IMPERIAL COLLEGE LONDON

DEPARTMENT OF SURGERY AND CANCER

COMPUTATIONAL AND SYSTEMS MEDICINE

An improved pipeline for LC-MS spectral processing and annotation

Elzbieta Lauzikaite

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London
June 29, 2020

Copyright Declaration

Declaration of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Abstract

Mass spectrometry coupled to liquid chromatography (LC-MS) is routinely used for metabolomics studies. While steps in data acquisition are fairly standardised and automated, structural metabolite identification still depends on manual curation and expert knowledge, forming a major bottleneck in LC-MS based pipelines. The work presented in this thesis represents a novel data processing strategy, which aids metabolite identification through deliberate us of the the correlation structure that exists between spectral features, as well as chromatographic profile and data acquisition order. This strategy aligns features originating from the same chemical entity across all samples as a group, ensuring that chemically-related features are accurately aligned despite fluctuations in the chromatographic and mass spectrometric measurements occurring during the experimental run time. Spectral features aligned in this way are consequently matched to in-house chemical standards databases more efficiently and accurately, on account of the retained and chemically-relevant spectral information. This pipeline has been developed and is presented as an open-source R package - massFlowR. This thesis demonstrates the utility of massFlowR with simulated data, as well as an open-source urine metabolomics study DEVSET, and a large-scale cohort study AIRWAVE, where the performance of massFlowR is compared with the widely-used package XCMS.

Acknowledgements

I would like to express my gratitude to my primary supervisor Dr Toby Athersuch for his enthusiastic support during the preparation of this thesis. I would also like to thank my secondary supervisors Dr Isabel Garcia-Perez, Dr Ioanna Tzoulaki and Professor Paul Elliott. I would like to acknowledge the funding that I received from the MRC-PHE Centre for Environment and Health.

A big thanks is due to Dr Jake Pearce and Dr Matthew Lewis for their scientific guidance and invaluable discussions that have led me on the right path. I am indebted to the whole Imperial Phenome Centre team, especially Elena Chekmeneva, Stephane Camuzeaux, Caroline Sands, Maria Gomez Romero, who have taught me so much about metabolite identification, to Gonçalo Correia and Benjamin Cooper, who have also contributed to the data analysis.

My friends from all around Europe have given me great support, as have my amazing colleagues in London. Above all, I am extremely grateful to my family at home.

This one is for the three greatest women in my life, Emilija, Raimonda and Janina.

Contents

List of Tables

List of Figures

Abbreviations

m/z	mass-to-charge ratio
APCI	atmospheric pressure chemical ionisation
BEH	bridged ethylsiloxane/silica hybrid
BPI	base peak intensity
EIC	extracted ion chromatogram
ESI	electrospray ionization
FT-ICR	fourier transform ion cyclotron resonance
FWHM	full width at half maximum
HETP	height equivalent to a theoretical plate
HILIC	hydrophilic interaction liquid chromatography
HPLC	high performance liquid chromatography
HRMS	high resolution mass spectrometry
LC	liquid chromatography
LOESS	locally estimated scatter-plot smoothing
LRP	reversed-phase chromatography for lipid analysis
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NEG	negative ionisation mode
NMR	nuclear magnetic resonance
NPC	MRC-NIHR National Phenome Centre
PC	principal components
PCA	principal components analysis
POS	positive ionisation mode

Q-TOF	quadrupole-time-of-flight
QA	quality assurance
QC	quality control
RP	reversed-phase
RSD	relative standard deviation
RT	retention time
TIC	total ion chromatogram
TOF	time-of-flight
TQMS	triple quadrupole mass spectrometer
UPLC	ultra performance liquid chromatography

Glossary

communities	densenly connected sets of nodes in a complex graph, here refers to structurally-related co-eluting features
feature	two-dimensional (*m/z* and retention time) LC-MS signal
feature alignment	the process of finding corresponding features across samples, relies on pseudo chemical spectra similarity analysis
feature grouping	the process of finding structurally-related features in a single sample, relies on the chromatographic shape analysis
peak	one-dimensional signal, either *m/z* centroid in the mass spectrum, or a chromatographic peak
pseudo chemical spectra	list of structurally related co-eluting features

Chapter 1

Introduction

1.1 Introduction

Over the last two decades, biological and biomedical sciences have become increasingly driven by high-throughput technologies, which generate enormous amount of information at different scales of organism organisation - from individual cells, to tissues and whole organ systems [1]. The development of high-throughput technologies has given rise to *omics* fields - principally *genomics*, *transcriptomics*, *proteomics* and *metabolomics*. The emergence of omics platforms has caused a paradigm shift from traditionally descriptive and reductionist approach [2], which was prevalent since the 19th century when experimental biological disciplines first emerged, to a more quantitative and holistic research framework. This new scientific approach to biological questions catalysed the formation of a discipline known as systems biology [3]. The central task of systems biology is to (1) gather comprehensive information from each distinct level of an individual biological system, and to (2) integrate these data to generate predictive mathematical models of the system [3]. Examples of such mathematical models include cell signalling pathways built using gene expression data [4, 5], pharmacokinetics-pharmacodynamics models for drug discovery and development [6], as well as models of of the perturbations to cell's metabolic pathways [7].

The two distinct systems biology methods - the bottom-up and the top-down approaches - have their own potential and limitations. Bottom-up systems biology typically emphasises the construction of mathematical models, which are heavily based on theory and hypotheses and only later validated experimentally [8] (Figure 1.1). Whereas top-down systems biology is heavily driven by experimental data, which is used to discover or to refine pre-existing models that would accurately describe the acquired data. Top-down systems biology studies are based on the use of large omics datasets, which provide a birds eye view of the behaviour of the system.

All omics platforms perform measurements of a large number of biological variables in parallel, however, each one of them focuses on a specific class of molecules. Genomics, for example, studies the *genomes* of organisms, i.e. the genomic information

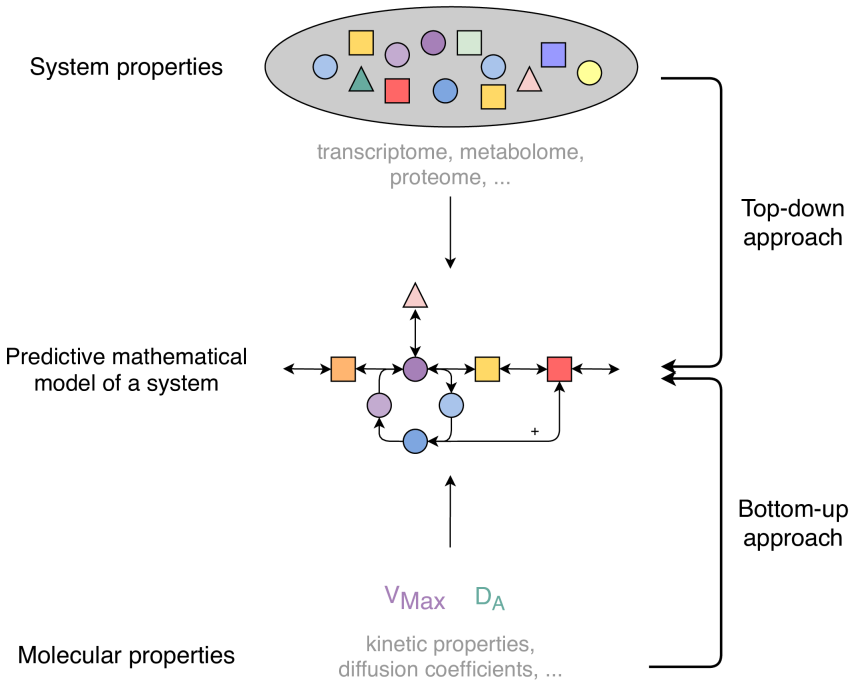


FIGURE 1.1: Schematic representation of the top-down and bottom-up approach to systems biology. Top-down systems biology provides insights into the functioning of living organisms by examining large and potentially complete experimental datasets (e.g. metabolomics datasets). The bottom-up systems biology is based on the modelling of the molecular properties of individual components of a system, which is derived using molecular methods (e.g. enzyme kinetic assays). Scheme is adapted from Bruggeman and Westerhoff, 2007[8].

encoded in the DNA. Transcriptomics, on the other hand, analyses the *transcriptome*, the set of all RNA molecules produced within a cell, tissue or an organism and thus provides insights into gene expression dynamics and regulation [9]. Together with proteomics, which focuses on the *proteome* - the entire set of proteins, their structure, expression and interactions networks [10], and metabolomics, which studies the *metabolome* - the set of all metabolites within, or secreted by an organism, or a cell/tissue, [11], these three omics platforms provide the quantitative measurement of the most dynamic aspects of a living system [12, 13].

The metabolome comprises of metabolites, which are low molecular weight molecules, produced as intermediates and end-products of all metabolic processes. Low molecular weight is generally understood as less than 1,500 Da, however, the molecular weight range for metabolites is very broad, averaging to 665 Da in humans (mean molecular weight for the Human Metabolome Database entries), while molecules of up to 1,200 Da are not uncommon in plants [14]. The estimations of the number of metabolites range from 7,800 [15] to 114,100 in humans (as reported for the Human Metabolome Database in 2018 [16]), and up to 200,000 [17] or even 500,000 [18] in plants. Metabolites are present in varying abundances, have diverse physico-chemical properties and are involved in numerous biochemical processes, including but not limited to:

- *Catabolic reactions* breaks down food nutrients, such as polysacharides, proteins

and fats, into metabolites - carbohydrates, amino acids, nucleic acids, lipids and fatty acids, and releases energy to run cellular processes.

- *Anabolic reactions* builds new and more complex compounds and require energy. For example, amino acids are implicated in the synthesis of new proteins and phenylpropanoids - a diverse group of phenolic plant compounds [19], while various forms of lipids are the building blocks of cell membranes [20].
- *Signalling* in response to stress in plants can be mediated by phenylpropanoid compounds, such as flavonoids, lignin and coumarins [21]. A number of metabolites secreted by human gut microbiota, such as short chain fatty acids and bile acids, have been shown to mediate signalling implicated in host cellular activity [22].
- *Regulation*, for example, through methylation (addition of methyl-groups, which can be delivered from dietary methyl donors, such as metabolites methionine, folate, betaine, and choline [23]) of DNA that is implicated in gene regulation and in turn can regulate metabolism itself [24].

The complex biological network in which metabolites are involved is influenced both by the genotype of the organism and its interactions with the surrounding environment and gut microflora. The dynamics of the metabolome, undermined by metabolic reactions operating at the timescale of seconds, is the second characteristic of the metabolome that makes it the most predictive measure of the phenotype [15, 17]. Consequently, the study of the metabolome through metabolomics, particularly when integrated with the other omics fields, provides exciting prospects in biological and biomedical research, leading to discovery of new drug targets and disease diganostics [12, 25].

Metabolomics uses modern techniques to analyse biological samples that are complex natural mixtures with unknown chemical composition. Depending on the objectives of the study, one of the two distinct analytical strategies can be followed: untargeted or targeted metabolomics (Table 1.1). The first approach, also known as untargeted metabolic profiling, aims to maximise the number of detected metabolites at the cost of quantification accuracy [26]. It is usually applied for hypothesis generation and discovery as it offers the possibility of detecting a wide range of metabolites with high sensitivity. The acquired spectral data is then subjected to statistical analysis to identify potential biomarkers, which then must be identified using authentic chemical standards and targeted approaches [27]. Targeted metabolomics assays focus on a set (tens to hundreds) of pre-defined metabolites. If samples are analysed and compared to authentic chemical standards over a broad dynamic concentration range using selective fragmentation methods, absolute quantification values can be obtained. Nevertheless, targeted approaches provides limited information about a given sample and therefore serves as a validation step in the study [28].

TABLE 1.1: Untargeted versus targeted metabolomics studies. Untargeted metabolomics is a discovery-based approach as it performs relative quantitation of globally detected metabolites. In contrast, targeted metabolomics is used for hypothesis validation since it quantifies sets of pre-defined metabolites of known identity.Table adapted from Schrimpe-Rutledge et al. 2016 [26].

Metabolomics	
Untargeted	Targeted
Discovery	Validation
Hypothesis generating	Hypothesis driven
Global analysis	Subset analysis
Qualitative identification	Known identification
Relative quantification	Absolute quantification

Among the few platforms that perform with sufficient specificity and sensitivity, the two that are employed most frequently are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) [29]. In general, both NMR and MS-based analytical platforms have distinct advantages and each are subject to a number of limitations. ^{1}H NMR spectroscopy is characterised by excellent analytical reproducibility and robustness [30]. Furthermore, NMR-based platforms can provide accurate quantification over a wide dynamic range [31], nevertheless, they are less sensitive than MS-based platforms [2]. When coupled to a liquid chromatography (LC) system, MS is a particularly attractive platform for biomarker discovery and untargeted metabolic profiling due to enhanced sensitivity, separation power and broader metabolite coverage [32, 33].

The application of untargeted LC-MS metabolic profiling to biomedical studies is a very active area of research. Since LC-MS-based methods allow analysis of urine samples without any pre-treatment other than removal of particulates [34], the early developments were undertaken primarily in the field of toxicology, such as screening of drug exposure [35, 36], heavy metal toxicity [37] and nephrotoxicity induced by aristolochic acid [38]. Characterising metabolic alterations in urine samples has also helped to improve clinical diagnosis of liver cancer [39]. Blood samples also require only minimal sample pre-treatment, such as removal of proteins, and have been successfully analysed to characterise the deregulation of fatty acids pathways in patients with chronic liver diseases [40], the nephrotoxicity induced by traditional medicine [41], as well as to discriminate animals with type II diabetes (the Zucker (fa/fa) obese strain) from the normal wild type individuals [42]. While tissue analysis require thorough sample extraction prior to analysis by LC-MS [43], the metabolic characterisation of human prostate tumour [44], pancreas tumour [45], as well as breast cancer [46] tissue extracts has been successfully performed.

The early LC-MS applications, including the examples listed above, were primarily case-control studies. Nevertheless, in order to determine the potential associations between common, low-level exposures, or biomarkers that exhibit high within-individual variability, such as blood metabolites levels, and risk of disease, large

study size is required [47]. Such large-scale metabolic profiling studies with thousands of samples have already successfully identified metabolic biomarkers of biological ageing [48], physical activity [49], diet [50], as well as metabolic predictors of the future development of diabetes [51] and coronary heart disease [52].

Nevertheless, extracting useful information from the large datasets produced through LC-MS analysis of complex samples, such as blood or urine, is highly challenging. Some of the LC-MS data characteristics, which are shared between epidemiological and smaller case-control studies, convolute the standard epidemiological data analysis procedures. These characteristics include high dimensionality and strong collinearity between variables, as well as data non-normality and a substantial degree of missing data [47, 53]. In addition, the number and the identity of most metabolites are unknown due to the untargeted mode of operation, which complicate statistical power calculations [47, 54]. Above all, the acquired LC-MS spectra first must be processed to identify metabolic features [55]. Spectral processing is challenged by numerous technical issues that take place during data acquisition. For example, peaks that correspond to the same analyte drift between sample runs due to physical changes to the chromatographic column and sample build-up in the system, as well as variation in the experimental conditions, e.g. temperature [56]. The acquired LC-MS spectra are misaligned and additional pre-processing steps are required to identify features corresponding to the same analyte across all samples [57]. The larger the sample size, the more profound the effect of such technical issues is on the accuracy of spectral processing. While feature annotation and subsequent metabolite identification can be aided by complex computational methods, inaccurately processed datasets are very hard to annotate, which represents the ultimate challenge in large-scale metabolic profiling studies.

This work will examine the computational tools that are used to process and annotate LC-MS spectra, assess their potential and limitations associated with the analysis of large-scale metabolic profiling studies, as well as suggest alternative strategies.

1.2 The fundamentals of LC-MS metabolic profiling

1.2.1 Analyte separation by liquid chromatography

Liquid chromatography (LC) is one of the most commonly applied chemical separation techniques, which has evolved from its early predecessors - paper and thin layer chromatography. LC operates on the principle that the mixture of interest is dissolved in a liquid mobile phase (eluent), which is passed through a solid stationary phase (chromatographic column) (Figure 1.2). During the process, the adsorbent material of the column interacts with the dissolved analytes differently depending on their physico-chemical properties. As a result, analytes elute from the column with different retention times, leading to the separation of the mixture into constituent components.

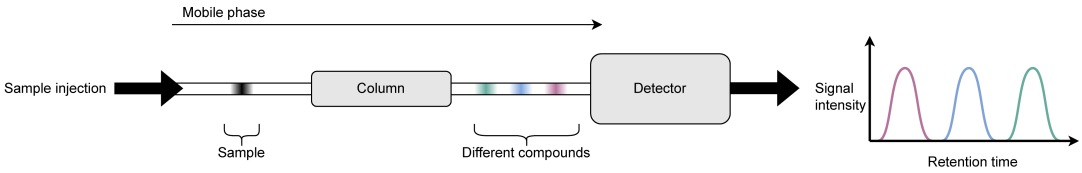


FIGURE 1.2: Liquid chromatography separation of a sample mixture. The sample is injected into the LC system through which mobile phase is flowing. The sample separates into its components that interact with the column differently. The eluted components are detected and a chromatogram is recorded.

A wide range of chromatographic columns can be used in the analysis of complex biological samples. The chemistry of the solid particles of which the column is made off determines by which property the analytes are separated. Therefore, the choice of the column largely depends on the question of interest that a given study is designed to investigate. In the field of metabolomics, two LC types are most often employed: reversed-phase (RP) and hydrophilic interaction liquid chromatography (HILIC). RP chromatography was first successfully developed by Howards and Martin in 1950 [58]. In contrast to earlier LC systems, in RP-LC the eluent is more polar than the stationary phase (i.e. phases are reversed in comparison to the earlier developed normal-phase chromatography). In RP-LC the chromatographic column is usually packed with porous silica particles that have straight octadecyl carbon chains (C18) covalently bound to them (Figure 1.3). The hydrophobic interactions between the alkyl chains and the non-polar moiety of the dissolved analyte determine for how long the analyte retains in the column (i.e. chromatographic retention time (RT)), with hydrophilic molecules eluting from the column first [59]. By increasing the percentage of non-polar solvent, such as acetonitrile, in the mobile phase, which is usually water, elution of less polar molecules is achieved. Such gradient elution allows clear separation of multiple chemical classes in one run. While RP chromatography is capable of measuring a wide range of chemical classes and has been successfully optimised for the analysis of urine [60–62] and blood [52, 63] samples, highly polar analytes are not retained well and elute with the solvent front [64]. Therefore, in order to achieve a comprehensive analytical coverage of the metabolites present in complex biological samples, multiple chromatography types are frequently used [62].

The second chromatography type frequently used in metabolic profiling studies is HILIC. HILIC, as first suggested by Alpert in 1990 [65], is based on the use of a polar stationary phase with a mobile phase similar to those employed in the RP-LC separation (Figure 1.3) [66]. A range of materials can be used for the HILIC columns, such as bare silica or silica gels modified with polar functional groups. The early HILIC columns were silica gels modified with diol or amide functional groups [66]. Modern HILIC columns provide increased stability over a broad range of pH and temperatures due to hybrid organic/inorganic packing materials, for example, bridged ethylsiloxane/silica hybrid (BEH) particles [67]. It is believed that the separation of

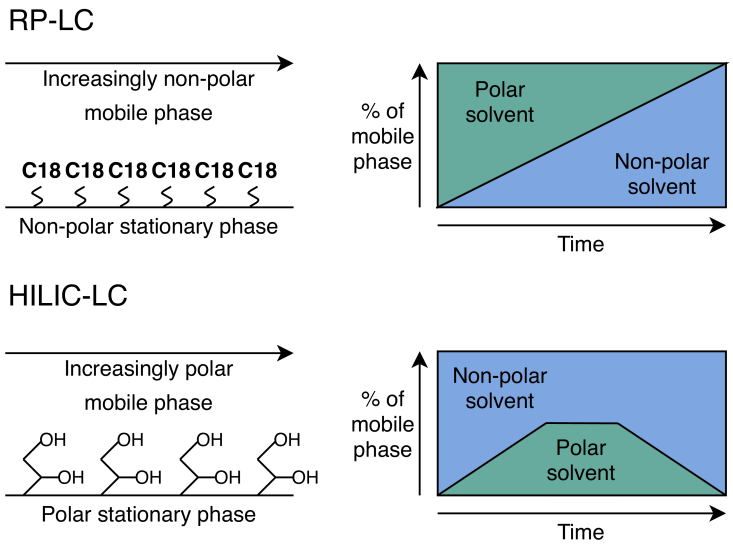


FIGURE 1.3: RP-LC and HILIC-LC are the two most frequently used chromatography types in metabolic profiling studies. In RP-LC, the mobile phase is more polar then the stationary phase, which is usually made of porous silica particles that have straight octadecyl carbon chains (C18) covalently bound to them. In HILIC-LC, polar chromatographic surfaces are used, such as silica gel bounded with dihydroxypropyl (diol) functional groups. Gradient elution is achieved by changing the mobile phase composition by mixing the polar and non-polar solvents in different proportions during the chromatographic separation.

analytes is achieved by the balance of multiple factors:

- Partitioning the analyte between two mobile phase layers of different polarity. When a mobile phase is composed of primarily acetonitrile and a minimum of 23% of water, an acetonitrile-rich and a water-enriched layer adsorbed onto the hydrophilic stationary phase develop [65]. Consequently, polar hydrophilic molecules are preferentially solubilized into the water layer, and thus, strongly retained (Figure 1.4).
- Weak electrostatic interactions between charged analytes and the ionized groups of the stationary phase [68].
- Hydrogen bonds, or direct interactions between the analyte and the stationary phase, have also been suggested to play a role in analyte retention in addition to the partitioning into the water layer [69].

The separation of analytes is achieved by decreasing the difference in polarity between the bulk and the adsorbed layer, i.e. by increasing the water content in the mobile phase (Figures 1.3 and 1.4) [69]. Due to the ability to separate polar analytes, HILIC has been applied to the analysis of polar metabolites, such as amino acids, organic acids and sugars, which elute closely together under RP conditions. Nevertheless, HILIC tends to produce broad and asymmetrical peaks, which complicate automatic peak detection and analysis, and suffers from irreproducible retention time and analytical drift when multiple samples are analysed [70].

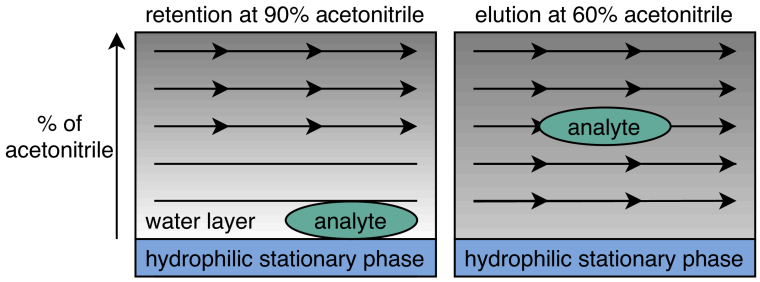


FIGURE 1.4: The HILIC partitioning of a polar analyte into the water layer of the mobile phase adsorbed on the surface of the hydrophilic phase. Scheme adapted from Greco and Letzel, 2013 [69].

The efficiency of a mixtures separation is improved when high performance liquid chromatography (HPLC) systems are employed in place of traditional chromatography. HPLC relies on pumps to pass mobile phase and the sample of interest through the column at high pressure (50 - 350 bar). While HPLC provides good separation, by decreasing the size of the column particles to $1.7\mu m$, a significantly increased peak separation can be achieved. In such chromatography systems, known as ultra performance liquid chromatography (UPLC), the pressure is raised even further to 800 bar to account for the increased particle resistance and achieve much shorter run times [71]. The efficiency of chromatographic separation is characterised by the plate count (N), which is a concept adapted for liquid chromatography from early work with distillation columns [72]. Plate count is a ratio of column length (L) to the theoretical plate height (H), as in:

$$N = \frac{L}{H} \tag{1}$$

where H is also defined as height equivalent to a theoretical plate (HETP). HETP relates to various flow and kinetic parameters leading to peak broadening, the relationship of which is captured in the van Deemter equation [73], as follows:

$$HETP = A + \frac{B}{u} + (C_s + C_m) \times u \tag{2}$$

where, A is eddy diffusion parameter (the dispersion of individual molecules in the column that relates to the particle bed of the stationary phase and therefore is characteristic of a given column); B is diffusion coefficient (the spread of eluting particles in the longitudinal direction, dependent on the column diameter and diffusion of the mobile phase); C is resistance to mass transfer coefficient of the analyte between mobile (m) and stationary phase (s), and u is the linear velocity of the mobile phase. Low $HETP$ values lead to higher separation resolution, as in Equation 1, as well as increased analysis speed. In UPLC, as illustrated in Figure 1.5, low HETP values are achieved by using columns packed with smaller, uniformly-sized particles and

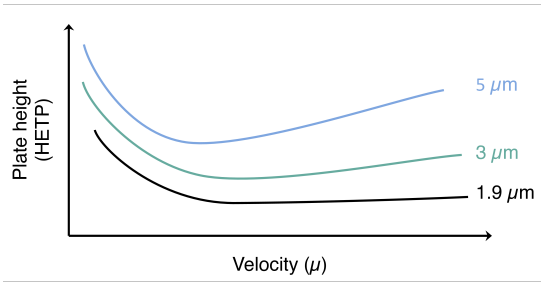


FIGURE 1.5: Van Deemter equation describes that chromatographic separation efficiency varies with the mobile phase velocity and particle size. The curves of theoretical plate height (HETP) values achieved with increasing mobile phase velocities (u) for thee particles of different sizes are demonstrated.

by increasing the flow rates of the mobile phase. However, as the particle size decreases, and/or the mobile phase velocity increases, higher force is required to move the mobile phase through the system, which represents one of practical limits on the performance of UPLC systems. Nevertheless, in the field of metabolomics, HPLC have been largely superseded by UPLC analyses, which will also be utilised in this work.

1.2.2 Analyte detection by mass spectrometry

While LC is capable of separating a sample mixture into its components - different compounds - according to their physicochemical properties, in order to identify and/or quantify the analytes eluting from the chromatography column, a detector system is required. Such a role can be performed by mass spectrometry (MS), which is an analytical technique based on the use of a mass spectrometer. MS analyses gas-phase ions by separating them according to their mass-to-charge ratio (m/z), which is the mass of an ion on the atomic scale divided by the number of charges the ion carries [74]. The intensities of the separated ions are recorded as a mass spectrum, which represents the distribution of m/z values in a given sample.

A typical mass spectrometer consists of three essential components - an ion source, a mass analyser and a detector. The first and the most critical step in MS-based analysis is ionisation since the ability to detect and quantify an analyte is determined by the degree of its ionisation. Early hyphenated LC-MS systems relied on the use of atmospheric pressure chemical ionisation (APCI) as the ion source. APCI is one of the few ionisation techniques that are suitable for coupling LC to MS which requires not only analyte ionisation, but also solvent desolvation, i.e. solvent vaporisation. Even though APCI provides a high dynamic range, is easy to operate, stable and tolerant to high buffer concentrations, in metabolite profiling it has been largely superseded by electrospray ionization (ESI) [75]. While the exact mechanism by which ions in solution are converted to ions in the gas phase is still debated, ESI generally works by forcing the solution of an analyte through a small capillary, to which a voltage

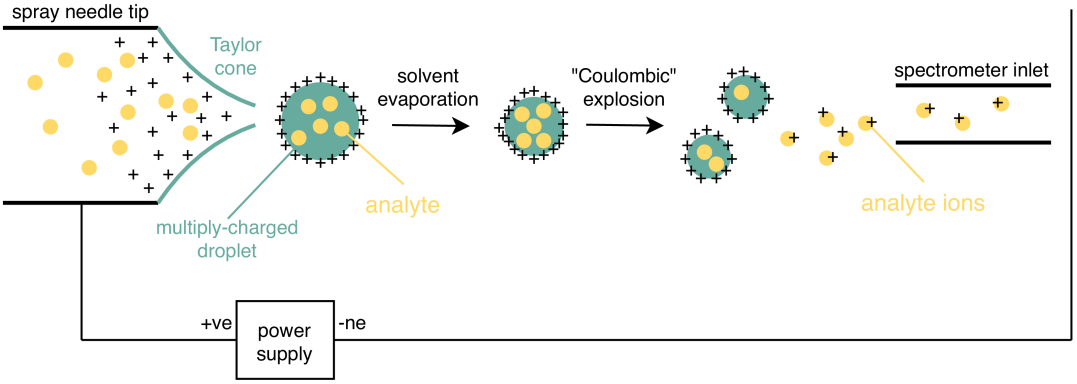


FIGURE 1.6: Scheme of the mechanisms of ion formation in electrospray ionisation. Diagram adapted from Gaskell 1997 [78].

is applied [76]. Dispersion of the solution through charged capillary results in production of charged droplets at the capillary tip. These solvent droplets move down an electric field imposed between the capillary tip and a metal plate, which prevents them from freezing and promotes solvent evaporation (Figure 1.6). As solvent evaporates, the droplet radius decreases, resulting in charge density build-up at the surface. Such droplets become unstable and eventually undergo so called coulomb fission, emitting analytes as ions. ESI ionisation efficiency is highly dependent on solvent (e.g. through mobile phase additives, such as ammonium salts), polarity mode and applied voltage, all of which contribute to droplet formation [77]. Similarly to APCI, ESI principally produce ions of the intact analyte molecule, usually in the form of the protonated molecule, for example, $[M+H]^+$, where M is the molecular mass of the analyte. However, due to adduct formation, different species can be formed through clustering with the solvent molecules, for example, potassiated and sodiated analyte ions $[M+K]^+$ and $[M+Na]^+$ [74]. In ESI, little fragmentation of the analyte can also occur.

Ions produced at the ion source are next separated according to their *m/z* by the the mass analyser. Multiple methods for ion separation have been developed, each of which has different accuracy and resolution performance. The term *resolution* relates to the separation of ions of two different *m/z* values and pertains to mass spectrometry data. Whereas *resolving power* is a function of a mass spectrometer and is defined as the difference in *m/z* values (ΔM) of ions that can be separated from one another by a given mass spectrometer, divided into a specific *m/z* value (M), as in the IUPAC definition:

$$R = \frac{M}{\Delta M} \tag{3}$$

Given the definition, R will be different at every *m/z* value. Therefore, both exact M and also the method for defining the peak width necessary for separation at this

TABLE 1.2: Typical analytical parameters of most commonly used mass analysers. The broad ranges of parameters indicate that the specifications are often instrument and brand-specific. For example, Agilent 6230B TOF-MS is capable of providing 22k FWHM resolution, whereas Waters LCT-Premier TOF-MS instrument can achieve >10k FWHM resolution.

Mass spectrometer	Resolving power (50% FWHM)	Mass accuracy (ppm)	Scan speed (scan/sec)	Mass range (*m/z*)
FT-ICR	100k - 10,000k	0.1 - 1	1	100 - 2k
Orbitrap	15k - 500k	0.5 - 1	<12	50 - 6k
Q-TOF	10k - 50k	1 - 5	20 - 50	50 - 30k
TOF	1k - 22k	1 - 5	50 - 500	25 - 20k

mass, ΔM, such as full width at half maximum (FWHM), must be given. High resolution mass spectrometry (HRMS) generally refers to mass analysers, which have a mass resolving power $M/\Delta M_{50\%}$ of > 10k [79]. Typical values for most commonly encountered mass analysers as summarised in Table 1.2. Since high mass resolution is required to deconvolute complex mixtures, such as human biofluids and whole-cell lysates, it is the key property when choosing a mass spectrometer for metabolic profiling studies[80].

The other essential criteria for a mass analyser is high *m/z* values measurement *accuracy*. The higher the accuracy of the measured mass, the fewer potential molecular formula can be assigned to it, which aids compound identification process. The accuracy of a mass measurement indicates the deviation of the instrument response from the calculated monoisotopic mass [81]. Accuracy is is often reported via the statistical error in parts per million (ppm), as in:

$$ppm = \frac{(measured - theoretical)}{(theoretical)} \times 10^6 \tag{4}$$

Finally, the other important criteria to consider when selecting a mass analyser for metabolic profiling studies are the ability for rapid data acquisition (*scan rate*) and detection of a wide range of *m/z* values (*mass range*) (Table 1.2).

Among the numerous mass analysers that have been developed up to date, the most common in the field of metabolomics are Orbitrap and time-of-flight (TOF) mass analysers. Orbitrap, invented by Makarov in 2000 [82], operates on a similar principle as other ion trap mass analysers, such as fourier transform ion cyclotron resonance (FT-ICR), invented by Comisarow and Marshall in 1974 [83]. In both FT-ICR instruments and Orbitrap, ions are trapped in a vacuum initiating harmonic axial oscillations. As the frequency of the oscillations are proportional to the mass of the ions, detected and Fourier-transformed oscillations are eventually converted into a mass spectrum [84]. Whereas in FT-ICR instruments ions are trapped in a strong magnetic field combined with a weak electric field, in Orbitrap ion motion is determined only by the electrostatic field [84]. The consequence of this difference is

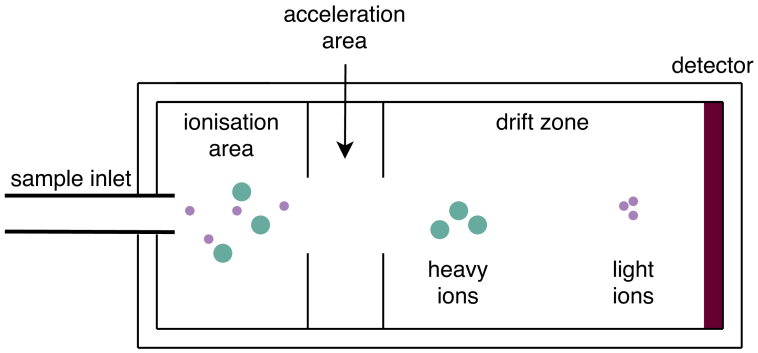


FIGURE 1.7: Schematic representation of a time-of-flight (TOF) mass spectrometer. All ions receive the same kinetic energy during acceleration by an electric field of known strength. However, ions have different *m/z* values and thus different velocities. Smaller masses (i.e. *m/z*) will have larger velocities and will reach the detector sooner. For simplicity of illustration, the scheme depicts a linear TOF mass spectrometer that does not have an ion mirror. Adapted from Watson and Sparkman, 2007 [74].

that Orbitrap has a significantly slower decrease of resolving power with increasing *m/z*. Nevertheless, both mass analysers are characterised by very high resolving power (Table 1.2). The main disadvantage of ion trap mass analysers with regards to metabolic profiling is low data acquisition rate since during the trapping phase data recording is off.

TOF mass analysers have been playing an increasing role in metabolic profiling since the first demonstration of the possibility of coupling ESI ion source to a TOF instrument in 1994 [85]. The basic operating principle of TOF mass analysers is to measure the time required for an ion to travel from the ion source to the detector (Figure 1.7). All ions at the electron gun are accelerated by an electric field of known strength. As a result, all ions have the same kinetic energy, therefore their velocities in the vacuum depends only on their *m/z* values, with heaver ions taking more time to reach the detector. A single mass spectrum is obtained by the acceleration and detection of a set of ions from the ion source to the detector. A complete spectrum can be obtained every few microseconds, with multiple transient spectra usually averaged into a single spectrum [86]. Such pulsed mode of operation is sufficient to detect metabolites with high sensitivity and virtually unlimited mass range at rapid data acquisition rate [87]. To improve the resolving power of a TOF mass spectrometer, an ion mirror, also known as the reflectron, can be employed (Figure 1.8). Such an ion mirror, operating in the form of an electric field, effectively doubles the flight distance in the same space and reflects ions of the same *m/z* that have different velocities [74]. Reflectron, can increase the resolving power from 1k FWHM to as much as 22k FWHM. Therefore, modern TOF mass analysers are ideally suited to profiling LC eluent.

A hybrid quadrupole-time-of-flight (Q-TOF) system has been increasing used in place of TOF-based analyses. The first element in such a mass analyser, the quadrupole, is traditionally used on its own in tandem mass spectrometry (MS/MS)

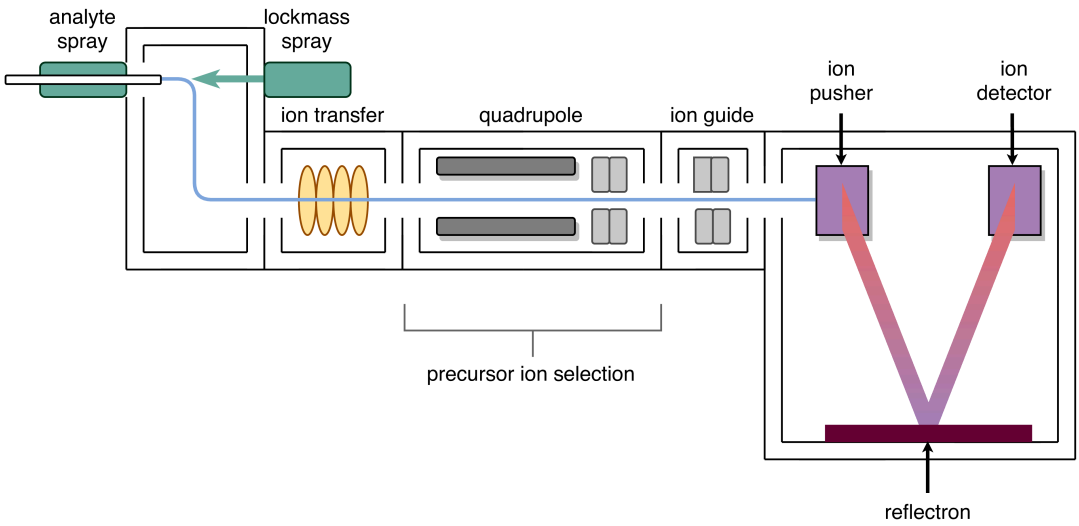


FIGURE 1.8: A simplified scheme of LC-Q-TOF mass spectrometer. Diagram adapted from Watson and Sparkman, 2007 [74].

to acquire a mass spectrum of a mass spectrum as part of a technique known as tandem mass spectrometry (MS/MS) [74]. The most common MS/MS spectrometer is triple quadrupole mass spectrometer (TQMS), which comprises of two quadrupole mass analysers and a collision cell (a quadrupole operating in radio frequency only) in between them. The first quadrupole in the tandem arrangement is used for selection of precursor ions derived from the ion source. The precursor ion is directed to the collision cell, in which it collides with the neutral atoms of the collision gas (usually nitrogen) to convert its kinetic energy to internal energy. This process drives precursor ion decomposition/fragmentation into the product ions, which are subsequently analysed by the third quadrupole. The resulting MS/MS spectrum contains new structural information that aids subsequent metabolite identification process.

In Q-TOF hybrid mass spectrometer, product ions generated in a collision cell are captured by a TOF mass analyser. Waters brand (Waters Corp., Milford MA, USA) Q-TOF-MS instrument Xevo G2-S also contains a module with ion transfer optics for increased ion transfer efficiency from ion source to the quadrupole analysers (Figure 1.8). Product ions leaving the quadrupole then enter an additional ion guide module that utilizes non-uniform, moving electric fields and/or voltage devices to separate ions on the basis of size, shape and charge. Ions leaving the ion guide are then transmitted to the pusher and are accelerated orthogonally by the pusher voltage. In the TOF flight tube, ions separate according to m/z and are focused by the reflectron grid voltage, while their arrival time at the detector is accurately measured. Such instrumental setup results in high mass accuracy and improved resolving power, significantly enhancing sample definition. Since precursor ion selection and fragmentation can be disabled, Q-TOF instruments can also be used to acquire MS rather than MS/MS spectrum. Therefore, Q-TOF mass spectrometer has been regarded as

TABLE 1.3: LC-MS-based analytical platforms suffer from unwanted analytical variantion, which falls into three broad categories. The examples of the most common sources of analytical variation are provided.

	Random	Run order dependent	Compound-specific
Measurements	*m/z*, RT	*m/z*, RT, intensity	RT, intensity
Causes		column contamination & ageing, experimental conditions (e.g. pressure)	ion suppression, ionisation efficiency

one of the superior tools for global metabolic profiling of complex biological samples, particularly when coupled to electrospray ionisation sources [80]. This work will focus on data acquired with an ESI-Q-TOF mass spectrometer.

1.2.3 Analytical variation in LC-MS experiments

The complexity of the LC-MS-based analytical techniques that are applied to metabolomics studies often lead to uncontrolled variance, which is not related to the desired biological variation resulting from the conducted experimental design [88]. Such variation is known as technical, or analytical, variation, which may distort or obscure subsequent data analysis. The most common sources of analytical variation are summarised in Table 1.3 [89–95]. While some of the observed variation is random, the most challenging is the run-order-dependent and analyte-specific variation, which cannot be modelled using monotonic functions. The phenomenon of the run order effect relates to the combined changes in the LC-MS system, such as the chromatographic column, ion source and MS detector, the performance of all of which deviates during continuous sample analysis. Furthermore, a vast proportion of the unwanted variation in LC-MS data is specific to a given analyte, or a class of related analytes. Such variation is particularly challenging in large-scale studies, where the combined run order-dependent and analyte-specific effects lead to convoluted spectra processing, producing inaccurate datasets. Such analytical variation patterns are investigated in detail in Chapter 2.

1.2.4 LC-MS spectra processing

The use of LC-MS systems to simultaneously separate and detect analytes generates complex data, which comprises of a large number of consecutively acquired mass spectra, or scans. The index of a scan is representative of retention time. The collection of scans for a single LC-MS experiment can be viewed a three-dimensional landscape of peaks, described by their *m/z*, RT and intensity (Figure 1.9). Such complex LC-MS data can be visualised in multiple ways. The chromatographic domain of the data can be evaluated by summing the intensities of all mass spectral peaks in a given MS scan and creating a total ion chromatogram (TIC). As TIC includes both analyte-derived signals and noise, peaks corresponding to analytes can be concealed in such data representation. Therefore, base peak intensity chromatograms (BPI),

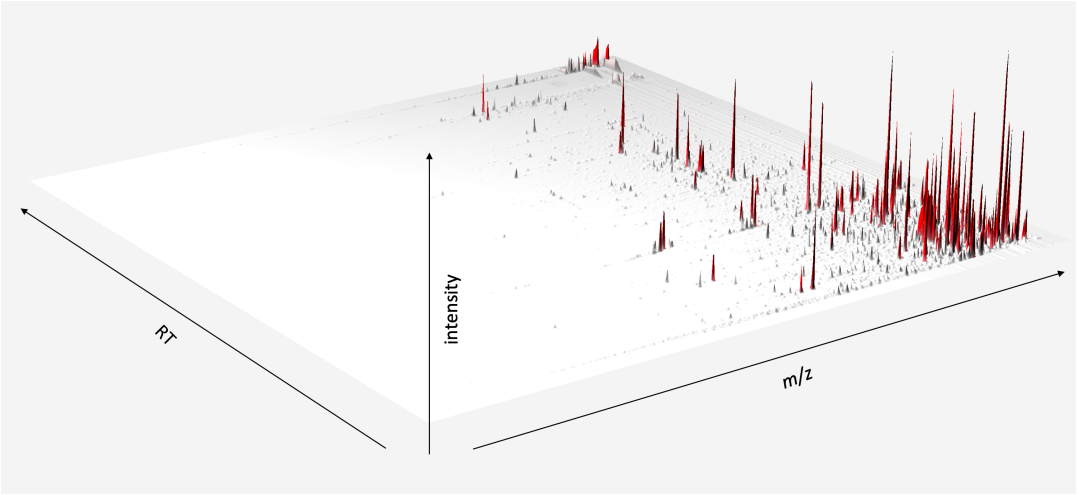


FIGURE 1.9: LC-MS analysis generates complex data that comprises of a large number of consecutively acquired mass spectra, or scans. The collection of scans obtained from a single LC-MS experiment can be interpreted as a three-dimensional landscape of peaks, described by *m/z*, RT and intensity.

representing the intensity of the most intense peak at every scan, are sometimes used instead. To visualise the chromatogram of one or more analytes of interest, the intensity of the signal arising from the corresponding *m/z* values are extracted from every scan and represented as an extracted ion chromatogram (EIC).

To quantify the analytes in a given metabolomics sample based on the information in the acquired LC-MS data, the signals corresponding to different ion species must be extracted. Such data pre-processing leads to significant data reduction and yields a list of *features*, each corresponding to a single species, characterized by *m/z*, RT and abundance (i.e. intensity). Features intensity measurements, correlating to relative concentration in the samples, can be subjected to statistical analyses in order to derive biological insights.

A wide range of pre-processing tools have been developed for LC-MS data. Among the most popular ones are open-source tools: XCMS [96], MZMine and MZmine2 [97, 98], OpenMS [99], MetAlign [100], as well as workflow-based systems, such as XCMS Online [101], MAVEN [102], Galaxy [103] and Workflow4Metabolomics [104]. While the sequence of pre-processing steps, as well as the exact algorithms applied at each stage, vary from one software to another, the generally accepted pre-processing pipeline comprises of the following steps:

- *Peak detection* identifies signals in the three-dimensional LC-MS data space (RT, *m/z*, intensity). Implemented algorithms are often based on continuous wavelet transformation to fit the shape of the chromatographic peak and integrate its area [105–107]. However, all of the highly cited algorithms are known to report false peaks [108]. In the case of the current gold standard, the *centWave* algorithm, false positive peaks are reported because of local noise underestimation [109].

- *Feature alignment/correspondence* finds corresponding peaks across all samples and joins them into features. Alignment is particularly sensitive to RT deviations [57, 95] and represents an unsolved issue in the field of metabolic profiling.
- *Retention time correction* aims to correct for RT drifts between samples, which improves feature alignment accuracy. Due to limited knowledge of how different chemical classes behave under different LC conditions, this pre-processing step is least standardised.
- *Missed peak filling* integrates raw data of a sample for a peak that was not detected by the first step in the pipeline. Peak filling sometimes is omitted since it relies on correct signal-to-noise estimation [96] and therefore can potentially integrate noise rather than a true signal. Furthermore, if false positive peaks are being subjected to peak filling, notoriously noisy datasets are produced.

The advantages and limitations of different pre-processing tools are discussed in detail in Chapter 3.

1.2.5 LC-MS spectra annotation and metabolite identification

Annotation and identification are two terms that are being used interchangeably, however, they represent very different concepts. Spectral feature annotation is its assignment to a potential chemical entity given its measured properties, while metabolite identification requires a comparison of experimental data to data acquired for authentic chemical standards [110].

Spectral features annotation and identification represents one of the most time-consuming steps in metabolomics studies. Nevertheless, without accurate metabolite identification, biological data interpretation is very convoluted. Several types of data are applied for metabolite annotation and identification: (1) accurate mass (AM); (2) chromatographic retention time (RT), (3) information about the sample and (4) fragmentation pattern (MS/MS) [111].

Routinely applied identification workflow for untargeted LC-MS data usually starts with matching the accurately measured *m/z* to molecular formulas in databases [112]. Even if prior knowledge, such as the chemical classes that are likely to elute under given LC conditions (i.e. at the particular RT) or be present in certain biological samples, can be applied to constrain the large search space composed of all potential molecular formulas, *m/z* matching alone can only yield putative annotations. The likely candidates are usually further investigated by recording MS/MS spectrum. The detected MS/MS fragments originating from the same molecule are used to elucidate the chemical formula and structure of the unknown metabolite [74].

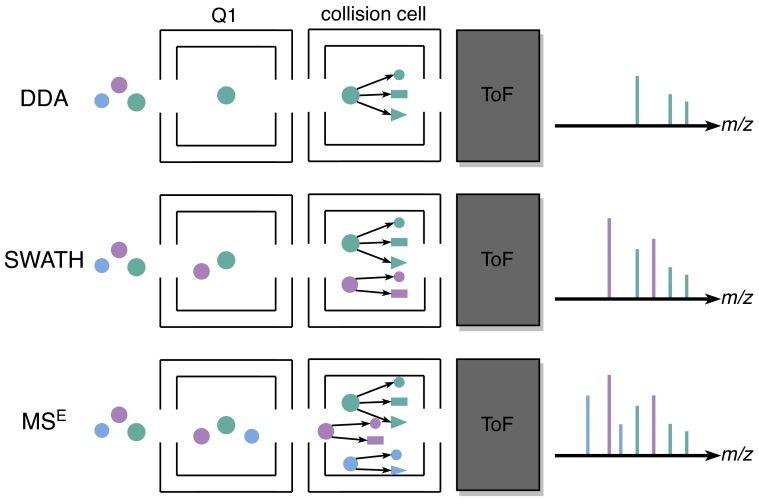


FIGURE 1.10: A hybrid Q-TOF mass spectrometer allows to acquire a MS/MS spectrum using multiple methods: (1) data-dependent acquisition (DDA); (2) sequential window acquisition of all theoretical fragment-ion spectra (SWATH) and (3) MS^E techniques.Figure adapted from Zhu et al. 2014 [113].

Hybrid Q-TOF-MS systems allow to acquire MS/MS fragmentation data in three ways. While high-resolution MS spectra is collected by disabling the precursor ion selection and fragmentation (Figure 1.8), MS/MS data can be acquired by enabling the precursor ion selection at quadrupole Q1 with either narrow, medium or wide pass mode, followed by fragmentation under high collision energy (Figure 1.10) [113]. During data-dependent acquisition (DDA), only ions meeting pre-defined criteria are selected and fragmented. DDA is normally operated to select ions within a narrow *m/z* window, typically 1 - 3 Da wide, or the most intense ions in a spectrum. DDA is usually performed as part of a standard untargeted metabolomics workflow during which acquisition of a full MS spectrum is followed by DDA. The method simultaneously obtains both quantitative and structural information for a given sample and therefore can aid both sample characterisation and metabolite identification [114]. Nevertheless, when precursor ion selection is based solely on the intensity in a given spectrum, the presence of any abundant contaminant is likely shift the results away from the unknown metabolite of interest. By contrast, data-independent (DIA) methods can acquire MS/MS fragmentation data for all precursor ions at the same time, significantly increasing the coverage of observed metabolites [115]. DIA methods include a technique known as SWATH (sequential window acquisition of all theoretical fragment-ion spectra), which operates under a medium pass mode. With SWATH, precursor ion selection occurs in cycles, during which ions within a medium *m/z* window, such as 20 Da wide, are selected and fragmented [113]. These selection windows cover the whole *m/z* range of interest and thus provide structural information on many more metabolites than DDA alone. The least selective MS/MS data acquisition mode is called MS^E, which records all product ions for all precursor ions and therefore produces highly complex MS/MS spectra that often

requires additional data deconvolution step in order to extract the original spectra for a given metabolite [115]. MS/MS spectra acquired for study samples aids compound identification process by providing information on metabolite chemical structure and functional groups, which is facilitated by a growing number of spectral databases and tools. Nevertheless, all likely candidates must be validated with authentic chemical standards [111].

Multiple computational strategies for can be taken to accelerate identification process, which include the analysis of naturally occurring isotopes, MS spectra adducts and fragments, as well as various correlation techniques. These are evaluated and discussed in detail in Chapter 4.

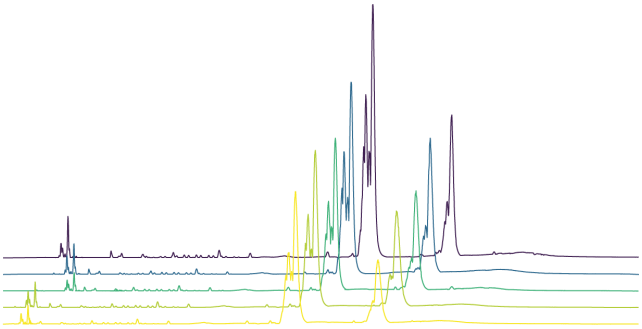
1.2.6 Thesis aims and structure

The overall aim of the work comprising this thesis is to deliver an improved pipeline for LC-MS spectral pre-processing and annotation capable of supporting the metabolic profiling of large-scale epidemiological studies.

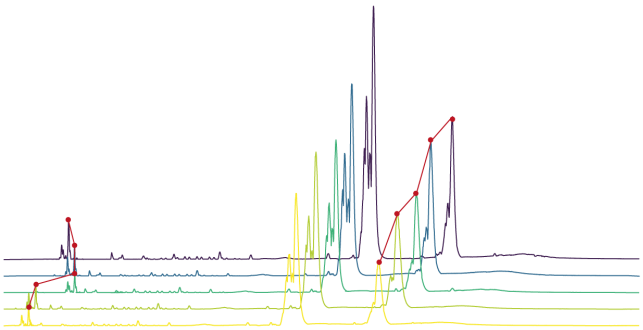
The specific objectives of this thesis are (Figure 1.11):

- To characterise the analytical variation observed in a large-scale metabolic profiling study (Chapter 2).
- To develop and implement a pre-processing pipeline that accommodates the observed analytical variance in large sample set profiling (Chapter 3).
- To demonstrate the utility of the suggested pre-processing approach for automated metabolite identification (Chapter 4).

Characterisation of analytical variation
Chapter 2



Feature alignment
Chapter 3



Automatic features annotation
Chapter 4

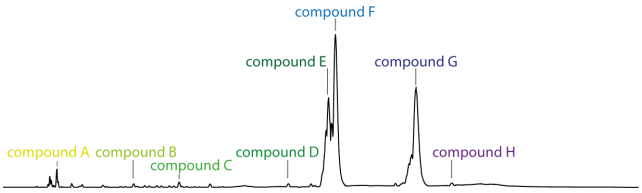


FIGURE 1.11: The development of LC-MS spectra processing and annotation pipeline can be divided into a number of conceptual steps. The results of each step are presented as a separate results chapter in this thesis.

