# metaboPipe: A Modular Pipeline for Metabolomic Data Preprocessing

Universitat Oberta de Catalunya

**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

June 5, 2024

UNIVERSITAT DE BARCELONA

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

| | |
|---|---|
| **Title of the work:** | metaboPipe: A Modular Pipeline for Metabolomic Data Preprocessing |
| **Name of the author:** | Eduard Pérez Méndez |
| **Name of the tutor:** | Alexandre Sánchez Pla |
| **Name of the SRP:** | Carles Ventura Royo |
| **Date of delivery:** | June 5, 2024 |
| **Studies or Prgram:** | Master's degree in Bioinformatics and Biostatistics |
| **Area or the Final Work:** | Statistical Bioinformatics and Machine Learning |
| **Language of the work:** | English |
| **Keywords:** | targeted metabolomics, preprocessing, pipeline |

**Abstract**

A maximum of 250 words, detailing the purpose, contexto of application, methodology, results and conclusiones of the work.

"*Never let your sense of morality stop you from doing the right thing*"

Isaac Asimov

# Contents

**1 Introduction** **9**
1.1 Problem description . . . 9
1.2 Context and justification . . . 9
1.2.1 Preprocessing of data . . . 10
1.2.2 Pretreatment of Data . . . 11
1.3 State of the art . . . 12
1.3.1 Hypothesis . . . 13

**2 Objectives** **14**
2.1 Main Objective . . . 14
2.2 Specific Objectives . . . 14

**3 Sustainable development goals** **15**

**4 Approach and methodology** **16**
4.1 Methodology . . . 16
4.2 Planning and calendar . . . 17
4.2.1 Tasks . . . 17
4.3 Risk analysis . . . 18
4.4 Final products . . . 18

**5 Materials and methods** **20**
5.1 Literature review . . . 20
5.2 Pipeline design . . . 20
5.3 Datasets . . . 20
5.4 Packages . . . 21

**6 Results** **23**
6.1 Pipeline . . . 23
6.2 Package Overview . . . 24
6.3 Data manipulation . . . 24
6.4 Target factories . . . 25
6.5 Documentation . . . 25
6.6 Shiny app . . . 26

**7 Discussion** **27**

**8 Conclusion and future vision** **29**

Glossary 31

Bibliography 35

# List of Figures

4.1 Gantt chart showing the project timeline and milestones. . . . . 17

6.1 Flow diagram showing the steps of the metabolomic data preprocessing pipeline proposed. . . . . 23

# 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. . . . . 18

5.1 List of R packages with their versions used to develop `metaboPipe` . . . . . 22

7.1 A data matrix generated by a metabolomics platform . . . . . 28
7.2 The sampleMetadata matrix expected by the package . . . . . 28

# 1. Introduction

## 1.1 Problem 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 significantly im-

pact 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 RT 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., 1H 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.

3. 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:

1. Assessing metabolite peak areas and comparing the ratio of mean to median, with the median often considered more robust in the presence of outliers.

2. Employing Principal Component Analysis (PCA) to identify outliers, followed by techniques such as Principal Component Partial R-square (PCPr2) and Analysis of Variance (ANOVA).

3. 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

Metabolomic data arrives to the researcher in different shapes and forms depending on the method, the instrument used and the company that analyses. Most of the time those companies do a preprocessing of the data, adjusting for baseline correction, peak picking, and alignment. The preprocessing of this data is a crucial step in the analysis, as it can significantly impact the results and the conclusions drawn from the data. For MS-based analysis, tools like XCMS

[24] and MZmine [25] facilitate preprocessing, while for NMR data, packages like BATMAN [26] and RAMSY [27] offer robust preprocessing capabilities. Pretreatment techniques include handling missing values, outlier detection, imputation and normalization.

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

Since there is no consensus on the order those pretreatment techinques should be applied, the researcher has to decide which steps to take and in which order. This can lead to inconsistencies in the results and the conclusions drawn from the data. Furthermore, the amount of packages that there are for performing the preprocessing and pretreatment of metabolomic data can be overwhelming. And thus preparing the data for analysis, changing the order of the steps, or even changing the parameters of the steps can be a time-consuming task.

The project aims to develop a modular pipeline for targeted metabolomic data pretreatment, implemented in R. The pipeline is designed to enhance efficiency and modularity compared to existing solutions. As a gift to the scientific community, this project is free and open source, with detailed documentation and code available in a public repository. It's encouraged continuous community efforts to improve and expand the package.

### 1.3.1 Hypothesis

- It is possible to design and develop a general pipeline for the pretreatment of targeted metabolomic data.
- A modular pipeline that allows researchers to customize preprocessing and pretreatment steps will improve the flexibility and efficiency of metabolomic data analysis.

# 2. Objectives

## 2.1 Main Objective

1. Develop a new pipeline for the preprocessing of targeted metabolomic data with the aim of improving efficiency and modularity compared to existing pipelines. This new pipeline will be implemented in R.

## 2.2 Specific Objectives

1. Design a new pipeline for the preprocessing of targeted metabolomic data.
2. Implement the preprocessing pipeline for targeted metabolomic data using the `targets` package to ensure replicability and efficient management of computational resources.
3. Select various targeted metabolomic datasets to validate and optimize the performance of the new pipeline, analyzing its quality and consistency.
4. Make the development accessible to the scientific community by creating detailed documentation and publishing the code in a public repository.

# 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

## 4.1 Methodology

In the context of enhancing the field of metabolomics and improving the efficiency and accuracy of metabolomics reports, the choice of creating a tool for targeted metabolomic data pretreatement is a strategic decision. This tool will be developed in R, a widely used programming language in the field of bioinformatics and biostatistics. The tool will be designed to streamline the pretreatement of targeted metabolomic data, ensuring that the data is ready for further analysis and interpretation.

The tool will be modular, allowing users to select and apply specific pretreatement steps according to their needs and preferences. This modularity design makes the tool flexible, adaptable and more important expandable, as new pretreatement methods can be easily added to the pipeline and thus to the package using existent methods and tools from other packages.

