statistical analysis, undergoes a series of processing steps. Each of these steps ensure the quality of the data and facilitate the accurate interpretation of the results. During the data acquisition stage, raw data is gathered from various sources, such as mass spectrometry or nuclear magnetic resonance spectroscopy. This raw data often contains noise and artifacts that might mask meaningful signals. Thus, preprocessing techniques are employed to clean and process the data, including baseline correction, spectra phasing, peak alignment, and normalization [90-91]. Baseline correction involves removing or adjusting the underlying signal, known as the baseline, from the raw data. The baseline represents the noise or background signal present in the data, such as solvent effects, chemical impurities, or instrumental artifacts which can obscure the true metabolite peaks of interest. Baseline correction methods aim to accurately estimate and subtract this baseline to improve the accuracy of peak detection and quantification. Common techniques include polynomial fitting, spline interpolation, and wavelet-based methods [92-93]. Peak alignment in NMR metabolomics involves aligning peaks across multiple spectra to correct for variations in chemical shift caused by differences in experimental conditions, sample preparation, or instrument settings. Since chemical shift values are sensitive to factors such as pH, temperature, and solvent composition, peak alignment ensures that corresponding peaks are aligned in all spectra, enabling accurate comparison and interpretation of metabolite profiles. Alignment algorithms in NMR data typically rely on the identification of reference peaks or spectral regions shared across samples, followed by alignment based on chemical shift values. NMR spectra may exhibit variations in signal intensity due to differences in sample concentration or total spectral area [94]. Normalization methods in NMR metabolomics aim to remove these systematic variations by scaling the spectral intensities across samples to a common reference. Common normalization approaches include total spectral area normalization, probabilistic quotient normalization (PQN), or normalization based on internal standards or reference peaks. By normalizing the spectra, impact of technical variability is minimized and the reliability of comparative analyses and biomarker discovery are enhanced [95]. Additionally, scaling techniques such as Pareto scaling, which reduces the influence of high-intensity metabolites while preserving variability, and unit variance scaling, which standardizes the data by dividing each variable by its standard deviation, are commonly used [96]. These methods are essential in metabolomics data analysis, as they ensure that differences observed are due to biological factors rather than technical artifacts, thereby improving the accuracy of downstream statistical and machine learning models.

Overall, preprocessing steps aim to enhance the quality and reliability of the data by minimizing technical variations and noise introduced during data acquisition. Following data preprocessing, the processed data undergoes statistical analysis to extract meaningful insights and identify significant metabolites or metabolic patterns. Statistical methods such as multivariate analysis, hypothesis testing, and machine learning algorithms are applied to unravel complex relationships within the data and discern relevant biological information [97].

I.3.3 Metabolomics data interpretation: univariate and multivariate perspectives

In spite of significant advancements in metabolomics and bioinformatics, challenges persist in maximizing the full potential of metabolomic data. One such challenge is the complexity of data interpretation, integration across multiple omics layers [98] and evaluating complex spectra [99]. Both univariate and multivariate analysis methods are commonly employed to extract meaningful insights from this data, aiming to reveal biological insights relevant to the investigated issue [100]. Univariate methods in metabolomics analysis offer a straightforward approach to investigating individual metabolites' behavior and their association with various experimental or clinical factors. These methods, such as t-tests, ANOVA, fold change analysis, and correlation analysis, focus on examining one variable at a time, making them suitable for initial exploration or hypothesis testing in metabolomics studies. For instance, t-tests and ANOVA are commonly employed to compare the means of metabolite levels between different groups, aiding in the identification of significant differences associated with factors like disease state or treatment response. Non-parametric alternatives like the Wilcoxon rank-sum test and Kruskal-Wallis test accommodate data that do not adhere to normality assumptions, ensuring robust statistical analysis [101-102]. Fold change analysis highlights metabolites that exhibit substantial alterations in abundance between experimental conditions, guiding towards potential biomarkers or biologically relevant molecules. Additionally, correlation analysis explores relationships between metabolite levels and clinical parameters, shedding light on metabolic pathways or physiological processes influenced by these associations. Univariate methods provide valuable insights into individual metabolite behavior and their relevance to biological phenomena in metabolomics research. However, they are often supplemented with multivariate approaches to capture the complex interactions and patterns present in high-dimensional metabolomics data comprehensively [103].

Concerning NMR-based metabolomics, the data typically manifest as spectra, which are then segmented into regions of predetermined width. Generally, these methodologies yield voluminous and complicated datasets. Initially, the number of observations (n) in metabolomics experiments is often significantly lower than the number of peaks (or variables) (p) in a spectrum $n \ll p$. In such instances, utilizing standard parametric statistical methods like regression is challenging due to insufficient data for parameter estimation. Furthermore, numerous metabolites may not

be correlated with the trait under examination, contributing irrelevant variation that complicates comprehensive data analysis [104]. Given these challenges, there exists a genuine necessity for multivariate dimension-reducing techniques capable of accommodating data complexities and unveiling any latent relationships. Multivariate analysis techniques, such as principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA), are commonly used for dimensionality reduction and visualization of complex metabolomics datasets. These methods facilitate the identification of patterns and clusters that differentiate between experimental groups or conditions [105]. Principal components analysis, an unsupervised multivariate statistical method, strategically employs orthogonal transformations to convert potentially correlated variables into linearly uncorrelated variables known as principal components. PCA is a prevalent technique for metabolomic data analysis, offering a straightforward non-parametric approach to projecting NMR or MS spectra into lower-dimensional space, revealing inherent data structure, and providing a compact representation of the original data. Despite its widespread adoption, PCA suffers from several limitations. Notably, it lacks an associated probabilistic model, making assessment of its fit to the data challenging and constraining its application scope. Additionally, PCA may fail to unveil underlying subject groups within the data, potentially presenting a distorted view of the data structure. Other drawbacks include its inability to appropriately handle missing data [106]. In metabolomics studies, alongside metabolomic data, other phenotypic data such as age, gender, BMI, or disease status are often generated. Incorporating these covariates into multivariate models is highly beneficial for comprehensive data analysis. Additionally, there is a struggle with data unification, as contributors often provide different answers and do not always read the questions carefully. Partial Least Squares Discriminant Analysis is a supervised multivariate statistical technique commonly used in chemometrics and various fields of data analysis, including omics data analysis. It is particularly useful when dealing with high-dimensional datasets where the number of variables (features) exceeds the number of samples or observations $p \gg n$, and when there is a need to classify or predict sample classes from phenotypic data based on the given features. PLS-DA can be viewed as a "supervised" counterpart to PCA, integrating dimensionality reduction with group information. Consequently, it not only reduces dimensionality but also enables feature selection and classification [107-108].

From a data analysis perspective, in order to address tasks such as feature selection and classification, metabolomics is beginning to see benefits from the adoption of machine learning (ML) and deep learning (DL) methods [109-111]. In feature selection, ML algorithms are used to identify a subset of relevant features (metabolites) from the large pool of measured variables. These algorithms assess the importance of each feature based on criteria such as predictive power or importance scores, enabling the selection of the most informative features for further analysis. Common ML techniques for feature selection in metabolomics include Recursive Feature Elimination (RFE) and LASSO (Least Absolute Shrinkage and Selection Operator) [112-113]. In classification, ML algorithms are employed to categorize samples into different groups based on their metabolic profiles. These

algorithms learn patterns from labeled training data and then use these patterns to predict the class labels of unseen samples. Popular ML classification algorithms in metabolomics include Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN) and Neural Networks (NN). These algorithms enable accurate classification of samples into predefined categories, such as healthy versus diseased or treated versus control [114]. Cross-validation techniques, such as k-fold cross-validation or leave-one-out cross-validation, are commonly used to evaluate the performance of ML models in feature selection and classification tasks [115]. These techniques partition the data into training and validation sets multiple times, allowing for robust assessment of model performance and generalization to unseen data. ML algorithms thanks to their ability to handle high-dimensional data can uncover hidden relationships between metabolites and disease phenotypes, predict disease outcomes, and identify novel biomarkers [116]. By leveraging machine learning, the limitations of traditional statistical methods can be overcome, uncovering novel insights into disease mechanisms.

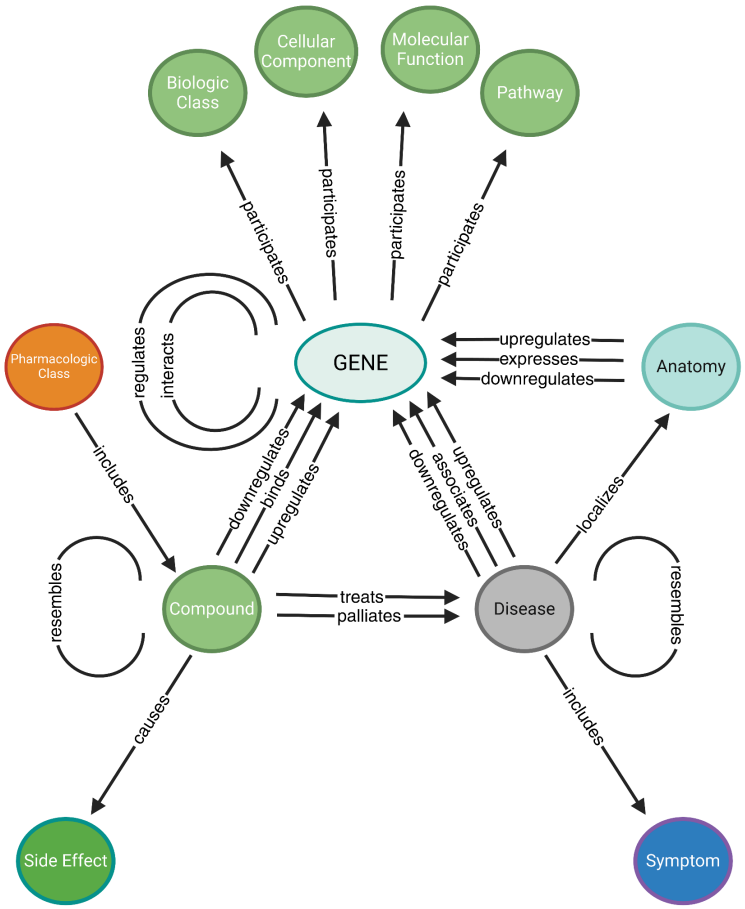
I.3.4 Metabolomics data interpretation: insights into biological processes

Pathway analysis is a computational method used in bioinformatics to gain insights into biological processes by analyzing sets of genes, proteins, or metabolites within the context of known biological pathways [117]. Pathway analysis begins with the identification of biological pathways relevant to the experimental context. This can be achieved through various pathway databases, literature mining, or computational methods that compile information on known pathways from published studies and curated databases [118-120]. It typically takes as input a list of genes, proteins, or metabolites that are differentially expressed or otherwise associated with a particular biological condition, such as a disease state or experimental treatment. These input molecules can be identified through high-throughput omics technologies such as microarrays, RNA sequencing, mass spectrometry, or metabolomics. The input list of molecules is then analyzed to determine whether they are significantly enriched in any particular biological pathways compared to what would be expected by chance. This is typically done using statistical methods such as Fisher's exact test, hypergeometric test, or gene set enrichment analysis (GSEA) [121]. Enrichment analysis helps identify pathways that are biologically relevant to the experimental condition under investigation. The results are often visualized using pathway diagrams or networks, where input molecules are mapped onto their respective pathways. This visualization allows researchers to interpret the biological significance of the findings and identify key pathways or biological processes that may be dysregulated in the experimental context [122-123].

Another captivating method in pathway analysis involves the utilization of knowledge graphs. These graphs serve as repositories of information, utilizing a graph-structured data model or topology to manage and illustrate data [124]. They aggregate data about metabolites from various sources,

including databases, literature, and experimental studies, providing extensive annotations that encompass chemical structures, biochemical properties, metabolic reactions, and biological functions. Through the integration of this information into a unified framework, knowledge graphs aid to annotate metabolites identified in metabolomics experiments and contextualize their roles within cellular metabolism [125-126]. Additionally, knowledge graphs can be integrated with other omics data, such as transcriptomics and proteomics, to construct multi-omics knowledge networks (see **Figure 7**). This integration enables the linking of metabolites to genes, proteins, and biological pathways, facilitating exploration of the relationships between metabolic changes and alterations in gene expression or protein abundance. Such an integrative approach enhances understanding of metabolic regulation and coordination across different molecular levels in complex biological systems [127].

Figure 7. Diagram of an exemplary multi-omics knowledge graph.



Source: Based on [128], created in Biorender.com.

Furthermore, knowledge graphs enable the inference of metabolic relationships and pathway associations using existing biological knowledge. By analyzing connectivity between metabolites, enzymes, and pathways within the graph, novel metabolic connections can be uncovered, and

potential metabolic pathways or pathway crosstalk can be predicted [129]. This capability enables generation of hypotheses about metabolic interactions underlying observed changes in metabolite levels or metabolic phenotypes in metabolomics experiments. Knowledge graphs encode detailed representations of metabolic pathways, illustrating the sequential steps of biochemical reactions involved in metabolite transformations. These pathway representations include information about enzyme-catalyzed reactions, substrate-product relationships, and regulatory interactions. Visualizing metabolic pathways as interconnected nodes and edges within the graph, can help to explore the flow of metabolites through different biochemical pathways and identify potential metabolic flux alterations associated with physiological conditions or perturbations.

I.4 Databases and state-of-the-art tools

In recent years, metabolomic research has significantly advanced, leading to an increasing number of records deposited in databases. This reflects the unified actions to document and share metabolomic data, enhancing accessibility and analysis worldwide. Metabolomic databases serve as repositories for storing extensive datasets generated from diverse experiments, providing curated information for querying, analyzing, and interpreting metabolomic data [130-131]. The FAIR principles, introduced in 2016 with a focus on making data Findable, Accessible, Interoperable, and Reusable, were initially aimed at fostering the reuse of scientific data and enhancing data management. Over time, these principles were expanded to encompass research software and workflows, aiming for greater uniformity and reusability across scientific endeavors [132]. FAIR initiative alongside continuously updated and expanded databases contribute significantly in advancing our understanding of metabolic pathways, biomarkers, and physiological processes [133]. In the following sections, state-of-the-art databases and bioinformatic tools are discussed. For easier reference, these materials are compiled in **Table 1**.

Table 1: Metabolomic state-of-the-art databases and bioinformatic tools.

Type	Name	Source [Reference]
Database	Biological Magnetic Resonance Data Bank	https://bmrb.io/ [134]
	CyanoCyc	https://cyanocyc.org/ [145]
	HumanCyc	https://humancyc.org/ [144]
	Human Metabolome Database	https://hmdb.ca/ [152]
	Kyoto Encyclopedia of Genes and Genomes	https://www.genome.jp/kegg/ [143]
	MetaboLights	https://www.ebi.ac.uk/metabolights/ [157]
	MetaCyc	https://metacyc.org/ [142]
	Protein Data Bank	https://www.rcsb.org/ [135]
	Reactome	https://reactome.org/ [148]
	UniPathway	https://www.uniprot.org/database/DB-0170 [147]
Software	Bruker TopSpin	https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html [166]
	Chenomx NMR Suite	https://www.chenomx.com/ [168]
	MetaboAnalyst	https://www.metaboanalyst.ca/ [175]
	Mnova	https://mnova.pl/ [167]
	NMRProcFlow	https://www.nmrprocflow.org/ [170]
R package	ASICS	https://bioconductor.org/packages/release/bioc/html/ASICS.html [173]
	metaboanalystR	https://github.com/xia-lab/MetaboAnalystR [177]
	PepsNMR	https://www.bioconductor.org/packages/release/bioc/html/PepsNMR.html [172]

Source: Based on literature research, own elaboration.

I.4.1 An overview of metabolomics data repositories

For the past three decades, the Biological Magnetic Resonance Data Bank (BMRB) has stood as the principal repository for spectral and derived data from nuclear magnetic resonance spectroscopy of biological systems. BMRB was established in 1988, distinguishing itself from other biophysical data banks, by housing primary time-domain data acquired by NMR spectrometers,

processed spectra, spectral peak attributes, assigned spectral peak chemical shifts, and derived data such as relaxation parameters [134]. Database also includes atomic coordinates for certain smaller molecules. BMRB has developed advanced technologies for annotating and processing chemical shift data archived within its repository, as well as the chemical shift and constraint data supporting NMR-based structures in the Protein Data Bank (PDB) – open-access digital repository housing three-dimensional structural data of biological macromolecules [135]. BMRB, adhering to FAIR data principles, serves as a repository for NMR experimental and derived data on biologically relevant molecular systems. It encompasses six primary data repositories: quantitative NMR spectral parameters and derived data, time-domain spectral data, atomic coordinates for small molecules, NMR constraints, CS-Rosetta structures, and a growing database of 1D and 2D NMR spectra for biological molecules. Validation reports for chemical shift entries and MolProbity reports for PDB entries are available on the BMRB website. Moreover, BMRB is affiliated with the Center for NMR Data Processing and Analysis, offering the NMRbox platform, a cloud-based computing platform facilitating access to existing NMR software tools and computational resources. NMRbox aims to enhance NMR data reproducibility, streamline depositions to BMRB and other public databases, and develop new data analysis tools. Various software services are provided for querying the archive, performing data visualizations, file format conversions, data validation, and structure calculations [136]. Data in BMRB are linked to literature citations and other public databases through Basic Local Alignment Search Tool (BLAST) searches algorithm [137]. BMRB acquires data through depositor submissions via multiple deposition systems, including OneDep, ADIT-NMR, and SMSDep [138-139]. The NMR-STAR data format is utilized, supported by tools for editing and handling NMR-STAR files. The NMR-STAR ontology facilitates data reusability by providing comprehensive information on archived experimental data [140].

The landscape of biomolecular NMR is continually evolving, with emerging NMR techniques presenting opportunities for studying diverse biological molecular systems, from larger proteins to nucleic acids, molecular machines, and membrane-bound biopolymers. BMRB supports metabolomics and natural products research by offering a library of 1D and 2D NMR spectra of pure compounds, including metabolites, natural products, drugs, and screening compounds. It adopts ALATIS compound and atom identifiers, based solely on the 3D structure of the compound and the InChI convention, alongside providing spin matrices in the GISSMO convention for an increasing number of small molecules [141]. Overall, its commitment to facilitating access to high-quality NMR resources and support services make it an indispensable resource for metabolome investigation worldwide.

MetaCyc and Kyoto Encyclopedia of Genes and Genomes (KEGG) are expansive metabolic pathway database projects with over twenty years of development history [142-143]. They both offer reference pathways utilized in predicting organismal metabolic pathways from annotated genomes. MetaCyc is a comprehensive database of metabolic pathways and enzymes, serving as a valuable

resource in bioinformatic, molecular biology, and related fields. MetaCyc contains curated information on biochemical reactions, metabolic pathways, enzymes, and metabolites across all domains of life. MetaCyc's significance extends to its utilization in predicting pathways within the BioCyc database collection and other Pathway/Genome Databases globally such as HumanCyc and CyanoCyc [144-145]. Meanwhile, KEGG consists of a reference pathway DB and projections of these pathways onto organisms with sequenced genomes, widely adopted in research. Initially, KEGG focused on genes and genomes, providing extensive resources for understanding biological pathways, functions, and interactions. Over time, KEGG expanded its scope to include metabolomics data, incorporating metabolic pathway maps, compound databases, and related information. The integration of metabolomics-related content into KEGG occurred gradually, with continuous updates and improvements to reflect advances in metabolomics research. In a study [146], a comprehensive comparison of the contents of the KEGG and MetaCyc pathway DBs was conducted in order to recognize the multidimensional nature of pathway databases, which encompass various types of data. Examination encompassed metabolites, reactions, and pathways, excluding other aspects like orthology data or genome-based pathway predictions. Notably, both databases boast substantial metabolic reactions and pathways compared to other similar resources like UniPathway and Reactome [147-148]. KEGG includes significantly more compounds compared to MetaCyc, while MetaCyc encompasses a substantially greater number of reactions and pathways than KEGG. However, the number of reactions within pathways is quite similar between the two databases.

MetaCyc and KEGG are frequently utilized for various purposes in metabolomics studies. Firstly, they are employed for pathway analysis, where the extensive metabolic pathway information available in these databases is utilized to map experimental metabolite data onto known pathways. This enables the identification of affected pathways in diverse biological contexts or experimental conditions. Moreover, databases serve as invaluable tools for metabolite annotation. Comparing experimental metabolomic data with entries in these databases allows to annotate detected metabolites based on their known chemical structures and metabolic pathways. Another common application in metabolomics research is pathway enrichment analysis [149]. Comparison lists of differentially expressed metabolites are created to known metabolic pathways in the databases, identifying overrepresented pathways and gaining insights into the biological significance of observed metabolomic changes. Additionally, MetaCyc and KEGG enable metabolic network reconstruction [150]. These databases provide the essential information needed to construct comprehensive metabolic networks, representing interconnected metabolic pathways within organisms. These networks are essential for systems-level analysis, modeling, and simulation of metabolic processes [151]. Furthermore, MetaCyc and KEGG are utilized for comparative metabolomics studies. These databases enable the comparison of metabolomic data across different organisms or experimental conditions, allowing the identification of conserved or divergent metabolic features across biological systems.

The Human Metabolome Database (HMDB) stands as the largest and most comprehensive metabolomic database dedicated to human organisms [152]. Created at the University of Alberta in Canada in 2007, HMDB offers an extensive array of data concerning the structures, chemical properties, concentrations in various biological fluids and tissues, and metabolic pathways associated with each metabolite. HMDB's information is meticulously curated from a diverse range of sources, including scientific literature, high-throughput experimental studies, and computational analyses, ensuring the reliability and accuracy of its content. Through its user-friendly interface, HMDB facilitates browsing, searching, and analysis of metabolite data, enhancing exploration of metabolic pathways, pinpointing potential biomarkers, and unraveling the connections between metabolites and human health or disease states. Unlike general metabolism databases or spectral repositories, such as BMRB, KEGG, MetaCyc or Reactome, HMDB serves as a dynamic and vividly illustrated online encyclopedia, providing in-depth coverage of human metabolites and metabolism.

Since its inception, HMDB has undergone significant evolution and enhancement to meet the evolving needs of the metabolomics community. Over the years, HMDB has expanded dramatically, from its initial release containing data on 2,180 human metabolites to its latest version, HMDB 5.0, which boasts more than 200,000 annotated metabolite entries, alongside over 1,5 mln unannotated derivatized metabolite entries for GC–MS [153-155]. This latest iteration introduces numerous improvements, including enhanced metabolite descriptions, a new Chemical Functional Ontology (ChemFOnt), and advanced visualization tools. This database is utilized in various ways to advance understanding of human metabolism and its implications for health and disease. One significant application of HMDB is in biomarker discovery. The database is used to identify potential biomarkers by comparing metabolomic data from biological samples to HMDB's extensive library of metabolites. By identifying molecules that are indicative of specific biological processes or disease states, improvements can be made in diagnostic and prognostic approaches, providing insights into disease mechanisms. Another key use of HMDB is in metabolite annotation. The database provides detailed information on metabolite structures, properties, and metabolic pathways, allowing to annotate metabolites detected in metabolomics experiments. This enables the verification of metabolite identities and provides significant insights into their biological functions and roles in metabolic pathways. Where identification of perturbed metabolic pathways associated with specific conditions or treatments, provides insights into underlying biological mechanisms. In pharmacology and drug discovery, HMDB is used for the examination of how drugs are metabolized, the identification of metabolites formed during drug metabolism, and the assessment of their pharmacological activities and potential toxicity. Additionally, HMDB data is integrated into computational models of biological systems for systems biology modeling. By incorporating HMDB's metabolite information, predictive models of cellular metabolism can be developed, metabolic fluxes investigated, and the effects of genetic or environmental perturbations on metabolic pathways explored [156].

In essence, HMDB serves as an invaluable tool in advancing our comprehension of human metabolism and its multifaceted implications for physiology, pathology, and personalized medical interventions. Its latest version represents a significant milestone in metabolite database development, with substantial enhancements in data quality, accessibility, and analytical capabilities. A comprehensive collection of selected metabolites included in the database enables innovative research and contributes to advancements in human health and biology.

MetaboLights serves as the open-access repository adapted for metabolomics investigations, offering a collection of raw experimental data and associated metadata [157]. Operated by a leading open-access data provider in molecular biology, MetaboLights serves as an resource for probing biological functioning and understanding systemic perturbations induced by factors like diseases, dietary habits, and environmental influences. The MetaboLights database was founded in 2012 by the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI). Since its creation, it has been widely utilized in the fields of metabolomics and bioinformatics. Its usage has steadily grown over the years as metabolomics research has gained momentum, with relying on MetaboLights as a primary resource for storing, sharing, and analyzing metabolomics data. The database continues to play a crucial role in advancing metabolomics research by providing a centralized platform for data management and collaboration within the scientific community [158]. MetaboLights operates on two primary fronts: as a repository for the global metabolomics community to exchange findings, data, and methodologies across various metabolomic studies, and as a reference layer containing meticulously curated information about metabolites. Rather than seeking to supplant specialized resources, MetaboLights is customized to complement existing databases and collaborates extensively with them to ensure seamless data exchange and address gaps in global knowledge. Close partnerships with key stakeholders, including the Metabolomics Society, Metabomeeting, and the Metabolomics Standards Initiative (MSI), underscore commitment to capitalizing on prior expertise in the field. MetaboLights actively pursues formal data-sharing agreements with major resources such as HMDB and Chemical Entities of Biological Interest (ChEBI), hence by referencing identified metabolites from external databases, it avoids redundant information and instead provides users with a unified, metabolite-centric perspective. As of September 2023, the database has attracted users from nearly 100 countries across the globe, with a significant increase in registered studies — from 1432 at the beginning 2020 to almost 8 times more in the third quartile of 2023. These studies are categorized into different stages, including public, in review, in curation, and submitted, indicating the platform's pivotal role in data storage and processing throughout the research lifecycle. The data hosted on MetaboLights have expanded substantially, now encompassing over 270,000 samples, 2,700 assays, 439,000 data files, and 1.6 million metabolites/ unknowns/ features [159]. This wealth of data covers a wide range of organisms, with a notable focus on human samples, followed by model organisms like mice and *Arabidopsis thaliana*. LC-MS dominates as the preferred analytical technique, highlighting the prevalence of untargeted studies and unassigned features. MetaboLights

reflects the evolving landscape of metabolomics research, with a growing emphasis on complex data types, advanced acquisition techniques, and multi-omics integration. Studies incorporate quality control measures and cutting-edge technologies, such as MS2 acquisition and data-independent acquisition. Efforts to streamline multi-omics integration are underway through collaborations with consortia like HoloFood, linking metabolomics data with genomic and metagenomics data from other EMBL-EBI resources [160]. These developments underscore the importance of MetaboLights as a central hub for metabolomics data management and collaboration within the scientific community.

I.4.2 State-of-the-art bioinformatics tools for metabolomics

Metabolomics has undergone a renaissance, largely driven by the integration of multi-omics analyses, aimed at uncovering insights beyond what individual omics studies can provide alone [161]. The extensive research efforts in metabolomics have led to a significant rise in data deposited in databases [162]. With advancements in metabolite detection accuracy and the ability to integrate multiple omics datasets, the field is rapidly evolving. As access to open data continues to increase and there is a growing demand for understanding the connections between metabolites and organism phenotypes, there is a crucial need to develop tools that allow transition from raw data to the biological interpretation of analyzed samples [163]. Bioinformatics tools for metabolomics analyses encompass a spectrum ranging from straightforward command-line utilities to more complex programs featuring graphical interfaces, as well as standalone web services [164-165]. These tools are developed by bioinformatics companies or public institutions, offered under open-source licenses or commercial arrangements (refer to **Table 1** for state-of-the-art databases and bioinformatic tools).

Bruker TopSpin is a software package designed for the processing, analysis, and visualization of nuclear magnetic resonance data. Developed by Bruker Biospin, it is specifically adapted for use with Bruker NMR spectrometers, providing users with a powerful suite of tools to extract valuable information from NMR spectra [166]. Bruker TopSpin has a user-friendly interface, offering an intuitive platform for engaging with NMR data in a straightforward manner. The software provides a wide range of functionalities for spectral processing, including baseline correction, phasing, and referencing, allowing enhancement of the quality of spectra for further analysis. In addition to basic spectral processing, Bruker TopSpin offers advanced analysis tools for peak picking, integration, and spectral fitting. These tools enable identifying and quantifying peaks corresponding to different metabolites or molecular components present in the sample. Furthermore, the software supports various 1D and 2D NMR experiments, facilitating the characterization of complex molecular structures and interactions. Software also includes features for data visualization and interpretation, allowing users to visualize their NMR data in various formats, such as stacked spectra, contour plots, or 2D spectra. The TopSpin supports customizable plotting options and interactive tools for exploring

spectral features, aiding in the identification of metabolites and the elucidation of biochemical pathways. One noteworthy feature is the availability of automation tools and customizable workflows, which efficiently streamline data processing and analysis tasks. Users have the flexibility to design custom processing scripts or employ built-in automation tools, ensuring the execution of routine tasks with reproducibility and ease.

Overall, Bruker TopSpin stands out as a top choice, offering essential functionalities to effortlessly process, analyze, and interpret NMR data with precision. The latest version of TopSpin is compatible with various operating systems, including Windows 10 and 11, Linux-based systems such as AlmaLinux 9.2 or CentOS 7, and macOS.

Another software broadly used in metabolomics analysis, Mnova (short for Mestrelab NMR and MS Analysis) is a package designed for the processing, analysis, and visualization of NMR and MS data [167]. Developed by Mestrelab Research, Mnova offers a wide range of tools and functionalities with an intuitive environment for users to interact with their metabolomics-derived data. The software provides a suite of tools for spectral processing, including phasing, baseline correction, peak picking, integration, and spectral alignment. In addition to basic spectral processing, Mnova offers advanced analysis capabilities, such as deconvolution, structure elucidation, quantification, and metabolite identification. The software supports a wide range of 1D and 2D NMR experiments, as well as various MS techniques, allowing users to analyze complex datasets and investigate molecular structures and interactions. Mnova also includes visualization tools for data exploration and interpretation. Users can visualize their NMR and MS data in different formats, such as stacked spectra, contour plots, and 2D spectra. The software supports customizable plotting options and interactive tools for exploring spectral features, facilitating the identification of metabolites, compounds, and biomarkers. Furthermore, Mnova offers integration with other software tools and databases, allowing users to import and export data in standard formats and access external resources for compound identification and spectral database searching. Like Bruker TopSpin, Mnova is a versatile software tool that goes beyond just NMR analysis. In addition to NMR spectra, Mnova is also capable of analyzing MS data, offering a comprehensive solution across both domains.

Chenomx NMR Suite, a specialized software designed for metabolomics research, offers a comprehensive platform for the analysis of NMR spectra [168]. At its core, Chenomx features an extensive metabolite library, providing a vast repository of reference spectra for accurate metabolite identification [169]. This library is continually updated and expanded to encompass a wide range of metabolites, ensuring comprehensive coverage across various biological samples and experimental conditions. In addition to quantification, Chenomx presents tools for quality control and validation of spectral data, empowering to assess the reliability and integrity of results. This ensures the accuracy and reproducibility of metabolite measurements, which are crucial in metabolomics experiments. Furthermore, Chenomx affords a variety of statistical analysis and metabolic pathway exploration, enabling them to uncover correlations, trends, and biological insights

within analyzed data. By linking identified metabolites to known metabolic pathways, the software helps elucidate the underlying biological mechanisms driving observed metabolic changes. Beyond its technical capabilities, Chenomx is supported through documentation, tutorials, and direct engagement of the developing team.

Bruker TopSpin, Mnova, and Chenomx play distinct yet complementary roles in NMR spectroscopy and metabolomics research, offering commercial solutions for data processing, analysis, and interpretation across various applications and experimental contexts. Chenomx complements both Bruker TopSpin and Mnova by providing specialized tools for metabolomics analysis. Users utilize Chenomx to precisely identify and quantify metabolites in NMR spectra, supplementing structural analysis from TopSpin and multi-omics integration from Mnova. The coexistence of commercial and open-source tools in metabolomics reflects diverse needs and preferences, alongside the evolving nature of the field. Commercial tools offer advanced functionalities, expert support, and compatibility with proprietary technologies, catering to users ranging from academic researchers to industry practitioners. Conversely, open-source tools promote accessibility, collaboration, and transparency, facilitating broader participation and knowledge exchange among users. NMRProcFlow exemplifies such open-source software, streamlining the processing and analysis of NMR spectroscopy data. It provides a workflow for various preprocessing and analytical tasks on NMR datasets, promoting reproducibility by granting users access to underlying algorithms and methodologies [170]. Key features include preprocessing functions like Fourier transformation, phase correction, baseline correction, and spectral alignment, enhancing signal-to-noise ratio and artifact correction [171]. The software enables spectral calibration by referencing chemical shifts to known standards, ensuring accurate reporting and facilitating comparison with databases. NMRProcFlow includes algorithms for automated peak detection and integration, quantifying peak intensity and converting to concentration values. Users can explore data, visualize spectra, compare samples, and identify patterns indicative of biological or chemical phenomena. Facilitating interpretation, NMRProcFlow aids in drawing conclusions about sample composition, structure, or properties. This involves comparing experimental spectra with references and conducting further analyses for validation. NMRProcFlow is designed to be compatible with multiple operating systems including Windows, macOS and Linux. NMRProcFlow emphasizes automation and reproducibility through customizable processing workflows implemented using Python scripts. While it does not provide a traditional graphical interface, users can create and execute processing pipelines using command-line interfaces or scripting environments.

PepsNMR and *ASICS* are additional examples of software that utilize command-line interfaces or custom scripts for operation [172-173]. *PepsNMR*, an innovative R package, specializes in robust data pre-processing for metabolomic studies, particularly in ^{1}H NMR spectra analysis. By offering transparency, automation, and a wide array of processing options, *PepsNMR* fills the gap left by proprietary software from instrument manufacturers such as Bruker TopSpin. It covers crucial

steps in analysis like solvent signal suppression, baseline correction, and bucketing, enhancing methodological transparency and efficiency. Contrary to the previously mentioned commercial software featuring graphical interfaces, PepsNMR operates through command-line interfaces or customized scripts for data analysis, making it primarily geared towards users with programming skill set. Nevertheless, its adaptability to automation renders it suitable for incorporation into automated workflows or pipelines. For instance, *ASICS* (Automatic Statistical Identification in Complex Spectra) provides a solution for automatic metabolite identification and quantification in NMR data analysis workflows. Based partially on the method proposed in *PepsNMR*, R package *ASICS* underwent refinement and enhancement to fine-tune parameters, improve results, and reduce computational costs. It integrates preprocessing of spectra, post-quantification statistical analyses, and comprehensive documentation, offering flexibility, ease of installation, and integration into existing workflows. *ASICS* introduced a collection of pure metabolite spectra that serves as a benchmark for identifying and measuring metabolite concentrations within the spectra of interest, which often contain complex mixtures. Moreover, the package offers functionalities to include or exclude certain spectra from the reference library or to utilize an alternative library provided by the user.

Together, *PepsNMR* and *ASICS* address the critical need for robust and automated data processing and analysis in metabolomics. They enhance the efficiency, reproducibility, and reliability of metabolomic studies, empowering researchers to extract meaningful insights from complex NMR spectra data. Noteworthy is the validation of both platforms' efficacy through metabolomic case studies, demonstrating superior information retrieval and predictive capability [174]. Nevertheless, despite the presence of numerous tools for NMR data analysis, the field of metabolomics continues to face a shortage of comprehensive tools that integrate bioanalytical, bioinformatics, and machine learning methodologies to offer a comprehensive understanding of metabolic dysregulation in diseases. The large volume of data generated by metabolomic studies necessitates advanced computational and statistical methods for processing, analyzing, and integrating multi-omics datasets. This work’s endeavor aims to address the gap in the availability of a comprehensive and standardized pipeline for translating raw ^{1}H 1D NMR data into biological insights, while providing access to the results of each individual analysis step.

An example worth showcasing is MetaboAnalyst – a widely utilized web-based platform for metabolomics analysis that offers a user-friendly interface accessible via web browsers, simplifying data uploading and analysis for users without specialized computational expertise [175]. MetaboAnalyst features a wide range of analysis modules covering data preprocessing of both targeted and untargeted metabolomics studies utilizing LC–MS. Through interactive visualization tools like heatmaps, volcano plots, pathway diagrams, and enrichment maps, the platform streamlines data exploration and interpretation. While primarily designed for mass spectrometry data, it is also capable of handling statistical analysis, pathway analysis, and metabolite set enrichment analysis

(MSEA) for NMR-derived preprocessed metabolites. Furthermore, it integrates with various metabolite and pathway databases such as HMDB or KEGG, enriching analyses with annotation and pathway information. Users can construct customized analysis workflows by linking multiple modules, streamlining the process of conducting complex analyses. Version 6.0 of MetaboAnalyst introduces a novel MS2 data processing workflow, providing a comprehensive, web-based platform for analyzing mass spectrometry metabolomics data. This workflow encompasses everything from processing raw MS spectra to identifying compounds to conducting functional analysis [176].

In addition to web-based solutions, there exists the R package *metaboanalystR*, providing users with the capability to access MetaboAnalyst's features programmatically from within R scripts or environments [177]. This simplifies the automation and customization of analyses. Being an R package, *metaboanalystR* integrates with other R packages and tools, enabling the incorporation of metabolomics analysis into pre-existing R-based data analysis pipelines. This ensures flexibility in customization, enhances the reproducibility of analyses, and facilitates the sharing of workflows via open-access repositories. MetaboAnalyst and *MetaboAnalystR* offer an extensive and adaptable toolkit for metabolomics data analysis, accommodating both those who favor a web-based interface and those who prefer programmatic access within the R environment.

II. Methodology: Evaluation of NASQQ pipeline on ^{1}H 1D NMR spectra

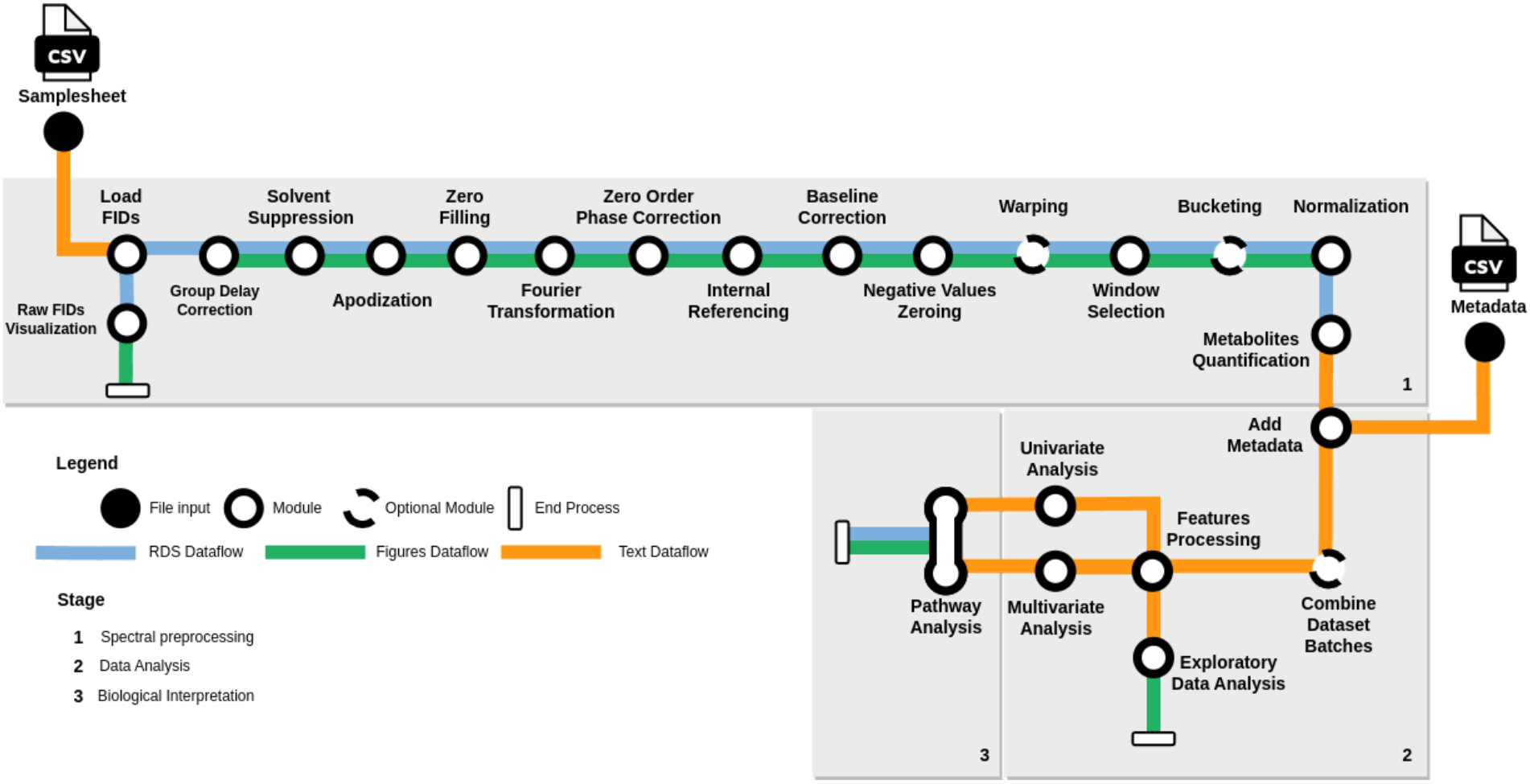
II.1 General overview of metabolomic workflow

Taking into account the considerations outlined in the introduction section, the doctoral thesis research was aimed to automate the analysis of proton nuclear magnetic resonance (^{1}H NMR) spectra for evaluating signals originating from metabolites found in urine and serum samples. This involved devising a systematic approach to streamline the interpretation of NMR data, particularly focusing on identification and quantification of metabolites. Through the automation of these processes, the goal was to enhance efficiency, reproducibility, and ultimately deepen comprehension of metabolomics in biological samples. Introducing NASQQ, an open-source Nextflow pipeline designed for the automated analysis of 1D ^{1}H NMR proton magnetic resonance spectra [178]. It aimed to combine existing methods to create a metabolomic workflow that encompasses data preprocessing of raw Bruker FIDs, feature extraction, and biological interpretation. By evaluating signals and conducting data analysis alongside exploration of biological pathways, the pipeline generates intuitive outcomes, facilitating the understanding of metabolomics in analyzed subjects without requiring extensive domain knowledge. This pipeline extends the capabilities of the existing R packages *PepsNMR* [172] and *ASICS* [173], representing state-of-the-art methods for converting raw signals from 1D ^{1}H NMR spectroscopy into a comprehensive set of metabolites. Furthermore, it introduces a novel approach for signal translation regularization, with the overarching goal of bridging the gap between raw spectral information and biological insights through the integration of machine learning methods.

