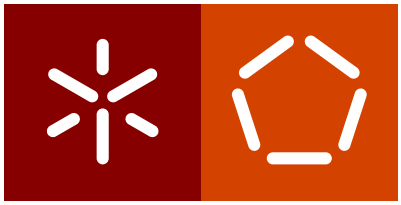


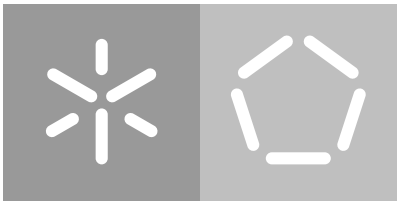
Universidade do Minho
Escola de Engenharia
Departamento de Informática

Bruno Miguel Marques Pereira

Development of computational tools for the analysis of 2D-nuclear magnetic resonance data

Master thesis

February 2021



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Bruno Miguel Marques Pereira

Development of computational tools for the analysis of 2D-nuclear magnetic resonance data

Master thesis

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Miguel Francisco Almeida Pereira da Rocha
Marcelo Maraschin

February 2021

COPYRIGHTS

ACKNOWLEDGMENTS

Quero deixar em primeiro lugar o meu agradecimento ao professor e orientador Miguel Rocha, por me ter proporcionado a oportunidade para desenvolver esta dissertação, por toda a formação instruída durante o mestrado, pelo acompanhamento e pelas oportunidades que advieram do trabalho realizado. Quero deixar o meu obrigado também ao Marcelo Marashin, da Universidade Federal de Santa Catarina, pela orientação referente aos dados espectrais e análise dos mesmos.

Em segundo lugar, quero agradecer aos meus colegas de mestrado que ao longo dos últimos tempos têm acompanhado o meu trabalho. O ambiente gerado por vós permitiu criar um espírito de aprendizagem saudável e enriquecedor, tanto a nível académico como social. A vossa companhia marca desta forma este momento da minha formação, tornando aqui explícito o meu desejo de boa sorte nos vossos percursos.

Quero agradecer também ao meu grupo de amigos de longa data por me apoiarem e acreditarem em mim, trazendo sempre uma mensagem de coragem e de esperança. De uma forma especial, quero agradecer ao Nuno e ao Bruno que mais diretamente manifestaram esta amizade singular que nos une a todos.

Por último e, não menos importante, quero agradecer aos meus pais. Sempre me apoiaram ao longo do meu percurso académico com especial atenção neste último ano, com o carinho e amor sempre presentes, tornando-se para mim exemplos de vida. São o meu orgulho pelo esforço que fizeram para que eu pudesse atingir este objetivo, e, por isso mesmo, obrigado por estarem ao meu lado nos bons momentos e nos menos bons também.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Metabolomics is one of the omics' sciences that has been gaining a lot of interest due to its potential on correlating an organism's biochemical activity and its phenotype. The applications of metabolomics are being extended as new techniques reveal new information on metabolic profiles and molecules, thus elucidating biological, chemical and functional knowledge. The main techniques that collect data are based on mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy. The last one has the advantage of analyzing a sample *in vivo* without damaging it and while its sensitivity is pointed out as a disadvantage, multidimensional NMR delivers a solution to this issue. It adds layers of information, generating new data that requires advanced bioinformatics methods in order to extract biological meaning.

Since multidimensional NMR has different approaches within itself, the need to establish an integrated framework that allows a researcher to load its data and extract relevant knowledge has become more imperative over the years. Also, establishing common data analysis pipelines on one-dimensional and multidimensional NMR remains a challenge in current scientific research hindering reproducibility across research groups.

In recent work from the host group, *specmine*, an R package for metabolomics and spectral data analysis/mining, has been developed to wrap and deliver key metabolomic methods that allow a researcher to perform a complete analysis.

In this dissertation, tools integrated in *specmine* were developed to read, visualize and analyze two-dimensional (2D) NMR. A new *specmine* structure was created for this type of data, easing interpretation and data visualization. In terms of visualization a novel approach towards three-dimensional environments enables users to interact with their data allowing peak hovering or identification of rich resonance regions. The selection of which samples to plot, when the user does not specify an input, is based on a signal-to-noise ratio scale which plots samples with opposite signal-to-noise ratios. A method to perform peak detection on 2D NMR based on local maximum search was implemented to obtain a data structure that best benefits from *specmine*'s functionalities. These include preprocessing, univariate and multivariate analysis as well as machine learning and feature selection methods.

The 2D NMR functions were validated using experimental data from two scientific papers, available on metabolomic databases and applying the necessary preprocessing steps to compare spectra and results. These data originated two case studies from different NMR sources, Bruker and Varian, which reinforces *specmine*'s flexibility. The case studies were carried out using mainly *specmine* and other packages for specific processing steps, such as, probabilistic quotient normalization. A pipeline to analyze 2D NMR was added to *specmine*, in a form of a vignette, to provide a guideline for the newly developed functionalities.

Keywords: 2D NMR; Metabolomics; Multivariate analysis; Nuclear Magnetic Resonance (NMR); Univariate analysis

RESUMO

A metabolómica é uma das ciências ómicas que tem vindo a ganhar muito interesse devido ao seu potencial para correlacionar a atividade bioquímica de um organismo com o seu fenótipo. As aplicações da metabolómica estão em constante crescimento à medida que novas técnicas revelam nova informação sobre perfis metabólicos e moleculares, elucidando conhecimento biológico, químico e funcional. As principais técnicas para recolher este tipo de dados são baseadas em espectrometria de massa e em ressonância magnética nuclear (RMN). Esta última tem a vantagem de analisar uma amostra *in vivo* sem a danificar e enquanto a sensibilidade da mesma tem sido apontada como uma desvantagem, surge a abordagem de RMN multidimensional melhorando a versão tradicional. Através da medição de outros núcleos adiciona camadas de informação, gerando um novo tipo de dados que requere métodos bioinformáticos avançados para se extrair significado biológico.

A existência de várias abordagens para realizar RMN multidimensional leva à crescente necessidade da existência de uma ferramenta que integre este tipo de dados, de forma a permitir ao investigador executar a sua análise de forma eficaz. Adicionalmente, a consolidação de pipelines comuns para analisar dados de RMN uni- e multidimensional permanece um desafio à investigação científica, dificultando a reprodutibilidade de resultados por diferentes grupos de investigação.

Em trabalhos recentes do grupo de acolhimento foi desenvolvido um package para o programa R focado na metabolómica e na análise/mineração de dados. Este package, *specmine*, tem sido melhorado desde o seu desenvolvimento funcionando como uma ferramenta que engloba diferentes métodos permitindo uma análise total a um determinado conjunto de dados. Baseado neste package, mais recentemente foi desenvolvida uma plataforma web integrada, *WebSpecmine*, com o mesmo propósito que providencia ao utilizador uma interface de utilizador mais fácil e amigável.

Nesta dissertação, ferramentas que permitem a leitura, visualização e análise de NMR bidimensional (2D) foram desenvolvidas tendo em conta a sua integração no *specmine*. Uma nova estrutura foi adicionada ao package, facilitando a interpretação e esquemetazição dos dados. Quanto à visualização, uma abordagem inovadora para

ambientes tridimensionais permite ao utilizador interagir com os seus dados através da identificação de regiões espectrais de interesse ou reconhecimento de picos. A visualização de espectros 2D, sem especificação por parte do utilizador, tem por base uma escala de relação sinal/ruído que permite numa primeira instância visualizar as amostras com uma maior e menor diferença entre sinal e ruído. Foi também implementado um método para realizar a deteção de picos em RMN 2D baseado na procura por valores máximos locais. Esta operação tem por objectivo obter uma estrutura de dados simplificada que melhor beneficia das funcionalidades do *specmine*. Estas incluem operações de pré-processamento, análises uni- e multivariada, métodos de seleção de variáveis e aprendizagem máquina.

As funções desenvolvidas para RMN 2D foram validadas com dados experimentais recolhidos de dois artigos científicos, disponíveis em bases de dados de metabolómica e sobre os quais foram aplicados os passos de pré-processamento que permitissem a comparação de resultados. Estes dados originaram dois casos de estudos que abordavam diferentes intrumentos utilizados em RMN, Bruker e Varian, reforçando desta forma a flexibilidade do *specmine* relativamente às tipologias de dados capazes de serem lidas. Estes casos foram realizados utilizando principalmente o *specmine*, no entanto, a utilização de packages externos foi necessária para passos de processamento específicos, como por exemplo, a normalização por quociente probabilístico. Uma pipeline para análise de dados RMN 2D foi adicionada ao *specmine*, sob a forma de vignette, um formato de documentação longa adequado a packages implementados no programa R. Desta forma é proporcionado ao utilizador um conjunto de procedimentos, orientados à utilização correta das funcionalidades implementadas.

Palavras-Chave: 2D NMR; Análise multivariada; Análise univariada; Metabolómica; Ressonância Magnética Nuclear (RMN)

CONTENTS

ACRONYMS

1D One-Dimensional

1D-NMR One-Dimensional Nuclear Magnetic Resonance Spectroscopy

1D-TOCSY One-Dimensional Total Correlation Spectroscopy

2D Two-Dimensional

2D-JRES-NMR Two-Dimensional *J*-Resolved Nuclear Magnetic Resonance Spectroscopy

2D-NMR Two-Dimensional Nuclear Magnetic Resonance Spectroscopy

2D-TOCSY Two-Dimensioanl Total Correlation Spectroscopy

3D Three-Dimensional

AMDIS Automated Mass Spectral Deconvolution and Identification System

ANOVA Analysis of Variance

APCI Atmospheric Pressure Chemical Ionization

AUC Area Under the Curve

BATMAN Bayesian Automated Metabolite Analyser for Nuclear Magnetic Resonance spectra

BQuant awBayesian Quantification

CI Confidence Intervals

CKD Chronic Kidney Disease

COSY Correlation Spectroscopy

CSF Cerebrospinal Fluid

CV	Coefficient of Variance
DNP	Dynamic Nuclear Polarization
DOSY	Diffusion Ordered Spectroscopy
DQF-COSY	Double Quantum Filtered Correlation Spectroscopy
EI	Electron Impact
ESI	Electrospray Ionization
FDR	False Discovery Rate
FID	Free Induction Decay
FMQ	Fast Metabolite Quantification
FT	Fourier Transformation
FTICR	Fourier Transform Ion Cyclotron Resonance
GC	Gas Chromatography
GC-MS	Gas Chromatography-Mass Spectrometry
HC	Hierarchical Clustering
HILIC	Hydrophilic Interaction Chromatography
HMBC	Heteronuclear Multiple Bond Correlation
HMDB	Human Metabolome Database
HMQC	Heteronuclear Multiple-Quantum Correlation/Coherence Spectroscopy
HPLC	High Performance Liquid Chormatography
HSD	Honestly Significant Difference
HSQC	Heteronuclear Single Quantum Correlation/Coherence Spectroscopy
INEPT	Insensitive Nuclei Enhanced by Polarization Transfer
IR	Infrared Spectroscopy

KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest Neighbour
LC	Liquid Chromatography
LC-MS	Liquid Chromatography - Mass Spectrometry
m/z	Mass-to-Charge Ratio
MALDI	Matrix-Assisted Laser Desorption Ionization
MCMC	Markov Chain Monte Carlo
MHz	Megahertz
MS/MS	Tandem Mass Spectrometry
MS	Mass Spectrometry
NA	Not Available
NLS	Nonlinear sampling
NMF	Non-negative Matrix Factorization
NMR	Nuclear Magnetic Resonance
O-PLS	Orthogonal Projections to Latent Structures
OPLS-DA	Orthogonal Partial Least Squares - Discriminant Analysis
PCA	Principal Component Analysis
PCs	Principal Components
PLS	Partial Least Squares
PLS-DA	Partial Least Squares - Discriminant Analysis
ppm	Parts per Million
Q	Quadrupole

QUIPU-HSQC QUantItative, Perfected and pUre shifted Heteronuclear Single Quantum Correlation/Coherence Spectroscopy

RdF Random Forests

RF Radio Frequency

RFE Recursive Feature Elimination

RMSE Root Mean Square Error

ROC Receiver Operating Characteristic

RPLC Reversed-Phase Liquid Chromatography

SFC Supercritical Fluid Chromatography

SFC-MS Supercritical Fluid Chromatography - Mass Spectrometry

SNR Signal-to-Noise Ratio

STOCSY Statistical Total Correlation Spectroscopy

SVM Support Vector Machine

TOCSY Total Correlation Spectroscopy

TOF Time of Flight

TQ Triple Quadrupole

UPLC Ultra-high Performance Liquid Chromatography

UV-vis Ultraviolet - visible

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1 CONTEXT

Metabolomics is a field of omics science that investigates the activity and cellular state through the study of small molecules known as metabolites [5]. It involves the quantification of low molecular weight ($<$1000 Da) biomolecules produced by cells, i.e., the quantification of the metabolome. These metabolites serve as direct signatures of biochemical activity within biological systems, providing an easier correlation with phenotypes [5].

In this sense, the characterization of the metabolome has gained more interest in several biomedical areas, being used, particularly, for diagnosing diseases, understanding disease mechanisms, identifying novel drug targets, customizing drug treatments and monitoring therapeutic outcomes [6, 7]. Nevertheless, there are other emerging applications such as biomarker discovery, food authentication and chemical characterization [7, 8].

Metabolomics studies can be divided into two categories, targeted metabolomics and untargeted metabolomics. The first attempts to describe a specific class of metabolites resulting in a higher sensitivity, caring with the quantification of the sample, whilst the other provides an identification of new compounds from a broader observation of metabolites [6].

The analytical techniques used in metabolomics are usually LC or Gas Chromatography (GC) coupled with MS, Infrared Spectroscopy (IR), Raman and NMR[6]. NMR detects molecular features by measuring an intrinsic magnetic property of the atomic nuclei that encodes information about the chemical environment, and thus its molecular structure [9]. MS is based on the principle that molecules can be ionized and sorted by their mass. Overall, NMR-based approaches are the gold standard in terms of structural eludication, biomarker detection or targeted metabolomics since reproducibility, easier sample preparation leading to low-cost studies and versatility are key advantages[9]. Despite NMR being less sensitive than MS, recent advances adding additional dimensions through different properties allows NMR techniques to push through in the field of metabolomics. Regardless of the technique used, there are lots of data generated by these platforms that need to be preprocessed and analyzed in order to infer biological meaning.

Nowadays, there are a lot of available tools for metabolomics data analysis. They differ regarding the origin of the data, mainly NMR and MS data[10], and the methods to perform certain preprocessing steps. Basic steps of data preprocessing include peak alignment, peak filtering, peak identification and metabolite identification [10]. In Metabolomics Society website http://www.metabolomicssociety.org, it is possible to access different software ranging from spectral alignment tools to data handling in metabolic fingerprinting studies. Nevertheless, the main common features of these tools are their free availability (open-source) and web-based services. Tools and software to analyze multidimensional NMR, such as, ChemoSpec2D[11], rNMR[12] and MetaboMiner[13] have novel functionalities that close the gap on sensitivity and overlapping resonances[9]. *MetaboAnalyst* is a web application to analyze metabolomics data and it is considered the most comprehensive tool with the option of structuring a code based pipeline since they have an R package (*MetaboAnalystR*)[14].

In recent work from the host group, *specmine*, a metabolomics and spectral data analysis/ mining framework, in the form of a package for the R system, has been

developed to address some of the issues involving the reproducibility of data analysis, the lack of frameworks to extract important information from metabolomics data, as well as their integration with previous knowledge [2]. This work aims to develop tools, integrated with *specmine*, that allow reading, visualization and analysis of 2D-NMR metabolomic data.

1.2 Objectives

Considering the context above, the aim of this work is to explore and extend the functionalities of the R package *specmine* developed to process and mine metabolomic data, namely addressing datasets of 2D-NMR, validating with case studies from experimental data. The functionalities addressed will be related to data reading, visualization and analysis, establishing a pipeline for this type of data.

This will encompass, in more detail, the following technological/scientific goals:

- Review the state of the art in metabolomics, with special focus on NMR techniques, and methods/tools for metabolomic data analysis;
- Extend the functionalities of *specmine* and related tools, integrating novel features for 2D-NMR data analysis, including a standard workflow to analyze this type of datasets;
- Validate the functionalities developed with real-world case studies, highlighting *specmine*'s capacity of reproducibility towards metabolomics;
- Write the thesis and related publications with the obtained results.

1.3 Structure

The present dissertation is divided into five chapters. The first one is an introductory one where the context is given alongside the motivation and the objectives aimed for this work. The second one gives insight on the state of the art of metabolomics including the principal NMR techniques used in the area, the preprocessing steps needed to follow a correct data analysis and the tools and methodologies needed to retrieve relevant information with biological meaning.

The following chapter describes the development strategy and the tools used on the present work to implement 2D-NMR spectra handling functions as well as establishing a pipeline for its analysis. They are complemented with detailed information on how they were integrated with *specmine* in order to close this gap on the package and extended tools.

The fourth chapter presents two case studies to show the applicability and flexibility of the methods developed. A brief introduction starts each case, followed by the results obtained using the tools developed as well as their interpretation in the context of each case's problem. Since both case studies' analysis are structured in the same way, a pipeline for this type of data was created in the form of a vignette.

The fifth and final chapter provides the main conclusions from this work and an insight on future work that could be done to extend *specmine*'s functionalities or to establish a new package, based on *specmine*, for 2D metabolomic data.

2

STATE OF THE ART

This chapter covers the state of the art in the field of metabolomics regarding its techniques with special focus on NMR, as well as the processes for the quantification of metabolites in NMR and the detection of biomarkers from the data generated by those technologies.

2.1 NUCLEAR MAGNETIC RESONANCE

2.1.1 1D-NMR SPECTROSCOPY

Together with Liquid Chromatography - Mass Spectrometry (LC-MS) and Gas Chromatography-Mass Spectrometry (GC-MS), NMR completes the set of the most used anaytical technologies in today's metabolomics [9].

Certain nuclei of those presented in a sample have the right magnetic properties making the basis for this technology: an odd or an even mass number, but an odd atomic number [15]. These special nuclei have a nuclear spin allowing them to be associated to a nuclear magnetic moment that interacts with the external magnetic field, β_0, applied by the NMR instrument. It is this interaction that is studied by NMR[16]. The nuclear spin is characterized by a quantum number, *I*, and this number has an influence in the produced spectra, i.e. an $I > 1/2$ producing broad lines [17]. There is also a magnetic quantum number,m_1, that indicates the number of energy states the nuclei has, which can be calculated by the formula $2I+1$[15]. The application of β_0 allows the separation of these states, and transitions between them can be accomplished through a photon in the Radio Frequency (RF) region leading ultimately to the slope of magnetization making it oscillate inducing a voltage, also called the Free Induction Decay (FID), which is then converted to a NMR spectrum using a Fourier Transformation (FT) [16].

This is conventional NMR where excitation and detection occur in different steps. There is also an NMR-spectroscopy technique in which samples are submitted to excitation and detection within the same hardware, an RF coil that is located within the NMR magnet[18]. It is called Flow NMR and it is able to enhance spectral intensity, improve signal detection and provide structural information between different moieties in a molecule[1, 19]. The most known Flow NMR technique is LC-NMR and it was used to identify tocotrienol isomers in palm-oil extracts[20] and aromatic compounds in several liquid foods[1]. Fig. 1 represents an NMR chromatogram from a wine phenolic extract which shows good compound separation and a selection of rows that represent the spectra of identified compounds in the sample.

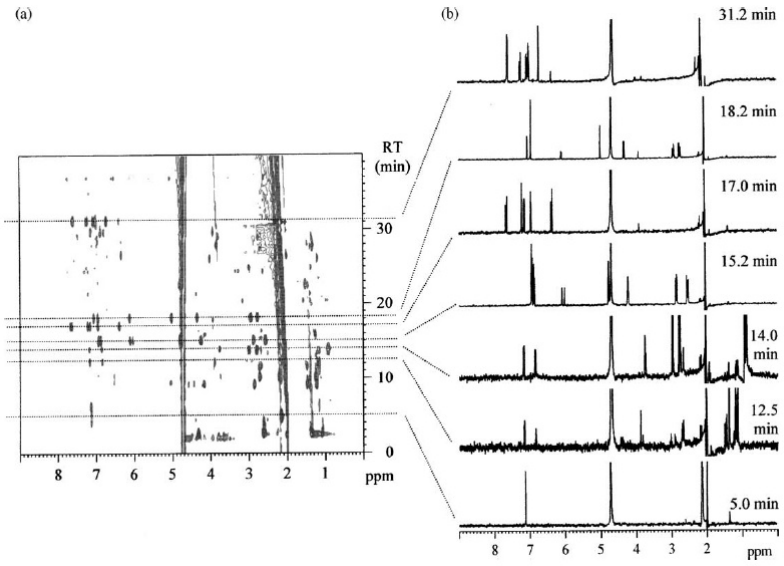


Figure 1: (a) LC-NMR/MS chromatogram resulting from the elution of the wine phenolic extract. (b) Spectra of the selected rows extracted from (a). Adapted from A. M. Gil, *et al.*[1].

NMR brings its uniqueness into metabolomics because it is nondestructive with minimum sample preparation, allowing for any biological analytes, even living ones, to be analyzed, for a long period of time, thus measuring free small molecules independently of their chemical nature [21]. Furthermore, with various nuclei that can be used in a NMR experiment, i.e. ^{1}H, ^{13}C, ^{15}N, ^{31}P [21], it is possible to assess different metabolite classes and establish a correlation between them with multidimensional NMR [9].

2.1.2 2D-NMR Spectroscopy

Multidimensional NMR, specially two-dimensional, can also treat overlapping peaks, thus performing a better peak assignment [22]. There are multiple forms of 2D-NMR that allow this task and they are summarized by the process of untangling these peaks into a second dimension, based on different physical properties that will define the bidimensional technique itself [9].

One of the oldest and fastest 2D-NMR techniques is Two-Dimensional *J*-Resolved Nuclear Magnetic Resonance Spectroscopy (2D-JRES-NMR) [23], where the second dimension is created by representing the coupling constant (*J* value) of each signal, and thus the overlapping peaks are viewed as singlets[24]. This reduces the complexity of the spectrum, allowing for a better peak assignment compared to One-Dimensional Nuclear Magnetic Resonance Spectroscopy (1D-NMR) experiments [25]. However, 2D-JRES-NMR provides insufficient structural information and the values of coupling constants cannot be used as a query to search in a database[22]. This means that metabolite identification with this technique can only be performed in an effective way with prior knowledge of the compound's chemical composition and reference spectra of the main constituents. Despite being a relatively quick technique, it is still a 2D experiment and when applied to metabolomics studies, which have a large number of samples, the acquisition time becomes a disadvantage [9]. Efforts have been made in order to reduce the acquisition time, optimizing single-scan 2D methods integrated with *J*-resolved NMR [26], but they lack sensitivity and require high metabolite concentrations [9].

Proton-decoupled projected One-Dimensional (1D) *J*-Resolved NMR is a technique that applies the base of coupling constants to 1D experiments, thus providing the speed of 1D experiments and key advantages of 2D-JRES-NMR [27]. This technique has been applied in the metabolomics field to resolve overlapped peaks of minor metabolites present in human urine and blood [28], solve superimposed peaks presented in 1D spectra that enabled peak assignment for 150 compounds in Cerebrospinal Fluid (CSF) [29] and clarify signal congestion of 1D-based metabolomics on eleven *Ilex* species making possible their chemotaxonomic classification [30].

Diffusion Ordered Spectroscopy (DOSY) [31] plots in the second dimension the diffusion coefficients associated to each NMR signal, instead of the coupling constant that 2D-JRES-NMR uses [22]. DOSY separates mixture components, without any prior treatment, by their molecular hydrodynamic properties which means that small molecules

decay faster than large molecules, having a higher diffusion coefficient [32]. However it has not been effective in isolating individual components' signals in spectral regions that have overlapping peaks, which means it is not effective on metabolomic studies due to the fact that separation in these regions is essential for peak assignment [33]. This lack of efficiency happens because an overlapped signal displays only one diffusion coefficient, which is the average of the metabolite's coefficients that take part in the intensity of the signal, making the only diffusion coefficient ambiguous [34].

To fight this issue, DOSY was coupled to other techniques, such as COSY and Heteronuclear Multiple-Quantum Correlation/Coherence Spectroscopy (HMQC), to spread the proton signals into another dimension, through a proton dimension or carbon dimension, respectively[22]. A. Sobolev, *et al.* [35] used a combination of DOSY with another 2D-NMR techniques to report for the first time the proton spectra of aqueous and organic extracts of lettuce leaves. This technique allowed to separate components of the complex patterns and assign a large number of water-soluble metabolites to their classes such as carbohydrates, organic acids and amino acids. S. Balayssac, *et al.*[36] used a three-dimensional DOSY-COSY experiment on seventeen herbal drugs for the enhancement of sexual function to investigate their formulations. This technique provided virtual separation and structural data as it allowed to observe COSY subsepctra for each component of interest [36].

