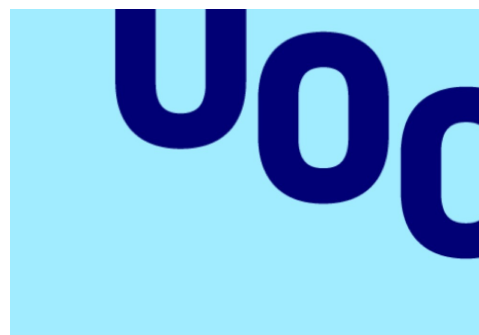
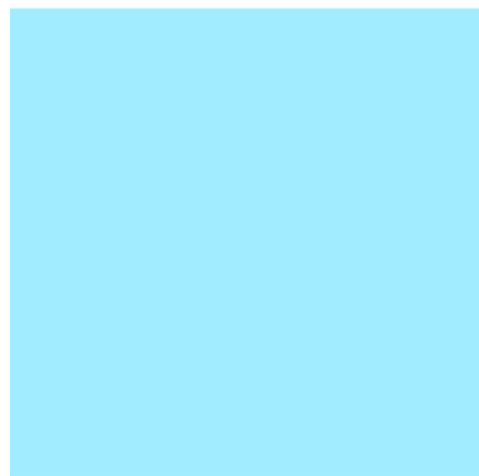


# metaboPipe: a Modular Pipeline for Metabolomic Data Preprocessing



Universitat Oberta  
de Catalunya



UNIVERSITAT DE  
BARCELONA

**Eduard Pérez Méndez**

Statistical Bioinformatics and  
Machine Learning

Master's degree in Bioinformatics  
and Biostatistics

Name of the tutor:

**Alexandre Sánchez Pla**

Name of the SRP:

Carles Ventura Royo

May 14, 2024



Except where otherwise noted, this work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike

<https://creativecommons.org/licenses/by-nc/4.0>

## Final Work Card

<b>Title of the work:</b>	metaboPipe: a Modular Pipeline for Metabolomic Data Preprocessing
<b>Name of the author:</b>	Eduard Pérez Méndez
<b>Name of the tutor:</b>	Alexandre Sánchez Pla
<b>Name of the SRP:</b>	Carles Ventura Royo
<b>Date of delivery:</b>	May 14, 2024
<b>Studies or Program:</b>	Master's degree in Bioinformatics and Biostatistics
<b>Area or the Final Work:</b>	Statistical Bioinformatics and Machine Learning
<b>Language of the work:</b>	English
<b>Keywords:</b>	targeted metabolomics, preprocessing, pipeline

### Abstract

A maximum of 250 words, detailing the purpose, context of application, methodology, results and conclusions of the work.

*“BIG MOTIVATIONAL QUOTE ”*

AUTHOR NAME

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
1.1	General description . . . . .	8
1.2	Context and justification . . . . .	8
1.2.1	Preprocessing of data . . . . .	9
1.2.2	Pretreatment of Data . . . . .	10
1.3	State of the art . . . . .	12
<b>2</b>	<b>Objectives</b>	<b>13</b>
2.1	Main objective . . . . .	13
2.2	Specific objectives . . . . .	13
<b>3</b>	<b>Sustainable development goals</b>	<b>14</b>
<b>4</b>	<b>Approach and methodology</b>	<b>15</b>
4.1	Methodology . . . . .	15
4.2	Planning and calendar . . . . .	15
4.2.1	Tasks . . . . .	15
4.3	Risk analysis . . . . .	15
4.4	Final products . . . . .	16
4.5	Chapters structure . . . . .	16
<b>5</b>	<b>Materials and methods</b>	<b>17</b>
5.1	Datasets . . . . .	17
5.2	Packages . . . . .	17
5.3	Modules . . . . .	17
5.4	Documentation . . . . .	17
5.5	Shiny app . . . . .	17
<b>6</b>	<b>Results</b>	<b>18</b>
<b>7</b>	<b>Conclusion and future vision</b>	<b>19</b>
	<b>Glossary</b>	<b>20</b>
	<b>Acronyms</b>	<b>21</b>
	<b>Bibliography</b>	<b>22</b>

