In-Situ NMR based Metabonomics of Microbial Secondary Metabolites

Sørensen, Mathies Brinks

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Sørensen, M. B. (2023). *In-Situ NMR based Metabonomics of Microbial Secondary Metabolites*. DTU Chemistry.

In-Situ NMR based Metabonomics of Microbial Secondary Metabolites

Mathies Brinks Sørensen

PhD thesis

May 2023

Center for Microbial Secondary Metabolites

&

Department of Chemistry

Kemitrovet building 207, 2800 Kgs. Lyngby

Technical University of Denmark

Preface

This thesis describes the work done for the PhD project titled ”In-Situ NMR based Metabonomics of Microbial Secondary Metabolites” at the Technical University of Denmark. The work was conducted at the Department of Chemistry in collaboration with The Center for Microbial Secondary Metabolites (CeMiSt), including external collaborators of DTU Compute and Copenhagen University department of Food science. The Project was supervised by associate professor Charlotte Held Gotfredesen and associate professor Mikael Lenz Strube. In addition, 3 months of external stay was included at the University of Copenhagen Department of Food science under the mentorship of Professor Rasmus Bro. All work was carried out in the period of December 2019 until May 2023.

The project was funded by the Danish National Research Foundation, DNRF137. In addition, The NMR Center • DTU and the Villum Foundation are acknowledged for access to the 800 MHz spectrometer.

Mathies Brinks Sørensen
Kgs. Lyngby May 2023

Abstract

Microbial communities are known to produce an abundance of natural products, including secondary metabolites. Parts of the microbial community have been studied under in vitro conditions, which has led to a greater understanding of specific elements attributed to set experimental conditions. However, in vitro conditions fail to mimic metabolites being produced as a consequence of microbial interactions within natural niches. Therefore, if one were to gain an understanding of natural microbial interactions, it may lead to the uncovering of novel secondary metabolites. However, when biological complexity increases, so does the chemical complexity, leading to an abundance of challenges within In-situ detection.

Throughout the work of this thesis, nuclear magnetic resonance (NMR) spectroscopy was applied to generate metabolomic data which was utilized within a targeted approach. Two specific challenges were undertaken within targeted NMR in-situ detection. The first challenge was to reduce operator bias by developing automatic detection and uncertainty evaluation of complex metabolomic spectral samples. The second challenge was to ensure that a robust workflow with respect to data generation and analysis of data could be set up via a standardized metabolomic pipeline.

The first challenge led to the creation of the Python/Pytorch-based NMR-Onion framework (paper 2). The framework allowed for automatic detection, quantification, and uncertainty evaluation of detected peaks within complex spectral NMR samples. It was concluded that NMR-Onion could detect and evaluate signals across multiple signal-to-noise ratio values. In addition, the algorithm would make users aware of potential sample-to-sample variations in the form of potentially resolved peaks, reducing the risk of drawing false conclusions. The program was developed with a targeted approach in mind but has the potential to be utilized within non-targeted studies as well.

The second challenge was addressed by creating a metabolomic workflow. The workflow was generated by combining design of experiments (DoE), statistical quality control (SQC), minimal preprocessing, automatic detection (NMR-Onion), and statistical analysis. Methods for DoE and SQC were reviewed in paper 1, which resulted in two recommend workflows for metabolomics experiments involving DoE and SQC. Finally, the generated metabolomics workflow was utilized within a case study of pseudo-in-situ data. Here it was found that deconvoluting DoE-based NMR spectral data with NMR-Onion and subsequently analyzing the deconvoluted results via general linear effect models could be utilized to link targeted amplitudes to that of specific amino acids.

In conclusion, this thesis has developed a new tool of NMR-Onion to be utilized within automatic detection, quantifying, and evaluation of NMR signals. When generating metabolomic data, the framework should be paired with steps of the generated workflow to ensure optimal results. The future of the workflow generated in this thesis may be utilized to explore metabolomic in-situ data from natural microbial communities, potentially aiding in uncovering the diversity and functions of secondary metabolites. In addition, the NMR-Onion framework may be utilized within any area of in-situ detection such as disease diagnostics, agriculture, or food science.

Resumé

Mikrobielle samfund er kendt for at kunne producere en overflod af naturstoffer, herunder sekundære metabolitter. Dele af det mikrobielle samfund er blevet undersøgt under in vitro betingelser, hvilket har ledt til en større forståelse af specifikke elementer, der leder tilbage til opsatte eksperimentelle forhold. Men in vitro betingelser ikke kan efterligner de betingelser under hvilket sekundære metabolitter bliver produceret som følge af mikrobielle interaktioner. Hvis man kan opnå en forståelse af naturlige mikrobielle interaktioner, kan dette lede til opdagelsen af nye hidtil ukendte sekundære metabolitter. Det gælder dog, at ved en øget biologisk kompleksitet, stiger den kemiske kompleksitet ligeledes, hvilket vanskeligøre in-situ detektion. I dette projekt blev kerne magnetisk resonans (NMR) spektroskopi anvendt til at genere metabolomic data, der blev anvendt til targted analyse. Der blev taget hånd om to specifikke udfordringer inden for targted NMR in-situ detektion. Den første udfordring bestod i at reducere operatør bias ved at udvilke automatisk detektion og usikkerheds evaluering af komplekse spektral data prøver. Den anden udfordring var at sikre et robust workflow kunne sættes op via en metabolomic data pipline, med henblik på generering og analyse af data. Den første udfordring førte til udviklingen af NMR-Onion, som er et Python/Pytorch baseret program (artikel 2). Programmet kan automatisk detektere, kvantificere samt vurdere usikkerheden af signaler fra komplekse NMR prøver. Konklusionen var at NMR-Onion kunne bruges til evaluere og detektere signaler under forskellige signal til støj forholds værdier. Ydermere giver algoritmen også brugere mulighed for at opdage potentielle signaler, hvilket reducere risikoen for falske konklusioner. Programmet blev udviklet med henblik på targted analyse, men har potentialet for at kunne blive anvendt til ikke targted analyse. Den anden udfordring blev adresseret ved at udvikle et metabolomic worflow. Workflowet blev lavet ved at kombinere design of experiments (DoE), statistical quality control (SQC), minimal præprocessering, automatisk detektion og statistisk analyse. DoE og SQC metoder blev reviewet i artikel 1, hvilket resultererede i to mulige workflows for anvendelsen af DoE samt SQC i metabolomic eksperimenter. Til slut blev det udviklet metabolomic workflow anvendt i et case study af psedou-in-situ data. I dette studie blev det fastslået, at dekonvolering af DoE baseret NMR spektral data, via NMR-Onion, efterfulgt af analyse via generelle lineære modeller, kunne anvendes til at sammenholde targted amplituder med specifikke aminosyrer. Konklusionen herpå var at et nyt værktøj, NMR-Onion, kunne anvendes til automatisk detektering, kvantificering og evaluering af NMR signaler. Ved generering af metabolomic data bør programmet sammensættes med det øvrige workflow, for at sikre optimale resultater. De fremtidige perspektiver for workflowet udviklet i projektet er at udforske metabolomic in-situ data fra naturlige mikrobielle samfund, med henblik på at forstå diversiteten samt rollen af sekundære metabolitter. Ydermere kan NMR-Onion programmet også anvendes inden for et hvilket som helst område af in-situ detektion, herunder sygdoms diagnostik, agerbrug og fødevarer videnskab.

Acknowledgements

First, I would like to thank the CeMiSt Center of Excellence as a whole for the collaboration during the project period. Here I would especially like to thank my co-supervisor Mikael Lenz Strube for helping me with data science, in particular for opening my eyes to the interplay between biology and statistical analysis.

I would like to give a special thanks to Michael Riis Andersen from DTU Compute, who has helped me tremendously throughout the project: Teaching me multiple statistical, machine learning, and mathematical methods within data science. Furthermore, he was always ready to provide a hand whenever there was a bug in the code or any other small problem.

I also want to thank Jan Kloppenborg Møller at DTU Compute for helping me out with the project. He always had the time for me to drop by and discuss statistics and mathematics helping me with all sorts of problems big and small, thank you.

I went to the Department of Food at the University of Copenhagen as part of my external stay. I sincerely want to thank my host supervisor, Professor Rasmus Bro, for letting me be a part of the chemometric group. I especially enjoyed the group meetings discussing various subjects within chemometrics. Also, I would like to thank Rasmus Bro for helping me develop the NMR-Onion algorithm and also for being a great and inspiring person - I always felt energized and full of inspiration after a meeting, thank you.

A heartfelt thank you to my good friends Mathilde Lerche and Magnus Karlssons for also supporting me during my time as a phd student, they were always there for a friendly chat whenever I needed it the most and their support have been invaluable. Mathilde and Magnus have helped me grow, not only as an academic but also as a person and I am grateful for knowing them.

I would also like to thank my fiance Karina, who have supported me during my studies even when times were hard. She was always there to help with our kids, making family life alongside a Ph.D. possible. In addition, I would like to thank my mother Sanne, who has always been there for me cheering me up whenever I needed it the most and supporting me emotionally throughout the whole project.

Finally, the last person I want to thank is my main supervisor Charlotte Held Gotfredsen. I would never have been able to complete this project without her help, guidance, and mentorship. She was always there when things were tough both academically and personally. I could not wish for a better mentor to guide me through the world of academia, I am truly grateful.

Papers included in the thesis

- **Paper 1**: **Sørensen, MB**, Møller, JK, Strube, ML and Gotfredsen, CH (2023). Designing optimal experiments in metabolomics. Status: To be submitted to Springer, Metabolomics
- **Paper 2**: **Sørensen, MB**, Andersen, MR, Siewertsen, MM, Bro, R, Strube,ML and Gotfredsen, CH (2023). NMR-Onion - a transparent multi-model based 1D NMR deconvolution algorithm. Journal: Elsevier, Journal of Magnetic Resonance, status: Submitted.

List of abbreviations

AD	Automatic differentiation
ACME	Automated phase correction based on minimization of entropy
AIC	Akaike information criterion
ANOVA	Analysis of variance
ARpls	Asymmetrically re-weighted penalized least squares smoothing
ASCA	ANOVA simultaneous component analysis
BIC	Bayesian information criterion
DFT	Discreet Fourier transformation
DNP	Dissolution dynamic nuclear polarization
DNSS	Sodium trimethylsilylpropanesulfonate
DoE	Design of experiments
FID	Free induction decay
FIR	Finite impulse response
FWHM	Full width at half maximum
GLM	Generalized linear model
GLMM	Generalized linear mixed model
GNPS	Global Natural Product Social
iid	Independent and identically distributed
IDFT	Inverse discreet Fourier transformation
IQR	Inter quantile range
LB	Line broadening
LBFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
LR	Learning rate
MS	Mass spectrometry
MSE	Mean sum of squares error
MSS	Mean sum of squares
NMR	Nuclear magnetic resonance
OVAT	One variable at a time
PCA	Principal component analysis
PC	Principal component
PC1	First principal component

PLS	Partial least squares
PLS-DA	Partial least squares discriminant analysis
QQ-plot	Quantile quantile plot
RA	Research area
RQ	Research question
RF	Radiofrequency
ROI	Region of interest
RSD	Residual standard deviation
SG-filter	Savitzky-Golay filter
SQC	Statistical quality control
SM	Secondary metabolite
SSE	Sum of squares error
SNR	Signal to noise ratio
TSP	Trimethylsilylpropanoic acid
w.r.t.	With respect to

List of symbols

B_0	Magnetic field strength
γ	Magnetogyric ratio
ω_1	Larmor frequency
M_0	Bulk magnetization vector
β	Tilt angle
M_x	Magnetization vector projected on the x-axis
M_y	Magnetization vector projected on the x-axis
t	time
T_2	T2 decay constant
ϕ_0	Single instantaneous phase
ω_0	Single frequency
A	Time domain amplitude
$y(t)$	Free induction decay as a function of time
A_c	Complex amplitude
j	Imagninary unit
Re	Real part of a complex number
Im	Imaginary part of a complex number
ψ	Time domain decay function
α	Decay rate of the free induction decay
K	Number of sinuoids or peaks
$S(\omega)$	Frequency domain signal post DFT
y_n	Discrete time series (free induction decay)
$\tilde{y_n}$	Zero-filled time domain sequence
N	Number of time and frequency points
ϕ	Phase angle
θ	Phase angle correction term
S_0	Intensity of frequency domain signal
θ_1	First order phase error
ω_d	Offset distance
b	Vector of smoothing coefficients
M	Number of smoothing coefficients

M^*	Total magnetization
N_{spin}	Total number of spins
P_r	Polarization ratio
N_α	Spin state 1
N_β	Spin state 2
S	Total signal
σ_N	Average noise amplitude
t_{aq}	Acquisition time
S_k	Signal of the k'th frequency bin
C_i	The i'th weighting coefficient
m	Degree of Savitky Golay filter smoothing
Δ	Frequency increments
V	Vandermonde matrix
SW_{Hz}	Sweep width in Hertz
n	Window length
P	Solutions to normal equations for the smoothing polynomial
$\boldsymbol{z}$	Normalized data point sequence vector
$\bar{S}$	Center point value
$\lfloor\rfloor$	Floor operator
$\tilde{S}$	Smoothing signal in the z-domain (normalized data)
$\boldsymbol{b}$	Vector of Savitky Golay coefficients
H_0	Null hypothesis
H_1	Alternative hypothesis
$\boldsymbol{\theta}$	Vector of experimental factors (treatments)
$\boldsymbol{Y}$	Data response vector
$f()$	Model function
$\boldsymbol{z_{ARpls}}$	Smoothed Arpls baseline
$\boldsymbol{W}$	Diagonal matrix containing asymmetric weights
D	Second order distance matrix
λ	Arpls penalty term
$y_{raw}(t)$	Free induction decay prior to apodization
R	Line broadening function coefficient

b_{off}	Accumulated signal decay offset constant
F	Rectangular filter function
S_l	Lower frequency value of filter
S_u	Upper frequency value of filter
$\boldsymbol{g(t)}$	Gaussian filter in time domain
$\boldsymbol{g(\omega)}$	Gaussian filter in frequency domain domain
σ_g	Gaussian filter scale parameter (filter standard error)
ω_c	Gaussian filter location parameter (center frequency)
$\boldsymbol{sg(\omega)}$	Super Gaussian filter in frequency domain
B	Bandwidth of super Gaussian filter
p	Filter order of super Gaussian filter
σ_{noise}	Spectral noise level
$\boldsymbol{W_s}$	Spectral noise vector
$N()$	Normal distribution
s_r	Spectral realizations in a noise region
σ_{ARpls}	Estimated noise level post ARpls correction
$\boldsymbol{H(\omega)}$	Sum of super Gaussian filter and average noise response vector
$\boldsymbol{\Psi_n}()$	Decay function for the n'th model
$\boldsymbol{\rho_k}$	Subset of parameters for the n'th decay function
γ_k^*	Skewing constant for the k'th sinusoid
η_k	Weighting constant of the k'th sinusoid
β_k	Stretch/compression constant for the k'th sinusoid
$\hat{\theta}$	Estimated model parameters
$\bar{\phi}$	Mean of all instantaneous phase values
$\boldsymbol{Z}$	Model matrix
$\boldsymbol{A}$	Complex amplitude vector
$()^H$	Hermitian transposed (complex conjugated)
W_s	Lortentizan weigthing function
W_r	Guass-to-Lortentzian weighing function
$d_I^{'}$	SG-filter first derivatives of imaginary part of spectrum
$d_R^{'}$	SG-filter first derivatives of real part of spectrum
$d_I^{''}$	SG-filter second derivatives of imaginary part of spectrum

d''_R	SG-filter second derivatives of real part of spectrum
$sd_{background}$	Standard deviation of background noise
lr_{epoch}	Adjusted learning rate per epoch
γ_{lr}	Learning rate decay
θ^*	Transformed parameters
$p(\boldsymbol{\theta} \mid y)$	Posterior distribution of time domain signals
$p(\boldsymbol{\theta})$	Parameter prior
$p(y \mid \boldsymbol{\theta})$	Likelihood function
$\frac{1}{g}$	Expected signal to noise
$U()$	Uniform distribution
$IG()$	Inverse gamma distribution
$\frac{1}{\alpha}$	Jefferys prior
$\boldsymbol{X}$	General linear model design matrix
$\boldsymbol{\beta_m}$	Vector of fixed effects
$\boldsymbol{\varepsilon}$	Vector of model residuals
$\boldsymbol{I}$	Identity matrix
σ	residual variance
μ	overall mean
α_i	predictor of factor E (Histidine)
β_j	predictor of factor B (Lysine)
γ_k	predictor of factor C (Argine)
δ_l	predictor of factor D (Glycine)
ψ_h	predictor of factor A (Cysteine)

Contents

Chapter 1

Introduction

Secondary metabolites (SM) have been an important part of many research fields and applications such as medicine[1][2], food[3][4] and agriculture[5][6]. It all began back in the early 20th century with the discovery of penicillin as the first antibiotic agent by Alexander Fleming[7], setting the stage for SMs to be vital for developing drugs. Within food production, many compounds such as coloring and flavoring agents originate from SMs. One of the earliest flavoring agents in food may be that of vanilla, which was isolated by Gobley as a pure compound in 1858[8]. As for SM-based coloring agents, the red carmine pigment (carminic acid) produced by insects has been known for centuries and was described in 1894 by Schunck[9]. Finally, for SMs within agriculture, a very important discovery was one of the first insecticides isolated from pyrethrum extracts (Pyrethrins) and identified by Fujitani in 1909[10]. This group of SMs is even utilized to this day, as the pyrethrins are of low toxicity for mammalians[11]. Hence SMs are highly diverse in function and origin, appearing in many types of biological environments.

Traditionally, SMs are detected within microbial organisms grown under in vitro conditions. The traditional technique may aid in understanding how to control the production of specific metabolites under controlled experimental conditions. However, the in vitro conditions fail in mimicking the ecological settings of an e.g. microbial community. This may lead to a loss of information with respect to metabolites produced as a result of microbial interactions. Therefore, if increased comprehensions of these interactions were to be gained, it may lead to novel discoveries of secondary metabolites which are only produced within specific natural microbial communities. In turn, this knowledge could improve the discovery of new antibiotic drugs, alternative food production, or agricultural enrichment, and increase the understanding of microbiology in general.
With an increasing biological diversity, an increased complexity of microbial chemistry follows, giving rise to the challenge of *In-Situ* detection.

1.1 In-situ detection

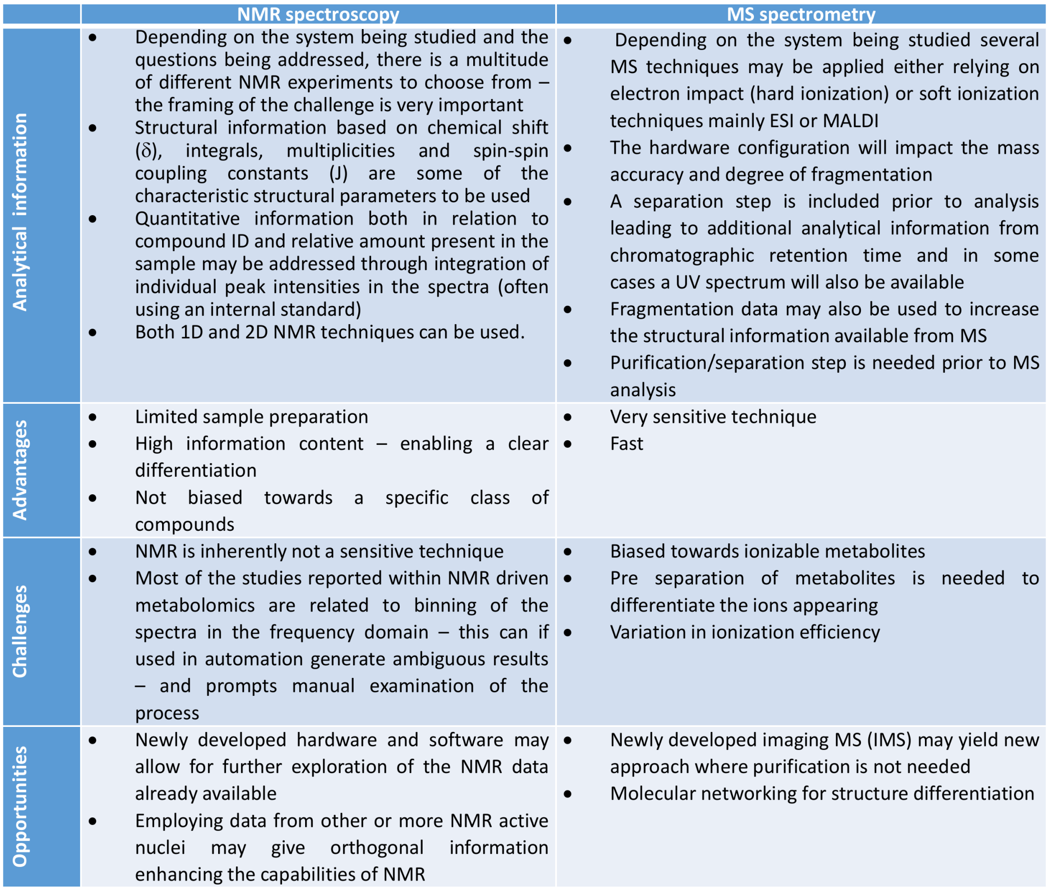
In-situ detection and analysis of metabolites is an area of increased focus and interest when the chemistry of natural environments is investigated while aiming for as little outside interference as possible. These studies are an important part within many fields of sciences such as ecology[12], food products[13], drug effect assessments[14] and disease diagnostics[15].

The aforementioned fields may seem diverse, but they share a common characteristic: highly intricate data sets containing mixtures of numerous metabolites, often numbering in the hundreds. The methods of detecting and/or quantifying metabolites within complex mixtures, may be divided into two approaches of metabolomic analysis. The first is a target analysis, identifying the presence of a specifically known metabolite conditioning on controlled experimental factors. The second approach focuses on identifying as many metabolites as possible within a given biological matrix, known as a non-targeted analysis. The latter has been used extensively in dietary response analysis, where one is attempting to identify new dietary biomarkers as seen in the work of Gambin[16]. Here, the compounds of trigonelline, 3-methylhistidine, dimethylglycine, trimethylamine, and lysine were identified as potential biomarkers linked to the intake of different types of beans. Another example of non-targeted metabolomic analysis is found in the field of ecology. In this study, the metabolomic response of soil supporting the growth of Burkea africana trees was investigated via both NMR and MS[17]. The goal of the study was to compare the metabolomic response of soil in which Burkea grows (Burkea soil) vs soil in which growth is not detected (non-Burkea soil). The results from applying Partial least squares discriminate analysis to NMR data showed that Trehalose and betaine were found in higher concentrations in Burkea soil, whereas non-Burkea soil exhibited higher acetate, lactate, and formate concentrations. From liquid chromatography MS it was further revealed that the presence of aspartic acid and glutamine was higher in Burkea soil. The study concluded that the metabolic response may be coupled to that of fungal variations within the different soil types (BLAST analysis was conducted). In addition, the bacterial interactions were also tested, where it was found that the bacterial species composition was identical in each soil type.

Throughout this thesis, targeted analysis is the primary focus, whilst the non-targeted analysis is only addressed in paper 2.

To fulfill the goal of targeted analysis, e.g. modeling the response of data requires statistical design of experiments (DoE) to be performed (see figure 1.2). The purpose of the DoE is to ensure that the data is statistically informative with respect to (w.r.t.) the design space of the study, whilst extracting maximum information with the smallest amount of samples. In addition to the design space, the aspect of data quality posses significant importance when modelling data, as measurement artifacts may alter conclusions even in the best-designed experiments. To ensure data quality, and lower the influence of nuisance factors, statistical quality control (SQC) is an important addition. Hence, how to statistically design experiments and ensure quality data is highlighted within the review presented in paper 1 and summarized in chapter 3, emphasizing the combined usage of DoE and SQC approaches that have been applied throughout the literature of metabolomics and related fields, e.g. medical science, genetics, agriculture (see paper 1 for more details).

The complexity of in-situ-based experiments does not occur solely due to the complex chemical matrix of samples nor their design, but also due to the nature of the data itself. In general two types of data are acquired when analyzing chemical compounds/metabolites, mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy data. The different types of data have their strength and weaknesses, which are shortly summarized within figure 1.1.



	NMR spectroscopy	MS spectrometry
Analytical information	• Depending on the system being studied and the questions being addressed, there is a multitude of different NMR experiments to choose from – the framing of the challenge is very important • Structural information based on chemical shift (δ), integrals, multiplicities and spin-spin coupling constants (J) are some of the characteristic structural parameters to be used • Quantitative information both in relation to compound ID and relative amount present in the sample may be addressed through integration of individual peak intensities in the spectra (often using an internal standard) • Both 1D and 2D NMR techniques can be used.	• Depending on the system being studied several MS techniques may be applied either relying on electron impact (hard ionization) or soft ionization techniques mainly ESI or MALDI • The hardware configuration will impact the mass accuracy and degree of fragmentation • A separation step is included prior to analysis leading to additional analytical information from chromatographic retention time and in some cases a UV spectrum will also be available • Fragmentation data may also be used to increase the structural information available from MS • Purification/separation step is needed prior to MS analysis
Advantages	• Limited sample preparation • High information content – enabling a clear differentiation • Not biased towards a specific class of compounds	• Very sensitive technique • Fast
Challenges	• NMR is inherently not a sensitive technique • Most of the studies reported within NMR driven metabolomics are related to binning of the spectra in the frequency domain – this can if used in automation generate ambiguous results – and prompts manual examination of the process	• Biased towards ionizable metabolites • Pre separation of metabolites is needed to differentiate the ions appearing • Variation in ionization efficiency
Opportunities	• Newly developed hardware and software may allow for further exploration of the NMR data already available • Employing data from other or more NMR active nuclei may give orthogonal information enhancing the capabilities of NMR	• Newly developed imaging MS (IMS) may yield new approach where purification is not needed • Molecular networking for structure differentiation

Figure 1.1: Outline of some of the differences between NMR spectroscopy and mass spectrometry

Each of these data types does not provide an immediate response as to which metabolites are present. Highly trained researchers are required to transform the data into useful chemical knowledge. However, the requirement for human interpretation often leads to operator bias. To mitigate the problem, usage of automation has been seen in MS data with the focus of molecular networking[18] combined with extensive usage of high-quality public databases found in the Global Natural Product Social Molecular Networking (GNPS) community, providing chemical information from individual components. As for NMR, the automation is less developed than the MS, though many contributions have gradually improved upon automation as discussed in paper 2. It should be mentioned that automatic detection in NMR does exist in the form of neural networks, for instance, the Caspar neural network[19] which is capable of detecting and classifying carbohydrates. Another example is the automatic chemical shift predictions of proteins[20] which is also well developed. However, a universal robust approach (e.g. not compound specific) for automatic detection of known metabolites requires better databases and data mining algorithms to run reliably without impact from operator bias. For the scope of this thesis, the main focus is on automatic data mining algorithms, whilst database integration is viewed as a future perspective. In short, the solutions developed within this project mainly emphasize the extraction of information from 1D

NMR data, within targeted DoE-based in-situ NMR spectroscopy experiments. The goal is to showcase how optimal experimental conditions may be set up w.r.t. design space, whilst minimizing operator bias through automatic detection. The complete workflow of the thesis is visually summarized in figure 1.2.

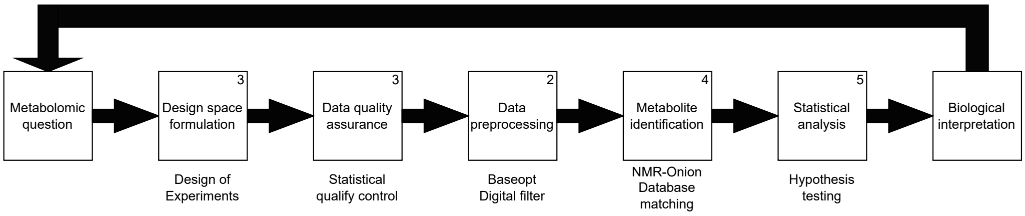


Figure 1.2: Visualization of the full workflow of the thesis, the numbers within each box refer to the chapters of the thesis.

1.2 The CeMiSt project

The Center for microbial secondary metabolites (CeMiSt) is a Center of Excellence funded by the Danish National Research Foundation. CeMiSt is engaged in researching the impact and roles of secondary metabolites (SM) from various bacterial and fungal species in natural microbial systems. The goals are to understand the ecological role of SMs within the natural environment of the species in question. Specifically, the research areas (RA) of function, diversity, and evolution of microbial communities are studied, and how these are linked to the production of SMs. Designated research questions (RQ) have been assigned to each RA, from which sub-conclusions may provide a bigger picture in understanding the three RA's. All of the research areas and research questions (seven in total) can be found on CeMiSt web page http://www.cemist.dtu.dk/. The ones related to the works of this thesis are listed below:

- RQ2: What is the diversity of microbial SM production in a natural environment?
- RQ3: How do SM-producing microorganisms or their products affect the diversity and functionality of a microbial niche?

From the RQs, the goals are to detect and investigate the role of SMs in the natural habitat of microbial communities. However, this is a very complex and very broad task. Therefore to reduce the complexity of the RQs, the next section emphasizes the scopes of the thesis linked to the two RQs.

1.2.1 Metabolomics vs metabonomics

The title of the thesis implies that metabonomics is the focus of this thesis rather than metabolomics. The terms are often used interchangeably when studying the metabolome of bio-samples[21]. The differences are very subtle and the chemometric methods for analyzing both types of data are identical[21]. In short, metabonomics aims at quantifying metabolomic profiles as a whole across populations (sometimes as a response to external experimental factors), whereas metabolomics aims at identifying individual usually smaller metabolites[21][22]. For this thesis, the aim was to perform targeted analysis of smaller individual metabolites whilst also quantifying the response of experimental factors. Hence, a mixture of both metabonomics and metabolomics was involved. However, within the literature metabonomics is often associated as a subcategory of metabolomics [22], therefore the choice of metabolomics was chosen as a broader term. It should be noted that metabolomics and metabonomics are often associated with Mass spectrometry or NMR spectroscopy respectively.

1.3 Purpose

This thesis aimed to investigate if specific compounds could be detected and analyzed within in-situ samples, applying the analytical technique of nuclear magnetic resonance spectroscopy. Previously, NMR spectroscopy has been utilized in tandem with chemometrics to study complex metabolomic samples ranging from urine samples[23] to even more complex biofluid extracts[24]. Hence, NMR is a suitable technique for analyzing complex in-situ mixtures. However, within the works of this thesis, the challenge of analyzing complex in-situ spectra is further enhanced, as the specific compounds targeted are present in very low concentrations and may be hidden underneath larger spectral peaks. Therefore, the aims of this thesis can be converted into the following objectives

1. Automatic in-situ detection and uncertainty evaluation of low signal-to-noise ratio (SNR) metabolites within complex mixtures

2. Ensure robust data generation and analysis within a metabolomic pipeline

To accommodate the first objective, the research of paper 2 was conducted, generating the novel automatic detection/quantification framework of NMR-Onion. In short, the algorithm ensured that even low SNR signals would be detected and most importantly, the uncertainty of each signal would be quantified utilizing a bootstrap re-sampling approach. The second objective was achieved by setting up the general workflow of figure 1.2, in which the second and third box is described within paper 1. Here it is reviewed how a design of experiments and statistical quality control may be implemented within a metabolomic setting, emphasizing why each concept is an important part of a solid metabolomic study. The remaining boxes of preprocessing and metabolite identification are embedded within NMR-Onion (paper 2), whilst the last box of statistical analysis is highlighted with a constructed example in Chapter 5.

1.4 Thesis outline

The thesis is comprised of six chapters in total, with each chapter being linked to the overall metabolomic workflow highlighted in figure 1.2 found section 1.1 of chapter 1. The **first chapter** serves as a general introduction to the aims of the Ph.D. project, In-situ detection background, and highlights the general aims of The center for microbial secondary metabolites linked to In-situ detection. The **second Chapter** is an introduction to the modelling of NMR signals, coupling the works of paper 2 with NMR theory, automatic signal detection, and spectral prepossessing methods. The **third chapter** consists of two parts, summarizing the works of paper 1, which emphasizes the importance of design of experiment (DoE) and statistical quality control (SQC) within metabolomics data generation. The **Fourth chapter** is meant as a short summary of the works done within the NMR-Onion framework (paper 2), whilst also covering additional details of the algorithm. This involves, how the underlying peak detection algorithm works in detail and aspects of the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) optimization algorithm within the PyTorch framework. Finally, the development history of NMR-Onion is included, covering how and why NMR-Onion ended up in its current form. **Chapter five** serves as the last piece of the metabolomic workflow (figure 1.2). The section highlights how statistical hypothesis testing methods can be applied to metabolomics data, analyzing the outputs of the NMR-Onion algorithm. The majority of methods highlighted throughout the entire thesis are then finally applied within a case study of pseudo-In-Situ NMR data. Here the results are analyzed and disused, with the emphasis of proving that the NMR-Onion framework may provide insights into analyzing in-situ samples in a reliant way. Within **chapter six**, conclusions and future perspectives are drawn as a whole, summarizing all the work conducted within this thesis.

Chapter 2

NMR and Metabolomics

A core function of the NMR-Onion algorithm (paper 2 and Chapter 4) is to model 1D NMR signals in the time domain. However, the link between signal generation and mathematical modelling is not included original paper.
Therefore, the goal of this chapter is to provide further insights into the modelling of 1D-NMR signals. The chapter focuses on the major aspects utilized within the NMR-Onion algorithm, highlighting time and frequency domain differences, detection of NMR signals, and the methods of sensitivity and resolution enchantments. In addition, the NMR acquisition schemes utilized in the work of this thesis are emphasized. Finally, the chapter links metabolomics and NMR spectroscopy through mathematical hypothesis formulations, highlighting the connection between the spectral prepossessing procedure and spectral outcomes. Here it is emphasized how the digital filters of NMR-Onion (paper 2) may aid in retaining more information than traditional prepossessing methods.

2.1 Theoretical NMR

NMR spectroscopy is an analytical technique that is based on the detection of magnetic resonance of nuclei. Specifically, nuclei that do not have an even number of protons and neutrons can be detected as resonances in an NMR spectrometer. The particular nuclei used in the works of this thesis are the ^{1}H, which is a spin $\frac{1}{2}$ nuclei, and may either be a spin $\frac{1}{2}$ or $-\frac{1}{2}$ state[25]. The signal strength and detection capabilities for NMR are dependent on field strength, magnetogyric ratio, and also on the natural abundance of the analyzed isotope. To fully understand the aforementioned relation, quantum mechanical models have been developed as showcased in the works of Cummings[26], but for the purpose of this thesis, the vector model[25] all-though much simpler, provides an excellent representation of how NMR signals may be understood.

2.1.1 Signal detection

The vector model describes the interactions of nucleis when placed in a magnetic field. An NMR magnet works by inducing alignment along the magnetic field direction for a population of nuclei. The nuclei align either with or opposite to the magnetic field creating a bulk vector effect for the population as a whole (which is parallel to the magnetic field). The bulk magnetization is thus a result of the sum of magnetic moments pointing toward the direction of the field vector.

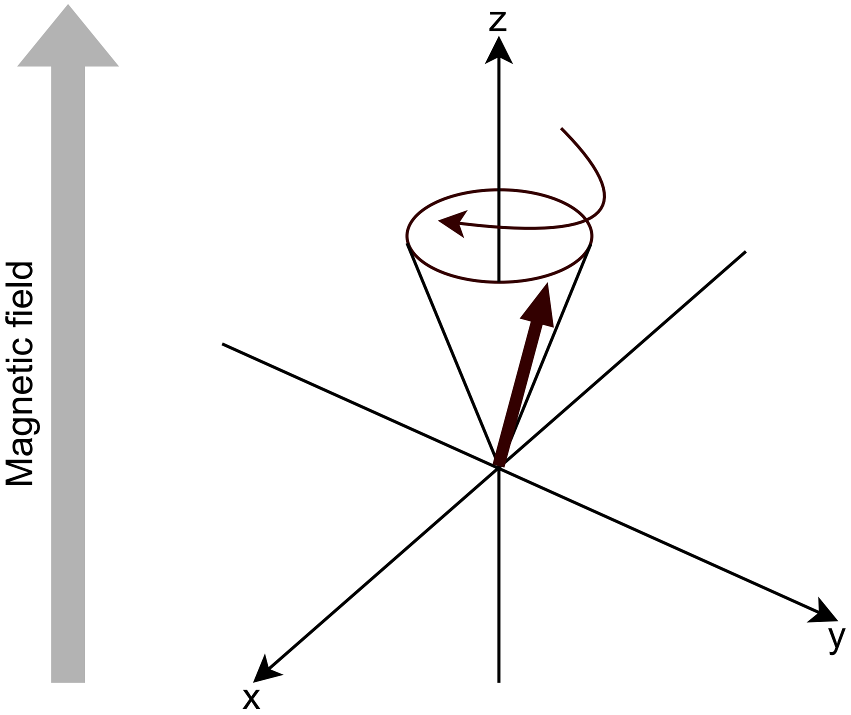


Figure 2.1: Visualization of the magnetization process where the magnetization vector (black arrow) has been tilted away from the z-direction

To generate a signal from the steady state bulk magnetization vector, the vector must be ”titled” away from the Z-axis (see figure 2.1) such that coils aligned in the xy-plane may read off the signal, generating a current which eventually turns into the free induction decay (FID) or time domain signal (see next subsection for details).

For the bulk magnetization vector to be ”tilted” away from the z-axis and into the xy-plane, a radio frequency (RF) pulse is applied. Specifically, the coil detecting the signal on the x-axis is used to generate the oscillations when being subjected to an RF, with a specific RF depending on the type of nuclei. In theory, if only a single nuclei is present, the RF pulse is applied at the Larmor frequency, which causes the nuclear spins to absorb the energy from the pulse and undergo resonance. The particular formulation of the Larmor frequency (see equation (2.1)) was discovered by Jopseh Larmor in the year of 1897[27] and can be expressed as the following

$$\omega_1 = -\gamma B_0. \tag{2.1}$$

Here, the Larmor frequency ω_1 is expressed in rad/s and is proportional to the magnetic field strength (B_0) whose slope is defined by the magnetogyric ratio (γ), corresponding to a specific value for given nuclei. The link between the Larmor frequency of equation (2.1) and the detection coil is further envisioned by how the magnetic field vector is projected onto the xy-plane. Given that the length of the bulk magnetization vector for one resonance is M_0 and the vector "tilted" at an angle of β, the x (M_x) and y (M_y) projection may be written as

$$M_x = M_0 \sin(\beta) \cos(\omega_0 t) \tag{2.2}$$

$$M_y = M_0 \sin(\beta) \sin(\omega_0 t) \tag{2.3}$$

Equation (2.2) and (2.3) give rise to two functions that are theoretically shifted by 90°. The relation implies that at t=0, the x projection would be equal to $M_0 \sin(\beta)$, while the y projection would be equal to 0. Subsequently, at t=1, the amplitudes of x and y would be $M_x = M_0 \sin(\beta) \cos(\omega_0)$ and $M_y = M_0 \sin(\beta) \sin(\omega_0)$ and so on, generating a sinusoidal pattern visualized in figure 2.2.

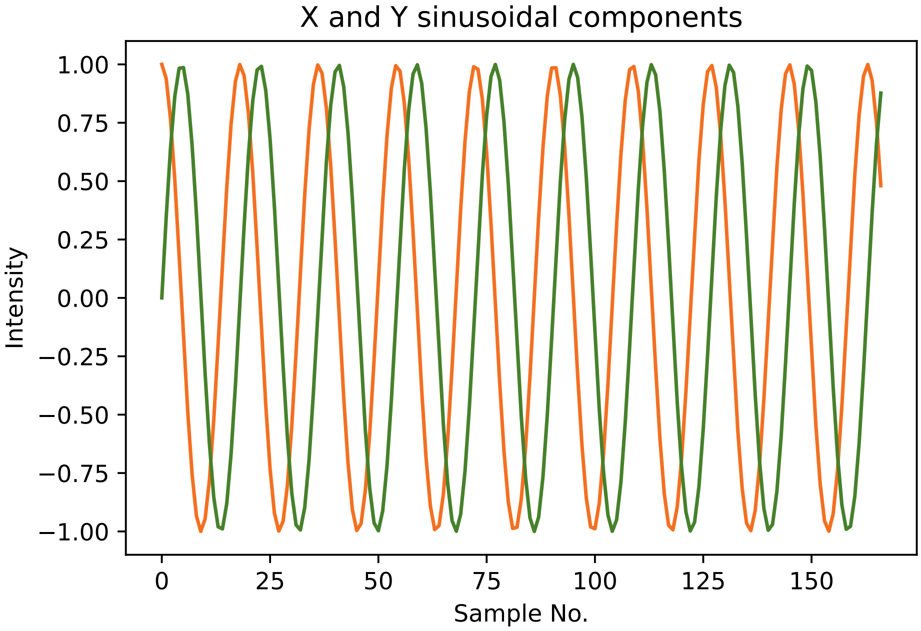


Figure 2.2: Visualization of the projections of the magnetic field vector in the x (orange) and y(green) over time for one nuclei.

It should be noted that ω_0 is the offset frequency (distance between RF and Larmor frequency), which makes it possible to detect more signals, as the offset would be different for each nuclei.

2.1.2 Signal generation and modelling

With the principle of signal detection covered, this subsection addresses the principles of how RF pulses generate an FID and how the time domain signals are modeled and subsequently transformed into frequency domain signals. The section also emphasizes why time-domain modelling may be a superior alternative to frequency-based models.

When applying a $\pi/2$ pulse, the magnetization vector is rotated from the z-axis towards the negative y-axis. Switching of the RF power causes magnetization to rotate in the transverse xy-plane, eventually returning to equilibrium. This would, due to the relaxation mechanism (see later in this section), in Cartesian coordinates, produce an upward narrowing spiral (see figure 2.3) which is akin to a spiral sink vector field in the form of

$$y(t) = M_0 \exp(-t/T_2) \begin{vmatrix} -\cos(\omega_0 t) \\ \sin(\omega_0 t) \end{vmatrix} \tag{2.4}$$

Where, M_0 is the magnetization, the term $\exp(-t/T_2)$ represents the T2 decay (see later in this section) and ω_0 is the frequency of the vector field projection.

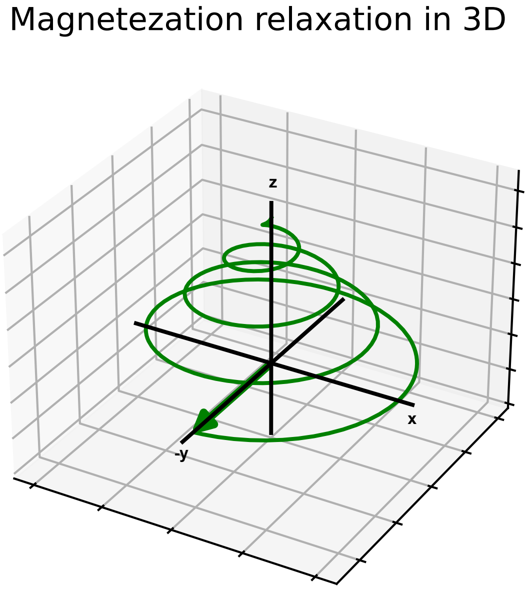


Figure 2.3: Visualization of the magnetization relaxation in 3D. The green vector represents the 90° flipped magnetization and over time the magnetization returns toward equilibrium in a spiral motion, becoming shorter in the x,y plane, returning to the positive z direction.

The x and y magnetization over time can thus be expressed (disregarding the decay term for now) as being proportional to the sine and cosine of the $\omega_0 t$ angle, defined as

$$M_y = -M_0 \cos(\omega_0 t) \tag{2.5}$$
$$M_x = \quad M_0 \sin(\omega_0 t) \tag{2.6}$$

Additionally, it is assumed that the pulses are generated such that the magnetization is rotated around the y-axis rather than the x-axis to avoid the negative y-component of equation (2.5). Furthermore, due to the receiver phase not being equal to the initial phase of the magnetization vector, a phase term (ϕ_0) must be added to the sinusoidal equations. The resulting harmonic of a single frequency (ω_0) rotated around the y-axis is thus defined as

$$M_y = M_0 \sin(\omega_0 t + \phi_0) \tag{2.7}$$

$$M_x = M_0 \cos(\omega_0 t + \phi_0) \tag{2.8}$$

The equations of (2.7) and (2.8) may be further simplified by viewing the x and y time series as projections of a complex sinusoidal, being the real and imaginary projection component respectively. Through Euler's formula, setting $M_0 = A$, the relation can be expressed as

$$y(t) = A(\cos(\omega_0 t + \phi) + j \sin(\omega_0 t + \phi)) \tag{2.9}$$

Here j is the imaginary unit and the plus sign of the imaginary part is chosen due to conventions[28]. The relation of equation (2.9) is further simplified by being written as a complex exponential function (in Hertz):

$$y(t) = A \exp(2j\pi\omega_0 t + j\phi) \tag{2.10}$$

Note that the formulation of equation (2.10) does not provide the magnitude (A) and phase (ϕ_0) directly. Instead, the complex amplitude of $A_c = A + j\phi$ is provided. Therefore to estimate time domain magnitude and phase the relation between A_c, A, and ϕ_0 is as stated by Smith[29] given as:

$$A = \sqrt{Re(A_c)^2 + Im(A_c)^2} \tag{2.11}$$

$$\phi_0 = arctan2(-Im(A_c), Re(A_c)), \tag{2.12}$$

where Re and Im indicates the real and imaginary part of the complex amplitudes. The estimations magnitude and phase through the means of equation (2.12) and (2.11) is not necessarily critical as later described in chapter 4, partly due to redundancy and the existence of a corresponding frequency height. The relations are still shown here for completeness and as an alternative for estimating amplitude ratios based on frequency domain amplitudes (see chapter 4).

Equation (2.10) is very important, as it forms the backbone for the harmonic term utilized in the NMR-Onion program described in chapter 4. The decay of the FID is mainly attributed to the relaxation[25], described in equation (2.13) as an exponential decay defined by an inverse T_2 rate:

$$\psi(T_2, t) = \exp(\frac{-t}{T_2}). \tag{2.13}$$

The shorter T_2, the faster signal decay. However, in general, the formulation of equation (2.13) is not favorable for modelling as an inverse constant lead to convergence issues, hence equation (2.13) is reformulated into:

$$\boldsymbol{\psi}(\alpha, t) = \exp(-\alpha t), \tag{2.14}$$

with $\alpha = \frac{t}{T_2}$. Here large α values would indicate a rapid decay (a short T_2), causing broad line shapes in the frequency domain. Combining equation (2.14) and (2.10) thus yields an exponentially dampened complex sinusoid:

$$y(t) = A \exp(2j\pi\omega_0 t + j\phi) \cdot \exp(-\alpha t) \tag{2.15}$$

However, equation (2.15) is still not enough to describe even the ideal FID, as nuclei subjected to magnetic resonance, omit more than one signal and the full FID is, as of such, a superposition/sum of all observed signals which may be expressed as

$$y(t) = \sum_{k=1}^{K} A_k \exp(2j\pi\omega_k t + j\phi_k) \cdot \exp(-\alpha_k t). \tag{2.16}$$

Here, K indicates the total number of sinusoids and $\omega_k, \phi_k, \alpha_k$ and A_k are vectors of parameters belonging to the k'th sinusoid (ω_k have had the 0 subscript removed to avoid dense confusing notation). Even so, the formulation of (2.16) does not account for none-ideal exponential decays and subsequent model modification has to be made (see chapter 4 and paper 2).

NMR spectroscopy is traditionally a visual science coupled to that of peak identification in the frequency domain, in which each local maximum in theory corresponds to the frequency of a dampened complex sinusoid in the time domain. The link between time and frequency domain is theoretically given by the discrete-time Fourier transformation (DTFT) of equation (2.17).

$$S(\omega) = \sum_{n=-\infty}^{\infty} y_n \exp(-j\omega n). \tag{2.17}$$

Though the signals are in theory continuous, they are in reality a result of finite digitized signal time points, and thus the continuous Fourier transformation of equation 2.17 does not apply in reality. However, the DTFT may be utilized to derive why the discrete Fourier transformation (DFT) applies and why zero filling is allowed for NMR signals in the time domain. For NMR an FID is not infinite, but rather sampled as a discrete periodic sequence (y_n) between 0 and N-1 discrete time points, which may be infinity zero-padded on each side, formulated as:

$$\tilde{y_n} = \begin{cases} y_n & 0 \leq \ n \leq N-1 \\ 0 & 0 > n \\ 0 & N-1 < n \end{cases} \tag{2.18}$$

While equation (2.18) mathematically make sense, NMR signals are purely causal w.r.t. sample points, and thus only causal zero padding[29] may be imposed (zeros may only be added when n>N-1). Hence by inserting equation (2.18) into equation

(2.17), substituting y_n with $\tilde{y_n}$, a suitable DFT for NMR signals with casual zero-padding (zero-filling) capabilities is obtained:

$$
\begin{aligned}
S(\omega) &= \sum_{n=-\infty}^{\infty} \tilde{y_n} \exp(-j\omega n) && (2.19)\\
&= \sum_{n=0}^{N-1} \tilde{y_n} \exp(-j\omega n) && (2.20)\\
&= \sum_{n=0}^{N-1} y_n \exp(-j\omega n) && (2.21)
\end{aligned}
$$

For this thesis, only uniform sampling acquisition schemes have been applied, which correspond to the points sampled being equally spaced. This results in the frequency (ω) in Hertz becoming $\omega = 2\pi\frac{k}{N}$, where k is the index for the bin in the frequency domain. Inserting into equation (2.21) yields the proper DFT for NMR signals

$$
\begin{aligned}
S(\omega) &= \sum_{n=0}^{N-1} y_n \exp(-j\omega n) && (2.22)\\
&= \sum_{n=0}^{N-1} y_n \exp(-2j\pi\frac{k}{N}n). && (2.23)
\end{aligned}
$$

The idea is that from the DFT an inverse DF transformation (IDFT) can be achieved in the form of:

$$y_n = \frac{1}{N}\sum_{k=0}^{N-1} S(\omega)\exp(2j\pi\frac{k}{N}n). \tag{2.24}$$

Mathematically, the relation of the IDFT and DFT would imply that signals detected in the time and frequency domain are completely invertible. While this is true, it should be noted that for NMR, pre-processing is carried out both prior to and post-DFT of the FID, which may disrupt one or the other domain in an unforeseen manner. Most notably prior DFT prepossessing is carried out by utilizing weight functions corresponding to the IDFT of the signal line shapes[30]. The simplest choice of weighting function would be the Lorentzian line shape, which in the time domain is equivalent to an exponential decaying weighting function. This would simply cause equation (2.23) and (2.24) to have an extra simple invertible function, which causes no alarm, except if the coefficient is too large signal which would disappear, but it will do so in a very predictable manner in both time and frequency domain (see more in the section 2.3).

