# MSDeconvolve: A new metabolomics fragmentation spectra resolver using statistics and machine learning


**Nikolaos Terzis**

School of Mathematics and Statistics

University of Glasgow

A thesis submitted for the degree of

*Master of Research*

October 2024

# Abstract

In research fields such as drug discovery and biomarker discovery in diseases, it is often important to identify metabolites present in samples. Metabolites are the end products of metabolic processes in cells, and metabolomics is the study of metabolites. In metabolomics, a widely used approach to identify metabolites in samples is to run liquid chromatography tandem mass spectrometry (LC-MS/MS) experiments. In those experiments, metabolites are fragmented, which means that energy is applied to break the chemical bonds of the metabolite, and produce fragment ions with different masses. The pattern with which a metabolite is fragmented provides valuable information that can lead to its identification. However, a compromise needs to be made between fragmenting a low number of metabolites but having good quality fragmentation patterns (Data Dependent Acquisition, or DDA), or fragmenting all metabolites but obtaining fragmentation patterns from multiple metabolites combined (Data Independent Acquisition, or DIA). In DIA, a deconvolution algorithm is required to determine which metabolite each fragment comes from, and then use the deconvoluted fragmentation patterns to perform identification. The current state-of-the-art deconvolution algorithms do not perform at the required level to produce reliable fragmentation spectra for identification. MSDeconvolve is a framework attempting to improve deconvolution of fragmentation patterns, using statistical methods and models. More specifically, it constructs a design matrix from extracted features from experiment results, where each covariate corresponds to the intensity of the signal of a metabolite. The response variables are constructed from the intensity of the signal of fragment ions. Then, by fitting a model for each response variable against the same design matrix, we obtain coefficient estimates for each fragment ion and each metabolite. By interpreting these coefficients as the proportion of the metabolite that results in a fragment, we can reconstruct the fragmentation patterns of the individual metabolites. Lasso regression was initially used, given its variable selection property that is appropriate in this problem. However, multiple linear regression, Ridge and Elastic Net regression were also evaluated. Further extensions to this modelling approach

were also explored. These extensions include applying a penalty such that the coefficients of each metabolite add approximately up to one, to reflect the fact that the signal intensity of fragments for each metabolite should approximately add up to the signal intensity of the original metabolite. Another extension was the combination of data from both a DDA and DIA experiment, in order to improve the deconvolution of metabolites which were not fragmented in the DDA experiment. Although MSDeconvolve performs very well on simulated metabolomics data, it was able to perform similarly, if not marginally better, than the state-of-the-art algorithm on real data. However, future work could potentially further improve its performance by overcoming the limitations of metabolomics data.

# Contents

**Contents** iii

**List of Tables** v

**List of Figures** vi

**1 Introduction** 1

**2 Background** 4

2.1 Metabolomics and Mass Spectrometry . . . 4
2.2 Data Format . . . 7
2.3 Data Acquisition Methods . . . 8
2.3.1 Data-Dependent Acquisition (DDA) . . . 9
2.3.2 Data-Independent Acquisition (DIA) . . . 10
2.3.3 Contaminant Noise . . . 10
2.4 Data Preprocessing and Analysis . . . 12
2.5 Peak Picking . . . 14
2.6 MS-DIAL . . . 15
2.7 Data . . . 17
2.7.1 Simulated Data . . . 17
2.7.2 Real Data . . . 18

**3 Statistical Methods** 19

3.1 Linear Regression . . . 19
3.2 Penalised Regression . . . 20
3.3 Ridge Regression as a Bayesian Approach . . . 22
3.4 B-splines . . . 22
3.5 Gaussian Mixture Model . . . 24
3.6 Non-negative Coefficients . . . 24
3.7 Cross Validation . . . 25

3.8 Evaluation Metrics . . . 27

**4 MSDeconvolve** **29**
4.1 Model Building . . . 29
4.1.1 Using Peaks to Extract Features . . . 30
4.1.2 Constructing the Response Variable from MS2 Spectra . . . 32
4.1.3 Regression . . . 32
4.2 Predictions . . . 33
4.3 Evaluation . . . 34
4.4 MSDeconvolve Extensions . . . 35

**5 Results** **37**
5.1 Simulation Results . . . 37
5.1.1 Lasso and Penalised Regression . . . 38
5.1.2 Linear Regression . . . 41
5.1.3 Filtering . . . 43
5.1.4 MS-DIAL comparison . . . 47
5.2 Real Data Results . . . 47
5.2.1 Real Data Challenges . . . 48
5.2.2 Linear and Lasso Regression Performance . . . 50
5.2.3 Ridge Regression Performance with Baseline Correction and Correlation Filtering . . . 53
5.2.4 Ensemble Method . . . 59
5.2.5 Peak Coefficients Additional Constraints . . . 60
5.2.6 Scaling by MS2 Bin Approach . . . 62
5.2.7 DDA and DIA Data Combination . . . 65

**6 Discussion** **67**
6.1 Conclusions and Limitations . . . 67
6.2 Improvements and Future Work . . . 69

**Bibliography** **71**

# List of Tables

2.1 Mathematical notation summary . . . . . 9
2.2 Simulated noise levels . . . . . 18

# List of Figures


2.1 Real mass spectrum example . . . . . 5
2.2 Example of experiment result data with TOPPView visualisation. . . . . 6
2.3 Synthetic example of DIA deconvolution of fragmentation spectra. . . . . 11

3.1 Lasso, Ridge and Elastic Net penalty visualisation. . . . . 21
3.2 Synthetic example of B-spline functions and curve fitting. . . . . 23
3.3 Data split strategy of $k$-fold cross validation. . . . . 26

4.1 TOPPView example of an MS1 scan intersecting multiple peaks. . . . . 30
4.2 TOPPView example of noise within an MS1 peak's boundaries. . . . . 31

5.1 Distribution of number of chemicals that are fragmented in the same MS2 scan in the five simulated samples with different number of chemicals. . . 38
5.2 Results of Lasso with cross validation on simulated data of 10, 100, 500, 1000 and 1500 chemicals. . . . . 39
5.3 Comparing the difference in cosine similarities of chemicals for the models using Lasso with all log transformation combinations, with 1500 chemicals and noise level 1 (Table 2.2). . . . . 40
5.4 Box plot comparison of performance of Lasso, Ridge and Elastic Net models on simulated datasets of noise level 1 (Table 2.2). . . . . 42
5.5 Results of Lasso and linear regression on the simulated datasets with no noise and 10 chemicals. . . . . 43
5.6 Results of linear regression (with and without log transformations) on dense simulated samples with no noise. . . . . 44
5.7 Comparison of multiple linear regression and Lasso regression on datasets with noise level 1 and noise level 2 (Table 2.2) with and without log transformations. . . . . 45
5.8 Box plot comparison of MSDeconvolve with and without noise filtering, on noise levels 1 and 2 (Table 2.2) using multiple linear regression. . . . . 46

5.9 Box plot comparison of the distribution of cosine similarities for MSDeconvolve using multiple linear regression and MS-DIAL. . . . . 48
5.10 MS1 contaminant example and peaks capturing it. . . . . 49
5.11 Heatmap of correlation matrix of peaks for the SWATH experiment of the real dataset, for the isolation window ranging from 170 m/z to 270 m/z. . 51
5.12 Box plots of cosine similarities of MS-DIAL and MSDeconvolve on the real dataset, with and without log transfromations and using linear regression. 52
5.13 Box plots comparing cosine similarity scores of MS-DIAL and MSDeconvolve modelling SWATH isolation windows separately and automatically filtering out covariates. . . . . 53
5.14 Box plots comparison of cosine similarity scores of MS-DIAL and MSDeconvolve for second SWATH isolation window of the real data, using linear regression, Ridge and Lasso. . . . . 54
5.15 MS2 bins with fragments of the contaminant examined in Figure 5.10. . . 55
5.16 Baseline correction example of an MS2 bin with a fragment of the contaminant of Figure 5.10. . . . . 56
5.17 Cosine similarity comparison of MSDeconvolve with and without baseline removal and MS-DIAL. . . . . 56
5.18 Comparison of predicted fragmentation spectrum against gold standard spectrum of specific peak for MSDeconvolve and MS-DIAL . . . . . 57
5.19 Partial residuals plots for model of MS2 bin with contaminant fragment. 58
5.20 Box plots of cosine similarities for MS-DIAL, and MSDeconvolve for models using Ridge and the correlation filter using different correlation threshold values. . . . . 59
5.21 Box plots of cosine similarity scores of ensemble methods and MS-DIAL. 61
5.22 Comparison of box plots of cosine similarity scores of MS-DIAL, MDeconvolve using Ridge regression, and MSDeconvolve using linear regression and the $[0, 1]$ bounds for the coefficients. . . . . 62
5.23 Box plots of cosine similarity scores for MS-DIAL and MSDeconvolve using linear regression, Ridge, and the sum to one constraint with different levels of regularization. . . . . 63
5.24 Box plot of cosine similarities of scaling approach versus MS-DIAL for all chemicals/peaks and the top 100 most intense peaks. . . . . 64
5.25 Box plots of models with combination of DDA and DIA information. . . 65

# Chapter 1

# Introduction

Metabolomics is the field studying molecular products of metabolism in cells. In combination with other fields such as proteomics, it can provide an analysis of chemical processes behind biological phenomena [2]. There are numerous applications and use cases of metabolomics, such as the discovery of biomarkers of diseases. This project focuses on attempting to improve the quality of data resulting from a specific type of experiment in metabolomics, which would in turn facilitate research conducted in this field.

Understanding the composition of a chemical sample and the identification of metabolites is often the main purpose in metabolomics experiments. Various methods are available to perform such a task, but in analytical chemistry, liquid chromatography (LC) coupled to tandem mass spectrometry (MS/MS) is a widely used method. Running such experiments is not a trivial task, with mass spectrometers being complex machines requiring expert knowledge to set up and maintain, and the data output of an experiment usually includes significant noise levels, making data processing difficult [30].

Before running such an experiment, a few options and settings need to be decided, usually depending on the sample and experiment goals. One of these options is the data acquisition mode; as metabolites from a sample pass through a mass spectrometer, there are a few possible strategies for making use of that information, each with advantages and disadvantages. In one of these data acquisition modes called Data-Dependent Acquisition (DDA), high quality data is collected for the identification of metabolites, but this is only for a small percentage of the total metabolites in the sample. On the other hand, in Data-Independent Acquisition (DIA) mode, a great coverage of metabolites in a sample is achieved. However in DIA mode, depending on the complexity of the sample, some of the metabolites' signals can be observed simultaneously, making it difficult to identify which metabolites each signal comes from [9, 10, 41]. As a result, identifying the chemical

compound of a metabolite becomes a more difficult and error-prone task. For this reason, deconvolution algorithms have been developed, in order to reconstruct the original signals from the multiplexed data and achieve high identification rates, with the most widely used deconvolution software being MS-DIAL [38].

In metabolomics DIA experiments, the currently available deconvolution algorithms are not yet performing at the required level to make it a trustworthy approach for the identification of metabolites in a sample [9, 10, 41]. In this project, we are aiming to develop statistical modelling techniques in order to achieve better deconvolution performance in metabolomics DIA LC-MS/MS experiments than the currently available state-of-the-art tools, where the aim of deconvolution is to accurately predict the fragmentation mass spectrum of each detected peak (more details introduced in Sections 2.1 and 2.3.2). The development process of such a tool is enhanced by the option of evaluating new deconvolution methods on simulated data [39, 40, 41], which is more accessible than real experiment data. Such a method would make DIA experiments more efficient and effective compared to other data-acquisition strategies, by inferring accurate fragmentation mass spectra for all metabolites in a sample.

In Chapter 2 we provide an overview of the background required for these experiments. We start by explaining in more detail the field of metabolomics and mass spectrometry, its importance and use cases, its limitations and the type of data we can obtain. We also compare different approaches to these experiments, giving emphasis on the data acquisition strategies. The two major categories of data acquisition strategies are DDA and DIA, and we explain the details of each and compare them. Although this project focuses on DIA experiments, it is important to be familiar with both, in order to have a better understanding of the problem we are addressing. Additionally, we mention common workflows and protocols used in the field, we explain what peaks are and the process of peak-picking, which is a very important step in the identification of metabolites. Finally, we look into the details of MS-DIAL which is the current state-of-the-art tool for deconvolution, and the data that has been used for evaluation in this project.

The work in this project heavily relies on statistical models and methods for the purpose of deconvolution, and thus in Chapter 3 we discuss the theory behind those. More specifically, we look into linear regression, penalised linear regression including Lasso [35], Ridge [14], and Elastic Net [45] regression and algorithms for enforcing non-negative coefficient constraints on these models. We also look at B-splines, cross validation, Gaussian mixture models and finally we provide an overview of the evaluation metric used to assess the quality of this project's methods.

In Chapter 4 we introduce the developed framework, MSDeconvolve, which makes use of the methods described in Chapter 3. This chapter describes how the data are processed,

organised and used in statistical models to obtain predicted fragmentation spectra, and how we use the evaluation metric to assess the quality of predictions. Finally, in Chapter 5 we present the results of applying our work on both simulated and real metabolomics data, and compare with MS-DIAL. Chapter 6 provides a conclusion of our work, with focus on the limitations and future work.

# Chapter 2

# Background

## 2.1 Metabolomics and Mass Spectrometry

Metabolomics is the field of attempted study of all metabolites in a system [17]. Many options are available for conducting metabolomics experiments. However, mass spectrometry is a popular choice for complex samples, allowing researchers to gain a better understanding of chemical processes and reactions in cells and organisms. Thus, it is a highly important and useful tool leading to research advancements of different kinds, such as disease diagnosis and understanding, drug discovery and oncology.

A mass spectrometer is one type of instrument that can be used to analyse metabolomics samples. They are complex machines, which make use of the physical properties of analytes to separate them by mass and charge. More specifically, analytes are given a charge through the process of ionisation (producing ions), which is necessary for the mass spectrometer to detect them. This results in a mass spectrum (Figure 2.1), a graph with two axes; the mass-to-charge ratio (m/z) axis and the intensity axis. The m/z axis represents the mass of ions divided by the acquired charge through ionisation, with the units of mass being daltons (where one dalton is equivalent to the mass of the one-twelfth of the carbon-12 atom when it is at rest and in its ground state [16]), and the charge is measured in absolute charge number (where a charge number of one is equivalent to the electric charge of a proton). The intensity axis represents the abundance of ions that are detected. Intensity is not an accurate measure of the absolute concentration of an analyte in a sample, due to a number of chemical and instrument factors that affect the final measurement. However, it does provide information about the relative abundance between analytes in a sample.

In complex samples, multiple metabolites can have the same m/z value, and thus m/z separation is usually not sufficient. For this reason, an extra separation step is

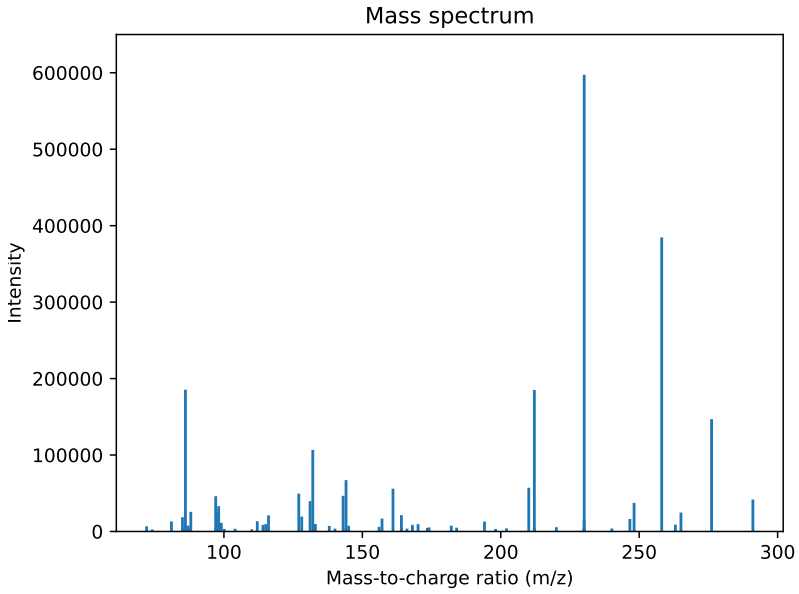


Figure 2.1: Real mass spectrum example obtained from beer data (introduced in Section 2.7.2). Each vertical line represents a detected ion, and its height its detected intensity.

performed. Popular choices are liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS), with LC-MS being more widely used in metabolomics experiments. These techniques allow metabolites to slowly be released into the mass spectrometer. This introduces a new dimension of time in the mass spectrum of Figure 2.1, which is called retention time (Figure 2.2). Depending on the chemical properties of a metabolite, it will be released into the mass spectrometer over a specific time, with its intensity values following a bell curve shape as it is being released. This is beneficial because it can separate metabolites with similar m/z values. Additionally, the retention time of a metabolite can provide information about the underlying chemical compound behind the signal.

Finally, there is a possible extension of LC-MS experiments, called liquid chromatography-tandem mass spectrometry (LC-MS/MS) experiments. In untargeted metabolomics experiments, the metabolites in a sample are not known prior to running the experiment. Therefore, it is often of interest to use the resulting data and signals to identify the metabolites present in the sample. However, this is not always possible from the m/z of an ion and its retention time alone, given also that retention time can vary between different experimental conditions and runs. Thus, a common strategy is to fragment an ion as it is being released into the mass spectrometer. This means that energy can be applied to break some of its chemical bonds, resulting in smaller fragment ions. This is displayed with a mass spectrum with the detected fragment ions. This spectrum with the

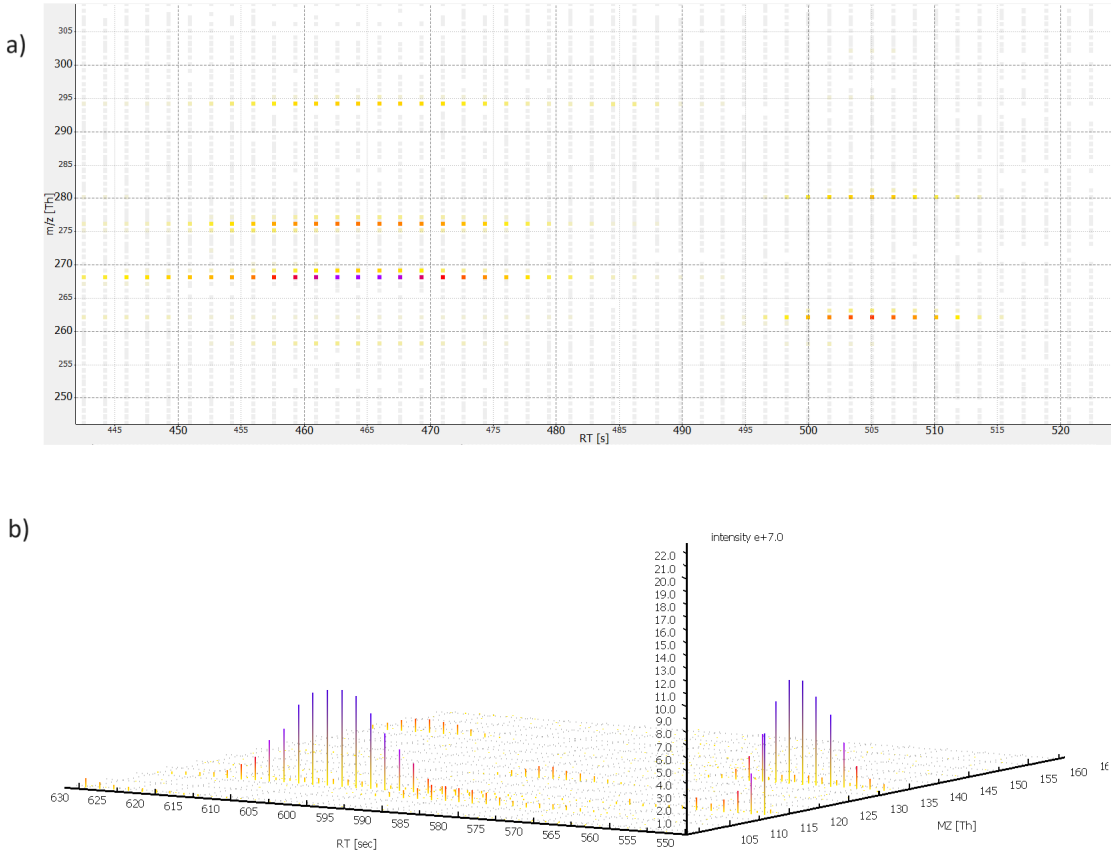


Figure 2.2: Example of experiment result data (zoomed in) taken from a beer sample. This is a screenshot from the TOPPView software [31], used to visualise the data. In screenshot a), the results are displayed as a two-dimensional plot, with the two axes being retention time and m/z, while the third dimension of intensity is encoded in the colour of data points. Screenshot b) shows similar data but in a three-dimensional plot. Metabolites usually appear in a bell curve shape, as they are being released from the chromatographic column into the mass spectrometer. As illustrated in the screenshots, each metabolite is only detectable during a specific retention time window.

fragmentation pattern provides valuable information about the ion that was fragmented (which is called the precursor ion) and can eventually lead to its identification, through the use of existing tools such as SIRIUS [7]. These type of scans are known as MS2 (level 2) or MS/MS scans. All other scans are called MS1 scans.

Another common issue in metabolomics experiments is the fact that there is not a one-to-one mapping between peaks and metabolites. Many ions can result from a single metabolite, for reasons such as the presence of isotopes, adducts and the unwanted fragmentation of metabolites during the ionisation process (at the MS1 scan level, before performing any intentional fragmentation through MS2 scans). Isotopes are two elements that have a different number of neutrons, but the same number of protons and electrons. Although isotopes share similar chemical properties, due to the different number of neutrons in the nucleus of the element, they have different mass. A common naturally occurring isotope is carbon-13. The carbon-13 atom has six protons and electrons, but seven neutrons instead of six. Therefore, its mass will be greater by approximately one dalton. For instance, the metabolite creatine has four carbon atoms, each of which could be a carbon-13 isotope. Therefore, in a LC-MS experiment, creatine can appear as multiple parallel peaks with a m/z difference of approximately one dalton. Depending on the frequency of isotope elements, peaks will have different abundances, with the peak with the lowest m/z ratio value being the most abundant in our example (since carbon-12 is more naturally abundant compared to carbon-13).

Adducts on the other hand can form during the ionisation process. Ionisation occurs by attaching a proton to metabolites when the experiment is run in positive mode, and by removing a proton from the metabolite when it is run in negative mode. However, the formation of different ions is possible. Common adducts are formed by the attachment of sodium or potassium elements instead of a proton. This results in different ion masses, which in turn results in different peaks showing up in the mass spectrometer. Finally, during the ionisation process, it is also possible for metabolites to get fragmented. This is a more difficult issue to deal with compared to isotopes and adducts. In the case of isotopes and adducts, there is an expected difference in mass between peaks. Fragmentation is a more unpredictable process, given that we would need to know the chemical structure of the metabolite in order to identify possible fragments, and even then, it is possible for a metabolite to have many different possible fragments.

## 2.2 Data Format

Due to the complex structure of mass spectrometry data, some preprocessing steps are required before they can be used for model formulation, which will be explored in Section

4.1. In this section, we will explore the format with which data is stored, and introduce some initial notation that is necessary to explain for the modelling process.

The results of an experiment are usually stored as a sequence of scans in the order they took place. Each scan has a retention time and an MS-level. In LC-MS/MS experiments there are two levels, MS1 and MS2 scans. Therefore, we can denote all MS1 scans as $\phi_r$ and all MS2 scans as $\psi_h$, with $r = 1, \ldots, R$ and $h = 1, \ldots, H$, and $R$ and $H$ being the total number of MS1 and MS2 scans respectively. We can refer to the retention time of a scan as $rt^{(\phi_r)}$ or $rt^{(\psi_h)}$ for MS1 and MS2 scans respectively. MS1 and MS2 scans are mixed in order, and only one scan can happen at a time. Therefore, the retention time of each scan will be different.

Each scan can be thought of as a mass spectrum, and thus it consists of pairs of m/z and intensity values. This is true for both MS1 and MS2 scans, and therefore they can share the same format, despite the difference of the data they hold. These pairs can be represented as two vectors of equal length, one for the observed m/z values, denoted as $\boldsymbol{m}^{(\phi_r)}$, and one for the observed intensity values, denoted as $\boldsymbol{\lambda}^{(\phi_r)}$ for MS1 scans (for example, the mass spectrum in Figure 2.1 consists of 62 vertical lines, translating to two vectors, each of length 62). For MS2 scans, the same notation is used, replacing $\phi_r$ with $\psi_h$. Each MS2 scan is linked to a specific MS1 scan (usually the most recent one). In the case of DDA experiments, each MS2 scan is also linked to a particular m/z region of its associated MS1 scan (which will be further explained in Section 2.3.1).

The number of MS1 and MS2 scans, $R$ and $H$ respectively, depend on the data acquisition mode of the experiment (which will be analysed in Section 2.3) as well as some of the machine settings. A summary of the notation introduced in this section is provided in Table 2.1.

## 2.3 Data Acquisition Methods

In LC-MS/MS experiments, scans can either be of MS1 or MS2 level. As ions are slowly released into the mass spectrometer, a choice has to be made between the two levels, since they cannot be performed simultaneously. Additionally, some time is required to perform each scan. Therefore, a strategy must be in place, to allow for a compromise between the advantages and disadvantages of each type of scan at each given time point.