In order to be modular, the package modules should take as inputs the same object type that outputs. Making it easier to chain the modules in any specific order. To achieve this objective, the package should use a main data structure to store the data. In omics data, the most used data structure is the `SummarizedExperiment` from the Bioconductor project [28]. However this data structure is not the most efficient for the pretreatement of the metabolomics data, as the `SummarizedExperiment` dataframes are designed to store the data features in a row-wise manner and the samples in a column-wise manner. Though this may fit well for other omics data, is not the most intuitive way to manipulate the data in the context of metabolomics.

Since targeted metabolomic data uses multiple dataframes to store the diferent types of data, the most intuitive and already developed data structure will be the `DatasetExperiment` from the `structToolbox` package [29]. This data structure is designed to store the data features in a column-wise manner and the samples in a row-wise manner, making it more intuitive to manipulate the data in the context of metabolomics. It also incorporate multiple methods for the manipulation of the data, making it easier to develop the package.

A deployment controller will be used to manage the order of the modules and the data flow between them. This controler will be created using the `targets` package [30], a pipeline toolkit for reproducible research. The `targets` package will ensure that the pipeline is reproducible, efficient and easy to manage.

The `targets` package only deploys the steps needed to complete the pipeline as previous steps that do not change its output will not be re-run. This makes the pipeline more efficient and faster, as only the steps that need to be re-run will be re-run.

In order to select the methods and tools that will be used in the pipeline, a deep review of the literature will be performed. This review will explore recent publications of metabolomics data and also comparative reviews of the most used methods and tools for the pretreatement of targeted metabolomic data to select the most used methods for metabolomic data pretreatement. This review will also explore and select datasets to validate and optimize the performance

of the pipeline.

## 4.2 Planning and calendar

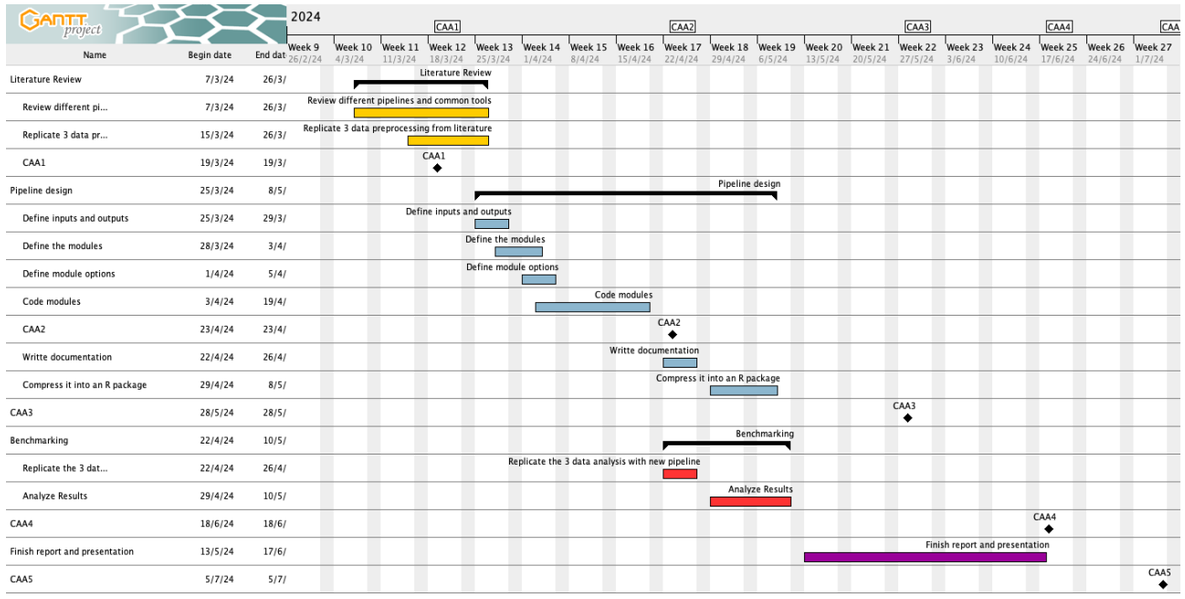


Figure 4.1: Gantt chart showing the project timeline and milestones.

### 4.2.1 Tasks

#### Main Tasks and prioritization

1. **Definition of the work plan:** Define the project's scope, objectives, methodology and expected outcomes. Create a project charter outlining the project's purpose and goals as well as a calendar with milestones and dates.

2. **Literature review:** Review multiple publications and replicate the data processing from 3 of them. Review most used methods and tools for metabolomics data processing like normalization, scaling, filtering, transformation, batch effect

3. **Pipeline design:** Propose and code in R a new pipeline to manage the processing of targeted metabolomics data.

4. **Benchmarking:** Replicate again the data processing of the 3 publications using the new pipeline and compare the results.

### Extra tasks

1. **Heavy-workload:** Optimize the pipeline to enable parallelization of the processing.
2. **Documentation:** Write a documentation for the package so its accessibility.
3. **R package implementation:** Pack the code for the pipeline in an R package to easy distribution.

## 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

- A **pipeline** for targeted metabolomic data preprocessing designed to streamline data pretreatment tasks, optimize workflows, and ensure reproducibility in metabolomic research endeavors.
- A **R package** crafted to facilitate the seamless integration and modular implementation of the pipeline within the R environment, empowering researchers with flexible and efficient tools for metabolomic data analysis.

- A detailed **documentation** accompanying the pipeline and R package.
- A user-friendly **Shiny app** that enables researchers with varying levels of computational expertise to effortlessly pretreat targeted metabolomic data.
- A **PDF report** detailing the project's processes, including investigation, development, results, conclusions, and discussions.
- A **virtual presentation** providing a comprehensive overview of the project. This includes a video recording with explanatory narration to further enhance understanding.