The pipeline is constructed in a modular fashion [refer to **Figure 8**], with its primary features comprising:

- Automated and scalable workflow: The process of metabolomic analysis is automated by using Nextflow version 23.10.1 framework and designed to scale and adjust to available computing resources, reducing manual intervention and ensuring reproducibility across multiple datasets simultaneously.
- Comprehensive analysis: The pipeline encompasses spectral preprocessing, metabolite identification, data analysis, and pathway enrichment, offering a comprehensive perspective of the metabolomic data.
- Machine learning integration: NASQQ integrates machine learning techniques to connect raw spectral information with biological insights, enhancing the interpretability and utility of the analysis results.

Figure 8. Graphical depiction of the NASQQ pipeline**.**



Source: https://github.com/ardigen/nasqq/, available under a CC-BY-4.0 license.

II.1.1 NextFlow implementation and containerized computing environment

The pipeline utilizes Docker, an open-source platform tailored for constructing, deploying, and executing applications within containers. This methodology encapsulates each task or process within the computational workflow into Docker containers. These containers provide a lightweight and portable solution for packaging software and its dependencies, ensuring consistency across different computing environments. With Docker, pipeline's modules and scripts can be packaged once and run on any target system, including macOS or Windows. While it's feasible to run individual scripts in dedicated Docker containers, it's recommended to execute the entire pipeline for enhanced management and efficiency. For pipeline execution, two containers were devised. The first container, *r_utils:1.0.0* [https://github.com/ardigen/nasqq/blob/main/docker/R/Dockerfile], encompasses the "*rutils*" package, established and installed using scripts *pathway_analysis_utils.R* [https://github.com/ardigen/nasqq/blob/main/docker/R/r_utils/R/preprocessing_utils.R] and *preprocessing_utils.R* [https://github.com/ardigen/nasqq/blob/main/docker/R/r_utils/R/pathway_analysis_utils.R], which contain custom supplementary functions essential for the "*Spectral Preprocessing*" and "*Metabolites Identification*" stages. Each of these functions serves a specific purpose in the workflow of processing, analyzing, and visualizing NMR data, ensuring efficient and consistent handling of the datasets. The subsequent container, *python_utils:1.0.0* [https://github.com/ardigen/nasqq/blob/main/docker/Python/Dockerfile], hosts the script *ml_helpers.py* [https://github.com/ardigen/nasqq/blob/main/docker/Python/ml_helpers.py] utilized in the "*Data Analysis*" stage. Functions prepared in *ml_helpers.py* collectively enable a comprehensive framework for preparing data, evaluating models, and visualizing results in a machine learning pipeline. Following clean code programmatic principles, all software dependencies are addressed within Docker containers (see **Table 2** for detailed versions of utilized packages and libraries), leveraging base images *rocker/r-base:4.2.2* and *python:3.9*. The Dockerfiles alongside instructions on how to build and utilize them can be found within the NASQQ Github repository [https://github.com/ardigen/nasqq/tree/main/docker].

Table 2: Software components in the NASQQ pipeline.

Programmic Language	Package	Version	Docker
R	ASICS	2.14.0	r_utils:1.0.0
	BiocManager	1.30.22	
	dplyr	1.1.4	
	FELLA	1.18.0	
	ggplot2	3.4.4	
	igraph	1.5.1	
	optparse	1.7.4	
	PepsNMR	1.16.0	
	Rnmr1D	1.3.2	
	svglite	2.1.2	
	visNetwork	2.1.2	
	testthat	3.2.1	
Python	scikit-learn	1.4.0	python_utils:1.0.0
	scipy	1.12.0	
	seaborn	0.13.0	
	shap	0.44.1	
	kaleido	0.2.1	
	matplotlib	3.4.3	
	numba	0.58.1	
	numpy	1.26.3	
	pandas	1.3.3	
	plotly	5.3.1	
	pyarrow	5.0.0	
	pytest	8.0.0	

Source: https://github.com/ardigen/nasqq.

Resource management within Nextflow involves utilizing containers and configuration files that define the execution environment and behavior of the entire workflow. In addition to the *nextflow.config* [https://github.com/ardigen/nasqq/blob/main/nextflow.config] file containing global parameters such as maximum available resources, four additional files were created for pipeline purposes: *base.config* [https://github.com/ardigen/nasqq/blob/main/conf/base.config], *modules.config* [https://github.com/ardigen/nasqq/blob/main/conf/modules.config], *profiles.config* [https://github.com/ardigen/nasqq/blob/main/conf/profiles.config], and *reports.config* [https://github.com/ardigen/nasqq/blob/main/conf/reports.config]. The *base.config* file contains base configuration settings for the workflow, encompassing global settings applicable across all processes and tasks. This configuration file establishes error handling and resource management within a Nextflow process. It includes directives for *errorStrategy*, *maxRetries*, *maxErrors*, cache, as well as resource requirements such as cpus, memory, and time. The *errorStrategy* determines how errors are managed based on the number of attempts and specified parameters. Cache is enabled to optimize performance by reusing previous results. Resource allocations for CPUs, memory, and time increase with each attempt of the task. The *modules.config* sets parameters for all modules in the NASQQ pipeline. Each stage, as follows "spectral_preprocessing", "data_analysis" and "pathway_analysis" is defined with empty arguments for special execution cases and designated directories for publishing results. Additionally, results are published by a unique identifier for each module. The *profiles.config* specifies settings related to Docker and execution profiles. Docker containers are run with user and group IDs matching the current user, with temporary directory management set to "auto". The execution profile named 'standard' specifies that tasks will be executed locally. Parameters are used to define directories for output, reports, working directory, and launch directory, providing flexibility in configuration. Lastly, *reports.config* orchestrates reporting features within the Nextflow pipeline. It includes settings for generating timeline, trace and Directed Acyclic Graph (DAG) visualizations. Timeline reports provide visual representations of task execution timelines within a computational pipeline. Trace reports in the context of computational pipelines provide detailed logs and records of the execution process. DAGs are used in workflow management systems to visualize the structure of pipeline, allowing to understand the order of task execution and identify potential dependencies between tasks. Each feature specifies its output file path using the current date and time, alongside the designated reports directory. If any existing reports or DAG files are present, they will be overwritten accordingly. Upon initiating the Docker environment setup and before execution pipeline, *manifest.csv* and *params.yml* files need to be created. The *manifest.csv* [https://github.com/ardigen/nasqq/blob/main/tests/manifest.csv] is a comma-separated file detailing the dataset information includes columns such as dataset name, batch name (defaulting to "None"), absolute paths to NMR datasets in Bruker format, metadata files for dataset merging, selected sample names, ppm value for the internal reference spectra (defaulting to 0), referencing range (if different from default, otherwise "None"), and the range of the informative part of the spectra. The *params.yml*

[https://github.com/ardigen/nasqq/blob/main/tests/params.yml] file outlines various inputs such as the absolute paths to the manifest.csv file, output directories for storing results and reports, intermediate work files directory, launch directory, maximum retry attempts for task processing, error handling strategy, pulse program for processing, bucketing and warping options, merging datasets for analysis, number of threads allocated for quantification, log1p normalization option, metadata column for analysis, and axis reversal option (please refer to **Table 3**). Moreover, the *run.sh* [https://github.com/ardigen/nasqq/blob/main/tests/run.sh] script containing execution commands is included in the workflow repository. Although, it is feasible to manually execute the workflow using the provided command directly in the command prompt (CMD).

Table 3: Required input parameters for configuring NASQQ pipeline execution.

Input	Description	Data type
manifest	Absolute path to the *manifest.csv* file containing metadata information for the analysis	string
outDir	Absolute path to the directory where the output files will be stored	string
reportsDir	Absolute path to the directory where the analysis reports will be generated	string
workDir	Absolute path to the directory where the intermediate work files will be stored	string
launchDir	Absolute path to the directory from which the pipeline is launched	string
maxRetries	Number of attempts the pipeline should make to process a task before giving up	integer
errorStrategy	The strategy to handle errors during pipeline execution (terminate/ignore/retry)	string
check_pulse_samples	The pulse program specified in the manifest file for processing	string
run_bucketing	Enable/disable bucketing for simplifying the density of peaks before metabolite quantification	boolean
run_warping	Enable/disable warping for spectra re-alignment based on a reference spectrum	boolean
run_combine_project_batches	Enable/disable merging datasets for data analysis where batch is not “None”	boolean
ncores	The number of threads allocated for the *ASICS* quantification task	integer
log1p	Enable/disable log1p normalization of metabolites before data analysis	boolean
metadata_column	The column containing binary state information for the data analysis module	string
reverse_axis_samples	Specifies whether to invert the axis for all samples or selected samples based on a threshold	string

Source: https://github.com/ardigen/nasqq.

Initial testing of the scripts for the "*Spectral Preprocessing*" and "*Metabolites Identification*" stages utilized raw FIDs from *Pseudomonas Aeruginosa* bacteria from our work [67]. While the NASQQ pipeline initially targeted human metabolomics, particularly serum samples, subsequent modular tests exclusively involved human experiments. The *Familial Dysautonomia* open dataset [73] served as a use case during the end-to-end testing phase. Detailed descriptions of the case study are provided in the results section "Results: Application of NASQQ pipeline on *Familial Dysautonomia* serum samples". Testing and results generation took place on a local machine, specifically the Lenovo Thinkpad T14 equipped with AMD Ryzen 7 PRO 4750U with Radeon Graphics and 32 GB RAM.

II.2 Spectral processing of raw 1D spectra and metabolites identification

The NASQQ pipeline consists of three primary stages: "*Spectral Preprocessing*", "*Data Analysis*" and "*Biological Interpretation*". Each stage has specific objectives, input/output requirements, and detailed procedures, which are summarized in **Table 4**. Due to the large size of the scripts, only descriptions are included here. However, on the NASQQ GitHub repository the full modules [https://github.com/ardigen/nasqq/tree/main/modules] and all scripts [https://github.com/ardigen/nasqq/tree/main/bin] are available. In the "*Spectral Preprocessing*" stage, 13 mandatory and 2 optional modules written in R programmatic language handle the preprocessing tasks, with metabolite quantification as the final step. While some established functions were utilized, the primary innovation involved the creation of custom scripts from scratch. These scripts were designed to efficiently manage the computational environment, handle I/O operations, and facilitate data preprocessing through wrapper functions. Additionally, the scripts focused on storing results, transferring data between modules, and effectively visualizing outcomes. The transformation from standalone scripts to structured Nextflow components represents a significant advancement in the field. The core functionalities of the original scripts were retained but adapted to integrate into Nextflow's architecture. This integration highlights efficient data flow management through channels, robust error handling, and the use of Nextflow-specific features. The workflow was designed based on thorough literature research to ensure alignment with current best practices and advancements in the field. Testing was performed on real laboratory data and open datasets, confirming the robustness and reliability of the workflow. This structured approach not only enhances the reproducibility and scalability of the analyses but also marks a novel contribution to the field of metabolomics, particularly in the context of unification of NMR spectral data preprocessing.

The intention behind elaborating on each module within the subsequent sections of Chapter II is to offer a comprehensive grasp of its functionality and objectives within the NASQQ pipeline. Each description focuses on a specific module, detailing its input parameters, processing steps, methodology and output formats. This methodical approach allows comprehension of the data flow and the efforts invested in the development of end-to-end solution. For detailed insights into

the complexities of the scripts, one can refer to dedicated readme files within the GitHub repository's auxiliary documents [https://github.com/ardigen/nasqq/tree/main/docs].

Table 4: Modules overview in the NASQQ Pipeline.

Stage	Module	Objective	Input	Output
Spectral Preprocessing	load_fids.nf	Load raw FIDs from a given path, select sample names and filter pulse program.	• raw FIDs path	• raw FIDs RDS file
	raw_fids_visualisation.nf	Visualise raw FIDs.	• raw FIDs RDS file	• plots RDS file • plots SVG files
	group_delay_correction.nf	Remove Bruker Group Delay.	• raw FIDs RDS file	• post group delay removal FIDs RDS file • plots RDS file • plots SVG files
	solvent_suppression.nf	Estimate and remove residual solvent signal from the FIDs.	• raw FIDs RDS file • post group delay removal FIDs RDS file	• post solvent suppression FIDs RDS file • plots RDS file • plots SVG files
	apodization.nf	Increase spectral signal-to-noise ratio.	• raw FIDs RDS file • post solvent suppression FIDs RDS file	• post apodization FIDs RDS file • plots RDS file • plots SVG files
	zero_filling.nf	Improve the visual representation of spectra by adding zeros.	• raw FIDs RDS file • post apodization FIDs RDS file	• post zero filling FIDs RDS file • plots RDS file • plots SVG files
	fourier_transformation.nf	Apply Fourier transformation, transition from time domain FIDs into the frequency domain spectra.	• raw FIDs RDS file • post zero filling FIDs RDS file	• spectra RDS file • plots RDS file • plots SVG files
	zero_order_phase_correction.nf	Phase spectra for the real part to be in pure	• raw FIDs RDS file	• post phasing spectra RDS file

Stage	Module	Objective	Input	Output
		absorptive mode.	• spectra RDS file	• plots RDS file • plots SVG files
	internal_referencing.nf	Align spectra with an internal reference compound.	• raw FIDs RDS file • post phasing spectra RDS file	• post internal referencing spectra RDS file • plots RDS file • plots SVG files
	baseline_correction.nf	Estimate and remove spectral baseline from the spectral profiles.	• raw FIDs RDS file • post internal referencing spectra RDS file	• post baseline correction spectra RDS file • plots RDS file • plots SVG files
	negative_values_zeroing.nf	Reduce to zero all negative values in spectra.	• raw FIDs RDS file • post baseline correction spectra RDS file	• post negative zeroing spectra RDS file • plots RDS file • plots SVG files
	warping.nf	(Optional) Warp and realign spectra based on Semi-Parametric Time Warping technique.	• raw FIDs RDS file • post negative zeroing spectra RDS file	• post warping spectra RDS file • plots RDS file • plots SVG files
	window_selection.nf	Select the informative part of spectra.	• raw FIDs RDS file • post negative zeroing spectra (or post warping) spectra RDS file	• post window selecting spectra RDS file
	bucketing.nf	(Optional) Simplify density of spectra peaks.	• post window selecting spectra RDS file	• post bucketing spectra RDS file • plots RDS file • plots SVG files
	normalization.nf	Normalize the spectra.	• raw FIDs RDS file • post window selecting (or post bucketing)	• post normalization spectra RDS file

Stage	Module	Objective	Input	Output
			spectra RDS file	• post normalization spectra TXT file • plots RDS file • plots SVG files
	metabolites_quantification.nf	Identify and quantify metabolites based on normalized spectra.	• post normalization spectra TXT file	• quantified metabolites RDS file • quantified metabolites TXT file
Data Analysis	add_metadata.nf	Merge metadata with quantified metabolites relative abundances.	• quantified metabolites TXT file • metadata CSV file	• quantified metabolites with metadata CSV file
	combine_dataset_batches.nf	(Optional) Combine batches from the dataset before performing data analysis.	• multiple quantified metabolites with metadata CSV files	• quantified metabolites with metadata batch combined CSV file
	features_processing.nf	Load, scaling and conduct sanity checks.	• quantified metabolites with metadata w/wo batch combined CSV file	• preprocessed metabolites PARQUET file
	exploratory_data_analysis.nf	Perform Principal Component Analysis, create exploratory analysis visualizations.	• preprocessed metabolites PARQUET file	• plots SVG files
	univariate_analysis.nf	Detect outliers, check data normality and perform univariate statistical tests.	• preprocessed metabolites PARQUET file	• outliers TXT file • univariate results CSV file
	multivariate_analysis.nf	Analyze metabolite data using various machine learning models.	• preprocessed metabolites PARQUET file	• multivariate results CSV file • plots SVG files
Biological Interpretation	pathway_analysis.nf	Create pathway enrichment using KEGG database entries.	• univariate/ or multivariate results TSV file	• pathway enrichment RDS file • pathway enrichment HTML file • plots PNG files

Source: Own elaboration.

II.2.1 Raw FIDs loading and visualization

The *load_fids.nf* module is configured with four input arguments, governing the project id, the path of provided spectra, the type of presaturation pulse program, and an optional parameter determining whether to remove duplicated sample names. This module reads FIDs from the specified path, verifies the correct passage of the pulse program, and ensures consistent sample names with the respective pulse programs. If the optional argument is provided, duplicated sample names are removed. Following this, a list object containing raw FIDs alongside their corresponding metadata is stored in the R Data Store (RDS) object. RDS is a file format utilized in the R programming language for storing serialized R objects. The resulting RDS object is then utilized in the subsequent module, *raw_fids_visualization.nf,* for plot generation. Plots are configured with the X-axis represented in microseconds [µs] and Y-axis showcasing intensity of signal. Raw FIDs figures are created and output is written to both Scalable Vector Graphics (SVG), an XML-based vector image format, and RDS objects.

II.2.2 Group delay correction

The *group_delay_correction.nf* module requires two input arguments: project ID and raw RDS. As Bruker FID signals commonly display a phenomenon called “group delay” - a time delay in the initial segment of the FID - the module executes a function to eliminate this portion of the FID. Information about this phase shift, resulting from a group delay, is accessible when loading the FID using *PepsNMR::ReadFids* function and is stored in a raw RDS file from the previous step of analysis. Module generates comparison plots pre- and post-correction, preserves SVG plots, compiles a list of plots in RDS, and stores the corrected FIDs in an RDS object.

II.2.3 Solvent suppression

The input parameters for the *solvent_suppression.nf* module include the project ID, as well as RDS files containing raw FID objects and FID post-Group Delay removal. FIDs naturally exhibit a wavy shape, and assuming water is the main compound in the analyzed samples, its signal can be represented by smoothing the FIDs. Subsequently, this smoothed signal, referred to as the solvent residuals resonance signal, is subtracted from the original FIDs. The presence of solvent residuals in the spectrum has the potential to obscure relevant signals from the molecules of interest. In this stage, the module employs a Whittaker smoother [179] implemented in *PepsNMR::SolventSuppression* function to estimate and eliminate the residual solvent signal from the FIDs in the time domain. The Whittaker smoother, a function represented by formula:

$$V + \lambda R \tag{1.5}$$

is composed of two terms – the sum of squared differences between the original and smoothed signals (V) and a measure of the roughness of the estimated signal (R). Where, lambda (λ) represents the penalty on roughness utilized in calculating the smoothed version of the FID. A higher value of lambda results in a smoother estimated solvent signal. The module generates comparison plots before and after solvent suppression, SVG plots, along with a list of plots stored in RDS format. Additionally, the corrected FIDs are saved in an RDS object.

II.2.4 Apodization

The *apodization.nf* module's input parameters encompass the project ID and RDS files containing raw FID objects, both raw FID and post-solvent suppression. Apodization is employed to enhance the spectral signal-to-noise ratio by multiplying the FID signal with a positive signal, typically exhibiting decay [180]. This technique leverages the fact that signal intensity diminishes over time, unlike noise, which maintains a constant amplitude, resulting in a noisy tail at the end of the FID. As the area under the spectral peak remains constant, a quicker decay leads to decreased peak heights in spectra, thereby diminishing spectral resolution. The FID captured by the ^{1}H NMR instrument comprises both real and imaginary components of the decaying signal, consisting of numerous data points. It reflects the aggregate of distinct signal components originating from various proton nuclei. A typical FID, represented as:

$$s_0 exp(i2\pi\nu t) exp(\frac{-t}{T}) \quad (1.6)$$

exhibits a spectral peak at frequency v with a width inversely proportional to T. Where, the equation describes a signal with an initial amplitude (s_0), oscillating sinusoidally at frequency (v), and decaying exponentially with a decay constant of $\frac{1}{T}$ over time (t). The spectral peak, referred to as a spectral line, and its width, termed spectral width, are crucial throughout the process. In the apodization function utilized in the module, exponential multiplication is implemented, where the decaying exponential is expressed as:

$$exp(-\ t(\frac{1}{T} + LB)) \quad (1.7)$$

A smaller decay time (T^*), which satisfies $\frac{1}{T^*} + LB = \frac{1}{T}$, leads to broader spectral lines. As time (t) progresses, the exponential term decreases, signifying the signal's decay. Hence, line broadening (LB) parameter, balances between enhancing SNR and degrading spectral resolution. To avoid substantial sensitivity and resolution losses, identifying optimal trade-off parameters for the apodization signal is crucial. As per the *PepsNMR* R package documentation [181], $LB = 0.3$ is recommended for NOESY, while $LB - 0.01$ is suitable for CMPG presaturation

pulse sequences. If the user hasn't specified the *check_pulse_samples* in the global parameters configuration (*params.yml* file), the default *LB* value is set to 0.3. After completing this procedure, the module generates comparison plots before and after apodization, SVG plots, and a list of stored plots in RDS format. Additionally, the corrected FIDs are stored in an RDS object.

II.2.5 Zero filling

The input parameters for the *zero_filling.nf* module include the project ID and RDS files containing both raw FID objects and those post-apodization. Zero filling enhances the visual representation by creating more distinct lines in the spectra. This involves appending a specified number of zeros at the end of the FIDs, rounded to the nearest power of 2 to aid the subsequent Fourier Transform. Following this, the module produces comparison plots before and after zero-filling, SVG plots, and a list of stored plots alongside the corrected FIDs in RDS format.

II.2.6 Fourier transformation

The input parameters for the *fourier_transformation.nf* module include the project ID and RDS files containing raw FID objects and those post-zero filling. The Fourier Transform, as described in the "General concepts of nuclear magnetic resonance" of the Introduction section, extracts signals from the time domain and translates them into peaks in a spectrum with specific characteristics. The transformation implemented in the module utilizes the *PepsNMR::FourierTransform* function. The resulting spectrum is initially calibrated in Hertz [Hz] based on spectrometer acquisition parameters and is then converted into a chemical shift scale in ppm. Output includes comparison plots before and after transformation in SVG format, along with a list of stored plots alongside the corrected FIDs in RDS format.

II.2.7 Phase correction

The input parameters of the module *zero_order_phase_correction.nf* consist of the project ID and RDS files containing both raw Fourier-transformed spectra objects and those post-transform. Due to technical factors like incorrect magnetization, the spectrum may display a zero-order phase shift error of a certain angle (φ_0) expressed as

$$F = F_{phased}\, exp(i\varphi_0) \tag{1.8}$$

where F_{phased} represents the ideally phased spectrum. This phase shift remains constant for all signal vectors, irrespective of spectral frequencies. Consequently, the real and imaginary components of the signal produce a blend of absorptive and dispersive mode line shapes, necessitating phase correction. Utilizing the principle that a perfectly phased signal (F_{phased}) exhibits a real part starting

at maximum and an imaginary part starting at 0, with the resulting spectrum featuring a positive real part in absorptive mode and an imaginary part in dispersive mode including positive and negative peaks, an optimal angle can be automatically determined. This is achieved by maximizing a suitable criterion related to the positivity of the spectrum's real part.
The *PepsNMR::ZeroOrderPhaseCorrection* function implemented in module quantifies positivity criteria using the root mean square (rms) criterion, representing the ratio between the sum of squares of positive intensities and the sum of squares of all intensities in the spectrum. The function identifies an optimal angle by maximizing criteria related to the positivity of the real part of the spectrum and rotates each spectrum in the spectral matrix, resulting in processed spectra with their respective rotation angles. The module generates comparison plots before and after the phasing process in SVG format. Additionally, it produces a list of stored plots and saves the phased spectra in RDS format.

II.2.8 Internal referencing

The module *internal_referencing.nf* requires project ID and RDS files containing both raw FIDs and post-phasing spectra as mandatory input parameters. Additionally, optional arguments include range_type (default “nearvalue”, with options: “nearvalue”, “all”, “window” based on the range from *PepsNMR::InternalReferencing* function), target_value (default 0), and fromto_RC (default NULL, relevant if range_type is set to “window”, representing a list with numerical vectors indicating the extremities of intervals to search for the referencing peak). For improved accuracy, it's advantageous to calibrate the scale using a well-established standard, usually an internal reference compound that remains stable against external influences such as temperature or concentration. Ideally, this reference compound should be situated outside the spectral region to ensure clear identification. Trimethylsilyl propionate (TSP) or 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) are commonly used internal reference compounds in NMR, with their standard peaks traditionally set at 0 ppm. For reference compounds other than TSP or DSS, a non-null ppm value in target_value requires manual adjustment by the user. After phase correction, residual artifacts from a first-order phase shift may persist, possibly resulting in inverted spectra. When over half of the spectra intensities are negative, the spectra are inverted to maintain accurate ppm values. The output includes visuals depicting spectra before and after internal referencing, saved in RDS and SVG files. Furthermore, the post-correction object containing spectra is stored in the RDS file.

II.2.9 Baseline correction

The module *baseline_correction.nf* input arguments include the project ID, RDS files containing raw FIDs objects, and post-referencing spectra. In addition to the mandatory inputs, there are two optional arguments: p_bc and lambda_bc. These parameters influence the behavior of the function employing asymmetric least squares (ASL) for baseline correction [182]. The baseline

correction aims to ensure that the signal predominantly comprises positive peaks, which represent metabolites present in the samples. This signal is typically superimposed on a baseline, ideally reflecting the absence of any metabolite and maintaining a consistent zero value. Given the assumption that

$$F^{*} = F - Z \tag{1.9}$$

where F denotes the initial spectrum and Z represents its estimated baseline, the corrected spectrum F^{*} is derived by subtracting Z from F once Z is approximated. In this module for each spectrum, its baseline is estimated and subsequently subtracted. Because negative signals pose issues, the estimated baseline aims to prevent negative values in the corrected spectrum, $F - Z$. Consequently, in the objective function to be minimized, the squared differences $(F - Z)$ are weighted by p if $F - Z > 0$ or by $1 - p$ if $F - Z < 0$. To prevent the baseline from matching the peaks exactly and to maintain smoothness, a penalty term on Z is incorporated into the objective function. The significance of this constraint is governed by λ. Hence, two additional parameters in module, p_bc and lambda_bc, represent p and λ respectively, with default values as follows: p_bc =0.0001, where a smaller value makes the baseline less reactive to peaks below the function while striving to remain beneath it. For lambda_bc, the default is 5e+06, with a larger value resulting in a smoother baseline. Setting lambda_bc to 0 renders the baseline identical to the signal, effectively zeroing the corrected signal. The output comprises visual representations of the spectra before and after the correction process, preserved in both RDS and SVG formats. Furthermore, the RDS file retains the post-baseline correction object containing the spectra.

II.2.10 Negative values Zeroing

The input parameters for the *negative_values_zeroing.nf* module include the project ID and pre-processed RDS files, containing both the raw FIDs object and the baseline-corrected spectra. If there are residual negative values in the spectrum due to incomplete phase or baseline correction, it can complicate further interpretation. To resolve this issue, any remaining negative intensities are adjusted to zero. The output comprises figures showcasing the spectra both before and after setting all negative values to zero, saved in RDS and SVG formats. Additionally, the RDS file stores the object containing the corrected spectra.

II.2.11 Warping

The optional module *warping.nf* takes input arguments such as the project ID, RDS files containing spectra after setting negative values to zero, and raw FID objects. Implemented within *PepsNMR::Warping*, this module applies a warping function $W(v)$ to a normalized spectrum F^{*}. This

function distorts the ppm axis through a combination of a polynomial term and a penalized B-splines (P-splines) term. Mathematically, the warping function $W(v)$ is defined as:

$$W(v) = \sum_{k=0}^{K} \beta k^{v^k} + \sum_{l=1}^{L} \alpha_l B_l(v) \quad (2.0)$$

In this equation, the first term $\sum_{k=0}^{K} \beta k^{v^k}$ represents a polynomial of order K, where βk are the corresponding polynomial coefficients. The second term $\sum_{l=1}^{L} \alpha_l B_l(v)$ is a weighted sum of B-splines, with L being the number of B-splines and α_l as the coefficient for the lth B-spline $B_l(v)$. These B-spline curves are assembled from polynomial segments and smoothly connected. Afterward, the normalized and distorted spectral profile is derived by interpolating from the discrete warping function. The essence of warping lies in constructing $W(v)$ to minimize the distance between a warped spectrum $F(W(v))$ and the reference spectrum. Once the similarity between the transformed spectrum and the chosen reference spectrum is improved, the module proceeds to realign the spectra to enhance profile resemblance. Without prior information, this robust reference selection process enables the selection of the spectrum that minimizes the sum of squared distances with all other spectra after warping. The output comprises figures illustrating the spectra before and after applying the warping function, saved in RDS and SVG formats. Additionally, the post-warping object is preserved in the RDS file.

II.2.12 Window selection

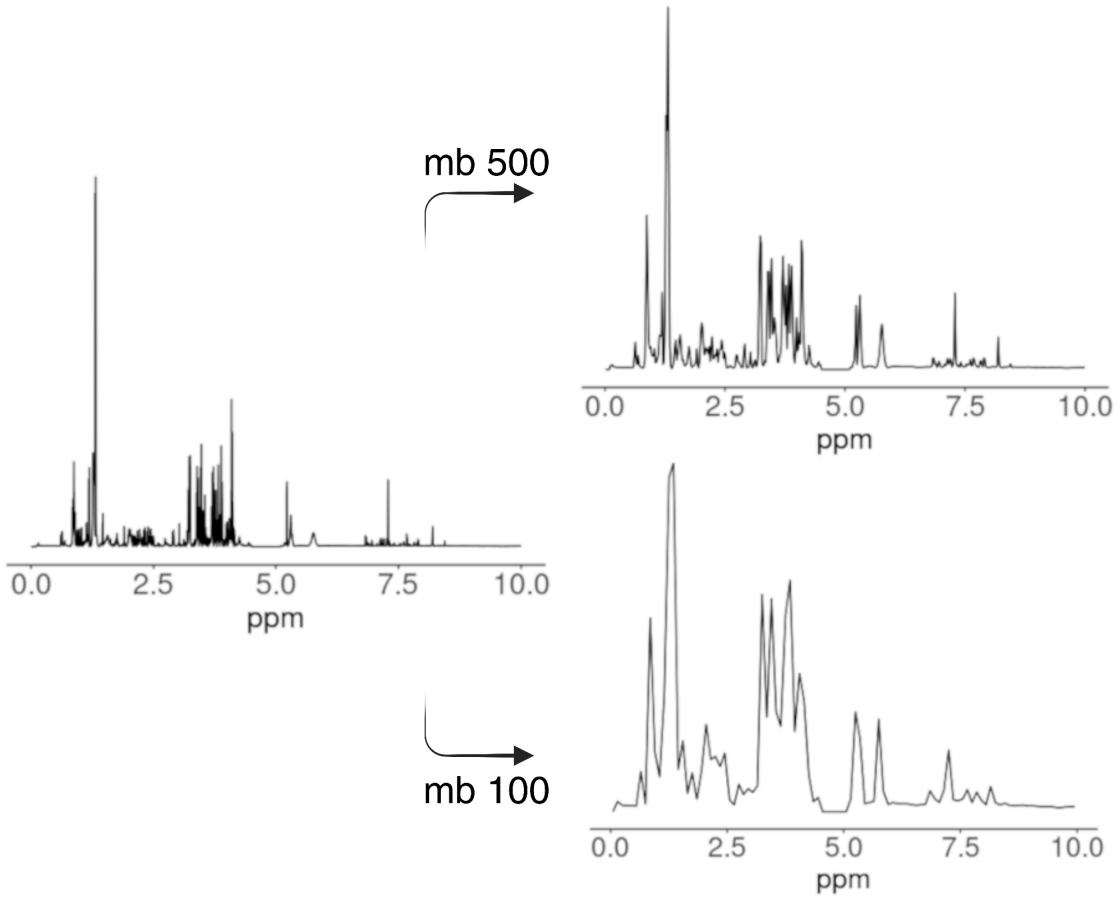
The *window_selection.nf* module takes as input the project ID and the RDS file containing spectra after either setting negative values to zero or applying warping. Optionally, it also accepts a ppm range specified as a list. The module utilizes a function to select the informative part of the spectra based on the provided ppm range and saves it to an RDS file. The default ppm range is set from 0 to 10. The output includes spectra after processing saved in RDS format.

II.2.13 Bucketing

In the optional module *bucketing.nf*, the input parameters include the project ID and an RDS file containing spectra after the selection of the informative part. The abundance of data points and subtle residual peak shifts can pose challenges for future multivariate analyses. To address this, bucketing is employed to condense the original spectral intensities into predefined intervals. This process, also known as data reduction, effectively decreases the number of data points from m (often >35,000) to mb (default value is configured to be 5000), where m represents the original number of points and mb denotes the number of buckets. The bucketing function within the module

utilizes the trapezoidal method, which involves creating trapezoidal shapes between adjacent data points and calculating the area under each trapezoid, thereby balancing the preservation of spectral information with the removal of peak shifts. The module stores visual representations of the spectra before and after bucketing in both RDS and SVG file formats. Additionally, the RDS file retains the object containing the spectra data itself. **Figure 9** illustrates the effects of bucketing on NMR spectral data. The left plot shows the original spectral intensities with a high number of data points. The right plots depict the spectra after applying bucketing to reduce the number of data points, with the top-right plot showing the data reduced to 500 buckets (mb 500) and the bottom-right plot showing the data reduced to 100 buckets (mb 100).

Figure 9. Impact of bucketing on NMR spectral data reduction using different numbers of buckets.



Source: Created in Biorender.com, own elaboration.

II.2.14 Normalization

The *normalization.nf* module requires specific inputs, such as the project ID and RDS files containing both raw FID objects and spectra selected post-windowing or post-bucketing. Furthermore, an optional parameter enables users to specify the desired normalization method. The available choices include: “mean”, “pqn”, “median”, “firstquartile”, or “peak”. The objective of normalization

is to ensure that the overall intensity distributions across samples are comparable, simplifying the identification of genuine biological differences between samples. Normalization methods like mean, median, and quartile normalization can be represented by a single equation:

$$X_{ij}^{'} = \frac{X_{ij}}{f(X_i)} \quad (2.1)$$

Here, $X_{ij}^{'}$ represents the normalized intensity of the *j*th variable in the *i*th spectrum, X_{ij} is the original intensity, and $f(X_i)$ stands for the normalization factor calculated based on the chosen method (mean, median, or quartile) for the *i*th spectrum. Peak normalization scales each spectrum by the maximum peak intensity within a specified interval. Probabilistic quotient normalization (PQN) corrects systematic biases by normalizing each spectrum against a reference spectrum, typically the median spectrum [183]. It follows the equation:

$$X_{ij}^{'} = X_{ij} \times \left(\frac{median(X_j)}{X_{ij}}\right) \quad (2.2)$$

where, $median(X_j)$ represents the median intensity across all spectra. By default, the water resonance area (4.5-5.1 ppm) is excluded from serum spectra before normalization. After the normalization process concludes, the module produces post-normalization RDS files containing comparison plots and spectra. It also generates stacked spectra plots for all samples using the manually adjusted *rnmr1d::plotSpecMat* function [170]. Lastly, the module outputs a TXT file containing the normalized spectra.

II.2.15 Metabolites quantification

The *metabolites_quantification.nf* module requires a project ID, a TXT file with the normalized spectra, and the number of processor cores for computation. Optionally, users can specify a quantification method from those available in the *ASICS::ASICS* function ("FWER", "Lasso", or "both", with "both" as the default if none is chosen) [184]. The *ASICS* method for metabolite quantification involves deconvoluting NMR spectra into individual metabolite signals via a reference library of pure spectra. Metabolite identification and quantification in the complex mixture spectrum are carried out using both the previously preprocessed complex mixture spectrum and the pure metabolite spectra from the reference library from the *ASICS* R package. The reference library consists of pure spectra of 191 known metabolites. Each spectrum in the library represents a unique metabolite, characterized by its chemical shifts and peak patterns. The process starts with finding the reference spectrum and involves utilizing FFT cross-correlation [185]. This technique helps identify the spectrum in the pure metabolite reference library that most closely matches the complex mixture spectrum. Once the reference spectrum is identified, alignment between

the mixture spectrum and the reference spectrum is achieved using FFT cross-correlation along with a hierarchical classification approach. The hierarchical classification approach systematically narrows down the list of potential matches, ensuring both accuracy and efficiency. Subsequently, individual peaks of each library spectrum undergo alignment via local linear regression centered around each peak, enhancing the robustness of metabolite identification through techniques like Least Absolute Shrinkage and Selection Operator (Lasso) [186] and Family-Wise Error Rate (FWER) [187]. These methods aid in navigating the complexity of NMR data by penalizing less probable solutions, thereby focusing on the most likely metabolites present in the sample. Lasso regression, characterized by shrinkage towards a central point like the mean, promotes simpler, sparser models with fewer parameters. The lasso regression minimizes the following cost function:

$$min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \quad (2.3)$$

where, $\frac{1}{2N}\sum_{i=1}^{N}(y_i - X_i\beta)^2$ is residual sum of squares (RSS), This term measures the difference between the observed values y_i and the predicted values $X_i\beta$. It is essentially the mean squared error of the model's predictions. The regularization term (L1 Norm) $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ adds the absolute values of the coefficients $\left|\beta_j\right|$. This L1 norm encourages sparsity by shrinking some coefficients to exactly zero, effectively performing feature selection. FWER is a method to control the probability of making one or more false discoveries (type I errors) when performing multiple hypothesis tests. In the context of metabolite quantification, this approach ensures that the identification of metabolites is statistically robust. The Bonferroni correction

$$\alpha_{corrected} = \frac{\alpha}{m} \quad (2.4)$$

is a simple and commonly used method to control the FWER. It adjusts the significance level (α) for multiple comparisons by dividing α by the number of tests (m). By controlling the FWER, function minimizes the risk of false positives, enhancing the reliability of the metabolite identification process. Once the metabolites are identified, their concentrations are quantified based on their signal intensities relative to the reference spectra. This quantification is done in a way that accounts for potential overlaps in spectral peaks. The module's output consists of a quantified *ASICS* object with metabolite quantification in RDS format. Additionally, a table containing the relative abundances of quantified metabolites is provided in both RDS and TXT formats.

To summarize, the FIDs are imported from specified paths, subjected to various corrections and noise removal processes, and then transformed via Fourier transformation to generate spectra.

This transformation converts the FIDs' time domain into the frequency domain, which is then expressed as the chemical shift. The spectra undergo phase correction and alignment to an internal reference compound. The spectral baseline is estimated and removed, with negative values set to zero. Subsequent modules offer an optional warping function, using polynomials and P-Splines to adjust the chemical shift axis, enhancing the similarity between the transformed and reference spectra. The final steps involve selecting informative spectral regions, optional bucketing, and normalization. Finally, the normalized signals are quantified into metabolites using a pure reference library object from the *ASICS* R package. After completing the spectral processing and metabolite identification cycle, a datatable with the relative abundances of detected metabolites is prepared for the next stage of the NASQQ pipeline.

II.3 Data analysis module - univariate tests

II.3.1 Features processing

The "*Data Analysis*" stage introduces an original approach to handling data and evaluating significant metabolites. This innovative method improves the efficiency and accuracy of metabolite analysis, providing a well-built framework for identifying and interpreting significant metabolites related to specified class within the datasets. In this stage, three modules utilizing Python scripts are executed concurrently: Exploratory Data Analysis (EDA), univariate analysis, and multivariate analysis. This stage starts with the *add_metadata.nf* module, which integrates additional metadata that includes patient state information (e.g., disease/healthy). This step is followed by an optional logarithmic (log1p) transformation of metabolite abundances. Batch information can also be incorporated if available. The module require several input parameters: "input_path", which is the directory path for the input data file; "disease_state", the column indicating the disease status of the samples; "batch", which specifies batch column for the samples; "log1p", a flag for applying the $log(1 + x)$ transformation; and "metadata", the path to the metadata file. The data can then be optionally merged by batch using the *combine_dataset_batches.nf* module, allowing for the data analysis of batches either separately or as a combined dataset.

The f*eatures_processing.nf* module is responsible for loading and preprocessing quantified metabolite data. It requires inputs including metadata merged with the relative abundances of metabolites, the column names for disease state and patient IDs, and optionally the column name indicating batch type. The module removes columns with a high percentage of zero or empty values, with a default threshold of 0.7 across all features. It also performs metadata checks to ensure class balance and feature availability. The sanity check criteria include a minimum count of 3 for each class and a minimum class percentage of 0.03 to prevent edge-case model training and numerical instability. The output is a PARQUET file containing the preprocessed features.