Homonuclear Approaches

COSY is considered to be the simplest of all 2D-NMR experiments, as it is based on the application of two ninety degree RF pulses that generate an evolution time (t_1) and a period time (t_2) which are then converted into a spectrum by FT [9, 32]. This allows cross peaks in this generated spectrum that indicate pairs of coupled nuclei connected by through-bond highlighting homonuclear correlations ($_1$H-^{1}H) [9]. Despite standard COSY's advantages, improvements on cross-peak resolution or signal filtering have been made to enhance COSY's results interpretation [32].

One of those improvements was applying another layer of ninety degrees RF pulse after the second one, ensuring that only signals with spin-spin coupled belonging to a double or higher quantum system are detected [37]. This is called Double Quantum Filtered Correlation Spectroscopy (DQF-COSY) and allows singlet signals, such as water, to be filtered out and to phase correct diagonal and cross-peaks efficiently [32].

COSY experiments, in general, suite metabolomics research because they are simple, fast and easy to interpret. However, the type of sample to analyze needs to be considered [9]. Samples such as urine, that have many small molecules, have spectral complexity high enough so that grouping information about individual spin-spin couplings into a molecular system is very difficult [32]. COSY experiments are then suitable for known and unknown metabolite identification, whereas in full assignment of signals or metabolite's quantification studies, COSY experiments underperform [22]. A protocol for NMR-based metabolomic analysis of plants describing sample preparation (drying, weighing and extraction) followed by NMR analysis with different techniques (1D-NMR, 2D-JRES-NMR,COSY and Heteronuclear Multiple Bond Correlation (HMBC)) followed by chemometric approaches was developed, being fast (40 minutes) and suitable for analysis of both primary and secondary metabolites [38]. A COSY experiment was also performed to analyze and measure the absolute metabolite concentration in three breast cancer cell line extracts, accessing fourteen relevant metabolites highlighting differences between cell lines [39].

To gather information about the network of atoms in a molecule that is not directly coupled, an extension of COSY experiment was developed, called Total Correlation Spectroscopy (TOCSY) or homonuclear Hartmann-Hahn (HOHAHA) [9, 32, 37]. This method is based on a spin-lock field that bounds magnetizations to that RF field, enabling coupled nuclei to share the same spin system even if they are not directly coupled [32, 37]. This spin-lock is performed applying a pulse train which creates an environment that removes chemical shift differences leaving *J*-coupling interactions intact, being MLEV-17 the most widely used spin-lock sequence [37]. With TOCSY one can analyze coupling networks that may represent groups of proton signals, understanding their correlations, while having in mind that intensity does not translate the number of bonds within coupled protons[32].

Despite its advantages, Two-Dimensioanl Total Correlation Spectroscopy (2D-TOCSY) is an experiment that takes a long time to acquire, so the TOCSY principle was applied to 1D, lowering the acquisition time and easing the analysis [9]. An One-Dimensional Total Correlation Spectroscopy (1D-TOCSY) spectrum is characterized by signals from nuclei that share the same spin system as the excited signal and is specially useful on metabolite quantification with signal overlapping [9]. P. Sandusky, *et al.* [40] used Pearson correlation applied to this technique to establish an approach to metabonomic analysis of rat and human urines, achieving a more sensitive and reliable method that the

standard 1D-NMR spectra. Selective 1D-TOCSY was also used to evaluate authenticity of *Turnera diffusa* extracts and commercial botanical remedies with the intent to support this method as a quality control proof towards bioactive compounds [41].

TOCSY experiments are also very good in metabolite assignment due to existing databases that have TOCSY results incorporated and can be used to be queried against [32]. COLMAR [42] was a web server that performed the analysis of complex mixtures from NMR, with emphasis on 2D-TOCSY. The method underneath consisted in reducing the 2D-TOCSY spectrum into a cluster of non-redundant sections through covariance, that represented individual metabolites in the mixture, which would be queried against NMR spectral databases. A new version of this tool emerged, called COLMARm [43], that allows co-analysis of up to three 2D-NMR spectra for metabolite identification purposes, with the advantage of validating manually possible metabolites by superimposing generated cross-peaks with experimental ones.

COSY and TOCSY perform homonuclear correlation spectroscopy, but the study of bond correlations can be applied to different nuclei. There are different heteronuclear correlation spectroscopy techniques that differ on how magnetization is transferred between nuclei and in the number of bonds between them.

Heteronuclear Approaches

Heteronuclear Single Quantum Correlation/Coherence Spectroscopy (HSQC) is based on the principle of transferring a nuclear spin polarization from a more sensitive nucleus (^{1}H) to one that has lower sensitivity (mainly ^{13}C and ^{15}N due to their biological importance) and then transferred back for detection [9, 32]. This is called Insensitive Nuclei Enhanced by Polarization Transfer (INEPT) and correlates chemical shifts between the directly and indirectly measured dimensions of coupled nuclei that are further plotted to produce a 2D spectrum [9]. This ability to produce one cross-peak for each chemical pair singularly bonded resolves and assigns metabolite signals arising from complex biofluid mixtures, as well as protein structures [44–46].

In terms of metabolite quantification, a new and fast strategy of 2D ^{1}H–^{13}C-HSQC, named Fast Metabolite Quantification (FMQ), has been developed quantifying, in twelve minutes, fourty metabolites from biological samples, being optimal for NMR metabolomic studies with samples around fifty milligrams of weight [47]. Together with COSY, HSQC can also complement the structural elucidation of metabolites by understanding the one-bond linkage between hydrogen and carbon in molecular specific studies [32].

HMQC is very similar to HSQC as its results highlight the same type of correlations [9]. The difference is in the evolution period when the pulse sequence is applied: in HMQC double and zero quantum coherences are maintained which means both coupled nuclei experience magnetization effects involving both spins, while in HSQC only single quantum coherences are maintained [32, 37]. HMQC has the advantage of using shorter pulse sequences which means it has fewer errors caused by the quality of the magnetic field. However, nowadays there are high standards regarding magnetic fields and pulses resulting on HSQC to be the preferred choice [32].

In the metabolomics field, HMQC has not been receiving much support because it produces broader peaks than HSQC thus presenting more peak overlaps which leads to worse resolution [9]. On the other hand, in proteomics, a technique was developed, SOFAST-HMQC [48], that can record correlation spectra of proteins of different sizes within seconds and higher sensitivity compared to standard ^{1}H–^{15}N correlation experiments [49]. HMQC and HSQC can also be used together, showing its usefulness when trying to assign NMR signals to metabolites because this enables new compound identification if their concentration levels are within spectrometer's boundaries [22]. This was proven by Y. S. Liang, *et al.* [50] when using both techniques, finding a new phenylpropanoid in methyl jasmonate treated *Brassica rapa* leaves.

Another technique of heteronuclear 2D-NMR, similar to HSQC and HMQC, is HMBC, which accounts for coupled nuclei that have two or more chemical bonds [9, 32]. It only selects small *J*-values, filtering out single bonded correlations, such as C–H, highlighting quaternary and carbonyl carbons [9, 37]. This advantage means that sometimes HMBC is the only technique that is able to establish the connection between protonated carbon structures separated by these quaternary carbons, such as cis- and trans-aconitic acid molecules [51]. Usually, this technique works best with other heteronuclear approaches complementing the chemical structure determination of unknown metabolites since there is the acknowledgement of the compound's backbone and side chains [32]. In a metabolomics study, if there is a large number of samples, heteronuclear approaches are not the best option because they have long acquisition times and their use is optimal for a subset of samples if further analysis is needed [32]. P. Bernini *et al.* proved ^{1}H–^{13}C-HMBC utility by combining it with other 2D-NMR techniques to improve metabolite identification in twelve urine samples [52]. HMBC was also used in a combination with 2D selective TOCSY to obtain NMR data for each impurity in active pharmaceutical

ingredient samples to distinguish responses, without prior chromatographic separation [53].

Multidimensional NMR is also a partial solution to fight the principal disadvantage of NMR, sensitivity, through the addition of chemoselective isotope labels, reacting with specific groups [54]. A larger magnetic field brings higher costs in an NMR experiment but can enhance the sensitivity and resolution of the final spectra. When employing a multidimensional experiment, the key is time because they require longer acquisition times due to sampling along with indirect frequency dimensions [54].

Recent advances in NMR metabolomics

Another way to fight the lack of sensitivity is through Dynamic Nuclear Polarization (DNP) methods in which the sample to be analyzed is previously frozen, submitted to microwave radiation and then melted to be transferred to an NMR spectrometry[55]. In the radiation step, the free-radicals presented in the microwaves induce a transfer of polarization from electrons to nuclei that is caused by the temporary hyperpolarization on spin-active nuclei, which results ultimately in over 10.000-fold increased sensitivity [56]. Another aspect to be improved in NMR metabolomics is the detection of low-concentrated metabolites, that are being acknowledged to be diagnostic biomarkers as important as high-abundance metabolites and yet only a few hundred have been detected in human fluids metabolomes (such as urine, CSF and serum) [9].

Besides DNP, which is a hyperpolarization method, fast NMR methods are also a new approach to NMR in the field of metabolomics [9]. As said before, multidimensional approaches usually take longer acquisition times, which is the disadvantage of 2D-NMR experiments, for example. Nonlinear sampling (NLS) is one of the methods that can ease the acquisition time by reconstruction of a complete spectrum from only a small number of optimally selected experimental data points, using nontraditional schemes, such as nonlinear FT, to process signals [57]. Another fast approach to 2D methods is Hadamard spectroscopy, in which direct irradiation at signal-bearing sites is done based on Hadamard matrices, encoded in multiplex excitation schemes that generate signals that need to be decoded using the same scheme [58]. Ultrafast 2D-NMR is based on the principle that a sample can be divided into slices, where different evolution periods are experimented simultaneously by the active nuclei on those slices in the same scan, being the signals processed by specific software capable of creating the 2D-NMR spectrum [59]. According to Emwas *et al.* [9], the majority of these methods are difficult

to implement in high-throughput metabolomic studies because they are not particularly quantitative, being ultrafast 2D-NMR the one that, together with hyperpolarized NMR techniques, holds the most promise for NMR-based metabolomics.

Recent food-related studies have been done using NMR-based approaches to characterize samples, evaluating the metabolic profiles or establishing new biomarkers. Blakebrough-Hall *et al.* [60] identified phenylalanine, lactate, hydroxybutyrate, tyrosine, citrate and leucine as high-importance metabolites to classify blood samples from animals that have or do not have Bovine Respiratory Disease. To prevent adulteration and to assure quality of sesame seeds, the development of a discrimination model and potential biomarker investigation for differentiation of the geographical origin was performed using NMR-based metabolic profiling [61]. Yang *et al.* [62] investigated the effect of wooden breast myodegeneration on the metabolite profile of chicken meat and found that WB-affected samples showed higher specific metabolites (such as leucine, valine, taurine and glutamate) and lower levels of other metabolites (such as histidine, creatine, acetate and serine), using ^{1}H NMR. Very recently, NMR-based metabolomics approaches have been used in areas, besides food-related ones, such as hazardous materials [63], environmental pollution [64] and biomedicine and pharmacotherapy [65].

2.2 Data Preprocessing

Preprocessing techniques are tasks performed to datasets to prepare them for knowledge extraction. These tasks intend to reduce data size, remove outliers, treat missing values, extract features, normalize data and transform them[66]. In order to achieve such objectives operations that clean, integrate, transform and reduce data have to be done. This supports that data preprocessing is a critical step in any metabolomics pipeline because raw data is usually complex and needs to be eased for subsequent data analysis and interpretation [67].

2.2.1 Missing Data and Outliers

The process to clean data is the first step in preprocessing, which normally detects incomplete records with missing values, atypical and inconsistent data points [66]. Missing data can occur mainly through three mechanisms [68] that characterize their produced missing values:

- missing completely at random - missing values are randomly distributed in a data matrix and do not depend on the known values [69].
- missing at random - missing values have an association due to possible variable dependency to other known variables (X) but not to response variables (Y) [70].
- not missing at random - there is a pattern and dependency within missing values in a variable. The dependency can be related to values on other variable and on the value of the missing data itself. There is no consensual method to treat these missing values[71].

Since data collection is expensive there are two categories of methods to handle missing data: deletion and imputation methods [72]. However, based on the work of Kapil[71], there are different categories of methods to handle missing values, having the imputation ones a lot of variations, such as parameter estimation approaches or machine learning models.

For data deletion there are two methods commonly used: listwise deletion and pairwise deletion[72]. Listwise deletion discards observations on one or multiple variables that have missing values. On the other hand, pairwise deletion only excludes missing data from variables when they are used in a statistical procedure. However, this method has limited application due to similar results when compared with listwise deletion and it is not supported in many statistical packages [72]. There is also an option to delete based on the extent of the missing values which means that if a feature or an observation have a high share (e.g. over 5 per cent) of missing values they can be discarded [71].

As for data imputation, there are a lot of methods to treat missing values. Mean imputation estimates the mean and can be done on numerical data to replace the missing values [71]. This method excels when dealing with continuous variables that are not related to other independent variables because the estimation performed does not affect other cases not leading to efficiency loss [73]. It is also possible to replace the missing value by estimating the median or other statistical measures. There is also the K-Nearest Neighbour (KNN) Imputation method to estimate and substitute missing data, by taking into account K number of observations that share similarity to the missing value considering other variables [71, 74]. KNN imputation can treat continuous and categorical missing data by considering the most frequent and the mean of the values in the k nearest neighbour, respectively[71]. Also, it takes into consideration the data structure and does not need to create a predictive model [75]. There are also Regression

methods, using predictive models, that take the observed data as input variables to create the model and a replacement for the missing value is predicted as a response variable [76].

The parameter estimation methods includes expectation maximization, maximum likelihood and Bayesian estimation and are more complex and considered to be more accurate than the previous ones [71]. Expectation maximization has two stages wherein the first it begins by estimating missing values based on observed data and the parameters underlined whereas in the second the values estimated are imputed to maximize the likelihood function to obtain new parameters for the next iteration [71]. The other two are model specific methods that try to optimize model parameters through complete data likelihood. All methods within parameter estimation are best used when the missing data is missing at random [71].

Outliers are data values that behave differently in comparison to other data values and are intended to be deleted for further analysis because on their own they can influence statistical metrics and univariate analysis. Outliers can emerge from different sources such as human error, instrument error or faults in systems and its detection has different methodologies based on statistics, neural networks, machine learning (with decision trees) and even hybrid systems [77].

2.2.2 Spectral Preprocessing

Spectral preprocessing is a key component in metabolomics because the instruments from MS and NMR techniques generate background noise, peaks that are not related to the biological sample and systematic variations to data [67]. To fix these issues, a set of preprocessing techniques can be applied to this type of data such as noise filtering, peak detection, deconvolution, deisotoping, alignment and normalization [67]. In NMR metabolomics, it is necessary to retain that the spectrum has to be transformed from FID into frequency spectra leading to preprocessing steps of zero-filling, apodization, Fourier transformation and phase correction before the techniques previously stated [78].

When analysing a biological sample in MS, there is a lot of background signaling coming from two sources, electronic and chemical. Electronic noise is a natural characteristic form the mass spectrometer and it is constant throughout the experience, whereas chemical noise originates primarily from the entire chromatographic system being the column bleeding or the presence of contaminants in the mobile phase the

main causes [79]. In a NMR experiment, the electronic component can also take part in producing noise, but the chemical noise has the most influence in the spectra itself. Sample parameters such as pH, salt type, salt concentrations, presence of paramagnetic ions, solvent chosen and dissolved oxygen content are just examples of causes to create noise and to induce variations on the spectra [80]. So, the process to filter the noise aims to eliminate spectral information non-related to the biological sample analyzed to clear the signal coming from the compounds presented there [67].

Peak detection is performed with specific algorithms that usually analyze each sample spectrum individually [81]. This process is comprised by two analytical steps: spectra smoothing and the actual detection [82, 83]. Smoothing a spectra intends to modify individual data points that are higher and/or lower than the immediately adjacent points, reducing and/or increasing them, respectively [84]. This aims to reduce the high frequency noise, as well as maintaining the ones with low frequency, thus improving the Signal-to-Noise Ratio (SNR) always depending on the frequency distribution of the noise [84].

These steps can be done by different filters such as Gaussian, Savitzky-Golay and moving average [83]. Despite higher computation time, Wavelet transform-based filters have higher performance because they can handle the usual unequal peak widths in metabolomic spectra [82]. SNR compares the level of a desired signal to the level of background noise [85] and has an important role in peak detection. These algorithms use local SNR, depending on the peak amplitude relative to the surrounding electronic and chemical noise levels as a reference and if there is a local maximum, above a certain threshold [86]. There is an opportunity to apply the threshold to other parameters (i.e., intensity, area of each peak) [83] and, in studies with a large number of samples, a frequency filter can be applied to peaks, being selected peaks that appear in a certain percentage of samples [82].

Deconvolution is a signal processing technique that allows to decompose a spectral region with multiple peaks by the relative area corresponding to each individual peak[82]. To be able to decompose such region, prior knowledge of the compounds in the mixture is needed (from a template library), so this technique is useful in targeted metabolomics [82]. For NMR data, the methods available are based on Bayesian model selection being *Bayesian Automated Metabolite Analyser for Nuclear Magnetic Resonance spectra (BATMAN)* [87] one of the open-source, user-friendly interfaces most frequently used and with very similar performance compared to NMR Suite (Chenomx Inc., Edmonton, AB,

Canada [88]), a proprietary software package and gold standard for NMR metabolomics [82].

2D-NMR provides the means to treat signal overlapping, which means an accurate identification and quantification of metabolites is possible using this technique. However, there is not a software that can provide a framework for the complete analysis of these spectra, being flexible for different 2D techniques and providing means to ease chemometric and quantifying methods. This task is particularly hard because each 2D experiment has specific data projected to the second dimension that is influenced by magnetic resonance characteristics. In addition to this, 2D-NMR has its own disadvantages which are being overcome with the recent advances in the area [9].

The process of spectral alignment allows for the correct position match of peaks, that belong to the same metabolic feature, in a multiple spectra study [82]. This is specially useful for metabolite identification and biomarker selection, because we cannot have variability on the position of peaks when they are associated within multiple spectra and to samples, leading to false biological meaning. The peaks suffer this change in position due to non-linear shifts and depending on if the study is NMR-based or MS-based, the causes to those shifts are different [82]. Non-linear shifts in NMR-based studies occur in the ppm axis and they are caused by ionic strength, pH or protein content [88], whereas in MS-based they occur in the retention time axis resulting from changes in the stationary phase of the chromatographic column [89].

Spectral alignment algorithms can be done before (spectral alignment methods) or after (peak-based alignment methods) the peak detection phase. The former use segments for the alignment, whereas the latter ones use peak coordinates [82]. Spectral alignment methods are classified into warping and segmentation methods, where the first applies a non-linear transformation to the respective axis (ppm for NMR and retention time to MS) and the second one applies a constant shift to all spectral points [82].

Warping methods aim for a correlation maximization between spectral segments by stretching or shrinking these segments, being correlation optimized warping and dynamic time warping the most commonly used and based in dynamic programming [90]. Segmentation methods also aims for correlation maximization between spectral segments but their segments are originated from spectra splits or they just consider the overall spectra, being the Icoshift algorithm [91] one of the commonly used and usually combined with automatic segmentation methods [92].

Peak-based methods are implemented in the XCMS software [93] and the algorithm computes, through a kernel density estimator, the retention time boundaries of the observed peaks to identify the ones that share the same metabolomic feature on different samples [82].

Data normalization is generally required to highlight biological differences in complex biofluids where there is a high entropy of metabolite concentration, in order to remove systematic biases [82]. It is defined as a row operation applied to each sample and allows all samples to be comparable to each other [94]. Regarding metabolomics data, there are numerous methods to normalize data, that can be grouped into two classes [95, 96]. The first class intends to reduce heteroscedasticity among metabolites and the second intends to remove sample-to-sample variations [97]. There are also normalization methods based on internal standards and/or quality control metabolites that are capable to remove unwanted variations [98].

There is also a method of normalization that sets each observation (spectrum) to have unit total intensity by expressing each data point as a fraction of the total spectral integral, also known as normalization to a constant sum [94]. This normalization method is specially useful to approximate the relative concentration of species in a series of spectra with highly similar internal peak ratios and differences in total intensity [94].

2.3 Metabolite Quantification with Nuclear Magnetic Resonance

Metabolite quantification, in general, is tightly related to metabolite identification, since both concepts are usually mentioned together and in some extent they both share similar methodologies, such as comparing experimental data to a reference [9, 99, 100]. Metabolite identification and quantification is possible in NMR due to the underneath principle that the observed spectrum is the sum of individual spectra for each of the metabolites presented in the sample analyzed [100, 101]. Each metabolite is then characterized as a set of peaks which are defined by three parameters (height, center and width) that are constant across different spectra of the same frequency, where concentration scales linearly with height [102].

There are different ways to perform metabolite quantification in NMR metabolomics, while their basis is usually the same which is comparing the NMR experimental spectrum to a spectral reference library, also called as targeted profiling [88]. The process involves

matching and fitting the reference peaks to the sample peaks (should be in the same or similar acquisition standards) which enables identification and quantification at the same time [100]. This methodology requires precise sample control, curve-fitting software and databases with pure metabolite information (including pH values and spectrometer frequencies used in data collection) [100].

Targeted profiling means that absolute metabolite quantification happens in targeted metabolomics, which aims to analyze a set of known metabolites, whereas in untargeted metabolomics the quantification procedure is relative (since there may not be a reference to compare) and aims to provide hypothesis for future investigations [103]. To optimize quantification processes efforts have been made to fully automate curve-fitting software and build comprehensive and complete databases [100].

Accordingly to C. Zheng *et al.* [99] there are three broad categories that classify the means to obtain metabolite concentrations: binning, curve-fitting without a database and curve-fitting with a database.

Binning is the process of partitioning the spectrum into individual equally or variably-sized contiguous segments, called bins, allowing for specific regions to be omitted thus reducing NMR spectra complexity and providing descriptors for further analysis [104]. These bins have their intensity (total Area Under the Curve (AUC)) measured and averaged to isolate distinct resonance signals and to be integrated for metabolite quantification [99, 100, 105, 106]. However, integrated intensity values of bins do not represent an individual metabolite due to overlapping peaks, chemical shift variations across samples, signal contamination and even fluctuations across bin boundaries that may lead to incorrect intensity values on adjacent bins [99, 105, 106]. These characteristics along with the lack of biological interpretability and improvements on software supported the development of deconvolution techniques to replace binning[9, 105].

Curve-fitting without a database attempts to recover non-negative source signals from individual metabolites directly from the spectra through algorithms [99]. To perform an accurate curve-fitting process, the spectra has to be phase and baseline corrected since water signals and metabolites itself can create phase distortions or non-zero baselines [100].

As said before, NMR spectra can be viewed as a linear combination, with unknown proportions, of individual spectra from each metabolite. Thus, specific algorithms (usually Non-negative Matrix Factorization (NMF) methods) are able to untangle this combination

and provide each metabolite's spectrum [99]. Bayesian spectral decomposition is a probabilistic approach that uses Markov Chain Monte Carlo (MCMC) to sample from the possible solutions to obtain a small number of basis spectra with their localized amplitudes [107]. It was successfully used to perform deconvolution in ^{1}H NMR spectra of urine rat samples, to understand metabolites involved in liver toxicity after an administration of the hepatotoxin hydrazine [108]. With time, new methods incorporating Bayesian approaches to identify and quantify metabolites have emerged, that will be presented later since they are based on curve-fitting with a database. Alternating least squares [109] was another NMF algorithm aimed at decomposing the data matrix as a product of a scores (concentration) matrix and a loadings (component spectra) matrix, estimating initial values for those matrices and executing a while loop until convergence. P. Soininen *et al.* [110] developed an integration strategy based on constrained total-line-shape fitting, using an optimization problem by minimizing the residual sum of the squares between the line-shape function [111] (sum of Lorentzian, dispersion, Gaussian and dispersive Gaussian functions) and the observed spectrum.

Curve-fitting with a database is the method that was above mentioned and enabled simultaneously metabolite identification and quantification. It interprets, usually ^{1}H NMR spectra, as linear combinations of reference peaks that are available in a library [100]. This method is the one that provides both interpretability and accuracy on obtaining real metabolite concentrations and for a long time they were performed manually which was inefficient [99].

Speeding up the process of deconvoluting an NMR spectrum and fitting reference spectra to it has been an active research area in NMR-based metabolomics [112], so numerous tools have been developed, the first one appearing in 2001 [113]. NMR Suite (Chenomx Inc., Edmonton, AB, Canada [88]) is a well-known commercial software that allows users to estimate metabolite concentrations with a comprehensive database, through manual deconvolution. Its accuracy is dependent on the correct peak assignment by spectroscopists which requires expertise and may lead to errors [106]. AMIX (Bruker) is another commercial spectral deconvolution software that is making advances in semi-automatic approaches for high-throughput analysis [114]. Recently, Bruker presented *WineScreener* and *JuiceScreener*, two commercial software tools to perform automated deconvolution on NMR spectra of wines, juices and honey. However, they require specifications on the NMR spectrometer which makes them expensive [114]. As any

given commercial software, its cost implies a great investment so research groups started to develop their own deconvolution and quantification approaches [100].