# List of Figures

4.1	Gantt chart showing the project timeline and milestones. . . . .	15
6.1	Error en función de la distancia en unidades arbitrarias. . . . .	18

# List of Tables

4.1 Risk analysis. This table presents various risks associated with the project,  
along with their severity, likelihood, and potential mitigation measures. . . . 16



# 1. Introduction

## 1.1 General description

Metabolomics, a powerful and evolving field within the realm of systems biology, plays a pivotal role in unraveling the intricate web of biochemical processes occurring within living organisms. As we delve into the molecular intricacies of biological systems, the generation of vast and complex datasets poses a significant challenge. Challenges in standardizing nutritional metabolomics include experimental design, sample preparation, and data analysis, which impact result validity and reproducibility. Efforts by the international community aim to establish standard procedures and infrastructure for advancing nutritional metabolomics research. This master thesis project aims for the creation of a modular pipeline designed to streamline the processing of targeted metabolomics data to a usable and meaningful dataset for further analysis and biological interpretation.

## 1.2 Context and justification

Metabolomics is a rapidly evolving field within biology that focuses on the comprehensive study of the metabolite composition of cell types, tissues, organs, or organisms [1–3]. It aims to measure, identify and (semi-)quantify those metabolites. Metabolites are chemical compounds that undergo analysis through conventional chemical assessment methods like [Mass Spectrometry \(MS\)](#) and [Nuclear Magnetic Resonance \(NMR\)](#) spectrometry. [MS](#) approaches are commonly integrated with [Gas Chromatography \(GC\)](#) and [Liquid Chromatography \(LC\)](#), leading to the development of two advanced techniques known as [Gas Chromatography-Mass Spectrometry \(GC-MS\)](#) and [Liquid Chromatography-Mass Spectrometry \(LC-MS\)](#). All of these analytical platforms and methodologies generate large amounts of high-dimensional and complex experimental raw data.

However, the statistical analysis of metabolomics data presents significant challenges, attributable not only to the inherent complexity of metabolomics as a research discipline but also to the intricate nature of the data itself. Notwithstanding that numerous studies have explored various methodologies for metabolomic data management, the field still lacks an accepted standard for preprocessing and pretreatment of such data.

One of the obstacles the field encounters is the lack of well defined terminology, as the terms “data preprocessing” and “data pretreatment” have not been used consistently in metabolomics literature [4].

The objectives of data preprocessing/pretreatment encompass two primary aims: firstly, to rectify or mitigate instrumental artifacts and extraneous biological variance, thereby amplifying the [Signal-to-Noise Ratio \(SNR\)](#); and secondly, to effectively transform the data into interpretable spectral profiles through processes such as centering, scaling, and dimensionality reduction [4, 5]. The choice of preprocessing and pretreatment methods can signifi-

cantly impact the downstream analysis and interpretation of metabolomic data [6] so the steps should be carefully selected based on the specific characteristics of the data and the research.

By establishing a standardized approach to preprocess and pretreat metabolomic data, the field can improve the quality, comparability, and reproducibility of metabolomic studies. This would facilitate data integration, enable the development of robust statistical models, and enhance our understanding of the complex metabolic processes underlying health and disease.

### 1.2.1 Preprocessing of data

Given the inherent dissimilarities in data acquisition techniques, unique preprocessing procedures are imperative before embarking on statistical analyses in metabolomics investigations. NMR spectra, for instance, often exhibit signal shifts along the axis due to factors like pH fluctuations [7]. Thus, meticulous preprocessing is indispensable to ensure robust statistical analyses and facilitate inter-spectral signal comparisons. This involves techniques such as binning, peak fitting with spectral databases, and exclusion of unstable or non-informative spectral regions (e.g., water peaks) [3, 4, 8]. By refining the dataset to a subset of relevant metabolites, statistical methods can effectively discern variations in signal intensity among sample groups [9].