# 5. Materials and methods

Metodes: Com Desenvolupo //
Materials: Amb que ho fas

## 5.1 Literature review

The literature selected for the review was obtained from multiple journals. The search was conducted using the following keywords: ”metabolomics”, ”data preprocessing”, ”data pretreatment”, ”metabolomics pipeline”, ”metabolomics tools”, ”metabolomics R packages”, ”targeted metabolomics”, ”nutrimetabolomics”, ”metabolomics proce”. The search was limited to articles published in the last 20 years, with a focus on metabolomics data preprocessing and pretreatment methodologies. The review aimed to identify the most commonly used methods and tools for metabolomic data preprocessing and pretreatment, as well as to explore recent advancements in the field. The review also sought to identify gaps in the existing literature and to inform the development of the new pipeline.

## 5.2 Pipeline design

The methods selected to be included in the pipeline were based on the results of the literature review. The pipeline was designed to be as modular as possible so a variety of methods are included to perform the same step, allowing users to select and apply specific pretreatment steps according to the data caracteristics and user needs and preferences.

## 5.3 Datasets

The datasets employed for the purpose of this study were obtained from public repositories. The datasets were selected based on the following criteria:

- Aim of the dataset
- Availability of raw data
- Targeted metabolomics data
- Diverse biological samples

The datasets selected were 2:

- **MTBLS79:** [31] This dataset represents a systematic evaluation of the reproducibility of a multi-batch direct-infusion mass spectrometry (DIMS)-based metabolomics study of cardiac tissue extracts. It comprises twenty biological samples (cow vs. sheep) that were analysed repeatedly, in 8 batches across 7 days, together with a concurrent set of Quality

control (QC) samples. Data are presented from each step of the data processing workflow and are available through MetaboLights. This dataset was selected due to its importance in the field of metabolomics and its data aviability.

- **ST000284:** [32] This dataset from MetaboWorkbench includes a study on colorectal cancer (CRC) using targeted liquid chromatography-tandem mass spectrometry. It examines 158 metabolites across 25 pathways in 234 serum samples (66 CRC patients, 76 polyp patients, 92 healthy controls). Blood samples were collected after fasting and bowel preparation. This dataset was selected due to its data aviability, its diversity (3 groups) and amount of samples.

## 5.4 Packages

The pipeline was developed using the R programming language [33] and the packages described in 5.1.

| Package | Version | Ref |
|---|---|---|
| arrow | *15.0.1* | [34] |
| base | *4.3.3* | [33] |
| BiocStyle | *2.30.0* | [35] |
| caret | *6.0.94* | [36] |
| cowplot | *1.1.3* | [37] |
| crew | *0.9.2* | [38] |
| datasets | *4.3.3* | [33] |
| dplyr | *1.1.4* | [39] |
| DT | *0.33* | [40] |
| fst | *0.9.8* | [41] |
| ggforce | *0.4.2* | [42] |
| graphics | *4.3.3* | [43] |
| grDevices | *4.3.3* | [33] |
| HotellingEllipse | *1.1.0* | [44] |
| impute | *1.76.0* | [45] |
| imputeLCMD | *2.1* | [46] |
| knitr | *1.46* | [47] |
| MetaboAnalystR | *4.0.0* | [48] |
| methods | *4.3.3* | [33] |
| missForest | *1.5* | [49] |
| pcaMethods | *1.94.0* | [50] |
| pmp | *1.14.1* | [51] |
| purrr | *1.0.2* | [52] |
| renv | *1.0.7* | [53] |
| reshape2 | *1.4.4* | [54] |
| rmarkdown | *2.27* | [55] |
| shiny | *1.8.1.1* | [56] |
| shinyFiles | *0.9.3* | [57] |
| stats | *4.3.3* | [33] |
| structToolbox | *1.14.0* | [58] |
| SummarizedExperiment | *1.32.0* | [28] |
| tarchetypes | *0.9.0* | [59] |
| targets | *1.7.0* | [60] |
| tidyverse | *2.0.0* | [61] |
| tinytex | *0.51* | [62] |
| tools | *4.3.3* | [33] |
| usethis | *2.2.3* | [63] |
| utils | *4.3.3* | [33] |
| VIM | *6.2.2* | [64] |
| withr | *3.0.0* | [65] |

Table 5.1: List of R packages with their versions used to develop `metaboPipe`

# 6. Results

## 6.1 Pipeline

The literature review provided valuable insights into the most commonly used methods and tools for metabolomic data preprocessing. Based on this information, the proposed pipeline steps can be viewed on Figure 6.1

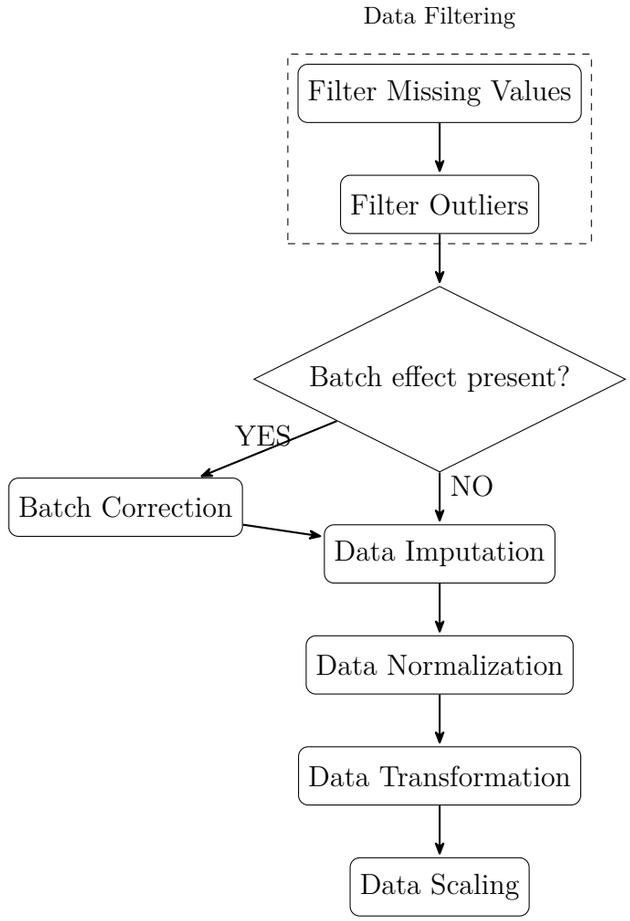


Figure 6.1: Flow diagram showing the steps of the metabolomic data preprocessing pipeline proposed.