2.3.1 Tools for 1D-NMR spectra analysis in metabolomics

a (BQuant) [99] is an R package that identifies and quantifies metabolites in a fully automated way, based on a Bayesian model selection and a database of candidate metabolites. It models the observed spectra as a combination of reference signals and candidate metabolites are viewed as variables, which are then submitted to a stochastic search to find highly likely matches [99]. Despite its good reported performance compared to the existing automated approaches by the time it was developed, *BQuant* has not been widely applied to NMR metabolomics, specially to cellular extracts and culture media [115]. This may be justified due to the prior requisites before identification and quantification, such as pre-aligned and partitioned into bins spectra, and due to peak shifts that are also not incorporated that influence overlapped peaks and quantification [87].

One of the well-known and open-access programs, also based on Bayesian approach, aimed specifically for an automated quantification of metabolites, was *BATMAN* [87]. It only focuses on ^{1}H NMR spectra and is available as an R package. *BATMAN* models the NMR spectrum as a two-component joint equation, where metabolite peaks are interpreted as catalogued if their characteristic patterns are known to the user and uncatalogued otherwise [87]. The catalogued peaks are modelled as a sum of all the template metabolites' resonance information (chemical shifts, *J*-couplings and intensity ratios) which is obtained from the Human Metabolome Database (HMDB) (having a freely available file with around 2500 peak patterns, corresponding to 600 metabolites [116]) and used to construct a spectrum that fits the data [117]. The uncatalogued peaks are modelled as a linear combination of wavelet functions that can be characterized by a probability density model whose distribution is further modelled by a truncated Gaussian that connects global precision to each wavelet [87]. A MCMC algorithm samples from the joint posterior distribution of the model parameters, with block updates and adaptation to peak shifts that are described in Astle *et al.* [106].

An advantage of *BATMAN* is its flexibility and adaptability since the user can set up prior spectra information deciding which resonance to fit and allowing for different spectra from different spectrometers to be integrated [105, 115]. *BATMAN* was successfully

applied to analyze 86 single-gene transposon insertion mutant strains of *Pseudomonas aeruginosa*, quantifying 25 extracellular metabolites [118]. For complex 1D-NMR spectra, there is also a protocol published on Nature Protocols using *BATMAN* for deconvolution and quantification purposes, that helps users to fit the software usage to their necessities and experiment specifications [117]. However, *BATMAN*'s fitting algorithm is quite slow (hours to complete) and limited to mixtures with 20-25 compounds [114].

To address the issue of analyzing complex mixtures (mainly biofluids), a software called *BAYESIL*[102] was developed and implemented through a web server, available at http://bayesil.ca/. It performs fully automated spectral processing and profiling, allowing users to quickly and accurately analyze complex mixtures (under two minutes and mixtures with 60 compounds), measuring reliable metabolite concentrations [114]. This software enhances metabolite quantification reproducibility and consistency by performing zero-filling, Fourier and Hilbert transformation, phasing and baseline correction, smoothing, chemical shift referencing and reference deconvolution without manual intervention [102].

BAYESIL's key innovation is their efficiency of solving the problem of fitting a set of reference compounds to the observed spectrum, finding the correct combination of concentrations and chemical shifts [102]. It divides the spectrum and the loss function into interrelated regions and functions that serve as convergence point for probabilistic graphical models (factor-graphs). This is possible due to *BAYESIL*'s *Gibbs* distribution [119] approach to evaluate shift variables and concentration assignments to those regions as probabilities, transforming the loss function into a probability distribution. Using a sequential Monte Carlo inference method [120], the distributions are narrowed in each iteration and when convergence is achieved, the distribution over each concentration and shift variable approximates the most probable assignment, relative compound quantification [102]. *BAYESIL*'s spectral library was obtained through 1D ^{1}H NMR reference spectra collection for each of the compounds using pure compound information from HMDB[102]. There are also specific sub-libraries for serum and CSF, thus improving spectral fitting performance.

There are other automated quantification tools based also on probabilistic models such as *ASICS* [121], *rDolphin* [122] and *AQuA* [123] which complement quantitative profiling of aqueous metabolites. Despite this great effort in automated quantification in 1D-NMR, this technique lacks spectral resolution limiting its accuracy, metabolome coverage and the ability to deal with spectral overlap [112].

In order to improve spectral resolution and resolve overlapping issues, 2D-NMR has been applied to quantify metabolites and is also a growing research area [124]. However, one has to take into account that in 2D-NMR peak volumes do not depend only on metabolite concentration since there are other resonance properties such as *J*-couplings and relaxation times that influence how signal and concentration is related [125]. Despite this, it was proven that for a specific biological sample, honey, a selective TOCSY-based quantification approach correlated peak intensities directly with concentration in amino acids samples, when there is not a high degree of variation in its composition [126].

2.3.2 2D-NMR for Metabolite Quantification

One solution and the most used is one that was already mentioned, FMQ. It involves the preparation of calibration standards of targeted previously identified metabolites, which allow the construction of a reference curve that relates signal intensity and concentration for each analyte [47]. The metabolites with unknown concentrations are then submitted to these standards, integrating the signals of interest, allowing the concentration of the targeted metabolites to be calculated from the regression of the calibration curves [47]. Despite this approach, instrument response can influence the sample calibration and complexity of biological samples are not always accounted for [125]. Using three calibration curves through AMIX software, R. Wedeking *et al.* were able to quantify the metabolic changes of roots and leaves of drought-stressed sugar beets throughout time [127]. This approach was also applied to ^{1}H NMR to quantify biodiesel and vegetable oil in diesel-biodiesel blends [128]. It was able to correctly quantify biodiesel content even in the presence of vegetable oil, thus providing an efficient alternative to check blends' adulteration.

One alternative that takes into account matrix effects is replacing external calibration by repeatedly analyzing the samples of interest with a mixture of metabolites in known concentrations [125]. This is a standard addition procedure, i.e., sample spiking, where a curve is fitted for each metabolite by a linear regression that relates the 2D spectrum peak volume with concentration [129]. This was used to absolutely quantify 15 metabolites that revealed significant differences between breast cancer cell lines from a 2D experiment [129]. However, for high throughput analysis, this approach is not optimal since it requires practical work on spiking each sample of interest with this mixture and getting the absolute metabolite quantification [125].

Another approach that does not require an identification of metabolites prior to quantification is based on getting the metabolite concentration information directly from the peaks [125]. This is possible by acquiring multiple ^{1}H–^{13}C HSQC spectra, repeating the pulse sequence block between excitation and acquisition, which will remove peak dependencies on *J*-coupling factors and relaxation times, a work developed by K. Hu *et al.* [130]. It is named time-zero ^{13}C HSQC and can provide absolute concentrations, if an internal standard of known concentration is added to sample; otherwise, relative concentrations are determined from cross peak intensities. This pulse sequence was applied by S. Halouska *et al.* [131] to study how D-Cycloserine affects *Mycobacterium tuberculosis* resistant strains metabolome, with a carbon-13 source supplement. This strategy is limited to this 2D technique and it takes a longer acquisition time due to the multiple pulse sequence block which lowers overall sensitivity [125].

Adapting the same technique, C. Mauve, *et al.* [132] developed a version of ^{1}H–^{13}C HSQC that aims to quantify low concentration metabolites from complex mixtures with a special focus on vegetal extracts. QUantItative, Perfected and pUre shifted Heteronuclear Single Quantum Correlation/Coherence Spectroscopy (QUIPU-HSQC) is based on a set of changes to the normal HSQC pulse sequence to ensure that signals are not modulated, thus providing an even proportion between peak volume and concentration [132]. Using this approach and an internal standard, the authors quantified key low concentrated metabolites involved in cellular death and Krebs cycle embedded in complex vegetal matrices. This was possible due to pure-shift elements in the acquisition phase that decoupled *J*-coupled protons, increasing sensitivity and efficiency [125]. A framework attempting to speed up QUIPU-HSQC was developed that was able to reduce the acquisition time through spectral aliasing, nonuniform sampling and variable repetition time while maintaining its resolution [133]. Such approach was applied to breast-cell extracts to prove its potential being sensitive to metabolites with submillimolar concentrations.

In terms of software tools aimed for 2D quantification, *MetaboQuant*[134] emerged as a tool that combined individual peak calibration and outlier detection for both 1D ^{1}H and 2D ^{1}H–^{13}C HSQC NMR spectra. This software tool requires the spectral data to be preprocessed with separate software for peak fitting and integration. It was developed under Matlab environment and is accessed via a graphical user interface. According to the authors, the novel feature implemented on their NMR quantification is a reliability check based on the proportion between visible and non-visible signals of a compound [134].

Other key components of the software is having a set of calibration factors incorporated and user defined threshold to ensure high likelihood of correct quantification.

Dolphin [135] is another software package that despite not being aimed to quantify 2D spectra, uses 2D-JRES-NMR spectra to find and integrate additional information to improve their line-shape fitting algorithm. This algorithm allows an automated quantification of a fixed target set of metabolites, achieving the best results when there is a referencing between the 1D and 2D-JRES-NMR spectra, usually alpha-D-glucose doublet [136]. The presence of each target metabolite is assessed by a subset of its spectral pattern in the 2D-JRES-NMR spectra, information that is obtained from public domain databases such as HMDB or BioMagResBank (BMRB) [137] or commercial packages such as NMR Suite and AMIX [135].

2D-NMR provides the means to treat signal overlapping, which means an accurate identification and quantification of metabolites is possible using this technique. However, there is not a software that can provide a framework for the complete analysis of these spectra, being flexible for different 2D techniques and providing means to ease chemometric and quantifying methods. This task is particularly hard because each 2D experiment has specific data projected to the second dimension that is influenced by magnetic resonance characteristics. In addition to this, 2D-NMR has its own disadvantages which are being overcome with the recent advances in the area [9].

The tools and softwares to quantify metabolites in NMR, described in this section, are summarized on Table 1. Some of the techniques applied to 2D-NMR were left out of the table because they were considered modifications to specific techniques.

2.4 Biomarker discovery

Metabolomics is connected to clinical research because metabolites hold high potential as biomarkers for diseases through early diagnostics or follow-up prognosis [138]. This potential is due to the assumption that diseases influence biochemical pathways in a cell that change its metabolic fingerprint to a new metabolic state that is disease-specific [139]. One other field is plant metabolomics where metabolites can be biomarkers of environmental or nutritional perturbations [140].

A metabolite is recognized as a biomarker when it is quantitatively capable of being detected, has high sensitivity and specificity, thus representing a measurable indicator of a physiological state [141]. This only happens when the reproducibility and validity

Table 1: Metabolite quantification tools.

Tool	Type	Spectral Data	Availability
BQuant	R package	1D-NMR	Free at https://www.stat.purdue.edu/~ovitek/BQuant-Web/
BATMAN	R package	1D-NMR	Free at https://cran.r-project.org/web/\packages/batman/index.html
BAYESIL	Web tool	1D-NMR	Free at http://www.bayesil.ca
ASICS	R package	Complex 1D-NMR	Free at https://www.bioconductor.org/packages/release/bioc/html/ASICS.html
rDolphin	GUI R package	1D-NMR	Free at http://github.com/danielcanueto/rDolphin
AQuA	MATLAB script	1D-NMR	Free through their article, here
MetaboQuant	Executable file and MATLAB script	1D-NMR and ^{1}H–^{13}C HSQC	Free at https://www.uni-regensburg.de/medicine/statistical-bioinformatics/software/software-from-gronwald-group/metaboquant/index.html
Dolphin	MATLAB software package	1D-NMR	Only by request. Link to article here
Chenomx NMR Suite	Software	1D-NMR	Commercial with a evaluation version at https://www.chenomx.com/products/
Bruker AMIX	Software	1D-NMR and HSQC	Commercial, available here

of the primary results are confirmed, through new independent samples, and then compared against traditional methods [142, 143]. However, metabolites' measurements rely on analytical methods [144] (mainly NMR and MS-based techniques) and different independent labs use different methods [142]. This translates into a difficulty when a biomarker study is published and needs to be validated and the protocols developed followed [145].

In metabolomics, we can either perform a targeted analysis or untargeted analysis based on the study we want to perform, and if there is or not an underlying hypothesis, respectively [146]. Usually, for biomarker research, the choice is untargeted metabolomics, in a scenario of hypothesis-generation research to extract biological meaning of a dataset with hundreds or thousands of metabolites [146, 147]. Due to the complexity and volume of the data generated in metabolomics, pattern recognition methods together with multivariate statistical approaches, allow for metabolic signature identification that can differentiate and discriminate certain cell's physiological patterns associated to pathophysiological frameworks[51].

Independently of the analytical method used in the study, unsupervised and supervised methods of multivariate statistical analysis are employed to reduce data dimensionality, find trends and search for discriminating features to give a clear interpretation of metabolome's alterations [148–151]. Univariate analyses may also be employed to find differential expressed metabolites. However, since metabolomics datasets have a lot of variables, it is fundamental to adjust the p-value [152].

The difference between unsupervised and supervised methods rely on the prior knowledge required. Unsupervised methods do not require prior knowledge and are used for data analysis, finding clusters of data classification and give an unbiased view of the data [148, 151]. On the other hand, supervised methods require prior knowledge about sample class and are used for optimal sample class segregation, generate clusters of patterns and new data prediction [139, 148]. To both methods it is highly recommendable to validate findings either by cross-validation or a second sets of samples to achieve higher truthfulness in the results obtained [148].

2.4.1 Unsupervised methods

Principal Component Analysis (PCA) is the most commonly used unsupervised method in metabolomics [153] and it is used to identify patterns and trends in large volumes of

data [149]. In order to do this, multidimensional data is reduced by creating new variables, called Principal Components (PCs), which result from linear combinations of the original variables and are orthogonal to each other [151]. These new variables are intended to explain the maximum amount of variance, not accounted for previous PCs [152].

When we perform PCA, two new matrices are created, scores and loadings, and they enable the interpretation of how spectra are related to each other and which variables contributed the most to principal component's variance, respectively [148, 151]. Usually, scores are plotted to evaluate the scatter of the samples and if they are clustered together they have similar metabolomics compositions, otherwise they are characterized by different properties. Clustering in these circumstances should be evaluated through loadings since potential biomarkers are normally characterized for having high variance. However, this does not guarantee a relevant biomarker and further statistical validation is required [148].

Clustering methods, such as Hierarchical Clustering (HC) or K-means, intend to subgroup heterogeneous data so that each subgroup has high homogeneity, through a measure of similarity [154, 155]. HC requires the user to choose the distance metric by which similarity will be calculated and the function that will link clusters, having both choices influence on the dendogram structure [151]. On the other hand, K-means only needs to be given an integer number, k, that defines the number of clusters that will be formed [155]. Clustering methods should be combined with PCA to provide an unbiased mechanism of analyzing groups of metabolites that share similar metabolic content, thus providing potential sets of biomarkers [148].

Regarding NMR, Statistical Total Correlation Spectroscopy (STOCSY)[156] is a method applied to complex 1D-NMR spectra to retrieve more information regarding spectral peak intensities [148]. It is based on a correlation matrix between all intensities of the NMR spectra and this matrix can then be plotted achieving a graphical representation similar to TOCSY, a 2D correlation NMR experiment on one sample [151]. STOCSY has numerous applications, such as, drug metabolite identification in human urine samples [156], cross-experiment analysis (NMR and MS simultaneously) [157], clustering combination for higher accuracy between peaks from the same molecule [158] and supervised technique combination for relevant metabolite signature linkage [156].

Unsupervised methods do not actually give us information on which metabolites can differentiate classes, but rather an overview of potential groups of metabolites that share similar properties and are somehow correlated, thus the need to combine with supervised

methodologies. Supervised methods can take advantage of these potential groups of metabolites to narrow the variance coming from other sources thus improving their performance and easing computational effort[82].

2.4.2 Supervised methods

The objective of supervised methods is to learn patterns and rules from the metabolomic dataset to predict new data, through an input matrix (i.e., NMR or MS spectra) and an output vector of responses (either continuous or discrete data) [151]. These methods rely on feature selection to identify a subset of variables that can discriminate classes with the objective to build a robust model for classification or regression [158].

There are three types of feature selection methods, based on how they are correlated to the model[159]. The first type are the filter methods that select features according to statistical tests independent of the machine learning algorithm, mainly univariate approaches such as Analysis of Variance (ANOVA) or chi-square[159]. Since so many tests are performed using, i.e., ANOVA, it is highly recommended a p-value adjustment to detect significant differently expressed metabolites [152]. However, univariate filtering methods do not account for feature dependency leading to worse performances which gave rise to a number of multivariate filter techniques [160].

The second one is named wrapper methods and they select features by training and testing a specific model, starting with a possible set of features and searching the ones that provide the best performance [160]. They are mainly defined by their searching methods that can be divided into three categories: exponential, sequential/deterministic and randomized [159]. Exponential is an exhuastive search that is impractical in metabolomics due to the dimension of the search space[159]. There a lot of different sequential/deterministic approaches being forward selection and backward selection the most well-known [161]. The first begins with an empty set and incorporates features with each iteration whereas backward selection begins with all possible features and eliminates the ones that affect performance the most [161]. There is an instance of backward selection, Recursive Feature Elimination (RFE), that ranks features with each iteration providing the user a feature ranking according to the elimination order [162, 163]. Randomized approaches perform heuristic searches in the possible feature space[159] through different algorithms such as genetic algorithm[164, 165] and particle swarm optimization. Wrapper methods perform better than filter

methods[159, 162], however there is a higher overfitting risk and processing computational demand [160].

The last type of feature selection methods are the embedded ones where the search for the optimal set of features is built into the model construction [160]. Features are selected at the same time that the model is trained easing computational effort and cost while improving model's performance [159]. This method can thus provide the advantages of the previous methods, the wrapper's feature-model interaction and the filter's lower computational demand (higher than filter methods but lower than wrapper ones). There different types of embedded methods being the regularization models the ones that are usually used due to their good performance[161]. Features are associated with coefficients that are forced to be small or zero (eliminating the respective feature) by objective functions that minimize fitting errors [161]. Examples of this models are Lasso Regularization[166], Adaptive
Lasso[167], Bridge Regularization[168, 169] and Elastic net regularization[170] that are reviewed by J. Tang, *et al*.[161].

For biomarker discovery purposes, building discriminant classification models is normally the path to go because the problem is based on class membership differentiation [151]. When using supervised classification methods, a fundamental step is validation because models that fit the data perfectly loose the ability to predict correctly new data and, in some cases, they can give a correct classification despite the lack of relationship in the data [151]. Therefore, it is necessary to validate the model's predictive performance either through cross-validation, double cross-validation, permutation, an external validation set or through a new independent validation set [151]. Since a new set of samples coming from an independent new experiment is rare, the use of an external validation set by means of data partitioning algorithms provide reliable predictive performance results [151].

Besides validation, evaluation of a classification model performance is also a key step when using machine learning approaches. This can be achieved through many ways such as classification accuracy rate, root-mean-squared estimates of model loadings, R^2/Q^2 plot and Receiver Operating Characteristic (ROC) curve analysis [171, 172].

ROC curves are a non-parametric measurement utility that assesses the performance through sensitivity and specificity rather than the prevalence of an outcome [173]. These curves plot one minus the specificity on the x-axis and the sensitivity on the y-axis and the classifier potential is often summarized by the AUC [172]. The AUC can be

interpreted as a probability that characterizes the classifier on its instances ranking capability of true positives before true negatives [173].

When taking into consideration AUC analysis in metabolomics Confidence Intervals (CI) should be calculated for the probability, since a potential biomarker intends to be applied on a larger population [173]. One way to achieve CI is through bootstrap percentile re-sampling [174], a method that constructs new samples from the original ones and where the 2.5 and 97.5 percentiles are taken to produce the 95% CI [173]. It has to be taken into consideration that when multiple comparisons are performed, a correction to the CI has to be employed such as Bonferroni or Benjamini-Hochberg False Discovery Rate (FDR). Since some machine learning methodologies can compute internal measurements of each metabolite's importance, it is possible to combine this information with the projection algorithm to produce a ROC curve that enables the selection of a useful metabolite subset [173]. In terms of biomarker performance evaluation, ROC curve analysis is considered to be the go-to statistically valid method [175, 176].

Linear methods

Partial Least Squares - Discriminant Analysis (PLS-DA), the classification approach of the regression technique Partial Least Squares (PLS), calculates latent variables (linear combinations of the original variables, similar to PCs in PCA) that maximize the covariation between observed data and categorical response variables [148, 149, 151, 152]. Since latent variables come from linear combinations of the original variables, the PLS-DA model can be expressed as $Y = Xb + r$, where Y is the class membership vector, X the data matrix, b a vector of regression coefficients and r a vector of residuals [151]. PLS-DA performs best when the variables are highly correlated and the potential markers can be obtained from latent variables and interpreting their contribution to the variation and correlation within the dataset [149, 151].

Despite the better separation of classes done in PLS-DA, compared to PCA, there is still variation present in the scores that is not correlated directly to the response classes [177]. To fix the problem of this direct and indirect correlated variation an orthogonal approach was developed, Orthogonal Projections to Latent Structures (O-PLS) [178], later applied to classification problems, Orthogonal Partial Least Squares - Discriminant Analysis (OPLS-DA). In order to obtain an orthogonal model, it is necessary to apply an orthogonal signal correction filter, separating variation linearly related to response from the one uncorrelated [177]. So, when the response estimation occurs, only the

variation linearly correlated is used, producing a model with higher interpretability and without structured noise [151]. It can happen that the original dataset has low or none uncorrelated variation naturally, which means the predictive performance of an OPLS-DA will be identical to PLS-DA.

Non-linear methods

Non-linear methods have been highly considered for pattern recognition due to the fact that biological processes are complex, with interactions that lead to a non-linear metabolic response [151]. They are also capable of being influenced by external factors which are not linearly related to different classes[151].

Support Vector Machine (SVM) is a kernel-based non-parametric machine learning technique that can be used for classification and regression problems [179]. It performs a kernel transformation to map data into a high-dimensional space allowing for different group separation [180]. Only a small fraction of the samples are identified, the support vectors, that enable the creation of a discriminative hyper-plane between the two classes [151, 180]. Depending on the problem to solve, the kernel transformation applied can be linear or non-linear providing this technique the necessary flexibility to different problems. There are different kernels available, such as, polynomial, Gaussian and sigmoid functions [181]. However, the results are not transparent because there is no easy way to visualize them[180].

SVM allowed for a classification of lung cancer cases versus control with an accuracy of 93.3%, suggesting that blood plasma metabolites analysis through Electrospray Ionization (ESI)-MS hold clinical potential in early-stage human lung cancer diagnosis [182]. Guan *et. al* [183] also applied SVM with RFE to LC-MS data in order to distinguish ovarian cancer samples from control, achieving over 90% accuracy. Allied to least squares, SVM with Gaussian kernel was applied to NMR data to classify patients with major depressive disorder achieving an accuracy of 96% in the test set, providing an auxiliary diagnostic tool for this disorder [184].

Random Forests (RdF) are a combination of decision trees, that sample different sets of random variables to build them [185]. The method uses bootstrapping with replacement to split the data, getting around 63% of all samples in the training sets and the remain on the test sets [186]. RdF have several advantages as they can be robust to over-fitting and outliers, handle missing data, deal with complex datasets without deleting variables, which results in a highly accurate classifier [180]. There are different

variable importance measures that may differentiate feature selection schemes, i.e., Gini importance and permutation importance[187]. Gini importance demonstrated to be biased when predictor variables vary in their scale or number of categories while permutation importance showed reliability when constructing decision trees based on subsampling wihtout replacement [188].

In 2013, T. Chen *et. al* [172] compared four classifiers (PLS, SVM, Linear Discriminant Analysis and RdF) on discriminating healthy patients versus patients with colorectal cancer from a GC-MS analysis, where RdF had better predictive ability than the other three classifiers. L. Zhao *et. al* [189] applied RdF models to investigate the syndromes associated to Coronary heart disease in NMR data, achieving a visual discrimination of two syndromes associated with this disease and obtaining twelve metabolites that could be considered as potential biomarkers.

3

DEVELOPMENT

This chapter covers the development process of the necessary functions to analyze 2D-NMR spectra that ended with a general workflow for this type of data, available through *specmine* package in the form of a vignette. Vignettes are long-form documentation guides to packages or even specific problems that the package is intended to solve[190].In this case, a vignette was done to guide the user through the new functionalities developed, regarding data input, visualizing, peak detection and further analysis.

All the tools and technologies used to achieve this goal are described, with special focus on the R package *specmine*, which integrated the work developed and served as a basis for the development. The code was developed using RStudio (version 1.3.1073)[191], an integrated development environment for R and Python. This work can be divided in the following processes of data analysis and structuring:

- Representation of two-dimensional data;
- Data reading;
- Data summary;
- Data visualization;
- Dimensional reduction.