II.3.2 Univariate tests

The univariate_analysis.nf module as an input requires a PARQUET file from the previous f*eatures_processing.nf* module along with the names of the disease state column and patient IDs column. The process begins by employing the Shapiro-Wilk test [188] to evaluate the normality distribution of individual features. Based on the results of this assessment, it proceeds to conduct either a Student t-test [189] for normally distributed data or a Mann-Whitney U test [190] for non-normal distributions. The Student t-test is utilized to ascertain whether there exists a significant difference between the means of two independent groups, assuming normality and equal variances. Conversely, the Mann-Whitney U test, a non-parametric alternative, compares two independent groups without relying on the assumption of normality, instead ranking all observations from both groups and evaluating differences in central tendency. Following either test, FDR correction [191] is applied to the states specified in the metadata groups column. Considering the increased likelihood of obtaining at least one significant result due to multiple statistical tests, FDR evaluates the expected proportion of falsely rejected null hypotheses (false positives) among the rejections. In addition, the module applies the Local Outlier Factor (LOF) outlier detection method [192]. The LOF algorithm is an unsupervised anomaly detection technique that calculates the local density deviation of a data point relative to its neighbors. Samples with significantly lower density compared to their neighbors are identified as outliers. The output is saved as a TXT file containing the names of outliers and a CSV file presenting the results of the univariate tests, with columns as follows: “Feature”, “Test”, “Statistic”, “p-value” and “FDR”.

II.4 Data analysis module - multivariate approach and machine learning models

II.4.1 Exploratory data analysis

Simultaneously to the *univariate_analysis.nf* module, the *exploratory_data_analysis.nf* and multi*variate_analysis.nf* modules analyze metabolite data. The *exploratory_data_analysis.nf* is performed in order to summarize the main characteristics of a dataset with visual methods. Essential inputs encompass metadata merged with metabolite relative abundances in PARQUET format, along with the disease state column name and patient IDs column. Optionally, a batch column name can be provided, leading to additional plots for investigating potential batch effects on the data. The module standardizes features by subtracting the mean and scaling to unit variance. In this process, the standard score of a sample x is computed as

$$z = \frac{(x-u)}{s}, \tag{2.5}$$

where u is the mean of the training samples and s is the standard deviation of the training samples. Following that, the module conducts PCA and produces visualizations for exploratory data analysis.

The resulting plots comprise: the explained variance of PCA (scree plot), a matrix of scatter plots depicting the first four principal components, box plots, and distribution plots for 10 randomly chosen features. Furthermore, cluster maps displaying feature-wise and patient-wise correlations are generated.

II.4.2 Multivariate analysis

The *multivariate_analysis.nf* module requires the PARQUET file previously used in the univariate module, as well as the specification of the columns for disease state and patient IDs. The multivariate module analyzes metabolite data using iterations of logistic regression and random forest models [193-194]. The models list include:

- "Logistic regression L1 (C=0)",
- "Logistic regression L1 (C=1)",
- "Logistic regression L2 (C=0)",
- "Logistic regression L2 (C=1)",
- "Logistic regression L1/L2 (*elastic net*), 3-fold cross-validation",
- and "Random Forest (num_trees = 100)".

Here, C represents the strength of the L1 or L2 regularization component, and num_trees is the number of decision trees used for predictions in the random forest classifier. The training and validation process is executed using a multi-split cross-validation method, specifically the *StratifiedShuffleSplit* function from the *sklearn* Python package. This function ensures that each fold of the cross-validation preserves the same proportion of samples for each class as the original dataset. *StratifiedShuffleSplit* works by randomly shuffling the data and then splitting it into a specified number of train/test splits. In this context, 30% of the samples are reserved for validation splits in each iteration. During the cross-validation process, multiple iterations are performed to ensure that the model's performance is robust and not dependent on a particular train/validation split. After evaluating the performance of various models across these splits, the model with the highest mean Receiver Operating Characteristic Area Under the Curve (ROC AUC) is selected [195]. This metric provides a comprehensive measure of the model's ability to distinguish between classes. Once the best-performing model is identified, Shapley values are computed to understand the contribution of each feature to the model's predictions [196]. Shapley values, derived from cooperative game theory, assign a value to each feature representing its contribution to the prediction [197]. These values are then used to calculate the relative importance of each feature according to the provided equation, offering insights into which features most significantly impact the model's decisions.

$$imp = \frac{1}{m}\sum_{i=1}^{m}\left|shap_i\right|, \qquad (2.6)$$

where, $shap$ denotes a matrix containing Shapley values, imp represents a vector indicating the mean absolute Shapley values for each feature (metabolite), and m corresponds to the total number of samples. An individual component j of the relative importance vector, designated as $relimp_j$, can be formulated as follows:

$$relimp_j = \frac{imp_j}{||imp||_1}, \tag{2.7}$$

where $||\cdot||_1$ is the L_1 norm. Subsequently, the selected features are determined by initially arranging the $relimp$ vector in descending order of the features with the most pronounced signal (highest absolute Shapley value). The maximum k features are then selected such that their cumulative absolute importance remains below the threshold of 0.95. The generated output is saved as a CSV file containing the results of the multivariate analysis. This file includes the outcomes of each model evaluated across stratified randomized folds of training and evaluation data, along with metabolites arranged according to their relative importance Shapley values from the best-performing model. If the optimal model happens to be a logistic regression model, a feature importance plot is also produced and stored as an SVG file.

For the analysis of spectral data, employing both univariate and multivariate methods offers significant advantages. Univariate analysis allows for the examination of data on a single-feature basis, which is particularly useful for identifying local changes and variations within individual metabolites in specific classes such as disease state. This approach is beneficial for pinpointing specific alterations that might be missed when considering the data in aggregate. In contrast, multivariate analysis considers the data as a whole, enabling the identification of the most important dimensions and patterns that emerge from the combined effects of multiple features. This holistic perspective is crucial for understanding the underlying structure and relationships within the data, providing insights that are not apparent from univariate analysis alone. Additionally, exploratory visualization helps in understanding the correlation between features and/or patients, revealing potential relationships and patterns that might otherwise go unnoticed. EDA visualizations also aid in detecting internal biases, such as batch effects, which can significantly impact the validity of the analysis. By identifying and accounting for these biases, more accurate and reliable results can be ensured. In summary, combining univariate and multivariate methods in spectral data analysis allows for a comprehensive understanding from both localized and holistic perspectives, while exploratory visualization enhances the interpretability and reliability of the analysis by revealing correlations and potential biases.

II.5 Biological interpretation of features derived from data analysis module

In the "*Biological Interpretation*" stage, the *FELLA* R package [198] and the KEGG database [143] are used to identify overlaps among biological pathways. The *pathway_analysis.nf* module is designed for metabolomics data interpretation, starting with a list of affected metabolites from either univariate or multivariate analysis modules, sorted by importance to potentially highlight key biological pathways. This module requires a KEGG organism identifier, with the default set to "hsa" for human samples, and optionally allows defining the number of metabolites for inclusion in pathway enrichment analysis, defaulting to 20 features (metabolites). Since the metabolite list generated during the "*Spectral Processing"* and "*Data Analysis*" stages comprises general compound names, it needs to be translated into KEGG database identifiers. To obtain KEGG identifiers, the metabolite list is matched against an internal dictionary of corresponding metabolites sourced from the KEGG database (http://rest.kegg.jp/list/compound, accessed on 29 Feb 2024). After standardizing the input data, the module builds a hierarchical representation of the organism's biochemistry using KEGG data. This knowledge graph spans various molecular levels and connects metabolites to pathways through intermediate entities like reactions, enzymes, and KEGG modules. These connections are directly sourced from KEGG annotations. The KEGG database-centric pathway enrichment is achieved by finding a sub-network from the whole KEGG graph that is statistically relevant for a list of input metabolites. The hypergeometric distribution is employed to evaluate if a biological pathway has a greater number of hits within the input metabolite list than expected by chance, considering its size. Pathways are then ranked based on their p-values after multiple testing corrections. Inclusion of a metabolite in a pathway is determined by its ability to be reached from the metabolite in the upward-directed KEGG graph, following the KEGG hierarchy from compound level to pathway. Consequently, metabolites associated with enzymes within a pathway will be considered part of the pathway, even if they were not originally defined in the KEGG pathway. After the initial data enrichment process, an enriched FELLA object is subjected to a propagation algorithm, namely an undirected heat diffusion model, to score the nodes within the graph [199]. This algorithm helps to propagate information across the graph, allowing for the assessment of the relevance and importance of each node. The input metabolites initiate a unitary flow within the network, which can only exit through pathway nodes. This requirement compels the flow to propagate through the intermediate entities, including reactions, enzymes, and modules, as well. To ensure statistical normalization of the scores obtained from the propagation algorithm, normality approximations are computed using parametric z-scores. This involves transforming the raw scores into z-scores, which represent the number of standard deviations a particular score is from the mean. Normalization through z-scores results in p-scores, which are defined as:

$$ps_i = 1 - \Phi(z_i), \quad (2.8)$$

where ps_i is the p-score of node i, z_i is its z-score and Φ is the cumulative distribution function of the standard gaussian distribution. This normalization process helps to standardize the scores and allows for more meaningful comparisons across different nodes in the graph, where nodes are ranked using increasing p-scores. The module outputs encompass a pathway enrichment object stored in an RDS file, along with a static plot and an interactive KEGG-based visualization saved in PNG and HTML formats, respectively.

The approach used in the "*Biological Interpretation*" stage has the potential to significantly improve understanding of enrichment results by offering new insights into the alterations occurring within the biochemistry of the studied organism. It can serve as a valuable starting point for further exploration and investigation into the underlying biological mechanisms driving these changes. By providing a more comprehensive and interactive visualization of pathway enrichment, it allows better identification of key pathways and potential metabolic pathways involved in the observed physiological or pathological processes. The deeper understanding can lead to more targeted research efforts and potentially uncover novel therapeutic targets or diagnostic biomarkers.

III Results: Application of NASQQ pipeline on *Familial Dysautonomia* serum samples

III.1 Open dataset raw spectra and corresponding metadata preparation

In PhD research work, the NASQQ pipeline was employed to transition from raw spectra to data analysis and pathway enrichment [178]. The main objective was to assess the effectiveness and functionality of the workflow using a publicly available dataset. The original dataset's authors also analyzed stool samples and conducted a detailed analysis utilizing mice models to investigate the interplay of the microbiome in FD pathology. However, our workflow focused solely on the serum samples, representing only a part of the original comprehensive analysis. The complexity of the original study, which included both human and animal models as well as various types of biological samples, highlights the multifaceted nature of FD research. Due to the rarity of the disease, samples were collected from various shipments around the world, adding another layer of complexity in terms of batch effects and variability. The approach that utilized the NASQQ pipeline in this work aimed to analyze a subset of this dataset, demonstrating the pipeline's potential for revealing significant metabolic alterations. More details about the disease and the broader scope of the original study can be found in the "Other examples of metabolomics importance" section of the Introduction chapter. The analyzed case study involved 101 serum samples from individuals with *Familial Dysautonomia*, distributed across five distinct batches, representing separate shipments: "*ACSC_HumanSerum_11_5_19*", "*ACSC_HumanSerumFD_11_1_19*", "*Human_Serum_2019_November_Shipment_Round_1*", "*Human_Serum_2020_Feb_March_Shipments*" and "*Human_Serum_2020_January_Shipment*". These included 53 samples from healthy individuals and 48 from FD patients. The raw serum spectra, obtained from a Bruker AVANCE III 600 MHz spectrometer, were retrieved from the Metabolights database [157] under MTBLS5138 reference number [73]. The corresponding metadata — namely *HumanFDProject_MasterMetaData.xlsx* and *README_HumanSerum_NMRFiles_MasterMapFile.xlsx* — were downloaded from https://ftp.ebi.ac.uk/pub/databases/metabolights/studies/public/MTBLS5138/ on February 29, 2024. Given that the original metadata files contained additional information irrelevant to pipeline execution, a new input file was generated, path to the file was included in *manifest.csv* and *params.yml* was adjusted. The *FD_Serum_Metadata_Curation.R* script used for manual metadata creation in the case studies with additional comments, may be provided upon request. Data processing via the NASQQ pipeline integrated the five originally deposited batches of raw spectra.

III.2 Spectral processing module outcomes

In accordance with the methodology described in source publication [73], proton NMR measurements were conducted using the "*zgesgp*" pulse sequence (in Bruker notation). The chemical shift was corrected relatively to the internal standard resonance signal at 0.000 ppm. Fourier transformation and phase correction were then performed. After selecting the informative segment of the spectra, an optional bucketing step was executed with the number of data points set to 5,000. Baseline correction was carried out using the default pipeline parameters (lambda_bc = 5e+06 and p_bc = 1e-04), followed by normalization using the "pqn" algorithm. All remaining steps in the "*Spectral Preprocessing*" stage were carried out using the default pipeline parameters, while the warping step was skipped.

The module *load_fids.nf* and subsequent *raw_fids_visualization.nf* module generated three outputs for each FD batch: raw FIDs RDS files, a plots RDS file, and plots SVG files. The raw FIDs RDS file is a list object containing two data frames: *FIDinfo* and *FIDdata*. *FIDinfo* includes metadata information, while *FIDdata* contains the real and imaginary parts of the FIDs. All this information was collected from the *acqu*, *acqus*, *fid*, and *pulseprogram* files in the main sample folder. The *FIDinfo* metadata is consistent across all batches. For a comprehensive description of *FIDinfo* metadata, please refer to the *PepsNMR* R package manual [181]. Below is an explanation of specific fields in *FIDinfo*, along with their values for all samples:

- TD: 65536 (Time domain size)
- BYTORDA: 0 (Determinant of the endianness [200] of stored data)
- DIGMOD: 1 (Digitization mode)
- DECIM: 2773.333 (Decimation rate of the digital filter)
- DSPFVS: 20 (DSP firmware version)
- SW_h: 7211.538 (Sweep width in Hz)
- SW: 12.01657 (Sweep width in ppm)
- O1: 2815.482 (Spectrometer frequency offset)
- GPRDLY: 67.98589 (Group delay)
- DT: 6.933333e-05 (Dwell time in microseconds)

Regarding the *FIDdata* and SVG plot files, *FIDdata* originally consisted of 32,768 points. Due to its size, only representative groups are presented in the results section. Specifically, data from sample number 320 representative FD patients from the dataset "*ACSC_HumanSerum_11_5_19*" and sample number 160 representatives of healthy relatives from the dataset "*Human_Serum_2020_Feb_March_Shipments*" are shown in figure format. However, all files are available upon request.

Figure 10 shows the real part of the Free Induction Decay signal in the time domain. The x-axis indicates the time scale in units of microseconds multiplied by 1000, while the y-axis indicates the signal intensity. The plot displays the FID signal as a line plot, where the intensity of the signal is plotted against time. Initially, there is a sharp and high-intensity signal that quickly decays and oscillates before stabilizing around zero. This behavior is characteristic of FID, where the intensity starts high and then decays over time due to relaxation processes in the sample. **Figure 11** consists of two subplots comparing the FID signals before and after group delay removal. The top subplot shows the real part of the FID signal with group delay. The x-axis and the y-axis represent time and intensity. The signal starts with minimal oscillation before exhibiting pronounced oscillations and gradually decaying in amplitude. This indicates the presence of group delay, where the initial part of the signal is delayed, causing the oscillations to appear later. The bottom part of the plot displays the real part of the FID signal after the group delay has been removed. In this subplot, the signal begins with immediate, significant oscillations and then decays over time. The removal of the group delay has resulted in the immediate onset of the expected oscillatory decay pattern. Both plots are zoomed in on the initial portion of the FID signal to highlight the effects of group delay and its removal. The comparison clearly shows how the removal of group delay corrects the initial delay, resulting in a more accurate representation of the FID signal's true behavior from the onset. **Figure 12** displays the comparison of the real part of the FID signal and the estimated solvent residuals signal over time. The x-axis and y-axis are labeled similarly as previous figures. The plot features a black line representing the FID signal and a red line representing the estimated solvent residuals signal. The FID signal starts with high intensity, quickly decays, and exhibits oscillations before stabilizing near zero. The estimated solvent residuals signal follows a smoother trajectory, illustrating the impact of solvent residuals on the FID signal. In the bottom subplot, similarly to the top subplot, the FID signal begins with high intensity and quickly decays, but the oscillations appear more immediate and pronounced. The estimated solvent residuals signal is displayed for comparison, showing the effect of the residuals post-processing. Both subplots focus on the initial portion of the FID signal to highlight the differences in signal behavior before and after accounting for solvent residuals. This comparison underscores the impact of solvent residuals on the accuracy of the FID signal's representation. **Figure 13** illustrates the comparison of FID signals over time after an apodization step, which aims to enhance the signal-to-noise ratio. The focus is on the real part of the signals in a zoomed-in view. The gray line represents the apodized FID signal, while the blue line shows the difference between the apodized and solvent-suppressed FID signals: $\Delta(FID\ Apod - FID\ SS)$. Although the changes in the presented samples are subtle, the figure demonstrates how the initial noise is smoothed out after processing, leading to an improved SNR and making the data more suitable for further analysis via Fourier transformation.

Figure 10. The real part of raw FID for exemplary samples numbers 320 and 160.

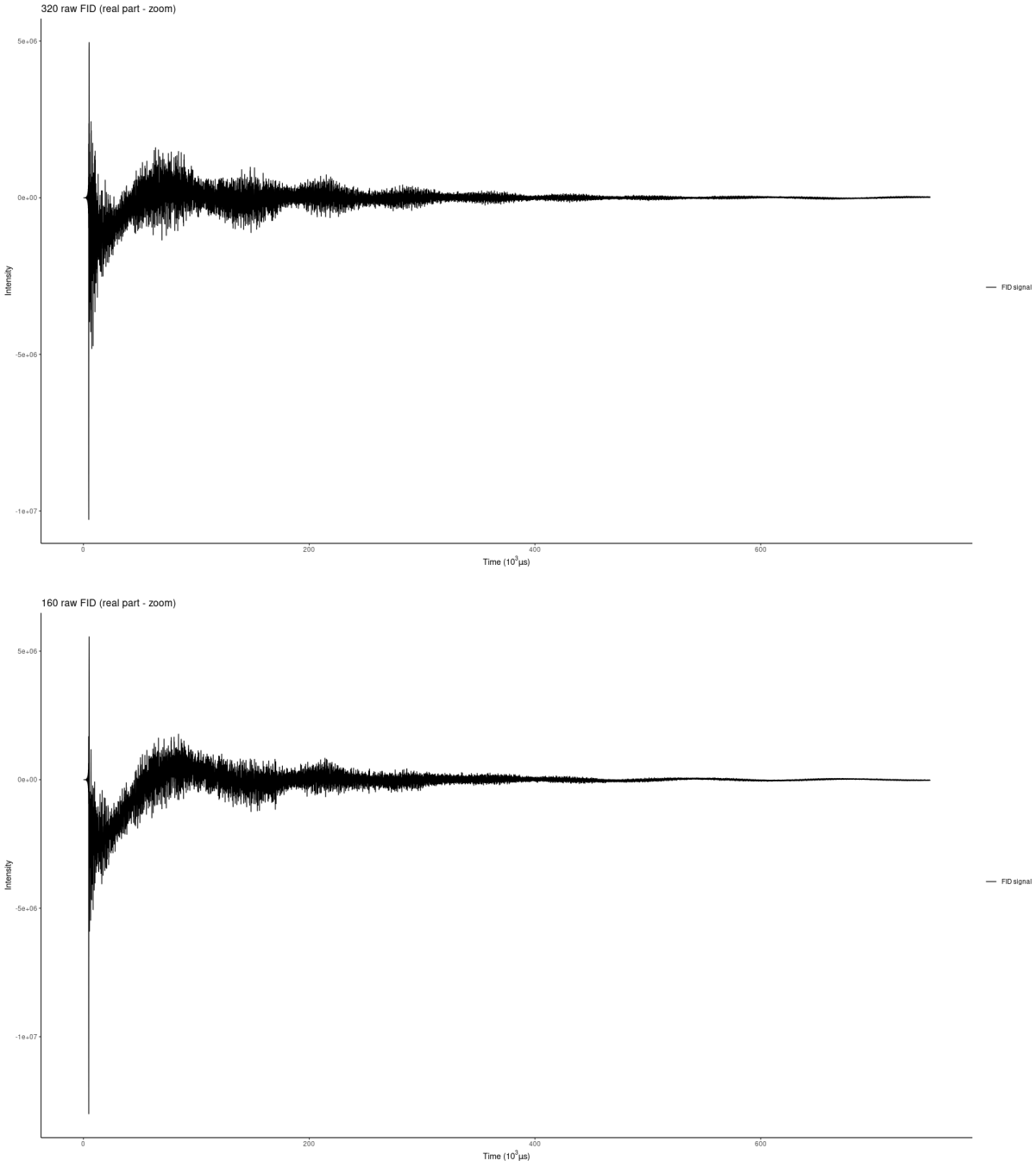


Figure 11. The comparison of FID pre and post-gdc removal FID for exemplary samples numbers 320 and 160.

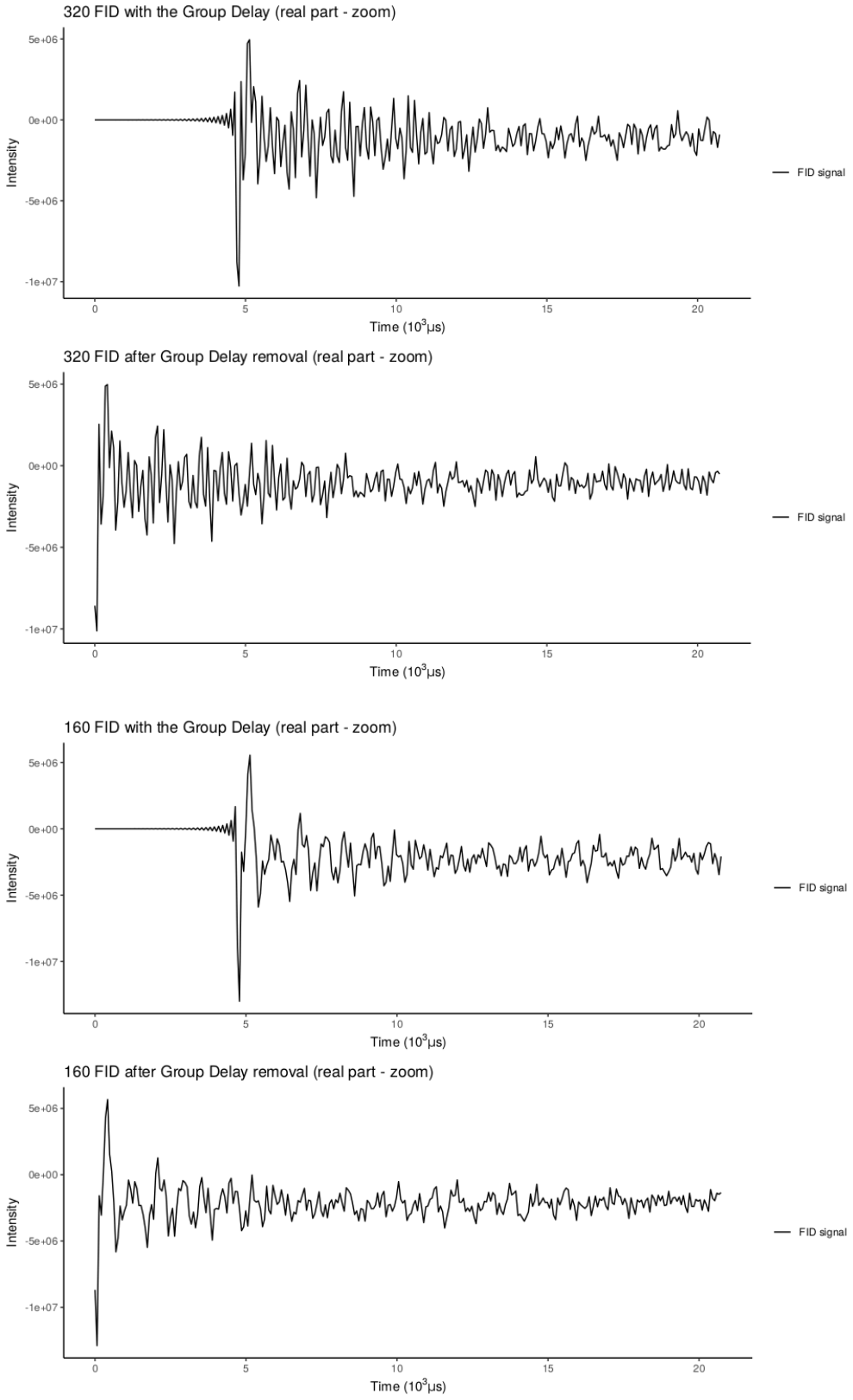


Figure 12. The comparison of FID pre and post-solvent removal FID for exemplary samples numbers 320 and 160.

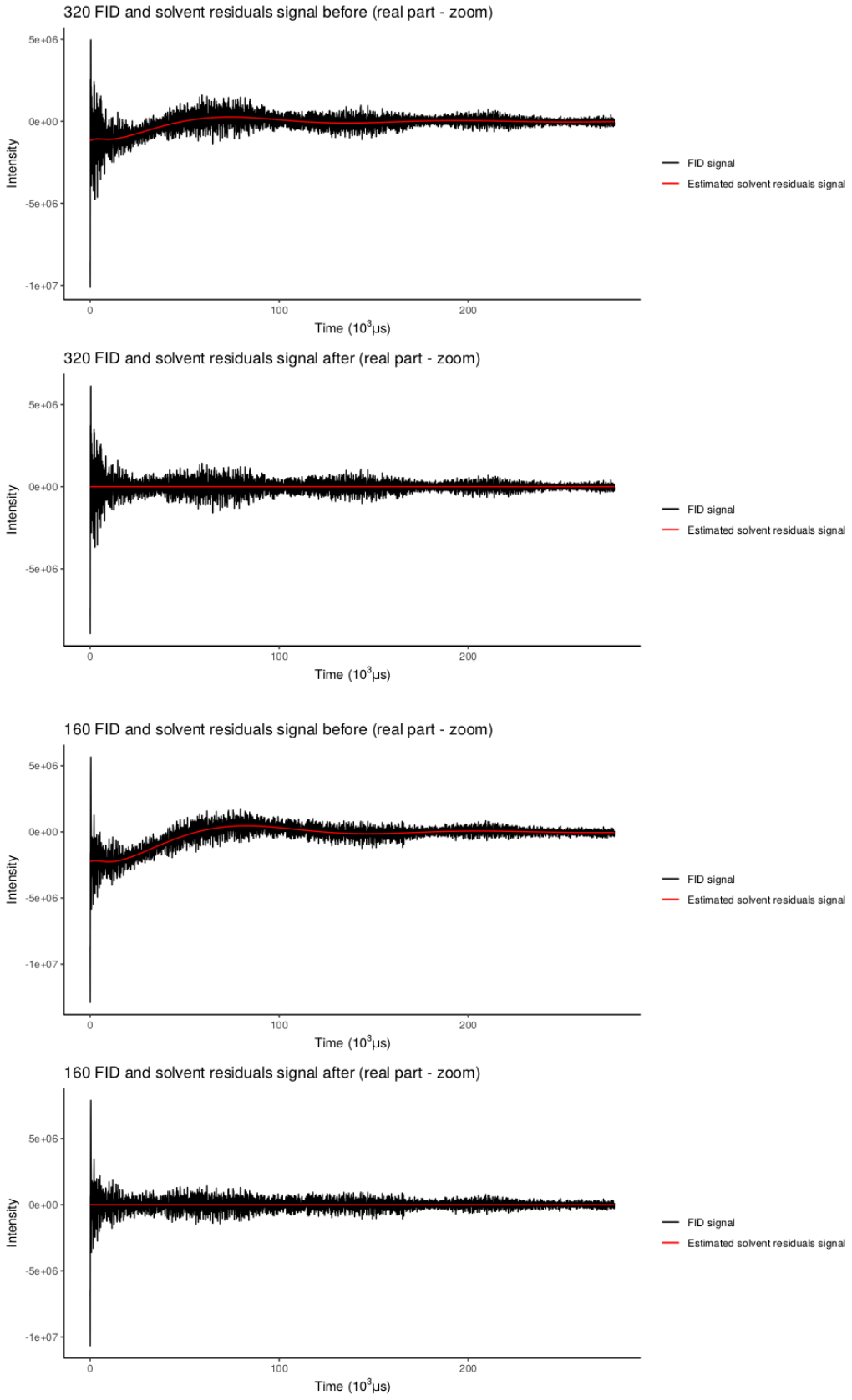


Figure 13. The comparison of post-apodization FID for exemplary samples numbers 320 and 160.

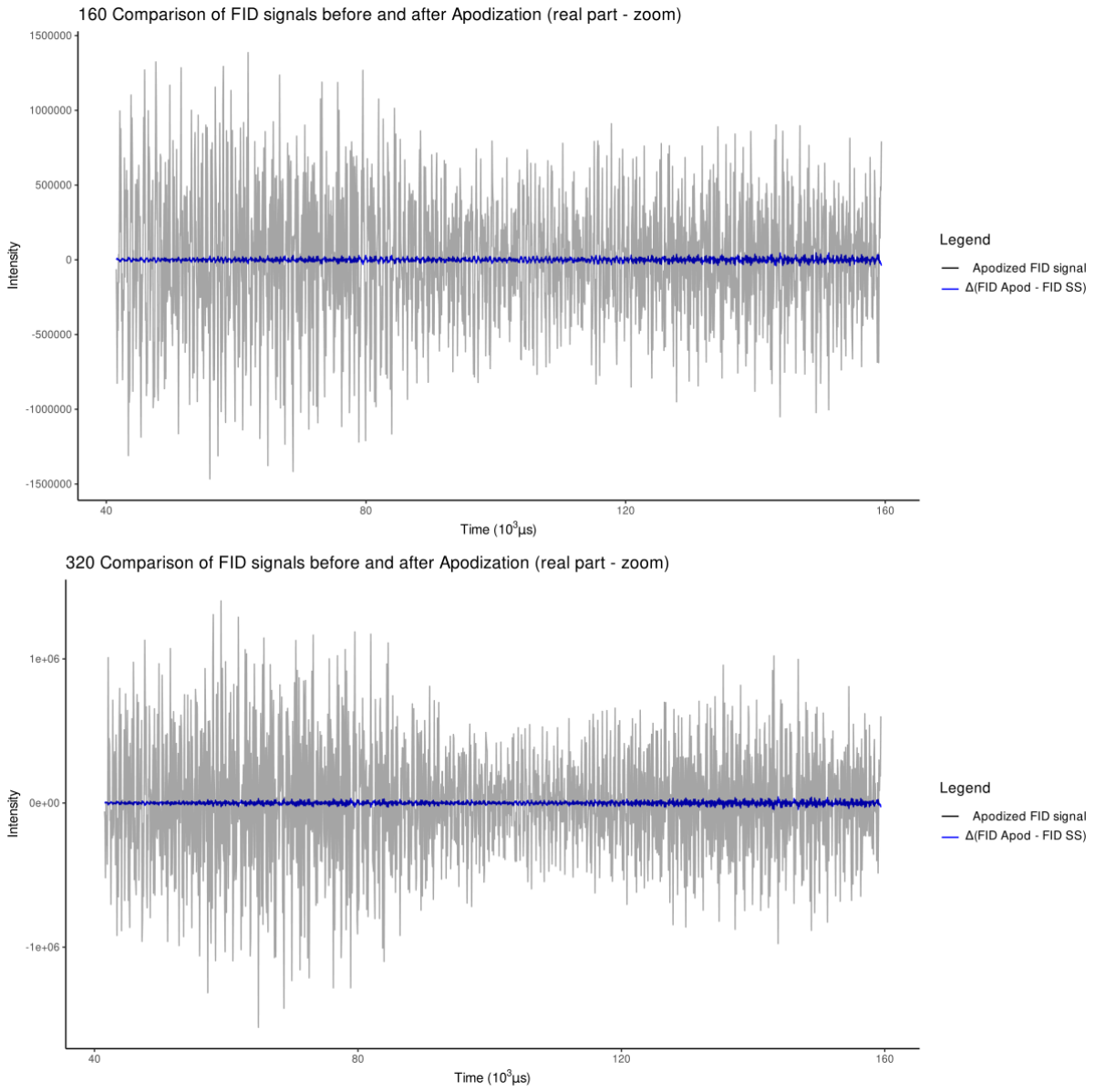
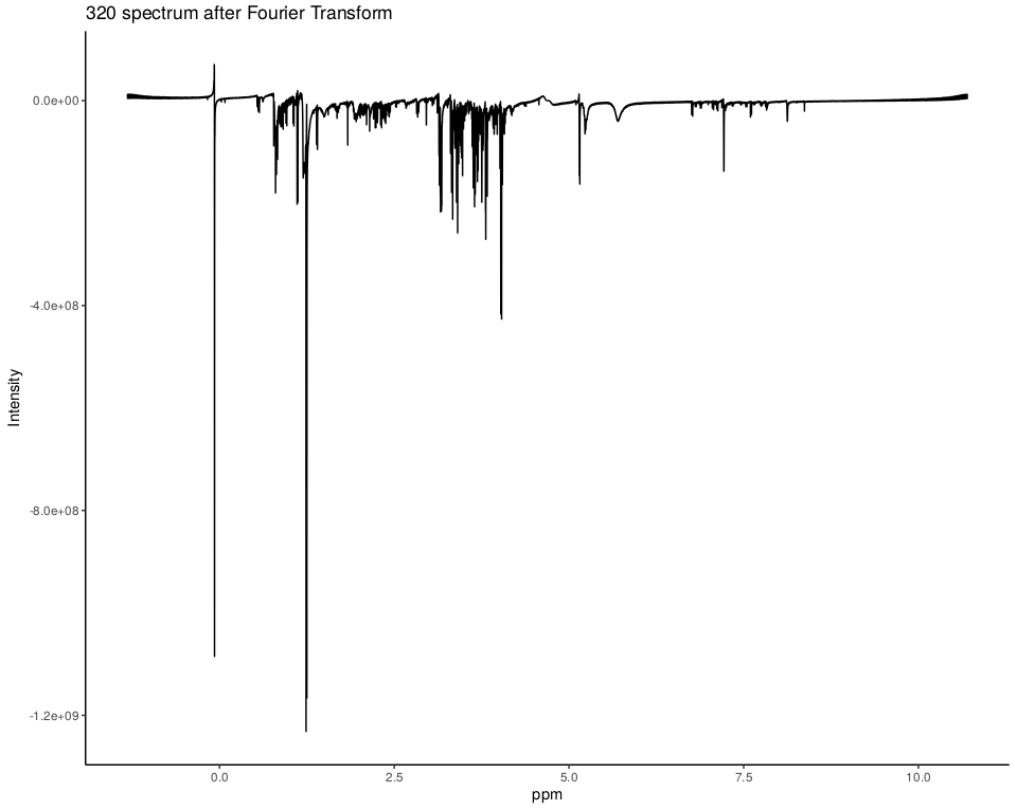


Figure 14 displays the spectrum after performing Fourier Transformation. The x-axis represents the chemical shift in parts per million, ranging from 0 to 10 ppm, while the y-axis shows the signal intensity. The spectrum reveals several peaks at different ppm values, indicating the presence of various components within the sample. However, the spectrum is plotted upside-down due to residual artifacts from a first-order phase shift, likely caused by a strong water signal. Incorrect phase adjustment results in an inverted spectrum where peaks appear negative instead of positive. Despite this, the transformation effectively converted time-domain data into the frequency-domain spectrum, with proper phase correction addressed in a subsequent module. **Figure 15** presents the spectrum after applying zero-order phase correction. Following this correction, the peaks in the spectrum are properly oriented with positive intensities, and the baseline is correctly aligned, enhancing the clarity and accuracy of the spectral data. This adjustment ensures that the peaks correspond to the correct chemical shifts, facilitating further analysis and interpretation of the sample's composition. **Figure 16** shows a spectral plot that represents data from spectra following internal referencing. The black lines denote the spectra, while the vertical red line signifies a reference peak position utilized during the internal referencing process. This red line, marked as "peak location" in the legend on the right side of the plot, is set at 0 ppm to ensure that the spectra are aligned for accurate comparison. **Figure 17** comprises two spectral plots, each depicting the results of baseline correction on the same set of spectra. In the top subplot the black lines represent the raw spectral data, while the red line indicates the baseline that has not been corrected. The red baseline is visible below the spectra, indicating unwanted background noise that can affect the accuracy of the peak measurements. The bottom subplot illustrates the same spectral data after the baseline correction process has been applied. In this subplot, the previously visible baseline noise has been removed, resulting in a cleaner spectrum where the peaks are more pronounced and clearly distinguishable. However, after automatic baseline correction, some values may still fall slightly below zero. This is because the method depends on the p_bc and lambda_bc parameters, as described in the "Spectral processing of raw 1D spectra and metabolites identification" section of the Methodology chapter. This issue is addressed in the subsequent module. **Figure 18** displays two spectral subplots: the top subplot shows the spectra before zeroing these values, while the bottom subplot shows the spectra after zeroing them. **Figure 19** consists of two spectral subplots, demonstrating the effects of window selection and bucketing. The top subplot shows the spectral data after applying window selection, highlighting specific regions of interest while excluding irrelevant areas. The bottom subplot presents the spectral data after applying the bucketing process, which groups the spectral data into discrete bins and averages the intensity values within each bin. This results in a smoother representation of the spectra, making it easier to identify and analyze trends. Comparing the two subplots, the spectra after bucketing (bottom subplot) appear less noisy and more consolidated than those after window selection (top subplot), facilitating a clearer interpretation of the data.

Figure 14. The spectrum after Fourier transformation for exemplary samples numbers 320 and 160.



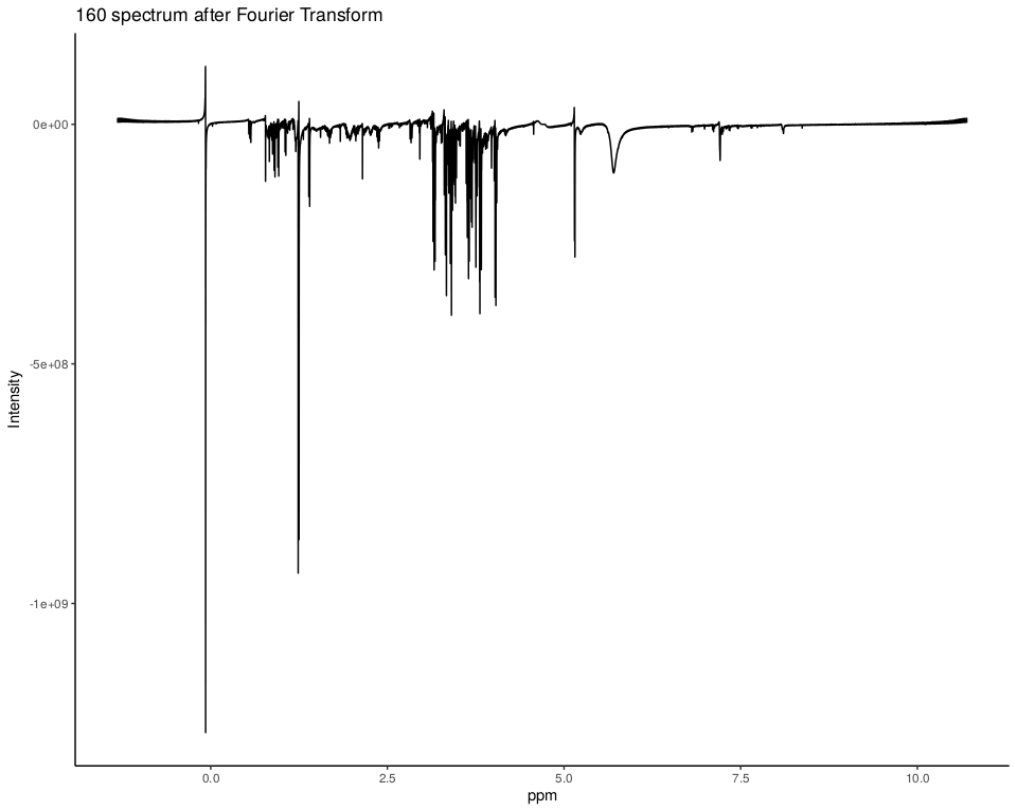
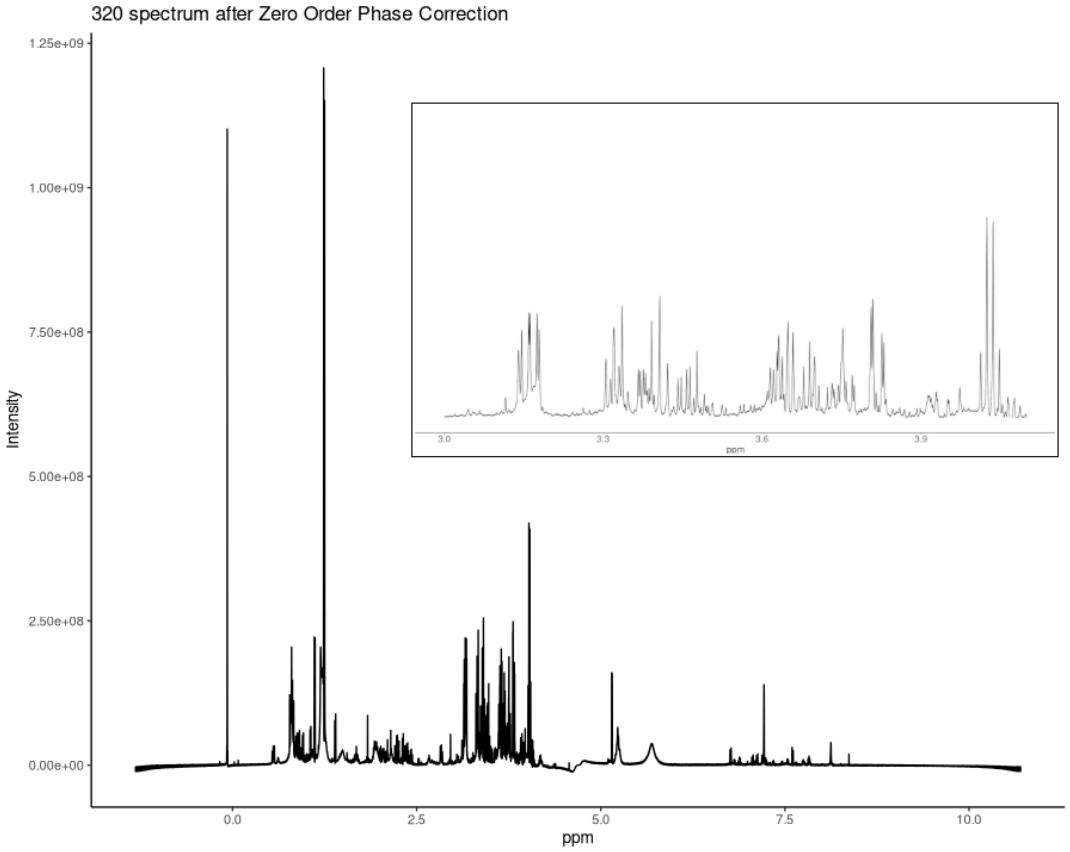


Figure 15. The spectrum after zero order phase correction for exemplary samples numbers 320 and 160.