An MS1 scan is valuable because it provides information about all detectable ions and their abundance in the sample at a given time point. However, if we wish to identify the chemical compound behind a signal, we need to perform an MS2 scan to capture its fragmentation pattern. In our project, we focus on Data-Independent Acquisition (DIA) modes, where the MS2 scans may include the fragmentation pattern of multiple

Table 2.1: Mathematical notation summary

| Notation | Description |
|---|---|
| $R$ | Total number of MS1 scans |
| $r = 1, \ldots, R$ | MS1 scan index |
| $H$ | Total number of MS2 scans |
| $h = 1, \ldots, H$ | MS2 scan index |
| $\phi_r$ | MS1 scan at index $r$ |
| $\psi_h$ | MS2 scan at index $h$ |
| $rt^{(\phi_r)}$ | Retention time (in seconds) of scan $\phi_r$ |
| $rt^{(\psi_h)}$ | Retention time (in seconds) of scan $\psi_h$ |
| $\boldsymbol{m}^{(\phi_r)}, \boldsymbol{\lambda}^{(\phi_r)}$ | m/z and intensity pair of vectors (of equal length) of scan $\phi_r$ |
| $\boldsymbol{m}^{(\psi_h)}, \boldsymbol{\lambda}^{(\psi_h)}$ | m/z and intensity pair of vectors (of equal length) of scan $\psi_h$ |
| $m_w^{(\phi_r)}, \lambda_w^{(\phi_r)}$ | m/z and intensity value pair of scan $\phi_r$ at index $w$ |

metabolites. The other option is Data-Dependent Acquisition (DDA) modes, which target and obtain MS2 scans with the fragmentation pattern of a single metabolite. Performance comparisons between the two have already been performed [9, 10, 41], and they will be explained in Sections 2.3.1 and 2.3.2.

### 2.3.1 Data-Dependent Acquisition (DDA)

In Data-Dependent Acquisition mode, the choice between MS1 and MS2 scans is performed based on the available data at any given time point. For instance, in a specific strategy called Top-N, an MS1 scan is performed to capture the current mass spectrum. Then, the m/z values where the top $N$ intensities are observed are selected in real-time (with $N$ being a positive integer parameter defined prior to running the experiment), and the next $N$ (or at most $N$, depending on whether there are enough ions) scans performed are MS2 scans, fragmenting the ions at the selected m/z values. This process is then repeated until the end of the experiment. This results in MS2 fragmentation spectra for selected ions. These can then be used to search a database of fragmentation patterns for known metabolites to identify them [2, 7, 8, 21, 42].

Although Top-N is a very popular choice as a data-acquisition mode, it has some drawbacks. Mainly, it is possible for many ions to not be fragmented (see Davies et al. [4] for more details). For instance, during the time required to obtain MS2 fragmentation spectra for some selected m/z values, a different ion might be released into the mass spectrometer, and by the time an MS1 scan is performed again, that ion is no longer detectable. Another drawback is that the time frame during which an ion is detectable varies, and therefore ions that are being released slower than others may be fragmented

multiple times, obtaining the same fragmentation pattern in each iteration. This is not ideal, because the time required to perform these repeated MS2 scans could be used to explore different ion signals. Adaptations to Top-N have been developed to overcome these limitations, such as the ones presented in Davies et al. [4] and McBride et al. [19].

### 2.3.2 Data-Independent Acquisition (DIA)

In data-independent acquisition methods, the sequence of MS1 and MS2 scans is already defined before running an experiment. The simplest DIA experiment type is All-Ion Fragmentation (AIF), where MS1 and MS2 scans are chosen alternatively, with MS2 scans fragmenting the entire m/z range. This means that MS2 mass spectra normally contain fragments from all metabolites eluting at that time, where some fragments could also have the same m/z value. Although this results in obtaining the fragments of all metabolites, we cannot determine which metabolite each fragment belongs to. This is the reason why deconvolution is a necessary step in DIA experiments, which aims at separating fragmentation patterns, and associate them with precursor ions, which can then be used for a database lookup. Figure 2.3 provides a visual example of the deconvolution process. Although this would in theory make DIA approaches a much more favourable choice compared to DDA, deconvolution of multiplexed MS2 spectra does not always provide mass spectra of the necessary quality for identification of metabolites.

Another type of DIA experiment is the Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH) method [18]. In this approach, instead of fragmenting the entire m/z range (as in AIF), m/z windows (which are predefined) are fragmented in sequence after every MS1 scan. If, for example, the m/z range is split in five windows, every MS1 scan is followed by five MS2 scans, one for each SWATH window, and then this pattern is repeated. Each window has a lower and upper m/z value, within which all ions are fragmented. This results in less complicated MS2 spectra compared to AIF, since fragments from fewer metabolites are observed in MS2 scans, making deconvolution an easier task.

### 2.3.3 Contaminant Noise

It is possible for contaminants to appear in metabolomics experiments, which can affect the results of both DDA and DIA experiments. With the term contaminants we refer to unwanted chemicals that are released into the mass spectrometer, therefore introducing noise into the resulting dataset. It is difficult to determine exactly the source of this noise, given the complexity of these experiments. However, contaminants can potentially occur from an unstable ionisation process, for instance due to background chemicals in

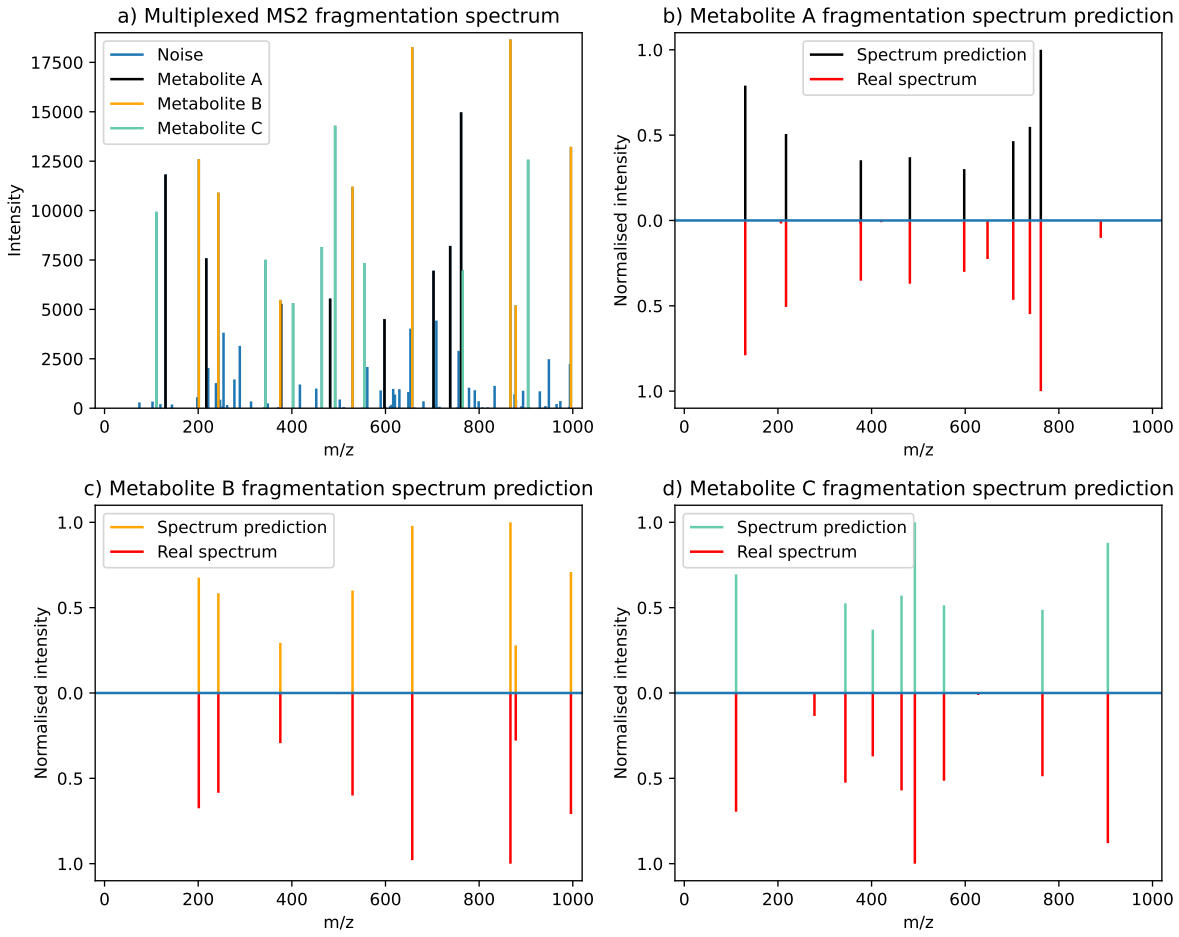


Figure 2.3: Synthetic example of spectra deconvolution in DIA experiments. Plot a) demonstrates a multiplexed MS2 scan, which consists of the fragments of all ions in a m/z range at a specific retention time. In our example, three metabolites were fragmented in this scan, and different colours have been used to distinguish the fragments of each metabolite, plus the noise. In a real DIA experiment, we would not know these associations. In fact, the goal of deconvolution is to make these associations, which then allows the identification of all three metabolites, using database lookup methods of their fragmentation patterns. This is illustrated in plots b), c) and d). After each spectrum is deconvoluted, a matching fragmentation pattern is found, which is linked to a known metabolite. Note that the normalised intensities are displayed in these plots, since it is the pattern of fragments, rather than their intensity that needs to be similar (since a metabolite can be more abundant in a different experiment, but the way its chemical bonds break in similar fragmentation conditions is the same).

the chromatography column [11, 37].

Although contaminants can influence experiments of both data acquisition modes, they do so in a different manner. In DDA experiments, signals from contaminants can be selected for fragmentation, therefore sacrificing time which could have been used to obtain fragmentation patterns of more valuable metabolites (as discussed in Section 2.3.1). In DIA experiments, fragments of contaminants can appear in MS2 scans, complicating deconvolution even further. Since such signals are unpredictable, they are not trivial to be removed. An approach to determining and removing such noise is through the use of blank samples. Blank samples are obtained following the same procedure as with normal samples, with the only difference being that blank samples do not contain chemicals of interest. The observed signals in experiments with blank samples can therefore be compared with the signals of experiments with the samples of interest, and similar signals can be removed [6]. Moreover, additional experiments could be performed to better understand the source of these contaminants and assist in removing their signals.

## 2.4 Data Preprocessing and Analysis

The workflow of a metabolomics experiment can vary depending on the purpose and the context of the experiment. However, it is common practice to follow some standard steps, often in a specific order [2, 22]. These steps can have a big influence on the final result and analysis of the experiment, and thus careful consideration is required for the selection of a workflow. Furthermore, due to the high complexity of metabolomics experiments and data, there are numerous methods and available software tools performing each or a subset of these steps, and for the majority of steps, there is no method that is guaranteed to completely eliminate the issues they are attempting to solve. In this section, we will briefly examine these steps, the data complexity issues they address and their importance for the final experiment results. Section 2.5 will then explore a crucial step in more detail, which is peak picking.

The first step of a metabolomics experiment is data collection. This step can include the choice between a mass spectrometry approach and other existing approaches, and sample collection and preparation techniques. Samples need to be processed following specific chemical procedures before being injected into a mass spectrometer. Furthermore, it is possible to inject many different samples before, between and after injecting the samples of interest. Each of these extra samples serve a specific purpose, with the ultimate goal of increasing the trustworthiness of the collected data. For instance, samples from previous experiments or with known chemical composition can be injected, to ensure that the resulting data are as expected and that the mass spectrometer is functioning

properly. Another common strategy is to use pooled samples, which contain a small sample from all samples of interest. These can be injected prior to injecting the samples of interest, in order for the liquid chromatography column to adapt to the biological and chemical processes of all the compounds present in the study. These can then also be injected between injections of samples of interest, and compare their resulting data with the initial runs, to ensure or to correct for deviations in the retention time axis or intensity values. Finally, it is common to conduct treatment versus control metabolomics studies. Such studies are useful in the discovery of biomarkers for diseases. Samples from subjects of both groups are collected and injected in random order, and differences in the group level can be identified through statistical analysis, which is one of the further steps of a metabolomics experiment workflow.

The next step after data collection is data preprocessing. Data preprocessing refers to the adjustments performed on the raw data in order to reduce the noise and improve the signal to noise ratio [32]. Common methods to achieve this include signal smoothing, baseline correction, peak picking and peak alignment between samples. Examples of possible signal smoothing approaches are splines, digital filters, or applying transformation functions such as the wavelet and the Fourier transformation [36]. Baseline correction aims at estimating and reversing changes in the background or baseline signal over time which can occur. A common method for tackling this issue is again using splines. Peak picking is the most important preprocessing step, and it is almost always part of a metabolomics experiment's workflow. It attempts to identify regions (called peaks) in the data corresponding to signals from metabolites. Signals from metabolites usually follow a specific pattern which differentiate them from noise signals (a bell curve shape, as displayed in Figure 2.2). However, peak picking is not a straightforward process, since not all metabolite peak signals will follow this exact pattern. This issue will be further explored in Section 2.5. Finally, peak alignment is also an important step in experiments with multiple samples. It attempts to identify metabolite peaks in different samples. Collecting data about the same metabolite peak across many samples facilitates statistical analysis and helps to eliminate the noise arising from each experimental run. Furthermore, in treatment versus control studies, it allows the comparison of metabolites between groups.

Data processing is then performed on the data obtained from the previous step. Each peak can be represented as a feature, and a data matrix can be constructed, where the rows are the different samples, the columns are the different peak features, and the value of each sample and peak feature combination can be the highest intensity value of the peak in that sample, or the total area under the peak. Data normalisation can also be performed, with different approaches aiming at making data between different

samples comparable, as well as improving the distribution of the values of peak features in each sample [5, 32]. However a very important data processing step in untargeted metabolomics experiments is metabolite identification. It is at this step that fragmentation patterns of ions are matched against database records of known metabolites. There are public databases with the fragmentation patterns (and other features) of metabolites under specific instrument and experiment conditions, such as METLIN, SIRIUS and HMDB [2, 7, 21, 42]. However, there is also the option of using an in-house library, which is a collection of fragmentation patterns constructed with injecting samples of known metabolites into the mass spectrometer prior to running the experiment. This can be useful, due to the fact that many different experiment parameters can produce different fragmentation patterns, and therefore not all public databases can capture all the combinations for all metabolites [3]. Metabolite identification can play a vital role in a metabolomics experiment, since observing a significant difference between the abundance of a specific metabolite in a treatment versus control study can help researchers draw important conclusions about the underlying chemical and biological processes. This is the part of the experiment that our work is aiming to improve. As discussed in Section 2.3, a compromise has to be made between running an experiment in DDA or DIA mode, with the former producing less but more accurate fragmentation patterns of metabolites, while the latter captures fragmentation patterns of more metabolites, but their accuracy directly depends on the effectiveness of the deconvolution algorithm. A perfectly effective deconvolution algorithm would result in accurate fragmentation patterns for all metabolites in the sample, which would in turn allow for their identification.

Finally, the last steps of a metabolomics experiment are statistical analysis of peak features and pathway analysis. The former uses statistical tests to identify significant deviations of metabolite abundances between samples or groups. These can be univariate tests. However, multivariate analysis is more common, which examines the simultaneous changes of peak features, given that many metabolites will be dependent through their associated reactions in different chemical contexts. When significant changes of identified metabolites have been concluded, pathway analysis attempts to infer the possible metabolic process which would explain the observed changes.

## 2.5 Peak Picking

In liquid chromatography-mass spectrometry, analytes are slowly released into the mass spectrometer at different times, depending on their physical properties. As mentioned in Section 2.1, this introduces the dimension of retention time, making the results of experiments three dimensional (with the other two dimensions being m/z and intensity).

When an ion is being released into the mass spectrometer, the detected MS1 intensity values usually follow the shape of a normal distribution, or a peak (Figure 2.2). Therefore, it is important to identify regions where this pattern is observed, and MS2 scans targeting these regions could be necessary for identifying metabolites behind peaks. The reason why this is so important is because there is usually a lot of background noise signal present, and the process of detecting peaks, also referred to as "peak-picking", attempts to separate possible metabolite signals from noise.

This is not a straightforward process, and there are no exact rules that can be followed to separate metabolite peaks from noise. It is common to observe bad shaped peaks, which can occur for various reasons, such as belonging to less abundant metabolites in the injected sample, or due to the influence of larger peaks eluting at the same time. Various software tools are available that implement peak picking algorithms. They make use of mathematical filters, such as the second-derivative Gaussian filter, approximations to the first and second derivatives of a signal and the wavelet transformation [27, 34, 38]. It can be possible to customise some of the settings for these algorithms. These settings can control how flexible or strict the filters are, where a compromise has to be made between excluding low quality peaks of metabolites, and including noise signals resembling a peak. Examples of such software are XCMS [27], MS-DIAL [38], MZMine [24, 26] and peakonly [20]. They return the m/z ratio and the retention time boundaries of each detected peak.

Finally, peak picking is considered to be a good dimensionality reduction technique, with minimal loss of information [44]. This is because a peak is expected to have a constant m/z ratio value, while the retention time and intensity axis vary. In reality, there are small fluctuations due to random errors in the m/z axis, but it remains constant overall. Performing this dimensionality reduction technique of collapsing the m/z axis results in a graph know as an extracted ion chromatogram (XIC), with the retention time as the $x$-axis and the intensity as the $y$-axis [28].

## 2.6 MS-DIAL

MS-DIAL is a software that was mentioned in Section 2.5 for peak-picking. However, peak-picking is one of the many tasks that MS-DIAL can perform, with deconvolution of MS2 spectra in DIA experiments being a very popular feature [33, 38]. MS-DIAL is considered to be the gold standard method for DIA deconvolution and also the most widely used, so we have selected MS-DIAL to compare the performance of our methods. The methods and metrics used to make these evaluations will be explained in more detail in Section 4.3 and Chapter 5. In this section, we will explore how MS-DIAL performs peak picking and deconvolution.

To perform peak picking, MS-DIAL first smooths the MS1 data [38]. Various methods are available to the user for this smoothing operation, with the default and most common being a linearly weighted smoothing average. Following this operation, then MS-DIAL looks at slices of the m/z axis in turn. For each m/z slice (of width 0.1 Daltons), MS-DIAL derives three different threshold values: the amplitude filter, the first derivative filter and the second derivative filter. The amplitude filter corresponds to the difference in the amplitude of two (smoothed) consecutive signal observations. The first and second derivative filters correspond to the numerically approximated first and second derivative of the signal. For each threshold type, the maximum value observed in the m/z slice is obtained. Then, the median of all values lower than the 5% of that maximum value is chosen to be the threshold value. When all three threshold values have been calculated for a m/z slice, peak edges are identified by locating points where both the first and second derivative values of the signal surpass their thresholds, and the peak apex is located where the sign of the first derivative of the signal changes.

For deconvolution of DIA spectra, MS-DIAL considers each detected MS1 peak in turn [38]. It first collects all MS2 chromatograms falling within the retention time range of the peak. Subsequently, each MS2 chromatogram is smoothed similarly to how MS1 data was smoothed prior to peak picking (for instance via a linearly weighted smoothing average). Subsequently, baseline correction is performed, in order to correct possible deviations in the background/baseline signal. Baseline correction is performed using the following steps: Each MS2 chromatogram is split into segments of the same width. In each segment, all local minima are located. Then, all minima lower than the median of all the minima in the segment are connected to form the baseline. When a baseline is determined for each segment, it is subtracted from the smoothed signal.

Following baseline correction, peak picking is again performed on all MS2 chromatograms, and two metrics are calculated for each peak, the ideal slope and the sharpness value [13, 38]. If the ideal slope value of a peak is lower than 95%, it is discarded. For the remaining peaks, the second Gaussian filter is fitted to an array of their sharpness values. Local maxima of the result of applying this filter indicate which of these peaks can be considered for deconvolution. Finally, the remaining peaks are used in least squares calculation to determine the fragments of the MS1 peak.

MS-DIAL has also introduced CorrDec, an extension to this deconvolution procedure when multiple samples are available [33]. It examines the correlation of ion abundances of MS1 peaks and fragment ions across samples. A metabolite will appear in different concentrations in different samples, but that relative difference will also be reflected in its fragment ions. By identifying this pattern, CorrDec is able to improve the effectiveness of deconvolution. In our project we have focused on the single sample case (with the

exception of combining data from a DIA and a DDA experiment, with more details in Section 4.4). However, our work could be extended to incorporate this useful information and also improve its deconvolution performance.

## 2.7 Data

To develop and evaluate our methods, two different types of data were used; simulated data and data from a real experiment. The simulated data were considered first, before moving on to the real dataset. More details about each type of datasets are explored in Sections 2.7.1 and 2.7.2.

### 2.7.1 Simulated Data

Simulating metabolomics experiment data facilitates the development of new methods and techniques, such as fragmentation strategies. This can be achieved using the Virtual Metabolomics Mass Spectrometer (ViMMS) software [39, 40]. It enables the creation of datasets where their level of complexity can be controlled, and running virtual experiments similarly to how they would be run in a real mass spectrometer setting, without the time and cost requirements of a real experiment. For this project specifically, ViMMS was used to develop new DIA MS2 deconvolution techniques, evaluating their performance on simulated experiment datasets varying the noise and the total number of metabolites in a sample. Initial experimentation of new deconvolution approaches on the simulated data was followed by the use of real metabolomics data, as described in Section 2.7.2.

The simulated datasets that were created for evaluation differ in the level of noise and the number of metabolites in the samples. More specifically, samples with 10, 100, 500, 1000 and 1500 metabolites were simulated. Furthermore, the noise was introduced by adding spike noise and intensity noise in MS2 data. Both these types of noise can be set prior to creating a dataset through ViMMS. For spike noise, intensity spikes are added to the data randomly and uniformly, with a predefined density of the spikes. Additionally, a maximum value of the intensity of each spike is set, and the intensity of a spike is randomly and uniformly selected between zero and that maximum value. Finally, for the noise of intensity values of MS2 data, a Gaussian distribution is chosen around the original values, choosing a value for the variance of that distribution.

We have used these settings to create three different noise levels, starting from no noise at all, to moderate and high noise levels. The exact settings for the noise are provided in Table 2.2. The three different noise levels multiplied by the five possible number of chemicals results in 15 simulated datasets which will be used in evaluation in Chapter 5.

Table 2.2: The three noise levels for the simulated data. In level 0, there is no noise at all. MS1 scans only include the peaks of metabolites, and the MS2 scans only contain the fragments of the metabolites. In level 1, spike noise and MS2 intensity noise are introduced. Spike noise density is set at 5%, which means that on average, in any interval of 100 m/z units, five spikes are to be expected. The intensity value of each spike follows a uniform distribution between zero and 100. The smallest peak in the sample can have an intensity value of 1000 at its apex, and therefore spike noise in this case will always be less that 10% of any peak in the sample. Similarly for level 2.

| Noise level | Spike noise density | Spike noise maximum value | Spike noise max. value as % of smallest peak | MS2 intensity noise standard deviation |
|---|---|---|---|---|
| 0 | 0% | 0 | 0% | 0 |
| 1 | 5% | 100 | 10% | 500 |
| 2 | 15% | 1000 | 100% | 1500 |

### 2.7.2 Real Data

For the real experiment dataset, the results of experiments with samples of different types of beers from McBride et al. [19] were used. Beer was chosen because it is easy to obtain, it is not sensitive data, and it provides rich metabolomic data for analysis. Samples were extracted and used for a number of runs, using different fragmentation methods. More specifically, a SWATH experiment was conducted, by splitting the m/z axis into 10 equal sized fragmentation windows. Additionally, 10 DDA experiment runs were conducted, using samples from the same beer, using a fragmentation strategy called Intensity Non-Overlap [19]. This strategy allows for obtaining high quality MS2 spectra for almost all metabolites, given that metabolites fragmented in previous runs are given low priority in subsequent runs, thus improving the final coverage. Although this is a very effective approach at obtaining high-quality fragmentation spectra for most metabolites, many experiment runs need to be performed, making it a time consuming process. However, these 10 runs allow us to extract MS2 spectra of metabolites that can be used as the gold standard fragmentation spectra, for comparison with deconvolved spectra from the SWATH experiment. To construct this collection of gold standard spectra, we process each run result in turn, and for each peak, we store its MS2 spectrum that corresponds to the highest intensity of the MS1 peak at the retention time of the fragmentation across runs. This approach is based on the assumption that fragmenting a metabolite when it is most abundant will produce the most accurate MS2 fragmentation spectrum.

# Chapter 3

# Statistical Methods

In this chapter we will explore the statistical models and methods that were utilised in this project for the deconvolution of DIA MS2 spectra. A brief overview of each method will be covered in each section, while in Section 3.8 we will look at the evaluation metric used for predicted deconvolved fragmentation spectra.

## 3.1 Linear Regression

The simplest model that was used is the multiple normal linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{3.1}$$

where $\boldsymbol{y}$ is the response vector, $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}$ is the parameter vector and $\boldsymbol{\epsilon}$ is the random error vector, with the random errors being independently and identically normally distributed. The parameter vector estimate is given by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{3.2}$$

which is the ordinary least squares (OLS) estimate.

The intuition behind this choice of model is that the response vector could represent the intensity of an MS2 fragment, the design matrix would provide the information of how the intensity values of metabolites vary by scan (with the number of rows of the design matrix representing the number of MS1 scans, and the number of columns representing the number of metabolites). The parameter vector would be of length equal to the number of metabolites, and each parameter estimate would provide insight into how much each metabolite contributes to the intensity of an MS2 fragment. However, for each MS2 fragment, we would not expect all metabolites to contribute to the fragment's intensity,

making this model a simple choice.

## 3.2 Penalised Regression

Apart from the multiple linear regression model, it was logical to consider regularised regression as well, given that not all metabolites will have the same MS2 fragments. Therefore, two different regularised regression models were considered; Lasso regression [35] and Ridge regression [14]. Both of these models can be expressed in the same way as the multiple linear regression model in (3.1). However, the objective function to be minimised (which is the residual sum of squares) is extended, and a penalty term is added to prioritise models with more coefficients closer to zero. In Lasso regression, the objective function becomes:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \alpha \sum_{j=1}^{p} |\beta_j| \tag{3.3}$$

where $p$ is the total number of covariates. For Ridge regression, the objective function is given by:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \alpha \sum_{j=1}^{p} \beta_j^2 \tag{3.4}$$

In (3.3), the penalty term is the sum of the absolute values of the parameters, also known as the L1 norm of the parameter vector. This forces some of the parameters to be equal to zero when $\alpha > 0$. For this reason, Lasso is used as a form for variable selection. In Ridge regression, the penalty term in (3.4) is the squared sum of the parameter estimates, which is also known as the L2 norm. Contrary to Lasso, Ridge regression tends to keep unimportant coefficients small, but not exactly equal to zero. The $\alpha$ term controls the regularisation effect on the parameter estimates, with a value of zero resulting in the normal linear regression model. The closed form solution for $\hat{\beta}$ in Ridge regression can be shown to be:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \alpha \boldsymbol{I}_p)^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{3.5}$$

where $\boldsymbol{I}_p$ is the identity matrix where its dimension has the size of the number of parameters in $\boldsymbol{\beta}$. The objective function of Lasso is not differentiable, and thus numerical optimisation methods must be used to obtain the parameter estimates.