Post-DFT processing operations on the other hand can make for some challenges, as the operations conducted in the frequency domain may have unforeseen consequences in the time domain. The specific operations are made such that ideally the peaks of the real spectrum are in absorption mode, that is a flat baseline is obtained and peaks do not exhibit phase errors. In the frequency domain the imaginary part of the spectrum is utilized, solving the problem of

$$S_0 \cdot [Re(S(\omega)) + jIm(S(\omega))] \cdot \exp(j[\phi + \theta]) \tag{2.25}$$

Here the goal is to minimize the distance between the phase angle ϕ and phase correction term θ such that only the $S_0 \cdot [Re(S(\omega)) + jIm(S(\omega))]$ remains. This is known as the frequency-independent phase or zero-order phase. In addition to zero order phase errors, frequency-dependent phase errors, known as first-order phase errors, occurs which are approximately proportional to offset distance (ω_d) from the transmitter[25], which can be expressed as:

$$\theta_1 = k \cdot \omega_d, \tag{2.26}$$

where the expression of θ_1 may be inserted into the same minimization problem as formulated in equation (2.25), replacing θ. It should be noted that the linear correct only works in nice cases, where the spectral lines are almost in phase.

From the phasing operations of equation (2.25) and (2.26), it is clear that an IDFT of the absorption spectrum would not yield an FID similar to the raw FID. This is not necessarily problematic, but while some algorithms like automated phase correction based on minimization of entropy (ACME)[31] do exist for automatic phasing, many NMR spectra are manually phased causing a large operator bias, leading to high spectral viability when attempting to formulate NMR based models. In addition, the absorption spectrum is based solely on the real part of the spectrum, which may cause a lowering of SNR from the lack of the imaginary spectrum[32]. Apart from phasing, baseline correction is a very common post DFT processing method. The method introduces a polynomial or spline correction to the spectrum (subtracting the polynomial or spline from the spectrum). This operation makes for a flat baseline in the frequency domain but may have unforeseen consequences in the time domain, such as Gibbs ringing effects, and phase and amplitudes distortions, potentially leading to loss of information w.r.t. original data. These effects occur because subtracting the baseline from the spectrum in the frequency domain is equivalent to convolving the baseline function with the time-domain FID signal[33]. Mathematically the operation of convolution may be written as the following:

$$y_{corrected} = IDFT(DFT(y_n) * DFT(b)), \tag{2.27}$$

where b is a polynomial baseline function of the from $b = b_1, b_2, ..., b_M$, with M number of coefficients attaining the values of $b_1, ..b_M$. The results of a baseline correction can be found in figure 2.4.

In figure 2.4, a linear baseline distortion was introduced to a synthetic dataset containing 5 signals each modeled as an ideal exponential decay (see equation (2.16)). The asymmetrically re-weighted penalized least squares (ARpls) algorithm was utilized to correct the baseline within the frequency domain (see section 2.3 for details). The simulation highlights how a correction within the frequency domain, may introduce nontrivial effects within the time domain (see figure 2.4 (C) and (D)). Summing up, the frequency and time domain are completely invertible if no prepossessing is

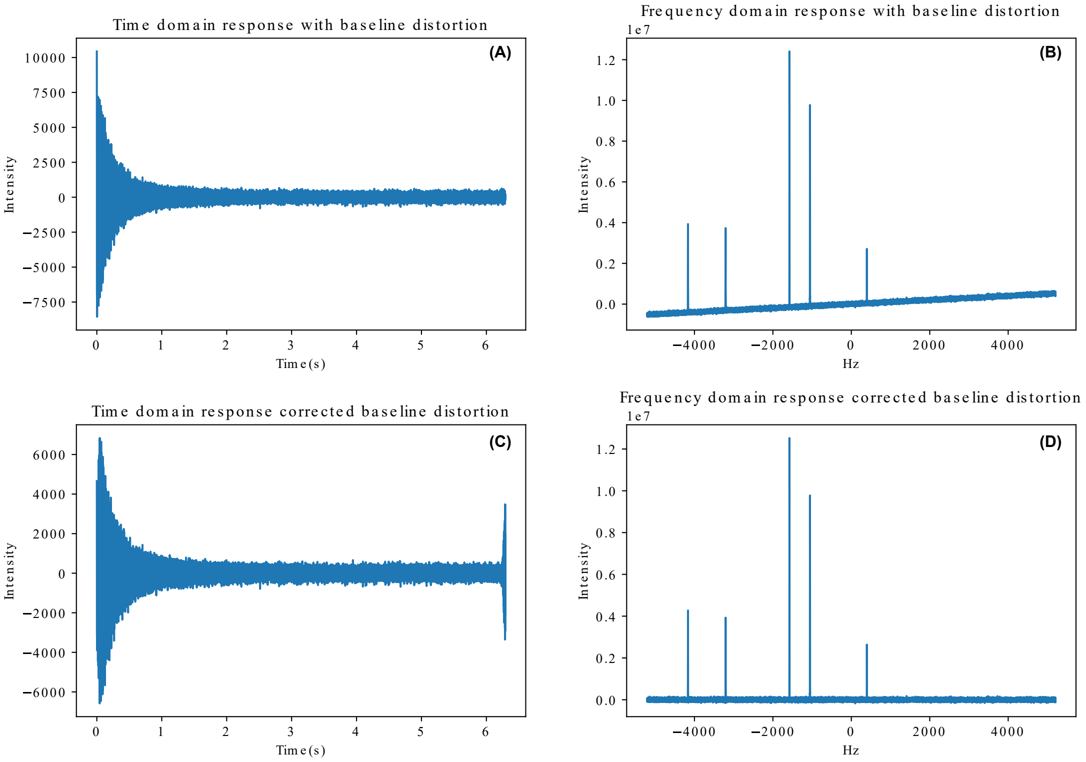


Figure 2.4: (A) The time domain response of a non-corrected baseline distortion. (B) The frequency domain response of a non-corrected baseline distortion. (C) The time domain response of a corrected baseline distortion. (D) The frequency domain response of a corrected baseline distortion.

carried out. However, as soon as post-DFT processing is introduced, the domains can be affected by operator bias and unforeseen consequences of correction methods. Therefore, if the goal is to be as close to raw data as possible a time domain model formulation should be favored with as little preprocessing as possible (see section 2.3 for more details).

With the reasoning of signal invertibility w.r.t. raw data in mind, a time domain model implementation was chosen as the approach for modelling NMR signals as described in 4. Here the signal is modeled from the raw time domain, where phase error and non-ideal shapes are accounted for, whilst signal detection is carried out in a resolution-enhanced processed spectrum (see the 2.1.3 section).

2.1.3 NMR sensitivity and resolution

When compared to techniques such as MS, NMR is as mentioned in the introduction a low sensitivity technique. This section clarifies the concepts of sensitivity, resolution, and noise while also emphasizing the usage of denoising (smoothing) and resolution/sensitivity boosting in the detection of peaks.

NMR spectroscopy has the ability to detect signals by recording an FID as stated in section 2.1.1 and 2.1.2, but as mentioned the magnetization relies heavily on which nuclei are investigated. For the ^{1}H nuclei utilized in this thesis, a natural abundance of 99.9844 % is found throughout all spin $\frac{1}{2}$ isotopes. For the 1/2 spin nuclei, the magnetization (M^*) is defined as:

$$M^* = \frac{\gamma \hbar N_{spin}}{2} P_r. \tag{2.28}$$

Here γ is the magnetogyric ratio, $\hbar$ is the reduced planks constant, N_{spin} is the total number of spins and P_r is the polarization ratio, which at thermal equilibrium may be expressed via the Boltzmann distribution

$$P_r = \frac{N_\alpha - N_\beta}{N_\alpha + N_\beta}. \tag{2.29}$$

The idea is that the higher the polarization ratio, the higher magnetization which in turn leads to higher signal levels. To put the ratio into perspective, the polarization obtained from an 800 MHz spectrometer in ^{1}H experiments is about 0.0065% at room temperature[34]. The polarization ratio may seem low, but the low ratio needs to be viewed in relation relative to the noise floor of the spectrum, that is the signal-to-noise ratio (SNR) matters more than the absolute signal. In other words, if no noise is present one still has a full signal despite being based on low magnetization. The signal-to-noise ratio of NMR signals can be expressed in many ways, but a common relation used in much of the literature defines the SNR as:

$$SNR = \frac{S}{2\sigma_N}, \tag{2.30}$$

or sometimes a version on based decibels (dB) is utilized

$$SNR_{dB} = 10 \log_{10}(\frac{S}{\sigma_N}), \tag{2.31}$$

where S is the signal amplitude and σ_N is the average noise amplitude. From the SNR equation, two options are presented to achieve higher levels of signal, either increasing the total signal or decreasing the noise level. The first option to increase signal strength is possible with the use of a technique such as dissolution dynamic nuclear polarization (DNP). The method has been developed to work around the Boltzmann distribution, artificially energizing protons via polarization[35]. The effects of DNP make it possible to track lower abundance nuclei (such as ^{13}C) and can enhance 1D analysis greatly as shown in the works of Frahm[36], where the targeted analysis of cancer metabolites in mice was made possible due to DNP. Despite the great results with DNP, the results come at the cost of having to have access to specialized equipment and expensive radicals for the process to work.

Apart from chemical solutions to increase sensitivity by signal boosting, other more affordable ”tricks” are also available which are based around signal processing improvements to gain higher sensitivity or SNR. The simplest SNR modification on a modern spectrometer may be the usage of signal averaging[37], in which the average of multiple spectral measurements is utilized such that the SNR is increased linearly with the square root of the number of spectra. The SNR is increased as the noise level goes towards its expected value of 0 with an increased number of scans, whilst the signal is constant throughout all spectra. The cost of this operation is acquisition time and if the tracked process is dynamic (e.g. reaction monitoring, generating new products during acquisition), a higher number of scans cannot be attained and different approaches must be employed. For the work of this thesis, no dynamic process was investigated, and as such a minimum of 128 scans were applied in every experiment. Another common method to increase sensitivity is through the use of weighting functions, which causes the noise part of the FID to be minimized. However, in this project targeted signals are small, and therefore great care must be taken when applying line-broadening functions (see section 2.3 for elaboration). A second very important concept for NMR is resolution, which is greatly attributed to the field strength of the equipment, the higher the field, the higher the resolution and the more peaks can be distinguished from one another. All spectral data generated for this thesis was acquired on a high field 800 MHz spectrometer equipped with a cryo-probe and thus are of high resolution. However, apart from increasing field strength, other signal processing tricks may enhance the overall resolution and thus increase the detection capabilities of spectra. Once again weighting functions are utilized, but instead of having purely negative functional values as the line broadening functions, a mixture of positive and negative functional values are applied in order to increase the beginning of the FID and lower the noise part in the later part of the FID. Some of the most common resolution enhancement functions are the Gauss-to-Lorentzian[38] or Sine-bell[39] functions, while sometimes more classical signal processing weightings functions such as Hamming[29] or Hanning[29] are applied. For this thesis, the Gauss-to-Lorentzian weighting function was chosen as the resolution enhancement method, being part of our own Python implementation of the R-based rNMRfind framework[40], which applies resolution enhancements, made for peak detection in the NMR-Onion algorithm (see chapter 4 for more information).

The last concept to introduce in this section is the usage of denoising algorithms also known as smoothing algorithms. These types of algorithms are useful when detecting peaks, as they aid in distinguishing random narrow spikes from actual peaks with a larger width. Many algorithms exist for smoothing and detection, some of the most popular within NMR are the usage of moving average filters as found in the Focus software[41], wavelets found within the Speaq software[42] and Savitzky-Golay (SG) filters applied for identifying first and second order derivatives to emphasize the finding of local maximas[43]. As part of the NMR-Onion framework highlighted in chapter 4, the derivatives off an SG-filter are utilized for peak detection, adapting the framework of rNMRfind[40] into our own Python implementation. The principles of the SG-filter, originally invented by Savitzky and Golay[44], are that the filter acts as a polynomial representation of a frequency or time domain response, which is smoothed or denoised to a varying degree, expressed with the following relation:

$$S_k = \sum_{i=\frac{1-n}{2}}^{\frac{n-1}{2}} C_i S_{k+i}. \tag{2.32}$$

Here S_k is the signal of the k'th frequency bin, C_i is the i'th the weighting coefficient for the S_{k+i} point, and n is the window length in which a polynomial of the m'th order may be applied. The specific polynomial order applied in this thesis was that of a third order and a minimum window length of five. Hence equation (2.33) can be written as:

$$S_k = \sum_{i=\frac{1-5}{2}}^{\frac{5-1}{2}} C_i S_{k+i} \tag{2.33}$$

$$= C_2 S_{k-2} + C_1 S_{k-1} + C_0 S_k + C_1 S_{k+1} + C_2 S_{k+2}. \tag{2.34}$$

In the case of evenly spaced point with Δ spacing, which occurs in uniform sampling used throughout this thesis, a closed form-solution to the normal equations of the smoothing polynomial ($\boldsymbol{P}$) does exist based on linear least squares fitting, given in matrix form as:

$$\boldsymbol{P} = (\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{V}^T\boldsymbol{S}. \tag{2.35}$$

Here $\boldsymbol{V}$ is a Vandermonde matrix of $n \times (m+1)$, where the n and m match the window and polynomial order length, such that the m'th column of $\boldsymbol{V}$ is given as

$$\boldsymbol{V}_m = \mathbf{1}, \boldsymbol{z}, \boldsymbol{z^2}, \boldsymbol{z^3}, ..., \boldsymbol{z^m}. \tag{2.36}$$

Here z is the normalized data point sequence of $z = \frac{S_k - \bar{S}}{\Delta}$, where $\bar{S}$ is the center point value and Δ is point increment distance. When $n = 5$, $m = 3$ and $\bar{S} = 0$, estimating $\boldsymbol{P}$ is straightforward as $z = \frac{1-n}{2}, .., \frac{n-1}{2}$, which leads to V being defined as:

$$\boldsymbol{V} = \begin{vmatrix} 1 & z_{1,1} & z_{1,2}^2 & z_{1,3}^3 \\ 1 & z_{2,1} & z_{2,2}^2 & z_{2,3}^3 \\ 1 & z_{3,1} & z_{3,2}^2 & z_{3,3}^3 \\ 1 & z_{4,1} & z_{4,2}^2 & z_{4,3}^3 \\ 1 & z_{5,1} & z_{5,2}^2 & z_{5,3}^3 \end{vmatrix} = \begin{vmatrix} 1 & -2 & 4 & -8 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{vmatrix} \tag{2.37}$$

After inserting $\boldsymbol{V}$ into equation (2.35) and subsequently estimating $\boldsymbol{P}$, the solution of $P_{0,k}$ can be expressed as:

$$P_{0,k} = C_2 S_{k-2} + C_1 S_{k-1} + C_0 S_k + C_1 S_{k+11} + C_2 S_{k+2} \tag{2.38}$$

$$= \frac{1}{35}(-3S_{k-2} + 12S_{k-1} + 17S_k + 12S_{k+1} - 3S_{k+2}). \tag{2.39}$$

Here, $P_{0,k}$ is the solution of the first row of equation (2.35), which corresponds to the smoothing polynomial[44]. The reaming solutions of $P_{1,k}..P_{3,k}$ may be found in the original paper of Savitky and Golay[44].

The smoothing polynomial of equation (2.33) is naturally altered with an increase of window length, the change may be coupled with the degree of denoising/smoothing and in the end linked to that of peak width. Now, the question remains of how window length and peak width are linked and how narrow a peak should be kept in data without being subjected to smoothing. There is most likely more than one way to express the relation between peak width and window length, but the linkage of the full width at half maximum (FWHM) with window length was chosen as it seemed a natural and simple choice linked to known NMR terminology. The relation is expressed as:

$$n = \frac{2N}{SW_{Hz}} \cdot FWHM, \tag{2.40}$$

where N is the number of frequency points, SW_{Hz} is the sample rate (or sweep width) in Hertz. To ensure the window length is odd equation (2.40) is tweaked with a floor operation:

$$n = \lfloor n/2 \rfloor + 1, \tag{2.41}$$

Where $\lfloor\rfloor$ indicates a floor operation. The key takeaway is that the FWHM is chosen such that any peak which has a width of less than the chosen FWHM would be subjected to smoothing. In practice, this would mean that the smaller the width (or FWHM), the fewer peaks would be subjected to smoothing. Appropriate values are set between 0.8 and 3 Hz (see chapter 4 for more details). With spectral smoothing covered (such that noise and signal can be separated), the next part addresses how a signal may be automatically detected based on a smoothed spectrum.

The main idea when detecting peaks is to identify the first and second derivatives of the SG-filtered spectrum for both the real and imaginary parts. The identification is performed by taking the derivative of equation (2.35) w.r.t. z, which for the smallest five-point window and third order polynomial at center point $\bar{S} = 0$ utilized in this thesis yields:

$$\tilde{S}|_{z=0} = P_0 + P_1 z + P_2 z^2 + P_3 z^3 = P_0 \tag{2.42}$$

$$\frac{d\tilde{S}}{dz}|_{z=0} = \frac{1}{\Delta} P_1 + 2P_2 z + 3P_3 z^2) = \frac{1}{\Delta} P_1 \tag{2.43}$$

$$\frac{d^2\tilde{S}}{d^2 z}|_{z=0} = \frac{1}{\Delta^2}(2P_2 + 6P_3 z) = \frac{1}{\Delta} 2P_2, \tag{2.44}$$

where P_0 of equation (2.42) is equal to equation (2.39) when transforming back from $\tilde{S}$ to S. Here it was utilized that S may be written into a general polynomial form of $\tilde{S} = P_0 + P_1 z + P_2 z^2 + P_3 z^3$. The first and second derivatives expressions of P_1 and P_2 can likewise be found from the solution of (2.35) or they can be looked up in the table of the original article of the SG-filter[44]. Here the solutions are

$$P_1 = \frac{1}{12\Delta}(S_{k-2} - 8S_{k-1} + 8S_{k+1} - S_{k+2}) \tag{2.45}$$

$$P_2 = \frac{1}{7\Delta^2}(2S_{k-2} - S_{k-1} - 2S_k - S_{k+1} + 2S_{k+2}). \tag{2.46}$$

Now that how to find the deviates have been covered, the next step is realizing why this information can be utilized to automatically identify peaks and why both real and imaginary spectral parts are needed. The latter is theoretically not needed as the imaginary and real part of the spectrum should consist of the same signals. However in reality small deviations may produce subtle, but important difference[28][40] and in addition, the overall SNR is improved from applying both parts[40]. When it comes to retrieval of information from the first and second derivatives, the first derivative of equation (2.43), will be zero at maximum peak height, while the second order derivative of equation (2.44) exhibits reduced FWHM compared to the original SG-filtered signal, such that resolution is enhanced and overlapping peaks can be resolved at local minima. The visual response for smoothing out noisy spectral data is shown in figure 2.5, detecting two overlapping peaks via the first and second-order SG-derivatives is exemplified in figure 2.6 and figure 2.7

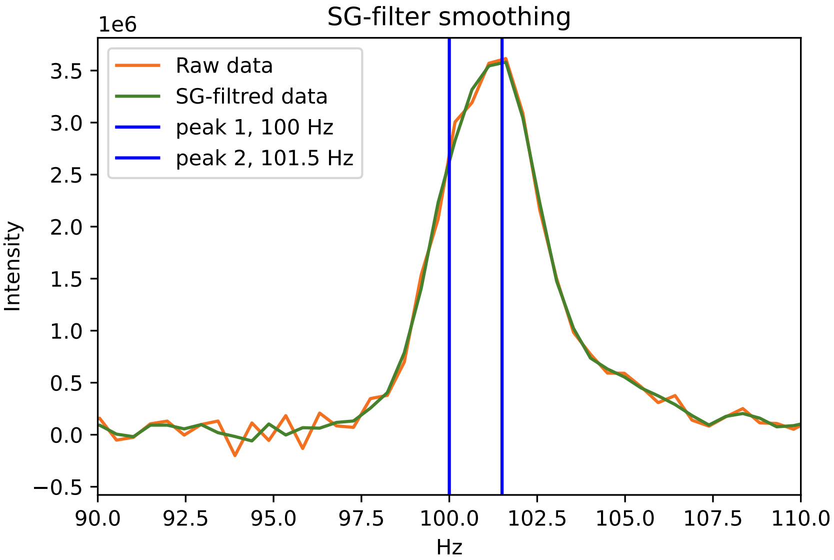


Figure 2.5: visualization of the SG-filter effects applying equation (2.38)

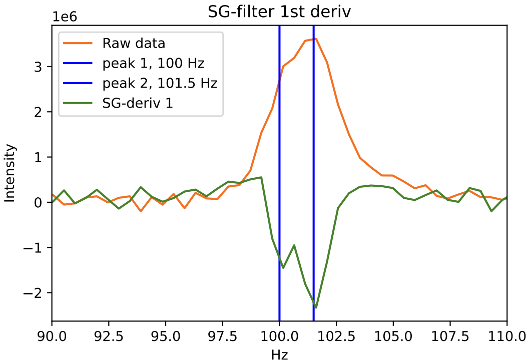


Figure 2.6: First derivatives of SG filtered spectral data, applying equation (2.45)

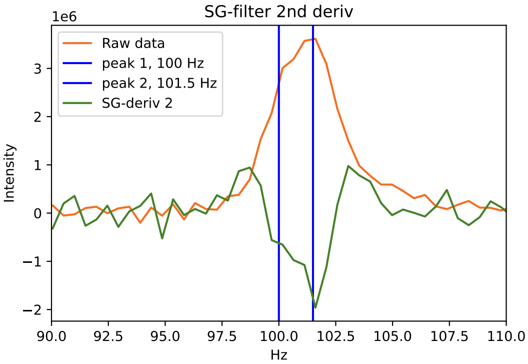


Figure 2.7: second derivatives of SG filtered spectral data,applying equation (2.46)

From figure 2.5 it is observed how the SG-filter is removing unwanted noise spikes around the targeted peak, minimizing the potential of false detection. The usage of the SG-filtered first and second derivatives successfully detects the overlapping frequencies, which would be very challenging to detect from visual inspection.

Summing up the sensitivity and resolution of an NMR spectrum can be enhanced by properer application of weighting functions. The enhanced spectrum can be coupled with an SG smoothing algorithm in order to automatically detect and separate peaks from noise. The detection and enhancement are very important within metabolomics as many signals are overlapping and are challenging to distinguish manually. In addition, the detection also serves a second purpose in being initial estimates for frequency estimation within the NMR-Onion algorithm of chapter 4.

2.2 NMR data acquisition

Three types of data acquisition were utilized within this thesis being, the zg, zg30, zgespg, and 1D nosey experimental setups. For the most part, the zg experiment was utilized as the default experiment. Here, if not otherwise stated, 32k complex time points were sampled at a sweep width (sample rate) of 13 ppm with a total acquisition time of 3.14 seconds. The relaxation delay (d1) was set to 2.0 seconds to ensure adequate relaxation before applying the excitation pulse and as previously stated a minimum of 128 scans were acquired to ensure higher SNR for each experiment.

The data from the zg30 experiment was also tested in this thesis. Here the pulse angle was set at 30° with equal sweep width and the number of points sampled as in the zg experiment. The zg30 experiment was included to test if the algorithm of NMR-Onion (paper 2) would be affected by a different pulse angle. The results can be found in case study 2 of paper 2.

Finally when water concentrations were very high (90% H_2O to 10% D_2O) the zgespg experiment was chosen over the zg30 and zg set up to suppress the very intense water peak. Here the effects of the shaped pulses are that targeted signals (in this case the water resonance) are suppressed in the spectrum and FID. The removal of the water regions was necessary as the models developed within this thesis do not account for the much larger resonances which interfere with the shape of FID in a non-trivial manner[45]. The downside of this type of experiment is that signals within close proximity to the water resonance are also impacted, making parts of the spectrum inadequate for analysis. In addition, the baseopt rectangular filter was utilized for every experiment, the filter is further discussed in the next subsection. Finally, the preprocessing of all experiments was carried out according to the scheme outlined in figure 2.8 found within the next subsection.

2.3 Linking NMR and metabolomics

Within targeted analysis, the goal is often to associate the presence and possibly concentration of specific metabolites based on a set of experimental conditions. Hence, a general hypothesis linking NMR spectral data ($\boldsymbol{Y}$) and vector of experimental factors ($\boldsymbol{\theta}$) would be set up as the following

$$H_0 : \boldsymbol{\theta_i} = \boldsymbol{\theta_j} \tag{2.47}$$

$$H_1 : \boldsymbol{\theta_i} \neq \boldsymbol{\theta_j} \tag{2.48}$$

Here examining if any of the factors levels (i and j or possibly more levels) within $\boldsymbol{\theta}$ are significantly different from one another, leading to a rejection of H_0. It should be noted that the experimental factor of $\boldsymbol{\theta}$ may originate from both sample preparations and other experimental parameters. To be able to conduct the hypothesis testing, a model needs to enable the link between the data ($\boldsymbol{Y}$) and $\boldsymbol{\theta}$. A very simple notation to express the relation may be formulated as

$$\boldsymbol{Y} = f(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \tag{2.49}$$

Where $f(\boldsymbol{\theta})$ represent the model linking $\boldsymbol{\theta}$ (factors) to the spectral data outcome and $\boldsymbol{\varepsilon}$ is a vector of model residuals.
The model function of $f()$ can be any function type in principle, but when conducting hypothesis testing, the models are often linked to that of response surface methodology (RSM)[46], spanning everywhere from simple linear fixed effect models[47] to complicated generalized non-linear mixed effect models[48]. Throughout the metabolomics literature, various examples of different kinds of RSM models have been applied when conducting targeted analysis. One example is medium optimization[49] in which RSM ensured the stability of the metabolomic response. Another example is the usage of RSM to optimize extracting methods and solvent composition in the metabolite profiling of Azadirachta indica plants[50]. Further elaboration of RSM w.r.t. DoE methodology in metabolomics is given in chapter 3 and paper 1, whilst the statistical analysis of DoE-based studies is emphasized in chapter 5.

A very unique feature of spectral NMR data is, as stated in the introduction, that no immediate spectral data outcome is provided from the raw data, spectral processing has to be performed in order to gain chemical insight, corresponding to metabolite identification highlighted in figure 1.2. In targeted analysis, the goal is to identify fingerprint signals of specific metabolites in order to prove the presence of the compound in question and if possible also the concentration of the targeted compound.

To get meaningful comparable spectral data outcomes from each spectrum acquired within a metabolomic study, prepossessing has to be carried out as stated in the workflow in figure 1.2. The goal is to ensure each spectrum is phased, the baseline is flat, proper resolution is obtained and all spectra are aligned within the targeted region. The order of the prepossessing steps utilized in this thesis is shown in figure 2.8, which is inspired by the route suggested by NMRprocflow[51].

For this thesis, all spectra were acquired on Bruker NMR spectrometers, which enabled the usage of the baseopt rectangular filter[52]. This filter fixes the group delay of the acquired FID (see figure 2.9), ensures a mostly flat baseline, and reduces 0 order phase influence, leaving only the 1st order phase to be corrected.

For additional baseline correction (outside Topspin as seen in figure 2.8), the asymmetrically re-weighted penalized least squares smoothing algorithm (ARpls)[53] was

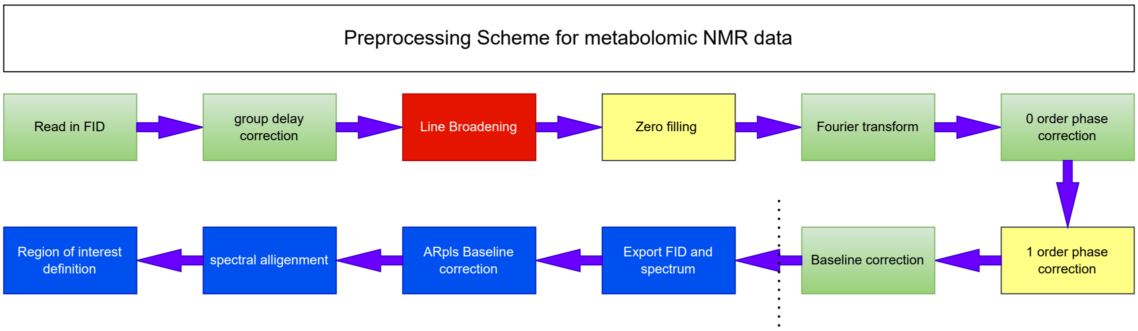


Figure 2.8: Visualization of the prepossessing steps for metabolomics NMR data. The green color indicates steps automatically handled by the Bruker Topspin Baseopt rectangular filter, yellow indicates user input needed in Topspin, while the dashed line indicates the programming environment outside Topspin with custom functions marked in blue implemented in Python. The red color on line broadening marks a step that generates symmetric peak shapes (Lorentzian, Gaussian etc.), but may cause small peaks to disappear

applied, which is formulated as

$$\boldsymbol{z_{ARpls}} = (\boldsymbol{W} + \lambda \boldsymbol{D}^T \boldsymbol{D})^{-1} \boldsymbol{W} \boldsymbol{y} \tag{2.50}$$

Here $\boldsymbol{z_{ARpls}}$ is the smoothed baseline, $\boldsymbol{D}$ is the second order distance matrix[53], $\boldsymbol{W}$ is a diagonal matrix containing asymmetric weights, such that regions with a signal is treated with a logistic weight and regions containing only noise has its weight set to 1. Finally, λ controls the degree of penalty, the higher value, the more influence of the weights. The algorithm was implemented in Python and the source code is found as part of the NMR-onion Algorithm (see chapter 4) deconvolution software package on GitHub of the author (www.github.com/Mabso1).

From figure 2.8, one should note the red box of line broadening (LB). The step would lead to visually more uniform peak shapes. However, this comes at the cost of masking closely spaced peaks originating from compounds present in minute amounts. In addition, smaller peaks close to the noise floor may also disappear if broadening is used too extensively. Even so, some LB is needed if the signals should have some shape consistency, therefore all data have been transformed with a very small exponential decay of 0.3 Hz during post-acquisition, such that all time series FIDs are expressed according to equation (2.51).

$$y(t) = y_{raw}(t) \cdot \exp(-R \cdot t) \tag{2.51}$$

Here $y_{raw}(t)$ are the FID data observed after acquisition and $y(t)$ is the FID after apodization with $R = 0.3$.

The last step prior to analyzing the region containing the targeted signals, is to perform spectral alignment for every sample, such that the sample-to-sample variance is reduced with respect to small spectral shifts. A popular automatic spectral

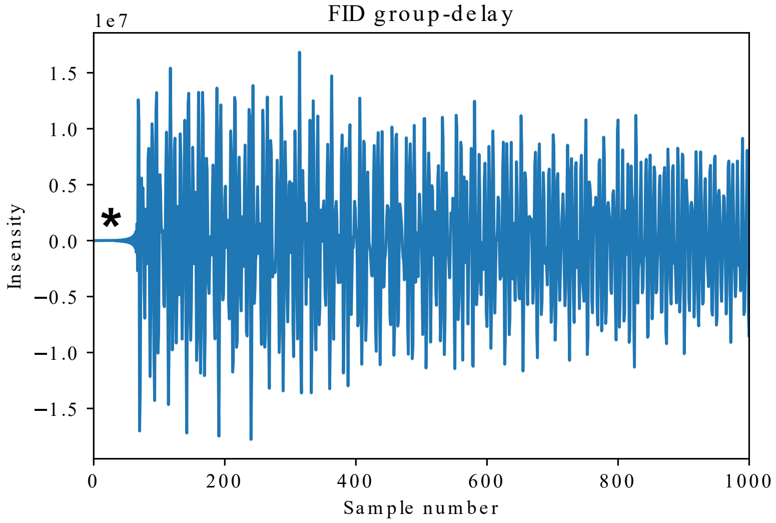


Figure 2.9: The FID group delay introduced by the experimental setup of the spectrometers is visibly shown at the beginning of the FID marked with a black * symbol. The delay is subsequently removed by the baseopt filter.

alignment method is the Icoshift method[54], capable of rapidly aligning hundreds of spectra. Following the alignment of all acquired spectra, the next step of spectral prepossessing comes in the form of isolating the targeted region of interest (ROI). Traditionally this has been done via a bucketing/binning process e.g. NMR-procflow[51] which has later been automated via intelligent bucketing[55]. The advantages of bucketing/binning are the rapid grouping of spectral signals along with well-described multivariate methods for analyzing the grouped spectral buckets (see Chapter 5 for more details). One example of hypothesis testing being conducted within targeted analysis based on bucketed/binned NMR data is found in the works of Zhu[56]. Here hypothesis testing was performed based on a linear mixed effect model utilized to distinguish metabolite profiles from cognitively normal patients and Alzheimer's diseased patients. Another example of hypothesis testing within targeted analysis is found in the works of West[57], where cardiac metabolism was investigated through supervised PLS modelling of the concentrations of proline and methyl-histidine found in healthy and dilated cardiomyopathy-stricken mice.

The downside of binning/bucketing comes in the form of sensitivity, as small peaks are often lost in the process[58]. The loss of information is not problematic if the targeted metabolite is linked to larger peaks, but when trying to capture low SNR and highly overlapping peaks, the method of binning/bucketing is inadequate. Therefore, in the case of low SNR overlapping targets, other types of regional isolation methods must be applied when prepossessing the spectral data.

One such method capable of retaining spectral information, of a ROI, to a higher extent than binning/bucketing comes in the form of digital band-pass filtering[29]. The main idea behind the bandpass filtering approach is to have a function that passes frequencies within a certain band range (B) while setting all other frequencies outside the boundaries of B to 0. The naive perfect way of achieving the perfect bandpass filter would be the application of the rectangular filter[29] which in the frequency domain would be equal to a rectangular function (F) (see figure 2.10). The function is 0 when the frequency ω is outside the lower (S_l) and upper (S_u) frequency band as stated in equation (2.52).

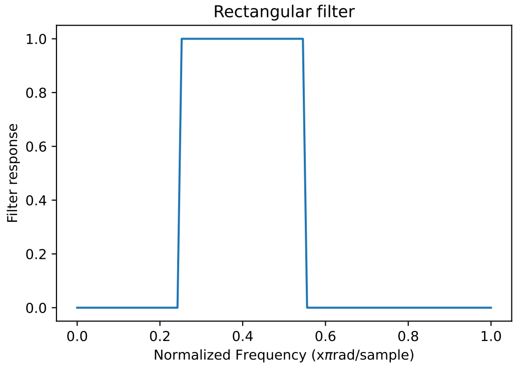


Figure 2.10: The ideal Rectangular band-pass filter response in the frequency domain, showing conceptual perfect response by having no transition bands and no leakage into any of the stop-bands

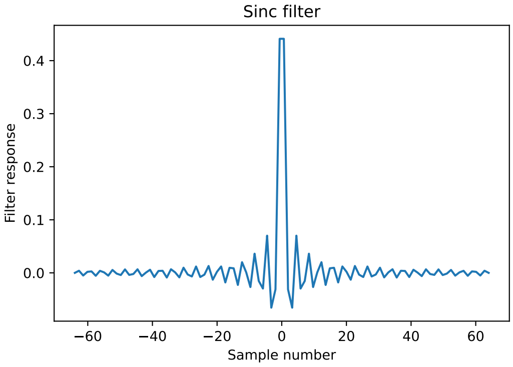


Figure 2.11: The IFT of the ideal Rectangular band-pass filter response in the time domain. The Gibbs ringing phenomenon is observed across all of the time-domain.

$$F = \begin{cases} \omega = 0 & S_l > \ \omega > S_u \\ \omega = 1 & S_l \leq \ \omega \leq S_u \end{cases} \tag{2.52}$$

Unfortunately invoking the filter of equation (2.52) would cause ringing artifacts known as the Gibbs phenomenon[29], which causes ripples to appear in the time domain. The explanation as to why the ripples appear lies in the fact that the inverse discrete Fourier transformation (IDFT) of equation (2.52) results in a sinc function as seen in (2.53) and plotted in figure 2.11

$$sinc(\omega) = \frac{\sin(\omega t)}{\omega t} \tag{2.53}$$

From equation (2.53) and figure 2.11 it is observed that the function spans all of the time domain, and thus introduces artifacts across all signals.

To mitigate the problem of ringing, several options within signal processing are available. One of the simplest options is to combine the sinc filter with a window function with a specific transition band. By convolving the windowed function and sinc filter, a finite impose response (FIR) filter is achieved, which is shown in figure 2.12 for the frequency domain and figure 2.13 for the time domain. When comparing the FIR filter with the previous ”ideal” filter, it is observed that the ringing artifacts have been reduced, but at the cost of the generation of a transition band. The FIR filter has been applied as part of the popular CRAFT[59] algorithm used heavily in metabolomics, where a Blackman window function has been combined with a presumably modified sinc function, emphasizing presumably, as the exact specifications of the filter coefficients are kept as a secret due to the program being part of commercial software.

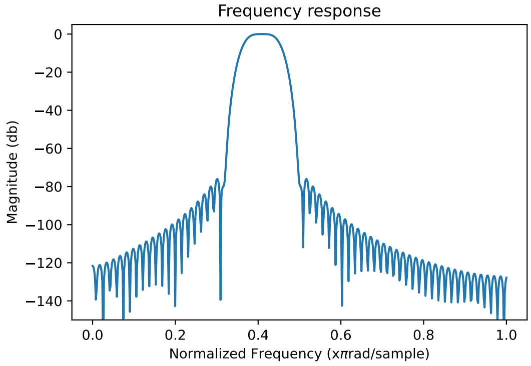


Figure 2.12: The response of the windowed FIR band-pass filter response in the frequency domain, showing transition bands and small leakages into the stop-bands.

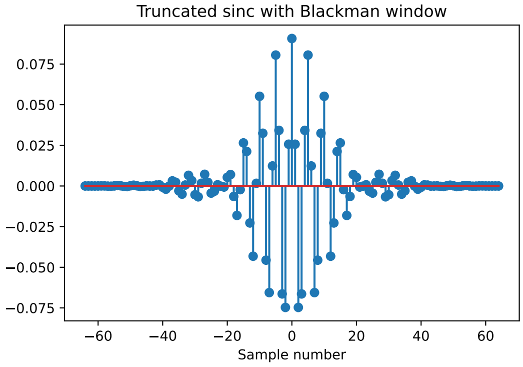


Figure 2.13: The IFT of the windowed FIR band-pass filter response in the time domain. The Gibbs ringing phenomenon is diminished when compared to the ideal filter.

For the isolation of ROIs, FIR-based filtering was originally attempted but ultimately failed. Instead other options were considered and from an extensive search of the literature, an example of a Gaussian filter combined with synthetic noise generation[60] was found, adapted, and modified to fit into our own prepossessing scheme as part of the NMR-Onion algorithm in chapter 4.

When comparing the Gaussian filter of equation (2.54) to the sinc function of equation (2.53), the filter does not impose wriggles in the time domain and thus does not cause ringing.

$$\boldsymbol{g}(t) = \sqrt{\frac{\sigma_g}{\pi}} \exp(-\sigma_g t^2) \tag{2.54}$$

However, unlike the sinc-based FIR filters, the Gaussian filter produces leaks in the stop band (unwanted signals are getting through) and does not filter as well in its native form of equation (2.54) as the FIR filter. Hence, modifications of equation (2.54) are made in order to gain the same filtering properties as the FIR filter with respect to stop band stability, while retaining the advantages of having no ringing artifacts. First, the discrete Fourier transformation of (2.54) is found, resulting in the following:

$$\boldsymbol{g}(\omega) = \frac{1}{\sigma_g\sqrt{2\pi}} \exp\left(-\frac{\omega^2}{2\sigma_g^2}\right) \tag{2.55}$$

The format of equation (2.55) suggests that the center frequency of the filter should be set at 0, while it is possible to Fourier shift targeted ROIs to the center, the addition of a location parameter ω_c makes for a simpler choice, resulting in (2.55) being modified into

$$\boldsymbol{g}(\omega) = \frac{1}{\sigma_g\sqrt{2\pi}} \exp\left(-\frac{\omega_c - \omega^2}{2\sigma_g^2}\right) \tag{2.56}$$

The filtering function of (2.56) is still in need of modification as leakages into the stop-bands are still a challenge in the current form. Hence, to mimic the rectangular filter, a super-Gaussian filter[61] is constructed based on (2.56), expressed as

$$\boldsymbol{sg}(\omega) = \exp\left(-2^{p+1} \cdot \left(\frac{\omega_c - \omega^2}{B}\right)^p\right) \tag{2.57}$$

Here, B corresponds to the bandwidth, ω_c is the center frequency and p is the filter order. In general the higher the filter order (see figure 2.14), the closer to the rectangular filter. For all ROI selections in this thesis, p is set to 40, as higher-order did not improve the filtering output. The effect of the filter order can be seen in figure 2.14, showcasing a bandpass filter from 2.40 ppm to 2.65 ppm.

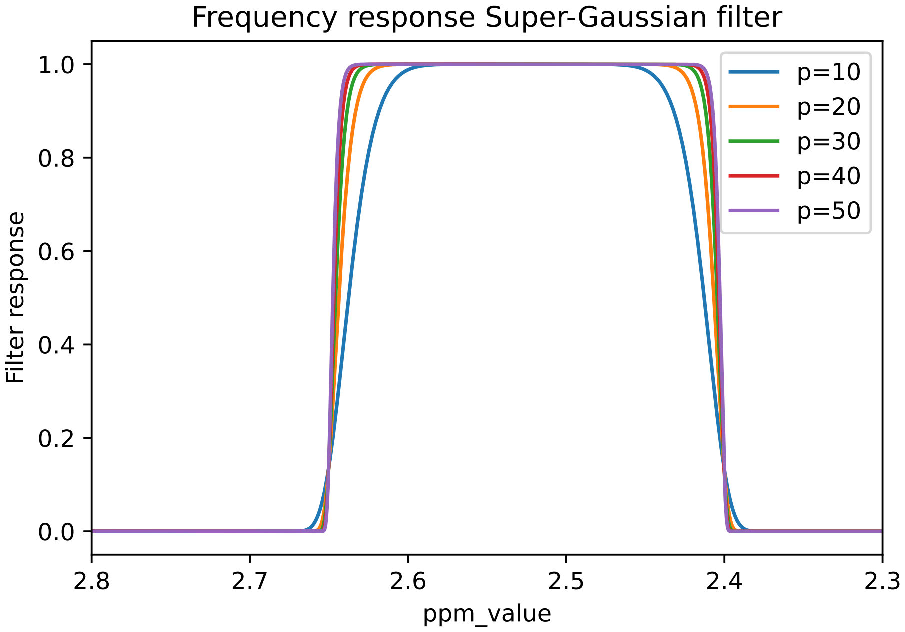


Figure 2.14: Visualization of the filter order of the super-Gaussian filter within the frequency domain. Exemplified with a bandpass from 2.65 to 2.4 ppm.

An additional very attractive property of this specific filter and the reason why it was chosen over other implementations, is as previously mentioned, its ability to incorporate a synthetic noise floor outside the band-pass region. The synthetic noise floor enables the differing of noise vs signal to be made within the band-pass region, such that when modelling the spectrum, the models can be evaluated by observing the residuals, comparing these to the overall noise floor (see chapter 4). To generate the synthetic noise of the filter, the idea is to take advantage of the spectral signal sparsity of a 1D NMR spectrum, picking a region containing only noise and from the noise region simulating the overall noise level (σ_{noise}) in the full spectrum. For the simulation, it is assumed that all noise ($\boldsymbol{W_s}$) of the spectrum may be represented as a vector of average white Gaussian noise (AWGN) described as

$$\boldsymbol{W_s} \sim N(0, \sigma_{noise}) \tag{2.58}$$

$\boldsymbol{W}$ is thus generated by sampling from equation (2.58), which conditions on the estimated noise level. In order to estimate the noise level careful steps must be taken not to overestimate the noise, as overestimation might lead to signals being seen as noise, potentially causing a loss of information. One particular effect causing the overestimation of the noise levels comes in the form of baseline artifacts causing unwanted variation. To mitigate baseline effects, we employ our implementation of the ARpls prior to estimating the variance. The effects of the ARpls can be seen in figure 2.16 vs the uncorrected spectrum in figure 2.15, in which the mean value is shifted towards 0 and overall variation goes down.

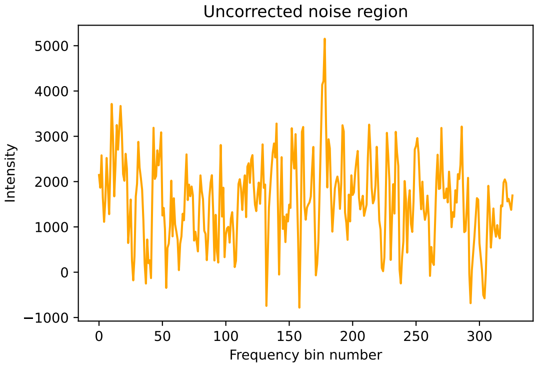


Figure 2.15: An example of a raw noise region selected from a signal-free region of a spectrum

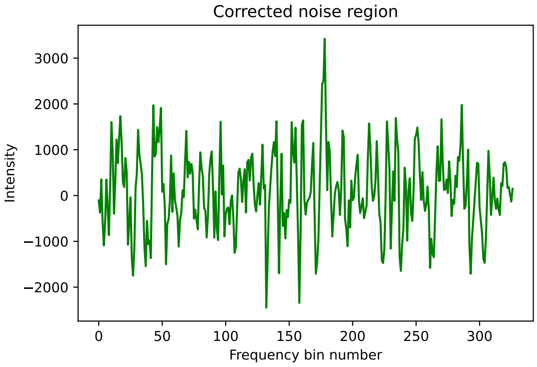


Figure 2.16: An example of an ARpls corrected noise region selected from a signal-free region of a spectrum

To ensure additional precision of estimation, a further re-sampling scheme is added drawing 1000 realizations of the noise spectrum, assuming a normal Gaussian distribution stated as

$$s_r \sim N(0, \sigma_{ARpls}) \tag{2.59}$$

where s_r is the r'th noise signal realization of the r'th draw from the distribution of s_r, with the standard deviated (σ_{ARpls}) based on the ARpls corrected noise ($r = 1, 2, ...1000$). The mean of the 1000 realizations are then computed, generating a more robust expression of the noise vector in equation (2.58) stated as

$$\boldsymbol{W_s} = \frac{1}{R}\sum_{r=1}^{R} s_r \tag{2.60}$$

The robust formulation of equation (2.60) is thus utilized as a synthetic noise floor for the stop-band regions of the filter. At last, the complete filter can be expressed, combining the robust noise of equation (2.60) and the super-Gaussian filter function of equation (2.57), resulting in

$$\boldsymbol{H(\omega)} = \boldsymbol{sg}(\omega) + \boldsymbol{W_s} \tag{2.61}$$

Like any FIR filter, the super-Gaussian $\boldsymbol{sg}(\omega)$ may be convolved with the frequency domain, as an example the bandpass filter of figure 2.14, may be applied to a spectrum (figure 2.17) resulting the in the frequency and time domain responses, shown in figure 2.18) and 2.19 respectively.

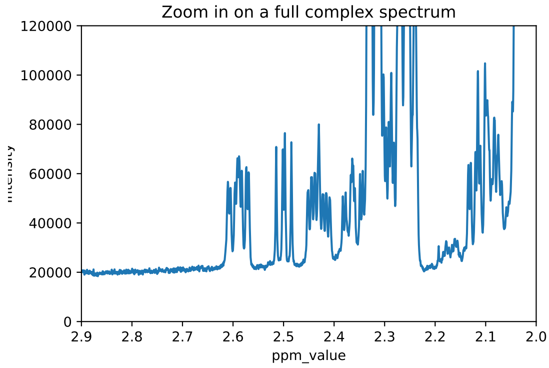


Figure 2.17: A zoom-in on the region close to the band-pass region of the super-Gaussian filter outlined figure 2.14

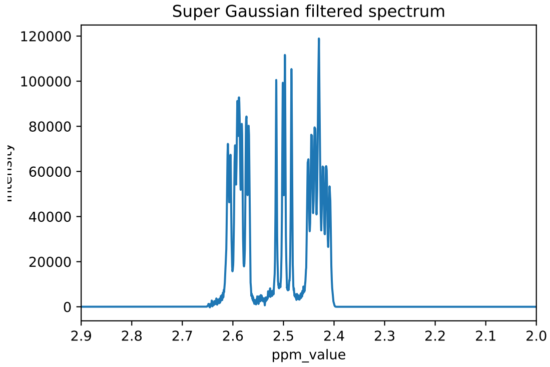


Figure 2.18: A zoom-in on the resulting convolution between the SG-filtered region

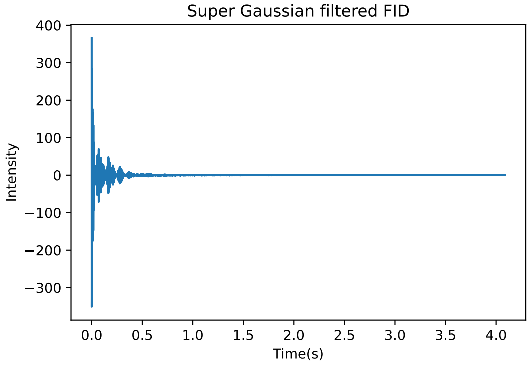


Figure 2.19: The time domain representation of the band-pass filtered region

The key point is that the time domain and frequency domain exhibits very little to no ringing artifacts nor any leaks in the stop bands, making the SG-filter a suitable filter for representing ROIs within a spectrum. Now the addition of the noise vector $\boldsymbol{W_s}$, is from a visual perspective obsolete, but from a computational perspective, the noise floor is required for the fitting algorithm of NMR-Onion to reach convergence[60]. The addition of the noise vector to the filtered region of figure 2.18 and 2.19 is visualized in figure 2.20 and 2.19 respectively

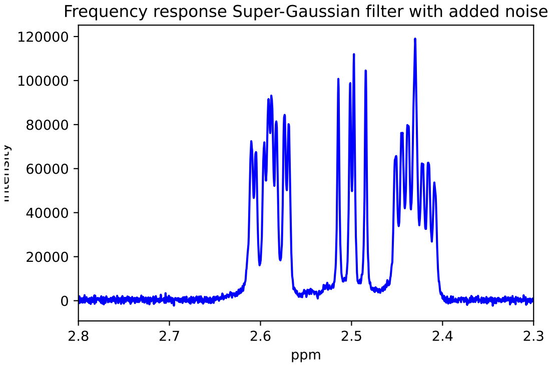


Figure 2.20: Noise added to the spectrum of figure 2.18, zoomed in for a better view.