3.1 *specmine*

In the past few years, the host group developed and improved an R package, called *specmine*[2], that provides methods for metabolomic data analysis in a user-friendly environment integrating functions from other R packages for a complete workflow. The

package has the capability to read data from different metabolomic techniques such as NMR, MS-based (GC and LC), Ultraviolet - visible (UV-vis) and IR[2]. This means different file formats are supported through functions from *xcms*[192] (NetCDF, mzDATA and mzXML data), *ChemoSpec*[193] for (J)DX data and R itself for Comma (or Tab) Separated Values (CSV or TSV) files[2]. Metadata files can be given to the data reading function in CSV/TSV file format.

Independently of the data format, *specmine* represents a dataset as an R list with several fields that characterize the data, presented graphically in Figure 2. This supports reproducibility in metabolomics as well as flexibility, since it provides the same data structure for different original data formats. The users can then establish their own framework of analysis by performing the necessary preprocessing steps, univariate, principal components, clustering and regression analysis as well as machine learning approaches to build predictive models [2]. It is also possible to perform feature selection to choose a subset of discriminatory variables with biological interest. Metabolite identification is possible through *MAIT*(https://www.bioconductor.org/packages/release/bioc/html/MAIT.html) which means it is only available for LC-MS data. In terms of visualizing any results obtained, the package implemented base graphic R functions, while, in some cases *specmine* relies on *ggplot2*(https://www.rdocumentation.org/packages/ggplot2/versions/3.3.0).

In terms of preprocessing steps, *specmine* is able to perform treatment of missing values, remove unecessary data or variables (accordingly or not to the presence of missing values), transform data, scaling, correction, smoothing interpolation, conversion of metadata variables to factors, mean centering, subset the dataset by samples and/or variables, data fusion, flat pattern filtering or data/metadata replacement. It is possible to treat a missing value by replacing it with a specific one (user-defined), according to the mean or median of the variables, using K-nearest neighbour averaging or with a linear approximation. One can perform a logarithmic or cubic root transformation.

In terms of scaling, there are four options available: auto, range, Pareto and interval. Background, offset and baseline corrections are available and the user can choose the baseline method when baseline correction is desirable. For smoothing interpolation the methods available are Bin, Loess and Savitzky-Golay. Data normalization can be achieved through the median, a reference sample/variable or the sum of a constant (user-defined). Low variance variables can be filtered out by specific functions such

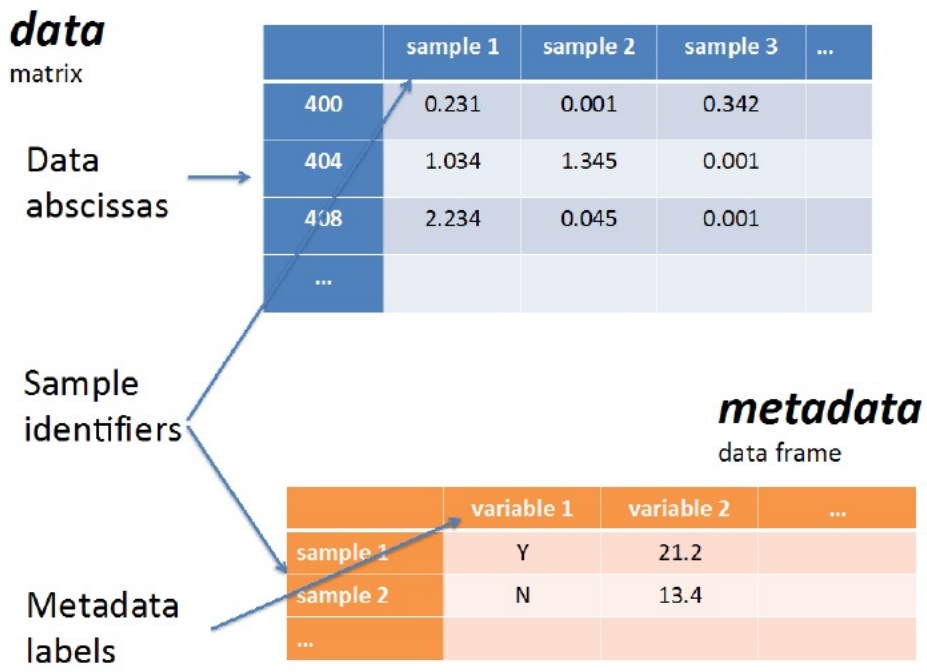


Figure 2: Graphical representation of a *specmine's* dataset structure. Retrieved from C. Costa, *et al.* [2]

as interquantile range, relative standard deviation, median absolute deviation, mean or median using a percentage or threshold given by the user.

Using the *stats* R package, *specmine* can provide several functions to perform univariate, correlation and regression analysis. The user can execute *t*-test, one-way and/or multifactorial ANOVA with the Tukey HSD post-hoc test and non parametric tests such as Kruskal-Wallis and Komolgorov-Smirnov. The parametric tests can handle multiple testing providing the ajusted p-values according to the FDR. Fold change analysis can be performed in each variable comparing two groups or in two specific variables for precise analysis and the results can be visualized in tabular and graphical forms.

Unsupervised multivariate analysis performed by *specmine* include PCA and two clustering methods. It is possible to perform classical and robust PCA, using this last one the R package *pcaPP*[194], and the user can visualize the results through scree plots, scores plots, biplots and pairs plots[2]. Regarding clustering methods, k-means and hierarchical are available with the option to the user to choose the different distance metric in the case of hierarchical one and the number of clusters on the other one.

Supervised and feature selection methods are provided through the application of functions from the R package *caret*. One can train, use and evaluate machine learning classification or regression approaches with validation methods such as k-fold cross-validation, leave-one-out cross-validation, resampling, repeated cross-validation and leave group out cross-validation. Error metrics estimated by these methods are also available including accuracy, AUC ROC analysis, Kappa statistic for classification, Root Mean Square Error (RMSE) and the coefficient of determination for regression. In terms of feature selection, both filter and wrapper methods are provided in order to determine which attributes present more value for the problem in question.

Since the release of the package, several improvements have been done to extend the capability analysis of the framework. Pathway analysis was implemented through connection to Kyoto Encyclopedia of Genes and Genomes (KEGG) and HMDB databases where it is possible to retrieve the organism and compounds' names. Another functionality was metabolite identification for NMR peak data, a work developed based on the R code proposed by Jacob *et al.* [195]. The method is based on peak clustering (through *igraph* package) according to a correlation value where each cluster is considered to be a potential metabolite [196]. After setting a library of reference metabolites, from HMDB, it is possible to compare each cluster to a reference metabolite using the Jaccard index.

In a recent work, developed by the host group, a web-based application of the package *specmine* was created to provide users, with no programming skills, means to perform metabolomic analysis. *WebSpecmine*(https://webspecmine.bio.di.uminho.pt/)[3] is an easy-to-use and freely accessible tool with the capability to be flexible towards different metabolomics experiments since there is not a fixed workflow for data preprocessing and following analysis. *WebSpecmine's* interoperability with *specmine* is overviewed on Figure 3 where it is also presented the tools that made possible this connection. In addition to this, *Webspecmine* is able to store data which *MetaboAnalyst*(https://www.metaboanalyst.ca/) cannot, also a website based on an R package (*MetaboAnalystR*[197]) with the same purposes.

The package *specmine* was used in four case studies to prove its capability and to provide useful pipelines for data analysis[2, 198]. Their entire analysis is available here, work developed by C. Costa [198].

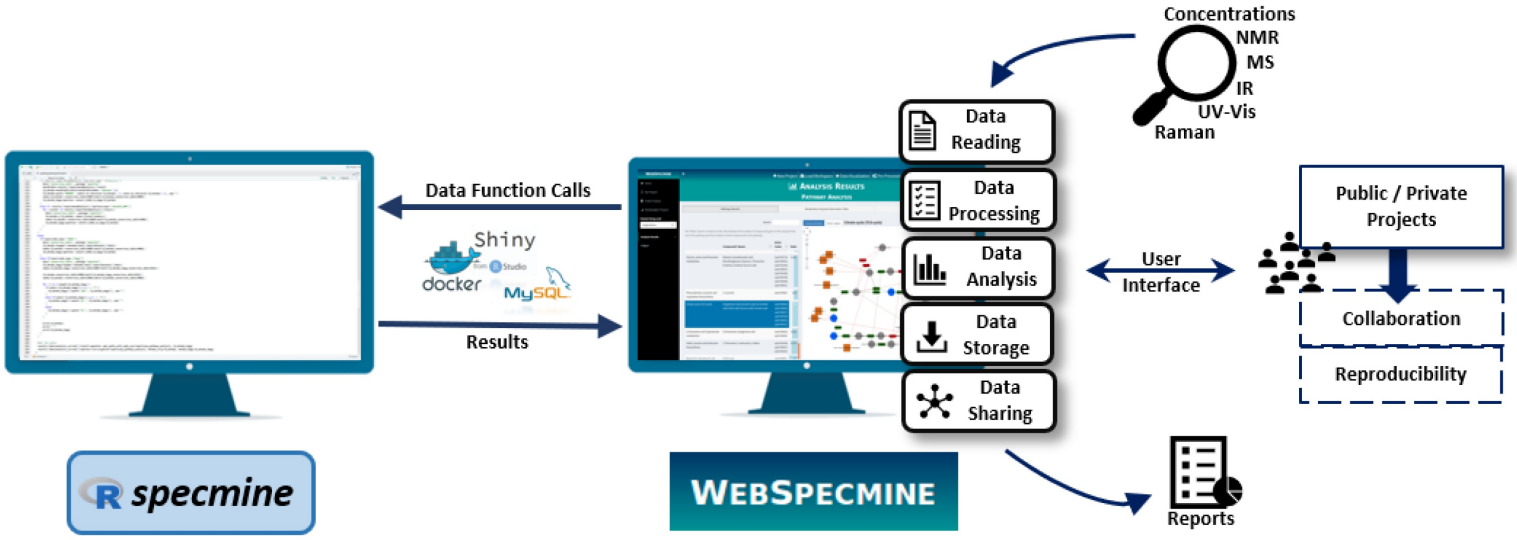


Figure 3: Overview of *WebSpecmine's* implementation and features based on *specmine* and the tools Shiny, MySQL and Docker. Retrieved from S. Cardoso, *et al.*[3].

3.2 Representation of two-dimensional data

In order to develop tools to analyze 2D-NMR metabolomic data it is needed an object that can represent this type of data in *specmine*. Currently, it is implemented the structure presented on Figure 2. However, a similar structure is not possible when a sample is considered a 2D matrix where the variables across the dataset are combinations of values from two different dimensions. In 1D, a single x axis can represent the resonance frequency of a 1D-NMR study and y axis values represent the intensity measured. In 2D-NMR, two x axis are necessary to provide the resonance frequencies of an experiment (one for each dimension) and intensity values shift from y axis to z axis, where each point results from two resonance frequencies. The directly measured dimension resonance frequencies' are present in the columns and the indirectly measured ones are present in the rows of a matrix. To ease user's interpretability it was considered the basic structure of a *specmine* dataset, an R list, with changes in the data field and addition of new fields. A 2D dataset on *specmine* will consist on the following fields:

- **data** - The metabolomics experiment data', stored in a list of numeric matrices where columns represent *ppm* values' from the direct dimension and rows represent *ppm* values' form the indirect dimension. Each matrix represents a sample and

values in the matrix represent the 1D-corresponding y axis, intensity values for the combination of resonance frequencies. The names of this field in the object keep the sample identifiers;

- **type** - String indicating the type of data. Currently it is only allowed ”2d-nmr” and ”undefined”;
- **description** - String that describes the dataset and possible pre-processing steps performed on it;
- **metadata** - Extra variables regarding information on samples, stored in a data frame (columns are variables and rows are samples);
- **F1_ppm** - A numerical vector that represents the resonance frequency for the indirect dimension (F1);
- **F2_ppm** - A numerical vector that represents the resonance frequency for the direct dimension (F2);
- **labels** - list that allows the user to define labels for each axis; defaults: x-axis ($x) and y-axis ($y) are ”*ppm*” and z-axis($val) is ”intensity”.

This new modified structure is presented on 4. As in *spemcine*’s 1D structure, the data and type fields cannot be NULL. Furthermore, data field has to be a list and names for both rows and columns have to be numerical, otherwise the dataset will be considered invalid. This R list can be generalized to other types of 2D metabolomics data, enhancing *specmine*’s flexibility.

3.3 Data reading

The function *create_2d_dataset* allows to create the structure above mentioned and it is called by the functions that read the 2D-NMR data. Following the work developed by the host group, it was necessary to read from two different NMR instruments: Bruker and Varian. Available data from Bruker is structured in a different way that it is when collected through Varian instruments. This demanded two different functions to allow a complete implementation of 2D-NMR data reading on *specmine*, increasing it’s readability of available metabolomic spectra.

Figure 4: Representation of the structure of 2D data in a *specmine* dataset.

3.3.1 Bruker Data

In the case of Bruker files, additional files regarding the second dimension are identifiable through the number two in their name. A simple scheme of how files from Bruker instrument are displayed is present on the Figure 5. The most important files to read 2D Bruker spectra are *2rr* (processed data), *procs* (processing parameters for the first dimension) and *proc2s* (processing parameters for the second dimension).

The implementation on *specmine* followed the code already developed for 1D-NMR from Bruker files, changing the function that reads only one NMR spectrum. The modification was calling the function *read_Bruker* from the package *mrbin*[199] assigning the folder parameter to each sample spectrum and the dimension parameter to "2D". This function returns an intensity matrix containing raw *ppm* values for the direct and indirect dimensions (columns and rows, respectively). The *ppm* values are then rounded to two decimal places and the matrix is appended to a list of matrices associated by sample id. The new function *read_Bruker_files_2d* includes also parameters that users can change, i.e., description, metadata file and labels for each dimension.

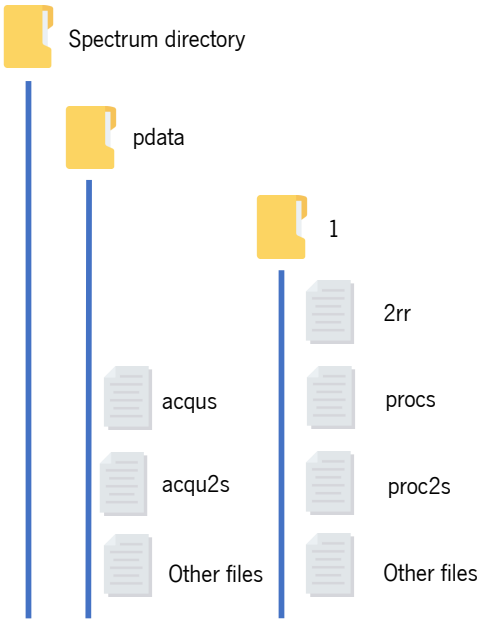


Figure 5: Organization of Bruker files from a 2D metabolomics experiment. Icons designed by DinosoftLabs from Flaticon.

3.3.2 Varian Data

Regarding Varian files, they follow the same scheme as 1D raw spectra. The files present in a Varian sample directory are the *fid* file (contains the raw spectrum, in binary format), *log* file (contains the recorded events on the acquisition process), *procpar* file (contains the parameters of the experiment) and *text* file (contains additional information of the experiment). The second dimension is provided by the *fid* file and key varian instrument parameters in *procpar* that allow a correct preprocessing of the spectra.

Following the code developed to read 1D Varian data, was developed a script in *Python3*[200], using the package *mrglue*[201]. As far as preprocessing the *fid* file, the example provided here shows that 2D Varian data should be preprocessed in both dimensions, performing for each dimension an apodization, zero-filling, a fourier-transformation, phase-correction and imaginary numbers removal. There are multiple ways to perform an apodization and *nmrglue* provides a set of them, including generic, exponential, lorentz-to-gauss and sine bell. The apodization that the developed script uses is exponential since this apodization was the one already used in the 1D script. The script consists in a

function that takes the same arguments as 1D (directory folder, *fid* file, *procpar* file and two boolean values to either perform or not apodization and zero filling) and returns a data matrix and a list of two vectors of *ppm* values. This script it is incorporated in an R function that is present in the general function *read_varian_2dspectra_raw*, whichs reads multiples sample spectra from a directory. In order for this function to work, the user must have *Python3* and the package *nmrglue* installed on the computer.

3.4 Data summary

In order to access some initial information on the dataset the functions *check_2d_dataset* and *sum_2d_dataset* were developed. Both functions were developed based on the existing corresponding functions for 1D. The first validates or not a 2D *specmine* dataset. This type of dataset is not valid if:

- data field is null;
- the number of sample spectra in data field does not match the number of rows in metadata (when it is not null);
- the type of data is not allowed for this type of dataset;
- the column/row names are not numeric values (it considers the first spectrum as reference).

The second function presents to the user a summary of the dataset with some statistics if the *stats* option is *TRUE*. It iterates over the multiple matrices and applies different functions to obtain different information. Currently, it prints if the dataset is valid, its description, type of data, number of samples, number of data points and information regarding metadata and labels, if they are not null. In case *stats* equals *TRUE* it prints the number of missing values, mean, median and standard deviation for each spectrum.

3.5 Data Visualization

The plotting of one or more 2D spectra is achieved by the incorporation of the package *plotly*[202]. This package is an R version of the open source *plotly* graphing libraries, helping to build interactive and high-quality graphs. It works similarly to the *ggplot2*[203]

$$f(x) = \frac{\overline{x_{i,j}}}{\sigma_{x_{a,b}}} \tag{1}$$

$$where,$$

$$x_{i,j} = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{i,j} \\ z_{2,1} & z_{2,2} & \cdots & z_{i,j} \\ \vdots & \vdots & \ddots & \vdots \\ z_{i,1} & z_{j,2} & \cdots & z_{i,j} \end{pmatrix}$$

$$a = \frac{\sigma_i}{\bar{i}} > 15$$

$$b = \frac{\sigma_j}{\bar{j}} > 15$$

Formula 1: Signal-to-noise ratio of a spectrum.

package for R, where a user can create a *plotly* object and add multiple customizable fields, such as, labels, colors, axes, text annotations, ... It was developed a function, *plot_2d_spectra*, that allows the user to visualize a single/group of spectra within the same plot, in an interactive way. Interactivity enables the user to zoom in/out, hover a peak (receiving information regarding its intensity and F1 and F2 dimension indexes), rotate the plot and select which spectra to plot. Besides interactivity, if the user does not give any samples to plot, the function will generate a plot with four spectra, the two with higher and lower SNR.

The SNR of a spectrum can be calculated as it shows in Formula 1.

It was formulated based on the work of Wang *et al.*[204], where they related the Coefficient of Variance (CV) ($\frac{\sigma}{\mu}$) and the SNR in NMR-based metabolomics studies. In their case, they used a specific noise region (9.5 - 10 ppm) to extract its standard deviation and divide the intensity of a peak by that value. In a generalized case the noise region has to be identified based on the data and it is supported by the literature that noise regions have a high CV (above 15 %)[204–206]. Following this information, noise regions are represented as a subset of the spectra, where ppm values of both dimensions are selected if their CV is higher than 15. Instead of looking for possible peaks across spectra with higher CV (in 2D-NMR metabolomics , combinations of ppm values) it looks for regions of the spectra which is computational quicker and easier to interpret. The mean of intensity values in the spectra is divided by the standard deviation of the calculated noise regions achieving the SNR of a spectra.

The function *plot_2d_spectra* can be divided in sections that represent the steps to build the entire *plotly* object. The first section involves sample selection, whether if it is done through SNR as described above or by user input (either through a character vector with sample names or numerical vector with indexes). The second section is the buttons list building, a necessary list that for each sample selected in the previous section creates a graphical button in the interactive plot, presented as a dropdown menu. The third section adds the selected sample' spectra into a single *plotly* object as surfaces and they are colored differently if a metadata variable is given or not. If this variable is not given, each spectrum has a different color. The fourth and last section adds the final layout to the plot, including the title (given by the user), axis names, the dropdown menu with the buttons and the legend.

3.6 Dimension reduction

Extracting information from the 2D-NMR *specmine* structure, mentioned in the first section of this chapter, is a computational challenge because any operation iterated over a list of large matrices will take a long time to obtain results. With this in mind the objective of the function *peak_detection2d* is to perform peak detection and build a 1D structured *specmine* dataset with the combination of ppm values as variables (rows) and samples in the columns.

The algorithm to perform peak detection is a search for local maxima, finding points in the matrix that are larger than all surrounding points, taking from the rNMR[12], an open source software for NMR data analysis. The user can establish the degree of a filter that correlates the defined threshold with the points in the spectra, influencing if all/none of the points should be above the threshold or row/column points should be searched for local maxes. The search itself is based on the *intersect* function, within R *base* package, where two vectors (in this case, the same matrix with added Not Available (NA)s) are compared to find points that belong to both vectors. An example of how this function works in the algorithm is represented on 6. It returns a vector with matrix indexes which is further used to obtain the corresponding intensity values and the pair row/column that characterizes the peak. Regarding the threshold, if it is not given by the user, it is determined by calculating the mean of the intensity values of the spectra.

For each spectra in the 2D *specmine* list, a new matrix is built filled with NAs. This will serve as the template for the new 1D dataset because it is easier to work with an

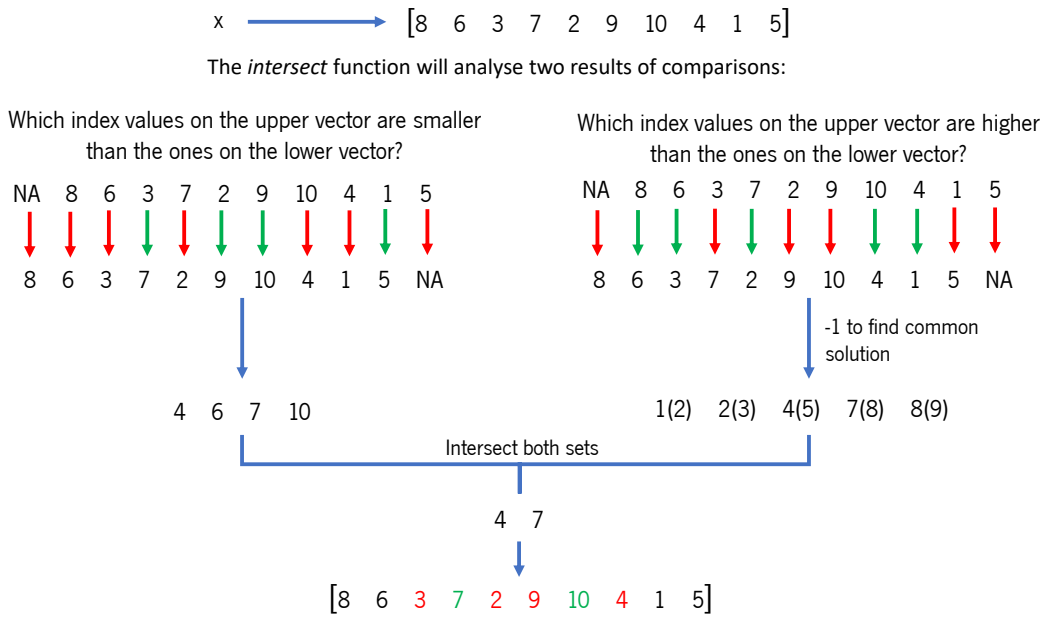


Figure 6: Example on a simple vector of how *intersection* function is implemented on the local max search algorithm.

empty matrix rather the original. From the peak list (result of the peak detection phase), each pair (row/column) that had a peak will have its intensity value in the empty matrix on the same pair. This will result in a matrix filled with NAs except on the points that were detected as peaks, for every sample. With the new list of recent matrices it is built a Three-Dimensional (3D) array and columns or rows that have negative ppm values are removed, according to the user input. The 3D array is then converted into a 2D matrix using the function *two.d.array*, from the R package *geomorph*[207]. Due to the existence of identical ppm values in different dimensions, it was necessary to establish unique names for the new variables which was possible through the base function *make.names*. The next step was to remove variables that had an empty expression of peaks, i.e., the row had only NAs throughout the samples. With this step done, it is now possible to create a 1D *specmine* dataset with the function *create_dataset*. This reduction of dimensionality will allow the use of the already developed functions for 1D datasets which eases the process of 2D-NMR analysis. One has to take into account that the newly created dataset has high percentage of NAs and a method for NA imputation should be used before analysis.

3.7 Further Analysis

After peak detection phase it is possible to apply the functionalities of *specmine* to analyze the standard 1D-NMR dataset, as mentioned on the last paragraph. In this section, the main options for uni- and multivariate analysis available in *specmine* will be summarized.

In terms of univariate parametric analysis it is possible to perform student's t-test and one-way and multifactor ANOVA. As for univariate non-parametric analysis it is possible to perform Kruskal-Wallis and Kolmogorov-Smirnov tests. It is also possible to perform fold change analysis on each variable of the dataset and on two specific variables for targeted differences of groups. Despite the test, the option to plot the results, highlighting different features, is always available providing meaningful tools to interpret any approach.