Both these methods can have advantages and disadvantages when applied to the type of data dealt with in this project. Lasso results in sparse coefficients, which is

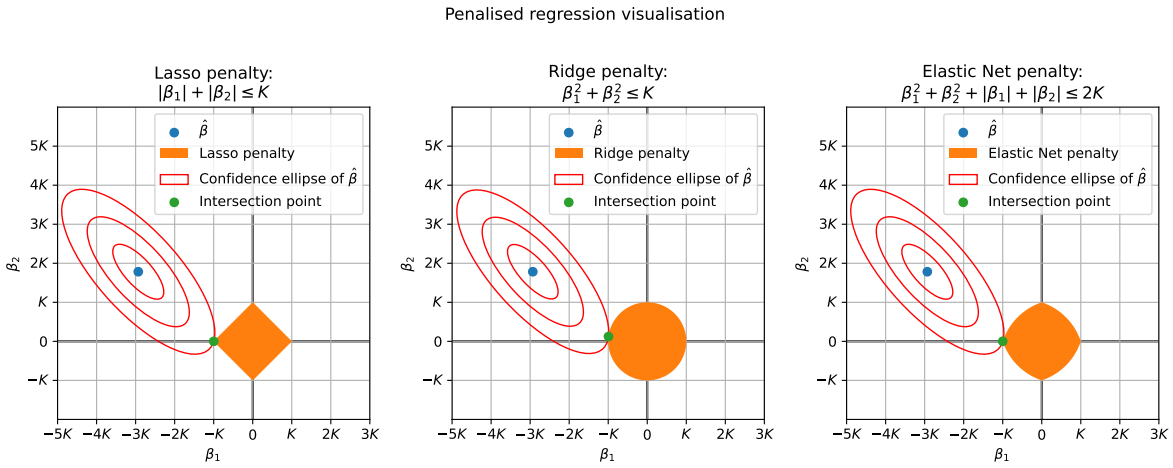


Figure 3.1: Example graphs of Lasso, Ridge and Elastic Net penalties on a synthetic dataset. The blue points show the ordinary least squares estimate, and the red ellipses show the contour graph of the distribution of the OLS estimate. The orange graphs show the shape of each penalty, and how each penalty approaches zero. Lasso can shrink coefficients exactly to zero, which is the case in the leftmost graph, while for ridge it is not equal, but close to zero. The shape of the Elastic Net penalty is a combination of the other two methods.

desirable, given that only a few of the total metabolites in an experiment will have an MS2 fragment with the same mass. However, the disadvantage of Lasso in this case is the expected presence of multicollinearity, due to the presence of isotopes, adducts and MS1 fragments, as described in Section 2.1. In multicollinear data, Lasso is unstable, as it arbitrarily chooses one of the correlated covariates [12]. However, Ridge regression can deal better with multicollinearity, by assigning similar weights to highly correlated covariates. This can work well in the context of isotopes, adducts and fragments, since for instance, it is a better approach to equally distribute the contribution of correlated peaks to a fragment ion's intensity (Ridge), rather than randomly assigning it to one of them (Lasso).

Finally, the Elastic Net model combines both types of regularisation techniques, adding both L1 and L2 penalties to the loss function [45]:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \alpha\rho \sum_{j=1}^{p} |\beta_j| + \alpha(1-\rho) \sum_{j=1}^{p} \beta_j^2 \tag{3.6}$$

where $\alpha$ is the regularisation parameter, and $\rho \in [0, 1]$ controls which type of penalty is more important. In the special case of $\rho = 0$, Elastic Net is equivalent to Ridge, and when $\rho = 1$ it is equivalent to Lasso. A visualisation of the effect of all three penalisation techniques are provided in Figure 3.1.

## 3.3 Ridge Regression as a Bayesian Approach

In this project we need to fit multiple models for each MS2 fragment. In the case of Ridge regression, using the same value of $\alpha$ in (3.4) does not necessarily result in the same amount of penalty being applied to the coefficients of each model. To better understand why this is true, we can interpret Ridge regression from a Bayesian perspective, where the coefficients are given a normal prior distribution with mean zero and variance which is related to the parameter $\alpha$ in (3.4) [14]. Thus, by using the following prior on the coefficient vector:

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \tau^2 I) \tag{3.7}$$

where $I$ is the identity matrix, and then assume the following normal distribution for the regression problem:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I) \tag{3.8}$$

then the posterior distribution of the coefficient vector is:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \tag{3.9}$$

from which we can obtain the maximum a posteriori probability estimate:

$$\hat{\boldsymbol{\beta}}_{MAP} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{3.10}$$

We can see that we can obtain (3.4) by multiplying this quantity with $2\sigma^2$ (since multiplying it with a constant does not change the value that minimises the expression) and setting $\alpha = \frac{\sigma^2}{\tau^2}$. Thus, by scaling the response vector and the design matrix so that $\sigma^2 = 1$ and applying the same $\alpha$ in each MS2 fragment model will result in applying the same amount of penalty on coefficients of all models. This allows us to make the same assumptions about the distribution of coefficients, independently of the MS2 fragment we are modelling.

## 3.4 B-splines

B-splines were used for the estimation of baselines in MS2 chromatograms, as described in Section 2.4. B-splines allow us to fit a smooth, non-parametric curve to data, such that we can remove complex chemical background noise which we cannot predict with

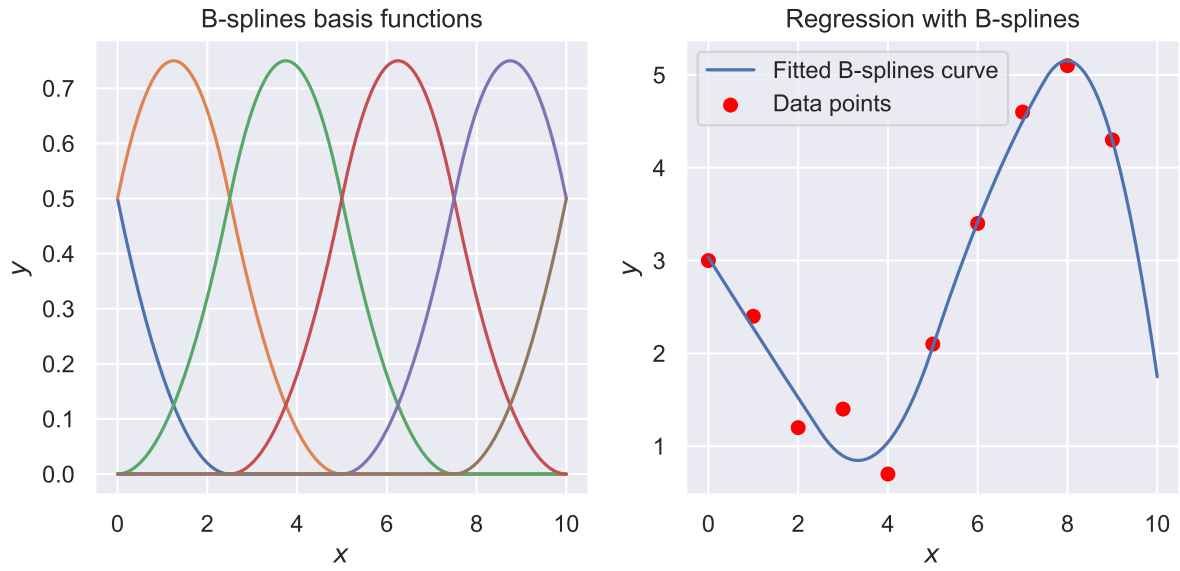


Figure 3.2: Demonstration of B-splines on synthetic data. On the left panel, we observe the B-splines basis functions for five knots and degree of two. For these settings, we obtain six basis functions (six lines in the left panel). If we treat those lines as covariates and fit a linear regression model on the synthetic data points on the right panel, we obtain the blue smooth curve.

our covariate variables. B-splines work by defining a number of points on the x-axis of the data called the knots. The more knots that are chosen, the smoother the final curve is. The final curve can be estimated by performing regression of the response variable against a number of covariate variables, which result from the recursive B-splines basis function definition:

$$B_{i,0}(t) = \begin{cases} 1 & \text{if } t_i \leq t < t_{i+1}, \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

$$B_{i,p}(t) = \frac{t - t_i}{t_{i+p} - t_i} B_{i,p-1}(t) + \frac{t_{i+p+1} - t}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}(t) \tag{3.12}$$

where $i$ in (3.11) and (3.12) is the index of knots, and $p$ is the degree of the splines. For instance, when choosing five knots and degree $p = 2$, then the basis functions are displayed in the left panel of Figure 3.2. The panel on the right displays the fitted curve on synthetic data using these basis functions.

## 3.5 Gaussian Mixture Model

To find the points which we are going to use to fit the B-splines curve described in Section 3.4, we have used a Gaussian mixture model. The idea behind this model is that points occur from a number of different normal distributions, where the number of normal distributions and the number of dimensions of points need to be specified prior to fitting the model [25]. In our case, we wish to distinguish intensity values between values originating from a signal and values originating from the baseline. Therefore, we need two normal distributions (known as components) and our points are one-dimensional. The mixture model is given by:

$$p(x_i|L) = \sum_{j=1}^{2} w_j g(x_i|\mu_j, \sigma_j^2) \tag{3.13}$$

where:

- $x_i$ is a data point, in our case an intensity value,
- $L = \{w_1, w_2, \mu_1, \mu_2, \sigma_1, \sigma_2\}$ is the set of parameters of the model,
- $w_j$ are the mixture weights, satisfying the constraint $w_1 + w_2 = 1$,
- $\mu_j$ is the mean of component $j$, and
- $\sigma_j$ is the standard deviation of component $j$.

By fitting this model we obtain parameter estimates for all the parameters in $L$, which in turn allows us to classify points in one of the two components, by choosing the one that yields the highest probability for each point.

## 3.6 Non-negative Coefficients

The models described in Sections 3.1 and 3.2 can be used to obtain coefficient estimates, which can be interpreted as the contribution of each metabolite to an MS2 fragment. These coefficients should be positive, given that metabolites cannot have negative fragments. Therefore, we need to enforce a non-negative coefficients constraint. This was achieved using the scikit-learn Python library [23], which uses efficient algorithms to calculate non-negative coefficient estimates that minimise each model's objective function.

In Lasso and Elastic Net regression, scikit-learn uses the coordinate descent algorithm for both the unconstrained and the non-negative coefficients constraint [43]. Coordinate

descent works by considering each coefficient in turn, while keeping all other coefficients fixed at values from previous iterations. The value of that coefficient is found that minimises the objective function. If this value is negative, then the coefficient is set to zero before moving on the next coefficient. This process is repeated until the objective function has converged to a minimum, or until the algorithm has exceeded a total number of iterations.

In Ridge regression, the algorithm used to find positive coefficients that minimise the objective function defined in (3.4) is L-BFGS-B [1]. This algorithm makes use of specific techniques to make memory efficient approximations to the hessian matrix, and to constraint variables between lower and upper bounds. In scikit-learn, the lower bound is set to zero, and the upper bound to infinity.

## 3.7 Cross Validation

In regularised linear regression, there are hyperparameters that need to be tuned. More specifically, in (3.4) of Ridge regression, $\alpha$ needs to be specified. Similarly in Lasso and Elastic Net regression in (3.3) and (3.6), specifying $\rho$ as well in Elastic Net. One way to derive values for the hyperparameters is to split the data into two categories; training and validation sets. Then, the model is fit on the training set, the loss or an evaluation metric is used on the validation set, and this process is repeated for different candidate values of the hyperparameters. The hyperparameter values that result in the best performance on the validation set can then be selected for the final model.

The problem with this approach is that the training and validation split might not be representative of the data. A model fitted and evaluated this way could perform well on the given split, but not generalise well. Thus, $k$-fold cross validation can be used to obtain a better understanding of the model's generalisation abilities [**?** ]. In $k$-fold cross validation, all observations are partitioned to obtain $k$ disjoint subsets of the data (Figure 3.3). At each iteration, one of these subsets is used as the validation set, and the remaining subsets are used as the training set. This is repeated until all subsets have been used as a validation set, and an evaluation score is produced at each iteration. The average (or another aggregation operation) score of all iterations is used as an indication of the generalisation performance of the model with the chosen hyperparameter values. This can be repeated for different candidate hyperparameter values, and the best scoring candidates can be used for the final model, which can be fitted using all data.

To define this process in general mathematical terms, first let $D_j$ denote one of the $k$ disjoint subsets, with $j = 1, \ldots, k$. Let each observation pair of covariate values and response value $(\boldsymbol{x_i}, y_i)$ be assigned to one of these subsets. We can also define the function

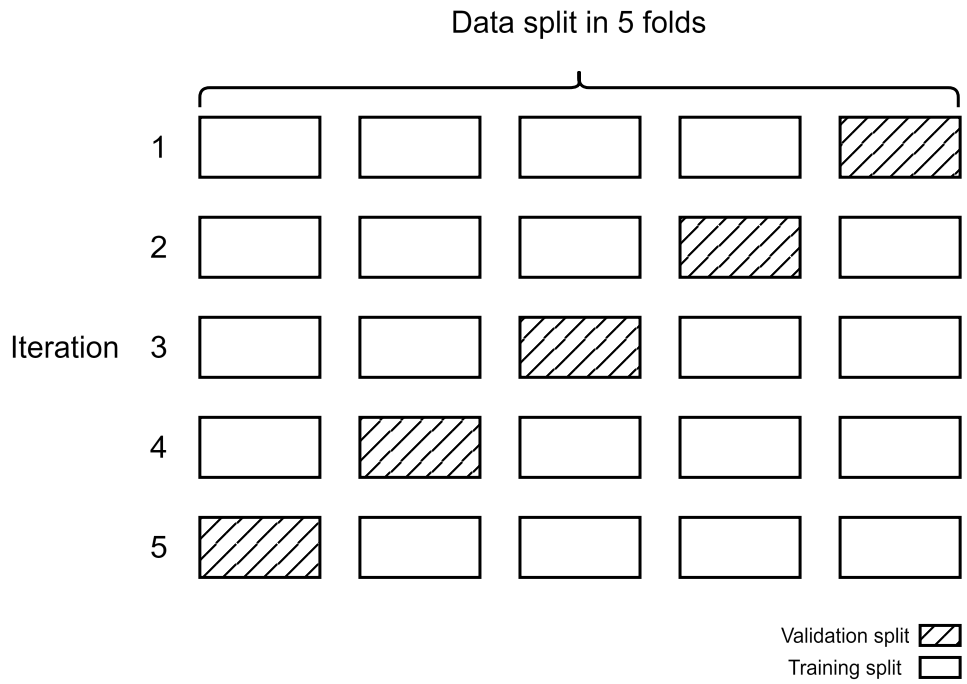


Figure 3.3: $k$-fold cross validation. All data are split into $k$ (in this example $k = 5$) disjoint subsets. At each iteration, one of the subsets is kept aside, and the remaining subsets are used to train the model of interest. An evaluation score is then calculated from the held-out subset. This is repeated until all subsets have been used as validation sets, and the average value of the calculated metrics is calculated to quantify the model's performance.

$g$ as the evaluation metric function which receives two inputs, the observed and the predicted response for a single observation. And finally, if we define $f_{-j}$ as the function of the model trained using all data not belonging to set $D_j$, then we can write the expression:

$$\text{Score}_{\text{cv}} = \frac{1}{k} \sum_{j=1}^{k} \underset{(\boldsymbol{x}_i, y_i) \in D_j}{AF} g(f_{-j}(\boldsymbol{x}_i), y_i) \tag{3.14}$$

where $AF$ is some aggregation function. More specifically, in the case of using the mean residual sum of squares, the expression in (3.14) becomes:

$$\text{Score}_{\text{cv}} = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{(\boldsymbol{x}_i, y_i) \in D_j} (y_i - f_{-j}(\boldsymbol{x}_i))^2 \tag{3.15}$$

## 3.8 Evaluation Metrics

After a model is fitted to the data, we would like to use that model to construct predicted MS2 fragmentation spectra for each metabolite. Thus, we can assume that we have a pair of a predicted fragmentation mass spectrum and a true fragmentation mass spectrum for each metabolite. Details of how each of these are obtained were provided in Section 2.7.2.

A scoring function is required to assess how close a predicted mass spectrum is to the true mass spectrum. The choice made for this was the cosine similarity metric. The reason behind this choice was that the intensity of fragments do not have to match, but rather the relative intensities and the fragmentation pattern is important. This is because when comparing two fragmentation spectra, the precursor intensity might differ significantly due to different experiment conditions, as well as different machine settings or baselines. It is therefore a metric widely used for searching and matching fragmentation spectra obtained from experiments with fragmentation spectra of known metabolites on a database [29].

Other metrics also exist to quantify the similarity of fragmentation spectra. For instance, when MS-DIAL matches an observed fragmentation spectrum against a reference framgentation spectrum from a database, it uses an adaptation to the cosine similarity, where it halves ion abundances in the measured spectrum for which the reference spectrum does not have a fragment [38]. This technique attempts to take into account the fact that there may be isotope fragments or other background noise present, which would deteriorate the performance of cosine similarity. It also calculates the same metric, but by taking the halves of the reference spectrum rather than the observed one. A combination of these two scores is used as the final evaluation metric.

Another spectral similarity metric is the Spec2Vec metric [15]. This aims at improving the association between high similarity scores and structural similarities of the underlying chemicals of the spectra being compared. This was achieved by deriving vector embeddings of fragmentation spectra from numerous molecules, which are lower dimensity vectors representing the structural information of a spectrum. However, we have not used this approach in our evaluation, mainly because it requires to fit a model using a large number of MS2 spectra, and we are also interested in how our methods compare to MS-DIAL, for which cosine similarity suffices.

To calculate the cosine similarity between two mass spectra, they need to be in the form of vectors. This can be achieved by splitting the m/z range into bins and treating each bin as a dimension of the vectors. In our deconvolution methods, the bins (their width and their range) have already been specified prior to model fitting and evaluation as part of the process of constructing the response variables (with more details about this process being introduced in Section 4.1.2). If we denote the vector of the predicted spectrum as $\hat{\boldsymbol{z}}$ and the vector of the true spectrum as $\boldsymbol{z}$, then we can calculate their cosine similarity score with:

$$S_C(\hat{\boldsymbol{z}}, \boldsymbol{z}) = \frac{\hat{\boldsymbol{z}} \cdot \boldsymbol{z}}{|\hat{\boldsymbol{z}}||\boldsymbol{z}|} \tag{3.16}$$

where the cosine similarity between two vectors will always be between zero and one (since the elements of the vectors are intensity values which are non-negative), with one indicating a perfect match.

# Chapter 4

# MSDeconvolve

In this section, we will combine the description of the data from Chapter 2 with the models introduced in Chapter 3 to describe how we have created a framework (MSDeconvolve) which allows for the deconvolution of MS2 mass spectra.

To build such a model, we make the assumption that we are working with data from a single experiment, with the only exception being the combination of DDA and DIA experiment data, which is described in Section 4.4. As described in Section 2.2, the raw data is organised in a sequence of scans, where a scan can either be an MS1 or an MS2 scan. We generally assume that at least one MS2 scan is followed after each MS1 scan, which is often the case in DIA experiments (but not necessarily in DDA experiments). According to Table 2.1, we represent MS1 scans with $\phi_r$ and MS2 scans as $\psi_h$. We denote the time when scans take place as $rt^{(\phi_r)}$ and $rt^{(\psi_h)}$ for an MS1 or an MS2 scan respectively.

In Sections 4.1 and 4.2 we will make use of this notation and introduce a few more terms to define our model, and show how it can be used to predict deconvoluted mass spectra.

## 4.1 Model Building

From the raw data that we have available, we would like to extract all the relevant information that we can use to create a statistical model. From the MS1 scans, we are only interested in data of extracted peaks. This will be used to construct covariate data, a process which will be explained in detail in Section 4.1.1. From MS2 scans, we are aiming to extract information in order to construct the response variable. To achieve this, we need to split the m/z axis of MS2 scans into bins. Then, each MS2 bin will be treated as a response variable, which will be explained by the MS1 peaks. This process will be thoroughly explained in Section 4.1.2. Finally, we will look at how these results

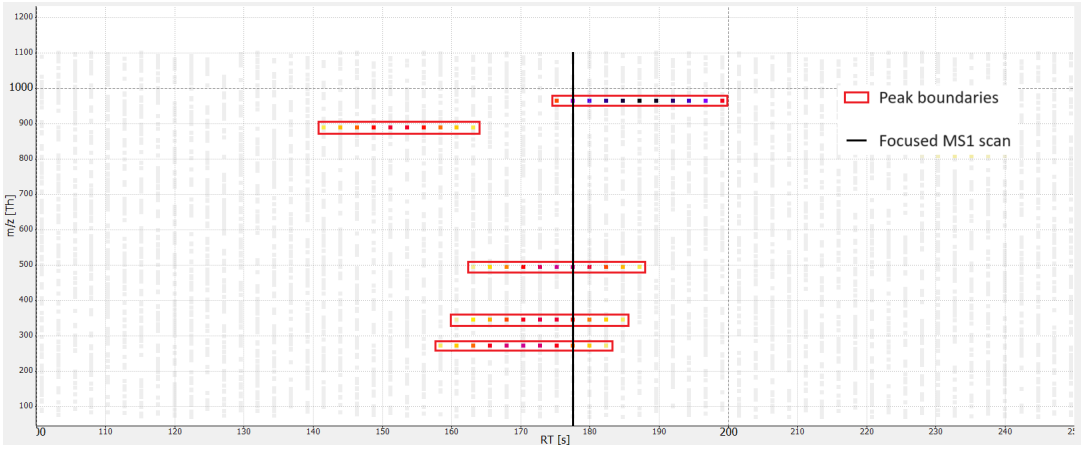


Figure 4.1: Example of an MS1 scan intersecting a set of peaks. Data is displayed as a two-dimensional plot (m/z and retention time), with the intensity dimension being represented by the color of data points. Peak-picking will provide the boundaries of detected peaks, which is displayed with the red boxes. Each MS1 scan will intersect a subset of the peaks. This image is a screenshot from the TOPPView software.

of data processing are combined and used in a model in Section 4.1.3.

### 4.1.1 Using Peaks to Extract Features

To extract the covariate variables, a peak-picking algorithm is applied to the MS1 scans of the experiment data, $\Phi = (\phi_1, ...\phi_R)$, as described in Section 2.5. The peak-picking algorithm then returns a list of peaks that we can use. By denoting the number of peaks as $J$, we can refer to each peak as $p_j$, where $j = 1, \ldots, J$. Each peak is associated with a two-dimensional rectangle, with one dimension being retention time, and the other being m/z (Figure 4.1). Therefore, each peak, $p_j$ has minimum and maximum retention time boundaries $(rt_{\min}^{(p_j)}, rt_{\max}^{(p_j)})$, and similar boundaries for m/z $(mz_{\min}^{(p_j)}, mz_{\max}^{(p_j)})$.

The goal is to treat the intensity values of each peak as a covariate for our model, and each MS1 scan represents an observation. Therefore, an algorithm to extract these values is in place, performing the following steps:

1. Each MS1 scan $(\phi_r)$ is processed in sequence
2. For each MS1 scan, we observe the retention time $rt^{(\phi_r)}$ it corresponds to.
3. We collect all the peaks that are intersected by the MS1 scan in terms of retention time (see example in Figure 4.1). Mathematically, this translates to the set of peaks for which the retention time boundaries include the retention time $rt^{(\phi_r)}$ that corresponds to the MS1 scan $\phi_r$. We can denote this set of peaks as

$$P_r = \{p_j, j \in [1, \ldots, J] : rt^{(\phi_r)} \in (rt_{\min}^{(p_j)}, rt_{\max}^{(p_j)})\} \tag{4.1}$$

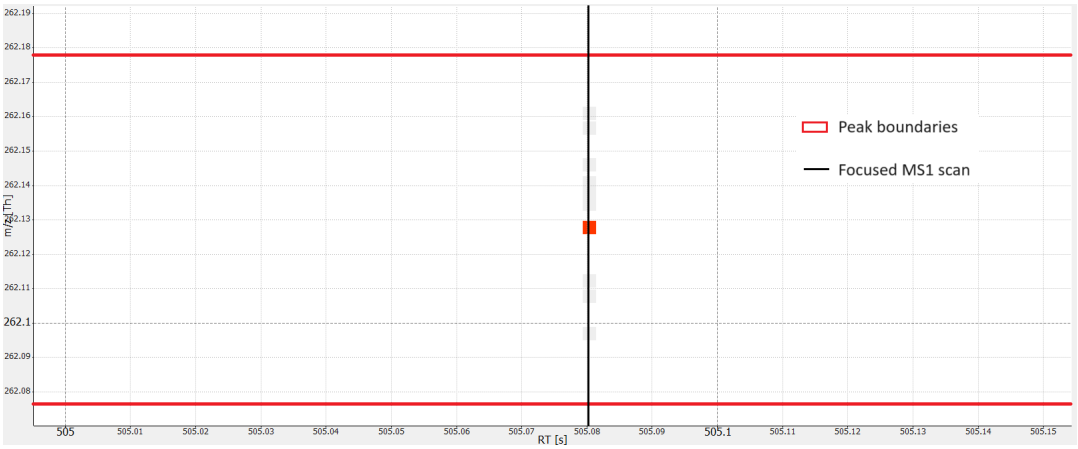


Figure 4.2: Zoomed in TOPPView screenshot of a peak similar to the ones presented in Figure 4.1. There is noise present within a peak's m/z peak boundaries (grey data points under black line, where the black line represents the time an MS1 scan is performed). The maximum of these values (red data point) is taken as the value representing this peak at this particular MS1 scan.