The preprocessing workflows diverge between MS-based and NMR-based metabolomic analyses. In MS-based profiling, data are presented as three-dimensional (3D) tables, in contrast to the two-dimensional (2D) tables derived from GC-MS data preprocessing [4, 8]. GC-MS preprocessing entails deconvolution and peak integration to generate intensity profiles for each sample feature corresponding to RT/ $m/z$  pairs. Notably, metabolite identification strategies differ between GC-MS and LC-MS methodologies. While GC-MS relies on reproducible mass spectra and extensive databases for metabolite identification based on characteristic fragment ions, MS-based methods prioritize automation, accuracy, peak identification, integration, and annotation [10, 11].

While the primary objective of preprocessing is to render data comparable across samples despite instrumental discrepancies, the strategies employed in MS-based methodologies differ from those in NMR-based approaches. Moreover, variations exist between preprocessing methodologies utilized in GC-MS and LC-MS metabolomic analyses, underscoring the intricate nature of metabolomics data preprocessing.

#### MS-based data preprocessing

MS-based analysis involves the measurement of Mass-to-Charge Ratio ( $m/z$ ). When combined with either LC or GC, the resulting raw GC/LC-MS data encompass three measured variables:  $m/z$ , chromatographic Retention Time (RT), and intensity count, thereby constituting a three-dimensional (3D) data structure. To streamline the data and eliminate spectral noise and irrelevant biological variability, a two-dimensional (2D) features table is generated through peak picking. This table encompasses all quantified metabolic features from the analyzed samples, with rows corresponding to samples and columns representing variables such as peak areas or intensities, characterized by  $m/z$  and retention time in minutes or

scan number (m/z-RT pairs). The preprocessing of MS data involves several steps: 1) denoising and baseline correction; 2) alignment across all samples; 3) peak picking; 4) merging the peaks; and 5) creating a data matrix [3, 4, 10, 12–17].

### NMR-based data preprocessing

Similar to MS-based analysis, NMR-based analysis generates a 2D structure of feature data matrix with the samples in the rows and the spectral data points in the columns. Also similar to MS-based analysis, the NMR-based analysis (e.g., <sup>1</sup>H NMR analysis) requires data preprocessing to mitigate non-biologically relevant effects. The following data preprocessing steps could be performed: 1) baseline correction; 2) peak binning; 3) peak alignment; 4) quality control; 5) create a data matrix [4, 5, 15–20]. Preprocessing by either MS or NMR constructs a data matrix containing the relative abundances of a set of mass spectra for a group of samples or subjects under different conditions. The metabolomics data matrix are typically constructed in such a way that each row of the data matrix represents a subject and each column represents the mass spectra (metabolite intensities or metabolite relative abundances, peak or peak intensities).

## 1.2.2 Pretreatment of Data

### Handling Missing Values

Within datasets, missing values or zeros can arise due to a variety of factors, both biological and technical in nature. Categorizations by Sun Xia delineate these zeros into four distinct categories: 1) Structural zeros, 2) Sampling zeros, 3) Values below the limit of detection (LOD), and 4) Zeros derived from negative values that are automatically transformed.

1. **Structural zeros** pertain to peaks absent from a sample or chromatogram due to genuine biological absence rather than technical errors. For instance, if a compound is not present in a biological sample, the corresponding peak for that compound is deemed a structural zero.
2. **Sampling zeros** refer to peaks present in samples but missed during peak picking.
3. **Values below LOD** represent intensities or abundances falling below the detection limit of the mass spectrometer.
4. **Negative value zeros** result from negative intensity or abundance values, considered spectral artifacts or noise, and subsequently transformed to zero.

Identifying the origins of these zeros poses a challenge, and their prevalence presents a significant obstacle for statistical analyses [4, 21]. Hence, practical approaches for managing zeros include:

1. **Filtering** based on a threshold, such as the 80% rule.
2. **Imputation** techniques, which can involve substituting zeros with the mean, minimum (or half of the minimum) of non-missing values, or simply zero.

- Utilizing **missing data estimation algorithms** to employ various methods for handling missing values.

However, it's crucial to recognize that valuable biological insights may be embedded within peaks containing missing values.