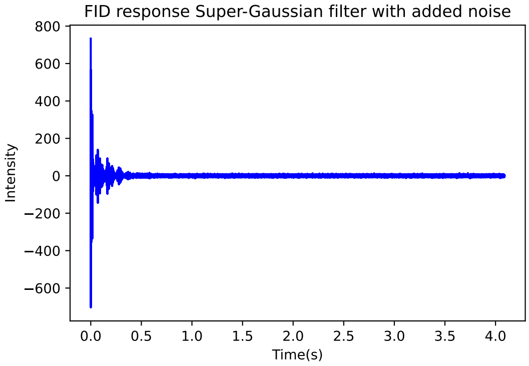


Figure 2.21: Inverse Fourier transformation of figure 2.20.

The above filtering process can then be carried out for multiple spectral samples within an experiment which are then aligned utilizing the icoshift algorithm and sent for further analysis as stated in the workflow of figure 1.2 found in the introduction chapter.

Apart from pre-processing data prior to analyzing and extracting spectral information. Crucial parts of the workflow in figure 1.2 regarding the construction of a proper informative design space (box 2 of the workflow) and assurance of data quality (box 3 of the workflow) have been left out. One might have all the right tools for analyzing and extracting information, but if crucial parts of a design space are missing or information derived from said space is heavily influenced by nuisances factors, masking the influence of the treatment effects placed upon the design space, little is gained and conclusions are meaningless. Therefore, the next section reviews how a combined DoE and SQC may aid in ensuring optimal design space and a reduced influence of nuisance factors within metabolomic studies.

Chapter 3

Designing metabolomics experiments

Along with paper 1, this chapter provides an insight into design of experiment (DoE) and statistical quality control (SQC) applied within metabolomics. Key elements underlining the importance of SQC and DoE are highlighted in the following paragraph, emphasizing their importance within metabolomics. It should be noted that DoE is not limited to NMR-based metabolomics and can be easily applied in the setting of other metabolomics techniques such as MS (see examples within paper 1).

3.1 Design of experiments

Metabolomics is a very broad discipline applied within many scientific areas ranging from ecology to personalized medicine. Within the area of metabolomics, the focus on DoE-based data generation has steadily increased over the last 20 years (see paper 1, see appendix E). The statistical methodology of DoE is utilized to optimize designs and improve the accuracy and reliability of data obtained from experiments[62]. DoE involves systematic variations of multiple experimental parameters and the analysis of their effects on the response variable, which in the context of metabolomics often involve a metabolite or a group of metabolites[63]. The method of DoE is known to exceed the one variable at a time (OVAT) method[64], as OVAT does not account for the effects of interactions nor shares the efficiency of design space sample organization as exploited by the DoE methods[64][62].
The specific design of an experiment is highly related to the metabolomic questions (see figure 1.2), but can as stated in paper 1 be generalized into the categories of either optimization or screening.

For NMR-based metabolomics, DoE can be used to optimize experimental factors such as the choice of solvent, pH, temperature, and acquisition parameters. By systematically varying these factors using DoE, one may identify the optimal conditions that provide the highest signal-to-noise ratio (SNR), resolution, and the most reproducible results. This can lead to improved sensitivity, resolution, and accuracy of NMR spectra, which in turn can improve the identification and quantification of metabolites in complex biological samples. The most common types of designs applied within metabolomics are either that of the Box-Behken design[46] or variations of the central composite design[65] (see paper 1 for more detail).

In addition to optimizing experimental factors using DoE, the screening of factors is another important application of statistical tools in metabolomics. Screening is used to identify the most important factors that contribute to the variability in metabolomic data. These can then subsequently be prioritized for further optimization using DoE.
During the screening process, a large number of factors are systematically varied using a fractional factorial design[62][66], Plackett-Burman design[67] or other screening design types (see paper 1). This allows for the efficient evaluation of a large number of factors with a limited number of experiments.
In NMR spectroscopy the response variable (Y), may be the NMR spectral amplitude response, which is then analyzed using statistical methods. The most common methods are principal component analysis (PCA)[68], partial least squares (PLS)[69], generalized linear models (GLM)[48] or ANOVA simulant component analysis (ASCA)[70], which all enable the identification of factors that have the greatest impact on the variability in the full data or specific regions within the spectrum. In other words, the aforementioned method links the design space of the experimental design with the hypothesis testing introduced in section 2.3 and in chapter 5, addressed in a case study in the preceding theoretical section.

In summary, by using a combination of screening and optimization techniques, researchers can develop robust and reliable experimental protocols for NMR-based metabolomics, which can facilitate the identification and quantification of metabolites in complex biological samples. Specifically, paper 1 recommends a workflow based around DoE, linking the choice of design to the task of either optimization or screening. Furthermore, the workflow also accounts for the number of factors, factor levels, and type of factor (eg categorical or numerical), matching appropriate design algorithms with the variables of the study at hand.

3.2 Statistical quality control

Another important field within the metabolomic workflow is that of statistical quality control(SQC). This is a powerful tool that can be used to monitor the quality of data generated in metabolomic experiments, allowing one to quickly identify and address any potential issues that may impact the validity and reproducibility of results. In this context, SQC involves the use of statistical methods to evaluate the precision, accuracy, and robustness of analytical methods used in metabolomics, as well as to assess the quality of data generated from those methods. This can help to ensure that the results obtained from metabolomics experiments are reliable, consistent, and of high quality, which is essential for advancing our understanding of biological systems. Some of the more common challenges affecting data reproducibility are highlighted in Paper 1 along with metrics for quantifying and evaluating these effects. The challenges highlighted in paper 1 may in short be summarized as batch-to-batch effects[71], internal standard effect on biological matrix[72], and sample preparation variations. In addition, paper 1 reviews methods for evaluating the reproducibility of metabolomic data, with the intent of uncovering the potential influence of the aforementioned challenges. Here it was found that the three most common methods of quality evaluation within metabolomics were that of pooled quality control samples coupled with PCA score plots, residual standard deviation (RSD) of peaks, and relative log abundance plots. It should be noted that the evaluation methods do not fix the lack of reproducibility but rather makes one aware that the results of a DoE might not have occurred due to treatment interventions, but rather other sources of technical variations (see paper 1 for details). To mitigate the potential lack of reproducibility (eg high RSD across quality control samples), calibration methods such as normalization, grouped batch profile calibrations, and regression techniques can be applied (see paper 1). However, it should be noted that there are no general methods for calibrating spectral data, as the problems encountered may be dependent on the specific study. To address this, paper 1 delivers an additional recommend workflow of SQC, which may aid in identifying if any lack of reproducibility is present in data, whilst also suggesting some strategies to solve common challenges highlighted within paper 1.

Chapter 4

NMR-Onion

4.1 NMR-Onion Summary

The aim of the research conducted within this thesis is to be able to detect secondary metabolite signals in-situ within an NMR spectrum. The challenge of conducting metabolomics analysis using in-situ NMR data comes in the form of spectral complexity, e.g. high signal abundance, large signal SNR differences, and extensive overlap occurring due to the presence of hundreds of compounds produced by a microbial community. The aforementioned circumstances, makes traditional spectral analysis nearly impossible, as visual inspection and analysis of very complex data would be time-consuming when having multiple spectral datasets. In addition, when the signals of targeted metabolites are of low concentration and possibly located in highly overlapping spectral regions, a high operator/analyst bias may occur both w.r.t. analysis and spectral processing.

To resolve the challenges outlined above, the deconvolution program of NMR-Onion was constructed (paper 2, see appendix F). The program is capable of automatic detection and model 1D NMR spectral data using a hybrid multi-model-based approach, combining the frequency and time domain. Furthermore, the program is also capable of evaluating the repeatability of its own results, minimizing the risk of drawing false conclusions based on experimental artifacts. For the purpose of this chapter, some key elements of NMR-Onion are highlighted, while additional details are expanded upon, such as the mathematics behind some parts of the algorithm, computational efficiency, and the development history of NMR-Onion.

In short, NMR-Onion can be summarized as an algorithm solving the inverse problem of estimating the shape, amplitude, frequency, and number of underlying signals found within a 1D NMR spectrum. The base model for mapping the sum of signals within the time domain is expressed in equation (2.16), the relation also shown below for the sake of the reader:

$$y(t) = \sum_{k=1}^{K} A_k \exp(2j\pi\omega_k t + j\phi_k) \cdot \exp(-\alpha_k t). \tag{4.1}$$

As previously stated in chapter 2, the model of equation (4.1) does not take into account non-ideal signals caused by numerous artifacts such as eddy-currents, shimming imperfects, temperature fluctuation, sample preparation differences, etc. Therefore, the model of equation (4.1) was reformulated into two novel time domain formulations, being defined as a weighted sum of sinusoids governed by a Gaussian/exponential mixture decay (see equation (4.4)) and a stretched/compressed exponential decay (see equation(4.5)) respectively. In addition, a skewing term was added to all models, such that asymmetric signals would be adequately captured. The resulting models are shown below

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_n(\boldsymbol{\rho_k}) \cdot \exp(j\gamma_k^*)t, \tag{4.2}$$

where the function of f represents the harmonic term in equation (4.1), Ψ_n is the decay function of the n'th model with a subset of $\boldsymbol{\rho_k}$ parameters (varying for each decay function) and γ_k^* is the skewing constant, not to be confused with the γ symbolizing gyromagnetic ratio (In the original paper,$\gamma_k^* = \gamma_k$). The specific formulation of Ψ_n is presented below

$$\Psi_1(\alpha_k) = \exp(-\alpha_k t) \tag{4.3}$$
$$\Psi_2(\alpha_k, \eta_k) = (1 - \eta_k) \exp(-\alpha_k t) + \eta_k \exp(-\alpha_k t^2) \tag{4.4}$$
$$\Psi_3(\alpha_k) = \exp(-\alpha_k t^{\beta_k}) \tag{4.5}$$

To view the effect of the skewing constant, simulated data is presented in figure 4.1 and 4.2, visualizing the effects of γ^* on pure Lorentzian line-shapes. The models are further visualized in figure 4.3 and 4.4 showcasing the effects of the two novel decay types (equation 4.4 and (4.5)). Finally, the combined effect of decay types and values of γ^* is visualized in figure 4.5 and 4.6.

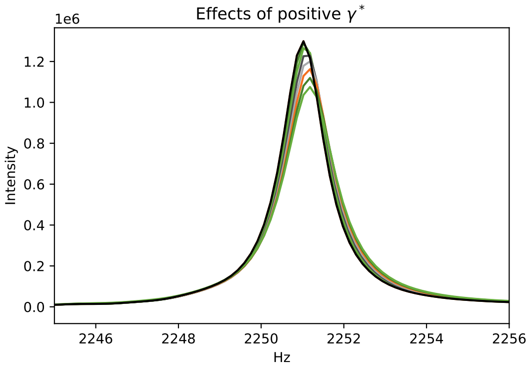


Figure 4.1: Visualization of the effect caused by increasing the values of γ^* from 0 to $\pi/2$. The black line is the pure Lortentizan line shape with $\alpha = 5$ and $\gamma^* = 0$.

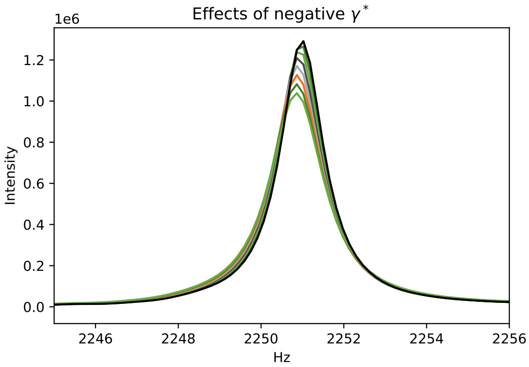


Figure 4.2: Visualization of the effect caused by decreasing the values of γ^* from 0 to $-\pi/2$. The black line is the pure Lortentizan line shape with $\alpha = 5$ and $\gamma^* = 0$.

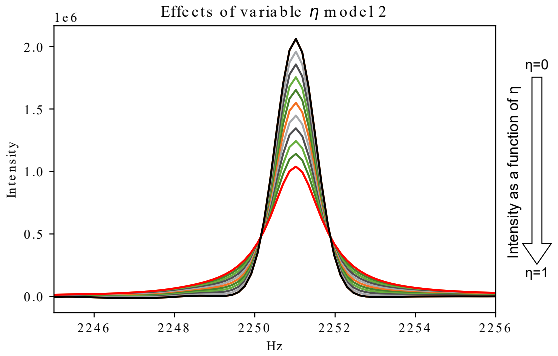


Figure 4.3: Visualization of the effect caused by increasing the values of η from 0 to 1 with $\alpha = 5$, the black line is the pure Gaussian line shape and the red is the pure Lortenzian line-shape

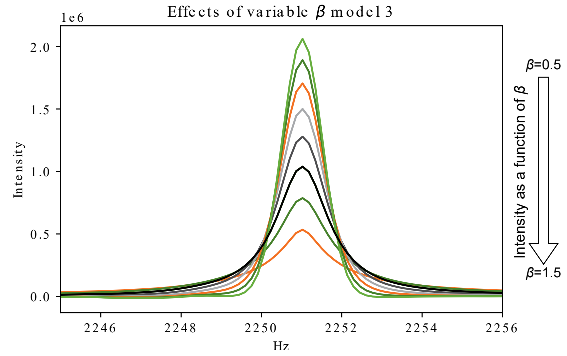


Figure 4.4: Visualization of the effect caused by increasing the values of β from 0.5 to 1.5 with $\alpha = 5$, the black line is the pure Lortentzian line shape.

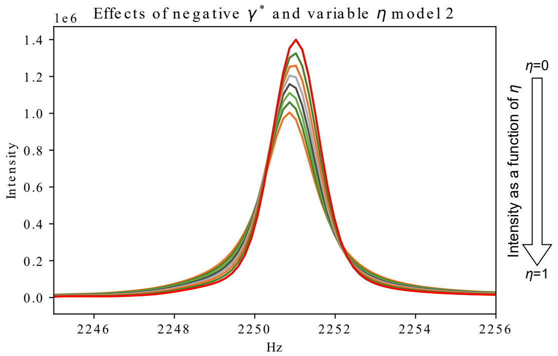


Figure 4.5: Visualization of the combined effect caused by decreasing γ^* from 0 to $-\pi/2$ and increasing values of η from 0 to 1 with $\alpha = 5$. The red line is the pure Gaussian line shape with $\gamma^* = 0$

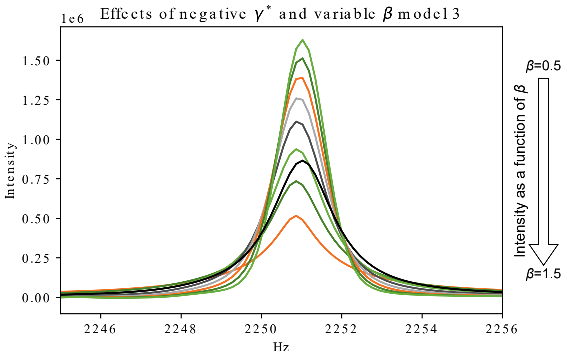


Figure 4.6: Visualization of the combined effect caused by decreasing γ^* from 0 to $-\pi/2$ and increasing values of β from 0.5 to 1.5 with $\alpha = 5$. The black line is the pure Lorentzian line shape with $\gamma^* = 0$

To solve the inverse problem of identifying parameter values and the number of signals, equation (4.1) was formalized into an optimization problem in which a penalized loss-function was minimized, formulated as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} SSE + \frac{1}{K}\sum_{k=1}^{K}(\phi_k - \bar{\phi}), \tag{4.6}$$

in which $\bar{\phi}$ is the mean instantaneous phase, estimated based on equation (2.12) and SSE is the sum of squared error found from the Hermitian transposed (H) residuals, e.g:

$$SSE = (\boldsymbol{Y} - \boldsymbol{ZA})^H(\boldsymbol{Y} - \boldsymbol{ZA}). \tag{4.7}$$

Here $\boldsymbol{Y}$ and $\boldsymbol{A}$ are vectors of $1 \times N$ and $1 \times K$, where K and N are the number of sinusoids and the number of data points respectively. The vectors of $\boldsymbol{Y}$ and $\boldsymbol{A}$ may be explicitly stated as:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}^T \tag{4.8}$$

$$\boldsymbol{A} = \begin{bmatrix} A_{c_1} & A_{c_2} & \dots & A_{c_K} \end{bmatrix}^T, \tag{4.9}$$

where A_c are the individual complex amplitudes corresponding to equation (2.10) found in chapter 2. Finally, Z is an $(N \times K)$ model matrix containing all time dependent terms of equation (4.2), given as:

$$\boldsymbol{Z} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,K} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,K} \\ \vdots & \vdots & \dots & \vdots \\ z_{N,1} & z_{N,2} & \cdots & z_{N,K} \end{bmatrix}. \tag{4.10}$$

The advantage of the above matrix formulation lies in the fact that the complex amplitudes may be represented by a least squares solution found from the model matrix, effectively removing two parameters from the optimization (see paper 2).

The application of the penalty term enables the elimination of possible spurious signals occurring due to the models being a superposition of sinusoids. In other words, the term prevents large variations in phases, which otherwise may cause large reversely phased signals of closely spaced frequencies to cancel out one another, generating a perfect sum, but meaningless individual signals.

Apart from the spurious signals, the loss function of equation (4.6) comes with three additional challenges:

- N and K is often at the size of >32k and >100
- The ω_k parameter is highly multimodal
- The loss function of equation (4.6) does not offer a way to identify the number of signals (K).

The first challenge was overcome by reducing parameter space size (K), applying the super-Gaussian filter described in paper 2 and section 2.3 of chapter 2. The reason why the lowering of K works can be found within the space-time complexity of the specific optimization algorithm. In the case of the LBFGS algorithm applied in NMR-Onion, the worst-case space-time complexity scales as $O(kN)$ per iteration[73], implying that lowering K or the input N would cause a less steep form of scaling. The LBFGS algorithm was implemented since the loss function is continuous, implying that an analytical gradient can be identified, speeding up the optimization, as numerical gradient computations are not needed at each iteration. The PyTorch back-end was applied for the purpose of optimization, as the automatic differentiation (AD) module ensured that optimal definition and usage of the gradient was applied. Pytorch was selected over a Scipy:autograd combination as it seemed in this particular case, the PyTorch implementation performed better. Applying less demanding algorithms such as Adam[74], which does not compute the inverse hessian, but relies on pure gradient descent, was also attempted. However, the initial results on real data revealed that first-order optimizations did not produce usable results.

The second and third challenge was solved jointly by implementing a peak detecting algorithm, utilizing resolution enhancement combined with the principles of the SG-filter and its derivatives highlighted in section 2.1.3 of chapter 2. The technique of resolution enhancement was chosen since the operation is essentially "free" (disregarding loss of potential resolution due to line broadening), as the resolution enhancements are carried out prior to DFT can easily be matched post-DFT. The method for choosing the correct weighting function and parameters for resolution enchantment is not a standardized procedure as mentioned in section 2.1.3. For NMR-Onion, the Gauss-to-Lorentzian (W_r) weighing function was chosen which may be expressed as:

$$W_r = \exp(Rt)\exp(-Rt^2/(2t_{max})), \tag{4.11}$$

where t_{max} is set at the time point where 99% of signal have been collected and R is the weighting parameter. The relation between the cumulative decay (from 0 to t_{max}) and R may be viewed as:

$$\sum_{n=0}^{N=\lfloor N_{0.99} \rfloor} y(t_n) = 1 - b_{off}\exp(-Rt_n), \tag{4.12}$$

where b_{off} is an offset constant and the $\lfloor N_{0.99} \rfloor$ operator indicates that the floor quantile is found when computing the 99 % quantile. To estimate R for equation (4.12), nonlinear optimization is utilized, in which a prior linearized fit is conducted such that reasonable initial values may be attained. The initial values are automatically evaluated by performing a three-point grid search. For the search, the linearized initial value is set at the initial estimated levels, 1.5 times the estimated level and 0.5 times the estimated level. Here the R-value which produces the highest number of detected peaks is chosen for resolution enchantment (the peaks are later subjected to deletions - see later in this section). For the SG-filter (principles covered in section 2.1.3), the SG-derivatives of first and second order are found for both the imaginary and real parts of the data stated as $d_I^{'}$, $d_I^{''}$ and $d_R^{'}$, $d_R^{''}$ for the imaginary and real SG-derivatives of first and second order respectively. The SG-derivatives are then transformed such that the absolute values are taken for $d_R^{'}$ and $d_I^{''}$, whilst zeroing all positives values for $d_I^{'}$ and $d_R^{''}$ respectively. Following the identification of the SG-derivatives and their respective transformations, PCA is applied to the transformed SG-derivatives, retaining only the first PC. The standard deviation of the inter-quantile range (IQR) for the first principal component (PC1) values is then utilized to filter out peak regions from noise regions. It is assumed that the IQR of PC1 values is mostly composed of noise[40][75] and therefore can be utilized to estimate the standard deviation of the background noise ($sd_{background}$).

The default cutoff, separating values containing signal and values containing noise, is set as $median(PC1) + 2 \cdot sd_{background}$, where the value of 2 is user-specified, the higher the value, the fewer peaks detected. In addition to ensure consistent results of the PCA, the rotational ambiguity of PCA is handled by setting the maximum deviation from zero to be in the positive direction. Finally, equation (2.40) ensures that peaks below a certain FWHM value are discarded (default 1.0 Hz), the higher the values, the more peaks are discarded. The only apparent downside to the algorithm comes in the form of non-automation, as the user needs to check the output of the algorithm when setting noise and width threshold. For the data analyzed in this thesis, the default width of 1.0 Hz was adequate, but the noise threshold had to be tuned ranging from 2 to 5 times the standard error. In future releases, this function will be automated, possibly combining the noise estimated by the digital filter (see section 2.3) and the peak detection.

4.2 Optimization

The purpose of this section is to get an overview of the optimization techniques utilized within NMR-Onion. The section is a further expansion of the supplementary material that comes with paper 2, providing a more in-depth explanation. A major challenge to overcome during the creation of NMR-Onion was to implement a stable optimization routine for solving the problem of equation (4.6). Originally, many of the algorithms found in the standard optimization library of Scipy[76][77] were considered. Inspired by Bretthorst[28] we originally attempted to have the optimization based on a derivative-free simplex approach[78] which was utilized with a different loss function than the one in equation (4.6) (see next section). However the derivative-free approaches did not show promising results, hence a choice was made to apply derivative-based optimization algorithms. Many options within derivative-based optimization algorithms were considered, investigating both Quasi-newton[79] and Newton[80] approaches. The main difference between the two approaches lies in how the Hessian (second derivative) matrix is computed. The Quasi-Newton methods approximate the Hessian matrix without explicitly computing it at each iteration via the gradient, whilst pure Newton methods explicitly compute the Hessian at each step. The advantages of the Quasi-Newton methods over pure Newton methods are in short that they can converge faster than pure gradient descent algorithms while avoiding the computational cost of computing the Hessian matrix. There are many types of Quasi-Newton algorithms, but in the end Limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS)[81] algorithm was chosen, as it is a very common and well-tested algorithm applied for many large optimization problems[82]. For an in-depth explanation of each of the aforementioned methods, the reader is referred to the book ”Numerical Optimization” by Nocedal and Wrigth[81], covering optimization theory in detail.

To further optimize Quasi Newton-based methods, the analytical gradient of the loss function (equation (4.6)) with respect to θ (the parameters of table 4.1) can be explicitly expressed and passed to the algorithm. When compared to that of the numerical gradient approximation, the application of analytical gradients has been shown to speed up the optimization process by many folds[83][84] and increase the robustness of results[81][85]. In spite of the opportunities presented by passing an analytical gradient to an optimization algorithm, the implementation may not always be straightforward, due to the following reasons:

1. Loss function complexity
2. Optimal input formation for the optimizer

To properly overcome the challenges of complexity and input formulation, the framework of automatic differentiation[85] was invented and implemented in many applications such as numpy [86], Tensorflow[87] and Pytorch[88] to mention some of the most popular.

To fully utilized the properties of automatic differentiation and LBFGS the loss function of (4.6) was implemented in PyTorch. However, further modifications needed to be added and implemented in order for the optimization to work properly. It turned out that the learning rate (step size) of the LBFGS prevented the optimization from reaching a suitable minima as the optimizer would often get trapped in a local minima. Initially, lowering the learning rate of the LBFGS was attempted and showed some promise, but resulted in convergence time being vastly increased. Unfortunately, LBFGS unlike optimizers such as Adam[74] does not have adaptive moments and therefore does not alter the learning rate as epochs increase (it does have adaptive stepsize within each epoch, but it resets after a completed epoch). To accommodate for the lack of learning rate adaptation, a learning rate scheduler was applied to the optimizer such that the learning rate[89] could be altered as the number of epochs (interaction cycles) increased. Different schedulers were tested but the exponential learning rate scheduler (see equation (4.13)) seemed to be the most consistently working scheduler. The function scheduler can be expressed as

$$lr_{epoch} = \gamma_{lr} \cdot lr_{epoch-1}, \tag{4.13}$$

where γ_{lr} is the learning rate reduction coefficient per epoch, the lower the γ_{lr}, the more reduction for each epoch. This is further visualized in figure 4.7.

Another aspect of the optimization comes in the form of parameter constraints. To get an overview of every model parameter of the NMR-Onion modelling framework, these are summarized in table 4.1 and paper 2.

Note from table 4.1 that A (amplitude) and ϕ (phase angle) are not part of the table. These are left out due to the implication of equation (4.7), stating that ϕ and A are nuisance parameters[90] not directly involved in the loss function of equation (4.6). However, A and ϕ can be computed from the parameters of table 4.1 by applying equation (2.11) and (2.12) to the resulting $\boldsymbol{A}$ vector (equation (4.7)). Despite being computable, ϕ and A do not serve any immediate purpose in NMR analysis. The time domain instantaneous amplitude (A) ratios are equivalent to the maximum peak intensity heights at a given frequency (which is given by the DFT of the individual deconvoluted signals), whilst the instantaneous phase values (ϕ) are not standard parameters used for any interpretation. Hence A and ϕ can be extracted but are not critical for any further analysis.

In its early form, the NMR-Onion Framework did follow the constraints set by

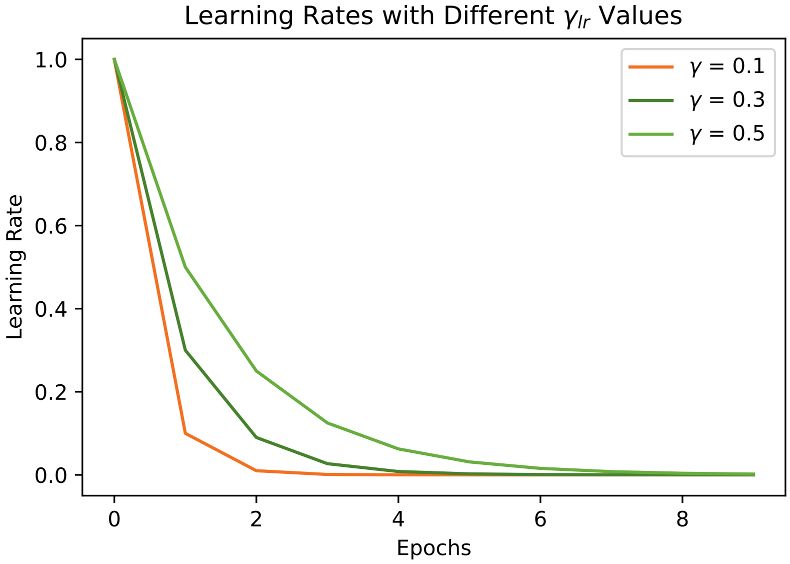


Figure 4.7: Visualization of the exponential learning rate scheduler with different values of γ_{lr} (γ in the figure legend)

table 4.1 very strictly by setting hard box-constraints[91]. Nonetheless, it was revealed that the box constraints approach often caused numerical instabilities with parameters being stuck at the boundaries. Therefore another approach had to be implemented which is described in the supplementary of paper 2, and for completeness partly covered here for some additional details.

The proposed method of paper 2, which solved the numerical instability issues, was very ”simple” - switch the constrained optimization to an unconstrained space. This sounds simple and could be achieved by removing the box constraints. Unfortunately removing constraints would cause the optimization to produce results that are not interpretable. Instead, the solution was to move the parameters to a different space in which the constraints are upheld by the properties of the transformation function rather than hard boundary constraints. The constraints which require the simplest transformation are the β parameter space, as positive values are the only required condition (though α shares the properties, a different transformation function was utilized - see paper 2 supplementary for details). A first choice would be an exponential transformation formulated as

$$\theta^* = \exp(\theta). \tag{4.14}$$

Here θ^* is the transformed parameter (θ is the input parameter) which is always positive and may easily be transformed back by applying the inverse transformation of $\log(\theta^* = \theta)$. Nevertheless, whilst equation (4.2) is mathematically correct, it causes numerical overflows in an optimization routine. Therefore, a different transformation was applied in the form of the Softplus function, expressed as

Table 4.1: Overview of model parameters for optimization

Model 1	Parameter	Description	Constraint
	ω	frequency	$[-\frac{SW_{Hz}}{2} < \omega < \frac{SW_{Hz}}{2}]$
	α	decay	$0 < \alpha$
	γ^*	skewness	$\frac{-\pi}{2} < \gamma^* < \frac{-\pi}{2}$
Model 2	Parameter	Description	Constraint
	ω	frequency	$[-\frac{SW_{Hz}}{2} < \omega < \frac{SW_{Hz}}{2}]$
	α	decay	$0 < \alpha$
	γ^*	skewness	$\frac{-\pi}{2} < \gamma^* < \frac{-\pi}{2}$
	η	weighting parameter	$0 < \eta < 1$
Model 3	Parameter	Description	Constraint
	ω	frequency	$[-\frac{SW_{Hz}}{2} < \omega < \frac{SW_{Hz}}{2}]$
	α	decay	$0 < \alpha$
	γ^*	skewness	$\frac{-\pi}{2} < \gamma^* < \frac{-\pi}{2}$
	β	stretch/compression	$0 < \beta$

$$\theta^* = \frac{1}{\tau} \log(1 + \exp(\tau\theta)). \tag{4.15}$$

The Softplus function of equation (4.2) mathematical achieve the goal as equation (4.2), transforming parameters to be positive. However, numerically the Softplus function is much more stable as numerical overflow cannot occur. For the purpose of NMR-Onion, τ was set at 1 (the default value of the Pytorch implementation).

For the reaming details surrounding the logistic Sigmoid transformation and combined Softplus scalar weighting transformation (applied for the η and α respectively), the reader is referred to the supplementary of paper 2.

4.3 NMR-Onion development history

The initial NMR-Onion algorithm was, as stated in the previous section, build on a very different framework compared to that the of final version. This section covers how NMR-Onion began and how elements were added or removed as development progressed. In general three major components were changed during development, being model formulation, optimization strategy (see previous subsection), and model selection. The section is meant as a guide for any future works within the NMR-Onion framework or as an aid for developing future NMR deconvolution methods, especially showcasing which approaches were less successful.

The initial model formulation started as a Bayesian model based on the works of Bretthorst[28], Andrieu[92] and Rubtsov[90]. The model was specified as the following

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{A} + \boldsymbol{\varepsilon}, \tag{4.16}$$

Where $\boldsymbol{Z}$ and $\boldsymbol{A}$ are identical to the formulation of equation (4.10) and (4.9) with $\Psi_n(\boldsymbol{\rho_k}) = \Psi_1(\alpha_k)$ not including the skewing term. In addition, the residuals ($\boldsymbol{\varepsilon}$) were assumed to be independent and identically distributed following a Gaussian distribution. The additional expansions of $\Psi_n(\boldsymbol{\rho_k})$ were not added until much later in the development process. To estimate the parameter of equation (4.16), the initial method was not that of equation (4.6), but rather a Bayesian inference approach applying the Metropolis-Hastings Algorithm[93]. The posterior from which the inference was generated, was for equation (4.16) formulated as

$$p(\boldsymbol{\theta}|y) \propto p(\boldsymbol{\theta})p(y|\boldsymbol{\theta}) = p(y|\omega, \sigma^2, \alpha, \boldsymbol{A})p(\boldsymbol{A}|\omega, \sigma^2, \alpha)p(\omega)p(\sigma^2)p(\alpha), \tag{4.17}$$

where $p(y|\boldsymbol{Z}, \boldsymbol{A})$ is the likelihood function and $p(\omega)$,$p(\sigma^2)$ and $p(\alpha)$ are the priors of each parameter expressed as $\boldsymbol{\theta} = (\omega, \sigma^2, \alpha)$ (frequency, variance and decay rate). Finally $p(\boldsymbol{A}|\omega, \sigma^2, \alpha)$ is the conditional distribution for the amplitudes. Inserting the distributions for each parameter, the posterior may be expressed as

$$\begin{aligned} p(\boldsymbol{\theta}|y) \propto p(\boldsymbol{\theta})p(y|\boldsymbol{\theta}) &\propto p(y|\omega, \sigma^2, \alpha, \boldsymbol{A})p(\boldsymbol{A}|\omega, \sigma^2, \alpha)p(\omega)p(\sigma^2)p(\alpha) && (4.18)\\ &= N(y; \boldsymbol{ZA}, \sigma^2) \times N(\boldsymbol{A}; \boldsymbol{0}, \boldsymbol{\sigma^2\Sigma}) \times U(\omega; 0, 2\pi) && (4.19)\\ &\times IG(\sigma^2; a, b) \times p(\alpha), && (4.20)\end{aligned}$$

where $\boldsymbol{\Sigma} = \frac{1}{g}\boldsymbol{Z^T Z}$, with g expressing expected signal to noise, $N()$ indicating a normal distribution, $U()$ uniform distribution, $IG()$ inverse gamma distribution and finally $p(\alpha) \propto \frac{1}{\alpha}$ being Jeffery's prior. In principle, no further reduction was needed as the Metropolis-Hastings Algorithm could draw inference from the posterior at this point, but to increase the inference speed, the complexity of equation (4.20) was reduced further integrating out the A and σ dependent terms[90]. This results in the following posterior distribution for the frequencies (ω) and decay rates (α)

$$p(\omega, \alpha|y) = p(\omega)p(\alpha)[\gamma_0 + y^H(I_N - \boldsymbol{Z\Sigma Z^H}))y]^{(-N-v_0/2)}, \tag{4.21}$$

from which $\boldsymbol{A}$ and σ^2 may be computed as their marginal distributions only depend on $\boldsymbol{Z}$ and the hyper parameters of γ_0, v_0 and g[90].

Even with reduced complexity, it is evident that the distribution of equation (4.21) is non-linear[90], multi-modal[28] and does not offer a way to estimate the number of components. The initial solution was to employ a greedy search strategy, combining the inference of equation (4.21) with a cutoff criterion (BIC in this case). The algorithm can be outlined in the following steps

1. Estimate initial ω value as the maximum value in the frequency domain
2. Draw inference from equation (4.21) and compute amplitudes and variance
3. Estimate BIC of current step k
4. Estimate residuals from $Y - ZA$

5. Estimate the next initial ω value as the maximum value of the residual signal in the frequency domain

6. Draw inference from equation (4.21) with k components using all initial values of ω and compute amplitudes and variance

7. Compare BIC of Current step k with previous step k-1

8. if BIC of Current step k> previous step k-1 stop else repeat step 5-8

The algorithm showed promise but suffered from two major flaws. The first was efficiency, as each Metropolis-Hastings run was slow especially as the size of Z grew. The second flaw was in the form of model imperfections, which caused data from real experiments to exhibit large residual signals. These large residuals would be mistaken for real signals causing a large degree of over-fitting within each spectrum. To make up for the aforementioned challenge caused by the greedy search, several methods for estimating both the number of sinusoids and initial frequency values were attempted. At first techniques such as the Prony method[94] and the Matrix Pencil method[95] were considered for estimating initial values, but these ultimately failed if the signals were not of pure exponential form and signals were highly overlapping. However, no better alternatives had been found at the time, and therefore the matrix pencil method was applied for generating initial values and estimating the number of components. The challenge of improving the speed of the Metropolis-Hastings approach was met by moving from a Bayesian approach to a frequentist approach, not relying on drawing samples, but rather optimization algorithms.

The optimization framework was, as stated in the 4.2 section, a derivative-free approach in its early phase, but was quickly converted into a derivative-based optimization approach, as the derivative-free approaches often failed, perhaps due to the function of equation (4.6) being non-convex. However, moving into gradient-based optimization called for explicit definitions of objective function gradients which proved problematic within the R programming language, as different optimizers required different gradient input formats.

Hence, the NMR-Onion framework was moved from R to Python with the aim of applying automatic differentiation (see previous section) in tandem with the Scipy optimization library. This approach did show promise with respect to speed increase but was still challenged by the fact that the sum of signals forming the FID does not consist solely of ideal exponential decaying sinusoids. To account for the non-ideal decay, the novel time domain models of equation (4.4) and (2.15) were formulated and eventually the SG-filter derivative-based method of rNMRfind[40] was discovered, adapted and implemented in a Python version within the NMR-Onion framework. Finally, the model framework was moved from Scipy to PyTorch which enabled an additional increase in speed. However, the greatest speed increase came from the digital filter implementation (see paper 2 and section 2.3).

An interesting future approach would be to implement the original Bayesian version in tandem with the digital filter, SG-filter derivatives detection, and the decay types of equation (4.4) and (2.15). However, as mentioned in paper 2, the approach would have to rely on variational inference[96] rather than full Monte Carlo Markov Chain inference to be efficient for a large Z matrix. In addition, suitable priors for the additional parameters need to be generated. Jefferys prior, which is non-informative prior, might be a suitable choice due to minimal influence on inference[97].

Chapter 5

Modelling Metabolomics

This chapter aims to provide the final part of the workflow highlighted within figure 1.2 of chapter 1. The chapter covers the central theoretical aspects of the statistical methods utilized for analyzing spectral outputs when applying deconvolution algorithms. In addition, the workflow of Chapter 1 is applied in practice, analyzing a case study in which targeted analysis of specific amino acids is conducted. The study highlights the interplay between DoE, deconvolution, and statistical analysis in particular.

5.1 Statistical modelling

With the detection of signals covered in the previous chapter and paper 2, the next step of the workflow (see figure 1.2 in chapter 1) is to conduct statistical analysis of the deconvoluted results, such that the effect of treatment interventions may be investigated. Within metabolomics, some of the most popular statistical modelling methods are that of partial least squares discriminant analysis (PLS-DA)[98] and principal component analysis (PCA)[99]. The main idea behind PCA is to have principal components as a linear combination of the original variables' orthogonality constraints that accounts for the maximum amount of variation in the data. The principal components consist of a matrix product of scores and loadings. Here the loadings are the weighting coefficients of each original variable and scores refer to the values of each observation (or sample) on the principal components. Scores represent the transformed data in the principal component space (seen in score plots), which can be used for further analysis or visualization. In particular, the application of score plots is observed extensively throughout the metabolomic literature. One example is found in a study of human plasma samples, where the effects of nutritional conditions are investigated via NMR[100]. Here PCA was applied such that the grouping of the responses could be investigated with respect to nutrient conditions. Another example is found in a study of bacterial resistance pathways using NMR[101]. In this study PCA was applied to view the groupings of bacterial metabolome responses, disguising wild types from mutated strains. However, despite PCA being heavily utilized to decompose NMR spectra, it is not guaranteed to produce meaningful results, as the first few components may be influenced by large peaks that do not hold any biological meaning[68]. Another factor is selecting the number of components, as classic PCA does not offer a clear method for the

selection process[68]. As an alternative, probabilistic PCA[102] can be utilized for selecting the number of components in a non-heuristic manner.
The second method of PLS-DA is a supervised method based on the concept of PLS regression. The method is used to model the relationship between two variables being X_s (the spectral data matrix) and Y (response variable), by constructing a set of latent variables, that explains the maximum covariance between the two variables. PLS-DA identifies the variables, or spectral regions, that are most important for discriminating between the different sample groups. It achieves this by selecting the principal components that maximize the separation between the groups while minimizing the variation within each group. The resulting PLS-DA model can be used to predict the class membership of new, unseen samples based on their NMR spectra. The performance of a PLS-DA model can be assessed using various metrics, such as classification accuracy, sensitivity, and specificity. In addition, the number of components can be estimated by cross-validation, preventing overfitting and ensuring valid models.
Numerous accounts of PLS-DA being applied in the metabolomics literature can be found. One example is found in a metabolomic study of Chrons diseases[103]. Here, metabolic profiling was done via proton NMR, resulting in the classification of healthy vs diseased based on PLS-DA. Specifically, the PLS-DA model was able to discriminate between diseased and healthy via the lipid profile of each test subject. Another study involving PLS-DA is found in the profiling of *Peganum harmala L* via NMR[104]. Within this study, PLS-DA was utilized such that the metabolome of *Peganum harmala L* could be classified during different stages of growth. The PLS-DA model revealed, that during the summer months the levels of vasicine, sucrose, choline, and valine were at their highest, whilst during winter proline, 4-hydroxyisoleucine and choline levels were much higher than any of the other seasons.

Despite PLS and PCA being popular methods, these may not necessarily be the obvious choice for analyzing a series of deconvoluted spectra. The reason is that the deconvolution table outputs consist of individual signals rather than bins or continuous signals. The single signals can prove to be a challenge, as different samples may have different numbers of signals and each sample may have slightly shifted signals. One possible solution for the latter would be to apply algorithms such as icoshift[54], but effects such as pH might be lost in the process. Another possible solution would be to store each signal in bins, but this would cause potential loss of information, as the amplitudes of many peaks may be merged into one, causing coupling patterns of smaller signals to vanish.

Another approach apart from PLS and PCA which does not require binning is that of traditional general linear modelling (GLM)[105][106]. The GLM may in general terms be set up as the following matrix notation:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta_m} + \boldsymbol{\varepsilon}. \tag{5.1}$$

Here, $\boldsymbol{Y}$ is the amplitude response, $\boldsymbol{X}$ is the design matrix related to that of the experimental design, and $\boldsymbol{\beta_m}$ is a vector of fixed effects. Note that the subscript of m is included such that the formulation may be disguised from previous uses of

β. A GLM makes it possible to investigate the significance of specific fixed effects (pH, temperature, concentrations of added compounds, etc) via multi-way ANOVA. For the random variable (residual term) of $\boldsymbol{\varepsilon}$ it is assumed to follow that of an iid (identically, independently distributed) Gaussian distribution (see the works of Bolker[106] or Madsen[105] for more details). This implies that the model of (5.1) is not valid without testing the model assumptions. The testing is carried out via model diagnostics which is utilized within the case study of section 5.3.

The general process of performing a multi-way ANOVA typically involves two main steps[107]. The first step aims to determine whether any main effects or interactions ($\boldsymbol{\beta_m}$) are statistically significant contributors to the overall variability in the data. To estimate the significance of effects, methods of either backward or forward selection[108] may be employed. The backward method starts by assuming a full model, gradually removing insignificant variables, whilst the forward method builds the model one variable at a time, retaining only significant terms. For a small number of variables, the backward selection is preferred[108], whilst, for a large number of predictors, the forward selection is usually applied[108].
However, algorithms such as leaps[109], applies a combination of both methods, which for larger complex datasets may provide more adequate results. For the evaluation of significance (e.g. model reduction), multiple tests may be conducted. Some of the more common evaluation metrics are the Akaike information criterion (AIC)[110], the Bayesian information criterion (BIC)[111], F-test[112] (testing sum of squares effects) or likelihood ratio test[113]. Apart from the F-test, all other methods are likelihood-based methods, which evaluate the goodness of fit vs the number of model parameters, thus these methods are used in model reduction. However, instead of evaluating likelihood, the F-test method evaluates the decomposition of the sum of squares, utilized in ANOVA to distinguish significant effects from non-significant effects. Here partitioning the sum of squares is an important concept, as different partitioning may lead to different results[114]. The most common types of partitioning are type I, II, or III ANOVA[105]. The main difference between type I and III is that type I depends on the order of effects, whilst type III is independent of order. Type II on the other hand assumes that the main effects of each predictor variable are independent of the other predictor variables, making it non suitable for testing interactions. For work of this thesis, type III is applied as it is considered the most robust of the portioning types[114][105].

The second step of hypothesis testing involves post hoc tests, which may include pairwise comparisons to confirm the significant contributions of different factor levels or to test for significant differences between factor levels. To lower the risk of type 1 errors for multiple comparisons of different factor levels, the post hoc tests are usually based on an adjusted p-value. The adjustment methods such as Bonferroni correction[115], Holms method[116] or false discovery rate[117] are some of the more common methods, with Bonferroni being the most strict adjustment method.

Within the literature of metabolomics multiple examples of GLMs have been applied, analyzing spectral responses. One example is found in a study profiling the

metabolomic response of anorexia nervosa in serum samples using 1D NMR[118]. Here serum sample bio-markers of 65 women with anorexia nervosa, 65 women recovering from anorexia nervosa, and 65 healthy women were compared using ANOVA. The mean signal peak areas of each group were compared, further including BMI as a covariate. The Bonferroni p-value adjusted post hoc comparisons revealed when comparing the three groups, methanol was significantly higher in the groups recovering from anorexia nervosa and with anorexia nervosa. Hence, the study concludes that higher levels of methanol can be seen as a trait biomarker of anorexia nervosa. An additional interesting detail was that when applying a different p-value adjusting method of false discovery rate, the metabolites of methanol, glutamine, threonine, glucose, and glycoproteins differed across the three groups.
A different example of general linear model-based hypothesis testing can be found in a study analyzing the metabolomics response of diabetes 2 rats treated with metformin[119]. In this study urine samples were analyzed via NMR. Hypothesis testing was conducted between four groups of rats being lean, obese, type 2 diabetic rats, and diabetic rats treated with metformin. The study applied ANOVA with Tukey's honest significant difference[120] for post hoc analysis. Here it was shown that the metabolomic profile of the four groups exhibited significantly different concentrations of trimethylaimne, trimethylaimne N-oxid, phenylacetateglycine indoxyl sulfate, and D-glucose. Specifically, the post hoc analysis indicated that trimethylamine, trimethylaimne N-oxid, phenylacetylglycine, and indoxyl sulfate levels were higher in metformin-treated rats when compared to diabetic rats. In addition, it was shown that D-glucose was lowered in diabetic rats treated with metformin. The study concluded based on the statistical evidence that metformin increased glucose metabolism, whilst the remaining metabolites showed an increase in the growth and activity of gut microbiota.

With the theory of statistical analysis and the ANOVA framework covered, a case study is presented in the next section. The study aims at applying the workflow presented in Chapter 1 figure 1.2, utilizing elements covered within this thesis in practice. Highlight how DoE, deconvolution, and statistical analysis may aid in linking specific treatment effects to specific spectral regions.

5.2 Programming environment

All scripts produced in this thesis are either made within the open source framework of R[121] or Python[122]. The environments were chosen not only due to non-propriety but also due to their many advantages. Python is excellent for development with low memory usage and its PyTorch library[123] served as the backbone for the NMR-Onion algorithm, enabling the usage of automatic differentiation and efficient optimization. R has the benefit of having vast statistical libraries capable of handling almost any analysis, hence R was chosen for all statistical data analysis of NMR-Onion output. All code is stored on the author's GitHub, found at the following link www.github.com/Mabso1.

5.3 Case Study

5.3.1 Introduction

In-situ spectra from biological samples are known to be very complex containing hundreds of known and unknown compounds at varying unknown concentrations[124][125]. This complexity imposes two main challenges within targeted spectral analysis, detection limitations and spectral resolution limitations[124]. The limit of detection is specifically linked to equation (2.30) in chapter 2, describing the ratio between a signal and noise floor of a spectrum. It should be noted that there are no specific definitions in NMR for high vs low SNR ranges[126], hence it is up to the operator to separate lower SNR signals from noise. In software such as craft[59], no signals beyond 5dB are considered to be valid (default value), whilst in NMR-Onion no default hard constraints are set, and it is up to the user to set the lower boundary (see chapter 4).
To increase the limit of detection, for in-situ data, the methods covered within Chapter 2 section 2.1.3 can be applied. However, as stated in section 2.3 care must be taken when applying line broadening to increase SNR, as smaller signals found within in-situ data may vanish.
The second challenge of spectral resolution is highly linked to spectral overlap, as spectral information may be hidden within the overlapping parts of a spectrum[40]. Overlapping signals are very common within in-situ data, as multiple metabolites may only be separated by a few Hz, resulting in highly convoluted regions. This makes the separation of signals by visual inspection nearly impossible and results highly biased.