4. For each peak $p_j \in P_r$, we would like to extract an intensity value. It is possible however that there will be more than one intensity value within the m/z boundaries of a peak, because of noise due to complex chemical processes (see example in Figure 4.2). Therefore, we would like to extract the maximum of these intensities, given that it most likely corresponds to the ion of interest, while the remaining intensities will usually be much smaller in magnitude. Mathematically, we obtain this information from vectors $\boldsymbol{m}^{(\phi_r)}$ and $\boldsymbol{\lambda}^{(\phi_r)}$. We find the range of indices of $\boldsymbol{m}^{(\phi_r)}$ for which their values lie within the peak boundaries $(mz_{\min}^{(p_j)}, mz_{\max}^{(p_j)})$, we fetch the corresponding values from the same indices in vector $\boldsymbol{\lambda}^{(\phi_r)}$, and we obtain the maximum of these values. We can express this as:

$$\max_{w \in V} \lambda_w^{(\phi_r)}, \text{where } V = \{w : mz_w^{(\phi_r)} \in (mz_{\min}^{(p_j)}, mz_{\max}^{(p_j)})\} \tag{4.2}$$

5. For peaks $p_j \notin P_r$, the extracted intensity value is zero.

6. This results in a single value for each peak for each MS1 scan. We can denote this as $x_{rj}$, which by combining steps four and five, can be expressed as:

$$x_{rj} = \begin{cases} \max_{w \in V} \lambda_w^{(\phi_r)}, \text{where } V = \{w : mz_w^{(\phi_r)} \in (mz_{\min}^{(p_j)}, mz_{\max}^{(p_j)})\} & \text{if } rt^{(\phi_r)} \in (rt_{\min}^{(p_j)}, rt_{\max}^{(p_j)}) \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

By repeating this process for each MS1 scan $\phi_r$ we obtain $x_{rj}$ for $r = 1, \ldots, R$ and

$j = 1, \ldots, J$. This allows us to construct a design matrix $\mathbf{X}$, which we will use in Section 4.1.3.

### 4.1.2 Constructing the Response Variable from MS2 Spectra

We would like to use the information provided by MS2 scans to construct a response variable. To achieve this, we first need to split the m/z axis into bins. The bins are of equal width (where the width needs to be manually specified), and we can denote them as $b_k$, for $k = 1, \ldots, K$. Furthermore, we can express the m/z boundaries of each bin as $(mz_{\min}^{(b_k)}, mz_{\max}^{(b_k)})$.

To split each MS2 scan, $\psi_h$, into bins, we simply look at the m/z boundaries of each bin, and sum all the intensities that lie within the boundaries. If there are no recorded intensities in a bin, the value for that bin becomes zero. We can refer to the intensity of a bin $b_k$ for a scan $\psi_h$ as $y_{hk}$. Therefore, we can express $y_{hk}$ as:

$$y_{hk} = \sum_{w \in I} \lambda_w^{(\psi_h)}, \text{ where } I = \{w : m_w^{(\psi_h)} \in (mz_{\min}^{(b_k)}, mz_{\max}^{(b_k)})\} \tag{4.4}$$

As a result, by collecting all the $y_{hk}$ value for all MS2 scans ($h = 1, \ldots, H$), we can express all intensity values of an MS2 bin as a vector $\boldsymbol{y}_k$. We can further define matrix $\mathbf{Y}$, with its columns being all vectors $\mathbf{y}_k$. However, as it will be explained in more detail in Section 4.1.3, we are interested in modelling each MS2 bin separately.

### 4.1.3 Regression

We can now combine all the extracted information as described in Sections 4.1.1 and 4.1.2 to fit regression models. This is done by modelling each MS2 bin, $b_k$, separately, using all MS1 peaks. The reason behind this modelling strategy is that each MS1 peak corresponds to an ion that is subsequently fragmented. Therefore, if that ion has a fragment in an MS2 bin, then that ion's intensity in MS1 scans will be positively associated with the intensity of the MS2 bin. However, multiple MS1 ions could have fragments in the same MS2 bin. Thus, a regression model could theoretically capture this relationship between MS1 ions and MS2 fragment ions. We use a linear model specifically, given that we expect the intensity of fragments to be directly proportional to the intensity of the precursor ion. The resulting regression equation for each $K$ bins, $b_k$, can be expressed as:

$$\boldsymbol{y}_k = \boldsymbol{X}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k \tag{4.5}$$

where:

- $\boldsymbol{y}_k$ is a vector of MS2 intensities for bin $b_k$ of length $R$.
- $\boldsymbol{X}$ is a design matrix with dimensions $(R, J+1)$, with its entries being $x_{rj}$ plus a column of ones as the first column. The same design matrix is used for all $K$ models.
- $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{Jk})^T$ the regression parameters for bin $b_k$. We enforce a non-negative constraint on these coefficients, given that it should not be possible for a peak to have a fragment with negative intensity.
- $\boldsymbol{\epsilon}_k = (\epsilon_{1k}, \epsilon_{2k}, \ldots, \epsilon_{Rk})^T \sim N(0, \sigma_k^2 I)$ the normal uncorrelated random errors of each model.

Then, for each bin we can use one of the regression models described in Chapter 3. For the penalised regression models introduced in Section 3.2, the hyper-parameter values can be chosen using cross validation (Section 3.7). Specifically, for the model of each bin, a list of possible hyper-parameter values is provided. For example, for Ridge and Lasso regression, a list of possible values for $\alpha$ (see (3.3) and (3.4)) is specified. Then, for each value of $\alpha$, cross validation is performed, which results in an evaluation metric (in this case it is the loss function). This is repeated for all the possible $\alpha$ values provided, and the value that results in the best cross validation score is selected. Finally, the model is fitted using all the data using the selected value of $\alpha$.

## 4.2 Predictions

Fitting the regression models presented in (4.5) will result in the parameter estimates $\hat{\boldsymbol{\beta}}_k$ for $k = 1, \ldots, K$. We could interpret these coefficients as the amount by which each MS1 ion contributes to the intensity of an MS2 bin, as a result of its fragmentation. Therefore, in order to produce a deconvolved spectrum for each MS1 ion, we need to make use of these coefficients.

For instance, assume that we wish to predict the fragmentation spectrum of a specific peak, $p_j$ for some $j = 1, ..., J$. We would need to consider each MS2 bin in turn. For each MS2 bin, $b_j$, there is a fitted model, for which there is an estimated coefficient corresponding to the peak of interest, $\hat{\beta}_{jk}$. This will be a non-negative value, given that we enforce such a constraint on all coefficients. By recording this value and then repeating this for all MS2 bins, we end up with a vector of length equal to the number of MS2 bins. This vector is the predicted MS2 fragmentation spectrum of peak $p_j$. This is because we expect these coefficients to represent the ratio of the original peak that results in a fragment in each bin. Therefore, by obtaining all coefficients of a peak across MS2 bins,

we can obtain a fragmentation pattern. We can then multiply all these coefficients by the maximum intensity of the MS1 peak $p_j$, to reconstruct the intensity values of fragments corresponding to the apex of the precursor ion. However, this step does not influence the evaluation of predicted fragmentation spectra, since the cosine similarity metric is used, and only the pattern of the fragmentation spectrum is important.

More specifically, to predict the deconvolved mass spectrum $\boldsymbol{z}_j$ of peak $p_j$, we multiply the maximum intensity of peak $p_j$ with the corresponding coefficient estimates:

$$\hat{z}_{jk} = \left( \max_{r \in [1,R]} x_{rj} \right) \hat{\beta}_{jk} \tag{4.6}$$

where $\hat{z}_{jk}$ is the predicted intensity of the deconvolved spectrum of peak $p_j$ for bin $b_k$. By collecting the predicted values for each bin, we can construct the vector $\hat{\boldsymbol{z}}_j$ for a peak $p_j$.

## 4.3 Evaluation

After model fitting is completed and we have obtained a predicted fragmentation spectrum for all peaks ($p_j$, for $j = 1, ..., J$) as described in Sections 4.1 and 4.2, we then need to evaluate the performance of our method, by measuring how similar the predicted fragmentation spectra are to the gold standard spectra. We achieve this by calculating the cosine similarity score of the vector of each predicted fragmentation spectrum $\hat{\boldsymbol{z}}_j$ against its gold standard fragmentation spectrum $\boldsymbol{z}_j$ using the formula defined in (3.16). The gold standard fragmentation spectra were obtained using data from an experiment using the Intensity Non-Overlap method [19], as described in Section 2.7.2. It is important to note that the gold standard spectra were not used or considered to be available during model fitting. They were solely used for the evaluation of our methods after obtaining the predictions.

In addition to reporting a single cosine similarity score, we visualise the distribution of the scores over the peaks of interest using box plots, to assess the performance of our method. This box plot of cosine similarity scores is plotted next to the corresponding box plot of cosine similarity scores of MS-DIAL predicted spectra against the gold standard spectra. Furthermore, we have also used box plots of the cosine similarity scores of the most intense peaks, using different thresholds (for instance, for the top 20 most intense peaks). This is useful since the most abundant peaks are usually the ones of interest, with lower intensity peaks having a higher chance of being noise instead of metabolite peaks.

Finally, we directly compare MS-DIAL and our method using a scatter plot, where each point corresponds to a peak. This allows us to further investigate the number of

peaks for which the two methods perform similarly (points close to the $y = x$ line) and the points for which each method performs better.

## 4.4 MSDeconvolve Extensions

We can further extend MSDeconvolve by applying a penalty on coefficients across MS2 bin models. As it will be explored in more detail in Chapter 5, this allows us to shrink the sum of the coefficients of each peak across all MS2 bins models to be close to one. The reason behind this approach is that we would expect the intensity of the fragments of an ion to approximately add up to the precursor intensity, prior to fragmentation. Consequently, we can no longer model each MS2 bin independently. The loss function to fit a model for all MS2 bins simultaneously is given by:

$$L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \sum_{k=1}^{K} (\boldsymbol{y}_k - \boldsymbol{X}\boldsymbol{\beta}_k)^T (\boldsymbol{y}_k - \boldsymbol{X}\boldsymbol{\beta}_k) + \lambda \sum_{j=1}^{J} \left[ (\sum_{k=1}^{K} \beta_{k,j}) - 1 \right]^2 \tag{4.7}$$

where the first term is the sum of the sum of squares of the residuals of each MS2 bin, and the second term is the penalty term, controlled by a parameter $\lambda$. The penalty term is an L2 penalty rather than L1, since we are not interested in forcing the sum of coefficients of a peak to be exactly equal to one.

Another extension of MSDeconvolve is the combination of data from a DIA and a DDA experiment for the same sample. This can be useful when we have data from both experiments available, and we could potentially improve the deconvolution performance for peaks which were not fragmented in the DDA experiment. The loss function in this case becomes:

$$\begin{aligned} L_k(\hat{\beta}_k) = &(\boldsymbol{y}_k^{(\text{DIA})} - \boldsymbol{X}^{(\text{DIA})}\boldsymbol{\beta}_k)^T (\boldsymbol{y}_k^{(\text{DIA})} - \boldsymbol{X}^{(\text{DIA})}\boldsymbol{\beta}_k) + \\ &\phi (\boldsymbol{y}_k^{(\text{DDA})} - \boldsymbol{X}^{(\text{DDA})}\boldsymbol{\beta}_k)^T (\boldsymbol{y}_k^{(\text{DDA})} - \boldsymbol{X}^{(\text{DDA})}\boldsymbol{\beta}_k) \end{aligned} \tag{4.8}$$

where we have two sets of response vectors and design matrices, one for the DDA and one for the DDA experiment. The response vectors can have different lengths and the design matrices different number of rows, since the number of observations can differ. The peaks remain the same between experiments, and therefore the parameter vector $\boldsymbol{\beta}_k$ remains the same in the two terms, as well as the number of columns of the design matrices. In DDA however, the design matrix will have columns with only zeros, since not all peaks are fragmented (and thus a response value cannot be determined). Furthermore, $\phi$ is a

hyperparameter which controls the influence of the information from the DDA experiment for coefficient estimation.

# Chapter 5

# Results

All the methods examined in Chapters 3 and 4 were used to develop and evaluate different approaches to deconvolution of DIA metabolomics data. Initially, the methods were applied and evaluated on the simulated data, introduced in Section 2.7.1. After obtaining results from the simulated data, we applied the same methods on the real dataset introduced in Section 2.7.2. Due to the increased complexity of the real dataset compared to the simulated one, further techniques were developed to address some of the challenges. The results of experiments on the simulated data are presented in Section 5.1, while the results on real data, the challenges encountered and the techniques used to overcome these challenges are presented in Section 5.2.

## 5.1 Simulation Results

For the simulated metabolomics dataset, the initial assumption was that a Lasso model would perform well, given its inherent covariate selection property. However, other models were also subsequently attempted. More specifically, we have used linear regression (Section 5.1.2), to assess whether a penalty on the coefficients is indeed necessary, Ridge regression, to examine whether an L2 penalty performs better and is thus more appropriate than an L1 penalty, and finally Elastic Net, which combines both types of penalties (Section 5.1.1). Furthermore, some filtering techniques were attempted (Section 5.1.3), to remove spike noise in the data. This would in turn reduce the noise in the signal and potentially improve the accuracy of the fitted models.

Given the simulated datasets with varying levels of noise and of number of chemicals, we can observe how these factors affect the performance of our methods. In samples with a low number of chemicals, we expect fewer chemicals to be fragmented at the same time. Thus, we should obtain MS2 scans with fragments from only a few chemicals, which

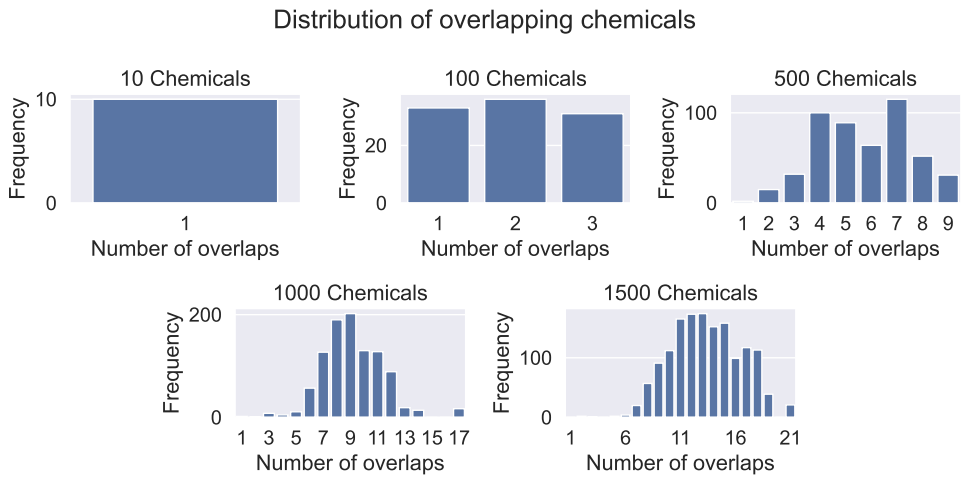


Figure 5.1: Distribution of number of chemicals that are fragmented in the same MS2 scan (overlap) in the five simulated samples with different number of chemicals. For instance, in the 10 chemical sample, the fragmentation spectra of all 10 chemicals do not overlap at all. In the 100 chemical sample, the fragmentation spectra of approximately one third of the chemicals do not overlap with any other, the fragmentation spectra of one third of the chemicals overlap with the fragmentation spectrum of one more chemical, and the fragmentation spectra of the remaining one third of chemicals overlap with the fragmentation spectra of two more chemicals. As the number of chemicals increases and the samples become more dense, chemicals overlap even more.

would make deconvolution an easy task. In Figure 5.1 we can observe bar plots for the number of chemicals being fragmented in the same MS2 scans (number of overlaps) for all our simulated samples. As expected, as the number of chemicals increases, the number of overlaps also increases. Therefore, MS2 scans typically include fragments from many metabolites, increasing the difficulty of deconvolution. Apart from examining how these factors influence our methods, we are ultimately concerned about how the performance we achieve on the simulated dataset compares with the performance of MS-DIAL, which is the focus of Section 5.1.4.

### 5.1.1 Lasso and Penalised Regression

Initially, Lasso regression was attempted using cross validation, as explained in Sections 3.7 and 4.1.3. Cross validation using five folds was performed, using 20 possible values of $\alpha$. Despite the fact that covariates and the response vectors are highly skewed, a log transformation was not necessarily appropriate, given that the relationship between the intensity of a fragment ion and the precursor ion is still expected to be additive. However, we have not completely ruled out using the log transformation, in order to examine how it compares to the models under the additive relationship hypothesis. Therefore,

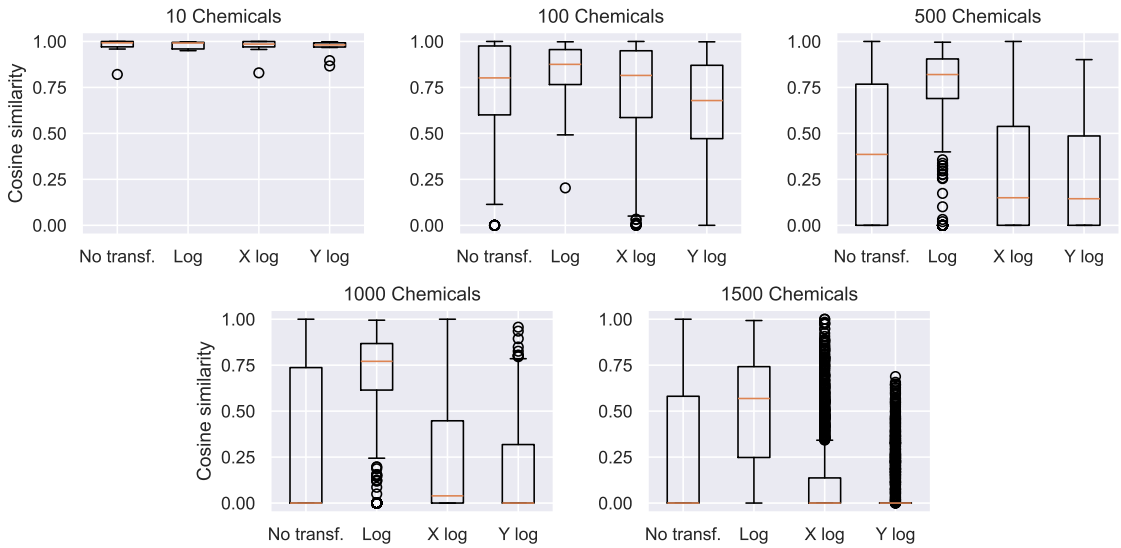


Figure 5.2: Results of Lasso with cross validation on simulated data of 10, 100, 500, 1000 and 1500 chemicals. In each box plot, all four combinations of log transformations is explored.

four different combinations of data transformations were explored, which include not transforming any data, log transforming either the design matrix or the response, and log transforming all the data.

The results on the moderate noise level dataset (noise level 1, from Table 2.2) can be seen in Figure 5.2. For all the models we are also enforcing the non-negative constraint (Section 3.6), given that we do not wish to obtain any negative coefficients (since coefficients are roughly interpreted as the proportion of the precursor ion that results into a fragment). In the 10 chemicals case we can observe that almost all transformation approaches perform well. This comes as no surprise, given that as observed in Figure 5.1, chemicals do not overlap, and thus deconvolution is trivial. In more dense samples, we observe that transforming only one of the design matrix or the response performs poorly with cosine similarities close to zero. However the no transformation and log transforming models both perform well, with the log model appearing to perform better.

From the box plots we can get an idea of how each model performs. However, we cannot compare how the cosine similarities of individual chemicals vary between the models, to examine how different the predictions of each method are. For this reason, we can look at Figure 5.3 for a better comparison of the models with the highest number of chemicals, 1500. We can see that in the last row of panels (the model with log transformation of both the design matrix and the response in the $y$-axis), the majority of chemicals sit above the $y = x$ line, meaning that indeed for most chemicals, the log model performs best. This is clearly the case in the last two panels. It is not as obvious, but still observable in the

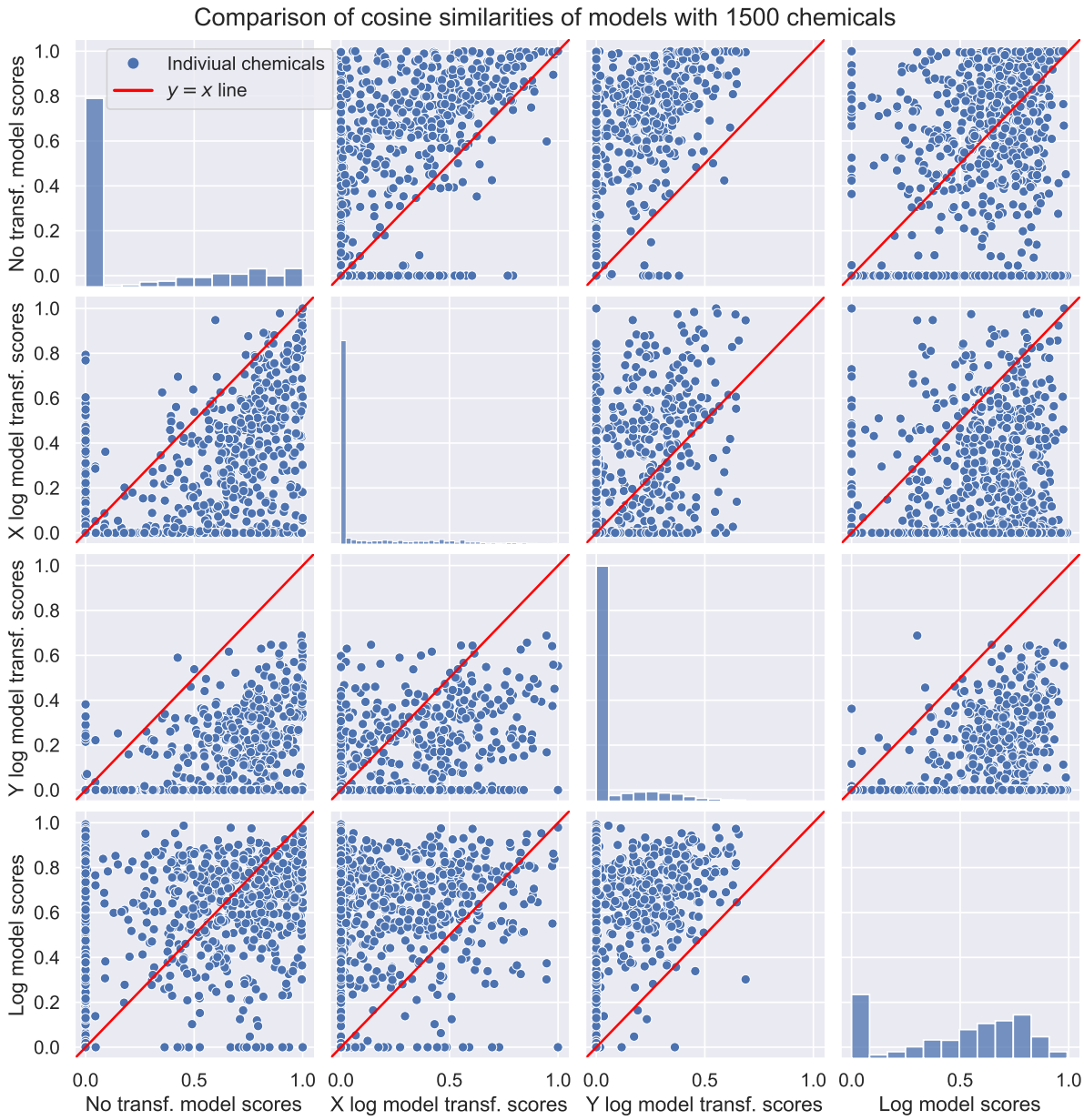


Figure 5.3: Comparing the difference in cosine similarities of chemicals for the models using Lasso with all log transformation combinations, with 1500 chemicals and noise level 1 (Table 2.2).

first panel, which compares it with the model with no transformations.

Although Lasso was expected to perform well with this type of data, it was also of interest to investigate the performance of other penalised linear models. Specifically, Ridge and the Elastic Net were subsequently used, and results are displayed in Figure 5.4. We have used models with no data transformations, and models with log transformation of both the design matrix and the response vector, given that models with a log transformation on only one of those performed poorly in the results of Figure 5.2. Similarly to the results of Figure 5.2, for the sample with 10 chemicals, all cosine similarities are close to one. It is for more dense samples that differences start to appear. Generally, log transformations appear to perform better, with the only exception of Ridge regression. Furthermore, Lasso and Elastic Net seem to perform better than Ridge in the log transformation column of panels.

### 5.1.2 Linear Regression

Apart from the penalised regression approaches examined in Section 5.1.1, it was also of interest to investigate the performance of linear regression without any penalties, to examine whether a penalty was necessary. The non-negative coefficient constraint was enforced, which is the case in all MSDeconvolve models. The linear regression approach was evaluated on the dataset with no added noise (noise level 0 in Table 2.2), and compared to Lasso in the same context for 10 chemicals. Relevant results can be seen in Figure 5.5. Specifically, we observe that multiple linear regression gives a perfect score, while Lasso does not, when no data transformations are applied. We therefore examine if this is the case for samples with more chemicals in Figure 5.6. We observe that we obtain a perfect score using linear regression without log transforming, no matter the density of the sample.

It was therefore of interest to examine how the multiple linear regression approach performs with noise compared to Lasso. For this reason, the corresponding models were fitted on all simulated datasets with noise levels 1 and 2 (Table 2.2). The results can be seen in Figure 5.7. By looking at the left column of panels, we observe that when no data transformations are applied, the multiple linear regression model yields superior performance compared to Lasso for both noise levels. On the left column of panels however, which looks at models with log transformations, we observe that for the dataset with only 10 chemicals, Lasso seems to perform better. This is not the case for more dense samples (which are more realistic for a typical experiment) since multiple linear regression outperforms Lasso in the other samples. However, on average, the multiple linear regression approach with no data transformations appears to have more cosine similarity scores close to one.