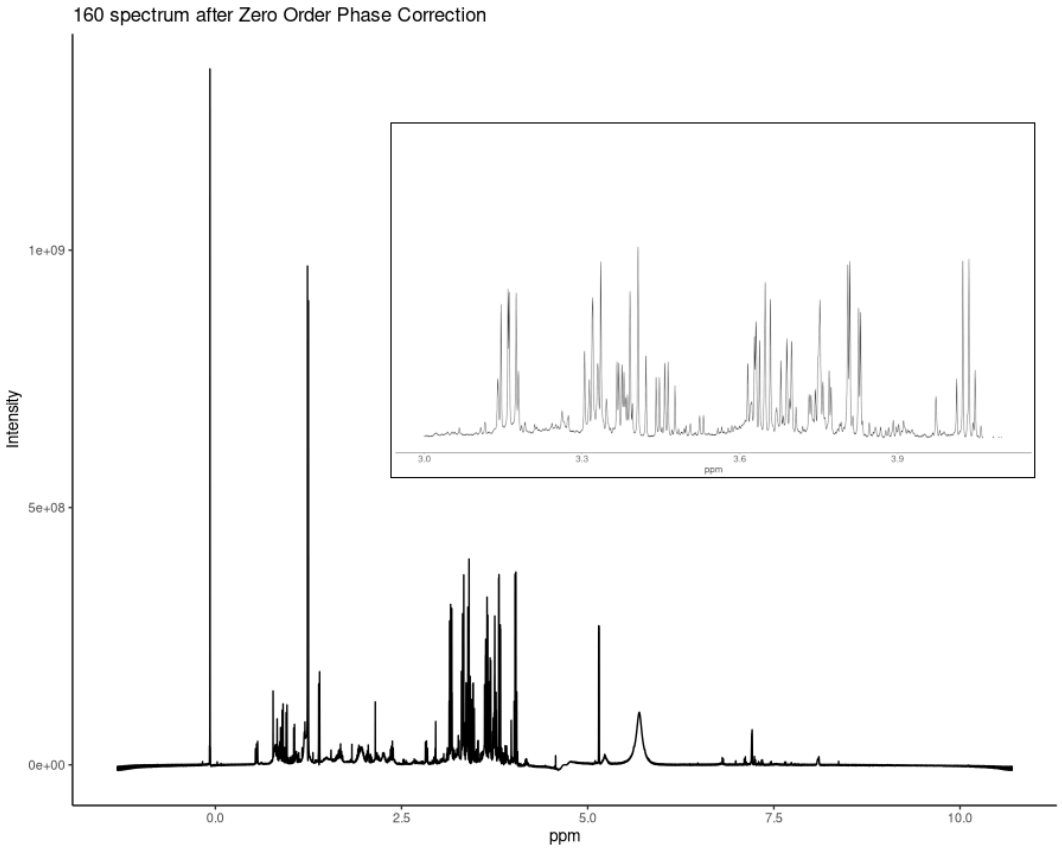


Figure 16. The spectrum after internal referencing for exemplary samples numbers 320 and 160.

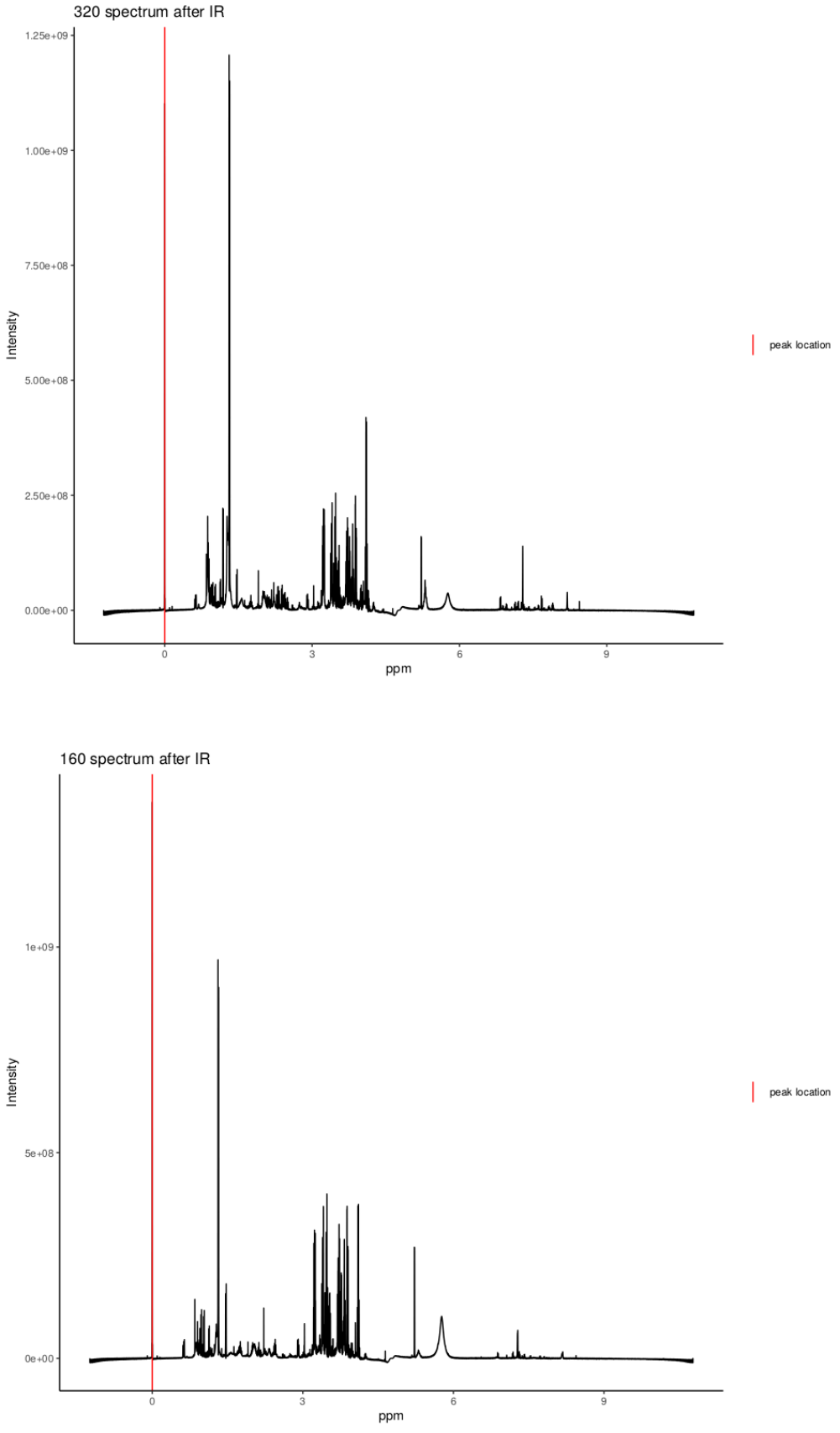


Figure 17. The comparison of spectrum pre and post baseline correction for exemplary samples numbers 320 and 160.

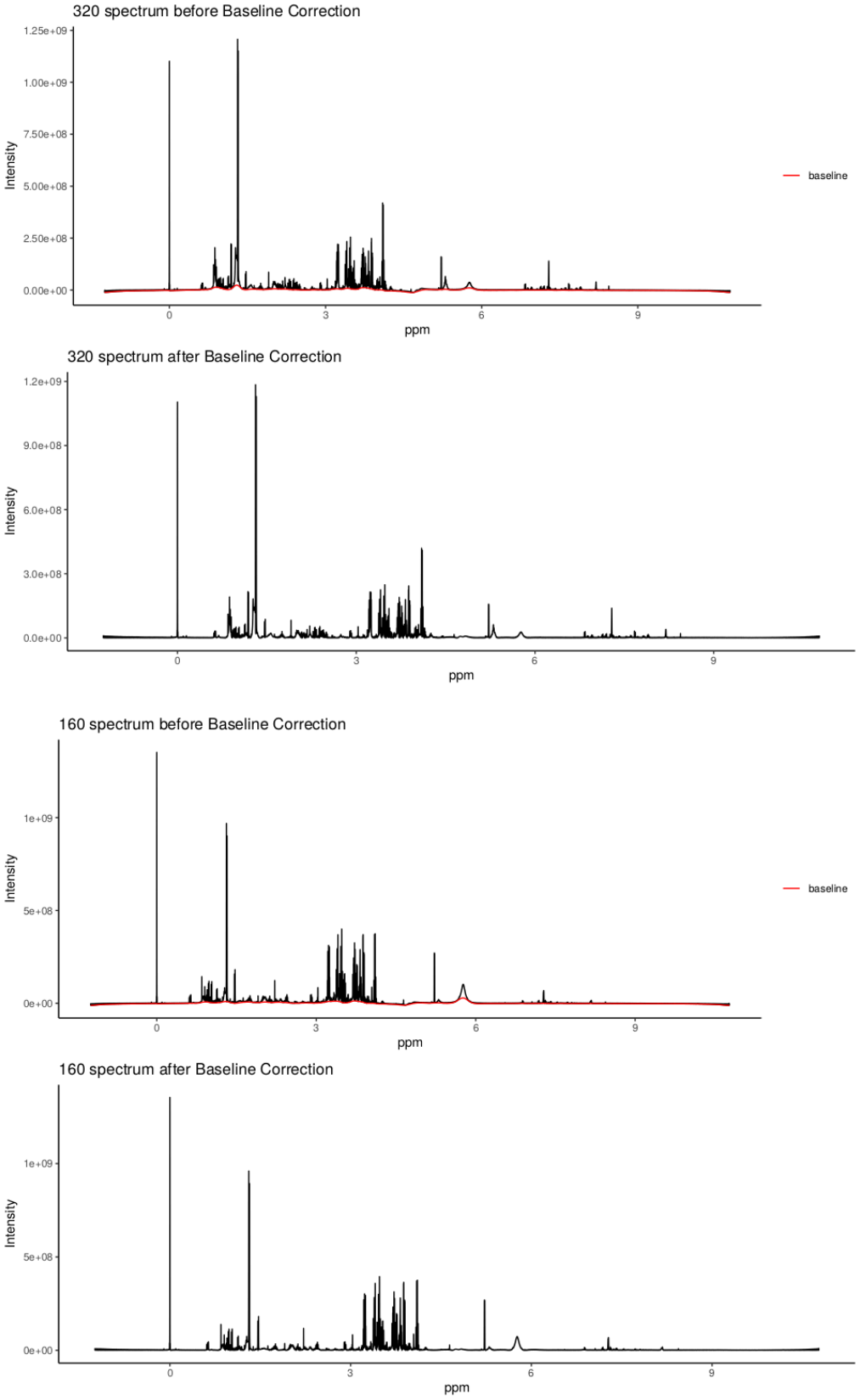
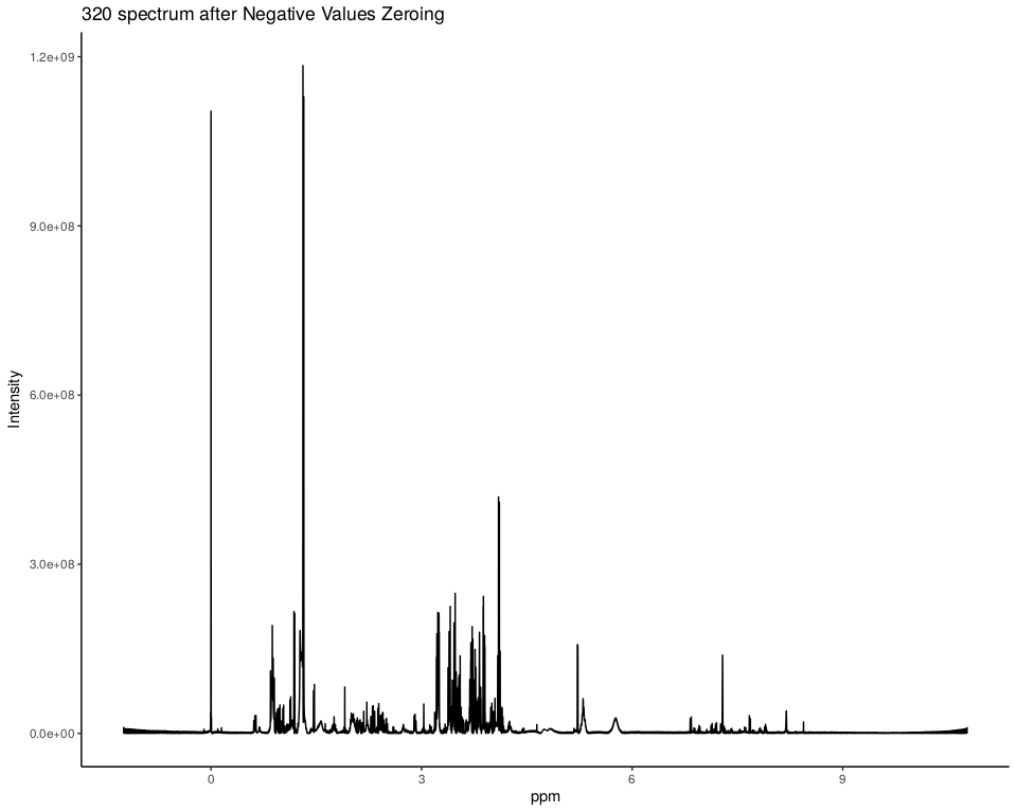


Figure 18. The spectrum after negative values zeroing for exemplary samples numbers 320 and 160.



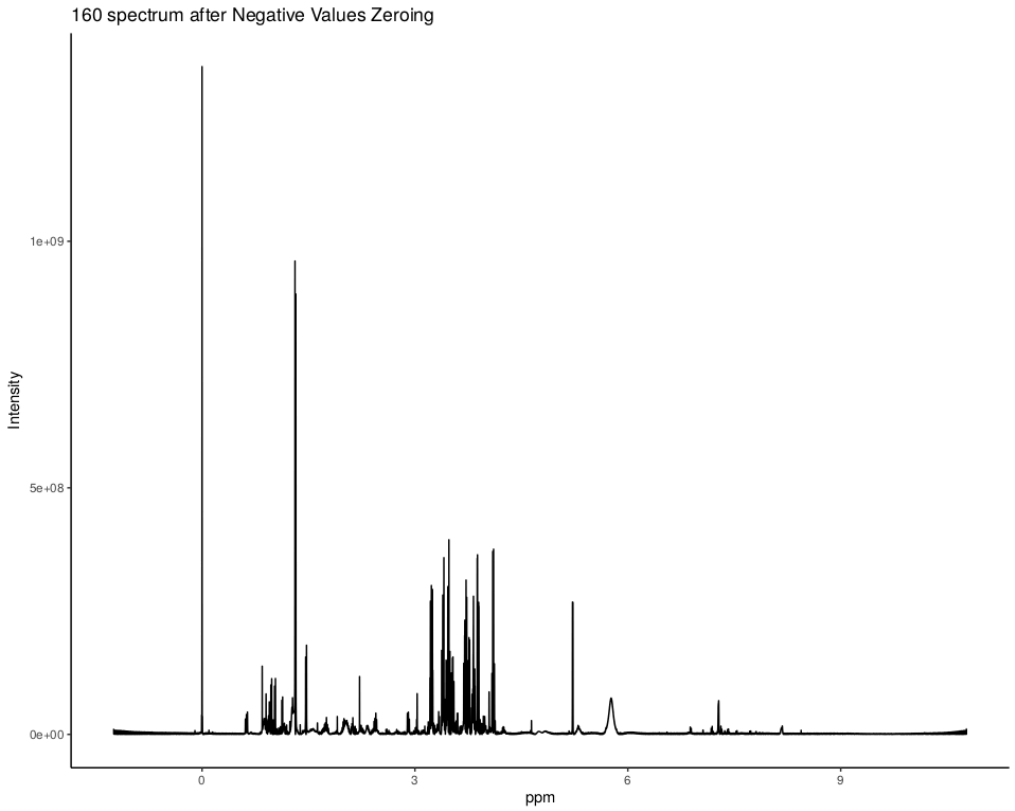


Figure 19. The comparison of spectrum pre and post bucketing for exemplary samples numbers 320 and 160.

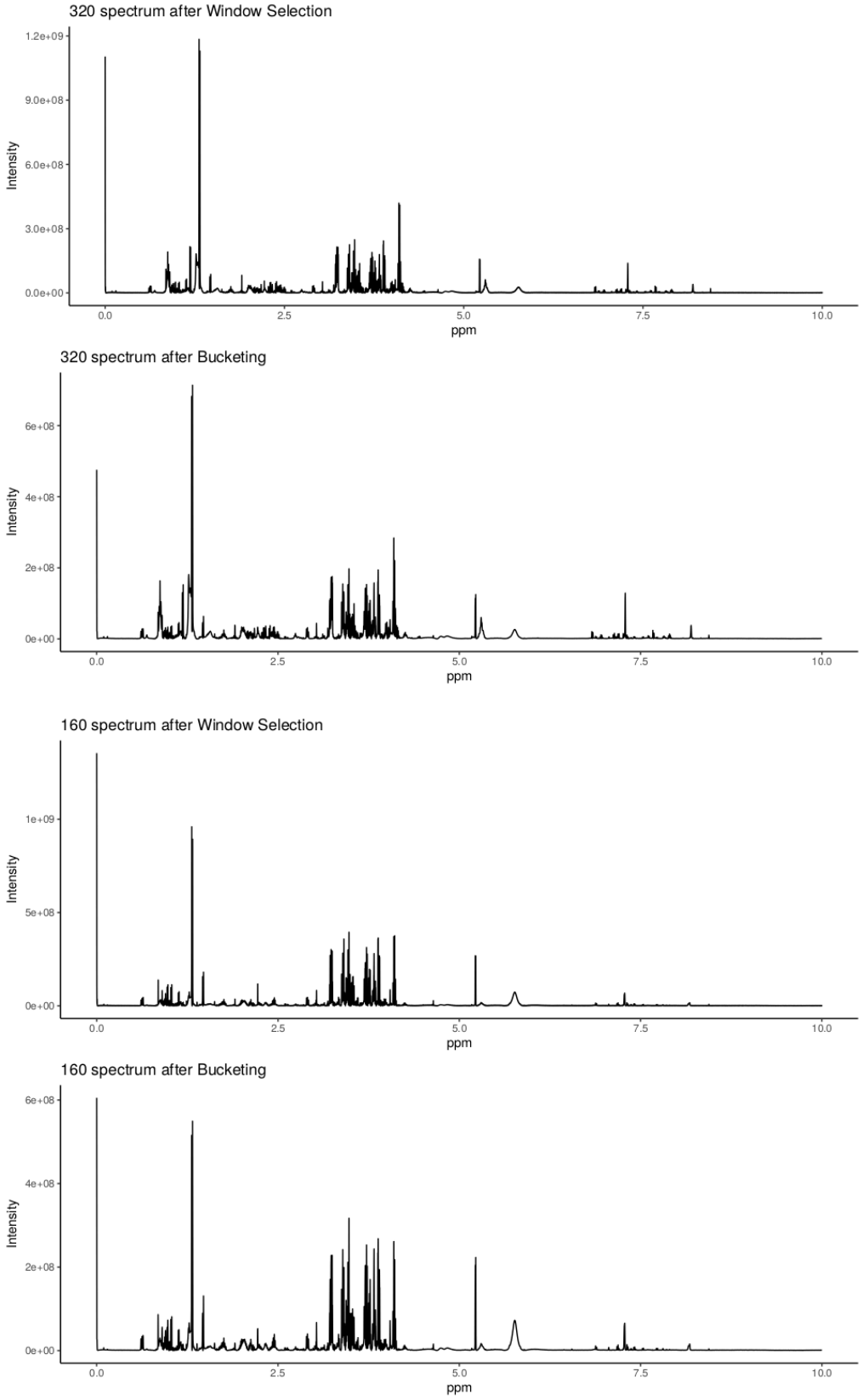


Figure 20 illustrates the intensity (y-axis) against the chemical shift in ppm (x-axis) after applying Probabilistic Quotient Normalization to the spectra. The black lines indicate the spectral data, ranging from 0 to 10 ppm on the x-axis. The normalization process has adjusted the intensity values, making the spectra more comparable across samples. This results in a more uniform baseline and consistent peak heights, allowing accurate comparisons and analyses of the spectral features. The peaks are more distinct and well-aligned, indicating that PQN has effectively mitigated the variability caused by different sample concentrations. To compare the quality of normalization, **Figures 21-25** present stacked spectra plots for all datasets after the complete preprocessing stage. Similar to other plots, these figures show intensity versus chemical shift in ppm, with values decreasing from 10 ppm to 0 ppm. The spectra are displayed with each spectrum stacked on top of the other, allowing for a clear visualization of multiple spectra simultaneously. The intensity values range from 0 to 8e+08. The plots use a rainbow color scheme to differentiate between individual spectra, highlighting the variability and patterns within the dataset. This visualization aids in identifying consistent peaks and patterns across the preprocessed spectra. **Figures 26-27** present a comparison of fully preprocessed spectra, highlighting the similarities in results obtained by the NASQQ in-house pipeline and the external methodology used by the authors of the original FD paper. The top spectrum, generated using the NASQQ pipeline, shows a clear representation of the data with distinct peaks and minimal baseline noise, spanning from 9 to 0 ppm. The peak at 0 ppm was deliberately removed as it represents the reference standard. The bottom spectrum was produced using the methodology described by the authors of the FD paper, also spanning from 9 to 0 ppm but presented in a slightly different style. Both spectra display similar features, offering a clean and refined representation of the spectral data. This comparison indicates that, despite potential differences in preprocessing techniques, the overall quality between the two methodologies is essentially equivalent, validating the quality of the NASQQ pipeline against the manually performed preprocessing.

Figure 20. The spectrum after PQN normalization for exemplary samples numbers 320 and 160.

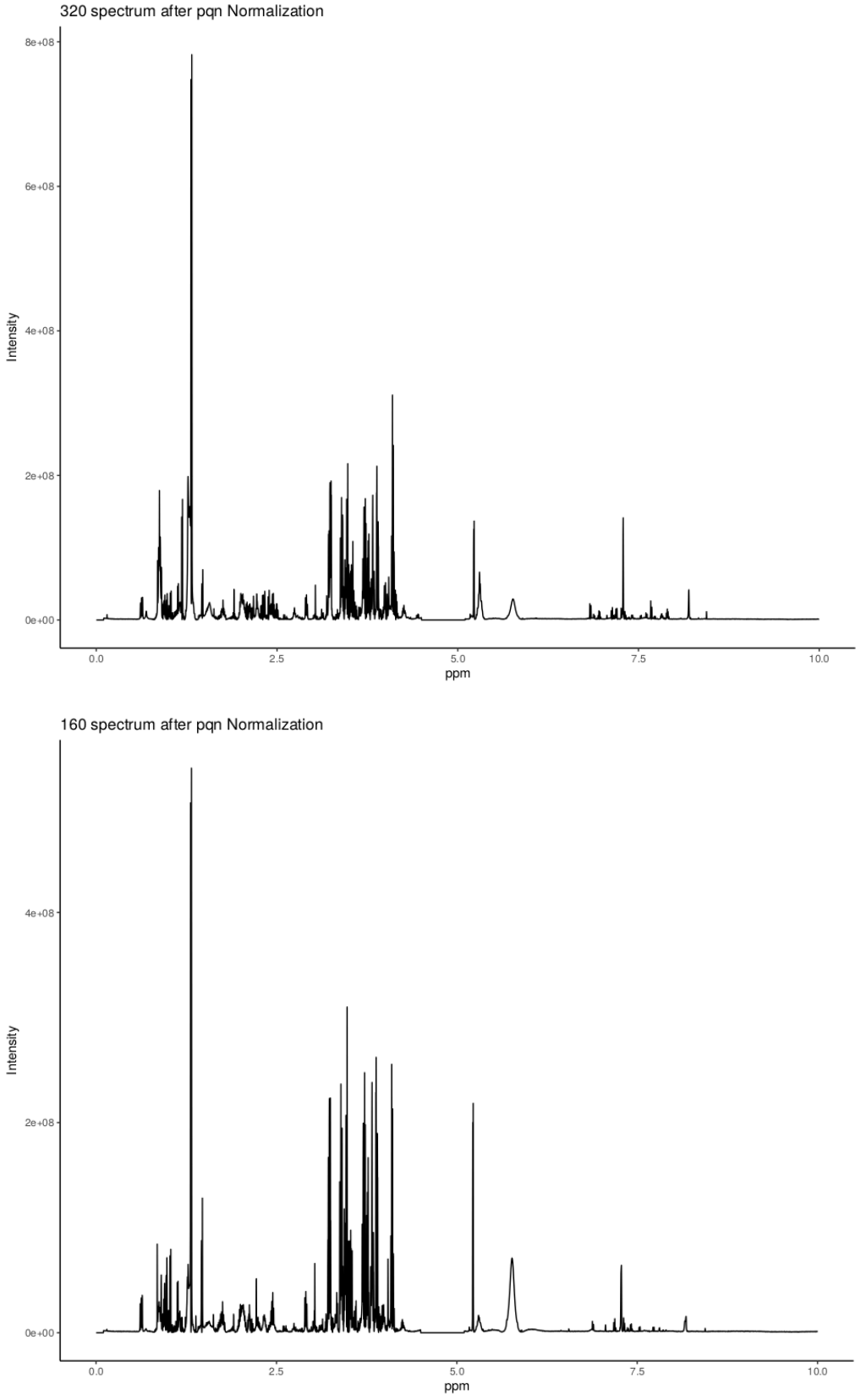


Figure 21. Stacked spectra of "*ACSC_HumanSerum_11_5_19*" dataset following complete preprocessing.

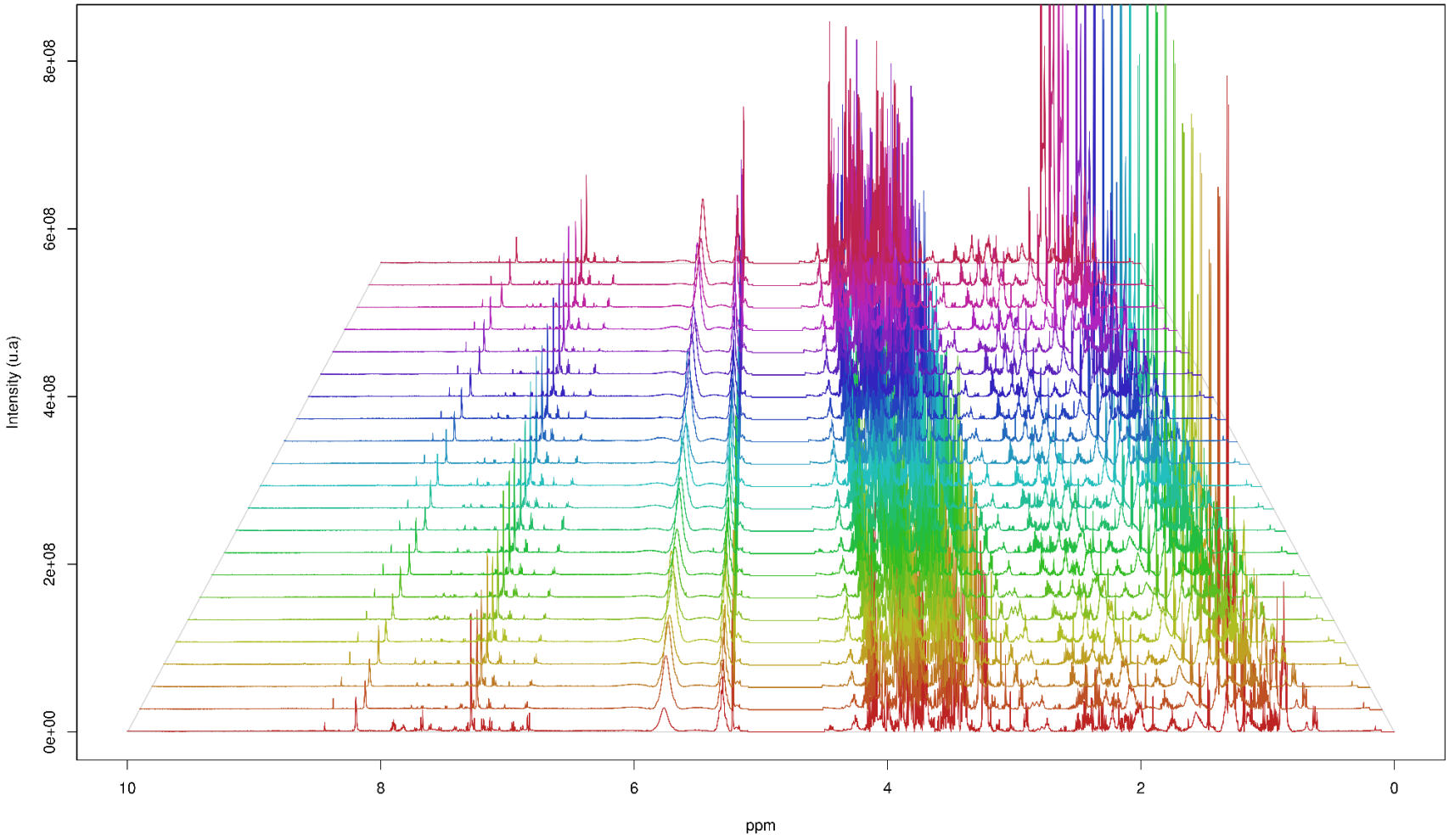


Figure 22. Stacked spectra of "*ACSC_HumanSerumFD_11_1_19*" dataset following complete preprocessing.

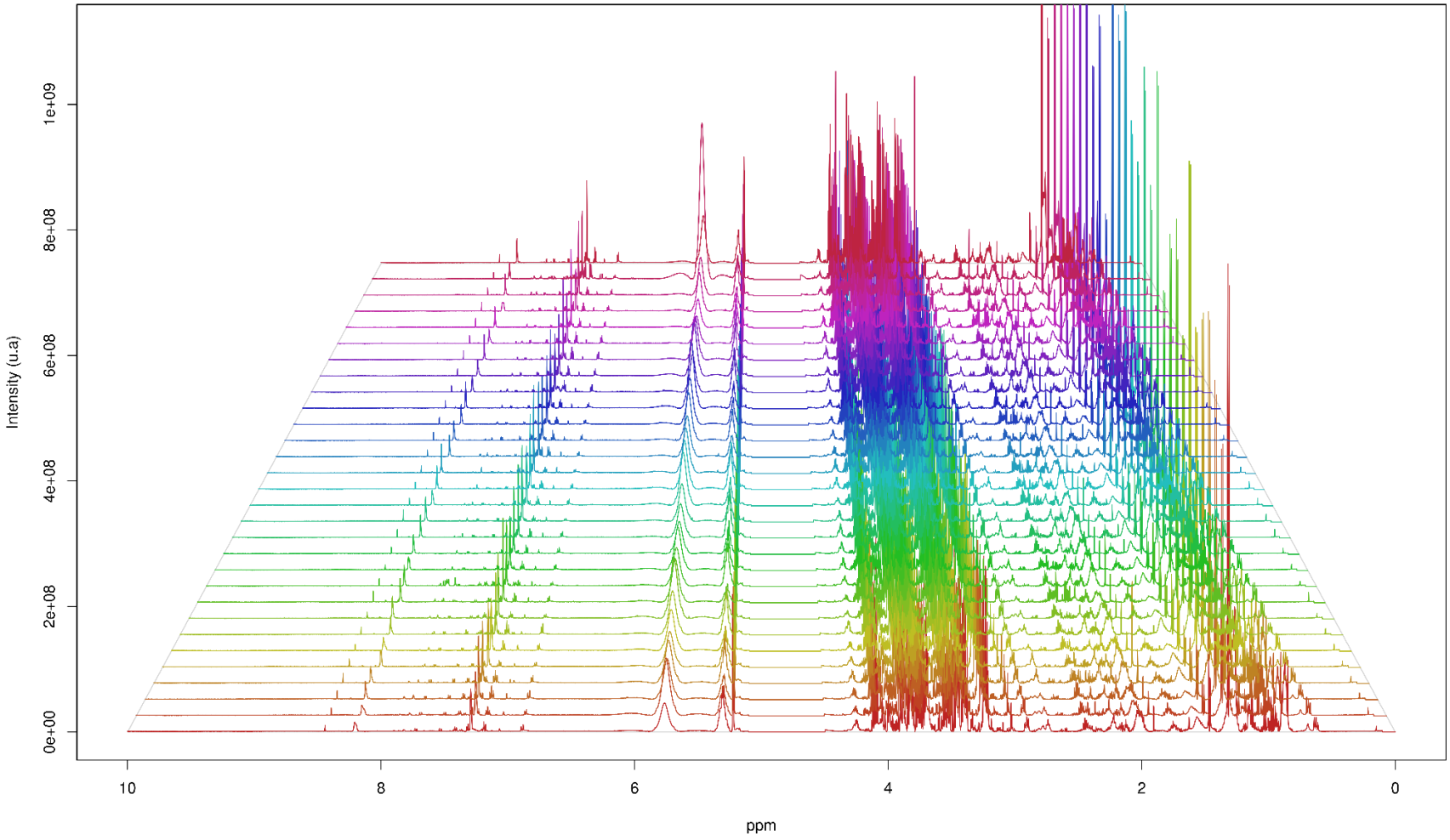


Figure 23. Stacked spectra of "*Human_Serum_2019_November_Shipment_Round_1*" dataset following complete preprocessing.

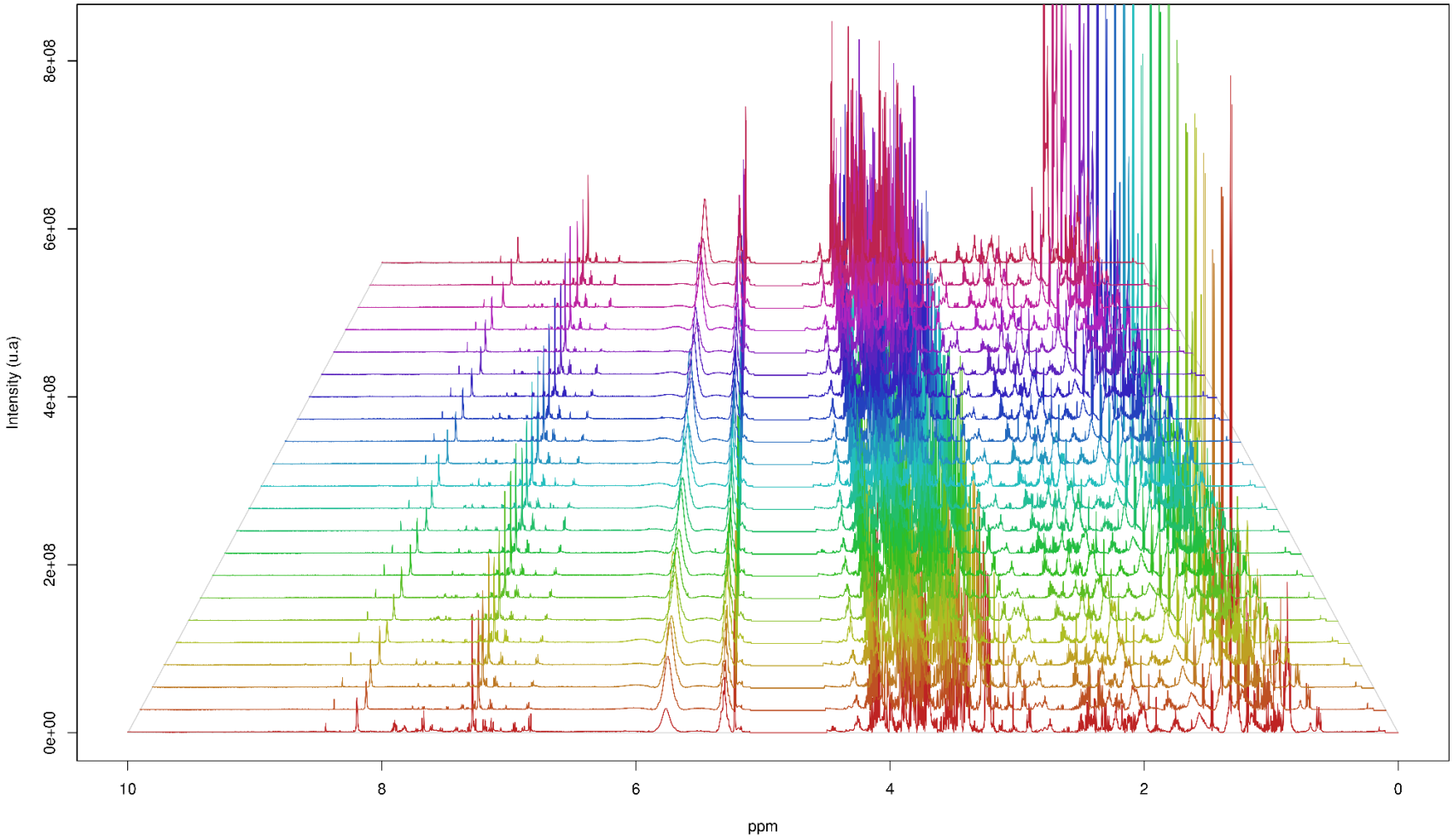


Figure 24. Stacked spectra of "*Human_Serum_2020_Feb_March_Shipments*" dataset following complete preprocessing.

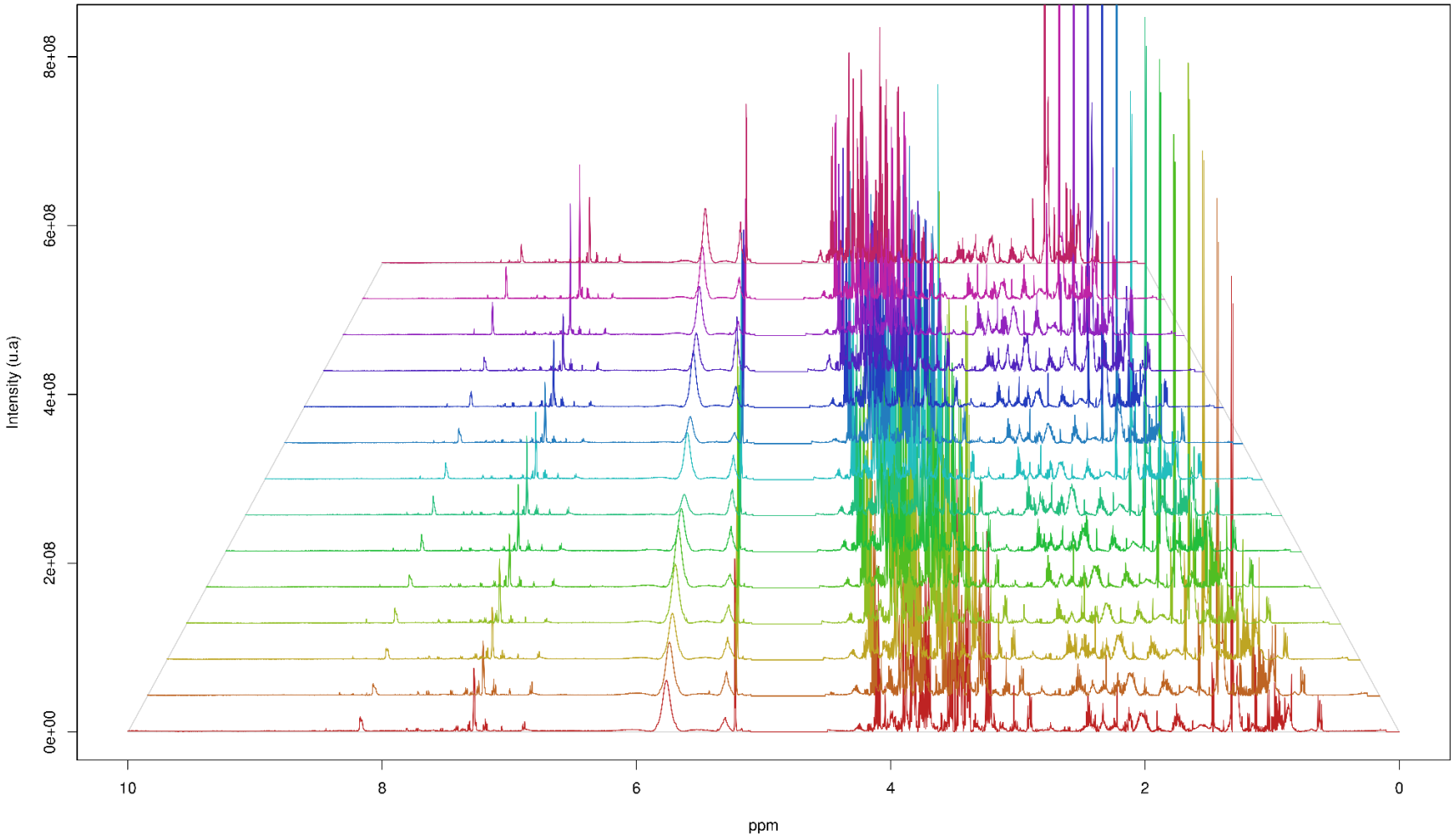


Figure 25. Stacked spectra of "*Human_Serum_2020_January_Shipment*" dataset following complete preprocessing.