To showcase how to accommodate the challenges presented in the previous paragraph, a case study is constructed. The study utilizes the in-situ metabolomics techniques presented throughout the entire thesis up until this point. The study will follow the workflow presented in figure 1.2 whilst highlighting how deconvolution methods developed through the NMR-Onion framework may aid in detecting targeted metabolites within 1D NMR spectra acquired from in-situ samples.

5.3.2 Study design

The purpose of this case study is to test how well the NMR-Onion framework works within targeted metabolomic analysis of 1D NMR data. This study is meant as a proof of concept and as such we wanted to make sure that samples contained the targeted compounds at a detectable SNR level. Hence, the data analysis within this study is constructed as simplified artificial in-situ samples. Complex real in-situ samples containing unknown compounds at unknown SNR levels are left out for now but will be analyzed in future studies.

The pseudo-in-situ samples generated for this study consisted of five amino acids at different mixing ratios (see table 5.1). Combinations of two-level concentration were investigated, being 0.1mM (low) and 1mM (high), thereby considering each amino acid a treatment factor of interest(see table 5.1).

With the treatments shown in table 5.1, the objective of this case study was to investigate if a change in spectral amplitude responses within specific regions of

Table 5.1: Overview of experimental setup

Amino acid	High conc (mM)	Low conc (mM)
Histidine	1	0.1
Arginine	1	0.1
Lysine	1	0.1
Cystine	1	0.1
Glycine	1	0.1

interest (ROIs) could be linked to the concentration of specific amino acids.

To accommodate the objective, of deconvolution of each spectrum was carried out by applying the NMR-Onion framework described in Chapter 4. However, blind deconvolution of every spectral signal without considering prior knowledge regarding peak positions would result in an overly complex analysis. Thus to incorporate prior knowledge, specific reporter groups of each amino acid were targeted with the aim of estimating amplitude responses conditioning on each factor. An overview of the amino acid reporter groups can be found in table 5.2, containing the expected chemical shift values and multiplicity for each reporter group.

Table 5.2: Overview of targeted amino acids reporter groups

Amino acid	Chemical shift(ppm)	Multiplicity	J (Hz)
Histidine	3.01	dd	15.2,4.8
Histidine	3.12	dd	15.4,$\sim$ 8
Histidine	3.86	dd	8, 4.8
Histidine	6.91	s	-
Histidine	7.26	s	-
Arginine	1.68	m	-
Arginine	3.90	t	6.9
Lysine	1.51	t	$\sim$ 7.5
Lysine	1.71	tt	$\sim$ 6.6
Lysine	2.93	t	$\sim$ 7.5
Cystine	3.98	dd	$\sim$ 8,4
Cystine	3.0	dd	15,$\sim$ 6.1
Glysine	3.47	s	-

It should be noted that the chemical shifts of table 5.2 are not exact but serves as an approximate location to construct targeted regions of interest which can be found in the results section.

To investigate the effects of treatment interventions (see table 5.1) a null hypothesis (H_0) was set up. The H_0 states that varying the concentrations of specific amino acids does not affect the spectral amplitude responses of the targeted reporter groups, which may be formulated as

$$H_0 : \boldsymbol{\theta_i} = \boldsymbol{\theta_j}, \tag{5.2}$$

which leads to the formulation of the reverse alternative hypothesis (H_1)

$$H_1 : \boldsymbol{\theta_i} \neq \boldsymbol{\theta_j} \tag{5.3}$$

Here $\boldsymbol{\theta}$ represents a vector of experimental factors ($k = 5$) and i and j are the two-factor levels (high and low concentration). In other words, if a significant experimental factor is detected, this would imply that the amplitude response ($\boldsymbol{Y}$) of a targeted ROI changes with the increase/decrease of a specific amino acid concentration. Therefore, one may be able to detect which amino acid is responsible for a change in a spectral amplitude response within a given ROI.

When conducting hypothesis testing of multiple factors, an analysis of variance (ANOVA - see section 5.2) is applied. Here the probability of observing a test statistic based on a mean sum of squares ratio (MSE_{factor}/MSE) is evaluated and the contribution of an effect is considered significant if the probability exceeds a preset significance level of α[106]. For this study, we set the significance level of α at 0.01.

Experimental design and design space modelling
To cover the entirety of the design space would require a total of $N = 2^5 = 32$ samples. The full design space would result in a design capable of resolving up to 5-factor interactions. However, much of the design resolution may be seen as redundant (see paper 1 of Chapter 3). Therefore, the design space may be reduced by applying the sparsity of effects principle[62], in which higher-order terms are neglected. The design space was reduced by choosing a fractional factorial resolution V design where main effects are not aliased with higher order interactions and two-factor interactions are not aliased with other two-factor interactions, but only with higher order terms. The resulting design matrix can be found in table 5.3.

Within table 5.3, the letters A, B, C, D, and E each represent an amino acid with A=Cystine, B=Lysine, C=Arginine, D=Glycine, and E=Histdine. Finally, the +1 and -1 indicates high and low concentration of the amino acid. It is worth mentioning that the design was randomized such that no bias may be introduced from the order of experiments[127].

To link the spectral amplitude response vector ($\boldsymbol{Y}$) to the treatment interventions of the design space, the following general fixed effect model formulation of equation (5.4) was applied

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta_m} + \boldsymbol{\varepsilon}. \tag{5.4}$$

Here, X is the design matrix of size $N \times k + 1$, which is almost equivalent to the responses given in table 5.3, with the exception of the first column consisting solely

Table 5.3: Overview of the 2^{k-1} fractional factorial design, the design generator was set as $E = ABCD$

Run no.	A	B	C	D	E
1	-1	-1	-1	-1	1
2	1	-1	-1	-1	-1
3	1	1	1	1	1
4	1	1	-1	-1	1
5	1	-1	1	1	-1
6	1	-1	1	-1	1
7	-1	-1	1	1	1
8	1	-1	-1	1	1
9	-1	1	1	1	-1
10	-1	1	-1	1	1
11	-1	1	-1	-1	-1
12	-1	-1	1	-1	-1
13	1	1	-1	1	-1
14	1	1	1	-1	-1
15	-1	-1	-1	1	-1
16	-1	1	1	-1	1

of ones such that an intercept is constructed. The second column of X corresponds to column A in table 5.3, the third column to column B, and so on. The remaining parameters of $\boldsymbol{\beta_m}$ and $\boldsymbol{\varepsilon}$ correspond to the predictors and vector residuals. For the residuals, it is assumed that these are independent and identically distributed (iid) with $\mathbf{0}$ mean vector and covariance matrix of $\sigma^2\boldsymbol{I}$, which may be expressed as:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}). \tag{5.5}$$

To better align with the hypothesis formulated within equation (5.3) and (5.2), the model of equation (5.4) is modified into equation (5.6), representing a traditional ANOVA vector formulation in which each predictor is explicitly stated.

$$Y_{ijkh} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \zeta_h + \varepsilon_{ijkh} \tag{5.6}$$

The predictors of equation (5.6) are each linked to the factors of A, B, C, D, and E as shown in table 5.4 along with the levels of each factor

To evaluate which amino acids have a significant impact, the predictors of the model found in table 5.4 are evaluated within the results section.

5.3.3 Experimental setup

The mixture ratios of the amino acids were set according to table 5.3, producing a total of 16 datasets. D2O was added to each sample prior to spectral acquisition. All spectra were acquired on a Bruker AVANCE III HD 800 MHz spectrometer equipped with a 5 mm TCI cryoprobe. The data was acquired by applying a zg

Table 5.4: Overview of model parameters

Amino acid	parameter	factor level
Histidine	α	i=1..2
Arginine	β	j=1..2
Lysine	γ	k=1..2
Cystine	δ	l=1..2
Glycine	ζ	h=1..2

pulse sequence, in addition, the digital filter was set as a baseopt rectangular filter such that baseline and first-order phase distortions were minimized. All spectra were acquired at 25° C, with a relaxation delay of 2 s, 128 scans, and 32K complex data points.

5.3.4 Spectral preprocessing

Following the acquisition of raw spectral data, preprocessing was carried out following the scheme highlighted within figure 2.8 of section 2.3. All spectral data were zero-filled to 64k and first-order phase correction was automatically carried out utilizing the automatic first-order phase correction in topspin. After exporting data from topspin to NMR-Onion, spectral alignment via the Icoshift algorithm and ARpls with $\lambda = 10^5$ was employed to make sure peaks were aligned and a flat baseline was achieved. It should be noted, that ARpls is only applied to aid in peak detection and therefore the data passed to NMR-Onion is not baseline corrected as the polynomial correction would alter the time domain in an unpredictable fashion[128]. No additional line broadening was applied apart from the 0.3Hz exponential decay window function.

To account for the amino acids reporter groups (see table 5.2), the bandpass filter of NMR-Onion was utilized to define regions of interest (ROIs) matching reporter groups' positions. The resulting ROIs are located within table 5.5 found in the results section.

5.3.5 Results

The band-pass filtering process of NMR-Onion resulted in a total of five regions of interest containing all reporter groups found in table 5.2. The range of the regions was selected such that the targeted ROIs were ensured to be contained within each region, whilst also ensuring that no boundary was set in the middle of a peak. Furthermore, the noise region (defining the filter noise floor - see section 2.3) was set at a constant reference interval with no visible peaks or random spikes. The resulting ROIs and noise ROI are found within table 5.5.

Each of the ROIs generated for the 16 datasets were deconvoluted generating three NMR-Onion model candidates (see Chapter 4) and subsequently, the best model from among the candidates was automatically chosen from the BIC. The resulting deconvolution for each ROI is visualized for one full dataset in figure 5.1.

Apart from the visual deconvolution, the resulting amplitudes (raw and ratios) and

Table 5.5: Overview of experimental regions of interest (ROI)

ROI No.	lower bound (ppm)	upper bound (ppm)
1	6.4	8.5
2	3.8	4.0
3	3.4	3.7
4	2.8	3.3
5	1.5	1.9
Noise ROI	-0.2	-0.1

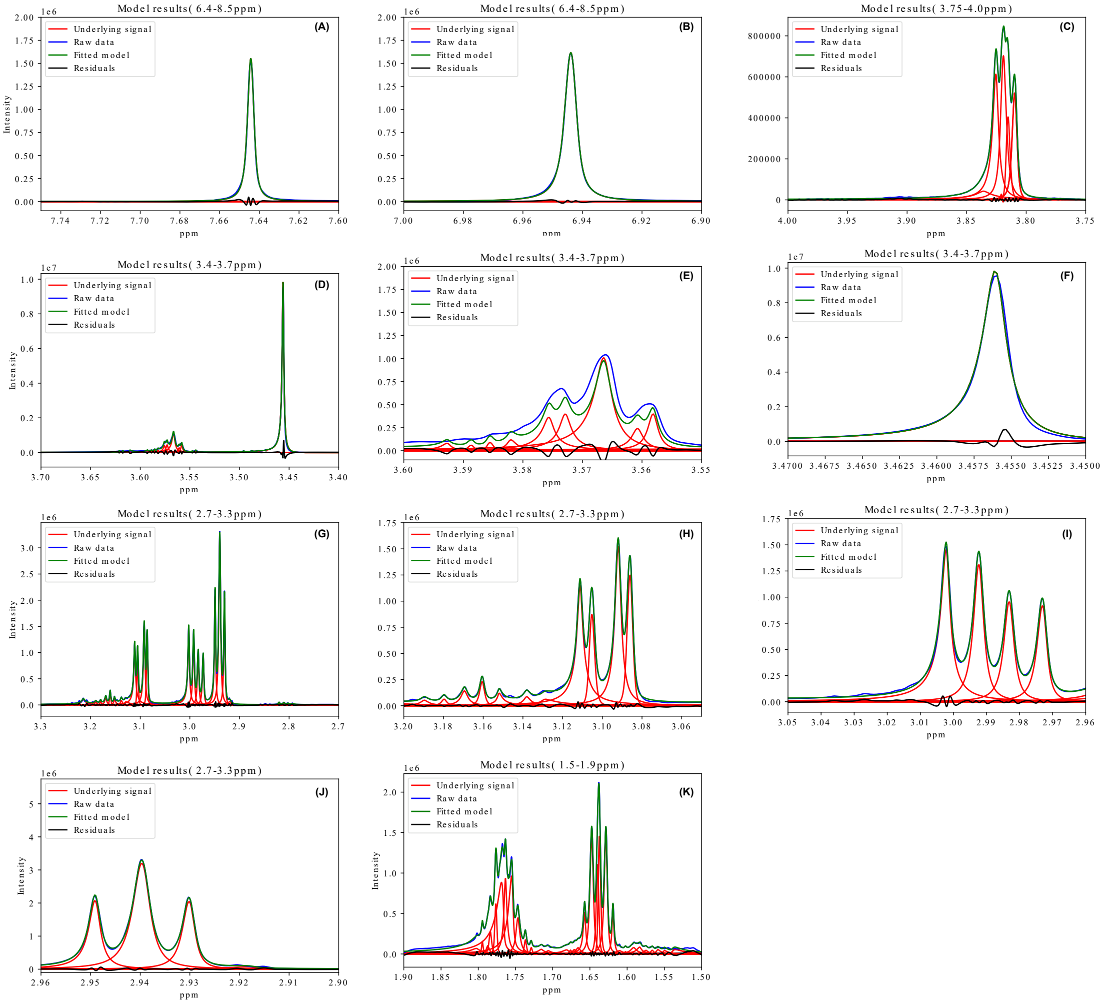


Figure 5.1: (A) deconvoultion of first sub-ROI in ROI1. (B) deconvolution results of the second sub-ROI of sub-ROI1 in ROI1. (C) Deconvoultion results of ROI 2. (D) Deconvolution results of ROI 3. (E) Zoom in on a sub-ROI in ROI3. (F) Zoom in on a second sub-ROI in ROI3. (G) Deconvolution results of ROI 4. (H) Zoom in on the first sub-ROI of ROI4. (I) zoom in on the second sub-ROI of ROI4. (J) zoom in on the third sub-ROI of ROI4. (K) deconvoultion results of ROI5. Larger individual plots can be found in appendix A

chemical shifts (ppm and Hz) are stored in a CVS file along with their uncertainty estimates. The latter were computed by applying the bootstrap model of NMR-Onion setting the number of bootstrap samples to 1000. With deconvolution, a list of peaks and amplitudes is generated, but they cannot solely identify which peaks belong to which compound. Therefore additional analysis had to be carried out, matching detected peaks with the expected response of the reporter groups found in 5.2. Within appendix A, an example of the NMR-Onion output is showcased for two sub-ROIs corresponding to figure 5.1A, figure 5.1B and figure 5.1C.

The findings reported in appendix A were employed in conjunction with the remaining deconvolution tables generated by NMR-Onion to construct appropriate sub-ROIs of deconvolution outputs. The sub-ROIs were generated such that the frequency ranges were aligned with the targeted peaks of table 5.2. The generated sub-ROIs are each outlined in table 5.6.

Table 5.6: Overview of experimental subregions of interest (sub-ROI)

sub-ROI No.	lower bound (ppm)	upper bound (ppm)
1	6.40	8.50
2	3.90	4.00
3	3.80	3.88
4	3.40	3.55
5	3.15	3.20
6	3.00	3.10
7	2.91	2.95
8	2.81	2.89
9	1.62	1.81
10	1.50	1.61

From the filtered sub-ROIs of deconvolution outputs (see table 5.6), statistical modelling of the amplitude responses ensued with the goal of identifying significant factors for each sub-ROI of table 5.6. The model of equation (5.6) was investigated by model diagnostics and reduced by applying a backward selection approach (see section 5.1). At each reduction step, the models were evaluated by applying subsequent F testing with the sum of squares partitioned in accordance with type III partitioning (see section 5.1).
From the model diagnostics analysis, it was revealed that the model assumptions of residual normality (equation (5.5)) were not met as shown in the quantile-quantile plot (QQ-plot) in figure 5.2A.

To mitigate the lack of normality, a transformation of data was applied using a log function for the amplitude responses. This resulted in the fixed effects becoming adequately normally distributed, which is evident from figure 5.2B. Note that the above QQ plots are the results of modelling the first sub ROI found in 5.6, but the pattern was mostly identical in each ROI (see appendix D). Note that the outliers of points 35 and 37 are removed, and outliers of all other spectra were also removed.

Having applied corrections such that the model assumptions are met, table 5.7

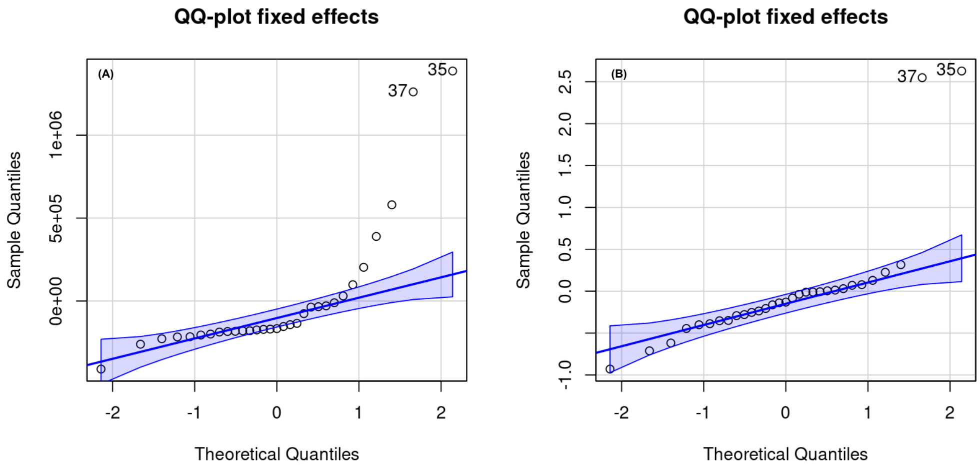


Figure 5.2: (A) QQ-plots for the fixed effects residuals before log-transformation. (B) QQ-plot of fixed effects residuals after log transformation.

presents the outcomes of each model. Here the significant factors of the fully reduced model for each region of interest (ROI) listed in table 5.6 are reported. The full table of 5.7 provides details regarding the sum of squares error (SSE), the mean sum of squares (MSS), degrees of freedom, F_0-value, and corresponding P-values for statistically significant factors. Note that the complete backward selection process is available in Appendix B, showcasing a graphical overview.

In addition to table 5.7 a visualization of the modelling results are found in figure 5.3. The first sub-ROI is showcased within this section, whilst the reaming sub-ROIs plots can be found in appendix C.

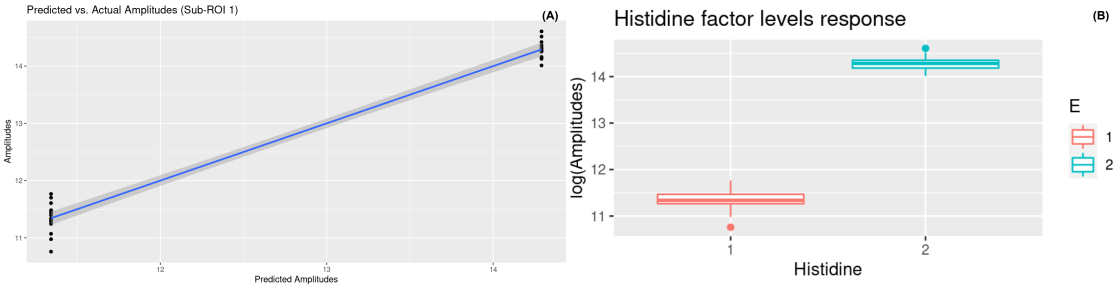


Figure 5.3: (A) visualization of model correlation. (B) Treatment effect responses, level 1 vs level 2.

To get a more exact estimate of the correlation plotted in 5.3A and the remaining sub-ROIs (see figure A for each sub-ROI in appendix C), numerical correlations were calculated and stored in table 5.8.

From the resulting factors of significance in table 5.7, the following information may be extracted. For the first, third, and sixth sub-ROI, altering the concentrations of amino acid E (histidine) is found to have a significant effect on the amplitude

Table 5.7: Overview of experimental regions of interest (ROI) modelling results. The table contains the results of the final model, containing only significant factors. Here A=Cystine, B=Lysine, C=Arginine,D=Glycine and E=Histidine.

sub-ROI 1				**(8.50:6.40 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
E	62.90	62.90	1	1305.40	$2.20 \cdot 10^{-16}$
sub-ROI 2				**(4.00:3.90 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
A	51.67	51.67	1	175.06	$2.72 \cdot 10^{-14}$
sub-ROI 3				**(3.88:3.80 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
E	7.52	7.52	1	7.76	$1.01 \cdot 10^{-2}$
sub-ROI 4				**(3.55:3.40 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
D	27.05	27.05	1	568.38	$8.06 \cdot 10^{-11}$
sub-ROI 5				**(3.20:3.15 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
C	16.91	16.91	1	36.28	$1.72 \cdot 10^{-6}$
sub-ROI 6				**(3.10:3.00 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
E	33.94	33.94	1	15.11	$1.69 \cdot 10^{-3}$
sub-ROI 7				**(2.95:2.91 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
B	15.40	15.40	1	15.71	$2.93 \cdot 10^{-4}$
sub-ROI 8				**(2.89:2.81 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
A	16.95	16.95	1	8.32	$7.78 \cdot 10^{-3}$
B	44.60	44.60	1	21.89	$7.83 \cdot 10^{-5}$
sub-ROI 9				**(1.81:1.62 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
B	43.30	43.30	1	36.72	$1.80 \cdot 10^{-8}$
C	17.20	17.30	1	14.62	$2.15 \cdot 10^{-4}$
sub-ROI 10				**(1.61:1.50 ppm)**	
Factor	SSE	MSE	DF	F_0	P-value
B	109.40	109.40	1	98.22	$2.20 \cdot 10^{-16}$

responses, whilst the remaining amino acids had no significant impact on the amplitude responses. As for the fifth sub-ROI, it was found that amino acid C (Argine) showed significant responses with respect to treatment interventions. For amino acid D (glycine) only the fourth sub-ROI emerged as having a significant response. Finally, for the remaining two amino acids of A (Cystine) and B (Lysine), sub-ROI two had a significant response for A, whilst sub-ROI seven and ten indicated significant responses with respect to B. Interestingly, sub-ROI 8 and 9 did not emit one, but two significant response, being linked to concentration alterations of A and B for sub-ROI 8 and B and C for sub-ROI 9. It should be noted that since factors

Table 5.8: Overview of Sub-ROI Pearson correlations

sub-ROI No.	Correlation
1	0.98
2	0.92
3	0.55
4	0.99
5	0.90
6	0.79
7	0.75
8	0.73
9	0.80
10	0.75

only have two levels each, no post hoc analysis was conducted, as no adjustment was needed.
The plots and table 5.8 also show, that in general a correlation above < 0.7 between observed and predicted was achieved for the most part, expected for sub-ROI 3, which exhibited much lower correlation than other sub-ROIs (see more in the discussion).

The results and methods of analysis are further interpreted and discussed in the next section.

5.3.6 Discussion

The first point of the discussion comes from comparing the expected responses of 5.2 with the factors of significance found in table 5.7. For the most part, the results make sense, as the expected responses seemed to match the pattern of the significant amplitude response of each sub-ROI with respect to treatment interventions. One of the more interesting results is found in sub-ROI nine and ten, where the responses are purely consisting of multiplets, making first-order assignment impossible. However, by combining deconvolution with a targeted approach and a DoE setup it was possible to gain value from these regions, identifying patterns within the responses which could be linked to the experimental factors. However, it was not possible to distinguish B from C in sub-ROI nine, which may be due to high overlaps and spectral shifts occurring from both Lysine and Arginine. Comparing the results with HMDB, it is found that a triple of triplets is reported at 1.71 ppm for Lysine whilst a multiplet is found for Arginine at 1.68 ppm which may cause the overlap. Nevertheless, the NMR-Onion algorithm in tandem with a GLM manages to identify the correct amino acids at each sub-ROI. This strategy may be transferred to more complex scenarios where specific targeted regions are consisting of purely mulitplets, such that the difference in amplitude responses may be linked to specific treatment interventions. The strategy in complex scenarios does not alone provide answers to ”what is in the mixture”, but it can aid in automatically identifying regions in which treatments have a significant impact. Another interesting result is the two significant responses found in the eighth sub-ROI. The result implies that the treatment interventions within the region are occurring due to significantly altering amplitudes of both lysine (B) and cystine (A). Compared with table 5.2, the result makes sense, as the cystine and lysine signals at 2.93 and 2.95 ppm are highly overlapping and thus cannot be separated solely from amplitude response. However, it is possible to manually investigate the output of a deconvoluted spectrum to identify exactly which signals belong to which amino acid. One method is to use the DoE setup to separate the signals by comparing samples from high concentrations of A vs low concentrations of B or vice versa to pinpoint the peak origin.

The statistical modelling of the deconvolution showed in short that the general linear model was able to adequately model the data responses (see model diagnostic D and model results B). The model successfully coupled the responses of each ROI to one or two amino acids. Despite the aforementioned results, the GLM is not perfect. This is especially observed for the amplitude responses found in sub-ROI 3, which shows the lowest correlation values and highest deviations from normality for the fixed effects (see QQ-plots (D) appendix D). One reason why the deviations from normality occur could be that the log transformation of amplitude responses is not generalizing well enough for all sub-ROIs. One option could be to employ a Box-Cox transformation[129] instead of the log transformation. This might be an improved transformation, as the log transformation is a special case of the Box-Cox transformation (when $\lambda = 0$)[129]. Apart from the model itself, the experiment might be set up differently such that replicates are included. This would enable the usage of statistical quality control, such that Quality control samples[71] can be generated and evaluated for each targeted ROI[130], identifying potential nuisance artifacts, possibly interfering with treatment variance[130]. In addition, sample intra-correlations may also be accounted for by introducing a linear mixed effect model[106]. Another modification could be the introduction of an internal standard[131] such as Sodium trimethylsilylpropanesulfonate (DSS) or Trimethylsilylpropanoic acid (TSP). The addition would cause the output of the deconvolution to become quantitative, relating amplitudes to concentrations.

As for the modelling of deconvolution outputs, there are other methods that may be suitable for detecting the impact of experimental factors. Some of the most popular methods are Partial least squares discriminant analysis (PLS-DA)[98] or ANOVA simultaneous component analysis (ASCA)[70]. The ASCA framework could be a particularly interesting framework to test out, as it is essentially an extension of the GLMM ANOVA framework at high dimensions (when the number of variables exceeds the number of observations). Also, the ASCA framework might be an obvious choice, as it requires an underlying DoE[70], which has been applied during this study. The reaming methods of PLS could work, but in this case, the study is not considered a classification problem (the deconvolution response is not aimed at investigating different metabolite classes, healthy vs sick, or other common classification problems) and therefore these methods may not be an obvious choice. In summary, it would for future studies, be interesting to combine that of ASCA with NMR-Onion, comparing the results of the hypothesis testing found with the GLMM approach to the ASCA multivariate approach.

5.3.7 Conclusion

From the results and discussion, it may be concluded that the effects of altering sample concentrations may be correlated with specific amino acids at given frequency ranges. This was evident as general linear effect modelling concluded that significant treatment interventions were found solely for one or few amino acids within a given sub-ROI matching the expected report group values of table 5.2. Furthermore, it is concluded that the fractional factorial design of the experiment lowered the experimental space significantly whilst still retaining the significant factors. Finally, it is concluded deconvolution using NMR-Onion successfully identifies targeted peaks found in various concentrations across different samples.

Chapter 6

Conclusion and Future perspectives

6.1 Conclusion

The first aim of automatic detection and uncertainty evaluation of targeted peaks was fulfilled by creating the novel detection/quantification framework of NMR-Onion (paper 2). The second aim of the thesis was to ensure robust data generation and analysis with a metabolomic pipeline. The aim was partly completed by establishing a robust workflow, highlighting the need for quality data generation, and combining design of experiments with statistical quality control (paper 1). Furthermore, the workflow generated within this thesis was applied to a case study investigating a mixture of amino acids, acting as pseudo-in-situ data. However, the apporch needs to be further tested for real complex NMR in situ data. A further elaboration of the conclusions for each part of the thesis is further emphasized in the next paragraphs.

In Paper 1 design of experiments (DoE) and statistical quality control (SQC) were reviewed in the context of a metabolomic workflow (eg. DoE, SQC, and hypothesis testing). The aim of the review was two folded. The first aim was to highlight some of the theoretical concepts of DoE, such that the choice of design would be adequately linked to the specific type of study (optimization or screening). The DoE part linked the theoretical aspects of the review to various studies of both NMR and MS-based metabolomics, in which DoE was applied such that maximum information could be attained from the smallest sample space possible, saving both time and effort. This led to a suggested baseline workflow for DoE within metabolomics, creating a road map for optimal data generation within screening and optimization studies.

The second aim of the review was to introduce the concepts of SQC, focusing on the quantification of experimental repeatability and possible calibration methods for common challenges. The SQC part of the review highlighted studies both within NMR and MS which employed SQC to quantify and calibrate metabolomic data. This led to the creation of an SQC workflow which may serve as a baseline for measuring and potentially correcting data within a metabolomic study.

The NMR-Onion algorithm of paper 2 sets itself apart from previous algorithms, as it considers multiple novel time domain models, and evaluates data repeatability via potential resolved peaks. From the two case studies of paper 2, the algorithm was shown to be capable of fitting multiple signals at different SNR values with high resonace/signal overlaps for various degrees of spectral complexity (eg. the number of signals). In addition, it was shown that the algorithm in some cases outperformed that of Mnova GSD, detecting more signals than the GSD algorithm. Summing up, paper 2 concluded that the NMR-Onion algorithm is a robust framework capable of detecting, deconvoluting, and evaluating the repeatability of specific targeted peak areas.

The case study of chapter 5 within this thesis demonstrated how NMR-Onion may be utilized within a larger metabolomic pipeline, applying the workflow highlighted in figure 1.2 for targeted analysis. The study was designed as a resolution V fractional factorial design, with each factor (five in total) attributed to the concentration of amino acid spanning two levels of high (1mM) and low (0.1mM). NMR-Onion was utilized to deconvolute specific targeted ROIs linked to the targeted reporter groups of each amino acid. Through the usage of general linear effect models, it was concluded that changes in spectral amplitudes within specifically selected ROIs across samples could be correlated with specific treatment interventions (changes in concentration levels) set up in the experimental design. In other words, the peaks and amplitudes identified by NMR-Onion could be linked to specific amino acids.

6.2 Future perspectives

For the research conducted within this thesis, various possibilities exist for the application of the developed workflow/pipeline in future research endeavors. The NMR-Onion program within the workflow may be utilized for any in-situ detection purpose such as drug discovery, disease diagnostics, or personalized medicine to mention a few areas. However, the main purpose for which the pipeline was built was to analyze the metabolomic profiles and responses of microbial communities, identifying the role of secondary metabolites. Hence for future endeavors, the program and workflow can be applied for analyzing microbial NMR data, uncovering the diversity (RQ2) and function of secondary metabolites (RQ3). For future expansions of the pipeline, the output generated by the NMR-Onion framework may be coupled with database matching. This would enable a more rapid analysis, as the automated detection would be coupled to that of automatic compound identification possibly via a deep learning approach. Another near-future update would be the addition of a graphical user interface for the program, as this would enable the usage for a wider audience as fewer programming skills are needed. The metabolomic data may even be combined with that of genomics data, such that the expression of genes may be linked to that of microbial metabolomic profiles or it could be coupled to that of MS data to gain border chemical insights. This could eventually lead to a broader understanding of microbial communities within different niches, as uncovering chemical complexity may aid in uncovering the complexity of microbial interactions and the role of secondary metabolites.

Bibliography

[1] T. C. Munedzimwe, R. L. Van Zyl, D. C. Heslop, A. L. Edkins, and D. R. Beukes, "Semi-synthesis and evaluation of sargahydroquinoic acid derivatives as potential antimalarial agents," *Medicines*, vol. 6, no. 2, p. 47, 2019.

[2] P. Vaishnav and A. L. Demain, "Unexpected applications of secondary metabolites," *Biotechnology Advances*, vol. 29, no. 2, pp. 223–229, 2011.

[3] N. Kallscheuer, T. Classen, T. Drepper, and J. Marienhagen, "Production of plant metabolites with applications in the food industry using engineered microorganisms," *Current Opinion in Biotechnology*, vol. 56, pp. 7–17, 2019.

[4] M. Swallah, H. Sun, R. Affoh, H. Fu, and H. Yu, "Antioxidant potential overviews of secondary metabolites (polyphenols) in fruits," *International Journal of Food Science*, vol. 2020, p. 9 081 686, 2020.

[5] S. Khatri, S. Dubey, Y. Shivay, L. Jelsbak, and S. Sharma, "Organic farming induces changes in bacterial community and disease suppressiveness against fungal phytopathogens," *Applied Soil Ecology*, vol. 181, p. 104 658, 2023.

[6] J. Mishra and N. K. Arora, "Secondary metabolites of fluorescent pseudomonads in biocontrol of phytopathogens for sustainable agriculture," *Applied Soil Ecology*, vol. 125, pp. 35–45, 2018.

[7] A. Fleming, "On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae," *British journal of experimental pathology*, vol. 10, no. 3, pp. 226–236, 1929.

[8] S. de Pharmacie (Paris), *Journal de pharmacie et de chimie: contenant les travaux de la Société de Pharmacie de Paris : une revue médicale* (Journal de pharmacie et de chimie: contenant les travaux de la Société de Pharmacie de Paris : une revue médicale vb. 33). Doin, 1858.

[9] E. Schunck and L. Marchlewski, "Zur kenntniss des isatins," *Berichte der deutschen chemischen Gesellschaft*, vol. 29, no. 1, pp. 194–203, 1896.

[10] J. Fujitani, "Beiträge zur chemie und pharmakologie des insektenpulvers," *Archiv für experimentelle Pathologie und Pharmakologie*, vol. 61, no. 1, pp. 47–75, 1909.

[11] A. F. Hernández, "Food safety: Pesticides," in *Encyclopedia of Human Nutrition (Fourth Edition)*, B. Caballero, Ed., Fourth Edition, Oxford: Academic Press, 2023, pp. 375–388.

[12] M. Á. García-Sevillano, T. García-Barrera, and J. L. Gómez-Ariza, "Environmental metabolomics: Biological markers for metal toxicity," *ELECTROPHORESIS*, vol. 36, no. 18, pp. 2348–2365, 2015.

[13] E. Schievano, M. Sbrizza, V. Zuccato, L. Piana, and M. Tessari, “NMR carbohydrate profile in tracing acacia honey authenticity,” *Food Chemistry*, vol. 309, p. 125 788, 2020.

[14] D. Ausgabe, E. Luchinat, L. Barbieri, M. Cremonini, A. Nocentini, C. T. Supuran, and L. Banci, “Drug Screening in Human Cells by NMR Spectroscopy Allows the Early Assessment of Drug Potency,” *Angewandte Chemie*, vol. 132, no. 16, pp. 6597–6601, 2020.

[15] C. M. Slupsky, H. Steed, T. H. Wells, K. Dabbs, A. Schepansky, V. Capstick, W. Faught, and M. B. Sawyer, “Urine metabolite analysis offers potential early diagnosis of ovarian and breast cancers,” *Clinical Cancer Research*, vol. 16, no. 23, pp. 5835–5841, 2010.

[16] F. Madrid-Gambin, C. Brunius, M. Garcia-Aloy, S. Estruel-Amades, R. Landberg, and C. Andres-Lacueva, “Untargeted 1H-NMR-based metabolomics analysis of urine and serum profiles after consumption of lentils, chickpeas, and beans: An extended meal study to discover dietary biomarkers of pulses,” *Journal of Agricultural and Food Chemistry*, vol. 66, no. 27, pp. 6997–7005, 2018.

[17] L. Nemadodzi, J. Vervoort, and G. Prinsloo, “NMR-Based Metabolomic Analysis and Microbial Composition of Soil Supporting Burkea africana Growth,” *Metabolites*, vol. 10, no. 10, pp. 1–17, 2020.

[18] R. A. Quinn, L.-F. Nothias, O. Vining, M. Meehan, E. Esquenazi, and P. C. Dorrestein, “Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy,” *Trends in Pharmacological Sciences*, vol. 38, no. 2, pp. 143–154, 2017.

[19] P.-E. Jansson, R. Stenutz, and G. Widmalm, “Sequence determination of oligosaccharides and regular polysaccharides using nmr spectroscopy and a novel web-based version of the computer program casper,” *Carbohydrate Research*, vol. 341, no. 8, pp. 1003–1010, 2006.

[20] D. Li and R. Brüschweiler, “PPM-One: A static protein structure based chemical shift predictor,” *Journal of biomolecular NMR*, vol. 62, pp. 403–409, 2015.

[21] J. Nicholson and J. Lindon, “Systems biology - metabonomics,” *Nature*, vol. 455, pp. 1054–1056, 2008.

[22] A. Kosmides, K. Kamisoglu, S. Calvano, S. Corbett, and I. Androulakis, “Metabolomic fingerprinting: Challenges and opportunities,” *Critical reviews in biomedical engineering*, vol. 41, no. 3, pp. 205–221, 2013.

[23] E. Kim, H. Kwon, H. Nam, J. Kim, S. Park, and Y. H. Kim, “Urine-NMR metabolomics for screening of advanced colorectal adenoma and early stage colorectal cancer,” *Scientific Reports*, vol. 9, no. 1, p. 4786, 2019.

[24] G. F. Giskeødegård, T. S. Madssen, L. R. Euceda, M.-B. Tessem, S. A. Moestue, and T. F. Bathen, “NMR-based metabolomics of biofluids in cancer,” *NMR in Biomedicine*, vol. 32, no. 10, e3927, 2019.

[25] J. Keeler, *Understanding NMR Spectroscopy.* John Wiley & Sons, Ltd, 2016, pp. 1–526.

[26] H. K. Cummins and A. Jones, “Nuclear magnetic resonance: A quantum technology for computation and spectroscopy,” *Contemporary Physics*, vol. 41, no. 6, pp. 383–399, 2000.

[27] J. S. b. Larmor, "Lxiii. on the theory of the magnetic influence on spectra; and on the radiation from moving ions," *Philosophical Magazine Series 1*, vol. 44, no. 271, pp. 503–512, 1897.

[28] G. L. Bretthorst, C. C. Hung, D. A. D'Avignon, and J. J. Ackerman, "Bayesian analysis of time-domain magnetic resonance signals," *Journal of Magnetic Resonance (1969)*, vol. 79, no. 2, pp. 369–376, 1988.

[29] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing.* California Technical Publishing, 1997.

[30] R. G. Spencer, "The Time-Domain Matched Filter and the Spectral-Domain Matched Filter in 1-Dimensional NMR Spectroscopy," *Concepts in magnetic resonance. Part A, Bridging education and research*, vol. 36A, no. 5, pp. 255–264, 2010.

[31] L. Chen, Z. Weng, L. Goh, and M. Garland, "An efficient algorithm for automatic phase correction of nmr spectra based on entropy minimization," *Journal of Magnetic Resonance*, vol. 158, no. 1, pp. 164–168, 2002.

[32] S. Sokolenko, T. Jézéquel, G. Hajjar, J. Farjon, S. Akoka, and P. Giraudeau, "Robust 1d nmr lineshape fitting using real and imaginary data in the frequency domain," *Journal of Magnetic Resonance*, vol. 298, pp. 91–100, 2019.

[33] L. R. Rabiner, B. Gold, and C. K. Yuen, "Theory and application of digital signal processing," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 2, pp. 146–146, 1978.

[34] J. H. Lee, Y. Okuno, and S. Cavagnero, "Sensitivity enhancement in solution NMR: Emerging ideas and new frontiers," *Journal of Magnetic Resonance*, vol. 241, pp. 18–31, 2014.

[35] J. H. Ardenkjær-Larsen, B. Fridlund, A. Gram, G. Hansson, L. Hansson, M. H. Lerche, R. Servin, M. Thaning, and K. Golman, "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*, vol. 100, no. 18, pp. 10 158–10 163, 2003.

[36] A. B. Frahm, D. Hill, S. Katsikis, T. Andreassen, J. H. Ardenkjær-Larsen, T. F. Bathen, S. A. Moestue, P. R. Jensen, and M. H. Lerche, "Classification and biomarker identification of prostate tissue from TRAMP mice with hyperpolarized 13C-SIRA," *Talanta*, vol. 235, p. 122 812, 2021.

[37] P. Kolar, M. S. Grbić, and S. Hrabar, "Sensitivity enhancement of nmr spectroscopy receiving chain used in condensed matter physics," *Sensors*, vol. 19, no. 14, p. 3064, 2019.

[38] W. D. Van Horn, A. J. Beel, C. Kang, and C. R. Sanders, "The impact of window functions on NMR-based paramagnetic relaxation enhancement measurements in membrane proteins," *Biochimica et biophysica acta*, vol. 1798, no. 2, pp. 140–149, 2010.

[39] R. E. Hoffman and G. C. Levy, "Modern methods of n m r data processing and data evaluation," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 23, no. 3, pp. 211–258, 1991.

[40] R. MacDonald and S. Sokolenko, "Detection of highly overlapping peaks via adaptive apodization," *Journal of Magnetic Resonance*, vol. 333, p. 107 104, 2021.

[41] A. Alonso, M. A. Rodríguez, M. Vinaixa, R. Tortosa, X. Correig, A. Julià, and S. Marsal, “Focus: A robust workflow for one-dimensional nmr spectral analysis,” *Analytical chemistry*, vol. 86, no. 2, pp. 1160–1169, 2014.

[42] C. Beirnaert, P. Meysman, T. N. Vu, N. Hermans, S. Apers, L. Pieters, A. Covaci, and K. Laukens, “Speaq 2.0: A complete workflow for high-throughput 1d nmr spectra processing and quantification,” *PLoS computational biology*, vol. 14, no. 3, e1006018, 2018.

[43] M. A. Bernstein, S. Sýkora, C. Peng, A. Barba, and C. Cobas, “Optimization and automation of quantitative nmr data extraction,” *Analytical chemistry*, vol. 85, no. 12, pp. 5778–5786, 2013.

[44] A. Savitzky and M. J. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[45] W. Hutton, G. Bretthorst, J. Garbow, and J. Ackerman, “High dynamic-range magnetic resonance spectroscopy (mrs) time-domain signal analysis,” *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, vol. 62, no. 4, pp. 1026–1035, 2009.

[46] G. E. P. Box and K. B. Wilson, “On the experimental attainment of optimum conditions,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 13, no. 1, pp. 1–45, 1951.

[47] K. Wanichthanarak, S. Jeamsripong, N. Pornputtapong, and S. Khoomrung, “Accounting for biological variation with linear mixed-effects modelling improves the quality of clinical metabolomics data,” *Computational and Structural Biotechnology Journal*, vol. 17, pp. 611–618, 2019.

[48] R. Drikvandi, “Nonlinear mixed-effects models for pharmacokinetic data analysis: assessment of the random-effects distribution,” *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 44, no. 3, pp. 223–232, 2017.

[49] J. Lin, X. Yi, and Y. Zhuang, “Medium optimization based on comparative metabolomic analysis of chicken embryo fibroblast df-1 cells,” *RSC Advances*, vol. 9, no. 47, pp. 27 369–27 377, 2019.

[50] B. Mudaser, M. W. Mumtaz, M. T. Akhtar, H. Mukhtar, S. A. Raza, A. A. Shami, and T. Touqeer, “Response surface methodology based extraction optimization to improve pharmacological properties and 1H NMR based metabolite profiling of Azadirachta indica,” *Phytomedicine Plus*, vol. 1, no. 2, p. 100 015, 2021.

[51] D. Jacob, C. Deborde, M. Lefebvre, M. Maucourt, and A. Moing, “NMR-ProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics,” *Metabolomics*, vol. 13, no. 4, p. 36, 2017.

[52] *TopSpin | NMR Data Analysis | Bruker.*

[53] Sung-June Baek, Aaron Park, Young-Jin Ahn, and Jaebum Choo, “Baseline correction using asymmetrically reweighted penalized least squares smoothing,” *Analyst*, vol. 140, no. 1, pp. 250–257, 2014.

[54] F. Savorani, G. Tomasi, and S. Engelsen, “Icoshift: A versatile tool for the rapid alignment of 1d nmr spectra,” *Journal of Magnetic Resonance*, vol. 202, no. 2, pp. 190–202, 2010.

[55] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson, "Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 144–154, 2007.

[56] H. Zhu and M. Luo, "Chemical structure informing statistical hypothesis testing in metabolomics," *Bioinformatics*, vol. 30, no. 4, pp. 514–522, 2013.

[57] J. A. West, A. Beqqali, Z. Ament, P. Elliott, Y. M. Pinto, E. Arbustini, and J. L. Griffin, "A targeted metabolomics assay for cardiac metabolism and demonstration using a mouse model of dilated cardiomyopathy," *Metabolomics*, vol. 12, no. 3, p. 59, 2016.

[58] T. N. Vu and K. Laukens, "Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data," *Metabolites*, vol. 3, no. 2, p. 259, 2013.

[59] K. Krishnamurthy, "CRAFT (complete reduction to amplitude frequency table) – robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR," *Magnetic Resonance in Chemistry*, vol. 51, no. 12, pp. 821–829, 2013.

[60] S. G. Hulse and M. Foroozandeh, "Newton meets Ockham: Parameter estimation and model selection of NMR data with NMR-EsPy," *Journal of Magnetic Resonance*, vol. 338, p. 107 173, 2022.

[61] L. Alacoque and K. James, "Topology optimization with variable loads and supports using a super-gaussian projection function," *Structural and Multidisciplinary Optimization*, vol. 65, no. 2, p. 50, 2022.

[62] D. C. Montgomery, "Design and Analysis of Experiments," in 8th ed. John Wiley & Sons, 2008.

[63] J. Trygg, J. Gullberg, A. I. Johansson, P. Jonsson, and T. Moritz, "Chemometrics in metabolomics — an introduction," in *Plant Metabolomics*, K. Saito, R. A. Dixon, and L. Willmitzer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 117–128.

[64] M. Tanco, E. Viles, and P. Lourdes, "Comparing Different Approaches for Design of Experiments (DoE)," in *Advances in Electrical Engineering and Computational Science*, S. Ao and L. Gelman, Eds. Springer, Dordrecht, 2008, vol. 39, pp. 611–621.

[65] S. Bhattacharya, "Central composite design for response surface methodology and its application in pharmacy," in *Response Surface Methodology in Engineering Science*, P. Kayaroganam, Ed., Rijeka: IntechOpen, 2021, ch. 5.

[66] S. Aoki and A. Takemura, "Design and analysis of fractional factorial experiments from the viewpoint of computational algebraic statistics," *Journal of Statistical Theory and Practice*, vol. 6, no. 1, pp. 147–161, 2010.

[67] K. Vanaja and R. S. Rani, "Design of experiments: Concept and applications of plackett burman design," *Clinical Research and Regulatory Affairs*, vol. 24, no. 1, pp. 1–23, 2007.

[68] K. Kjeldahl and R. Bro, "Some common misunderstanding in chemometrics," *Journal of Chemometrics*, vol. 24, no. 7-8, pp. 558–564, 2010.

[69] H. Wold, *Partial Least Squares*, S. Kotz, C. Read, N. Balakrishnan, B. Vidakovic, and N. Johnson, Eds. John Wiley Sons, Ltd, 2006.

[70] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, and M. E. Timmerman, "ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data," *Bioinformatics*, vol. 21, no. 13, pp. 3043–3048, 2005.

[71] D. Broadhurst, R. Goodacre, S. N. Reinke, J. Kuligowski, I. D. Wilson, M. R. Lewis, and W. B. Dunn, "Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies," *Metabolomics*, vol. 14, no. 6, p. 72, 2018.

[72] D. Grasso, S. Pillozzi, I. Tazza, M. Bertelli, D. A. Campanacci, I. Palchetti, and A. Bernini, "An improved NMR approach for metabolomics of intact serum samples," *Analytical Biochemistry*, vol. 654, p. 114 826, 2022.

[73] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[74] D. P. Kingma, J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.*, 2014.

[75] I. A. Lewis, S. C. Schommer, and J. L. Markley, "Rnmr: Open source software for identifying and quantifying metabolites in nmr spectra," *Magnetic Resonance in Chemistry*, vol. 47, no. 1, S123–S126, 2009.

[76] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[77] D. M. Cooke, *minimize(method='L-BFGS-B') — SciPy v1.9.0 Manual*, 2004.

[78] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[79] J. Nocedal and S. Wright, "Quasi-newton methods," in *Numerical Optimization* (Springer series in operations research and financial engineering), Springer series in operations research and financial engineering. New York, NY: Springer New York, 2006, pp. 135–163.

[80] B. Polyak, "Newton's method and its use in optimization," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1086–1096, 2007.

[81] J. Nocedal and S. Wright, *Numerical optimization* (Springer series in operations research and financial engineering), 2. ed. New York, NY: Springer, 2006, XXII, 664.

[82] X. Yunhai, W. Zengxin, and W. Zhiguo, "A limited memory bfgs-type method for large-scale unconstrained optimization," *Computers Mathematics with Applications*, vol. 56, no. 4, pp. 1001–1009, 2008.

[83] J. Wurts, J. L. Stein, and T. Ersal, "Increasing computational speed of nonlinear model predictive control using analytic gradients of the explicit integration scheme with application to collision imminent steering," in *2018*

IEEE Conference on Control Technology and Applications (CCTA), 2018, pp. 1026–1031.

[84] C. Nita, S. Vandewalle, and J. Meyers, "On the efficiency of gradient based optimization algorithms for dns-based optimal control in a turbulent channel flow," *Computers Fluids*, vol. 125, pp. 11–24, 2015.

[85] A. Baydin, B. Pearlmutter, A. Radul, and J. Siskind, "Automatic differentiation in machine learning: A survey," *Journal of Machine Learning Research*, vol. 18, no. 153, pp. 1–43, 2018.