Regarding multivariate analysis, the main tool available is PCA. The user can perform either classical or robust PCA which means it has the option to customize how this analysis is conducted, being able to choose how to center and scale the dataset. In terms of visualization there is a wide range of methods to plot the PCA results. Ranging from a scree plot to k-means pairs plot, it is possible to understand the variance explained by each principal component and the contribute of each samples/variable to the result.

There are also methods to perform clustering, regression and correlation analysis. The user has the option to perform hierarchical and k-means clustering within the same function, being able to select which metrics to use for hierarchical and the number of clusters for k-means. The methods implemented for regression and correlation analysis allow to test all variables at once and in the case of correlations tests it is possible to visualize the result through heatmap.

In more specific cases there are methods implemented in *specmine* for feature selection and machine learning. The methods are wrappers for functions and models from package *caret* and there are several customizable options which provides users the tools to try different approaches.

4

CASE STUDIES

In this chapter, the new developed functions will be tested with real data, creating pipeline examples of how this type of data can be analyzed, using *specmine*. The case studies will be divided according to NMR recording instrument. The first case study involves two MetaboLights[208] studies (MTBLS131 and MTBLS132) regarding an absolute quantification experiment of tomato fruit extracts using different frequencies. These will also serve as examples of how further analysis can be done by incorporating both datasets using the frequency as a metadata variable. The second case study is from a Varian Metabolomics Workbench[209] study (ST000103) where a new untargeted NMR approach was applied to compare the exometabolome and endometabolome of worms.

Negative chemical shift values are present in both case studies. One of the steps within signal processing is direct current correction which is usually performed by the instrument software. In terms of 2D experiments the multiple FID's are short and have fewer data points which means this correction is more challenging and less accurate, causing offsets on chemical shifts[210]. Manual phase cycling procedures before the first FT are recommended to remove offsets[210]. The results presented in this chapter do not include manual processing of these offsets since available data is already processed and remaining chemical shifts should not hinder visualization or analysis.

4.1 TOMATO FRUIT EXTRACTS

4.1.1 INTRODUCTION

One-dimensional proton NMR protocols have been widely used as an untargeted approach due to its efficiency and post-analysis comparison[211, 212]. However, efficient

discrimination of metabolite quantities is not possible due to signal overlapping and a more targeted strategy (identification and quantification of a low number of metabolites) becomes more reliable when searching for biomarkers[4]. Despite the advantages of targeted approaches on the quantification of metabolites, complex biological samples limit the deconvolution process that allow the necessary fitting of individual metabolite 1D spectra[213]. Complex spectral patterns, lack of universal reliable internal standards, sample size or limited dynamic range are some factors that prevent precise quantification through 1D-NMR and 2D-NMR emerged as one way of reducing spectral overlap[4, 214].

The use of 2D-NMR for quantitative metabolomics is hampered by the long experiment duration which can influence unstable samples or provide noisier data and most of them rely on a calibration procedure[4]. Ways to overcome these challenges comes from reducing the acquisition time, using an ultrafast method where a 2D spectrum is obtained from a single scan[26]. This methodology has been evolving during the last decade, in order to provide a higher resolution spectra, emerging new hybrid strategies where single-scan experiments are repeated several times to improve the quality of the spectra[59]. These hybrid strategies have been applied to different studies of metabolomics, such as breast cancer cell extracts[39] or pig lipid serum[215], and guidelines have been made to ease its implementation and use[216, 217]. The most simple hybrid approach is the Multi-Scan Single Shot (M3S) acquisition method and by accumulating ultrafast scans over a period of time (determined by the duration of the pulse sequence and the number of accumulations) it is possible to achieve higher sensitivity per unit of time[218].

In the work that originated this case study, M3S COSY experiments were applied to tomato fruit pericarps in order to validate this technique as a method for quantification of major metabolites in biologically relevant samples. Tomato fruit pericarps were chosen as biologically relevant samples because their major metabolites signals' are overlapped on 1D proton spectra[4], it has a well characterized metabolism[219], its compositional changes across fruit development are also well characterized[220] and it is considered as a model species for the study of fleshy fruits[221]. The authors performed the experiment on two different acquisition frequencies, with calibration standards, to provide insight on the method's applicability and compare quantitative results regarding key metabolites on different stages of fruit development.

The aim of this case study is to access these major metabolites using the functions developed and achieve consistent results with metabolic behaviour of tomato fruit by comparing with the authors' results.

4.1.2 2D-NMR DATA

The tomato samples (*Solanum lycopersicum* L. cv Moneymaker) used in this study were harvested at four different stages post anthesis (DPA) and on three different trusses, with three biological replicates for each truss on each development stage. A total of 36 samples were collected and each stage (8, 21, 34 and 55 days) has 9 samples. These 9 samples are divided by the three different trusses (5, 6 and 7). The ultrafast COSY experiments were recorded on two different spectrometers (500 and 700 MHz), having a final spectral width of 4.5 ppm in both dimensions and 5 ppm (F2) x 5.4 ppm (F1), respectively.

The entire data for the experiment is available via MetaboLights under the accession numbers MTBLS131 (tomato samples recorded in 500 MHz), MTBLS132 (tomato samples recorded in 700 MHz), MTBLS133 (calibration samples at 500 MHz) and MTBLS134 (calibration samples at 700 MHz).

Using *specmine*'s function *get_metabolights_files_assay* it is possible to retrieve all data files from one assay, related to a study. After retrieving data files, using the newly developed function to read Bruker NMR data will generate an output in R identical to the one presented below. Warnings may arise from the lack of a metadata file or the existence of a temporary folder that needs to be replaced when data files are zipped.

```
Reading Metadata file
Reading sample P21.5.1_700_66 in C:/temp/P21.5.1_700_66/66/pdata/1
Reading sample P21.5.2_700_67 in C:/temp/P21.5.2_700_67/67/pdata/1
(...)          (...)              (...)
Creating dataset (this may take a while)
Done.
```

Listing 4.1: Example of an output of correctly read data files in R.

4.1.3 DATA SUMMARY AND VISUALIZATION

Initial accessment to the datasets allows to identify their key aspects, both for MTBLS131 and MTBLS132. In the example code below, a description of the dataset MTBLS132 is presented where it is possible to assess the number of missing values, mean, median

and standard deviation of each sample spectrum. Both datasets were considered valid datasets and none of the samples in each dataset had missing values. Standard deviation values should be high, considering intensity matrices where peak values regarding molecular interactions should be significant to identify spectral metabolite signatures. This condition was validated on both datasets which will increase the quality of peak detection.

```
Dataset summary:
Valid dataset
Description: MTBLS132 - 700 MHz tomato samples
Type of data: 2d-nmr
Number of samples: 36
Number of data points 524288
Number of metadata variables: 2
Label of x-axis values: ppm
Label of y-axis values: ppm
Label of pair'(x,y) values: intensity
Number of missing values in data:
P21.5.1_700_66 P21.5.2_700_67 (...)
             0              0
P34.5.3_700_68 P34.6.1_700_66 (...)
             0              0
(...)          (...)
Mean of data values:
P21.5.1_700_66 P21.5.2_700_67 (...)
      9354.143       9426.771
P34.5.3_700_68 P34.6.1_700_66 (...)
     11636.760      11302.586
(...)          (...)
Median of data values:
P21.5.1_700_66 P21.5.2_700_67 (...)
      183.9023       207.4102
P34.5.3_700_68 P34.6.1_700_66 (...)
      211.2031       213.7734
(...)          (...)
Standard deviation:
P21.5.1_700_66 P21.5.2_700_67 (...)
      99801.96       99816.54
P34.5.3_700_68 P34.6.1_700_66 (...)
     123293.76      119494.93
```

```
(...)                (...)
```

Listing 4.2: Example of a 2D dataset summary.

In terms of data visualization, the sample spectra corresponding to extract 514 was used to assess the capabilities of the developed function. Figure 7 shows the direct comparison between the reference spectra obtained by the authors for both frequencies, 500 and 700 MHz, which correspond to sub-figures **a** and **b**, respectively. The plot obtained on sub-figures **c** and **d** allow to dynamically visualize the spectrum, possible peaks of interest and it is not limited to the static sub-figures presented.

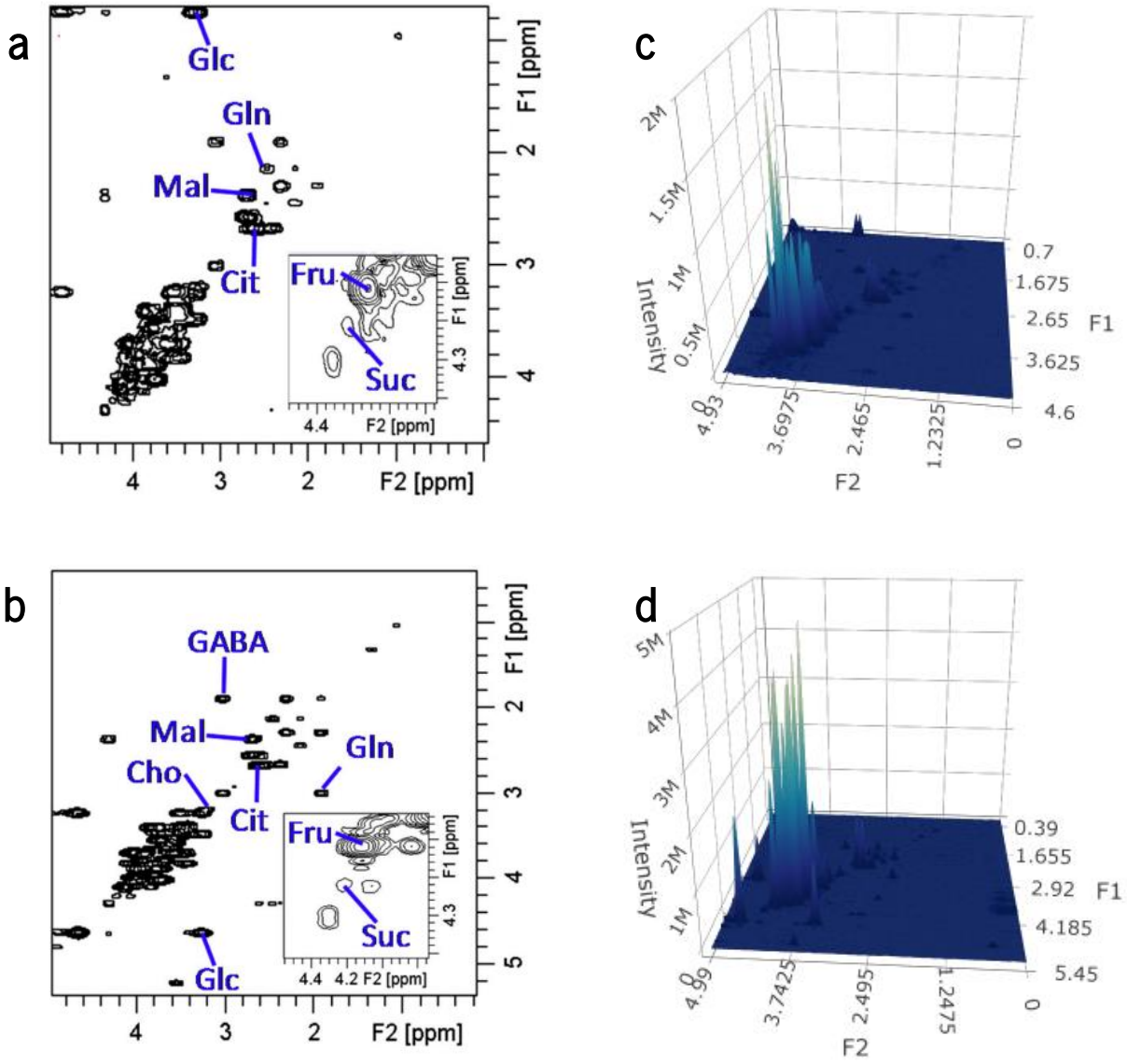


Figure 7: Fast M3S COSY spectra of a tomato fruit pericarp extract (extract 514, 34 days post anthesis) recorded in 5 min at 298 K on 500 (a) and 700 (b) MHz Bruker NMR spectrometers equipped with cryogenically cooled probes, taken from Jézéquel, *et al.*[4]. Plot of the second biological replicate of the same extract on 500 (c) and 700 (d) MHz, using *specmine* package.

It is possible to identify the main interactions between nuclei which are present on the diagonal of the spectra, for both frequencies. The layout of these interactions and ppm scales are consistent with the reference aswell as the difference in resolution from

Table 2: UF COSY NMR peaks used for the quantification of the 8 targeted metabolites. Adapted from Jézéquel, *et al.*[4].

Name	Abbreviation	UF COSY peak used for quantification			
		500		700 Mhz	
		F2 (ppm)	F1 (ppm)	F2 (ppm)	F1 (ppm)
Glucose	Glc	3.31	0.74	3.25	4.61
Fructose	Fru	4.12	4.12	4.12	4.12
Glutamine	Gln	2.47	2.15	1.91	3.01
Citric Acid	Cit	2.57	2.69	2.57	2.69
Sucrose	Suc	4.22	4.22	4.22	4.22
GABA	GABA	-	-	3.01	1.91
Malic acid	Mal	2.67	2.38	2.67	2.38
Choline	Cho	-	-	3.20	3.20

spectra with different frequencies. Peaks that are further away from the diagonal and have biological meaning, such as glucose (Glc), are also identifiable which represents a validation mark for this tool. The figures generated will always be colored with a gradient since the object is a surface, however, gradients can hinder discrimination of regions and peaks with low intensity values, such as the region where citric acid (Cit), glutamine (Gln) and malic acid (Mal) are detected for this experience. Despite this characteristic, the available option for the user to zoom in/out coupled with hover feature allows one to clarify these regions.

4.1.4 Statistical analysis

The search for peaks in the two datasets allowed to test the local search algorithm by comparing the combinations of ppm obtained with the reference ppm values for 8 targeted metabolites that were quantified using calibration samples. The reference values can be found on table 2. Table 3 has the number of hits for the combinations of ppm values across all samples. Regarding the count process, peaks with more or less 0.10 ppm than the reference on either dimensions were discarded. The threshold to detect the peak was calculated by the function, as described on the previous chapter.

There were detected 1529 peaks across all samples for the dataset MTBLS131 and 868 peaks for the dataset MTBLS132. This means that most of the noise from the dataset was removed and from a search space with more than one million variables (MTBLS131)

Table 3: Number of samples for which peaks were detected using *specmine*'s function for 2D-NMR spectra. Accounted peaks with reference $+/-$ 0.10 ppm.

Name	**Number of samples for each metabolite**	
	500 Mhz	700 Mhz
Glucose	9	7
Fructose	13	10
Glutamine	8	7
Citric Acid	6	21
Sucrose	9	5
GABA	-	8
Malic acid	8	5
Choline	-	14

and more than half million variables (MTBLS132) it was possible to get 1529 and 868 variables, respectively. Considering the results on Table 3, the developed functions for peak detection were validated for Bruker NMR data since it was possible to identify more than one peak for each metabolite' peak reference information. Overall the MTBLS132 dataset should present a higher number of peaks since it has higher resolution and peaks are better separated. This was not validated for most of the metabolites (glucose, fructose, glutamine, sucrose and malic acid) although the number of peaks obtained is similar between datasets. This can be explained by the rounding step of ppm values done when reading spectral data and intensity values immediately close to each other which makes intersect function not validate such values as individual peaks.

Regarding analysis, the first step after any procedure of peak detection done with the developed function is to treat the resulting NA values. In dataset MTBLS131, 51296 NA values were replaced and in dataset MTBLS132 were 27419, both sets of values were replaced with $5e-04$, using *specmine*'s function to perform missing values imputation. This replacement is intended to mimic noise throughout the samples, since a replacement by mean or other positive value could shadow peaks that were detected with low intensity values. Negative intensity values that passed through the peak detection step are also being transformed to $5e-04$ because further analysis require non-negative values and the purpose of the original study (quantification of metabolite concentration) require positive values.

Table 4: ANOVA results from dataset MTBLS131 after peak detection with development stage metadata.

Combination of ppm (X.F1ppm.F2ppm)	pvalues	logs	fdr	tukey
X3.72.4.06	1.584e-06	5.800	1.866e-04	21-8; 34-8; 55-8
X4.12.4.14 (Fru)	1.586e-06	5.800	1.866e-04	21-8; 34-8; 55-8
X0.75.3.24 (Glc)	1.597e-06	5.797	1.866e-04	21-8; 34-21; 55-21
X0.75.3.33 (Glc)	1.618e-06	5.791	1.866e-04	21-8; 34-21; 55-21
X3.83.3.85	1.622e-06	5.790	1.866e-04	21-8; 34-21; 55-21
X3.81.3.95	1.623e-06	5.790	1.866e-04	21-8; 34-8; 55-8
X3.57.3.77	1.632e-06	5.787	1.866e-04	21-8; 34-8; 55-8
X3.63.3.62	1.637e-06	5.786	1.866e-04	21-8; 34-8; 55-8
X3.88.3.95	1.663e-06	5.779	1.866e-04	21-8; 34-8; 55-8
X3.41.3.51	1.713e-06	5.766	1.866e-04	21-8; 34-8; 55-8
X0.76.3.3 (Glc)	1.785e-06	5.748	1.866e-04	21-8; 34-8; 55-8
X3.26.3.34	1.793e-06	5.746	1.866e-04	21-8; 34-8; 55-8

In terms of univariate analysis, the one-way ANOVA and Tukey's Honestly Significant Difference (HSD) post-hoc test results can be seen in Table 4 for dataset MTBLS131 after peak detection.

The first column of the Table 4 refer to the variables of the dataset MTBLS131 after peak detection, i.e., *X3.72.4.06* is the variable formed by the 3.72 ppm and 4.06 ppm on the indirect and direct dimension, respectively. The results from this table indicate that, by cross-referencing to Table 2, variable *X4.12.4.14* (Fructose) have significant effect in discriminating the tomato samples harvested at 8 days of development over the other stages, while variable *X0.75.3.24* (Glucose) have significant effect in discriminating the tomato samples harvested at 21 days of development over other stages.

Figure 8 shows the dendrogram plot resulting from hierarchical clustering (Euclidean distance and average linkage) with the different development stages as label colors, using dataset MTBLS131.

This dendrogram indicates that the samples are reasonably grouped on the development stages 8, 21 and 34 while samples from development stage 55 are spread throughout the clusters. The third biological replicate samples from 55 days of development stage, independent of the truss number, seem to have an average of detected peak intensities similar to development stages 21 and 34. This can be explained by combinations of ppm values that may represent metabolites produced on later stages of

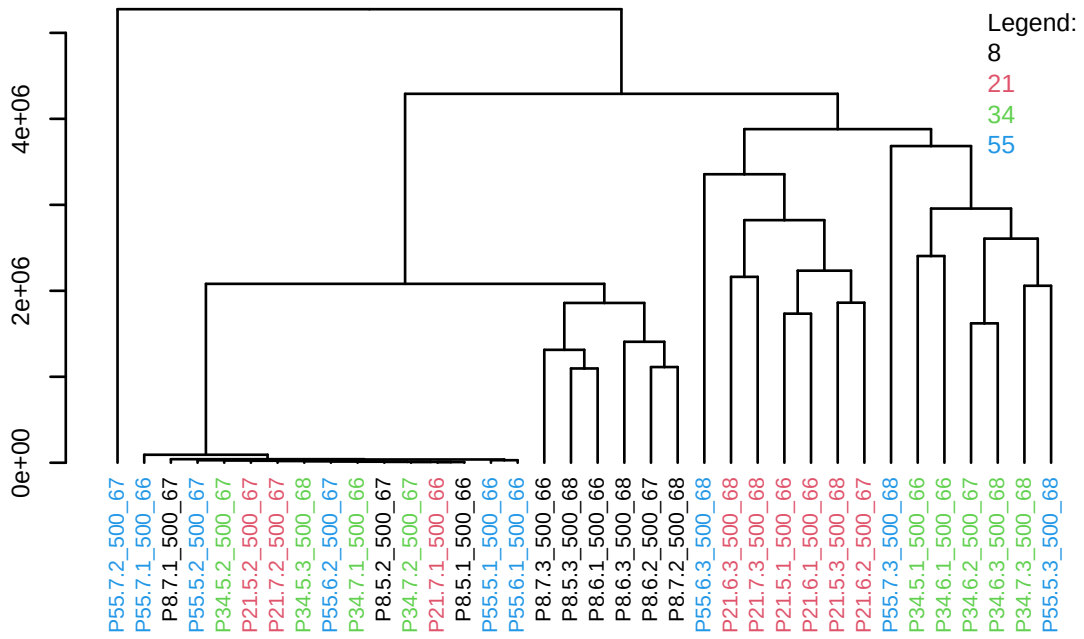


Figure 8: Dendrogram plot with development stages as label colors. Data from dataset MTBLS131 after peak detection.

development. On the other hand, combinations that may represent metabolites produced on earlier stages can explain the grouping of samples from all stages with the cluster from 8 days.

To further validate the tools developed, a graphical representation similar to the one presented in Figure 9 was plotted using data from both datasets after peak detection. The objective was to reproduce the changes in the pericarp metabolite contents throughout tomato fruit development. Figure 10 shows the multiple plots obtained through *ggplot2* and *gridExtra* package.

The results present on Figure 10 show that the approach to group combinations of ppm, which could be reasonably considered a metabolite signature, is not efficient on producing the same results overall. It is necessary to take into account that the Figure 9 was done using calibration samples and the metabolite content represented in y-axis is a measured concentration, where as in case of Figure 10, the values on y-axis are signal intensities that came from a mean of all the variables that were considered peaks from

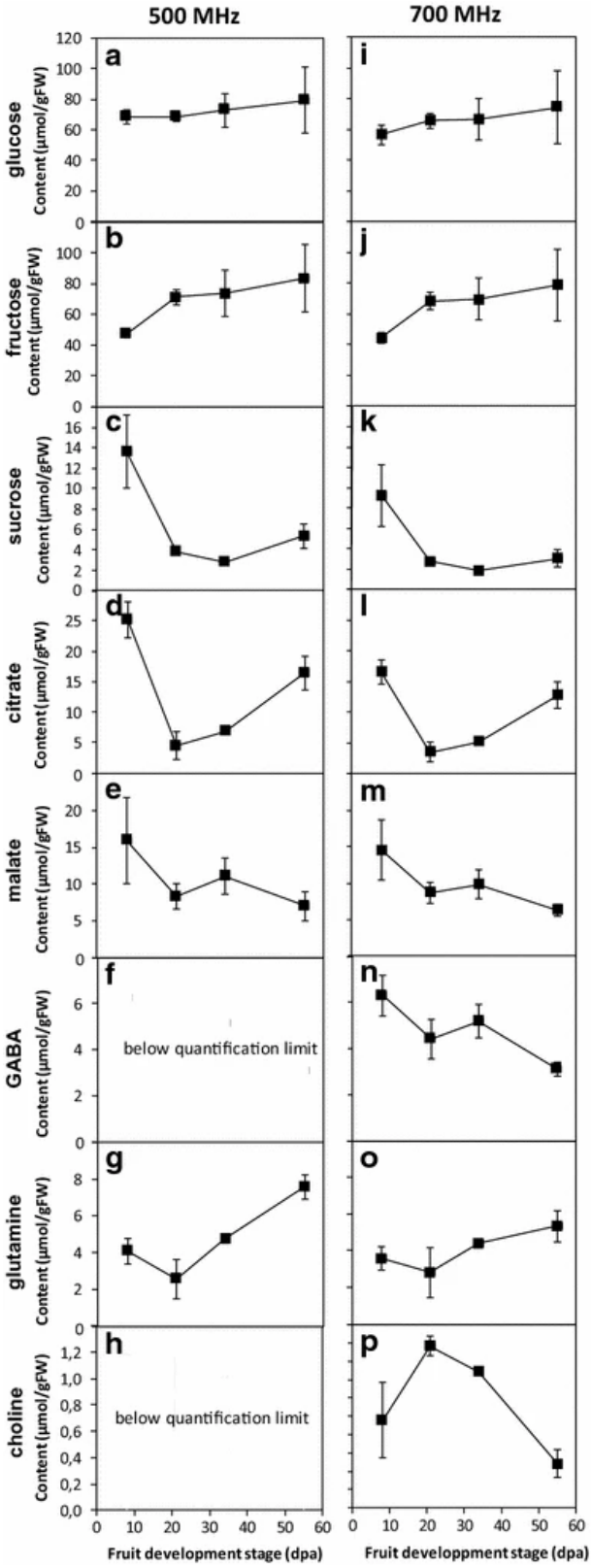


Figure 9: Changes of choline, glutamine, GABA, malate, citrate, sucrose, fructose, and glucose contents throughout tomato fruit development. Results obtained with the fast quantitative COSY at 500 MHz (a–h) and 700 MHz (i–p) on polar extracts. Taken from Jézéquel, *et al.*[4]

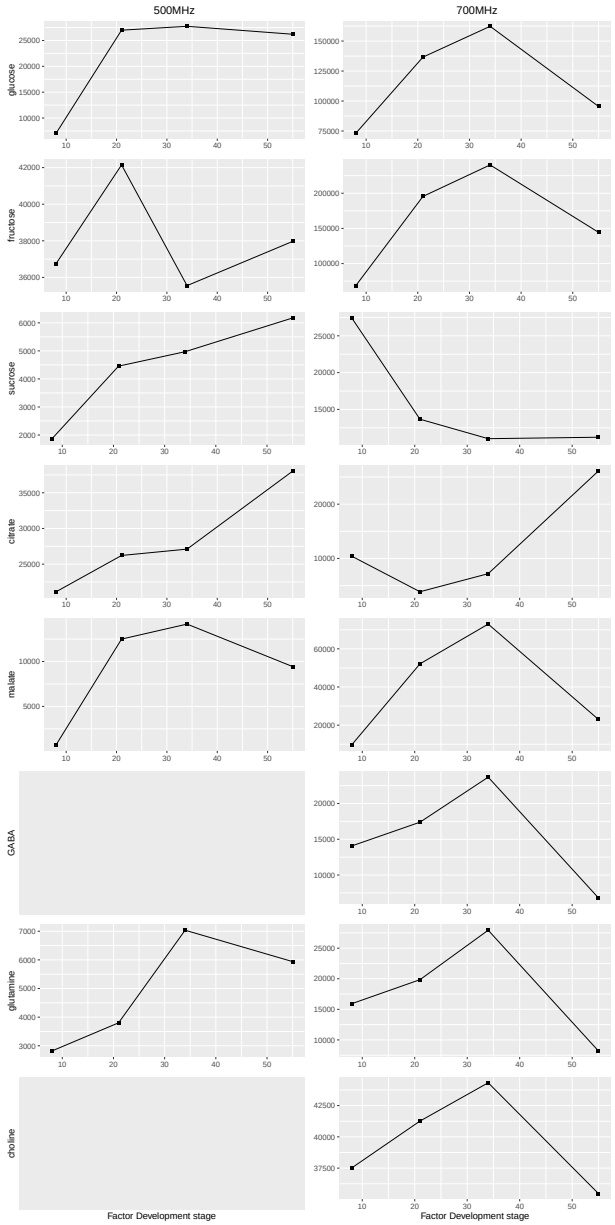


Figure 10: Changes of choline, glutamine, GABA, malate, citrate, sucrose, fructose, and glucose contents throughout tomato fruit development, using *ggplot2* and *gridExtra* on the data from MTBLS131 and MTBLS132 after peak detection, 500 and 700 MHz, respectively. In this figure, the signal intensity (y-axis) is plotted as a function of fruit development stage (x-axis), for each metabolite in both frequencies.

each metabolite. Nonetheless there are changes throughout tomato fruit development obtained with *specmine* that are similar to the ones found with precise measurements. Sucrose, citrate, choline on the dataset MTBLS132 and glucose on MTBLS131 present similar patterns to the reference. This can be explained by the quality of the peaks detected that translated to the quality of the group selected downstream to represent the corresponding metabolite. A deeper analysis on group selection could allow a better characterization of the changes on each development stage by each metabolite, reproducing patterns identical to the reference.