In this project, an R package named `metaboPipe` was developed to preprocess metabolomic data efficiently and accessibly. Metabolomic data preprocessing is a crucial step in ensuring the

quality and reliability of downstream analyses. The transformations applied during preprocessing are primarily intended to remove unwanted effects unrelated to the study, but these tools can inadvertently delete meaningful biological data if used without caution. The `metaboPipe` package leverages the `targets` and `structToolbox` packages to create a modular and reproducible pipeline for preprocessing. This section details the components and functionality of `metaboPipe`.

## 6.2 Package Overview

The `metaboPipe` package is designed to streamline the preprocessing of metabolomic data by creating a series of user-accessible modules that translate into interconnected steps. These steps can be easily added, modified, and rearranged. The package utilizes the `targets` package to orchestrate task deployment and dependencies, ensuring that each step is executed in the correct order. The `structToolbox` package provides the `DatasetExperiment` object, which serves as the primary data structure throughout the pipeline, along with multiple functions for data transformation.

A total of 2400 lines of code were written to develop the `metaboPipe` package, without counting the vignette and the documentation.

The package is hosted on GitHub: https://github.com/eperezme/metaboPipe. And its main functionalities can be divided into three main categories:

- **Data manipulation:** These functions are the core of the package, as they transform the data into the desired formats and apply the diversity of methods to perform the preprocessing steps including the visualization.
- **Target factories:** They are the pipeline contruction functions for the users to interact with, as they use the targets package to create pipeline nodes that use the data manipulation functions. Each factory is intended to represent a specific pretreatement step.
- **Shiny app:** This is a user-friendly interface that allows users to create and run the pipeline without the need for R programming knowledge.

## 6.3 Data manipulation

For the pretreatement of the data, the package includes the following functions:

- A function to create a `DatasetExperiment` object from a list of dataframes.
- A function to filter the data based on the percentage of missing values both in the features and the samples.
- A function to filter the outliers based on a Hotelling's T2 distribution ellipse.
- 8 functions to impute the missing values based on the following methods: *Mean, Median, Random Forest (RF), Quantile Regression Imputation of Left-Censored data (QRILC), k-Nearest Neighbors (kNN), Singular Value Decomposition (SVD), Bayesian Principal Component Analysis (BPCA) and Probabilistic Principal Component Analysis (PPCA).*

- 2 functions to remove the batch effect using either the *Quality Control-Robust Spline Correction (QC-RSC)* method or the `ComBat` method [66, 67].
- A function to normalize, scale and transform the data based on the methods provided by the `metaboAnalystR` package.
- 12 data visualization functions to create plots.
- 17 utility functions to perform different tasks.

## 6.4 Target factories

For the pipeline construction, the package includes the following 9 functions:

- The `load_data` function is implemented to efficiently load and read data from specified files into data frames. It dynamically creates six different nodes based on the presence of the `variableMetadata` parameter.
- The `create_experiment` function is the pivotal component in the data processing pipeline, enabling the creation of a `DatasetExperiment` S4 object that will be used in almost all of the pipeline. This facilitates efficient modularity.
- The `factorize_cols` function that makes the specified columns of the `sampleMetadata` as factors.
- The `filter_step` function that creates up to 5 nodes depending on the parameters provided. These nodes filter the data based on the percentage of missing values in the features and samples, filter outliers using a Hotelling's T2 distribution ellipse and creates a plot for visualizing the data missing values before and after the filtering is done.
- The `batch_correct` function that creates a node to remove the batch effect using the `QC-RSC` method.
- The `impute` function that creates a node to impute the missing values using the method provided.
- The `normalize` function that creates a node to normalize, scale and/or transform the data using the methods provided.
- The `export_data` function that creates a node to export the data to a specified directory.

## 6.5 Documentation

The package is distributed with documentation that consists of three parts:

- The help documentation for the methods and functions of the package.

- A PDF manual that includes all the aforementioned methods and functions of the package. add the url
- A tutorial vignette that details the processing of the ST000284 dataset using the package to create and execute a pipeline. add the url

## 6.6 Shiny app

A shiny app was developed to allow users to interact with the pipeline without the need for R programming knowledge. The app provides 3 main tabs:

- The `Data upload` tab allows users to either upload their data or use one of the example data provided.
- The `Data Config` tab allows users to set the global parameters for the data, like the sample_ID column, the column to be used as QC or Sample identifier, and the main study factor. It also allows users to visualize the data tables they are working with to ensure the data is correctly loaded.
- The `Process Selector` tab allows users to select and configure the different steps of the pipeline. Each step is represented by a card that can be expanded to show the parameters of the step. The user can select the steps they want to include in the pipeline and configure them according to their needs. The app will then generate the `_targets.R` file and execute the pipeline from the app.

# 7. Discussion

Different methods for preprocessing metabolomic data are essential to accommodate the diverse types of data encountered in this field. This variety arises due to the numerous factors influencing metabolomic data, which can be biological, technical, or experimental. Each preprocessing method offers unique advantages and disadvantages, therefore choosing the appropriate method can be challenging and is guided by the specific biological question, the nature of the data, and the desired analytical outcomes.

In this work, a general workflow for targeted metabolomics data preprocessing has been proposed. This workflow is not specific as the selection of preprocessing methods affects their order. For example, when using Singular Value Decomposition (SVD) as an imputation method, it is recommended to scale and center the data beforehand, but if we impute the data using Quantile Regression Imputation of Left-Censored Data (qRILC), log-transforming the data beforehand is recommended for better accuracy [68]. When normalizing Quantile normalization (QN) is not recomended for datasets with $n < 50$ samples [69] and Probabilistic Quotient Normalization (PQN) is not adequate to use when the number of metabolites is $>$ than the number of samples [70]. These considerations must be factored in when selecting preprocessing methods, making it impractical to establish a one-size-fits-all pipeline

pq el ordre de la pipeline

Due to the diversity of methods and their interactions, this work focuses on developing a tool that facilitates the easy and interactive creation of customized data preprocessing sequences. This tool, the `metaboPipe` package, offers functions to perform common preprocessing steps and enables the creation of tailored pipelines using these steps. Additionally, the package includes a shiny app that allows users to interact with the pipeline without requiring R programming skills.