[86] D. Maclaurin, D. Duvenaud, and R. P. Adams, "Autograd: Effortless gradients in numpy," in *ICML 2015 AutoML workshop*, vol. 238, 2015.

[87] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015.

[88] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[89] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*, IEEE, 2017, pp. 464–472.

[90] D. V. Rubtsov and J. L. Griffin, "Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy," *Journal of Magnetic Resonance*, vol. 188, no. 2, pp. 367–379, 2007.

[91] P. L. De Angelis, P. M. Pardalos, and G. Toraldo, "Quadratic programming with box constraints," in *Developments in Global Optimization*, I. M. Bomze, T. Csendes, R. Horst, and P. M. Pardalos, Eds. Boston, MA: Springer US, 1997, pp. 73–93.

[92] C. Andrieu and A. Doucet, "Joint bayesian model selection and estimation of noisy sinusoids via reversible jump mcmc," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2667–2676, 1999.

[93] M. D. Hoffman and A. Gelman, "The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo," *arXiv:1111.4246*, 2011.

[94] A. Fernández Rodríguez, L. Santiago, E. López-Guillén, J. Rodriguez-Ascariz, J. M. Miguel-Jiménez, and L. Boquete, "Coding prony's method in matlab and applying it to biomedical signal filtering," *BMC Bioinformatics*, vol. 19, no. 1, p. 451, 2018.

[95] Y. Hua and T. Sarkar, "Matrix pencil method and its performance," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 1988, pp. 2476–2479.

[96] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[97] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman and Hall/CRC, 2004.

[98] C. Gavaghan, I. Wilson, and J. Nicholson, "Physiological variation in metabolic phenotyping and functional genomic studies: Use of orthogonal signal correction and pls-da," *FEBS Letters*, vol. 530, no. 1-3, pp. 191–196, 2002.

[99] R. Bro and A. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[100] H.-W. Cho, S. Kim, M. Jeong, Y. Park, N. Gletsu-Miller, T. Ziegler, and D. Jones, "Discovery of metabolite features for the modelling and analysis of high-resolution nmr spectra," *International journal of data mining and bioinformatics*, vol. 2, no. 2, pp. 176–192, 2008.

[101] M. Aries and M. Cloninger, "NMR metabolomic analysis of bacterial resistance pathways using multivalent quaternary ammonium functionalized macromolecules," *Metabolomics*, vol. 16, no. 8, p. 82, 2020.

[102] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[103] F. Fathi, A. Arefi Oskouie, M. Tafazzoli, N. Naderi, K. Sohrabzedeh, S. Fathi, M. Norouzinia, and M. Rostami-Nejad, "Metabonomics based NMR in crohn's disease applying PLS-DA," *Gastroenterology and hepatology from bed to bench*, vol. 6, S82–S86, 2013.

[104] Y. Li, Q. He, Z. Geng, S. Du, Z. Deng, and E. Hasi, "NMR-based metabolomic profiling of peganum harmala l. reveals dynamic variations between different growth stages," *Royal Society Open Science*, vol. 5, no. 7, p. 171 722, 2018.

[105] H. Madsen and P. Thyregod, *Introduction to General and Generalized Linear Models.* Publisher: Chapman Hall/CRC, 2011.

[106] B. Bolker, M. Brooks, C. Clark, S. Geange, J. Poulsen, H. Stevens, and J.-S. White, "Generalized linear mixed models: A practical guide for ecology and evolution," *Trends in ecology evolution*, vol. 24, no. 3, pp. 127–135, 2009.

[107] Y. Hochberg and A. C. Tamhane, *Multiple comparison procedures.* John Wiley & Sons, Inc., 1987.

[108] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer New York Inc., 2001.

[109] T. Lumley and A. J. Miller, "Leaps: Regression subset selection.," 2004.

[110] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[111] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[112] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[113] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, no. 2, pp. 307–333, 1989.

[114] K. G. Brown, “On analysis of variance in the mixed model,” *The Annals of Statistics*, vol. 12, no. 4, pp. 1488–1499, 1984.

[115] C. Bonferroni, “Teoria statistica delle classi e calcolo delle probabilita,” *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, vol. 8, pp. 3–62, 1936.

[116] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian journal of statistics*, no. 2, pp. 65–70, 1979.

[117] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate - a practical and powerful approach to multiple testing,” *J. Royal Statist. Soc., Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[118] A. Salehi M., I. A. Nilsson, J. Figueira, L. M. Thornton, I. Abdulkarim, E. Pålsson, C. M. Bulik, and M. Landén, “Serum profiling of anorexia nervosa: A 1H NMR-based metabolomics study,” *European Neuropsychopharmacology*, vol. 49, pp. 1–10, 2021.

[119] M. Maulidiani, F. Abas, R. Rudiyanto, N. H. A. Kadir, N. K. Z. Zolkeflee, and N. H. Lajis, “Analysis of urinary metabolic alteration in type 2 diabetic rats treated with metformin using the metabolomics of quantitative spectral deconvolution 1H NMR spectroscopy,” *Microchemical Journal*, vol. 153, p. 104 513, 2020.

[120] J. W. Tukey, “The problem of multiple comparisons,” *Princeton University*, 1953.

[121] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020.

[122] G. Van Rossum and F. L. Drake Jr, “Python reference manual,” in Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[123] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.

[124] N. Gowda and D. Raftery, “Overview of NMR spectroscopy-based metabolomics: Opportunities and challenges,” in G. Gowda and D. Raftery, Eds. Methods in molecular biology (Clifton, N.J.), 2019, vol. 2037, pp. 3–14.

[125] A. A. Crook and R. Powers, “Quantitative NMR-based biomedical metabolomics: Current status and applications,” *Molecules*, vol. 25, no. 21, p. 5128, 2020.

[126] P. Kolar, L. Blažok, and D. Bojanjac, “Nmr spectroscopy threshold signal-to-noise ratio,” *tm - Technisches Messen*, vol. 88, no. 9, pp. 571–580, 2021.

[127] S. Kp, “An overview of randomization techniques: An unbiased assessment of outcome in clinical research,” *Journal of human reproductive sciences*, vol. 4, no. 1, pp. 8–11, 2011.

[128] Y. Matviychuk, E. von Harbou, and D. J. Holland, “An experimental validation of a Bayesian model for quantification in NMR spectroscopy,” *J. Magn. Reson.*, vol. 285, pp. 86–100, 2017.

[129] G. E. Box and D. R. Cox, “An analysis of transformations,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[130] F. M. van der Kloet, I. Bobeldijk, E. R. Verheij, and R. H. Jellema, “Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping,” *Journal of proteome research*, vol. 8, no. 11, pp. 5132–5141, 2009.

[131] T. Rundlöf, M. Mathiasson, S. Bekiroglu, B. Hakkarainen, T. Bowden, and T. Arvidsson, “Survey and qualification of internal standards for quantification by 1H NMR spectroscopy,” *Journal of Pharmaceutical and Biomedical Analysis*, vol. 52, no. 5, pp. 645–651, 2010.

Appendix A

NMR-Onion deconvoultion output

Table A.1: NMR-Onion output for two experimental sub-region in one dataset. The confidence interval (CI) for the chemical shifts was set a 95%. Overlapping CIs are marked with a star symbol. Note that the multiplicity is manually assigned whilst the rest of the columns are taken directly from NMR-Onion

sub-ROI No.	Chemical shift (ppm)	lower CI (ppm)	upper CI(ppm)	J (Hz)	multiplicity	Amplitude height
1	6.821*	6.816	6.824	-	-	11188.280
1	6.823*	6.821	6.825	-	-	21588.194
1	6.942	6.939	6.943	-	s	1616350.182
1	7.060	7.057	7.063	-	-	21835.350
1	7.511	7.508	7.513	-	-	27241.110
1	7.642	7.640	7.644	-	s	1550771.714
1	7.772	7.768	7.775	-	-	26343.050
1	7.955	7.951	7.957	-	-	28056.470
1	8.380	8.377	8.384	-	-	11676.760
2	3.810	3.809	3.811	4.6,8.3	m	521104.728
2	3.814	3.813	3.815	4.6,8.5	m	404279.189
2	3.817	3.817	3.82	5.2,8.3	m	702890.560
2	3.821*	3.820	3.822	-	-	27793.815
2	3.822*	3.822	3.822	5.2,8.5	m	613178.961
2	3.836	3.836	3.836	-	-	42411.997

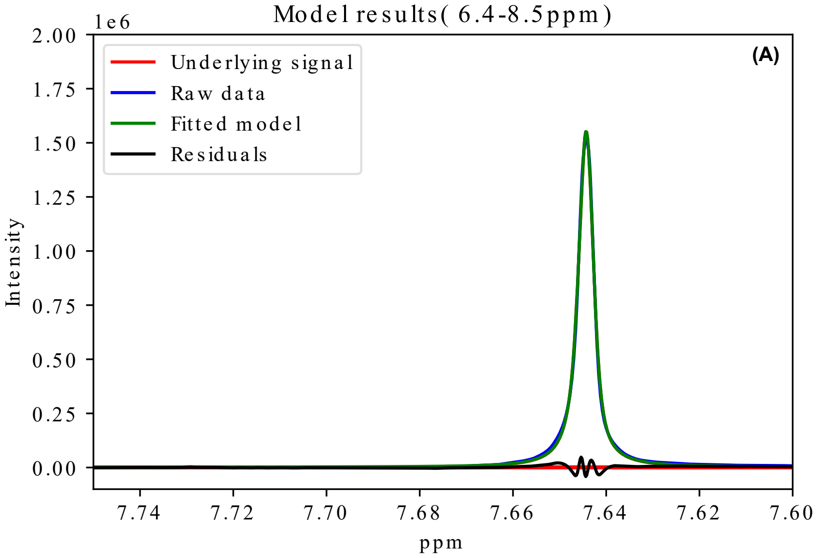


Figure A.1: Zoom in on figure A of the deconvoultion output in chapter 5.

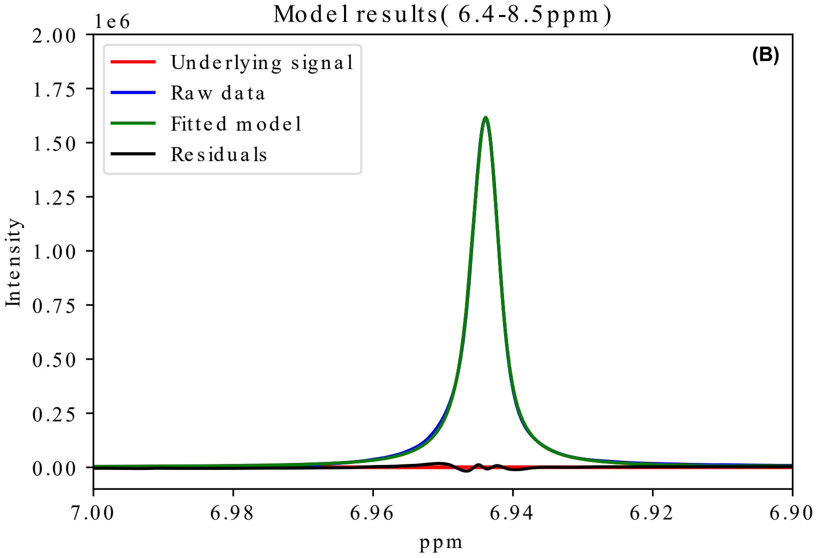


Figure A.2: Zoom in on figure B of the deconvoultion output in chapter 5.

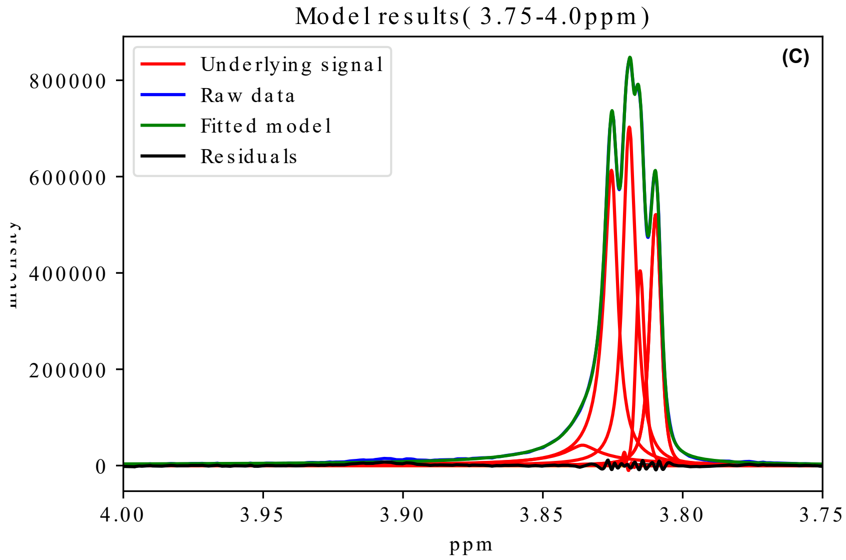


Figure A.3: Zoom in on figure C of the deconvoultion output in chapter 5.

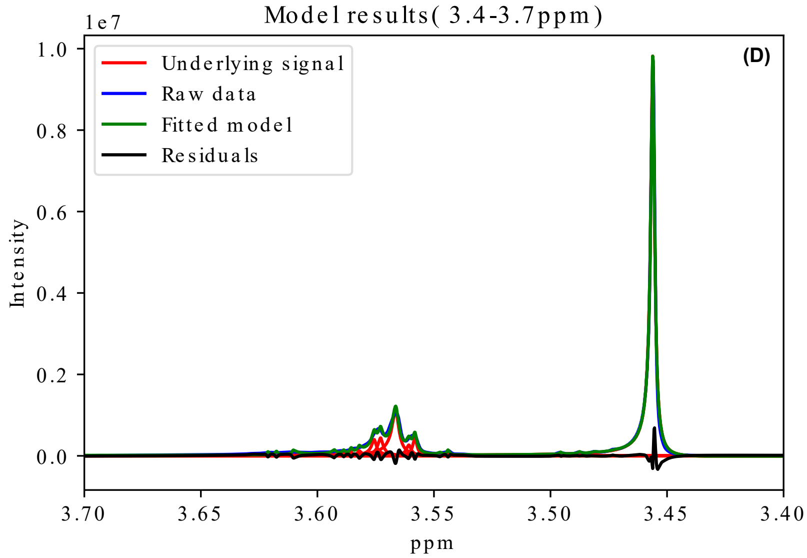


Figure A.4: Zoom in on figure D of the deconvoultion output in chapter 5.

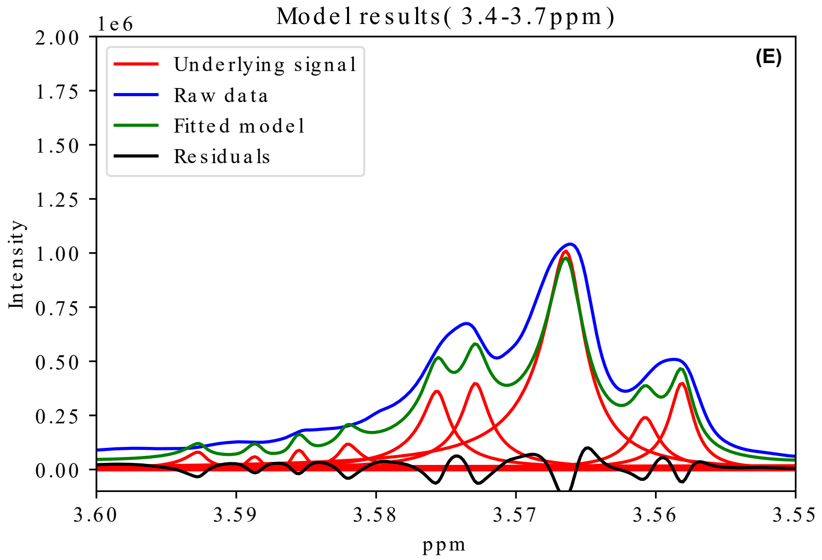


Figure A.5: Zoom in on figure E of the deconvoultion output in chapter 5.

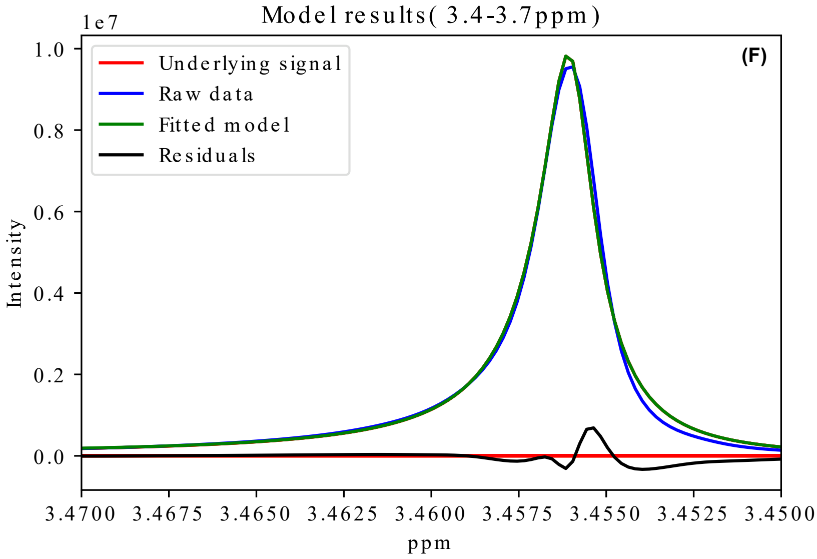


Figure A.6: Zoom in on figure F of the deconvoultion output in chapter 5.

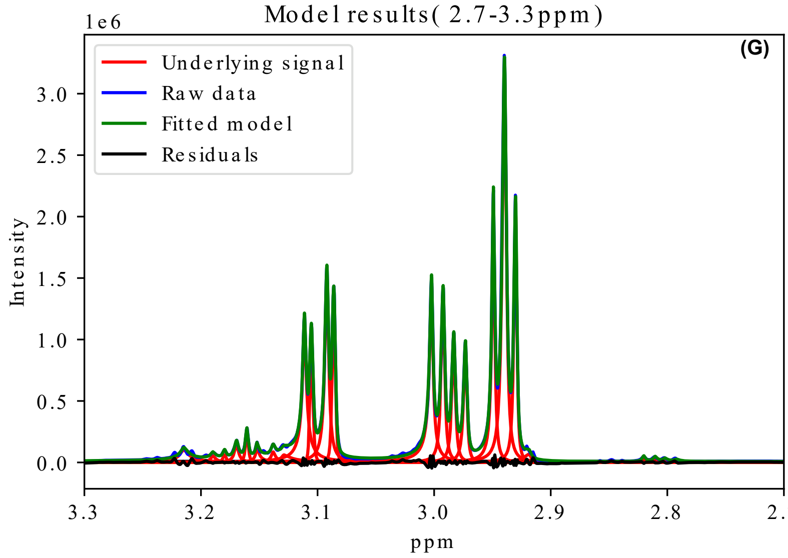


Figure A.7: Zoom in on figure G of the deconvoultion output in chapter 5.

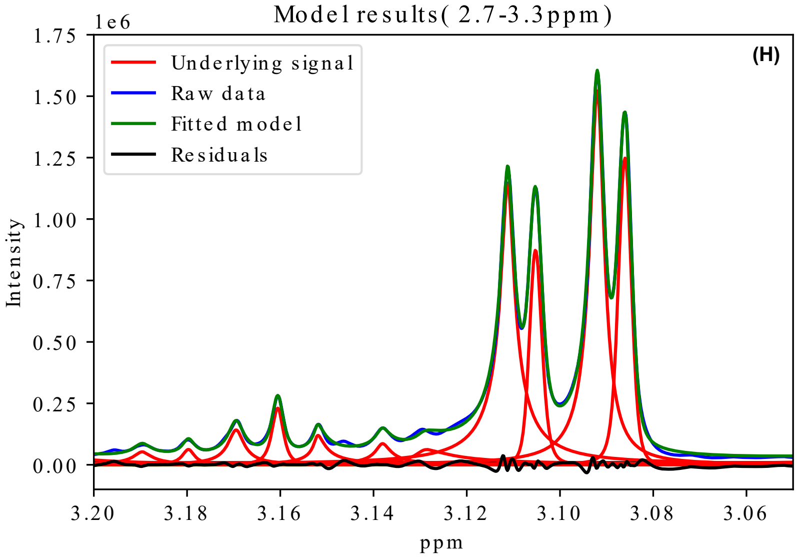


Figure A.8: Zoom in on figure H of the deconvoultion output in chapter 5.

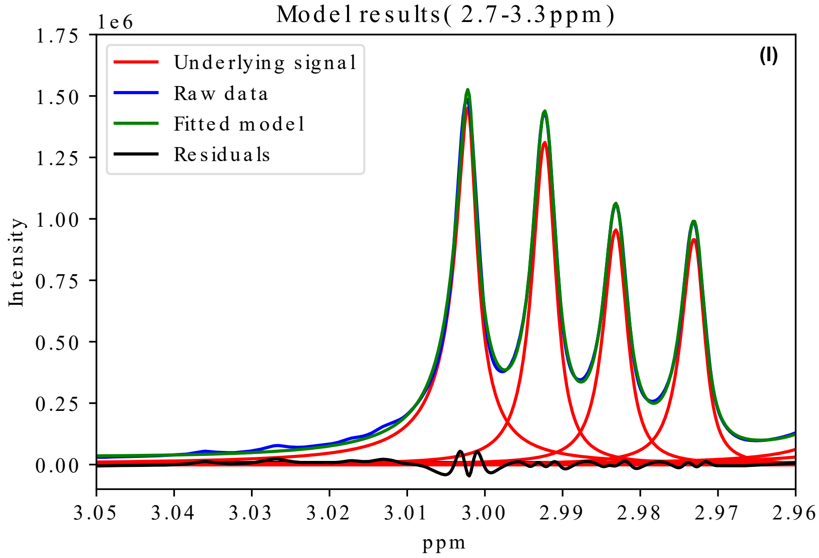


Figure A.9: Zoom in on figure I of the deconvoultion output in chapter 5.

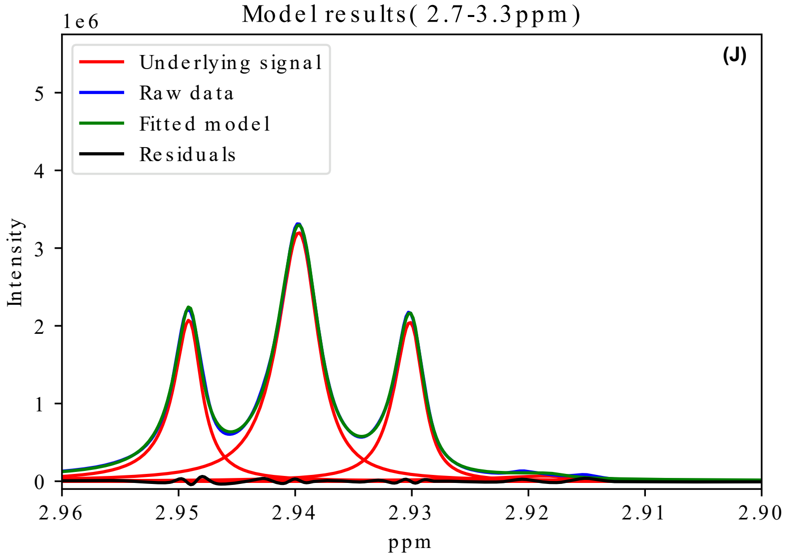


Figure A.10: Zoom in on figure J of the deconvoultion output in chapter 5.

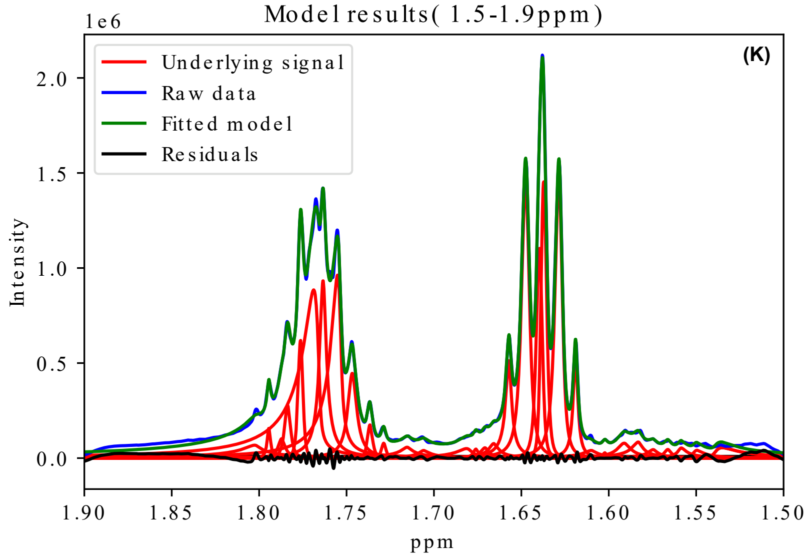


Figure A.11: Zoom in on figure K of the deconvoultion output in chapter 5.

Appendix B

Backwards selection

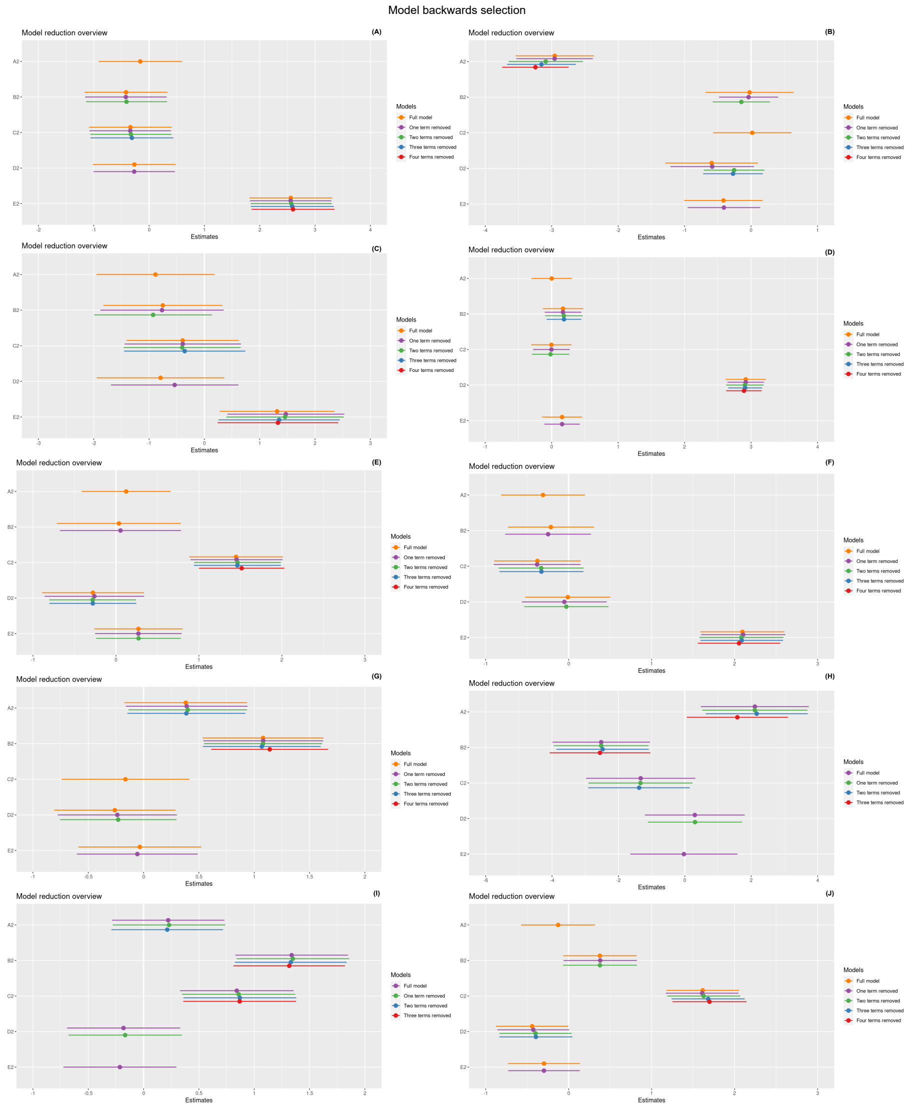


Figure B.1: Figure A-J shows the graphical backwards selection process of each Sub-ROI (A=Sub-ROI1, B=Sub-ROI2 ... J=Sub-ROI10). Each line show the effect size and confidence interval (CI) for each of the coefficients for the general linear effect models at a given model size. A non-significant term is found and removed if the coefficient confidence interval contain 0 (the CI overlaps with 0).

Appendix C

Model visualizations

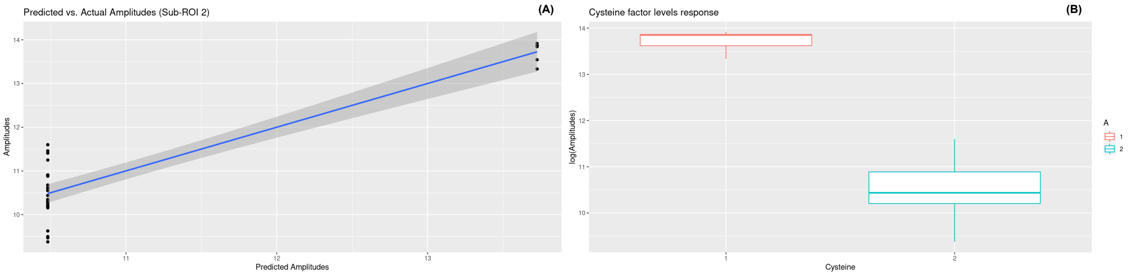


Figure C.1: (A) visualization of model correlation of sub-ROI 2. (B) fixed effect responses of sub-ROI 2.

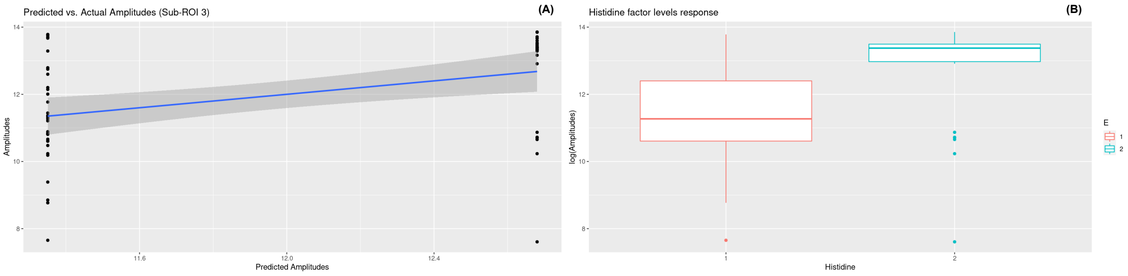


Figure C.2: (A) visualization of model correlation of sub-ROI 3. (B) fixed effect responses of sub-ROI 3.

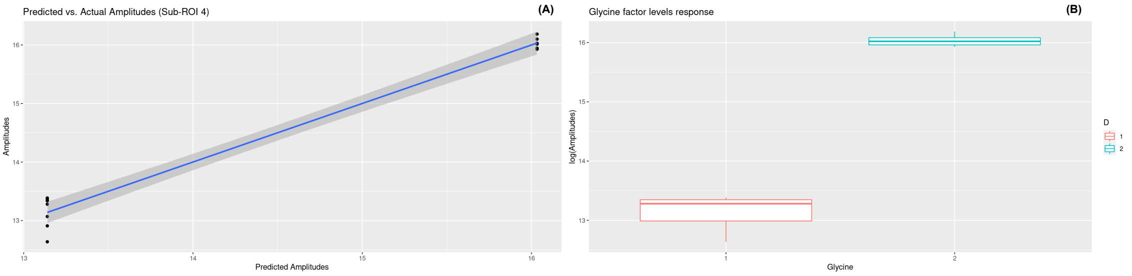


Figure C.3: (A) visualization of model correlation of sub-ROI 4. (B) fixed effect responses of sub-ROI 4.

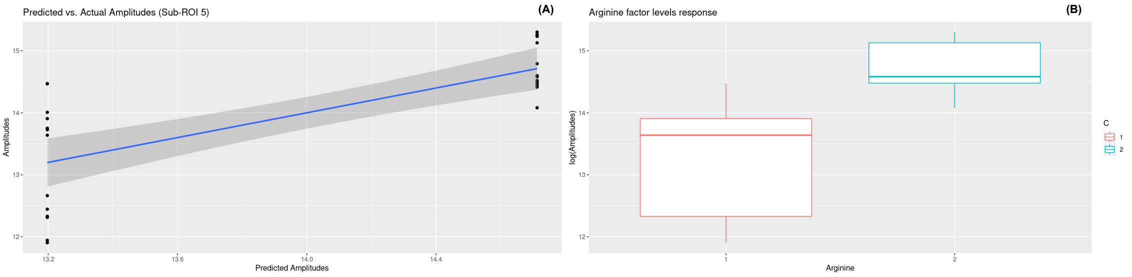


Figure C.4: (A) visualization of model correlation of sub-ROI 5. (B) fixed effect responses of sub-ROI 5.

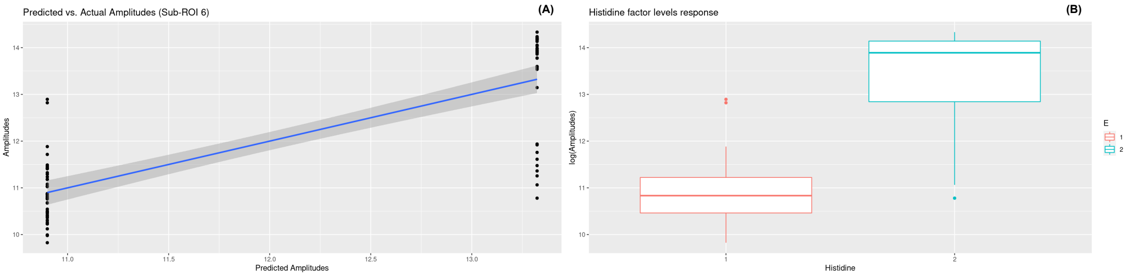


Figure C.5: (A) visualization of model correlation of sub-ROI 6. (B) fixed effect responses of sub-ROI 6.

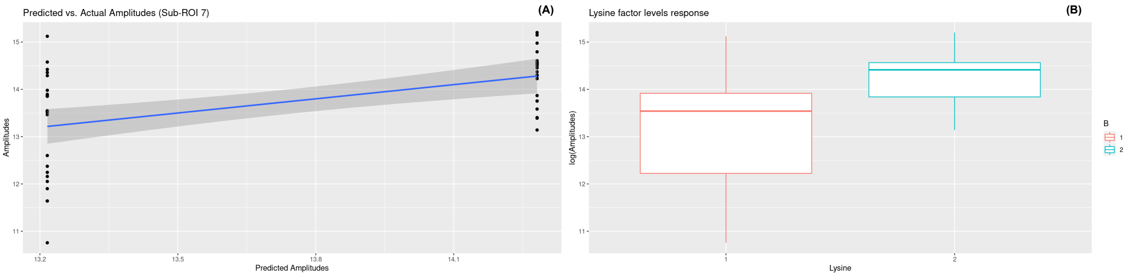


Figure C.6: (A) visualization of model correlation of sub-ROI 7. (B) fixed effect responses of sub-ROI 7.

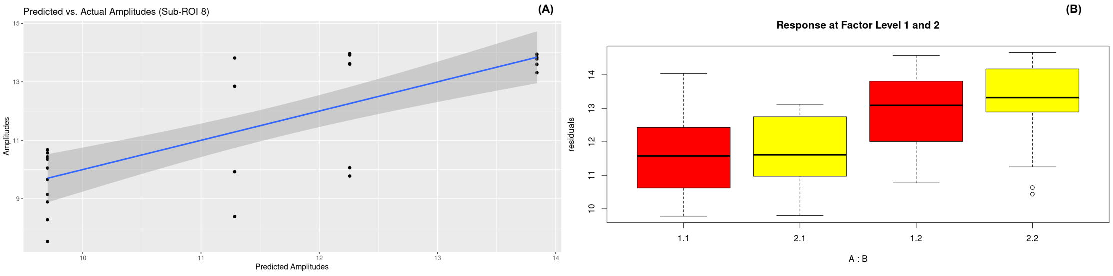


Figure C.7: (A) visualization of model correlation of sub-ROI 8. (B) fixed effect responses of sub-ROI 8, B=Lysine (yellow) and A=Cystiene (red).

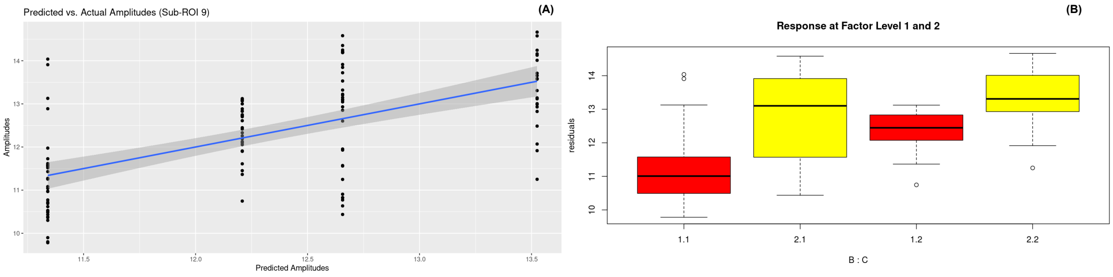


Figure C.8: (A) visualization of model correlation of sub-ROI 9. (B) fixed effect responses of sub-ROI 9, B=Lysine (red) and A=Argininie (yellow).

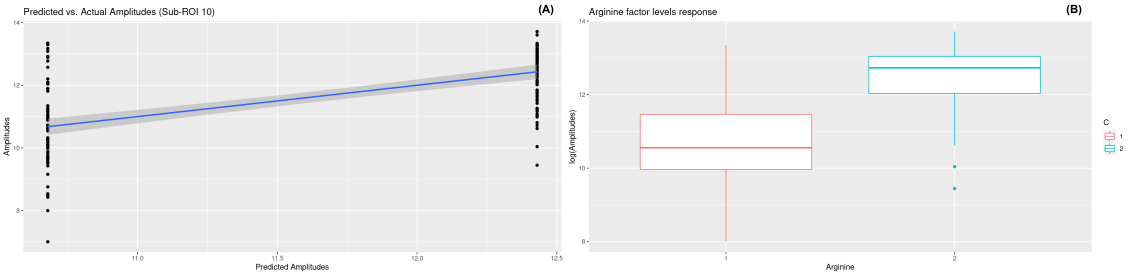


Figure C.9: (A) visualization of model correlation of sub-ROI 10. (B) fixed effect responses of sub-ROI 10.

Appendix D

Model diagnostics

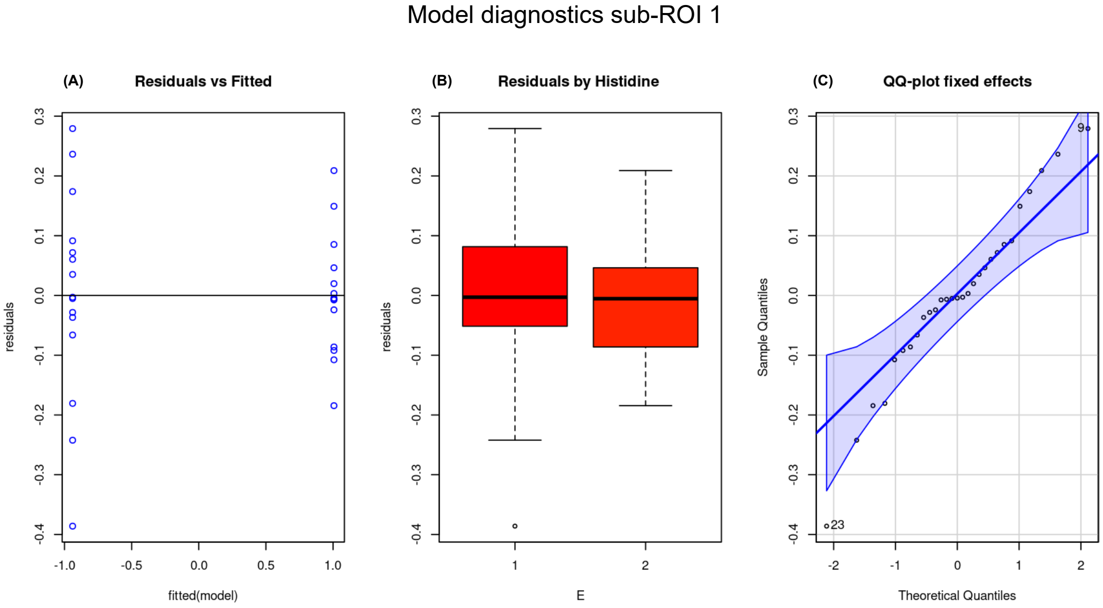


Figure D.1: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

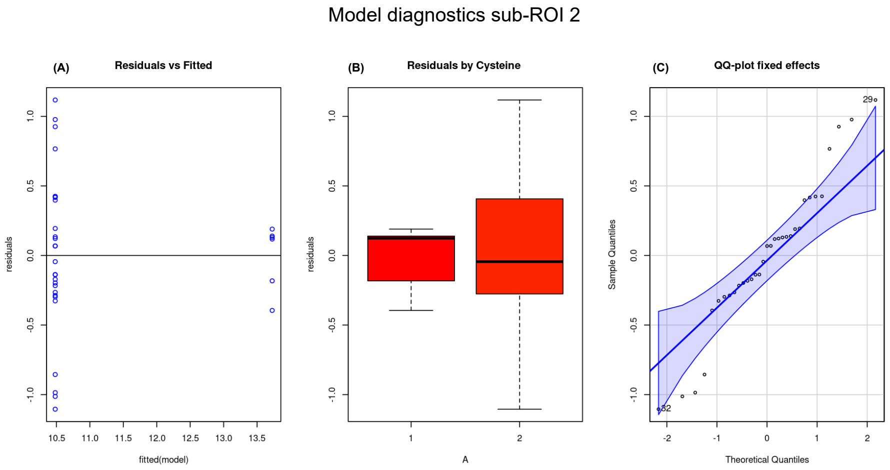


Figure D.2: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

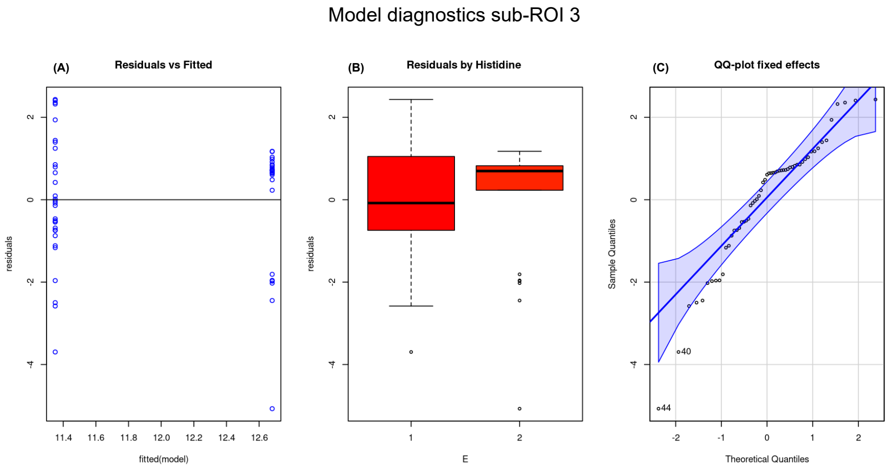


Figure D.3: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

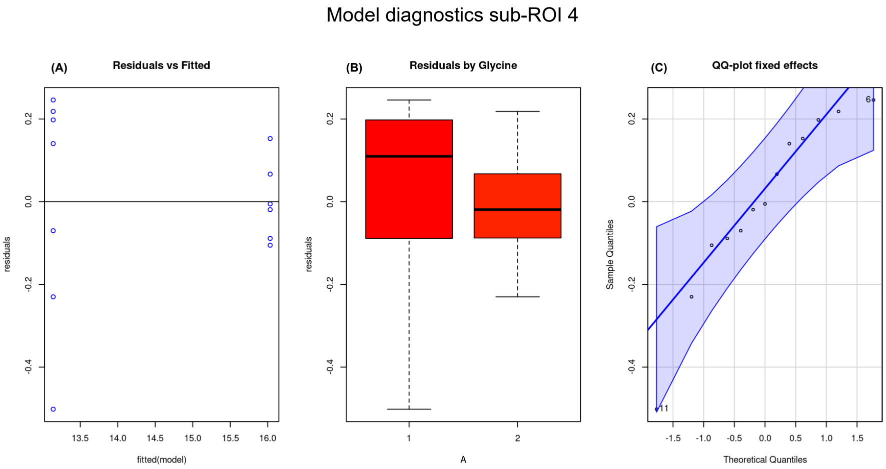


Figure D.4: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

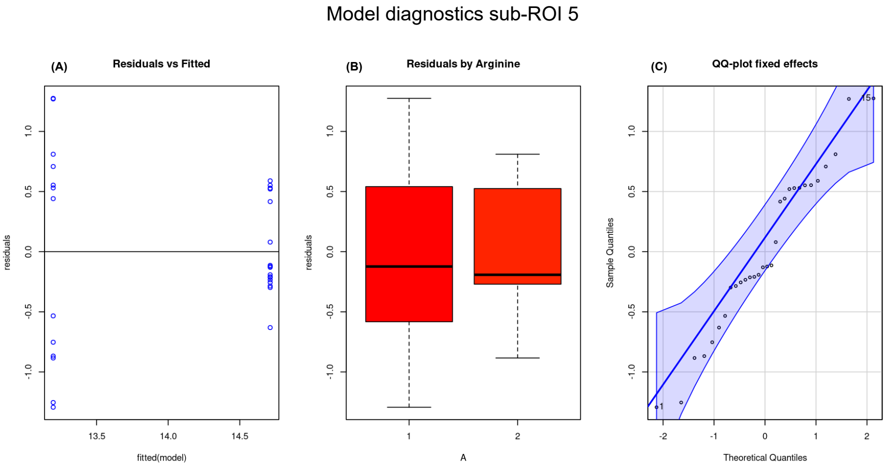


Figure D.5: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

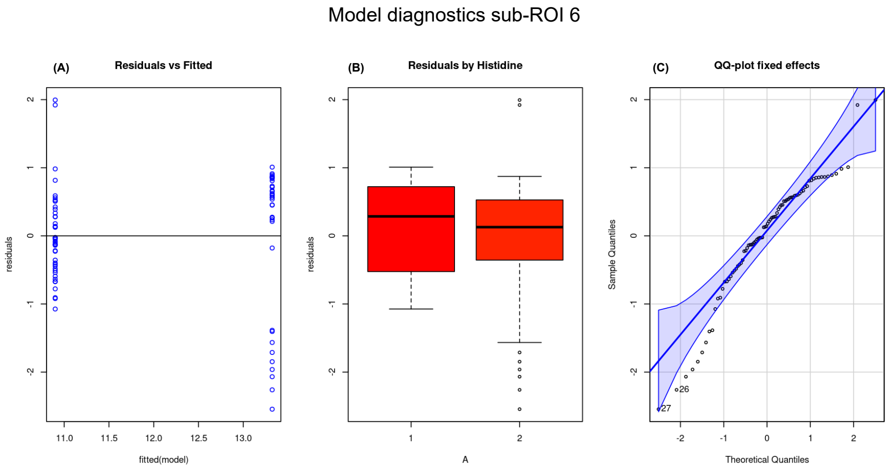


Figure D.6: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

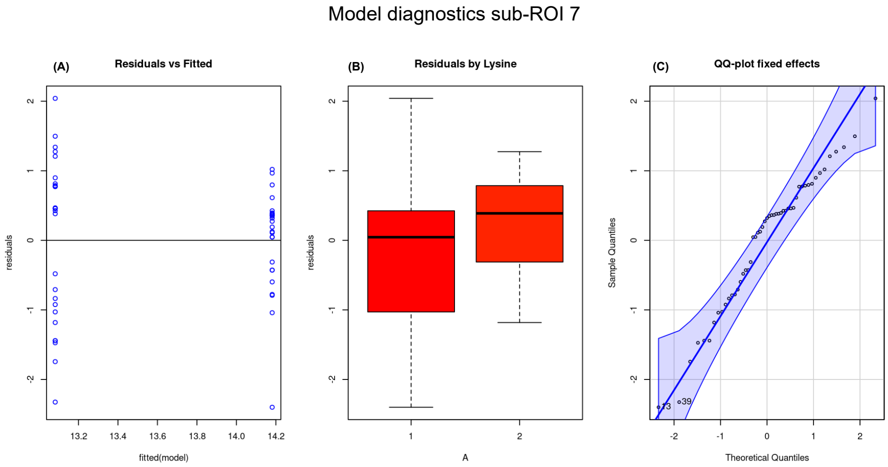


Figure D.7: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

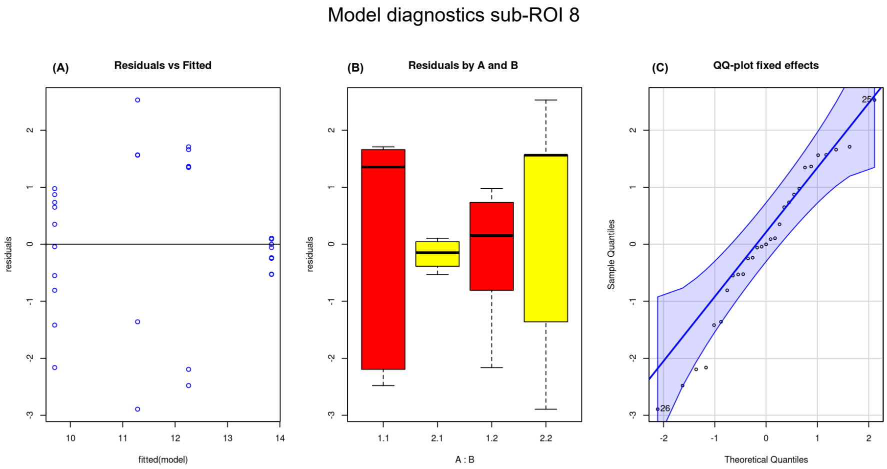


Figure D.8: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses (B=Lysine and A=Cystine). (C) Normality visualization for fixed effects.