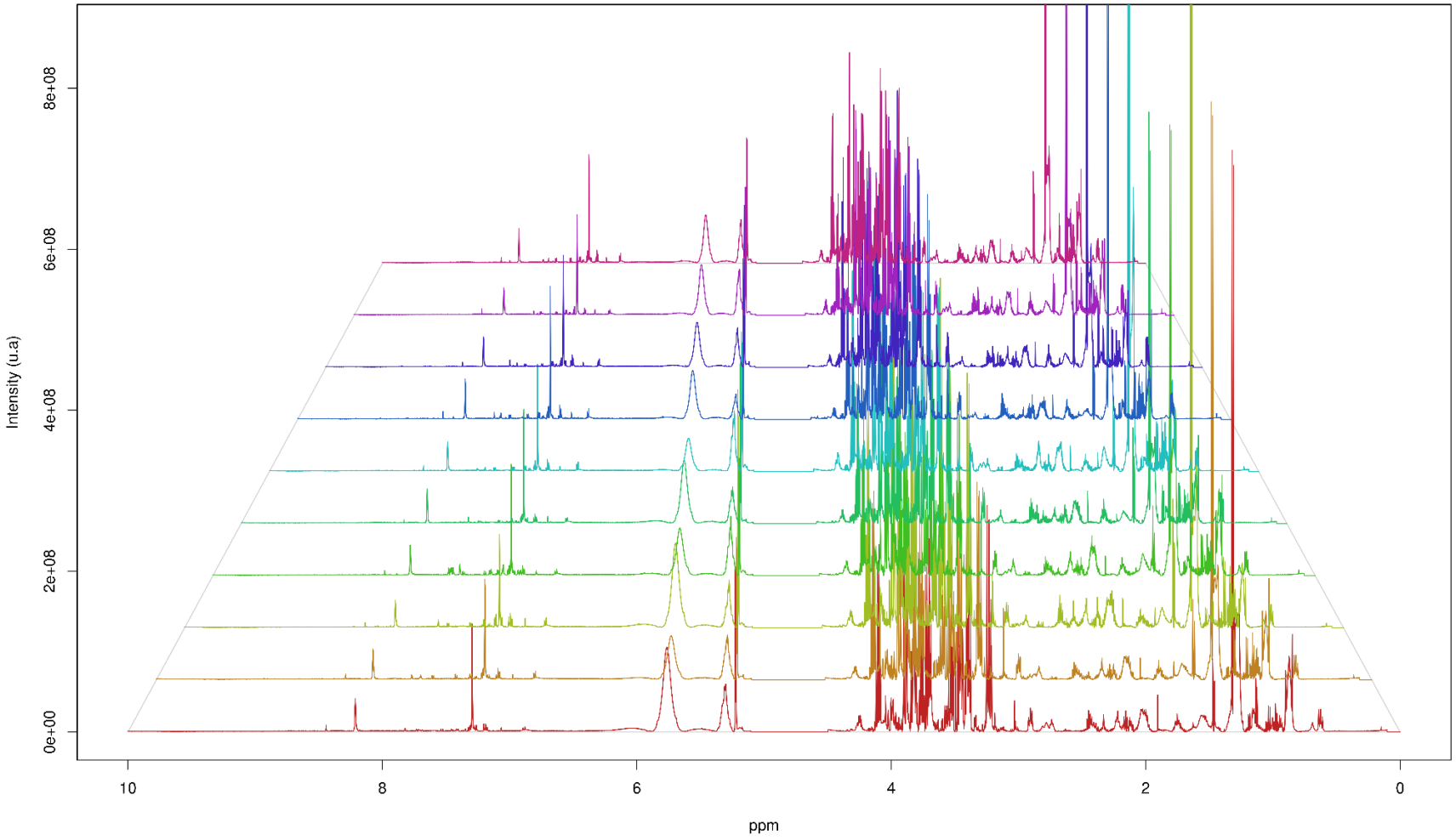


Figure 26. A comparison of preprocessed spectra between the original FD paper and NASQQ pipeline for exemplary sample number 320.

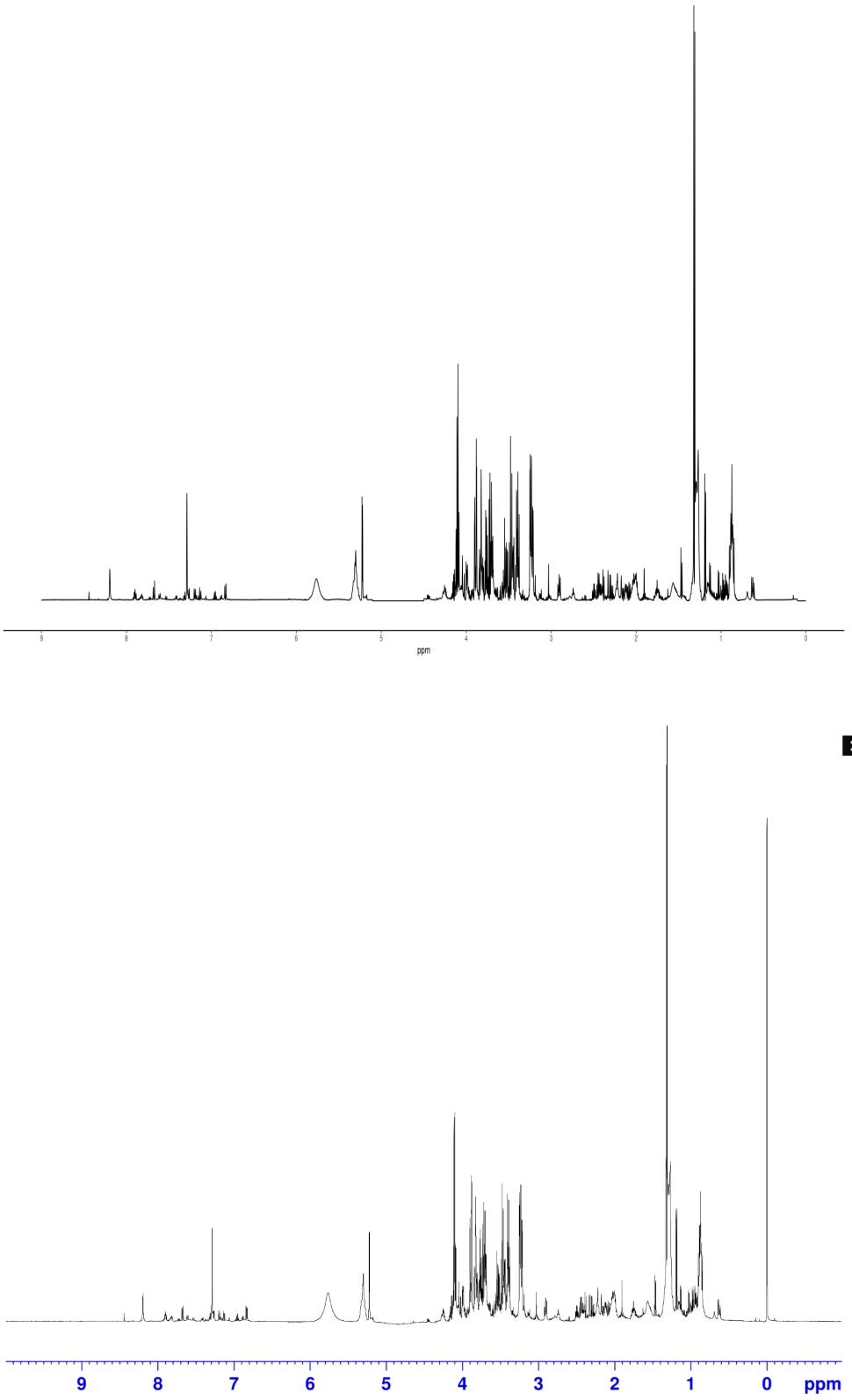
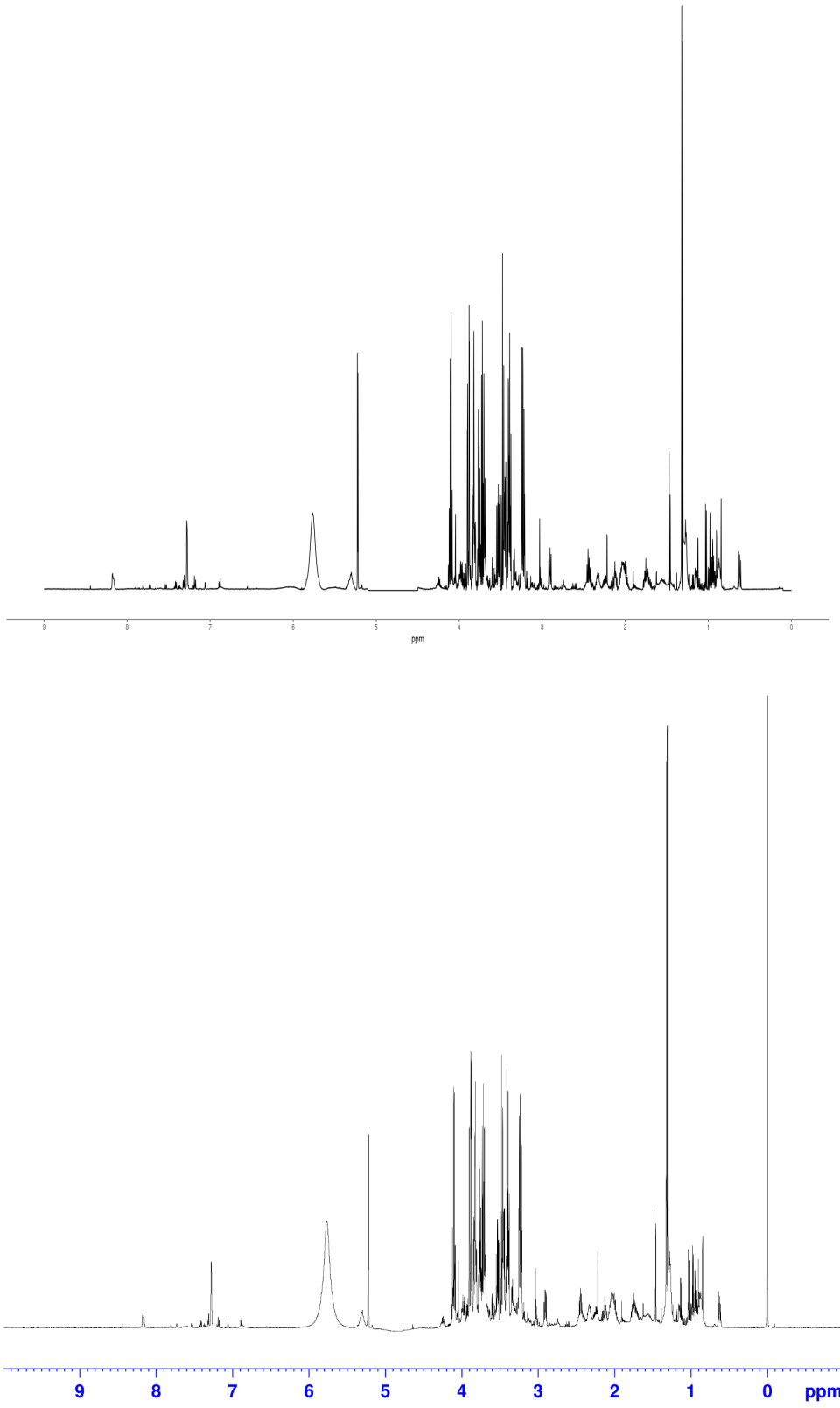


Figure 27. A comparison of preprocessed spectra between the original FD paper and NASQQ pipeline for exemplary sample number 160.



The results from the metabolite quantification module are presented in two main files: "*grouped_Spectrum_data_Quantified.rds*" and "*asics_normalized_metabolites.txt*". Due to the batch-wise nature of the analysis, metabolite quantification was also conducted in batches, resulting in five dataset-specific outcomes. A total of 140 metabolites were consistently quantified across all batches. Despite each batch highlighting slightly different most concentrated metabolites, D-glucose, D-fucose, and glyceric acid were frequently observed across all batches.

III.3 Results of univariate module tests

The entire flow of the "Data Analysis" stage followed the default pipeline parameters. After integrating metadata with all batches, the resulting table comprised 189 metabolites, 3 metadata columns ("patient_no", "batch", and "Disease.state"), and 101 samples. During feature processing, 7 metabolites were excluded, and the relative abundances were log1p transformed. The outcomes of the univariate module produced two files: *outliers.txt* and *univariate_analysis.csv*. The outliers identified by the LOF algorithm included samples numbers 180, 240, 50, 140, and 180. In terms of the univariate statistical tests, a normal distribution was observed for 18 metabolites, while 164 metabolites exhibited a non-normal distribution. Among the metabolites deemed significant, as indicated by a p-value below the threshold of 0.05, only D-fucose showed significance in the T-test. In contrast, the Mann-Whitney U test identified several significant metabolites, including uracil, β-hydroxyisovaleric acid, L-aspartate, hypoxanthine, and 5-amino valeric acid. Despite D-fucose showing a significant p-value of 0.0212, its FDR value of 0.9727 suggests a high likelihood of false discoveries when adjusting for multiple testing. Similarly, for the other metabolites in the T-test, their significance is questionable after applying multiple testing corrections. In the Mann-Whitney U test, while several metabolites exhibit p-values below 0.05, only uracil remains significantly different in abundance between FD patients and their relatives after accounting for multiple testing, with an FDR value of 0.1210. Detailed results of the univariate tests are available in **Supplementary Table 1** and **Supplementary Table 2**.

III.4 Assessment of metabolites classified by machine learning models

The EDA module generated multiple plots to characterize samples and features. Since a batch column name was provided, additional plots were created to investigate potential batch effects on the data. **Figure 28** shows a series of boxplots comparing the distribution of randomly chosen metabolites' abundances across all five available batches. Randomization is handled by setting a seed to ensure that resulting metabolites are consistent across different runs. Each boxplot represents a random, yet consistently selected 10 metabolites, with the x-axis listing the metabolite names and the y-axis showing their relative abundances. The central line within each boxplot indicates the median value, while the boxes represent the interquartile range (IQR). Whiskers extend

to the smallest and largest values within 1.5 times the IQR from the quartiles, and outliers are displayed as individual points beyond the whiskers. Notable observations include the varying distributions and presence of outliers among the batches, with some metabolites such as L-glutamic acid and glycerol showing significant differences in variability across the batches. The same box plot comparisons were performed for classification groups, specifically disease status (see **Figure 29**). Although there are no major changes visible, the median for both statuses varies slightly across metabolite abundances.

Another dataset characteristic depiction is presented in **Figures 30** and **31** in the form of density plots. The purpose of these density plots is to visualize the distribution and variability of metabolite abundances across different batches and disease statuses. Each subplot shows the density distribution of a specific metabolite, with the x-axis representing the metabolite values and the y-axis representing the density. In **Figure 30**, the colors in each plot correspond to different batches as indicated in the legend. Notable observations include varying peak heights and spreads across the batches for metabolites such as L-aspartate, lactate, 2-hydroxyphenylacetic acid, uracil, argininosuccinic acid, D-galactose, L-glutamic acid, methylamine, glycerol, and 3-methyl-L-histidine. Differences in the distributions highlight batch effects on metabolite levels, with some metabolites showing significant variability between batches. Interestingly, in **Figure 31**, it is visible that relatives have higher density in multiple metabolite abundances, indicating potential differences in metabolite profiles between FD patients and their healthy relatives.

Figure 28. Box plot comparison of 10 randomly selected metabolites' log1p abundances across batches.

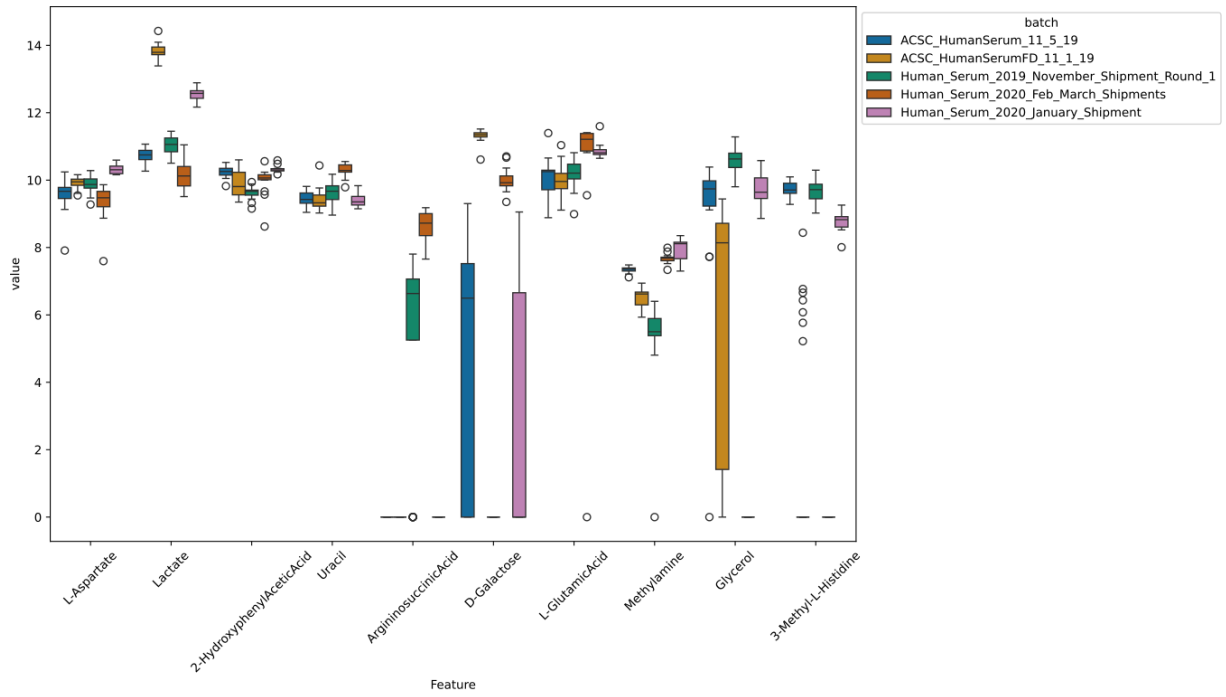


Figure 29. Box plot comparison of 10 randomly selected metabolites' log1p abundances across disease status.

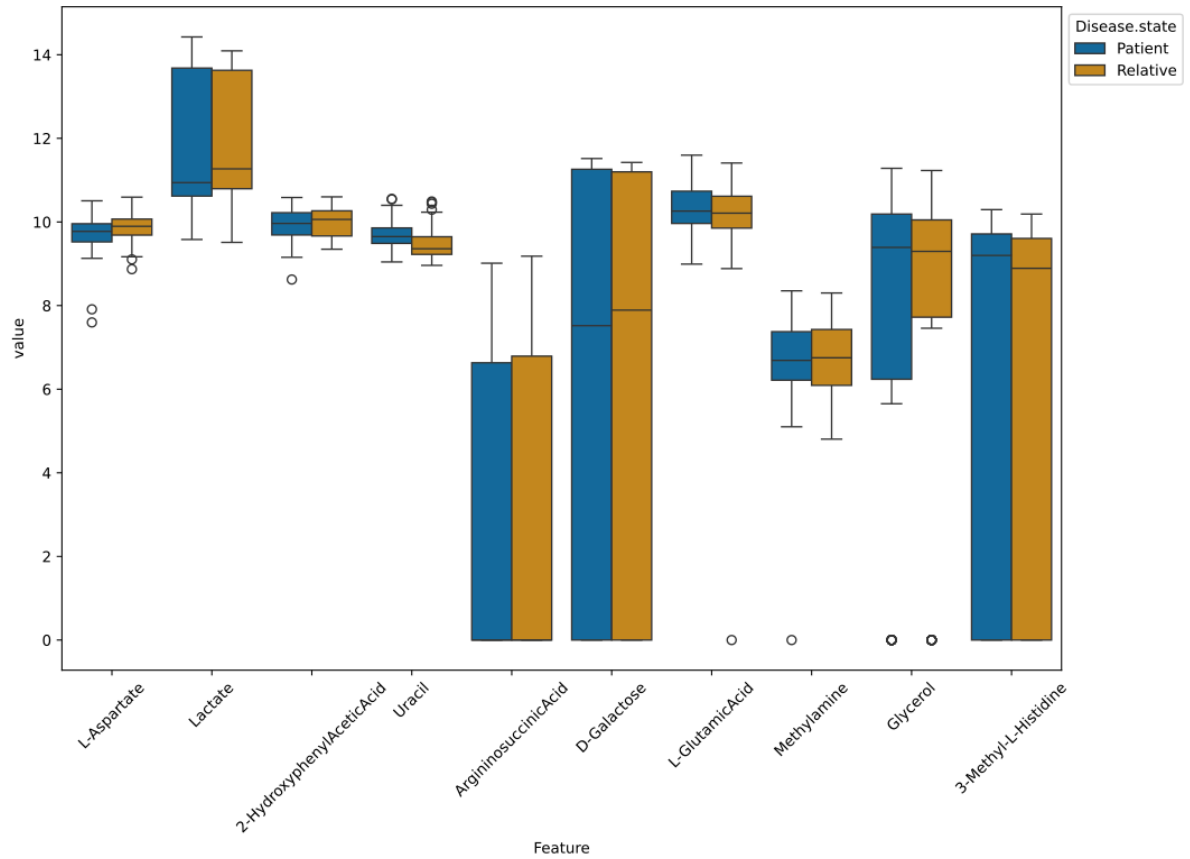


Figure 30. Density plot comparison of 10 randomly selected metabolites' log1p abundances across batches.

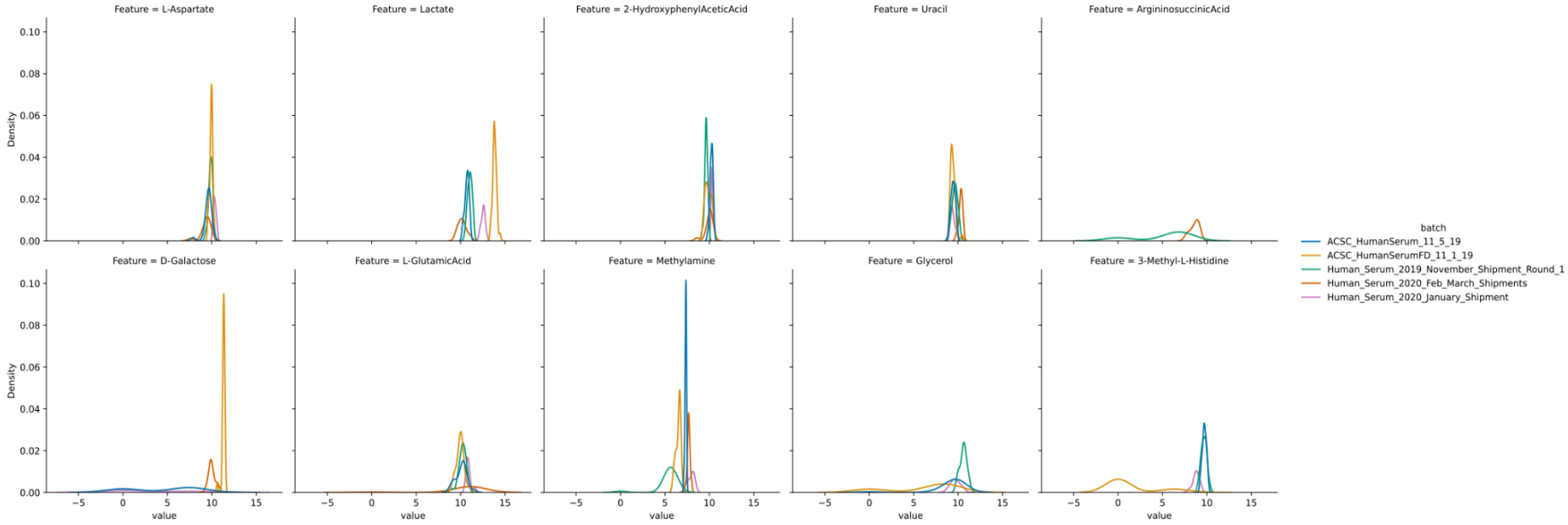


Figure 31. Density plot comparison of 10 randomly selected metabolites' log1p abundances across disease status.

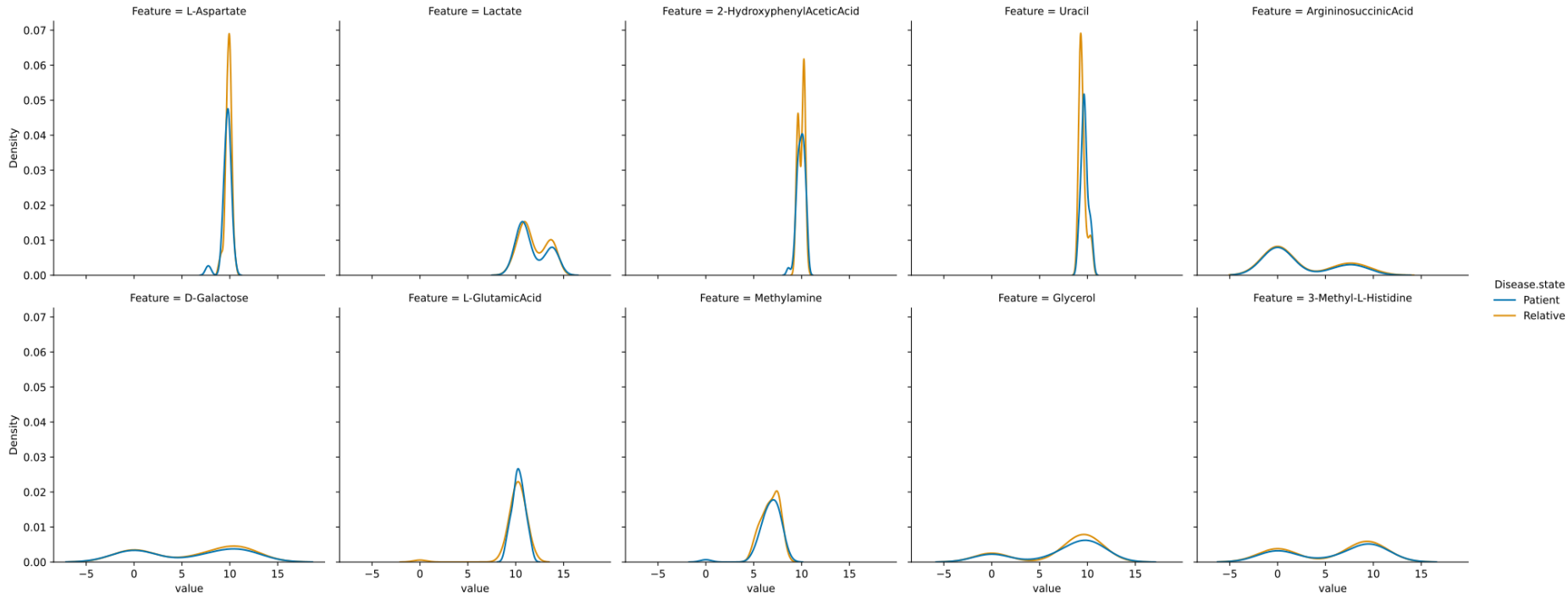


Figure 32 displays a heatmap featuring hierarchical clustering applied to both rows and columns, providing a visual representation of the relationship and clustering of metabolites. The color gradient, ranging from blue to red on the left side, signifies negative to positive correlation values respectively. Darker shades denote stronger correlations. Dendrograms situated at the top and left side of the heatmap illustrate the hierarchical clustering of metabolites, with closer proximity in the dendrogram indicating greater similarity in correlation patterns. The axes list the names of the metabolites, and each cell within the heatmap denotes the Pearson correlation between the corresponding metabolites of its row and column. The diagonal line represents the perfect correlation of each metabolite with itself. Clusters of metabolites exhibiting similar correlation patterns are evident in the heatmap. Strong correlations are highlighted by darker red or blue shades. Hierarchical clustering aids in identifying groups of metabolites with akin correlation profiles. The top-left cluster comprises metabolites such as 2-propanol, trans-aconitic acid, methylamine, L-valine, and 2-oxobutyrate, demonstrating strong positive correlations. Their close grouping in the dendrogram suggests shared biochemical pathways or co-regulation. A distinct middle cluster consists of metabolites like malonate, succinate, pantothenic acid, glycerophosphocholine, and ethylmalonic acid, displaying varying degrees of positive correlations. The bottom cluster encompasses metabolites like glutaconic acid, argininosuccinic acid, L-carnosine, L-proline, and phenethylamine, showing strong positive correlations within the cluster. Their tight clustering in the dendrogram implies involvement in closely related metabolic processes. Prominent strong positive correlations are observed in the top-left and middle clusters, while lighter colors, particularly around the diagonal, indicate weak or no correlation between certain metabolite pairs.

Figure 32. Heatmap of metabolites' Pearson correlations alongside hierarchical clustering.

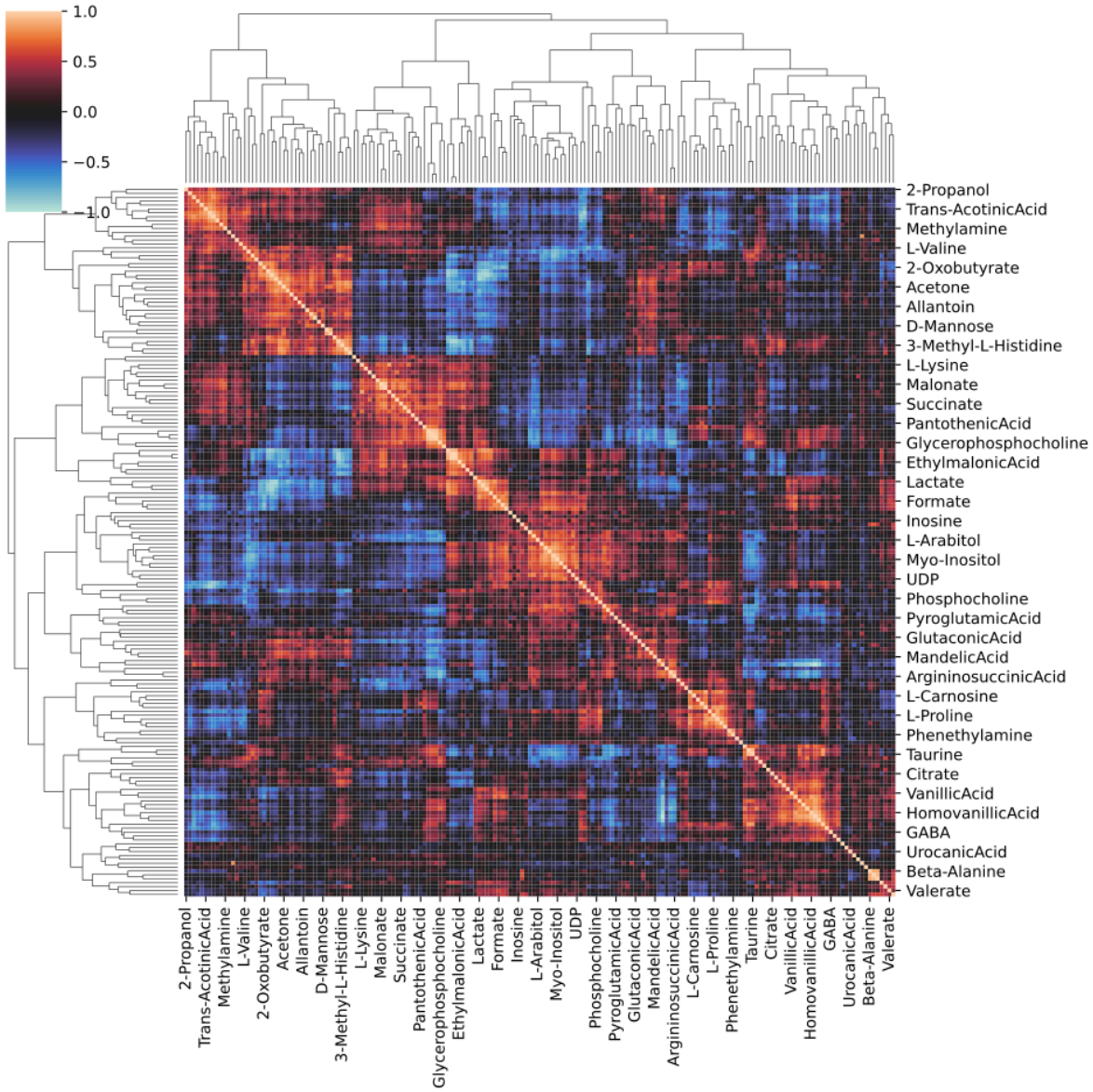
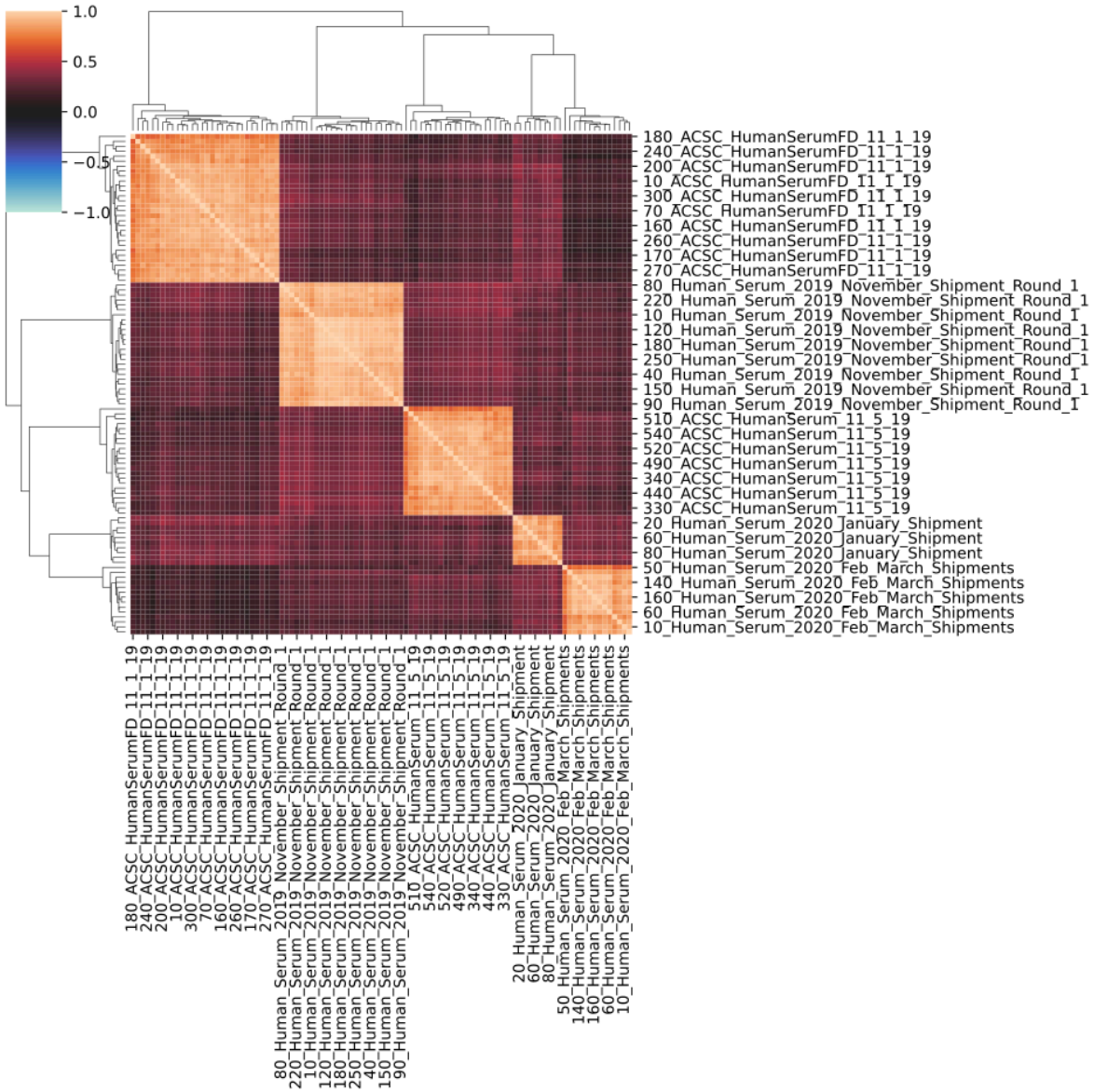


Figure 33 presents a heatmap illustrating the correlation matrix of samples, organized by batch, with hierarchical clustering applied to both rows and columns. The color spectrum spans from blue (denoting strong negative correlations) to red (indicating strong positive correlations), with intermediate shades indicating varying degrees of correlation. Samples are annotated with batch-specific details, such as "*180_ACSC_HumanSerumFD_11_1_19*" or "*510_Human_Serum_2020_January_Shipment*". Hierarchical clustering reveals that samples from the same batch tend to cluster together, displaying higher intra-batch correlations. For instance, samples from the "*ACSC_HumanSerumFD_11_1_19*" batch form a cohesive cluster, indicating robust correlations within this batch. Disparities in correlation patterns between different batches and shipments are noticeable. While some batches exhibit more uniform clustering (e.g. "*Human_Serum_2020_Feb_March_Shipments*"), others demonstrate greater variability, hinting at potential batch effects or discrepancies in sample processing and handling across shipments. The branching structures illustrate the hierarchical relationships between samples, with closely related samples merging earlier in the tree. The robust grouping of samples within identical batches underscores the existence of batch effects in the dataset, while the dense clusters within particular batches indicate substantial sample homogeneity.

Figure 33. Heatmap of samples' Pearson correlations across batches alongside hierarchical clustering.



Following generating heatmaps, the EDA module produced Scree plots and conducted PCA matrices. Scree plot shown in **Figure 34** displays the explained variance for the first 10 principal components from a PCA. The plot includes two elements: blue bars representing the individual explained variance for each principal component and a blue line indicating the cumulative explained variance. The x-axis represents the principal components, numbered from 1 to 10, while the y-axis represents the explained variance ratio, which is the proportion of the dataset's total variance explained by each principal component. The first principal component explains the highest variance, as indicated by the tallest bar. The explained variance decreases for subsequent principal components, with the first four components showing relatively higher values than the rest. The cumulative explained variance line starts at the explained variance of the first principal component and increases as additional components are included. The cumulative variance approaches 0.7 by the tenth principal component, indicating that these ten components together explain about 70% of the total variance in the dataset. Scree plot suggests that the first few principal components capture the majority of the variance, while additional components contribute progressively less.

Figure 34. Scree plot showing explained variance by principal components.

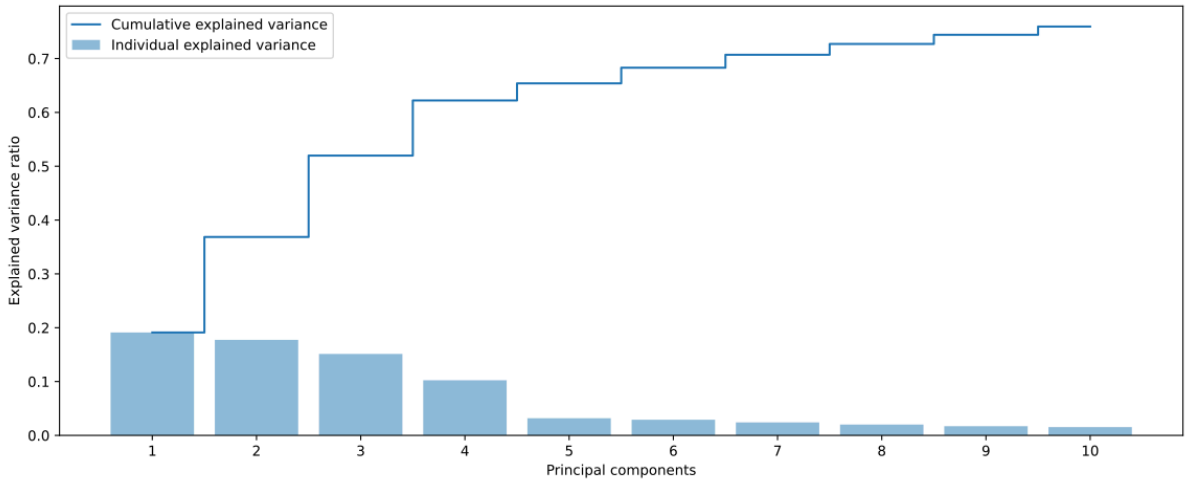


Figure 35 and **Figure 36** are a scatter plots matrix representing the first four principal components (PCs) from a principal component analysis of metabolite abundances. Each subplot displays the pairwise relationships between the principal components, allowing for the visualization of clustering patterns among the samples. The x-axis and y-axis of each subplot are labeled with the principal components (PC 1, PC 2, PC 3, and PC 4), with the percentage of variance explained by each component indicated in parentheses. In **Figure 35** different colors represent different batches of serum samples: "*ACSC_HumanSerum_11_5_19*" (blue), "*ACSC_HumanSerum_11_11_19*" (green), "*Human_Serum_2019_November_Shipment_Round_1*" (red), "*Human_Serum_2020_Feb_March_Shipments*" (purple), and "*Human_Serum_2020_January_Shipment*" (orange). Distinct clustering of samples based on their batch is observed. Samples from the same batch are grouped together, indicating similar variance captured by the principal components. Notably, PC 1, explaining 19.1% of the variance, delineates clear separations among several batches. Further differentiation is observed through PC 2, which accounts for 17.7% of the variance. PC 3 and PC 4, contributing 15.1% and 10.2% of the variance respectively, reveal even finer separations within the clusters. In **Figure 36** the data points are color-coded to distinguish between two groups: "*Patient*" (blue) and "*Relative*" (red). This coloring aids in visualizing clustering and separation patterns among the samples across disease status. In several subplots, particularly those involving PC 1 and PC 2, there are distinct clusters of blue and red points. The separation between the groups suggesting these components are significant in distinguishing between the two groups, clustering patients and relatives into separate regions.