The `metaboPipe` package is designed to work with the `targets` and `structToolbox` packages. To create a pipeline using the `metaboPipe` package, users must define a `_targets.R` file that specifies the steps of the pipeline. A pre-configured template is aviable using the `create_pipeline()` function that creates a `_targets.R` file in the working directory with the necessary packages configured. Then users only have to add the targets or commands to the list in the file and run the pipeline using the `tar_make()` function.

To load the data, users need to import two tables in `.csv` format: `dataMatrix` and `sampleMetadata`. The `dataMatrix` table contains the relative abundance or concentration data (Table 7.1), while the `sampleMetadata` table holds the sample-related data (Table 7.2).

For filtering missing values, the modified 80% rule is adopted by default: if a column or row has more than 20% missing values, that column or row is removed. Users can manually adjust the threshold for missing values.

To identify and filter outliers, a Hotelling's T2 distribution ellipse is used with a default significance level of 95%, adjustable to 99% if desired. This method is chosen for its robustness and flexibility in detecting outliers in multivariate data, considering the covariance between variables [71, 72].

Batch effects are a common issue in metabolomics studies, introducing unwanted variability

Table 7.1: A data matrix generated by a metabolomics platform

| $n \times m$ | $Metabolite_1$ | $Metabolite_2$ | ... | $Metabolite_m$ |
|---|---|---|---|---|
| $Sample_1$ | | | | |
| $Sample_2$ | | | | |
| ... | ... | ... | ... | ... |
| $Sample_n$ | | | | |

Table 7.2: The sampleMetadata matrix expected by the package

| $n \times m$ | $Factor_1$ | $Factor_2$ | ... | $Factor_m$ |
|---|---|---|---|---|
| $Sample_1$ | | | | |
| $Sample_2$ | | | | |
| ... | ... | ... | ... | ... |
| $Sample_n$ | | | | |

into the data. To correct this effect, two methods have been implemented: the `QC-RSC` method and the `ComBat` method, which are widely used .

add citations

Missing values are imputed using the eight most common methods: Mean of the feature, Median of the feature,Random Forest (RF), qRILC, k-Nearest Neighbors (k-NN), SVD, Bayesian Principal Component Analysis (BPCA), and Probabilistic Principal Component Analysis (PPCA). These methods were selected based on their popularity and effectiveness in imputing missing values in metabolomic data [4, 68].

For normalization, scaling, and transformation of the data, it has been decided to use the methods provided by the `metaboAnalystR` package as it offers a wide variety of methods, some of which are difficult to implement manually, such as normalization to an internal standard or a physiological constant. However, several functions are implemented to normalize the data using other methods not present in `metaboAnalystR` such as Probabilistic Quotient Normalization (PQN) or Vector Length Normalization (VLN). These methods, although implemented in `metaboPipe`, are not incorporated into the shiny application interface nor the `normalize()` function.

The implementation of the PC-PR2 method described by Viallon *et al.* is also considered. This method first identifies sources of variation using PCA combined with multiple linear regression and then corrects unwanted variations using a random effects model for each metabolite. However, due to the complexity of robustly integrating this method into the package within the available time, it is deferred to future revisions. Nevertheless, the `pcpr2()` function can be used to visualize the data and identify sources of variation, although correction is not yet possible.

As `metaboPipe` is based on the `structToolbox` package, users can easily access the functions provided by the package to perform additional steps not included in `metaboPipe` like the filtering of QC samples, the removal of metabolites with greater specific Relative Standard Deviation (RSD) and more.

# 8. Conclusion and future vision

In this work, an attempt has been made to propose a general pipeline for the preprocessing of targeted metabolomics data. However, the diversity of methods and their interactions make it impossible to establish a pipeline that fits all cases in the timespan of a master's thesis. Therefore, a tool has been developed that facilitates the creation of customized data preprocessing sequences. By doing it this way, a tool for automation has been created for both researchers and data analysts, regardless of their level of programming knowledge.

Data preprocessing is a complex task that requires significant attention and expertise to avoid introducing unwanted variance into the data. Until a universal pipeline is developed, the `metaboPipe` package can serve as a valuable tool to address this need and to compare various data preprocessing methods, thereby aiding in the development of a comprehensive pipeline.

The project's planning underwent significant changes as new challenges emerged, requiring adjustments to ensure the continuation of the work and the creation of a useful product by the end of May. Nevertheless, it was possible to adhere to the planning for the most part, and the main objective was nearly completed.

Overall, the positive impacts anticipated in section 3 have been largely achieved, while potential negative impacts have been actively mitigated through careful planning and ethical considerations. Our project has made significant contributions to sustainable development goals, fostering a more inclusive and equitable scientific community.

The planned methodology was sufficiently adequate in terms of its strategic decisions, choice of tools, and overall design. It successfully created a modular, flexible, and efficient tool for metabolomics data preprocessing. However, time constraints and the need for methodological adjustments during the project development phase indicate areas where the methodology could have been further optimized or expanded given more time and resources. Despite these challenges, the project has laid a strong foundation for future development and refinement.

Despite the efforts made, while a new pipeline has been proposed to improve the preprocessing of metabolomics data, the lack of time to generate and analyze comparative results means that its effectiveness has not been demonstrated. Consequently, it cannot be stated that the objectives have been fully met.

Several lines of future work have been identified as a result of this project. These areas for further exploration include:

- **Comparison with existing pipelines:** While a new pipeline has been created that potentially enhances the preprocessing of metabolomics data, its effectiveness has yet to be demonstrated. It would be beneficial to compare this pipeline with existing ones to evaluate its impact on data preprocessing.

- **Implementation of additional preprocessing methods:** Although the most common preprocessing methods have been implemented, there are many more that could be incorporated. Future versions of the package could include these additional methods to further enhance its capabilities.