Chapter 2

Characterisation of analytical variation in a large-scale metabolic profiling dataset

2.1 Introduction

In metabolic profiling studies, a snapshot of the metabolome is obtained, which can be used to derive observations on the effects of the experimental conditions on the biological system studied. However, extracting relevant biological information from metabolic profiling data represents a major challenge.

Data acquired with untargeted LC-MS assays is firstly processed to detect chromatographic peaks in each sample, which then must be aligned/grouped between all samples into so called "features", each representing a unique ion. The relative concentrations of these ions can be analysed mathematically to model alterations in metabolism in response to the experimental conditions studied, ultimately identifying the metabolites inflicted in the changes. However, the process of detecting and grouping chromatographic peaks, as well as identifying the corresponding metabolites is highly complex. Besides the desired biological variation resulting from the conducted experimental design, metabolomics data also contain other types of variation. While removal of unknown biological variation, such as varying biofluid sample concentration [116] or differing cell sizes [117], is commonly attempted through various scaling/normalisation methods [118], a huge proportion of unwanted variation originates from errors associated with sample handing (e.g. during sample preparation) and data acquisition [119]. Multiple sources of analytical variation occurring during data acquisition are known: (1) changes in the LC column performance over time, for example, due to stationary phase contamination by the injected samples [120]; (2) changes in the MS system, such as decreasing MS detection sensitivity or ion suppression from co-eluting compounds [88]; (3) alterations in experimental conditions, such as temperature, pressure, mobile phase composition [121].

Analytical variation is often systematic and therefore different statistical methods are required to remove it in order to preserve the meaningful biological information.

To reduce the levels of unwanted variation and ensure high data quality, a number of quality assurance (QA) and quality control (QC) processes must be followed. QA processes focus on providing confidence that quality requirements will be fulfilled during data acquisition and therefore mostly relate to procedures applied in preparation for the study, such as staff training, laboratory and instruments maintenance and calibration, methods validation and documentation [122]. QC processes, on the other hand, concentrate on fulfilling quality requirements and thus relate to steps taken during and after data acquisition, including, but not limited to, analysis of measurement standards, spiked samples, blank samples and QC samples [123].

Clear quality guidelines have been established for targeted assays, which are utilised in drug development projects and therefore are tightly regulated by various national and international regulatory bodies, such as the Food and Drug Administration (FDA) in the United States [124]. However, the objective of untargeted assays is to measure as many metabolites as possible, including unknowns. The lack of community agreed-upon QA/QC guidelines for untargeted assays is reflected in the responses to the Metabolomics Society questionnaire undertaken in 2015 [125]. Following the recommendations outlined by the questionnaire respondents, a two-day Think Tank on QA and QC for untargeted metabolomic studies was held and the metabolomics QA and QC consortium was established in 2017 [126].

While community-level guidelines for untargeted metabolomics QA are not available at the moment, numerous efforts have been made recently to promote good practices that are already being applied in the field (for example, [62, 64, 92]). The common strategies include the use of a pooled QC sample, which is often created by taking a small aliquot of each biological sample in the study [124]. Such a pooled QC sample should represent the aggregate metabolite composition of all of the samples in a study. Its primary role is therefore to enable the estimation of the variability in the measurements of each metabolite in the study. Multiple applications of a pooled QC sample have been suggested. Firstly, pooled QC samples are frequently used to "condition" the analytical system prior to analysing the study samples [63, 127]. Usually, the QC sample is injected 5 to 10 times at the beginning of the run to stabilise the retention time and detector response. Furthermore, QC sample is analysed periodically throughout the analytical batch, usually at least every 10th sample. Repeated measures enable to obtain quality metrics, such as measurement precision, for metabolites that are present in the QC samples, as well as to detect and correct for systematic intensity variation associated with injection order. Multiple mathematical models have been proposed for the correction of run order effect using QC samples, such as linear regression [93], locally estimated scatterplot smoothing (LOESS) [92] and cubic spline function [119]. Even though the use of pooled QC sample is being increasingly advised, alternative QC strategies have been suggested as well. For

example, within-batch drift correction was performed through Bayesian clustering of features in [128], whereas feature quality assessment based on the missing rate pattern along the injection order was performed in [129].

Other standard QC procedures for untargeted assays include the use of internal standards, which are added to every sample before and after metabolite extraction so as to be present at the same concentration. Internal standards are used to model unwanted variation and subsequently normalise acquired data by, for example, subtracting the log abundance of a single standard from the log metabolite abundances in each sample [117]. Such approach is based on the assumption that every metabolite experiences the same amount of unwanted variation, however, compound-specific variation has been observed [130]. Therefore, the use of multiple internal standards is strongly advised to cover a range of retention time and m/z values, as well as different physico-chemical properties [118, 131]. The choice of the internal standards ultimately depends on the chromatographic assay of choice, the metabolome of the samples studied, availability and cost [124]. Additionally, endogenous metabolites can also be used as internal standards [117].

Besides the pooled QC samples and dilution series, each analytical batch in an untargeted LC-MS experiment also contains blank samples. Two types of blank samples are typically utilised in metabolomics studies. At the start of each analysis, LC-MS system suitability is inspected by injecting saline blank samples [92]. These samples, usually comprised of authentic chemical standards solution, can reveal any instrument performance problems arising due to system contamination [124]. Similarly, such blank samples can be injected after a series of study samples, or at the end of an analytical batch, to inspect whether any carryover metabolites or standards take place between samples. The second type of blank samples commonly used in metabolomics experiments are known as process, or extraction blanks, as they are prepared using the same sample preparation and analysis procedures as the study samples, however, are comprised of only the solvent mixture and/or water, as well as any internal standards [124]. Extraction blank samples represent the signals that arise from chemicals and contaminants in the mobile phase solvents and sample processing materials. These samples are therefore utilised to adjust data for unwanted technical variation since metabolite features present in the extraction blanks may be removed [43, 117, 119]. The use of blank samples therefore greatly improves the quality of the data, which in turn helps to distinguish real biological variation in the data.

A single unified QA/QC procedure is unlikely to suit all laboratories and analytical assays. Following good practices most suitable in a particular situation decreases the levels of unwanted variation. Nevertheless, analytical variation cannot be completely avoided. Therefore post-acquisition data quality control and normalisation procedures must be followed as well.

2.1.1 Aims and objectives

The purpose of this chapter was to explore open source pre-processing and quality control tools for untargeted metabolic profiling studies. More specifically, this aims of this chapter were three-fold:

- To pre-process a large-scale untargeted LC-MS study using open-source tools.
- To evaluate the observed analytical variation using post-acquisition QC procedures.
- To assess the suitability of the applied tools for the analysis of the study of interest.

2.2 Methods

2.2.1 Analytical data acquisition

A large-scale cohort study, the Airwave Health Monitoring Study (AIRWAVE), was used as the source of extensive metabolic profiling data [132]. The AIRWAVE study is an observational cohort study of the British police forces undertaken to evaluate the link between the use of terrestrial trunked radio (TETRA) and a wide range of health outcomes. AIRWAVE now has more than 36,000 participants with extensive data on lifestyle and clinical measurements, and a wide range of biological samples. It has been adopted as a Tissue Biobank within the Imperial College Human Tissue Biobank.

The cohort provides two random samples of specimens known as AIRWAVE1 and AIRWAVE2 (Table 2.1). AIRWAVE1 comprises of a random sample of 3,000 urine and 3,000 serum specimens, which were analysed by UPLC-MS and ^{1}H NMR using standardised protocols [133, 134] at the MRC-NIHR National Phenome Centre (NPC) prior to the start of this project. AIRWAVE2 subset comprises of additional 2,250 plasma samples, which were analysed by commercial data provider Metabolon using proprietary methods [135]; 1,000 of those were also analysed by the NPC.

TABLE 2.1: Datasets available within the AIRWAVE cohort.

AIRWAVE1		AIRWAVE2	
n=3,000*		n=1,000*	n=2,250†
Serum	Urine	Plasma	Plasma
HILIC-POS-MS	HILIC-POS-MS	HILIC-POS-MS	RP1-POS-MS
LRP-NEG-MS	RP-NEG-MS	LRP-NEG-MS	RP2-POS-MS
LRP-POS-MS	RP-POS-MS	LRP-POS-MS	RP-NEG-MS
^{1}H NMR	^{1}H NMR	^{1}H NMR	HILIC-NEG-MS

* samples were analysed at the National Phenome Centre.

† samples were analysed at Metabolon.

Three AIRWAVE1 serum LC-MS datasets - reversed-phase chromatography for lipid analysis (LRP) (positive ionisation mode (POS) and negative ionisation mode (NEG)) and HILIC-POS - were subjected to pre-processing and quality analyses. In this thesis, the analytical details and pre-processing results are provided for the HILIC-POS-MS dataset only. Since observations and conclusions were similar for all three assays, it was decided to focus on the most challenging dataset, which is characterised by more significant RT drifts, as well as lipid elution interference at around four to five minutes, which challenge metabolite annotation.

Upon sample delivery, sorting and overnight thawing at 4°C, 39 sets of 80 samples were formatted in 96-deep-well plates with two columns left empty for addition of pooled QC and external reference samples. The order of the samples was randomised using an in-house laboratory information management system. Each plate here is referred to as sample batch. On the day of the LC-MS analysis, a single batch of samples was prepared according to the standard NPC operating procedure with minor modifications [136]. In brief, samples were thawed at 4°C for 2h. Subsequently, samples were spiked with HILIC internal standards (e.g. adenine-2-d1, visualised in Figure 2.6). To aid protein precipitation, acetonitrile was added to spiked sample. The plate was centrifuged to separate the homogenous supernatant from the precipitated protein. The resulting supernatant was dispensed into a 96-well plate and centrifuged for 5 min prior to UPLC-MS analysis.

A HILIC UP-LC protocol optimised for large-scale metabolic profiling was used to analyse the AIRWAVE cohort samples, as reported in Lewis et al. [62]. The chromatographic separation of analytes was achieved using a 2.1 Œ 150 mm Acquity BEH HILIC column (Waters Corp., Milford, MA, USA) on an ACQUITY UPLC (Waters Corp., Milford, MA, USA) chromatography system. The solvent mixtures used for the mobile phase were: (A) 20mM ammonium formate in water with 0.1% formic acid; (B) acetonitrile with 0.1% formic acid. The initial flow rate of 0.6 mL/min was used during sample loading and gradient elution, which was swiftly increased to 1.0 mL/min at 7.8 min in order to accelerate chromatography system equilibration. The extended equilibration step, which takes almost half of the total run time, was optimised for HILIC in particular, which is known to benefit from longer equilibration stage in comparison to other chromatography methods. The initial isocratic separation, during which the composition of the mobile phase remained constant (95% solvent B), for the first 0.1 min was followed by a two-stage gradient: (1) a shallow gradient between 95% and 80% of solvent B; (2) rapid gradient from 80% to 50% of solvent B. Such a two-stage chromatographic protocol was shown to achieve approximately uniform peak shape for both early and late eluting polar analytes and help avoid broad peaks [62]. The detailed gradient conditions are provided in Table 2.2. The chromatography system was connected to Xevo G2-S Q-ToF mass spectrometer (Waters Corp., Manchester, UK) with Zspray electrospray ionization (ESI) source.

TABLE 2.2: Chromatographic gradient of the HILIC method showing the duration and mobile phase composition of each step. Solvents were as follows: A - 20mM ammonium formate in water + 0.1% formic acid; B - acetonitrile + 0.1% formic acid.

Time (min)	Flow rate (mL/min)	A(%)	B(%)
Initial	0.6	5	95
0.1	0.6	5	95
4.6	0.6	20	80
5.5	0.6	50	50
7	0.6	50	50
7.1	0.605	5	95
7.2	0.61	5	95
7.3	0.62	5	95
7.4	0.65	5	95
7.5	0.7	5	95
7.6	0.8	5	95
7.7	0.9	5	95
7.8	1	5	95
12.5	1	5	95
12.65	0.6	5	95

HILIC-POS-MS data was acquired in analytical batches of 1,000 samples, each of which was made of study samples, external reference samples, study pool samples and two dilution series (Table 2.3). External reference samples came from a large pool of serum that is maintained and used as an independent sample reference throughout all studies within the NPC. To create the pool, 10L of bulk serum were purchased from Seralab, homogenised, and aliquoted for long term storage [136]. A pooled QC sample was prepared for the whole study by thawing each study sample and combining their contents. The dilution series of the study pool samples was prepared for each UPLC assay separately at 1%, 10%, 20%, 40%, 60%, 80% and 100% of the original concentration. A single column was used for the duration of the whole experiment and ionisation source was cleaned in between batches.

TABLE 2.3: Summary of samples acquired in each analytical batch of HILIC-POS-MS dataset of the AIRWAVE1 serum cohort.

	Batch 1	Batch 2	Batch 3
Dilution series	92	92	92
External reference	104	100	96
Study sample	1,027	995	953
Quality control	103	99	96
All	1,326	1,286	1,237

2.2.2 Raw data conversion

Acquired proprietary Waters files were converted to open-source file format mzML with ProteoWizard software (version 3.0.18302) using *msConvert* command line tool with the following arguments:

msconvert –zlib –filter "scanEvent 1" –filter "threshold absolute 100 most-intense"

where argument *–zlib* refers to zlib compression for improved data storage efficiency; *–filter "scanEvent 1"* refers to the Waters MS function to be extracted from Waters .RAW format files; *–filter "threshold absolute 100 most-intense"* specifies the removal of data points with less than 100 absolute intensity.

The same *msConvert* parameters were applied to all of the datasets analysed within this thesis.

2.2.3 Endogenous metabolites detection and integration

A set of endogenous and xenobiotic metabolites known to be commonly present and detectable in human serum were identified in the AIRWAVE samples. Metabolite identification was performed by Dr Goncalo Correia and Mr Benjamin Cooper at the NPC. In total, 46 metabolites and their main adducts and in-source fragments were identified in the LC-MS spectra of all AIRWAVE samples using *m/z* and RT regions established in previous annotation projects at NPC. The *m/z* and RT integrations regions for each spectral feature were optimised for every AITRWAVE sample using R package peakPantheR that is available on the public GitHub repository: `https://github.com/phenomecentre/peakPantheR`.

The broad integration regions for the validated metabolites are available in Appendix A. Exemplar annotation plots obtained by G. Correia and B. Cooper are provided in Appendix A.

2.2.4 Pre-processing and data quality assessment

Chromatographic peak detection, grouping and subsequent features filling was performed using XCMS (version 3.0.0) [96, 107] running on R version 3.4.0. The parameters for *centWave* detection were attempted for optimisation using R package IPO (version 1.4.0) [137]. The final set of parameters are listed in Table 2.7.

XCMS-generated datasets were subjected to post-processing data quality assessment and correction according to standardised QC procedures for metabolic profiling [62, 92]. QC procedures were performed using Python library nPYc-toolbox [138], which is available on GitHub at `https://github.com/phenomecentre/nPYc-Toolbox`.

The recommended dataset quality assessment is based on: (1) intensity correlation to the matrix concentration in serial dilution samples, (2) relative standard deviation (RSD) and (3) retention time. The correlation to dilution value for feature i is

the Pearson correlation coefficient between the features measured concentration and the expected concentration of the serial dilution sample [139]. Correlation to dilution, ranging between -1 and 1, therefore accurately reflects feature's measurements quality. As only features that respond to dilution linearly provide meaningful information, features with low correlation values are removed.

The relative standard deviation (RSD) for each feature i is calculated from repeated measurements, the pooled QC samples, and is defined as the ratio of the standard deviation and the mean:

$$rsd(i) = \frac{\sigma_i}{\mu_i} \times 100 \quad (1)$$

Feature filtering was performed according to the standard procedures at the NPC [62]. A given feature must meet the following quality control criteria to be retained in the dataset:

- Pearson correlation to dilution > 0.7
- RSD in study pool samples < 30
- RSD in study samples * 1.1 > RSD in study pool samples
- Elution after 0.6 min (analytes eluting as a single peak in the solvent front are subjected to ion suppression effects)
- Elution before 10.5 min (few analytes elute near the end of the chromatographic gradient)

Unsupervised multivariate analyses were employed to evaluate run order effect. Filtered datasets were unit-variance scaled [140] and subjected to principal components analysis (PCA). The optimal number of principal components was estimated using 7-fold cross-validation. K-fold cross-validation is one of the most common strategies for the assessment of the quality of a model [141]. A K number of test sets are made from non-overlapping subsets of the original dataset with $\frac{1}{K}$ of the samples. For each of the K-folds, a model is trained using all of the samples that are not part of a given test set. The model is then evaluated with the test set and the evaluation scores are retained. K is set to 7 by default within the Python library nPYc-toolbox that was used for data quality assessment [138].

Potential associations between the latent structures in the data and analytical and biological sources of variation were evaluated by either correlating (for continuous variables) or applying a Kruskal-Wallis test (for categorical variables) to each analytical parameter and PCA scores of every principal component. The investigated sources of analytical variation are explained in Table 2.4.

TABLE 2.4: Analytical variation sources were investigated for potential association with the latent structures in the AIRWAVE data. Definitions for short variable names are provided.

Continuous		Categorical	
TOF	MS chamber vacuum pressure	Plate	96-well plate number
Run order	Sample injection order	Well	Position on a 96-well plate
Detector	MS detector voltage	Sample batch	Sample preparation batch
Collision	MS collision energy	Sample position	Position in sample preparation batch
Backing	MS vacuum backing pump flow		

Run-order/batch correction method was based on the locally estimated scatter-plot smoothing (LOESS) approach, defined in [92]. LOESS function, applied to each feature independently, is fitted to QC samples, taking a subset of samples at a time and fitting the correction curve locally, rather than to the whole data. Intensity values in each sample are corrected by dividing original value by the interpolated value of the correction curve at sample's position. The LOESS estimator smoothing parameter was set to 11, which is the number of QC samples to be used for local curve fitting.

All scripts used within this and other chapters are available on the public GitHub repository: `https://github.com/lauzikaite/PhD_thesis_code`.

2.3 Results

2.3.1 Analytical batches characterisation

Three AIRWAVE1 serum LC-MS datasets - HILIC-POS, lipid RP-NEG and lipid RP-POS - were subjected to pre-processing and quality analyses. In this thesis, however, the analytical details and pre-processing results are provided for the HILIC-POS-MS dataset only. Since observations and conclusions were similar for all three assays, it was decided to focus on the most challenging dataset, which is characterised by more significant RT drifts, as well as lipid elution interference at around four to five minutes, which challenge metabolite annotation.

The quality of the acquired HILIC-POS-MS data was examined prior to XCMS processing. The potential drifts in intensity and retention time were investigated using pooled QC samples. Only the QC samples that were analysed in between the study samples were used for this task, as the first QC samples in the experiment are used to condition the system and therefore would not accurately reflect the technical variation to which the study samples are subjected. 10 QC samples were selected from each analytical batch to cover the full experimental run of the whole batch equally. Therefore, in the following figures these samples are named according to their relative order in such a subset of QC samples of a single batch.

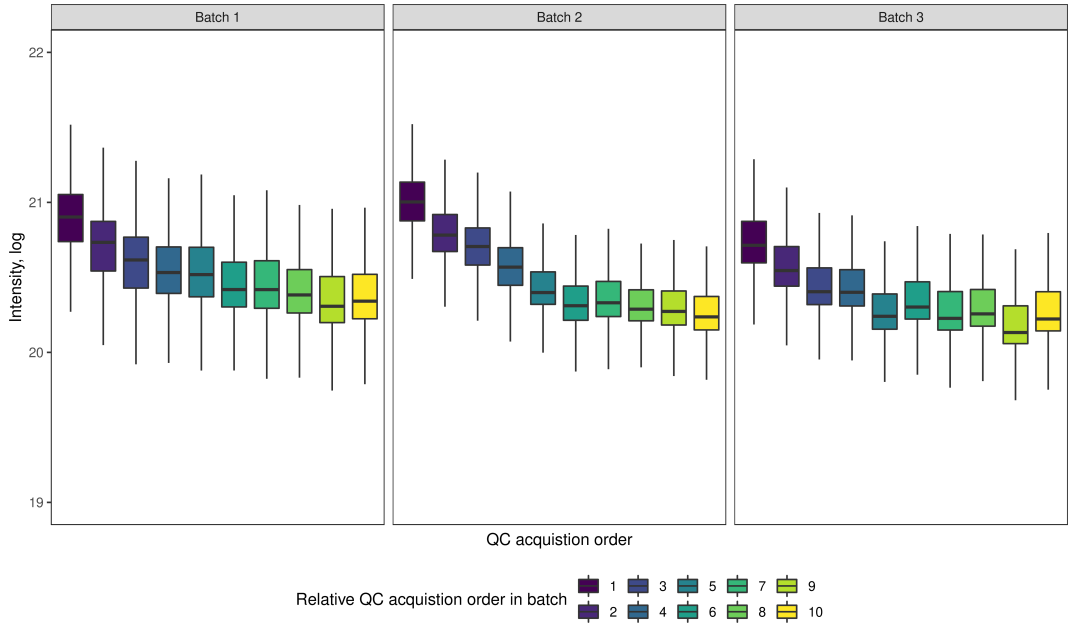


FIGURE 2.1: The distribution of total ion current signal from all detected ions in each mass spectrometer scan versus sample acquisition time. Total ion current is plotted for 30 QC samples in three analytical batch of the AIRWAVE serum HILIC-POS-MS dataset. QC samples were selected to cover the full experimental run of each batch equally.

A substantial signal intensity drift occurred in each analytical batch of the serum HILIC-POS-MS experiment. This is visualised by the distribution of total ion currents (i.e. the summed intensity of all ions per single mass spectrum) in the QC samples (Figure 2.1). The drop in the total ion current between the first and the last QC samples in each analytical batch was clearly observed. The presence of intensity drift represents an important issue since its effects can be large enough to mask the subtle biological variation, particularly for low abundance metabolites, which could fall below the limit of detection in the samples analysed later in the run. The effect of the intensity drop on the quality of the final processed data is discussed in the later sections of this chapter, particularly in Figure 2.10. A similar observation is provided by the base peak intensity (BPI) chromatograms (Figure 2.2), in which a drop in intensity over experimental run is noticeable. The most intense peak in each spectrum tends to drop with each sample in the run.

It is important to note that intensity drift followed a similar pattern in each analytical batch - ion intensities fell in the beginning of the run and reached a plateau in the middle of the analytical batch (Figure 2.1). This pattern can be attributed to loss of instrument sensitivity due to initial contamination of the ion source with the sample materials, as well as conditioning of the MS detector. Nevertheless, this does not suggest that larger analytical batches could be easily acquired, as other types of technical variation are introduced during extended sample analysis, mainly chromatographic RT deviation, which is discussed in the next section and in Figures 2.5, 2.3, 2.4 and Appendix A. LC-MS system is restored between the batches through

thorough cleaning of the ion source components, re-calibration, conditioning and even column replacement, if needed [43, 63, 123]. Nevertheless, BPI chromatograms (Figure 2.2) indicate that the third analytical batch differs from the first two even though column was not replaced during the acquisition of the HILIC-POS-MS data. Unfortunately, such variation between batches is commonly observed [128] and is the reason why data pre-processing is usually applied batch-wise.

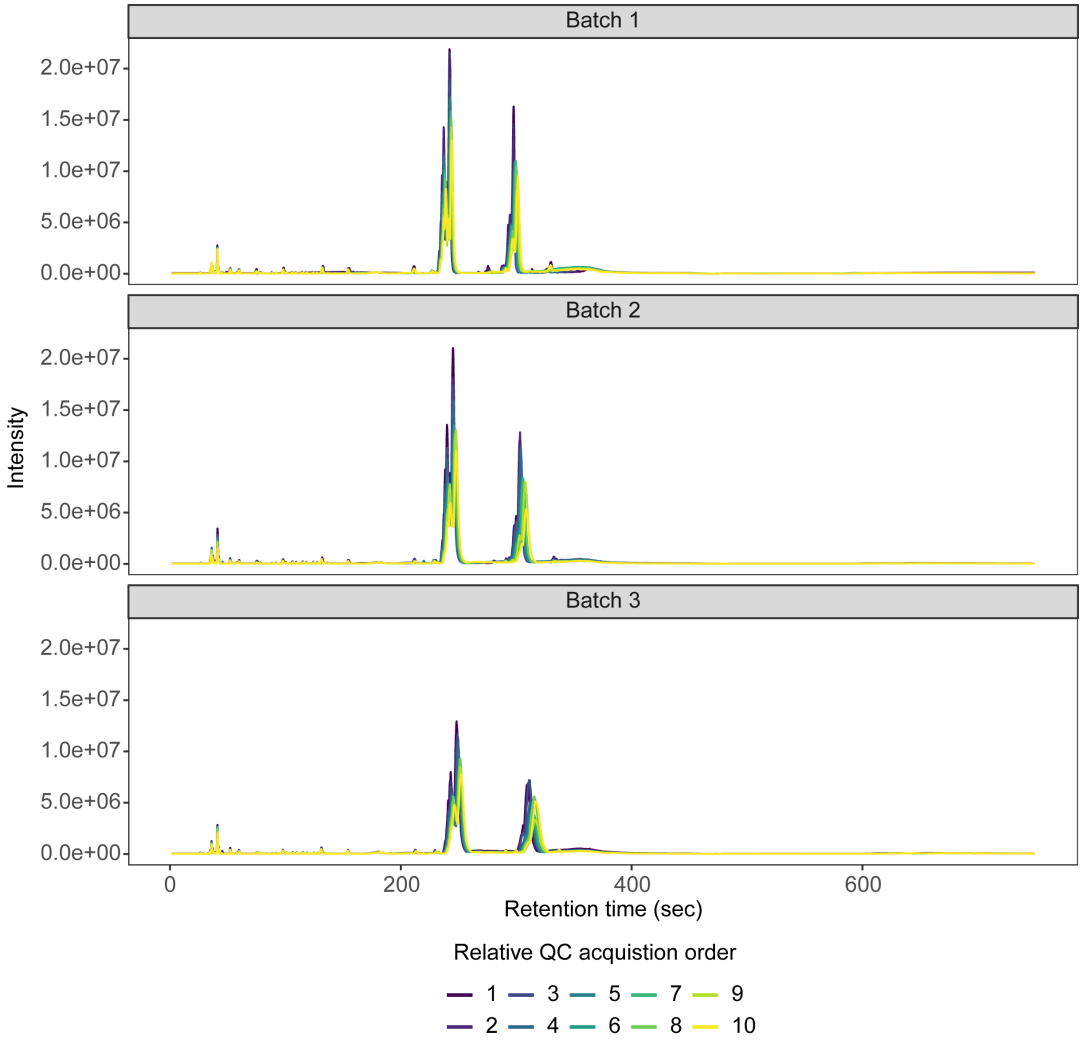


FIGURE 2.2: Base-peak intensity (BPI) chromatograms of 30 QC samples in the three analytical batches of the serum HILIC-POS-MS dataset.

To evaluate the retention time drift that could have occurred during data acquisition, a set of 46 endogenous and xenobiotic metabolites were identified in raw LC-MS spectra. EIC chromatograms for ions corresponding to their main adducts and in-source fragments were obtained for all samples analysed with HILIC-POS-MS. The extraction of features corresponding to carnitine (Figure 2.3) and a-glycerophosphocholine ions (Figure 2.4) indicate that chromatographic peaks shifted between samples and in between batches significantly. Summary plots for

other validated metabolites are provided in Appendix A. The *m/z* and RT integrations regions for 46 ions corresponding to validated metabolites' adducts and in-source fragments are available in Appendix A, Table A.1.

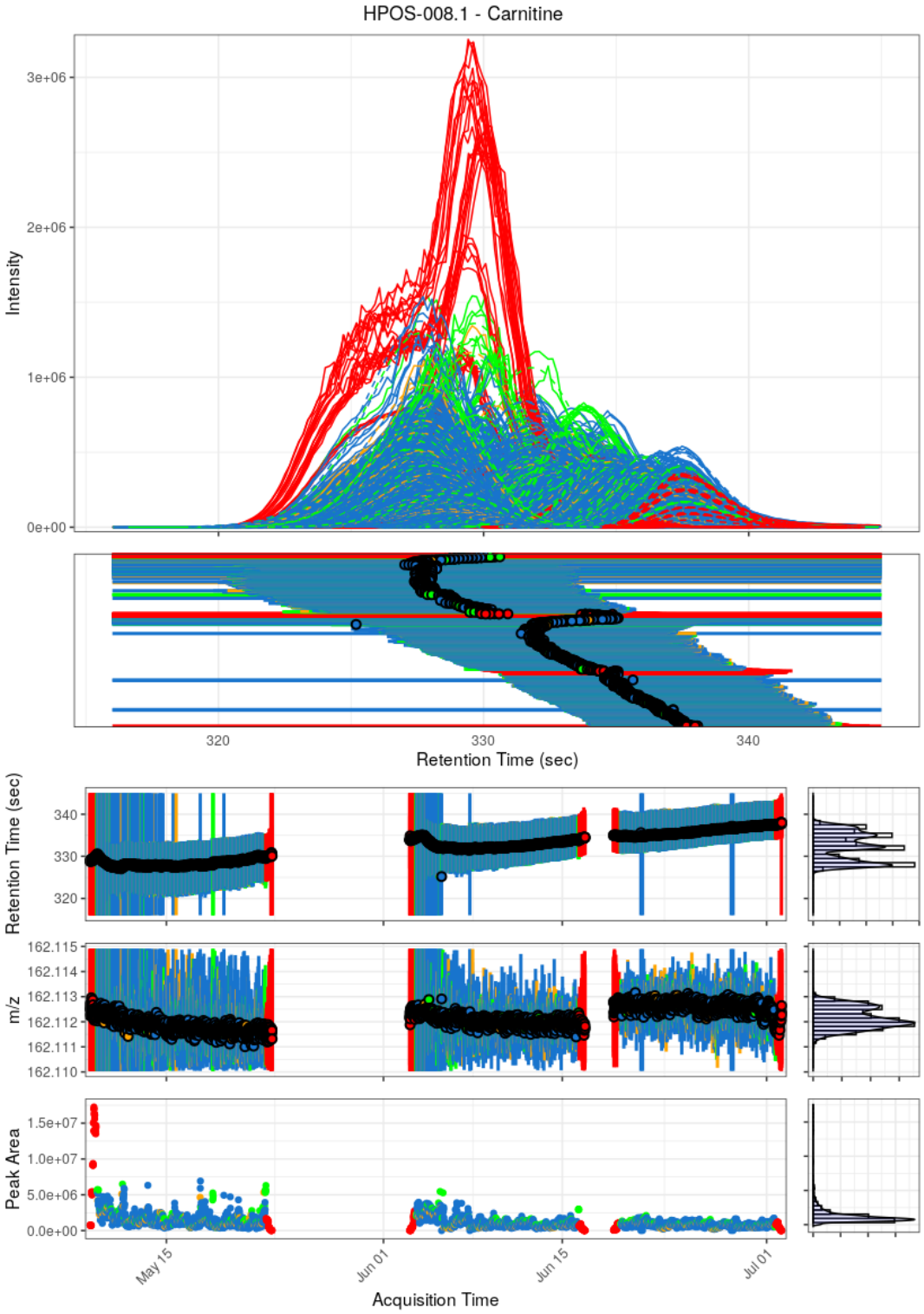


FIGURE 2.3: Detection and integration of carnitine main adduct ion in all AIRWAVE serum HILIC-POS-MS samples. Different colours indicate sample type: red - QC dilution series, green - QC sample, orange - external reference sample, blue - study sample.

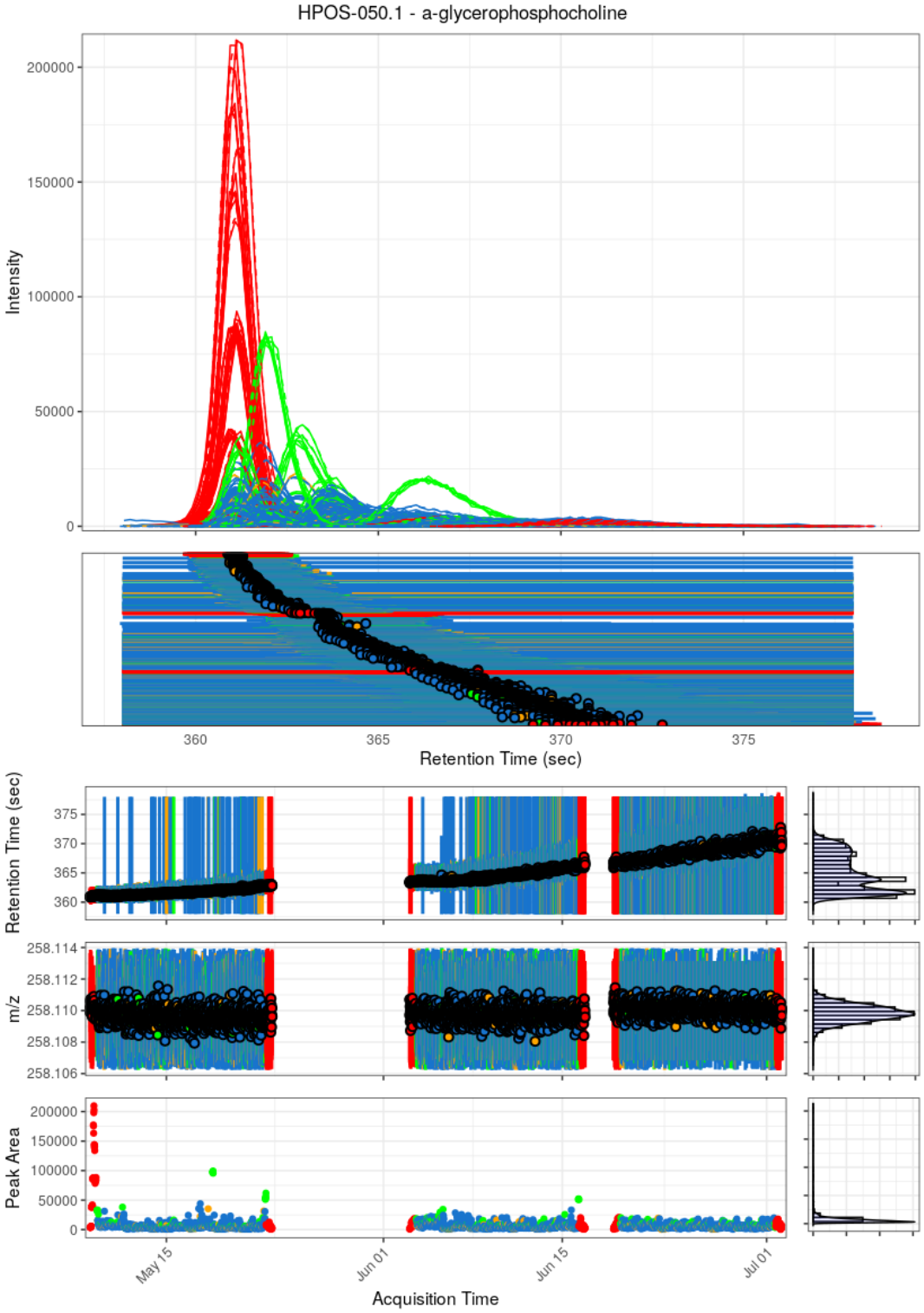


FIGURE 2.4: Detection and integration of a-glycerophosphocholine main adduct ion in all AIRWAVE serum HILIC-POS-MS samples. Different colours indicate sample type: red - QC dilution series, green - QC sample, orange - external reference sample, blue - study sample.

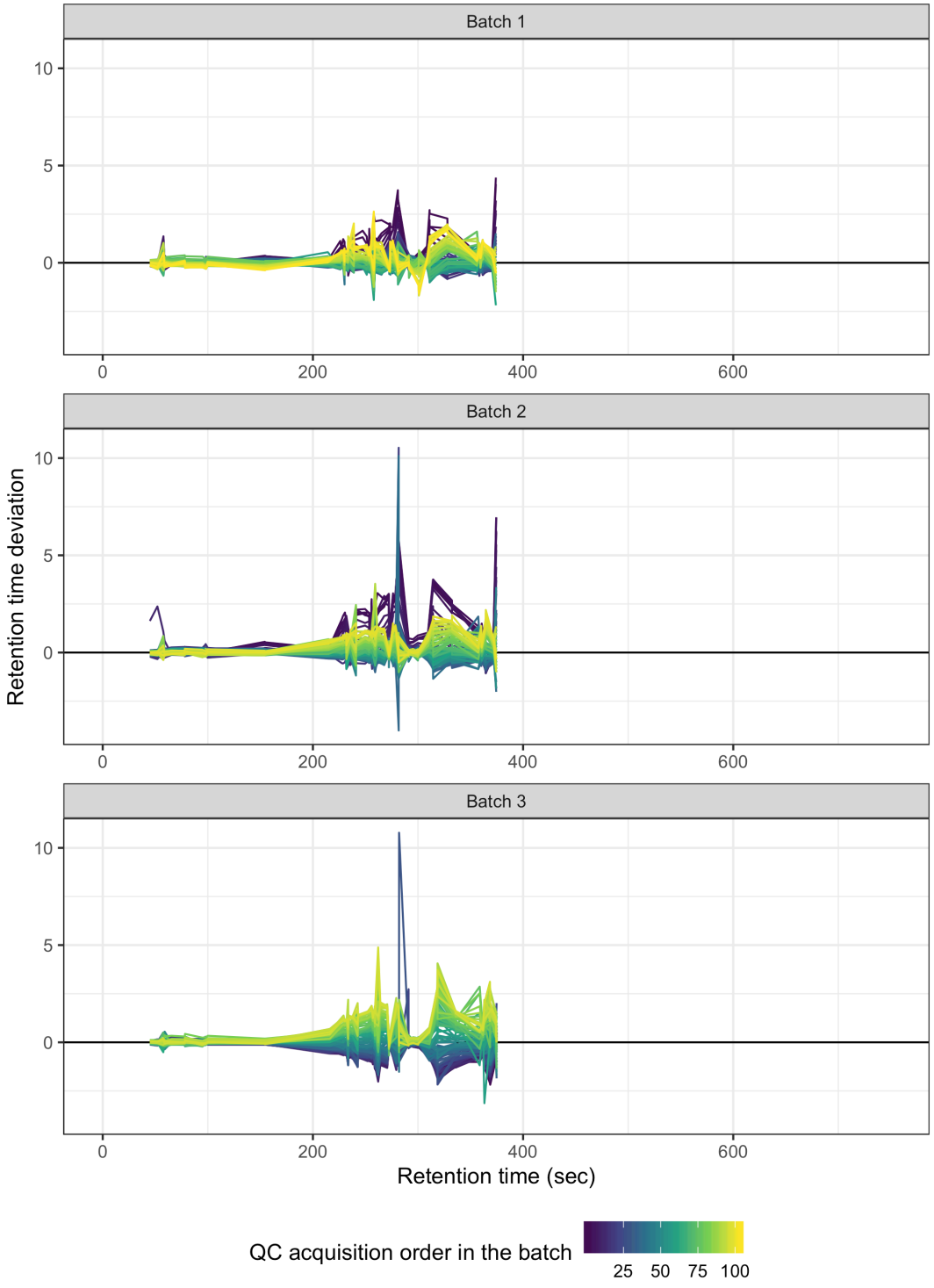


FIGURE 2.5: AIRWAVE serum HILIC QC samples retention time (RT) deviation (in seconds) from the analytical batch median is plotted for 46 validated metabolites. Each line represents a single QC sample that is coloured according to its acquisition order in the analytical batch.

The extracted RT values for all validated metabolites were used to model the RT drift observed during experimental run time. For each metabolite, the observed RT in a given sample was compared to the median RT for that metabolite in the analytical batch. The obtained RT deviations for every QC sample are plotted against the average metabolite RT. Figure 2.5 illustrates that RT drift varies largely between metabolites eluting at different times. The scale of the deviation also changes during sample acquisition and there is gradual RT deviation from the beginning to the end of the batch.

2.3.2 XCMS parameters optimisation

It was attempted to optimise *centWave* peak-picking parameters with an open-source package IPO [137]. The IPO parameter optimisation approach is based on design of experiments (DoE). A designed experiment comprises of a number of tests, in each of which specific alterations are maded to the input variables. In the context of the IPO package, DoE optimises *centWave* parameters by evaluating the quality of peaks detected in each experiment. Peak picking quality in each experiment is quantified by peak picking score (PPS), which is defined as the ratio between the number of reliable peaks and the number of non-reliable peaks. Reliable peaks here are defined as the peaks corresponding to stable ^{13}C isotopic peaks, whereas non-reliable peaks are those that do not belong to the ^{13}C isotope cluster. Once PPS are obtained for each experiment of the DoE, response surface models are estimated and applied to identify the combinations of parameters resulting in the highest PPS score.

Parameter optimisation for AIRWAVE datasets was performed with a varying number of QC samples to evaluate parameters robustness. QC samples were selected to evenly cover the full experimental run of a single batch. To reduce the computational time, starting values were set close to the expected values. Lower and upper starting values for parameters that were optimised for HILIC-POS-MS QC samples are listed within parentheses in Table 2.5. Parameters which were not optimised by IPO are listed as a single starting value.

TABLE 2.5: *centWave* peak-picking parameter optimisation was performed using IPO package. Optimisation was repeated using varying number of HILIC-POS-MS QC samples to evaluate optimisation robustness. Parameters that were subjected to optimisation are listed within parentheses, which contain lower and upper starting values. Parameters that were provided for the IPO experiments but were not subject to optimisation are listed as a single value.

	Starting values	Optimised values			
Parameter	All tests	5 QC	10 QC	20 QC	30 QC
peakwidth, min	c(1.5, 3)	3	3	3	3
peakwidth, max	c(5, 20)	26.75	26	26	26.75
prefilter, k	c(4, 10)	3.4	3.1	3.4	3.1
prefilter, I	c(500, 10000)	1	1	1	1
noise	c(200, 1000)	1	1	1	1
snthresh	c(3, 5)	4.7	4.6	4.7	4.7
fitgauss	FALSE	FALSE	FALSE	FALSE	FALSE
integrate	2	2	2	2	2
mzCenterFun	wMean	wMean	wMean	wMean	wMean
mzdiff	0.01	0.01	0.01	0.01	0.01
ppm	25	25	25	25	25
verboseColumns	FALSE	FALSE	FALSE	FALSE	FALSE

While IPO optimisation experiments returned similar parameter values with varying numbers of QC samples, suggesting a certain level of optimisation robustness and stability, the returned values do not satisfy our expectations. First of all, suggested peakwidth of 3 to approximately 26 seconds does not reflect the peaks observed in the raw spectra. A few examples of much narrower chromatographic peaks are demonstrated by the extracted ion chromatograms for known and unknown metabolites (Figures 2.6 and 2.7). IPO tendency for much broader peaks can be appreciated given that a minimum of eight to ten data points across a chromatographic peak are required to be able to define its shape. The high-throughput Waters LC-MS system that was used to analyse the AIRWAVE cohort acquires 0.084 scan per second (0.07 second scan time + 0.014 second interscan time). At such speed, a peak with 1.5 seconds peakwidth width would have approximately 17 data points. Typically, 10 to 15 data points are sufficient to achieve high quantitative reproducibility and to distinguish the shapes of co-eluting peaks [142]. Therefore, given the observed narrow peaks and high data acquisition rate, a peakwidth of minimum 1.5 seconds would reflect the data more accurately that the value returned by IPO. Secondly, the maximum peakwidth value of 26 seconds optimised by IPO experiments pose a danger of merging multiple peaks together. The in-house HILIC protocol typically produces peaks of up to 10 seconds in peakwidth. The widest peak at 230 - 270 seconds, as seen in the BPC chromatogram in the earlier Figure 2.2, is actually comprised of multiple overlapping peaks, each of which is around 10 seconds in

peakwidth (Figure 2.8). Therefore, IPO returned value of 26 seconds is unlikely to facilitate accurate peak-picking.

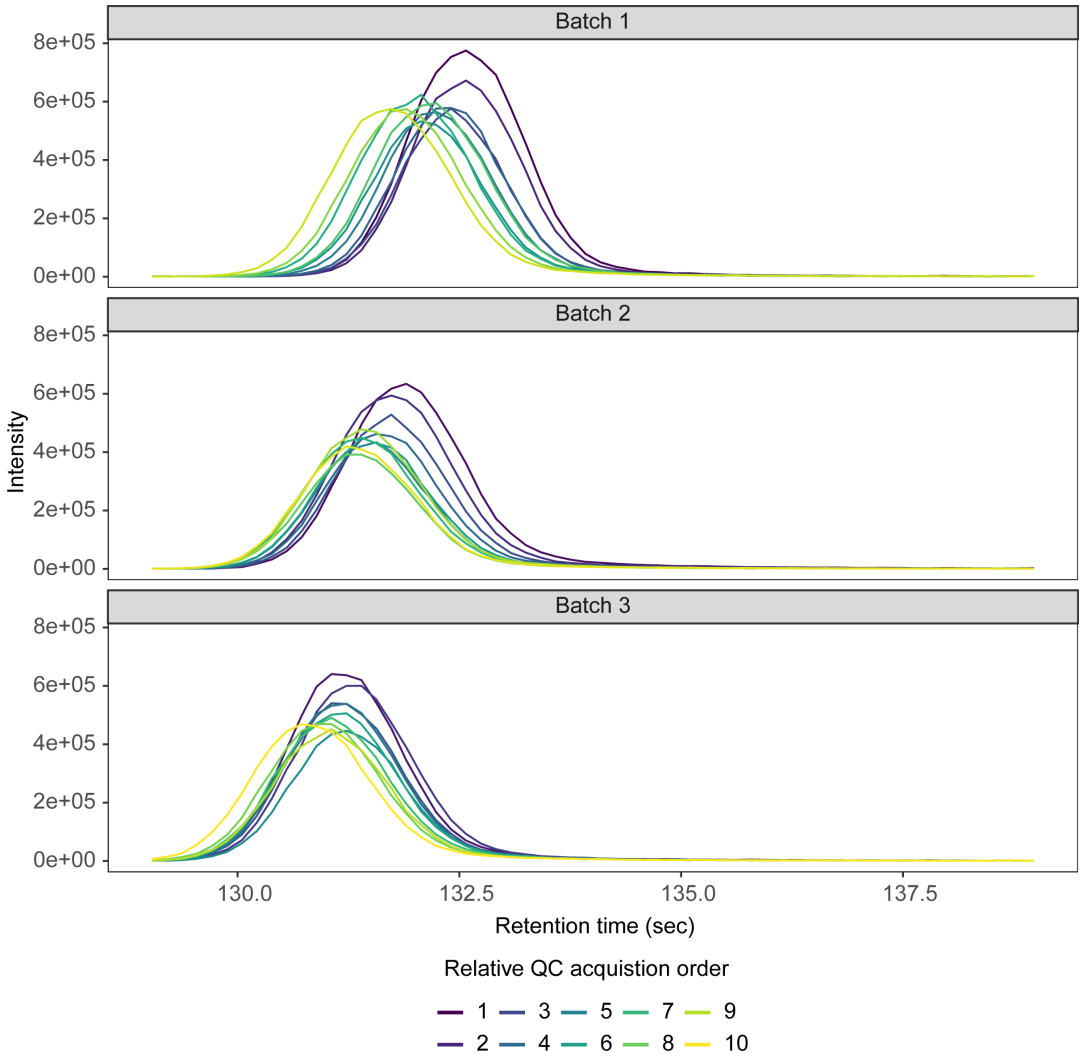


FIGURE 2.6: Extracted ion chromatogram (EIC) of spiked-in internal standard adenine-2-d1 from 30 QC samples in three analytical batches of the serum HILIC-POS-MS dataset. The average peakwidth for this analyte was 2.5 seconds.

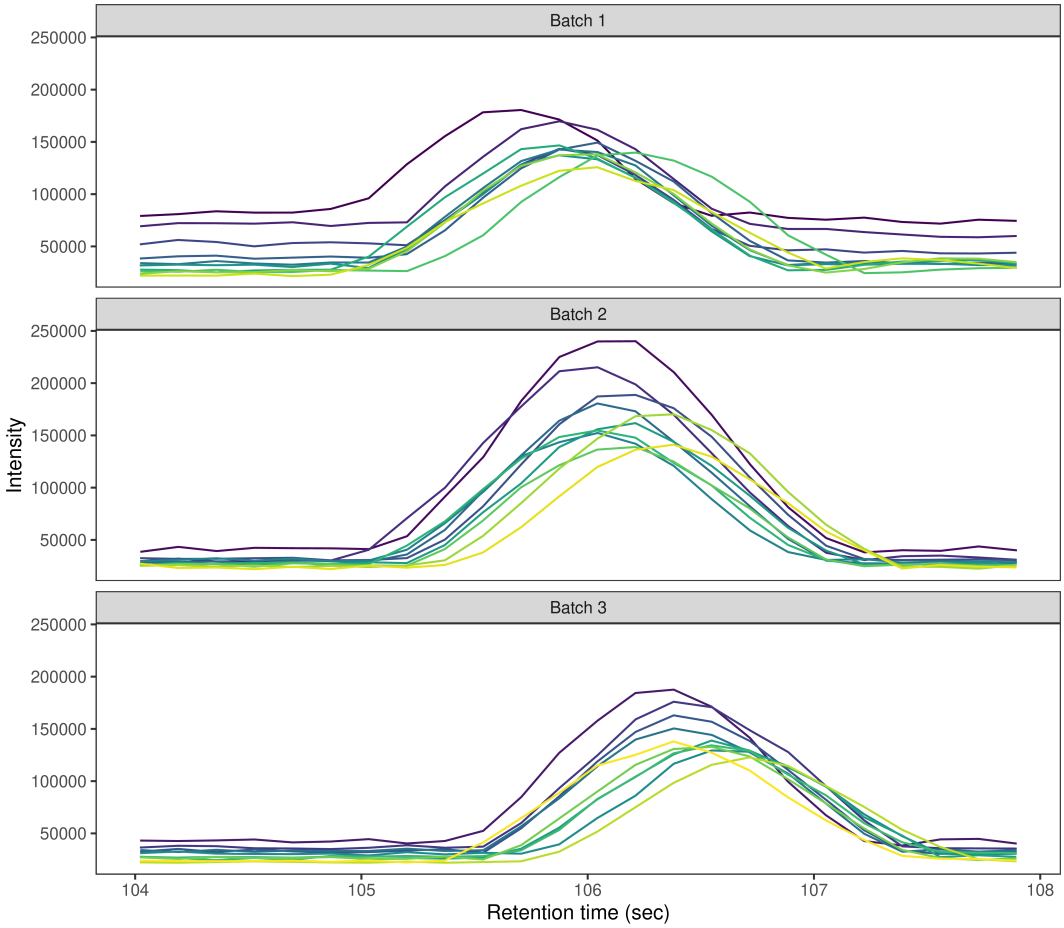


FIGURE 2.7: Extracted ion chromatogram (EIC) of unidentified metabolite from 30 QC samples in three analytical batches of the serum HILIC-POS-MS dataset. The average peakwidth for this analyte was 1.5 seconds.

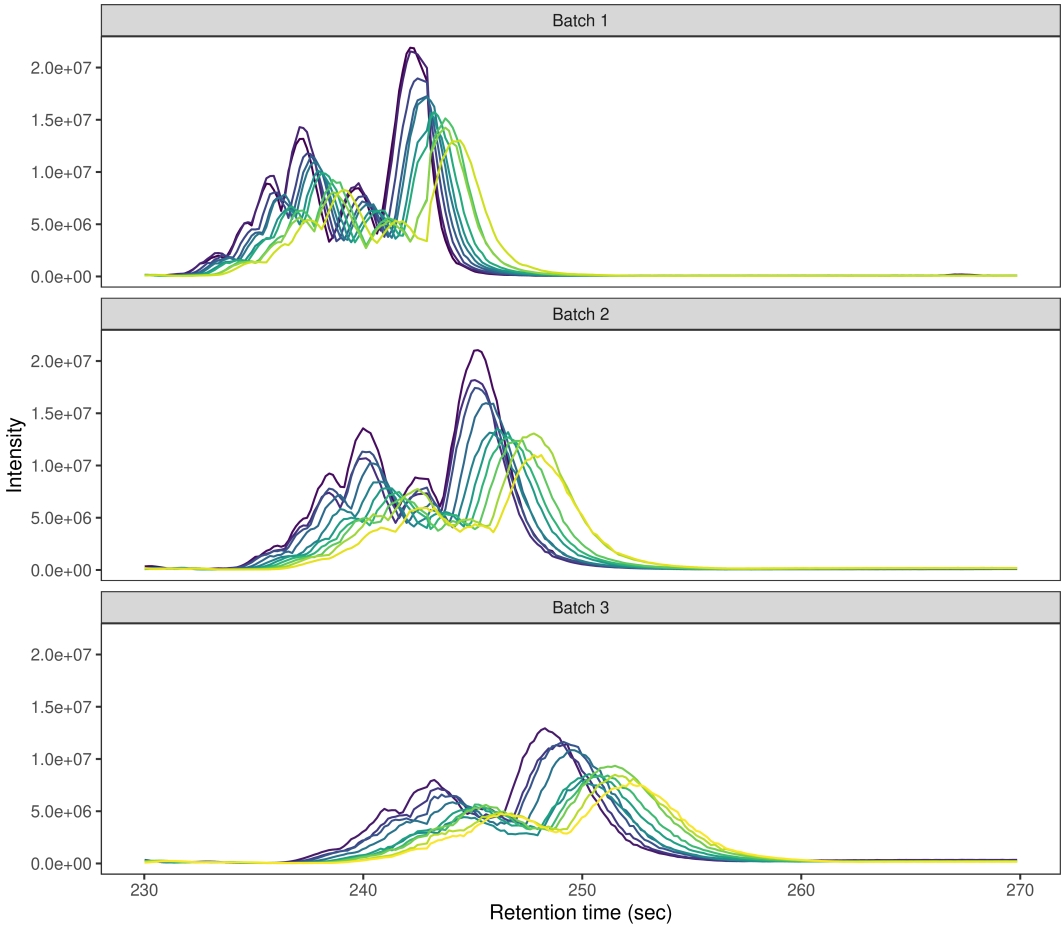


FIGURE 2.8: Extracted ion chromatogram (EIC) of unidentified metabolite from 30 QC samples in three analytical batches of the serum HILIC-POS-MS dataset. The average peakwidth for this analyte was 1.5 seconds.