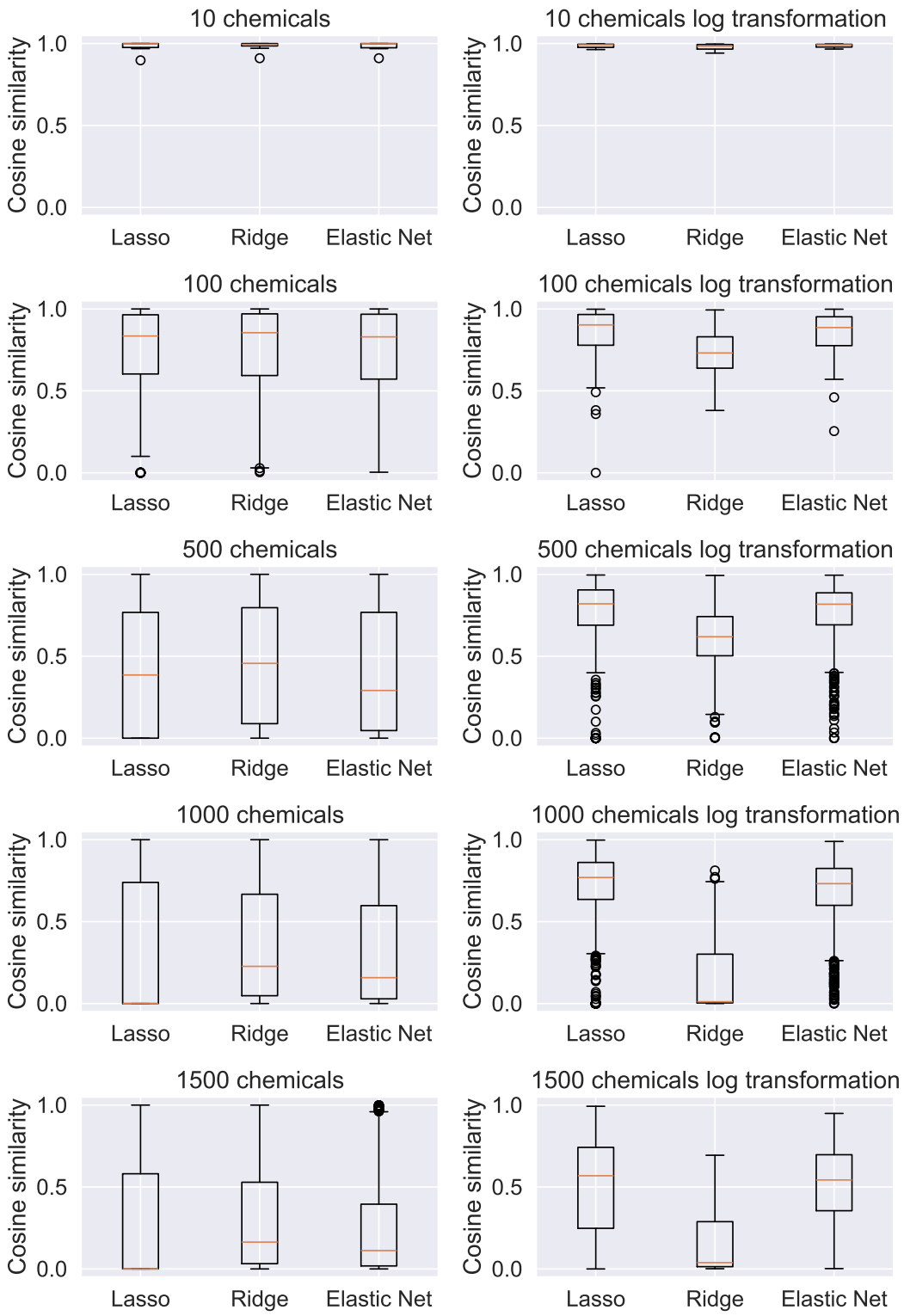


Figure 5.4: Box plot comparison of performance of Lasso, Ridge and Elastic Net models (with cross validation for the selection of hyper-parameters and enforcing the non-negative coefficients constraint) on simulated datasets of noise level 1 (Table 2.2). The left column of panels includes models without any transformations, and the right column has all models which use the log transformation on both the design matrix and the response vector.

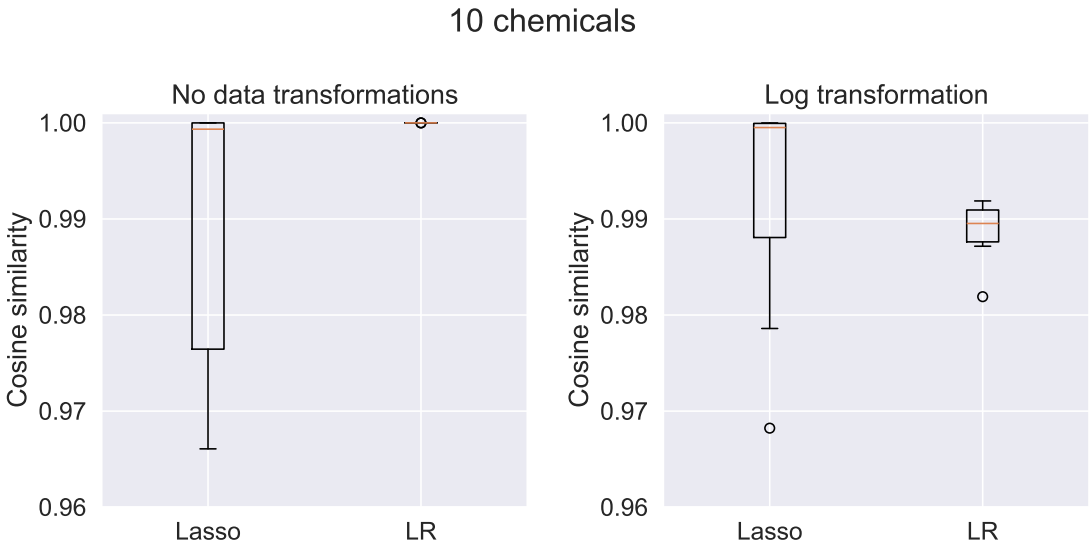


Figure 5.5: Results of Lasso (using cross validation) and linear regression on the simulated datasets with no noise and 10 chemicals. The left panel compares the two models without performing any transformation on the data, while the right panel log transforms both the response and the covariates prior to model fitting.

The final conclusion from all the model comparisons performed here and Section 5.1.1 is that although Lasso performs similarly, if not better, compared to other penalised regression approaches on the simulated datasets, multiple linear regression with no penalty applied appears to be the best model choice. A log-transformation seems to be appropriate only when it is applied to both the dependent and independent variables, which does improve the performance of penalised models. For the multiple linear regression model however, not log-transforming any variable performs slightly better, although both approaches perform well.

### 5.1.3 Filtering

Prior to using the results obtained this far to compare with the performance of MS-DIAL on the simulated datasets, an attempt was made to improve the performance by filtering and reducing the noise of the data prior to model fitting. Since noise can be introduced in the form of random intensity spikes, the assumption was that the identification and removal of those spikes could be possible, resulting in a cleaner dataset and better cosine similarity scores. This type of noise is introduced in the simulations of ViMMS [39, 40]. However, we are also interested in applying this technique in the real data.

The approach implemented to reduce noise in the data was the following. For each MS2 bin, a rolling window algorithm was implemented to identify intensity values surrounded by many zero values. The window width was set to five, and if the total number

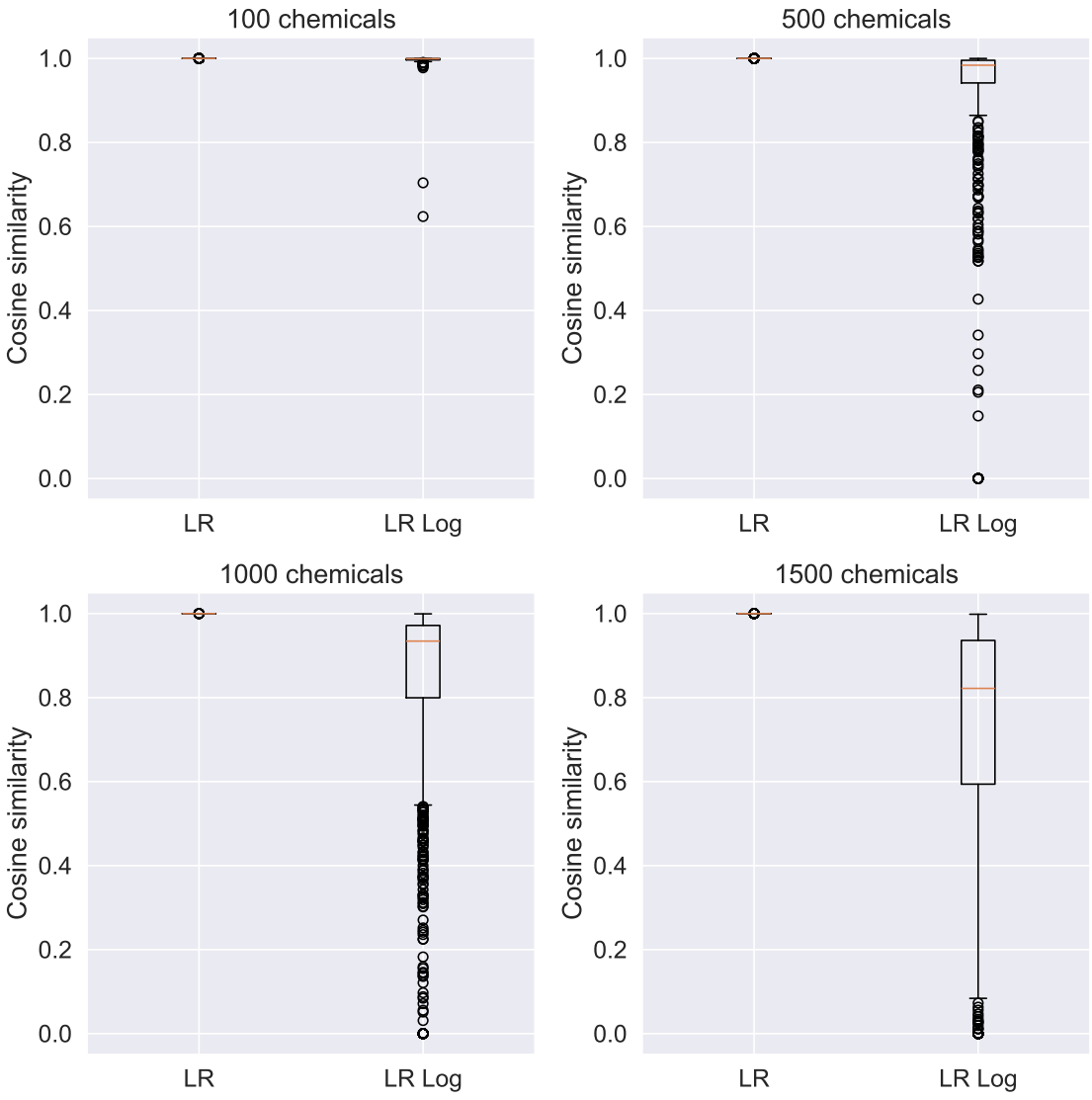


Figure 5.6: Results of linear regression on simulated datasets with no noise. The four panels evaluate the performance of linear regression (with and without performing log transformations) on dense samples.

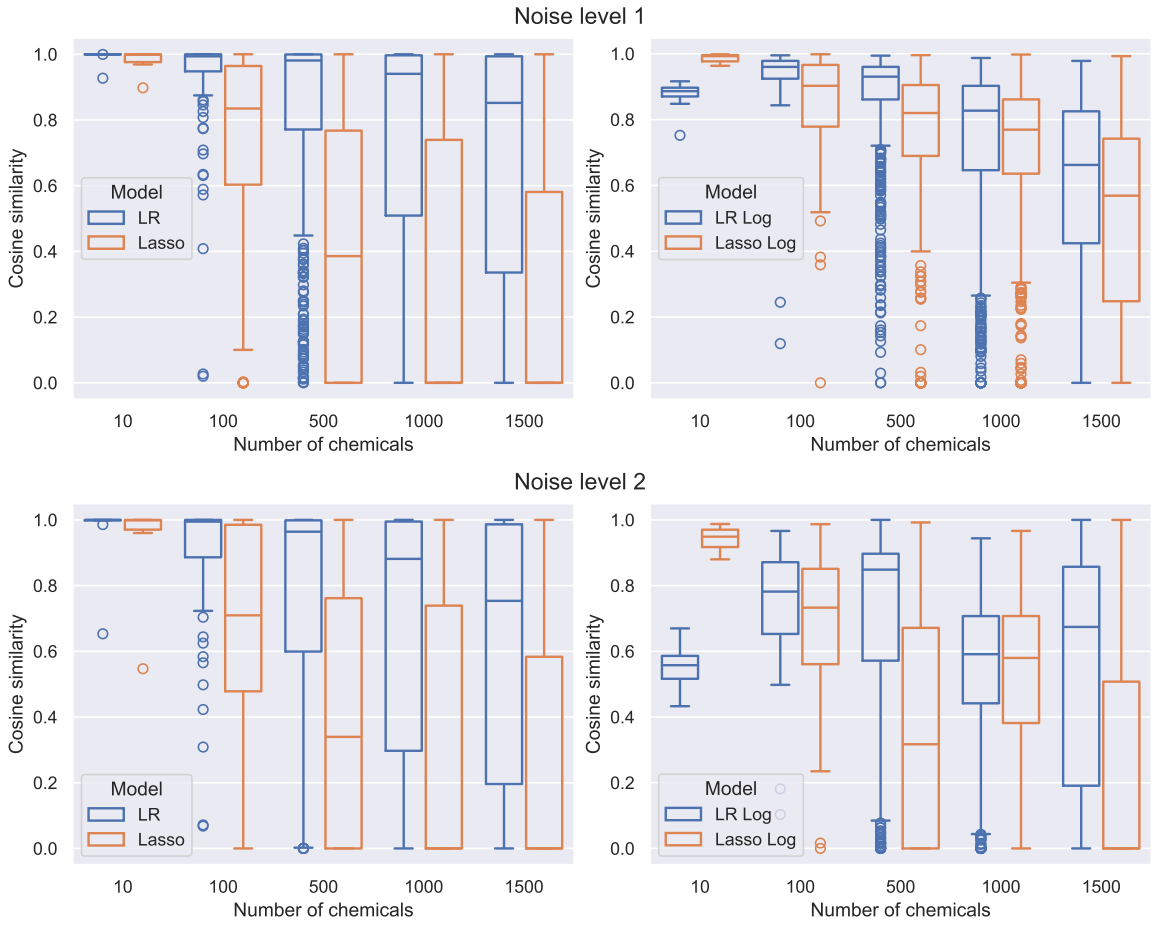


Figure 5.7: Comparison of multiple linear regression and Lasso regression (with cross validation for the selection of hyper-parameters and enforcing the non-negative coefficient constraint) on datasets with noise level 1, which corresponds to moderate noise levels (Table 2.2) and is displayed in the first row of panels, and noise level 2 (high noise levels, second row of panels). The right column of panels compares the models using log transformation of both covariates and the response, while the left column of panels compares model with no data transformations.

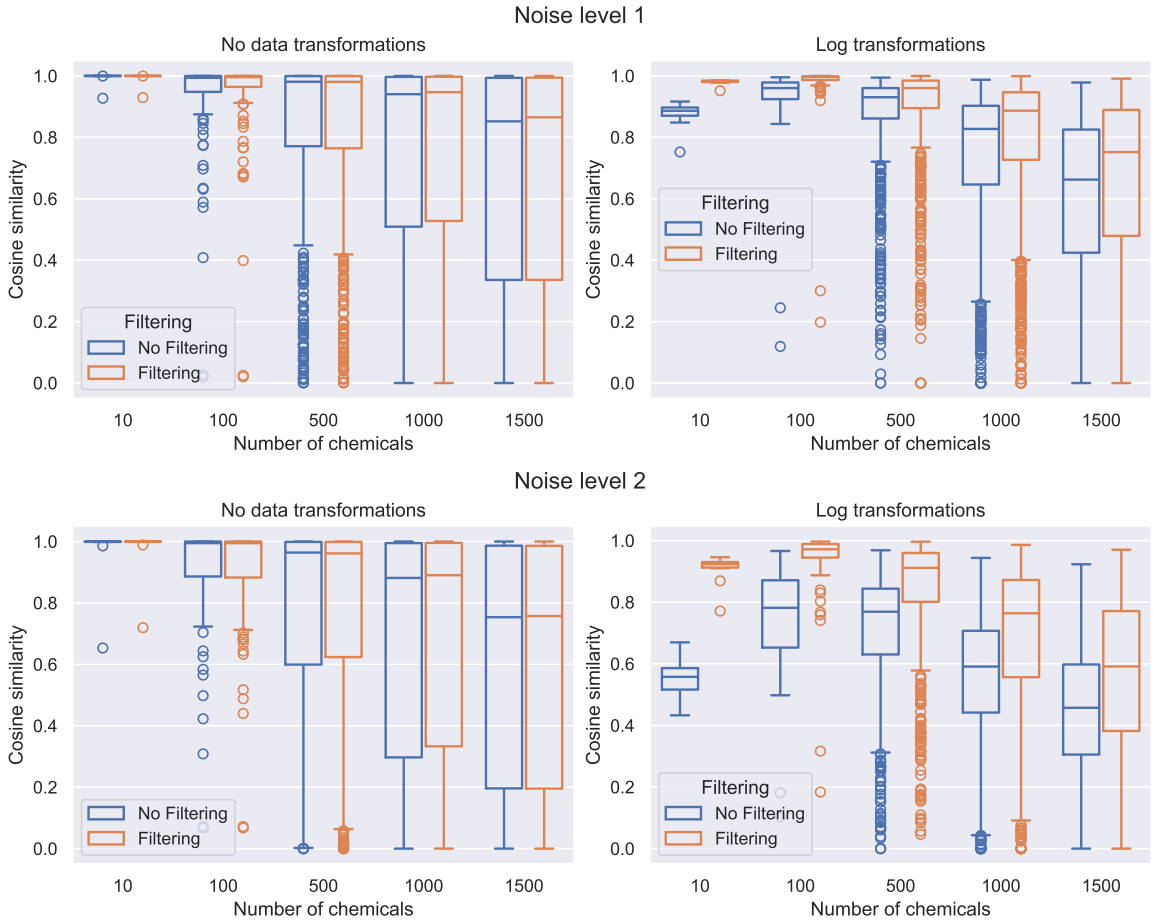


Figure 5.8: Box plots comparing the performance of our methods using filtering and without using filtering to preprocess the data. The top row of panels looks at the datasets with noise level 1 (medium levels of noise, Table 2.2) and the second row looks at datasets with noise level 2 (high noise levels). For MSDeconvolve methods, the multiple linear regression without any penalty is used.

of non-zero intensity values of a window was less then three, then the value at the center of the window was considered to be spike noise, and was therefore removed. Since chemicals appear in the form of peaks with multiple intensity values appearing close to each other, this approach aims at removing isolated intensity values, which are less likely to belong to a signal due to the presence of a metabolite. Furthermore, all removed intensity values were collected, from which their third quartile value was calculated and used as a threshold to further eliminate noise values.

The effect that filtering had on performance can be seen in Figure 5.8. There is a clear improvement in the scores of the models with the log transformations. In the models without any data transformations however, there is almost no difference in the distribution of the cosine similarity scores.

### 5.1.4 MS-DIAL comparison

Using the results from Sections 5.1.1 and 5.1.3, we will now compare in this section MS-DIAL with the best performing model of MSDeconvolve, which is multiple linear regression with noise filtering (from results of Figure 5.8). To achieve this, MS-DIAL was given the noise level 1 and 2 (Table 2.2) simulated datasets to perform peak-picking and deconvolution.

In order for MS-DIAL to perform deconvolution, it first needs to perform peak-picking. However, this puts MS-DIAL in a disadvantage against our methods for the simulated data. This is because we have access to all the peaks from the simulation, while peak-picking can miss some of them. Therefore, instead of using the true peaks in our methods, we also use the peaks found by MS-DIAL. However, we still need to match the peaks found by MS-DIAL to simulated chemicals, because it is through this link that we can obtain true fragmentation spectra we can compare predicted spectra against. For all samples with 10 and 100 chemicals, MS-DIAL was successful in identifying all peaks correctly, and thus comparison was straightforward. For the samples with more chemicals, MS-DIAL was able to identify almost all of them, missing only a few. Therefore, we have fitted our models using only the peaks that MS-DIAL was able to identify, which would be the case in a real experiment.

A box plot comparison of MS-DIAL and the best method from MSDeconvolve using linear regression and noise filtering can be examined in Figure 5.9. In all cases, our method outperforms MS-DIAL, and as the number of chemicals (and thus the number of overlaps) increases, the difference between the two methods becomes greater. This is an encouraging result, which however needs to be confirmed on data of a real experiment, which is the main focus of Section 5.2.

## 5.2 Real Data Results

Similar methods to the ones mentioned in Section 5.1 were attempted on the real dataset (which was introduced in Section 2.7.2). However, the real dataset comes with an increased level of complexity, and thus further methods were examined, which will be presented in detail in this section. In Section 5.2.1 we will explore what makes real metabolomics experiments data highly complex, presenting examples from the dataset we have available. In Section 5.2.2 we will explore how methods similar to the ones we attempted in the simulated experiments perform with the real data, and compare that performance to MS-DIAL. Section 5.2.3 then examines the performance of Ridge regression in combination with other techniques. Those techniques are baseline correction and

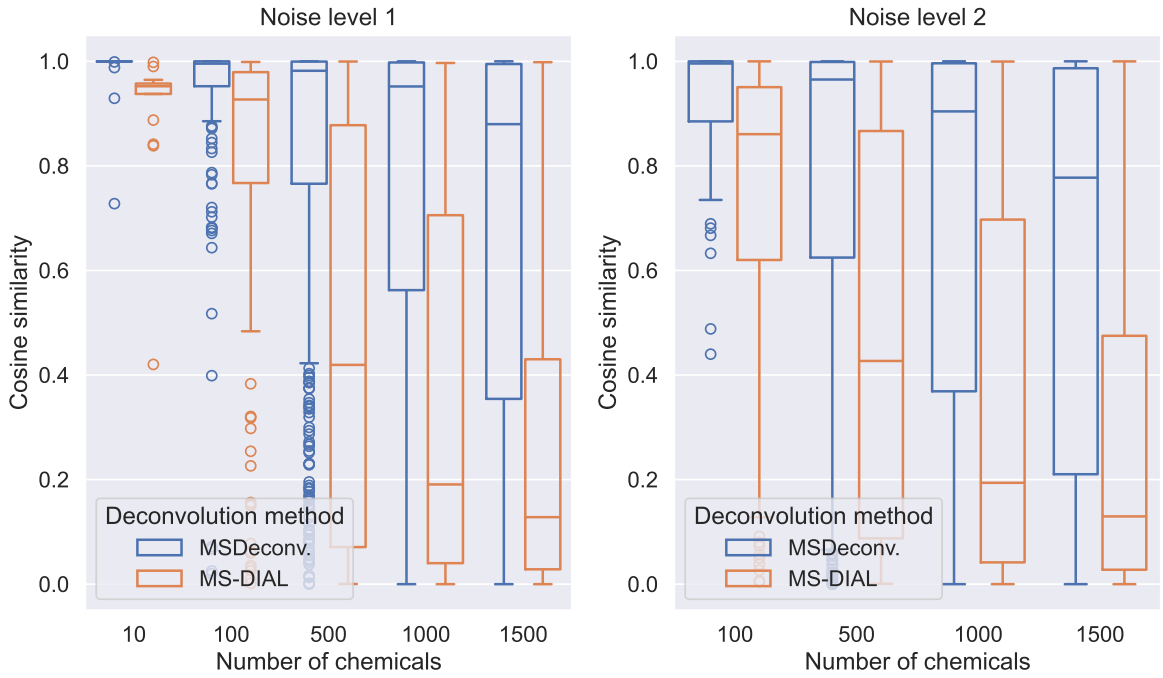


Figure 5.9: Box plot comparison of the distribution of cosine similarities for MSDeconvolve using multiple linear regression and MS-DIAL.

filtering chemicals to include in the model of each bin based on the correlation of its MS1 peak and the MS2 fragment. In Section 5.2.4 we attempt to combine the fragmentation spectra predictions of MSDeconvole and MS-DIAL to observe if the combination of these methods can improve the performance of MS-DIAL. Subsequently, Section 5.2.5 extends the methods already mentioned by enforcing an additional constraint on the coefficients, and Section 5.2.6 attempts an approach were the same L2 penalty is applied in all MS2 bins of the Ridge regression approach. Finally, Section 5.2.7 examines whether the combination of information from both a DDA and a DIA experiment can improve deconvolution of peaks which are not fragmented in the DDA experiment.

### 5.2.1 Real Data Challenges

Although MSDeconvolve methods on the simulated dataset perform well, it is crucial to confirm these results on the real data. However, metabolomics data are highly complex, and thus they present with some big challenges. A mass spectrometry experiment is a complicated chemical process, and thus it inevitably comes with many possible sources of noise, which are difficult to determine exactly. In Section 2.1 some of these challenges are described, such as the presence of isotopes, adducts and the fragmentation of ions in the MS1 level. The appropriate preparation of samples and the adoption of a workflow using

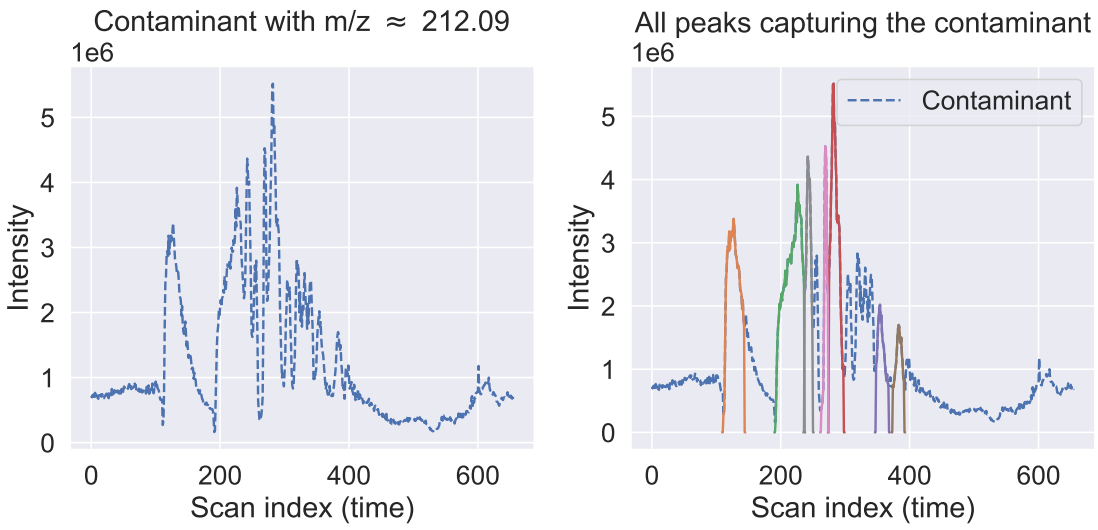


Figure 5.10: Example of a contaminant in the real dataset. In the left panel, the intensity of the contaminant across all MS1 scans is observed. In the right panel, the same graph is plotted, but with all the peaks identified by MS-DIAL that include the contaminant (where each peak is a solid line with a different color).

many samples as described in Section 2.4 can reduce significantly the effects of noise, however, it is impossible to completely eliminate it. Furthermore, despite the fact that DDA acquisition methods have been proven to be consistent across simulations and real experiments [4, 39], it has not been explored how realistic simulated datasets actually are.