## Managing Outliers

Various methods exist for addressing outliers, including:

- Assessing metabolite peak areas and comparing the ratio of mean to median, with the median often considered more robust in the presence of outliers.
- Employing [Principal Component Analysis \(PCA\)](#) to identify outliers, followed by techniques such as [Principal Component Partial R-square \(PCPr2\)](#) and [Analysis of Variance \(ANOVA\)](#).
- Recent advancements have introduced specialized algorithms for outlier identification in metabolomic data, such as cellwise outlier diagnostics using robust pairwise log ratios (cell-rPLR) and a kernel weight function-based biomarker identification technique.

## Normalization

Normalization is a crucial step in data preprocessing that seeks to eliminate unwanted variations between samples. By doing so, it ensures that samples can be directly compared to each other by eliminating or reducing systematic errors, biases, and experimental variance [22].

Normalization of data within metabolomic workflows can occur either during sample analysis (preanalytical normalization) or during postanalytical data processing. Normalization of samples is essential due to variations in composition influenced by factors like time of day, health status, and dietary intake.

For instance, blood samples may not require normalization due to the body's control over blood volume and composition. However, urine samples may necessitate normalization due to potential concentration variations [23].

## Centering and Scaling

Centering aims to shift metabolite concentrations to fluctuate around zero, while scaling adjusts for fold-change differences between metabolites. Both steps are crucial in data preprocessing.

## Transformation

Transformation becomes necessary to address data variance after scaling, aiming to correct for heteroscedasticity, convert multiplicative relations into additive ones, and normalize skewed distributions.

## 1.3 State of the art

Punto de partida del trabajo (¿Cuál es la necesidad a cubrir? ¿Por qué es un tema relevante? ¿Cómo se resuelve el problema de momento?) y aportación realizada (¿Qué resultado se quiere obtener?).

Metabolomic data preprocessing involves denoising, baseline correction, peak picking, and alignment. For **MS**-based analysis, tools like XCMS and MZmine facilitate preprocessing, while for **NMR** data, packages like BATMAN and RAMSY offer robust preprocessing capabilities. Pretreatment techniques include handling missing values, outlier detection, and normalization using methods like imputation, robust statistical measures, and scaling techniques.

Nevertheless the field lacks a standardized approach to metabolomic data preprocessing, with inconsistencies in terminology and methodologies. Stanstrup *et al.* in their “The metaRbolomics Toolbox in Bioconductor and Beyond” made an extensive revision of both the scientific literature and the R landscape for packages relevant for metabolomic research.

En algun lloc és important emfasitzar que descartem la metabolòmica no dirigida.

## 2. Objectives

Escriure-ho en angles

### 2.1 Main objective

1. Desenvolupar un nou pipeline per al preprocés de dades de metabolòmica dirigida amb l'objectiu de millorar la eficiència i la modularitat respecte als pipelines existents. Aquest nou pipeline s'implementarà en R.

### 2.2 Specific objectives

1. Implementar el pipeline de preprocés de dades de metabolòmica dirigida utilitzant el paquet "targets" per a garantir la replicabilitat i la gestió eficient dels recursos computacionals.
2. Seleccionar diversos conjunts de dades de metabolòmica dirigida per validar i optimitzar el rendiment del nou pipeline, analitzant-ne la qualitat i la consistència.
3. Facilitar l'ús de l'eina desenvolupada a la comunitat científica, mitjançant la creació d'una documentació detallada i la publicació del codi en un repositori públic.

En els objectius específics és on hauries de parlar de la forma en que ho faras (per exemple, fent servir el paquet "targets") i també introduir la selecció i anàlisi d'alguns datasets ilustratius. En aquest camp sovint els datasets ho son tot. Ja he vist que després parles deL "approach taken to achieve this objective", però es que això és el que haurien de ser les tasques i la metodologia.

Resumint, valdria la pena incloure lo del targets i els datasets en les tasques i deliverables i sobretot tenir en compte que caldrà agafar que ja estigui fet o que caldria fer, i per això serà bo que en la revisió de la literatura" hi incloguis, no només, treballs de metabolòmica sino paquets de R que ja estan implementant moltes d'aquestes coses.