Another argument against the use of IPO optimisation for this particular dataset is that the returned noise and prefilter parameter values are unreasonably low and inconsistent with the raw data. The noise parameter value was optimised to 1, whereas prefilter k and prefilter I values were set to approximately 3.1 and 1 respectively (Table 2.5). The filtering step in the *centWave* algorithm allows to remove noise-level features by retaining only the peaks that contain at least k consecutive values with intensity of $> I$ [107]. The IPO parameters-driven filtering step would effectively retain all regions of interest in the m/z domain since peaks which appear in 3.1 scans with intensity of > 1 would not be discarded. Visualisation of the mass spectrum at 8 minutes, which was shown to be the least intense chromatographic region in the earlier BPC chromatogram (Figure 2.2), suggests substantially high baseline levels (Figure 2.9). The red segments indicate m/z signals with intensity of < 500, whereas the black segments are m/z signals with intensity of > 500. A dense floor of ion signals suggest a noise level that is clearly higher than 1, as suggested by the IPO experiments. Therefore, higher noise and prefilter values are required in order to achieve high quality data.

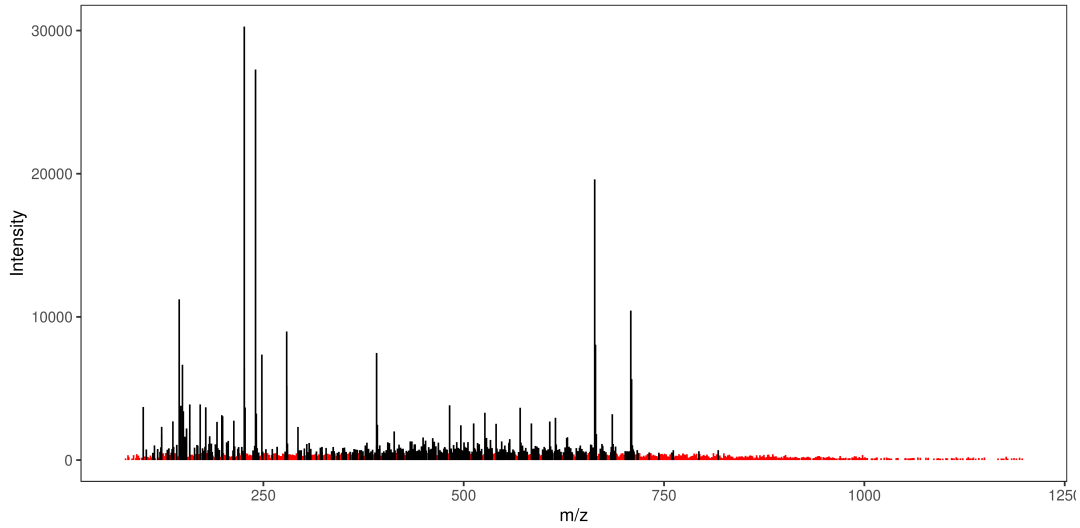


FIGURE 2.9: Mass spectrum of a representative QC sample at 8 min visualises high baseline noise level even at the least intense chromatographic region. Segments in red indicate *m/z* signals with intensity of < 500. Segments in black are signals with intensity of > 500.

To test the validity of IPO optimisation, *centWave* peak-picking was performed using the parameters returned by the IPO experiments. The number of features detected per sample are summarised in Table 2.6. As suspected given the low noise and prefilter parameter values, a large number of peaks were detected per each sample, ranging from 16k to almost 30k. Such large number of peaks is unusual for serum samples analysed by HILIC-POS-MS in-house. More importantly, such large peak tables would significantly increase the computational time taken by subsequent XCMS steps, which are already computationally expensive and a cluster computing environment was required to process such a large study. Similar and other disadvantages of using IPO for automatic parameter selection was discussed in Alboniga et al. [143] where it was concluded that IPO might lead to unrealistic parameters when challenging datasets are investigated.

TABLE 2.6: *centWave* peak-picking was performed using the parameters optimised by IPO. The number of detected peaks per single HILIC QC sample are listed for each each experiment, which was run with varying number of QC samples.

	Number of detected peaks			
QC sample	5 QC samples	10 QC samples	20 QC samples	30 QC samples
1	24678	24933	24650	24678
2	16912	20684	19977	19927
3	21975	17072	20429	26295
4	20527	21323	21765	20469
5	23332	22202	16883	21762
6	-	21125	20844	19451
7	-	20719	21091	16912
8	-	16774	22591	21074
9	-	23555	21997	29000
10	-	29367	28175	21098
11	-	-	20902	15628
12	-	-	22842	22876
13	-	-	20530	21975
14	-	-	19683	24687
15	-	-	16595	20222
16	-	-	21486	20921
17	-	-	23315	24634
18	-	-	21976	15978
19	-	-	29020	20527
20	-	-	20787	16962
21	-	-	-	18733
22	-	-	-	16608
23	-	-	-	19932
24	-	-	-	21152
25	-	-	-	23332
26	-	-	-	22239
27	-	-	-	19995
28	-	-	-	29044
29	-	-	-	21462
30	-	-	-	21187
Min	16912	16774	16595	15628
Max	24678	29367	29020	29044

2.3.3 XCMS pre-processing

Following unsatisfactory IPO optimisation, it was decided to select XCMS parameters by manually investigating raw LC-MS spectra. Final XCMS parameter values are listed in Table 2.7. Retention time adjustment using the OBI-warp method would be a preferred strategy [144]. However, the corresponding *retcor.obiwarp* method in XCMS version 3.0.0, which was available at the time of analysis, could not handle that many samples at once, all of which have slightly varying number of scans. While this bug had been fixed in the new XCMS 3 interface and the underlying methods, the peak-picking method that was required for the version 3 interface, *findChromPeaks*, took enormous amount of memory: 2 - 3 TB of RAM for a single analytical batch of 1,300 samples. Therefore, it was decided to use the original XCMS interface with *retcor.peakgroups* method for retention time adjustment (Table 2.7). Given that XCMS feature alignment methods do not distinguish within-batch and between-batch variation [96] and the observed batch effect in intensity and retention time drift, XCMS was applied to each batch separately using the same parameters.

XCMS pre-processing was applied to each of the analytical batches separately due to significant intensity drift and detectable chromatographic retention time drift between the batches, as discussed in the earlier section. First, *centWave* peak-picking method was applied. The number of detected peaks varies between the batches (Figure 2.10). The third analytical batch stands out from the first two since all types of samples have fewer *centWave*-detected peaks per sample. Such clear differences are mostly explained by the intensity drift, as illustrated in Figures 2.1 and 2.2. The batch-wise differences are also visible in the total ion chromatogram (TIC) of XCMS reported features (Figure 2.11).

centWave-detected peaks were grouped into features using the XCMS *density* method, which is a kernel density estimation algorithm applied to slices of *m/z* to group the peaks close in retention time. To correct for RT deviations between samples, RT adjustment was performed using XCMS *retcor.peakgroups* method, which is based on a local regression model that uses so called "well-behaved" peak goups as anchors. In these groups, fewer than 10 samples have no peaks assigned and fewer than 10 samples have more than one peak assigned to the group (parameters *missing* and *extra* respectively). The well-behaved groups that were identified after the initial round of grouping are demonstrated in the ions maps in Figure 2.12. A small number of peak groups pass the selection criteria: 42, 28 and 25 in the three analytical batches respectively. These automatically selected and unevenly distributed features were used to align the RT of all features across all samples.

The results of the RT correction are visualised in Figure 2.13. In comparison to the observed RT drift for the annotated metabolites (Figure 2.5), XCMS-mediated RT

correction in a given sample was applied much more evenly across the chromatographic domain. Furthermore, no RT regions were corrected for more than 3 seconds, whereas the observed RT drift in some cases was more than 10 seconds.

TABLE 2.7: XCMS methods and their parameters used in the pre-processing of AIRWAVE1 serum HILIC-POS-MS datasets. Note that original XCMS interface was employed in the pre-processing. Corresponding methods have different names in the newest XCMS version 3 interface. Methods are listed in the order of use.

XCMS		
Method	Parameter	Value
xcmsSet	peakwidth	c(1.5, 14)
	prefilter	c(10, 3000)
	noise	500
	snthresh	5
	fitgauss	FALSE
	integrate	2
	mzCenterFun	wMean
	mzdiff	0.01
	ppm	25
group	method	density
	minfrac	0
	minsamp	0
	bw	2
	mzwid	0.01
retcor.peakgroups	plottype	none
	smooth	loess
	missing	10
	extra	10
	span	10
group	method	density
	minfrac	0
	minsamp	0
	bw	2
	mzwid	0.01
fillPeaks	method	chrom

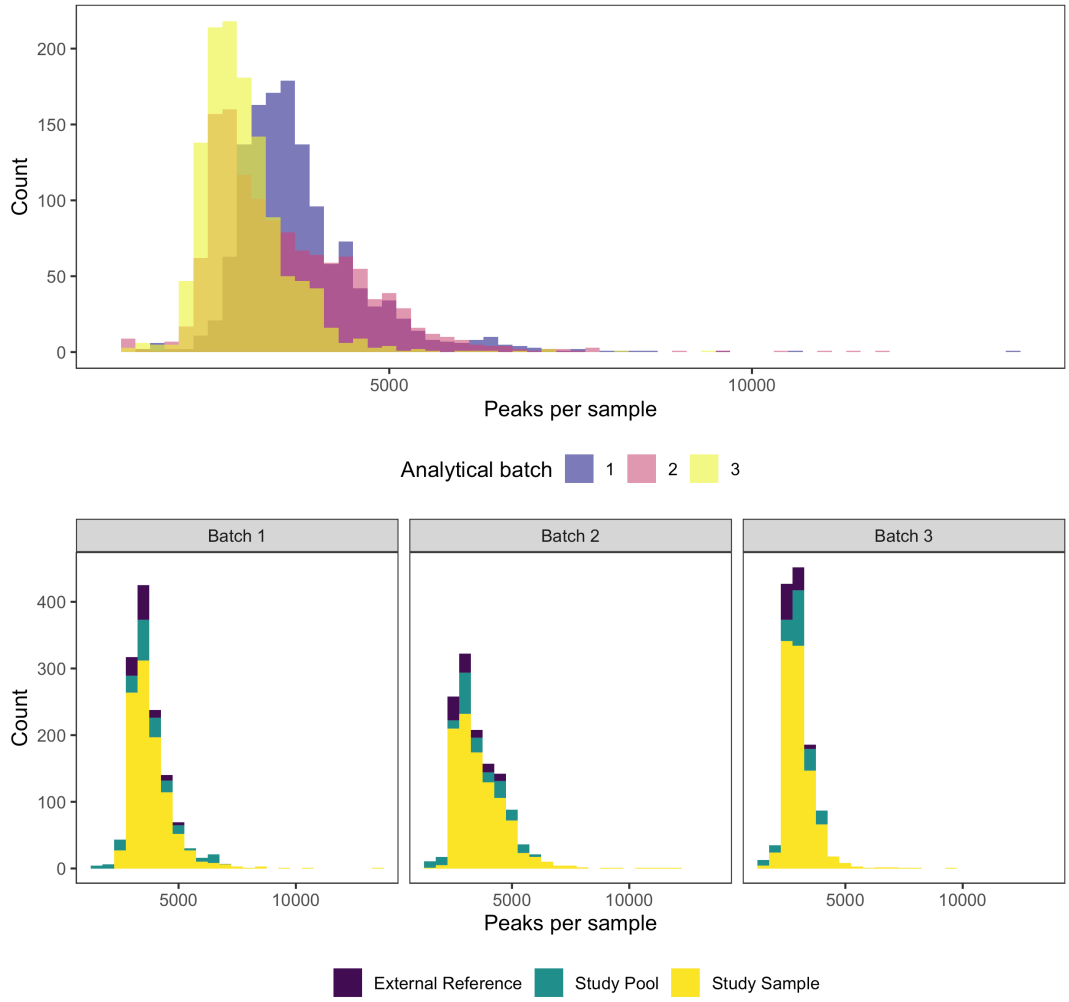


FIGURE 2.10: Number of *centWave*-detected peaks differs between different types of samples in three AIRWAVE HILIC-POS-MS analytical batches.

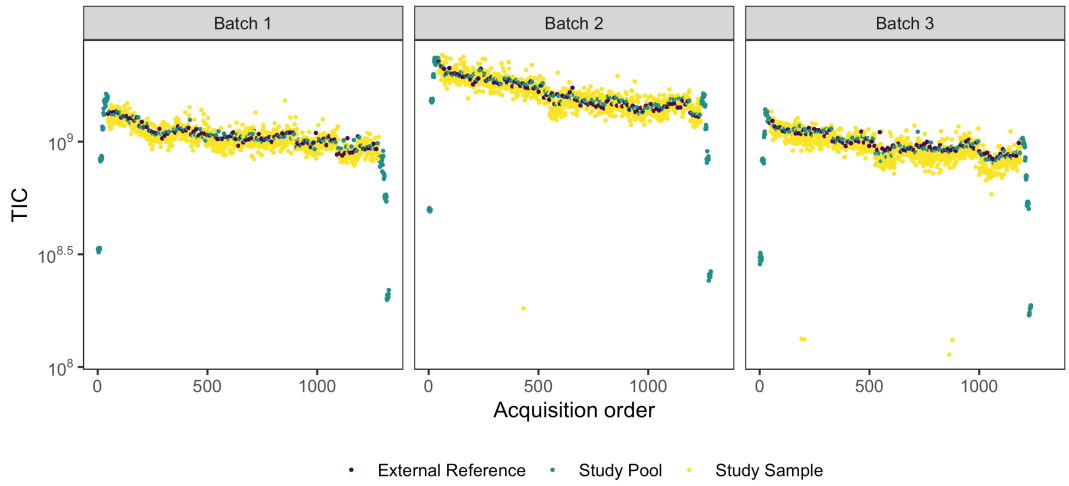


FIGURE 2.11: Total ion chromatograms of XCMS reported features for all analysed AIRWAVE samples.

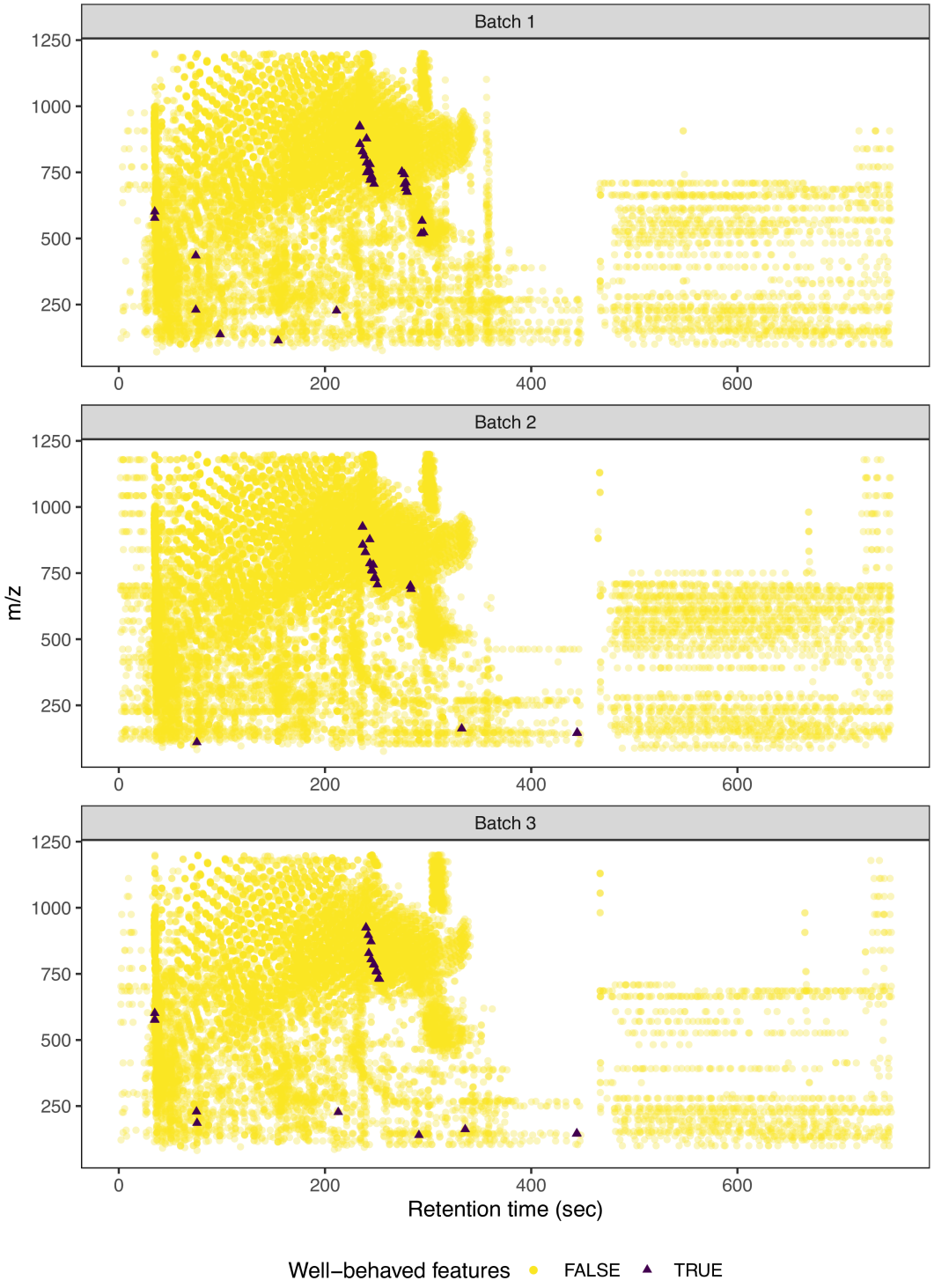


FIGURE 2.12: Ions maps of *centWave*-detected and *density*-grouped features for three AIR-WAVE analytical batches. The dark triangles represent the well-behaved features that were selected by the XCMS algorithm to act as anchors for kernel density based RT correction. These well-behaved features are few and clustered together, leaving a large proportion of the ion maps uncovered during RT correction.

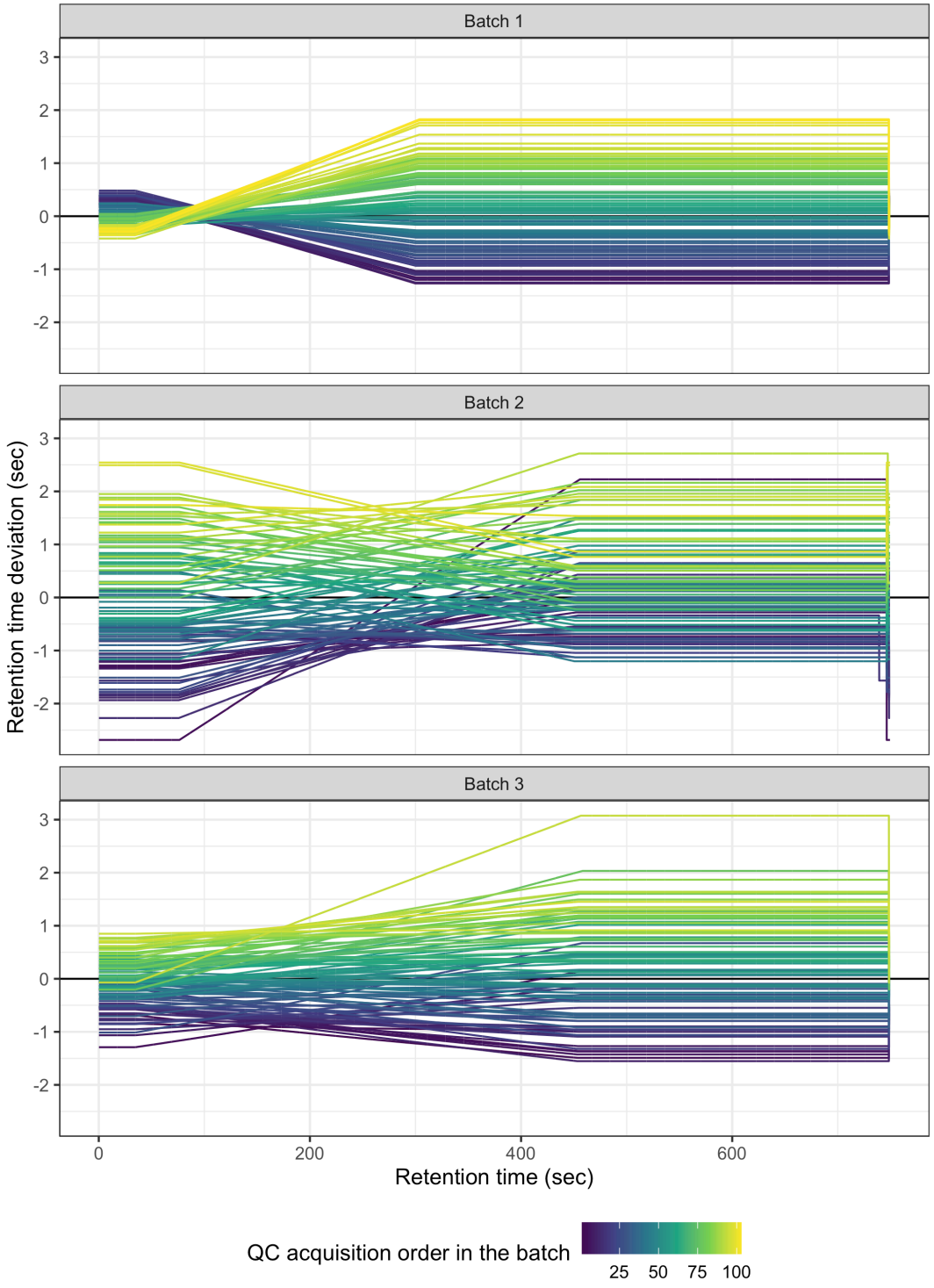


FIGURE 2.13: XCMS *retcor* retention time correction method was applied during AIRWAVE serum HILIC pre-processing. The deviation between the raw and the XCMS-corrected RT values in all QC samples in three analytical batches is plotted against the corresponding mass spectrometer scan (here presented as retention time in seconds). Each line represents a single QC sample that is coloured according to its acquisition order in the analytical batch.

2.3.4 Processed data quality assessment

To evaluate the quality of the XCMS-processed AIRWAVE data, a standard quality control procedure was applied using Python library nPYc [138]. First, the analytical precision of the obtained features was examined. The distribution of relative standard deviation (RSD) values estimated for all features across pooled QC samples is concentrated around 19% and is similar for all three batches (Figure 2.14). The distribution of correlation to dilution coefficients varies slightly between the batches, with the first one having a higher proportion of features with Pearson coefficients around 0.

Removal of low quality features according to the QC standards described in Section 2.2.4 produced datasets of relativity similar sizes even though the initial number of XCMS features varies largely between the three analytical batches, as summarised in Table 2.8. Overall, up to 63% of XCMS features are removed due to low analytical precision and low correlation to dilution.

TABLE 2.8: Number of total XCMS-reported AIRWAVE features and features that meet the quality control assessment criteria.

Analytical batch	Total features	Filtered features
1	24,771	11,178
2	30,155	10,881
3	19,745	10,015

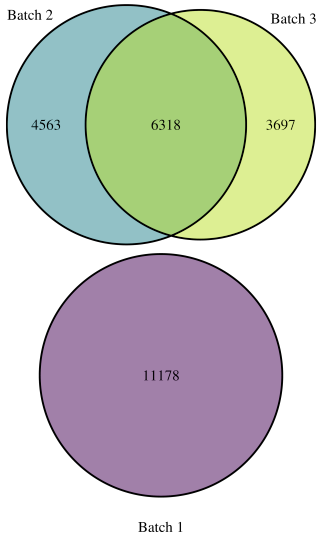
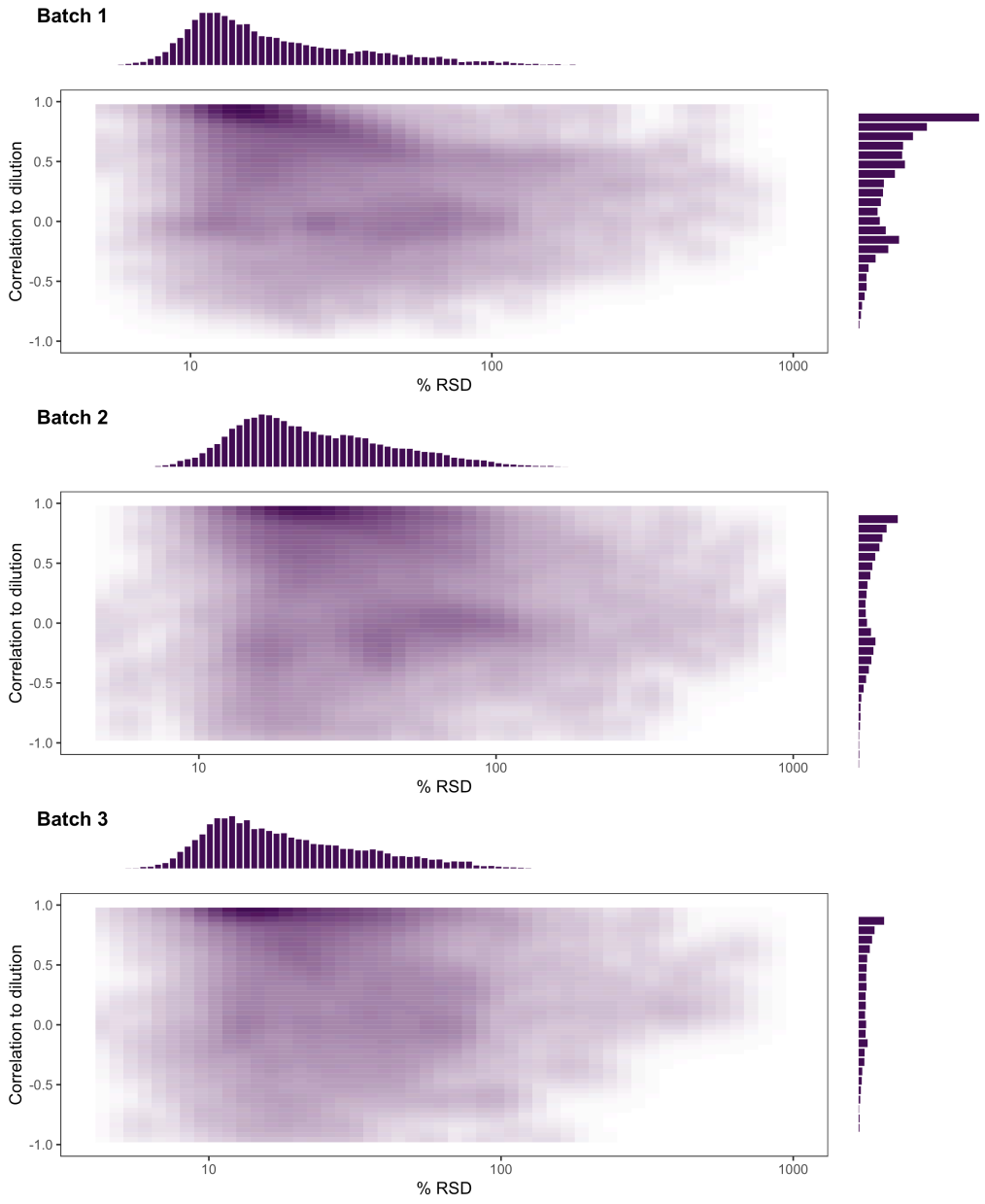


FIGURE 2.15: The number of XCMS features in the three AIRWAVE HILIC-POS-MS analytical batches after QC feature filtering. Generous *m/z* and retention time windows of 0.001 and 10 seconds were used to find matching features between the three batches.

Common features between the three analytical batches were identified using generous *m/z* and retention time matching windows of 0.001 and 10 seconds. Figure 2.15 shows Venn diagrams with sections representing the number of features unique in



[H]

FIGURE 2.14: The analytical precision (expressed as relative standard deviation, RSD) and linearity of response (correlation to dilution) of XCMS-detected features are visualised for the three AIRWAVE serum HILIC batches. All three batches have a large number of features with poor linearity of response.

individual batches, with the overlapping regions corresponding to features common between batches. These results demonstrate that batches 2 and 3 are more similar to each other than batch 1, which does not share any of its XCMS features. These observations indicate that either batch 1 suffered from stronger technical variation than batches 2 and 3, or the other way around. However, given the distribution of RSD and correlation to dilution coefficients (Figure 2.14), it is more likely that stronger analytical biases are present in the data for batch 1, which has a higher proportion of features poorly correlated to dilution. The high number of common features after the QC feature filtering between batches 2 and 3 is a strong indication that the applied QC pipeline is capable of removing such analytical biases to an extent. However, it is important to note that features matched using a 2-dimensional window of *m/z* and retention time errors do not necessary correspond to ions arising from the chemical compounds. A more in-depth analysis and validation with chemical standards would be needed to confirm the identities of such features. Nevertheless, such matching approach has been used in previous studies to evaluate the performance of peak-pickers [109, 145]. Here it serves as an additional exercise that brings us to the earlier observation that if peak detection and alignment algorithms are applied to samples subjected to analytical biases, the reported features will be different for each analytical batch of the same LC-MS experiment.

To further assess the quality of the dataset and to identify potential sources of analytical variation, multivariate analyses were performed. The scores of the calculated principal components (PC) were tested for association with analytical variables and clinical information (Figure 2.16, definitions of variable names are provided in Table 2.4). The strength of association between the first two PC and categorical variables - well and sample position on 96-well plate, sample batch, plate number, as well as subject's gender and BMI category, was evaluated using Kruskal-Wallis test. Tests with p-value of < 0.001 were denoted as significant. The strength of correlation between the first two PC and continuous variables - MS parameters, such as chamber vacuum pressure, detector voltage, collision energy and backing pump flow, as well as sample run order and subject's age - was tested using Pearson correlation. Variable pairs with correlation coefficient > 0.5 or < -0.5 were denoted as significant (marked with an asterisk in Figure 2.16). PCA analysis and PC association with technical and biological variables were performed with both raw features, generated by the XCMS pipeline, and batch-corrected features obtained using the earlier described feature removal and LOESS smoothing procedures.

Significant associations were identified between some of the analytical variation sources and the first two PCs in the raw XCMS features. First of all, Figure 2.16 indicates that the important sources of analytical variation in batch 1 and 3 are mainly MS detector voltage, sample preparation batch, as well as sample plate number and run order. These were explained by the second PC, which accounts for 4% and 5% of total variance respectively. That suggests that up to 5% of total variance in these

raw XCMS datasets arose from run order effect (MS voltage, sample run order and plate number) and sample preparation bias (sample preparation batch). Nevertheless, a different pattern is observed in batch 2, where a strong association between both PCs and most of the analytical variables, including well and sample position on the well, plate number and sample preparation batch, as well as MS detector voltage and sample run order, was detected. In contrast to batch 1 and 3, analytical variance in batch 2 is explained by both PCs, with the first PC accounting for 68% of total variance, suggesting stronger analytical bias in this batch.

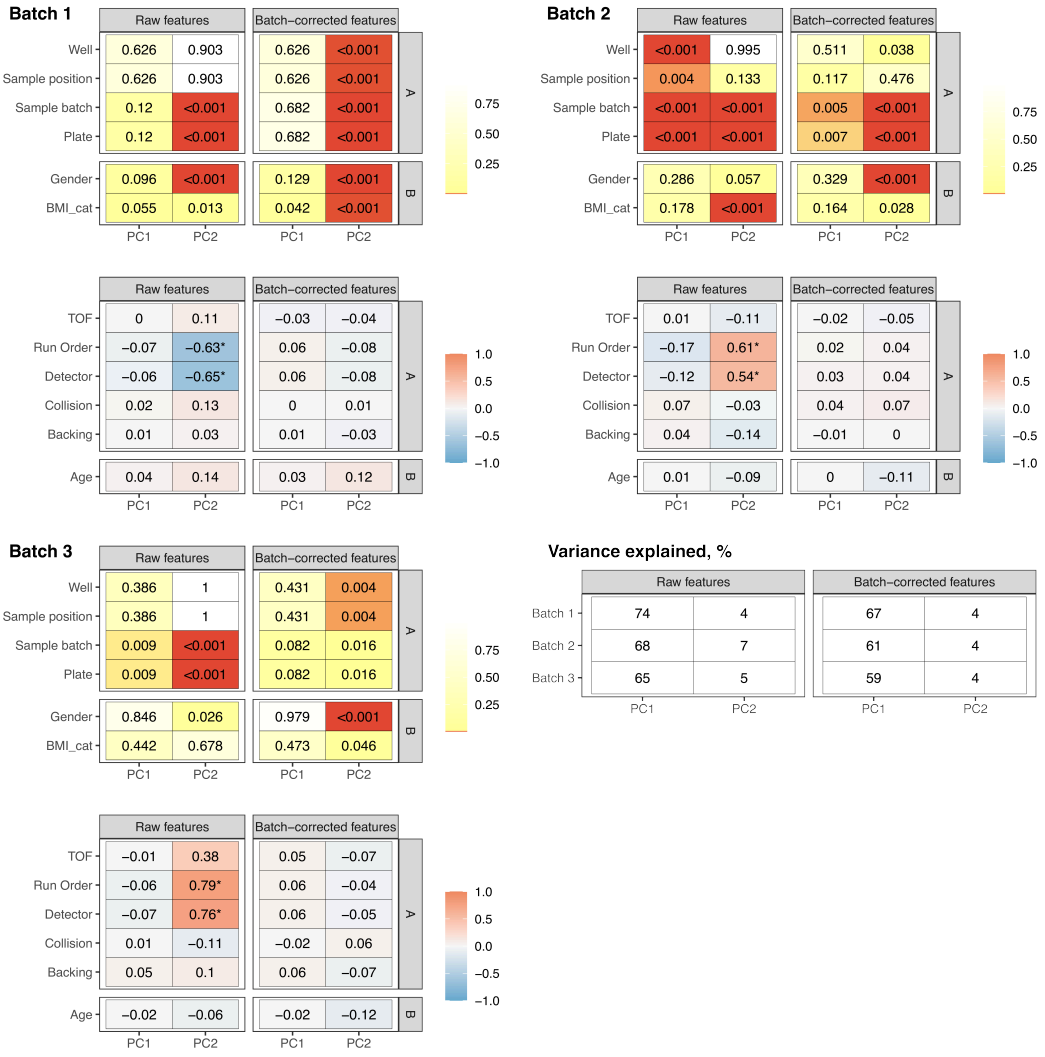


FIGURE 2.16: Principal components (PC) scores association with analytical (A) and biological (B) sources of variation in AIRWAVE data generated by XCMS pre-processing pipeline. Potential associations between the scores of every PC and sample metadata was determined by either Kruskal-Wallis test (categorical data, upper panel for each batch) or Pearson correlation (continuous data, lower panel for each batch). Asterisks indicate strong associations (Kruskal-Wallis p-value < 0.001, correlation > 0.5 or < -0.5). Variable names are provided in Table 2.4. The number of principal components was estimated using 7-fold cross-validation. The variance explained by each PC is listed in the right bottom panel of the figure.

Correction of the run order effect using LOESS smoothing removed association between some of the analytical variables and the PC scores, but correction was not equally successful for all batches, nor for all sources of variance. Among the successfully eliminated sources of variance that were significant before the batch correction are the continuous variables - sample run order and MS (Figure 2.16). Batch correction was not highly effective with regards to the categorical type of analytical variation sources - well and plate number, sample position on the plate and sample preparation batch. Given the observed results, several conclusions can be drawn. First of all, it is important to note the differences between the two tests used to analyse the strength of association. Kruskal-Wallis test is a non-parametric test based on the use of ranks of values, whereas Pearson correlation is a parametric test, which assumes data normality. Therefore, the discrepancies in batch correction performance success for variables analysed with the different tests may stem from the underlying differences between the two statistical methods. Nevertheless, a clear conclusion can be made that each batch suffers from different analytical biases, some of which can be corrected for using batch correction techniques and feature filtering. Nevertheless, a universal QC and batch correction pipeline cannot account for the differences between the batches. This represents a significant issue in large-scale multi-batch metabolic profiling studies, when datasets obtained for individual batches are not directly comparable.

2.4 Conclusions

Within this chapter, the analytical variation observed in a large-scale, multi-batch, untargeted LC-MS metabolic profiling study, the AIRWAVE, was investigated. The results indicate that despite of carefully implemented experimental design and standardised protocols for sample preparation and analysis, analytical variation was observed in the acquired data. The unwanted variation in the chromatographic retention time of the endogenous metabolites and spiked internal standards was identified in the pooled QC samples. Furthermore, clear run-order effect was observed in the detected ion intensity patterns.

Next, a set of open source pre-processing and quality control tools was assessed for their ability to extract information from the study data. The widely-used open source tool XCMS was applied to pre-process data acquired for each of the analytical batch. Investigating the individual steps in the XCMS pipeline, including the IPO-driven parameters optimisation, suggested that these methods were not designed to process data of such scale and complexity. The underlying XCMS algorithms do not take into account the complex retention time drift patterns, which take place during an analytical batch of a thousand samples. The variation in RT between samples is not only compound-specific, but also strongly associated with the injection order. Therefore, RT drifts are difficult, if not impossible, to model and correct for using a small set of peak-groups, which is a method implemented within the XCMS

pipeline. Un-adjusted RT drifts affect peak grouping and potentially introduce a new layer of unwanted variation. Furthermore, applying XCMS pipeline to such a large study is computationally expensive. Some of the analytical variation was successfully removed using post-processing QC measures, such as run-order effect correction based on pooled QC samples, followed by low-quality feature removal. Nevertheless, the analysis of the final datasets indicate that analytical variation is still one of the major sources of differences between samples and further data normalization would be required prior to statistical data modelling.

Chapter 3

Development of a novel LC-MS spectral pre-processing tool

3.1 Introduction

Untargeted LC-MS employed in metabolic profiling of biological samples produces complex data that requires significant pre-processing before samples can be analysed statistically. To enable relative metabolite concentration comparison, distinct two-dimensional features, defined by their *m/z* and retention time (RT), are first identified in each LC-MS spectra of the study. These features then must be matched across all samples in the study. The process of finding corresponding features is sometimes called *correspondence*, while here it will be referred to as *feature alignment*, since features in one sample are aligned to features detected in the next, which simultaneously corrects for RT deviation between the two samples for a given set of features.

A vast number of LC-MS peak alignment methods relying on very different underlying assumptions are available today. In the most up-to-date review focusing on feature alignment, 50 algorithms available as implemented software were described [95]. These can be broadly divided into two main categories: (1) warping and (2) direct matching algorithms. Warping algorithms aim to fit a RT correction function between samples before finding corresponding features, while direct matching algorithms skip RT correction and focus on computing feature similarity instead.

Among the most commonly used LC-MS pre-processing tools, as reported in the Metabolomics Society community survey [146], are XCMS (70% of respondents), mzMine and mzMine2 (26% of respondents together). All three tools use warping-based algorithms for feature alignment. In XCMS, a kernel estimation procedure is used to cluster features with similar *m/z* values and RTs [96]. First, features are divided into overlapping *m/z* bins. Groups of features with similar RTs within the same *m/z* bin are resolved by dynamically estimating the boundaries of RT regions to which corresponding features fall. While mzMine implements a simple alignment method, which assigns each feature to the closest match in the master list [147],

mzMine 2 uses an altogether different algorithm referred to as RANSAC [98]. RANdom SAmple Consensus (RANSAC) algorithm, together with locally-weighted scatterplot smoothing (LOESS) regression, estimate the optimal alignment window for feature matching non-deterministically.

Direct matching algorithms are fewer, but as much varied as warping algorithms. Early solutions, such as RTAlign [148], were simplistic methods that merge features from all samples and aligns those that are within user-defined RT tolerance window. These were followed by more sophisticated solutions that employ feature clustering. DeSouza et al. [149] proposed a two-step hierarchical clustering of features based on RT. First, features are clustered in sets of samples of different experimental groups, obtained clusters are then pooled across all sample groups and clustered again. Similarly to DeSouza et al., MassUntangler performs nearest-distance matching of features using *m/z* and RT dimensions [150]. In contrast to earlier described direct matching algorithms, MassUntangler, performs alignment in a pairwise fashion and also checks for the same charge state.

Even the most controlled LC-MS experiments will experience fluctuations in the chromatographic and mass spectrometric measurements during extended periods of continuous analysis, as described in details in Chapter 2. Some of the LC-MS variation sources generate shifts in *m/z*, RT and instrument sensitivity at the system-level and therefore can be modelled using monotonic functions [95]. For example, column ageing effect is applicable to the whole run and was modelled in numerous studies [91–94]. However, a vast proportion of the unwanted variance in LC-MS data is specific to a given analyte, or a class of related analytes. Such variance complicates features alignment due to ambiguous matching scenarios. Three basic cases of ambiguous feature matching are known (Figure 3.1):

- Peak A in sample 1 can be aligned with either of the two peaks in sample 2. Final alignment will depend on the RT shift model of choice.
- Two peaks in sample 1 can be aligned with two peaks in sample 2 if RT for both peaks is shifted equally. However, peak B in sample 1 can be aligned with peak A in sample 2 with no RT shift as well.

Currently employed feature alignment methods are not capable of resolving ambiguity. Warping-based tools, including the most popular tools, such as XCMS and mzMine 2, cannot correct for analyte-specific shifts since warping functions are fitted to the whole LC-MS spectra.

Most importantly, most alignment methods completely neglect the structural relationship between co-eluting features. In ESI-LC-MS, multiple ions are produced for a given metabolite [74]. These include isotopes, adducts and fragments ions, which co-elute with the main metabolite ion and have similar chromatographic shapes [151]. One of the two tools that do take into account the structural relationship between features is MET-COFEA [152]. In MET-COFEA, features with similar peak shapes,

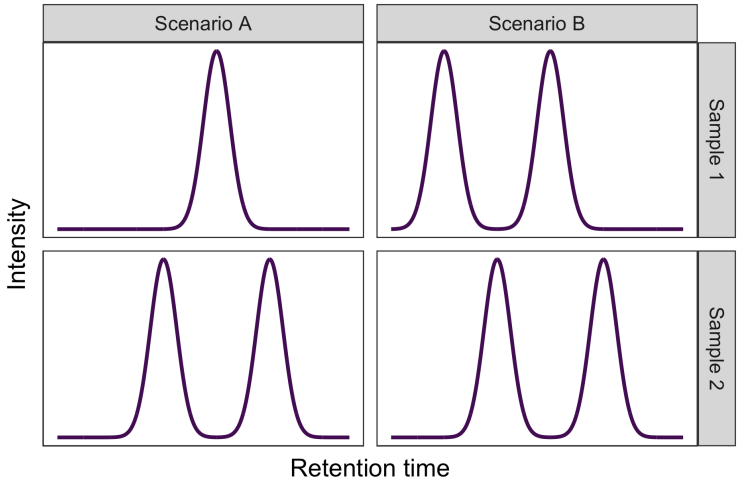


FIGURE 3.1: Three scenarios of ambiguous assignments. (A) A peak in the first sample can match to either of the two peaks in the second sample. (B) Either one or two peaks from the first sample can match to peaks in the second sample.

which is estimated by calculating the dot product of two peak pairs, are grouped together and then annotated using pre-defined adduct formation rules. Annotated groups of features are then aligned between samples by comparing their RT and annotated mass. Similarly, Wandy et al. aligns groups of features by estimating a pair-wise peak similarity score that takes use of their *m/z* and RT values [153]. Both of these tools dismiss spectral intensity information. However, one of the key LC-MS principles is that intensity ratios between the mass fragments from the same metabolite are relatively constant [154]. Features intensities therefore represent invaluable information that should be incorporated into alignment algorithm in order to capture individual metabolites behaviour - *m/z* and RT shifts - in an LC-MS experiment.

Furthermore, none of the currently employed tools, including XCMS, emphasise the importance of feature alignment according to sample injection order. As discussed earlier, clear batch effects arising from analytical variation are known to introduce systematic correlations into the noise. Experimental study design and data acquisition order therefore should be taken into account during data pre-processing.

We must note that any feature alignment algorithm is inherently dependent on the accuracy of the initial peak detection step. As discussed in Chapter 1, even the highly cited pre-processing tools, including XCMS, are prone to reporting false peaks or missing peaks in some of the samples. The issue has been widely discussed in the field [108, 109, 155], however, it is beyond the scope of this thesis, which focuses on the methods for accurate feature alignment that accommodate the analytical variation typically observed in large-scale metabolic profiling studies. Nevertheless, to address the issue of missing peaks, a method for raw data re-integration that is robust to run order effects is investigated within this Chapter.

3.1.1 Most commonly used terms

- *Structurally related features* refer to *centWave* detected features with highly correlated peak shapes. The underlying assumption is that the ions corresponding to such features originate from the same chemical compound at the electrospray ionisation source and thus co-elute with highly similar peak shapes.
- *Pseudo chemical spectra* refers to a group of structurally related features originating from the same chemical compound. A single pseudo chemical spectra is essentially a list of features, which can be represented in a two-dimensional space using their retention time, *m/z* and/or intensity (Figure 3.2).

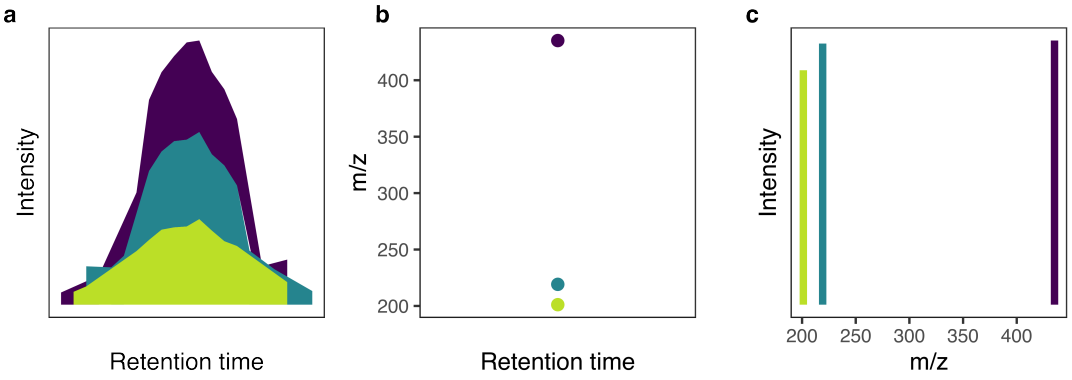


FIGURE 3.2: massFlowR functionality is based on the use of pseudo chemical spectra. Pseudo chemical spectra (PCS) is comprised of (a) co-eluting chromatographic peaks with highly correlated peak shapes. (b) Each PCS can be visualised in a two-dimensional space using the retention time and *m/z* of the corresponding *centWave* detected features. (c) A fingerprint for each PCS is made by taking the *m/z* and intensity values for the corresponding *centWave* features.

3.1.2 Hypothesis

The key hypothesis of this chapter is that groups of structurally related co-eluting features are preserved across samples. Information on how features are related in an individual LC-MS sample therefore will help to correctly align features across samples.

The second hypothesis is that incorporation of sample acquisition order information together with feature grouping information will help to capture metabolite-specific shifts in RT and *m/z*.

3.1.3 Aims and objectives

To address the hypotheses presented above, the purpose of this chapter was to develop and implement an LC-MS data pre-processing pipeline that:

- Groups structurally-related features in each LC-MS sample.

- Aligns features across samples in their acquisition order incorporating features grouping information.
- Re-integrates raw data for features not picked by the peak-picker by adjusting integration regions for each sample.

The aim of the second part of this chapter was to:

- Evaluate the accuracy of the developed feature alignment algorithm in comparison to the gold standard open source tool.
- Demonstrate the application of the pipeline to an open source dataset.

3.2 Methods

3.2.1 Pre-processing pipeline

Overview

A three-stage LC-MS pre-processing pipeline was developed and implemented as an R package massFlowR, source code for which is available on GitHub repository: `https://github.com/lauzikaite/massFlowR`. A high-level overview of the functionality of the pipeline is provided in Figure 3.3.

Pseudo chemical spectra generation

In the first stage of the massFlowR pipeline, each individual raw LC-MS file in the study is processed independently. Chromatographic peaks detected by the *centWave* algorithm are grouped together with structurally related features that originate from the same chemical compound: adducts and isotopes. Feature grouping is based on the expectation that features resulting from in-source transformations of the same molecule exhibit an identical chromatographic retention pattern. Therefore, to identify related features, peak shape similarity analysis is employed (Figure 3.2a).

The selected similarity function here is Pearson correlation between the extracted ion chromatogram (EIC) of two co-eluting features. Pearson correlation was selected for the task as it is a measure of the strength of a linear association between two variables. As it has been demonstrated that adducts and fragment ions of the same chemical compound have the same intensity ratio in each scan of the LC-MS experiment [156], the underlying Pearson correlation assumption of normal distribution is met. Figure 3.4 illustrates that intensities of extracted ion chromatograms, generated for *centWave* detected features, follows a normal distribution.

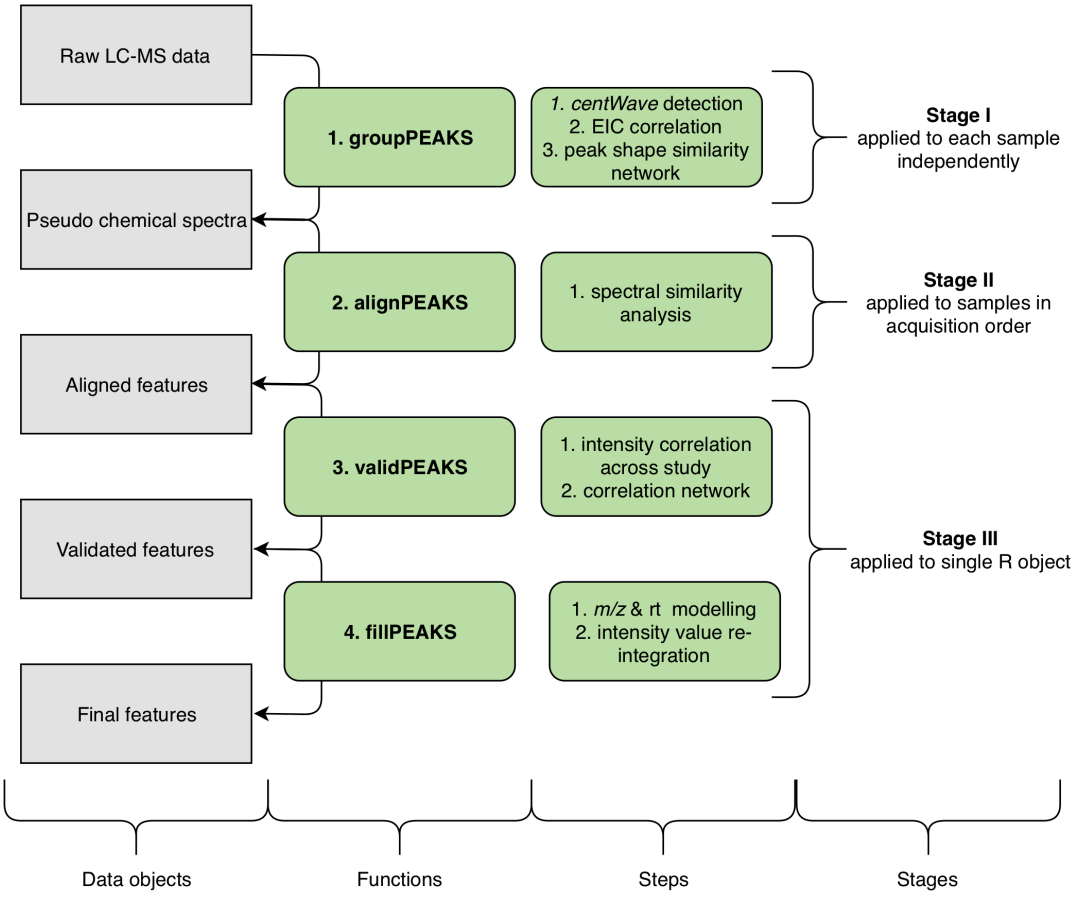


FIGURE 3.3: The massFlowR pipeline comprises of four main functions that are applied at three stages: (1) Chromatographic feature detection and EIC correlation to generate pseudo chemical spectra in each LC-MS sample. (2) Feature alignment across samples is performed in original data acquisition order. (3) Aligned features intensity correlation is applied across all samples to identify features that truly belong to the same PCS. (4) Missing data points integration using raw LC-MS files.