- **Enhancement of the graphical interface:** The current graphical interface is functional but could be improved by adding more features. For example, more configuration options for different preprocessing methods or a tab for visualizing resulting graphs could greatly enhance usability.
- **Expansion of documentation:** While the package documentation is comprehensive, it could benefit from more examples and tutorials to assist users. Additionally, more detailed explanations of how different preprocessing methods work could be included.
- **Inclusion in CRAN:** Presently, the package is available on GitHub, which, although convenient for programmers, can pose an additional barrier for those with limited programming knowledge. Including the package in CRAN would make installation and usage easier for a broader audience.
- **Integration of graphics creation methods as pipeline nodes:** Currently, graphics creation methods are independent functions. Integrating these as pipeline nodes would allow users to visualize results at each step of the pipeline, enhancing the overall functionality and user experience in the Shiny graphical interface.

Although the initial objectives have not been entirely met, a tool has been created that is expected to help achieve them in the future. An unforeseen function has emerged with its creation: the ability to use the package as a pretreatment method comparison tool. Given its ease of use and flexibility, users can test different pretreatment methods and sequences to compare the resulting changes in the data.

# Glossary

**Analysis of Variance (ANOVA)**

A statistical technique used to analyze the differences between group means. ANOVA tests whether there are any statistically significant differences between the means of three or more independent groups. It decomposes the total variability in the data into variability between groups and within groups, allowing researchers to determine if the group membership explains a significant portion of the variability. ANOVA is widely used in experimental design and research across various scientific disciplines. 12

**Bayesian Principal Component Analysis (BPCA)**

A probabilistic approach to principal component analysis that incorporates prior knowledge. BPCA models the data as a probabilistic distribution and uses Bayesian inference to estimate the principal components. This approach allows for more robust dimensionality reduction, especially in cases with small sample sizes or missing data. BPCA is useful in fields such as bioinformatics and signal processing, where incorporating prior knowledge can improve the analysis and interpretation of complex datasets. 28

**Gas Chromatography (GC)**

A separation technique in which a gaseous mobile phase is passed through a column packed with a stationary phase. Compounds in the sample are vaporized and carried by an inert gas through the column, where they interact with the stationary phase. This causes them to separate based on their boiling points and affinities to the stationary phase. Gas chromatography is commonly used for the analysis of volatile substances and can be coupled with detectors like mass spectrometry for enhanced identification and quantification. 9, 10

**Gas Chromatography-Mass Spectrometry (GC-MS)**

A hyphenated technique that combines gas chromatography with mass spectrometry. GC-MS is used for the separation and identification of volatile and semi-volatile compounds. In this technique, gas chromatography separates the components of a mixture, and mass spectrometry provides detailed molecular information about each component. GC-MS is widely used in forensic science, environmental analysis, and the pharmaceutical industry for its ability to provide both qualitative and quantitative data on complex mixtures. 9, 10

**k-Nearest Neighbors (k-NN)**

A non-parametric classification algorithm that classifies new data points based on their proximity to known data points. KNN assigns a class to a data point by finding the k

nearest neighbors in the feature space and taking a majority vote among their classes. This method is simple, intuitive, and effective for many types of data. KNN is widely used in pattern recognition, data mining, and machine learning applications, particularly when the decision boundaries are complex and non-linear. 28

### Liquid Chromatography (LC)

A separation technique in which a liquid mobile phase is pumped through a column packed with a stationary phase. The different components in the sample interact with the stationary phase to varying degrees, causing them to elute at different times. This method is widely used for the separation and analysis of complex mixtures, such as those found in pharmaceuticals, environmental samples, and biological materials. The efficiency and specificity of the separation can be influenced by the choice of stationary and mobile phases. 9, 10

### Liquid Chromatography-Mass Spectrometry (LC-MS)

A hyphenated technique that combines liquid chromatography with mass spectrometry. LC-MS integrates the separation capabilities of liquid chromatography with the detection and identification power of mass spectrometry. This combination allows for the separation, identification, and quantification of complex mixtures, making it invaluable in proteomics, metabolomics, and pharmaceutical research. The technique provides high sensitivity and specificity, enabling detailed analysis of biomolecules and small organic compounds. 9, 10

### Mass Spectrometry (MS)

An analytical technique that measures the mass-to-charge ratio of ions. It involves ionizing chemical compounds to generate charged molecules or molecule fragments and measuring their mass-to-charge ratios. Mass spectrometry is highly sensitive and can be used to determine the molecular weight of compounds, identify unknown substances, and study the structure and composition of complex mixtures. The technique is widely applied in fields such as chemistry, biochemistry, and pharmacology. 9–12

### Mass-to-Charge Ratio (m/z)

The ratio of the mass of an ion to its charge. This parameter is fundamental in mass spectrometry, where it is used to characterize and identify ions based on their mass and charge states. The mass-to-charge ratio is denoted as m/z, and analyzing the m/z values allows scientists to determine the molecular weight and structure of compounds. This information is crucial for applications in proteomics, metabolomics, and chemical analysis. 10

### Nuclear Magnetic Resonance (NMR)

A spectroscopic technique that exploits the magnetic properties of atomic nuclei. NMR involves applying a magnetic field to a sample and then measuring the absorption of radiofrequency radiation by the nuclei. This provides detailed information about the molecular structure, dynamics, and environment of the atoms within the sample. NMR

is widely used in chemistry and biochemistry for the elucidation of molecular structures and in medical imaging as the basis for magnetic resonance imaging (MRI). 9–11, 13

**Principal Component Analysis (PCA)**

A statistical technique used to reduce the dimensionality of data. PCA transforms the original variables into a new set of uncorrelated variables called principal components, which capture the most variance in the data. This technique simplifies the data while retaining its essential patterns, making it easier to visualize and analyze. PCA is widely used in fields such as bioinformatics, finance, and image processing for data compression, noise reduction, and feature extraction. 12