Figure 35. PCA scatter plot matrix of first four components of metabolite abundances across all batches.

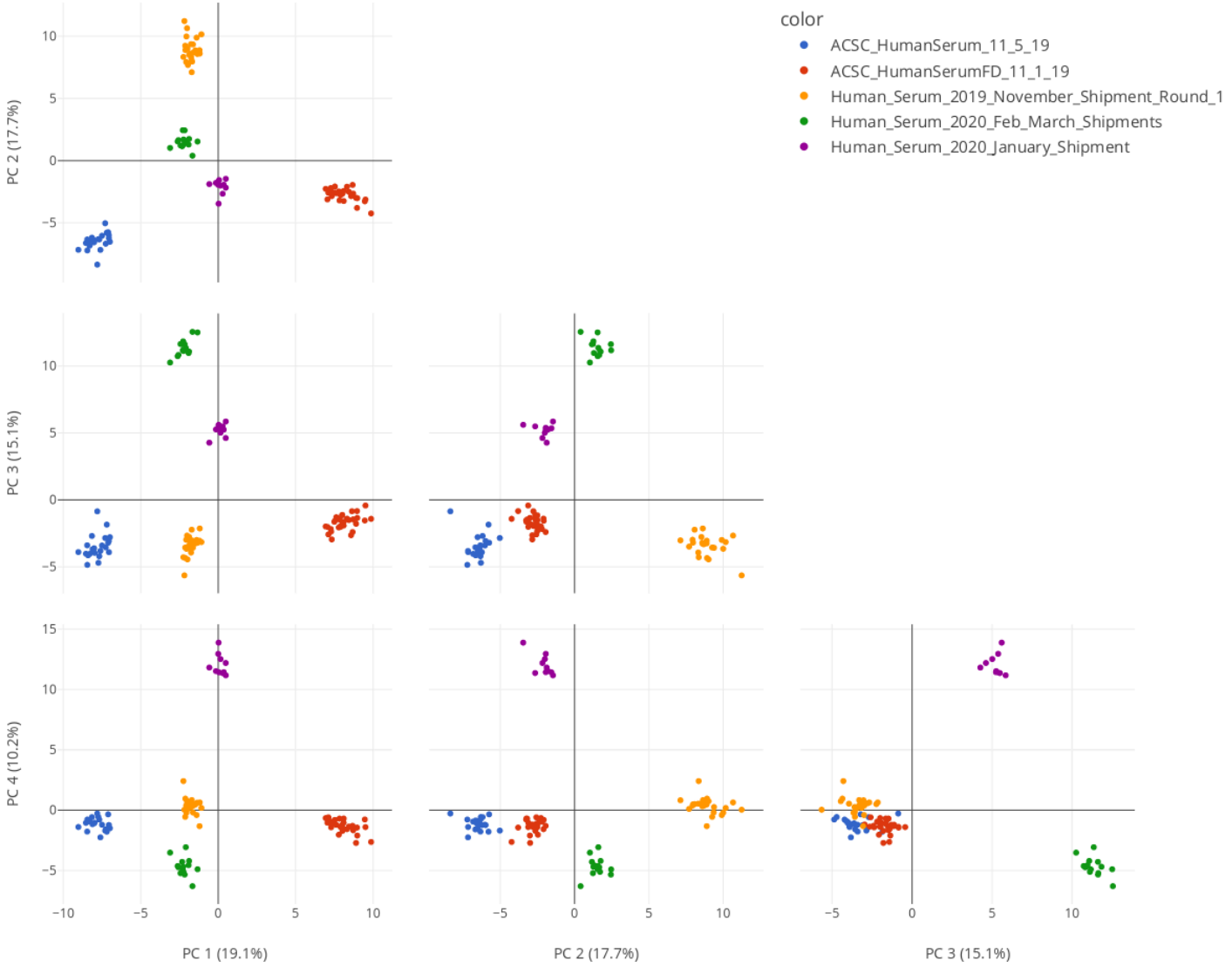
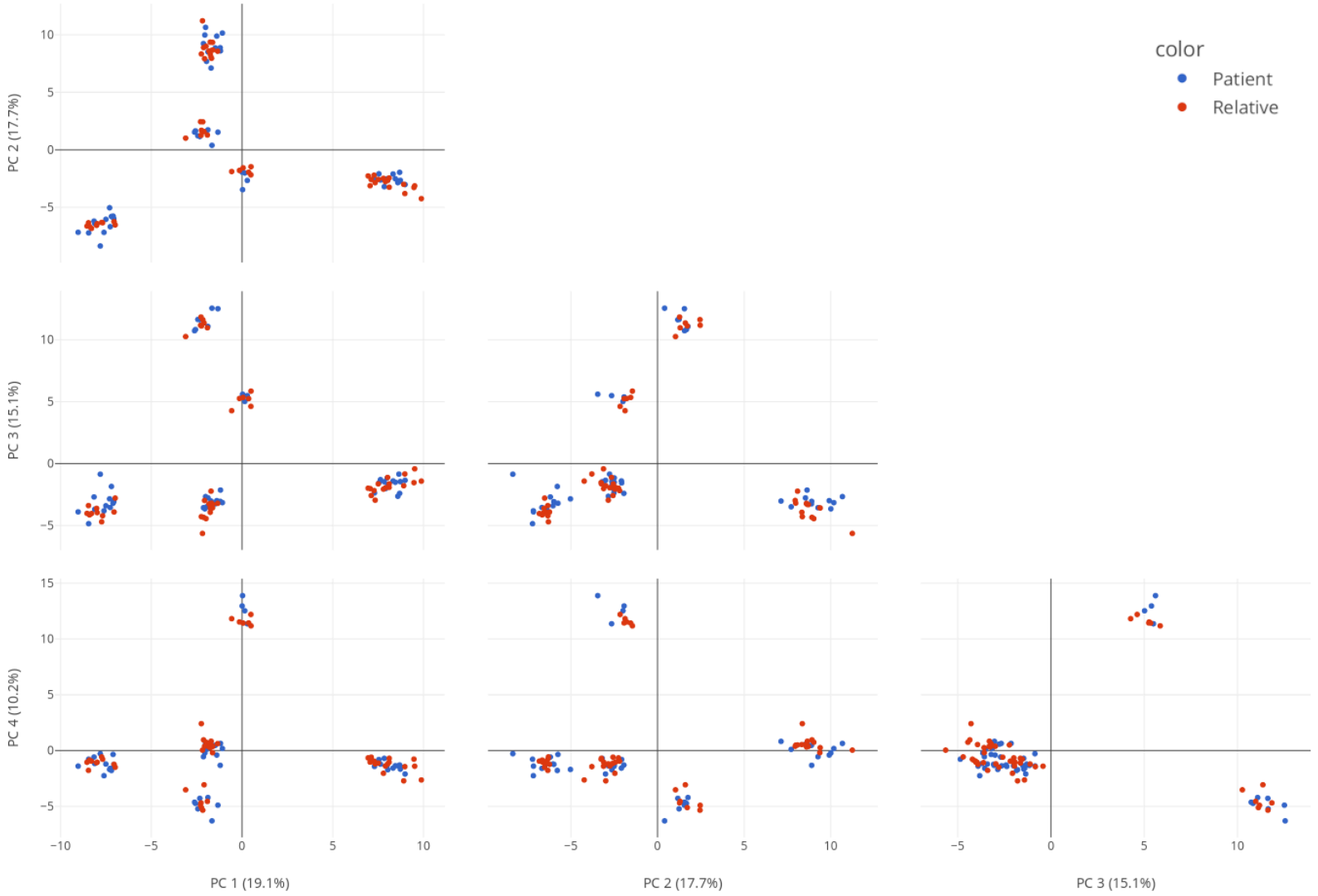


Figure 36. PCA scatter plot matrix of first four components of metabolite abundances across disease status.



The multivariate analysis generated two key files: *models_stratification.csv* and *multivariate_analysis_logistic_regression_c_1__features_relative_importance.csv*. The first file provides a comprehensive table with performance metrics and configurations for various machine learning models. This table comprises 1600 rows, each representing a specific model configuration and detailing its performance across several metrics. The six columns include:

- **Model**: Specifies the machine learning model along with its configuration.
- **Cycle**: Indicates the number of iterations performed during the fitting process.
- **Fit Time**: Represents the time taken to fit the model, measured in seconds.
- **Score Time**: Represents the time taken to score the model, measured in seconds.
- **Estimator**: Describes the specific estimator used, including any relevant hyperparameters.
- **Test Score**: Indicates the test score achieved by the model, reflecting its performance (ROC AUC statistic).

The best-performing model was Logistic Regression with C=1, achieving the highest mean ROC AUC of 0.6425 and a standard deviation of 0.0855 across 200 iterations of shuffled test/train dataset. The selection criteria for the best-performing model are detailed in Chapter II, section "Multivariate analysis". **Table 5**, extracted from the *models_stratification.csv* file, lists the top 20 models sorted in descending order of mean ROC AUC. This table demonstrates the variability in model performance across different shuffles of the data and underscores the choice of the best-performing model based on mean ROC AUC.

Table 5: Performance metrics and configurations for top 20 machine learning models.

Model	Cycle	Fit time	Score time	Estimator	Test score
Logistic regression (C=1)	49	0.0132	0.0011	LogisticRegression(random_state=0)	0.8709
Logistic regression (C=0)	173	0.0119	0.0013	LogisticRegression(C=0.1, random_state=0)	0.8387
Logistic regression (C=1)	173	0.0136	0.0014	LogisticRegression(random_state=0)	0.8387
Logistic regression L1 (C=1)	166	0.0101	0.0008	LogisticRegression(penalty='l1', random_state=0, solver='saga', tol=0.01)	0.8387
Logistic regression L1 (C=1)	173	0.0098	0.0008	LogisticRegression(penalty='l1', random_state=0, solver='saga', tol=0.01)	0.8387

Model	Cycle	Fit time	Score time	Estimator	Test score
Logistic regression L2 (C=0)	173	0.0062	0.0010	LogisticRegression(C=0.1, random_state=0, solver='saga', tol=0.01)	0.8387
Logistic regression L2 (C=1)	173	0.0069	0.0007	LogisticRegression(random_state=0, solver='saga', tol=0.01)	0.8387
Random forest	122	0.0962	0.0057	ExtraTreesClassifier(random_state=0)	0.8387
Logistic regression (C=1)	120	0.0090	0.0011	LogisticRegression(random_state=0)	0.8064
Logistic regression L1 (C=1)	4	0.0093	0.0008	LogisticRegression(penalty='l1', random_state=0, solver='saga', tol=0.01)	0.8064
Logistic regression L1 (C=1)	146	0.0093	0.0007	LogisticRegression(penalty='l1', random_state=0, solver='saga', tol=0.01)	0.8064
Logistic regression L1 (C=1)	182	0.0076	0.0006	LogisticRegression(penalty='l1', random_state=0, solver='saga', tol=0.01)	0.8064
Logistic regression L1 (C=1)	186	0.0087	0.0008	LogisticRegression(penalty='l1', random_state=0, solver='saga', tol=0.01)	0.8064
Logistic regression L2 (C=0)	49	0.0076	0.0011	LogisticRegression(C=0.1, random_state=0, solver='saga', tol=0.01)	0.8064
Logistic regression L2 (C=1)	49	0.0067	0.0008	LogisticRegression(random_state=0, solver='saga', tol=0.01)	0.8064
Logistic regression (CV=2)	146	5.1095	0.0011	LogisticRegressionCV(cv=2, l1_ratios=array([0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]), max_iter=10000, penalty='elasticnet', random_state=0, solver='saga')	0.8064
Logistic regression (CV=2)	173	5.3120	0.0014	LogisticRegressionCV(cv=2, l1_ratios=array([0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]), max_iter=10000, penalty='elasticnet', random_state=0, solver='saga')	0.8064
Random forest	161	0.1026	0.0054	ExtraTreesClassifier(random_state=0)	0.8064
Logistic regression (C=0)	49	0.0069	0.0011	LogisticRegression(C=0.1, random_state=0)	0.7741
Logistic regression (C=0)	120	0.0082	0.0013	LogisticRegression(C=0.1, random_state=0)	0.7741

Source: Based on [178], outcome from NASQQ pipeline execution.

The second key file from multivariate analysis contains a table listing the features' relative importance in descending order, as determined by the best performing machine learning model. Through multivariate analysis, 50 metabolites that significantly contributed to the best model's feature stratification across all samples, covering 70% of the relative importance of all Shapley values, were identified (refer to **Supplementary Table 3**). These quantified metabolites may be categorized into distinct groups, including their derivatives: carbohydrates (D-glucuronic acid, D-galactose, xylitol, D-glucose-6-phosphate), amino acids (L-anserine, 3-methyl-L-histidine, creatinine, L-arginine, L-cysteine, L-glutamine, L-serine, L-methionine), organic acids (trans-4-hydroxy-L-proline, threonic acid, succinate, pyruvic acid, citrate, isobutyrate, saccharic acid, glycerol, isocitric acid, malic acid, glycolic acid, isovaleric acid, oxypurinol, methanol, butyrate, acetoacetate), nucleotides (UMP, argininosuccinic acid, 7-methylxanthine, dihydrothymine, IMP, dAMP, GMP), and miscellaneous metabolites (spermidine, dimethylamine, vanillic acid, 2-aminobutyric acid, glycerophosphocholine, 2-propanol, trans-acotinic acid, TMAO, sarcosine, ascorbic acid).

III.5 KEGG-based metabolomic pathways intersection

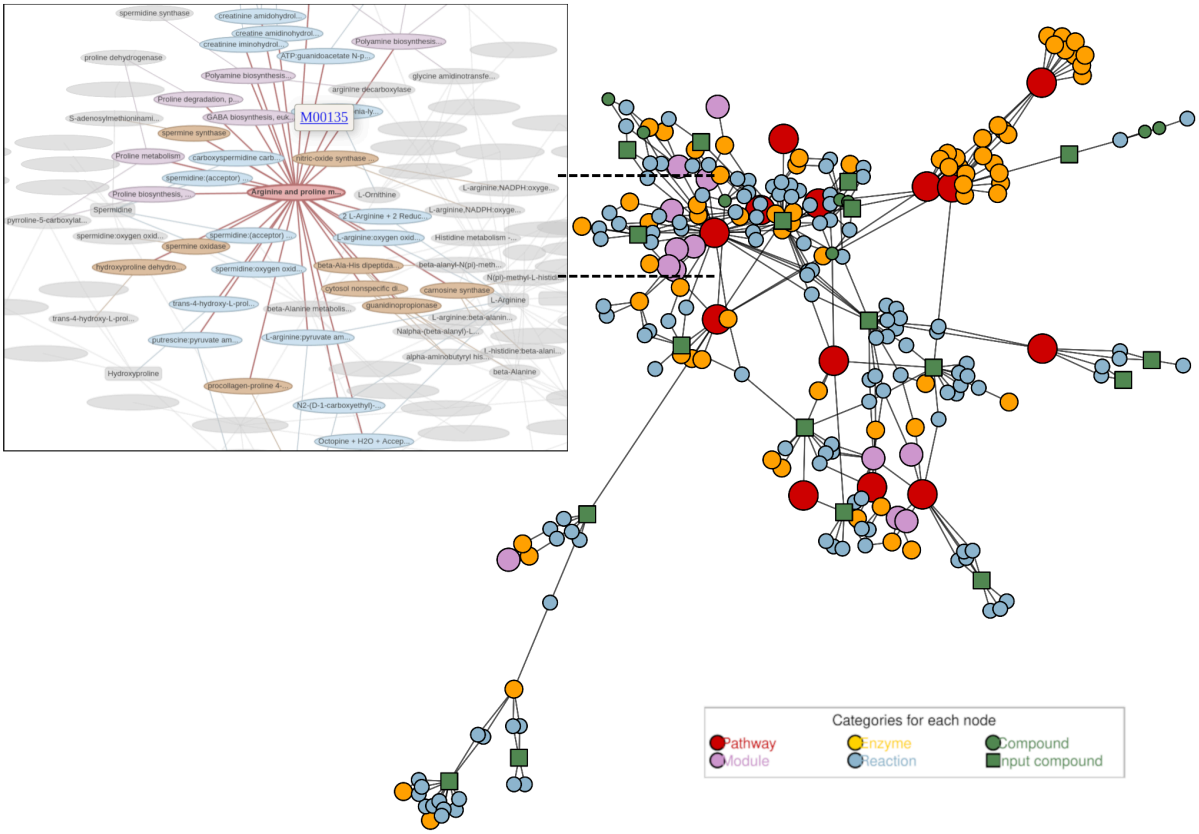
During the "*Biological Interpretation*" stage, the top 20 metabolites from the multivariate results underwent a FELLA-like pathway analysis. The univariate module, however, did not yield enough significant metabolites (N=6) to provide robust statistical power for further pathway analysis (N=20) and is therefore not included in the results section. The pathway analysis outcomes offered valuable insights into various metabolic pathways, enzyme functions, and reactions significantly associated with the FD dataset. This analysis highlighted potential alterations in metabolomic pathways, including the citrate cycle, arginine and proline metabolism, beta-alanine metabolism, glucagon signaling pathway, ABC transporters, mTOR signaling pathway, and carbon metabolism, as well as the carbon metabolism in cancer pathway (refer to **Supplementary Table 4**). The identified pathways encompassed a range of metabolic processes crucial for cellular function and health. Several pathways exhibited notably low p-value scores, indicating their strong correlation with the dataset. Particularly noteworthy are the citrate cycle (TCA cycle), ascorbate and aldarate metabolism, arginine and proline metabolism, histidine metabolism, and beta-alanine metabolism. Numerous enzymes with significant p-scores, underscoring their vital roles in the associated metabolic pathways. Major enzymes included L-xylulose reductase, aldose reductase, malate dehydrogenase, isocitrate dehydrogenase, nitric-oxide synthase, and various peptidases and oxidoreductases. These enzymes are involved in critical biochemical reactions, such as redox reactions, amino acid catabolism, and peptide processing, reflecting their diverse functions in cellular metabolism. The highlighted reactions with significant associations provided detailed insights into specific biochemical transformations. Reactions such as ATP phosphotransferase, L-alanine:2-oxoglutarate aminotransferase, succinate + oxidoreductase, and various aminotransferases represent key metabolic steps in glycolysis, the TCA

cycle, and amino acid metabolism. The involvement of reactions like L-cysteine oxidoreductase and gamma-L-glutamyl-L-cysteinyl-glycine oxidoreductase underscored the importance of redox balance and polyamine metabolism. The pathway analysis underscored metabolic reactions, offering an in-depth comprehension of the biochemical transformations and metabolic alterations within the FD dataset.

Supplementary Table 4 is visually represented in **Figure 37**, with the results also accessible as an interactive HTML object (provided upon request). This figure illustrates a network visualization derived from a KEGG enrichment analysis, encompassing metabolic pathways, enzymes, reactions, and compounds. In the visualization, nodes symbolize various biological entities: pathways are depicted as large red circles, modules as pink circles, enzymes as orange circles, reactions as yellow circles, compounds as small blue circles, and input compounds as green squares. The edges connecting these nodes delineate the relationships among them, providing insight into the connectivity within the metabolic network. Dense clusters of nodes indicate areas of heightened connectivity and interaction, serving as focal points within the metabolic network. Larger nodes, such as pathways and modules, offer a broader categorical context, while smaller nodes, like compounds and input compounds, provide detailed insights into specific biochemical substances. **Figure 37** offers a comprehensive view of the intricate interactions and relationships within metabolic pathways, pinpointing key areas of metabolic activity and potential avenues for further investigation.

The pathway analysis showcases the metabolic background linked to the FD dataset, underscoring the vital functions of energy metabolism, amino acid processing, redox homeostasis, and enzymatic activities in maintaining cellular health and responding to metabolic demands. These results provide a valuable foundation for further investigation into the specific biological processes and their implications in the studied context. Understanding these metabolic pathways and enzyme functions could lead to new insights into the metabolic reprogramming associated with the FD dataset. This knowledge can help in identifying potential therapeutic targets and developing strategies to modulate these metabolic pathways for better disease management and treatment outcomes. The detailed insights into specific biochemical transformations offer a promising direction for future research, aiming to unravel the complex metabolic networks and their roles in cellular health and disease progression.

Figure 37. Network visualization of KEGG pathways intersection for top 20 metabolites identified by multivariate analysis.



Source: Based on [178], created in GIMP.

IV Discussion

IV.1 General conclusions of pipeline usage for metabolomic analysis

As demonstrated in the results section for the *Familial Dysautonomia* use case, the NASQQ pipeline offers users a comprehensive, end-to-end solution, providing control over each step of the analysis process. It ensures access to processed data tables and visual representations in figures, allowing users to easily track progress and identify discrepancies in the data flow. These discrepancies can be identified either visually or by loading specific RDS objects into the R environment. It was proved that this approach provides a flexible framework for loading objects, adjusting parameters, and reiterating the analysis. The comparison between the *Familial Dysautonomia* originally preprocessed spectra and those obtained using the NASQQ pipeline reveals that, with appropriate parameter configuration, the pipeline can produce results of comparable quality. The pipeline's capability to deliver high-quality results, ensures that the spectra quality from NASQQ closely matches that of the benchmark study.

In the context of the metabolite identification module, identifying too many metabolites can lead to poorer quality quantifications for metabolites which are at the low concentration level. This issue often arises because these low-abundance metabolites are difficult to distinguish from noise. The problem may stem not only from the quantification process itself but also from preceding preprocessing steps meant to enhance the quality of the complex spectrum being analyzed. A crucial preprocessing step in the *ASICS* method implemented in the pipeline involves aligning every pure spectrum from the reference library with the analyzed complex mixture. However, identification and quantification tools, regardless of the method used, are typically designed to process complex spectra individually. This approach is inefficient when spectra come from the same experiment under similar conditions and share similarities [201]. Consequently, many metabolites were quantified at insignificant abundances and discarded during the data analysis stage. While the foundational *ASICS* method relies on predefined libraries, the tool's design inherently supports the use of custom reference libraries. This adaptability makes it highly suitable for various research applications, allowing results to be validated both *in silico* and in the laboratory. This versatility was a crucial factor in choosing *ASICS* as the leading method, influencing the final choice of methodology used in the pipeline during the design stage.

The data analysis approach employed in NASQQ integrates traditional univariate methods with advanced machine learning-based multivariate techniques. This comprehensive methodology has been thoroughly reviewed based on existing literature and has resulted in a published review on data analysis methods in immune checkpoint therapy (ICT) [202]. Figures from exploratory data analysis, such as distribution plots or PCA, presented in the results section, revealed batch effects across multiple shipments in the original FD dataset. To address these issues and mitigate the impact

of lower-quality quantifications of metabolites, the multivariate module is designed to handle large metabolomics datasets, encompassing various time points, patient conditions, and diseases. It employs an iterative process to select the optimal model by computing the mean ROC AUC to assess model performance. The method determines the relative importance of metabolites using Shapley values, which provide a measure of each metabolite's contribution to the outcome. This helps differentiate significant metabolites from artifacts. In the final stage of the pipeline, Shapley values are integrated into KEGG-based pathway analysis, linking identified metabolites to specific biological pathways. This approach ensures the accurate identification of biomarkers while minimizing the inclusion of irrelevant data.

In the original study on *Familial Dysautonomia*, metabolites like xanthine, urea, and methanol showed significantly different levels in serum samples of affected patients compared to their healthy relatives [73]. In a NASQQ pipeline case study, methanol as a metabolite with significantly increased abundance was also identified. However, the identification of other metabolites was hindered by potential discrepancies in the reference database used for quantification during our analysis (xanthine and urea were not part of the pure library reference, hence could not be identified). Since the study's focus diverged from replicating or correcting published findings—due to our fundamentally different analysis workflow design, especially at the metabolites identification stage—this open dataset was chosen as a case study because of its thorough raw data preparation and well-organized metadata. These discrepancies highlight the sensitivity of results to specific data processing workflows employed, which can significantly influence the outcome. Furthermore, univariate approaches often lack the statistical power to detect subtle yet significant changes in metabolite levels, particularly when handling high-dimensional metabolomics data. This limitation was evident in the current analysis, where the univariate testing did not yield a sufficient number of metabolites for a valid KEGG-based pathway analysis. This shortfall underscores the necessity of using machine learning models in parallel. The ML approach enhanced the sensitivity and specificity of metabolite detection, allowing for the identification of a wider range of biologically significant metabolites and the integration of information from multiple metabolites to identify more reliable biomarkers. Consequently, it enabled a more comprehensive and accurate mapping of metabolic pathways, offering deeper insights into the biological processes underlying the condition under chosen case study. With results from the multivariate analysis module, further biological interpretation revealed potential alterations in several metabolomic pathways, including the citrate cycle, arginine and proline metabolism, beta-alanine metabolism, glucagon signaling pathway, ABC transporters, mTOR signaling pathway, carbon metabolism, and the carbon metabolism in cancer pathway. The altered metabolites identified by NASQQ, despite their structural and functional diversity, may share common roles within these cellular processes and metabolic pathways. For example, amino acids such as L-glutamine, L-serine, and L-arginine, which are involved in neurotransmitter synthesis [203], are also integral to several of the identified pathways.

L-glutamine can feed into the citrate cycle through conversion to alpha-ketoglutarate, influencing energy production and biosynthesis. While L-arginine plays a key role in arginine and proline metabolism, contributing to the synthesis of polyamines like spermidine, which are crucial for cell growth, differentiation, and regulation of apoptosis. L-serine is involved in the biosynthesis of beta-alanine, important for muscle function and neurotransmission. Polyamines like spermidine, essential for cell growth, differentiation, and regulation of apoptosis [204-205], also intersect with these pathways. Spermidine can influence the mTOR signaling pathway, vital for cell growth and metabolism regulation, and serve as substrates for certain ABC transporters involved in cellular detoxification and homeostasis. Dysfunctions in neurotransmitter function, including defective release or signaling, along with dysregulation of polyamine metabolism, could impact neuronal development and function, contributing to the observed pathophysiology in *Familial Dysautonomia* [206]. The glucagon signaling pathway, which influences glucose metabolism, may also involve regulatory roles for amino acids like L-glutamine and L-arginine. Additionally, pathways such as carbon metabolism are linked to overall cellular metabolic activity, with metabolites like L-glutamine serving as key intermediates.

In conclusion, the discovery of altered metabolites and their associated pathways through the NASQQ pipeline provides a comprehensive understanding of FD's functional background. This identification and interpretation of metabolic alterations offer valuable insights into disease mechanisms and suggest potential therapeutic targets. With further refinement and calibration, the NASQQ pipeline could facilitate more detailed investigations into the metabolic signatures of various diseases, advancing both diagnostic and therapeutic strategies. While the pipeline's capacity for automated analysis of complex metabolic data remains promising, it is important to note the limitations and challenges, which are detailed in the section below.

IV.2 Utilized methodologies limitations

The NASQQ pipeline, designed specifically for 1D ^{1}H NMR analysis, has several limitations. First, the workflow begins with raw data in Bruker format and currently does not support other data formats. This limitation is mostly by design of version 1.0.0. of the pipeline, even though the baseline function used for loading raw FIDs includes support for other formats, during testing, it was found that other formats were not as straightforward for extracting metadata and required more error handling. Additionally, the pipeline is configured for experiments using a consistent presaturation pulse program. If the pulse program varies among samples within an experiment directory, the pipeline may encounter issues, either failing to execute when the pulse program is provided incorrectly by the user or including only those sample names that match the pulse program specified in the configuration file. The impact of the pulse program on the LB parameter during the apodization

step is crucial, as described in the “Spectral processing of raw 1D spectra and metabolites identification” section.

The complexity of preparing and processing NMR spectra and the potential pitfalls of a fully automated solution should not be overlooked. However, unifying the pipeline should enhance the reproducibility of results in cross-laboratory projects. Parameters in modules such as zero-order phase correction and baseline correction should be carefully considered, as the automated nature of the pipeline does not allow for real-time flexible adjustments, unlike web-based or standalone tools with built-in graphical user interfaces. While default parameters are sufficient for basic analysis, Nextflow allows rerunning sessions with different parameters if adjustments are needed. The tables and plots generated at each step of the analysis make it easy to identify any issues during preprocessing. Additionally, the prepared individual scripts and Docker environment enable users to load outcome objects and make manual corrections or test multiple parameter settings. The pipeline is ideally suited for analyzing samples derived from human biofluids such as plasma, serum, or urine. Feces, however, have a much more complex and variable composition, containing a high diversity of microorganisms, undigested food particles, and a wide range of metabolites. This complexity can complicate the analysis, and therefore, fecal samples were not included as a primary target for pipeline validation in this work. Metabolite quantification is based on a reference object from the *ASICS* R package, which includes 191 compounds and serves as input for the “Biological Interpretation” stage. Additionally, the pathway enrichment process relies on the KEGG database. If a metabolite identified during quantification does not have a corresponding KEGG identifier, it is excluded from the enrichment analysis—this affects 11 metabolites in the current dataset.

Due to the workflow's sequential execution order, NextFlow's potential for parallelization appears to be underutilized. Although the framework supports simultaneous execution of multiple processes once preceding steps are completed and resources are available, the pipeline primarily achieves maximum efficiency during metabolite quantification and machine learning stages. Furthermore, the absence of a GUI may pose a challenge for users who lack basic programming and command line skills. The manuals found in the GitHub repository offer guidance on setting up the workflow across multiple operating systems, preparing inputs, and using individual scripts. However, they do not provide thorough explanations of NextFlow's operational mechanics and debugging methods. This knowledge gap may initially hinder less technically proficient users from effectively using the workflow, particularly when starting first analyses.

In summary, the NASQQ pipeline efficiently processes 1D ^{1}H NMR data, but while effective for standardized experiments, its dependency on Bruker format and consistent presaturation pulse programs may limit adaptability across diverse experimental setups. Automation simplifies basic analysis but requires manual intervention for nuanced parameter adjustments, highlighting the need for user proficiency with command-line interfaces and basic NMR processing knowledge. Future

improvements in interface usability and workflow flexibility could broaden its application in metabolomics research.

IV.3 Future directions and perspectives for further development

Considering the vast and complex nature of the metabolomic analysis, even with the development of a comprehensive pipeline, there remains room for further additions and refinements. Additionally, as an open system, the NASQQ pipeline is easily adaptable for incorporating new modules as needed. Its modular design facilitates the expansion of current functionalities and the integration of additional modules in the future. For instance, the "Spectral Preprocessing" stage could be extended to support 2D ^{1}H-^{13}C NMR spectra, offering improved peak resolution and revealing additional cross-peak information, particularly useful for analyzing complex mixtures. Another valuable enhancement would be the capability to analyze other metabolomics techniques, such as mass spectrometry. Integrating with pipelines like metaboigniter (https://github.com/nf-core/metaboigniter/tree/2.0.0) would further enhance the workflow's versatility and usability in the metabolomics field. Furthermore, incorporating custom subworkflows that exclude the preprocessing of spectra could allow for more flexible handling of semi-preprocessed input data, thereby accommodating varying preprocessing needs and improving the adaptability of the pipeline.

Creating a custom reference library of pure metabolites can significantly enhance the identification and quantification of metabolites in preprocessed NMR spectra. By validating this library in wet laboratories, it is possible to include specific metabolites relevant to the targeted study, which may be absent in standard libraries. This approach ensures that the analysis aligns more closely with the unique metabolic profile of the samples, thereby improving the accuracy and precision of metabolite identification and quantification.

To enhance the statistical power and performance of the "*Data Analysis*" stage, it would be beneficial to expand the range of machine learning models to include deep learning techniques such as Convolutional Neural Networks (CNNs). These advanced models can handle complex patterns and interactions within the data, potentially leading to more accurate and insightful results compared to traditional models. Additionally, incorporating feature selection methods across different models can improve the robustness of the analysis by identifying the most relevant variables. Implementing automated batch correction processes would further ensure data integrity by addressing batch effects utilizing methods such as ComBat [207], thereby enhancing the overall reliability and validity of the results.

Enhancing the workflow with additional downstream analyses using R packages such as *ROTS* [208] for differential expression analysis and *metaboAnalystR* [177] for metabolite set enrichment analysis (MSEA) would significantly improve result interpretability. Expanding the workflow to incorporate multi-omics tools and approaches, along with leveraging pathview

for comprehensive biochemical pathway representation, would provide a more holistic understanding of the data. Improvements could also include additional graphical representations with interactive plots and the creation of PDF reports or markdown documents consolidating all results, thereby enhancing the user experience. Augmenting the pipeline's computing capabilities by integrating other executors, such as Kubernetes (K8s) or cloud computing platforms like AWS Batch and Azure Batch, is a promising direction for future development. Collaboration with communities like NF-core remains crucial to address potential bugs and further optimize the NASQQ pipeline's performance.

Lastly, improving the accessibility of the tool should be a key focus. More detailed manuals and guides for users with limited programming and NMR analysis knowledge should be incorporated into the GitHub repository. Ideally, long-term support for the project should include developing a graphical user interface for the pipeline, a standalone application, or at least a web interface such as MaCWorP [209]. Such enhancements would significantly increase the likelihood of the pipeline being adopted by various scientific groups.

In summary, the NASQQ pipeline offers a production-wise approach compared to established and widely used methods like *NMRprocFlow* [170] and *Metaboanalyst* [175]. Its functionality extends beyond preprocessing and univariate analysis, incorporating machine learning approaches to elucidate biological interactions within pathway analysis modules. NASQQ serves as a valuable alternative to existing tools, particularly suitable for automated analysis implementations. Unlike web-based solutions that require a connection to external servers—often impractical due to data confidentiality—pipeline operates locally, ensuring data security. Additionally, compared to GUI-based solutions, NASQQ allows for programmatic adjustments and leverages Docker containers to provide a stable environment, accelerating the analysis of multiple spectra and ensuring tidy data output. The continuous data processing capabilities of NASQQ are especially important for enterprises and industrial implementations where maintaining data integrity and confidentiality is paramount. This local, automated pipeline ensures efficient handling of large-scale data while safeguarding sensitive information, making it a robust choice for production environments. Despite their differences, both NASQQ and traditional existing approaches contribute significantly to the advancement of metabolomics, each offering unique strengths and applications.

IV.4 Final remarks of the thesis

In conclusion, the main and secondary objectives outlined in the "Motivation and thesis outline" section—namely, the automated bioinformatic analysis of metabolite-derived signals in blood serum spectra obtained through 1D ^{1}H NMR proton magnetic resonance—have been successfully achieved, as demonstrated in the FD case study. The study also validated the reliability of the proposed hypotheses for this type of analysis.

Firstly, at nearly every stage of the NASQQ pipeline, a set of user-defined parameters significantly impacts the final outcome. In the "*Spectral Preprocessing*" stage, parameters such as the pulse program, which influences subsequent apodization line broadening settings, internal referencing parameters (range_type and target_value), baseline correction parameters (p_bc and lambda_bc), the number of buckets (mb) in bucketing, and the type of normalization used, all shape the data analysis process and the resulting insights.

Secondly, custom machine learning methods markedly enhance the identification and accuracy of significant metabolites in disease progression. A comparison of significant metabolites identified through univariate methods corrected by multiple testing and those identified through multivariate machine learning approaches supports this hypothesis. Machine learning models handle high-dimensional data and capture complex, nonlinear interactions and dependencies between metabolites that univariate methods might overlook. Models proposed in the "*Data Analysis*" stage effectively manage correlated features, avoiding issues like multicollinearity, and advanced techniques such as Shapley values make them more interpretable, revealing the contribution of each metabolite to the model's predictions. Additionally, machine learning models may facilitate the integration of NMR data with other omics data types, enabling comprehensive multi-omics analyses.

Lastly, the efficiency in computation time is notable. For average-power personal computers, as described in the "NextFlow implementation and containerized computing environment" section, the analysis of 101 FD serum samples took approximately 5 hours and 30 minutes, with metabolite quantification consuming the bulk of this time. This level of efficiency is difficult to achieve manually, even for experienced chemists using standalone state-of-the-art tools. Moreover, with more computing power, such as a dedicated server, the scalability and speed of the analysis increase significantly.

Overall, the study successfully demonstrates the feasibility and effectiveness of automated metabolite analysis using advanced bioinformatics pipelines and machine learning methods, providing a robust framework for future research and clinical applications

List of Figures

List of Tables

List of Equations

References

[1] Nicholson JK, Lindon JC. Systems biology: Metabonomics. Nature. 2008 Oct 23;455(7216):1054-6. doi: 10.1038/4551054a. PMID: 18948945.

[2] Fiehn O. Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol. 2002 Jan;48(1-2):155-71. PMID: 11860207.

[3] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015 Feb;16(2):85-97. doi: 10.1038/nrg3868. Epub 2015 Jan 13. PMID: 25582081.

[4] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L et al.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. doi: 10.1038/35057062. Erratum in: Nature 2001 Aug 2;412(6846):565. Erratum in: Nature 2001 Jun 7;411(6838):720. Szustakowki, J [corrected to Szustakowski, J]. PMID: 11237011.

[5] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51. doi: 10.1126/science.1058040. Erratum in: Science 2001 Jun 5;292(5523):1838. PMID: 11181995.

[6] Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, Schneider VA, Potapova T, Wood J, Chow W, Armstrong J, Fredrickson J et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020 Sep;585(7823):79-84. doi: 10.1038/s41586-020-2547-7. Epub 2020 Jul 14. PMID: 32663838; PMCID: PMC7484160.

[7] Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007 Feb 23;128(4):669-81. doi: 10.1016/j.cell.2007.01.033. PMID: 17320505.

[8] Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol. 2010 Oct;28(10):1057-68. doi: 10.1038/nbt.1685. PMID: 20944598.

[9] Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017 May 18;13(5):e1005457. doi: 10.1371/journal.pcbi.1005457. PMID: 28545146; PMCID: PMC5436640.

[10] Al-Amrani S, Al-Jabri Z, Al-Zaabi A, Alshekaili J, Al-Khabori M. Proteomics: Concepts and applications in human medicine. World J Biol Chem. 2021 Sep 27;12(5):57-69. doi: 10.4331/wjbc.v12.i5.57. PMID: 34630910; PMCID: PMC8473418.

[11] Nicholson JK, Connelly J, Lindon JC, Holmes E. Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov. 2002 Feb;1(2):153-61. doi: 10.1038/nrd728. PMID: 12120097.

[12] Weckwerth W, Morgenthal K. Metabolomics: from pattern recognition to biological interpretation. Drug Discov Today. 2005 Nov 15;10(22):1551-8. doi: 10.1016/S1359-6446(05)03609-3. PMID: 16257378.

[13] Aharoni A, Goodacre R, Fernie AR. Plant and microbial sciences as key drivers in the development of metabolomics research. Proc Natl Acad Sci U S A. 2023 Mar 21;120(12):e2217383120. doi: 10.1073/pnas.2217383120. Epub 2023 Mar 17. PMID: 36930598; PMCID: PMC10041103.

[14] Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. Nat Biotechnol. 2000 Nov;18(11):1157-61. doi: 10.1038/81137. Erratum in: Nat Biotechnol 2000 Feb;19(2):173. PMID: 11062433.

[15] Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN, Nicholls AW, Wilson ID, Kell DB, Goodacre R; Human Serum Metabolome (HUSERMET) Consortium. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011 Jun 30;6(7):1060-83. doi: 10.1038/nprot.2011.335. PMID: 21720319.

[16] Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, Kopka J. GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett. 2005 Feb 28;579(6):1332-7. doi: 10.1016/j.febslet.2005.01.029. Epub 2005 Jan 28. PMID: 15733837.

[17] Dunn WB, Ellis DI. Metabolomics: Current analytical platforms and methodologies. Trends in Analytical Chemistry, 24(4), 285-294. doi: 10.1016/j.trac.2004.11.021

[18] Cajka T, Fiehn O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. Anal Chem. 2016 Jan 5;88(1):524-45. doi: 10.1021/acs.analchem.5b04491. Epub 2015 Dec 16. PMID: 26637011.

[19] Plumb RS, Johnson KA, Rainville P, Smith BW, Wilson ID, Castro-Perez JM, Nicholson JK. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid Commun Mass Spectrom. 2006;20(13):1989-94. doi: 10.1002/rcm.2550. Erratum in: Rapid Commun Mass Spectrom. 2006;20(14):2234. PMID: 16755610.

[20] Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. The Orbitrap: a new mass spectrometer. J Mass Spectrom. 2005 Apr;40(4):430-43. doi: 10.1002/jms.856. PMID: 15838939.

[21] Faßbender S, von der Au M, Koenig M, Pelzer J, Piechotta C, Vogl J, Meermann B. Species-specific isotope dilution analysis of monomethylmercury in sediment using GC/ICP-ToF-MS and comparison with ICP-Q-MS and ICP-SF-MS. Anal Bioanal Chem. 2021

Sep;413(21):5279-5289. doi: 10.1007/s00216-021-03497-z. Epub 2021 Jul 23. PMID: 34302182; PMCID: PMC8405517.

[22] Letertre MPM, Giraudeau P, de Tullio P. Nuclear Magnetic Resonance Spectroscopy in Clinical Metabolomics and Personalized Medicine: Current Challenges and Perspectives. Front Mol Biosci. 2021 Sep 20;8:698337. doi: 10.3389/fmolb.2021.698337. PMID: 34616770; PMCID: PMC8488110.

[23] Emwas AH. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. Methods Mol Biol. 2015;1277:161-93. doi: 10.1007/978-1-4939-2377-9_13. PMID: 25677154.

[24] Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. Trends Biotechnol. 1998 Sep;16(9):373-8. doi: 10.1016/s0167-7799(98)01214-1. Erratum in: Trends Biotechnol 1998 Oct;16(10):447. PMID: 9744112.

[25] Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999 Nov;29(11):1181-9. doi: 10.1080/004982599238047. PMID: 10598751.