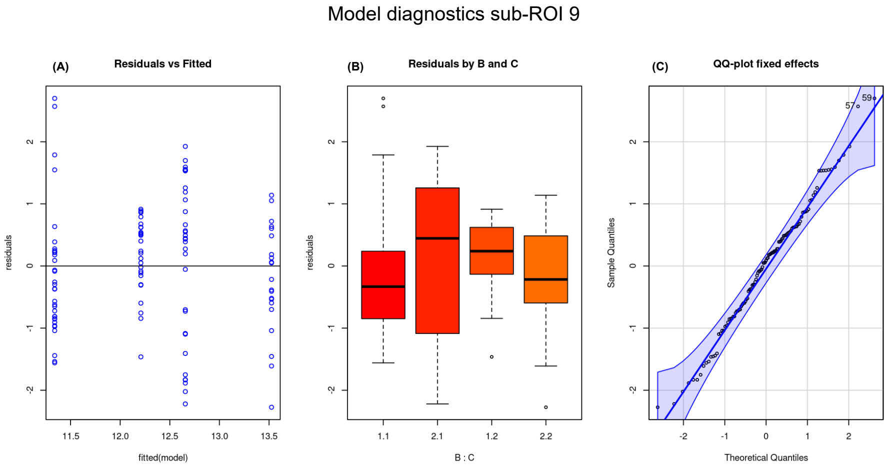


Figure D.9: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses (B=Lysine and C=Arginine). (C) Normality visualization for fixed effects.

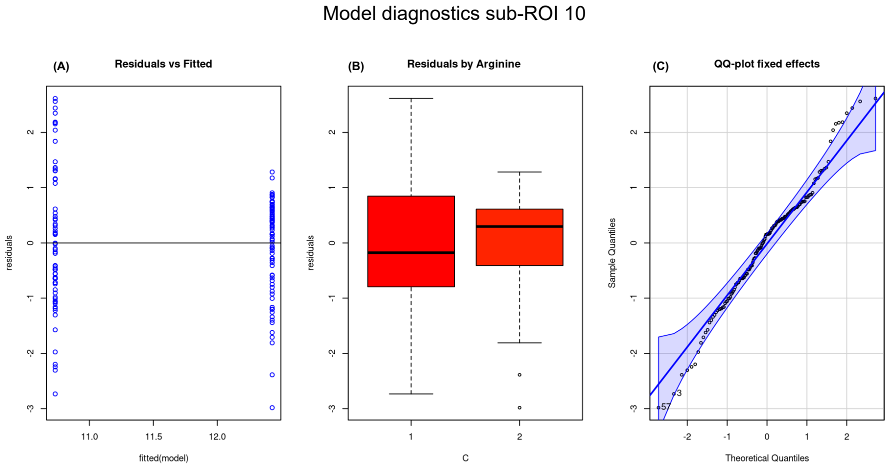


Figure D.10: (A) Visualization of variance homogeneity. (B) Residuals vs factor responses. (C) Normality visualization for fixed effects.

Appendix E

Paper 1

Designing optimal experiments in metabolomics

Mathies Brinks Sørensen[1], Jan Kloppenborg Møller[2], Mikael Lenz Strube[3] and Charlotte Held Gotfredsen[1*]

[1]Department of Chemistry, Technical University of Denmark, Kemitorvet, Kgs Lyngby, 2800, Hovedstaden, Denmark.

[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Asmussens Allé, Kgs Lyngby, 2800, Hovedstaden, Denmark.

[3]DTU BIOENGINEERING Department of Biotechnology and Biomedicine, Technical University of Denmark, Søltofts Plads, Kgs Lyngby, 2800, Hovedstaden, Denmark.

*Corresponding author(s). E-mail(s): chg@kemi.dtu.dk;
Contributing authors: mabso@kemi.dtu.dk; jkmo@dtu.dk; milst@dtu.dk;

Abstract

Background Metabolomics data is often complex due to the high number of metabolites, chemical diversity, and dependence on sample preparation. This makes it challenging to detect significant differences between factor levels and to obtain accurate and reliable data. To address these challenges, the use of Design of Experiments (DoE) and Statistical Quality Control (SQC) techniques in the setup of metabolomic experiments is crucial. DoE techniques can be used to optimize the experimental design space, ensuring that the maximum amount of information is obtained from a limited sample space. SQC techniques can be used to identify and monitor sources of variation, enabling researchers to control and reduce the variability in their experiments.

Aim of Review This review aims at providing a baseline workflow for applying design of experiment (DoE) and statistical quality control (SQC) when generating metabolomics data.

Key scientific concepts of review The review provides insights into the theory of DoE and SQC techniques. The review showcases the theory being put into practice by highlighting different examples of SQC and DoE being applied in metabolomics throughout the literature, considering both targeted and untargeted metabolomic studies in which the data was acquired using both nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry techniques. In addition, the review presents DoE concepts not currently being applied in metabolomics, highlighting these as potential future prospects

Keywords: Design of Experiments · Metabolomics · Nuclear magnetic resonance · Mass spectrometry · Statistical quality control

1 Introduction

Metabolomics is an important tool for analyzing the metabolomic response in many biological systems. The metabolomic response may be evaluated through a targeted or untargeted approach focusing on determining the presence/absence and/or concentration of specific metabolites, related to specific experimental settings being the end goal. Whereas a non-targeted analysis, there is no specific target present, but the metabolome as a whole is evaluated with respect to the impact of the experimental condition, the goal being observing and examining the difference in responses. Examples of targeted and non-targeted analysis thought the literature are many, but some of the major areas are food authenticity e.g. meat (von Bargen et al, 2014) or wine (Pinu, 2018), personalized medicine (Song et al, 2020; Zhang et al, 2020; Breier et al, 2014) or within ecology (Vetting et al, 2015; Poulin and Pohnert, 2019).

The aforementioned studies may be from diverse scientific fields, but they all have the same scope in mind.

- *Can the acquired metabolomics data explain a certain biological/sensory/chemical outcome given different experimental conditions?*
- *Would certain conditions produce a different metabolomics outcome?*

Therefore, the metabolomic outcome can either be seen as the response or a covariate, explaining a different response such as in diagnostics of diseases (Geng et al, 2020) or sensory investigations (Castro-Alves et al, 2021). Despite the goal, of uncovering the effect of the metabolome within the aforementioned fields, metabolomics studies are highly driven by mathematical hypothesis testing, that is can the significance of experimental factors within the design space of the study be identified? Or can the response be predicted from the metabolomic outcome (based on an experimental design space)? The relation for investigating the significance of experimental factors within a design linked to a metabolomic outcome may be mathematically written as:

$$H_0 : \boldsymbol{\theta_i} = \boldsymbol{\theta_j} \tag{1}$$

$$H_1 : \boldsymbol{\theta_i} \neq \boldsymbol{\theta_j}, \tag{2}$$

Here $\boldsymbol{\theta}$ represents a vector of experimental factors such as temperature, media composition, pH, species, etc., found at the factor levels of j and i (or possibly more). As an example, a two-level factor setting could be the recording of responses based on pH at $i = 7$ and $i = 5$. The Null hypothesis H_0 states that the response of the factor levels (i and j) are identical. This implies that a significant difference between the response recorded at the $i'th$ and $j'th$ level would only occur if the probability (p-value) of H_0 being true is lower than a given level of significance (traditionally set at $\alpha = 0.05$). The reversely is stated in the alternative hypothesis H_1, given that factor levels are assumed to be different. The goal of hypothesis testing is thus falsifying the H_0 at certain levels, indicating a factor or sub-level of the factor would have a significant impact on the design space. The most common statistical method of hypothesis testing, within a design space, is that of analysis of variance (ANOVA) (Girden, 1992), while the design space itself is set up through design of experiments (DoE) (Madsen and Thyregod, 2011). In addition to the DoE, the aspect of statistical quality control (SQC) (Montgomery, 2009), ensures the reproducibility and validity of data, increasing certainty that a possible significant effect does not occur due to instrumental error or other measurement artifacts (Broadhurst et al, 2018; van der Kloet et al, 2009).

To emphasize how an optimal design space may be set up and analyzed within a metabolomic framework, we have reviewed, recent literature within DoE and SQC, highlighting the ongoing efforts in the metabolomic field to secure the generation of optimal informative data which is of key importance in an increasingly data-driven world.

2 Statistical study design

Within metabolomics, experiments are often formulated to determine the effects of treatment interventions based on data. However, the data is not unlimited, as experiments can be tedious, time-consuming, and expensive. Therefore, effective usage of sample size and replicates needs to be considered when generating data, that is optimal organization of samples within a design space such that maximal information is achieved with respect to (w.r.t.) treatments effects interventions, utilizing as few samples as possible. To achieve optimal design space organization, the concept of DoE plays a key role in ensuring proper data generation. The principles of DoE are heavily used in classical industries such as maintenance and repair (Hill et al, 2017) as part of the six-sigma principle (Montgomery, 2009), but has also found its way into many diverse fields such as surgery optimization (Bertolaccini et al, 2015) or laboratory efficiency improvements (Inal et al, 2018). The increased usage is also evident in metabolomics in which the usage of DoE has been steadily increasing over the past 20 years (see figure 1). However, blindly applying DoE leads to poor results, as no one size fit all design type exists and different design types are suited for different types of experiments. Therefore, to aid in deciding which types of design may be helpful in common experiments, we emphasize some of the more applicable designs for metabolomics in the next paragraphs.

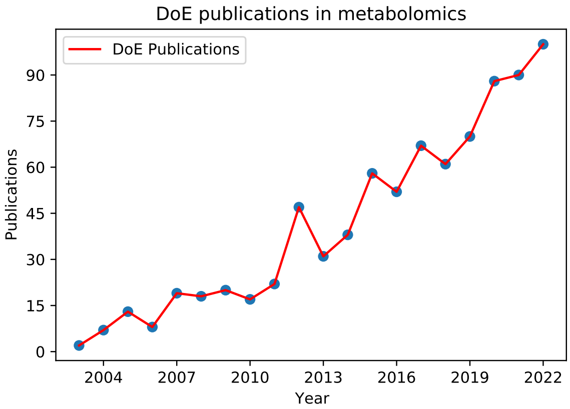


Fig. 1 Journal papers within metabolomics involving DoE over the past 20 years. Results found within Web of Science, searching the phrase: "metabolomics design of experiments"

It should be noted that there are many other types of design principles than presented in this review, and for the more curious readers or those not too familiar with DoE, we recommend the books of Montgomery (Montgomery, 2008) and Quinn (Quinn and Keough, 2002). Montgomery explains DoE aimed at an audience with mathematical/engineering background, while Quinn explains DoE based on a biological/chemical background.

2.1 Design types

There are many experimental designs, and choosing the correct one can often be difficult. In general, the choice of design should be made based on the resolution level required to answer a specific experimental question. In metabolomics the experimental questions vary greatly, spanning everywhere from cancer diagnostics to cheese quality optimization. However, despite differing fields, two types of settings within DoE are commonly encountered throughout all the metabolomics literature, being either screening or optimization investigation studies.

2.1.1 Screening designs

When screening for factors of importance without a proper design, the number of possible combinations quickly becomes an issue. As an example, investigating k=5 factors at p=2 levels with r=3 replicates would require 96 samples for the design space to have full resolution, resulting in a p^k full factorial design (FFD). The term full resolution refers to the design, being able to capture information w.r.t all factor effects and their interactions. In the case of 5 factors, the design would be able to capture and test the significance of all main effects and up to 5 order interaction effects. The FFD is by no means wrong for screening, but is it suitable? Here, the concept of achieved resolution vs needed resolution should be considered, that is one should take into account resolution redundancy of higher order terms when screening for significant main effects or possibly including 2-factor interactions.
The sparsity of effects principle (Wu and Hamada, 2011) that comes with the factorial design can be applied to accommodate resolution redundancy. The principle implies that interactions above a certain resolution level do not contain significant information and may safely be aliased with lower order terms when modelling the data, leading to a fractional factorial design (Frac-FD) of 2^{k-p}. To clarify the concept of aliasing and resolution, we consider the previously mentioned five-factor example, consisting of factors A, B, C, D, and E. In a screening context, as often occurs in metabolomics, the goal would be to assess the significance of main effects and possibly two-factor interactions. To do so efficiently, the goal would be to make sure the two-factor interactions and main effect are not aliased with one another (the significance of an effect is not the result of an effect being aliased/confounded). The simplest approach would be a half-fractional factorial design of 2^{5-1}, making the resolution V design of table 1.

The resulting effects of imposing the generated design of table 1 onto an experimental set would be a sample reduction from 32 to 16 samples per replicate. How does this make sense? The idea is to use a generator for the design $I = ABCDE$,

Table 1 Overview of 2^{5-1} fractional factorial design, including treatments (Tr.) and aliasing structure.

Run	A	B	C	D	E	Tr.	Alias
1	-	-	-	-	+	E	$E = ABCD$
2	+	-	-	-	-	A	$A = BCDE$
3	-	+	-	-	-	B	$B = ACDE$
4	+	+	-	-	+	ABE	$ABE = CD$
5	+	-	+	-	+	ACE	$ACE = BD$
6	+	+	+	-	+	BCE	$BCE = AD$
7	+	+	+	-	-	ABC	$ABC = DE$
8	-	-	-	+	-	D	$D = ABCE$
9	+	-	-	+	+	ADE	$ADE = BC$
10	-	+	-	+	-	BDE	$BDE = AC$
11	+	+	-	+	-	ABD	$ABD = CE$
12	-	-	+	+	+	CDE	$CDE = AB$
13	-	-	+	+	+	ABD	$ABD = CE$
14	+	-	+	+	-	ACD	$ACD = BE$
15	-	+	+	+	-	BCD	$BCD = AE$
16	+	+	+	+	+	$ABCD$	$ABCD = I$

with I being the identity column. Hence expressing the missing factor E as $E = ABCD$, therefore, E is now aliased with $ABCD$. Consequently, as higher order terms have less impact, the point is E is effectively measured from I. One, may find the aliasing structure for the rest of the effects and interactions as $I \cdot A = A^2BCDE = BCDE$, hence A is aliased with the four order effect of $BCDE$. A continuation of the calculation would eventually generate all rows of table 1.

Ultimately, the defining relation in table 1 can be modified further as $I = ABD = ACE = BCDE$ such that only main effects become distinguishable, imposing the design generator of $D = AB$ and $E = AC$, resulting in a resolution III 2^{5-2} frac-FD design stated in table 2

Table 2 Overview of 2^{5-2} fractional factorial design, including treatments (Tr.) and aliasing structure.

Run	A	B	C	D	E	Tr.	Alias
1	+	-	-	-	-	A	$A = BD = CE$
2	-	+	-	-	-	B	$B = AD = CDE$
3	-	-	+	-	-	C	$C = AE = BDE$
4	-	-	-	+	-	D	$D = AB = BCE$
5	-	-	-	-	+	E	$E = AC = BCD$
6	-	+	+	-	-	BC	$BC = DE = ACD = ABE$
7	-	-	+	+	-	CD	$CD = BE = ABC = ADE$
8	+	+	+	+	+	$ABCDE$	$ABCDE = I$

The design based on table 2, results in 8 samples per replicate. No further meaningful fractions can be adapted, as main effects would be aliased with main effects making for an indistinguishable response. Generally, all designs based on factorial design can be found in tables or generated via computer programs such as R (RStudio Team, 2020) or SAS JMP (Institute, 1985). For none programming experts we recommend SAS JMP over R as a GUI is provided, though the program is proprietary software.

The general two-level resolution screening design types are visualized for 3,4,5 and 6 factors in figure 2, stating both the number of runs (N) and the effects of choosing a specific design.

Another design type extensively applied within metabolomics is the designated screen design method of the Plackett Burman design (PBD) (Vanaja and Rani, 2007). This type of design can be used for certain scenarios where the number of runs is equal to a multiple of 4, testing up to k=N-1 factors with the largest smallest design being N=12 and k=11 (when N is a power of 2, a fractional factorial design is achieved (Vanaja and Rani, 2007)). Unlike the resolution III design, the PBD has a complicated aliasing structure between main and quadratic effects, in which partial aliasing is present. For instance when N=12, the main effects are aliased with two-factor interactions not involving the effect itself. Hence, care must be taken with this type of design, as it is excellent for determining if a main effect has a significant contribution, but not for determining the exact contribution of the effect. Therefore, PBD is an efficient design, but it has a risk of introducing type 1 errors (detecting false positives). However, during early phase screening type 1 errors are not crucial, as non-significant factors would eventually be detected through further testing. An upside is that the risk of type 2 errors (false negative) is lowered with this type of design, as potential significant effects are not discarded, though they can be masked by partial aliasing. In summary, PBD is suitable for removing the most insignificant factors during initial screening but should be backed up with a different design for a more exact estimation of effects (eg frac-FFD or FFD when sample space has been lowered).

In addition to screening of identical factor levels, the need for identifying factors of importance within a mixed-level scenario is at times required within metabolomics (eg. one factor may have 2

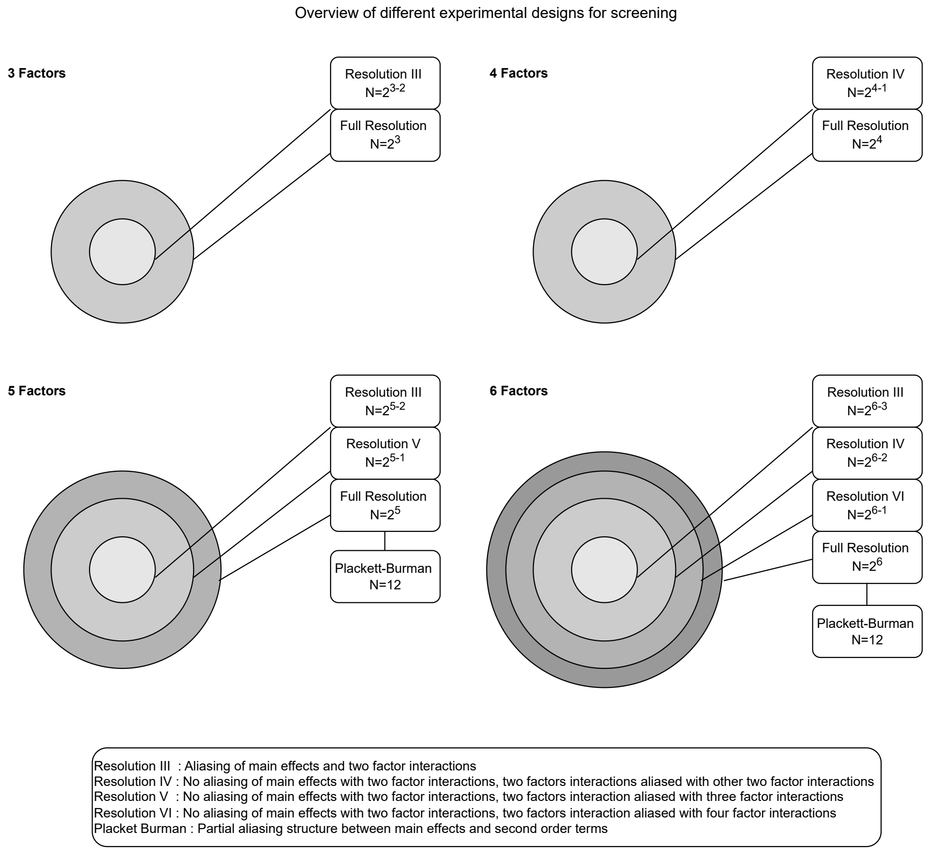

Overview of different experimental designs for screening
3 Factors
Resolution III N=2^{3-2}
Full Resolution N=2^3
4 Factors
Resolution IV N=2^{4-1}
Full Resolution N=2^4
5 Factors
Resolution III N=2^{5-2}
Resolution V N=2^{5-1}
Full Resolution N=2^5
Plackett-Burman N=12
6 Factors
Resolution III N=2^{6-3}
Resolution IV N=2^{6-2}
Resolution VI N=2^{6-1}
Full Resolution N=2^6
Plackett-Burman N=12
Resolution III : Aliasing of main effects and two factor interactions
Resolution IV : No aliasing of main effects with two factor interactions, two factors interactions aliased with other two factor interactions
Resolution V : No aliasing of main effects with two factor interactions, two factors interaction aliased with three factor interactions
Resolution VI : No aliasing of main effects with two factor interactions, two factors interaction aliased with four factor interactions
Placket Burman : Partial aliasing structure between main effects and second order terms

Fig. 2 Overview of different screening designs for 3,4,5 and 6 factors including the effects of choosing a specific design resolution.

levels whilst another has 3 levels). Efficient screening designs for mixed-level experiments may be generated by applying the Taguchi designs (TD) (Kacker et al, 1991). The main idea of TD is to construct orthogonal arrays such that factors may be evaluated independently, despite being of a highly fractional design. Table 3 shows an example of a mixed-level TD design of 4 two-level factors (A, B, C, D) and 1 four-level factor (E), resulting in N=8 runs, whereas a full combination would yield N=20 runs. The design is denoted to be $L8(2^{4\times}4^1)$, indicating eight experimental trials capable of investigating up to 4 two-level factors and 1 four level factor.

The side effect of the TD being highly orthogonal is that limited information about interaction effects may be extracted, making TD suitable for identifying main effects. We note that within metabolomics the TD mixed-level design is rarely if ever utilized (perhaps due to the usage of optimal design algorithm instead), nevertheless, the technique is included, as it is a standard method implemented within most software.

Table 3 Overview of $L8(2^{4\times}4^1)$ Taguchi design, numbers within each factor column indicate factor level. A, B, C, and D can either be 1 or 2 whereas E can be 1, 2, 3, or 4

Run	A	B	C	D	E
1	1	1	1	1	1
2	2	2	2	2	1
3	1	1	2	2	2
4	2	2	1	1	2
5	1	2	1	2	3
6	2	1	2	1	3
7	1	2	2	1	4
8	2	1	1	2	4

2.1.2 Optimization

Having identified significant main effects (possibly including 2-factor interactions) during screening, the subsequent step would often be the optimization of experimental conditions. To achieve an optimized design, relaying the effects of higher-order terms such as quadratic effects (A^2,B^2, etc.) is often crucial. However, blindly applying a simple FFD design to extract quadratic information be very costly. The increased cost arises from the inclusion of an additional factor level needed for proper investigation, that is a 2^k design becomes a 3^k design, increasing from 8 samples per replicate to 27 samples per replicate (see figure 3A). To overcome the challenges of an additional factor level, a popular method is the use of response surface methodology (RSM) designs (Box and Wilson, 1951), in which the Box-Bekhen design (BBD) figure 3B and central composite design (CCD) (Bhattacharya, 2021) figure 3C are very popular choices in metabolomics.

In principle, BBD and CCD achieve the same goal of generating quadratic information, therefore it can be difficult to know when to apply which of the two designs. In general, the BBD should be applied if there is a high confidence in the design space limits, that is no settings make sense outside a certain concentration, temperature, pH, etc within the design space. The CCD enables investigating settings outside a strictly defined region (the range of factors in the factorial design) via axial points (or star points) and includes the corners of a design space - see figure 3C.

The concept of organizing axial points in the CCD may not be very clear, for what is a suitable range outside the factor range? To clarify, the concept of design rotatability is introduced. For a design to be rotatable, the response prediction variance should be the same for all points with an equal distance from the design center, such that equal estimation precision is achieved in all directions (Draper, 2008). To achieve rotatability, some of the most common organizing methods are the spherical design (SD) or the general rotatable design (GRD). The SD provides a uniform distance of $\alpha = \sqrt{k}$ from the center with k=number of factors and are thus rotatable since they have an equal prediction variance. The GRD has a distance of $\alpha = (2^k)^{1/4}$ this has the same properties regarding rotatability, but at larger values of k produces much larger distances than the SD. Thus in many real applications, the GRD-based star points may be difficult to measure due to value constraints of the design space, whilst the SD-based star points are more obtainable. A special case occurs when $\alpha = 1$, resulting in axial points being on the boundary of the design space known as a face-centered central composite design (CCF) (Montgomery, 2008). The CCF can be applied if one is uncertain of the extremeness of axial points or if certain conditions cannot be outside the range of the design space. Comparing CCF with BBD, the difference is still the extremeness, as CCF includes corner points, whilst BBD does not. So, BBD produces a more narrow design than CCF and can be used if one is very certain of the design space.

2.2 Computer aided designs

For optimization, all of the above techniques only work within a regular symmetric design space (factors are varied systematically and uniformly across samples). However, cases with irregular design space (factors are not varied uniformly across samples, or the experimental conditions are not the same for all samples) often occur, where the design space is not a cube or sphere. The cause of irregularities can be experimental constraints, for instance, a specific combination of too high or too low pH and temperature may cause the experiment to fail. A second cause would be if one of the factors in the experiment is categorical (non-quantitative factors, such as solvent types, bacterial species, blood type, etc), as no standard design exists (Montgomery, 2008) involving categorical factors.
To overcome the aforementioned challenges, computer-aided optimal designs were developed, presenting a wide array of criteria for evaluating the quality of an optimal design (Montgomery, 2008). There are multiple algorithms to perform the optimal design for irregular design space. Some of the most widely used are the D, A, G, or V optimal design (Montgomery, 2008). Each of these designs has different properties but is essentially built using the same principle, relying on the evaluation of candidate points for optimal treatment combinations through computer algorithms. One example is the Fedorov exchange algorithm (Miller and Nguyen, 1994) which can be applied (with some modifications) to generate the D, A, and V optimal design. The A, V, and G designs are not, to the best of our knowledge, used

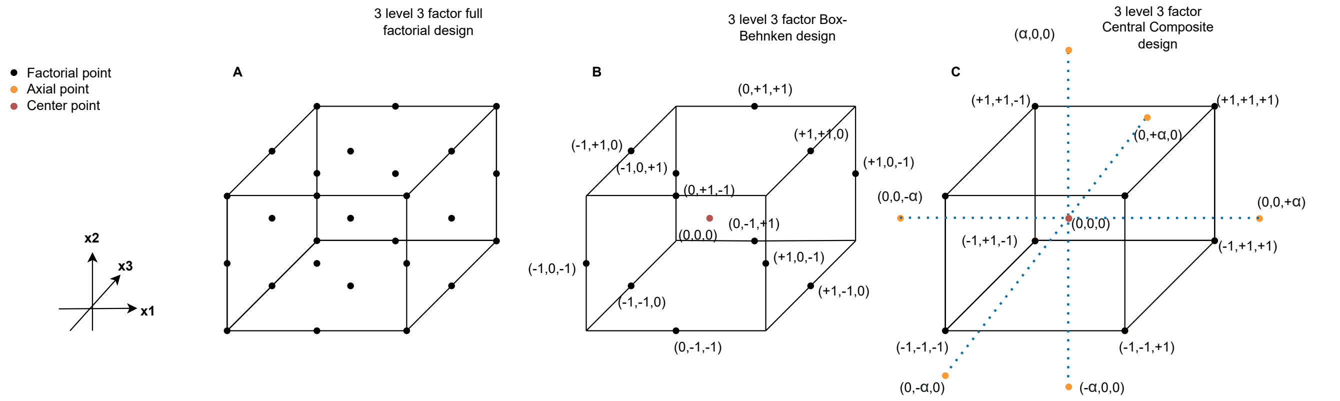


Fig. 3 Overview of different optimization designs for 3 three-level factors including design space spacial coordinates (-1=low,0=medium,+1=high), α represents the distance from the center point

in metabolomics, this may be due to the G design being computationally expensive (Hernandez and Nachtsheim, 2018), whilst V and A might suffer from the information matrix not being inevitable (Shahmohammadi and McAuley, 2019). In this review, we will mainly focus on the D-optimal design and refer to (Montgomery, 2008) for an extensive explanation of the rest of them.

The main idea of the D-optimal design is to construct a design space, such that the variance of the estimated coefficients is minimized. This is achieved by selecting the design matrix X, such that maximizing the determinant of the information matrix occurs, resulting in an optimal design matrix ($X_{optimal}$), that is

$$X_{optimal} = \arg\max_{X} \mid X^T X \mid \tag{3}$$

The design matrix X is defined within the scope of the study and can in principle be of any size, programs such as SAS JMP will let the user define the specific matrix of interest by defining factors and their level of complexity (eg. quadratic, interaction terms, etc.). To evaluate the efficiency of a resulting design vs another ((X_1) vs (X_2)), the D-criterion can be utilized (see (Montgomery, 2008)). The criterion produces a number between 0 and 1, expressing how close a resulting optimal design may be to the nearest frac FFD, where values closer to 1 indicate higher efficiency (less correlated design). The D-optimal design may be applied for all kinds of studies and data, but should not be a default choice over the classic methods as designs generated from the D-optimal design may be correlated in an unpredictable manner.

Another, often overlooked option, of the optimal designs, are the augment or repair design types (Federer and Raghavarao, 1975; Radson and Herrin, 1995). These build on previous experiments and lead to a way of expanding the design space by identifying complementary points based on previously recorded experiments (Lu and Anderson-Cook, 2021). The augmentation strategy has the advantage of expanding already existing experimental data and thus increasing statistical power. The augmentation design has been applied in areas such as agriculture (Federer et al, 2001) and plant bacterial interaction studies (Montañez et al, 2012), but the design is seemingly overlooked in metabolomics. Despite the lack of usage, we believe that the augmented design has a place in the design of non-temporal metabolomic experiments (e.g. all studies which do not have a time-dependent response) as the augmentation may save time and effort by expanding designs in a statistically efficient manner.

2.3 Application of DoE in metabolomics

Within metabolomics, many cases of DoE are found being applied within either a screening or optimization setting. We have collected a wide array of such studies which are based on the techniques described in the previous paragraphs, highlighting the practical use of DoE.

Table 4 Overview of DoE-based metabolomics studies involving screening or optimization.

Screening			
DoE Technique	Analytical technique	Type of study	Reference
fractional factorial design	HPLC-PDA	targeted metabolomic	(Zanatta et al, 2022)
	SELDI-TOF-MS	protomeics profiling	(Szalowska et al, 2007)
	UHPLC-IM-MS	untargeted metabolomics	(Tebani et al, 2016)
	MS/MS	targeted metabolomics	(Yon et al, 2021)
Full Factorial design	LC/MS	untargeted metabolomics	(Pezzatti et al, 2019)
Plackett-Burman	NMR	untargeted metabolomic analysis	(Sokolenko et al, 2014)
	LC/MS	untargeted metabolomic analysis	(Macioszek et al, 2021)
	LC/MS	untargeted metabolomic analysis	(Zhao et al, 2017)
D-optimal design	GC-TOF-MS	untargeted metabolomics	(Gullberg et al, 2004)
	UPLC–MS	untargeted metabolomics	(Kellogg et al, 2017)
Optimization			
Central Composite design	HPLC	targeted metabolomics	(Guo et al, 2019)
	GC-MS	untargeted metabolomics	(Silva et al, 2019)
Box-Bekhen	HPTLC	targeted metabolomic	(Alam et al, 2021)
	UHPLC–QTOF-ESI-MS	targeted metabolomics	(Zhou et al, 2009)
	LC-MS	Pharmacometabolomics	(He et al, 2022)
D-optimal design	GC/MS	targeted metabolomics	(Danielsson et al, 2012)
	HS-SPME-GC-TOF	untargeted metabolomic analysis	(Fedrizzi et al, 2012)
	LC-MS/MS	targeted metabolomics	(Njumbe Ediage et al, 2012)

2.3.1 Screening

To emphasize DoE-based screening in metabolomics, we have collected various studies (see table 4). Each study in table 4 exemplifies how the sparsity of effect principle may be applied to reduce the design space in various screening settings, emphasizing acquired effect resolution vs needed effect resolution, minimizing experimental sample generation.

A case of applying frac-FFD for identifying significant main effects is found in the works of Zanatta (Zanatta et al, 2022), here seven parameters affecting the chromatographic separation process were screened to identify the most influential factors. A IV resolution 2^{7-2} design was utilized, reducing the number of samples from 128 for the FFD to 32 per replicate whilst ensuring main effects are not aliased with higher order terms. The results concluded that flow rate, run time, the final concentration of ethanol, and column choice were significant factors.

A different example investigating the significance of main effects is found in a study investigating LC/MS optimization for untargeted metabolomics (increasing number of total peaks and reliable peaks) (Tebani et al, 2016). Here, the optimization was split into two different experiments, with the screening part investigating the influence of 7 UHPLC-MS parameters applying a IV resolution 2^{7-3} design, including center runs to test for potential curvature. The design resulted in the number of samples being reduced from 128 per replicate to 16 + 4 center runs per replicate. The addition of center runs was successful in concluding that no curvature was present, whilst it was found that desolvation temperature, desolvation gas flow, extraction voltage, and sample cone voltage had a significant effect on the total number of peaks. In addition, only desolvation gas flow had a negative influence on the number of reliable peaks.

Successful employment of the PBD for screening is reported in the analysis of complex Mixtures, such as cell culture analysis though 1D ^{1}H NMR (Sokolenko et al, 2014) and in sample preparation of gastrointestinal stromal tumor samples utilized in untargeted metabolomic analysis via LC/MS (Macioszek et al, 2021). In the first case, a mixture of 28 compounds was made to mimic a cell culture metabolomic environment (Sokolenko et al, 2014). 16 of the compounds were set at two levels. This resulted in treatment intervention estimations being based on a total of 16 factors, while the remaining 16 compounds served as background. The PBD was an effective choice of investigation, as k=16 resulted in a sample space of N=20 whilst a FFD would require $2^{16} = 65536$ for a full resolution. For the latter case, a PBD was generated based on 10 factors resulting in the 12 samples including 3 center runs. Interestingly, the analysis of the experimental data was based on a Bayesian approach applying hierarchical models instead of the traditional ANOVA approach.

An example of utilizing a FFD for screening during sample preparation in LC/MS is showcased in the study of Caulobacter crescentus (Pezzatti et al, 2019). Here three factors were investigated being cell retrieval, quenching and extraction solvents, and cell disrupting mechanisms, resulting in a 2^3 FFD. In conclusion, it was identified that all main effects had a significant contribution to the overall variance whilst no interactions were of significance. The study is an excellent showcase that FFD can be a valid design choice when the design space is sufficiently small.

At last, we are highlighting an example of a screening design based on the D-optimal design utilized in the study of accelerated solvent extraction of catechins in tea (Kellogg et al, 2017) to be analyzed via untargeted UPLC-MS. The study involved 3 factors (extraction temperature, cycle time, and solvent ratio), which each had three levels. The D-optimal design was applied resulting in a 15 experiment run with 3 center points. The results showed that only solvent ratio had a significant impact on the measured concentration levels of catechins, whilst the reaming factors emitted no significant impact.

2.3.2 Optimization

For optimization of various processes within metabolomics, the CCD and BBD dominate throughout most of the literature. One example of optimization comes in the form of the BBD being applied to optimize the extraction of parthenolide from stems of tarconanthus camphoratus (Alam et al, 2021). Here the BBD was employed to maximize parthenolide extraction, based on 3 factors (extraction temperature, extraction time, and microwave power) of 3 levels. The resulting design had 17 runs with a total of 5 center points. The analysis of the design space generated by BBD resulted in a model capable of predicting parthenolide and from the model an optimum for the three factors was identified.

Another example of the BBD is found in a study setting up a model system for drug monitoring in plasma (He et al, 2022) using LC/MS. The BBD was specifically utilized to optimize a three-phase electroextraction step, optimizing pH, extraction time, and voltage. The results of the BBD identified a model where the main effects of extraction time, pH and voltage, and quadratic effect of extraction time were all of significantly contributing to the prediction of optimal enrichment factor.

A different approach to optimization via the CCD was successfully utilized to identify potential volatile biomarkers of breast cancer by applying GC-MS to urine samples (Silva et al, 2019). The CCD was constructed such that the process of solid-phase microextraction for isolating volatile metabolomics could be optimized. The experiment was set up such that the extraction capability of 3 different fiber coatings was investigated with their own CCD-based model. In total 3 factors were investigated for each fiber coating being salt amount, extraction time, and extraction temperature. Each factor had 5 different levels being center level(middle), corner points (high, low), and star points (extreme high, extreme low) with the star point axial distance set to $\alpha = 2k/4$ resulting in 45 sample points. From the study design, extraction temperature, and extraction time was found to be of significant influence, whilst salt amount had no effect in the given range. Additionally, no quadratic effects nor interactions were found to be of significance.

Apart from the classical methods of CCD and BBD, cases of none symmetric designs have also been observed throughout the metabolomic literature (see table 4). One such example is found in the works of Njumbe (Njumbe Ediage et al, 2012). Here an assessment of mycotoxins in urine samples was investigated via LC/MS, in which optimization of analyte recovery and method sensitivity was investigated via a D-optimal design. In total five factors (one qualitative factor and four quantitative factors) with three levels each was considered for optimizing liquid–liquid extraction, being extraction solvent, acid levels, the volume of extraction solvent, and evaporation temperature. The D-optimal design resulted in a 37-sample experiment for investigation. The results showed that the percentage of acid, volume of extraction solvent, and extraction time, whilst evaporation temperature and higher order terms were of no significance. For the qualitative factor, ethyl acetate showed the highest yield compare to the remaining solvents. The study is an excellent demonstration of how to incorporate qualitative factors within the design itself.

3 Statistical quality control

Many different SQC methods can be used in metabolomics, including control charts (Lörchner et al, 2022), batch calibrations (Broadhurst et al, 2018), and normalization techniques (Biswapriya B, 2020). These methods can help to ensure data is reliable and accurate, which is essential for identifying biomarkers, understanding metabolic pathways, and developing diagnostic tools for a range of diseases. In short, SQC plays an important role in ensuring the quality and reliability of metabolomics data, which is essential for advancing our understanding of metabolomic systems.

3.1 Calibration

Calibration employing SQC is an essential step in metabolomics research to ensure accurate and reliable measurement of analytes. One common approach for calibration is the use of pooled QC samples (Broadhurst et al, 2018). Pooled QC samples are prepared by combining small aliquots of each sample throughout the entire experiment and are analyzed along with the individual samples to monitor the quality of the data. The pooled QC samples are used to assess the precision, accuracy, and reproducibility of the instrument, and to detect any drift or bias in the data over time (van der Kloet et al, 2009).

Another important calibration technique is batch-to-batch calibration (Kuligowski et al, 2015; van der Kloet et al, 2009). This involves analyzing reference standards or quality control (QC) samples at the beginning and end of each batch or run to ensure consistency and accuracy across different batches. By comparing the results obtained from the reference standards or QC samples, any drift or variation in the instrument performance can be detected and corrected (van der Kloet et al, 2009).

Lowering nuisance factors is another concept to improve calibration, minimizing the impact of variation not related to the analytes being measured. This includes using internal standards (Trygg et al, 2006), optimizing sample preparation and extraction protocols, and minimizing the impact of matrix effects (Broadhurst et al, 2018).

Finally to correct intra-batch variations data, mathematical techniques such as regression analysis (van der Kloet et al, 2009), support vector regression (Kuligowski et al, 2015), and spline regression (Dunn et al, 2011) can be utilized. The aforementioned correction methods may be applied to samples and QC samples by performing linear or more complicated non-linear corrections. The goal of the correction is to achieve a higher level of homoscedasticity, which results in a lowering variability. The methods may be evaluated either visually via PCA plots, or by computing the residual standard deviation (RSD). It should be noted that to tune the hyperparameters of the non-linear correction methods, cross-validation Kuligowski et al (2015) should be employed to improve model robustness. Within the next subsection, examples from the literature on how to apply the aforementioned SQC techniques in metabolomics are showcased.

3.2 Application of statistical quality control

A diverse range of studies utilizing the methods described in the preceding paragraph and showcasing the practical application of SQC has been gathered. For a complete short-form summary of every study, we refer to table 5.

An example showcasing the usage of pooled QC sampling is found in a study investigating the urinary metabolome of children using NMR (Maitre et al, 2017). Pooled QC was done by pooling samples from all individuals for the urine, summing a total of 24 QC samples made at regular intervals during the experiment. To analyze the coefficients of variation (CV) for each detected metabolite, the pooled QC was utilized. The purpose of applying CVs with QC samples was to investigate the reliability of the measurements (eg. technical replicates). The mean and median CV showed a total variation of 7.2 % and 7.7 % respectively which was further refined by splitting QC samples for nighttime and daytime samples.

Table 5 Overview of SQC-based metabolomics studies.

SQC Technique	Analytical technique	Type of study	Reference
pooled QC-sampling	LC/MS	untargeted metabolomic	(Godzien et al, 2014)
	NMR	untargeted metabolomic	(Maitre et al, 2017)
	NMR and LC/MS	untargeted metabolomic	(Lau et al, 2018)
	LC/MS	untargeted metabolomic	(Liu et al, 2022)
	UPLC–Q-TOF-MS	untargeted metabolomic	(Liu et al, 2013)
Internal standards	NMR	untargeted metabolomic	(Grasso et al, 2022)
	LC/MS	targeted metabolomic	(Ulvik et al, 2021)
	LC/ESI-MS/MS	targeted metabolomic	(Stokvis et al, 2005)
Normalization	CE-TOF-MS	targeted metabolomic	(Cuevas-Delgado et al, 2020)
	NMR	untargeted metabolomic	(Dieterle et al, 2006)
	LC/MS	untargeted metabolomic	(Lee et al, 2012)
	LC/MS	untargeted metabolomic	(Veselkov et al, 2011)
Batch to batch calibration	NMR	targeted metabolomic	(Fages et al, 2013)
	LC/MS	targeted metabolomic	(Yue et al, 2022)

A second example of utilizing pooled QC-sampling was showcased in another study analyzing the urine metabolome by applying UPLC–Q-TOF-MS (Liu et al, 2013). Here, PCA was utilized to analyze measurement repeatability. The clustering of QC samples showed evidence that the QC samples were adequately clustered indicating accurate repeatability for the measurements.

For internal standards, a study evaluating the effects of different types of standards in quantitative bioanalytical LC-MS via relative standard deviations (Stokvis et al, 2005) is highlighted. Here it was concluded based on relative standard deviations that cheaper structural analogs may in some cases be an alternative to expensive SIL internal standards.

Another interesting study testing a range of internal standards for intact serum samples is found in the works of Grasso (Grasso et al, 2022). The purpose was to identify internal standards which did not exhibit binding properties towards proteins (as DDS and TSP are known for). The suggested standards were evaluated by applying a T2 filter with different delays. Here it was shown that formate showed very little intensity loss as relaxation time delays increased, compared to the remaining internal standards.

Apart from internal standards and QC samples, normalization is another very important technique utilized in calibration. One example in which normalization was applied with the aim of reducing the influence of nuisance factors is found in the study of Chronic kidney disease applying CE-TOF-MS for untargeted metabolomics (Cuevas-Delgado et al, 2020). The study involved six different normalization methods applying Probabilistic quotient normalization (PQN) (Dieterle et al, 2006), internal standard (IS) normalization (Sysi-Aho et al, 2007), Median fold change (MFC) (Veselkov et al, 2011), total protein content normalization (Silva et al, 2011), total useful signal (Warrack et al, 2009) and quantile normalization (QN) (Lee et al, 2012). Each method was tested using residual standard deviation (RSD), within-group relative log abundance (RLA) plots (Livera et al, 2012) and Partial least squares discriminant analysis (PLS-DA) (Brereton and Lloyd, 2014; Trygg and Wold, 2002) revealed that the choice of method did change the overall variation of data. Specifically, only MFC and QN exhibited similar results whilst the remaining normalization methods exhibited responses not matching any other responses.

Sample matrix effects are as mentioned also a challenge to be solved as these highly contribute to masking results. Apart from applying internal standards as proposed in the works of Grasso and Stovik (Grasso et al, 2022; Stokvis et al, 2005). Examples of filtering during sample preparation or during data acquisition are seen throughout the literature. One example of filtration during sampling preparation is found in the study of Human Plasma applying LC-MS (Sitnikov et al, 2016). Within the study, seven methods of solid-phase extraction and solvent-based were evaluated using standard analytes spiked into Human plasma and buffer. The methods were evaluated by estimating the RSD for pooled QC samples containing all targeted analytes.

The final technique to be covered is batch-to-batch calibrations. Here Fages (Fages et al,

2013) reported an interesting study on this topic. The study applied grouped-batch profile (GBP) calibration for NMR-based metabolomics data used within a larger epidemiological study. The results showed that by applying GBP to data, the statistical predictive power, modelling cell infection status increased significantly. Specifically, the effects of applying grouped-batch profile calibration were observed by investigating the second and first principal components of QC samples taken from each batch. Here it was found that the GBP enabled more tightly clustered QC samples (one QC sample per batch), indicating a lowering of batch-to-batch variation.

A second example of batch calibration is found in the works of Yue (Yue et al, 2022), investigating intra-batch variations in LC-MS-MS. The purpose of the study was to correct and quantify intra-batch variations for targeted metabolomics by investigating three methods being reversed-phase liquid chromatography (RP-LC), ion-pair liquid chromatography (IP-LC), and hydrophilic interaction liquid chromatography (HILIC). To evaluate the intra-batch variation of each targeted metabolite for each method, an assessment of the mean relative standard deviation (RSD) of peak areas for QC samples was analyzed for continuous runs. In addition, QC samples were investigated via PCA score plots, assessing the clustering density of each method. The correction was done by modelling signal drift of the QC samples in which a correction factor was chosen based upon an exponential model. The results of the correction were assessed using RSD and PCA score plots, in which fewer intra-batch variations were observed.

4 Recommendations

To summarize there are mainly two types of scenarios encountered within the DoE part of metabolomics, that is screening and optimization for which different techniques excels in achieving maximum information from a limited sample space. For the SQC part, the main goal is to minimize the impact of variation sources which may mask the effect of specific treatment interventions set up via the experimental design. We have set up two tables highlighting every DoE (see table 6) and SQC techniques (see table 7) covered in this review that is currently utilized within metabolomics and marked potential new techniques which may be applicable in future studies.

Table 6 Summary of DoE techniques for screening and optimization within metabolomics. Potential applicable DoE techniques not current applied in metabolomics are marked with *.

Technique	Screening	Optimization
Full Factorial design	yes	yes
Fractional factorial design	yes	no
D-optimal design	yes	yes
Box-Bekhen design	no	yes
Central composite design	no	yes
Taguchi designs*	yes	yes
Plackett-Burman design	yes	no
Augmentation design*	yes	yes

Table 7 Summary of SQC techniques for calibration.

Assessment subject	**Quantify effects**
Batch to batch variations	PCA score plot, RSD, QC-samples
Pooled QC samples	PCA score plots, RSD
Internal standards quality	PCA score plots, RSD
normalization	PCA score plots, RSD, RLA plots
Problem	**correction method**
Batch to batch variations	regression, spline regression, GBP, normalization
Nuisance effects	Internal standards optimization, normalization

In addition to the summary tables (table 6 and 7) we have generated two figures highlighting a recommended workflow of setting up and ensuring data quality of metabolomics experiments. Figure 4 visualizes the DoE workflow and Figure 5 the SQC workflow.

The recommendations within DoE (see figure 4) are based on the concepts of acquired resolution vs needed resolution, whilst also distinguishing between screening and optimization tasks. The setups are usable for both NMR-based and various MS techniques. It does not distinguish between targeted or untargeted metabolomics analysis but relies purely on generating an optimal design space based on the analysis treatment intervention (e.g. the impact of different factors levels - see table 4 for examples). The application of DoE can further enrich the quantification of treatment interventions, carried out via analysis

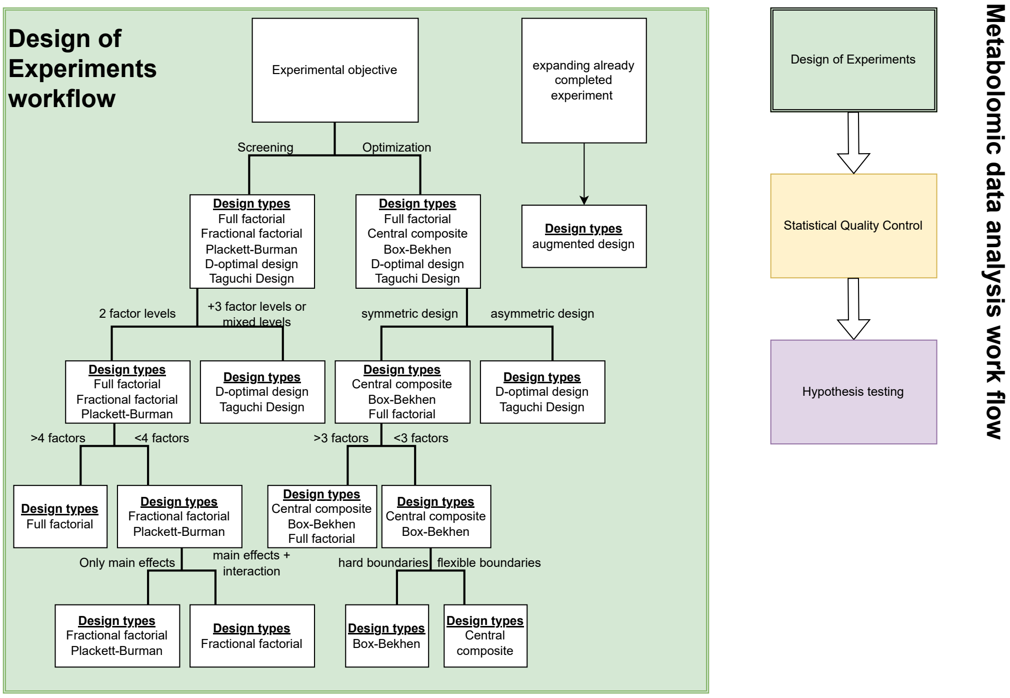


Fig. 4 Visualization of recommended workflow for applying Design of Experiments for screening and optimization tasks in metabolomics.