### 3. Sustainable development goals

Our project aligns with multiple crucial Sustainable Development Goals (SDGs) set by the United Nations, fostering global sustainability and development. The primary objectives of our project focus on developing a pipeline to modulate the pretreatment of metabolomics data and creating an R implementation. This solution holds significant potential to support the following SDGs:

#### **SDG 3: Good Health and Well-being**

The use of our pipeline has the potential to reduce the time required for metabolomic data research, accelerating the investigation of rare diseases, cancer, and other medical conditions. By expediting research processes, our project contributes to advancing medical science, improving healthcare outcomes, and ultimately enhancing global health and well-being.

#### **SDG 9: Industry, Innovation, and Infrastructure:**

Our focus on developing an open-source, well-documented, and user-friendly pipeline fosters innovation and infrastructure development. By opening access to metabolomics research tools, our project empowers individuals from diverse backgrounds to engage in scientific inquiry and innovation, thus promoting inclusive economic growth and technological progress.

#### **SDG 10: Reduced Inequalities:**

Through our implementation, we prioritize inclusivity and accessibility, ensuring that individuals regardless of sex, gender, race, wealth, or ability can utilize, learn from, and contribute to our pipeline. By reducing barriers to entry and promoting equal opportunities for participation in scientific endeavors, our project contributes to reducing inequalities and promoting social inclusion.

While our project aims to bring about positive change, it is essential to consider potential negative impacts and ethical considerations. These may include concerns about data privacy and security, particularly in handling sensitive information. Additionally, there may be unintended consequences such as exacerbating existing inequalities in access to technology or inadvertently reinforcing biases in data analysis. Therefore, it is imperative to approach the development and implementation of our pipeline with careful consideration of ethical principles, transparency, and accountability to mitigate potential risks and maximize societal benefits.



## 4. Approach and methodology

Mención de cuáles son las posibles estrategias para llevar a cabo el trabajo y cuál es la estrategia elegida (desarrollar un producto nuevo, adaptar un producto existente...). Hay que incluir una valoración de por qué esta es la estrategia más apropiada para conseguir los objetivos.