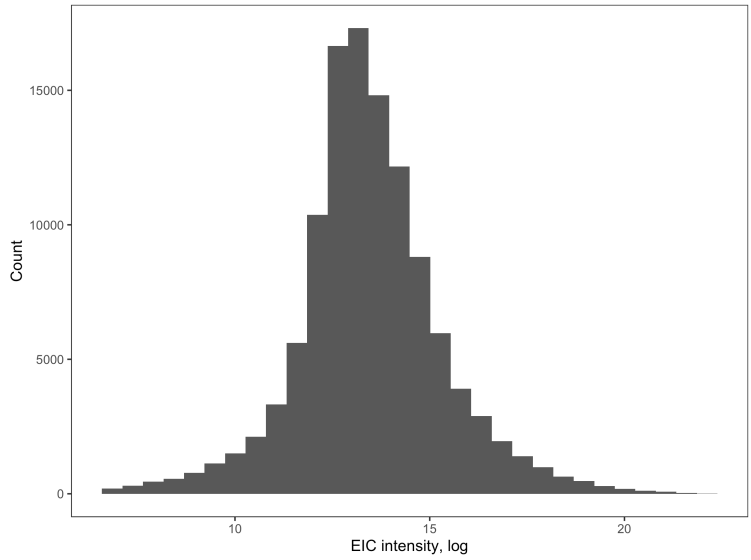


FIGURE 3.4: The distribution of the intensities of extracted ion chromatograms, generated for all *centWave* detected features in a representative DEVSET QC sample.

The obtained Pearson correlation coefficients are then used to construct a weighted undirected network in which nodes correspond to co-eluting features and weights of the edges to the pair-wise EIC correlation coefficients. The resulting network is not fully connected as only the edges with weights above a user-selected threshold are retained. The default threshold value was set to 0.95 through optimisation with experimentally validated metabolite annotations in quality control samples.

To identify groups of features with a similar peak shape, constructed networks are subjected to label propagation algorithm using the IGRAPH package [157]. The algorithm implemented within the IGRAPH assigns a unique label to each node in a given network. At every subsequent iteration, each node adopts a label that a maximum number of its neighbours have. In such a manner, labels propagate through the network and densely connected groups of nodes, called *communities*, with the same label are formed. Detected communities comprised of more than one feature in massFlowR are denoted as pseudo chemical spectra (PCS). Such PCS, which are essentially lists of structurally related features, are generated for each LC-MS sample. A schematic representation of a single PCS is provided in Figure 3.2.

The complexity of generated PCS is demonstrated in Figure 3.5, where the number of *centWave* detected features grouped into the same PCS according to this procedure is plotted across all DEVSET samples.

Pseudo spectra generation step for a given sample is summarised in Figure 3.6.

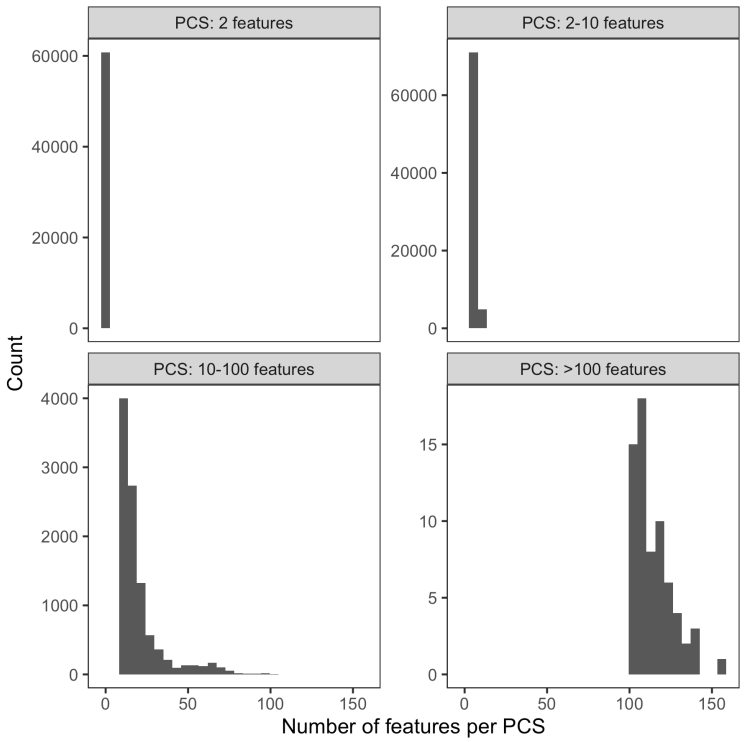


FIGURE 3.5: The number of *centWave* detected features per pseudo checmial spectra (PCS) across all DEVSET samples. Distribution of PCS size is visualised over four sub-figures tp account for very different y-axis scales.

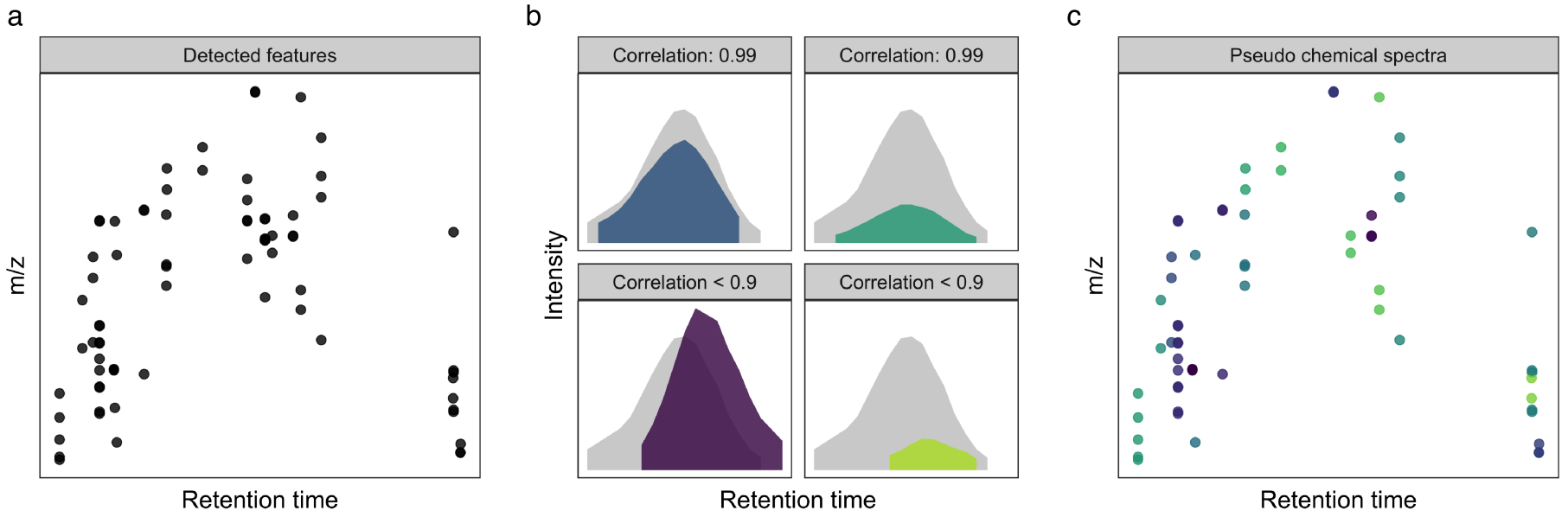


FIGURE 3.6: In the first stage of the pipeline, each individual raw LC-MS sample is processed independently in three steps: (a) chromatographic peak detection facilitated by the *centWave* algorithm is followed by (b) EIC correlation estimation between all co-eluting chromatographic peaks, eluting at +/- 1 scan as the feature-of-interest. Correlation coefficients are then used to build networks of co-eluting features, identifying groups of features that are correlated to the feature-of-interest above the user-defined threshold. Such EIC correlation-based network analysis is performed for each feature. (c) Only groups comprising more than one feature are retained for further pre-processsing steps.

Feature alignment across samples

During the second stage of the massFlowR pipeline, corresponding features are matched across samples of the study using pseudo chemical spectra built for each sample independently. Matches for features of a PCS in the sample-of-interest are found in the template (list of features from all earlier samples) through *m/z* and RT search (Figure 3.8a). The quality of the match between the target PCS and all matching PCS in the template is evaluated using dot product function (Figure 3.7). Dot product, also know as cosine correlation, is a measure of correlation between two sequences of intensities and is widely used in spectral library search [158] and data alignment [159] algorithms. Dot product function is defined as:

$$cos\theta = \frac{t \times m}{\|t\| \cdot \|m\|} \tag{1}$$

where $t \times m = \sum_i^n s_i m_i$ and $\|t\| = \sqrt{\sum_i^n t_i^2}$. t and m are spectral vectors for target and match PCS respectively.

PCS is a list of features which can be written as $\{f_1, f_2, ...\}$, where *m/z*(f_i) and *intensity*(f_i) are the *m/z* and intensity values for feature f_i. Spectral vectors are obtained by placing scaled intensity values into equally-spaced *m/z* bins and normalising these bins intensities to total magnitude of the vector [160].

Dot product is the measure of agreement between two spectral vectors t, PCS comprised of target features in the sample-of-interest, and m, PCS containing matching features in the template. Target-match pair with the highest dot product value is considered the most similar and their features are merged. Features that are not matched directly by *m/z* and RT are added to the template as part of the same PCS. Templates *m/z* and RT for the matching features are updated to the average between the current sample and the template. This assures that template stores the moving averages of *m/z* and RT values.

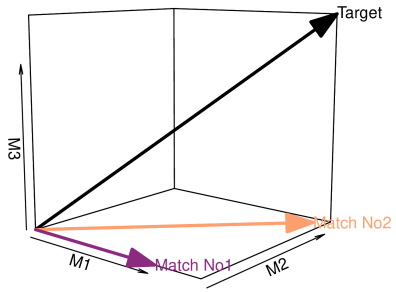


FIGURE 3.7: Vector representation of a hypothetical three-peak pseudo chemical spectrum (target) and two potential matches in the three-dimensional space corresponding to the target's *m/z* peaks. Dot product function between the target and each of the matches distinguishes best-matching pseudo chemical spectra in the template.

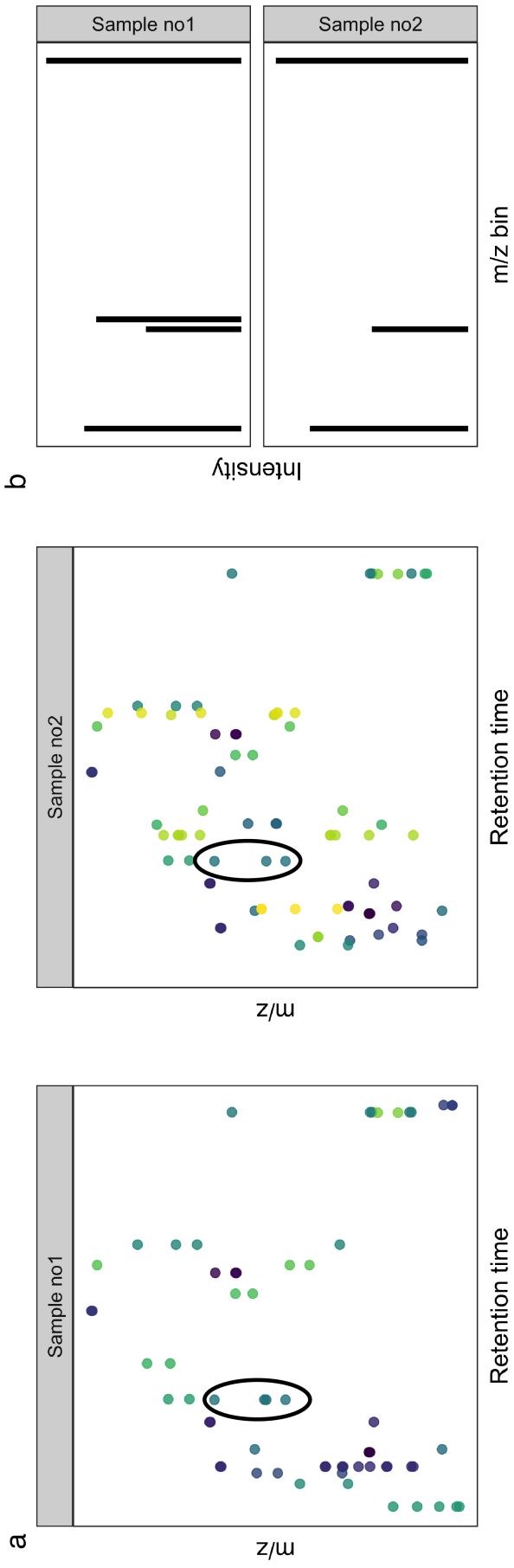


FIGURE 3.8: During the second stage of the massFlowR pipeline, features are aligned across all samples of the study in data acquisition order. (a) Features of a given pseudo chemical spectra in sample no2 are matched against all features in previous samples (in this case template comprises of features from sample no1 only) using *m/z* and RT window. (b) The level of agreement between a given PCS in sample no2 and matching PCS in sample no1 is estimated using dot product function, which is a measure of spectral similarity. PCS from sample no2 is then aligned with the most similar PCS from sample no1.

Feature alignment validation

Once features are aligned across all samples, the obtained PCS are validated. Intensity values for each feature in a group are correlated across all samples. Obtained correlation estimates are used to build a similarity network for each PCS (Figure 3.9). Similarly to the feature grouping method described in 3.2.1, label propagation algorithm identifies groups, or, communities, of features that exhibit similar intensity pattern across a study. These communities represent the validated PSC.

Filling in missing data

The final step in the pipeline is to re-integrate raw LC-MS files to fill in intensity values for each of the missing features in validated PCS. In contrast to XCMS, *m/z* and RT values for integration are estimated for each sample separately (Figure 3.10). *m/z* and RT values for each feature are modelled and interpolated using cubic smoothing spline. While local regression smoothing can also be applied for non-linear data modelling, cubic smoothing was found to be less sensitive to parameter fluctuation in LC-MS data modelling in [128].

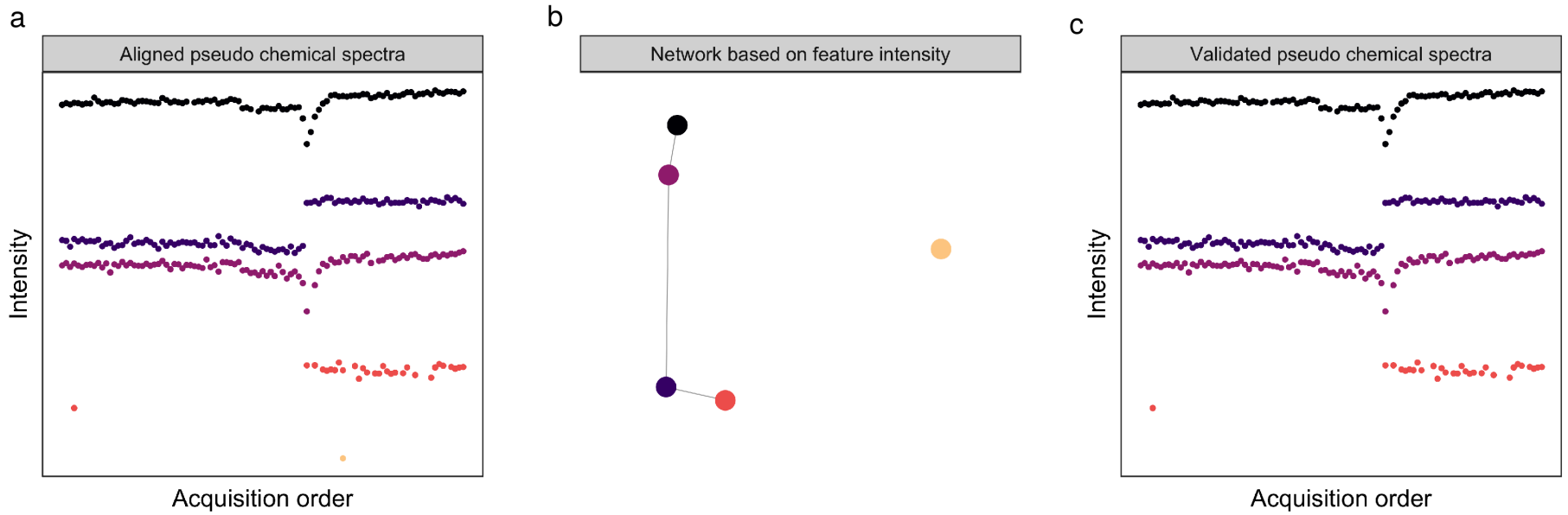


FIGURE 3.9: In the last stage of the pipeline, aligned features are validated. (a) Intensity values of features aligned to the same PCS (as indicated by different colors) are correlated pair-wise using values from samples in which features were detected. (b) Obtained correlation values are used to build a similarity network, where each node represents a feature and edges are Pearson correlation coefficients. Only edges with coefficients above a user-defined threshold are retained. Label propagation algorithm is applied to identify feature assignment to communities. (c) Only communities with more than one feature are retained.

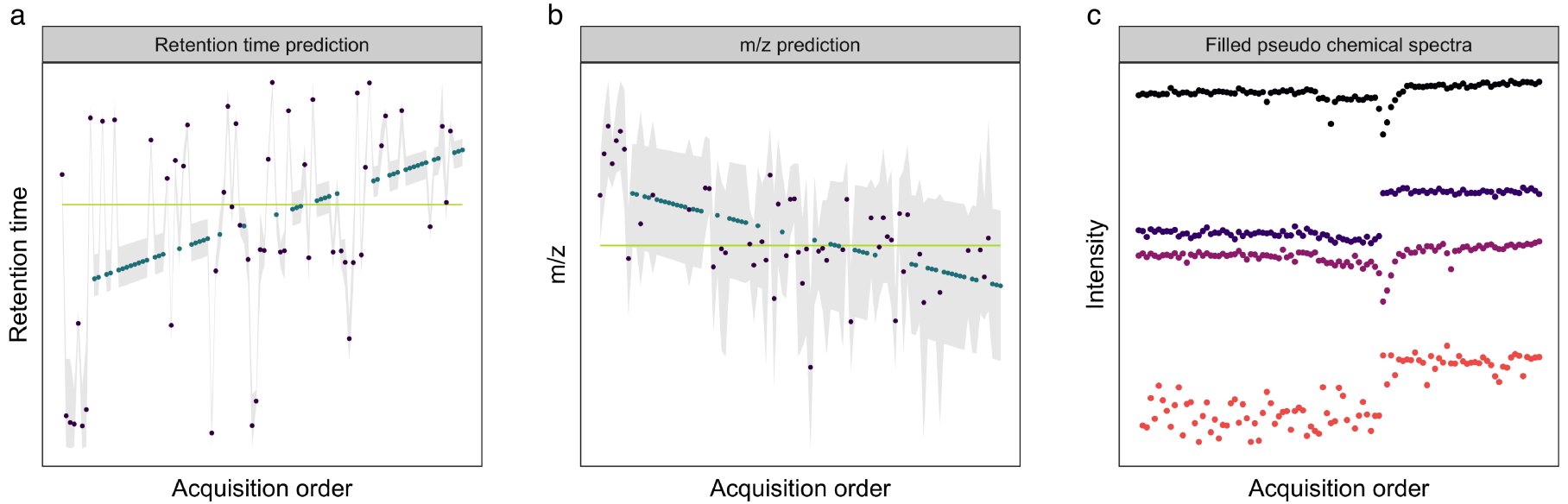


FIGURE 3.10: In the last step of the pipeline, raw LC-MS spectra are re-integrated for features which were missed in samples during initial peak picking. (a) RT and (b) *m/z* values for spectral integration are estimated for each sample separately using cubic smoothing spline intrapolation. Purple dots indicate RT and *m/z* values in samples in which feature was detected; Blue dots indicate values modelled by the spline function; Green line represents the median across samples in which feature was detected (this value would be used for spectral integration in XCMS).

3.2.2 Analytical data acquisition

An open source metabolic phenotyping study DEVSET was chosen for method development and validation. The study comprises of three unique samples of human urine, which were mixed in known proportions according to {3,2} simplex lattice experimental design [161] such that six unique samples of pooled urine were prepared:

- DevSet1
- DevSet2
- DevSet3
- DevSet1v2 (50:50 mix of DevSet1 and DevSet2)
- DevSet1v3 (50:50 mix of DevSet1 and DevSet3)
- DevSet2v3 (50:50 mix of DevSet2 and DevSet3)

These samples were then split into 13 equivalent aliquots. Together with the pooled quality control(QC) samples and independent external long-term reference samples (which is a QC sample composed of urine specimens that are completely independent of the study and are routinely used across multiple studies at the NPC), the study comprised of 201 samples in total.

Samples were previously prepared and analysed by the ESI-LC-MS according to the standartised National Phenome Centre (NPC) protocols [62, 133]. Briefly, the ACQUITY UPLC (Waters Corp., Milford, MA, USA) chromatography system was connected to Xevo G2-S Q-TOF mass spectrometer (Waters Corp., Manchester, UK) with Zspray electrospray ionization (ESI) source. The mobile phases employed were: (A) water, (B) acetonitrile, each supplemented with 0.1% formic acid. The gradient can be summarised as follows: 0 min - 0.1 min isocratic separation at initial conditions (99% A); 0.1 min - 10.0 min a linear gradient elution (99% A to 45% A); 10.0 min - 10.7 min a rapid gradient elution (45% A to 0% A); followed by fast column washing with 99% A until 15.0 min.

Raw LC-MS data is available for download on the MetaboLights server (study identifier MTBLS694). Here, Waters .RAW files were converted to open-source format mzML using ProteoWizard software as specified in the section 2.2.2 in Chapter 2.

3.2.3 Synthetic data generation

To benchmark and evaluate algorithm performance, its output must be compared against expected results. Such comparison to the known true answer enables quantification of algorithm performance, and thus, direct comparison with other algorithms [57].

To assess the performance of the developed feature alignment algorithm, synthetic datasets were generated by systematically introducing noise into the LC-MS features

table obtained for a real sample. The original features table was obtained by peak-picking a representative quality control sample from the DEVSET study using *centWave* algorithm. Features were assigned into pseudo chemical spectra (PCS) using method described in Section 3.2.1. The resulting PCS table was then used to generate features tables, representing different types of variance commonly observed in LC-MS experiments.

Intensity values were drawn from a multivariate normal distribution such that (A) features in the same PCS are highly correlated across all generated tables, and (B) log-transformed values are normally distributed. The arithmetic means were set to the values in the original feature table. The covariance matrix was simulated such that features in the same PCS have a correlation value of > 0.8. Correlation matrix obtained with simulated intensity values of five randomly selected PCS is visualised in Figure B.16. High correlation is observed only between features that belong to the same PCS.

Three experiments were designed to simulate different types of noise in the *m/z* and RT domains (Table 3.1). Every further experiment includes the noise introduced previously and adds a new type of variance. In each experiment, 100 data tables were simulated. Each experiment was replicated three times. Overview of generated noise is visualised for all of the features of a single PCS in Figure 3.11.

In Experiment A, random noise was introduced into both the *m/z* and RT values (normal distribution with a standard deviation of 0.001 and 2 seconds respectively).

In Experiment B, systematic RT drift was applied to all features. Run-order dependent RT drift was first modelled on real-word QC samples using 15 endogenous compounds by fitting a cubic smoothing spline (Figure 3.12). Obtained model parameters were then applied to generate a non-linear drift to all features in the synthetic data by allocating a particular spline model to all features of a single PSC at random.

In Experiment C, the effect of missing values on the performance of the alignment tools was evaluated. As discussed in Di Guida et al., missingness can arise from either: a non-random biological response, or from variation in the observed variables, or finally, can be completely random [162]. While the first cause represents the purely biological differences between samples and therefore can provide meaningful information in biological data interpretation, the latter two causes of missing values are undesired and can hide the subtle yet real differences between samples. Such undesired value missingness can arise because of various reasons, including but not limited to (1) metabolite concentration falling below the instrument's limit of detection; (2) sample matrix effects, such as ion suppression; (3) run-order dependent changes in the LC-MS system; and (4) pre-processing software failure to detect and align spectral peaks [163]. To simulate a typical metabolomics dataset where a combination of different sources of technical variation are present, features

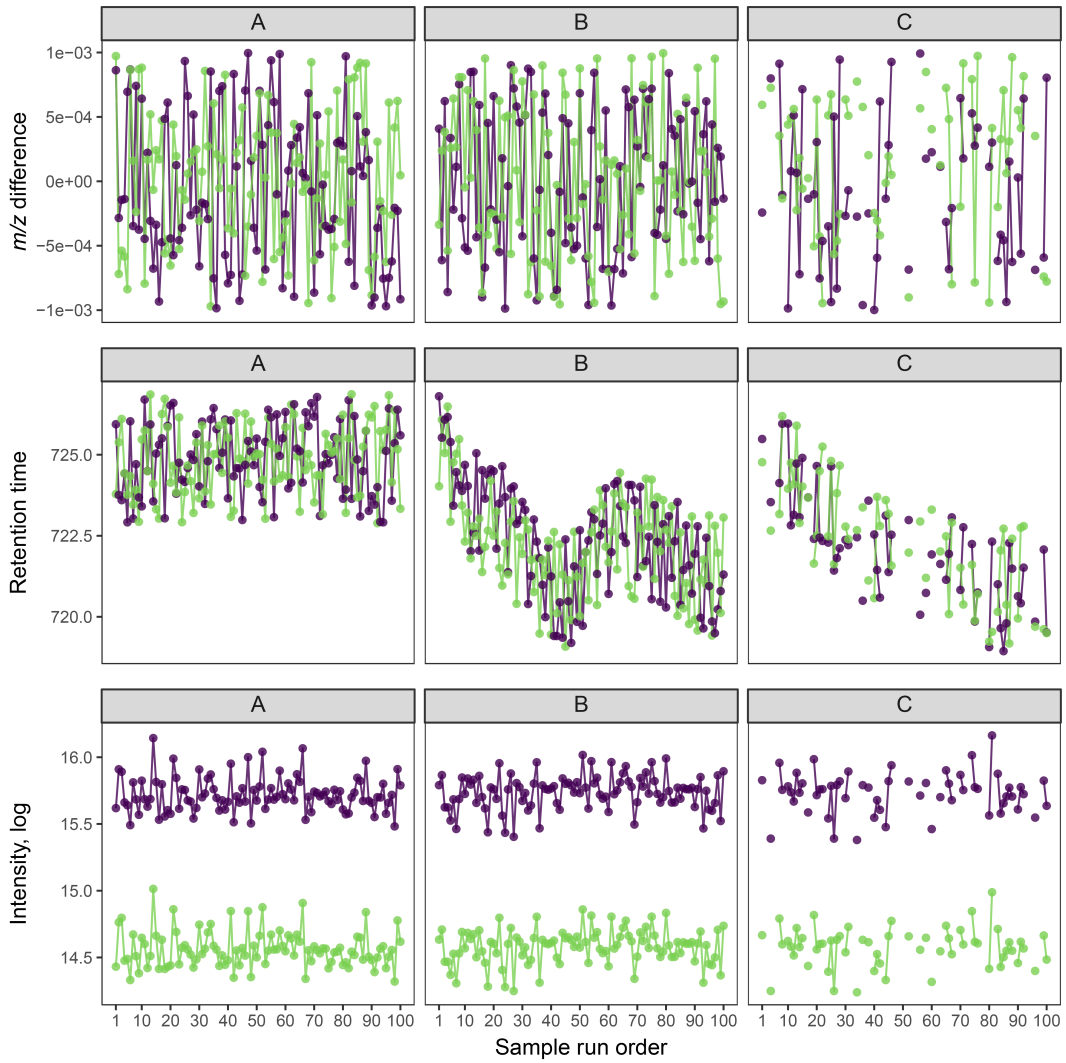


FIGURE 3.11: Simulated datasets representing different types of noise were used in feature alignment performance assessment. Introduced *m/z*, RT and intensity variance is visualised for all features (as indicated by different colours) of a single PCS. In experiment A, only random noise was added to *m/z* and RT values. In experiment B, RT follows a non-linear drift across the 100 synthetic samples. In experiment C, features are removed from samples using a probabilistic model. Intensity values were simulated such that features of a single PCS are highly correlated across all samples in a single experiment.

TABLE 3.1: Three experiments, representing commonly observed LC-MS experimental noise, were performed to evaluate feature alignment algorithm performance.

	Introduced noise	Target missingness, %	Total missingness, % (average)
Experiment A	Random *m/z* and *rt*	-	-
Experiment B	Systematic *rt* drift	-	-
Experiment C	Missing features	5, 10, 20, 30, 40	9, 18, 34, 49, 61

Random *m/z* variance (Gaussian noise with a standard deviation of 0.001 *m/z*) was generated for every feature using original *m/z* values.
Random *rt* variance (Gaussian noise with a standard deviation of 2 seconds) was generated for every feature, using either original *rt* values (experiment A), or systematically altered *rt* values (experiments B, C).
Systematic *rt* drift was generated using non-linear splines obtained on reference compounds in quality control samples.
Features were removed at random, where the probability of its missingness inversely depends on its intensity value. If the most intense feature of the PSC, or one of the two features of the PSC was removed, the remaining features of the PSC were removed from the sample as well.

were removed from samples at random with a probability of missingness inversely dependent on intensity. Such correlation is based on the assumption that missing values occur due to metabolite concentrations falling below the limit of detection. The probabilistic model for missingness was based on logistic function formulated by Do et al. [163] and represented as $P(x_i missing) = logistic(\beta_0 + \beta_1 \times x_i)$ with logistic function $logistic(a) = \frac{\exp(a)}{1+\exp(a)}$. Coefficient β_1 was set to -10 and intercept β_0 was found by numerically solving the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} P(x_i missing) = miss \tag{2}$$

where *miss* is the desired proportion of missing features in the table, n is the number of features in the table. If the most intense feature or one of the two-features PSC was selected to be removed, the remaining features of the PSC were omitted from the sample as well, as summarised in Figure 3.13.

3.2.4 Feature alignment algorithm comparison

Before describing the experimental design that was used for algorithm performance assessment, some of the definitions that are used throughout this section must be introduced. As before, *feature* is a two-dimensional LC-MS signal. For each feature, a given alignment algorithm finds a suitable match in each sample (if any) and groups them together. These groups here are referred to as *consensus features* since each individual feature in it should correspond to the same ion of the same chemical entity. All consensus features together are referred to as *consensus map*, which stores the alignment information of all detected features in all LC-MS feature maps.

In theory, each feature should be allocated to one consensus feature and each consensus feature should include one feature per map. Such optimal consensus map is

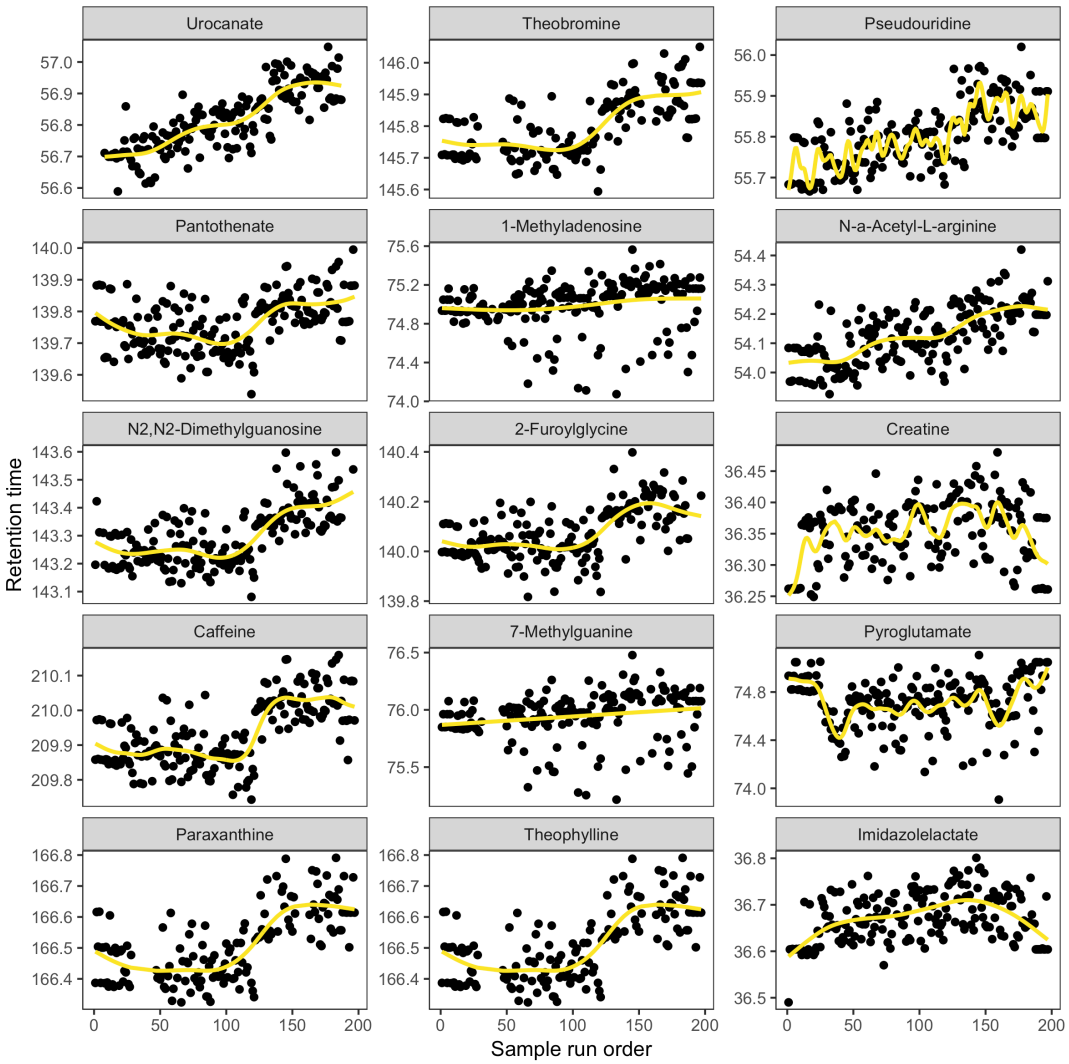


FIGURE 3.12: Cubic smoothing splines fitted to the retention time of the features corresponding to 15 endogenous chemical compounds in urine quality control samples in DEVSET study. The obtained splines were used to model RT drift in synthetic data.

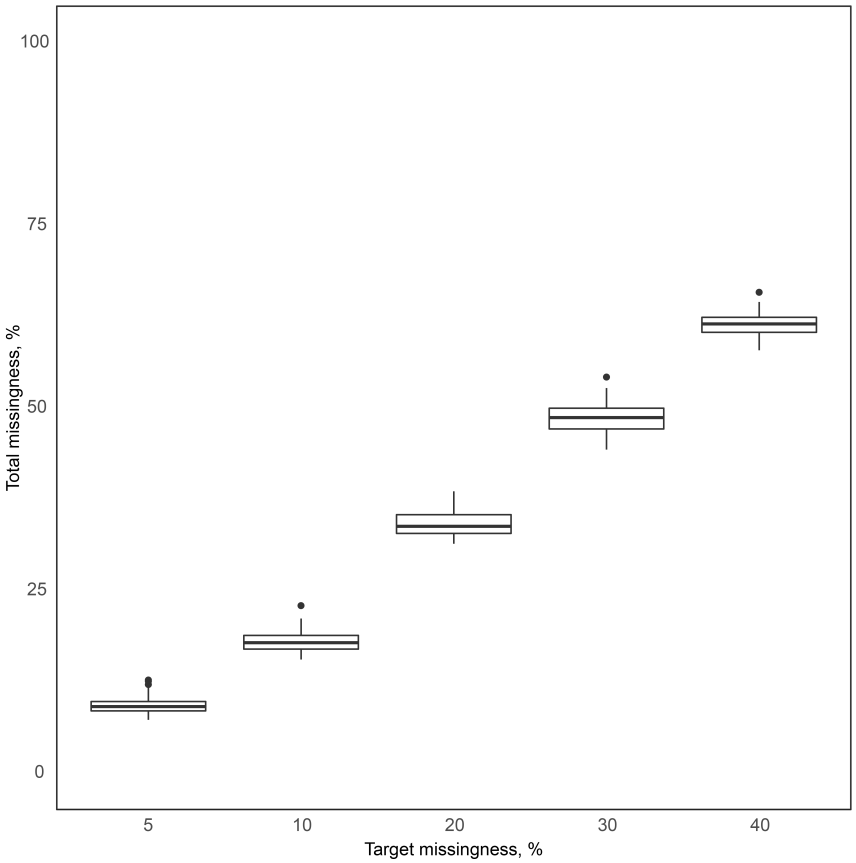


FIGURE 3.13: The proportion of missing features in the 100 synthetic datasets of the three replicates of experiment C. 100 datasets were generated with a varying proportion of feature missingness. Probabilistic removal mechanism selected features at random at desired target proportion (5 % to 40%). If the most intense feature or one of the two-features PSC was selected, the remaining features of the PSC were removed from the sample as well. Thus, feature removal resulted in higher total missingness (9% to 61%).

called *ground truth*. In practice, however, fluctuations in the chromatographic and mass spectrometric measurements occurring during the experimental run time will, as well as biological sample-to-sample variation lead to consensus features that do not appear in all LC-MS maps. As a result, alignment tool splits related features across multiple groups and/or assigns unrelated features to the same consensus feature.

Alignment algorithm performance here is evaluated by calculating precision and recall, which are defined as $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$ respectively. Precision and recall adaptation for the assessment of feature alignment algorithms was provided by Lange et al. [57]. The mathematical representation is as follows:

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{|g_i \cap t_i|}{|t_i|} \tag{3}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{|g_i \cap t_i|}{|M_i| \times |g_i|} \tag{4}$$

where g is consensus features in the ground truth of length N, t is consensus features obtained by the alignment tool of length M. For each consensus feature in the ground truth, g_i, a set of corresponding consensus features from the tool is denoted as t_i. The set of consensus features from the tool that contain at least two features and intersect with ground truth feature g_i is denoted as M_i. The length of M_i corresponds to the number of sub-groups into which a consensus feature was assigned by the tool. Thus, the more times it was split, the lower the recall value is.

Generated synthetic datasets were subjected to the developed feature alignment algorithm, as well as the "density" method from XCMS package. Applied XCMS and massFlowR parameters are summarised in Table 3.2.

3.2.5 Quality control assessment

DEVSET datasets obtained by massFlowR and XCMS pre-processing - detected, aligned and filled features - were subjected to further post-processing steps according to standardised quality control (QC) procedures for metabolic profiling, described in details in Chapter 2, Section 2.2.4.

All scripts used within this and other chapters are available on the public GitHub repository: `https://github.com/lauzikaite/PhD_thesis_code`. The source code for the developed massFlowR package is available on another GithHub repository: `https://github.com/lauzikaite/massFlowR`.

TABLE 3.2: XCMS and massFlowR parameters used in the pre-processing of DEVSET study and synthetic data.

XCMS		massFlowR	
Parameter	Value	Parameter	Value
centWave		***groupPEAKS***	
peakwidth	c(1, 5)	peakwidth	c(1, 5)
prefilter	c(10, 5000)	prefilter	c(10, 5000)
noise	200	noise	200
snthresh	5	snthresh	5
ppm	25	ppm	25
density		***alignPEAKS***	
minfrac	0	rt_err	10
minsamp	0	mz_err	0.01
bw	2	cutoff	0.3
mzwid	0.01	***validPEAKS***	
		cor_thr	0.7

Unlisted parameters were set to defaults.

3.3 Results and discussion

3.3.1 Pipeline development

Pseudo chemical spectra generation

The proposed pre-processing strategy builds on and combines previously developed methods and ideas in MS analysis. Pseudo chemical spectra (PCS) generation is based on chromatographic peak shape similarity analysis, which is used in metabolite identification [164–166] and mass spectrometry data reduction algorithms [151]. The assumption of peak shape similarity analysis is that adduct and fragment ions originating from the same compound will have the same intensity ratio in every LC-MS scan [156]. In theory, their EICs are linearly dependent, and as a result, their correlation can be used as an indicator of peaks origin.

To evaluate how well EIC correlation distinguishes similar peaks among co-eluting peaks, a simulation was first performed with peaks of identical shape. A characteristic chromatographic peak, missing a few scans and fitting a Gaussian curve well, was identified in the raw data of a representative urine sample (Figure 3.14A). Its self-EIC correlation was observed as the apexes moved further apart scan by scan. Correlation between two identical chromatographic peaks quickly dropped from 0.934 for peaks within one scan distance to 0.755 for peaks within two scans distance (Figure 3.14B). These results indicate that peak shape similarity analysis is inherently sensitive to peak alignment and thus co-elution. In light of the results of this simulation, a conclusion was drawn that EIC correlation analysis should only be applied to chromatographic peaks eluting at one scan distance at most. Such restriction greatly reduces search space and thus computational time.

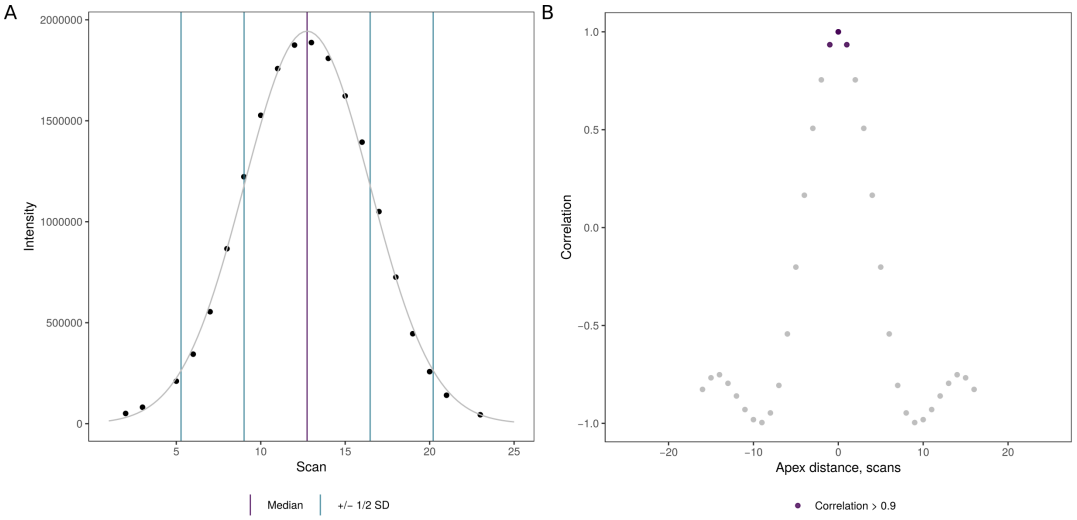


FIGURE 3.14: EIC correlation of a representative chromatographic peak with itself was observed. (A) peak with five missing scans and a Gaussian-like shape was selected. Gaussian fitting (grey line) was performed using median and standard deviation values derived from the *centWave* output. (B) Self-correlation decreases as apexes of two identical peaks move further away.

To evaluate how successfully EIC correlation identifies similar peaks in a complex spectrum, an experiment with DEVSET dataset was performed. Endogenous metabolites that are detectable in urine samples using standard LC-MS assays were investigated. 15 metabolites and their main adducts and in-source fragments were identified in the LC-MS spectra of all DEVSET samples using *m/z* and RT regions kindly provided by the NPC team. Detection of metabolites in spectra was performed using R package peakPantheR, available at `https://github.com/phenomecentre/peakPantheR`. Detection regions of the validated ions, as well as summary plots are available in Appendix B.

EIC correlation between the features corresponding to the main ion and all of its daughter ions (adducts and in-source fragments) of each of the 15 metabolites was performed in every DEVSET sample. EIC correlation distribution indicates that structurally related chromatographic peaks exhibit high peak shape similarity (Figure 3.15). Therefore, EIC correlation coefficient cut-off of 0.95 is used during pseudo chemical spectra generation in the massFlowR pipeline (Figure 3.6). Only features with EIC correlation of > 0.95 are considered as part of the same pseudo chemical spectra.

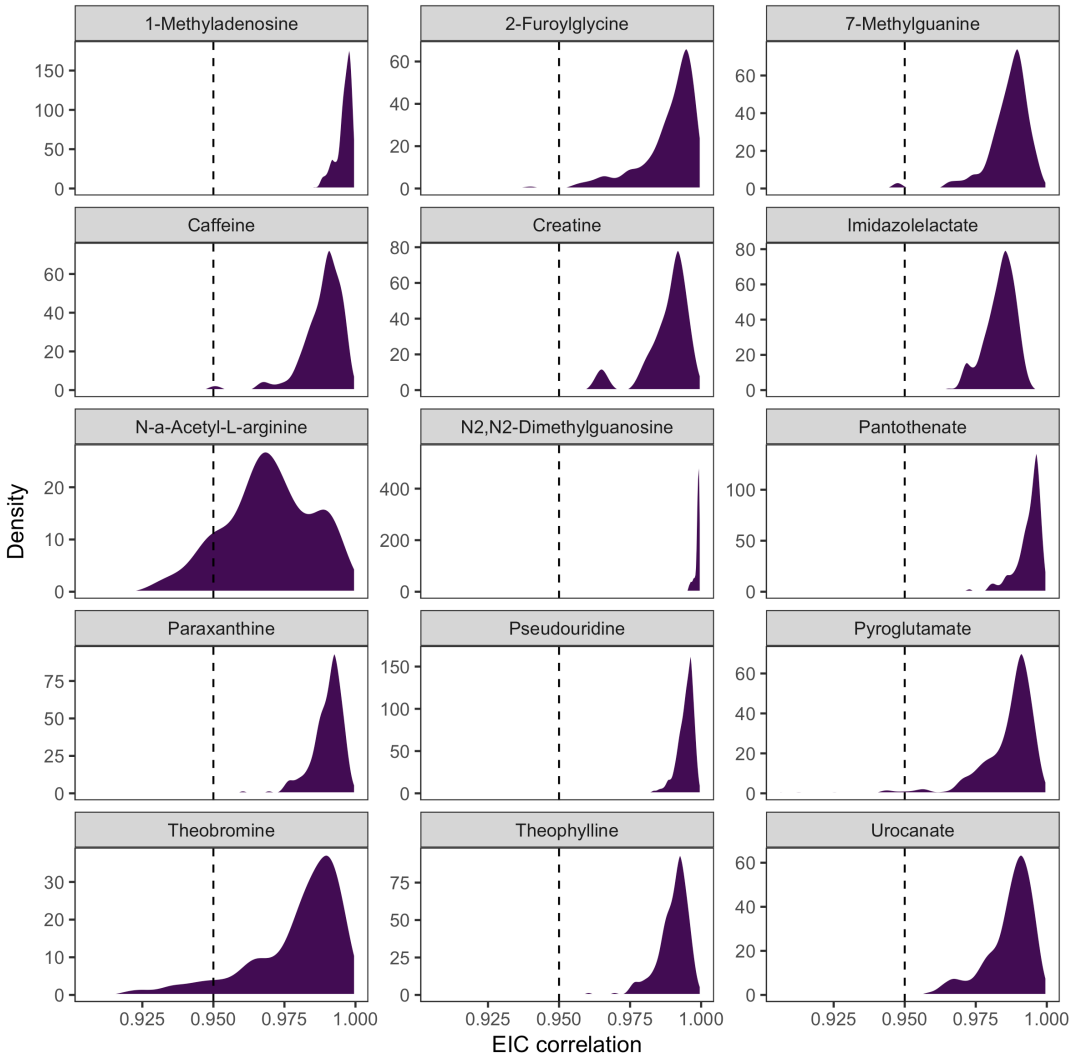


FIGURE 3.15: EIC correlation was performed between features corresponding to the main adduct and its adducts/in-source fragments of 15 validated metabolites. Distribution of correlation coefficients obtained in all DEVSET samples is shown for each metabolite. Most of the adduct pairs exhibit correlation above 0.95, which was therefore selected as the default threshold value pseudo chemical spectra generation in massFlowR.

Feature alignment across samples

Pseudo chemical spectra is employed in the alignment of features across samples. For each feature in the sample-of-interest, matches are found in the next sample through *m/z* and RT search. All PCS with matching features are compared with the PCS comprising the target feature in the sample-of-interest. PCS with the highest spectral similarity is the most appropriate match for the target feature. PCS spectral similarity comparison is based on widely used MS/MS library search methods. First, target and match PCS spectra are pre-processed: (1) raw intensity values for each ion are scaled, (2) scaled intensities are placed into 0.001 *m/z* wide bins to generate a spectral vector, (3) vector values are normalised. Next, dot product function is applied to the target and match spectral vectors. PCS with the highest dot product value is considered the most similar to the target PCS.

As MS/MS fragmentation pattern plays a crucial role in metabolite identification, methods for accurate spectra matching to MS/MS libraries have been investigated in a multifold of studies. While each MS/MS database employs a slightly different method to evaluate the quality of the spectral match between experimental and database fragments [111], certain aspects of spectral matching algorithms are widely accepted. For example, heavier ions are considered to be more important and informative in MS/MS spectra. Thus, spectra is often scaled by giving more weight to heavier ions [158, 167], as in:

$$W = into_i^m \times m/z_i^n \tag{5}$$

where W is the weighted-intensity vector, m and n represent the weight factors of peak intensity and *m/z* value, 0.6 and 3 respectively, as optimised by [158].

It is important to emphasise that pseudo chemical spectra is composed of ESI-LC-MS features and therefore exhibit different properties from MS/MS spectra. In contrast to MS/MS experiments, ESI-LC-MS principally produce adduct ions of intact analyte molecule with little in-source fragmentation. While ion pattern is inherently dependent on the LC assay protocol and solvents composition, a metabolite is likely to produce the same ions in all samples analysed in one LC-MS experiment. The emphasis therefore should be placed on the most intense ions in the spectra, which represent the most consistently produced form of ion for the particular metabolite.

To evaluate which spectra pre-processing is most suitable for PCS spectral similarity analysis, an experiment was performed using DEVSET study. Earlier described 15 metabolites - their adducts and in-source fragments - were identified among features detected in DEVSET samples. Spectral similarity between PCS that comprise of corresponding features was analysed using three intensity scaling strategies: (A) no scaling, (B) square-root intensity scaling, as in [160], (C) weight-based intensity scaling, as in Equation 5. Scaled spectra was normalised to the total magnitude of

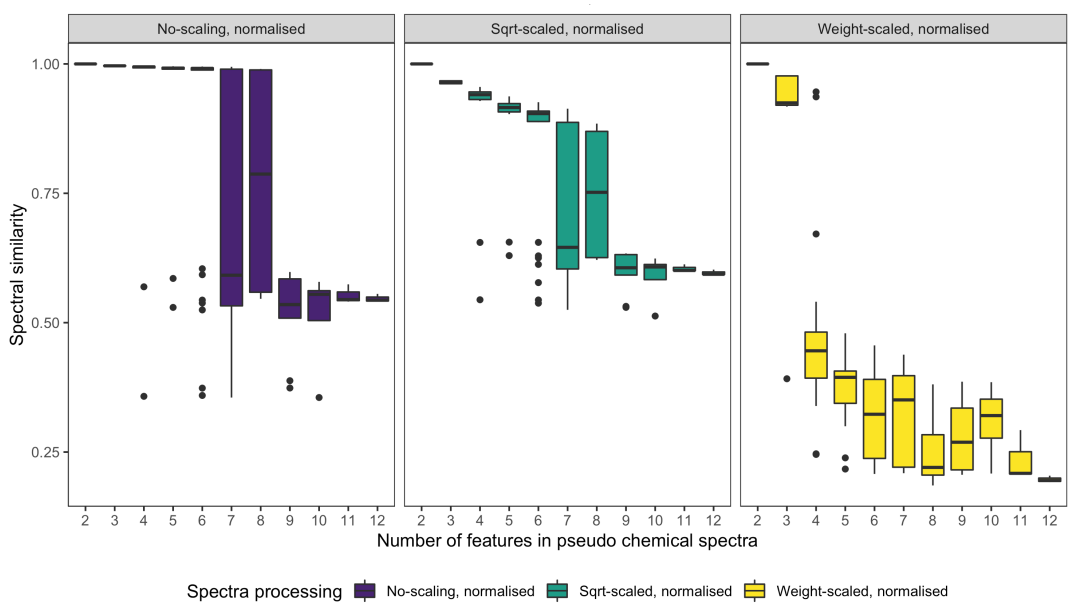


FIGURE 3.16: Features corresponding to imidazolelactate adducts were grouped into PCS in all DEVSET samples. Spectral similarity between the reference PCS and all of the PCS with the imidazolelactate adducts was evaluated using three intensity scaling strategies: left plot - no scaling, middle plot - square-root scaling, right plot - weight-based scaling.

the spectral vector. A simple spectral dot product function (Equation 1) was then applied to determine spectral similarity between the PCS in the first DEVSET sample in which adducts of metabolite-of-interest were detected and all following samples.