One of the complexities of real metabolomics data is the presence of contaminants, as discussed in Section 2.3.3. The use of blank samples for the removal of noise from contaminants has not been applied, due to the added complexity and time constraints. However, it is discussed as a possible extension to MSDeconvolve in Section 6.2. An example of a contaminant in our dataset is presented in Figure 5.10. From the left panel, we can observe that the MS1 intensity values of this chemical spans across nearly the entire experiment, and it does not follow the expected shape of a peak. However, it does not fit the description of random spike noise, and thus it introduces some complications. More specifically, the peak-picking algorithm of MS-DIAL identifies multiple peaks (as shown in the right panel of Figure 5.10), while in reality this is likely to be the same chemical, since it produces the same fragmentation spectrum for all these peaks. Furthermore, the fragments of this chemical will be present in all DIA MS2 scans with an m/z range that include the m/z of this contaminant. To put this into perspective with MSDeconvolve's modelling strategy, where each MS2 bin is modelled separately, if the fragmentation spectrum of this contaminant includes 10 fragments, then this translates to 10 MS2 bins

that will have their intensity values affected. This is an issue because there is not the required information in our design matrix to model this relationship. This is because the parts of the contaminant that are not captured by a peak are not included as covariates in the model.

Another complexity that is found in the real data is the fact that not all peaks have a good shape, and thus peak picking cannot be completely accurate. A compromise must be made between annotating many peaks, where many of them could potentially be just noise, or annotating fewer peaks to avoid including noise, which can also miss other valid peaks. Both sides of this problem can degrade the performance of our models. Including many peaks increases the number of features of each model, and more features need to be eliminated. Including fewer peaks however, can result in missing important covariates in our models. Furthermore, peak picking can accidentally result in capturing the same peak twice, with slightly different m/z and retention time boundaries. This needs to be treated with caution, because including the intensity of the same peak twice in our design matrix of our models increases multicollinearity significantly. Thus, prior to model fitting, duplicate peaks are identified by manually examining covariates with a correlation equal or very close to one, and also by finding peaks for which the same DDA MS2 scan has been used to construct their gold standard fragmentation spectrum.

Despite the removal of duplicate peaks, multicollinearity cannot be eliminated due to isotopes, adducts and fragments which are included in the design matrix. A single metabolite can result in multiple peaks which have very similar chromatograms, with the only major difference being their m/z value. Peaks belonging to isotopes for example, usually differ from the original peak by a few m/z units, but are otherwise very similar to the original peak, resulting in covariates with very high correlation. For example, isotopes which result from the carbon-13 atom instead of the carbon-12 atom will differ in mass by approximately 1.0033 daltons. Figure 5.11 shows an example of the correlation matrix for one of the SWATH isolation windows in our dataset, where many values can be spotted which are close to one.

### 5.2.2 Linear and Lasso Regression Performance

The first model used to fit on the real data was the linear regression model with the non-negative coefficients constraint. Additionally, the log transformation was taken of all variables and used to fit another model. MS-DIAL outperformed this approach on the real data, as clearly seen in the results of Figure 5.12. The distribution of the cosine similarity scores of MS-DIAL is higher than the distribution of scores of our method without any data transformations, while our model with the log transformation performs very poorly. However, due to the sensitivity of the results to the peak picking performed prior to

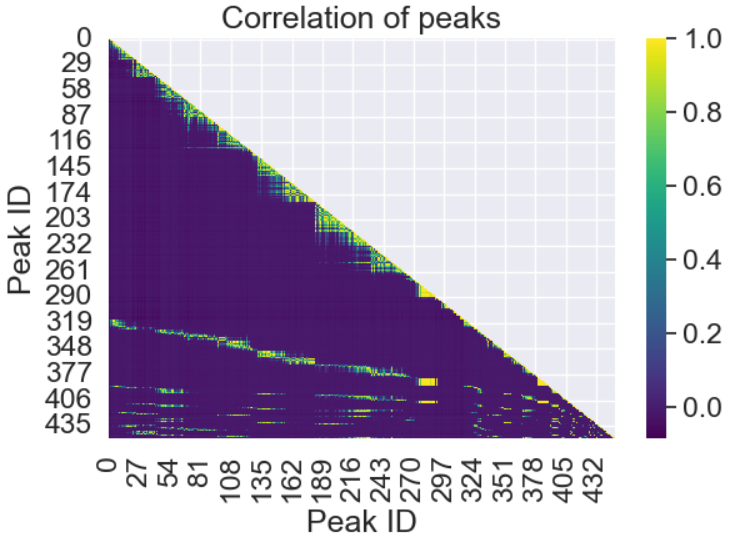


Figure 5.11: Heatmap of correlation matrix of peaks for the SWATH experiment of the real dataset, for the isolation window ranging from 170 m/z to 270 m/z.

model fitting (as discussed in Section 5.2.1), peaks with smaller maximum intensities have a higher chance of being noise rather than a signal from a real peak. Therefore, it is useful to compare the performances of each method for the most intense peaks, given that they have a higher chance of originating from actual metabolites. We thus compare the same models in the lower panel of Figure 5.12 but only for the scores of the top 100 most intense peaks. Similarly to the above panel, MS-DIAL outperforms both our methods, with the one without any data transformations performing substantially better than the log transformed model.

These results were contradicting the results of the simulated data, shown in Figure 5.9. Therefore, some further processing steps were taken to simplify our models without discarding any useful information, to observe if we could improve performance. The first step was to consider each SWATH isolation window separately for model fitting. In our previous approach, all MS2 bins from all SWATH isolation windows were modelled concurrently. However, in this new approach, we significantly reduce the number of covariates for each model, to only include the peaks that fall within each isolation window. The next step to simplify our models was to further eliminate some peaks from each model using two criteria. The first criterion is that all peaks for which their MS1 m/z value is smaller than the m/z value of the MS2 bin that is currently being modelled are discarded. The reasoning behind this approach is that when a metabolite is fragmented, all fragments generally have mass smaller or equal to the original metabolite.

The second criterion was determined as follows; For each peak, the MS1 scan (of the

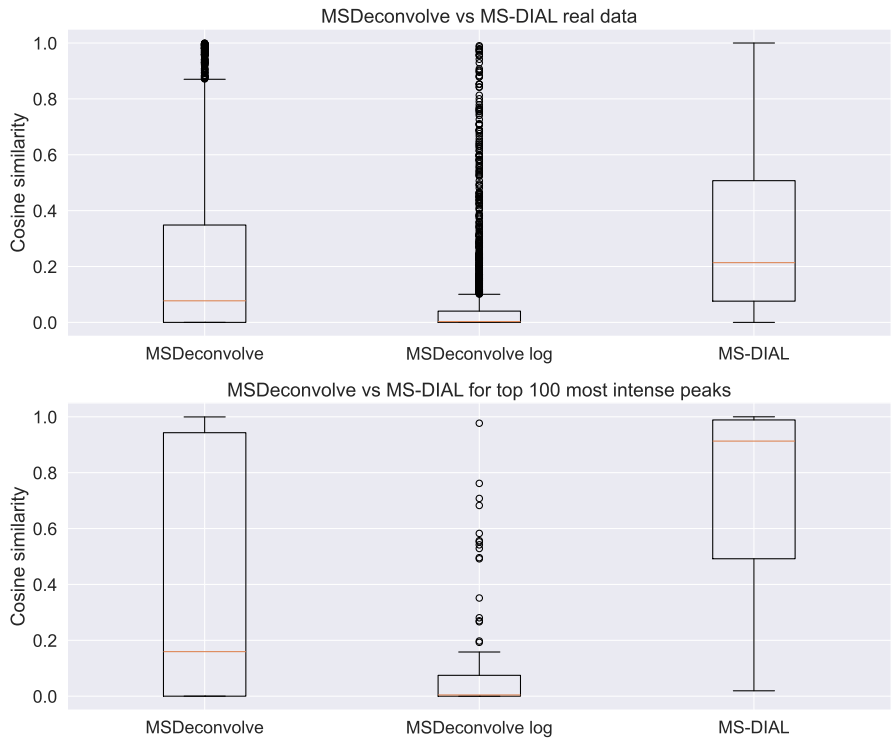


Figure 5.12: Box plots of cosine similarities of MS-DIAL and MSDeconvolve. In the top panel, the cosine similarities of all chemicals in the dataset is plotted. In the lower panel, only the cosine similarities of the top 100 chemicals with the highest intensity at the apex of their peaks. MSDeconvolve methods include linear regression, and linear regression of the logged covariates and response.

DIA experiment) which captures its apex is located. Then, the very next MS2 scan that includes that chemical is obtained. This MS2 scan can include multiple fragments from multiple other metabolites from the same fragmentation window. However, if we bin the m/z axis of that MS2 scan using our predefined MS2 bins, then we will see that not all MS2 bins have a signal. It is impossible for the peak to have a fragment in those MS2 bins. Therefore, for the MS2 bins where the signal is zero, the peak under investigation is not included as a covariate. The combination of these two criteria and the modelling of each isolation window separately in the case of SWATH, significantly reduces the number of peaks included as covariates in our models, potentially improving the performance as well as reducing the time required to fit the models.

An overview of the results of taking these extra steps is shown in Figure 5.13. We can clearly observe an improvement in the performance of linear regression when these steps are performed to filter out unnecessary covariates from the model of each bin. Thus, all subsequent MSDeconvolve methods will make use of this covariate filtering approach. Lasso seems to perform very poorly on average. However, in the lower panel of the figure,

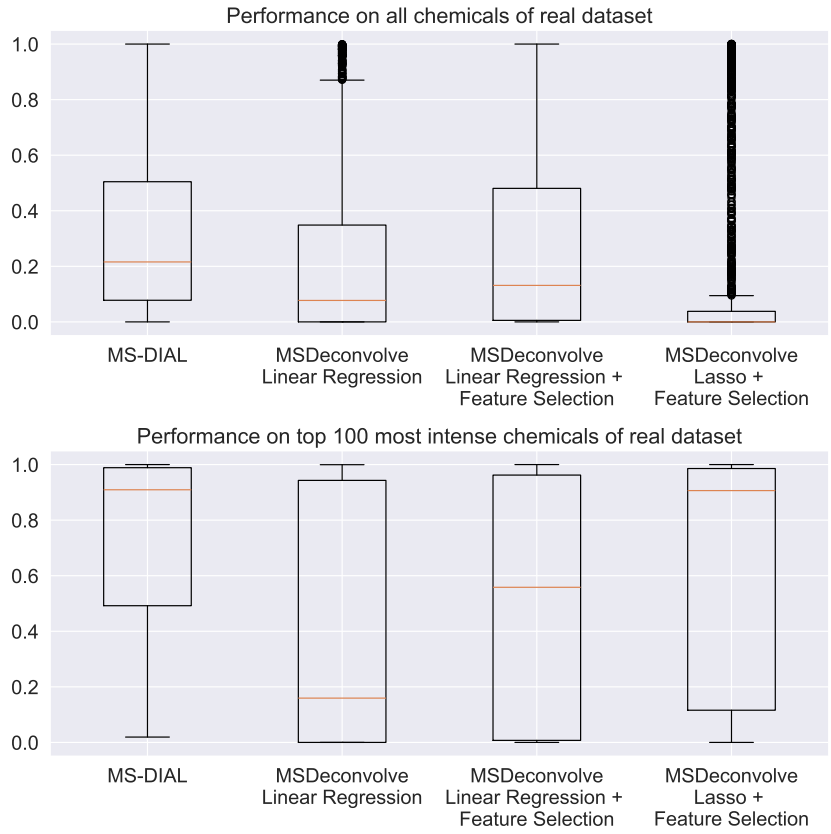


Figure 5.13: Box plots of the cosine similarity scores of our methods and MS-DIAL. The first MSDeconvolve model is using linear regression (same model as in 5.12). The second model is again linear regression, but each SWATH isolation window is modelled separately, and covariates are discarded based on specific criteria. Finally, the third model is using Lasso instead of linear regression, each SWATH isolation window is modelled separately, and covariates are again discarded based on the same criteria.

where the top 100 most intense chemicals are examined, Lasso performs better than linear regression. Despite these improvements in both performance and the time required to fit the models, MS-DIAL is still performing better than MSDeconvolve. Therefore, new approaches needed to be introduced and attempted, which is the focus of the remaining sections of this chapter.

### 5.2.3 Ridge Regression Performance with Baseline Correction and Correlation Filtering

Due to the added complexity of multicollinearity in the real dataset, our next hypothesis was that Ridge regression could perform well, given that it can deal better with highly correlated features. This is because, as discussed in Section 3.2, Ridge converges to a stable solution for the coefficients of highly correlated data, in contrast to Lasso which

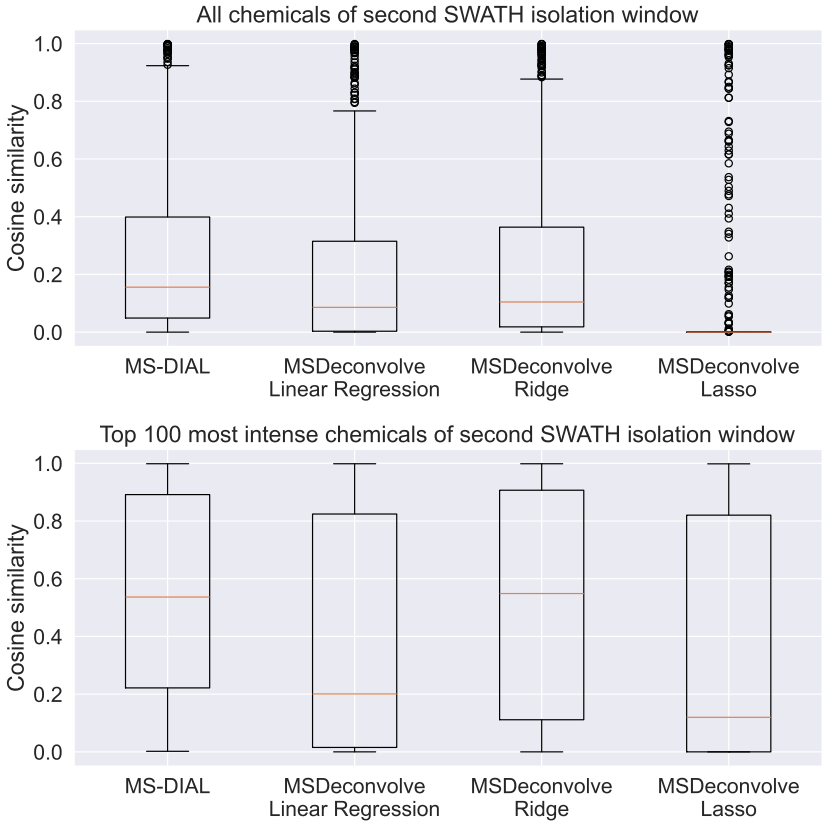


Figure 5.14: Box plots of cosine similarity scores of different models for all the peaks in the experiment and for the top 100 most intense peaks (top and bottom panels) for only the second SWATH isolation window (range from 170m/z to 270m/z, which is the SWATH isolation window with the most peaks). For our methods in this figure, we implement the automatic covariate selection procedure prior to fitting each MS2 bin model, as explained in Section 5.2.2.

can be unstable. Therefore, the results of using Ridge regression were compared against MS-DIAL, linear regression and Lasso for the chemicals of a specific SWATH isolation window are shown in Figure 5.14. Again, we are using the automatic covariate selection technique explained in Section 5.2.2, as well as using cross validation for the selection of hyper-parameters and enforcing a non-negative coefficients constraint. We can indeed observe an improvement in the Ridge approach in both panels of the figure. Comparing our initial attempt at competing with MS-DIAL in the real dataset in Figure 5.12, we can conclude that the adoption of the automatic covariate filtering technique and using Ridge regression instead of linear regression yields a performance which is similar to that of MS-DIAL. However, MS-DIAL still seems to obtain better scores overall. Therefore, further investigation was required to understand what the limitations of our approach was, and to experiment with new techniques to overcome these limitations.

An initial hypothesis was that we could potentially improve performance by controlling

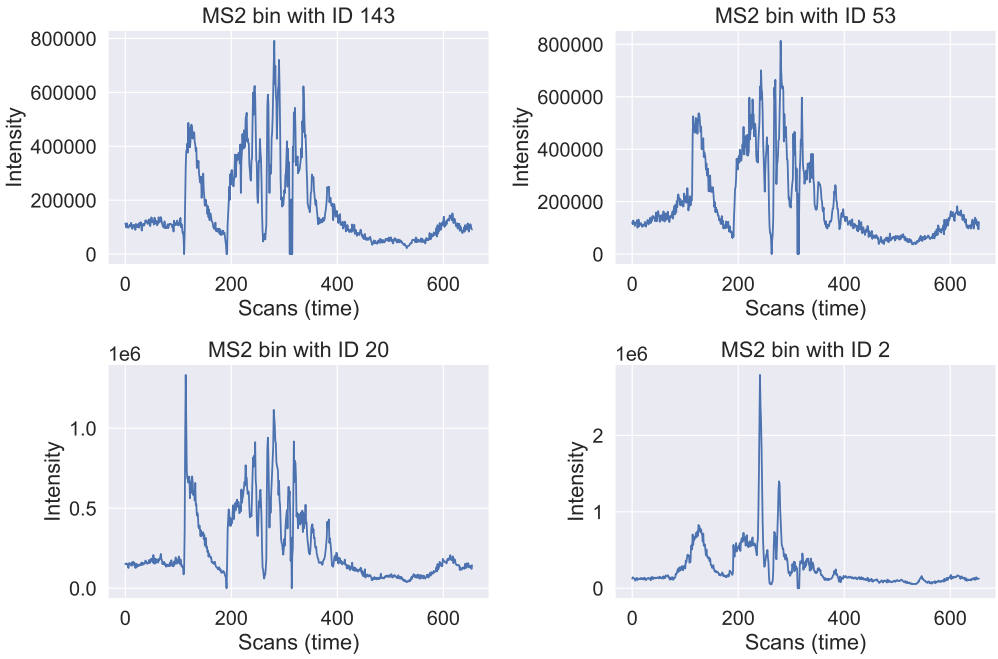


Figure 5.15: MS2 bins with fragments of the contaminant examined in Figure 5.10.

the effect of contaminants. The contaminant in Figure 5.10 for example, could have multiple MS2 fragments, which will inevitably affect the models of the MS2 bins. In Figure 5.15, we have identified some of the MS2 bins that contain a fragment for this contaminant. We can clearly observe a very similar pattern to the intensity of all these bins, which does confirm our hypothesis that from only one contaminant, multiple models can be significantly affected.

Therefore, we attempted to remove this pattern in each bin using B-splines (Section 3.4), prior to fitting the models. We achieved this by fitting a Gaussian mixture model (Section 3.5) with two components for the intensity values on each bin. One component would correspond to the baseline, which would be the intensity of the fragments of the contaminant, while everything else would originate from different fragments of different peaks. Then, all intensity values belonging to the baseline group were used to fit B-splines, and the fitted line was used as the baseline of the bin. Finally, this baseline was subtracted from the original intensity values of the MS2 bin, to obtain the updated response vector for that model. An example of this process is illustrated in Figure 5.16, for one of the MS2 bins with a fragment of the contaminant we have been investigating. The top left panel shows the original signal, while the bottom right panel shows the final signal after removing the baseline. This process is applied to all MS2 bins prior to fitting a model.

A new model was fitted to the data, which used this new baseline removal approach,

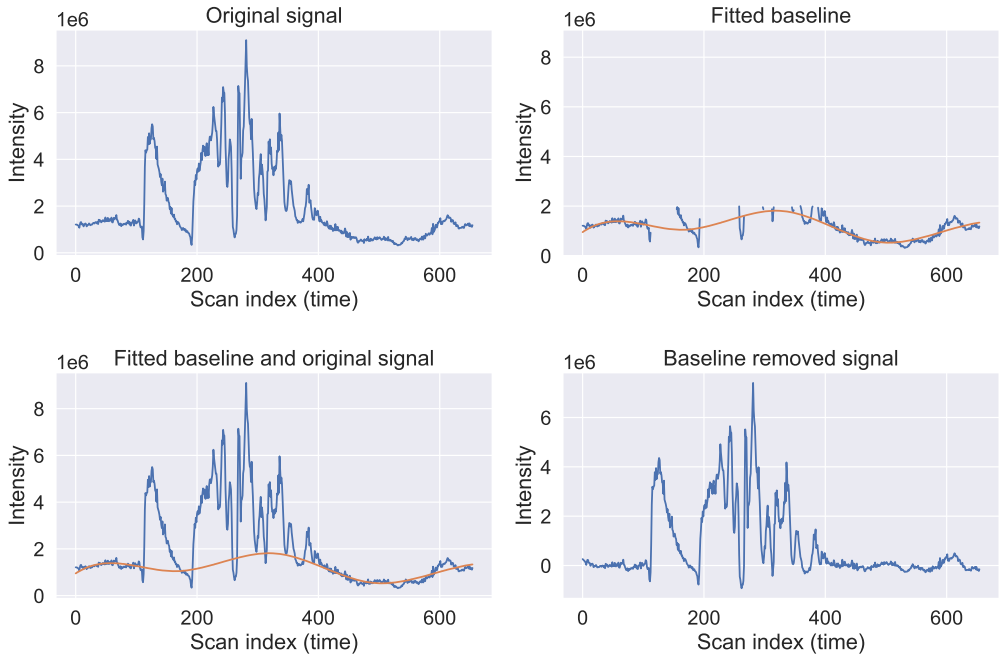


Figure 5.16: Baseline correction example of an MS2 bin with a fragment of the contaminant of Figure 5.10. The top left panel is a graph of the original intensity values of the bin. The top right panel is a graph of the intensity values which belong to the baseline group (after applying a Gaussian mixture model), and the fitted B-splines curve, where the number of knots is five and the degree is three. The bottom left panel is graph of the original signal with the baseline derived from the top right panel. Finally, the bottom right panel is the corrected signal.

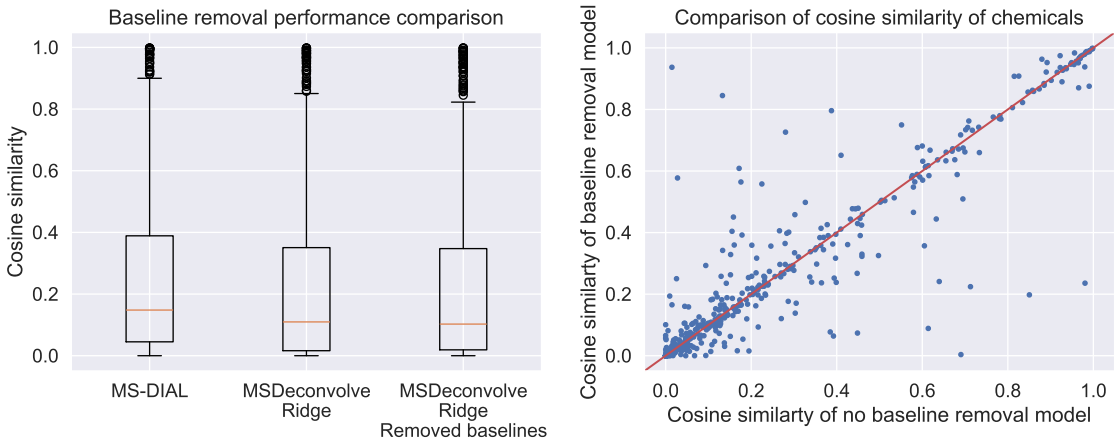


Figure 5.17: Comparison of cosine similarity scores of baseline removal method with Ridge against MS-DIAL for the chemicals in the second SWATH isolation window (left panel). The right panel is a scatter plot of the cosine similarities for the two MSDeconvolve models of the left panel.

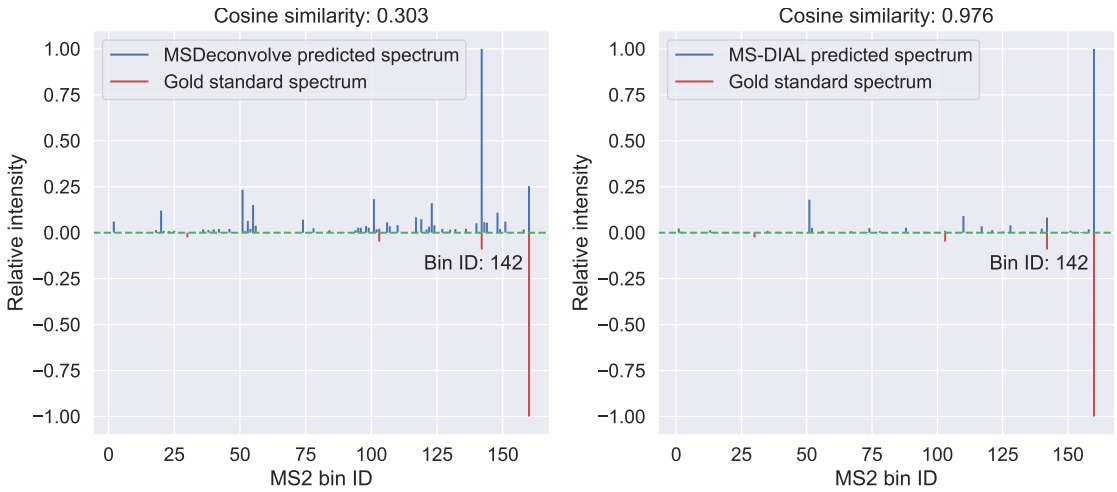


Figure 5.18: Comparison of predicted fragmentation spectrum against the gold standard fragmentation spectrum for a specific peak (with ID 0) for MSDeconvolve (using Ridge and automatically selecting the covariates) on the left panel, and MS-DIAL on the right panel.

and the results are shown in Figure 5.17. We have used Ridge with the automatic covariate selection method (with the non-negative coefficients constraint and using cross validation for hyper-parameter selection), and compare the models with and without removing the baselines. There is not a significant difference between the two box plots, with both of these approaches performing slightly worse than MS-DIAL. To examine the effect of removing the baselines, the scatter plot of the right panel examines the difference in cosine similarities for each chemical between our two models. We observe that most chemicals are scattered around the $y = x$ line, meaning that for most chemicals, the two models perform very similarly. There are a few chemicals that are not close to this line, and have a very high cosine similarity score in one model, while they perform very poorly in the other. This is however only a minority of the chemicals, and this is observed on both sides of the line.