**Principal Component Partial R-square (PCPr2)**

A measure of the contribution of each principal component to the explained variance in a regression model. PC-PR2 helps in identifying which principal components are most significant in predicting the dependent variable. This metric is useful in regression analysis when dealing with high-dimensional data, as it aids in selecting the most relevant components and improving model interpretability and performance. 12

**Probabilistic Principal Component Analysis (PPCA)**

A probabilistic approach to principal component analysis that models the data as a probabilistic distribution. PPCA assumes that the observed data are generated by a linear transformation of latent variables with added Gaussian noise. This approach provides a more flexible and robust framework for dimensionality reduction and allows for handling missing data and uncertainty in the analysis. PPCA is widely used in machine learning, bioinformatics, and image processing. 28

**Quality control (QC)**

In metabolomics studies, quality control (QC) samples are internal checks to ensure data reliability. They are analyzed alongside biological samples and act as a reference point. The most common type is a pooled QC, created by mixing a tiny bit of each biological sample. By analyzing the QC throughout the experiment, researchers can monitor for issues like instrument drift or errors in sample preparation. This helps ensure the data accurately reflects what's happening in the biological samples themselves. 20, 26, 28

**Quantile Regression Imputation of Left-Censored Data (qRILC)**

A statistical method for estimating missing values in left-censored data using quantile regression. Left-censored data occur when measurements fall below a detection limit, resulting in a large number of values reported as below this threshold. QRILC uses quantile regression to estimate the distribution of the censored data and imputes missing values accordingly. This technique is particularly useful in environmental and biomedical studies where such data are common. 27, 28

**Random Forest (RF)**

An ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest is built from a random subset of the data and features, and the

final prediction is obtained by aggregating the predictions of all trees. Random forests are robust to overfitting and can handle large datasets with high dimensionality. They are widely used for classification, regression, and feature selection tasks in various fields, including bioinformatics, finance, and marketing. 28

**Retention Time (RT)**

The time it takes for a compound to travel through a chromatographic column. Retention time is a key parameter in chromatography and is used to help identify and quantify the components in a mixture. Each compound has a characteristic retention time under specific chromatographic conditions, allowing for its separation and analysis. Accurate measurement of retention time is essential for reproducibility and comparison across different experiments and laboratories. 10

**Signal-to-Noise Ratio (SNR)**

A measure of the strength of a signal relative to the background noise. SNR is a crucial parameter in various fields, including telecommunications, imaging, and analytical chemistry. A higher SNR indicates a clearer and more reliable signal, making it easier to detect and analyze the information of interest. Improving SNR can involve increasing the signal strength, reducing noise, or both, and is often a key focus in the design and optimization of measurement and detection systems. 9

**Singular Value Decomposition (SVD)**

A matrix factorization technique that decomposes a matrix into singular values and singular vectors. SVD expresses a matrix as the product of three matrices: a diagonal matrix of singular values and two orthogonal matrices of singular vectors. This decomposition is used in many applications, including data compression, noise reduction, and solving linear systems. SVD is a fundamental tool in numerical linear algebra and is widely used in machine learning, signal processing, and bioinformatics. 27, 28

# 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 1H 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 (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).

24. Smith, C. A. *et al. xcms: LC-MS and GC-MS Data Analysis* R package version 4.0.2 (2024). https://bioconductor.org/packages/xcms.

25. Schmid, R. *et al.* Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. *Nature Biotechnology* **41,** 447–449. ISSN: 1546-1696 (Apr. 2023).

26. Keyes, O. *et al. batman: Convert Categorical Representations of Logicals to Actual Logicals* R package version 0.1.0 (2015). https://CRAN.R-project.org/package=batman.

27. Gu, H., Gowda, G. A. N., Neto, F. C., Opp, M. R. & Raftery, D. RAMSY: Ratio Analysis of Mass Spectrometry to Improve Compound Identification. *Analytical Chemistry* **85,** 10771–10779. ISSN: 0003-2700 (Nov. 19, 2013).

28. Morgan, M., Obenchain, V., Hester, J. & Pagès, H. *SummarizedExperiment: SummarizedExperiment container* R package version 1.32.0 (2023). https://bioconductor.org/packages/SummarizedExperiment.

29. Lloyd, G. R., Jankevics, A. & Weber, R. J. M. struct: an R/Bioconductor-based framework for standardized metabolomics data analysis and beyond. *Bioinformatics* **36,** 5551–5552. https://doi.org/10.1093/bioinformatics/btaa1031 (2020).

30. Landau, W. M. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software* **6,** 2959. https://doi.org/10.21105/joss.02959 (2021).

31. Kirwan, J. A., Weber, R. J. M., Broadhurst, D. I. & Viant, M. R. Direct Infusion Mass Spectrometry Metabolomics Dataset: A Benchmark for Data Processing and Quality Control. *Scientific Data* **1,** 140012. ISSN: 2052-4463 (June 10, 2014).

32. Zhu, J. *et al.* Colorectal Cancer Detection Using Targeted Serum Metabolic Profiling. *Journal of Proteome Research* **13,** 4120–4130. ISSN: 1535-3907. pmid: 25126899 (Sept. 5, 2014).

33. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2024). https://www.R-project.org/.

34. Richardson, N. *et al. arrow: Integration to Apache 'Arrow'* R package version 15.0.1 (2024). https://github.com/apache/arrow/.

35. Oleś, A. *BiocStyle: Standard styles for vignettes and other Bioconductor documents* R package version 2.30.0 (2023). https://bioconductor.org/packages/BiocStyle.

36. Kuhn, M. *caret: Classification and Regression Training* R package version 6.0-94 (2023). https://github.com/topepo/caret/.

37. Wilke, C. O. *cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2* R package version 1.1.3 (2024). https://wilkelab.org/cowplot/.

38. Landau, W. M. *crew: A Distributed Worker Launcher Framework* R package version 0.9.2 (2024). https://wlandau.github.io/crew/.

39. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *dplyr: A Grammar of Data Manipulation* R package version 1.1.4 (2023). https://dplyr.tidyverse.org.

40. Xie, Y., Cheng, J. & Tan, X. *DT: A Wrapper of the JavaScript Library DataTables* R package version 0.33 (2024). https://github.com/rstudio/DT.