Results obtained with metabolite imidazolelactate are decribed in details below, while results for other 14 metabolites are available in Appendix C. Two validated ions of imidazolelactate - protonated molecule at *m/z* 157.0608 and in-source fragment at *m/z* 111.0546 - were grouped into PCS with varying number of other features. The largest PCS containing these two adducts was made of 12 features in total. Spectral similarity between the PCS comprising features corresponding to imidazolelactate is summarised in Figure 3.16. An example of discrepancy between different scaling methods is visualised in Figure 3.17. Weight-scaling undoubtedly performed worse than the other two tested methods with all metabolites since it generated most variation in the spectral similarity values. However, cosine values dropped for larger, i.e. more complex, PCS independently of the scaling method. The differences between no-scaling and square-root scaling are harder to conclude since their performance varied more between different metabolites (Appendix C). For imidazolelactate, it could be argued that no scaling retained cosine values closer to 1 for more of the PCS. Nevertheless, spectral similarity for unscaled imidazolelactate PCS was more varied, particularly for larger PCS with more than 5 features. In the light of these simulations, it could be concluded that no scaling or square-root could be performed. Square-root scaling was selected for further method development and testing both because of its more stable behaviour with 6 of the 14 tested metabolites and previously reported success with a spectral searching tool [160, 168].

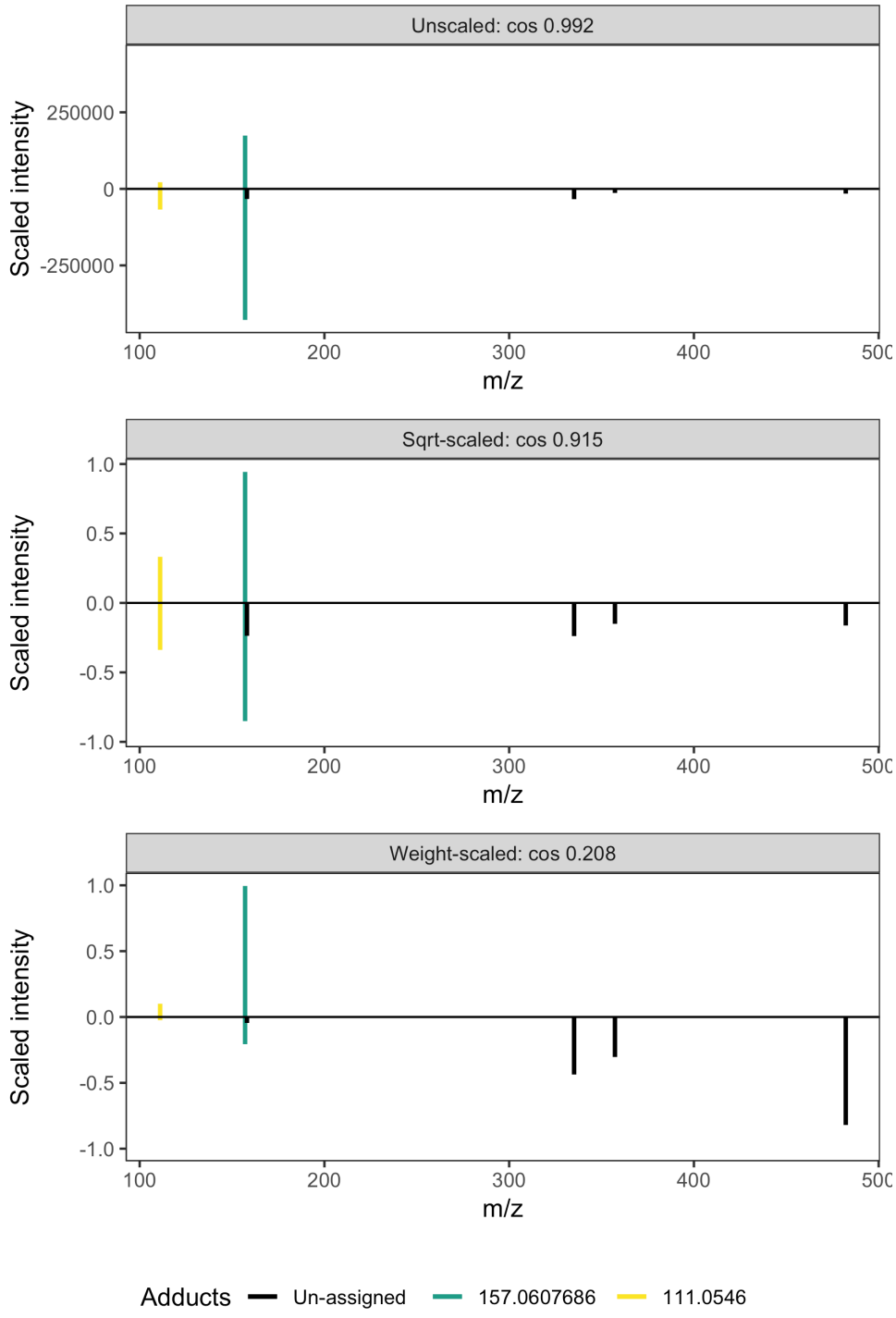


FIGURE 3.17: Spectral similarity score between PCS containing two imidazole ions (adducts at *m/z* 157.0607 and *m/z* 111.0546) largely depends on spectra scaling method.The top panel in each plot represents the spectra of the target PCS, whereas the bottom panel is the spectra of the same matching PCS, scaled using a different method in each plot.

Feature alignment validation

One of the major assumptions behind massFlowR pipeline design is that intensities of peaks from the same metabolites are correlated across all samples in the dataset. Intensity correlation is implemented in multiple metabolite annotation workflows, such as CAMERA [156, 164], PUTMEDID-LCMS [169] and RAMClustR [170]. Here correlation analysis is used to denoise aligned pseudo chemical spectra, which tend to grow in size as features from similar PCS are added from each sample. PCS features are correlated to each other, the resulting correlation matrix is clustered using network analysis via IGRAPH package [157]. Identified communities of features are considered to be ions of the same metabolite and are marked as part of the same PCS.

TABLE 3.3: Intensity correlation between the main ions and corresponding adducts/in-source fragments of 15 metabolites across DEVSET samples was analysed. Obtained Pearson correlation coefficients between all ion pairs is listed.

Metabolite	**RT**	**Main ion m/z**	**Adduct/fragment m/z**	**Intensity correlation**
Urocanate	57.54	139.0502	121.0395	0.986
Theobromine	145.14	181.0720	163.0618	0.971
			138.0665	0.997
Pseudouridine	56.94	245.0768	209.0560	0.978
			191.0447	0.971
			179.0446	0.979
			155.0440	0.979
Pantothenate	141.06	220.1179	202.1090	0.993
			184.0980	0.982
1-Methyladenosine	75.3	282.1197	150.0780	0.994
N-a-Acetyl-L-arginine	54.96	217.1295	200.1040	0.963
N2,N2-Dimethylguanosine	143.22	312.1302	180.0890	0.998
2-Furoylglycine	141.66	170.0448	124.0390	0.995
			95.0130	0.998
Creatine	36.48	132.0768	90.0550	0.773
Caffeine	208.92	195.0877	138.0668	0.997
7-Methylguanine	75.66	166.0723	149.0455	1.000
			124.0500	0.985
Pyroglutamate	76.38	130.0499	84.0450	0.996
Paraxanthine	166.26	181.0720	124.0515	0.997
Theophylline	168.6	181.0720	124.0515	0.997
Imidazolelactate	36.6	157.0608	111.0546	0.981

To illustrate the validity of intensity correlation for feature alignment validation, 15 metabolites were identified in DEVSET study, as described earlier in section 3.3.1. The main ion of each metabolite was correlated with all its validated adducts/in-source fragments. Obtained Pearson correlation coefficients (Table 3.3) indicate high correlation between all structurally related ions. While correlation coefficients for adduct/fragment pairs of all metabolites but one were > 0.97, the single exception - creatine with correlation of 0.773 - suggested that a correlation threshold must be drawn with great care.

In order to assess how robust Pearson correlation is in different studies, a large-scale metabolic profiling study AIRWAVE was investigated using validated metabolites. Metabolite detection and integration procedure is described in detail in Chapter 2, Section 2.2.3 and Appendix A. As with the DEVSET study, Pearson correlation coefficients were obtained between the adducts and in-source fragment pairs for the 35 validated metabolites (Table 3.4). These results demonstrate that correlation between structurally related ions vary widely depending on the study. As discussed in Chapter 2, AIRWAVE1 serum HILIC-POS-MS experiment experienced a great deal of analytical variation. As intensity correlation was investigated using raw LC-MS spectra rather than processed and corrected features, the introduced biases greatly affected correlation coefficients. In AIRWAVE, obtained correlation coefficients range from as low as -0.13 to 0.995 with a slight variation between analytical batches (Figure 3.18). In the light of these results, it was concluded that intensity correlation threshold for validation of pseudo chemical spectra should be selected for each dataset independently.

Given the single outlier in the correlation results for the DEVSET study, a cut-off of 0.75 was selected for feature alignment validation step for pre-processing of this study. A cut-off of 0.75 was also selected as the default parameter for feature alignment validation step in the massFlowR pipeline (Figure 3.9). Nevertheless, it can be adjusted by the user when calling the corresponding massFlowR function.

TABLE 3.4: Intensity correlation between the main ions and corresponding adducts/in-source fragments of 35 metabolites across AIRWAVE1 serum HILIC samples was analysed. Obtained Pearson correlation coefficients between all ion pairs are listed for each analytical batch separately.

Metabolite	**RT**	**Main ion m/z**	**Adduct/Fragment m/z**	**Pearson correlation Batch 1**	**Batch 2**	**Batch 3**
Adenosine	104.41	268.1040	136.0620	-0.060	0.097	0.003
Carnitine	320.26	162.1125	184.0950	0.979	0.990	0.985
			103.0386	0.911	0.893	0.799
Laurylcarnitine (C12:0)	234.37	344.2795	366.2620	0.243	0.572	0.449
			285.2080	0.208	0.071	-0.008
Histidine	369.04	156.0768	178.0580	0.842	0.778	0.277
			110.0711	0.931	0.925	0.742
N6,N6,N6-Trimethyllysine	369.77	189.1598	130.0852	0.730	0.707	0.395
			84.0800	0.072	0.384	0.157

TABLE 3.4: Intensity correlation between the main ions and corresponding adducts/in-source fragments of 35 metabolites across AIRWAVE1 serum HILIC samples was analysed. Obtained Pearson correlation coefficients between all ion pairs are listed for each analytical batch separately.

Metabolite	**RT**	**Main ion**	**Adduct/Fragment**	**Pearson correlation**		
		m/z	**m/z**	**Batch 1**	**Batch 2**	**Batch 3**
Trigonelline	291.00	138.0550	94.0650	0.125	0.711	0.469
Betaine	286.97	118.0863	140.0682	0.931	0.984	0.918
Warfarin	40.37	251.0703	163.0398	0.271	0.508	0.319
Caffeine	50.98	195.0877	138.0680	0.508	0.582	0.736
Creatinine	150.41	136.0481	227.1250	0.945	0.959	0.919
1,1-Dimethylbiguanide	205.41	130.1087	113.0810	0.071	0.324	0.199
Tryptophan	229.34	205.0972	188.0710	0.745	0.242	0.420
Phenylalanine	226.78	166.0863	103.0550	0.850	0.726	0.754
Methionine	246.74	150.0583	194.0220	0.829	0.793	0.499
			133.0310	0.848	0.846	0.309
Trimethylamine N-oxide	250.72	76.0757	151.1441	0.827	0.841	0.408
Proline	267.45	116.0706	160.0350	0.988	0.992	0.984
			365.0670	0.958	0.968	0.953
Alanine	273.94	134.0188	313.0355	0.955	0.950	0.901
			356.0780	0.802	0.753	0.036
Creatine	303.58	132.0768	154.0587	0.937	0.990	0.651
			176.0415	0.928	0.984	0.668
Glutamine	303.32	147.0764	191.0403	0.726	0.907	0.812
			130.0510	0.701	0.911	0.763
Citrulline	344.21	176.1030	198.0850	0.796	0.694	0.265
			159.0770	0.784	0.733	0.126
Arginine	355.10	175.1190	219.0860	0.944	0.956	0.899
			158.0920	0.798	0.690	0.368
a-glycerophosphocholine	361.04	258.1101	280.0920	0.940	0.984	0.853
			184.0720	0.653	0.661	0.060
			104.1070	0.970	0.982	0.906
3-methylhistidine	365.50	170.0924	126.1020	0.122	0.258	0.212
Hypoxanthine	95.52	159.0277	119.0360	0.848	0.948	0.857
			110.0350	0.795	0.927	0.761
Pantothenate	67.97	220.1179	242.0990	0.442	0.497	0.267
			202.1070	0.078	0.318	0.088
Urocanate	78.23	139.0502	95.0600	0.157	0.101	0.129
5′-Methylthioadenosine	78.07	298.0968	136.0630	0.074	0.373	0.141
Pipecolate	268.72	130.0863	174.0500	0.931	0.952	0.812
Thiamine	329.55	265.1123	122.0710	0.441	0.761	0.320
4-Guanidinobutanoate	230.41	146.0930	168.0740	-0.044	0.174	-0.009
			86.0600	0.078	-0.056	-0.130
N,N-Dimethylglycine	278.32	104.0712	148.0330	0.696	0.666	0.248
Inosine	97.31	291.0700	313.0520	0.990	0.995	0.977
Cortisol	44.46	363.2166	345.2040	0.303	0.706	0.526
			327.1970	0.264	0.779	0.284
1-Methylnicotinamide	249.34	137.0713	94.0660	0.682	0.687	0.431
Sucrose	138.21	365.1053	381.0790	0.587	0.839	0.446

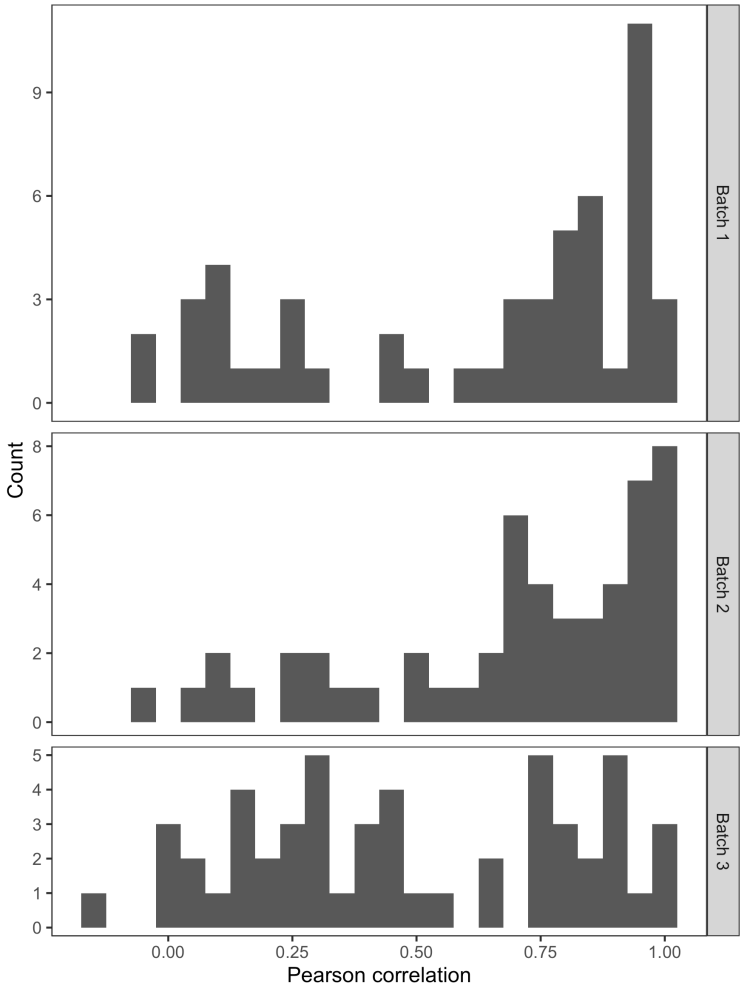


FIGURE 3.18: Intensity correlation between the main ions and corresponding adducts/in-source fragments of 35 validated metabolites across AIRWAVE1 serum HILIC samples was analysed. The distribution of the obtained Pearson correlation coefficients between all ion pairs are visualised for each analytical batch separately.

3.3.2 Comparison to other tools

To assess the performance of the developed feature alignment algorithm, synthetic datasets were generated by systematically introducing noise into the LC-MS features table obtained for a real sample. Such a testing approach provides numerous advantages over the use of experimentally derived or purely synthetic data. First of all, the absolute ground truth can be established. Secondly, introducing noise into a real sample, rather than creating entirely artificial ones, ensures that assessment is performed in as much realistic settings as possible. Finally, incremental introduction of noise allows to test the effect of different types of noise on the accuracy of algorithms.

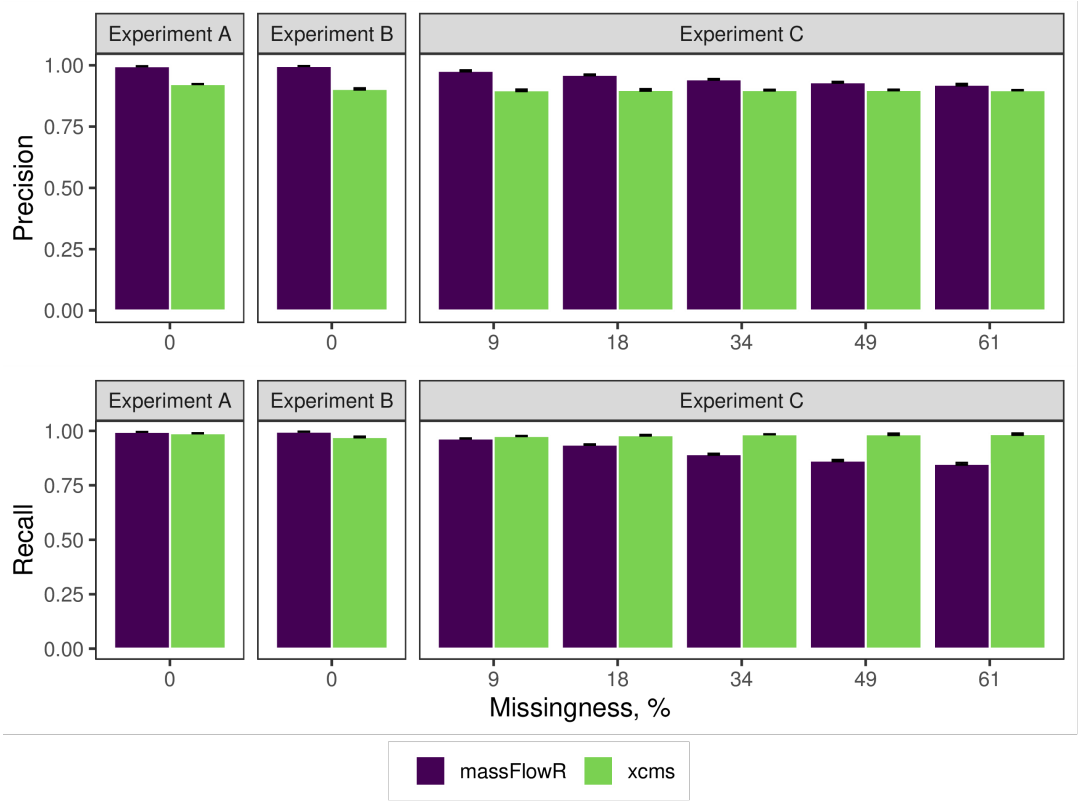


FIGURE 3.19: Precision and recall values obtained with massFlowR and XCMS on synthetic data, generated with three different types of experimental noise in experiments A, B and C. Summaries of results obtained with three replicates in each experiment are shown.

To benchmark and compare the performance of the developed feature alignment algorithm with the most commonly used tool, "density" method within XCMS, simulation experiments were performed. The obtained precision and recall values are reported in Figure 3.19 and Table 3.5. While both methods performed with generally high scores, higher alignment precision was achieved with massFlowR in all three experiments. Introduction of systematic RT noise in experiment B did not affect the precision and recall values for massFlowR. In contrast, XCMS suffered from a drop in both precision and recall in response to systematic RT drift. However, the performance of XCMS alignment did not deteriorate with incremental removal of features in further experiment C. On the contrary, massFlowR was less robust to feature missingness introduced in experiment C. While its performance precision dropped only slightly from 0.997 in experiment B to 0.921 in the final experiment C, the recall values dropped from 0.996 to 0.893.

These results indicate that XCMS performance is relatively optimised. Its sensitivity to non-linear RT drift and robustness to feature missingness is not surprising given the underlying algorithm. The "density" method within XCMS finds matching features across samples by calculating the overall distribution of peaks' RT in a given *m/z* bin [96]. The employed kernel density estimator is not meant to be applied to a dataset where features do not deviate around a "central" RT, but rather drift from

TABLE 3.5: Precision and recall values obtained with massFlowR and XCMS on synthetic data in three experiments. Values correspond to the mean of the scores obtained with three replicate datasets in each experiment.

Experiment		**A**	**B**			**C**		
Target missingness, %		-	-	5	10	20	30	40
Total missingness, %		-	-	9	18	34	49	61
Precision	**massFlowR**	0.996	0.997	0.977	0.961	0.943	0.931	0.921
	xcms	0.923	0.904	0.898	0.899	0.899	0.899	0.899
Recall	**massFlowR**	0.995	0.996	0.965	0.937	0.893	0.863	0.849
	xcms	0.989	0.972	0.976	0.98	0.984	0.984	0.985

sample to sample in a non-linear manner, as in experiments B and C. The removal of random features in experiment C do not affect the success of feature matching as all similar RTs within one *m/z* bin are drawn together from all samples and the order of samples is lost in the process. Consequently, XCMS feature alignment performed relatively stably with the simulated datasets which comprised of 100 samples each, with its performance dropping slightly in experiment B but not C. In contrast, the performance of the newly developed massFlowR algorithm was more varied, with much higher precision than XCMS in experiments A and B, and lower recall values in experiment C. These results are likely a symptom of the unsupervised approach employed within massFlowR feature alignment algorithm. The current algorithm ensures that only the most similar PCS are grouped together, whereas PCS with more deviated features are added as new, initiating a new chain of features. This attribute contributed to the drop in recall values in experiment C, during which features were grouped correctly but into multiple separate PCS. Nevertheless, a more considerable drop in recall value to 0.893 occurred only when 34% of all features were removed from datasets. Such ratio of missingness is higher than the typically observed 20 - 30% missing values in untargeted LC-MS data [163]. It is important to note that the developed algorithm compares samples in their original acquisition order, which contributed to its robustness to non-linear RT drift introduced in experiment B. Therefore, it could be concluded that XCMS and massFlowR alignment algorithms operate in a fundamentally opposite manner and each has its own weaknesses and strengths.

3.3.3 Proof-of-concept

To demonstrate that massFlowR pipeline aligns complex samples with high precision, the DEVSET study was utilised. The simplex lattice experimental design applied in DEVSET study provides an opportunity to evaluate how well a given data pre-processing method preserves expected biological variation. Theoretically, in a mixture design study the observed response of a metabolite follows the mixture design [171]. In DEVSET study, three unique urine samples were mixed in all possible combinations, producing six unique samples in total. Thus, each sample is a combination of the three unique samples at three possible concentrations (0, 1/3, 2/3,

TABLE 3.6: Number of features in the massFlowR and XCMS datasets for DEVSET study before and after features removal based on RSD and correlation to dilution assessment. Two version of filtered datasets were produced: using raw intensity values and using batch-corrected intensity values.

	Before filtering	After filtering	
	Raw features	**Raw features**	**Batch-corrected features**
massFlowR	6,849	3,771 (44.9% removed)	4,175 (39 % removed)
XCMS	13,088	7,006 (46.5 % removed)	7,723 (41 % removed)

1). DEVSET samples are therefore expected to diverge in the sample compositional space accordingly to their mixture design.

To evaluate how well massFlowR pipeline preserves the expected sample composition, raw DEVSET data was subjected to massFlowR, as well as XCMS pre-processing using parameters specified in Table 3.2. The quality of the obtained datasets was investigated using a number of metrics.

First, the analytical precision of the obtained features was examined. The distribution of relative standard deviation (RSD) values estimated for all features across pooled QC samples is relatively similar in massFlowR and XCMS datasets (Figure 3.20). Nevertheless, the peak of the distribution is shifted more to the right in the XCMS dataset with a median of 19.7%, in contrast to 14.6% in the massFlowR dataset. The distribution of correlation to dilution coefficients is also comparable between massFlowR and XCMS features, however, XCMS reported more negatively correlated features, as indicated by the second peak around 0.0 to -0.2.

Removal of low quality features according to the QC standards described in Section 2.2.4 produced datasets of different sizes, as summarised in Table 3.6. While massFlowR reported visibly smaller number of features than XCMS, fewer of the raw massFlowR features were removed by the QC feature filtering pipeline (44.9%, in contrast to 46.5% of XCMS raw features). These results are in line with the expectations given the distribution of RSD values and correlation to dilution coefficients in Figure 3.20. This suggest that XCMS reports more noisy features that do not meet the QC criteria and are removed in the post-processing. Nevertheless, as XCMS produced more batch corrected features than massFlowR (7723 and 4175 respectively), it indicates that massFlowR removed some of these reproducible features reported by XCMS in its pre-processing.

To assess filtered datasets quality and determine any potential analytical associations with the main sources of variance, multivariate analyses were employed, as described in Section 2.2.4. The scores of the calculated principal components were correlated with analytical parameters. Raw massFlowR features (Figure 3.21) were indeed highly correlated with run order and MS detector voltage, both of which are highly interlinked analytical parameters. On the other hand, such analytical variance association with principal components was not observed in XCMS dataset.

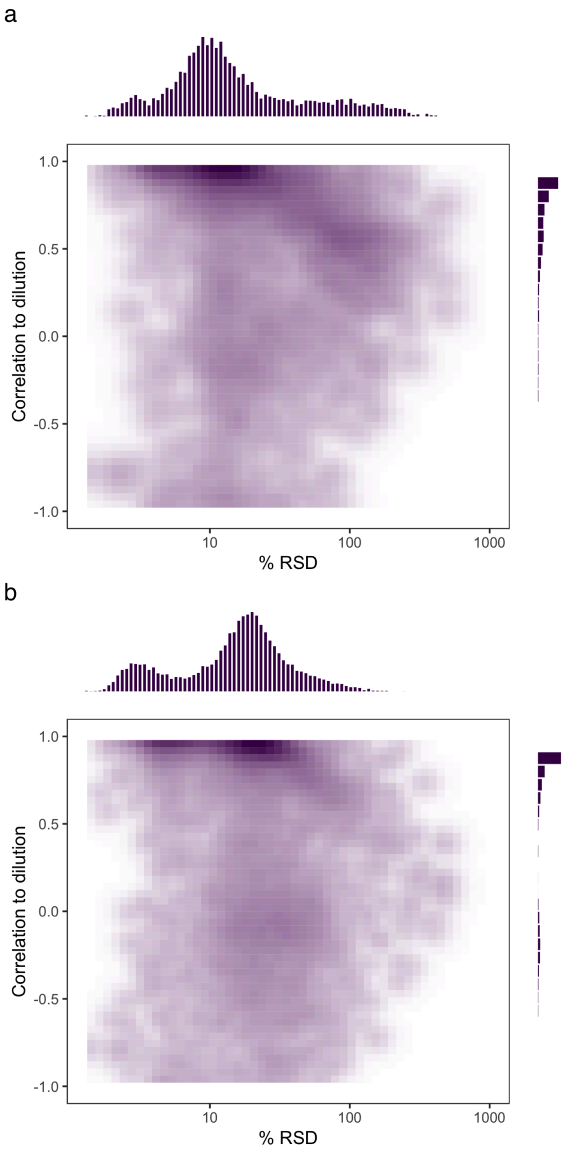


FIGURE 3.20: The analytical precision (expressed as relative standard deviation, RSD) and linearity of response (correlation to dilution) of features detected and reported in the DEVSET samples by (a) massFlowR and (b) XCMS pre-processing pipelines. The RSD values for XCMS-reported features are generally higher than for the massFlowR features. Similarly, the distribution of correlation coefficients of the XCMS-reported features has a peak around 0.0 to -0.5, indicating a higher proportion of noisy features in the dataset.

These results are highly positive, since they indicate that massFlowR feature alignment preserves the first expected variance in the data - the variance arising from sample acquisition order. Biases dependent on the run order, such as changes in retention time, peak shape, sensitivity or MS accuracy, cannot be fully eliminated even by the strictest control over the experimental conditions [172]. Among these biases arising due to e.g. column ageing or ion source contamination, time-dependent intensity drift is the main source of unwanted variation in LC-MS data [88, 91–94]. Intensity drift was also observed in the DEVSET study, illustrated by the decreasing total ion current for each acquired sample (Figure 3.22).

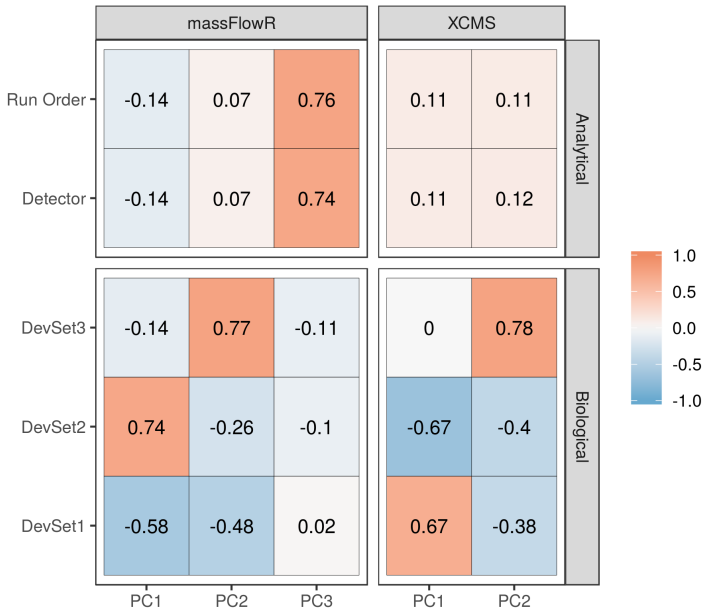


FIGURE 3.21: Principal components scores correlation with analytical and biological variance in DEVSET data produced by massFlowR and XCMS pre-processing pipelines. The number of principal components to be subjected to correlation analyses was estimated using 7-fold cross-validation.

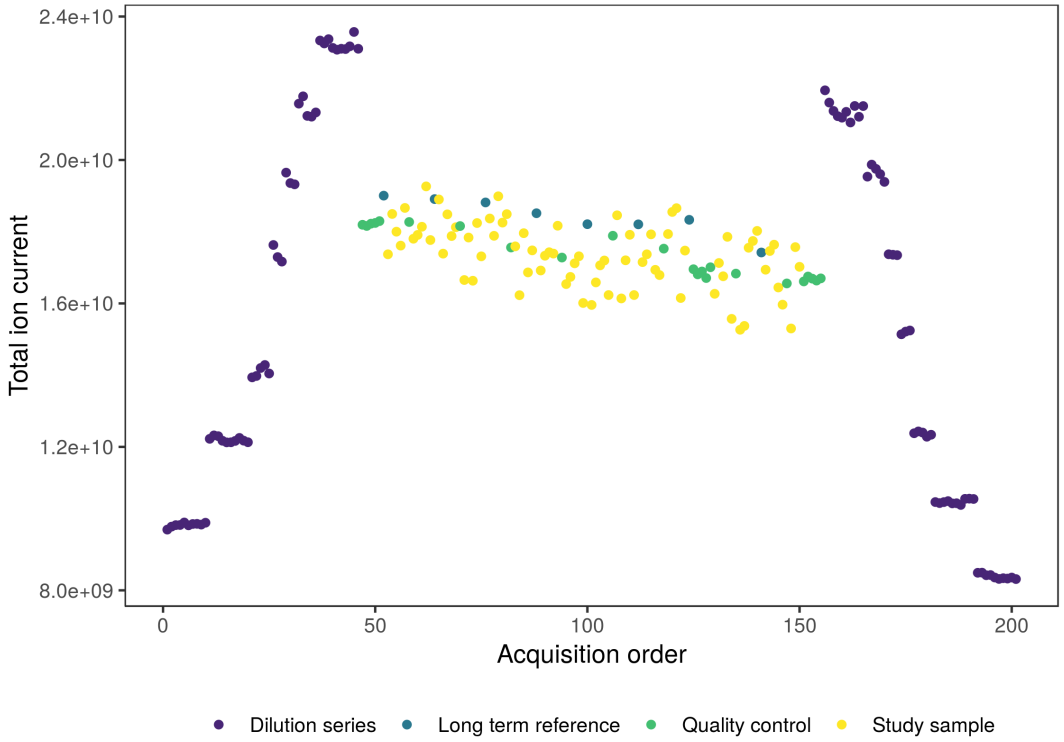


FIGURE 3.22: Total ion current for each sample in the DEVSET study depicts that detected ion intensity dropped with each acquired sample.

To account for the observed intensity drift, batch correction was applied to DEVSET datasets. The LOESS-based intensity smoothing [92] was followed by QC feature filtering. These final datasets were subjected to PCA analysis. Principal components correlation with analytical parameters indicated that batch-correction removed the unwanted analytical variance (Figure 3.23). Association with the sample origin - the three original urine samples (DevSet1, DevSet2 and DevSet3) of which each sample in the study was made off - was very strong in both massFlowR and XCMS datasets. Each of the three urine samples was clearly associated with one of the calculated principal components. Samples also clustered according to their origin in the PCA scores plots (Figure 3.24). The PCA scores plots show that both tools produced similar results for a biological dataset even though XCMS reported more features.

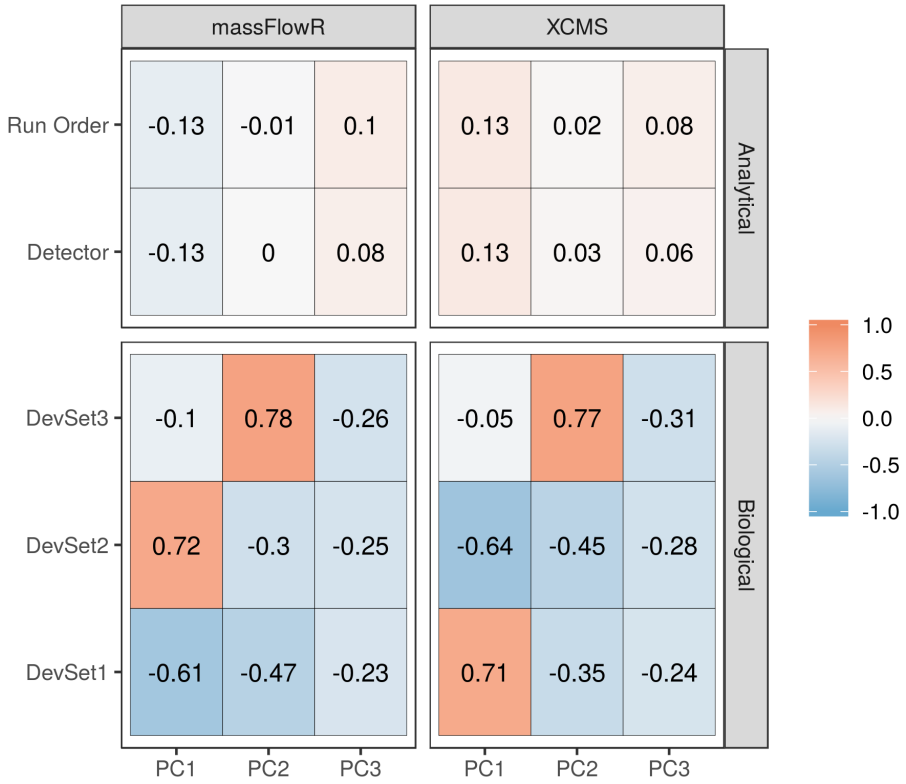


FIGURE 3.23: PCA scores correlation with analytical and biological sources of variation in the DEVSET study datasets generated by either massFlowR or XCMS pre-processing pipelines. The colors and numbers indicate Pearson correlation coefficient for each principal component and source of variation combination. Batch-corrected and QC-filtered datasets were analysed.

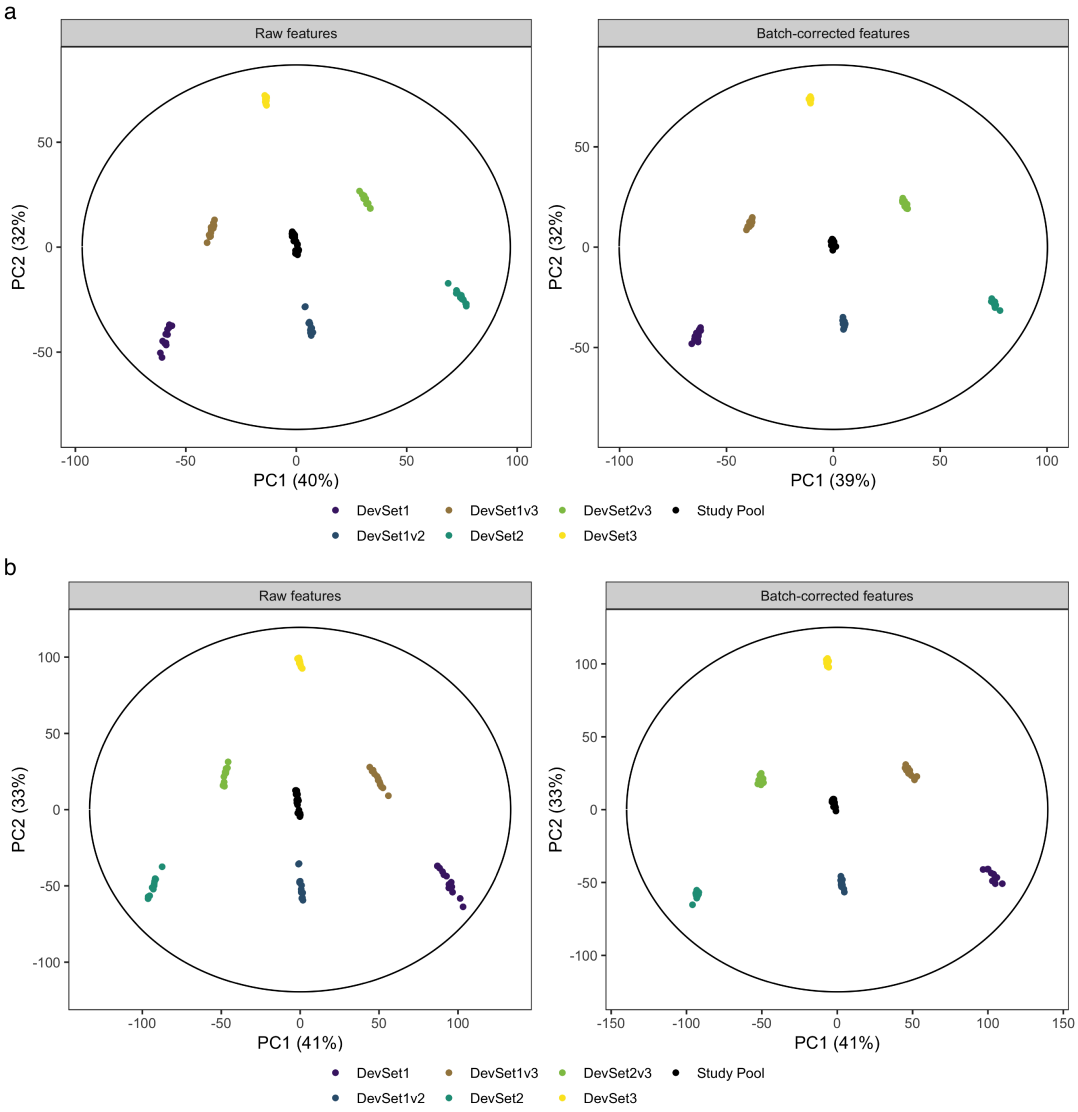


FIGURE 3.24: DEVSET samples segregation into neat clusters according to their sample class was achieved by both (a) massFlowR and (b) XCMS pre-processing.

3.4 Conclusions

In this chapter, a novel feature alignment method massFlowR that incorporates information on features grouping was proposed. The method builds pseudo chemical spectra comprised of structurally related co-eluting features in a an individual LC-MS sample and finds the best match for each of the features in the next acquired sample by evaluating the overall spectral similarity of matching pseudo chemical spectra.

The method fits into the category of direct matching algorithms as RT is not corrected prior features alignment. Results with synthetic and real metabolomics datasets demonstrate the potential of this approach. While recall values were lower for massFlowR than for XCMS with synthetic datasets with > 18% of all features removed, precision values were higher for massFlowR in all experiments. Furthermore, application to a real metabolomics dataset indicates that massFlowR accurately captures the underlying sources of variance, such as the expected time-dependent intensity drift, as well as biological sample origin. Most of all, information on how features are related to one another is available in the final massFlowR output, which will be discussed in details in Chapter 4.

Chapter 4

Strategies for automatic LC-MS features annotation

4.1 Introduction

Mass spectrometry coupled to liquid chromatography (LC-MS) is being increasingly utilised for molecular profiling of biological samples. Advances in instrumentation, particularly development of UPLC [71] coupled to high resolution mass spectrometry (HRMS) [79], which provided substantial enhancement in detection sensitivity, secured LC-MS the central role in the field of metabolomics [128, 173]. The range of untargeted LC-MS applications [92] was further expanded by the standardisation of experimental protocols for large-scale population studies with thousands of samples [52, 62, 174]. However, the key aspect of metabolomics is spectral feature annotation to the corresponding metabolite. Despite of depth of metabolic information acquired for each sample, structural metabolite identification still represents a major challenge in LC-MS based profiling [79]. Traditionally, only a subset of all detected features that were found to be statistically significant are annotated using laborious methods [175], which heavily depend on manual data analysis [27] and often yield multiple putative identifications [112].

To reach a consensus about reporting standards, four levels of confidence in metabolite identification were proposed by the Metabolomics Standards Initiative (MSI) of the Metabolomics Society [176]. A new 'Level 0' was proposed at the 2017 annual meeting of the Metabolomics Society (Brisbane, Australia) (Table 4.1) [177], which requires confirmation of the 3D structure and stereochemistry of isolated pure metabolite.

4.1.1 Challenges and current standards

Several analytical parameters can be used for LC-MS metabolite annotation: accurate mass (AM), chromatographic retention time (RT), fragmentation pattern (MS/MS) and information about the sample, such as abundance in certain biological groups [111].

TABLE 4.1: Summary of levels of confidence in metabolite identification, as proposed at the 2017 annual meeting of the Metabolomics Society (Brisbane, Australia).

Confidence	Definition	Minimum data requirements
Level 0	"Unambiguous 3D structure"	As level 1, and, Full stereochemistry of isolated pure metabolite.
Level 1	"Confident 2D structure "	As level 2, and, Two orthogonal analytical techniques applied to the analysis of both the metabolite and the chemical standard: accurate mass and RT, MS/MS, isotopic pattern, 2D NMR spectra,...
Level 2	"Putatively annotated structure"	As level 3, and, Spectral (LC-MS and/or NMR) similarity with libraries and/or literature data.
Level 3	"Putatively characterised structure and/or class"	As level 4, and, Spectral (NMR) and/or chromatographic (LC-MS) features consistent with the characterised class.
Level 4	"Unknown"	A reproducible and quantifiable signal in a sample.

While high mass accuracy is achieved with modern time-of-flight (TOF) mass spectrometers (four decimal places, < 5 parts per million), on its own it is not enough to determine unambiguous elemental composition of a compound given its accurate mass [178]. Formula determination is especially complex for molecules containing common elements C, H, N, S, O and P, which are particularly common in the field of metabolomics. For example, glutamine's formula of $C_5H_{10}N_2O_3$ can have over one million theoretical structures [111]. An added challenge in metabolite identification is the presence of structural isomers, which are very common among organic analytes. Structural isomers are compounds with the same molecular formula but different physical and chemical properties. For example, glucose-6-phosphate, fructose-6-phosphate and glucose-1-phosphate all have identical parent ion mass and even similar MS/MS fragmentation pattern [179]. Nevertheless, they play very different roles in glucose metabolism and therefore should be interpreted separately during statistical data analyses. In lipidomics chemical space is expanded ever more by the presence of multiple fatty acid chains, which also frequently contain unsaturated bonds [180]. To detect and quantify each of the structural isomers separately, they must be resolved prior to MS analysis. Multiple methods have been suggested, including ion mobility-MS, since it separates ions based on their mass, charge and cross section, which is linked to ion size and shape. But the most explored option is coupling MS to different LC systems [181], which separates isomers in the chromatographic space.

Chromatographic separation provides additional information essential for resolving ambiguous metabolite annotations made through AM alone. Chromatographic RT is a chemical structure-specific property, therefore, matching unknown analyte's RT with candidate's standard RT provides a high level of annotation confidence (Table 4.1). Nevertheless, in addition to the need of procuring authentic standards, which can be economically infeasible for large-scale annotation efforts, RT matching driven annotation is also challenged by RT shifts. As described in details in Chapter 2, RT varies not only between different LC-MS instruments operated following the same protocol, but also during a single LC-MS experimental run. Even minor differences

in chromatographic conditions, such as pH or sample-induced matrix effects, can lead to RT shifts. As a consequence, using AM and RT alone does not always lead to unambiguous compound annotation.

In order to discriminate closely eluting compounds, as well identify compounds for which RT is not known, tandem mass spectrometry (LC-MS/MS) can be performed. Acquired MS/MS spectra provide increased confidence in metabolite annotations through spectral comparison to authentic chemical standards or to *in silico* fragmentation patterns. To assist the structural annotation of a metabolite, MS/MS spectra are matched against a reference spectral library. Open-source databases, such as METLIN [182], HMDB [183] and LIPID MAPS [184], are widely used due to their accessibility and high data quality. To achieve a broad metabolite coverage, a combination of multiple databases is often required, for example, HMDB does not contain lipids and therefore LIPID MAPS should be included in the analysis of human biofluids [169]. Furthermore, spectra for specific classes of chemicals sometimes have to be obtained from commercial alternatives, such as the National Institute of Standards and Technology (NIST) library.

Experimental MS/MS data can be matched against reference spectra using various spectral comparison methods. One of the most popular methods is the dot product, which computes the cosine of the angle between the unknown and the reference spectra vector representations (Figure 3.7). Originally proposed in 1978 by Sokolow et al. [185], dot product has been proven to be the most reliable method for library search and spectral comparison [158, 186]. Dot product based algorithm is implemented in the MS/MS Spectrum Match Search tool in METLIN [182] and the NIST mass spectral library [167]. Other commonly applied similarity measures include Tanimoto coefficient, also known as Jaccard index, which measures the similarity between finite sample sets. It is defined as the number of elements in common between the two sets divided by the total number of elements. The Tanimoto coefficient is a particularly intuitive similarity measure, which in the field of mass spectrometry accounts for the number of ions (i.e. spectral fingerprints) that might be in common relative to the number of ions that are common [187]. Tanimoto similarity index is included in the MetFrag workflow, which identifies small molecules using in silico fragmentation [187].

Even with such a broad selection of open source and commercial reference libraries, an annotated MS/MS spectrum is not always available for the compound of interest. While MS/MS spectrum prediction algorithms have been developed for such cases [187–189], fragmentation rules are not fully understood yet and prediction success is still erroneous for many chemical classes [190]. Furthermore, acquisition of MS/MS spectra for each unknown metabolite in a large-scale profiling study is often infeasible due to limited time and resources. Therefore, a solution for automatic metabolite annotation based on LC-MS spectra alone is desirable.

4.1.2 Automatic LC-MS spectra annotation

Soft ionization techniques, such as electrospray-ionization (ESI) which is explored in this thesis, principally produce ions of the intact analyte molecule, most often in the form of the protonated molecule, e.g. $[M+H]^+$, where M is the molecular mass of the analyte. In ESI, ions are formed in solution, as a result, multiply charged ions are produced with little or even no fragmentation of the analyte [74]. In addition, in ESI analytes can form adduct ions through solvent-analyte clustering, for example, potassiated and sodiated analyte ions $[M+K]^+$ and $[M+Na]^+$. Analytes can sometimes also form protonated multimers, such as $[2M+H]^+$. While the exact mechanism of adduct formation is not fully understood, high abundances of the most commonly encountered adducts, mainly $[M+K]^+$, $[M+Na]^+$ and $[M+NH_4]^+$ in positive ionisation mode, are widely reported [191]. Therefore, ESI produces a highly complex spectra, the highest *m/z* value in which is usually not the protonated analyte molecule, but an analyte-mobile phase cluster ion, or an analyte multimer ion.

Algorithms for automated untargeted LC-MS data annotation have been previously developed. Most of them are based on clustering features corresponding to ions originating from the same compound into a spectra.

A range of computational tools for LC-MS annotations are based on pairwise intensity correlation analysis across multiple samples. One of the earlier tools is MSClust, which operates on two main assumptions: (1) the chromatographic peaks of structurally related ions have similar RT span; (2) the intensity patterns across samples are similar for ions originating from a single metabolite [192]. Similarly to MSClust, the RAMClust feature clustering algorithm assumes that two features derived from the same compound will exhibit similar retention time and quantitative trend across samples [170]. While features can be correlated at the level of either: (i) MS vs MS, (ii) MS vs indiscriminate high-collision MS/MS (idMS/MS), or (iii) idMS/MS vs idMS/MS; algorithm is based on XCMS-detected and aligned features and does not make use of raw data. Other examples include AStream [193], MS-FLO [194], xMSannotator [195] and PUTMEDID-LCMS workflow [169]. All these tools rely on already aligned features and therefore are intrinsically sensitive to pre-processing errors that take place due to experimental noise, as described in Chapter 2.

An alternative school of thought suggests that features clustering should be based not only on intensity patterns across samples, but on chromatographic peak shape correlation as well [156]. The approach was implemented as a data-reduction tool [151], as well as an annotation tool CAMERA [164]. CAMERA is the most widely used annotation tool, as reported in the recent Metabolomics Society survey [173]. CAMERA performs clustering by selecting the most intense feature not yet assigned to a compound and adding all features within a RT window around its centroid into a new compound spectrum [164]. This spectrum is later refined by correlating the

extracted ion chromatograms (EIC) from the sample in which the feature-of-interest is the most intense. As a result, CAMERA is biased toward the most abundant features [170]. Furthermore, CAMERA was benchmarked with 48 LC-MS spectra and therefore may not be suitable for large-scale profiling studies.

A third parameter that is often used to deconvolute and annotate complex LC-MS spectra is the analysis of *m/z* differences between co-eluting features [112]. As certain forms of adducts are more likely to be produced in a given LC-MS experiment, rules on specific *m/z* differences can be applied in order to identify the structurally related features among the co-eluting features. For example, the *m/z* difference between the protonated $[M + H]^+$ and potassiated $[M + Na]^+$ analyte molecules is 21.9820. Such design was implemented in a number of tools, among which is the earlier described CAMERA [164], as well as lipid identification software LipiDex [196], LC-MS data processing and analysis platform MET-COFEA [152] and workflow PUTMEDID-LCMS [169]. Most of the platforms allow the user to specify the list of expected adducts, which accounts for the differences in the adduct formation processes among the varied LC-MS systems and protocols. Nevertheless, such approach is inherently targeted and will miss the previously uncharacterised adducts and fragments.

The newest edition to the features clustering and annotation toolbox is CliqueMS [166]. CliqueMS clusters XCMS features with similar chromatographic peak shapes, as denoted by their cosine similarity, together. Isotopes, adducts and in-source-fragments (ISF) are identified among the clustered features. Despite of improved annotation efficiency in comparison to CAMERA, CliqueMS only produces annotations for individual samples, preventing large-scale annotation efforts.

4.1.3 Hypothesis

Complex spectra generated by ESI provides an advantage in metabolite identification process. We hypothesise that annotating LC-ESI-MS adducts and ISF ions originating from the same metabolite as an MS/MS spectra through direct LC-ESI-MS spectra matching to a reference database will help make accurate putative annotations.

4.1.4 Aims and objectives

The purpose of this chapter was to annotate the AIRWAVE cohort LC-MS data and to explore multiple automatic data annotation strategies. This chapter is therefore organised in three main sections:

- Application and validation of the most popular open-source annotation tool CAMERA.

- Development and validation of a feature-to-spectra matching algorithm to annotate XCMS generated data to a reference database.
- Application and validation of massFlowR-driven pseudo chemical spectra matching to a reference database.

4.2 Methods

4.2.1 Data acquisition and pre-processing

Serum samples collected as part of the AIRWAVE cohort were analysed by HILIC-POS-MS at the National Phenome Centre (NPC), as described in details in Chapter 2, Section 2.2.1. A single batch of 1,326 samples, which includes pooled quality control, dilution series and external long-term reference samples, as well as 1,027 study samples (Table 2.3) were subjected to both XCMS and massFlowR pre-processing.

XCMS pre-processing methods are described in details in Chapter 2, Table 2.7. Parameters for massFlowR pre-processing and annotation are summarised in Table 4.2.

TABLE 4.2: CAMERA and massFlowR parameters used in the automatic annotation of AIRWAVE serum HILIC-POS-MS dataset. Functions are listed in the order of use.

CAMERA

Function	**Parameter**	**Value**
groupFWHM	perfwhm	0.6
findIsotopes	mzabs	0.01
groupCorr	cor_eic_th	0.75
findAdducts	polarity	"positive"

massFlowR

Function	**Parameter**	**Value**
groupPEAKS	peakwidth	c(1, 5)
	prefilter	c(10, 5000)
	noise	200
	snthresh	5
	ppm	25
alignPEAKS	rt_err	10
	mz_err	0.01
	cutoff	0
validPEAKS	cor_thr	0.5
	min_samples_prop	0.1
fillPEAKS	fill_value	"into"
annotateDS	rt_err	15
	mz_err	0.01

Unlisted parameters were set to defaults.