To better understand the reason why MS-DIAL was still outperforming our methods, we chose to investigate specific peaks in more detail, for which our method was performing poorly compared to MS-DIAL. Such an example is shown is Figure 5.18. MSDeconvolve achieves a low cosine similarity score of approximately 0.303, while MS-DIAL achieves a score of 0.976. In the prediction of our method, the MS2 bin with ID 142 is given a very high intensity value, while MS-DIAL makes a more accurate prediction for this bin. This MS2 bin turned out to be the MS2 bin including the main fragment of the contaminant we have been examining from Figure 5.10.

A more detailed inspection of the fitted model for this bin was conducted. More specifically, it was of interest to compare the partial residuals for the covariates with

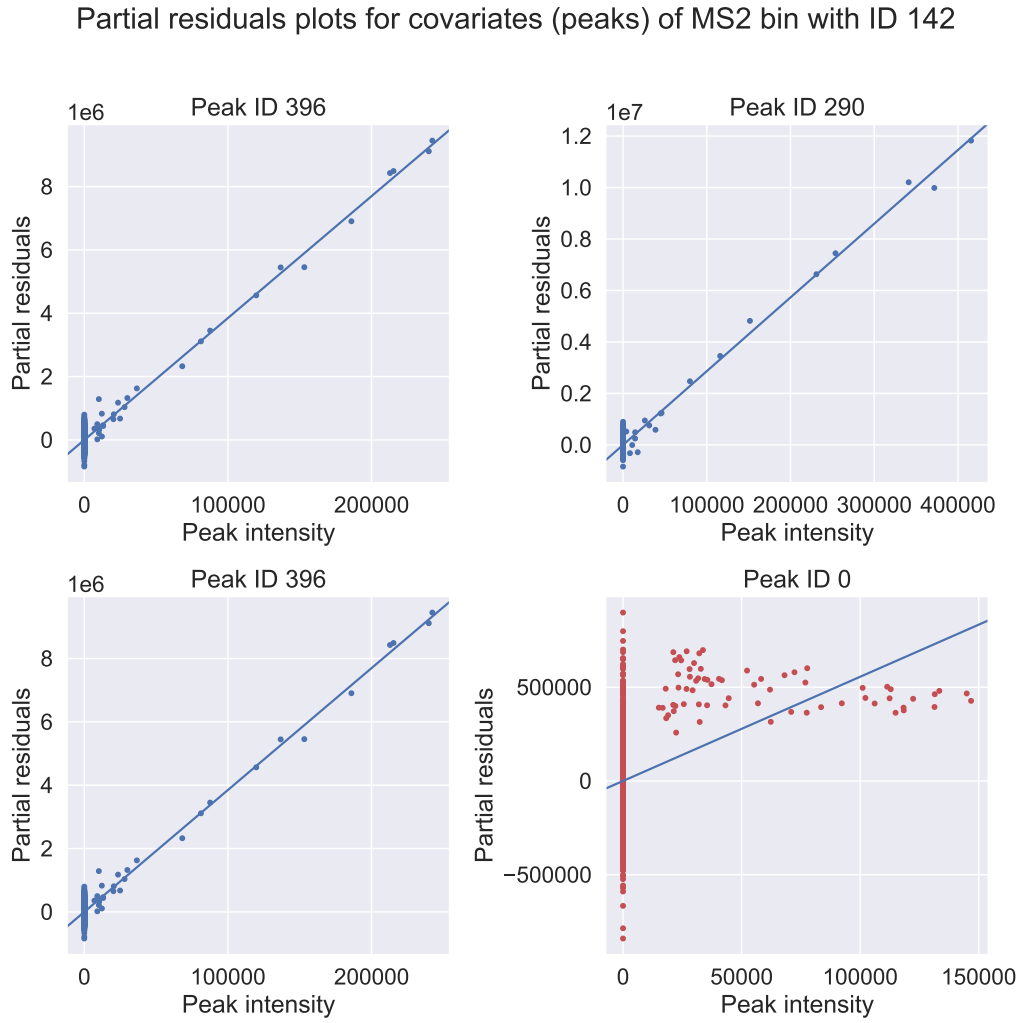


Figure 5.19: Partial residuals plots for model of MS2 bin with ID 142 for the covariates/peaks with the three highest estimated coefficients (blue points), and the peak of Figure 5.18 (panel with red points). For each covariate, a line with a slope equal to the fitted coefficient is also plotted on the data.

the highest estimated coefficients, and observe how these compare with the peak we are examining in Figure 5.18. The partial residuals for the covariates/peaks with the top 3 highest estimated coefficients are shown in Figure 5.19, along with the peak examined in Figure 5.18 in the scatter plot with the red points. The main observation to be made from these plots is that for all other covariates apart from the one we are investigating, a linear relationship appears to be sensible. The lines plotted on the data correspond to lines with slope equal to the fitted coefficient of each covariate. These lines describe the data well. However, the peak we are interested in is the only exception, where there is not a clear linear relationship from the data.

Inspired from these findings, we attempted to further limit the number of peaks

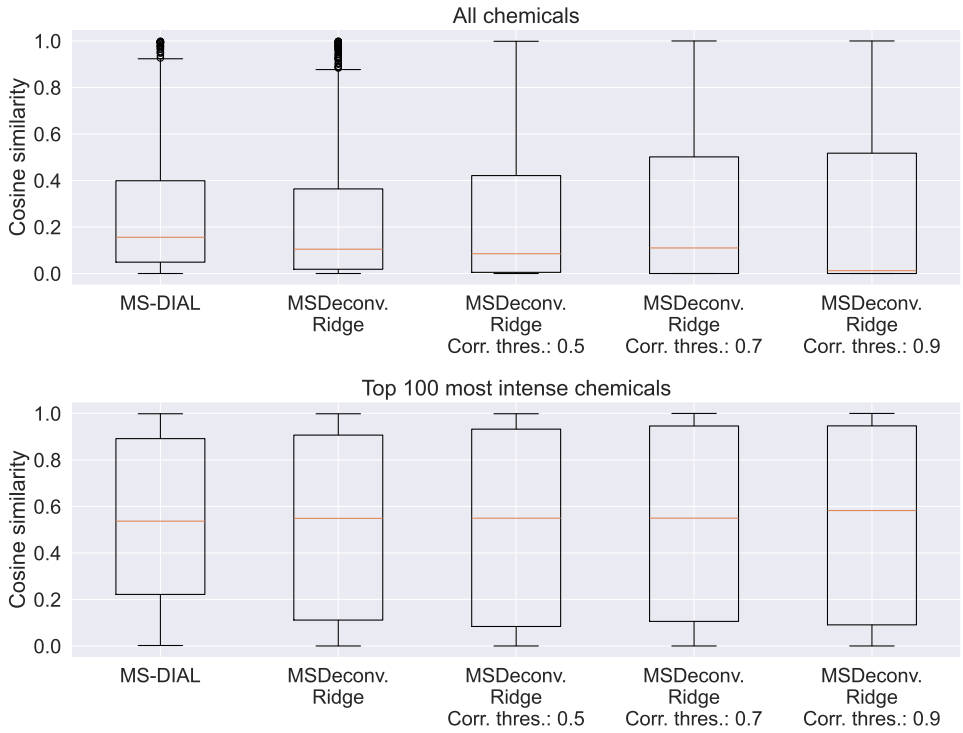


Figure 5.20: Box plots of cosine similarities for MS-DIAL, and MSDeconvolve for models using Ridge and the correlation filter using different correlation threshold values.

selected for each MS2 bin model, by using a correlation filter after the automatic selection procedure introduced in Section 5.2.2. For each selected peak, the data points ($x$ and $y$ pairs, where $y$ is the response) for which the peak has intensity greater than zero is obtained, and the correlation of these data points is calculated. A correlation threshold is then used to discard peaks with a correlation with the response which is lower than that threshold. Various correlation thresholds were tested, and the corresponding results are shown in Figure 5.20. Although there is a visible difference in the box plots with the correlation filter, it is not clear if they are an improvement to our original method. The chemicals that score well do not appear as outliers anymore. However, the median and the lower quartiles appear to shift towards zero. The bottom panel of the figure examines the top 100 most intense peaks, to obtain a better understanding of how these models perform for chemicals that are likely to be important. The differences of the box plots in this panel are minimal, with MS-DIAL still outperforming our method.

### 5.2.4 Ensemble Method

In all models of MSDeconvolve that were attempted, there were always peaks for which MSDeconvolve performed better than MS-DIAL and vice versa. Therefore, combining

the predictions of MSDeconvolve and MS-DIAL could potentially improve the overall performance of cosine similarity scores.

Fragmentation spectra predictions from three methods (two MSDeconvolve methods and MS-DIAL) were combined together in a new 'ensemble' method. For the two MSDeconvolve methods, we have used Ridge regression with and without the correlation filter approach (both with the automatic covariate selection procedure and the non-negative coefficients constraint, and the correlation threshold being set to 0.7, given the positive results of these methods in Figure 5.20).

To combine three predicted fragmentation spectra for a peak, we only include in the final prediction MS2 bins which are included in at lest two predictions, or in all three predictions. For those bins, we obtain the median or mean intensity value.

Results of this ensemble method are displayed in Figure 5.21. We do observe a slight improvement, especially in methods where MS2 bins that are found in the predictions of at least two of the three individual methods are used. This improvement is more obvious in the lower panel for the most intense peaks. We can therefore observe that combining the predictions of MSDeconvolve methods and MS-DIAL could slightly improve results. However, these improvements need to be validated on new independent data to conclude that this method performs better than MS-DIAL alone.

### 5.2.5 Peak Coefficients Additional Constraints

A different approach to the problem that was adopted was focused around the fact that fragments should generally not be greater than the precursor in intensity. Given that coefficients are approximately interpreted as the proportion of the original peak that results into a fragment, it would be appropriate to further restrict those. In our penalised models, we do enforce coefficients to shrink towards zero. However, it is still possible for coefficients to take large values. Therefore, we attempted to use a model which enforces coefficients to lie within the $[0, 1]$ interval. The lower bound was set to zero, since negative coefficients should not be possible, and the upper limit is set to one, to reflect the fact that a fragment's intensity should not surpass the intensity of the original metabolite.

Fitting this new model resulted in the box plot in Figure 5.22, alongside the corresponding box plots of MS-DIAL, and the Ridge approach. At a first glance, it seems that this new method performs better than the previous approach, and is very similar to the performance of MS-DIAL. However, on the right panel where only the cosine similarities of the top 100 most intense peaks are plotted, the previous method is still better.

This idea was then taken a step further, by not only enforcing a fragment to be less or equal than the original metabolite, but rather to encourage the sum of the intensities of the fragments of a metabolite to not surpass the precursor intensity. In terms of modelling

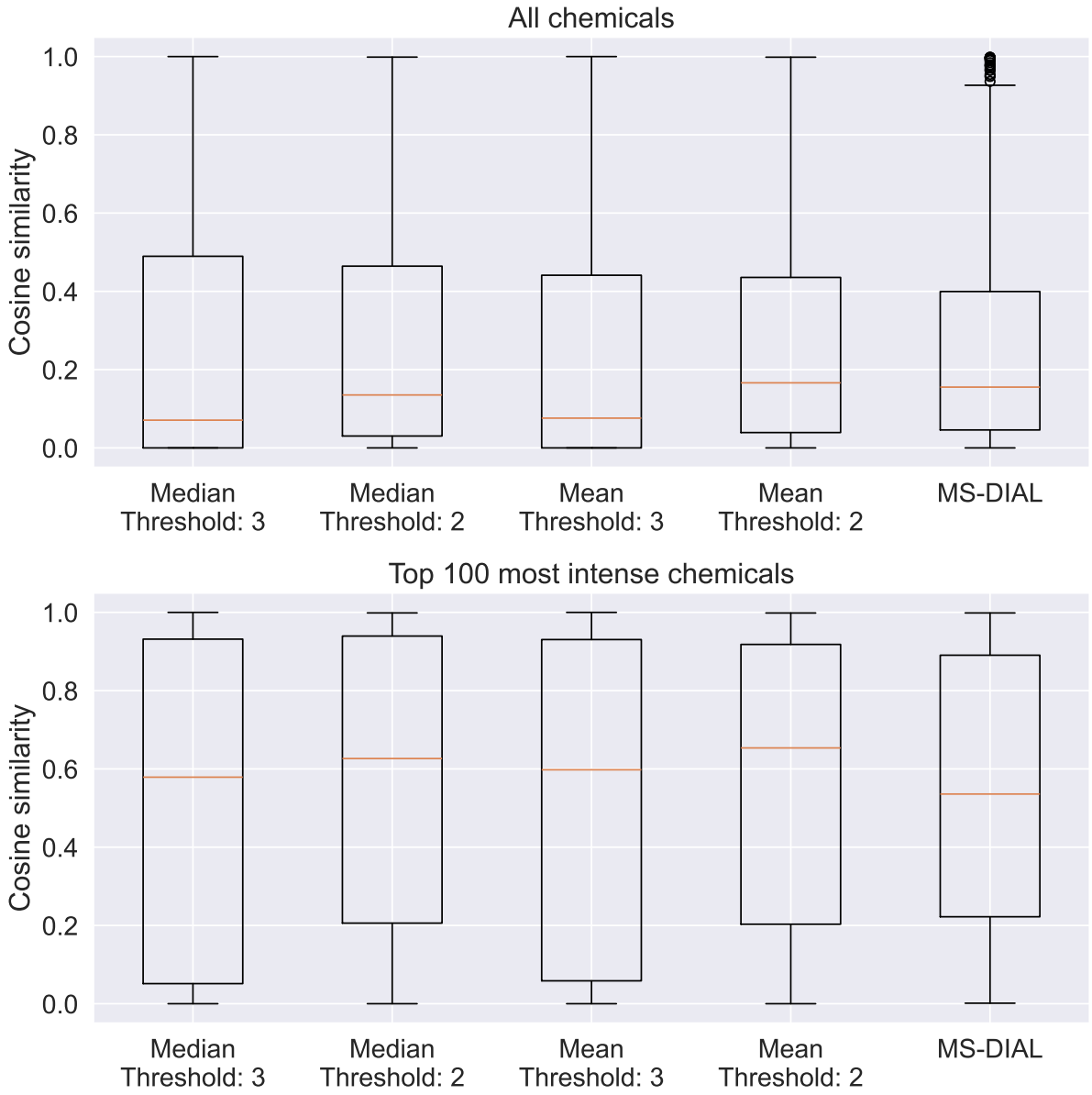


Figure 5.21: Box plots of cosine similarity scores of the combination of three methods: two MSDeconvolve methods and MS-DIAL. For example, the predicted fragmentation spectra of the 'mean, threshold: 2' method contains MS2 bins which appear in at least two of the three individual methods, and the mean intensity value is obtained. The two MSDeconvolve methods use Ridge regression with and without the correlation filter. The top panel is constructed using all peaks, while the bottom panel is constructed for only the top 100 most intense peaks.

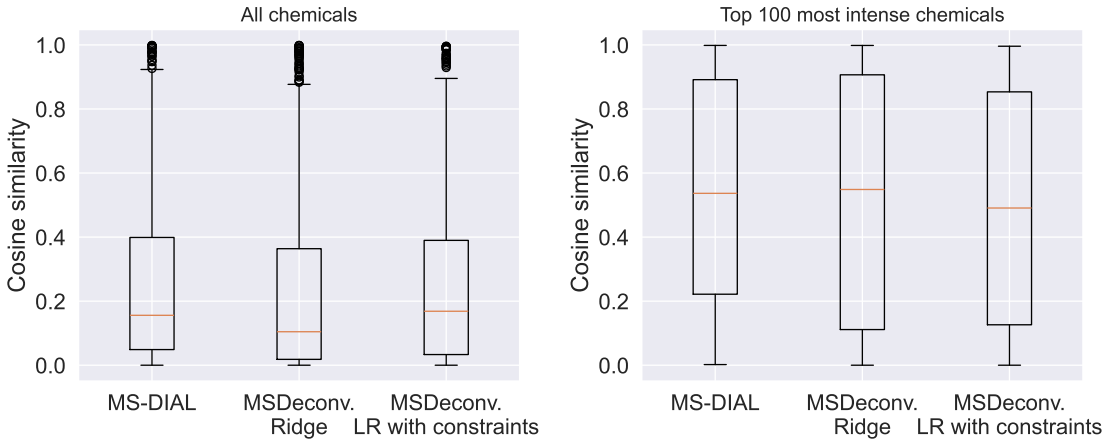


Figure 5.22: Comparison of box plots of cosine similarity scores of MS-DIAL, MDeconvolve using Ridge regression, and MSDeconvolve using linear regression and the $[0, 1]$ bounds for the coefficients. Left panel is constructed using all covariates/peaks, while the right panel is constructed using only the top 100 most intense peaks.

this idea, an L2 penalty was applied to shrink the sum of the coefficients of each peak towards one (Section 4.4, Equation 4.7). This was more complicated than the models we have used so far. This is because, we have used independent models for each MS2 bins, where the covariates are the MS1 peaks. However, in this approach, we need to model all MS2 bins simultaneously, so that we can then apply a penalty to the sum of the coefficients of each MS1 peak across MS2 bins.

In the penalty term of (4.7), $\lambda$ is a hyperparameter that controls the amount of penalty that is added to the loss function. The non-negative constraint is also applied to the coefficients, as with the previous models. Various values for $\lambda$ were used, and the performance results are displayed in Figure 5.23. Unfortunately, we do not observe any of these methods to be improvements to either MS-DIAL or the Ridge approach, which is the best performing method of MSDeconvolve. In both panels of the figure, the cosine similarities of MS-DIAL are still greater than any of the MSDeconvolve methods.

### 5.2.6 Scaling by MS2 Bin Approach

So far in Ridge regression models of MSDeconvolve, we have used cross validation to establish the value of $\alpha$ for each MS2 bin model (which controls the amount of penalty applied to the coefficients, refer to (3.4)). This results in different penalties being applied to the coefficients in each MS2 bin. However, it was of interest to explore the effect of attempting to apply the same penalty in each MS2 bin, which from a Bayesian perspective would correspond to applying the same prior distribution on all coefficients of all MS2 bin models (see Section 3.3).

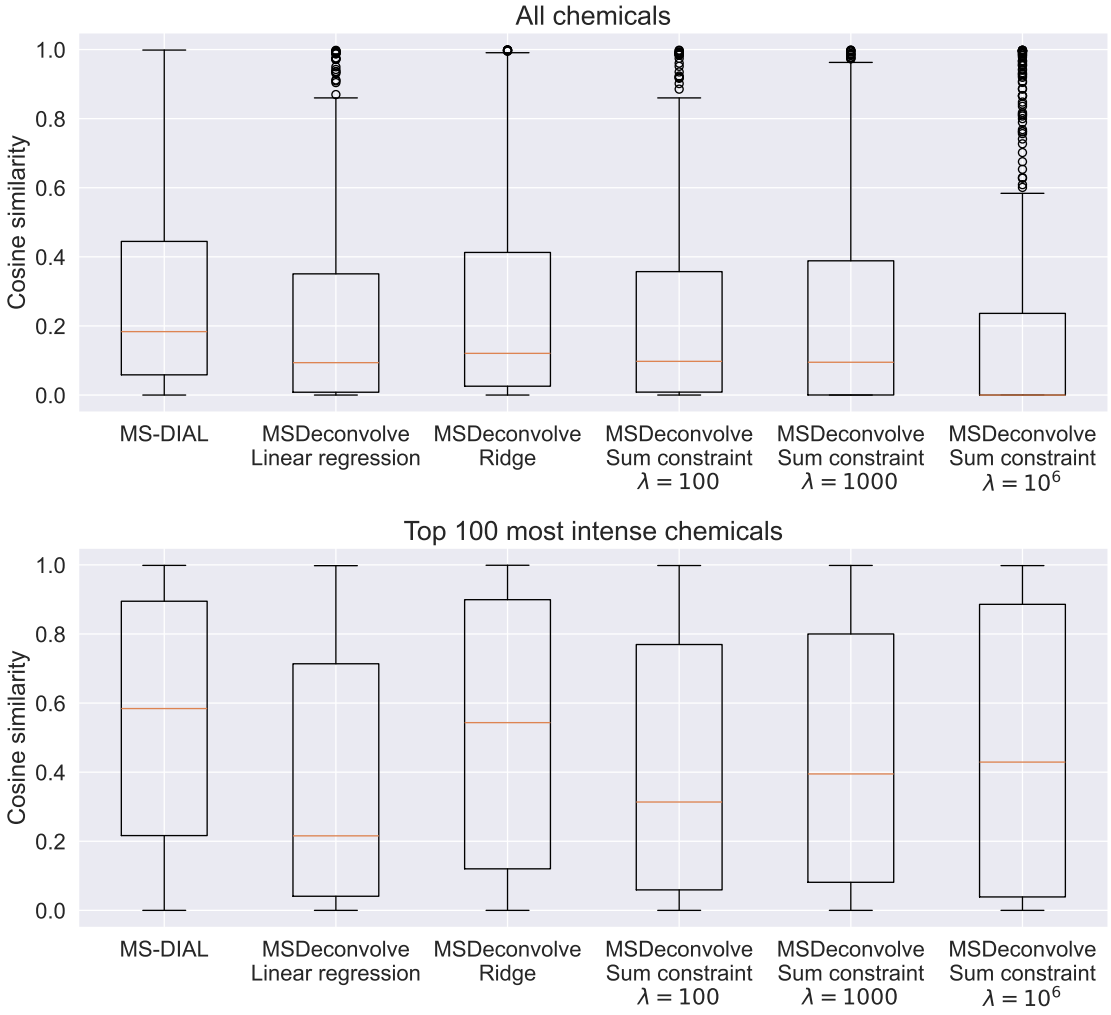


Figure 5.23: Box plots of cosine similarity scores for MS-DIAL and MSDeconvolve using linear regression, Ridge, and the sum to one constraint with different values of $\lambda$. The top panel is constructed using all chemicals, while the bottom panel is constructed using only the top 100 most intense chemicals.

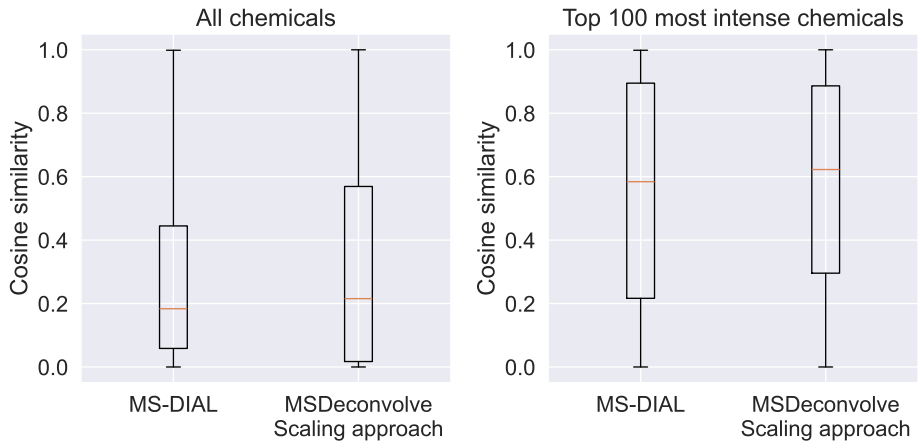


Figure 5.24: Box plot of cosine similarities of scaling approach versus MS-DIAL for all chemicals/peaks and the top 100 most intense peaks.

Thus, for each MS2 bin model, we first filter out covariates/peaks based on the automatic selection criteria introduced in Section 5.2.2 and the correlation filter introduced in Section 5.2.3. We then proceed by scaling both the design matrix and the response vector, in order to make the residuals have variance approximately equal to one. By making residuals have variance close to one, we can use the same value of the penalty coefficient $\alpha$ (refer to Section 3.3) and achieve our goal of applying the same amount of penalty on the coefficients of all MS2 bin models. To calculate the scaling factor, we first fit a linear regression model (with the non-negative constraint) and use the residuals of this model to calculate their standard deviation. This calculated standard deviation is used as an approximation to the standard deviation of the residuals of our model. It is then used as the scaling factor of both the design matrix matrix and the response vector, and this process is repeated for all MS2 bins. We can then fit all models using Ridge regression, and using the same penalty coefficient $\alpha$.

The box plots of the cosine similarity results of this approach can be seen in Figure 5.24. Various values for the penalty coefficient $\alpha$ were tested, and the one which produced the best results ($\alpha = 1000$) were used in the final comparison. We can observe in the left panel a better performance compared to MS-DIAL, with a median cosine similarity score which is slightly higher, and the upper third quartile also being higher. In the right panel, where the top 100 most intense peaks are plotted however, the cosine similarity scores of the two methods are more similar. Despite the positive results of this approach, they need to be validated on new independent test data before drawing final conclusions.