4.2 Worm (*Caenorhabditis elegans*) Metabolome

4.2.1 Introduction

In the previous case study, the 1D proton NMR was mentioned as the standard for an untargeted approach due to its efficiency which comes from the natural abundant ^{1}H and the highest frequency, leading to higher sensitivity. Another alternative is to consider the nuclei ^{13}C due to its large chemical shift dispersion, ability to detect quaternary carbons and direct measurement of metabolites backbone structure'[222]. The use of this nuclei is hampered by its low sensitivity, that comes from low natural abundance and decreased gyromagnetic ratio γ and in addition to these factors, solutions such as isotopic labelling can be expensive and challenging[222]. One effort that has been made to solve this issue and extend the use of ^{13}C-based approaches is the use of 2D techniques and creation of specific metabolite databases. In example, the database TOCCATA[223] stores information on ^{13}C chemical shifts that allow the identification of metabolites, their spin systems and isomeric states that come from ^{13}C–^{13}C TOCSY experiments of complex mixtures.

One approach that was developed as an alternatative to NMR untargeted metabolomics is using INADEQUATE (incredible natural abundance double quantum transfer experiment) network analysis[224]. This 2D technique records ^{13}C chemical shifts in the direct dimension and ^{13}C–^{13}C double quantum correlations in the indirect dimension, providing carbon correlated networks that allow direct metabolite identification[224]. It has been shown that INADEQUATE-like experiments could be used to obtain direct ^{13}C–^{13}C correlations in complex spectra[225–228] which led to the development of a software package to automatically identify these correlations. The package's name is INETA (INADEQUATE

network analysis) and it was developed by the authors that generated the data for this case study. They tested this package with samples from the endo- and exometabolome of *C.elegans* that were heat-shocked or maintained at room temperature (control).

This case study has the objective to identify metabolites and their changes on the samples from *C.elegans*' endo- and exometabolome submitted to a heat-shock condition. To achieve the results the same processing steps and PCA analysis will be carried out using *specmine*'s functions. The peak detection results obtained using *specmine* were not compared directly with the peak detection done by the authors that lead to metabolite identification. In their work (Clendinen, *et al.*[224]), a new approach to identify metabolites through correlated networks was employed to INADEQUATE spectra. The double quantum frequency in an INADEQUATE experiment allows to establish a double quantum diagonal with slope of 2 that correlates pairs of coupled ^{13}C nuclei. These pairs are symmetric to the diagonal at a given double quantum frequency which allows to establish connections between peaks through INADEQUATE rules. If the peaks satisfy this inequation $|(CS1 + CS2) - \overline{[DQ1 + DQ2]}| < SDT$, where the pairs *(DQ1*, *CS1*) and (*DQ2*, *CS2*) represent the coordinates of the peaks (indirect dimension, direct dimension) and *SDT* is the user defined symmetric/diagonal tolerance, then the peaks are considered double quantum correlated. Vertical chemical shift correlations are satisfied when the difference between the peaks' chemical shifts is less than a user defined threshold. With both correlations (double quantum frequency and chemical shifts) checked, peaks can be connected translating their ^{13}C nuclei into a network that identifies the backbone of a metabolite. The identification process is possible because peaks that form a network match their metabolite 1D ^{13}C spectra.

4.2.2 2D-NMR Data

This study comprises 8 million 99% ^{13}C labeled young adult organisms of *Caenorhabditis elegans* splitted in 8 samples. These 8 samples were equally divided into control and condition. Four replicates were directly stored at room temperature (22ºC) where the other four were heat shocked at 33ºC for 30 minutes and then stored at room temperature. After incubation both endo- and exometabolome were extracted from all replicates and submmited to an INADEQUATE experiment. This originated two datasets, one for each metabolome, where the direct dimension (F2) has a final spectral width of 202 ppm and the indirect dimension (F1) has 404 ppm.

The data to this study is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, https://www.metabolomicsworkbench.org where it has been assigned Project ID PR000095. The data can be accessed directly via it's Project DOI: https://doi.org/10.21228/M8S88T. The work is also supported by NIH grant, U2C- DK119886.

In this case, the developed function to read Varian data had to be changed in terms of the apodization function (exponential to Lorentz-to-Gauss) in order to match the spectral preprocessing done by the authors. Besides this step, after reading the data into R and perform peak detection, four different sets of data were tested to compare results with the authors. Two datasets of endo- and exometabolome were tested to compare results with the authors, one was submitted to probabilistic quotient normalization and to a log scale after peak detection and the other without these preprocessing steps.

4.2.3 Data Summary and Visualization

The datasets from this case study are very large with more than 16 million data points for each sample, as shown in the code below. Each sample is a matrix of 4096 rows and 4096 columns and in the entire endometabolome dataset there is not a single missing value across samples. However, the sample *N2_Control1_INAD* in the exometabolome dataset has 3898 columns instead of 4096. In the author's work they identified one control sample, using PCA analysis, in the exometabolome dataset that was considered an outlier. Since there is not a specification of the control sample by the author's, the removal of the sample *N2_Control1_INAD* before peak detection was not considered. Instead, the integration of this sample in the peak detection phase may lead to the propagation of the peaks that characterize this sample as an outlier in this case study' PCA analysis. Both endo- and exometabolome datasets have zero missing values across samples and standard deviations are high which means the values are not closer to the mean, thus enhancing peak detection.

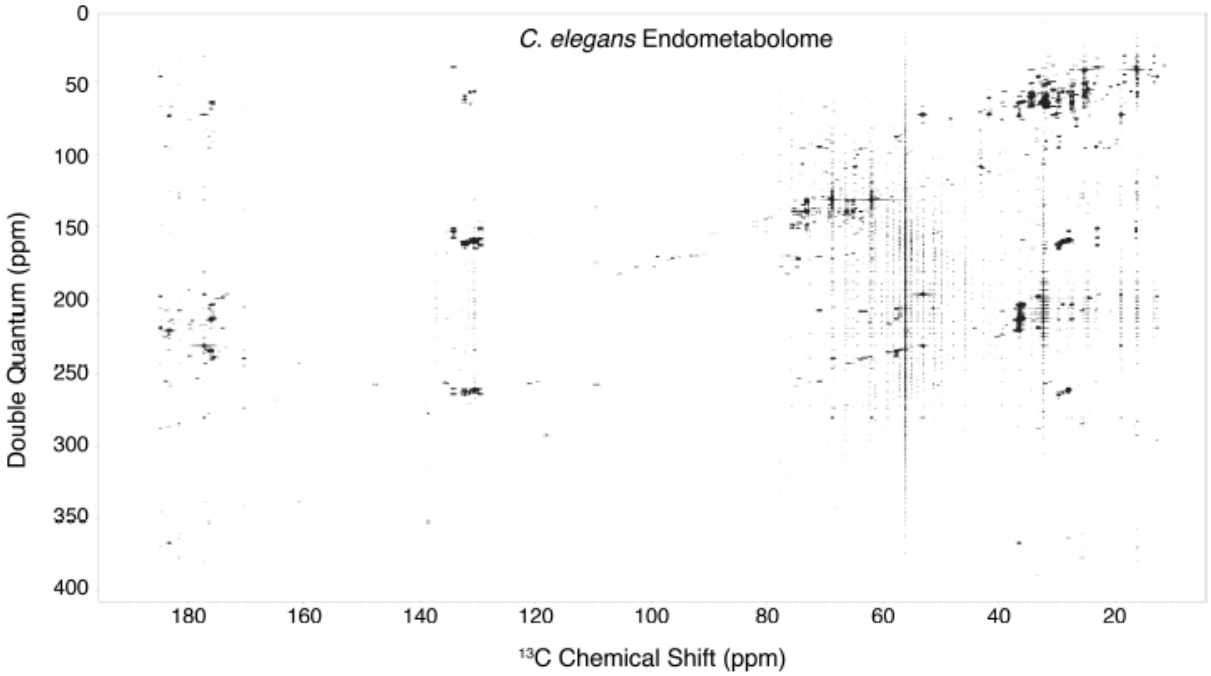
```
Dataset summary:
Valid dataset
Description:  Endometabolome INAD
Type of data:  2d-nmr
Number of samples:  8
Number of data points 16777216
```

```
Number of metadata variables:  1
Label of x-axis values:  ppm
Label of y-axis values:  ppm
Label of pair'(x,y) values:  intensity
Number of missing values in data:
N2_Control_WP1_INAD N2_Control_WP2_INAD N2_Control_WP3_INAD
                  0                   0                   0
N2_Control_WP4_INAD       N2_HS_WP1_INAD
                  0                   0
     N2_HS_WP2_INAD       N2_HS_WP3_INAD       N2_HS_WP4_INAD
                  0                   0                   0
Mean of data values:
N2_Control_WP1_INAD N2_Control_WP2_INAD N2_Control_WP3_INAD
           75369.68            71522.44            75187.47
N2_Control_WP4_INAD       N2_HS_WP1_INAD
           74703.18            81922.96
     N2_HS_WP2_INAD       N2_HS_WP3_INAD       N2_HS_WP4_INAD
           81413.00            70712.09            81845.13
Median of data values:
N2_Control_WP1_INAD N2_Control_WP2_INAD N2_Control_WP3_INAD
           44485.50            50967.11            43562.82
N2_Control_WP4_INAD       N2_HS_WP1_INAD
           40650.65            42538.45
     N2_HS_WP2_INAD       N2_HS_WP3_INAD       N2_HS_WP4_INAD
           45765.93            41042.20            44326.70
Standard deviation:
N2_Control_WP1_INAD N2_Control_WP2_INAD N2_Control_WP3_INAD
           593004.9            272944.5            638784.8
N2_Control_WP4_INAD        N2_HS_WP1_INA
           747510.5            827350.3
     N2_HS_WP2_INAD       N2_HS_WP3_INAD       N2_HS_WP4_INAD
           708588.2            688280.4            771142.3
```

Listing 4.3: Summary of Endometabolome's 2D dataset.

In terms of data visualization, there is not a specific reference sample spectra to compare a specific plot result. Since there is no information on what sample was used, the sample *N2_Control_WP1_INAD* was plotted to compare its INADEQUATE spectra to the one published by the authors, shown on Figure 11.

a



b

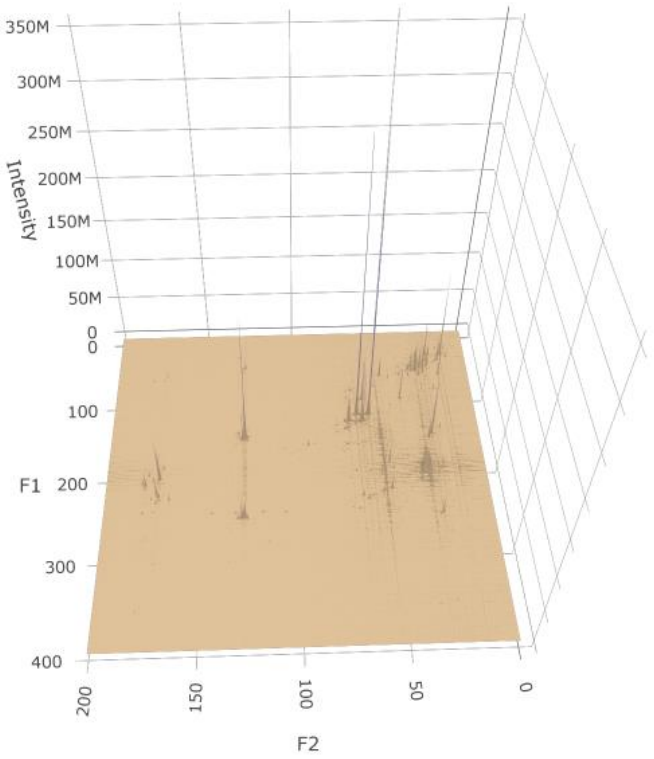


Figure 11: INADEQUATE of one replicate *C. elegans* endometabolome, retrieved from Clendinen, *et al.* (a). Plot of the 2D spectra of sample *N2_Control_WP1_INAD* from the endometabolome dataset, using *specmine* (b).

Visualization of a spetrum with 4096 rows and columns was not possible due to the size of the surface. To achieve the plot presented on Figure 11, half the rows and columns were merged, having the new rows and columns the mean values of their originals. This means that the sample plotted has 2048 rows and columns, however, it is possible to compare the results since spectra are similar. The high values of some peaks hinders the comparison between spectra because the color gradient is applied to all values and not regions. Despite this difficulty, the result obtained allows to identify the same regions of metabolite resonances found on the published spectrum. A darker blue on the obtained plot allows to recognize regions of interest, i.e., 100-150 double quantum ppm and 50-80 ^{13}C chemical shift where the peaks with higher intensity values are located. Considering that the user can zoom in and out with greater image quality than the one presented on Figure 11, the function to plot 2D spectra is validated for both Bruker and Varian datasets.

4.2.4 Reduce dimensionality and Analysis

In this particular case study, the results in this subsection will be limited to the analysis of the datasets after peak detection. The directly comparable results will be the PCA analysis where similar scores plots will validate the peak detection phase. As said in the subsection 4.2.2 this analysis will be done with and without preprocessing on either datasets, however, the results presented on this subsection are from datasets with preprocessing.

Since there is no information on the double quantum frequencies or chemical shifts of the peaks that were picked it is not possible to compare directly the results obtained when detecting peaks through the functions developed for this purpose. Despite the lack of information reagrding which peaks were picked, there is information on thresholds (minimum and maximum) that were used in the process. The mean of these values ($5e7$ and $7e7$) for endo- and exometabolome datasets, respectively, were used as thresholds for peak detection and a noise filter 1 was used for the endo- and exometabolome datasets. A mild filter was applied due to the high number of peaks detected without filter in earlier stages which hinders future analysis. In the example code below it is shown the output of the peak detection function for the endometabolome dataset. A total of 401 peaks were detected for this dataset where as for the exometabolome dataset were detected 908 peaks. The number of peaks is relevant because a higher number of peaks

translates into a larger dataset which usually leads to higher data variance that needs to be explained by each principal component in PCA analysis.

```
> endo_peaks <- peak_detection2d(endometab, baseline_thresh = 5e7, noiseFilt
    = 1)
Sample: N2_Control_WP1_INAD has 38 peaks
Sample: N2_Control_WP2_INAD has 7 peaks
Sample: N2_Control_WP3_INAD has 50 peaks
Sample: N2_Control_WP4_INAD has 61 peaks
Sample: N2_HS_WP1_INAD has 74 peaks
Sample: N2_HS_WP2_INAD has 51 peaks
Sample: N2_HS_WP3_INAD has 59 peaks
Sample: N2_HS_WP4_INAD has 61 peaks
```

Listing 4.4: Example of a 2D peak detection; Endometabolome dataset.

In terms of univariate analysis, t-tests were performed on all variables and the results for both the endo- and exometabolome are presented in Table 5 and Table 6, respectively.

Table 5: t-test results from Endometabolome dataset after peak detection and preprocessing with Temperature metadata.

Combination of ppm (X.F1ppm.F2ppm)	pvalues	logs	fdr
X257.49.77.54	0.023	1.630	0.316
X52.29.77.54	0.024	1.629	0.316
X28.51.18.99	0.024	1.620	0.316
X128.55.68.02	0.024	1.620	0.316
X58.7.23.04	0.024	1.620	0.316
X33.25.23.04	0.024	1.620	0.316
X55.35.29.2	0.024	1.620	0.316
X60.18.27.82	0.024	1.620	0.316
X59.69.23.04	0.024	1.620	0.316
X56.33.79.17	0.024	1.620	0.316
X51.99.77.54	0.024	1.620	0.316
X56.23.79.17	0.024	1.620	0.316

The tables allow to compare the different peaks selected for each metabolome. The main difference is that peaks detected on the endometabolome dataset have statiscally different means regarding the Temperature metadata variable. Only one peak (36 x 58 ppm) from the exometabolome dataset was able to achieve the same result. In Table 5 it

Table 6: t-test results from Exometabolome dataset after peak detection and preprocessing with Temperature metadata.

Combination of ppm (X.F1ppm.F2ppm)	pvalues	logs	fdr
X36.5.58.8	2.289e-10	9.640	1.513e-07
X25.75.12.43	1.340e-01	0.873	4.490e-01
X36.5.58.9	1.340e-01	0.873	4.490e-01
X25.75.12.53	1.340e-01	0.873	4.490e-01
X31.37.190.23	1.340e-01	0.873	4.490e-01
X264.8.133.73	1.340e-01	0.873	4.490e-01
X25.95.128.99	1.340e-01	0.873	4.490e-01
X264.89.188.57	1.340e-01	0.873	4.490e-01
X25.95.116.12	1.340e-01	0.873	4.490e-01
X161.11.167.86	1.340e-01	0.873	4.490e-01
X25.95.129.09	1.340e-01	0.873	4.490e-01
X19.73.92.1	1.340e-01	0.873	4.490e-01

is possible to consider the regions between 30-60 and ppm on indirect dimension and 14-30 ppm on direct dimension as set of variables that could discriminate the heat shock condition on *C.elegans*. It is important to note that although Table 6 does not highlight statistically significant peaks, it does not follow the same exact peak region as Table 5, which could mean that different metabolites are present on different metabolomes under a specific condition. On the other hand, similar peaks detected on both datasets allow to study the metabolic changes of an organism under an environmental condition, by comparing to it's default behaviour (control).

In terms of PCA analysis, three results were obtained that compare the scores plots (PC1 and PC2) from the endo- and exometabolome. These plots are shown on Figure 12 and Figure 13, respectively.

In Figure 12 **(b)** it was not possible to observe a good separation along the PC1 for all samples as the authors obtained. Nonetheless, the majority of the samples can be separated along the PC1 following the same distribution regarding to Temperature metadata, i.e. most control samples have negative values on PC1 axis where heat shock samples have positive values. This means that the heat shock condition has an effect on *C.elegans*' endometabolome. The variation explained by each PC is similar to the one obtained by the authors, a difference of 2-3%, which suggests that the eigenvalues for each PC are identical. Despite the lack of information regarding which peaks were picked by the authors' pipeline, the results obtained after peak detection suggest that this step selected relevant information for future analysis in the endometabolome dataset.

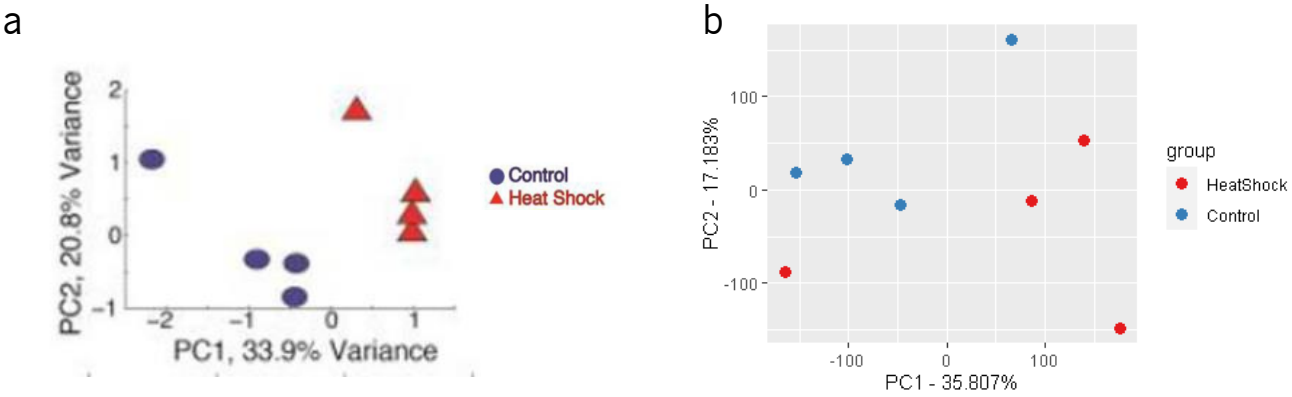


Figure 12: **(a)** Scores plot from the PCA analysis done by Clendinen, *et al*, on the endometabolome data. **b** Scores plot from the PCA analysis for the endometabolome dataset using *specmine* after peak detection and preprocessing.

The other two samples that do not follow the values for the first principal component could be a result of differences in spectral preprocessing, the hierarchical alignment of 2D spectra[229] performed by the authors and the quality of the peaks detected for those samples.