41. Klik, M. *fst: Lightning Fast Serialization of Data Frames* R package version 0.9.8 (2022). http://www.fstpackage.org.

42. Pedersen, T. L. *ggforce: Accelerating ggplot2* R package version 0.4.2, https://github.com/thomasp85/ggforce (2024). https://ggforce.data-imaginist.com.

43. Gentleman, R., Whalen, E., Huber, W. & Falcon, S. *graph: A package to handle graph data structures* R package version 1.80.0 (2023). https://bioconductor.org/packages/graph.

44. Goueguel, C. L. *HotellingEllipse: Hotelling T-Square and Confidence Ellipse* R package version 1.1.0 (2022). https://github.com/ChristianGoueguel/HotellingEllipse.

45. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. *impute: Imputation for microarray data* R package version 1.76.0 (2023). https://bioconductor.org/packages/impute.

46. Lazar, C. & Burger, T. *imputeLCMD: A Collection of Methods for Left-Censored Missing Data Imputation* R package version 2.1 (2022). https://CRAN.R-project.org/package=imputeLCMD.

47. Xie, Y. *knitr: A General-Purpose Package for Dynamic Report Generation in R* R package version 1.46 (2024). https://yihui.org/knitr/.

48. Xia, J., Chong, J. & Pang, Z. *MetaboAnalystR: An R Package for Comprehensive Analysis of Metabolomics Data* R package version 4.0.0 (2024).

49. Stekhoven, D. J. *missForest: Nonparametric Missing Value Imputation using Random Forest* R package version 1.5 (2022). https://www.r-project.org.

50. Stacklies, W., Redestig, H. & Wright, K. *pcaMethods: A collection of PCA methods* R package version 1.94.0 (2023). https://bioconductor.org/packages/pcaMethods.

51. Jankevics, A., Lloyd, G. R. & Weber, R. J. M. *pmp: Peak Matrix Processing and signal batch correction for metabolomics datasets* R package version 1.14.1 (2024).

52. Wickham, H. & Henry, L. *purrr: Functional Programming Tools* R package version 1.0.2 (2023). https://purrr.tidyverse.org/.

53. Ushey, K. & Wickham, H. *renv: Project Environments* R package version 1.0.5 (2024). https://rstudio.github.io/renv/.

54. Wickham, H. *reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package* R package version 1.4.4 (2020). https://github.com/hadley/reshape.

55. Allaire, J. *et al. rmarkdown: Dynamic Documents for R* R package version 2.26, https://pkgs.rstudio.com/rmarkdown/ (2024). https://github.com/rstudio/rmarkdown.

56. Chang, W. *et al. shiny: Web Application Framework for R* R package version 1.8.1.1 (2024). https://shiny.posit.co/.

57. Pedersen, T. L., Nijs, V., Schaffner, T. & Nantz, E. *shinyFiles: A Server-Side File System Viewer for Shiny* R package version 0.9.3 (2022). https://github.com/thomasp85/shinyFiles.

58. Lloyd, G. R. & Weber, R. J. M. *structToolbox: Data processing & analysis tools for Metabolomics and other omics* R package version 1.14.0 (2023). https://bioconductor.org/packages/structToolbox.

59. Landau, W. M. *tarchetypes: Archetypes for Targets* R package version 0.9.0, https://github.com/ropensci/tarchetypes (2024). https://docs.ropensci.org/tarchetypes/.

60. Landau, W. M. *targets: Dynamic Function-Oriented Make-Like Declarative Pipelines* R package version 1.6.0, https://github.com/ropensci/targets (2024). https://docs.ropensci.org/targets/.

61. Wickham, H. *tidyverse: Easily Install and Load the Tidyverse* R package version 2.0.0, https://github.com/tidyverse/tidyverse (2023). https://tidyverse.tidyverse.org.

62. Xie, Y. *tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents* R package version 0.51 (2024). https://github.com/rstudio/tinytex.

63. Wickham, H., Bryan, J., Barrett, M. & Teucher, A. *usethis: Automate Package and Project Setup* R package version 2.2.3 (2024). https://usethis.r-lib.org.

64. Templ, M., Kowarik, A., Alfons, A., de Cillia, G. & Rannetbauer, W. *VIM: Visualization and Imputation of Missing Values* R package version 6.2.2 (2022). https://github.com/statistikat/VIM.

65. Hester, J. *et al. withr: Run Code With Temporarily Modified Global State* R package version 3.0.0 (2024). https://withr.r-lib.org.

66. Kirwan, J. A., Broadhurst, D. I., Davidson, R. L. & Viant, M. R. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. *Analytical and Bioanalytical Chemistry* **405,** 5147–5157. ISSN: 1618-2650. https://doi.org/10.1007/s00216-013-6856-7 (2024) (June 1, 2013).

67. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **8,** 118–127. ISSN: 1465-4644 (Jan. 1, 2007).

68. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports* **8,** 663. ISSN: 2045-2322 (Jan. 12, 2018).

69. Kohl, S. M. *et al.* State-of-the Art Data Normalization Methods Improve NMR-based Metabolomic Analysis. *Metabolomics* **8,** 146–160. ISSN: 1573-3882. pmid: 22593726 (Suppl 1 2012).

70. Dong, J., Cheng, K.-K., Xu, J., Chen, Z. & Griffin, J. L. Group Aggregating Normalization Method for the Preprocessing of NMR-based Metabolomic Data. *Chemometrics and Intelligent Laboratory Systems* **108,** 123–132. ISSN: 0169-7439 (Oct. 15, 2011).

71. Wu, J. *et al.* Using Nontargeted LC-MS Metabolomics to Identify the Association of Biomarkers in Pig Feces with Feed Efficiency. *Porcine Health Management* **7,** 39. ISSN: 2055-5660 (June 2, 2021).

72. Viallon, V. *et al.* A New Pipeline for the Normalization and Pooling of Metabolomics Data. *Metabolites* **11,** 631. ISSN: 2218-1989 (9 Sept. 2021).