[26] Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol. 2001 Jan;19(1):45-50. doi: 10.1038/83496. PMID: 11135551.

[27] Wishart DS, Cheng LL, Copié V, Edison AS, Eghbalnia HR, Hoch JC, Gouveia GJ, Pathmasiri W, Powers R, Schock TB, Sumner LW, Uchimiya M. NMR and Metabolomics-A Roadmap for the Future. Metabolites. 2022 Jul 23;12(8):678. doi: 10.3390/metabo12080678. PMID: 35893244; PMCID: PMC9394421.

[28] Griffin JL, Nicholls AW. Metabolomics as a functional genomic tool for understanding lipid dysfunction in diabetes, obesity and related disorders. Pharmacogenomics. 2006 Oct;7(7):1095-107. doi: 10.2217/14622416.7.7.1095. PMID: 17054419.

[29] Griffin JL. Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. Curr Opin Chem Biol. 2003 Oct;7(5):648-54. doi: 10.1016/j.cbpa.2003.08.008. PMID: 14580571.

[30] Zhang A, Sun H, Yan G, Wang P, Wang X. Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research. Biomed Chromatogr. 2016 Jan;30(1):7-12. doi: 10.1002/bmc.3453. Epub 2015 Mar 4. PMID: 25739660.

[31] Gika HG, Theodoridis GA, Earll M, Wilson ID. A QC approach to the determination of day-to-day reproducibility and robustness of LC-MS methods for global metabolite profiling in metabonomics/metabolomics. Bioanalysis. 2012 Sep;4(18):2239-47. doi: 10.4155/bio.12.212. PMID: 23046266.

[32] Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov. 2016 Jul;15(7):473-84. doi: 10.1038/nrd.2016.32. Epub 2016 Mar 11. PMID: 26965202.

[33] Keeler J. Understanding NMR spectroscopy (2nd ed.). 2011. John Wiley & Sons.

[34] Levitt MH. Spin dynamics: basics of nuclear magnetic resonance. 2008. John Wiley & Sons.

[35] Fotopoulou E, Ronconi L. Application of Heteronuclear NMR Spectroscopy to Bioinorganic and Medicinal Chemistry. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. 2018:B978-0-12-409547-2.10947-3. doi: 10.1016/B978-0-12-409547-2.10947-3. Epub 2014 Feb 6. PMCID: PMC7157447.

[36] Hansen PE, Spanget-Larsen J. NMR and IR Investigations of Strong Intramolecular Hydrogen Bonds. Molecules. 2017 Mar 29;22(4):552. doi: 10.3390/molecules22040552. PMID: 28353675; PMCID: PMC6154318.

[37] McRobbie DW, Moore EA, Graves MJ, Prince MR. MRI: From Picture to Proton (3rd ed.). 2017. Cambridge University Press.

[38] Serai SD. Basics of magnetic resonance imaging and quantitative parameters T1, T2, T2*, T1rho and diffusion-weighted imaging. Pediatr Radiol. 2022 Feb;52(2):217-227. doi: 10.1007/s00247-021-05042-7. Epub 2021 Apr 15. PMID: 33856502.

[39] Bojorquez JZ, Bricq S, Acquitter C, Brunotte F, Walker PM, Lalande A. What are normal relaxation times of tissues at 3 T? Magn Reson Imaging. 2017 Jan;35:69-80. doi: 10.1016/j.mri.2016.08.021. Epub 2016 Sep 2. PMID: 27594531.

[40] Brown RW, Cheng YCN, Haacke EM, Thompson MR, Venkatesan R. Magnetic resonance imaging: Physical principles and sequence design (2nd ed.). 2014. John Wiley & Sons.

[41] Brink HF, Buschmann MD, Rosen BR. NMR chemical shift imaging. Comput Med Imaging Graph. 1989 Jan-Feb;13(1):93-104. doi: 10.1016/0895-6111(89)90081-5. PMID: 2538222.

[42] Jameson CJ. Understanding NMR Chemical Shifts. Annu Rev Phys Chem. 1996 Oct;47:135-169. doi: 10.1146/annurev.physchem.47.1.135.

[43] Lambert JB, Mazzola E, Ridge CD. Nuclear Magnetic Resonance Spectroscopy: An Introduction to Principles, Applications, and Experimental Methods (2nd ed.). 2018. Wiley Global Research (STMS). ISBN 9781119295280.

[44] Becker ED, Farrar TC. Fourier transform spectroscopy. Science. 1972 Oct 27;178(4059):361-8. doi: 10.1126/science.178.4059.361. PMID: 5077326.

[45] Rankin NJ, Preiss D, Welsh P, Burgess KE, Nelson SM, Lawlor DA, Sattar N. The emergence of proton nuclear magnetic resonance metabolomics in the cardiovascular arena as viewed from a clinical perspective. Atherosclerosis. 2014 Nov;237(1):287-300. doi: 10.1016/j.atherosclerosis.2014.09.024. Epub 2014 Sep 30. PMID: 25299963; PMCID: PMC4232363.

[46] Antcliffe D, Gordon AC. Metabonomics and intensive care. Crit Care. 2016 Mar 16;20:68. doi: 10.1186/s13054-016-1222-8. PMID: 26984158; PMCID: PMC4794854.

[47] Ernst RR, Bodenhausen G, Wokaun A. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. 1987. Clarendon Press.

[48] Lindon JC. NMR Spectrometers. In J. C. Lindon (Ed.), Encyclopedia of Spectroscopy and Spectrometry (Second Edition) (pp. 1872-1880). Academic Press. 2010. ISBN 9780123744135. DOI: 10.1016/B978-0-12-374413-5.00079-8.

[49] Bingol K, Brüschweiler R. Multidimensional approaches to NMR-based metabolomics. Anal Chem. 2014 Jan 7;86(1):47-57. doi: 10.1021/ac403520j. Epub 2013 Nov 22. PMID: 24195689; PMCID: PMC4467887.

[50] Lewis IA, Schommer SC, Hodis B, Robb KA, Tonelli M, Westler WM, Sussman MR, Markley JL. Method for determining molar concentrations of metabolites in complex solutions from two-dimensional 1H-13C NMR spectra. Anal Chem. 2007 Dec 15;79(24):9385-90. doi: 10.1021/ac071583z. Epub 2007 Nov 7. PMID: 17985927; PMCID: PMC2533272.

[51] Boiteau RM, Hoyt DW, Nicora CD, Kinmonth-Schultz HA, Ward JK, Bingol K. Structure Elucidation of Unknown Metabolites in Metabolomics by Combined NMR and MS/MS Prediction. Metabolites. 2018 Jan 17;8(1):8. doi: 10.3390/metabo8010008. PMID: 29342073; PMCID: PMC5875998.

[52] Corsaro C, Vasi S, Neri F, Mezzasalma AM, Neri G, Fazio E. NMR in Metabolomics: From Conventional Statistics to Machine Learning and Neural Network Approaches. Applied Sciences. 2022; 12(6):2824. https://doi.org/10.3390/app12062824

[53] Xia J, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics. 2010 Sep 15;26(18):2342-4. doi: 10.1093/bioinformatics/btq418. Epub 2010 Jul 13. PMID: 20628077.

[54] Jendoubi T. Approaches to Integrating Metabolomics and Multi-Omics Data: A Primer. Metabolites. 2021 Mar 21;11(3):184. doi: 10.3390/metabo11030184. PMID: 33801081; PMCID: PMC8003953.

[55] Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, Raftery D, Wishart DS. The future of NMR-based metabolomics. Curr Opin Biotechnol. 2017 Feb;43:34-40. doi: 10.1016/j.copbio.2016.08.001. Epub 2016 Aug 28. PMID: 27580257; PMCID: PMC5305426.

[56] Newgard CB. Metabolomics and Metabolic Diseases: Where Do We Stand? Cell Metab. 2017 Jan 10;25(1):43-56. doi: 10.1016/j.cmet.2016.09.018. Epub 2016 Oct 27. PMID: 28094011; PMCID: PMC5245686.

[57] Mochel F, Haller RG. Energy deficit in Huntington disease: why it matters. J Clin Invest. 2011 Feb;121(2):493-9. doi: 10.1172/JCI45691. Epub 2011 Feb 1. PMID: 21285522; PMCID: PMC3026743.

[58] Griffin JL, Shockcor JP. Metabolic profiles of cancer cells. Nat Rev Cancer. 2004 Jul;4(7):551-61. doi: 10.1038/nrc1390. PMID: 15229480.

[59] Zhang A, Sun H, Yan G, Wang P, Wang X. Metabolomics for Biomarker Discovery: Moving to the Clinic. Biomed Res Int. 2015;2015:354671. doi: 10.1155/2015/354671. Epub 2015 May 19. PMID: 26090402; PMCID: PMC4452245.

[60] Wahren-Herlenius M, Dörner T. Immunopathogenic mechanisms of systemic autoimmune disease. Lancet. 2013 Aug 31;382(9894):819-31. doi: 10.1016/S0140-6736(13)60954-X. PMID: 23993191.

[61] Liao KP, Alfredsson L, Karlson EW. Environmental influences on risk for rheumatoid arthritis. Curr Opin Rheumatol. 2009 May;21(3):279-83. doi: 10.1097/BOR.0b013e32832a2e16. PMID: 19318947; PMCID: PMC2898190.

[62] Cui Y, Sheng Y, Zhang X. Genetic susceptibility to SLE: recent progress from GWAS. J Autoimmun. 2013 Mar;41:25-33. doi: 10.1016/j.jaut.2013.01.008. Epub 2013 Feb 6. PMID: 23395425.

[63] Dendrou CA, Fugger L, Friese MA. Immunopathology of multiple sclerosis. Nat Rev Immunol. 2015 Sep 15;15(9):545-58. doi: 10.1038/nri3871. Epub 2015 Aug 7. PMID: 26250739.

[64] Bogunia-Kubik K, Wojtowicz W, Swierkot J, Mielko KA, Qasem B, Wielińska J, Sokolik R, Pruss Ł, Młynarz P. Disease Differentiation and Monitoring of Anti-TNF Treatment in Rheumatoid Arthritis and Spondyloarthropathies. Int J Mol Sci. 2021 Jul 9;22(14):7389. doi: 10.3390/ijms22147389. PMID: 34299006; PMCID: PMC8307996.

[65] Martin FP, Collino S, Rezzi S, Kochhar S. Metabolomic applications to decipher gut microbial metabolic influence in health and disease. Front Physiol. 2012 Apr 26;3:113. doi: 10.3389/fphys.2012.00113. PMID: 22557976; PMCID: PMC3337463.

[66] Kok M, Maton L, van der Peet M, Hankemeier T, van Hasselt JGC. Unraveling antimicrobial resistance using metabolomics. Drug Discov Today. 2022 Jun;27(6):1774-1783. doi: 10.1016/j.drudis.2022.03.015. Epub 2022 Mar 24. PMID: 35341988.

[67] Mielko KA, Jabłoński SJ, Pruss Ł, Milczewska J, Sands D, Łukaszewicz M, Młynarz P. Metabolomics Comparison of Drug-Resistant and Drug-Susceptible Pseudomonas aeruginosa Strain (Intra- and Extracellular Analysis). Int J Mol Sci. 2021 Oct 6;22(19):10820. doi: 10.3390/ijms221910820. PMID: 34639158; PMCID: PMC8509183.

[68] Lyczak JB, Cannon CL, Pier GB. Lung infections associated with cystic fibrosis. Clin Microbiol Rev. 2002 Apr;15(2):194-222. doi: 10.1128/CMR.15.2.194-222.2002. PMID: 11932230; PMCID: PMC118069.

[69] Ghini V, Saccenti E, Tenori L, Assfalg M, Luchinat C. Allostasis and Resilience of the Human Individual Metabolic Phenotype. J Proteome Res. 2015 Jul 2;14(7):2951-62. doi: 10.1021/acs.jproteome.5b00275. Epub 2015 Jun 24. PMID: 26055080.

[70] Axelrod FB, Gold-von Simson G. Hereditary sensory and autonomic neuropathies: types II, III, and IV. Orphanet J Rare Dis. 2007 Oct 3;2:39. doi: 10.1186/1750-1172-2-39. PMID: 17915006; PMCID: PMC2098750.

[71] Gold-von Simson G, Axelrod FB. Familial dysautonomia: update and recent advances. Curr Probl Pediatr Adolesc Health Care. 2006 Jul;36(6):218-37. doi: 10.1016/j.cppeds.2005.12.001. PMID: 16777588.

[72] Costello SM, Cheney AM, Waldum A, Tripet B, Cotrina-Vidal M, Kaufmann H, Norcliffe-Kaufmann L, Lefcort F, Copié V. A Comprehensive NMR Analysis of Serum and Fecal Metabolites in Familial Dysautonomia Patients Reveals Significant Metabolic Perturbations. Metabolites. 2023 Mar 16;13(3):433. doi: 10.3390/metabo13030433. PMID: 36984872; PMCID: PMC10057143.

[73] Cheney AM, Costello SM, Pinkham NV, Waldum A, Broadaway SC, Cotrina-Vidal M, Mergy M, Tripet B, Kominsky DJ, Grifka-Walk HM, Kaufmann H, Norcliffe-Kaufmann L, Peach JT, Bothner B, Lefcort F, Copié V, Walk ST. Gut microbiome dysbiosis drives metabolic dysfunction in Familial dysautonomia. Nat Commun. 2023 Jan 13;14(1):218. doi: 10.1038/s41467-023-35787-8. PMID: 36639365; PMCID: PMC9839693.

[74] Beheshti A, Chakravarty K, Fogle H, Fazelinia H, Silveira WAD, Boyko V, Polo SL, Saravia-Butler AM, Hardiman G, Taylor D, Galazka JM, Costes SV. Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver. Sci Rep. 2019 Dec 16;9(1):19195. doi: 10.1038/s41598-019-55869-2. Erratum in: Sci Rep. 2020 Jan 27;10(1):1517. PMID: 31844325; PMCID: PMC6915713.

[75] Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016 Jul;17(4):628-41. doi: 10.1093/bib/bbv108. Epub 2016 Mar 11. PMID: 26969681; PMCID: PMC4945831.

[76] Chen Y, Li EM, Xu LY. Guide to Metabolomics Analysis: A Bioinformatics Workflow. Metabolites. 2022 Apr 15;12(4):357. doi: 10.3390/metabo12040357. PMID: 35448542; PMCID: PMC9032224.

[77] Gagniuc P. Algorithms in Bioinformatics: Theory and Implementation. 2021. doi:10.1002/9781119698005.

[78] Lesk AM. "bioinformatics". Encyclopedia Britannica, 12 Apr. 2024, https://www.britannica.com/science/bioinformatics. Accessed 30 April 2024.

[79] Moco S. Studying Metabolism by NMR-Based Metabolomics. Front Mol Biosci. 2022 Apr 27;9:882487. doi: 10.3389/fmolb.2022.882487. PMID: 35573745; PMCID: PMC9094115.

[80] O'Shea K, Misra BB. Software tools, databases and resources in metabolomics: updates from 2018 to 2019. Metabolomics. 2020 Mar 7;16(3):36.

[81] Misra BB. New software tools, databases, and resources in metabolomics: updates from 2020. Metabolomics. 2021;17(5):49. Published 2021 May 11. doi:10.1007/s11306-021-01796-1.

[82] Djaffardjy M, Marchment G, Sebe C, Blanchet R, Bellajhame K, Gaignard A, Lemoine F, Cohen-Boulakia S. Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. Comput Struct Biotechnol J. 2023 Mar 7;21:2075-2085. doi: 10.1016/j.csbj.2023.03.003. PMID: 36968012; PMCID: PMC10030817.

[83] Luigi documentation website. https://luigi.readthedocs.io/en/stable/. Accessed 30 April 2024.

[84] Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. Sustainable data analysis with Snakemake. F1000Res. 2021 Jan 18;10:33. doi: 10.12688/f1000research.29032.2. PMID: 34035898; PMCID: PMC8114187.

[85] Tommaso PD, Floden EW, Magis C, Palumbo E, Notredame C. Nextflow : un outil efficace pour l'amélioration de la stabilité numérique des calculs en analyse génomique [Nextflow, an efficient tool to improve computation numerical stability in genomic analysis]. Biol Aujourdhui. 2017;211(3):233-237. French. doi: 10.1051/jbio/2017029. Epub 2018 Feb 7. PMID: 29412134.

[86] Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux journal. 2014;2014(239):2.

[87] Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLOS ONE. 2017 May;12(5):e0177459. doi: 10.1371/journal.pone.0177459.

[88] Jackson M, Kavoussanakis K, Wallace EWJ. Using prototyping to choose a bioinformatics workflow management system. PLoS Comput Biol. 2021 Feb 25;17(2):e1008622. doi: 10.1371/journal.pcbi.1008622. PMID: 33630841; PMCID: PMC7906312.

[89] Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020 Mar;38(3):276-278. doi: 10.1038/s41587-020-0439-x. PMID: 32055031.

[90] Karaman I. Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis. Adv Exp Med Biol. 2017;965:145-161. doi: 10.1007/978-3-319-47656-8_6. PMID: 28132179.

[91] Schiffman C, Petrick L, Perttula K, Yano Y, Carlsson H, Whitehead T, Metayer C, Hayes J, Rappaport S, Dudoit S. Filtering procedures for untargeted LC-MS metabolomics data. BMC Bioinformatics. 2019 Jun 14;20(1):334. doi: 10.1186/s12859-019-2871-9. PMID: 31200644; PMCID: PMC6570933.

[92] Xi Y, Rocke DM. Baseline correction for NMR spectroscopic metabolomics data analysis. BMC Bioinformatics. 2008 Jul 29;9:324. doi: 10.1186/1471-2105-9-324. PMID: 18664284; PMCID: PMC2516527.

[93] Zhang F, Tang X, Tong A, Wang B, Wang J. An Automatic Baseline Correction Method Based on the Penalized Least Squares Method. Sensors (Basel). 2020 Apr 3;20(7):2015. doi: 10.3390/s20072015. PMID: 32260258; PMCID: PMC7181009.

[94] Vu TN, Laukens K. Getting your peaks in line: a review of alignment methods for NMR spectral data. Metabolites. 2013 Apr 15;3(2):259-76. doi: 10.3390/metabo3020259. PMID: 24957991; PMCID: PMC3901265.

[95] Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem. 2006 Jul 1;78(13):4281-90. doi: 10.1021/ac051632c. PMID: 16808434.

[96] van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006 Jun 8;7:142. doi: 10.1186/1471-2164-7-142. PMID: 16762068; PMCID: PMC1534033.

[97] Korman A, Oh A, Raskind A, Banks D. Statistical methods in metabolomics. Methods Mol Biol. 2012;856:381-413. doi: 10.1007/978-1-61779-585-5_16. PMID: 22399468.

[98] Flores JE, Claborne DM, Weller ZD, Webb-Robertson BM, Waters KM, Bramer LM. Missing data in multi-omics integration: Recent advances through artificial intelligence. Front Artif Intell. 2023 Feb 9;6:1098308. doi: 10.3389/frai.2023.1098308. PMID: 36844425; PMCID: PMC9949722.

[99] Pandya A, Howard MJ, Zloh M, Dalby PA. An Evaluation of the Potential of NMR Spectroscopy and Computational Modelling Methods to Inform Biopharmaceutical Formulations. Pharmaceutics. 2018 Sep 21;10(4):165. doi: 10.3390/pharmaceutics10040165. PMID: 30248922; PMCID: PMC6320905.

[100] Karaman I, Ferreira DL, Boulangé CL, Kaluarachchi MR, Herrington D, Dona AC, Castagné R, Moayyeri A, Lehne B, Loh M, de Vries PS, Dehghan A, Franco OH, Hofman A, Evangelou E, Tzoulaki I, Elliott P, Lindon JC, Ebbels TM. Workflow for Integrated Processing of Multicohort Untargeted 1H NMR Metabolomics Data in Large-Scale Metabolic Epidemiology. J Proteome Res. 2016 Dec 2;15(12):4188-4194. doi: 10.1021/acs.jproteome.6b00125. Epub 2016 Oct 6. PMID: 27628670.

[101] Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. Metabolites. 2012 Oct 18;2(4):775-95. doi: 10.3390/metabo2040775. PMID: 24957762; PMCID: PMC3901240.

[102] Xu S, Bai C, Chen Y, Yu L, Wu W, Hu K. Comparing univariate filtration preceding and succeeding PLS-DA analysis on the differential variables/metabolites identified from untargeted LC-MS metabolomics data. Anal Chim Acta. 2024 Jan 25;1287:342103. doi: 10.1016/j.aca.2023.342103. Epub 2023 Dec 7. PMID: 38182346.

[103] Saccenti E, Hoefsloot H, Smilde A, Westerhuis J & Hendriks M. Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics. 2013 Oct 15. doi: 10.1007/s11306-013-0598-6

[104] Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006 Nov 28, 171–196. doi: 10.1007/s11306-006-0037-z

[105] Worley B, Powers R. Multivariate Analysis in Metabolomics. Curr Metabolomics. 2013;1(1):92-107. doi: 10.2174/2213235X11301010092. PMID: 26078916; PMCID: PMC4465187.

[106] Cheriyadat, A. & Bruce, Lori. (2003). Why principal component analysis is not an appropriate feature extraction method for hyperspectral data. 3420 - 3422 vol.6. 10.1109/IGARSS.2003.1294808.

[107] Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. Comput Struct Biotechnol J. 2013 Mar 22;4:e201301009. doi: 10.5936/csbj.201301009. PMID: 24688690; PMCID: PMC3962125.

[108] Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, Goodacre R. A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. Anal Chim Acta. 2015 Jun 16;879:10-23. doi: 10.1016/j.aca.2015.02.012. Epub 2015 Feb 11. PMID: 26002472.

[109] Galal A, Talal M, Moustafa A. Applications of machine learning in metabolomics: Disease modeling and classification. Front Genet. 2022 Nov 24;13:1017340. doi: 10.3389/fgene.2022.1017340. PMID: 36506316; PMCID: PMC9730048.

[110] Chen D, Wang Z, Guo D, Orekhov V, Qu X. Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy. Chemistry. 2020 Aug 17;26(46):10391-10401. doi: 10.1002/chem.202000246. Epub 2020 Jun 25. PMID: 32251549.

[111] Pomyen Y, Wanichthanarak K, Poungsombat P, Fahrmann J, Grapov D, Khoomrung S. Deep metabolome: Applications of deep learning in metabolomics. Comput Struct Biotechnol J. 2020 Oct 1;18:2818-2825. doi: 10.1016/j.csbj.2020.09.033. PMID: 33133423; PMCID: PMC7575644.

[112] Grissa D, Pétéra M, Brandolini M, Napoli A, Comte B, Pujos-Guillot E. Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. Front Mol Biosci. 2016 Jul 8;3:30. doi: 10.3389/fmolb.2016.00030. PMID: 27458587; PMCID: PMC4937038.

[113] Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019 Sep;112:103375. doi: 10.1016/j.compbiomed.2019.103375. Epub 2019 Jul 31. PMID: 31382212.

[114] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x. Epub 2021 Mar 22. PMID: 33778771; PMCID: PMC7983091.

[115] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006 Feb 23;7:91. doi: 10.1186/1471-2105-7-91. PMID: 16504092; PMCID: PMC1397873.

[116] Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Lê Cao KA. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019 Sep 1;35(17):3055-3062. doi: 10.1093/bioinformatics/bty1054. PMID: 30657866; PMCID: PMC6735831.

[117] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375. doi: 10.1371/journal.pcbi.1002375. Epub 2012 Feb 23. PMID: 22383865; PMCID: PMC3285573.

[118] Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. 2020 Jan;29(1):28-35. doi: 10.1002/pro.3711. Epub 2019 Aug 29. PMID: 31423653; PMCID: PMC6933857.

[119] Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, Evelo CT. PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol. 2015 Feb 23;11(2):e1004085. doi: 10.1371/journal.pcbi.1004085. PMID: 25706687; PMCID: PMC4338111.

[120] Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, Mélius J, Waagmeester A, Sinha SR, Miller R, Coort SL, Cirillo E, Smeets B, Evelo CT, Pico AR. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016 Jan 4;44(D1):D488-94. doi: 10.1093/nar/gkv1024. Epub 2015 Oct 19. PMID: 26481357; PMCID: PMC4702772.

[121] Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan;37(1):1-13. doi: 10.1093/nar/gkn923. Epub 2008 Nov 25. PMID: 19033363; PMCID: PMC2615629.

[122] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284-7. doi: 10.1089/omi.2011.0118. Epub 2012 Mar 28. PMID: 22455463; PMCID: PMC3339379.

[123] Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013 Jul 15;29(14):1830-1. doi: 10.1093/bioinformatics/btt285. Epub 2013 Jun 4. PMID: 23740750; PMCID: PMC3702256.

[124] Liu T, Feenstra KA, Huang Z, Heringa J. Mining literature and pathway data to explore the relations of ketamine with neurotransmitters and gut microbiota using a knowledge-graph. Bioinformatics. 2024 Jan 2;40(1):btad771. doi: 10.1093/bioinformatics/btad771. PMID: 38147362; PMCID: PMC10769815.

[125] Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, Coscia F, Albrechtsen NJW, Mundt F, Jensen LJ, Mann M. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol. 2022 May;40(5):692-702. doi: 10.1038/s41587-021-01145-6. Epub 2022 Jan 31. PMID: 35102292; PMCID: PMC9110295.

[126] Delmas M, Filangi O, Paulhe N, Vinson F, Duperier C, Garrier W, Saunier PE, Pitarch Y, Jourdan F, Giacomoni F, Frainay C. FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases. Bioinformatics. 2021 Nov 5;37(21):3896-3904. doi: 10.1093/bioinformatics/btab627. PMID: 34478489; PMCID: PMC8570811.

[127] Can H, Chanumolu SK, Nielsen BD, Alvarez S, Naldrett MJ, Ünlü G, Otu HH. Integration of Meta-Multi-Omics Data Using Probabilistic Graphs and External Knowledge. Cells. 2023 Aug 4;12(15):1998. doi: 10.3390/cells12151998. PMID: 37566077; PMCID: PMC10417344.

[128] Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020;18:1414-1428. Published 2020 Jun 2. doi:10.1016/j.csbj.2020.05.017

[129] Renaux A, Terwagne C, Cochez M, Tiddi I, Nowé A, Lenaerts T. A knowledge graph approach to predict and interpret disease-causing gene interactions. BMC Bioinformatics. 2023 Aug 29;24(1):324. doi: 10.1186/s12859-023-05451-5. PMID: 37644440; PMCID: PMC10463539.

[130] Banimfreg BH, Shamayleh A, Alshraideh H. Survey for Computer-Aided Tools and Databases in Metabolomics. Metabolites. 2022 Oct 21;12(10):1002. doi: 10.3390/metabo12101002. PMID: 36295904; PMCID: PMC9610953.

[131] Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, Sumner S, Subramaniam S. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 2016 Jan 4;44(D1):D463-70. doi: 10.1093/nar/gkv1042. Epub 2015 Oct 13. PMID: 26467476; PMCID: PMC4702780.

[132] van Rijswijk M, Beirnaert C, Caron C, Cascante M, Dominguez V, Dunn WB, Ebbels TMD, Giacomoni F, Gonzalez-Beltran A, Hankemeier T, Haug K, Izquierdo-Garcia JL, Jimenez RC, Jourdan F, Kale N, Klapa MI et al. The future of metabolomics in ELIXIR. F1000Res. 2017 Sep 6;6:ELIXIR-1649. doi: 10.12688/f1000research.12342.2. PMID: 29043062; PMCID: PMC5627583.

[133] Zulfiqar M, Crusoe MR, König-Ries B, Steinbeck C, Peters K, Gadelha L. Implementation of FAIR Practices in Computational Metabolomics Workflows-A Case Study. Metabolites. 2024 Feb 10;14(2):118. doi: 10.3390/metabo14020118. PMID: 38393009; PMCID: PMC10891576.

[134] Hoch JC, Baskaran K, Burr H, Chin J, Eghbalnia HR, Fujiwara T, Gryk MR, Iwata T, Kojima C, Kurisu G, Maziuk D, Miyanoiri Y, Wedell JR, Wilburn C, Yao H, Yokochi M. Biological

Magnetic Resonance Data Bank. Nucleic Acids Res. 2023 Jan 6;51(D1):D368-D376. doi: 10.1093/nar/gkac1050. PMID: 36478084; PMCID: PMC9825541.

[135] Velankar S, Burley SK, Kurisu G, Hoch JC, Markley JL. The Protein Data Bank Archive. Methods Mol Biol. 2021;2305:3-21. doi: 10.1007/978-1-0716-1406-8_1. PMID: 33950382.

[136] Baskaran K, Craft DL, Eghbalnia HR, Gryk MR, Hoch JC, Maciejewski MW, Schuyler AD, Wedell JR, Wilburn CW. Merging NMR Data and Computation Facilitates Data-Centered Research. Front Mol Biosci. 2022 Jan 17;8:817175. doi: 10.3389/fmolb.2021.817175. PMID: 35111815; PMCID: PMC8802229.

[137] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

[138] Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, Sen S, Gutmanas A, Armstrong DR, Berrisford JM, Chen L, Chen M, Di Costanzo L, Dimitropoulos D, Gao G, Ghosh S et al. OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive. Structure. 2017 Mar 7;25(3):536-545. doi: 10.1016/j.str.2017.01.004. Epub 2017 Feb 9. PMID: 28190782; PMCID: PMC5360273.

[139] Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL. BioMagResBank. Nucleic Acids Res. 2008 Jan;36(Database issue):D402-8. doi: 10.1093/nar/gkm957. Epub 2007 Nov 4. PMID: 17984079; PMCID: PMC2238925.

[140] Ulrich EL, Baskaran K, Dashti H, Ioannidis YE, Livny M, Romero PR, Maziuk D, Wedell JR, Yao H, Eghbalnia HR, Hoch JC, Markley JL. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. J Biomol NMR. 2019 Feb;73(1-2):5-9. doi: 10.1007/s10858-018-0220-3. Epub 2018 Dec 22. PMID: 30580387; PMCID: PMC6441402.

[141] Dashti H, Wedell JR, Westler WM, Markley JL, Eghbalnia HR. Automated evaluation of consistency within the PubChem Compound database. Sci Data. 2019 Feb 19;6:190023. doi: 10.1038/sdata.2019.23. PMID: 30778259; PMCID: PMC6380220.

[142] Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 2020 Jan 8;48(D1):D445-D453. doi: 10.1093/nar/gkz862. PMID: 31586394; PMCID: PMC6943030.

[143] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30. doi: 10.1093/nar/28.1.27. PMID: 10592173; PMCID: PMC102409.

[144] Trupp M, Altman T, Fulcher CA, Caspi R, Krummenacker M, Paley S, Karp PD. Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. Genome Biol. 2010;11(Suppl 1):O12. doi: 10.1186/gb-2010-11-s1-o12. Epub 2010 Oct 11. PMCID: PMC3026227.

[145] Moore LR, Caspi R, Campbell DA, Casey JR, Crevecoeur S, Lea-Smith DJ, Long B, Omar NM, Paley SM, Schmelling NM, Torrado A, Zehr JP, Karp PD. CyanoCyc cyanobacterial web portal. Front Microbiol. 2024 Jan 31;15:1340413. doi: 10.3389/fmicb.2024.1340413. PMID: 38357349; PMCID: PMC10864581.

[146] Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics. 2013 Mar 27;14:112. doi: 10.1186/1471-2105-14-112. PMID: 23530693; PMCID: PMC3665663.

[147] Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, Bridge A, Bougueleret L, Xenarios I, Viari A. UniPathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Res. 2012 Jan;40(Database issue):D761-9. doi: 10.1093/nar/gkr1023. Epub 2011 Nov 18. PMID: 22102589; PMCID: PMC3245108.

[148] Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, Haw R, Jassal B, Matthews L, May B, Petryszak R, Ragueneau E, Rothfels K, Sevilla C, Shamovsky V, Stephan R et al. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res. 2024 Jan 5;52(D1):D672-D678. doi: 10.1093/nar/gkad1025. PMID: 37941124; PMCID: PMC10767911.

[149] Karp PD, Billington R, Holland TA, Kothari A, Krummenacker M, Weaver D, Latendresse M, Paley S. Computational Metabolomics Operations at BioCyc.org. Metabolites. 2015 May 22;5(2):291-310. doi: 10.3390/metabo5020291. PMID: 26011592; PMCID: PMC4495374.

[150] Latendresse M. Efficiently gap-filling reaction networks. BMC Bioinformatics. 2014 Jun 28;15:225. doi: 10.1186/1471-2105-15-225. PMID: 24972703; PMCID: PMC4094995.

[151] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012 Jan;40(Database issue):D109-14. doi: 10.1093/nar/gkr988. Epub 2011 Nov 10. PMID: 22080510; PMCID: PMC3245020.

[152] Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S et al. HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007 Jan;35(Database issue):D521-6. doi: 10.1093/nar/gkl923. PMID: 17202168; PMCID: PMC1899095.

[153] Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F et al. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D801-7. doi: 10.1093/nar/gks1065. Epub 2012 Nov 17. PMID: 23161693; PMCID: PMC3531200.

[154] Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, Sayeeda Z, Lo E, Assempour N, Berjanskii M, Singhal S, Arndt D, Liang Y et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D608-D617. doi: 10.1093/nar/gkx1089. PMID: 29140435; PMCID: PMC5753273.

[155] Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, Dizon R, Sayeeda Z, Tian S, Lee BL, Berjanskii M, Mah R, Yamamoto M, Jovel J, Torres-Calzada C, Hiebert-Giesbrecht M et al.

HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D622-D631. doi: 10.1093/nar/gkab1062. PMID: 34986597; PMCID: PMC8728138.

[156] Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E et al. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. 2009 Jan;37(Database issue):D603-10. doi: 10.1093/nar/gkn810. Epub 2008 Oct 25. PMID: 18953024; PMCID: PMC2686599.

[157] Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, González-Beltrán A, Sansone SA, Griffin JL, Steinbeck C. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res. 2013 Jan;41(Database issue):D781-6. doi: 10.1093/nar/gks1004. Epub 2012 Oct 29. PMID: 23109552; PMCID: PMC3531110.

[158] Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res. 2020 Jan 8;48(D1):D440-D444. doi: 10.1093/nar/gkz1019. PMID: 31691833; PMCID: PMC7145518.

[159] Yurekten O, Payne T, Tejera N, Amaladoss FX, Martin C, Williams M, O'Donovan C. MetaboLights: open data repository for metabolomics. Nucleic Acids Res. 2024 Jan 5;52(D1):D640-D646. doi: 10.1093/nar/gkad1045. PMID: 37971328; PMCID: PMC10767962.

[160] Nyholm L, Koziol A, Marcos S, Botnen AB, Aizpurua O, Gopalakrishnan S, Limborg MT, Gilbert MTP, Alberdi A. Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research. iScience. 2020 Aug 21;23(8):101414. doi: 10.1016/j.isci.2020.101414. Epub 2020 Jul 25. PMID: 32777774; PMCID: PMC7416341.

[161] Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018 Mar 15;24(6):1248-1259. doi: 10.1158/1078-0432.CCR-17-0853. Epub 2017 Oct 5. PMID: 28982688; PMCID: PMC6050171.

[162] Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. Proteomics. 2016 Mar;16(5):741-58. doi: 10.1002/pmic.201500396. PMID: 26677817.

[163] Liang A, Kong Y, Chen Z, Qiu Y, Wu Y, Zhu X, Li Z. Advancements and applications of single-cell multi-omics techniques in cancer research: Unveiling heterogeneity and paving the way for precision therapeutics. Biochem Biophys Rep. 2023 Nov 29;37:101589. doi: 10.1016/j.bbrep.2023.101589. PMID: 38074997; PMCID: PMC10698529.

[164] Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010 Jul 23;11:395. doi: 10.1186/1471-2105-11-395. PMID: 20650010; PMCID: PMC2918584.

[165] Domingo-Almenara X, Brezmes J, Vinaixa M, Samino S, Ramirez N, Ramon-Krauel M, Lerin C, Díaz M, Ibáñez L, Correig X, Perera-Lluna A, Yanes O. eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. Anal Chem. 2016 Oct 4;88(19):9821-9829. doi: 10.1021/acs.analchem.6b02927. Epub 2016 Sep 14. PMID: 27584001.

[166] Products and software of Bruker. https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html. Accessed 30 April 2024

[167] Manual of MestReNova 15.0.01, https://mestrelab.com/downloads/mnova/manuals/MestReNova-15.0.1_Manual.pdf. Accessed 30 April 2024

[168] Website of Chenomx. https://www.chenomx.com/. Accessed 30 April 2024

[169] Libraries of Chenomx. https://www.chenomx.com/libraries/. Accessed 30 April 2024

[170] Jacob D, Deborde C, Lefebvre M, Maucourt M, Moing A. NMRProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics. Metabolomics. 2017;13(4):36. doi: 10.1007/s11306-017-1178-y. Epub 2017 Feb 17. PMID: 28261014; PMCID: PMC5313591.

[171] Overview of NMRProcFlow. https://nmrprocflow.org/b_introduction. Accessed 30 April 2024

[172] Martin M, Legat B, Leenders J, Vanwinsberghe J, Rousseau R, Boulanger B, Eilers PHC, De Tullio P, Govaerts B. PepsNMR for 1H NMR metabolomic data pre-processing. Anal Chim Acta. 2018 Aug 17;1019:1-13. doi: 10.1016/j.aca.2018.02.067. Epub 2018 Mar 12. PMID: 29625674.

[173] Lefort G, Liaubet L, Canlet C, Tardivel P, Père MC, Quesnel H, Paris A, Iannuccelli N, Vialaneix N, Servien R. ASICS: an R package for a whole analysis workflow of 1D 1H NMR spectra. Bioinformatics. 2019 Nov 1;35(21):4356-4363. doi: 10.1093/bioinformatics/btz248. PMID: 30977816.

[174] Tardivel P, Canlet C, Lefort G, Tremblay-Franco M, Debrauwer L, Concordet D, Servien R. ASICS: An automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. Metabolomics. 2017;13 doi: 10.1007/s11306-017-1244-5.

[175] Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W652-60. doi: 10.1093/nar/gkp356. Epub 2009 May 8. PMID: 19429898; PMCID: PMC2703878.

[176] Pang Z, Lu Y, Zhou G, Hui F, Xu L, Viau C, Spigelman A, MacDonald P, Wishart D, Li S, Xia J. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. Nucleic Acids Res. 2024. doi: 10.1093/nar/gkae253.

[177] Pang Z, Xu L, Viau C, Lu Y, Salavati R, Basu N, Xia J. MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics. Nat Commun. 2024. doi: 10.1038/s41467-024-48009-6.

[178] Pruss Ł, Gniewek O, Jetka T, Wojtowicz W, Milanowska-Zabel K, Młynarz P. NASQQ: Nextflow automatization and standardization for 1H NMR metabolomics data preparation and analysis [manuscript under resubmission]. Oxford GigaScience, Technical note. 2024.

[179] Frasso G, Eilers PH. L- and V-curves for optimal smoothing. Stat Model. 2015;15(1):91-111. doi:10.1177/1471082X14549288.

[180] Goldstein A, Powis RL. Medical Ultrasonic Diagnostics. In: Thurston RN, Pierce AD, Papadakis EP, editors. Physical Acoustics. Academic Press; 1999. Volume 23. Pages 43-195. ISSN 0893-388X. ISBN 9780124779235. doi:10.1016/S0893-388X(99)80012-8.

[181] PepsNMR R package documentation. https://www.bioconductor.org/packages/release/bioc/manuals/PepsNMR/man/PepsNMR.pdf. Accessed 30 May 2024.

[182] Eilers P, Boelens H. Baseline Correction with Asymmetric Least Squares Smoothing. Unpubl Manuscr. 2005.

[183] Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem. 2006 Jul 1;78(13):4281-90. doi: 10.1021/ac051632c. PMID: 16808434.

[184] ASICS R package documentation. https://bioconductor.org/packages/release/bioc/vignettes/ASICS/inst/doc/ASICSUsersGuide.html. Accessed 30 May 2024.

[185] Wong JW, Durante C, Cartwright HM. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. Anal Chem. 2005 Sep 1;77(17):5655-61. doi: 10.1021/ac050619p. PMID: 16131078.

[186] Daneshvar A, Mousa G. Regression shrinkage and selection via least quantile shrinkage and selection operator. PLoS One. 2023 Feb 16;18(2):e0266267. doi: 10.1371/journal.pone.0266267. PMID: 36795659; PMCID: PMC9934385.

[187] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300. PMID: 7832670; PMCID: PMC1336731.

[188] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. PMID: 0205384.

[189] Gosset WS. The Probable Error of a Mean. Biometrika. 1908;6(1):1–25. doi: 10.2307/2331554.

[190] Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics. 1947;18(1):50–60. doi: 10.1214/aoms/1177730491.

[191] Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001 Nov 1;125(1-2):279-84. doi: 10.1016/s0166-4328(01)00297-2. PMID: 11682119.

[192] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: Identifying Density-based Local Outliers (PDF). In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD. 2000. pp. 93–104. doi:10.1145/335191.335388. ISBN 1-58113-217-4.

[193] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1-22. PMID: 20808728; PMCID: PMC2929880.

[194] Breiman L. Random forests. Machine Learning. 2001;45(1):5-32. doi: 10.1023/A:1010933404324.

[195] Junge MRJ, Dettori JR. ROC Solid: Receiver Operator Characteristic (ROC) Curves as a Foundation for Better Diagnostic Tests. Global Spine J. 2018 Jun;8(4):424-429. doi: 10.1177/2192568218778294. Epub 2018 May 23. PMID: 29977728; PMCID: PMC6022965.

[196] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 2017. pp. 4765-4774.

[197] SHAP Python package documentation. https://shap.readthedocs.io/en/latest/index.html. Accessed 30 May 2024.