As it is shown on Figure 13 **(b)** there is no good separation along PC1 as it happens on the first plot of the figure. There is no possible separation between conditions which indicates that the exometabolome is not affected heat shock. However, this scores plot was done using an exometabolome dataset with an outlier. According to the authors this outlier was identified in their PCA analysis and Figure 13 **(b)** shows that there is one control sample with the highest positive value along PC1, opposing the negative values found in other control samples. This indicates that the outlier spectra propagated through *specmine* analysis and was also identifiable in the PCA analysis, validating

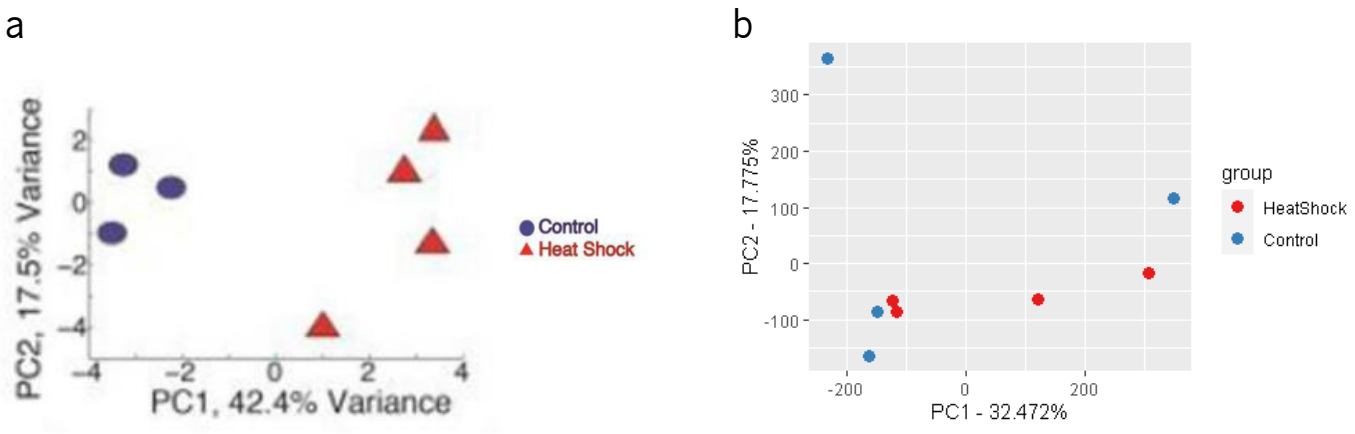


Figure 13: **(a)** Scores plot from the PCA analysis done by Clendinen, *et al*, on the exometabolome data. **b** Scores plot from the PCA analysis for the exometabolome dataset using *specmine* after peak detection and preprocessing.

the functions developed. The sample was the second replicate *N2_Control2_INAD* and presented the highest number of peaks detected in the this dataset (206). Figure 14 shows the scores plot (PC1 and PC2) from the exometabolome dataset after removing the second control replicate before peak detection. The results are different from the first scores plot, however, the clear separation on Figure 13 **(a)** is not achieved. In fact, two samples from heat shock condition have negative loadings for PC1 which indicate similarity with control samples. Therefore it is not possible to state through this plot that the heat shock condition has an effet on *C.elegans*' exometabolome. Despite this result, the peak detection on this dataset was validated. As it is shown on Figure 14, all control samples have negative loadings for PC1 and loadings for PC2 are similar to the ones shown in Figure 13 **(a)**. The other heat shock replicates have also correct loadings for PC1 and similar loadings for PC2. The variance explained by the first two principal

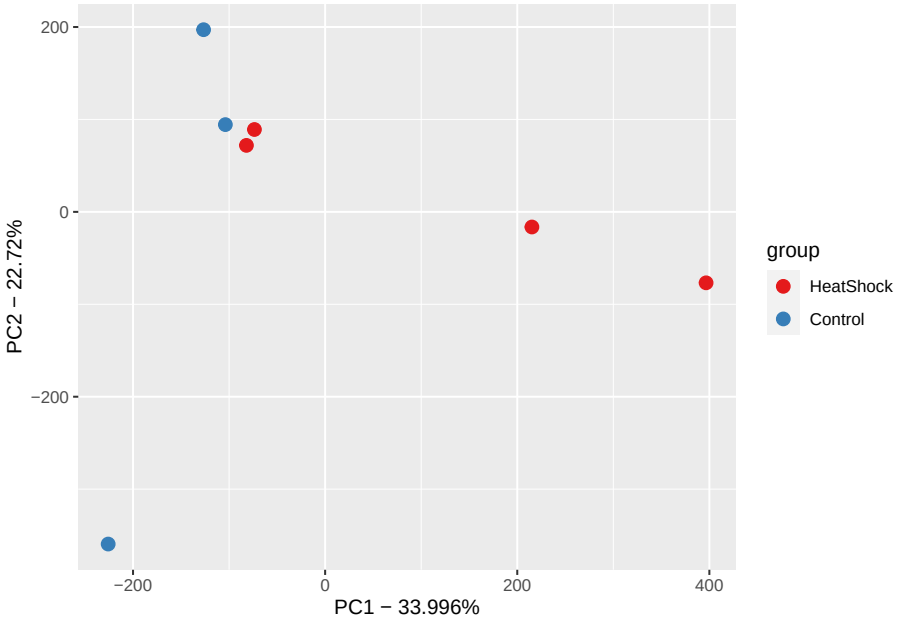


Figure 14: Scores plot from the PCA analysis for the exometabolome dataset without second replicate, using *specmine* after peak detection and preprocessing.

components is identical to the authors' work, however, PC1 explains less 8% variance which overall does not affect the outcome of those two specific heat shock replicates along PC1 axis due to the same reasons explained in the previous paragraph.

5

CONCLUSION AND FUTURE WORK

In order to update the functionality and capability of the R package *specmine* towards metabolomics, new functions were developed to provide tools for a new type of data, 2D-NMR. Since 1D-NMR lacks the sensitivity to treat overlapping resonances on more complex samples, 2D-NMR has been applied and adapted to provide easier to interpret and more informative data. This led to the development of key functions that enable 2D-NMR analysis with *specmine*, supporting the purpose of providing tools for metabolomic data analysis in a complete and user-friendly environment.

The functions developed start with reading data into a new *specmine* structure (list of matrices), following by initial assessment of the data (missing values, data points, number of metadata variables, ...) and data visualization. In this work, the user is capable of visualizing one or multiple spectra in a 3D interactive environment, presenting a novel feature for 2D-NMR analysis with R. A peak detection method allows to reduce the size of the data, changing it to a standard 1D structure with relevant information enabling further analysis using *specmine*'s functionalities. At this point the user can apply different methods of preprocessing and data analysis, such as, univariate, clustering or PCA. The package flexibility was ensured through easy-to-use functions with configurations that can be changed according to the user input. Validation of the tools was done using real-world case studies, where results from specific pipelines were compared to the ones presented on this dissertation, thus helping to find where to improve the methods developed.

This new feature implemented on *specmine* will enhance its flexibility towards metabolomics since 2D-NMR is relatively new and with scope for growth. Therefore, *specmine* has also scope for growth in this area, giving its users an up-to-date tool they can use with or without informatics background. Extending its functionalities while

maintaining easy-to-interpret data structures helps researchers link data from different experiments without the usual need for multiple packages.

The most recent version of *specmine* available on CRAN will have the developed functionalities as well as a vignette which provides a pipeline on 2D-NMR analysis with information on each step of this process. Unfortunately the vignette does not show outputs due to data storage issues.

For future work, there are different paths in which *specmine* can be improved regarding 2D metabolomics, either by improving the functions developed, adding new features to analyze 2D spectra or integrating the functions in the web version, *WebSpecmine*. The future work includes:

- Support to more types of 2D metabolomics data, i.e., 2D MS;
- If this support is given, considering a new package (based on *specmine*) is a good idea focusing only on 2D metabolomics data. The package would integrate the functions already developed for NMR and the new ones for MS;
- Implement preprocessing methods specific for 2D spectra which takes into account the properties of variables in this type of data. These methods should have the flexibility to be applied on specific samples, according to user input;
- Add more visualization methods for data exploration and reproducibility, i.e. 2D representation (on an xOy cartesian plane) of a sample;
- Improve/Optimize the peak detection method since it can take a long time for large datasets and an increase in efficiency eases computational demand;
- Integrate the fundamentals of the work developed with *WebSpecmine*, i.e., data reading, visualization and peak detection;

6

BIBLIOGRAPHY

[1] Gil, A. M., Duarte, I. F., Godejohann, M., Braumann, U., Maraschin, M., & Spraul, M. (2003). Characterization of the aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection. *Analytica Chimica Acta*, *488*(1), 35–51.

[2] Costa, C., Maraschin, M., & Rocha, M. (2016). An R package for the integrated analysis of metabolomics and spectral data. *Computer Methods and Programs in Biomedicine*, *129*, 117–124.

[3] Cardoso, S., Afonso, T., Maraschin, M., & Rocha, M. (2019). WebSpecmine: A website for metabolomics data analysis and mining. *Metabolites*, *9*(10), 237.

[4] Jézéquel, T., Deborde, C., Maucourt, M., Zhendre, V., Moing, A., & Giraudeau, P. (2015). Absolute quantification of metabolites in tomato fruit extracts by fast 2D NMR. *Metabolomics*, *11*(5), 1231–1242.

[5] Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, *13*(4), 263–269.

[6] Maraschin, M., R. Somensi Zeggio, A., M. Tomazzoli, M., K. Oliveira, S., Ramlov, F., Beatriz Veleirinho, M., & Rocha, M. (2017). Metabolómica e Quimiometria como Ferramentas para Análises Quimio(bio) diversas. In *Biotecnologia aplicada à agroindústria* (pp. 18–43).

[7] Kim, S. J., Kim, S. H., Kim, J. H., Hwang, S., & Yoo, H. J. (2016). Understanding metabolomics in biomedical research. *Endocrinology and metabolism (Seoul, Korea)*, *31*(1), 7–16.

[8] Cevallos-Cevallos, J. M., Reyes-De-Corcuera, J. I., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science: a review. *Trends in Food Science and Technology*, *20*(11-12), 557–566.

[9] Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., ... Wishart, D. S. (2019). NMR spectroscopy for metabolomics research. *Metabolites*, *9*(7), 123.

[10] Lamichhane, S., Sen, P., Dickens, A. M., Hyötyläinen, T., & Orešič, M. (2018). An overview of metabolomics data analysis: Current tools and future perspectives. *Comprehensive Analytical Chemistry*, *82*, 387–413.

[11] Hanson, B. A. [Bryan A.]. (2020). *Chemospec2d: Exploratory chemometrics for 2d spectroscopy*. R package version 0.4.176.

[12] Lewis, I. A., Schommer, S. C., & Markley, J. L. (2009). rNMR: Open source software for identifying and quantifying metabolites in NMR spectra. *Magnetic Resonance in Chemistry*, *47*(SUPPL. 1), S123.

[13] Xia, J., Bjorndahl, T. C., Tang, P., & Wishart, D. S. (2008). MetaboMiner - Semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, *9*(1), 507.

[14] Chong, J., & Xia, J. (2018a). MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics*, *34*(24), 4313–4314.

[15] Darbeau, R. (2006). Nuclear magnetic resonance (NMR) spectroscopy: A review and a look at its use as a probative tool in deamination chemistry. *Applied Spectroscopy Reviews*, *41*(4), 401–425.

[16] Hatzakis, E. (2019). Nuclear Magnetic Resonance (NMR) spectroscopy in food science: A c omprehensive review. *Comprehensive Reviews in Food Science and Food Safety*, *18*(1), 189–220.

[17] Levitt, M. H. (2008). *Spin dynamics*. Chichester: John Wiley & Sons.

[18] Keifer, P. A. [Paul A.]. (2013). Flow NMR. In *Encyclopedia of biophysics* (pp. 778–779).

[19] Keifer, P. A. [P. A.]. (2007). Flow techniques in NMR spectroscopy. In *Annual reports on nmr spectroscopy* (Vol. 62, pp. 1–47).

[20] Strohschein, S., Rentel, C., Lacker, T., Bayer, E., & Albert, K. (1999). Separation and identification of tocotrienol isomers by HPLC-MS and HPLC-NMR coupling. *Analytical Chemistry*, *71*(9), 1780–1785.

[21] Takis, P. G., Ghini, V., Tenori, L., Turano, P., & Luchinat, C. (2019). Uniqueness of the NMR approach to metabolomics. *TrAC - Trends in Analytical Chemistry*, *120*.

[22] Mahrous, E. A., & Farag, M. A. (2015). Two dimensional NMR spectroscopic approaches for exploring plant metabolome: A review. *Journal of Advanced Research*, *6*(1), 3–15.

[23] Nagayama, K., Wüthrich, K., Bachmann, P., & Ernst, R. R. (1977). Two-dimensional J-resolved 1H n.m.r. spectroscopy for studies of biological macromolecules. *Biochemical and Biophysical Research Communications*, *78*(1), 99–105.

[24] Ludwig, C., & Viant, M. R. (2010). Two-dimensional J-resolved NMR spectroscopy: Review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis*, *21*(1), 22–32.

[25] Aue, W. P., Karhan, J., & Ernst, R. R. (1976). Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy. *The Journal of Chemical Physics*, *64*(10), 4226–4227.

[26] Frydman, L., Scherf, T., & Lupulescu, A. (2002). The acquisition of multidimensional NMR spectra within a single scan. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(25), 15858–15862.

[27] Viant, M. R. (2003). Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochemical and Biophysical Research Communications*, *310*(3), 943–948.

[28] Foxall, P. J., Parkinson, J. A., Sadler, I. H., Lindon, J. C., & Nicholson, J. K. (1993). Analysis of biological fluids using 600 MHz proton NMR spectroscopy: application of homonuclear two-dimensional J-resolved spectroscopy to urine and blood plasma for spectral simplification and assignment. *Journal of Pharmaceutical and Biomedical Analysis*, *11*(1), 21–31.

[29] Lutz, N. W., Maillet, S., Nicoli, F., Viout, P., & Cozzone, P. J. (1998). Further assignment of resonances in 1H NMR spectra of cerebrospinal fluid (CSF). *FEBS Letters*, *425*(2), 345–351.

[30] Kim, H. K., Saifullah, Khan, S., Wilson, E. G., Kricun, S. D., Meissner, A., ... Verpoorte, R. (2010). Metabolic classification of South American Ilex species by NMR-based metabolomics. *Phytochemistry*, *71*(7), 773–784.

[31] Aguilar, J. A., Adams, R. W., Nilsson, M., & Morris, G. A. (2014). Suppressing exchange effects in diffusion-ordered NMR spectroscopy. *Journal of Magnetic Resonance*, *238*, 16–19.

[32] Čuperlović-Culf, M. (2013). Experimental methodology. In *Nmr metabolomics in cancer research* (Chap. 3, pp. 193–213).

[33] Khera, S., Grillo, M., Schnier, P., & Hollis, S. (2010). Application of diffusion-edited NMR spectroscopy for the structural characterization of drug metabolites in mixtures. *Journal of Pharmaceutical and Biomedical Analysis*, *51*(1), 164–169.

[34] Novoa-Carballal, R., Fernandez-Megia, E., Jimenez, C., & Riguera, R. (2011). NMR methods for unravelling the spectra of complex mixtures. *Natural Product Reports*, *28*(1), 78–98.

[35] Sobolev, A. P., Brosio, E., Gianferri, R., & Segre, A. L. (2005). Metabolic profile of lettuce leaves by high-field NMR spectra. *Magnetic Resonance in Chemistry*, *43*(8), 625–638.

[36] Balayssac, S., Trefi, S., Gilard, V., Malet-Martino, M., Martino, R., & Delsuc, M. A. (2009). 2D and 3D DOSY 1H NMR, a useful tool for analysis of complex mixtures: Application to herbal drugs or dietary supplements for erectile dysfunction. *Journal of Pharmaceutical and Biomedical Analysis*, *50*(4), 602–612.

[37] Lindon, J. C. (2016). Nuclear magnetic resonance spectroscopy — Multidimensional NMR spectroscopy. In *Encyclopedia of analytical science* (pp. 248–256).

[38] Kim, H. K., Choi, Y. H., & Verpoorte, R. (2010). NMR-based metabolomic analysis of plants. *Nature Protocols*, *5*(3), 536–549.

[39] Le Guennec, A., Tea, I., Antheaume, I., Martineau, E., Charrier, B., Pathan, M., ... Giraudeau, P. (2012). Fast determination of absolute metabolite concentrations by spatially encoded 2D NMR: Application to breast cancer cell extracts. *Analytical Chemistry*, *84*(24), 10831–10837.

[40] Sandusky, P., & Raftery, D. (2005a). Use of semiselective TOCSY and the Pearson correlation for the metabonomic analysis of biofluid mixtures: Application to urine. *Analytical Chemistry*, *77*(23), 7717–7723.

[41] Lucio-Gutiérrez, J. R., Delgado-Montemayor, C., Coello-Bonilla, J., Waksman-Minsky, N., & Saucedo, A. L. (2019). Selective 1D-TOCSY and chemometrics to evaluate authenticity of Turnera diffusa and related botanical extracts. *Phytochemistry Letters*, *30*, 62–68.

[42] Zhang, F., Robinette, S. L., Bruschweiler-Li, L., & Brüschweiler, R. (2009). Web server suite for complex mixture analysis by covariance NMR. *Magnetic Resonance in Chemistry*, *47*(SUPPL. 1), S118.

[43] Bingol, K., Li, D. W., Zhang, B., & Brüschweiler, R. (2016). Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server. *Analytical Chemistry*, *88*(24), 12411–12418.

[44] Öman, T., Tessem, M. B., Bathen, T. F., Bertilsson, H., Angelsen, A., Hedenström, M., & Andreassen, T. (2014). Identification of metabolites from 2D H-C HSQC NMR using peak correlation plots. *BMC Bioinformatics*, *15*(1).

[45] Yi, Q., Scalley-Kim, M. L., Alm, E. J., & Baker, D. (2000). NMR characterization of residual structure in the denatured state of protein L. *Journal of Molecular Biology*, *299*(5), 1341–1351.

[46] Lee, S. H., Cha, E. J., Lim, J. E., Kwon, S. H., Kim, D. H., Cho, H., & Han, K. H. (2012). Structural characterization of an intrinsically unfolded mini-HBX protein from hepatitis B virus. *Molecules and Cells*, *34*(2), 165–169.

[47] Lewis, I. A., Schommer, S. C., Hodis, B., Robb, K. A., Tonelli, M., Westler, W. M., ... Markley, J. L. (2007). Method for determining molar concentrations of metabolites in complex solutions from two-dimensional 1H-13C NMR spectra. *Analytical Chemistry*, *79*(24), 9385–9390.

[48] Schanda, P., & Brutscher, B. (2005). Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. *Journal of the American Chemical Society*, *127*(22), 8014–8015.

[49] Schanda, P., Kupĉe, E., & Brutscher, B. (2005). SOFAST-HMQC experiments for recording two-dimensional deteronuclear correlation spectra of proteins within a few seconds. *Journal of Biomolecular NMR*, *33*(4), 199–211.

[50] Liang, Y. S., Kim, H. K., Lefeber, A. W., Erkelens, C., Choi, Y. H., & Verpoorte, R. (2006). Identification of phenylpropanoids in methyl jasmonate treated Brassica rapa leaves using two-dimensional nuclear magnetic resonance spectroscopy. *Journal of Chromatography A*, *1112*(1-2), 148–155.

[51] Dona, A. C., Kyriakides, M., Scott, F., Shephard, E. A., Varshavi, D., Veselkov, K., & Everett, J. R. (2016). A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Computational and Structural Biotechnology Journal*, *14*, 135–153.

[52] Bernini, P., Bertini, I., Luchinat, C., Nepi, S., Saccenti, E., Schäfer, H., ... Tenori, L. (2009). Individual human phenotypes in metabolic space and time. *Journal of Proteome Research*, *8*(9), 4264–4271.

[53] Hwang, T. L., Ronk, M., & Milne, J. E. (2013). Application of two-dimensional selective-TOCSY HMBC for structure elucidation of impurities in mixture without separation. *Magnetic Resonance in Chemistry*, *51*(2), 89–94.

[54] Bingol, K., & Brüschweiler, R. (2014). Multidimensional APPROACHES to NMR-based metabolomics. *Analytical Chemistry*, *86*(1), 47–57.

[55] Powers, R., & Riekeberg, E. (2017). New frontiers in metabolomics: From measurement to insight. *F1000Research*, *6*.

[56] Ardenkjær-Larsen, J. H., Fridlund, B., Gram, A., Hansson, G., Hansson, L., Lerche, M. H., ... Golman, K. (2003). Increase in signal-to-noise ratio of $10,000 times in liquid-state NMR$. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(18), 10158–10163.

[57] Jaravine, V. A., Zhuravleva, A. V., Permi, P., Ibraghimov, I., & Orekhov, V. Y. (2008). Hyperdimensional NMR spectroscopy with nonlinear sampling. *Journal of the American Chemical Society*, *130*(12), 3927–3936.

[58] Kupče, E., & Freeman, R. (2003). Fast multi-dimensional Hadamard spectroscopy. *Journal of Magnetic Resonance*, *163*(1), 56–63.

[59] Giraudeau, P., & Frydman, L. (2014). Ultrafast 2D NMR: An emerging tool in analytical spectroscopy. *Annual Review of Analytical Chemistry*, *7*(1), 129–161.

[60] Blakebrough-Hall, C., Dona, A., D'occhio, M. J., McMeniman, J., & González, L. A. (2020). Diagnosis of bovine respiratory disease in feedlot cattle using blood 1H NMR metabolomics. *Scientific Reports*, *10*(115).

[61] Kim, S. Y., Kim, E. B., Shin, B. K., Seo, J. A., Kim, Y. S., Lee, D. Y., & Choi, H. K. (2020). NMR-based metabolic profiling discriminates the geographical origin of raw sesame seeds. *Food Control*, *112*.

[62] Wang, Y., Yang, Y., Pan, D., He, J., Cao, J., Wang, H., & Ertbjerg, P. (2020). Metabolite profile based on 1H NMR of broiler chicken breasts affected by wooden breast myodegeneration. *Food Chemistry*, *310*.

[63] Yan, S., Wang, D., Teng, M., Meng, Z., Yan, J., Li, R., ... Zhu, W. (2020). Perinatal exposure to 2-Ethylhexyl Diphenyl Phosphate (EHDPHP) affected the metabolic homeostasis of male mouse offspring: Unexpected findings help to explain dose- and diet- specific phenomena. *Journal of Hazardous Materials*, *388*.

[64] Yuan, P., Dong, M., Lei, H., Xu, G., Chen, G., Song, Y., ... Zhang, L. (2020). Targeted metabolomics reveals that 2,3,7,8-tetrachlorodibenzofuran exposure induces hepatic steatosis in male mice. *Environmental Pollution*, *259*.

[65] Chen, H., Zhang, F., Li, R., Liu, Y., Wang, X., Zhang, X., … Yao, Q. (2020). Berberine regulates fecal metabolites to ameliorate 5-fluorouracil induced intestinal mucositis through modulating gut microbiota. *Biomedicine and Pharmacotherapy*, *124*.

[66] Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102–4107.

[67] Castillo, S., Gopalacharyulu, P., Yetukuri, L., & Ore, M. (2011). Algorithms and tools for the preprocessing of LC – MS metabolomics data. *Chemometrics and Intelligent Laboratory Systems*, *108*, 23–32.

[68] McKnight, P. E., McKnight, K. M., Sidani, S., & Figueiredo, A. J. (2007). *Missing data : a gentle introduction*. Guilford Press.

[69] Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd Edition*. New York: John Wiley & Sons.

[70] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data* (Fisrt). Chapman & Hall.

[71] Kapil, A. R. (2018). Methods of missing value treatment and their effect on the accuracy of classification models. (October).

[72] Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487–508.

[73] Harrell, F. E. (2001). *Regression Modeling Strategies* (1st ed.).

[74] Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. *Frontiers in Artificial Intelligence and Applications*, *87*(June), 251–260.

[75] Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. *Proceedings of the Meeting of the International Federation of Classification Societies (IFCS)*.

[76] van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd). Chapman and Hall/CRC.

[77] Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

[78] Morris, G. A. (2017). NMR data processing. In *Encyclopedia of spectroscopy and spectrometry* (3rd, pp. 125–133). New York: Elsevier.

[79] Wehrens, R., & Salek, R. (2019). *Metabolomics: Practical Guide to Design and Analysis*. Chapman and Hall/CRC Computational Biology Series.

[80] Emwas, A. H., Saccenti, E., Gao, X., McKay, R. T., dos Santos, V. A., Roy, R., & Wishart, D. S. (2018). Recommended strategies for spectral processing and post-processing of 1D 1 H-NMR data of biofluids with a particular focus on urine. *Metabolomics*, *14*(3), 1–23.

[81] Pluskal, T., Castillo, S., Villar-Briones, A., & Orešič, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, *11*(1), 395.

[82] Alonso, A., Marsal, S., & Julià, A. (2015). Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, *3*(MAR), 23.

[83] Yang, C., He, Z., & Yu, W. (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics*, *10*(1), 4.

[84] O'haver, T. (2020). *Pragmatic Introduction to Signal Processing Applications in scientific measurement*.

[85] Gragido, W., Pirc, J., Selby, N., & Molina, D. (2013). Signal-to-Noise Ratio. In *Blackhatonomics* (pp. 45–55).

[86] Du, P., Kibbe, W. A., & Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, *22*(17), 2059–2065.

[87] Hao, J., Astle, W., De Iorio, M., & Ebbels, T. M. D. (2012). BATMAN-an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, *28*(15), 2088–2090.

[88] Weljie, A. M., Newton, J., Mercier, P., Carlson, E., & Slupsky, C. M. (2006). Targeted pofiling: Quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry*, *78*(13), 4430–4442.

[89] Burton, L., Ivosev, G., Tate, S., Impey, G., Wingate, J., & Bonner, R. (2008). Instrumental and experimental effects in LC-MS-based metabolomics. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *871*(2), 227–235.

[90] Tomasi, G., van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, *18*(5), 231–241.

[91] Savorani, F., Tomasi, G., & Engelsen, S. B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, *202*(2), 190–202.

[92] Veselkov, K. A., Lindon, J. C., Ebbels, T. M., Crockford, D., Volynkin, V. V., Holmes, E., ... Nicholson, J. K. (2009). Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Analytical Chemistry*, *81*(1), 56–66.