4.2.2 Standard annotation workflow

First, the most widely used open-source automatic annotation software, CAMERA, was applied to AIRWAVE data processed with XCMS. CAMERA workflow consists of three steps (corresponding functions are written in brackets):

- Features grouping into pseudo chemical spectra using RT (*groupFWHM*).

- Grouping validation using chromatographic peak shape similarity analysis (*groupCorr*).
- Detection of natural isotopes and adducts (*findIsotopes* and *findAdducts* respectively).

Applied CAMERA parameters are listed in Table 4.2.

4.2.3 Annotation to in-house database

Feature-to-spectra matching algorithm

In order to annotate data processed with standard peak detection software, e.g. XCMS, to an in-house chemical standards database, a feature-to-spectra matching algorithm was developed (Figure 4.1). First, dataset-of-interest features corresponding to a database compound are selected, as depicted in Figure 4.1.f. The match between selected dataset features and database compound is evaluated using a scoring method as follows:

$$score = \frac{\sum_i^n int(DS_j)}{\sum_j^k int(DB_i)} \times \frac{k}{n} \tag{1}$$

where n and k are the number of database compound features and matching features in the dataset of interest respectively. $int(DB_i)$ is the intensity of database compound's feature i; $int(DS_j)$ is the intensity of database compound's feature j that have a match in the dataset of interest. The ratio of sum of intensities of features with a match to the sum of intensities of all database features is proportional to how many of the the most intense features in the spectrum are matched. Similarly, ratio of the number of matching features to the total number of features in the compound spectrum is indicative of overall spectral similarity. Thus, proposed scoring method provides a value between 0 and 1, with 1 being a perfect match.

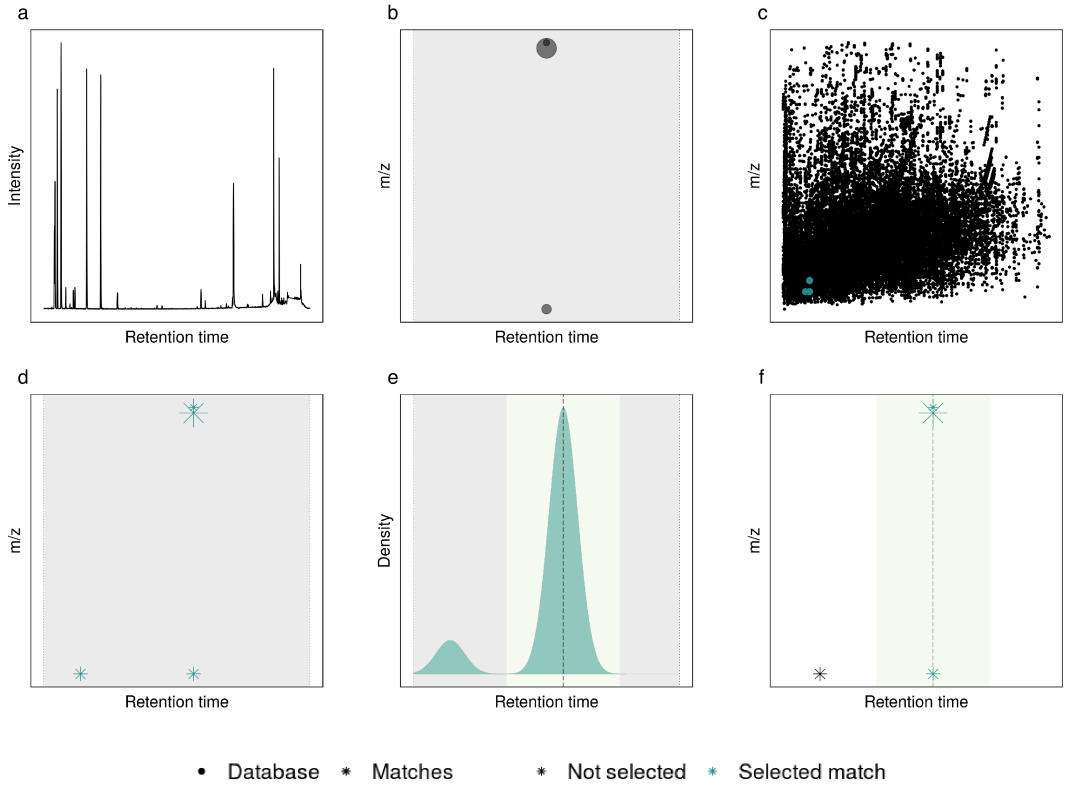


FIGURE 4.1: Feature-to-spectra matching enables dataset annotation using a chemical standards database. First, (a) raw LC-MS spectra is acquired for every chemical standard in the database. (b) Raw spectra is processed to extract features specific to this compound. (c),(d) Features matching to compound's *m/z* and RT values are identified in the dataset using *m/z* and RT windows. (e) Density estimate indicates the central RT value among the matching dataset features (dashed line). (f) Matches within two standard deviations of this central value (green area) are retained, only the closest match for each database feature is selected.

Spectra-to-spectra matching algorithm

In order to annotate data processed with massFlowR pipeline, which is discussed in details in Chapter 3, a spectra-to-spectra matching algorithm was developed. The algorithm is based on the massFlowR PCS alignment method. However, instead of aligning a sample-of-interest to the template, the final massFlowR output containing aligned, validated and filled PCS are aligned to the database. Since database standards LC-MS spectra were processed to group structurally related and standard-specific features together, they are equivalent to PCS generated by massFlowR. Therefore, the dataset-of-interest is matched to the database PCS using spectral similarity comparison method, as in PCS alignment in massFlowR pipeline. Obtained matches are ordered by the similarity score and the highest scoring chemical standard is suggested as the top annotation.

4.2.4 Database generation

An in-house chemical reference database, acquired at the National Phenome Centre (NPC) prior to the start of this project, was used to annotate both XCMS and massFlowR generated datasets. An empirical and non-deterministic approach was implemented when preparing chemical standards for analysis: an small amount of standard was dissolved in water and a 1:10 dilution series was created to achieve a wide range of four orders of magnitude. Every concentration was analysed and the sample with the highest concentration that produced an un-saturated standard-specific signal was retained.

Standards were analysed with the ACQUITY UPLC (Waters Corp., Milford, MA, USA) chromatography system was connected to Xevo G2-S Q-ToF mass spectrometer (Waters Corp., Manchester, UK) with Zspray electrospray ionization (ESI) source. Three complementary chromatographic assays (HILIC, RP and lipid RP) were used according to the standard NPC protocols. MS data was recorded as three independent functions, both low-collision (LC) and high-collision (HC) energy acquisitions were performed.

Every chemical standard (i.e target sample) acquisition was preceded and followed by a blank sample. Raw spectral files were de-noised and centroided using proprietary Waters software. All nine MS functions (i.e. three per sample) were peak-picked using XCMS *centWave* function, detected features were aligned using XCMS *density* grouping function. Features that were present in at least two of the three MS functions of the target sample, as well as were at least 10-fold more intense than corresponding features in the blanks, were marked as the "seed" features. These "seed" features were then EIC-correlated to every other feature in the target sample spectra. Highly correlated features were grouped together, these groups here are referred to as pseudo chemical spectra to account for the similarly with the massFlowR implementation and terminology discussed in depth in Chapter 3.

Obtained groups of features were recorded as a separate data file for each standard. Database generation was performed by Dr M.R.Lewis at the NPC.

4.2.5 Annotation validation

46 metabolites and their main adducts and in-source fragments were identified in the LC-MS spectra of all AIRWAVE samples using *m/z* and RT regions kindly provided by the NPC team. Detection of metabolites in spectra was performed using R package peakPantheR. Detection regions of the validated ions are available in Appendix A.

All scripts used within this and other chapters are available on the public GitHub repository: `https://github.com/lauzikaite/PhD_thesis_code`.

4.3 Results

4.3.1 XCMS features annotation

The AIRWAVE serum HILIC dataset was pre-processed with XCMS. The details on the pre-processing and downstream quality control filtering results are available in Chapter 2, which discusses the XCMS-based workflow in greater detail.

The obtained XCMS features were subjected to the CAMERA automated annotation workflow, which assigns features originating from the same metabolite, such as adducts, natural isotopes and in-source fragment ions, into groups called pseudogroups [164]. CAMERA annotation results are visualised in Figure 4.2. Features that were annotated to an adduct or an isotope (yellow dots) are distributed among the un-annotated features (purple dots) in the *m/z* and retention time space of the assay relatively equally. While CAMERA workflow ensures that every feature is assigned to a pseudogroup, the number of features grouped to a single pseudogroup varies (Figure 4.3). Majority of the generated pseudogroups contain just a single feature and therefore do not provide meaningful information for metabolite identification validation. One of potential explanations for the observed results lies in the design of the CAMERA annotation algorithm, which: (1) clusters XCMS features to compound spectra using just their RT values, (2) detects isotopes within each compound spectrum by checking the intensity ratios for features with *m/z* difference of 1.0033; (3) refines compound spectrum by performing chromatographic peak shape similarity analysis in a set of selected samples and pairwise Pearson correlation of intensities across all samples ; (4) assigns features to molecular formulas using a user-provided list of *m/z* differences for adducts [164]. As CAMERA pipeline largely depends on RT values to form the initial features clusters, incorrectly pre-processed and aligned features are unlikely to be accurately annotated using this approach. Given the significant RT deviation observed in this study (Chapter 2, Section 2.3.1),

XCMS features with similar RT values may not be truly co-eluting in individual samples. Such initial clusters would therefore not exhibit a typical isotopic pattern, high EIC correlation in a single sample or intensity correlation across all samples. Inaccurately formed initial clusters may therefore result in pseudogroups comprised of just a single feature, as observed in Figure 4.3. Nevertheless, pseudogroups comprised of two to ten features are highly abundant in the CAMERA output and further analyses will focus on them, as they represent the most informative pseudogroups for the purpose of metabolite annotation.

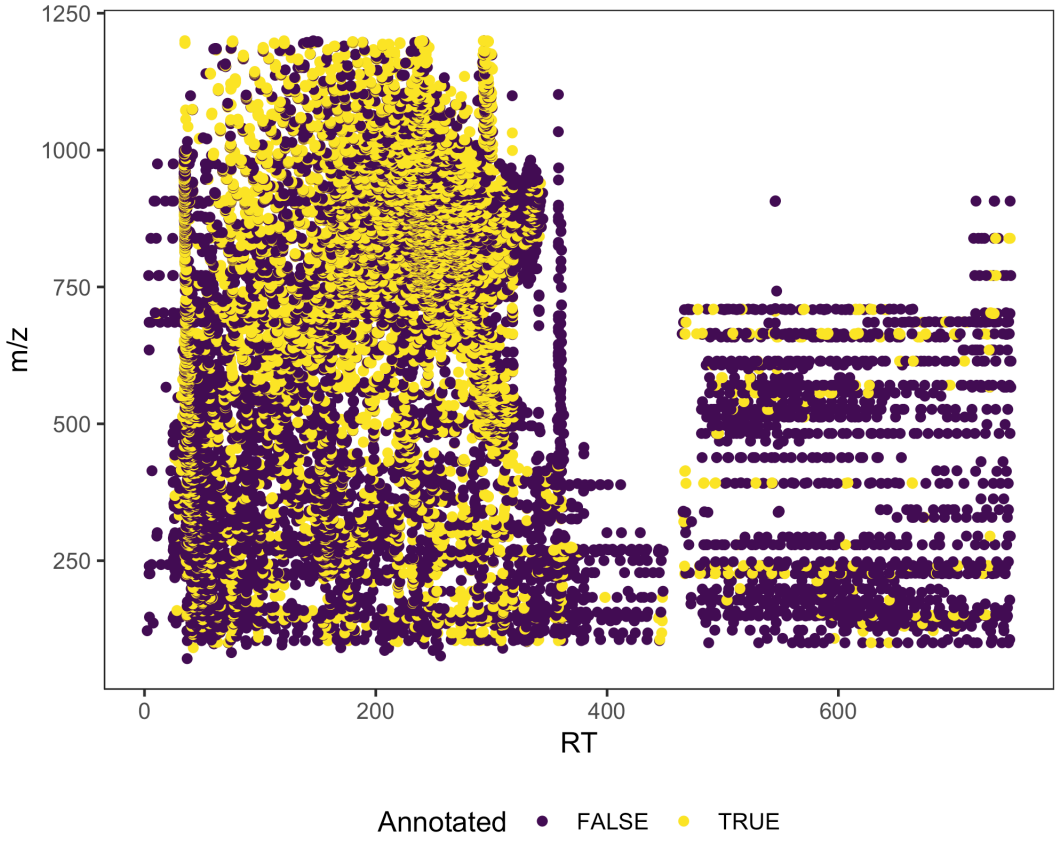


FIGURE 4.2: The ion map of features detected by XCMS in AIRWAVE serum HILIC POS assay dataset. The color of a feature indicates whether it was annotated to an adduct or a natural isotope by the CAMERA workflow.

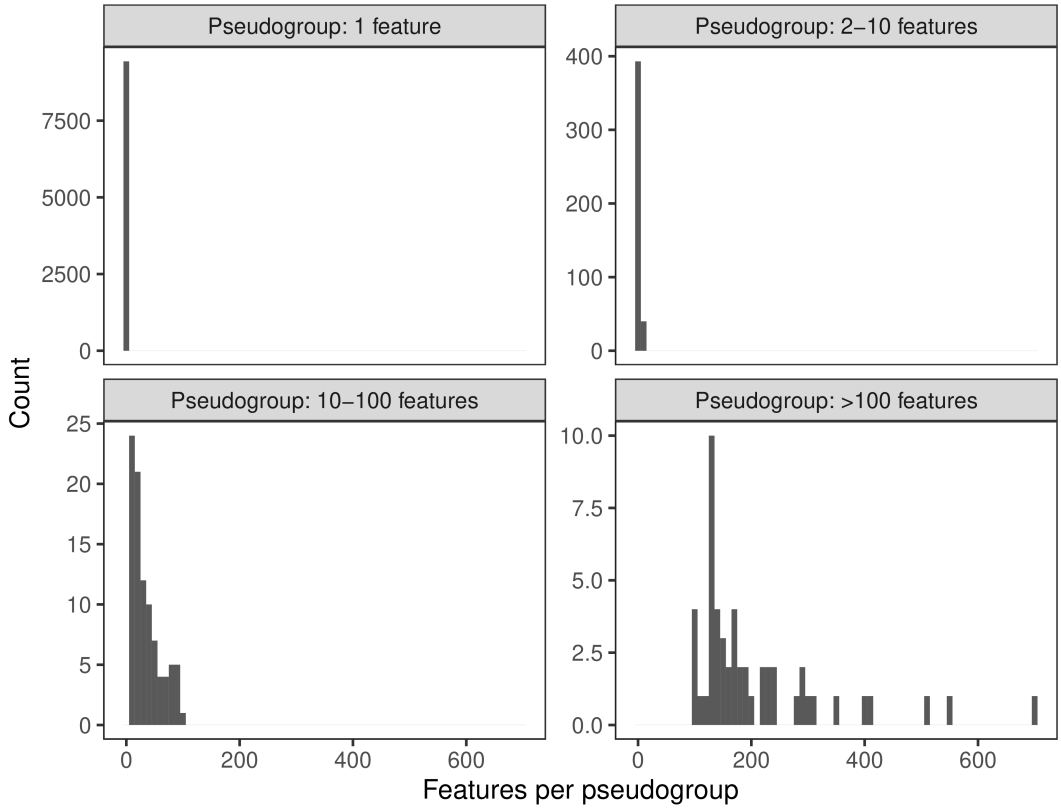


FIGURE 4.3: The number of XCMS features per pseudogroups, obtained by CAMERA workflow applied to AIRWAVE serum HILIC POS dataset. Distribution of pseudogroups size is visualised over four sub-figures to account for very different scales.

Ten most commonly detected and reported CAMERA adduct types are listed in Table 4.3 (Mr Stephane Camuzeaux, personal communication, August 2019). Among those are ions, such as $[M + Na]^+$, $[M + K]^+$ and $[M + H]^+$, that are commonly detected in NPC HILIC assay. However, this list also includes rarely encountered ions, such as $[M + H + NH3]+$ and $[M + H - H20]+$. These results indicate that running CAMERA in an untargeted mode (i.e. without a pre-defined adduct list) can be potentially dangerous and lead to misleading annotations.

TABLE 4.3: AIRWAVE dataset was annotated using XCMS and CAMERA. The 10 most commonly detected and reported adducts are listed. The frequency corresponds to the number of times a given adduct was reported by CAMERA annotation workflow. Top four most frequently reported adducts are also commonly observed in NPC assays (Mr Stephane Camuzeaux, personal communication, August 2019).

Ion	Frequency	Common in NPC assays
[M+K]+	1363	Yes
[M+Na]+	1091	Yes
[M+H]+	841	Yes
[M+Na+NaCOOH]+	238	Yes
[M+Na+HCOOH]+	144	
[M+H+NH3]+	140	
[M+H+HCOOH]+	135	
[M+Na+NH3]+	135	
[M+K+NaCOOH]+	131	
[M+H-H20]+	116	

Annotation validation

Here a set of 46 validated endogenous metabolites, which are commonly detected in serum samples, were used to visualise the accuracy of CAMERA annotation workflow. The *m/z* and RT values for the adducts and in-source fragments corresponding to the validated metabolites are provided in Appendix A. Each metabolite is characterised by one to three adducts and/or in-source fragments.

Features corresponding to the ions of 40 target metabolites were identified in the XCMS output. A summary for CAMERA results is illustrated in Figure 4.4, full details are provided in Table 4.4. While the selected endogenous metabolites had up to three unique ions, XCMS reported duplicated features for some of these ions, as indicated by the number of features per metabolite in Figure 4.4A. For example, two betaine adducts ($[M+H]^+$ and $[M+Na]^+$) were each detected and reported three times (Table 4.4). This and other metabolites with duplicated features represents an example of the ambiguity created by a pre-processing pipeline that treats features as independent entities. It cannot be single-handedly concluded which of these XCMS features correspond to the ions of interest as the raw spectral information is lost during the pre-processing. Next, the number of pseudogroups per metabolite was investigated (Figure 4.4B). 24 out of the 40 detected metabolites had their features assigned to the same pseudogroup, while the other 16 metabolites were grouped either across multiple pseudogroups (e.g. carnitine) or with other metabolites (e.g. urocanate and 5-methylthioadenosine adducts were grouped into the same pseudogroup). These results are particularly worrying as it suggests that the EIC of the adducts/in-source fragments for these metabolites were not found to be correlated in individuals samples, nor were their intensities across all samples. As it

was shown in details in Chapter 3, such correlation structures are highly expected in LC-MS data. A further inconsistency that was noticed in the CAMERA output is the highly varied RT values for features originating from the same metabolite (Figure 4.4C). Nevertheless, this mostly accounts for the multiple pseudogroups into which features of the same metabolite were assigned to (Figure 4.4D). Given that the initial input for the CAMERA algorithm are the RT values reported by XCMS, these results are not surprising. The duplicated XCMS features with varying RT values lead to multiple pseudogroups generation. This resulted in ambiguous annotation and multiple neutral masses for a single metabolite were suggested for many of the investigated metabolites (Table 4.4).

Among the investigated 46 metabolites, six were annotated correctly since their adduct ions were assigned to the correct molecular formulas. For example, features corresponding to creatinine $[M + Na]^+$ and $[2M + H]^+$ ions were both assigned to the correct neutral mass of 113.059. The number of correctly assigned features to the total number of features are provided for these six metabolites below:

- Creatinine (2/2);
- L-Proline (2/3);
- 1,2-Dimyristoyl-sn-glycero-3-phosphocholine (1/1);
- Hypoxanthine (3/3);
- Pipecolate (2/2);
- Inosine (2/2).

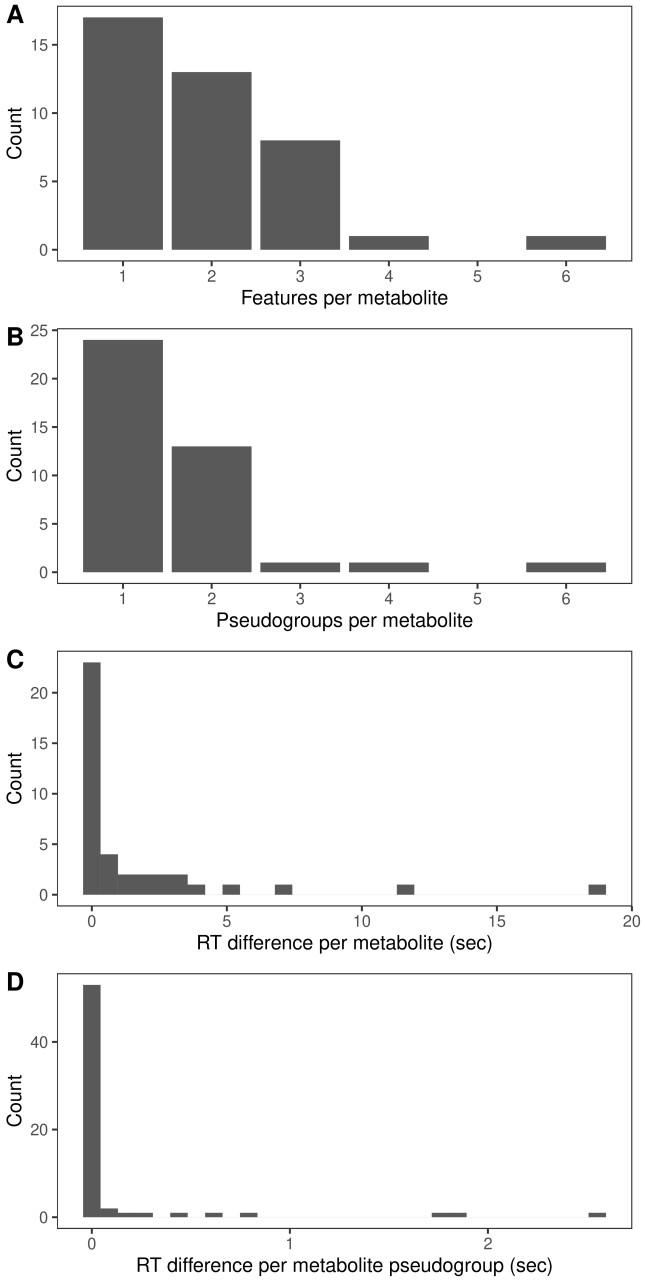


FIGURE 4.4: XCMS features annotation using CAMERA workflow was validated using 46 endogenous metabolites. CAMERA output was summarised to visualise: (A) the number of features per metabolite; (B) the number of pseudogroups into which a given metabolite was grouped; (C) RT difference between the features of a single metabolite; (D) RT difference between the features of a single metabolite grouped to the same pseudogroup.

TABLE 4.4: AIRWAVE dataset was annotated using XCMS and CAMERA. Validated metabolites: their adducts and in-source fragments (ISF) were identified among the reported features. CAMERA output - identified pseudogroups, isotopes and adducts - are provided for each of the feature.

	Chemical standard				CAMERA output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Pseudogroup**	**Isotopes**	**Adduct**
Choline	HPOS-007.1	M	104.1067	239.69	4		
Carnitine	HPOS-008.1	M+H	162.1118	328.16	11	[40][M]+	
	HPOS-008.2	M+Na	184.0937	328.12	1047	[47][M]+	
	HPOS-008.3	ISF	103.0390	330.71	11		[M+H]+ 102.032
Laurylcarnitine	HPOS-013.1	M+H	344.2788	239.14	4	[157][M+2]+	
(C12:0)	HPOS-013.2	M+Na	366.2631	239.25	4		
Histidine	HPOS-014.1	M+H	156.0762	377.20	209		
	HPOS-014.2	M+Na	178.0582	379.53	8415		
	HPOS-014.3	ISF	110.0710	379.88	8414		
N6,N6,N6-	HPOS-016.1	M+H	189.1591	371.83	5496		
Trimethyllysine	HPOS-016.2	ISF	130.0862	360.17	829		
Taurine	HPOS-023.1	M+H	126.0215	160.86	113		[M+3Na-H]2+ 184.09
Paraxanthine	HPOS-024.1	M+H	181.0717	58.51	62	[46][M]+	
Trigonelline	HPOS-025.1	M+H	138.0542	295.58	649	[17][M]+	
Betaine	HPOS-027.1	M+H	118.0857	291.39	2089	[6][M]+	
	HPOS-027.1	M+H	118.0856	280.41	89		
	HPOS-027.1	M+H	118.0856	273.72	93		[M+H-C6H10O5]+ 279.144
	HPOS-027.2	M+Na	140.0675	282.58	7151		
	HPOS-027.2	M+Na	140.0674	273.01	141		
	HPOS-027.2	M+Na	140.0675	291.73	33		
Warfarin	HPOS-031.2	ISF	251.0678	41.65	1483		
	HPOS-031.3	ISF	163.0389	38.52	1730		
Caffeine	HPOS-033.1	M+H	195.0875	52.42	38	[50][M]+	
Niacinamide	HPOS-034.1	M+H	123.0546	65.65	115	[7][M]+	
Creatinine	HPOS-036.1	M+Na	136.0476	154.51	59	[14][M]+	[M+Na]+ 113.059

TABLE 4.4: AIRWAVE dataset was annotated using XCMS and CAMERA. Validated metabolites: their adducts and in-source fragments (ISF) were identified among the reported features. CAMERA output - identified pseudogroups, isotopes and adducts - are provided for each of the feature.

Chemical standard					CAMERA output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Pseudogroup**	**Isotopes**	**Adduct**
	HPOS-036.2	2M+H	227.1249	154.49	59		[2M+H]+ 113.059
1,1-Dimethylbiguanide	HPOS-038.1	M+H	130.1080	210.01	43		
	HPOS-038.2	ISF	113.0815	210.12	2780		
Tryptophan	HPOS-039.1	M+H	205.0968	231.58	37		[M+2H-CO]2+ 436.167
	HPOS-039.2	ISF	188.0705	235.46	10		
Phenylalanine	HPOS-040.1	M+H	166.0854	232.65	2326		
Methionine	HPOS-041.1	M+H	150.0582	256.65	14		[M+Na+K]2+ 238.167
	HPOS-041.1	M+H	150.0581	249.95	97		
	HPOS-041.2	M+2Na-H	194.0224	249.94	5412		
	HPOS-041.2	M+2Na-H	194.0221	256.89	1254		
Trimethylamine N-oxide	HPOS-042.1	M+H	76.0757	255.97	1251		
	HPOS-042.2	2M+H	151.1431	257.03	1252		
L-Proline	HPOS-043.1	M+H	116.0699	272.65	141		[M+H]+ 115.063 [M+H-C4H8]+ 171.127
	HPOS-043.2	M+2Na-H	160.0338	272.84	141		[M+2Na-H]+ 115.063
	HPOS-043.3	2M+H+2HCOONa	365.0667	273.57	93		[M+K]+ 326.102
L-Alanine	HPOS-044.1	M+2Na-H	134.0185	283.18	147		[M+2H-HCOOH]2+ 312.028
	HPOS-044.2	2M+H+2HCOONa	313.0355	282.57	147	[142][M]+	[M+H]+ 312.028
	HPOS-044.3	3M+Na+2HCOONa	356.0775	285.49	42		
Creatine	HPOS-045.1	M+H	132.0760	308.44	25	[13][M]+	
	HPOS-045.2	M+Na	154.0583	309.92	25		[M+2H]2+ 306.104
	HPOS-045.3	M+2Na-H	176.0403	310.32	25		[M+2Na]2+ 306.104
L-Glutamine	HPOS-046.1	M+H	147.0760	314.96	152		
	HPOS-046.2	M+2Na-H	191.0399	314.18	152		[M+2Na]2+ 336.101 [M+H-C6H10O4]+ 336.101
	HPOS-046.3	ISF	130.0494	314.80	152		
L-Citrulline	HPOS-047.2	M+Na	198.0850	341.87	1165		

TABLE 4.4: AIRWAVE dataset was annotated using XCMS and CAMERA. Validated metabolites: their adducts and in-source fragments (ISF) were identified among the reported features. CAMERA output - identified pseudogroups, isotopes and adducts - are provided for each of the feature.

Chemical standard					CAMERA output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Pseudogroup**	**Isotopes**	**Adduct**
Arginine	HPOS-048.1	M+H	175.1185	357.18	8		
	HPOS-048.2	M+2Na-H	219.0877	358.17	840		
Lysine	HPOS-049.1	M+H	147.1128	360.08	8		[M+H]+ 146.107
a-glycerophosphocholine	HPOS-050.1	M+H	258.1100	362.26	8	[101][M+2]2+	
	HPOS-050.2	M+Na	280.0921	362.29	856	[117][M]+	
	HPOS-050.4	ISF	104.1065	361.83	8		
3-methylhistidine	HPOS-051.1	M+H	170.0921	370.06	5497		
N-Acetyl-	HPOS-053.1	M+Na	244.0814	106.47	26		[M+Na+NH3]+ 204.062
D-mannosamine	HPOS-053.1	M+Na	244.0794	111.74	129		
1,2-Dimyristoyl-sn-glycero-3-phosphocholine	HPOS-054.1	M+H	678.5070	250.10	157	[813][M]+	[M+H]+ 677.501
Hypoxanthine	HPOS-055.2	M+Na	159.0273	98.29	71	[38][M]+	[M+Na]+ 136.039
	HPOS-055.3	ISF	119.0347	98.40	71		[M+H-H20]+ 136.039
	HPOS-055.4	ISF	110.0349	98.53	71		[M+2Na+K-H]2+ 136.132 [M+H-C2H4]+ 137.067
Urocanate	HPOS-057.1	M+H	139.0499	81.03	247		
5′-Methylthioadenosine	HPOS-058.1	M+H	298.0954	80.81	247		
Pipecolate	HPOS-061.1	M+H	130.0855	273.65	93	[9][M]+	[M+H]+ 129.078
	HPOS-061.2	M+2Na-H	174.0498	275.45	93		[M+2Na-H]+ 129.078 [M+H]+ 173.043
Thiamine	HPOS-072.1	M+	265.1116	332.88	5859		
4-Guanidinobutanoate	HPOS-073.1	M+H	146.0910	235.96	601		
N,N-Dimethylglycine	HPOS-074.1	M+H	104.0697	282.21	147		[M+2H-C2H4]2+ 234.159
	HPOS-074.2	M+2Na-H	148.0334	282.04	7152		
Inosine	HPOS-079.1	M+Na	291.0699	99.93	32	[126][M]+	[M+Na]+ 268.081
	HPOS-079.2	M+2Na-H	313.0516	99.87	32	[143][M]+	[M+2Na-H]+ 268.081

TABLE 4.4: AIRWAVE dataset was annotated using XCMS and CAMERA. Validated metabolites: their adducts and in-source fragments (ISF) were identified among the reported features. CAMERA output - identified pseudogroups, isotopes and adducts - are provided for each of the feature.

Chemical standard					CAMERA output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Pseudogroup**	**Isotopes**	**Adduct**
Cortisol	HPOS-086.1	M+H	363.2166	44.90	15		[M+K]+ 324.252 [2M+Na+2K]3+ 494.363
1-Methylnicotinamide	HPOS-089.1	M+	137.0702	253.25	44	[16][M]+	[M+Na]+ 114.082
	HPOS-089.2	ISF	94.0650	254.07	2808		
Sucrose	HPOS-091.1	M+Na	365.1052	142.30	197		

XCMS features annotation using feature-to-spectra approach

To enable data annotation to an in-house chemical reference database, a feature-to-spectra matching algorithm was developed (Figure 4.1). The algorithm evaluates the confidence of a match between a feature and a database compound by estimating: (1) the complexity of the compound's spectra; (2) how many features in the dataset correspond to compound's spectra. When combined, these two measures provide a score ranging from zero to one, where one indicates a perfect match (within the limits of the suggested method), as illustrated in Equation 1.

XCMS features obtained for the AIRWAVE HILIC dataset were subjected to feature-to-spectra matching algorithm using a previously built in-house chemical standards database. The database contains 1,782 unique entries across multiple chemical classes, the most common of which are organic acids and derivatives, organoheterocyclic compounds and benzenoids (Figure 4.5 summarises HMDB super-classes distribution for entries with HMDB accession number only). The most abundant class - organic acids - comprises mostly of amino acids, peptides, dicarboxylic acids, short-chain keto acids and their derivative compounds.

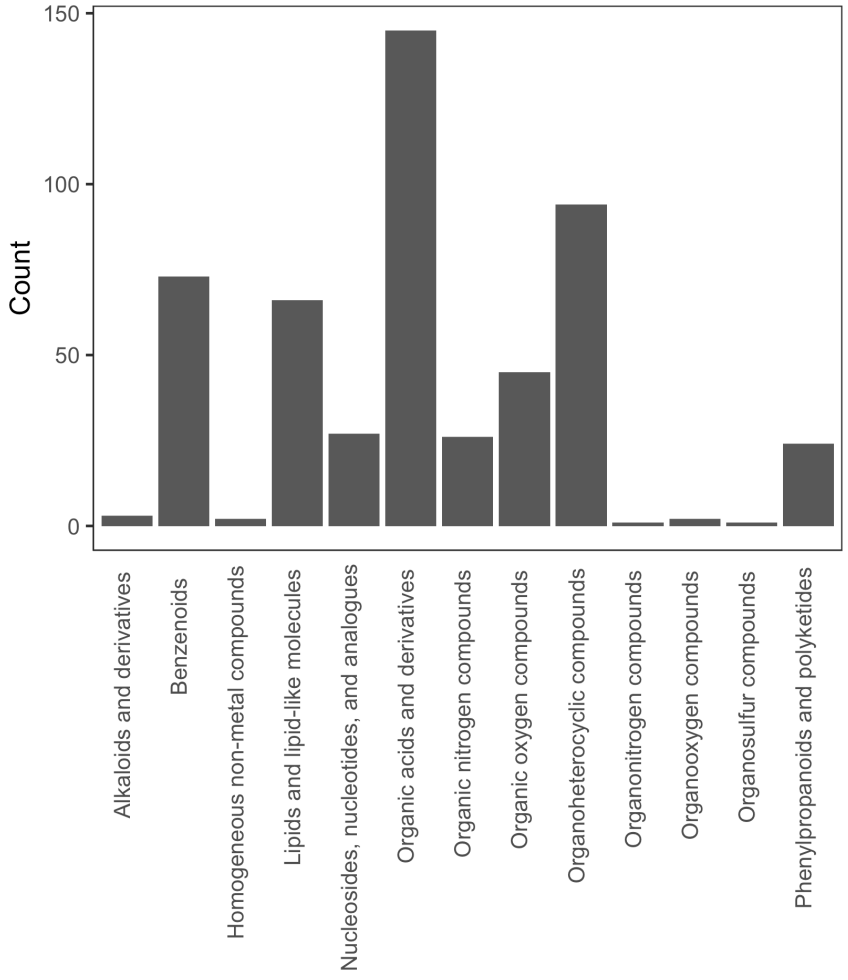


FIGURE 4.5: Distribution of HMDB super-classes for in-house chemical standards database entries with HMDB accession number.

Due to the untargeted approach implemented when acquiring the database, some chemical compounds have multiple entries since more than one different chemical standard was analysed (Figure 4.6). For example, hypoxanthine has four database entries (Figure 4.7) which correspond to four standards acquired from different suppliers.

Feature-to-spectra matching algorithm automatically annotated 6% of all features to a database compound. 45% of annotated features match only a single chemical compound (Figure 4.8), whereas 93% of features match 10 or less different compounds. The relatively low number of annotated compounds per feature suggests that this method provides reasonably unambiguous identifications.

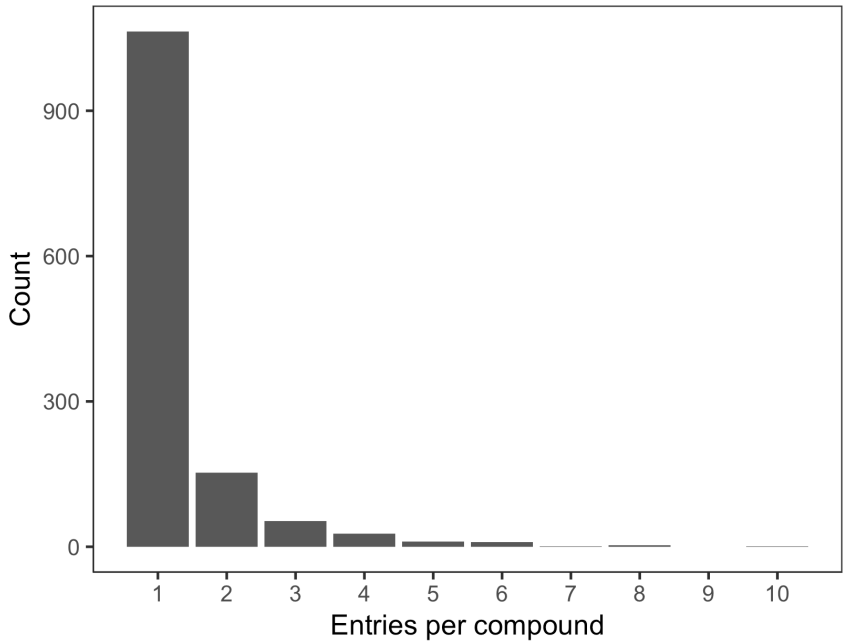


FIGURE 4.6: Some chemical compounds in the in-house reference database have multiple entries since more than one different chemical standard was analysed.

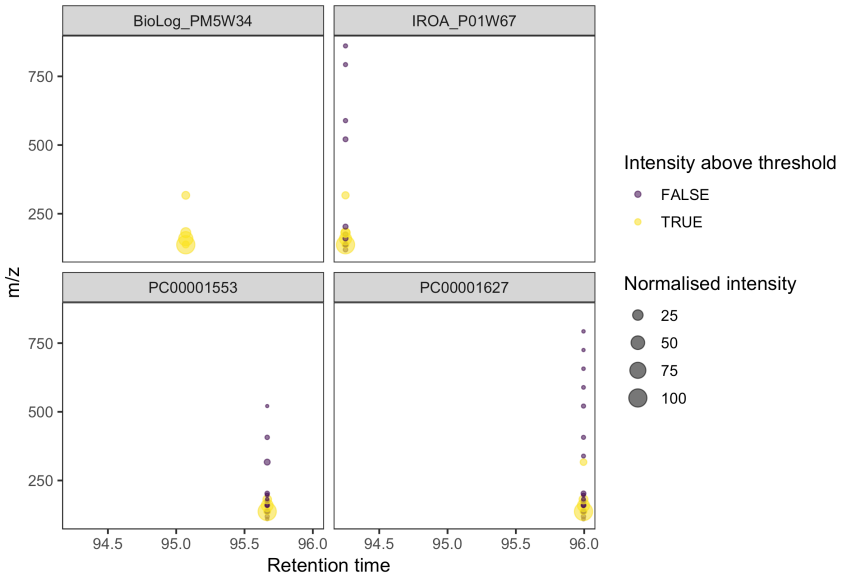


FIGURE 4.7: Four chemical standards of hypoxanthine were acquired with the HILIC LC-MS for the in-house chemical standards database. There is some variation in the retention time and the intensities of the features detected in standards spectra. Purple dots indicate features that were removed from the spectra during database built due to intensity lower than the user-selected threshold, which here is 5 % of the base peak to which every feature is normalised to.

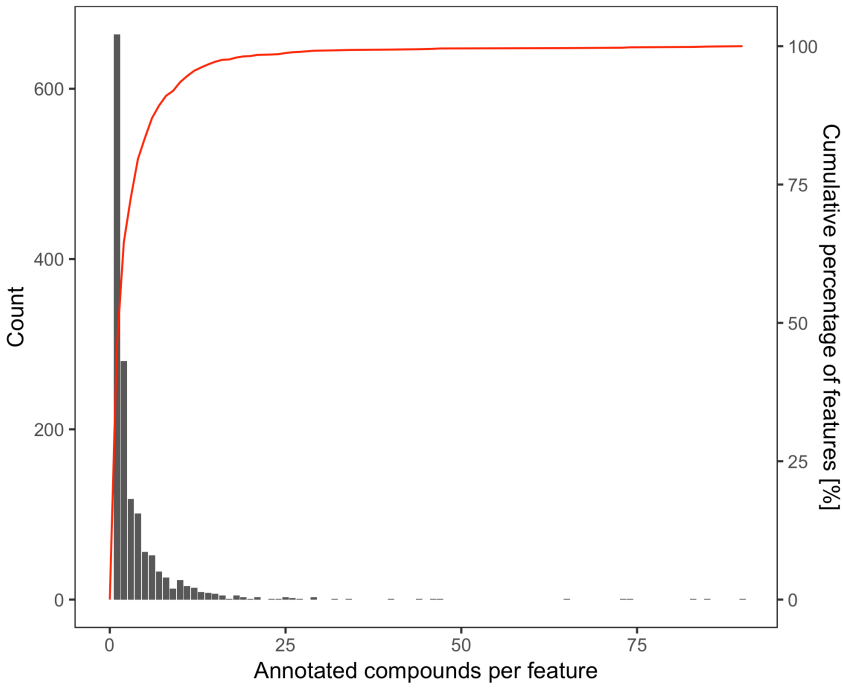


FIGURE 4.8: XCMS features obtained for a HILIC dataset were automatically annotated to an in-house chemical reference database using a feature-to-spectra matching algorithm. The distribution of the number of database compounds per feature illustrates that most features are annotated to a few compounds.

Annotation validation

Similarly as before, the accuracy of the feature-to-spectra matching algorithm was evaluated using 40 endogenous metabolites, identified among the XCMS features, as described in Appendix A. The results of of the feature-to-spectra matching of XCMS features to the standards database are provided in Table 4.5. Annotations for each feature were ranked according to the obtained matching score. If the highest-scoring database compound (i.e. top annotation's rank equals to 1) corresponds to the correct metabolite, then the automatic annotation for the feature was considered correct. 40 validation metabolites were categorised according to how many of their adducts and/or in-source fragment ions were correctly annotated (Table 4.6). 34 metabolites were annotated to the correct chemical standard for all of their features.

TABLE 4.5: AIRWAVE dataset was annotated using XCMS and in-house chemical reference database (DB). XCMS features were matched to the DB chemicals using a novel feature-spectra matching algorithm. Annotation results were validated using validated metabolites - their adducts and in-source fragments (ISF) were identified among XCMS features. Feature-wise annotations were ranked according to the obtained matching score. If the top annotation (i.e. rank = 1) corresponds to the correct metabolite, then algorithm automatically assigned feature to the correct chemical compound.

Chemical standard					Feature-to-spectra matching output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Correct annotation, rank**	**Correct annotation, score**	**Incorrect top annotation**
Choline	HPOS-007.1	M	104.1067	239.69	1	0.656	
Carnitine	HPOS-008.1	M+H	162.1118	328.16	3	0.806	Tripolyphosphate
	HPOS-008.2	M+Na	184.0937	328.12	1	0.806	
	HPOS-008.3	ISF	103.0390	330.71	-	-	*
Laurylcarnitine	HPOS-013.1	M+H	344.2788	239.14	1	0.425	
	HPOS-013.2	M+Na	366.2631	239.25	1	0.425	
Histidine	HPOS-014.1	M+H	156.0762	377.20	1	0.389	
	HPOS-014.2	M+Na	178.0582	379.53	1	0.389	
	HPOS-014.3	ISF	110.0710	379.88	1	0.389	
N6,N6,N6-	HPOS-016.1	M+H	189.1591	371.83	1	1	
Trimethyllysine	HPOS-016.2	ISF	130.0862	360.17	-	-	D-Lysine*
Taurine	HPOS-023.1	M+H	126.0215	160.86	1	0.913	
Paraxanthine	HPOS-024.1	M+H	181.0717	58.51	1	1	
Trigonelline	HPOS-025.1	M+H	138.0542	295.58	1	0.319	
Betaine	HPOS-027.1	M+H	118.0857	291.39	1	1	
	HPOS-027.2	M+Na	140.0675	282.58	1	1	

TABLE 4.5: AIRWAVE dataset was annotated using XCMS and in-house chemical reference database (DB). XCMS features were matched to the DB chemicals using a novel feature-spectra matching algorithm. Annotation results were validated using validated metabolites - their adducts and in-source fragments (ISF) were identified among XCMS features. Feature-wise annotations were ranked according to the obtained matching score. If the top annotation (i.e. rank = 1) corresponds to the correct metabolite, then algorithm automatically assigned feature to the correct chemical compound.

Chemical standard					Feature-to-spectra matching output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Correct annotation, rank**	**Correct annotation, score**	**Incorrect top annotation**
Warfarin	HPOS-031.2	ISF	251.0678	41.65	-	-	
	HPOS-031.3	ISF	163.0389	38.52	1	0.614	
Caffeine	HPOS-033.1	M+H	195.0875	52.42	1	1	
Niacinamide	HPOS-034.1	M+H	123.0546	65.65	1	1	
Creatinine	HPOS-036.1	M+Na	136.0476	154.51	1	1	
	HPOS-036.2	2M+H	227.1249	154.49	1	1	
Metformin	HPOS-038.1	M+H	130.1080	210.01	1	1	
	HPOS-038.2	ISF	113.0815	210.12	1	1	
Tryptophan	HPOS-039.1	M+H	205.0968	231.58	1	0.227	
	HPOS-039.2	ISF	188.0705	235.46	1	0.386	
Phenylalanine	HPOS-040.1	M+H	166.0854	232.65	1	0.844	
Methionine	HPOS-041.1	M+H	150.0582	256.65	1	0.201	
	HPOS-041.2	M+2Na-H	194.0224	249.94	1	0.201	
Trimethylamine-N-oxide	HPOS-042.1	M+H	76.0757	255.97	1	1	
	HPOS-042.2	2M+H	151.1431	257.03	1	1	
Proline	HPOS-043.1	M+H	116.0699	272.65	1	0.936	

TABLE 4.5: AIRWAVE dataset was annotated using XCMS and in-house chemical reference database (DB). XCMS features were matched to the DB chemicals using a novel feature-spectra matching algorithm. Annotation results were validated using validated metabolites - their adducts and in-source fragments (ISF) were identified among XCMS features. Feature-wise annotations were ranked according to the obtained matching score. If the top annotation (i.e. rank = 1) corresponds to the correct metabolite, then algorithm automatically assigned feature to the correct chemical compound.

Chemical standard					Feature-to-spectra matching output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Correct annotation, rank**	**Correct annotation, score**	**Incorrect top annotation**
	HPOS-043.2	M+2Na-H	160.0338	272.84	1	0.936	
	HPOS-043.3	2M+H+2HCOONa	365.0667	273.57	1	0.936	
Alanine	HPOS-044.1	M+2Na-H	134.0185	283.18	1	1	
	HPOS-044.2	2M+H+2HCOONa	313.0355	282.57	1	1	
	HPOS-044.3	3M+Na+2HCOONa	356.0775	285.49	1	1	
Creatine	HPOS-045.1	M+H	132.0760	308.44	2	0.631	Phosphocreatine
	HPOS-045.2	M+Na	154.0583	309.92	2	0.631	Phosphocreatine
	HPOS-045.3	M+2Na-H	176.0403	310.32	2	0.631	Phosphocreatine
Glutamine	HPOS-046.1	M+H	147.0760	314.96	1	1	
	HPOS-046.2	M+2Na-H	191.0399	314.18	1	1	
	HPOS-046.3	ISF	130.0494	314.80	1	1	
Citrulline	HPOS-047.2	M+Na	198.0850	341.87	1	0.19	
Arginine	HPOS-048.1	M+H	175.1185	357.18	1	1	
	HPOS-048.2	M+2Na-H	219.0877	358.17	1	1	
Lysine	HPOS-049.1	M+H	147.1128	360.08	1	0.774	
a-glycerophosphocholine	HPOS-050.1	M+H	258.1100	362.26	2	0.518	1-Oleoyl-sn-glycero-3-phosphocholine

TABLE 4.5: AIRWAVE dataset was annotated using XCMS and in-house chemical reference database (DB). XCMS features were matched to the DB chemicals using a novel feature-spectra matching algorithm. Annotation results were validated using validated metabolites - their adducts and in-source fragments (ISF) were identified among XCMS features. Feature-wise annotations were ranked according to the obtained matching score. If the top annotation (i.e. rank = 1) corresponds to the correct metabolite, then algorithm automatically assigned feature to the correct chemical compound.

Chemical standard					Feature-to-spectra matching output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Correct annotation, rank**	**Correct annotation, score**	**Incorrect top annotation**
	HPOS-050.2	M+Na	280.0921	362.29	2	0.518	1-Oleoyl-sn-glycero-3-phosphocholine
	HPOS-050.4	ISF	104.1065	361.83	2	0.518	1-Oleoyl-sn-glycero-3-phosphocholine
3-methylhistidine	HPOS-051.1	M+H	170.0921	370.06	1	0.202	
N-Acetyl-D-mannosamine	HPOS-053.1	M+Na	244.0814	106.47	2	0.365	N-Acetyl-DGalactosamine
	HPOS-053.1	M+Na	244.0794	111.74	3	0.134	N-Acetyl D-Glucosamine
1,2-Dimyristoyl-sn-glycero-3-phosphocholine	HPOS-054.1	M+H	678.5070	250.10	1	0.584	
Hypoxanthine	HPOS-055.2	M+Na	159.0273	98.29	1	1	
	HPOS-055.3	ISF	119.0347	98.40	-	-	*
	HPOS-055.4	ISF	110.0349	98.53	-	-	*
Urocanate	HPOS-057.1	M+H	139.0499	81.03	1	0.394	
5′-Methylthioadenosine	HPOS-058.1	M+H	298.0954	80.81	1	0.599	
Pipecolate ǀ N-methyl proline	HPOS-061.1	M+H	130.0855	273.65	1	0.196	

TABLE 4.5: AIRWAVE dataset was annotated using XCMS and in-house chemical reference database (DB). XCMS features were matched to the DB chemicals using a novel feature-spectra matching algorithm. Annotation results were validated using validated metabolites - their adducts and in-source fragments (ISF) were identified among XCMS features. Feature-wise annotations were ranked according to the obtained matching score. If the top annotation (i.e. rank = 1) corresponds to the correct metabolite, then algorithm automatically assigned feature to the correct chemical compound.

Chemical standard					Feature-to-spectra matching output		
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**Correct annotation, rank**	**Correct annotation, score**	**Incorrect top annotation**
	HPOS-061.2	M+2Na-H	174.0498	275.45	1	0.196	
Thiamine	HPOS-072.1	M+	265.1116	332.88	2	0.206	Thiamine pyrophosphate
4-Guanidinobutanoate	HPOS-073.1	M+H	146.0910	235.96	1	1	
N,N-Dimethylglycine	HPOS-074.1	M+H	104.0697	282.21	1	0.028	
	HPOS-074.2	M+2Na-H	148.0334	282.04	2	0.028	3- Aminoisobutanoate
Inosine	HPOS-079.1	M+Na	291.0699	99.93	1	0.765	
	HPOS-079.2	M+2Na-H	313.0516	99.87	1	0.765	
Cortisol	HPOS-086.1	M+H	363.2166	44.90	1	0.243	
1-Methylnicotinamide	HPOS-089.1	M+	137.0702	253.25	1	0.489	
	HPOS-089.2	ISF	94.0650	254.07	-	-	*
Sucrose	HPOS-091.1	M+Na	365.1052	142.30	1	0.608	

* Corresponding DB feature was removed from DB due to user-selected intensity threshold

+ Corresponding DB feature was omitted from match because its RT distance to the central RT was more than 2 standard deviations

TABLE 4.6: XCMS features obtained for the AIRWAVE dataset were annotated to the in-house chemical reference database. Features corresponding to the adduct and/or in-source fragment ions of 40 validated metabolites were identified in the automatic annotation output. These metabolites were categorised according to how many of their features were correctly annotated. The number of metabolites that have either all of its features, at least one of its features or none of its features correctly annotated are shown.

	All	**>1**	**None**
Metabolites	34	2	4

4.3.2 Pseudo chemical spectra annotation

AIRWAVE processing with massFlowR

AIRWAVE serum HILIC data was processed with massFlowR using parameters listed in Table 4.2. The size of the obtained PCS is visualised in Figure 4.9. The design of massFlowR algorithm only reports features which are grouped into PCS with at least one more feature and therefore the smallest PCS is comprised of two features. Most of the generated PCS are comprised of two to ten features, similarly as with CAMERA (Figure 4.3). In contrast to CAMERA, which produced pseudogroups comprised of up to 600 features for this dataset, the largest massFlowR PCS contains 42 features. This represents an important difference between CAMERA and massFlowR outputs, as pseudogroups with hundreds of features are unlikely to be useful for annotation purposes.

The obtained pseudo chemical spectra table was further investigated before proceeding to annotation. First, the analytical precision of the obtained features was examined. The median of the relative standard deviation (RSD) values estimated for all features across pooled QC samples is 25.2% (Figure 4.10). The distribution of correlation to dilution coefficients indicates that most of the features respond to dilution well (with a median value of 0.68).

To further assess the quality of the dataset and determine any potential analytical associations with the main sources of variance, multivariate analyses were performed. The scores of the calculated principal components were tested for association with analytical parameters and basic clinical information, such as age, gender and BMI category (Figure 4.11). Raw massFlowR features were indeed highly correlated with run order and MS detector voltage, but this association was removed by the batch correction procedure. On the other hand, significantly strong PCA scores association with sample batch and plate number was not completely removed by the batch correction. Nevertheless, the underlying biological variance, particularly gender and BMI category, was clearly detectable in both batch-corrected and raw datasets. This is also illustrated by the sample clusters the PCA scores plot (Figure 4.12).