[198] Picart-Armada S, Fernández-Albert F, Vinaixa M, Yanes O, Perera-Lluna A. FELLA: an R package to enrich metabolomics data. BMC Bioinformatics. 2018 Dec 22;19(1):538. doi: 10.1186/s12859-018-2487-5. PMID: 30577788; PMCID: PMC6303911.

[199] Picart-Armada S, Fernández-Albert F, Vinaixa M, Rodríguez MA, Aivio S, Stracker TH, Yanes O, Perera-Lluna A. Null diffusion-based enrichment for metabolomics data. PLoS One. 2017 Dec 6;12(12):e0189012. doi: 10.1371/journal.pone.0189012. PMID: 29211807; PMCID: PMC5718512.

[200] Cohen D. On Holy Wars and a Plea for Peace. Computer. 1981 Oct;14(10):48-54. doi: 10.1109/C-M.1981.220208.

[201] Lefort G, Liaubet L, Marty-Gasset N, Canlet C, Vialaneix N, Servien R. Joint Automatic Metabolite Identification and Quantification of a Set of 1H NMR Spectra. Anal Chem. 2021;93(5):2861-2870. doi:10.1021/acs.analchem.0c04232

[202] Wojciechowski S, Majchrzak-Górecka M, Biernat P, Odrzywołek K, Pruss Ł, Zych K, Jan Majta, Milanowska-Zabel K. Machine learning on the road to unlocking microbiota's potential for boosting immune checkpoint therapy. Int J Med Microbiol. 2022 Oct;312(7):151560. doi: 10.1016/j.ijmm.2022.151560. Epub 2022 Sep 9. PMID: 36113358.

[203] Hutson SM, Berkich D, Drown P, Xu B, Aschner M, LaNoue KF. Role of branched-chain aminotransferase isoenzymes and gabapentin in neurotransmitter metabolism. J Neurochem. 1998;71(2):863-874. doi:10.1046/j.1471-4159.1998.71020863.x. PMID: 9681479.

[204] Thomas SD, Jha NK, Ojha S, Sadek B. mTOR Signaling Disruption and Its Association with the Development of Autism Spectrum Disorder. Molecules. 2023;28(4):1889. doi:10.3390/molecules28041889. PMID: 36838876; PMCID: PMC9964164.

[205] Puleston DJ, Buck MD, Klein Geltink RI, Kyle RL, Caputa G, O'Sullivan D, Cameron AM, Castoldi A, Musa Y, Kabat AM, Zhang Y, Flachsmann LJ, Field CS, Patterson AE, Scherer S, Alfei F, Baixauli F, Austin SK, Kelly B, Matsushita M, Curtis JD, Grzes KM, Villa M et al. Polyamines and eIF5A Hypusination Modulate Mitochondrial Respiration and Macrophage Activation. Cell Metabolism. 2019 Aug 6;30(2):352-363.e8. DOI: 10.1016/j.cmet.2019.05.003. PMID: 31130465; PMCID: PMC6688828.

[206] Axelrod FB. Familial dysautonomia: a review of the current pharmacological treatments. Expert Opin Pharmacother. 2005;6(4):561-567. doi:10.1517/14656566.6.4.561.

[207] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118-27. doi: 10.1093/biostatistics/kxj037. Epub 2006 Apr 21. PMID: 16632515.

[208] Suomi T, Seyednasrollah F, Jaakkola MK, Faux T, Elo LL. ROTS: An R package for reproducibility-optimized statistical testing. PLoS Comput Biol. 2017 May 25;13(5):e1005562. doi: 10.1371/journal.pcbi.1005562. PMID: 28542205; PMCID: PMC5470739.

[209] MaCWorP - Massive aCcessible Workflow Platform GitHub repository https://github.com/cubimedrub/macworp. Accessed 08 July 2024.

Supplementary Materials

Supplementary Table 1: Univariate results of T-test of metabolites abundances between FD patients and their relatives (significant metabolite with a p-value below the threshold of 0.05 are marked with an asterisk [*]).

No	Feature	Test	Statistic	p-value	FDR
1	*D-fucose**	T-test	-2.3416	0.0212	0.9727
2	methylguanidine	T-test	-1.9139	0.0585	0.9727
3	L-citrulline	T-test	-1.4594	0.1476	0.9727
4	CMP	T-test	1.3700	0.1737	0.9727
5	formate	T-test	-1.3466	0.1811	0.9727
6	phenylglyoxylic acid	T-test	1.2604	0.2104	0.9727
7	L-alanine	T-test	-1.2529	0.2131	0.9727
8	4-hydroxyphenyl acetic acid	T-test	1.0714	0.2865	0.9727
9	adipic acid	T-test	-1.0384	0.3016	0.9727
10	ethylmalonic acid	T-test	-0.9263	0.3565	0.9727
11	2-oxoisovalerate	T-test	-0.8659	0.3885	0.9727
12	D-glucose	T-test	0.7849	0.4343	0.9727
13	UTP	T-test	0.4754	0.6355	0.9727
14	pyroglutamic acid	T-test	0.3328	0.7399	0.9727
15	hippuric acid	T-test	-0.3067	0.7596	0.9727
16	1-methylhydantoin	T-test	0.2001	0.8417	0.9727
17	nicotinuric acid	T-test	-0.0530	0.9577	0.9727
18	L-asparagine	T-test	0.0453	0.9639	0.9727

Supplementary Table 2: Univariate results of Mann-Whitney U test of metabolites abundances between FD patients and their relatives (significant metabolite with a p-value below the threshold of 0.05 are marked with an asterisk [*]).

No	Feature	Test	Statistic	p-value	FDR
1	*uracil**	Mann-Whitney U	1773	0.0006	0.1210
2	*beta-hydroxyisovaleric acid**	Mann-Whitney U	944	0.0259	0.9727
3	*L-aspartate**	Mann-Whitney U	962	0.0353	0.9727
4	*hypoxanthine**	Mann-Whitney U	977	0.0452	0.9727
5	*5-amino valeric acid**	Mann-Whitney U	978	0.0459	0.9727
6	valerate	Mann-Whitney U	1004	0.0688	0.9727
7	choline chloride	Mann-Whitney U	1020	0.0872	0.9727
8	2-picolinic acid	Mann-Whitney U	1026	0.0950	0.9727
9	cytosine	Mann-Whitney U	1514	0.1005	0.9727
10	methylmalonic acid	Mann-Whitney U	1031	0.1008	0.9727
11	levulinic acid	Mann-Whitney U	1031	0.1019	0.9727
12	L-arabitol	Mann-Whitney U	1038	0.1041	0.9727
13	2-methylglutaric acid	Mann-Whitney U	1038	0.1123	0.9727
14	glyceric acid	Mann-Whitney U	1503	0.1170	0.9727
15	L-tyrosine	Mann-Whitney U	1047	0.1268	0.9727
16	allantoin	Mann-Whitney U	1486	0.1463	0.9727
17	glycogen	Mann-Whitney U	1484	0.1503	0.9727
18	syringic acid	Mann-Whitney U	1483	0.1522	0.9727
19	L-cysteine	Mann-Whitney U	1079	0.1703	0.9727
20	L-leucine	Mann-Whitney U	1074	0.1792	0.9727
21	7-methylxanthine	Mann-Whitney U	1083.5	0.2007	0.9727
22	trigonelline	Mann-Whitney U	1086	0.2069	0.9727
23	2-amino adipic acid	Mann-Whitney U	1087	0.2095	0.9727
24	indoxyl sulfate	Mann-Whitney U	1088	0.2120	0.9727

No	Feature	Test	Statistic	p-value	FDR
25	adenosine	Mann-Whitney U	1455	0.2145	0.9727
26	3-hydroxybutyrate	Mann-Whitney U	1091	0.2196	0.9727
27	putrescine	Mann-Whitney U	1452	0.2222	0.9727
28	D-fructose	Mann-Whitney U	1450	0.2274	0.9727
29	acetoacetate	Mann-Whitney U	1101	0.2459	0.9727
30	L-arginine	Mann-Whitney U	1413	0.2470	0.9727
31	L-valine	Mann-Whitney U	1102	0.2490	0.9727
32	L-threonine	Mann-Whitney U	1103	0.2518	0.9727
33	lactose	Mann-Whitney U	1104	0.2546	0.9727
34	1,3-diaminopropane	Mann-Whitney U	1439	0.2575	0.9727
35	3-methyladipic acid	Mann-Whitney U	1106	0.2603	0.9727
36	trans-acotinic acid	Mann-Whitney U	1432	0.2687	0.9727
37	2-deoxyadenosine	Mann-Whitney U	1430	0.2841	0.9727
38	L-isoleucine	Mann-Whitney U	1114	0.2841	0.9727
39	GABA	Mann-Whitney U	1115	0.2872	0.9727
40	GTP	Mann-Whitney U	1116	0.2903	0.9727
41	D-glucose-6-phosphate	Mann-Whitney U	1426	0.2961	0.9727
42	4-ethyl phenol	Mann-Whitney U	1121	0.3060	0.9727
43	citrate	Mann-Whitney U	1125	0.3069	0.9727
44	trans-4-hydroxy-L-proline	Mann-Whitney U	1418	0.3081	0.9727
45	L-carnosine	Mann-Whitney U	1421	0.3125	0.9727
46	CDP	Mann-Whitney U	1420	0.3158	0.9727
47	kynurenic acid	Mann-Whitney U	1417	0.3257	0.9727
48	lactate	Mann-Whitney U	1131	0.3393	0.9727
49	pimelic acid	Mann-Whitney U	1134	0.3497	0.9727
50	N-acetyl-L-aspartic acid	Mann-Whitney U	1137	0.3603	0.9727

No	Feature	Test	Statistic	p-value	FDR
51	adenine	Mann-Whitney U	1407	0.3603	0.9727
52	L-serine	Mann-Whitney U	1140	0.3634	0.9727
53	glycerophosphocholine	Mann-Whitney U	1402	0.3664	0.9727
54	dimethylglycine	Mann-Whitney U	1401	0.3822	0.9727
55	ethanolamine	Mann-Whitney U	1400	0.3859	0.9727
56	myo-inositol	Mann-Whitney U	1146	0.3906	0.9727
57	pantothenic acid	Mann-Whitney U	1396	0.4009	0.9727
58	IMP	Mann-Whitney U	1394	0.4083	0.9727
59	levoglucosan	Mann-Whitney U	1390	0.4242	0.9727
60	dehydroascorbic acid	Mann-Whitney U	1388	0.4321	0.9727
61	ATP	Mann-Whitney U	1388	0.4321	0.9727
62	dihydro thymine	Mann-Whitney U	1160.5	0.4472	0.9727
63	L-ornithine	Mann-Whitney U	1161	0.4480	0.9727
64	spermidine	Mann-Whitney U	1381	0.4518	0.9727
65	threonic acid	Mann-Whitney U	1364.5	0.4529	0.9727
66	L-tryptophane	Mann-Whitney U	1162	0.4564	0.9727
67	dimethyl sulfone	Mann-Whitney U	1177	0.4602	0.9727
68	3-phenylpropionic acid	Mann-Whitney U	1381	0.4606	0.9727
69	L-glutathione-reduced	Mann-Whitney U	1374	0.4633	0.9727
70	CTP	Mann-Whitney U	1168	0.4815	0.9727
71	nicotinic acid	Mann-Whitney U	1170	0.4900	0.9727
72	betaine	Mann-Whitney U	1170.5	0.4921	0.9727
73	trimethylamine	Mann-Whitney U	1372	0.4984	0.9727
74	trans-ferulic acid	Mann-Whitney U	1369.5	0.5026	0.9727
75	acetone	Mann-Whitney U	1176	0.5160	0.9727
76	quinolinic acid	Mann-Whitney U	1176	0.5160	0.9727

No	Feature	Test	Statistic	p-value	FDR
77	L-histidine	Mann-Whitney U	1365	0.5293	0.9727
78	phenethylamine	Mann-Whitney U	1179	0.5293	0.9727
79	2-deoxyguanosine	Mann-Whitney U	1365	0.5293	0.9727
80	guanidinoacetic acid	Mann-Whitney U	1362.5	0.5293	0.9727
81	homovanillic acid	Mann-Whitney U	1363	0.5376	0.9727
82	L-cystine	Mann-Whitney U	1363	0.5382	0.9727
83	TMAO	Mann-Whitney U	1183	0.5413	0.9727
84	3-methyl-L-histidine	Mann-Whitney U	1357	0.5556	0.9727
85	methanol	Mann-Whitney U	1189	0.5672	0.9727
86	L-glycine	Mann-Whitney U	1355	0.5747	0.9727
87	propylene glycol	Mann-Whitney U	1355	0.5747	0.9727
88	alpha-hydroxyisobutyric acid	Mann-Whitney U	1191	0.5840	0.9727
89	taurine	Mann-Whitney U	1351	0.5934	0.9727
90	azelaic acid	Mann-Whitney U	1195.5	0.6047	0.9727
91	2-oxoglutarate	Mann-Whitney U	1196	0.6074	0.9727
92	succinate	Mann-Whitney U	1347	0.6100	0.9727
93	UMP	Mann-Whitney U	1347	0.6121	0.9727
94	L-proline	Mann-Whitney U	1197	0.6124	0.9727
95	isovaleric acid	Mann-Whitney U	1201	0.6260	0.9727
96	D-mannose	Mann-Whitney U	1200	0.6268	0.9727
97	uridine	Mann-Whitney U	1202.5	0.6389	0.9727
98	L-glutamine	Mann-Whitney U	1339.5	0.6464	0.9727
99	creatine	Mann-Whitney U	1205	0.6494	0.9727
100	isocitric acid	Mann-Whitney U	1338	0.6556	0.9727
101	UDP	Mann-Whitney U	1207	0.6609	0.9727
102	hypotaurine	Mann-Whitney U	1208	0.6658	0.9727

No	Feature	Test	Statistic	p-value	FDR
103	L-glutamic acid	Mann-Whitney U	1336	0.6658	0.9727
104	fumaric acid	Mann-Whitney U	1333	0.6807	0.9727
105	D-gluconic acid	Mann-Whitney U	1332	0.6857	0.9727
106	xylitol	Mann-Whitney U	1328	0.6872	0.9727
107	dAMP	Mann-Whitney U	1216.5	0.7076	0.9727
108	urocanic acid	Mann-Whitney U	1217	0.7109	0.9727
109	beta-alanine	Mann-Whitney U	1219	0.7210	0.9727
110	inosine	Mann-Whitney U	1325	0.7210	0.9727
111	NADP	Mann-Whitney U	1221	0.7312	0.9727
112	argininosuccinic acid	Mann-Whitney U	1231.5	0.7441	0.9727
113	D-glucuronic acid	Mann-Whitney U	1317	0.7567	0.9727
114	2-hydroxyphenyl acetic acid	Mann-Whitney U	1226	0.7570	0.9727
115	L-anserine	Mann-Whitney U	1226.5	0.7586	0.9727
116	ascorbic acid	Mann-Whitney U	1227	0.7616	0.9727
117	1-methyl-L-histidine	Mann-Whitney U	1316	0.7673	0.9727
118	mandelic acid	Mann-Whitney U	1228.5	0.7695	0.9727
119	L-phenylalanine	Mann-Whitney U	1229	0.7723	0.9727
120	GDP	Mann-Whitney U	1311	0.7934	0.9727
121	O-acetyl-L-carnitine	Mann-Whitney U	1310	0.7986	0.9727
122	4-amino hippuric acid	Mann-Whitney U	1234	0.7987	0.9727
123	glycerol	Mann-Whitney U	1308	0.8081	0.9727
124	L-lysine	Mann-Whitney U	1305	0.8250	0.9727
125	2-hydroxybutyric acid	Mann-Whitney U	1303	0.8356	0.9727
126	2-oxobutyrate	Mann-Whitney U	1303	0.8356	0.9727
127	D-maltose	Mann-Whitney U	1243	0.8463	0.9727
128	glutaconic acid	Mann-Whitney U	1300.5	0.8489	0.9727

No	Feature	Test	Statistic	p-value	FDR
129	phosphocholine	Mann-Whitney U	1244	0.8516	0.9727
130	GMP	Mann-Whitney U	1245.5	0.8595	0.9727
131	benzoic acid	Mann-Whitney U	1245.5	0.8596	0.9727
132	dimethylamine	Mann-Whitney U	1297	0.8643	0.9727
133	butyrate	Mann-Whitney U	1249	0.8718	0.9727
134	S-acetamido-methylcysteine	Mann-Whitney U	1248	0.8730	0.9727
135	ADP	Mann-Whitney U	1295	0.8783	0.9727
136	galactitol	Mann-Whitney U	1249	0.8783	0.9727
137	acetaminophen	Mann-Whitney U	1293.5	0.8789	0.9727
138	glycolic acid	Mann-Whitney U	1251	0.8831	0.9727
139	3-hydroxyphenyl acetic acid	Mann-Whitney U	1292	0.8842	0.9727
140	cadaverine	Mann-Whitney U	1251	0.8891	0.9727
141	sebacic acid	Mann-Whitney U	1293	0.8891	0.9727
142	isobutyrate	Mann-Whitney U	1253	0.8907	0.9727
143	creatinine	Mann-Whitney U	1252	0.8926	0.9727
144	oxypurinol	Mann-Whitney U	1291	0.8951	0.9727
145	L-methionine	Mann-Whitney U	1255	0.9047	0.9727
146	3-methylxanthine	Mann-Whitney U	1290	0.9052	0.9727
147	L-carnitine	Mann-Whitney U	1289	0.9106	0.9727
148	malonate	Mann-Whitney U	1257	0.9206	0.9727
149	pyruvic acid	Mann-Whitney U	1286.5	0.9226	0.9727
150	2-propanol	Mann-Whitney U	1286.5	0.9240	0.9727
151	2-deoxycytidine	Mann-Whitney U	1286	0.9268	0.9727
152	2-aminobutyric acid	Mann-Whitney U	1259	0.9285	0.9727
153	NAD	Mann-Whitney U	1285.5	0.9295	0.9727
154	D-galactose	Mann-Whitney U	1260	0.9356	0.9727

No	Feature	Test	Statistic	p-value	FDR
155	N-(2-furoyl)glycine	Mann-Whitney U	1260	0.9376	0.9727
156	methylamine	Mann-Whitney U	1284	0.9376	0.9727
157	saccaric acid	Mann-Whitney U	1260.5	0.9400	0.9727
158	sarcosine	Mann-Whitney U	1261	0.9430	0.9727
159	malic acid	Mann-Whitney U	1262	0.9481	0.9727
160	UDPG	Mann-Whitney U	1282	0.9483	0.9727
161	N-acetylglycine	Mann-Whitney U	1264	0.9593	0.9727
162	pyrocatechol	Mann-Whitney U	1280	0.9593	0.9727
163	AMP	Mann-Whitney U	1278.5	0.9674	0.9727
164	vanillic acid	Mann-Whitney U	1269	0.9864	0.9864

Supplementary Table 3: Multivariate results of the best-performing model for metabolite abundances between FD patients and their relatives, arranged by Shapley values-based relative importance.

No	Feature	Relative importance
1	D-glucuronic acid	0.03535
2	trans-4-hydroxy-L-proline	0.02681
3	L-anserine	0.02428
4	threonic acid	0.02326
5	spermidine	0.02221
6	3-methyl-L-histidine	0.02169
7	succinate	0.02105
8	pyruvic acid	0.02090
9	creatinine	0.02028
10	L-arabitol	0.01940
11	xylitol	0.01796
12	citrate	0.01790
13	dimethylamine	0.01747

No	Feature	Relative importance
14	isobutyrate	0.01728
15	D-galactose	0.01560
16	saccaric acid	0.01468
17	guanidinoacetic acid	0.01454
18	L-arginine	0.01437
19	L-cysteine	0.01420
20	vanillic acid	0.01393
21	glycerol	0.01393
22	isocitric acid	0.01365
23	malic acid	0.01290
24	2-aminobutyric acid	0.01278
25	UMP	0.01253
26	glycerophosphocholine	0.01246
27	glycolic acid	0.01232
28	2-propanol	0.01226
29	trans-acotinic acid	0.01218
30	TMAO	0.01193
31	argininosuccinic acid	0.01102
32	7-methylxanthine	0.01065
33	sarcosine	0.01064
34	D-glucose-6-phosphate	0.01055
35	L-serine	0.01049
36	homovanillic acid	0.01043
37	L-glutamine	0.01009
38	isovaleric acid	0.01007
39	oxypurinol	0.00978
40	dihydrothymine	0.00976
41	L-methionine	0.00924
42	IMP	0.00907

No	Feature	Relative importance
43	methylmalonic acid	0.00903
44	methanol	0.00899
45	ascorbic acid	0.00877
46	butyrate	0.00837
47	acetoacetate	0.00837
48	dAMP	0.00814
49	GMP	0.00814
50	dimethyl sulfone	0.00811

Supplementary Table 4: KEGG-based FELLA enrichment analysis results for significant metabolites from multivariate module.

KEGG.id	Entry.type	KEGG.name	p.score
hsa00020	pathway	citrate cycle (TCA cycle) - Homo sapiens (human)	0.0066
hsa00053	pathway	ascorbate and aldarate metabolism - Homo sapiens (human)	0.0008
hsa00330	pathway	arginine and proline metabolism - Homo sapiens (human)	0.0000
hsa00340	pathway	histidine metabolism - Homo sapiens (human)	0.0088
hsa00410	pathway	beta-alanine metabolism - Homo sapiens (human)	0.0003
hsa01200	pathway	carbon metabolism - Homo sapiens (human)	0.0226
hsa02010	pathway	ABC transporters - Homo sapiens (human)	0.0082
hsa04150	pathway	mTOR signaling pathway - Homo sapiens (human)	0.0005
hsa04614	pathway	renin-angiotensin system - Homo sapiens (human)	0.0000
hsa04922	pathway	glucagon signaling pathway - Homo sapiens (human)	0.0004
hsa04972	pathway	pancreatic secretion - Homo sapiens (human)	0.0001
hsa04974	pathway	protein digestion and absorption - Homo sapiens (human)	0.0000
hsa05230	pathway	central carbon metabolism in cancer - Homo sapiens (human)	0.0000

KEGG.id	Entry.type	KEGG.name	p.score
M00010	module	citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate	0.0013
M00012	module	glyoxylate cycle	0.0019
M00014	module	glucuronate pathway (uronate pathway)	0.0172
M00015	module	proline biosynthesis, glutamate => proline	0.0004
M00029	module	urea cycle	0.0216
M00047	module	creatine pathway	0.0000
M00133	module	polyamine biosynthesis, arginine => agmatine => putrescine => spermidine	0.0001
M00134	module	polyamine biosynthesis, arginine => ornithine => putrescine	0.0000
M00135	module	GABA biosynthesis, eukaryotes, putrescine => GABA	0.0096
M00171	module	C4-dicarboxylic acid cycle, NAD - malic enzyme type	0.0181
M00173	module	reductive citrate cycle (Arnon-Buchanan cycle)	0.0003
M00970	module	proline degradation, proline => glutamate	0.0002
M00972	module	proline metabolism	0.0000
1.1.1.10	enzyme	L-xylulose reductase	0.0000
1.1.1.14	enzyme	L-iditol 2-dehydrogenase	0.0033
1.1.1.21	enzyme	aldose reductase	0.0000
1.1.1.38	enzyme	malate dehydrogenase (oxaloacetate-decarboxylating)	0.0033
1.1.1.40	enzyme	malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+)	0.0202
1.1.1.41	enzyme	isocitrate dehydrogenase (NAD+)	0.0220
1.13.11.20	enzyme	cysteine dioxygenase	0.0083
1.14.11.16	enzyme	peptide-aspartate beta-dioxygenase	0.0000
1.14.11.18	enzyme	phytanoyl-CoA dioxygenase	0.0002
1.14.11.2	enzyme	procollagen-proline 4-dioxygenase	0.0000
1.14.13.39	enzyme	nitric-oxide synthase (NADPH)	0.0004
1.2.1.24	enzyme	succinate-semialdehyde dehydrogenase (NAD+)	0.0201

KEGG.id	Entry.type	KEGG.name	p.score
1.5.1.2	enzyme	pyrroline-5-carboxylate reductase	0.0000
1.5.3.13	enzyme	N1-acetyl polyamine oxidase	0.0000
1.5.3.16	enzyme	spermine oxidase	0.0000
1.5.5.2	enzyme	proline dehydrogenase	0.0004
1.5.5.3	enzyme	hydroxyproline dehydrogenase	0.0000
2.1.1.2	enzyme	guanidinoacetate N-methyltransferase	0.0000
2.1.1.22	enzyme	carnosine N-methyltransferase	0.0000
2.1.4.1	enzyme	glycine amidinotransferase	0.0000
2.3.3.1	enzyme	citrate (Si)-synthase	0.0000
2.3.3.8	enzyme	ATP citrate synthase	0.0000
2.5.1.16	enzyme	spermidine synthase	0.0081
2.5.1.22	enzyme	spermine synthase	0.0000
2.6.1.19	enzyme	4-aminobutyrate---2-oxoglutarate transaminase	0.0046
2.7.11.2	enzyme	[pyruvate dehydrogenase (acetyl-transferring)] kinase	0.0106
2.7.3.2	enzyme	creatine kinase	0.0000
3.2.1.22	enzyme	alpha-galactosidase	0.0020
3.2.1.23	enzyme	beta-galactosidase	0.0050
3.4.11.3	enzyme	cystinyl aminopeptidase	0.0000
3.4.11.7	enzyme	glutamyl aminopeptidase	0.0000
3.4.11.9	enzyme	Xaa-Pro aminopeptidase	0.0000
3.4.13.18	enzyme	cytosol nonspecific dipeptidase	0.0000
3.4.13.20	enzyme	beta-Ala-His dipeptidase	0.0000
3.4.14.5	enzyme	dipeptidyl-peptidase IV	0.0000
3.4.15.1	enzyme	peptidyl-dipeptidase A	0.0214
3.4.16.2	enzyme	lysosomal Pro-Xaa carboxypeptidase	0.0000
3.4.16.5	enzyme	carboxypeptidase C	0.0012

KEGG.id	Entry.type	KEGG.name	p.score
3.4.17.1	enzyme	carboxypeptidase A	0.0000
3.4.17.15	enzyme	carboxypeptidase A2	0.0000
3.4.17.2	enzyme	carboxypeptidase B	0.0000
3.4.17.20	enzyme	carboxypeptidase U	0.0000
3.4.17.23	enzyme	angiotensin-converting enzyme 2	0.0001
3.4.21.1	enzyme	chymotrypsin	0.0000
3.4.21.26	enzyme	prolyl oligopeptidase	0.0000
3.4.21.35	enzyme	tissue kallikrein	0.0037
3.4.21.39	enzyme	chymase	0.0001
3.4.21.4	enzyme	trypsin	0.0000
3.4.21.70	enzyme	pancreatic endopeptidase E	0.0000
3.4.21.71	enzyme	pancreatic elastase II	0.0000
3.4.23.1	enzyme	pepsin A	0.0000
3.4.23.15	enzyme	renin	0.0119
3.4.24.11	enzyme	neprilysin	0.0000
3.4.24.16	enzyme	neurolysin	0.0000
3.4.24.18	enzyme	meprin A	0.0000
3.4.24.63	enzyme	meprin B	0.0000
3.5.3.1	enzyme	arginase	0.0000
3.5.3.17	enzyme	guanidinopropionase	0.0000
4.1.1.11	enzyme	aspartate 1-decarboxylase	0.0003
4.1.1.17	enzyme	ornithine decarboxylase	0.0001
4.1.1.19	enzyme	arginine decarboxylase	0.0005
4.2.1.3	enzyme	aconitate hydratase	0.0001
5.6.1.6	enzyme	channel-conductance-controlling ATPase	0.0133
6.3.2.11	enzyme	carnosine synthase	0.0000

KEGG.id	Entry.type	KEGG.name	p.score
6.4.1.1	enzyme	pyruvate carboxylase	0.0013
7.2.2.13	enzyme	Na+/K+-exchanging ATPase	0.0005
7.4.2.14	enzyme	ABC-type antigen peptide transporter	0.0215
7.6.2.4	enzyme	ABC-type fatty-acyl-CoA transporter	0.0187
7.6.2.8	enzyme	ABC-type vitamin B12 transporter	0.0187
R00199	reaction	ATP:pyruvate,water phosphotransferase	0.0075
R00200	reaction	ATP:pyruvate 2-O-phosphotransferase	0.0093
R00206	reaction	ATP:pyruvate,phosphate phosphotransferase	0.0207
R00217	reaction	oxaloacetate carboxy-lyase (pyruvate-forming)	0.0010
R00220	reaction	L-serine ammonia-lyase	0.0128
R00258	reaction	L-alanine:2-oxoglutarate aminotransferase	0.0030
R00344	reaction	pyruvate:carbon-dioxide ligase (ADP-forming)	0.0155
R00351	reaction	acetyl-CoA:oxaloacetate C-acetyltransferase (thioester-hydrolysing)	0.0000
R00352	reaction	acetyl-CoA:oxaloacetate C-acetyltransferase [(pro-S)-carboxymethyl-forming, ADP-phosphorylating]	0.0000
R00369	reaction	L-alanine:glyoxylate aminotransferase	0.0105
R00396	reaction	L-alanine:NAD+ oxidoreductase (deaminating)	0.0033
R00400	reaction	L-alanine:oxaloacetate aminotransferase	0.0103
R00402	reaction	succinate:NAD+ oxidoreductase	0.0027
R00405	reaction	succinate:CoA ligase (ADP-forming)	0.0008
R00489	reaction	L-aspartate 1-carboxy-lyase (beta-alanine-forming)	0.0002
R00551	reaction	L-arginine amidinohydrolase	0.0000
R00552	reaction	L-arginine iminohydrolase	0.0000
R00557	reaction	L-arginine,NADPH:oxygen oxidoreductase (nitric-oxide-forming)	0.0003

KEGG.id	Entry.type	KEGG.name	p.score
R00558	reaction	L-arginine,NADPH:oxygen oxidoreductase (N-(omega)-hydroxyarginine-forming)	0.0001
R00562	reaction	N2-(D-1-carboxyethyl)-L-arginine:NAD+ oxidoreductase (L-arginine-forming)	0.0133
R00565	reaction	L-arginine:glycine amidinotransferase	0.0000
R00566	reaction	L-arginine carboxy-lyase (agmatine-forming)	0.0027
R00567	reaction	arginine racemase	0.0008
R00646	reaction	ascorbate + oxygen + H2O <=> threonate + oxalate	0.0000
R00670	reaction	L-ornithine carboxy-lyase (putrescine-forming)	0.0031
R00671	reaction	L-ornithine ammonia-lyase (cyclizing; L-proline-forming)	0.0024
R00782	reaction	L-cysteine hydrogen-sulfide-lyase (deaminating; pyruvate-forming)	0.0000
R00891	reaction	L-serine hydro-lyase (adding hydrogen sulfide, L-cysteine-forming)	0.0001
R00892	reaction	L-cysteine:NAD+ oxidoreductase	0.0058
R00893	reaction	L-cysteine:oxygen oxidoreductase	0.0084
R00894	reaction	L-glutamate:L-cysteine gamma-ligase (ADP-forming)	0.0009
R00897	reaction	O3-acetyl-L-serine:hydrogen-sulfide 2-amino-2-carboxyethyltransferase	0.0158
R00901	reaction	L-cysteine hydrogen-sulfide-lyase (adding sulfite; L-cysteate-forming)	0.0045
R00904	reaction	3-aminopropanal:NAD+ oxidoreductase	0.0151
R00907	reaction	L-alanine:3-oxopropanoate aminotransferase	0.0001
R00908	reaction	beta-alanine:2-oxoglutarate aminotransferase	0.0004
R00912	reaction	L-arginine:beta-alanine ligase (ADP-forming)	0.0027
R01001	reaction	L-cystathionine cysteine-lyase (deaminating; 2-oxobutanoate-forming)	0.0035
R01093	reaction	galactitol:NAD+ 1-oxidoreductase	0.0016
R01095	reaction	galactitol:NADP+ 1-oxidoreductase	0.0022
R01098	reaction	D-galactose:oxygen 6-oxidoreductase	0.0185

KEGG.id	Entry.type	KEGG.name	p.score
R01101	reaction	melibiose galactohydrolase	0.0128
R01103	reaction	raffinose galactohydrolase	0.0041
R01159	reaction	S-adenosyl-L-methionine:L-histidine N-methyltransferase	0.0000
R01164	reaction	L-histidine:beta-alanine ligase (ADP-forming)	0.0002
R01166	reaction	Nalpha-(beta-alanyl)-L-histidine hydrolase	0.0000
R01194	reaction	3-O-alpha-D-galactosyl-1D-myo-inositol galactohydrolase	0.0156
R01248	reaction	L-proline:NAD+ 5-oxidoreductase	0.0063
R01251	reaction	L-proline:NADP+ 5-oxidoreductase	0.0099
R01252	reaction	L-proline,2-oxoglutarate:oxygen oxidoreductase (4-hydroxylating)	0.0000
R01322	reaction	citrate:CoA ligase (ADP-forming)	0.0020
R01324	reaction	citrate hydroxymutase	0.0000
R01325	reaction	citrate hydro-lyase (cis-aconitate-forming)	0.0001
R01431	reaction	xylitol:NADP+ oxidoreductase	0.0000
R01566	reaction	creatine amidinohydrolase	0.0008
R01588	reaction	dimethylamine:electron-transferring flavoprotein oxidoreductase	0.0000
R01686	reaction	L-arginine:taurine amidinotransferase	0.0057
R01758	reaction	L-arabitol:NAD+ 1-oxidoreductase	0.0000
R01759	reaction	L-arabitol:NADP+ 1-oxidoreductase	0.0000
R01881	reaction	ATP:creatine N-phosphotransferase	0.0009
R01883	reaction	S-adenosyl-L-methionine:guanidinoacetate N-methyltransferase	0.0000
R01884	reaction	creatinine amidohydrolase	0.0000
R01896	reaction	xylitol:NAD+ 2-oxidoreductase (D-xylulose-forming)	0.0000
R01903	reaction	L-arabinitol:NAD+ 4-oxidoreductase (L-xylulose-forming)	0.0000
R01904	reaction	xylitol:NADP+ 4-oxidoreductase (L-xylulose-forming)	0.0000

KEGG.id	Entry.type	KEGG.name	p.score
R01914	reaction	spermidine:(acceptor) oxidoreductase	0.0000
R01915	reaction	spermidine:(acceptor) oxidoreductase	0.0000
R01917	reaction	gamma-L-glutamyl-L-cysteinyl-glycine:spermidine ligase (ADP-forming)	0.0000
R01918	reaction	gamma-L-glutamyl-L-cysteinyl-glycine:spermidine amidase	0.0000
R01920	reaction	S-adenosylmethioninamine:putrescine 3-aminopropyltransferase	0.0000
R01989	reaction	L-arginine:4-aminobutanoate amidinotransferase	0.0000
R01992	reaction	alpha-aminobutyryl histidine hydrolase	0.0188
R02144	reaction	S-adenosyl-L-methionine:carnosine N-methyltransferase	0.0000
R02164	reaction	succinate:quinone oxidoreductase	0.0181
R02441	reaction	L-arabitol:NAD+ 2-oxidoreductase (L-ribulose-forming)	0.0000
R02509	reaction	N,N-dimethylformamide amidohydrolase	0.0043
R02511	reaction	trimethylamine:electron-transferring flavoprotein oxidoreductase(demethylating)	0.0000
R02512	reaction	trimethylamine-N-oxide formaldehyde-lyase	0.0000
R02575	reaction	ATP:guanidoacetate N-phosphotransferase	0.0000
R02752	reaction	D-glucarate hydro-lyase	0.0000
R02754	reaction	5-dehydro-4-deoxy-D-glucarate tartronate-semialdehyde-lyase	0.0197
R02869	reaction	S-adenosylmethioninamine:spermidine 3-aminopropyltransferase	0.0003
R02922	reaction	creatinine iminohydrolase	0.0000
R02957	reaction	D-glucuronolactone:NAD+ oxidoreductase	0.0000
R03260	reaction	O-succinyl-L-homoserine succinate-lyase (adding cysteine)	0.0052
R03277	reaction	2-Hydroxy-3-oxopropanoate + Pyruvate <=> 2-Dehydro-3-deoxy-D-glucarate	0.0198
R03286	reaction	N(pi)-methyl-L-histidine:beta-alanine ligase (ADP-forming)	0.0000

KEGG.id	Entry.type	KEGG.name	p.score
R03288	reaction	beta-alanyl-N(pi)-methyl-L-histidine hydrolase	0.0000
R03291	reaction	trans-4-hydroxy-L-proline:NAD+ 5-oxidoreductase	0.0000
R03293	reaction	trans-4-hydroxy-L-proline:NADP+ 5-oxidoreductase	0.0000
R03295	reaction	trans-4-hydroxy-L-proline:quinone oxidoreductase	0.0000
R03296	reaction	trans-4-hydroxy-L-proline 2-epimerase	0.0000
R03355	reaction	beta-D-galactosyl-1,4-beta-D-glucosylceramide galactohydrolase	0.0095
R03617	reaction	D-galactosyl-N-acylsphingosine galactohydrolase	0.0014
R03618	reaction	globotriosylceramide galactohydrolase	0.0222
R03733	reaction	L-threonate:NAD+ 3-oxidoreductase	0.0000
R05364	reaction	2-hydroxy-6-oxo-7-methylocta-2,4-dienoate acylhydrolase	0.0000
R05377	reaction	2-hydroxy-3-carboxy-6-oxo-7-methylocta-2,4-dienoate carboxy-lyase	0.0096
R05831	reaction	xylitol:NAD oxidoreductase	0.0000
R05961	reaction	H2O + globotriaosylceramide <=> D-galactose + lactosylceramide	0.0000
R06010	reaction	GM1 + H2O <=> GM2 + D-galactose	0.0000
R07152	reaction	xylitol:oxygen oxidoreductase	0.0000
R07274	reaction	O-phospho-L-serine:hydrogen-sulfide 2-amino-2-carboxyethyltransferase	0.0007
R07420	reaction	phosphocreatine <=> creatinine + orthophosphate.	0.0000
R07807	reaction	G01977 + H2O <=> G13073 + D-galactose	0.0000
R08056	reaction	D-glucarate hydro-lyase	0.0000
R08197	reaction	L-arginine:pyruvate aminotransferase	0.0000
R08346	reaction	citalopram:oxygen oxidoreductase(deaminating)(flavin-containing)	0.0002
R08714	reaction	putrescine:pyruvate aminotransferase	0.0119
R09076	reaction	spermidine:oxygen oxidoreductase (spermidine-forming)	0.0000
R09077	reaction	spermidine:oxygen oxidoreductase (3-aminopropanal-forming)	0.0000

KEGG.id	Entry.type	KEGG.name	p.score
R09081	reaction	carboxyspermidine carboxy-lyase (spermidine-forming)	0.0002
R09124	reaction	trimethylamine:coenzyme M methyltransferase	0.0000
R09477	reaction	xylitol:NAD+ oxidoreductase	0.0000
R09999	reaction	dimethylamine:coenzyme M methyltransferase	0.0000
R10090	reaction	citrate:N6-acetyl-N6-hydroxy-L-lysine ligase (AMP-forming)	0.0200
R10343	reaction	succinyl-CoA:acetate CoA-transferase	0.0042
R11031	reaction	L-arginine:NAD+ oxidoreductase (deaminating)	0.0000
R11032	reaction	L-arginine:NADP+ oxidoreductase (deaminating)	0.0000
R11033	reaction	L-arginine <=> N(omega)-hydroxyarginine	0.0005
R11604	reaction	L-arginine:oxygen oxidoreductase (deaminating)	0.0000
R11711	reaction	L-arginine,reduced-flavodoxin:oxygen oxidoreductase (nitric-oxide-forming)	0.0081
R11712	reaction	2 L-arginine + 2 reduced flavodoxin + 2 oxygen <=> 2 N(omega)-hydroxyarginine + 2 oxidized flavodoxin + 2 H2O	0.0053
R11819	reaction	trans-4-hydroxy-L-proline hydro-lyase	0.0000
R12212	reaction	oxaloacetate carboxy-lyase (pyruvate-forming)	0.0141
R12308	reaction	L-2,3-diaminopropanoate:citrate ligase (AMP, 2-[(L-alanin-3-ylcarbamoyl)methyl]-2-hydroxybutanedioate-forming)	0.0116
R12353	reaction	D-ornithine:citrate ligase (AMP-forming)	0.0017
R12707	reaction	octopine + H2O + acceptor <=> L-arginine + pyruvate + reduced acceptor	0.0172
R13137	reaction	succinate:ferricytochrome-c oxidoreductase	0.0207
C00022	compound	pyruvate	0.0000
C00042	compound	succinate	0.0000
C00062	compound	L-arginine	0.0000
C00077	compound	L-ornithine	0.0070
C00097	compound	L-cysteine	0.0000

KEGG.id	Entry.type	KEGG.name	p.score
C00099	compound	beta-alanine	0.0000
C00124	compound	D-galactose	0.0000
C00135	compound	L-histidine	0.0133
C00158	compound	citrate	0.0000
C00300	compound	creatine	0.0007
C00315	compound	spermidine	0.0000
C00379	compound	xylitol	0.0000
C00386	compound	carnosine	0.0212
C00532	compound	L-arabitol	0.0000
C00543	compound	dimethylamine	0.0000
C00581	compound	guanidino acetate	0.0000
C00596	compound	2-hydroxy-2,4-pentadienoate	0.0005
C00791	compound	creatinine	0.0000
C00818	compound	D-glucarate	0.0000
C01152	compound	N(pi)-methyl-L-histidine	0.0000
C01157	compound	hydroxyproline	0.0000
C01262	compound	beta-alanyl-N(pi)-methyl-L-histidine	0.0000
C01620	compound	threonate	0.0000
C02305	compound	phosphocreatine	0.0085
C02632	compound	2-methylpropanoate	0.0000
C06582	compound	2-hydroxy-6-oxo-7-methylocta-2,4-dienoate	0.0004