### 4.1 Methodology

### 4.2 Planning and calendar

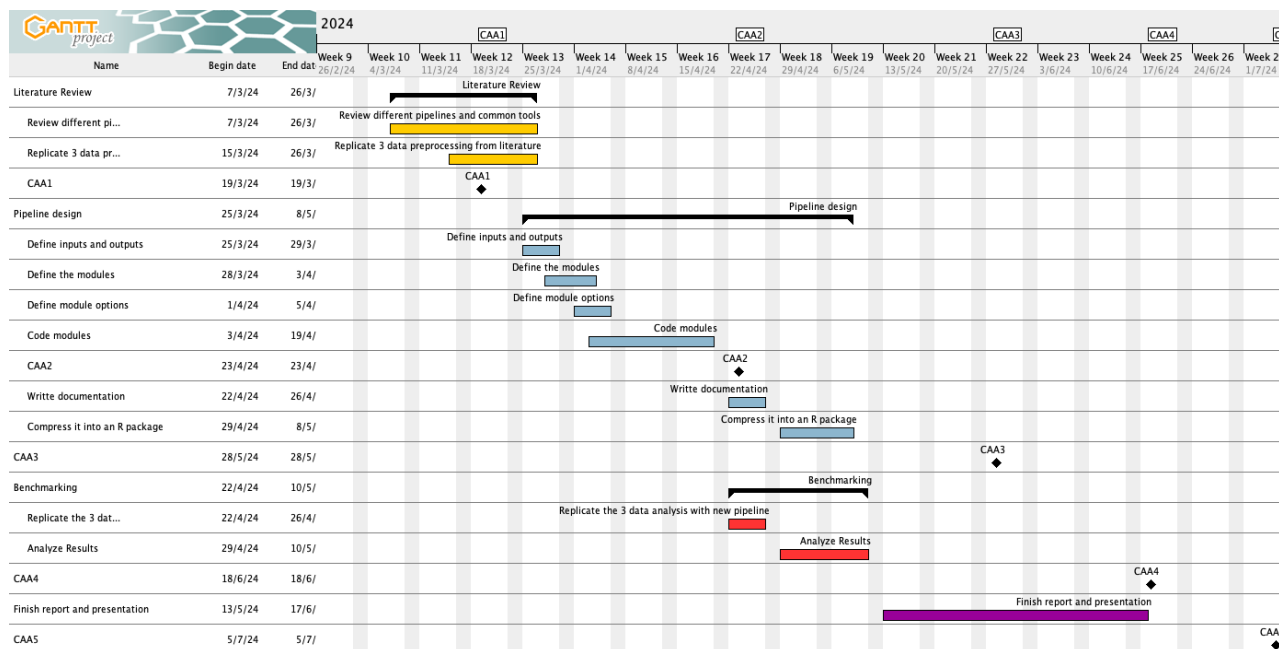


Figure 4.1: Gantt chart showing the project timeline and milestones.

#### 4.2.1 Tasks

#### Main Tasks and prioritization

#### Extra tasks

### 4.3 Risk analysis



Risk	Severity	Likelihood	Mitigation
Resource constraints	Moderate	Moderate	Develop a clear project timeline, incorporating milestones and allocating adequate time for each phase. Ensure contingency measures are in place to address unforeseen challenges or changes.
Technical challenges	Moderate	High	Perform proper exploration of packages and software and seek guidance and mentorship from professors or experts in relevant fields.
User adoption and awareness	High	Moderate	Be sure to incorporate appropriate cautions regarding the correct application of the chosen modules and data.
Pipeline branching	Low	Moderate	Adopt new methods to interactively select the branching

Table 4.1: Risk analysis. This table presents various risks associated with the project, along with their severity, likelihood, and potential mitigation measures.

## 4.4 Final products

Expand on every item

- A pipeline for targeted metabolomic data preprocessing.
- A package for the modular implementation of that pipeline in R.
- A detailed documentation of the pipeline.
- A Shiny app for the accessibility of the pipeline.
- A public repository with the code and documentation.

## 4.5 Chapters structure

Dejar para el final

Breve explicación de los contenidos de cada capítulo y su relación con el proyecto global.

## 5. Materials and methods

Los aspectos más relevante del diseño y desarrollo del trabajo.

La metodología elegida para hacer este desarrollo, describiendo las alternativas posibles, las decisiones tomadas, y los criterios utilizados para tomar estas decisiones.

Los productos obtenidos

### 5.1 Datasets

The datasets employed for the purpose of this study were obtained from the Metabolomics Workbench database, a public repository for metabolomics data.

### 5.2 Packages

The pipeline was developed using the R programming language and various packages .

Table of packages used in the project

### 5.3 Modules

add a summary of 1: Number of functions and 2: Number of lines of code

### 5.4 Documentation

### 5.5 Shiny app

Describir los módulos que forman parte del pipeline, la idea es que cada modulo representa una etapa del preprocesamiento de datos. (e.g., filtering, imputation, normalization, scaling, etc.)

## 6. Results

Detallad en este apartado los resultados obtenidos utilizando la metodología descrita en el apartado anterior.

Las figuras tienen que estar explicadas y citadas en el texto, como la 6.1, en la cual se muestra el error en función de la distancia, en unidades arbitrarias. En todas las gráficas tiene que haber el título de los ejes.

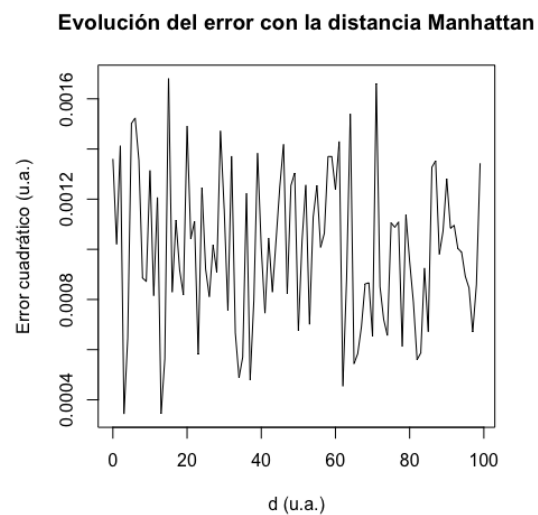


Figure 6.1: Error en función de la distancia en unidades arbitrarias.