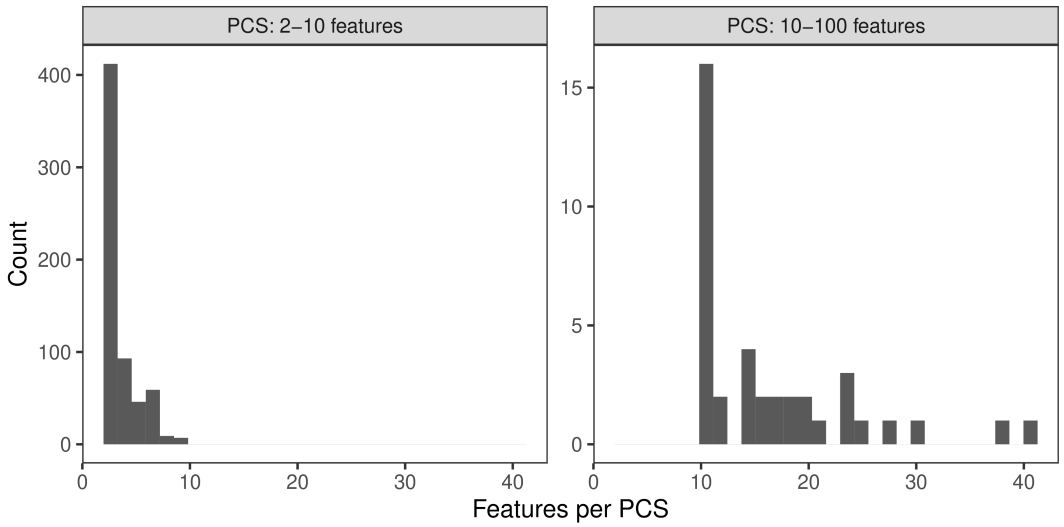


FIGURE 4.9: The number of features per pseudo chemical spectra (PCS), obtained by massFlowR pre-processing applied to AIRWAVE serum HILIC POS dataset. Distribution of PCS size is visualised over two sub-figures to account for very different scales.

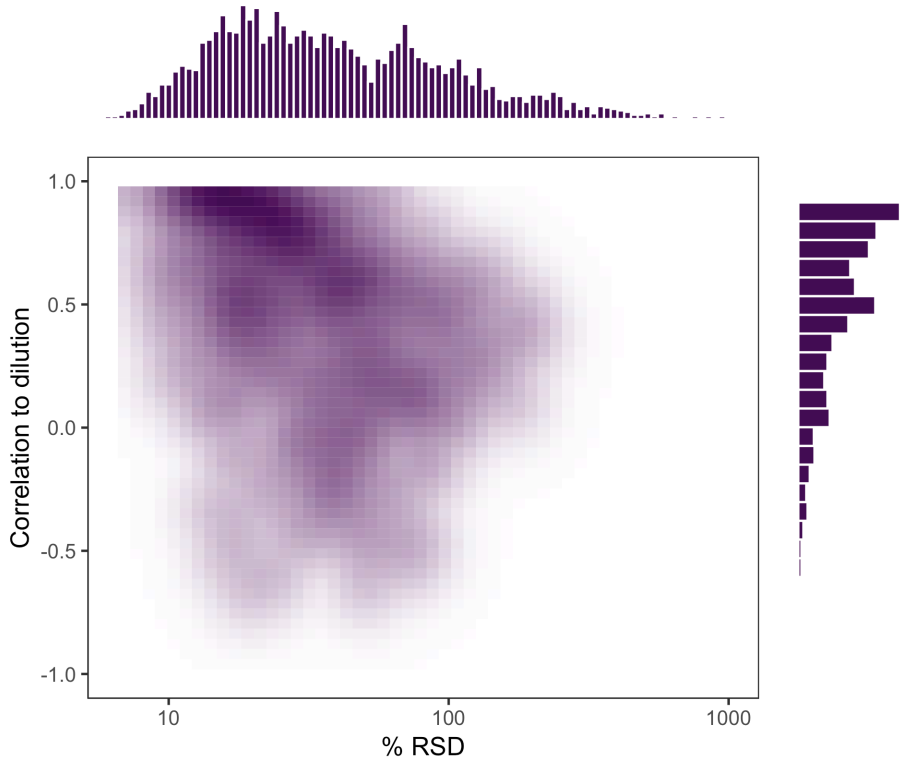


FIGURE 4.10: Analytical precision (relative standard deviation, RSD) and linearity of response (correlation to dilution) of AIRWAVE features reported by massFlowR.

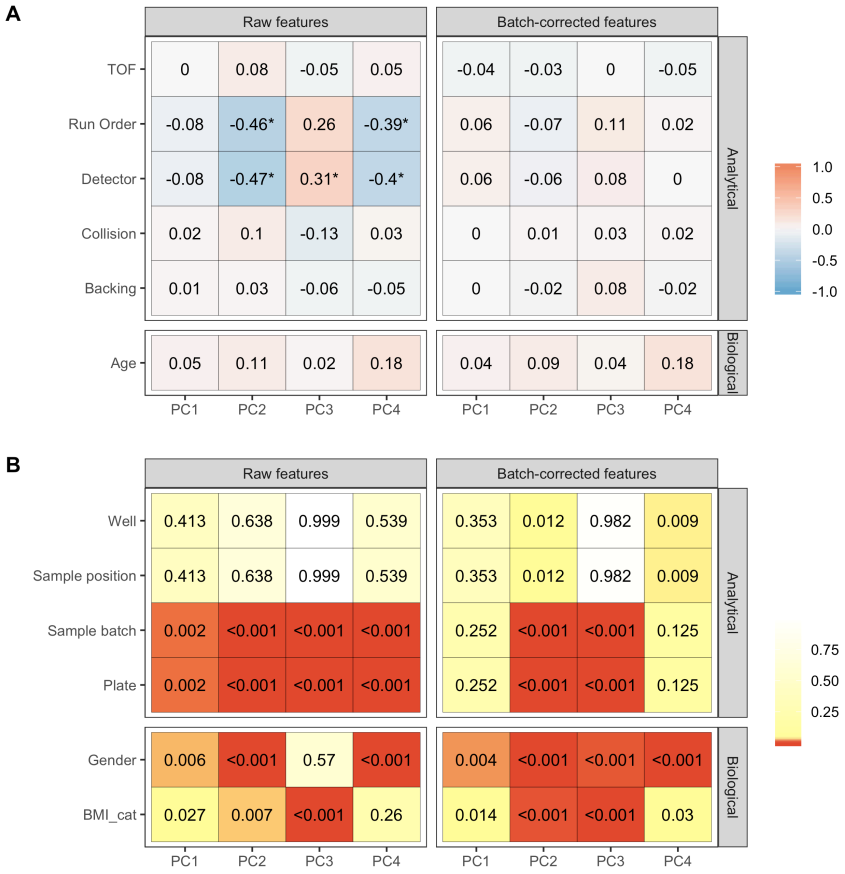


FIGURE 4.11: Principal components scores association with analytical and biological variance in AIRWAVE data generated by massFlowR pre-processing pipeline. Potential associations between the scores of every principal component (PC) and sample metadata was determined by (A) Pearson correlation (continuous data) or (B) Kruskal-Wallis test (categorical data) before and after batch-correction of intensity values. Asterisks denote pairs with correlation coefficient > 0.3 or < -0.3.

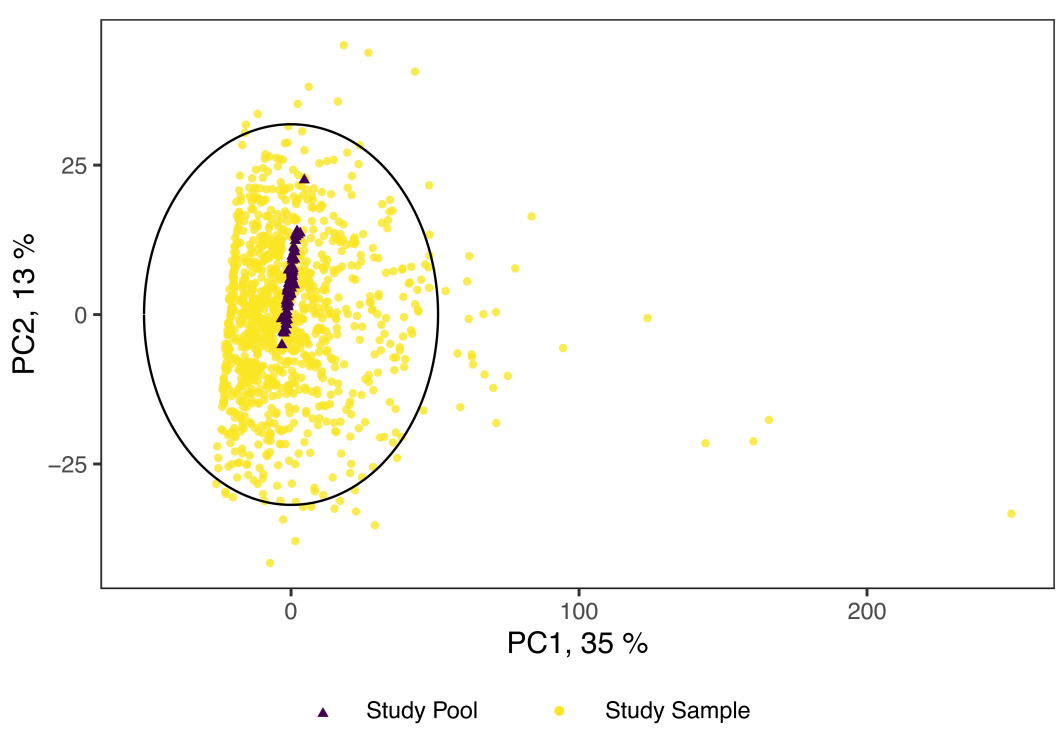


FIGURE 4.12: AIRWAVE samples cluster according to their type in the multivariate space, here illustrated by the scores of the first two principal components. massFlowR generated dataset is visualised.

Annotation validation

Aligned and filled pseudo chemical spectra were subjected to automatic annotations to the in-house standards database, which is described in Section 4.3.1. Ions corresponding to 10 of the 40 endogenous metabolites were identified among the massFlowR generated PCS, as described in Appendix A. The annotation results for these metabolites are provided in Table 4.7.

Some of the metabolite ions were reported by massFlowR more than once, contributing to several PCS. For example, inosine $[M+Na]$ and $[M+2Na-H]$ ions comprise two independent PCS, both of which were correctly annotated. In nine out of ten cases, all of the duplicated PCS were annotated to the correct chemical standard. While carnitine ions were annotated as cinnamic acid, the correct chemical standard was second among the suggested annotations. Most importantly, none of the PCS comprise ions from different compounds.

TABLE 4.7: AIRWAVE dataset was annotated using massFlowR and in-house chemical standards database. Features corresponding to known metabolites' ions were identified among reported features. massFlowR results for each feature were analysed: are ions of the same metabolite assigned to the same pseudo chemical group (PCS) and does top annotation (i.e. rank = 1) correspond to the correct metabolite.

Chemical standard					massFlowR output			
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**PCS**	**Correct annotation, rank**	**Correct annotation, score**	**Incorrect annotation**
Carnitine	HPOS-008.1	M+H	162.1122	327.55	13	5	0.87	Cinnamic Acid
	HPOS-008.2	M+Na	184.0942	327.72	13	2	0.87	Cinnamic Acid
	HPOS-008.1	M+H	162.1118	328.05	173	5	0.89	Cinnamic Acid
	HPOS-008.2	M+Na	184.0937	328.05	173	2	0.89	Cinnamic Acid
Paraxanthine	HPOS-024.1	M+H	181.0718	58.79	273	1	0.96	
	HPOS-024.1	M+H	181.0717	58.79	502	1	0.96	
Betaine	HPOS-027.1	M+H	118.0856	291.25	373	1	0.92	
	HPOS-027.2	M+Na	140.0674	291.25	373	1	0.92	
	HPOS-027.1	M+H	118.0859	291.42	374	1	0.86	
Caffeine	HPOS-033.1	M+H	195.0878	52.39	103	1	0.94	
	HPOS-033.1	M+H	195.0874	52.39	525	1	0.94	
Creatinine	HPOS-036.1	M+Na	136.0475	154.47	108	1	0.87	
	HPOS-036.2	2M+H	227.1249	154.47	108	1	0.87	
	HPOS-036.1	M+Na	136.0475	154.47	356	1	0.98	
	HPOS-036.2	2M+H	227.1248	154.47	356	1	0.98	
	HPOS-036.2	2M+H	227.1251	154.64	48	1	0.84	
Proline	HPOS-043.1	M+H	116.0700	272.72	179	1	0.83	

TABLE 4.7: AIRWAVE dataset was annotated using massFlowR and in-house chemical standards database. Features corresponding to known metabolites' ions were identified among reported features. massFlowR results for each feature were analysed: are ions of the same metabolite assigned to the same pseudo chemical group (PCS) and does top annotation (i.e. rank = 1) correspond to the correct metabolite.

Chemical standard					massFlowR output			
Metabolite	**cpdID**	**Ion**	**m/z**	**RT**	**PCS**	**Correct annotation, rank**	**Correct annotation, score**	**Incorrect annotation**
	HPOS-043.2	M+2Na-H	160.0338	272.72	179	1	0.83	
1,2-Dimyristoyl-sn-glycero-3-phosphocholine	HPOS-054.1	M+H	678.5072	250.15	233	1	0.64	
Hypoxanthine	HPOS-055.2	M+Na	159.0273	98.29	61	1	0.96	
	HPOS-055.2	M+Na	159.0273	98.29	285	1	0.98	
	HPOS-055.2	M+Na	159.0271	98.13	304	1	0.97	
	HPOS-055.2	M+Na	159.0273	98.29	507	1	0.96	
Inosine	HPOS-079.1	M+Na	291.0699	99.98	231	1	0.90	
	HPOS-079.2	M+2Na-H	313.0516	99.98	231	1	0.90	
	HPOS-079.1	M+Na	291.0697	99.98	641	1	0.90	
	HPOS-079.2	M+2Na-H	313.0515	99.98	641	1	0.90	

TABLE 4.8: AIRWAVE dataset was annotated using multiple strategies: (1) XCMS followed by CAMERA; (2) XCMS features matching to an in-house database; (3) massFlowR pseudo chemical spectra annotation to an in-house database. Validation of the annotations was performed using 46 chemical standards. PCS here stands both for CAMERA pseudogroups and pseudo chemical spectra generated by massFlowR.

Tool			Number of metabolites			
	Uses database	**Groups into PCS**	**Total annotated**	**Correctly annotated**	**Ions in same PCS**	**PCS contain other metabolites**
CAMERA	No	Yes	40	-	21, 52.5%	4, 10%
Feature-to-spectra matching	Yes	No	40	34, 85%	-	-
massFlowR	Yes	Yes	10	9, 90%	10, 100%	0

4.4 Conclusions

Automatic metabolite annotations of the AIRWAVE dataset were performed using three approaches: (1) CAMERA annotation of XCMS features; (2) feature-to-spectra matching to a chemical standards database; and (3) massflowR annotation of pseudo chemical spectra to a chemical standards database. In order to validate these different strategies, features corresponding to 46 endogenous metabolites adducts and in-source fragment ions were identified in the processed and annotated datasets. The number of endogenous metabolites which were detected and annotated using these tools are summarised in Table 4.8. The three applied annotation strategies are different in nature. Hoverer, some of the 46 target metabolites could be identified in the datasets generated by all three approaches (40 both for CAMERA and the developed feature-to-spectra matching algorithm, 10 for massFlowR) and thus could be used to compare the tools. Nevertheless, each tool should be evaluated on its own too given its unique attributes.

CAMERA annotation workflow differs from the other two approaches by not relying on any database. Instead, it groups features into pseudo chemical spectra through EIC correlation in individual sample(s). Grouped features are then annotated by identifying potential adducts based on pre-defined rules for mass differences and recognising putative isotopes using KEGG database statistics. Due to the absence of direct database matching, the annotations obtained for the AIRWAVE dataset could not be truly validated. However, most LC-MS researchers do not have access to internally acquired databases and cannot annotate data through direct chromatographic retention time matching. Therefore, other aspects of CAMERA annotation should be considered in more depth instead. 47.5% of the annotated metabolites had their features assigned to multiple pseudogroups and 10% of the metabolites were grouped into pseudogroups with features of other metabolites. Such high assignment ambiguity and duplication may lead to misleading putative annotations and further prolong identification efforts. This is clearly illustrated by the multiple neutral masses suggested for each metabolite (Table 4.4). For example, pipecolate with a molecular weight of 129.157 is assigned to a pseudogroup for which two neutral masses are suggested: 129.078 and 173.043.

In contrast to CAMERA, the other two applied tools were developed to enable data annotation to an existing in-house database. It is important to note that adducts obtained for a given compound through the analysis of a biological sample and a pure authentic standard may be different due to matrix effects. It is widely known that mass spectrometric response for an analyte is different in a biological matrix from its response in a standard solution [197]. Matrix effects result from ion suppression or ion enhancement caused by co-eluting compounds present in a biological sample, as well as impurities associated with sample preparation and mobile phase additives [198]. As the linearity of ESI response can be lost, this phenomena mostly causes errors in the accuracy and precision of bioanalytical methods. Nevertheless, it can also produce different adducts for a given compound, which challenges metabolite annotation procedures, including the spectral matching algorithms developed as part of this thesis. As PCS generated for a metabolite in a biological sample may be comprised of different adducts and/or demonstrate different adduct intensity ratios, it becomes hard to directly compare it with a PCS obtained with a pure standard. Nevertheless, matrix effects are highly varied and unpredictable, therefore, analyses in this thesis were based on a simplified assumption that the PCS from a biological sample and a pure standard should be comparable. A more in-depth investigation on how different these PCS are would be required to further validate the developed automatic annotation procedures.

A feature-to-spectra matching algorithm comparing XCMS features directly to spectra obtained for chemical standards in a database was developed and evaluated. Since it relies on XCMS for feature detection, alignment and filling, the number of total annotated metabolites is the same as in the CAMERA output (Table 4.8). Out of the 40 annotated metabolites, 34 (85%) were annotated to the correct chemical standard (Table 4.6). Only 4 metabolites (10%) had none of their ions correctly annotated.

The final applied approach was massFlowR-generated PCS annotation to an in-house database. Due to the different pre-processing and annotation strategy, massFlowR output differs from XCMS/CAMERA output in many ways. Firstly, fewer validation metabolites are in the final output (Table 4.8). However, out of the 10 annotated metabolites, 9 (90%) were assigned to the correct chemical standards. Furthermore, all of the metabolites had their ions assigned to a single PCS, which also did not include any other metabolites. Nevertheless, some of the metabolites features were reported more than once, contributing to several PCS. Therefore, while massFlowR provides a more accurate annotation method, it is prone to missing features that were not consistently detected and thus were not aligned across samples. In contrast to massFlowR, CAMERA-driven XCMS features annotation detected more metabolites, but is inherently limited by its post-hoc nature and reliance on XCMS reported RT values, which contributes to higher level of annotation ambiguity.

Chapter 5

General discussion

This thesis reflects the growing importance of metabolic profiling in the field of biological and biomedical research. While LC-MS based metabolic profiling is applied to studies of various scale and experimental design, the statistically-powerful studies that are capable of discovering subtle changes between conditions comprise of thousands [49, 51, 52, 199], sometimes even tens of thousands [132] of samples. This thesis is concerned with the analytical and informatics challenges represented by such studies.

5.1 The importance of sensible data processing

The acquisition of large-scale untargeted metabolic profiling data inevitably introduces various types of unwanted analytical variation, as demonstrated in details in Chapter 2 of this thesis. Such variation contributes to a latent noise structure that can conceal subtle, yet meaningful, biological variation. Nevertheless, most common sources of analytical variation, such as chromatographic column ageing or ion source contamination, introduce systematic, rather than random correlations into the noise. For example, samples that are close together will have more similar retention time deviations than those that are acquired further apart. Nevertheless, most of the LC-MS data processing tools available today treat such data correlation structures as random. One of such examples investigated in details throughout this thesis is XCMS feature alignment method "density", which aims to find the corresponding peaks by pooling them from all samples at once. XCMS algorithms have mostly been developed and tested having simpler high performance (HP), rather than ultra-high performance (UP) LC systems in mind. The latter, particularly when combined with high resolution mass spectrometry (HRMS), provides improved sensitivity, generating much more complex spectra with higher information content per sample. Additionally, the computational resources required by the XCMS methods are proportional to study size and quickly outgrow the capabilities of a standard desktop computer. Therefore, while XCMS methods may be well-designed

for small-scale HPLC-MS experiments in which the effect of analytical drift is negligible, they are inherently incapable of processing next generation metabolomics studies that are of interest in this thesis.

We have hypothesised that incorporating the knowledge of a typical noise structure and underlying structural relationships between ions into LC-MS pre-processing algorithms will improve the quality of the generated datasets. In Chapter 3 I introduced an LC-MS pre-processing pipeline that takes into account such data correlation structures through the following four steps (Figure in 3.3):

1. *groupPEAKS* - chromatographic feature grouping into pseudo chemical spectra (PCS) in each LC-MS sample.
2. *alignPEAKS* - feature alignment across samples in original data acquisition order.
3. *validPEAKS* - intensity correlation across samples to identify features that belong to the same PCS.
4. *fillPEAKS* - Missing data points integration using raw LC-MS files.

This pipeline builds on previously developed methods:

- *Chromatographic peak shape correlation* is employed to identify structurally-related chromatographic peaks. This method is implemented in several other annotation tools, such as CAMERA [164] and CliqueMS [166] for the same task.
- *Pearson intensity correlation* here is used to denoise aligned PCS. Pairwise intensity correlation in a generally accepted method to identify features derived from the same compound [169, 170, 192].

Other steps in the developed pipeline are incremental improvements to established practices:

- *Dot product function* here is applied to evaluate the similarity of PCS during sample alignment. Even though dot product similarity score is extensively used in MS/MS spectral matching tools and MS/MS databases [160, 168, 196], such approach has not been used before to compare pseudo-spectra comprised of MS adducts and in-source fragments.
- *Missing data points re-integration* here is performed using sample-specific *m/z* and RT values. The widely accepted pre-processing pipelines, including the gold standard XCMS, fill the intensity values for missing features using *m/z* and RT values averaged across all samples as the integration regions.

Most importantly, the focus on feature groups, in the form of PCS, rather than individual features, is a generally novel approach to LC-MS data processing. Such pipeline represents a paradigm shift from reductionism to a systematic understanding of LC-MS operating principles.

The potential and the limitations of the proposed pipeline were investigated in Chapter 3 using both synthetic and real-world data. Highly promising results were produced in both cases. Firstly, it is worth noting that the developed pipeline is more computationally efficient than XCMS due to multiple facts: (a) each LC-MS spectra is reduced to only its meaningful components, i.e. the PCS, early in the process, with information stored in a simple text file format; (b) PCS alignment does not require all data in the study to be acquired/processed in order to be initiated and can be applied to one sample at a time, as well as re-started at any time. This is a rather unique design that would enable real-time deployment as data is being generated.

The developed pipeline outperformed XCMS when applied to synthetic datasets that had up to 18% of features missing (Figure 3.19). When features were removed at higher rates, the performance recall scores started to drop indicating that the current algorithm tends to sacrifice recall over precision. Only the most similar PCS are grouped together, whereas PCS with deviated features are added to the template as new PCS. This represents a limitation that is unlikely to be solved using the currently employed untargeted approach. Originally an unrestricted feature alignment method that would detect and report as many metabolites as possible was desired. Nevertheless, alignment precision was higher for the developed massFlowR pipeline than XCMS in all tested datasets. Furthermore, massFlowR application to multiple metabolomics datasets indicated that massFlowR accurately captures the underlying sources of variance, such as the expected time-dependent intensity drift, as well as biological sample origin (Figures 3.24, 4.12).

5.2 The utility of annotatable data

Metabolite identification represents one of the greatest hurdles in the field of metabolomics. Due to the vast chemical space covered by molecules involved in metabolic reactions, every unknown LC-MS spectral feature can be matched to hundreds, if not thousands, of potential chemical formulas given just the measured m/z value. While multiple strategies can be taken to reduce the search space, for example, by acquiring MS/MS fragmentation data, which reveals more information about the chemical structure of the unknown metabolite, we have suggested to make better use of the MS data instead. By shifting the focus from individual spectral features to spectra components (i.e. PCS comprised of structurally related chromatographic peaks), we aimed to make data directly annotatable.

The utility of LC-MS data processed in such a novel manner was investigated in Chapter 4. The obtained results indicate that high annotation accuracy can be achieved when suggested processing pipeline is used together with an in-house chemical standards database. Furthermore, even if such a database is not available, the processed datasets contain the chemical information that was originally embedded in the raw LC-MS spectra. By contrast to standard processing software, such

as XCMS, the proposed pipeline relies on the underlying chemical relationships between adducts and in-source fragment ions. The knowledge of such ion-to-ion correlations, preserved during processing, makes the final dataset directly annotatable since critically important structural information is provided for each feature.

It is important to acknowledge, however, that the success of the automatic annotations enabled by the pipeline depends on the alignment step. While it was demonstrated that current alignment algorithm is highly precise, it tends to produce lower recall values for datasets with a high proportion of missing values, which was also the case for the cohort study investigated in Chapter 4. Nevertheless, even in such cases, the generated datasets provide more information on each feature than XCMS as every one of them is grouped into PCS with structurally-related ions. Since the *m/z* differences between features in a given PCS can be examined, a list of potential molecular formulas can be determined by an experienced analyst as part of a manual metabolite annotation and identification pipeline, accelerating the overall process.

5.3 Wider scope

There are several directions which could be followed to continue the work presented in this thesis.

In order to improve the performance of the developed pipeline, a few changes could be implemented and evaluated. First of all, alternative spectral scaling of PCS could be employed. The evaluation of three scaling methods in Chapter 3, Section 3.3.1, indicated that performance of two of them - no-scaling and square-root scaling - varied a lot between different metabolites. While it was decided to continue with square-root scaling, which is a more widely accepted method for spectral searching tools, it would be beneficial to further investigate whether no-scaling method could improve the performance of the PCS alignment algorithm. Next, the feasibility of restricting the growth of the template during the alignment of a metabolic profiling study could be evaluated. Currently employed algorithm allows user to select a threshold for the cosine similarity score for aligning matching PCS (*cutoff*). While in this thesis, a generally low *cutoff* value was selected, such as 0.3 for the synthetic datasets and DEVSET study and 0 for the AIRWAVE study, a further sensitivity analysis on this parameter would be beneficial. To begin with, simulated datasets could be aligned using a range of *cutoff* values to identify whether precision and recall values change.

In order to encourage the use of massFlowR within a broader metabolomics community, several steps could be taken. While it already is an attractive alternative to XCMS due to its low computational resources requirements, its ability to annotate the generated PCS relies on the use of in-house chemical standards databases. With a growing understanding of the importance of data FAIRness (findability, accessibility, interoperability and reusability) [200], more and more datasets are being

released to open-source databases as part of the standard publication process, such as on MetaboLights [201]. Publishing not just the study data, but also the data that was used for metabolite annotation would be the next step in terms of improving the FAIRness of metabolomics data. Therefore, it is highly likely that the in-house databases generated at the NPC and used for the work presented in this thesis will be published in the future, making massFlowR a more attractive pre-processing tool than XCMS due to its automated annotation functionality.

Another route for further development includes a move towards a more targeted pre-processing strategy. The developed pipeline could potentially be altered to align features in a study sample using the preceding pooled quality control (QC) sample as the template. Such alignment procedure would remove metabolites not present in the preceding QC sample. Such a change in the algorithm design may improve the alignment performance as fewer PCS would be compared in each round of alignment. This alternative approach would remove metabolites unique to a given sample, nevertheless, a similar filtering effect is already achieved by using the standardised post-processing QC procedures, discussed in Chapter 2. The most commonly applied QC procedures remove features that do not meet the quality criteria estimated using the measurements in the pooled QC samples. A more targeted PCS alignment design may generate datasets that require less post-processing filtering and correction and thus is likely to be accepted within the field.

Finally, massFlowR ability to process samples during sample acquisition can be extended even further to perform real-time QC monitoring. Usually QC procedures are applied once all of the data is acquired - samples are pre-processed together and generated datasets are subjected to analyses, such as PCA decomposition, in order to detect any underlying trends in the data. If the analytical system experienced issues that would require to re-run some of the study samples, it would only be determined long after the data was actually acquired. Such samples then would be re-analysed as part of a separate LC-MS experiment, which could potentially lead to batch effects, as discussed in Chapter 2. An alternative approach would be to monitor changes in the signal intensity and chromatographic RT in real-time by observing the deviation between adjacent QC samples. Real-time tracking of generated PCS and how well they match between subsequent samples could help identify if chromatographic drift or signal intensity drift have gone beyond the acceptable range. Such information would help to decide whether a sample has to be re-analysed while the given LC-MS experiment is still running. Real-time deployment thus represents a very important aspect for further development that is also likely to raise the standards of data quality in the field.

5.4 Concluding remarks

To conclude, within this thesis I have developed a novel untargeted LC-MS data processing pipeline, which aids metabolite identification through deliberate use of spectral features correlation structure, chromatographic profile and data acquisition order. Since the focus has been set towards real-world applicability, in order to test and validate the suggested approach, I applied it to synthetic data, as well as an open-source dataset and a large-scale cohort study. These real-world studies analysed different biological samples - urine and blood plasma, each of which represents different analytical and informatics challenges.

Overall, the findings of this thesis imply that the proposed approach holds potential for applications in the field of metabolomics. The approach could be further developed towards a more targeted metabolic profiling strategy, for example, by retaining only the PCS that are detected in pooled QC samples or in external reference samples. Such samples could be extensively characterised and annotated, allowing to not only automatically annotate the untargeted data obtained for study samples, but also to integrate data produced for different studies. Nevertheless, metabolomics studies have historically focused on untargeted measurements in order to capture the full set of metabolites present within an organism/tissue/cell at a given time. As we still need to improve the coverage of metabolism, untargeted approaches will play an important role in the field. In order to achieve this, current strategies for untargeted data pre-processing and annotation must be improved. This thesis demonstrates the potential of a spectral-knowledge driven pre-processing pipeline. By enhancing the quality of each of the generated datasets and making them more directly annotatable, we improve our knowledge about each of the sample in a study. Consequently, important conclusions about relationships between metabolic changes and health and disease can be drawn.

Bibliography

1. Tavassoly, I., Goldfarb, J. & Iyengar, R. Systems biology primer: The basic methods and approaches. *Essays in Biochemistry* **62,** 487–500. ISSN: 00711365 (2018).
2. Fearnley, L. G. & Inouye, M. Metabolomics in epidemiology: from metabolite concentrations to integrative reaction networks. *International Journal of Epidemiology*, dyw046. ISSN: 0300-5771. `http://ije.oxfordjournals.org/lookup/doi/10.1093/ije/dyw046` (2016).
3. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics* **2,** 343–372. ISSN: 0163-7525. `https://doi.org/10.1146/annurev.genom.2.1.343` (2001).
4. Neves, S. R. & Iyengar, R. Modeling of signaling networks. *BioEssays* **24,** 1110–1117. ISSN: 02659247 (2002).
5. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications* **9.** ISSN: 20411723. `http://dx.doi.org/10.1038/s41467-017-02391-6` (2018).
6. Van Der Graaf, P. H. & Benson, N. Systems pharmacology: Bridging systems biology and Pharmacokinetics- Pharmacodynamics (PKPD) in drug discovery and development. *Pharmaceutical Research* **28,** 1460–1464. ISSN: 07248741 (2011).
7. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934. ISSN: 00368075 (2001).
8. Bruggeman, F. J. & Westerhoff, H. V. The nature of systems biology. *Trends in Microbiology* **15,** 45–50. ISSN: 0966842X (2007).
9. Oliver, S. G., Winson, M. K., Kell, D. B. & Baganz, F. Systematic functional analysis of the yeast genome. *Trends in Biotechnology* **16,** 373–378. ISSN: 01677799 (1998).
10. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355. ISSN: 14764687 (2016).
11. Fiehn, O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics* **2,** 155–168. ISSN: 15316912 (2001).
12. Fiehn, O. Metabolomics - The link between genotypes and phenotypes. *Plant Molecular Biology* **48,** 155–171. ISSN: 01674412 (2002).

13. Weckwerth, W. & Morgenthal, K. Metabolomics: from pattern recognition to biological interpretation. *Drug Discovery Today* **10,** 1551–1558. ISSN: 1359-6446 (2005).
14. Kirchmair, J. *et al.* How do metabolites differ from their parent molecules and how are they excreted? *Journal of Chemical Information and Modeling* **53,** 354–367. ISSN: 15499596 (2013).
15. Mamas, M., Dunn, W. B., Neyses, L. & Goodacre, R. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of Toxicology* **85,** 5–17. ISSN: 03405761 (2011).
16. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research* **46,** D608–D617. ISSN: 13624962 (2018).
17. Weckwerth, W. Metabolomics in Systems Biology. *Annual Review of Plant Biology* **54,** 669–689. ISSN: 1543-5008 (2003).
18. Bhalla, R., Narasimhan, K. & Swarup, S. Metabolomics and its role in understanding cellular responses in plants. *Plant Cell Reports* **24,** 562–571. ISSN: 07217714 (2005).
19. Vogt, T. Phenylpropanoid biosynthesis. *Molecular Plant* **3,** 2–20. ISSN: 17529867. `http://dx.doi.org/10.1093/mp/ssp106` (2010).
20. Cullis, P. R. & de Kruijff, B. Lipid Polymorphism and the Functional Roles of Lipids in. *Biochimica et Biophysica Acta* **559,** 399–420 (1979).
21. Dixon, R. A. & Paiva, N. L. Stress-induced phenylpropanoid metabolism. *Plant Cell* **7,** 1085–1097. ISSN: 10404651 (1995).
22. Gribble, F. M. & Reimann, F. Signalling in the gut endocrine axis. *Physiology and Behavior* **176,** 183–188. ISSN: 1873507X. `http://dx.doi.org/10.1016/j.physbeh.2017.02.039` (2017).
23. Obeid, R. The metabolic burden of methyl donor deficiency with focus on the betaine homocysteine methyltransferase pathway. *Nutrients* **5,** 3481–3495. ISSN: 20726643 (2013).
24. Petersen, A. K. *et al.* Epigenetics meets metabolomics: An epigenome-wide association study with blood serum metabolic traits. *Human Molecular Genetics* **23,** 534–545. ISSN: 14602083 (2014).
25. Nicholson, J. K.; Wilson, I. D. Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discovery.* **2,** 668–676. ISSN: 1474-1776 (2003).
26. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted Metabolomics StrategiesChallenges and Emerging Directions. *Journal of the American Society for Mass Spectrometry* **27,** 1897–1905. ISSN: 18791123. `http://dx.doi.org/10.1007/s13361-016-1469-y` (2016).
27. Benton, H. P. *et al.* Autonomous Metabolomics for Rapid Metabolite Identification in Global Profiling. *Analytical Chemistry* **87,** 884–891. ISSN: 0003-2700. `http://pubs.acs.org/doi/10.1021/ac5025649` (Jan. 2015).

28. Putri, S. P., Yamamoto, S., Tsugawa, H. & Fukusaki, E. Current metabolomics: Technological advances. *Journal of Bioscience and Bioengineering* **116,** 9–16. ISSN: 13891723. http://dx.doi.org/10.1016/j.jbiosc.2013.01.004 (2013).
29. Nicholson, J. K. & Lindon, J. C. Systems Biology - Metabonomics. *Nature* **455,** 1054–1056. ISSN: 1742-464X. http://www.ncbi.nlm.nih.gov/pubmed/21083101 (2008).
30. Keun, H. C. *et al.* Analytical reproducibility in 1H NMR-based metabonomic urinalysis. *Chemical Research in Toxicology* **15,** 1380–1386. ISSN: 0893228X (2002).
31. Ravanbakhsh, S. *et al.* Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE* **10,** 1–15. ISSN: 19326203. arXiv: 1409.1456 (2015).
32. Büscher, J. M., Czernik, D., Ewald, J. C., Sauer, U. & Zamboni, N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Analytical Chemistry* **81,** 2135–2143. ISSN: 00032700 (2009).
33. Lenz, E. M. & Wilson, I. D. Analytical strategies in metabonomics. *Journal of Proteome Research* **6,** 443–458. ISSN: 15353893 (2007).
34. Theodoridis, G., Gika, H. G. & Wilson, I. D. LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics. *TrAC - Trends in Analytical Chemistry* **27,** 251–260. ISSN: 01659936 (2008).
35. Plumb, R. S. *et al.* Metabonomics: The use of electrospray mass spectrometry coupled to reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development. *Rapid Communications in Mass Spectrometry* **16,** 1991–1996. ISSN: 09514198 (2002).
36. Idborg-Björkman, H., Edlund, P. O., Kvalheim, O. M., Schuppe-Koistinen, I. & Jacobsson, S. P. Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. *Analytical Chemistry* **75,** 4784–4792. ISSN: 00032700 (2003).
37. Lafaye, A. *et al.* Metabolite profiling in rat urine by liquid chromatography/electrospray ion trap mass spectrometry. Application to the study of heavy metal toxicity. *Rapid Communications in Mass Spectrometry* **17,** 2541–2549. ISSN: 09514198 (2003).
38. Chen, M. *et al.* Metabonomic study of aristolochic acid-induced nephrotoxicity in rats. *Journal of Proteome Research* **5,** 995–1002. ISSN: 15353893 (2006).
39. Yang, J. *et al.* Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. *Journal of Chromatography B* **813,** 59–65. ISSN: 15700232 (2004).
40. Zhou, L. *et al.* Serum metabolomics reveals the deregulation of fatty acids metabolism in hepatocellular carcinoma and chronic liver diseases. *Analytical and Bioanalytical Chemistry* **403,** 203–213. ISSN: 16182642 (2012).
41. Ma, C. *et al.* Serum and kidney metabolic changes of rat nephrotoxicity induced by Morning Glory Seed. *Food and Chemical Toxicology* **48,** 2988–2993. ISSN: 02786915. http://dx.doi.org/10.1016/j.fct.2010.07.038 (2010).

42. Loftus, N. *et al.* Profiling and biomarker identification in plasma from different Zucker rat strains via high mass accuracy multistage mass spectrometric analysis using liquid chromatography/mass spectrometry with a quadrupole ion trap-time of flight mass spectrometer. *Rapid Communications in Mass Spectrometry* **22,** 2547–2554. ISSN: 09514198. http://doi.wiley.com/10.1002/rcm.3640 (Aug. 2008).
43. Want, E. J. *et al.* Global metabolic profiling of animal and human tissues via UPLC-MS. *Nature Protocols* **8,** 17–32. ISSN: 1754-2189. http://www.nature.com/articles/nprot.2012.135 (Jan. 2013).
44. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457,** 910–914. ISSN: 14764687 (2009).
45. Kaur, P. *et al.* Metabolomic profiling for biomarker discovery in pancreatic cancer. *International Journal of Mass Spectrometry* **310,** 44–51. ISSN: 13873806. http://dx.doi.org/10.1016/j.ijms.2011.11.005 (2012).
46. Hilvo, M. *et al.* Novel theranostic opportunities offered by characterization of altered membrane lipid metabolism in breast cancer progression. *Cancer Research* **71,** 3236–3245. ISSN: 00085472 (2011).
47. Tzoulaki, I., Ebbels, T. M. D., Valdes, A., Elliott, P. & Ioannidis, J. P. A. Design and analysis of metabolomics studies in epidemiologic research: A primer on-omic technologies. *American Journal of Epidemiology* **180,** 129–139. ISSN: 14766256 (2014).
48. Swann, J. R. *et al.* Microbial-mammalian cometabolites dominate the age-associated urinary metabolic phenotype in Taiwanese and American populations. *Journal of Proteome Research* **12,** 3166–3180. ISSN: 15353893 (2013).
49. Ding, M. *et al.* Metabolome-Wide Association Study of the Relationship Between Habitual Physical Activity and Plasma Metabolite Levels. *American journal of epidemiology* **188,** 1932–1943. ISSN: 14766256 (2019).
50. Acar, E. *et al.* Biomarkers of Individual Foods, and Separation of Diets Using Untargeted LCMS-based Plasma Metabolomics in a Randomized Controlled Trial. *Molecular Nutrition and Food Research* **63,** 1–10. ISSN: 16134133 (2019).
51. Wang, T. J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nature Medicine* **17,** 448–453. ISSN: 1078-8956. http://www.nature.com/doifinder/10.1038/nm.2307 (2011).
52. Ganna, A. *et al.* Large-scale Metabolomic Profiling Identifies Novel Biomarkers for Incident Coronary Heart Disease. *PLoS Genetics* **10.** ISSN: 15537404 (2014).
53. Chadeau-Hyam, M. *et al.* Metabolic profiling and the metabolome-wide association study: Significance level for biomarker identification. *Journal of Proteome Research* **9,** 4620–4627. ISSN: 15353893 (2010).
54. Blaise, B. J. *et al.* Power Analysis and Sample Size Determination in Metabolic Phenotyping. *Analytical Chemistry* **88,** 5179–5188. ISSN: 15206882 (2016).

55. Alonso, A., Marsal, S. & JuliÃ, A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Frontiers in Bioengineering and Biotechnology* **3,** 1–20. ISSN: 2296-4185. http://www.frontiersin.org/Bioinformatics%7B%5C_%7Dand%7B%5C_%7DComputational%7B%5C_%7DBiology/10.3389/fbioe.2015.00023/abstract (2015).
56. Johnson, C. H., Ivanisevic, J., Benton, H. P. & Siuzdak, G. Bioinformatics: The next frontier of metabolomics. *Analytical Chemistry* **87,** 147–156. ISSN: 15206882 (2015).
57. Lange, E., Tautenhahn, R., Neumann, S. & Gröpl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **9,** 375. ISSN: 1471-2105. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-375 (Dec. 2008).
58. Howard, G. A. & Martin, A. J. P. The separation of the C6-C12 fatty acids by reversed-phase partition chromatography. *The Biochemical journal* **46,** 532–538 (1950).
59. Molnar, I. & Horvath, C. Reverse phase chromatography of polar biological substances: separation of catechol compounds by high performance liquid chromatography. *Clinical Chemistry* **22,** 1497–1502. ISSN: 00099147 (1976).
60. Guy, P. A., Tavazzi, I., Bruce, S. J., Ramadan, Z. & Kochhar, S. Global metabolic profiling analysis on human urine by UPLC-TOFMS: Issues and method validation in nutritional metabolomics. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **871,** 253–260. ISSN: 15700232 (2008).
61. Hodson, M. P., Dear, G. J., Griffin, J. L. & Haselden, J. N. An approach for the development and selection of chromatographic methods for high-throughput metabolomic screening of urine by ultra pressure LC-ESI-ToF-MS. *Metabolomics* **5,** 166–182. ISSN: 15733882 (2009).
62. Lewis, M. R. *et al.* Development and Application of UPLC-ToF MS for Precision Large Scale Urinary Metabolic Phenotyping. *Analytical Chemistry* **88,** 9004–9013. ISSN: 0003-2700. http://pubs.acs.org/doi/abs/10.1021/acs.analchem.6b01481 (2016).
63. Zelena, E. *et al.* Development of a robust and repeatable UPLC - MS method for the long-term metabolomic study of human serum. *Analytical Chemistry* **81,** 1357–1364. ISSN: 00032700 (2009).
64. Want, E. J. *et al.* Global metabolic profiling procedures for urine using UPLC-MS. *Nature Protocols* **5,** 1005–1018. ISSN: 17502799 (2010).
65. Alpert, A. J. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of Chromatography A* **499,** 177–196. ISSN: 00219673 (1990).

66. Buszewski, B. & Noga, S. Hydrophilic interaction liquid chromatography (HILIC)-a powerful separation technique. *Analytical and Bioanalytical Chemistry* **402,** 231–247. ISSN: 16182642 (2012).
67. New, L. S. & Chan, E. C. Evaluation of BEH C18, BEH HILIC, and HSS T3 (C18) column chemistries for the UPLC-MS-MS analysis of glutathione, glutathione disulfide, and ophthalmic acid in mouse liver and human plasma. *Journal of Chromatographic Science* **46,** 209–214. ISSN: 00219665 (2008).
68. Cubbon, S., Antonio, C., Wilson, J. & Thomas-Oates, J. Metabolomic applications of HILIC- LC-MS. *Mass spectrometry reviews* **29,** 671–684 (2010).
69. Greco, G. & Letzel, T. Main interactions and influences of the chromatographic parameters in HILIC separations. *Journal of Chromatographic Science* **51,** 684–693. ISSN: 00219665 (2013).
70. Tang, D., Zou, L., Yin, X. & Ong, C. HILICMS for metabolomics: An attractive and complementary approach to RPLCMS. *Mass spectrometry reviews* **35,** 574–600 (2016).
71. Wilson, I. D. *et al.* High Resolution Ultra Performance Liquid Chromatography Coupled to oa-TOF Mass Spectrometry as a Tool for Differential Metabolic Pathway Profiling in Functional Genomic Studies. *Journal of Proteome Research* **4,** 591–598. ISSN: 1535-3893. `http://pubs.acs.org/doi/abs/10.1021/pr049769r` (Apr. 2005).
72. Martin, A. J. & Synge, R. L. A new form of chromatogram employing two liquid phases. 1. A theory of chromatography 2. Application to the microdetermination of the higher monoamino-acids in proteins. *Biochemical Journa* **35,** 1358–1368 (1941).
73. Van Deemter, J., Zuiderweg, F. & Klinkenberg, A. Longitudinal diffusion and resistance to mass transfer as causes of non ideality in chromatography. *Chemical Engineering Science* **5,** 271–289 (1956).
74. Watson, J. Throck & Sparkman, O. D. *Introduction to Mass Spectrometry: Instrumentation, Applications and Strategies for Data Interpretation* 4th (Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2007).
75. Want, E. J., Cravatt, B. F. & Siuzdak, G. The expanding role of mass spectrometry in metabolite profiling and characterization. *ChemBioChem* **6,** 1941–1951. ISSN: 14394227 (2005).
76. Kebarle, P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *Journal of Mass Spectrometry* **35,** 804–817. ISSN: 10765174 (2000).
77. Gao, S., Zhang, Z. P. & Karnes, H. T. Sensitivity enhancement in liquid chromatography/atmospheric pressure ionization mass spectrometry using derivatization and mobile phase additives. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **825,** 98–110. ISSN: 15700232 (2005).

78. Gaskell, S. J. Electrospray : Principles and Practice. *Journal of Mass Spectrometry* **32,** 677–688 (1997).
79. Rathahao-Paris, E., Alves, S., Junot, C. & Tabet, J.-C. High resolution mass spectrometry for structural identification of metabolites in metabolomics. *Metabolomics* **12,** 10. ISSN: 1573-3882. `http://link.springer.com/10.1007/s11306-015-0882-8` (Jan. 2016).
80. Raftery, D. (*Mass Spectrometry in Metabolomics. Methods and Protocols* 1st ed. 20 (New York, NY : Springer New York : Imprint: Humana, 2014).
81. Balogh, M. P. Debating Resolution and Mass Accuracy in Mass Spectrometry. *Spectroscopy* **19,** 34–40 (2004).
82. Makarov, A. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry* **72,** 1156–1162. ISSN: 00032700 (2000).
83. Comisarow, M. B. & Marshall, A. G. Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters* **25,** 282–283. ISSN: 00092614 (1974).
84. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Analytical Chemistry* **85,** 5288–5296. ISSN: 00032700 (2013).
85. Dodonov, A. F., Chernushevich, I. V. & Laiko, V. V. Electrospray Ionization on a Reflecting Time-of-Flight Mass Spectrometer. *American Chemical Society Symposium Series,* 108–123 (1994).
86. Wiley, W. C. & McLaren, I. H. Time-of-flight mass spectrometer with improved resolution. *Review of Scientific Instruments* **26,** 1150–1157. ISSN: 00346748 (1955).
87. Rousu, T., Herttuainen, J. & Tolonen, A. Comparison of triple quadrupole, hybrid linear ion trap triple quadrupole, time-of-flight and LTQ-Orbitrap mass spectrometers in drug discovery phase metabolite screening and identification in vitro - amitriptyline and verapamil as model compounds. *Rapid Communications in Mass Spectrometry* **24,** 939–957. ISSN: 09514198. `http://doi.wiley.com/10.1002/rcm.4465` (Apr. 2010).
88. Burton, L. *et al.* Instrumental and experimental effects in LC-MS-based metabolomics. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **871,** 227–235. ISSN: 15700232 (2008).
89. Eilers, P. H. Parametric Time Warping. *Analytical Chemistry* **76,** 404–411. ISSN: 00032700 (2004).
90. Lange, E. *et al.* A geometric approach for the alignment of liquid chromatographymass spectrometry data. *Bioinformatics* **23,** i273–i281. ISSN: 1460-2059. `https://academic.oup.com/bioinformatics/article/23/13/i273/233877` (July 2007).
91. Veselkov, K. a. *et al.* Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. *Analytical Chemistry* **83,** 5864–5872 (2011).
92. Dunn, W. B. *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled

to mass spectrometry. *Nature Protocols* **6,** 1060–1083. ISSN: 1754-2189. http://www.nature.com/articles/nprot.2011.335 (July 2011).

93. Wang, S.-Y., Kuo, C.-H. & Tseng, Y. J. Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Met. *Analytical Chemistry* **85,** 1037–1046. ISSN: 0003-2700. https://pubs.acs.org/doi/10.1021/ac302877x (Jan. 2013).
94. Fernández-Albert, F. *et al.* Intensity drift removal in LC/MS metabolomics by common variance compensation. *Bioinformatics* **30,** 2899–2905. ISSN: 1460-2059. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu423 (Oct. 2014).
95. Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in Bioinformatics* **16,** 104–117. ISSN: 1467-5463. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbt080 (Jan. 2015).
96. Smith, C. A., Want, E. J., Maille, G. O., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **78,** 779–787. ISSN: 0003-2700 (2006).
97. Katajamaa, M., Miettinen, J. & Orei, M. MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **22,** 634–636. ISSN: 13674803 (2006).
98. Pluskal, T., Castillo, S., Villar-Briones, A. & Orei, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11,** 395. ISSN: 1471-2105. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-395 (Dec. 2010).
99. Sturm, M. *et al.* OpenMS - An open-source software framework for mass spectrometry. *BMC Bioinformatics* **9,** 1–11. ISSN: 14712105 (2008).
100. Lommen, A. MetAlign : Interface-Driven , Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* **81,** 3079–3086. ISSN: 0003-2700 (2009).
101. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry* **84,** 5035–5039. ISSN: 0003-2700. http://pubs.acs.org/doi/abs/10.1021/ac300698c (June 2012).
102. Melamud, E., Vastag, L. & Rabinowitz, J. D. Metabolomic analysis and visualization engine for LC - MS data. *Analytical Chemistry* **82,** 9818–9826. ISSN: 00032700 (2010).
103. Davidson, R. L., Weber, R. J. M., Liu, H., Sharma-Oates, A. & Viant, M. R. Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion

and liquid chromatography mass spectrometry-based metabolomics data. *GigaScience* **5.** ISSN: 2047-217X. http://dx.doi.org/10.1186/s13742-016-0115-8%20https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-016-0115-8 (Dec. 2016).

104. Giacomoni, F. *et al.* Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics* **31,** 1493–1495. ISSN: 14602059 (2015).
105. Du, P., Kibbe, W. A. & Lin, S. M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **22,** 2059–2065. ISSN: 13674803 (2006).
106. Lange, E., Gropl, C., Reinert, K., Kohlbacher, O. & Hildebrandt, A. High-accuracy peak picking of proteomics data using wavelet techniques. *Proceedings of the Pacific Symposium on Biocomputing 2006, PSB 2006,* 243–254. ISSN: 2335-6928 (2006).
107. Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9,** 16. ISSN: 1471-2105 (2008).
108. Coble, J. B. & Fraga, C. G. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. *Journal of Chromatography A* **1358,** 155–164. ISSN: 18733778. http://dx.doi.org/10.1016/j.chroma.2014.06.100 (2014).
109. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Analytical Chemistry* **89,** 8689–8695. ISSN: 0003-2700. http://pubs.acs.org/doi/10.1021/acs.analchem.7b01069 (Sept. 2017).
110. Salek, R. M., Steinbeck, C., Viant, M. R., Goodacre, R. & Dunn, W. B. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience* **2,** 2–4. ISSN: 2047217X (2013).
111. Chaleckis, R., Meister, I., Zhang, P. & Wheelock, C. E. Challenges, progress and promises of metabolite annotation for LCMS-based metabolomics. *Current Opinion in Biotechnology* **55,** 44–50. ISSN: 18790429. https://doi.org/10.1016/j.copbio.2018.07.010 (2019).
112. Dunn, W. B. *et al.* Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9,** 44–66. ISSN: 15733882 (2013).
113. Zhu, X., Chen, Y. & Subramanian, R. Comparison of information-dependent acquisition, SWATH, and MS All techniques in metabolite identification study employing ultrahigh-performance liquid chromatography-quadrupole time-of-flight mass spectrometry. *Analytical Chemistry* **86,** 1202–1209. ISSN: 00032700 (2014).