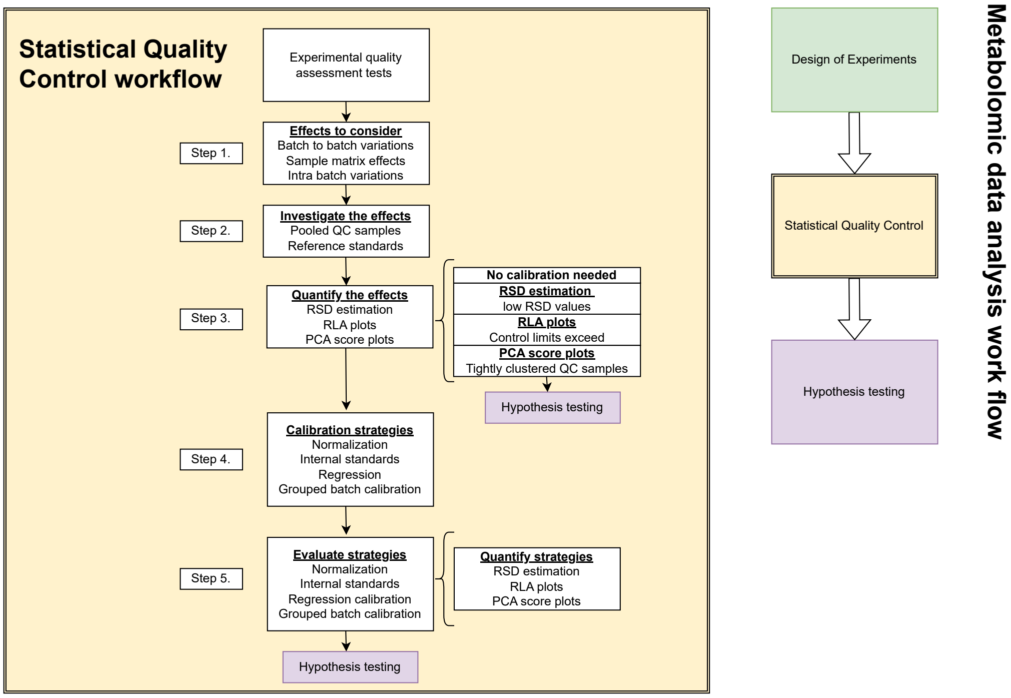


Fig. 5 Visualization of recommended workflow for applying Statistical Quality control quantification and calibration of experimental accuracy in metabolomics.

of variance (ANOVA) methods. The enrichment comes from the fact that a DoE framework will ensure a controlled sum of squares allocation such that significant treatments are not occurring due to aliasing of effects. Hence DoE reduces the chance of false positives (type 1 error) whilst ensuring sample space is optimal, saving time and effort in the process.

The recommendations set for SQC (see figure 5) are assembled such that common challenges encountered within metabolomics may be quantified and potentially solved via calibration techniques. The most important concept of this review is to understand how to quantify experimental accuracy and as such, the recommendation reflects methods and metrics to make the accuracy of experiments quantifiable. The quantification of accuracy makes it possible to discover potential nuisance factors affecting the analysis of data and the methods of calibration make for potential solutions, mitigating the nuisance effects (see table 7 for cases).

The DoE framework presents optimal sampling space, but one challenge that remains to be solved within metabolomics and DoE in general is the settings of design space boundaries (e.g. factor levels such as temperature and pH). The boundaries or factor ranges are currently set through prior knowledge for specific systems and the DoE itself does not guarantee successful results without proper domain knowledge. One possible solution in terms of metabolomics would be to work in tandem with genomics. The connection may be beneficial for setting experimental boundaries, as apriori genetic profiling may indicate the presence/absence of specific gene clusters responsible for expressing a metabolite of interest. The genetic information could be a key as the genes may only be expressed within specific ranges of abiotic factors, thus limiting the range of treatment interventions during screening.

5 Conclusion

In conclusion, this paper has provided an overview of Design of Experiments (DoE) and Statistical Quality Control (SQC) techniques applicable in metabolomics experiments. We have summarized the different DoE and SQC techniques currently used in metabolomics and highlighted potential new techniques that could be applicable in future studies. Furthermore, recommendations are provided for setting up metabolomics experiments using DoE and SQC techniques that will ensure optimal sampling space and experimental accuracy.

The highlighted DoE techniques are useful for both screening and optimization tasks and can be applied to NMR and various MS techniques, for both targeted and untargeted metabolomics analyses. By applying DoE, the experimental design space can be optimized to ensure maximum information is obtained from a limited sample space while minimizing the chances of false positives due to the controlled sum of squares allocation.

Finally, the use of SQC techniques in metabolomics experiments is crucial for obtaining accurate and reliable data, whilst also minimizing significance occurring due to technical variation. The recommendations provided in this review can serve as a guide for researchers in the field not too familiar with DoE and SQC. The guide enables experimental designs that are more efficient through the DoE recommendations. In addition, the SQC recommendations provide tools for quantifying and potentially correcting the effects of non-treatment-related variations, lowering the chance of detecting false positives due to nuisance variation.

Conflict of interest. All authors declare that they have no conflicts of interest.

Acknowledgments. We would like to thank Danish National Research Foundation for the Center for Microbial Secondary Metabolites Grant Number: DNRF137 for funding this project

References

Alam P, Siddiqui NA, Rehman MT, Hussain A, Akhtar A, Mir SR, Alajmi MF (2021) Box–Behnken Design (BBD)-Based Optimization of Microwave-Assisted Extraction of Parthenolide from the Stems of Tarconanthus camphoratus and Cytotoxic Analysis. *Molecules* 26(7):1876. https://doi.org/10.3390/molecules26071876

von Bargen C, Brockmeyer J, Humpf HU (2014) Meat Authentication: A New HPLC–MS/MS Based Method for the Fast and Sensitive Detection of Horse and Pork in Highly Processed Food. *Journal of Agricultural and Food Chemistry* 62(39):9428–9435. https://doi.org/10.1021/jf503468t

Bertolaccini L, Viti A, Terzi A (2015) The Statistical point of view of Quality: the Lean Six

Sigma methodology. *Journal of Thoracic Disease* 7(4):E66–E68. https://doi.org/10.3978/j.issn.2072-1439.2015.04.11

Bhattacharya S (2021) *Central Composite Design for Response Surface Methodology and Its Application in Pharmacy*. In: Kayaroganam P (ed) Response Surface Methodology in Engineering Science. IntechOpen, Rijeka, chap 5, https://doi.org/10.5772/intechopen.95835

Biswapriya B M (2020) Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *European Journal of Mass Spectrometry* 26(3):165–174. https://doi.org/10.1177/1469066720918446

Box GEP, Wilson KB (1951) On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society Series B (Methodological)* 13(1):1–45. https://doi.org/10.1111/j.2517-6161.1951.tb00067.x

Breier M, Wahl S, Prehn C, Fugmann M, Ferrari U, Weise M, Banning F, Seissler J, Grallert H, Adamski J, Lechner A (2014) Targeted Metabolomics Identifies Reliable and Stable Metabolites in Human Serum and Plasma Samples. *PLoS One* 9(2):e89728. https://doi.org/10.1371/journal.pone.0089728

Brereton R, Lloyd G (2014) Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics* 28(4):213–225. https://doi.org/10.1002/cem.2609

Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, Dunn WB (2018) Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* 14(6):72. https://doi.org/10.1007/s11306-018-1367-3

Castro-Alves V, Kalbina I, Nilsen A, Aronsson M, Rosenqvist E, Jansen MA, Qian M, Åsa Öström, Hyötyläinen T, Åke Strid (2021) Integration of non-target metabolomics and sensory analysis unravels vegetable plant metabolite signatures associated with sensory quality: A case study using dill (anethum graveolens). *Food Chemistry* 344:128714. https://doi.org/10.1016/j.foodchem.2020.128714

Cuevas-Delgado P, Dudzik D, Miguel V, Lamas S, Barbas C (2020) Data-dependent normalization strategies for untargeted metabolomics—a case study. *Analytical and Bioanalytical Chemistry* 412(24):6391–6405. https://doi.org/10.1007/S00216-020-02594-9

Danielsson AP, Moritz T, Mulder H, Spégel P (2012) Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics* 8(1):50–63. https://doi.org/10.1007/S11306-011-0283-6

Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1 H NMR Metabonomics. *Analytical chemistry* 78(13):4281–4290. https://doi.org/10.1021/ac051632c

Draper N (2008) *Rotatable Designs and Rotatability*. In: Ruggeri F, Kenett R, Faltin F (eds) Encyclopedia of Statistics in Quality and Reliability. John Wiley & Sons, Ltd, p 1–1800, https://doi.org/10.1002/9780470061572.eqr034

Dunn W, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles J, Halsall A, Haselden J, Nicholls A, Wilson I, Kell D, Goodacre R, Consortium T (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols* 6(7):1060–1083. https://doi.org/10.1038/nprot.2011.335

Fages A, Pontoizeau C, Jobard E, Lévy P, Bartosch B, Elena-Herrmann B (2013) Batch profiling calibration for robust NMR metabonomic data analysis. *Analytical and Bioanalytical Chemistry* 405(27):8819–8827. https://doi.org/10.1007/S00216-013-7296-0

Federer WT, Raghavarao D (1975) On augmented designs. *Biometrics* 31(1):29–35. https://doi.org/https://doi.org/10.2307/2529707

Federer WT, Reynolds M, Crossa J (2001) Combining results from augmented designs over sites. *Agronomy Journal* 93(2):389–395. https://doi.org/10.2134/agronj2001.932389x

Fedrizzi B, Carlin S, Franceschi P, Vrhovsek U, Wehrens R, Viola R, Mattivi F (2012) D-optimal design of an untargeted hs-spme-gc-tof metabolite profiling method. *The Analyst* 137(16):3725–31. https://doi.org/10.1039/C2AN16309H

Geng C, Cui C, Guo Y, Wang C, Zhang J, Han W, Jin F, Chen D, Jiang P (2020) Metabolomic Profiling Revealed Potential Biomarkers in Patients With Moyamoya Disease. *Frontiers in Neuroscience* 14:308. https://doi.org/10.3389/fnins.2020.00308

Girden ER (1992) *ANOVA: Repeated measures*. In: Lewis-Beck MS (ed) Quantitative Applications in the Social Sciences. 84, Sage Publications, p 88, https://doi.org/10.4135/9781412983419

Godzien J, Alonso Herranz V, Barbas C, Armitage E (2014) Controlling the quality of metabolomics data: new strategies to get the best out of the QC sample. *Metabolomics* 11:518–528. https://doi.org/10.1007/s11306-014-0712-4

Grasso D, Pillozzi S, Tazza I, Bertelli M, Campanacci DA, Palchetti I, Bernini A (2022) An improved nmr approach for metabolomics of intact serum samples. *Analytical Biochemistry* 654:114826. https://doi.org/10.1016/j.ab.2022.114826

Gullberg J, Jonsson P, Nordström A, Sjöström M, Moritz T (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Analytical Biochemistry* 331(2):283–295. https://doi.org/https://doi.org/10.1016/j.ab.2004.04.037

Guo Z, Zhu Y, Xu W, Luo K, Xiao H, Wang Z (2019) Alteration of amino acid profiling influenced by the active ingredients of danhong injection after prescription optimization. *Drug Design, Development and Therapy* 13:3939–3947. https://doi.org/10.2147/DDDT.S220314

He Y, Drouin N, Wouters B, Miggiels P, Hankemeier T, Lindenburg PW (2022) Development of a fast, online three-phase electroextraction hyphenated to fast liquid chromatography–mass spectrometry for analysis of trace-level acid pharmaceuticals in plasma. *Analytica Chimica Acta* 1192:339364. https://doi.org/10.1016/j.aca.2021.339364

Hernandez LN, Nachtsheim CJ (2018) Fast computation of exact g-optimal designs via i λ-optimality. *Technometrics* 60(3):297–305. https://doi.org/10.1080/00401706.2017.1371080

Hill J, Thomas AJ, Mason-Jones RK, El-Kateb S (2017) The implementation of a Lean Six Sigma framework to enhance operational performance in an MRO facility. *Production & Manufacturing Research* 6(1):26–48. https://doi.org/10.1080/21693277.2017.1417179

Inal TC, Goruroglu Ozturk O, Kibar F, Cetiner S, Matyar S, Daglioglu G, Yaman A (2018) Lean six sigma methodologies improve clinical laboratory efficiency and reduce turnaround times. *Journal of Clinical Laboratory Analysis* 32(1):e22180. https://doi.org/10.1002/jcla.22180

Institute S (1985) SAS user's guide: Statistics, vol 2. Sas Inst

Kacker RN, Lagergren ES, Filliben JJ (1991) Taguchi's Orthogonal Arrays Are Classical Designs of Experiments. *Journal of Research of the National Institute of Standards and Technology* 96(5):577. https://doi.org/10.6028/JRES.096.034

Kellogg JJ, Wallace ED, Graf TN, Oberlies NH, Cech NB (2017) Conventional and accelerated-solvent extractions of green tea (camellia sinensis) for metabolomics-based chemometrics. *Journal of Pharmaceutical and Biomedical Analysis* 145:604–610. https://doi.org/10.1016/j.jpba.2017.07.027

van der Kloet FM, Bobeldijk I, Verheij ER, Jellema RH (2009) Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *Journal of proteome research* 8(11):5132—5141. https://doi.org/10.1021/pr900499r

Kuligowski J, Sánchez-Illana A, Sanjuán-Herráez D, Vento M, Quintás G (2015) Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (qc-svrc). *Analyst* 140(22):7810–7817. https://doi.org/10.1039/C5AN01638J

Lau CH, Siskos A, Maitre L, Robinson O, Athersuch T, Want E, Urquiza J, Casas M, Vafeiadi M, Roumeliotaki T, McEachan R, Azad R, Haug L, Meltzer H, Andrusaityte S, Petraviciene I, Grazuleviciene R, Thomsen C, Wright J, Coen M (2018) Determinants of the urinary and serum metabolome in children from six european populations. *BMC Medicine* 16(1):202. https://doi.org/10.1186/s12916-018-1190-8

Lee J, Park J, Lim Ms, Seong SJ, Seo JJ, Park SM, Lee HW, Yoon YR (2012) Quantile normalization approach for liquid chromatography–mass spectrometry-based metabolomic data from healthy human volunteers. *Analytical Sciences* 28(8):801–805. https://doi.org/10.2116/analsci.28.801

Liu L, He Y, Lu H, Wang M, Sun C, Na L, Li Y (2013) Metabonomic analysis of urine from rats after low-dose exposure to 3-chloro-1,2-propanediol using uplc–ms. *Journal of chromatography B, Analytical technologies in the biomedical and life sciences* 927:97–104. https://doi.org/10.1016/j.jchromb.2013.01.038

Liu X, Tian X, Qinghong S, Sun H, Jing L, Tang X, Guo Z, Liu Y, Wang Y, Ma J, Na R, He C, Song W, Sun W (2022) Characterization of LC-MS based urine metabolomics in healthy children and adults. *PeerJ* 10:e13545. https://doi.org/10.7717/peerj.13545

Livera AMD, Dias DA, Souza DD, Rupasinghe T, Pyke J, Tull D, Roessner U, McConville M, Speed TP (2012) Normalizing and integrating metabolomics data. *Analytical Chemistry* 84(24):10768–10776. https://doi.org/10.1021/ac302748b

Lu L, Anderson-Cook CM (2021) Strategies for sequential design of experiments and augmentation. *Quality and Reliability Engineering International* 37(5):1740–1757. https://doi.org/10.1002/qre.2823

Lörchner C, Horn M, Berger F, Fauhl-Hassek C, Glomb MA, Esslinger S (2022) Quality control of spectroscopic data in non-targeted analysis – development of a multivariate control chart. *Food Control* 133:108601. https://doi.org/10.1016/j.foodcont.2021.108601

Macioszek S, Dudzik D, Jacyna J, Wozniak A, Schöffski P, Markuszewski MJ (2021) A robust method for sample preparation of gastrointestinal stromal tumour for LC/MS untargeted metabolomics. *Metabolites* 11(8):1740–1757. https://doi.org/10.1002/qre.2823

Madsen H, Thyregod P (2011) Introduction to General and Generalized Linear Models, CRC Press, pp 1–316. https://doi.org/10.1201/9781439891148

Maitre L, Lau CH, Vizcaino E, Robinson O, Casas M, Siskos A, Want E, Athersuch T, Slama R, Vrijheid M, Keun H, Coen M (2017) Assessment of metabolic phenotypic variability in children's urine using 1h nmr spectroscopy. *Scientific Reports* 7(1):46082. https://doi.org/10.1038/srep46082

Miller AJ, Nguyen NK (1994) Algorithm as 295: A fedorov exchange algorithm for d-optimal design. *Journal of the royal statistical society series c (applied statistics)* 43(4):669–677. https://doi.org/10.2307/2986264

Montañez A, Blanco AR, Barlocco C, Beracochea M, Sicardi M (2012) Characterization of cultivable putative endophytic plant growth promoting bacteria associated with maize cultivars (zea mays l.) and their inoculation effects in vitro. *Applied Soil Ecology* 58:21–28. https://doi.org/10.1016/j.apsoil.2012.02.009

Montgomery DC (2008) Design and Analysis of Experiments, 8th edn., John Wiley & Sons, pp 1–725

Montgomery DC (2009) Statistical Quality Control A Modern Introduction, 6th edn., John Wiley & Sons, pp 1–704

Njumbe Ediage E, Diana Di Mavungu J, Song S, Wu A, Van Peteghem C, De Saeger S (2012) A direct assessment of mycotoxin biomarkers in human urine samples by liquid chromatography tandem mass spectrometry. *Analytica Chimica Acta* 741(5):58–69. https://doi.org/10.1016/j.aca.2012.06.038

Pezzatti J, Bergé M, Boccard J, Codesido S, Gagnebin Y, Viollier PH, González-Ruiz V, Rudaz S (2019) Choosing an optimal sample preparation in caulobacter crescentus for untargeted metabolomics approaches. *Metabolites* 9(10):193. https://doi.org/10.3390/METABO9100193

Pinu FR (2018) Grape and wine metabolomics to develop new insights using untargeted and targeted approaches. *Fermentation* 4(4):92. https://doi.org/10.3390/fermentation4040092

Poulin RX, Pohnert G (2019) Simplifying the complex: metabolomics approaches in chemical ecology. *Analytical and bioanalytical chemistry* 411(1):13–19. https://doi.org/10.1007/s00216-018-1470-3

Quinn GP, Keough MJ (2002) Experimental Design and Data Analysis for Biologists, 2nd edn., Cambridge University Press, pp 1–553. https://doi.org/10.1017/CBO9780511806384

Radson D, Herrin GD (1995) Augmenting a factorial experiment when one factor is an uncontrollable random variable: A case study. *Technometrics* 37(1):70–81. https://doi.org/10.1080/00401706.1995.10485891

RStudio Team (2020) RStudio: Integrated Development Environment for R. RStudio, PBC., Boston, MA, URL http://www.rstudio.com/

Shahmohammadi A, McAuley KB (2019) Sequential model-based a- and v-optimal design of experiments for building fundamental models of pharmaceutical production processes. *Computers & Chemical Engineering* 129(4):106504. https://doi.org/10.1016/j.compchemeng.2019.06.029

Silva A, Cordeiro-da Silva A, Coombs G (2011) Metabolic variation during development in culture of leishmania donovani promastigotes. *PLos neglected tropical diseases* 5(12):e1451. https://doi.org/10.1371/journal.pntd.0001451

Silva CL, Perestrelo R, Silva P, Tomás H, Câmara JS (2019) Implementing a central composite design for the optimization of solid phase microextraction to establish the urinary volatomic expression: a first approach for breast cancer. *Metabolomics* 15(4):66. https://doi.org/10.1007/S11306-019-1525-2

Sitnikov DG, Monnin CS, Vuckovic D (2016) Systematic assessment of seven solvent and solid-phase extraction methods for metabolomics analysis of human plasma by lc-ms. *Scientific Reports 2016 6:1* 6:1–11. https://doi.org/10.1038/srep38885

Sokolenko S, Blondeel EJM, Azlah N, George B, Schulze S, Chang D, Aucoin MG (2014) Profiling convoluted single-dimension proton nmr spectra: A plackett–burman approach for assessing quantification error of metabolites in complex mixtures with application to cell culture. *Analytical Chemistry* 86(7):3330–3337. https://doi.org/10.1021/ac4033966

Song JW, Lam SM, Fan X, Cao WJ, Wang SY, Tian H, Chua GH, Zhang C, Meng FP, Xu Z, et al (2020) Omics-driven systems interrogation of metabolic dysregulation in covid-19 pathogenesis. *Cell metabolism* 32(2):188–202. https://doi.org/10.1016/j.cmet.2020.06.016

Stokvis E, Rosing H, Beijnen JH (2005) Stable isotopically labeled internal standards in quantitative bioanalysis using liquid chromatography/mass spectrometry: necessity or not? *Rapid Communications in Mass Spectrometry* 19(3):401–407. https://doi.org/10.1002/rcm.1790

Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* 8(1):93. https://doi.org/10.1186/1471-2105-8-93

Szalowska E, Van Hijum SA, Roelofsen H, Hoek A, Vonk RJ, Te Meerman GJ (2007) Fractional Factorial Design for Optimization of the SELDI Protocol for Human Adipose Tissue Culture Media. *Biotechnology Progress* 23(1):217–224. https://doi.org/10.1021/BP0602294

Tebani A, Schmitz-Afonso I, Rutledge DN, Gonzalez BJ, Bekri S, Afonso C (2016) Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Analytica Chimica Acta* 913:55–62. https://doi.org/10.1016/j.aca.2016.02.011

Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* 16(3):119–128. https://doi.org/10.1002/cem.695

Trygg J, Gullberg J, Johansson A, Jonsson P, Moritz T (2006) *Chemometrics in Metabolomics - An Introduction*. In: Saito K, Dixon R, Willmitzer L (eds) Plant Metabolomics. Biotechnology in Agriculture and Forestry, vol 57. Springer, Berlin, Heidelberg, p 117–128, https://doi.org/10.1007/3-540-29782-0_9

Ulvik A, McCann A, Midttun O, Meyer K, Godfrey K, Ueland P (2021) Quantifying precision loss in targeted metabolomics based on mass spectrometry and nonmatching internal standards. *Analytical Sciences* 93(21):7616–7624. https://doi.org/10.1021/acs.analchem.1c00119

Vanaja K, Rani RS (2007) Design of experiments: Concept and applications of plackett burman design. *Clinical Research and Regulatory Affairs* 24(1):1–23. https://doi.org/10.1080/10601330701220520

Veselkov K, Vingara L, Masson P, Robinette S, Want E, Li J, Barton R, boursier neyret C, Walther B, Ebbels T, Pelczer I, Holmes E, Lindon J, Nicholson J (2011) Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical chemistry* 83(15):5864–5872. https://doi.org/10.1021/ac201065j

Vetting MW, Al-Obaidi N, Zhao S, San Francisco B, Kim J, Wichelecki DJ, Bouvier JT, Solbiati JO, Vu H, Zhang X, Rodionov DA, Love JD, Hillerich BS, Seidel RD, Quinn RJ, Osterman AL, Cronan JE, Jacobson MP, Gerlt JA, Almo SC (2015) Experimental strategies for functional annotation and metabolism discovery: Targeted screening of solute binding proteins and unbiased panning of metabolomes. *Biochemistry* 54(3):909–931. https://doi.org/10.1021/bi501388y

Warrack BM, Hnatyshyn S, Ott KH, Reily MD, Sanders M, Zhang H, Drexler DM (2009) Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B* 877(5-6):547–552. https://doi.org/10.1016/j.jchromb.2009.01.007

Wu CJ, Hamada MS (2011) *Planning, analysis, and optimization*. In: David J. Balding, Noel A.C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein,Geert Molenberghs, David W. Scott, Adrian F.M. Smith, and Ruey S. Tsay (ed) Experiments, 3rd edn. John Wiley & Sons, p 1–736, https://doi.org/10.1002/9781119470007

Yon T, Sibat M, Réveillon D, Bertrand S, Chinain M, Hess P (2021) Deeper insight into Gambierdiscus polynesiensis toxin production relies on specific optimization of high-performance liquid chromatography-high resolution mass spectrometry. *Talanta* 232(1):122400. https://doi.org/10.1016/j.talanta.2021.122400

Yue Y, Bao X, Jiang J, Li J (2022) Evaluation and correction of injection order effects in lc-ms/ms based targeted metabolomics. *Journal of Chromatography B* 1212(1):123513. https://doi.org/10.1016/j.jchromb.2022.123513

Zanatta AC, S. Borges M, Mannochio-Russo H, Heredia-Vieira SC, Campaner dos Santos L, Rinaldo D, Vilegas W (2022) Green chromatography as a novel alternative for the quality control of serjania marginata casar. leaves. *Microchemical Journal* 181(10):107671. https://doi.org/10.3390/ph15101289

Zhang Xw, Li Qh, Xu Zd, Dou Jj (2020) Mass spectrometry-based metabolomics in health and medical science: a systematic review. *RSC Advances* 10(6):3092–3104. https://doi.org/10.1039/C9RA08985C

Zhao A, Chen F, Ning C, Wu H, Song H, Wu Y, Chen R, Zhou K, Xu X, Lu Y, Gao J (2017) Use of real-time cellular analysis and Plackett-Burman design to develop the serum-free media for PC-3 prostate cancer cells. *Plos ONE* 12(9):0185470. https://doi.org/10.1371/JOURNAL.PONE.0185470

Zhou Y, Song JZ, Choi FFK, Wu HF, Qiao CF, Ding LS, Gesang SL, Xu HX (2009) An experimental design approach using response surface techniques to obtain optimal liquid chromatography and mass spectrometry conditions to determine the alkaloids in meconopsi species. *Journal of Chromatography A* 1216(42):7013–7023. https://doi.org/10.1016/j.chroma.2009.08.058

Appendix F

Paper 2

NMR-Onion - a transparent multi-model based 1D NMR deconvolution algorithm

Mathies Brinks Sørensen[a], Michael Riis Andersen[b], Mette-Maya Siewertsen[a], Rasmus Bro[c], Mikael Lenz Strube[d] and Charlotte Held Gotfredsen[a,*]

[a]*Department of Chemistry, Technical University of Denmark, Kgs Lyngby, DK-2800, Denmark*
[b]*Department of Applied Mathematics and Computer Science, Kgs Lyngby, DK-2800, Denmark*
[c]*Department of Food Science, University of Copenhagen, Frederiksberg, DK-1958, Denmark*
[d]*Department of Biotechnology and Biomedicin, Kgs Lyngby, DK-2800, Denmark*

ARTICLE INFO

Keywords:
Statistical evidence
Time domain models
Open source
Deconvolution
High sensitivity
Extensive overlaps
Computationally efficiency

ABSTRACT

We present an open source computationally efficient Python/PyTorch based algorithm, dubbed NMR-Onion, capable of automatically assisting in deconvolution of 1D NMR spectra. The NMR-Onion framework presents two novel time domain models capable of handling asymmetric non-Lorentzian line shapes. The core modules for resolution enhanced peak detection and the digital filtering of user selected key regions ensure high precision peak estimation and computational time efficiency. NMR-Onion provides the user with three built-in statistical models, which are automatically selected using the BIC criterion. Furthermore, NMR-Onion also quantifies the repeatability of the results by evaluating post modelling uncertainty. Applying the NMR-Onion algorithm will aid in reducing excessive peak detection.

1. Introduction

Deconvolution of especially 1D spectra in NMR spectroscopy is a key step when elucidating complex 1D ^{1}H NMR spectra. These spectra contain a vast amount of structural, quantitative and dynamic information. Information may be extracted even from 1D spectra and are important in many areas of science where complex mixtures are being studied such as metabolomics [1] and *in situ* samples. As complex spectra often contain hundreds of highly overlapping signals from compounds present in different concentrations, the traditional manual spectral analysis is inadequate due to the complexity of the spectra. However, applying different methodologies to achieve automated extraction of spectral information, and subsequent comparison with spectral databases can facilitate this process. The most commonly used approaches to extract data are based on either deconvolution [2][3] or binning of frequency buckets [4][5]. The latter has been extensively applied in metabolomics following the invention of intelligent bucketing [6] capable of automatically bucketing various frequency bins. There are many successful applications of this approach in many scientific areas such as disease diagnotics[7][8], natural product identification [9][10], foodomics [11] and drug discovery [12]. Unlike binning, deconvolution aims at resolving single peaks, potentially uncovering more information from small peaks otherwise lost in the binning process. However, increased information content leads to an increase in complexity both with respect to spectral interpretation, mathematical modelling and computational demands. To tackle the three issues of complexity, multiple contributions have advanced the field of spectral deconvoultion during the last 30 years. It started with the pioneering research of Bretthorst, who developed a probabilistic framework for modelling the shape and number of NMR signals based on free induction decay (FID) data [2]. On top of the mathematical framework of Bretthorst, the method of Craft [13] was introduced, combining probabilistic modelling with digital filtering, lowering the computational complexity of model estimation. These approaches are focusing on time domain data, where peaks are assumed to be a sum of exponentially damped sinusoids, corresponding to the Lorentzian line shapes in the frequency domain[14]. In recent research, a weighted sum of Gaussian and Lorentzian line shapes was combined into a frequency domain model known as pseudo Voigt line shapes [15, 16]. This approach has been implemented in the commercial MNOVA GSD software [17] and also in the R software package of rNMRfit [18]. In rNMRfit a baseline correction method producing robust results was further implemented. The frequency domain methods have been incorporated into algorithms capable of matching deconvoluted NMR signals with databases. This has led to the automatic detection of compounds as seen in the popular frameworks of NMRbatman [19], Bayesil [20] and the commercial program of Chenomx [21].

Following 30 years of research, the field of deconvolution has come a long way, but issues still remain. The challenges being statistical evidence of model quality when compared with other models (model selection), parameter uncertainty, and the statistical evidence for the presence or absence of highly overlapping peaks.

Here we propose a novel five-step process dubbed NMR-Onion, which evaluates model quality by selecting the best model for estimating frequencies, coupling constants (within a frequency distance matrix), amplitudes, and parameter uncertainty within a user-specified region of interest (ROI).

[*] Corresponding author.
E-mail address: chg@kemi.dtu.dk (Charlotte Held Gotfredsen).

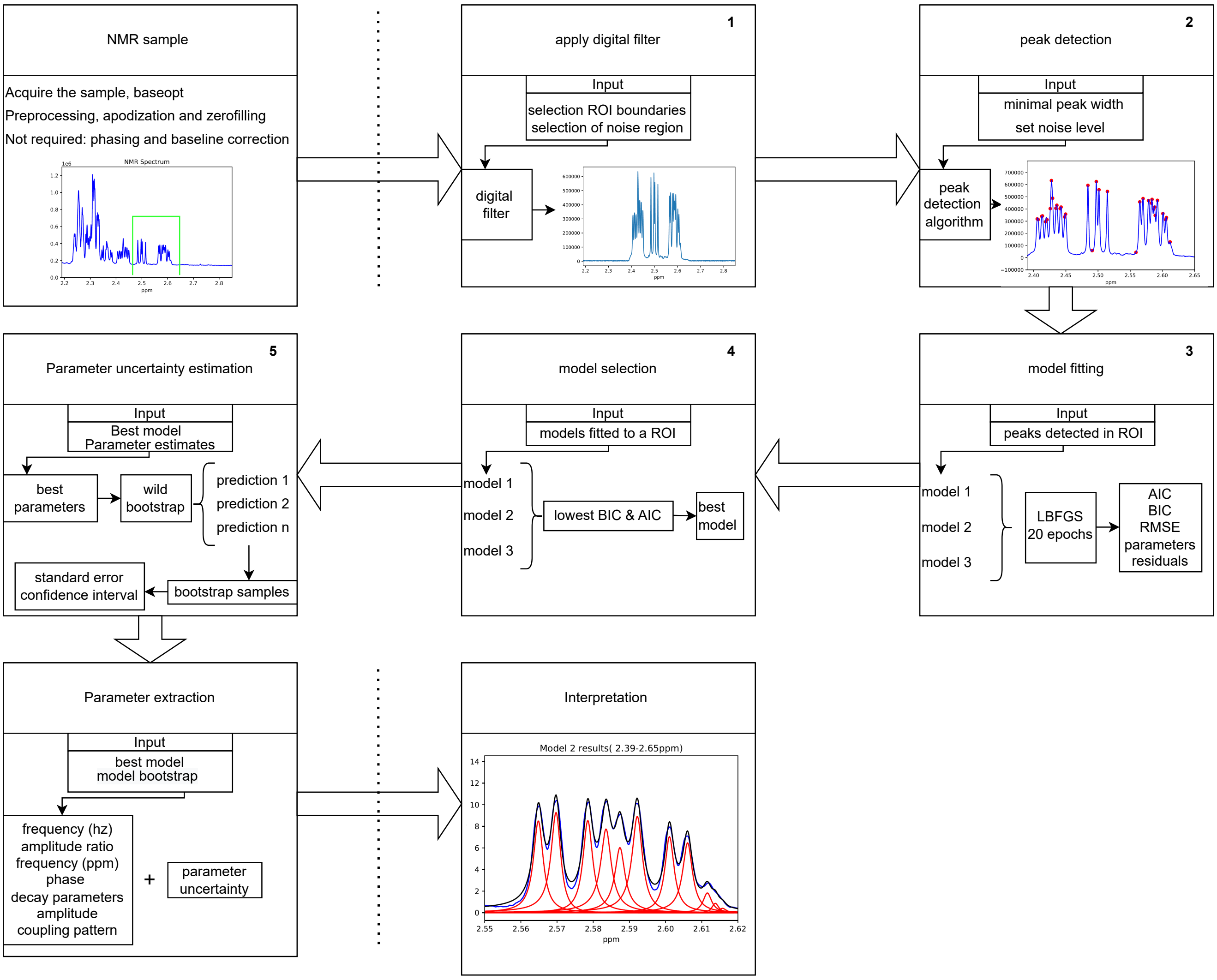


Figure 1: Visual representation of the NMR-Onion algorithm outlined in the above five steps. The individual steps described below are marked with numbers 1-5 in the figure.

Furthermore, during uncertainty evaluation, confidence intervals are generated enabling a statistical-based evaluation of overlapping signals. The five-step process making up the NMR-Onion algorithm is outlined below and visually presented in figure 1.

- First step: The computational burden of the algorithm during the fitting process is reduced by applying a digital band-pass filter, generating a user-specified ROI.

- Second step: Peaks within a ROI are identified by applying the first and second-order derivative Savitzky Golay filter [22] in tandem with resolution enhancement. [14]

- Third step: Multiple times domain models are applied to the ROI. To avoid multimodality of the frequency estimations, the detected peak values are used as initial parameter input.

- Fourth step: The best model is selected using a likelihood-based information criterion[23]

- Fifth step: The uncertainty of each parameter is found using the wild bootstrap algorithm. [24]

2. Theory

2.1. Model formulation

Outlined in figure 1, one of the goals of the algorithm is identifying chemical shifts, intensities, distances within a given multiplet (to obtain J - spin-spin coupling constant information) and the number of underlying signals found in an NMR spectrum. Using the time domain, the spectrum may be mathematically formulated as a sum of damped complex sinusoids:

$$y(t) = \sum_{k=1}^{K} A_k \cdot exp(2j\pi\omega_k t + j\phi_k) \cdot \Psi_n(\boldsymbol{\rho_k}), \qquad (1)$$

where the term of $\Psi_n(\boldsymbol{\rho_k})$ in equation (1) corresponds to the n'th decay function and $\boldsymbol{\rho_k}$ is a vector representation of all parameters belong to the n'th decay. The decay function can be represented by a negative exponential function, $\Psi_1(\boldsymbol{\rho_k}) = exp(-\alpha_k t)$, as described by Keeler[14] and in equation (2).

$$y(t) = \sum_{k=1}^{K} A_k \cdot exp(2j\pi\omega_k t + j\phi_k) \cdot \Psi_1(\boldsymbol{\rho_k}). \qquad (2)$$

The parameters of equation (2) are defined as ω=frequency (Hz), A=amplitude, ϕ=phase, α=decay rate for the k'th sinusoid. The relation of equation (2) holds true if no artifacts impact the experimental data. However, in reality, shimming, eddy current, temperature fluctuations, receiver gain set too high/low, sample conditions, etc. may impact the acquired spectrum in an unpredictable fashion[25]. To the best of our knowledge, no one has managed to take into account all distortions. Hence, we seek to expand the damping term of $\Psi(\boldsymbol{\rho_k})$ such that some of the aforementioned distortions may be accounted for by introducing a flexible decay rather than a pure exponential decay. To achieve this flexibility we introduce two novel time domain models. The first being a weighted sum of Gaussian and exponential decays, also known as a pseudo-Voigt shape in the frequency domain, and the other being an exponential power law model. The pseudo-Voigt will lead to a reformulation of the decay term in equation (1) formulated as:

$$\Psi_2(\boldsymbol{\rho_k}) = (1 - \eta_k) \cdot \exp(-\alpha_k t) + \eta_i \cdot \exp(-\alpha_k t^2) \quad (3)$$

resulting in the full model of the weighed sums being

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_2(\boldsymbol{\rho_k}) \qquad (4)$$

With f being equal to the harmonic term of equation (1):

$$f(A_k, \omega_k, \phi_k) = A_k \cdot \exp(2j\pi\omega_k t + j\phi_k). \qquad (5)$$

The addition of the η term produces a weighting between an exponential and Gaussian decay type. When $\eta = 0$ a pure exponential decay is produced and equation 4 is reduced to equation (2). If $\eta = 1$ a pure Gaussian decay is achieved as the $\exp(-\alpha_k t)$ term becomes zero. The second model is a further generalization of (2), introducing a power term $\Psi_3(\boldsymbol{\rho_k}) = \exp(-\alpha_k t^{\beta_k})$ as the decay function, resulting in:

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_3(\boldsymbol{\rho_k}) \qquad (6)$$

The addition of the β term creates a stretched exponential decay when $\beta > 1$ and a compressed exponential when $0 < \beta < 1$. Finally, when $\beta = 1$ equation (6) is reduced to the classic exponential decay shown in equation (2).

The advantage of the power law decay model of equation (6) and exponential mixture model of equation (4) comes from their flexibility to describe non-Lorentzian peak shapes. To add further flexibility to the models of equation (2), (4) and (6), asymmetric line shapes are included by following the same line of thought as presented in works of Matviychuk [25], introducing a complex skewing term of $exp(j\gamma_k)$ for each signal. This results in equation (2), (4) and (6) being formulated as:

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_1(\boldsymbol{\rho_k}) \cdot \exp(j\gamma_k)t \qquad (7)$$

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_2(\boldsymbol{\rho_k}) \cdot \exp(j\gamma_k)t \qquad (8)$$

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_3(\boldsymbol{\rho_k}) \cdot \exp(j\gamma_k)t \qquad (9)$$

respectively. The $exp(j\gamma_k) \cdot t$ term results in the peak being skewed to the left if $\gamma > 0$ and to the right $\gamma < 0$, where γ is constrained within $[-\frac{\pi}{2} : \frac{\pi}{2}]$. With the models formulated, a routine for estimating the parameters can be generated by turning equation (7), (8) and (9) into a non-linear least squares optimization (NLS-opt) problem. For the NLS-opt a matrix formulation of all models, including a residual term E, is set up

$$Y = ZA^T + E \qquad (10)$$

Here, A is a $1 \times K$ vector with each element defined as a complex amplitude $a_k = A_k \cdot \exp(2j\pi\phi_k)$ and Z is a $N \times K$ matrix containing the time-dependent terms of equation (7),(8) and (9). Each column of the Z-matrix represents a single sinusoid/signal with its own subset of parameters, while the rows represent the signal value at the n'th time point. Y is a $1 \times N$ vector with each element being a measured time point in the FID. Finally, E is a $1 \times N$ residual vector which is assumed to be identically, independently distributed via a Gaussian distribution of $E \overset{\text{i.i.d}}{\sim} N(0, \sigma)$. The model of equation (10) is reduced by integrating out the time independent terms, following the same method originally suggested by Bretthorst [2], the A matrix can be expressed as a function of the Z matrix:

$$A = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{YZ} \qquad (11)$$

This enables us to turn the model into an NLS-opt problem, by rewriting equation (10) as minimization of the sum of squared errors (SSE) loss function only depending on the Z-matrix.

$$E = Y - AZ = Y - ((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{YZ})Z \qquad (12)$$

where

$$\text{SSE} = E^H E \qquad (13)$$

and

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}(\text{SSE}) \qquad (14)$$

Here, H refers to the complex conjugated matrix (Hermitian transposed matrix) and $\hat{\theta}$ is a vector of estimated parameters. Ideally, the SSE formulation of equation (14) would be a simple standardized criterion for minimizing. However, due to the model being a superposition of sinusoids, the possibility of spurious signals occurring is very high. Hence a penalty term minimizing the phase variance is introduced into equation (14)

$$\hat{\theta} = \arg\min_{\theta} SSE + \frac{1}{K}\sum_{k=1}^{K}(\phi_k - \bar{\phi}) \tag{15}$$

The penalty term of equation (15), where $\bar{\phi}$ is the mean phase, ensures that the phases do not get out of control, preventing large reversed phased peaks from occurring. One should note that the magnitude of a phase is very small, being $[-\pi : \pi]$, compared to that of the SEE of an NMR spectrum. So, for the penalty criterion to have an effect, all FID data is normalized via the Frobenius norm [26]. The loss function of equation (15) may seem simple enough. However, estimation of a global minimum producing optimal parameters for (7), (8) and (9) is notoriously difficult, as the model has an unknown number of components (K), is non-linear, multi-modal with respect to the frequencies (ω_k) and computationally expensive. Therefore the next sections provide insights as how to reduce computational burdens, estimation of model order, and handling multimodality. The numerical constraints details for each parameter can be found in the constraining parameters supplementary

2.2. Computational bottlenecks

When applying the NMR-Onion algorithm to a metabolomic 1D ^{1}H NMR spectrum, it is common to have more than 1000 peaks present. This leads to a computationally expensive problem, as the Z-matrix of equation (10) becomes very large with $K > 1000$ and $32768 < N < 131072$ time points, depending on the sample acquisition scheme. Fortunately, 1D ^{1}H NMR data is sparse (none overlapping regions are independent), often containing large regions of redundant noise and typically only specific regions of the spectrum are of interest. Therefore, the Z-matrix dimensions can be greatly reduced by selecting smaller regions of interest (ROI's), resulting in a lower number of columns (K). Reduction is achieved by using a digital band-pass filter as outlined in step 2 of figure 1. There are multiple approaches to implement a digital filter, for instance, the CRAFT algorithm applied a finite impulse response (FIR) filter in combination with a Blackman window function [13], while wavelet packet-like filter banks were used in the works of Djermoune [27] to make an adaptive subband filtering scheme. In NMR-Onion we have adapted the super-Gaussian band-pass digital filter presented in the works of Hulse and Foroozandeh [28] with some modifications for how to handle baseline artifacts and noise estimation of data. This approach was chosen, as the idea of incorporating prior knowledge about the noise level is ideal for NMR deconvolution, making it much easier to separate signals from noise.

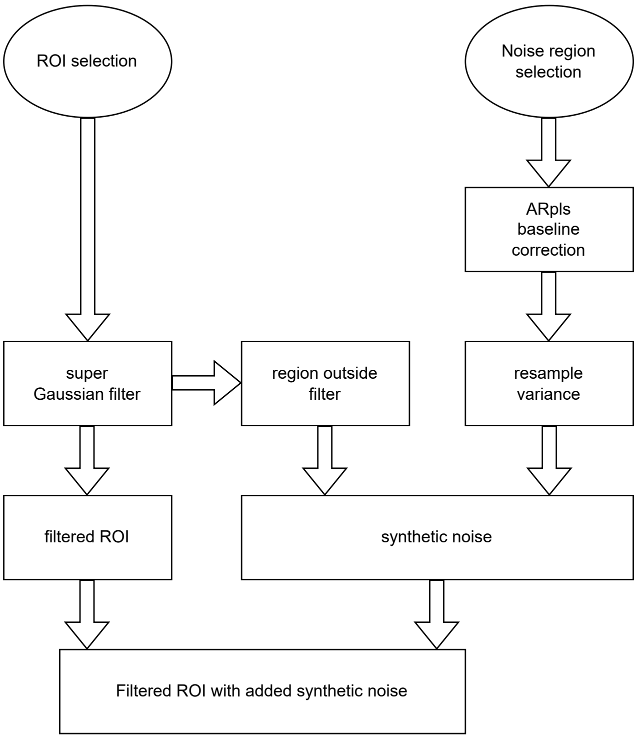


Figure 2: Visual representation of the digital filter workflow. The ellipses represent user input which should be the targeted ROI and noise region (in ppm). The output is a digitally filtered ROI with only signal remaining at the target range while the reaming part of the spectrum will contain synthetic noise

The first modification to the filter was the incorporation of a baseline correction step before estimating the noise level found in a noise region. For baseline correction, the algorithm of asymmetrically re-weighted penalized least squares smoothing (ARpls)[29] was applied. Additionally, the noise level was determined using, a re-sample scheme which was constructed to achieve a more robust noise level estimation. Assuming the noise is average Gaussian white noise, 1000 samples are randomly drawn from a Gaussian distribution with 0 mean and variance found from the baseline corrected noise region. Subsequently, the mean of the re-samples is utilized to generate an artificial noise floor for the filter. The modified filtering process is shown in figure 2.

Another bottleneck, apart from model dimensions, originates from the optimization technique, programming language, and specification of loss function derivatives. To address these challenges, we apply the modern tool of Pytorch implementing the loss function of equation (15). We rely on the quasi-newton optimizer based on the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm[30], which for the NMR-Onion algorithm, exhibited much faster convergence than the Scipy implementation [31] and has the capability of applying automatic differentiation (AD) through the autograd module [32]. AD provides information on both the gradient and Hessian needed to adequately estimate the parameters of the model in a fast and efficient way, not relying on manual identification and implementation of the first (gradient) and second (Hessian) order derivative (*vide infra* discussion section).

2.3. Peak detection

Apart from computational bottlenecks, handling the multi-modality of the frequencies and estimating the number of signals for the model is notoriously difficult. This challenge was originally addressed by Bertthorst [2] by applying a search pattern algorithm[33] coupled with the maximum of power spectral density to identify initial values for the frequencies, sequentially fitting peaks, until the number of components was identified by a Bayesian generalized likelihood criterion. Rubtsov and Griffin applied a different approach by using reversible Monte Carlo Markov Chain (MCMC) jump for model order selection and parameter estimation [34]. For frequency-based methods, the combined use of the first and second-order derivative Savitzky Golay (SG) filters [22] can be applied for detection, as high field NMR data is generating high-resolution data [35]. The NMR-Onion algorithm utilizes the derivatives of a third-order SG filter for peak detection, combined with resolution enhancements[14] and a clustering algorithm for the potential merging of peaks. We adopted the rNMRfind algorithm[36] and implemented it within Python, forming the third building block of the algorithm outlined in figure 1. The detection algorithm revolts around having a user input defining the minimal peak width parameter. In general, as shown in the original work of rNMRfind[36], setting this parameter higher would decrease the number of detected peaks, offering a more conservative detection. The default noise threshold is set as 2 times the standard deviation found by the inter quantile range (IQR) of the first principle component from the SG-filtered first and second-order derivatives of the real and imaginary spectrum.

2.4. Model selection

Post model fitting, the model selection outlined in the fourth building block of figure 1 is carried out. The selection criterion may be addressed using different approaches, but for the NMR-Onion algorithm we have chosen to use the Akaike Information Criterion (AIC) [37] or the Bayesian information criterion (BIC) [38] and defined it as

$$AIC = -2 \cdot \mathcal{L}(\theta) + 2p \cdot K \tag{16}$$

$$BIC = -2 \cdot \mathcal{L}(\theta) + 2 \cdot p \cdot K \log(N) \tag{17}$$

Here p = number of parameters per sinusoid, K is the number of sinusoids and $\mathcal{L}(\theta)$ refers to the log-likelihood of equation (4) which according to the works of Nadler [39] may be formulated as

$$\mathcal{L}(\theta) = \frac{-N}{2}(1 + \log(2\pi)) - \frac{N}{2}\log[E^H E] \tag{18}$$

corresponding to a Gaussian log-likelihood by applying the maximum likelihood estimator (root mean squared error) for the variance. The final model is the one which most adequately describes the data with the fewest components, here reflected in the lowest value obtained for equation (16) or (17). The main difference between the approach using (16) or (17) is the strictness of the penalty imposed related to the number of model parameters. BIC imposes a harder penalty for having more model parameters than AIC. Hence, BIC would result in simpler models as opposed to AIC. The NMR-Onion algorithm considers both options to be valid and provides the option for testing both for different scenarios. Other criteria have also been developed such as the kmap criterion [40]. However, this has only been applied to none decaying sinusoids and does not explicitly state how the decay rate and flexibility constant should be weighted.

2.5. Parameter uncertainties

A key result, often not addressed within deconvolution algorithms, are the estimation of uncertainties. In particular, the uncertainty regarding the repeatability of a peak and the estimation procedure of parameter uncertainties. The uncertainty estimation may be addressed in several ways, a popular method being the application of the second order derivative of the model negative expected log-likelihood function, producing the fisher information which may be used in the Wald approximation confidence interval [41]. However, as shown by Wilson [3], the profile likelihood of the parameters, particularly the frequencies, is far from representing a second-order curvature. Therefore, to reliably quantify uncertainties, different methods should be imposed. A popular solid method is the use of various Monte Carlo Markov chain (MCMC) schemes, as exemplified by Jie Hao [19], but the increase in information comes at a price. In the case of the MCMC schemes high precision is achieved, but at the cost of heavy computational burdens, other alternatives are addressed in the discussion.