## 7. Conclusion and future vision

Este capítulo tiene que incluir:

- Una descripción de las conclusiones del trabajo:
  - Una vez se han obtenido los resultados, ¿qué conclusiones se extraen?
  - ¿Estos resultados son los esperados? ¿O han sido sorprendentes? ¿Por qué?
- Una reflexión crítica sobre el logro de los objetivos planteados inicialmente:
  - ¿Hemos logrado todos los objetivos? Si la respuesta es negativa, ¿por qué motivo?
- Un análisis crítico del seguimiento de la planificación y metodología a lo largo del producto:
  - ¿Se ha seguido la planificación?
  - ¿La metodología prevista ha sido suficientemente adecuada?
  - ¿Ha habido que introducir cambios para garantizar el éxito del trabajo? ¿Por qué?
- De los impactos previstos en 3, ético-sociales, de sostenibilidad y de diversidad, evaluar/mencionar si se han mitigado (si eran negativos) o si se han conseguido (si eran positivos).
- Si han aparecido impactos no previstos a 3, evaluar/mencionar cómo se han mitigado (si eran negativos) o que han aportado (si eran positivos).
- Las líneas de trabajo futuro que no se han podido explorar en este trabajo y han quedado pendientes.
- Glossary test: [LaTeX](#)

# Glossary

**LaTeX** A typesetting system used for document preparation. [19](#)

# Acronyms

**ANOVA** Analysis of Variance. [11](#)

**GC** Gas Chromatography. [8, 9](#)

**GC-MS** Gas Chromatography-Mass Spectrometry. [8, 9](#)

**LC** Liquid Chromatography. [8, 9](#)

**LC-MS** Liquid Chromatography-Mass Spectrometry. [8, 9](#)

**m/z** Mass-to-Charge Ratio. [9](#)

**MS** Mass Spectrometry. [8–10, 12](#)

**NMR** Nuclear Magnetic Resonance. [8–10, 12](#)

**PCA** Principal Component Analysis. [11](#)

**PCPr2** Principal Component Partial R-square. [11](#)

**RT** Retention Time. [9](#)

**SNR** Signal-to-Noise Ratio. [8](#)

# Bibliography

1. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: The Apogee of the Omics Trilogy. *Nature Reviews Molecular Cell Biology* **13**, 263–269. ISSN: 1471-0080 (Apr. 2012).
2. Zhang, A., Sun, H. & Wang, X. Serum Metabolomics as a Novel Diagnostic Approach for Disease: A Systematic Review. *Analytical and Bioanalytical Chemistry* **404**, 1239–1245. ISSN: 1618-2650 (Sept. 1, 2012).
3. Chen, Y., Li, E.-M. & Xu, L.-Y. Guide to Metabolomics Analysis: A Bioinformatics Workflow. *Metabolites* **12**, 357. ISSN: 2218-1989 (4 Apr. 2022).
4. Sun, J. & Xia, Y. Pretreating and Normalizing Metabolomics Data for Statistical Analysis. *Genes & Diseases* **11**, 100979. ISSN: 2352-3042 (May 1, 2024).
5. Martin, M. *et al.* PepsNMR for <sup>1</sup>H NMR Metabolomic Data Pre-Processing. *Analytica Chimica Acta* **1019**, 1–13. ISSN: 1873-4324. pmid: [29625674](#) (Aug. 17, 2018).
6. Karaman, I. in *Metabolomics: From Fundamentals to Clinical Applications* (ed Sussulini, A.) 145–161 (Springer International Publishing, Cham, 2017). ISBN: 978-3-319-47656-8.
7. Bhinderwala, F., Roth, H., Noel, H., Feng, D. & Powers, R. Chemical Shift Variations in Common Metabolites. *Journal of magnetic resonance (San Diego, Calif. : 1997)* **345**, 107335. ISSN: 1090-7807. pmid: [36410060](#) (Dec. 2022).
8. Stanstrup, J. *et al.* The metaRbolomics Toolbox in Bioconductor and Beyond. *Metabolites* **9**, 200. ISSN: 2218-1989 (10 Oct. 2019).
9. Qiu, S. *et al.* Small Molecule Metabolites: Discovery of Biomarkers and Therapeutic Targets. *Signal Transduction and Targeted Therapy* **8**, 1–37. ISSN: 2059-3635 (Mar. 20, 2023).
10. Xiao, J. F., Zhou, B. & Ressom, H. W. Metabolite Identification and Quantitation in LC-MS/MS-based Metabolomics. *Trends in analytical chemistry : TRAC* **32**, 1–14. ISSN: 0165-9936. pmid: [22345829](#) (Feb. 1, 2012).
11. Kiseleva, O., Kurbatov, I., Ilgisonis, E. & Poverennaya, E. Defining Blood Plasma and Serum Metabolome by GC-MS. *Metabolites* **12**, 15. ISSN: 2218-1989. pmid: [35050137](#) (Dec. 24, 2021).
12. Defernez, M. & Le Gall, G. in *Advances in Botanical Research* (ed Rolin, D.) 493–555 (Academic Press, Jan. 1, 2013).
13. Troisi, J., Troisi, G., Scala, G. & Richards, S. M. in *Metabolomics Perspectives* (ed Troisi, J.) 287–379 (Academic Press, Jan. 1, 2022). ISBN: 978-0-323-85062-9.
14. Burton, L. *et al.* Instrumental and Experimental Effects in LC-MS-based Metabolomics. *Journal of Chromatography B. Hyphenated Techniques for Global Metabolite Profiling* **871**, 227–235. ISSN: 1570-0232 (Aug. 15, 2008).