[93] Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, *84*(11), 5035–5039.

[94] Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., & Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, *78*(7), 2262–2267.

[95] De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., ... Speed, T. P. (2012). Normalizing and integrating metabolomics data. *Analytical Chemistry*, *84*(24), 10768–10776.

[96] Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R., & Gronwald, W. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, *8*(S1), 146–160.

[97] Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., ... Zhu, F. (2017). NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Research*, *45*.

[98] De Livera, A. M., Sysi-Aho, M., Jacob, L., Gagnon-Bartsch, J. A., Castillo, S., Simpson, J. A., & Speed, T. P. (2015). Statistical methods for handling unwanted variation in metabolomics data. *Analytical Chemistry*, *87*(7), 3606–3615.

[99] Zheng, C., Zhang, S., Ragg, S., Raftery, D., & Vitek, O. (2011). Identification and quantification of metabolites in 1H NMR spectra by Bayesian model selection. *Bioinformatics*, *27*(12), 1637–1644.

[100] Wishart, D. S. [David S.]. (2008). Quantitative metabolomics using NMR. *TrAC - Trends in Analytical Chemistry*, *27*(3), 228–237.

[101] Wishart, D. S. [David S.]. (2009). Computational strategies for metabolite identification in metabolomics. *Bioanalysis*, *1*(9), 1579–1596.

[102] Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., ... Wishart, D. S. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLOS ONE*, *10*(5), e0124219.

[103] Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted metabolomics strategies — Challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, *27*(12), 1897–1905.

[104] Spraul, M., Neidig, P., Klauck, U., Kessler, P., Holmes, E., Nicholson, J. K., ... Lindon, J. C. (1994). Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *Journal of Pharmaceutical and Biomedical Analysis*, *12*(10), 1215–1225.

[105] Padayachee, T., Khamiakova, T., Louis, E., Adriaensens, P., & Burzykowski, T. (2019). The impact of the method of extracting metabolic signal from 1H-NMR data on the classification of samples: A case study of binning and BATMAN in lung cancer. *PLOS ONE*, *14*(2), e0211854.

[106] Astle, W., De Iorio, M., Richardson, S., Stephens, D., & Ebbels, T. (2012). A Bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, *107*(500), 1259–1271. arXiv: 1105.2204

[107] Ochs, M. F., Stoyanova, R. S., Arias-Mendoza, F., & Brown, T. R. (1999). A new method for spectral decomposition using a bilinear bayesian approach. *Journal of Magnetic Resonance*, *137*(1), 161–176.

[108] Stoyanova, R., Nicholson, J. K., Lindon, J. C., & Brown, T. R. (2004). Sample classification based on Bayesian spectral decomposition of metabonomic NMR data sets. *Analytical Chemistry*, *76*(13), 3666–3674.

[109] Eads, C. D., Furnish, C. M., Noda, I., Juhlin, K. D., Cooper, D. A., & Morrall, S. W. (2004). Molecular factor analysis applied to collections of NMR spectra. *Analytical Chemistry*, *76*(7), 1982–1990.

[110] Soininen, P., Haarala, J., Vepsäläinen, J., Niemitz, M., & Laatikainen, R. (2005). Strategies for organic impurity quantification by 1H NMR spectroscopy: Constrained total-line-shape fitting. *Analytica Chimica Acta*, *542*(2), 178–185.

[111] Laatikainen, R., Niemitz, M., Malaisse, W. J., Biesemans, M., & Willem, R. (1996). A computational strategy for the deconvolution of NMR spectra with multiplet structures and constraints: Analysis of overlapping 13C-2H multiplets of 13C en-

riched metabolites from cell suspensions incubated in deuterated media. *Magnetic Resonance in Medicine*, *36*(3), 359–365.

[112] Bingol, K. (2018). Recent advances in targeted and untargeted metabolomics by NMR and MS/NMR methods. *High-Throughput*, *7*(2).

[113] Wishart, D. S. [David S], Querengesser, L. M. M., Lefebvre, B. A., Epstein, N. A., Greiner, R., & Newton, J. B. (2001). Magnetic resonance diagnostics: A new technology for high-throughput clinical diagnostics. *Clinical Chemistry*, *47*(10), 1918–1921.

[114] Baxevanis, A., Bader, G., & Wishart, D. (2020). *Bioinformatics*. Wiley.

[115] Kostidis, S., Addie, R. D., Morreau, H., Mayboroda, O. A., & Giera, M. (2017). *Quantitative NMR analysis of intra- and extracellular metabolism of mammalian cells: A tutorial*.

[116] Wishart, D. S. [David S.], Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., . . . Forsythei, I. (2009). HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, *37*(SUPPL. 1).

[117] Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J. G., & Ebbels, T. M. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature Protocols*, *9*(6), 1416–1427.

[118] Behrends, V., Bell, T. J., Liebeke, M., Cordes-Blauert, A., Ashraf, S. N., Nair, C., . . . Bundy, J. G. (2013). Metabolite profiling to characterize disease-related bacteria: Gluconate excretion by pseudomonas aeruginosa mutants and clinical isolates from cystic fibrosis patients. *Journal of Biological Chemistry*, *288*(21), 15098–15109.

[119] Gibbs, J. W. (2010). *Elementary Principles in Statistical Mechanics : Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Cambridge University Press.

[120] Doucet, A., Freitas, N., & Gordon, N. (2001). An Introduction to Sequential Monte Carlo Methods. In *Sequential monte carlo methods in practice* (pp. 3–14).

[121] Tardivel, P. J., Canlet, C., Lefort, G., Tremblay-Franco, M., Debrauwer, L., Concordet, D., & Servien, R. (2017). ASICS: an automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. *Metabolomics*, *13*(10), 1–9.

[122] Cañueto, D., Gómez, J., Salek, R. M., Correig, X., & Cañellas, N. (2018). rDolphin: a GUI R package for proficient automatic profiling of 1D 1 H-NMR spectra of study datasets. *Metabolomics*, *14*(3), 1–5.

[123] Röhnisch, H. E., Eriksson, J., Müllner, E., Agback, P., Sandström, C., & Moazzami, A. A. (2018). AQuA: An Automated Quantification Algorithm for high-throughput NMR-based metabolomics and its application in human plasma. *Analytical Chemistry*, *90*(3), 2095–2102.

[124] Bingol, K., Zhang, F., Bruschweiler-Li, L., & Brüschweiler, R. (2013). Quantitative analysis of metabolic mixtures by two-dimensional 13-C constant-time TOCSY NMR spectroscopy. *Analytical Chemistry*, *85*(13), 6414–6420.

[125] Marchand, J., Martineau, E., Guitton, Y., Dervilly-Pinel, G., & Giraudeau, P. (2017). *Multidimensional NMR approaches towards highly resolved, sensitive and high-throughput quantitative metabolomics*.

[126] Sandusky, P., & Raftery, D. (2005b). Use of selective TOCSY NMR experiments for quantifying minor components in complex mixtures: Application to the metabonomics of amino acids in honey. *Analytical Chemistry*, *77*(8), 2455–2463.

[127] Wedeking, R., Maucourt, M., Deborde, C., Moing, A., Gibon, Y., Goldbach, H. E., & Wimmer, M. A. (2018). 1H-NMR metabolomic profiling reveals a distinct metabolic recovery response in shoots and roots of temporarily drought-stressed sugar beets. *PLoS ONE*, *13*(5), 1–21.

[128] Shimamoto, G. G., Bianchessi, L. F., & Tubino, M. (2017). Alternative method to quantify biodiesel and vegetable oil in diesel-biodiesel blends through 1H NMR spectroscopy. *Talanta*, *168*(March), 121–125.

[129] Martineau, E., Tea, I., Akoka, S., & Giraudeau, P. (2012). Absolute quantification of metabolites in breast cancer cell extracts by quantitative 2D 1H INADEQUATE NMR. *NMR in Biomedicine*, *25*(8), 985–992.

[130] Hu, K., Westler, W. M., & Markley, J. L. (2011). Simultaneous quantification and identification of individual chemicals in metabolite mixtures by two-dimensional extrapolated time-zero 1 H-13 C HSQC (HSQC 0). *Journal of the American Chemical Society*, *133*, 5.

[131] Halouska, S., Fenton, R. J., Zinniel, D. K., Marshall, D. D., Barletta, R. G., & Powers, R. (2014). Metabolomics analysis identifies d-alanine-d-alanine ligase as the primary lethal target of d-cycloserine in mycobacteria. *Journal of Proteome Research*, *13*(2), 1065–1076.

[132] Mauve, C., Khlifi, S., Gilard, F., Mouille, G., & Farjon, J. (2016). Sensitive, highly resolved, and quantitative 1H-13C NMR data in one go for tracking metabolites in vegetal extracts. *Chemical Communications*, *52*(36), 6142–6145.

[133] Farjon, J., Milande, C., Martineau, E., Akoka, S., & Giraudeau, P. (2018). The FAQUIRE Approach: FAst, QUantitative, hIghly Resolved and sEnsitivity Enhanced 1H, 13C Data. *Analytical Chemistry*, *90*(3), 1845–1851.

[134] Klein, M. S., Oefner, P. J., & Gronwald, W. (2013). MetaboQuant: A tool combining individual peak calibration and outlier detection for accurate metabolite quantification in 1D 1-H and 1H-13C HSQC NMR spectra. *BioTechniques*, *54*(5), 251–256.

[135] Gómez, J., Brezmes, J., Mallol, R., Rodríguez, M. A., Vinaixa, M., Salek, R. M., ... Cañellas, N. (2014). Dolphin: A tool for automatic targeted metabolite profiling using 1D and 2D 1 H-NMR data. *Analytical and Bioanalytical Chemistry*, *406*(30), 7967–7976.

[136] Pearce, J. T., Athersuch, T. J., Ebbels, T. M., Lindon, J. C., Nicholson, J. K., & Keun, H. C. (2008). Robust algorithms for automated chemical shift calibration of 1D 1H NMR spectra of blood serum. *Analytical Chemistry*, *80*(18), 7158–7162.

[137] Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., ... Markley, J. L. (2008). BioMagResBank. *Nucleic Acids Research*, *36*(Database issue), D402–8.

[138] Zhang, A., Sun, H., Yan, G., Wang, P., & Wang, X. (2015). Metabolomics for biomarker discovery: moving to the clinic. *BioMed Research International*, *2015*, 354671.

[139] Monteiro, M., Carvalho, M., Bastos, M., & Guedes de Pinho, P. (2012). Metabolomics analysis for biomarker discovery: advances and challenges. *Current Medicinal Chemistry*, *20*(2), 257–271.

[140] Wolfender, J.-L., Rudaz, S., Hae Choi, Y., & Kyong Kim, H. (2013). Plant metabolomics: from holistic data to relevant biomarkers. *Current Medicinal Chemistry*, *20*(8), 1056–1090.

[141] Van Gestel, C. A., & Van Brummelen, T. C. (1996). Incorporation of the biomarker concept in ecotoxicology calls for a redefinition of terms. *Ecotoxicology*, *5*(4), 217–225.

[142] Wang, X., Chen, S., & Jia, W. (2016). Metabolomics in cancer biomarker research. *Current Pharmacology Reports*, *2*(6), 293–298.

[143] Koulman, A., Lane, G. A., Harrison, S. J., & Volmer, D. A. (2009). From differentiating metabolites to biomarkers. *Analytical and Bioanalytical Chemistry*, *394*(3), 663–670.

[144] Qiu, Y., Cai, G., Su, M., Chen, T., Zheng, X., Xu, Y., ... Jia, W. (2009). Serum metabolite profiling of human colorectal cancer using GC-TOFMS and UPLC-QTOFMS. *Journal of Proteome Research*, *8*(10), 4844–4850.

[145] Ni, Y., Xie, G., & Jia, W. (2014). Metabonomics of human colorectal cancer: New approaches for early diagnosis and biomarker discovery. *Journal of Proteome Research*, *13*(9), 3857–3870.

[146] Ghanbari, R., & Sumner, S. (2018). Using metabolomics to investigate biomarkers of drug addiction. *Trends in Molecular Medicine*, *24*(2), 197–205.

[147] Steuer, A. E., Brockbals, L., & Kraemer, T. (2019). Metabolomic strategies in biomarker research-new approach for indirect identification of drug consumption and sample manipulation in clinical and forensic toxicology? *Frontiers in Chemistry*, *7*(May 10), 319.

[148] Shanaiah, N., Zhang, S., Desilva, M. A., & Raftery, D. (2008). NMR-Based Metabolomics for Biomarker Discovery. In *Biomarker methods in drug discovery and development* (pp. 341–368).

[149] Zhang, A., Sun, H., Yan, G., Wang, P., & Wang, X. (2016). Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research. *Biomedical Chromatography*, *30*(1), 7–12.

[150] Gebregiworgis, T., & Powers, R. (2012). Application of NMR metabolomics to search for human disease biomarkers. *Combinatorial Chemistry & High Throughput Screening*, *15*(8), 595–610.

[151] Smolinska, A., Blanchet, L., Buydens, L. M., & Wijmenga, S. S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta*, *750*, 82–97.

[152] Carneiro, G., Radcenco, A. L., Evaristo, J., & Monnerat, G. (2019). Novel strategies for clinical investigation and biomarker discovery: A guide to applied metabolomics. *Hormone Molecular Biology and Clinical Investigation*, *38*(3).

[153] Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., ... Gerszten, R. E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine*, *17*(4), 448–453.

[154] Eick, C. F., Zeidat, N., & Zhao, Z. (2004). Supervised clustering - Algorithms and benefits. In *Proceedings - international conference on tools with artificial intelligence, ictai* (pp. 774–776).

[155] Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The application of unsupervised clustering methods to alzheimer's disease. *Frontiers in Computational Neuroscience*, *13*, 31.

[156] Cloarec, O., Dumas, M. E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., ... Nicholson, J. (2005a). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry*, *77*(5), 1282–1289.

[157] Cloarec, O., Dumas, M. E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., ... Nicholson, J. (2005b). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry*, *77*(5), 1282–1289.

[158] Zou, X., Holmes, E., Nicholson, J. K., & Loo, R. L. (2014). Statistical HOmogeneous cluster spectroscopY (SHOCSY): An optimized statistical approach for clustering of 1H NMR spectral data to reduce interference and enhance robust biomarkers selection. *Analytical Chemistry*, *86*(11), 5308–5315.

[159] Venkatesh, B., & Anuradha, J. (2019). A review of Feature Selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3–26.

[160] Saeys, Y., Aki Inza, I., & Larrañ Aga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics Review*, *23*, 2507–2517.

[161] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In *Data classification: Algorithms and applications*.

[162] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324.

[163] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1-3), 389–422.

[164] Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning* (1st). USA: Addison-Wesley Longman Publishing Co., Inc.

[165] Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization: An overview. *Swarm Intelligence*, *1*(1), 33–57.

[166] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

[167] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

[168] Huang, J., Horowitz, J. L., & Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, *36*(2), 587–613. arXiv: 0804.0693

[169] Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*, 1356–1378.

[170] Zou, H., & Hastie, T. (2005). *Regularization and variable selection via the elastic net* (tech. rep. No. 2).

[171] Vu, T., Siemek, P., Bhinderwala, F., Xu, Y., & Powers, R. (2019). Evaluation of multivariate classification models for analyzing NMR metabolomics data. *Journal of Proteome Research*, *18*(9), 3282–3294.

[172] Chen, T., Cao, Y., Zhang, Y., Liu, J., Bao, Y., Wang, C., . . . Zhao, A. (2013). Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-Based Complementary and Alternative Medicine*, *2013*.

[173] Xia, J., Broadhurst, D. I., Wilson, M., & Wishart, D. S. (2013). Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, *9*(2), 280–299.

[174] Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

[175] Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577.

[176] Sø, K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *Journal of Clinical Pathology*, *62*(1), 1–5.

[177] Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. *Current Metabolomics*, *1*(1), 92–107.

[178] Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, *16*(3), 119–128.

[179] Vapnik, V. N. (1998). *Statistical learning theory*. New York: John Wiley & Sons.

[180] Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, *879*, 10–23.

[181] Xu, Y., Zomer, S., & Brereton, R. G. (2006). Support vector machines: A recent method for classification in chemometrics. *Critical Reviews in Analytical Chemistry*, *36*(3-4), 177–188.

[182] Lokhov, P. G., Kharybin, O. N., & Archakov, A. I. (2012). Diagnosis of lung cancer based on direct-infusion electrospray mass spectrometry of blood plasma metabolites. *International Journal of Mass Spectrometry*, *309*, 200–205.

[183] Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., Walker, L. D. E., Gray, A., ... Fernández, F. M. (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, *10*(1), 259.

[184] Zheng, H., Zheng, P., Zhao, L., Jia, J., Tang, S., Xu, P., ... Gao, H. (2017). Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine. *Clinica Chimica Acta*, *464*, 223–227.

[185] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

[186] Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*(1), 36.

[187] Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307.

[188] Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25.

[189] Zhao, L.-L., Qiu, X.-J., Wang, W.-B., Li, R.-M., & Wang, D.-S. (2019). NMR metabolomics and random forests models to identify potential plasma biomarkers of blood stasis syndrome with coronary heart disease patients. *Frontiers in Physiology*, *10*, 1109.

[190] Wickham, H., & Bryan, J. (2015). Vignettes: long-form documentation. In *R packages* (2nd, Chap. 11). O'Reilly Media, Inc.

[191] RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA.

[192] Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, *78*(3), 779–787.

[193] Hanson, B. A. [Bryan A], & Keinsley, J. (2012). *ChemoSpec: An R Package for Chemometric Analysis of Spectroscopic Data and Chromatograms (Package Version 1.51-0) with contributions from Matt*.

[194] Croux, C., Filzmoser, P., & Oliveira, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *87*(2), 218–225.

[195] Jacob, D., Deborde, C., & Moing, A. (2013). An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data. *Analytical and Bioanalytical Chemistry*, *405*(15), 5049–5061.

[196] Cardoso, S. (2017). *Development of web-based tools for metabolomics data analysis and mining* (Master's thesis, University of Minho, Braga).

[197] Chong, J., & Xia, J. (2018b). MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics (Oxford, England)*, *34*(24), 4313–4314.

[198] Borges Costa, C. (2014). *Development of an integrated computational platform for metabolomics data analysis and knowledge extraction Master dissertation* (Master's thesis, University of Minho, Braga).

[199] Klein, M. (2021). *Mrbin: Magnetic resonance binning, integration and normalization*. R package version 1.5.0.

[200] Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

[201] Helmus, J. J., & Jaroniec, C. P. (2013). Nmrglue: An open source Python package for the analysis of multidimensional NMR data. *Journal of Biomolecular NMR*, *55*(4), 355–367.

[202] Inc., P. T. (2015). Collaborative data science. Retrieved from https://plot.ly

[203] Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

[204] Wang, B., Goodpaster, A. M., & Kennedy, M. A. (2013). Coefficient of variation, signal-to-noise ratio, and effects of normalization in validation of biomarkers from NMR-based metabonomics studies. *Chemometrics and Intelligent Laboratory Systems*, *128*, 9–16.

[205] Dumas, M. E., Maibaum, E. C., Teague, C., Ueshima, H., Zhou, B., Lindon, J. C., ... Holmes, E. (2006). Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: The INTERMAP study. *Analytical Chemistry*, *78*(7), 2199–2208.

[206] Parsons, H. M., Ekman, D. R., Collette, T. W., & Viant, M. R. (2009). Spectral relative standard deviation: A practical benchmark in metabolomics. *Analyst*, *134*(3), 478–485.

[207] Adams, D. C., & Otárola-Castillo, E. (2013). Geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, *4*(4), 393–399.

[208] Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., & O'Donovan, C. (2020). MetaboLights: A resource evolving in response to the needs of its scientific community. *Nucleic Acids Research*, *48*(D1), D440–D444.

[209] Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., ... Subramaniam, S. (2016). Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, *44*(D1), D463–D470.

[210] Atta-ur-Rahman, Choudhary, M. I., & Atia-tul-Wahab. (2016). Chapter 5 - the second dimension. In Atta-ur-Rahman, M. I. Choudhary, & Atia-tul-Wahab (Eds.), *Solving problems with nmr spectroscopy (second edition)* (Second Edition, pp. 191–225).

[211] Kruger, N. J., Troncoso-Ponce, M. A., & Ratcliffe, R. G. (2008). 1H NMR metabolite fingerprinting and metabolomic analysis of perchloric acid extracts from plant tissues. *Nature Protocols*, *3*(6), 1001–1012.

[212] Le Gall, G., Colquhoun, I. J., Davis, A. L., Collins, G. J., & Verhoeyen, M. E. (2003). Metabolite profiling of tomato (Lycopersicon esculentum) using 1H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *Journal of Agricultural and Food Chemistry*, *51*(9), 2447–2456.

[213] Tredwell, G. D., Behrends, V., Geier, F. M., Liebeke, M., & Bundy, J. G. (2011). Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. *Analytical Chemistry*, *83*(22), 8683–8687.

[214] Crook, A. A., & Powers, R. (2020). Quantitative NMR-based biomedical metabolomics: current status and applications. *Molecules*, *25*(21), 5128.

[215] Marchand, J., Martineau, E., Guitton, Y., Le Bizec, B., Dervilly-Pinel, G., & Giraudeau, P. (2018). A multidimensional 1 H NMR lipidomics workflow to address chemical food safety issues. *Metabolomics*, *14*(5).

[216] Pathan, M., Charrier, B., Tea, I., Akoka, S., & Giraudeau, P. (2013). New practical tools for the implementation and use of ultrafast 2D NMR experiments. *Magnetic Resonance in Chemistry*, *51*(3), 168–175.

[217] Júnior, L. H., Ferreira, A. G., & Giraudeau, P. (2013). Optimization and practical implementation of ultrafast 2D NMR experiments. *Quimica Nova*, *36*(4), 577–581.

[218] Pathan, M., Akoka, S., Tea, I., Charrier, B., & Giraudeau, P. (2011). "Multi-scan single shot" quantitative 2D NMR: A valuable alternative to fast conventional quantitative 2D NMR. *Analyst*, *136*(15), 3157–3163.

[219] de Vos, R. C. H., Hall, R. D., & Moing, A. (2011). Metabolomics of a model fruit: Tomato. In *Annual plant reviews volume 43* (Vol. 43, pp. 109–155).

[220] Carrari, F., Baxter, C., Usadel, B., Urbanczyk-Wochniak, E., Zanor, M. I., Nunes-Nesi, A., ... Fernie, A. R. (2006). Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiology*, *142*(4), 1380–1396.

[221] Giovannoni, J. (2001). *Molecular Biology of Fruit Maturation and Ripening*.

[222] Clendinen, C. S., Lee-McMullen, B., Williams, C. M., Stupp, G. S., Vandenborne, K., Hahn, D. A., ... Edison, A. S. (2014). 13C NMR metabolomics: Applications at natural abundance. *Analytical Chemistry*, *86*(18), 9242–9250.

[223] Bingol, K., Zhang, F., Bruschweiler-Li, L., & Brüschweiler, R. (2012). TOCCATA: A customized carbon total correlation spectroscopy NMR metabolomics database. *Analytical Chemistry*, *84*(21), 9395–9401.

[224] Clendinen, C. S., Pasquel, C., Ajredini, R., & Edison, A. S. (2015). 13C NMR metabolomics: INADEQUATE network analysis. *Analytical Chemistry*, *87*(11), 5698–5706.

[225] Oh, B. H., Westler, W. M., Darba, P., & Markley, J. L. (1988). Protein carbon-13 spin systems by a single two-dimensional nuclear magnetic resonance experiment. *Science*, *240*(4854), 908–911.

[226] Takeuchi, K., Sun, Z. Y. J., & Wagner, G. (2008). Alternate 13C- 12C labeling for complete mainchain resonance assignments using Cα Direct-detection with appli-

cability toward fast relaxing protein systems. *Journal of the American Chemical Society*, *130*(51), 17210–17211.

[227] Sumner, S. C., Williams, C. C., Snyder, R. W., Krol, W. L., Asgharian, B., & Fennell, T. R. (2003). Acrylamide: A comparison of metabolism and hemoglobin adducts in rodents following dermal, intraperitoneal, oral, or inhalation exposure. *Toxicological Sciences*, *75*(2), 260–270.

[228] Sumner, S. C., Stedman, D. B., Clarke, D. O., Welsch, F., & Fennell, T. R. (1992). Characterization of Urinary Metabolites from [1,2,methoxy-13C]-2-Methoxyethanol in Mice Using 13C Nuclear Magnetic Resonance Spectroscopy. *Chemical Research in Toxicology*, *5*(4), 553–560.

[229] Robinette, S. L., Ajredini, R., Rasheed, H., Zeinomar, A., Schroeder, F. C., Dossey, A. T., & Edison, A. S. (2011). Hierarchical alignment and full resolution pattern recognition of 2D NMR Spectra: Application to nematode chemical ecology. *Analytical Chemistry*, *83*(5), 1649–1657.