Algorithm 1 Wild bootstrap algorithm

```
compute residuals based on best model fit
$\varepsilon_i = y_i - \hat{y}_i$
draw bootstrap samples
**for** $b = 1, 2, .., B$ **do**
    **for** $i = 1, 2..., N$ **do**
        $\tilde{\varepsilon}_i^b \sim Z_i^b \cdot \varepsilon_i,\ Z_i^b \sim N(0, 1)$
        $\tilde{y}_i^b = \hat{y}_i + \tilde{\varepsilon}_i^b$
        $\boldsymbol{\theta_b} = \arg\min_{\boldsymbol{\theta_b}} SSE(\tilde{y}_i^b)$
    **end for**
**end for**

Generate $\alpha$ level confidence interval for the k'th parameter
$\boldsymbol{\theta_{kCI}} = [\boldsymbol{\theta_{k_{b_{\alpha/2}}}}, \boldsymbol{\theta_{k_{b_{1-\alpha/2}}}}]$.

Here $\boldsymbol{\theta_{k_b}}$ referrers to the k'th parameter CI of b'th estimation at the upper and lower CI value of $\alpha$
Generate sample variance for the k'th parameter
$\bar{\boldsymbol{\theta_k}} = \frac{1}{b}\sum_{b=1}^{B} \theta_{k_b}$
$\boldsymbol{\theta_{k_{var}}} = \frac{1}{b}\sum_{b=1}^{B}(\boldsymbol{\theta_{k_b}} - \bar{\boldsymbol{\theta_k}})^2$
**return** $\boldsymbol{\theta_b}, \boldsymbol{\theta_{k_{CI}}}, \boldsymbol{\theta_{k_{std}}}$
```

In the NMR Onion algorithm, we have not applied a Bayesian model but rather a frequency approach, applying the ad hoc method of wild bootstrapping [24] to estimate parameter uncertainties. The scheme is outlined in figure 1 step six and formally outlined in algorithm 1 above. The idea behind the bootstrap scheme shown in algorithm 1 is to enable the estimation of the confidence interval (CI) for every parameter. This is of particular importance for the frequencies as CI overlaps would indicate that highly overlapping peaks have a significant probability of not being detected in replicates, as within a replicate the resolved peak may have been fused into one peak. We classify peaks with overlapping CIs as potential resolved peaks (PRPs). To evaluate the repeatability of the PRPs, independent experimental replicates should be made, investigating if the overlapping peaks are consistent or occurring due to random sample variations.

3. Data acquisition

Here different spectra were acquired with two different experimental setups. The first experiment consisted of sample mixtures between phenol and isopropanol dissolved in 90:10 H_2O : D_2O and D_2O. The mixture ratios of phenol: isopropanol was set according to Table1, producing a total of 5 data sets. The second experimental set consisted of a sample containing the complex molecule of phytosteroid Diosgenin dissolved in chloroform. Two samples of the same final concentration (4 mM) were made. All spectra were acquired on a Bruker AVANCE III HD 800 MHz spectrometer equipped with a 5 mm TCI cryoprobe. The ^{1}H pulse programs depended on the solvent used, where spectra acquired with only 10% D2O in water a zgespg pulse sequence was used whereas for samples dissolved in CDCl3 a zg and zg30 pulse sequence scheme was used, for all spectra the baseopt rectangular filter setting was used to minimize baseline and first-order phase distortions. All spectra were acquired at 25° C, with a relaxation delay of 2 s, 128 scans, and 64K data points. For the data analysis, the NMR-Onion program was run on a virtual-box Ubuntu (64-bit) Linux operating system with a Processor Intel(R) Core(TM) i9-9880H CPU @ 2.30GHz, 2304 MHz, 8 Core(s), 16 Logical Processor(s). In addition, all experimental data were normalized according to the Forbinous norm prior to executing any model fitting as part of the NMR-Onion algorithm. The modelling part of NMR-Onion was performed using a learning rate (lr) of 0.1 and 20 epochs (iteration cycles) with an exponential lr schedule reducing the lr by 30% per epoch.

Table 1
Experiment 1 sample compositions of phenol:isopropanol mixtures

Sample No.	phenol (mM)	isopropanol (mM)
1	1	1
3	2	1
3	1	2
4	0.5	0.5
5	0.1	0.1

Following data acquisition, pre-processing was carried out in Bruker TopSpin version 4.0.7 [42]. This included apodization using exponential line broadening of 0.3 Hz followed by automatic phase and baseline correction (using a polynomial 5). The preprocessed data was then transferred from Topspin to the Python environment using the NMR-Glue[43] package which enables importing of Bruker data along with preprocessing and acquisition parameters.

All acquired data (raw and processed) for this paper can be found in the attached supplementary material and is also available for download on our GitHub: `https://www.github.com/Mabso1/NMR-onion`

4. Results

4.1. Case Study 1

For the first experiment setup, the goal is to validate the NMR-Onion algorithm against easily identifiable peak frequencies and coupling patterns, covering a large area of the proton spectrum. The main peaks of the phenol:isopropanol composition are outlined in Table 2

Table 2
Experiment 1 theoretical peak locations and experimental coupling patterns

Compound	δ (ppm)	Coupling pattern	J (Hz)
Phenol	6.84	broad doublet	~8.9
Phenol	6.91	triplet of triplets	8.0, 0.95
Phenol	7.24	doublet of doublets	7.7, 8.6
Isopropanol	1.16	doublet	6.4
Isopropanol	4.01	septet	6.4

In addition to peak validity, the dilution series (see table 1) was made such that the lower limit of detection could be identified. Hence, regions of interest (ROIs) aligning with the theoretical peak locations were constructed, by applying the filtering process described in section 2.2, generating four ROIs in total (see table 3).

Table 3
Experiment 1 region of interest and noise region

Region No.	lower cutoff (ppm)	higher cutoff (ppm)
1	0.9	1.2
2	3.8	4.1
3	6.6	7.0
4	7.1	7.5
Noise region	-0.1	-0.2

For a better description of detection capabilities, the SNR of each ROI at all levels of concentrations was calculated, as $SNR = 10 \log_{10} \frac{S}{N}$ and used to compare the performance of the algorithm (see table 4)

A graphical result is presented for the highest and lowest SNR region samples in 3 and 4. The best model for each ROI was automatically selected by the BIC of equation (17) (AIC gave similar results)

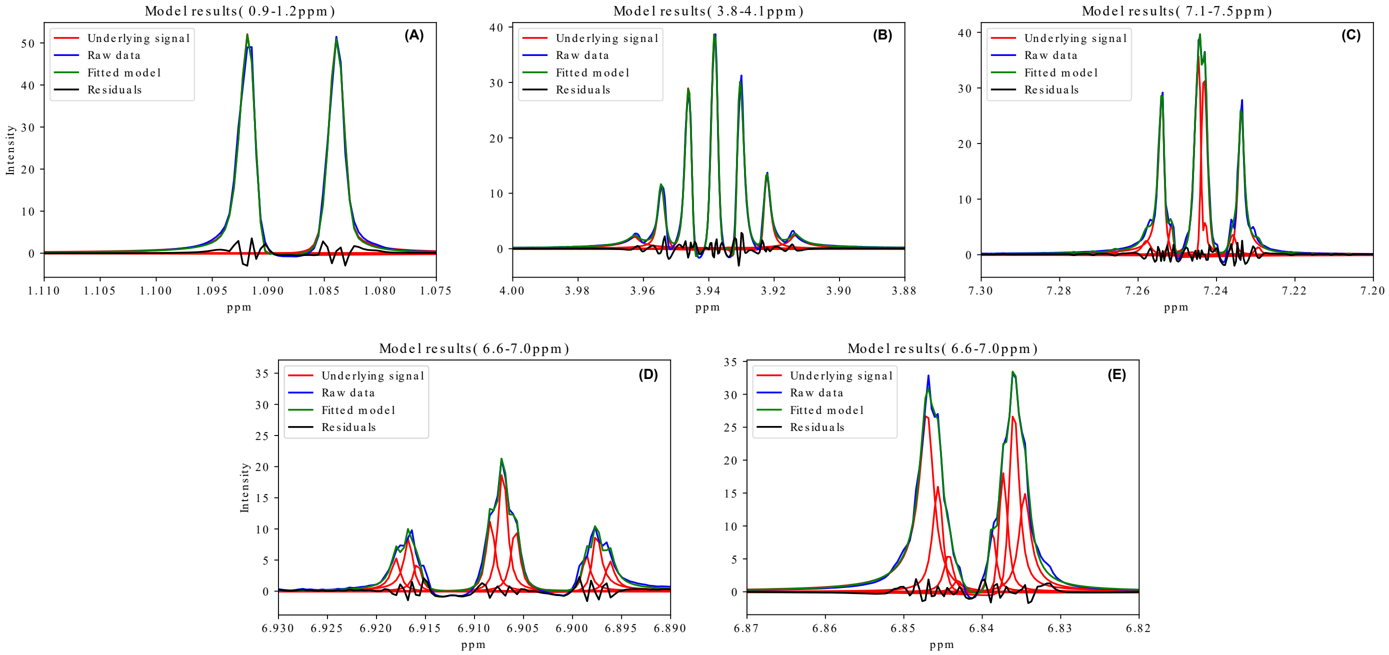


Figure 3: Visual model deconvoultion of sample No. 1 (concentrated sample - see table 4 for SNR values). (A) zoom in on the targeted doublet of Region 1. (B) zoom in on the targeted septet of Region 2. (C) zoom in on the targeted doublet of doublet region 4. (D) Zoom in on the triplet of triplets subpart of region 3. (E) zoom in on the board doublet subpart of region 3.

Table 4
Experiment 1 SNR values

Region No.	Sample No.	SNR (dB)
1	1	46.9
2	1	35.0
3	1	38.2
4	1	38.2
1	2	47.2
2	2	34.7
3	2	34.3
4	2	34.0
1	3	34.4
2	3	30.7
3	3	37.0
4	3	36.9
1	4	39.7
2	4	27.7
3	4	30.4
4	4	30.3
1	5	34.2
2	5	22.2
3	5	23.9
4	5	23.5

From figure 3 A and figure 4 A, the targeted doublet is clearly identified, likewise the targeted septet is found within figure 3 B and figure 4 B. For figure C, for both the lowest and the highest SNR sample doublets of doublets were identified.

The third ROI contains two sub-ROI that exhibit second-order effects, and for this reason, caution using a first-order multiplicity analysis is needed. The first sub-region (figure 3 D and 4 D), is observed to be a triplet of triplets as expected if applying a 1st order multiplicity analysis and disregarding J_{para}, whereas for sample 5 broader signals than for sample 1 are observed and the small J coupling constant is not resolved when visually expected but only after deconvolution. For the second sub-ROI of the third region, sample one (figure 3 E) is a doublet of multiples, whereas sample 5 (figure 4 E) shows different splits making a 1st order multiplicity analysis non-applicable. As for the residuals of each plot, none fulfills the assumptions of being white noise and this will be further discussed within section 4.3 and 5.
In addition to individual deconvolution analysis, model stability across the 5 identical regions of different samples was investigated, summarizing the best models selected across all 4 regions of 4 in table 5.

Table 5
Experiment 1 model summary across all regions of interests (ROI) for 5 different samples

ROI	model 1	model 2	model 3
1	0	2	3
2	0	1	4
3	1	1	3
4	0	0	5
Total	1	4	15

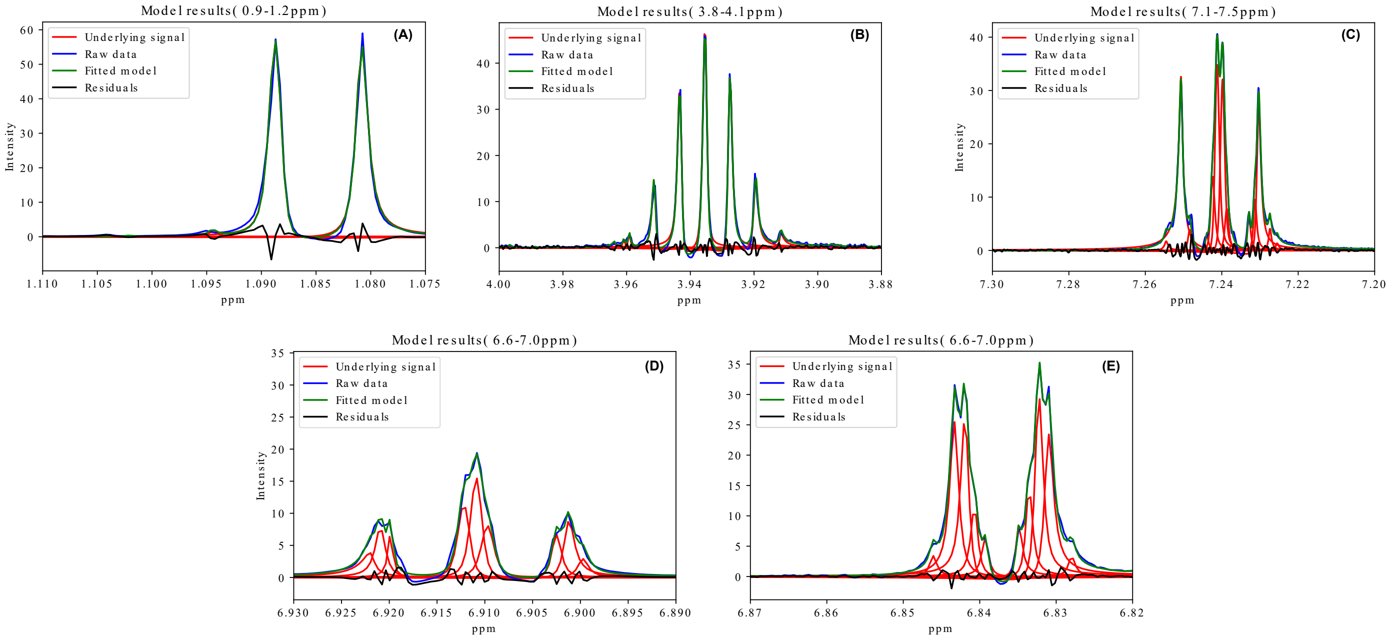


Figure 4: Visual model deconvolution of sample No. 5 (lowest concentration sample- see table 4 for SNR values). (A) zoom in on the targeted doublet of Region 1. (B) zoom in on the targeted septet of Region 2. (C) zoom in on the targeted doublet of doublet region 4. (D) Zoom in on the triplet of triplets subpart of region 3. (E) zoom in on the board doublet subpart of region 3.

From table 5, it is evident that model 3, the power law model (equation (9)) serves as the general better model, while in some cases, model 2, the mixture model (equation (8)) is the second best, but model 1, the traditional model of exponential decay (equation (7)) only occurs in 1 instance. It should be noted that each model tested on the data has had the skew term included (see equation (7), (8) and (9)).

A key feature from NMR-Onion is the ability to detect PRPs in highly overlapping signals. The feature was utilized to investigate how many of the total peaks detected in each region may be less likely to appear in replicates as they originate from very high overlapping peaks. The results are summed up in table 6 for all 5 experiments including both targeted peaks (see table 2) and peaks from ^{13}C satellites:

Table 6
Experiment 1, summary of potential false peaks (PRPs) found across all datasets

ROI	Total PRPs	Targeted peaks
1	2	0
2	3	0
3	130	66
4	39	9

From table 6, it is observed that the regions containing the majority of PRPs are the third and fourth region, whilst the first and second region barely contain any, this matches well with the visual results of figure 3 and 4, as peaks are highly overlapping and exhibit second-order effects and the presence of small unresolved *J* coupling constants. However, it should be noted that many of the PRPs do not come from the targeted peaks of table 2, but rather from the smaller ^{13}C satellites and some impurities having CI overlaps in sample 1-3 where they could be detected. This was particularly evident in ROI 3 as second-order effects caused different multiplicity patterns within the signals of ROI 3. As for ROI 4, it was revealed that the targeted peaks, where PRPs were identified, occurred only within sample 5 and sample 2 (see more in the discussion section).
Finally, it was observed that across each sample, the consistently detected peaks all appeared, within the range of CIs of the first sample (or any other sample CI), indicating that the model is adequately predicting peak location consistently across concentration.

The analysis of the PRPs within this case may not enrich the analysis too much, as it would take replicates of the same concentrations to pinpoint specific PRPs occurring due to sample-to-sample variation. Furthermore, experiment 1 also has very distinguishable regions, making PRPs detection less impactful. Hence, to truly showcase the value of the PRP feature, a second case study was constructed utilizing a more complicated molecule.

4.2. Case Study 2

The second case study revolts around analyzing the sample containing the complex phytosteroid diosgenin molecule. The goal of this experiment is to demonstrate how NMR-Onion reliably identifies peaks and detects PRPs across two,

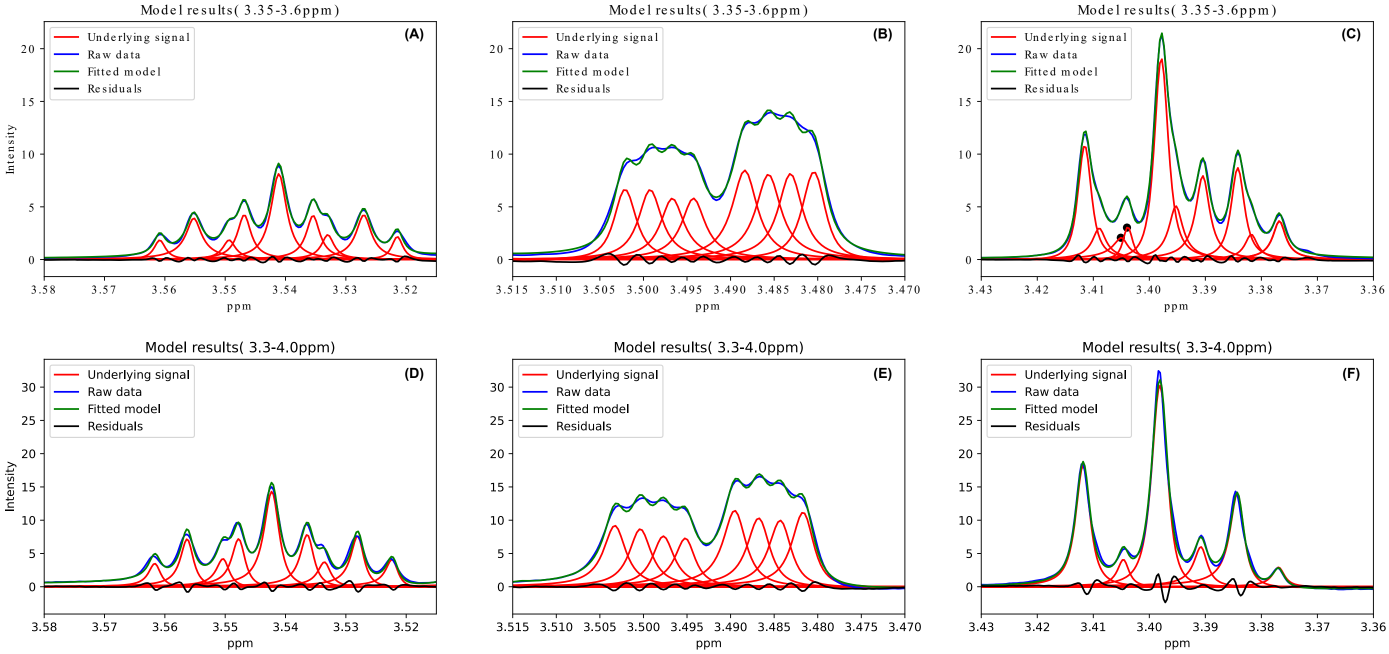


Figure 5: Visual model deconvolution of two samples within case study 2, the black dots indicate that one peak is resolved to potentially two peaks. (A) zoom in on sub-region 1, sample 1. (B) zoom in on sub-region 2, sample 1. (C) Zoom in on sub-region 3, sample 1. (D) Zoom in on sub-region 1, sample 2. (E) zoom in on sub-region 2, sample 2. (F) zoom in on sub-region 3, sample 2.

in principle, identical samples. We selected one ROI and noise region outlined in table 7.

Table 7
Experiment 2 region of interest and noise region

Region No.	lower cutoff (ppm)	higher cutoff (ppm)
1	3.35	3.60
Noise region	-0.1	-0.2

The underlying model of the ROI shown in table 7 was selected by applying the same approach as stated within section 4.1, resulting in both datasets being based on the exponential power law model (equation (9)). The graphical result based on the exponential power law in each replicate is presented within figure 5, which gives rise to 3 different small sub-regions.

Two of the sub-regions shown figure 5 A, B and D, E, revealed no obvious difference as the same peaks were detected within both samples. However, the third sub ROI revealed a PRP might be present within the first sample (figure 5 C), as the same peaks cannot be found within (figure 5 F). The detected PRP indicates that a replicate may have low probability of resolving the same peaks. We also do note that the second sample does have as many peaks detected as the first sample, but the residuals do reveal that signals missing in the second sample match the signals found in the first sample (see more in the discussion section).

4.3. Comparison with other software

To better evaluate the results of the NMR-Onion algorithm we choose to compare it with one of the most popular and widely applied algorithms of MNOVA GSD. The same experiments as within case study 1 and 2 are run with the MNOVA GSD algorithm and the results are visually shown in figure 6 and figure 7. Unfortunately, it is not possible to directly compare metrics such as root mean squared error or BIC/AIC as the internal normalization of data in MNOVA and loss function cannot be extracted, therefore only visual comparisons of residuals are evaluated here.

Comparing the Mnova output of figure 6 with the NMR-Onion results of figure 4, it is evident that the resulting residuals of both programs do not represent a normal distributed white noise pattern (see more in the discussion section). The results are summarized in table 8, counting the number of detected peaks in each ROI (each region Number has been marked with a letter corresponding to the sub-plot numbering of figure 4 and 6).

We note that when comparing the number of peaks detected (see table 8), NMR-Onion and MNOVA mostly detect the same number of peaks, though it seems MNOVA detects peaks with negative amplitudes as well (one example seen in 6 A). Despite many similarities, there are some differences, one example being the underlying signals of the peaks found around 7.24 where the signal is more highly resolved (more peaks are detected) in NMR-Onion than MNOVA. Note that, the peaks in this particular region are marked as PRP, indicating that these may have low repeatability and would have occurred due to sample-to-sample variations (see more

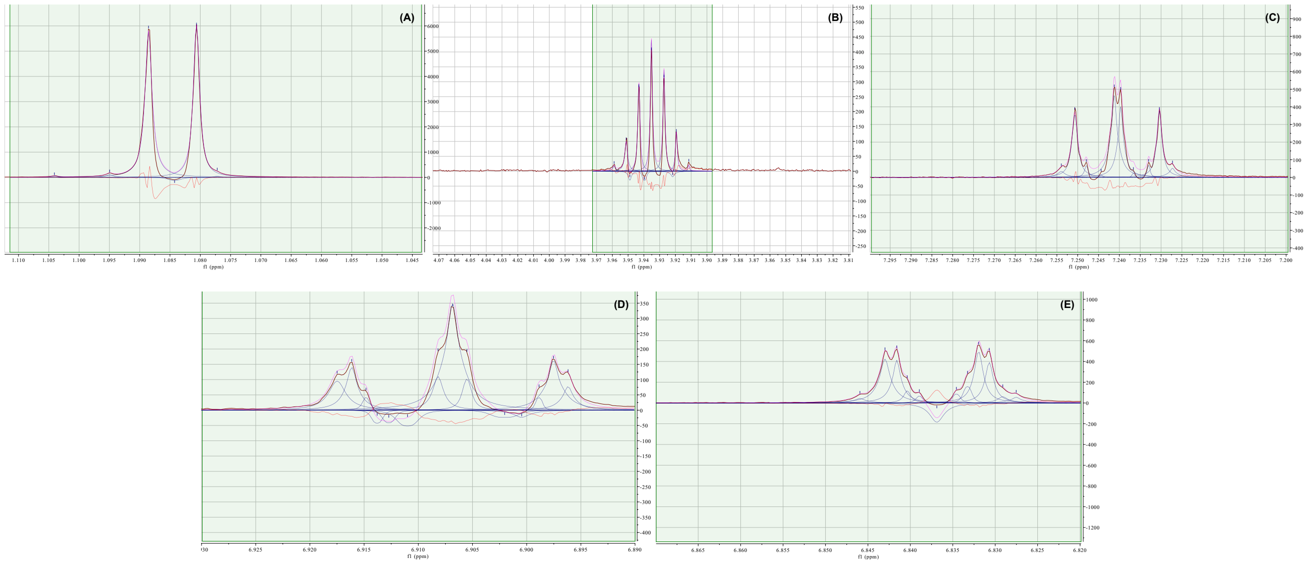


Figure 6: Visual model deconvoultion of sample 5 found in table 1 and 4. The dark red lines indicate the original spectrum, the purple lines are the fitted spectrum, the blue line underlying signals, the orange lines are the residuals and the black ticks are detected peaks. (A) Zoom in on the targeted doublet of Region 1. (B) Zoom in on the targeted septet of Region 2. (C) Zoom in on the targeted doublet of doublet region 4. (D) Zoom in on the triplet of triplets sub-part of region 3. (E) zoom in on the doublet sub-region of region 3.

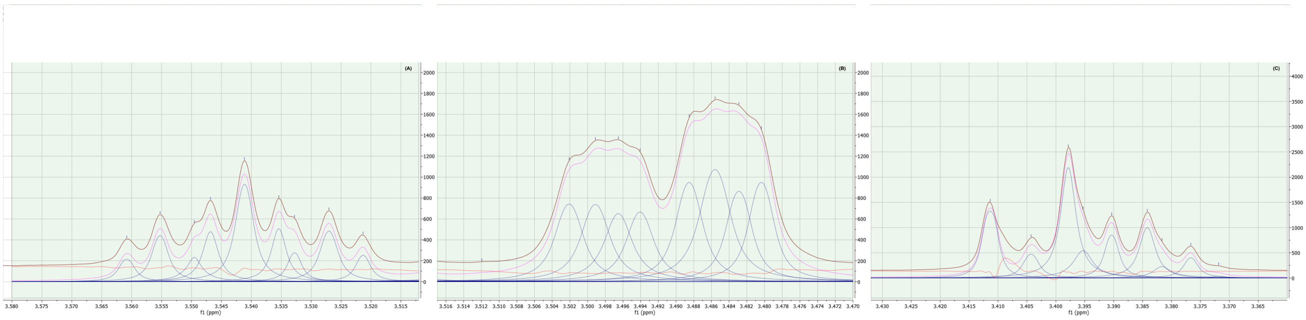


Figure 7: Visual model deconvoultion of sample 1 found in table 7 and plotted in figure 5 A,B and C. The dark red lines indicate the original spectrum, the purple lines are the fitted spectrum, the blue line underlying signals, the orange lines are the residuals and the black ticks are detected peaks. (A) zoom in on sub-region 1. (B) zoom in on sub-region 2. (C) zoom in on sub-region 3.

Table 8
Experiment 1, MNOVA and NMR-onion comparisons of sample 5

Region No.	No. Peaks(MNOVA)	No. Peaks(NMR-Onion)
1(A)	5	4
2(B)	9	9
3(D)	14	9
3(E)	12	10
4(C)	10	14
Total: -	50	46

in the discussion section).

For the second experiment, the results of MNOVA are plotted in figure 7. Generally from visual inspection, MNOVA and NMR-Onion are mostly consistent in their results for sub-region 1 and 2, but within the third sub-region (figure 5 C and 7 C) differences are identified, as NMR-onion detects a highly overlapping peak shoulder around 3.4 ppm whilst Mnova does not. In addition, the models of both MNOVA and NMR-Onion seem to be much closer to fulfilling the model assumptions of white noise, than in the first experiment (see more in the discussion).

5. Discussion

5.1. Case Study 1

From the results of section 4.3, regarding the first experiment, it was in general observed that MNOVA and NNR-Onion detect an almost identical amount of peaks across all the range of SNR set up in 4, but in some cases, NMR-Onion is capable of detecting more peaks than MNOVA, as exemplified in figure 4 C. Furthermore, we observed from table 5, that our novel models of equation(8) and (9) generally

outperformed the traditional model of pure exponential decay, making it suitable for fitting non-Lorentzian line shape. Though our models are time domain based, we have not shown the results in the time domain, in this paper, primarily due to two reasons. One reason is that NMR is typically evaluated in the frequency domain. The second reason was that we had to rely on visual comparisons between software results, which is difficult in the time domain as the fit is much more convoluted (The time domain output can be generated if needed and is done so in the tutorial of NMR-Onion at the GitHub site - see supporting information).
Another result of interest is outlined in table 6 where ROI 3 was shown to have many confidence interval overlaps suggesting that the peaks found around the targeted resonances may be potentially resolved peaks and therefore should be further investigated for consistency within independent replicates. However, as presented in the results, the presence of the targeted peaks was consistent across all samples, but these had different underlying multiplet structure. We believe this is due to the second-order effects occurring particularly in ROI 3. In addition to the many PRPs in ROI 3, ROI 4 also indicated a high amount of PRPs. The detailed pinpointing of consistent PRPs were not pursued in detail for this study. However, it may be addressed by adding more replicates at the same concentration, revealing which peaks are consistent but highly overlapping and which are occurring due to the sample-to-sample variations.

Finally, it should be noted that in the first case study, the residuals of both MNOVA and NMR-Onion are far from fulling the model assumptions of having white noise. We believe this to be caused by the fact that the data is very imperfect with respect to model formulations, occurring both in MNOVA and NMR-Onion (non-flat baseline) and very little preprocessing has been carried out. The reasoning behind the imperfections was to investigate how our approach could handle complex data without correction, which is based on none automated procedures, heavily relying on the operator. Interestingly, the minimal preprocessing scheme is applied in the second experiment and here residuals are much in line with the model assumptions which may be owed to a higher SNR.

5.2. Case Study 2

For the second experiment, the results of section 4.3 and 4.2, revealed that one PRP was identified within the first sample (figure 5 C), as the overlap was to present within the replicate (figure 5 F), indicating that this particular peak might have occurred due to sample variations. In addition, the results from MNOVA and NMR-Onion were very similar for the most part, but the last sub-region (3.36-3.43 ppm) did indicate a difference in detection, as NMR onion was able to detect a peak shoulder around 3.4 ppm which MNOVA could not (see figure 7 C vs figure 5 C). Finally, it should be noticed that figure 5 F does not have as many peaks detected as figure 5 C, which is clearly shown in the residuals indicating that some peaks are missing. It is possible to detect these peaks by decreasing the peak width cutoff, but we chose not to do this, as having the same parameters for each experiment makes much more sense when comparing output.

The last observation to address is that in case study 2, the residuals are much closer to fulfilling the model assumptions in this study than in the previous case study, but it should be noted that perfect white noise is achieved in neither MNOVA nor NMR-Onion. This makes sense, as this is real data and one cannot create a perfect model accounting for every type of distortion without introducing severe overfitting. Still, we have managed to formulate a model and a framework capable of accurately representing a spectrum in which residual signals are very small without containing many obviously missed peaks. A possible improvement could be introducing that of a random effect to the model, resulting in a non-linear mixed effect type of model that might be capable of capturing random distortions. This has to the best of our knowledge never been done and could potentially capture stochastic sample variation.

5.3. The NMR-Onion Algorithm

An essential aspect of the NMR-Onion algorithm comes from the ability to detect potential resolved peaks via overlapping CIs based on the wild bootstrap method (see algorithm 1). The downside of this method comes from the high computational time required as the model essentially has to be refitted 1000 times (default value) or more. This challenge was resolved by decimating the time series signal[44] which works given that proper initial parameter values from the fit of the none decimated ROI was estimated prior to executing the bootstrap. Alternative methods to the wild bootstrap algorithm (see algorithm 1) does exist, these are based on a Bayesian approach, however as addressed by Wilson[3] the sampling schemes are often much slower for pure MCMC approaches. Therefore, one might consider a variational Bayesian (VB) inference sampling scheme[45] as a possible alternative to the wild bootstrap and model fitting, though how to specify appropriate priors for the parameters should be considered. The Zellener prior has shown good results for fitting sinusoids in general as seen in the works by Rubtsov and Griffth [34] and could gain efficient speed with a VB inference. Another essential aspect of the NMR-Onion algorithm lies in the implementation of the models and optimization routine utilizing the modern framework of PyTorch. One of the main advantages of applying Pytorch comes from the automatic differentiation (AD) properties, that is when defining a loss function such as equation (15) the Hessian and gradient are automatically optimally defined making optimization much faster and more robust. Hence AD enables robust models to be developed much easier as one does not have to manually implement the derivatives. We believe that in tandem with the peak detection and digital filter modules, other models may be easily implemented and tested using the Pytorch core optimization framework of NMR-Onion making both time and frequency domain model development much more accessible for all developers. We attempted other non-quasi-newton approaches such as

ADAM [46] and RSM-prop [47] algorithms, but these ultimately failed (results not included) compared to that of the LBFGS both in Pytorch [30] and Scipy[31].
In making the optimization routine using LBFGS, computationally feasible, alternatives to the digital band-pass filter were also attempted. Inspired by deep learning methods, we attempted mini-batch stochastic optimization, which is capable of handling much larger data sets. However, this did not show promising results on real data or simulated data. We believe this was largely due to the LBFGS algorithm of Pytorch not being able to properly handle the mini-batch implementation rather than the method itself. As for the simulated data, the non-quasi-newton approaches (Adam and RSM-prop) worked properly with mini-batches for simulated data but failed for the real data. Hence, an attractive improvement to our algorithm would be to implement mini batches when the LBFGS algorithm of Pytorch is further developed with stochastic optimization, as this would make manual ROI selection obsolete, fitting the full spectrum, all at once. In the current state of NMR-Onion, the multiplets have to be manually assigned based on the estimated amplitude ratios and coupling constants. This can be problematic, especially in untargeted studies. Hence for the future of NMR-Onion, the implementation of automatic assignment of multiples based on amplitude ratios and coupling constants is a desired feature making NMR-Onion suited for faster targeted as well as untargeted studies.

6. Conclusion

From the results and discussion, it can be concluded that the NMR-Onion framework is a robust tool for analyzing 1D ^{1}H NMR spectra, capable of targeting specific ROIs within a spectrum for targeted analysis at a wide range of SNR values. Additionally, we conclude that our new novel time domain models are capable of fitting and detecting highly overlapping signals. We believe that with the NMR-Onion being open source, model improvements and further development can be rapidly added due to the AD library, while the core modules of digital filtering ensure computational feasibility and the peak detection algorithm enabling multi-modality of the frequencies to be handled. Furthermore, it can be concluded that the detection of potentially resolved peaks in combination with replicates will ensure that the risk of false conclusions will be reduced significantly. This would be very relevant for large metabolomics samples where many signal overlaps are present and sample-to-sample variations would potentially play a significant role. With the NMR-Onion algorithm one would be made aware of the potential artifacts and hence draw fewer false conclusions. With NMR-Onion we have built an algorithm capable of statistically evaluating the uncertainties of the results in a manner such that the user will become aware of potentially resolved peaks appearing, which coupled with replicates, would aid in reassuring that highly overlapping peaks are consistent throughout samples and not occurring due to sample to sample variations.

7. Acknowledgements

We would like to thank Danish National Research Foundation for the Center for Microbial Secondary Metabolites Grant Number: DNRF137 for funding this project and also thank the NMR Center • DTU and the Villum Foundation are acknowledged for access to the 800 MHz spectrometers.

8. Conflict of interest

All authors declare that they have no conflicts of interest.

9. Supporting Information

Github: `https://www.github.com/Mabso1/NMR-onion`. Additional information can be found in the online version of the article

References

[1] A. H. Emwas, E. Saccenti, X. Gao, R. T. McKay, V. A. P. M. dos Santos, R. Roy, D. S. Wishart, Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine, Metabolomics 14 (3) (2018) 31. doi:10.1007/S11306-018-1321-4.

[2] G. L. Bretthorst, C. C. Hung, D. A. D'Avignon, J. J. H. Ackerman, Bayesian analysis of time-domain magnetic resonance signals, J. Magn. Reson. (1969) 79 (2) (1988) 369–376. doi:10.1016/0022-2364(88)90233-8.

[3] A. G. Wilson, Y. Wu, D. J. Holland, S. Nowozin, M. D. Mantle, L. F. Gladden, A. Blake, Bayesian inference for nmr spectroscopy with applications to chemical quantification, arXiv:1402.3580: Applications (2014) doi:10.48550/arXiv.1402.3580.

[4] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, J. C. Wilson, Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform, Chemom. Intell. Lab. Syst. 85 (1) (2007) 144–154. doi:10.1016/j.chemolab.2006.08.014.

[5] S. A. A. Sousa, A. Magalhães, M.M.C. Ferreira, Optimized bucketing for nmr spectra: Three case studies, Chemom. Intell. Lab. Syst. 122 (2013) 93–102. doi:10.1016/j.chemolab.2013.01.006.

[6] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, W. Van Criekinge, NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm, Anal. Chem. 80 (10) (2008) 3783–3790. doi:10.1021/AC7025964.

[7] C. Piras, M. Pibiri, V. P. Leoni, F. Cabras, A. Restivo, J. L. Griffin, V. Fanos, M. Mussap, L. Zorcolo, L. Atzori, Urinary 1H-NMR metabolic signature in subjects undergoing colonoscopy for colon cancer diagnosis, Appl. Sci. (Switzerland) 10 (16) (2020) 5401. doi:10.3390/APP10165401.

[8] F. Probert, V. Ruiz-Rodado, D. Te Vruchte, E. R. Nicoli, T. D. W. Claridge, C. A. Wassif, N. Farhat, F. D.

Porter, F. M. Platt, M. Grootveld, NMR analysis reveals significant differences in the plasma metabolic profiles of Niemann Pick C1 patients, heterozygous carriers, and healthy controls, Sci. Rep. 2017 7 (1) (2017) 6320. doi:10.1038/s41598-017-06264-2.

[9] F. M. M. Ocampos, L. R. A. Menezes, L. M. Dutra, M. F. C. Santos, S. Ali, A. Barison, NMR in chemical ecology: An overview highlighting the main NMR approaches, eMagRes 6 (2) (2017) 325–342. doi:10.1002/9780470034590.EMRSTM1536.

[10] R. X. Poulin, G. Pohnert, Simplifying the complex: metabolomics approaches in chemical ecology, Anal Bioanal Chem. 411 (1) (2019) 13–19. doi:10.1007/S00216-018-1470-3.

[11] U. K. Sundekilde, N. Eggers, H. C. Bertram, NMR-Based Metabolomics of Food, Springer New York, G. A. N. Gowda and D. Raftery (eds), 2019. doi:10.1007/978-1-4939-9690-2_18.

[12] M. Cuperlovic-Culf, A. S. Culf, Applied metabolomics in drug discovery, Expert Opin. Drug Discovery 11 (8) (2016) 759–770. doi:10.1080/17460441.2016.1195365.

[13] K. Krishnamurthy, CRAFT (complete reduction to amplitude frequency table) – robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR, Magn. Reson. Chem. 51 (12) (2013) 821–829. doi:10.1002/MRC.4022.

[14] J. Keeler, Understanding NMR Spectroscopy, John Wiley & Sons, Ltd, 2010, p. 526.

[15] I. Marshall, J. Higinbotham, S. Bruce, A. Freise, Use of Voigt lineshape for quantification of in vivo 1H spectra, Magn. Reson. Med. 37 (5) (1997) 651–657. doi:10.1002/MRM.1910370504.

[16] M. Niklasson, R. Otten, A. Ahlner, C. Andresen, J. Schlagnitweit, K. Petzold, P. Lundström, Comprehensive analysis of NMR data using advanced line shape fitting, J. Biomol. NMR 69 (2) (2017) 93–99. doi:10.1007/s10858-017-0141-6.

[17] Mestrelab, Global Spectral Deconvolution (GSD) - Mestrelab Resources (2017).
URL https://resources.mestrelab.com/gsd/

[18] S. Sokolenko, T. Jézéquel, G. Hajjar, J. Farjon, S. Akoka, P. Giraudeau, Robust 1D NMR lineshape fitting using real and imaginary data in the frequency domain, J. Magn. Reson. 298 (2019) 91–100. doi:10.1016/J.JMR.2018.11.004.

[19] J. Hao, M. Liebeke, W. Astle, M. De Iorio, J. Bundy and T. M, D. Ebbels, Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN, Nat. Protoc. 9 (6) (2014) 1416–1427. doi:10.1038/nprot.2014.090.

[20] S. Ravanbakhsh, P. Liu, T. C. Bjorndahl, R. Mandal, J. R. Grant, M. Wilson, R. Eisner, I. Sinelnikov , X. Hu , C. Luchinat , R. Greiner, D. S. Wishart, Accurate, fully-automated NMR spectral profiling for metabolomics, PLoS One 10 (5) (2015) e0124219. doi:10.1371/journal.pone.0124219.

[21] P. Mercier, M. J. Lewis, D. Chang, D. Baker, D. S. Wishart, Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra, J. Biomol. NMR 49 (3-4) (2011) 307–323. doi:10.1007/S10858-011-9480-X.

[22] A. Savitzky, M. J. E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (8) (1964) 1627–1639. doi:10.1021/ac60214a047.

[23] N. Narisetty, Bayesian model selection for high-dimensional data, in: A. S. S. Rao, C. Rao (Eds.), Handbook of Statistics, Vol. 43, Elsevier, 2020, pp. 207–248. doi:10.1016/bs.host.2019.08.001.

[24] E. Mammen, Bootstrap and Wild Bootstrap for High Dimensional Linear Models, The Ann. of Statist. 21 (1) (1993) 255 – 285. doi:10.1214/aos/1176349025.

[25] Y. Matviychuk, E. von Harbou, D. J. Holland, An experimental validation of a Bayesian model for quantification in NMR spectroscopy, J. Magn. Reson. 285 (2017) 86–100. doi:10.1016/J.JMR.2017.10.009.

[26] R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, 1985, p. 662. doi:10.1017/CBO9780511810817.

[27] E. H. Djermoune, M. Tomczak, D. Brie, NMR data analysis: A time-domain parametric approach using adaptive subband decomposition, Oil & Gas Science and Technology - Revue d'IFP Energies nouvelles 69 (2) (2014) 229–244. doi:10.2516/OGST/2012092.

[28] S. G. Hulse, M. Foroozandeh, Newton meets Ockham: Parameter estimation and model selection of NMR data with NMR-EsPy, J. Magn. Reson. 338 (2022) 107173. doi:10.1016/J.JMR.2022.107173.

[29] S. J. Baek, A. Park, Y. J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (1) (2014) 250–257. doi:10.1039/C4AN01061B.

[30] A. Paszke, LBFGS — PyTorch 1.12 documentation (2022).
URL https://pytorch.org/docs/stable/generated/torch.optim.LBFGS

[31] D. M. Cooke, minimize(method='L-BFGS-B') — SciPy v1.9.0 Manual (2004).

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, NIPS 2017 Workshop on Autodiff.

[33] R. Hooke, T. A. Jeeves, "Direct Search" Solution of Numerical and Statistical Problems, Journal of the ACM (JACM) 8 (2) (1961) 212–229. doi:10.1145/321062.321069.

[34] D. V. Rubtsov, J. L. Griffin, Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy, J. Magn. Reson. 188 (2) (2007) 367–379. doi:10.1016/J.JMR.2007.08.008.

[35] M. U. A. Bromba, H. Ziegler, Application hints for Savitzky-Golay digital smoothing filters, Anal. Chem.

53 (11) (1981) 1583–1586. doi:10.1021/ac00234a011.

[36] R. MacDonald, S. Sokolenko, Detection of highly overlapping peaks via adaptive apodization, J. Magn. Reson. (Calif. 1997) 333 (2021) 107104. doi:10.1016/J.JMR.2021.107104.

[37] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723. doi:10.1109/TAC.1974.1100705.

[38] G. Schwarz, Estimating the dimension of a model, The Ann. of Statist. 6 (2) (1978) 461–464. doi:10.1214/AOS/1176344136.

[39] B. Nadler, L. A. Kontorovich, Model selection for sinusoids in noise: Statistical analysis and a new penalty term, IEEE Trans. Signal Process 59 (4) (2011) 1333–1345. doi:10.1109/TSP.2011.2105482.

[40] P. M. Djurić, A model selection rule for sinusoids in white gaussian noise, IEEE Trans. Signal Process 44 (7) (1996) 1744–1751. doi:10.1109/78.510621.

[41] A. Wald, Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large, Trans. Amer. Math. Soc. 54 (3) (1943) 426. doi:10.2307/1990256.

[42] TopSpin | NMR Data Analysis | Bruker (2023). URL https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html

[43] J. J. Helmus, C. P. Jaroniec, Nmrglue: An open source Python package for the analysis of multidimensional NMR data, J. Biomol. NMR 55 (4) (2013) 355–367. doi:10.1007/S10858-013-9718-X.

[44] R. Crochiere, L. Rabiner, Interpolation and decimation of digital signals - A tutorial review, Proceedings of the IEEE 69 (3) (1981) 300–331. doi:10.1109/PROC.1981.11969.

[45] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational Inference: A Review for Statisticians, J. Am. Stat. Assoc. 112 (518) (2016) 859–877. doi:10.1080/01621459.2017.1285773.

[46] D. P. Kingma, J. L. Ba, Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. doi:10.48550/arxiv.1412.6980.

[47] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural networks for machine learning 4 (2) (2012) 26–31.

Transparent multi model-based 1D NMR deconvolution: Supplementary 1

Mathies Brinks Sørensen[a], Michael Riis Andersen[b], Mette-Maya Siewertsen[a], Rasmus Bro[c], Mikael Lenz Strube[d] and Charlotte Held Gotfredsen[a,*]

[a]*Department of Chemistry, Technical University of Denmark, Kgs Lyngby, DK-2800, Denmark*
[b]*Department of Applied Mathematics and Computer Science, Kgs Lyngby, DK-2800, Denmark*
[c]*Department of Food Science, University of Copenhagen, Frederiksberg, DK-1958, Denmark*
[d]*Department of Biotechnology and Biomedicin, Kgs Lyngby, DK-2800, Denmark*

1. Constraining specifications

The numerical challenges of fitting the parameters of every model of the NMR-Onion paper are listed here and the solution to the challenge is provided. The models of the paper are outlined below for a better overview

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_1(\boldsymbol{\rho_k}) \cdot exp(1j\gamma_k)t \tag{1}$$

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_2(\boldsymbol{\rho_k}) \cdot exp(1j\gamma_k)t \tag{2}$$

$$y(t) = \sum_{k=1}^{K} f(A_k, \omega_k, \phi_k) \cdot \Psi_3(\boldsymbol{\rho_k}) \cdot exp(1j\gamma_k)t \tag{3}$$

Where $f(A_k, \omega_k, \phi_k)$ and $\Psi(\rho_k)$ are the harmonic and decay terms of each model. These are formulated as

$$f(A_k, \omega_k, \phi_k) = A_k \cdot exp(2j\pi\omega_k t + \phi_k) \tag{4}$$

$$\Psi_1(\boldsymbol{\rho_k}) = exp(-\alpha_k t) \tag{5}$$

$$\Psi_2(\boldsymbol{\rho_k}) = (1 - \eta_k) \cdot exp(-\alpha_k t) + \eta_i \cdot exp(-\alpha_k t^2) \tag{6}$$

$$\Psi_3(\boldsymbol{\rho_k}) = exp(-\alpha_k t^{\beta_k}) \tag{7}$$

The first challenge comes from constraining parameters to be positive, particularly the decay rate (α), power law constant (β), and mixing constant (η) are all positive parameters. To avoid numerical overflows in the optimization routine the softplus function (SPF) is applied

$$SPF(\rho_k) = \frac{1}{\beta_{sp}} \log(1 + \exp(\beta_{sp}\rho_k)) \tag{8}$$

where β_{sp} is set to 1 per default and the linearity threshold was set as $\beta_{sp}\rho_k > 20$. We found that applying the SPF worked well for the power law constant of β and also partially for the decay rate, but additional twists had to be made for the decay rates not rely heavily on initial values. For the initial values of β we set $\beta_{start} = 0.54$ which corresponds to $SPF(\eta_{start}) \approx 1$. Prior to showing the decay rate constraining method, we introduce the logistic sigmoid function (LSF) which ensured the stability of the mixing constant and later the decay rate as well. From equation (6) it is evident that $0 \geq \eta \geq 1$ is a hard constraint for the model. To avoid the problems of having hard constraints on parameters, we transformed η via the LSF such that the transformed space is formulated as

$$LSF(\eta) = \frac{1}{1 + \exp(-\eta)} \tag{9}$$

Equation (9) constrains the parameters such that the function values are always between 0 and 1 no matter how large or small the values of η. For the initial values of η we set $\eta_{start} = 0.0$ which corresponds to $LSF(\eta_{start}) = 0.5$.

The decay parameter was not as simple to constrain, as peaks with very different heights would have very different decay rates. Hence, we combined the SPF with the LSF resulting in the transformation function Q seen in equation (10)

$$Q(\alpha_k, S) = SPF(S)LSF(\alpha_k) \tag{10}$$

The Q transformation of equation (10) works such that each decay rate is constrained between 0 and 1 and is scaled by a positive scalar S. The value of S would then represent the peak with the highest decay rate which is scaled down by $LSF(\alpha_k)$ for the remaining peaks. For initial values of S and α we set $S_{start} = 0.0$ and $\alpha_{start} = 0.0$. We attempted different values and found that setting S_{start} to low values between $[SP(-1) : SP(1)]$ worked well for all tested datasets and therefore we chose the middle value of $SP(0.0)$. Likewise, we also choose the middle value of the LSF at $LSF(\alpha_{start}) = 0.5$. Finally, for the skewing term of γ we set the initial value at $\gamma_{start} = 0.0$ no special transformation was needed as the parameter seemed to be self-constrained between $[\frac{-\pi}{2} : \frac{\pi}{2}]$ regardless of data input.

[*] Corresponding author.
E-mail address: chg@kemi.dtu.dk (Charlotte Held Gotfredsen).