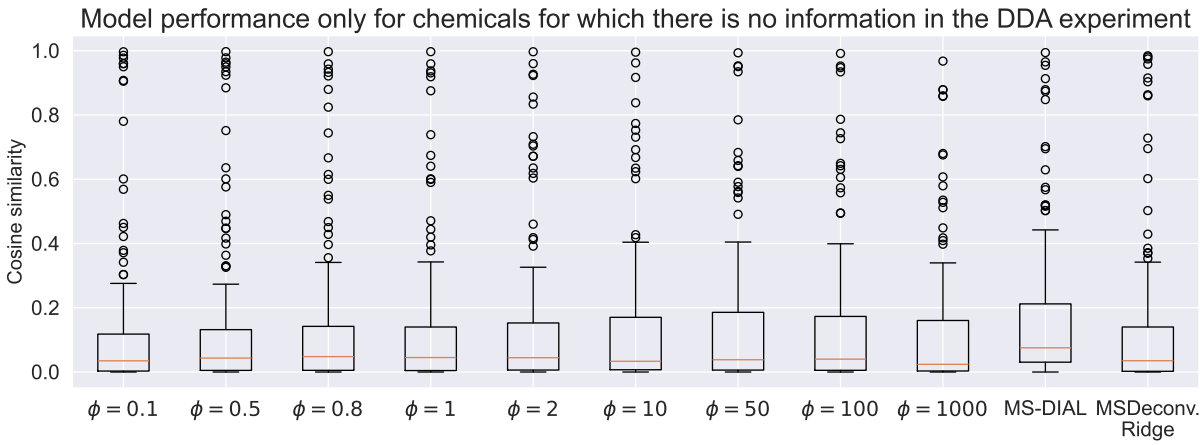


Figure 5.25: Box plots of models with combination of DDA and DIA information. The box plots were constructed for only the chemicals for which there was no MS2 information in the DDA experiment, for different values of $\phi$ (refer to (4.8)).

### 5.2.7 DDA and DIA Data Combination

On top of the methods we had attempted so far, we were also interested in answering the following question; in a scenario where a sample was used for both a DDA and a DIA experiment, would the combination of data from both experiments improve the deconvolution performance for peaks which were not fragmented in the DDA experiment? We therefore used one of the non-intensity overlap DDA experiment results in combination with the SWATH DIA experiment results we have been using so far, to attempt to answer this question.

In order to incorporate this extra information in MSDeconvolve models, we have used (4.8), which was introduced in Section 4.4. An evaluation of this approach was conducted where various values of $\phi$ were attempted, which is the hyper-parameter that controls the influence of the DDA data on coefficient estimation. Results were used to construct the box plots of Figure 5.25. The box plots were constructed for only the chemicals for which there was no MS2 information in the DDA experiment, in order to determine whether the extra information assists with the deconvolution of unknown chemicals. We do observe a possible improvement for $\phi = 50$ compared to our method with Ridge. However, this improvement is not enough to make this a better approach than MS-DIAL, even when MS-DIAL does not have the extra DDA information. It should be noted however, that the chemicals that were used to construct these box plots are generally given low cosine similarity scores anyway. This is no surprise, given that in the DDA experiment, where a Top-N fragmentation strategy is used, the metabolites with higher intensities are prioritised, and thus the peaks that are left for deconvolution from the DIA experiment

are low intensity peaks, and are more likely to be of low quality.

# Chapter 6

# Discussion

## 6.1 Conclusions and Limitations

In metabolomics LC-MS/MS experiments, the identification of metabolites is often of great importance. A common approach in identifying metabolites is to obtain their fragmentation pattern, which can then be used to find a match in a metabolite database. In DDA experiments, the fragmentation patterns are of high quality, and thus the identification of metabolites is reliable. However, not all metabolites are fragmented in DDA experiments. In DIA experiments, all metabolites are fragmented, and fragmentation patterns of multiple metabolites can appear in the same scan. Therefore, fragmentation patterns in DIA experiments require deconvolution prior to searching a database for a possible match. Deconvolution algorithms however do not perform at the required level to make DIA experiments a reliable alternative to DDA.

In this project, a new deconvolution framework has been developed for metabolomics LC-MS/MS experiment data called MSDeconvolve. MSDeconvolve uses peaks that are identified in the experiment data to construct a design matrix. The design matrix contains the intensity values of peaks, where the columns of the design matrix correspond to the peaks, and the rows of the design matrix correspond to the MS1 scans of the experiment. For MS2 scans, the m/z axis is split into equal-width bins, and intensity values within each MS2 bin are summed. Then, for each MS2 bin, a vector of intensity values is created across MS2 scans. By matching MS1 with MS2 scans, MSDeconvolve can fit regression models between each MS2 bin vector and the design matrix.

In each of those MS2 bin models, we obtain parameter estimates for all peaks. By interpreting each of those parameters as the proportion of the original peak that results in a fragment in the current MS2 bin, then we can use them to reconstruct the fragmentation spectrum of each peak. This reconstruction can be achieved for any peak by obtaining

the parameter estimates of that peak from all MS2 bin models, and multiply them with the maximum intensity of the peak. This deconvoluted fragmentation spectrum can then be used for identification through a database lookup.

Various regression methods and approaches were hypothesised to perform well and were first evaluated on simulated datasets, before attempting to validate these results on the real dataset. The ViMMS software was used for the generation of the simulated datasets [39, 40]. Apart from the simulated DIA experiment data, ViMMS provides access to the true fragmentation spectra of the simulated metabolites. Thus, evaluation of MSDeconvolve methods was possible through the use of the cosine similarity scores between MSDeconvolve predicted fragmentation spectra and true fragmentation spectra. Lasso regression was initially expected to perform well given its inherent variable selection property, which was desirable given that an MS2 bin is not expected to contain fragments from all peaks. However, other types of penalties were also explored, which were Ridge regression and Elastic Net, as well as using multiple linear regression without any penalty added to the objective function (Sections 5.1.1, 5.1.2 and 5.1.4). A noise filtering technique was also attempted, to remove spike noise and facilitate the estimation of regression parameters (Section 5.1.3).

The deconvolution methods of MSDeconvolve perform very well compared to MS-DIAL, the current state-of-the-art deconvolution software [38], on the simulated data. Unfortunately, these positive results were not reflected on the real dataset, due to the high complexity of metabolomics and mass spectrometry data, and the consequent difficulty in modelling it (Sections 5.2.1 and 5.2.2). The real dataset was obtained from McBride et al. [19], which provides multiple DDA runs of the same sample in order to capture the fragmentation pattern of almost all peaks in the sample. These fragmentation spectra were then used as the gold standard fragmentation spectra of peaks to compare with MSDeconvolve predicted spectra.

Different approaches were developed for the real dataset to obtain results similar to those of MS-DIAL. These methods include Ridge regression with a correlation filter (Section 5.2.3), applying a new constraint and L2 penalty to reflect the fact that the intensity of fragment ions should generally not exceed the precursor ion intensity (Section 5.2.5) and combining information from both a DDA and a DIA experiment to improve the deconvolution of peaks that were not fragmented in the DDA experiment (Section 5.2.7). We were also able to obtain results that were slightly better than MS-DIAL on the real dataset, such as the scaling method (Section 5.2.6) and the ensemble method (Section 5.2.4). However, these results need to be evaluated further on new independent datasets to draw definitive conclusions about MSDeconvolve performing better than MS-DIAL. Even if these results are validated on new data, there are still improvements required for

MSDeconvolve methods to make the identification of metabolites through deconvolution of DIA data a reliable approach compared to DDA methods.

There are many possible factors that could be affecting the performance of MSDeconvolve. A possible factor is the fact that peak picking algorithms cannot be completely accurate, given that there is not a clear definition between an actual metabolite peak and noise. Therefore, since MSdeconvolve uses detected peaks as covariates, different peak picking algorithms can result in different design matrices for our models, inevitably affecting results. Apart from including all metabolite peaks however, there is also the problem of background noise not belonging to a peak, which can also result in fragments in MS2 scans (such as the case of contaminants examined in Section 5.2.1).

More generally, it is important to investigate into more detail the noise present in these experiments, and how to better model it. We have assumed an additive relationship between the response variable and the covariates with constant noise, which might not be appropriate. However, noise that is a result of signals outside of peaks might significantly limit the effectiveness of MSDeconvolve.

## 6.2 Improvements and Future Work

To extend MSDeconvolve, a new method could be added to handle multiple DIA samples, similar to the approach of CorrDec in MS-DIAL [33]. We would expect approximately the same peaks to be present in each sample, but the noise could be different, individual to each experimental run. We could thus better differentiate signals from noise, and improve deconvolution performance. Furthermore, additional techniques could be incorporated into the workflow of MSDeconvolve to eliminate background noise, such as the use of blank samples as discussed in Section 2.3.3, and the use of similar signal patterns of contaminants in MS1 and MS2 scans of DIA experiments.

Upon promising results of MSDeconvolve, a further extension could be to develop a new data acquisition method, as has been done in DDA methods [19]. Rather than being restricted to the choice between DDA and DIA methods, and thus between sacrificing either metabolite coverage or fragmentation spectra quality, a decision between fragmenting a single ion or an m/z range in each MS2 scan can be made in real time, based on the available MS1 scans. This new dynamic data acquisition method would be a more balanced approach and allow for a more flexible compromise between metabolite coverage and fragmentation spectra quality.

The effectiveness of a Bayesian modelling framework for MSDeconvolve models could also be explored. Currently, MSDeconvolve only obtains coefficient estimates. However, with the use of prior and posterior distributions on the coefficients we could achieve two

different results. The first one is that by deriving informative priors, we could improve the accuracy of coefficient estimates. Furthermore, by obtaining posterior distributions for coefficient estimates, we could quantify the uncertainty in our results, as well as quantify the uncertainty in predicted fragments of predicted fragmentation spectra.

Finally, another potential direction for future work is the experimentation with different data acquisition parameters and their effect on deconvolution performance. Data acquisition parameters are determined prior to performing a metabolomics LC-MS experiment, and therefore potentially not set optimally. Controlling factors such as the m/z boundaries of SWATH windows in DIA SWATH experiments, mass accuracy, and scan resolution in a statistically optimal manner could improve performance of the deconvolution methods.

# Bibliography


[1] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069. URL https://doi.org/10.1137/0916069. 25

[2] Yang Chen, En-Min Li, and Li-Yan Xu. Guide to Metabolomics Analysis: A Bioinformatics Workflow. *Metabolites*, 12(4):357, April 2022. ISSN 2218-1989. doi: 10.3390/metabo12040357. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9032224/. 1, 9, 12, 14

[3] Katyeny Manuela Da Silva, Maria Van De Lavoir, Rani Robeyns, Elias Iturrospe, Lisa Verheggen, Adrian Covaci, and Alexander L. N. Van Nuijs. Guidelines and considerations for building multidimensional libraries for untargeted MS-based metabolomics. *Metabolomics*, 19(1):4, December 2022. ISSN 1573-3890. doi: 10.1007/s11306-022-01965-w. URL https://link.springer.com/10.1007/s11306-022-01965-w. 14

[4] Vinny Davies, Joe Wandy, Stefan Weidt, Justin J. J. Van Der Hooft, Alice Miller, Rónán Daly, and Simon Rogers. Rapid Development of Improved Data-Dependent Acquisition Strategies. *Analytical Chemistry*, 93(14):5676–5683, April 2021. ISSN 0003-2700, 1520-6882. doi: 10.1021/acs.analchem.0c03895. URL https://pubs.acs.org/doi/10.1021/acs.analchem.0c03895. 9, 10, 49

[5] Alysha M. De Livera, Daniel A. Dias, David De Souza, Thusitha Rupasinghe, James Pyke, Dedreia Tull, Ute Roessner, Malcolm McConville, and Terence P. Speed. Normalizing and Integrating Metabolomics Data. *Analytical Chemistry*, 84(24):10768–10776, December 2012. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac302748b. URL https://pubs.acs.org/doi/10.1021/ac302748b. 14

[6] Danuta Dudzik, Cecilia Barbas-Bernardos, Antonia García, and Coral Barbas. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. *Journal of Pharmaceutical and Biomedical Analysis*, 147:149–173, January

2018. ISSN 07317085. doi: 10.1016/j.jpba.2017.07.044. URL https://linkinghub.elsevier.com/retrieve/pii/S0731708517315911. 12

[7] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16(4):299–302, April 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0344-8. URL https://www.nature.com/articles/s41592-019-0344-8. 7, 9, 14

[8] Madeleine Ernst, Kyo Bin Kang, Andrés Mauricio Caraballo-Rodríguez, Louis-Felix Nothias, Joe Wandy, Christopher Chen, Mingxun Wang, Simon Rogers, Marnix H. Medema, Pieter C. Dorrestein, and Justin J.J. Van Der Hooft. MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites*, 9(7):144, July 2019. ISSN 2218-1989. doi: 10.3390/metabo9070144. URL https://www.mdpi.com/2218-1989/9/7/144. 9

[9] Jian Guo and Tao Huan. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Analytical Chemistry*, 92(12):8072–8080, June 2020. ISSN 0003-2700, 1520-6882. doi: 10.1021/acs.analchem.9b05135. URL https://pubs.acs.org/doi/10.1021/acs.analchem.9b05135. 1, 2, 9

[10] Jian Guo and Tao Huan. Evaluation of significant features discovered from different data acquisition modes in mass spectrometry-based untargeted metabolomics. *Analytica Chimica Acta*, 1137:37–46, November 2020. ISSN 0003-2670. doi: 10.1016/j.aca.2020.08.065. URL https://www.sciencedirect.com/science/article/pii/S0003267020309120. 1, 2, 9

[11] Xinghua Guo, Andries P. Bruins, and Thomas R. Covey. Characterization of typical chemical background interferences in atmospheric pressure ionization liquid chromatography-mass spectrometry. *Rapid Communications in Mass Spectrometry*, 20(20):3145–3150, October 2006. ISSN 0951-4198, 1097-0231. doi: 10.1002/rcm.2715. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/rcm.2715. 12

[12] Netti Herawati. Regularized Multiple Regression Methods to Deal with Severe Multicollinearity. *International Journal of Statistics and Applications*, 8(4):167–172, May 2018. doi: 10.5923/j.statistics.20180804.02. 21

[13] Karsten Hiller, Jasper Hangebrauk, Christian Jäger, Jana Spura, Kerstin Schreiber, and Dietmar Schomburg. MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis. *Analytical Chemistry*, 81(9):3429–3439, May 2009. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac802689c. URL https://pubs.acs.org/doi/10.1021/ac802689c. 16

[14] Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, February 1970. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.1970.10488634. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634. 2, 20, 22

[15] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):1–18, February 2021. doi: 10.1371/journal.pcbi.1008724. URL https://doi.org/10.1371/journal.pcbi.1008724. 28

[16] International Bureau of Weights and Measures. The Internation System of Units (SI), 2019. URL https://www.bipm.org/documents/20126/41483022/SI-Brochure-9.pdf/fcf090b2-04e6-88cc-1149-c3e029ad8232. 4

[17] Xiaojing Liu and Jason W. Locasale. Metabolomics: A Primer. *Trends in Biochemical Sciences*, 42(4):274–284, April 2017. ISSN 0968-0004. doi: 10.1016/j.tibs.2017.01.004. URL https://www.sciencedirect.com/science/article/pii/S096800041730018X. 4

[18] Christina Ludwig, Ludovic Gillet, George Rosenberger, Sabine Amon, Ben C Collins, and Ruedi Aebersold. Data-independent acquisition-based \textlessspan style="font-variant:small-caps;"\textgreaterSWATH\textless/span\textgreater - \textlessspan style="font-variant:small-caps;"\textgreaterMS\textless/span\textgreater for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 14(8):e8126, August 2018. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20178126. URL https://www.embopress.org/doi/10.15252/msb.20178126. 10

[19] Ross McBride, Joe Wandy, Stefan Weidt, Simon Rogers, Vinny Davies, Rónán Daly, and Kevin Bryson. TopNEXt: automatic DDA exclusion framework for multi-sample mass spectrometry experiments. *Bioinformatics*, 39(7):btad406, June 2023. ISSN 1367-4803. doi: 10.1093/bioinformatics/btad406. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10336026/. 10, 18, 34, 68, 69

[20] Arsenty D. Melnikov, Yuri P. Tsentalovich, and Vadim V. Yanshole. Deep Learning for the Precise Peak Detection in High-Resolution LC–MS Data. *Analytical Chemistry*, 92(1):588–592, January 2020. ISSN 0003-2700, 1520-6882. doi: 10.1021/acs.analchem.9b04811. URL https://pubs.acs.org/doi/10.1021/acs.analchem.9b04811. 15

[21] J. Rafael Montenegro-Burke, Carlos Guijas, and Gary Siuzdak. METLIN: A Tandem Mass Spectral Library of Standards. In Shuzhao Li, editor, *Computational Methods and Data Analysis for Metabolomics*, pages 149–163. Springer US, New York, NY, 2020. ISBN 978-1-07-160239-3. doi: 10.1007/978-1-0716-0239-3_9. URL https://doi.org/10.1007/978-1-0716-0239-3_9. 9, 14

[22] Abzer K. Pakkir Shah, Axel Walter, Filip Ottosson, Francesco Russo, Marcelo Navarro-Diaz, Judith Boldt, Jarmo-Charles J. Kalinski, Eftychia Eva Kontou, James Elofson, Alexandros Polyzois, Carolina González-Marín, Shane Farrell, Marie R. Aggerbeck, Thapanee Pruksatrakul, Nathan Chan, Yunshu Wang, Magdalena Pöchhacker, Corinna Brungs, Beatriz Cámara, Andrés Mauricio Caraballo-Rodríguez, Andres Cumsille, Fernanda De Oliveira, Kai Dührkop, Yasin El Abiead, Christian Geibel, Lana G. Graves, Martin Hansen, Steffen Heuckeroth, Simon Knoblauch, Anastasiia Kostenko, Mirte C. M. Kuijpers, Kevin Mildau, Stilianos Papadopoulos Lambidis, Paulo Wender Portal Gomes, Tilman Schramm, Karoline Steuer-Lodd, Paolo Stincone, Sibgha Tayyab, Giovanni Andrea Vitale, Berenike C. Wagner, Shipei Xing, Marquis T. Yazzie, Simone Zuffa, Martinus De Kruijff, Christine Beemelmanns, Hannes Link, Christoph Mayer, Justin J. J. Van Der Hooft, Tito Damiani, Tomáš Pluskal, Pieter Dorrestein, Jan Stanstrup, Robin Schmid, Mingxun Wang, Allegra Aron, Madeleine Ernst, and Daniel Petras. Statistical analysis of feature-based molecular networking results from non-targeted metabolomics data. *Nature Protocols*, September 2024. ISSN 1754-2189, 1750-2799. doi: 10.1038/s41596-024-01046-3. URL https://www.nature.com/articles/s41596-024-01046-3. 12

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.48550/ARXIV.1201.0490. URL https://arxiv.org/abs/1201.0490. 24

[24] Tomáš Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1):395, December 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-395. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-395. 15

[25] Douglas A. Reynolds. Gaussian Mixture Models. In *Encyclopedia of Biometrics*, volume 741, pages 659–663. 2009. 24

[26] Robin Schmid, Steffen Heuckeroth, Ansgar Korf, Aleksandr Smirnov, Owen Myers, Thomas S. Dyrlund, Roman Bushuiev, Kevin J. Murray, Nils Hoffmann, Miaoshan Lu, Abinesh Sarvepalli, Zheng Zhang, Markus Fleischauer, Kai Dührkop, Mark Wesner, Shawn J. Hoogstra, Edward Rudt, Olena Mokshyna, Corinna Brungs, Kirill Ponomarov, Lana Mutabdžija, Tito Damiani, Chris J. Pudney, Mark Earll, Patrick O. Helmer, Timothy R. Fallon, Tobias Schulze, Albert Rivas-Ubach, Aivett Bilbao, Henning Richter, Louis-Félix Nothias, Mingxun Wang, Matej Orešič, Jing-Ke Weng, Sebastian Böcker, Astrid Jeibmann, Heiko Hayen, Uwe Karst, Pieter C. Dorrestein, Daniel Petras, Xiuxia Du, and Tomáš Pluskal. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology*, 41(4):447–449, April 2023. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-023-01690-2. URL https://www.nature.com/articles/s41587-023-01690-2. 15

[27] Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 78 (3):779–787, February 2006. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac051437y. URL https://pubs.acs.org/doi/10.1021/ac051437y. 15

[28] Rob Smith, Andrew D. Mathis, Dan Ventura, and John T. Prince. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC Bioinformatics*, 15(7):S9, May 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-S7-S9. URL https://doi.org/10.1186/1471-2105-15-S7-S9. 15

[29] Stephen E. Stein and Donald R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, September 1994. ISSN 1879-1123. doi: 10.1016/1044-0305(94)87009-8. URL https://doi.org/10.1016/1044-0305(94)87009-8. 27

[30] Paolo Stincone, Abzer K. Pakkir Shah, Robin Schmid, Lana Graves, Stilianos P. Lambidis, Ralph Torres, Shu-Ning Xia, Vidit Minda, Allegra Aron, Mingxun Wang, Chambers C. Hughes, and Daniel Petras. Evaluation of Data Dependent MS/MS Acquisition Parameters for Non-targeted Metabolomics and Molecular Networking of Environmental Samples - Focus on the Q Exactive Platform, March 2023. URL https://chemrxiv.org/engage/chemrxiv/article-details/6416f0bfaad2a62ca1017b3f. 1

[31] Marc Sturm and Oliver Kohlbacher. TOPPView: An Open-Source Viewer for Mass Spectrometry Data. *Journal of Proteome Research*, 8(7):3760–3763, July 2009. ISSN 1535-3893, 1535-3907. doi: 10.1021/pr900171m. URL https://pubs.acs.org/doi/10.1021/pr900171m. 6

[32] Jun Sun and Yinglin Xia. Pretreating and normalizing metabolomics data for statistical analysis. *Genes & Diseases*, 11(3):100979, May 2024. ISSN 2352-3042. doi: 10.1016/j.gendis.2023.04.018. URL https://www.sciencedirect.com/science/article/pii/S2352304223002246. 13, 14

[33] Ipputa Tada, Romanas Chaleckis, Hiroshi Tsugawa, Isabel Meister, Pei Zhang, Nikolaos Lazarinis, Barbro Dahlén, Craig E. Wheelock, and Masanori Arita. Correlation-Based Deconvolution (CorrDec) To Generate High-Quality MS2 Spectra from Data-Independent Acquisition in Multisample Studies. *Analytical Chemistry*, 92(16):11310–11317, August 2020. ISSN 0003-2700, 1520-6882. doi: 10.1021/acs.analchem.0c01980. URL https://pubs.acs.org/doi/10.1021/acs.analchem.0c01980. 15, 16, 69

[34] Ralf Tautenhahn, Christoph Böttcher, and Steffen Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9(1):504, December 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-504. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-504. 15

[35] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, December 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. 2, 20

[36] J. Trygg, J. Gabrielsson, and T. Lundstedt. Background Estimation, Denoising, and Preprocessing. In Steven Brown, Romà Tauler, and Beata Walczak, editors, *Comprehensive Chemometrics (Second Edition)*, pages 137–141.

Elsevier, Oxford, January 2009. ISBN 978-0-444-64166-3. doi: 10.1016/B978-0-444-64165-6.02022-X. URL https://www.sciencedirect.com/science/article/pii/B978044464165602022X. 13

[37] Martin Trötzmüller, Xinghua Guo, Alexander Fauland, Harald Köfeler, and Ernst Lankmayr. Characteristics and origins of common chemical noise ions in negative ESI LC–MS. *Journal of Mass Spectrometry*, 46(6):553–560, June 2011. ISSN 1076-5174, 1096-9888. doi: 10.1002/jms.1924. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jms.1924. 12

[38] Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, 12(6):523–526, June 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3393. URL https://www.nature.com/articles/nmeth.3393. 2, 15, 16, 27, 68

[39] Joe Wandy, Vinny Davies, Justin J. J. van der Hooft, Stefan Weidt, Rónán Daly, and Simon Rogers. In Silico Optimization of Mass Spectrometry Fragmentation Strategies in Metabolomics. *Metabolites*, 9(10):219, October 2019. ISSN 2218-1989. doi: 10.3390/metabo9100219. URL https://www.mdpi.com/2218-1989/9/10/219. 2, 17, 43, 49, 68

[40] Joe Wandy, Vinny Davies, Ross McBride, Stefan Weidt, Simon Rogers, and Ronan Daly. ViMMS 2.0: A framework to develop, test and optimise fragmentation strategies in LC-MS metabolomics. *Journal of Open Source Software*, 7 (71), March 2022. ISSN 2475-9066. doi: 10.21105/joss.03990. URL https://eprints.gla.ac.uk/268790/. 2, 17, 43, 68

[41] Joe Wandy, Ross McBride, Simon Rogers, Nikolaos Terzis, Stefan Weidt, Justin J. J. van der Hooft, Kevin Bryson, Rónán Daly, and Vinny Davies. Simulated-to-real benchmarking of acquisition methods in untargeted metabolomics. *Frontiers in Molecular Biosciences*, 10, 2023. ISSN 2296-889X. URL https://www.frontiersin.org/articles/10.3389/fmolb.2023.1130781. 1, 2, 9

[42] David S Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L Lee, Mark Berjanskii, Robert Mah, Mai Yamamoto, Juan Jovel, Claudia Torres-Calzada, Mickel Hiebert-Giesbrecht, Vicki W Lui, Dorna Varshavi, Dorsa Varshavi, Dana Allen, David Arndt,

Nitya Khetarpal, Aadhavya Sivakumaran, Karxena Harford, Selena Sanford, Kristen Yee, Xuan Cao, Zachary Budinski, Jaanus Liigand, Lun Zhang, Jiamin Zheng, Rupasri Mandal, Naama Karu, Maija Dambrova, Helgi B Schiöth, Russell Greiner, and Vasuk Gautam. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1062. URL https://doi.org/10.1093/nar/gkab1062. 9, 14

[43] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1), March 2008. ISSN 1932-6157. doi: 10.1214/07-AOAS147. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-1/Coordinate-descent-algorithms-for-lasso-penalized-regression/10.1214/07-AOAS147.full. 24

[44] Bin Zhou, Jun Feng Xiao, Leepika Tuli, and Habtom W. Ressom. LC-MS-based metabolomics. *Molecular bioSystems*, 8(2):470–481, February 2012. ISSN 1742-206X. doi: 10.1039/c1mb05350g. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3699692/. 15

[45] Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67 (2):301–320, April 2005. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL https://academic.oup.com/jrsssb/article/67/2/301/7109482. 2, 21