15. Trygg, J., Gabrielsson, J. & Lundstedt, T. in *Comprehensive Chemometrics* (eds Brown, S. D., Tauler, R. & Walczak, B.) 1–8 (Elsevier, Oxford, Jan. 1, 2009). ISBN: 978-0-444-52701-1.
16. Alonso, A., Marsal, S. & Julià, A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Frontiers in Bioengineering and Biotechnology* **3**. ISSN: 2296-4185 (Mar. 5, 2015).
17. Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R. & Buydens, L. M. C. Warping Methods for Spectroscopic and Chromatographic Signal Alignment: A Tutorial. *Analytica Chimica Acta* **781**, 14–32. ISSN: 0003-2670 (June 5, 2013).
18. Bork, C., Ng, K., Liu, Y., Yee, A. & Pohlscheidt, M. Chromatographic Peak Alignment Using Derivative Dynamic Time Warping. *Biotechnology Progress* **29**, 394–402. ISSN: 1520-6033 (2013).
19. Veselkov, K. A. *et al.* Recursive Segment-Wise Peak Alignment of Biological (1)h NMR Spectra for Improved Metabolic Biomarker Recovery. *Analytical Chemistry* **81**, 56–66. ISSN: 1520-6882. pmid: [19049366](https://pubmed.ncbi.nlm.nih.gov/19049366/) (Jan. 1, 2009).
20. Sawall, M. *et al.* Multi-Objective Optimization for an Automated and Simultaneous Phase and Baseline Correction of NMR Spectral Data. *Journal of Magnetic Resonance* **289**, 132–141. ISSN: 1090-7807 (Apr. 1, 2018).
21. Martín-Fernández, J. A., Palarea-Albaladejo, J. & Olea, R. A. in *Compositional Data Analysis* 43–58 (John Wiley & Sons, Ltd, 2011). ISBN: 978-1-119-97646-2.
22. Zacharias, H. U., Altenbuchinger, M. & Gronwald, W. Statistical Analysis of NMR Metabolic Fingerprints: Established Methods and Recent Advances. *Metabolites* **8**, 47. ISSN: 2218-1989 (3 Sept. 2018).
23. Ulaszewska, M. M. *et al.* Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. *Molecular Nutrition & Food Research* **63**, 1800384. ISSN: 1613-4133 (2019).



Listado de apartados que son demasiado extensos para incluir dentro de la memoria y tienen un carácter autocontenido (por ejemplo, manuales de usuario, manuales de instalación, etc.)

Dependiendo del tipo de trabajo, es posible que no haya que añadir algún anexo.