114. Hu, Y., Cai, B. & Huan, T. Enhancing Metabolome Coverage in Data-Dependent LC-MS/MS Analysis through an Integrated Feature Extraction Strategy. *Analytical Chemistry* **91,** 144331444. ISSN: 15206882 (2019).
115. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods* **12,** 523–526. ISSN: 1548-7091. http://www.nature.com/articles/nmeth.3393 (June 2015).
116. Tsuchiya, Y., Takahashi, Y., Jindo, T., Furuhama, K. & Suzuki, K. T. Comprehensive evaluation of canine renal papillary necrosis induced by nefiracetam, a neurotransmission enhancer. *European Journal of Pharmacology* **475,** 119–128. ISSN: 00142999 (2003).
117. De Livera, A. M. *et al.* Normalizing and Integrating Metabolomics Data. *Analytical Chemistry* **84,** 10768–10776. ISSN: 0003-2700. https://pubs.acs.org/doi/10.1021/ac302748b (Dec. 2012).
118. Livera, A. M. D. *et al.* Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Analytical Chemistry* **87,** 3606–3615. ISSN: 0003-2700. arXiv: 15334406. https://pubs.acs.org/doi/10.1021/ac502439y (Apr. 2015).
119. Sánchez-Illana, Á. *et al.* Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling. *Analytica Chimica Acta* **1019,** 38–48. ISSN: 18734324 (2018).
120. Lai, L. *et al.* Methodological considerations in the development of HPLC-MS methods for the analysis of rodent plasma for metabonomic studies. *Molecular BioSystems* **6,** 108–120. ISSN: 1742206X (2009).
121. Watrous, J. D. *et al.* Visualization, Quantification, and Alignment of Spectral Drift in Population Scale Untargeted Metabolomics Data. *Analytical Chemistry* **89,** 1399–1404. ISSN: 0003-2700. http://pubs.acs.org/doi/10.1021/acs.analchem.6b04337 (Feb. 2017).
122. Barwick, V. (*Eurachem/CITAC Guide: Guide to quality in analytical chemistry: An aid to accreditation (3rd ed.)* ISBN: 1522-6514 (Print)\r1522-6514 (Linking). http://www.ncbi.nlm.nih.gov/pubmed/25174426 (2016).
123. Dudzik, D., Barbas-Bernardos, C., García, A. & Barbas, C. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. *Journal of Pharmaceutical and Biomedical Analysis* **147,** 149–173. ISSN: 1873264X. https://doi.org/10.1016/j.jpba.2017.07.044 (2018).
124. Broadhurst, D. *et al.* Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **14,** 1–17. ISSN: 15733890. http://dx.doi.org/10.1007/s11306-018-1367-3 (2018).
125. Dunn, W. B. *et al.* Quality assurance and quality control processes: summary of a metabolomics community questionnaire. *Metabolomics* **13,** 1–6. ISSN: 15733890 (2017).

126. Beger, R. D. *et al.* Towards quality assurance and quality control in untargeted metabolomics studies. *Metabolomics* **15,** 1–5. ISSN: 15733890. http://dx.doi.org/10.1007/s11306-018-1460-7 (2019).
127. Gika, H. G., Theodoridis, G. A., Wingate, J. E. & Wilson, I. D. Within-Day Reproducibility of an HPLCMS-Based Method for Metabonomic Analysis: Application to Human Urine. *Journal of Proteome Research* **6,** 3291–3303. ISSN: 1535-3893. https://pubs.acs.org/doi/abs/10.1021/pr070183p (Aug. 2007).
128. Brunius, C., Shi, L. & Landberg, R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* **12,** 173. ISSN: 1573-3882. http://link.springer.com/10.1007/s11306-016-1124-4 (Nov. 2016).
129. Cao, L. *et al.* genuMet: distinguish genuine untargeted metabolic features without quality control samples. *bioRxiv,* 837260. https://www.biorxiv.org/content/10.1101/837260v1 (2019).
130. Sysi-Aho, M., Katajamaa, M., Yetukuri, L. & Orei, M. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics* **8,** 93. ISSN: 14712105. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-93 (2007).
131. Bijlsma, S. *et al.* Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry* **78,** 567–574. ISSN: 00032700 (2006).
132. Elliott, P. *et al.* The Airwave Health Monitoring Study of police officers and staff in Great Britain: Rationale, design and methods. *Environmental research* **134,** 280–285. ISSN: 1096-0953. http://www.sciencedirect.com/science/article/pii/S0013935114002564 (2014).
133. Dona, A. C. *et al.* Precision High-Throughput Proton NMR Spectroscopy of Human Urine, Serum, and Plasma for Large-Scale Metabolic Phenotyping. *Analytical Chemistry* **86,** 9887–9894. ISSN: 0003-2700 (2014).
134. Lewis, M. R. *Development of an Advanced Molecular Profiling Pipeline for Human Population Screening* PhD thesis (Imperial College London, 2014).
135. Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. *Analytical Chemistry* **81,** 6656–6667. ISSN: 0003-2700. http://pubs.acs.org/doi/abs/10.1021/ac901536h (Aug. 2009).
136. Izzi-Engbeaya, C. *et al.* The effects of kisspeptin on β-cell function, serum metabolites and appetite in humans. *Diabetes, Obesity and Metabolism* **20,** 2800–2810. ISSN: 14631326 (2018).

137. Libiseller, G. *et al.* IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16,** 118. ISSN: 1471-2105. http://www.biomedcentral.com/1471-2105/16/118 (2015).
138. Sands, C. J. *et al.* The nPYc-Toolbox, a Python module for the pre-processing, quality-control and analysis of metabolic profiling datasets. *Bioinformatics,* 1–2. ISSN: 1367-4803 (2019).
139. Croixmarie, V. *et al.* Integrated Comparison of Drug-Related and Drug-Induced Ultra Performance Liquid Chromatography/Mass Spectrometry Metabonomic Profiles Using Human Hepatocyte Cultures. *Analytical Chemistry* **81,** 6061–6069. ISSN: 0003-2700. https://pubs.acs.org/doi/10.1021/ac900333e (Aug. 2009).
140. Van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* **7.** ISSN: 14712164 (2006).
141. Triba, M. N. *et al.* PLS/OPLS models in metabolomics: The impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Molecular BioSystems* **11,** 13–19. ISSN: 17422051 (2015).
142. Kuracina, M. & Schreiber, A. Considerations when using LC-MS/MS Systems with Fast and High Resolution Liquid Chromatography. *Applied Biosystems.* https://sciex.com/Documents/Downloads/Literature/mass-spectrometry-High-Resolution-Liquid%20Chromatography-1282110.pdf (2008).
143. Albóniga, O. E., González, O., Alonso, R. M., Xu, Y. & Goodacre, R. Optimization of XCMS parameters for LC MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics* **16.** ISSN: 1573-3890. https://doi.org/10.1007/s11306-020-1636-9 (2020).
144. Prince, J. T. & Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry* **78,** 6140–6152. ISSN: 00032700 (2006).
145. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Analytical Chemistry* **89,** 8696–8703. ISSN: 15206882 (2017).
146. Weber, R. J. M. *et al.* Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. *Metabolomics* **13,** 12. ISSN: 1573-3882. http://link.springer.com/10.1007/s11306-016-1147-x (Feb. 2017).
147. Katajamaa, M. & Orei, M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* **6.** ISSN: 14712105 (2005).

148. Duran, A. L., Yang, J., Wang, L. & Sumner, L. W. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* **19,** 2283–2293. ISSN: 13674803 (2003).
149. De Souza, D. P., Saunders, E. C., McConville, M. J. & Liki, V. A. Progressive peak clustering in GC-MS Metabolomic experiments applied to Leishmania parasites. *Bioinformatics* **22,** 1391–1396. ISSN: 13674803 (2006).
150. Ballardini, R., Benevento, M., Arrigoni, G., Pattini, L. & Roda, A. MassUntangler: A novel alignment tool for label-free liquid chromatographymass spectrometry proteomic data. *Journal of Chromatography A* **1218,** 8859–8868. ISSN: 00219673. http://dx.doi.org/10.1016/j.chroma.2011.06.062%20https://linkinghub.elsevier.com/retrieve/pii/S0021967311008776 (Dec. 2011).
151. Scheltema, R. *et al.* Simple data-reduction method for high-resolution LCMS data in metabolomics. *Bioanalysis* **1,** 1551–1557. ISSN: 1757-6180. http://www.future-science.com/doi/10.4155/bio.09.146 (Dec. 2009).
152. Zhang, W. *et al.* MET-COFEA: A Liquid Chromatography/Mass Spectrometry Data Processing Platform for Metabolite Compound Feature Extraction and Annotation. *Analytical Chemistry* **86,** 6245–6253. ISSN: 0003-2700. https://pubs.acs.org/doi/10.1021/ac501162k (July 2014).
153. Wandy, J., Daly, R., Breitling, R. & Rogers, S. Incorporating peak grouping information for alignment of multiple liquid chromatography-mass spectrometry datasets. *Bioinformatics* **31,** 1999–2006. ISSN: 1367-4803. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv072 (June 2015).
154. Gu, H., Gowda, G. a. N., Neto, F. C., Opp, M. R. & Raftery, D. RAMSY: Ratio Analysis of Mass Spectrometry to Improve Compound Identification. *Analytical chemistry* **85,** 10771–10779. ISSN: 00032700. http://pubs.acs.org/doi/abs/10.1021/ac4019268 (2013).
155. Rafiei, A. & Sleno, L. Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Communications in Mass Spectrometry* **29,** 119–127. ISSN: 10970231 (2014).
156. Tautenhahn, R., Bottcher, C. & Neumann, S. Annotation of LC ESI-MS mass signals. *Proceedings of BIRD 2007 1st International Conference on Bioinformatics Research and Development*, 371–380 (2007).
157. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* **76,** 036106. ISSN: 1539-3755. arXiv: 0709.2938. https://link.aps.org/doi/10.1103/PhysRevE.76.036106 (Sept. 2007).
158. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society*

for Mass Spectrometry **5,** 859–866. ISSN: 1044-0305. http://link.springer.com/10.1016/1044-0305(94)87009-8 (Sept. 1994).

159. Prakash, A. *et al.* Signal Maps for Mass Spectrometry-based Comparative Proteomics. *Molecular & Cellular Proteomics* **5,** 423–432. ISSN: 1535-9476. http://www.mcponline.org/lookup/doi/10.1074/mcp.M500133-MCP200 (2006).
160. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–667. ISSN: 16159853. http://doi.wiley.com/10.1002/pmic.200600625 (Mar. 2007).
161. Gorman, J. W. & Hinman, J. E. Simplex Lattice Designs for Multicomponent Systems. *Technometrics* **4,** 463–487. ISSN: 15372723 (1962).
162. Di Guida, R. *et al.* Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **12,** 1–14. ISSN: 15733890 (2016).
163. Do, K. T. *et al.* Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **14,** 128. ISSN: 15733890. http://dx.doi.org/10.1007/s11306-018-1420-2%20http://link.springer.com/10.1007/s11306-018-1420-2 (Oct. 2018).
164. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Analytical Chemistry* **84,** 283–289. ISSN: 0003-2700. arXiv: NIHMS150003. http://pubs.acs.org/doi/abs/10.1021/ac202450g%20http://pubs.acs.org/doi/10.1021/ac202450g (Jan. 2012).
165. Borràs, S., Kaufmann, A. & Companyó, R. Correlation of precursor and product ions in single-stage high resolution mass spectrometry. A tool for detecting diagnostic ions and improving the precursor elemental composition elucidation. *Analytica Chimica Acta* **772,** 47–58. ISSN: 00032670. http://dx.doi.org/10.1016/j.aca.2013.02.012 (2013).
166. Senan, O. *et al.* CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics* (ed Stegle, O.) 1–9. ISSN: 1367-4803. https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz207/5418951 (Mar. 2019).
167. Kim, S., Koo, I., Wei, X. & Zhang, X. A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry. *Bioinformatics* **28,** 1158–1163. ISSN: 13674803 (2012).
168. Depke, T., Franke, R. & Brönstrup, M. Clustering of MS2 spectra using unsupervised methods to aid the identification of secondary metabolites from Pseudomonas aeruginosa. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **1071,** 19–28. ISSN: 1873376X. https://doi.org/10.1016/j.jchromb.2017.06.002 (2017).

169. Brown, M. *et al.* Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics* **27,** 1108–1112. ISSN: 13674803 (2011).
170. Broeckling, C. D., Afsar, F. A., Neumann, S. & Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral- Matching-Based Annotation for Metabolomics Data. *Analytical Chemistry* **86,** 6812–6817 (2014).
171. Athersuch, T. J., Malik, S., Weljie, A., Newton, J. & Keun, H. C. Evaluation of 1H NMR metabolic profiling using biofluid mixture design. *Analytical Chemistry* **85,** 6674–6681. ISSN: 00032700 (2013).
172. Surowiec, I. *et al.* Quantification of run order effect on chromatography - mass spectrometry profiling data. *Journal of Chromatography A* **1568,** 229–234. ISSN: 00219673. `https://doi.org/10.1016/j.chroma.2018.07.019%20https://linkinghub.elsevier.com/retrieve/pii/S0021967318308616` (Sept. 2018).
173. Spicer, R., Salek, R. M., Moreno, P., Cañueto, D. & Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics* **13,** 106. ISSN: 1573-3882. `http://link.springer.com/10.1007/s11306-017-1242-7` (Sept. 2017).
174. Drogan, D. *et al.* Untargeted Metabolic Profiling Identifies Altered Serum Metabolites of Type 2 Diabetes Mellitus in a Prospective, Nested Case Control Study. *Clinical Chemistry* **61,** 487–497. ISSN: 0009-9147. `http://www.clinchem.org/cgi/doi/10.1373/clinchem.2014.228965` (Mar. 2015).
175. Athersuch, T. Metabolome analyses in exposome studies: Profiling methods for a vast chemical space. *Archives of Biochemistry and Biophysics* **589,** 177–186. ISSN: 10960384. `http://dx.doi.org/10.1016/j.abb.2015.10.007` (2016).
176. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3,** 211–221. ISSN: 15733882 (2007).
177. Blaenovi, I., Kind, T., Ji, J. & Fiehn, O. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* **8.** ISSN: 22181989 (2018).
178. Kind, T. & Fiehn, O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **7.** ISSN: 14712105 (2006).
179. Xu, Y. F., Lu, W. & Rabinowitz, J. D. Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. *Analytical Chemistry* **87,** 2273–2281. ISSN: 15206882 (2015).
180. Aicheler, F. *et al.* Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches. *Analytical Chemistry* **87,** 7698–7704. ISSN: 15206882 (2015).

181. Boudah, S. *et al.* Annotation of the human serum metabolome by coupling three liquid chromatography methods to high-resolution mass spectrometry. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **966,** 34–47. ISSN: 1873376X (2014).
182. Guijas, C. *et al.* METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Analytical Chemistry* **90,** 3156–3164. ISSN: 15206882 (2018).
183. Wishart, D. S. *et al.* HMDB: The human metabolome database. *Nucleic Acids Research* **35,** 521–526. ISSN: 03051048 (2007).
184. Fahy, E., Sud, M., Cotter, D. & Subramaniam, S. LIPID MAPS online tools for lipid research. *Nucleic Acids Research* **35,** 606–612. ISSN: 03051048 (2007).
185. Sokolow, S., Karnofsky, J. & Gustafson, P. *The Finnigan Library Search Program, Finnigan Application Report 2* tech. rep. (Finnigan Corp.; San Jose, CA, 1978), 1–45.
186. Wan, K. X., Vidavsky, I. & Gross, M. L. Comparing similar spectra: From similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry* **13,** 85–88. ISSN: 10440305 (2002).
187. Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **11.** ISSN: 14712105 (2010).
188. Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Analytical Chemistry* **88,** 7946–7958. ISSN: 15206882 (2016).
189. Aguilar-Mogas, A., Sales-Pardo, M., Navarro, M., Guimerà, R. & Yanes, O. iMet: A Network-Based Computational Tool to Assist in the Annotation of Metabolites from Tandem Mass Spectra. *Analytical Chemistry* **89,** 3474–3482. ISSN: 15206882 (2017).
190. Vinaixa, M., Schymanski, E. L. & Neumann, S. Mass spectral databases for LC/MS-and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in* **78,** 23–35. `http://www.sciencedirect.com/science/article/pii/S0165993615300832` (2016).
191. Mortier, K. A., Zhang, G. F., Van Peteghem, C. H. & Lambert, W. E. Adduct formation in quantitative bioanalysis: Effect of ionization conditions on paclitaxel. *Journal of the American Society for Mass Spectrometry* **15,** 585–592. ISSN: 10440305 (2004).
192. Tikunov, Y. M., Laptenok, S., Hall, R. D., Bovy, A. & de Vos, R. C. MSClust: A tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics* **8,** 714–718. ISSN: 15733882 (2012).
193. Alonso, A. *et al.* AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics* **27,** 1339–1340. ISSN: 1460-2059. `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr138` (May 2011).

194. DeFelice, B. C. *et al.* Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography-Mass Spectroscopy (LC-MS) Data Processing. *Analytical Chemistry* **89,** 3250–3255. ISSN: 15206882 (2017).
195. Uppal, K., Walker, D. I. & Jones, D. P. xMSannotator: An R package for network-based annotation of high-resolution metabolomics data. *Analytical Chemistry* **89,** 1063–1067. ISSN: 15206882 (2017).
196. Hutchins, P. D., Russell, J. D. & Coon, J. J. LipiDex: An Integrated Software Package for High-Confidence Lipid Identification. *Cell Systems* **6,** 621–625.e5. ISSN: 24054720. `https://doi.org/10.1016/j.cels.2018.03.011` (2018).
197. Van Eeckhaut, A., Lanckmans, K., Sarre, S., Smolders, I. & Michotte, Y. Validation of bioanalytical LC-MS/MS assays: Evaluation of matrix effects. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **877,** 2198–2207. ISSN: 15700232 (2009).
198. Chambers, E., Wagrowski-Diehl, D. M., Lu, Z. & Mazzeo, J. R. Systematic and comprehensive strategy for reducing matrix effects in LC/MS/MS analyses. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **852,** 22–34. ISSN: 15700232 (2007).
199. Würtz, P. *et al.* Metabolite profiling and cardiovascular event risk: A prospective study of 3 population-based cohorts. *Circulation* **131,** 774–785. ISSN: 15244539 (2015).
200. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018. ISSN: 2052-4463. `http://www.nature.com/articles/sdata201618` (Dec. 2016).
201. Haug, K. *et al.* MetaboLights: A resource evolving in response to the needs of its scientific community. *Nucleic Acids Research* **48,** D440–D444. ISSN: 13624962 (2020).

Appendix A

TABLE A.1: Detection of validated metabolites (adducts/in-source fragments (ISF) ions) was performed in AIRWAVE samples using specified *m/z* and RT regions kindly provided by the National Phenome Centre team.

Metabolite	cpdID	Ion	rtMin	rtMax	mzMin	mzMax
Adenosine	HPOS-005.1	M+H	89.41	119.41	268.1000	268.1081
	HPOS-005.2	ISF	89.41	119.41	136.0600	136.0640
Choline	HPOS-007.1	M	221.36	251.36	104.1060	104.1091
Carnitine	HPOS-008.1	M+H	305.26	335.26	162.1100	162.1149
	HPOS-008.2	M+Na	305.26	335.26	184.0922	184.0978
	HPOS-008.3	ISF	305.26	335.26	103.0371	103.0401
Laurylcarnitine (C12:0)	HPOS-013.1	M+H	219.37	249.37	344.2744	344.2847
	HPOS-013.2	M+Na	219.37	249.37	366.2565	366.2675
	HPOS-013.3	ISF	219.37	249.37	285.2037	285.2123
Histidine	HPOS-014.1	M+H	354.04	384.04	156.0744	156.0791
	HPOS-014.2	M+Na	354.04	384.04	178.0553	178.0607
	HPOS-014.3	ISF	354.04	384.04	110.0694	110.0728
N6,N6,N6-Trimethyllysine	HPOS-016.1	M+H	354.77	384.77	189.1569	189.1626
	HPOS-016.2	ISF	354.77	384.77	130.0832	130.0872
	HPOS-016.3	ISF	354.77	384.77	84.0787	84.0813
Taurine	HPOS-023.1	M+H	141.54	171.54	126.0200	126.0238
Paraxanthine	HPOS-024.1	M+H	41.88	71.88	181.0693	181.0747
Trigonelline	HPOS-025.1	M+H	276.00	306.00	138.0529	138.0570
	HPOS-025.2	ISF	276.00	306.00	94.0636	94.0664
Betaine	HPOS-027.1	M+H	271.97	301.97	118.0845	118.0880
	HPOS-027.2	M+Na	271.97	301.97	140.0661	140.0703
Warfarin	HPOS-031.2	ISF	25.37	55.37	251.0665	251.0741
	HPOS-031.3	ISF	25.37	55.37	163.0374	163.0422
Caffeine	HPOS-033.1	M+H	35.98	65.98	195.0847	195.0906
	HPOS-033.2	ISF	35.98	65.98	138.0659	138.0701
Niacinamide	HPOS-034.1	M+H	49.05	79.05	123.0534	123.0571
Creatinine	HPOS-036.1	M+Na	135.41	165.41	136.0461	136.0502
	HPOS-036.2	2M+H	135.41	165.41	227.1216	227.1284
1,1-Dimethylbiguanide	HPOS-038.1	M+H	190.41	220.41	130.1068	130.1107
	HPOS-038.2	ISF	190.41	220.41	113.0793	113.0827
Tryptophan	HPOS-039.1	M+H	214.34	244.34	205.0941	205.1002
	HPOS-039.2	ISF	214.34	244.34	188.0682	188.0738
Phenylalanine	HPOS-040.1	M+H	211.78	241.78	166.0838	166.0887
	HPOS-040.3	ISF	211.78	241.78	103.0535	103.0565
Methionine	HPOS-041.1	M+H	231.74	261.74	150.0561	150.0606
	HPOS-041.2	M+2Na-H	231.74	261.74	194.0191	194.0249

TABLE A.1: Detection of validated metabolites (adducts/in-source fragments (ISF) ions) was performed in AIRWAVE samples using specified *m/z* and RT regions kindly provided by the National Phenome Centre team.

Metabolite	**cpdID**	**Ion**	**rtMin**	**rtMax**	**mzMin**	**mzMax**
	HPOS-041.3	ISF	231.74	261.74	133.0290	133.0330
Trimethylamine N-oxide	HPOS-042.1	M+H	235.72	265.72	76.0745	76.0768
	HPOS-042.2	2M+H	235.72	265.72	151.1418	151.1464
Proline	HPOS-043.1	M+H	252.45	282.45	116.0689	116.0723
	HPOS-043.2	M+2Na-H	252.45	282.45	160.0326	160.0374
	HPOS-043.3	2M+H+2HCOONa	252.45	282.45	365.0615	365.0725
Alanine	HPOS-044.1	M+2Na-H	258.94	288.94	134.0168	134.0209
	HPOS-044.2	2M+H+2HCOONa	258.94	288.94	313.0308	313.0402
	HPOS-044.3	3M+Na+2HCOONa	258.94	288.94	356.0727	356.0833
Creatine	HPOS-045.1	M+H	288.58	318.58	132.0748	132.0787
	HPOS-045.2	M+Na	288.58	318.58	154.0564	154.0610
	HPOS-045.3	M+2Na-H	288.58	318.58	176.0389	176.0441
Glutamine	HPOS-046.1	M+H	288.32	318.32	147.0742	147.0786
	HPOS-046.2	M+2Na-H	288.32	318.32	191.0374	191.0432
	HPOS-046.3	ISF	288.32	318.32	130.0490	130.0530
Citrulline	HPOS-047.1	M+H	329.21	359.21	176.1003	176.1056
	HPOS-047.2	M+Na	329.21	359.21	198.0820	198.0880
	HPOS-047.3	ISF	329.21	359.21	159.0746	159.0794
Arginine	HPOS-048.1	M+H	340.10	370.10	175.1163	175.1216
	HPOS-048.2	M+2Na-H	340.10	370.10	219.0827	219.0893
	HPOS-048.3	ISF	340.10	370.10	158.0896	158.0944
Lysine	HPOS-049.1	M+H	343.81	373.81	147.1106	147.1150
a-glycerophosphocholine	HPOS-050.1	M+H	346.04	376.04	258.1062	258.1140
	HPOS-050.2	M+Na	346.04	376.04	280.0878	280.0962
	HPOS-050.3	ISF	346.04	376.04	184.0692	184.0748
	HPOS-050.4	ISF	346.04	376.04	104.1054	104.1086
3-methylhistidine	HPOS-051.1	M+H	350.50	380.50	170.0899	170.0950
	HPOS-051.2	ISF	350.50	380.50	126.1001	126.1039
	HPOS-052.1	M+H	366.01	396.01	170.0899	170.0950
N-Acetyl-D-mannosamine	HPOS-053.1	M+Na	93.48	123.48	244.0755	244.0828
1,2-Dimyristoyl-sn-glycero-3-phosphocholine	HPOS-054.1	M+H	235.75	265.75	678.4967	678.5170
Hypoxanthine	HPOS-055.2	M+Na	80.52	110.52	159.0253	159.0301
	HPOS-055.3	ISF	80.52	110.52	119.0342	119.0378
	HPOS-055.4	ISF	80.52	110.52	110.0333	110.0367
Pantothenate	HPOS-056.1	M+H	52.97	82.97	220.1146	220.1213
	HPOS-056.2	M+Na	52.97	82.97	242.0954	242.1026
	HPOS-056.3	ISF	52.97	82.97	202.1040	202.1100
Urocanate	HPOS-057.1	M+H	63.23	93.23	139.0481	139.0523
	HPOS-057.4	ISF	63.23	93.23	95.0586	95.0614
5′-Methylthioadenosine	HPOS-058.1	M+H	63.07	93.07	298.0924	298.1013
	HPOS-058.2	ISF	63.07	93.07	136.0610	136.0650
Cytosine	HPOS-059.1	M+H	128.53	158.53	112.0489	112.0522
Pipecolate	HPOS-061.1	M+H	253.72	283.72	130.0843	130.0882
	HPOS-061.2	M+2Na-H	253.72	283.72	174.0474	174.0526
Thiamine	HPOS-072.1	M+	314.55	344.55	265.1083	265.1163

TABLE A.1: Detection of validated metabolites (adducts/in-source fragments (ISF) ions) was performed in AIRWAVE samples using specified *m/z* and RT regions kindly provided by the National Phenome Centre team.

Metabolite	**cpdID**	**Ion**	**rtMin**	**rtMax**	**mzMin**	**mzMax**
	HPOS-072.2	ISF	314.55	344.55	122.0692	122.0728
4-Guanidinobutanoate	HPOS-073.1	M+H	215.41	245.41	146.0908	146.0951
	HPOS-073.2	M+Na	215.41	245.41	168.0715	168.0765
	HPOS-073.4	ISF	215.41	245.41	86.0587	86.0613
N,N-Dimethylglycine	HPOS-074.1	M+H	263.32	293.32	104.0696	104.0727
	HPOS-074.2	M+2Na-H	263.32	293.32	148.0308	148.0352
L-prolyl-L-proline	HPOS-076.1	M+H	311.06	341.06	213.1202	213.1266
N6-Acetyl-L-lysine	HPOS-077.1	M+H	308.98	338.98	189.1205	189.1262
Inosine	HPOS-079.1	M+Na	82.31	112.31	291.0656	291.0744
	HPOS-079.2	M+2Na-H	82.31	112.31	313.0473	313.0567
Cortisol	HPOS-086.1	M+H	29.46	59.46	363.2112	363.2220
	HPOS-086.2	ISF	29.46	59.46	345.1988	345.2092
	HPOS-086.3	ISF	29.46	59.46	327.1921	327.2019
1-Methylnicotinamide	HPOS-089.1	M+	234.34	264.34	137.0693	137.0734
	HPOS-089.2	ISF	234.34	264.34	94.0646	94.0674
Sucrose	HPOS-091.1	M+Na	123.21	153.21	365.0998	365.1107
	HPOS-091.2	M+K	123.21	153.21	381.0733	381.0847

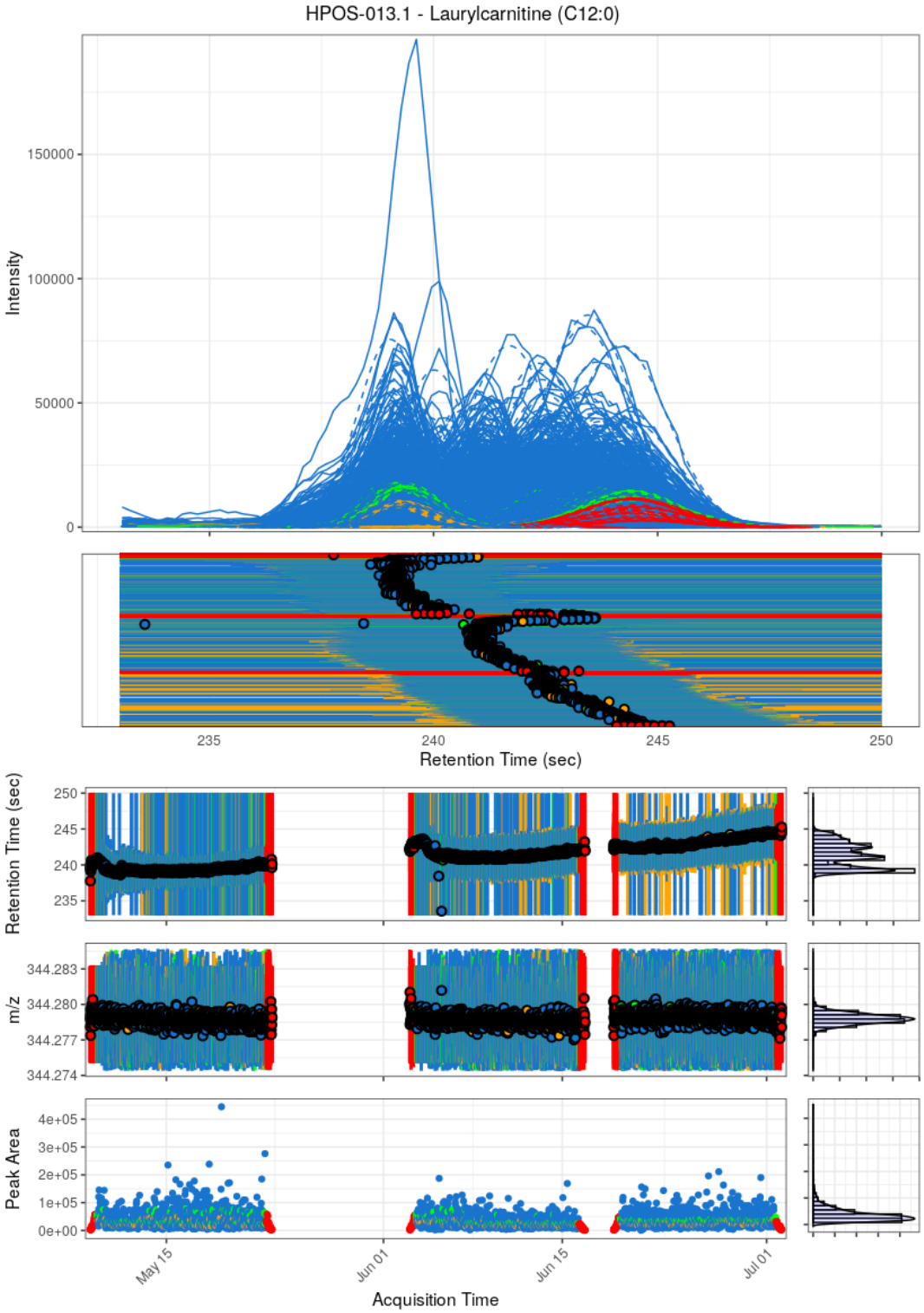


FIGURE A.1: Detection and integration of laurylcarnitine (C12:0) main adduct ion in all AIRWAVE serum HILIC-POS-MS samples. Different colours indicate sample type: red - QC dilution series, green - QC sample, orange - external reference sample, blue - study sample.

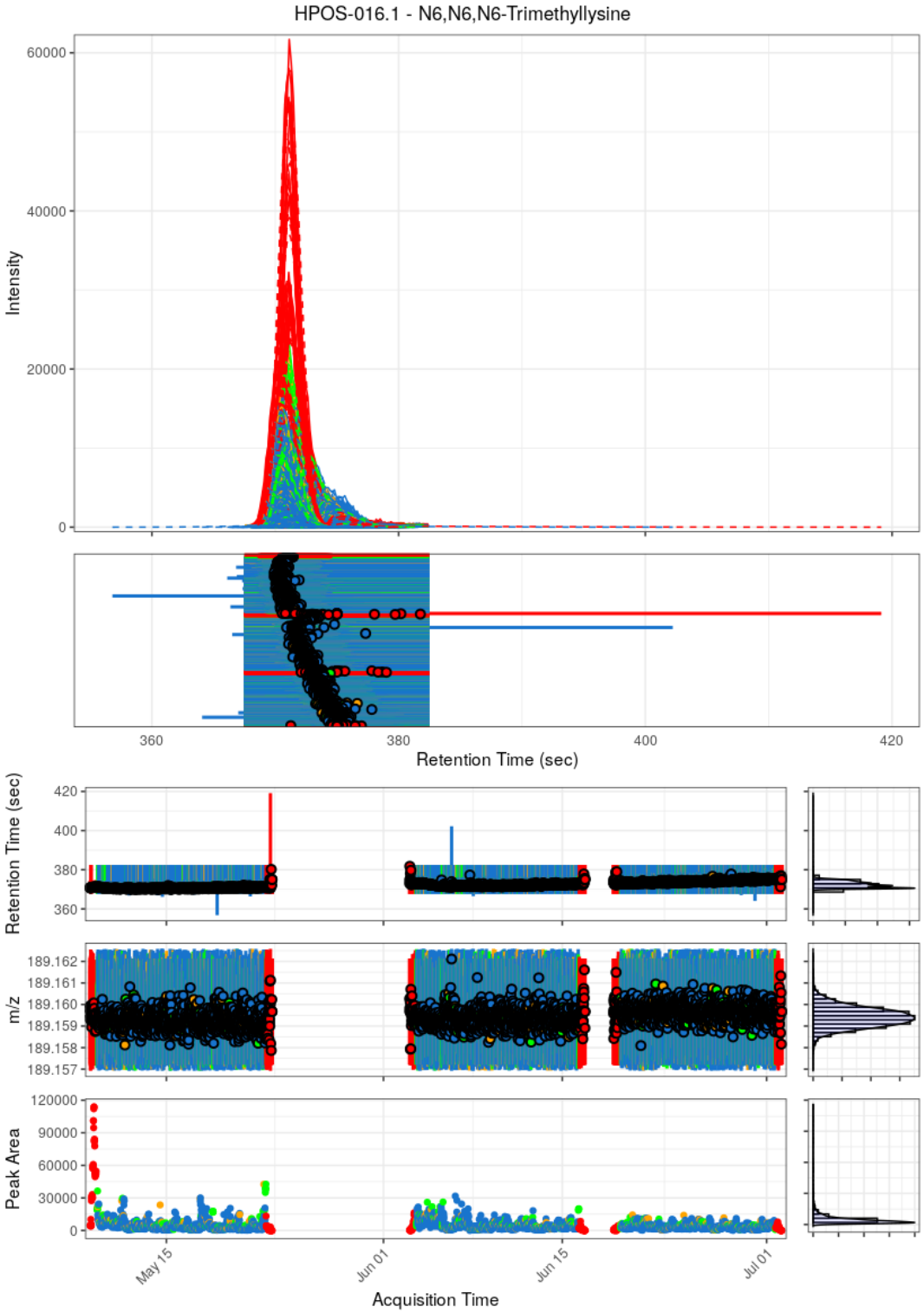


FIGURE A.2: Detection and integration of N6,N6,N6-Trimethyllysine main adduct ion in all AIRWAVE serum HILIC-POS-MS samples. Different colours indicate sample type: red - QC dilution series, green - QC sample, orange - external reference sample, blue - study sample.

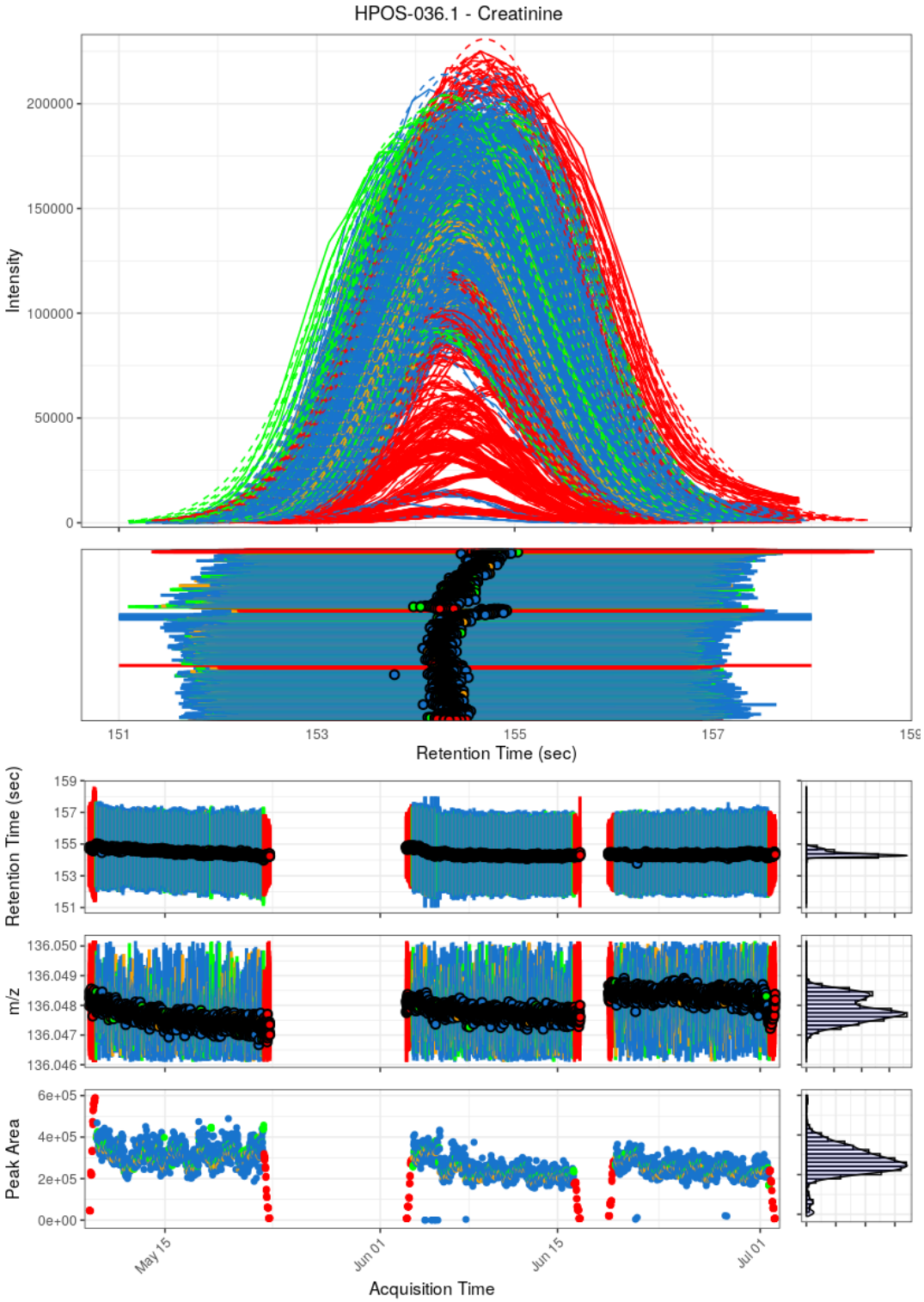


FIGURE A.3: Detection and integration of creatinine main adduct ion in all AIRWAVE serum HILIC-POS-MS samples. Different colours indicate sample type: red - QC dilution series, green - QC sample, orange - external reference sample, blue - study sample.

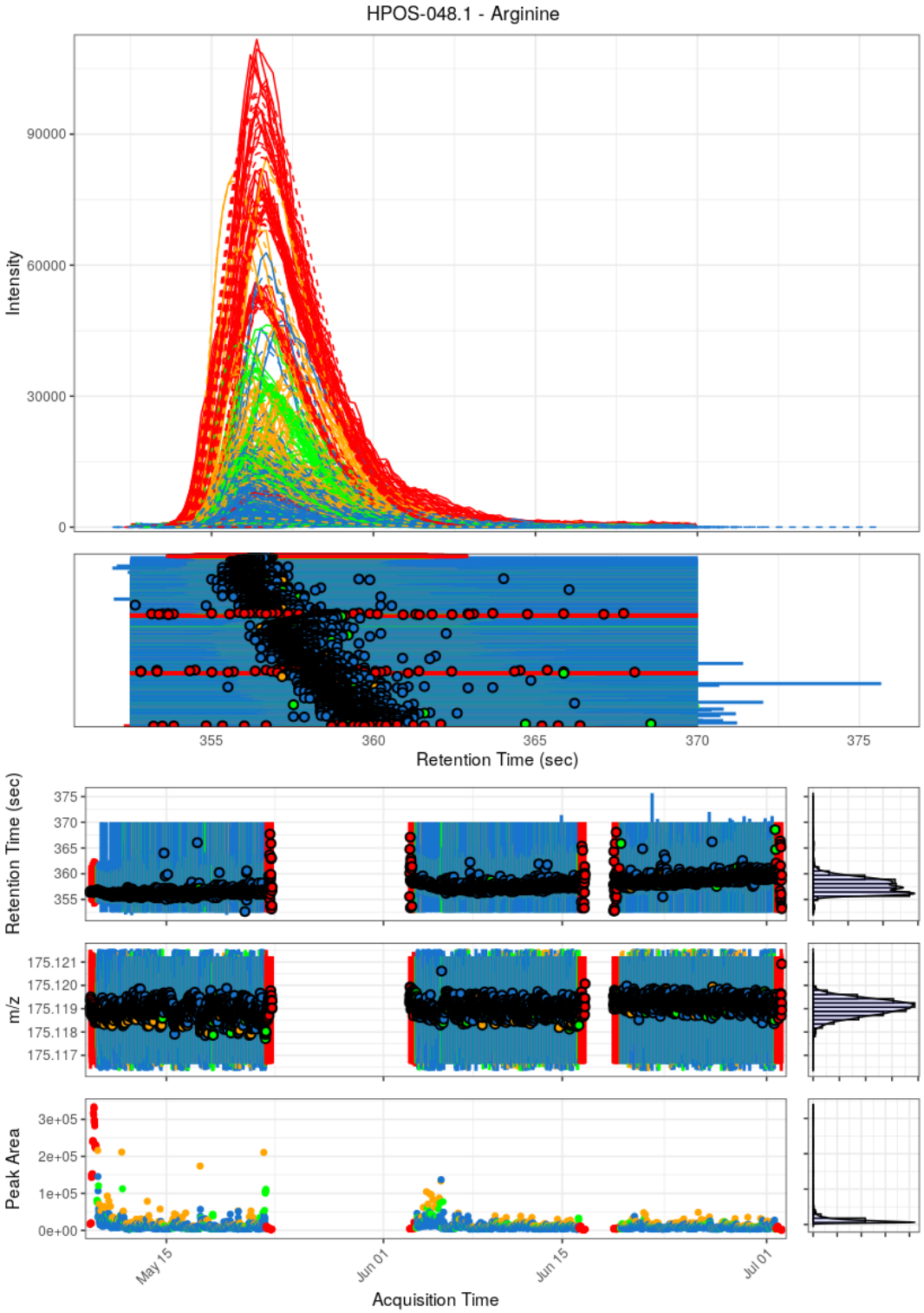


FIGURE A.4: Detection and integration of arginine main adduct ion in all AIRWAVE serum HILIC-POS-MS samples. Different colours indicate sample type: red - QC dilution series, green - QC sample, orange - external reference sample, blue - study sample.

Appendix B

Endogenous metabolites detection in DEVSET samples

TABLE B.1: Detection of validated metabolites and their adducts/in-source fragments (ISF) was performed in DEVSET samples using specified *m/z* and RT regions kindly provided by the IPC team.

Metabolite	**cpdID**	**Ion**	**rtMin**	**rtMax**	**mzMin**	**mzMax**
Urocanate	RPOS-002.1	M+H	54.11	59.92	139.0484	139.0515
	RPOS-002.2	ISF	54.01	59.92	121.0377	121.0411
Theobromine	RPOS-005.1	M+H	142.78	148.74	181.0706	181.0730
	RPOS-005.2	ISF	142.69	148.88	163.0594	163.0642
	RPOS-005.3	ISF	142.89	148.70	138.0645	138.0685
Pseudouridine	RPOS-008.1	M+H	53.00	58.78	245.0739	245.0795
	RPOS-008.2	ISF	53.11	58.72	209.0536	209.0565
	RPOS-008.3	ISF	53.06	58.74	191.0419	191.0474
	RPOS-008.4	ISF	53.08	58.75	179.0420	179.0464
	RPOS-008.5	ISF	53.07	58.75	155.0422	155.0459
Pantothenate	RPOS-015.1	M+H	136.82	142.77	220.1153	220.1203
	RPOS-015.2	ISF	136.81	142.79	202.1060	202.1109
	RPOS-015.3	ISF	136.60	142.93	184.0953	184.1008
1-Methyladenosine	RPOS-018.1	M+H	71.66	77.86	282.1162	282.1211
	RPOS-018.2	ISF	71.71	77.79	150.0760	150.0786
N-a-Acetyl-L-arginine	RPOS-025.1	M+H	51.26	57.11	217.1264	217.1324
	RPOS-025.2	ISF	51.26	57.31	200.1010	200.1069
N2,N2-Dimethylguanosine	RPOS-026.1	M+H	140.49	146.15	312.1269	312.1313
	RPOS-026.2	ISF	140.49	146.13	180.0863	180.0896
2-Furoylglycine	RPOS-027.1	M+H	136.52	143.36	170.0432	170.0471
	RPOS-027.2	ISF	136.49	143.46	124.0372	124.0408
	RPOS-027.3	ISF	136.56	143.37	95.0117	95.0137
Creatine	RPOS-041.1	M+H	33.91	39.22	132.0754	132.0775
	RPOS-041.2	ISF	33.84	39.23	90.0537	90.0563
Caffeine	RPOS-044.1	M+H	206.79	213.03	195.0868	195.0890
	RPOS-044.2	ISF	206.79	213.20	138.0647	138.0688
7-Methylguanine	RPOS-051.1	M+H	72.89	78.88	166.0716	166.0735
	RPOS-051.2	ISF	72.98	78.92	149.0434	149.0475
	RPOS-051.3	ISF	73.01	78.82	124.0482	124.0518
Pyroglutamate	RPOS-052.1	M+H	71.07	78.29	130.0482	130.0509
	RPOS-052.2	ISF	71.00	78.37	84.0437	84.0462
Paraxanthine	RPOS-061.1	M+H	163.39	169.61	181.0704	181.0735
	RPOS-061.2	ISF	163.48	169.52	124.0497	124.0532
Theophylline	RPOS-074.1	M+H	163.39	169.61	181.0704	181.0735
	RPOS-074.2	ISF	163.48	169.52	124.0497	124.0532
Imidazolelactate	RPOS-086.1	M+H	34.25	39.35	157.0591	157.0631
	RPOS-086.2	ISF	34.28	39.49	111.0531	111.0561

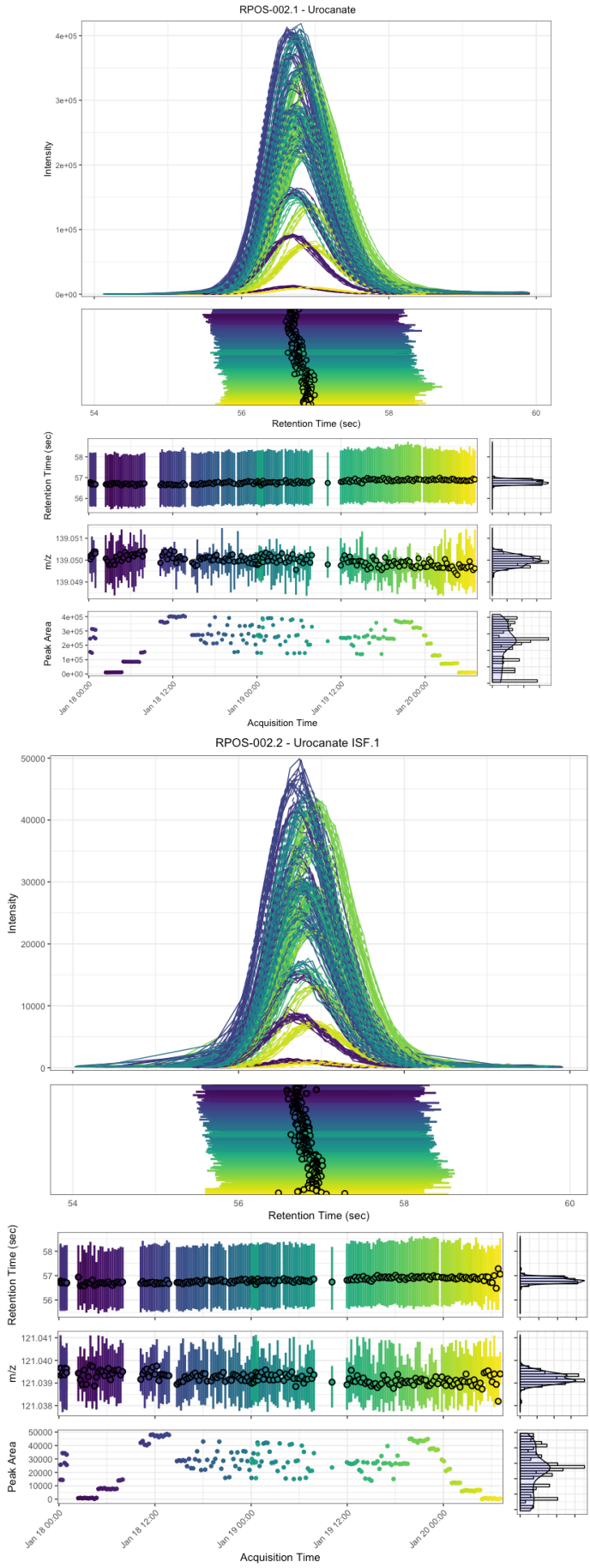


FIGURE B.1: Detection and integration of Urocanate main ion and its in-source fragment.

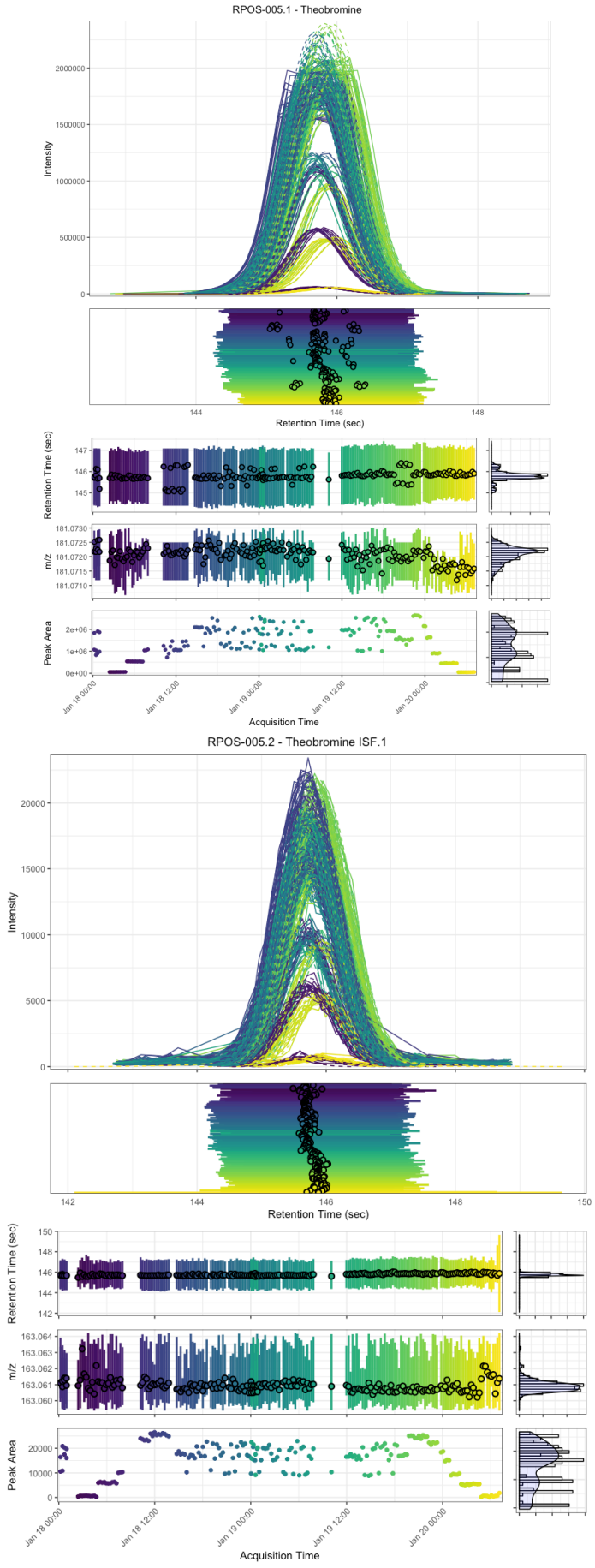


FIGURE B.2: Detection and integration of theobromine main ion and its in-source fragment.

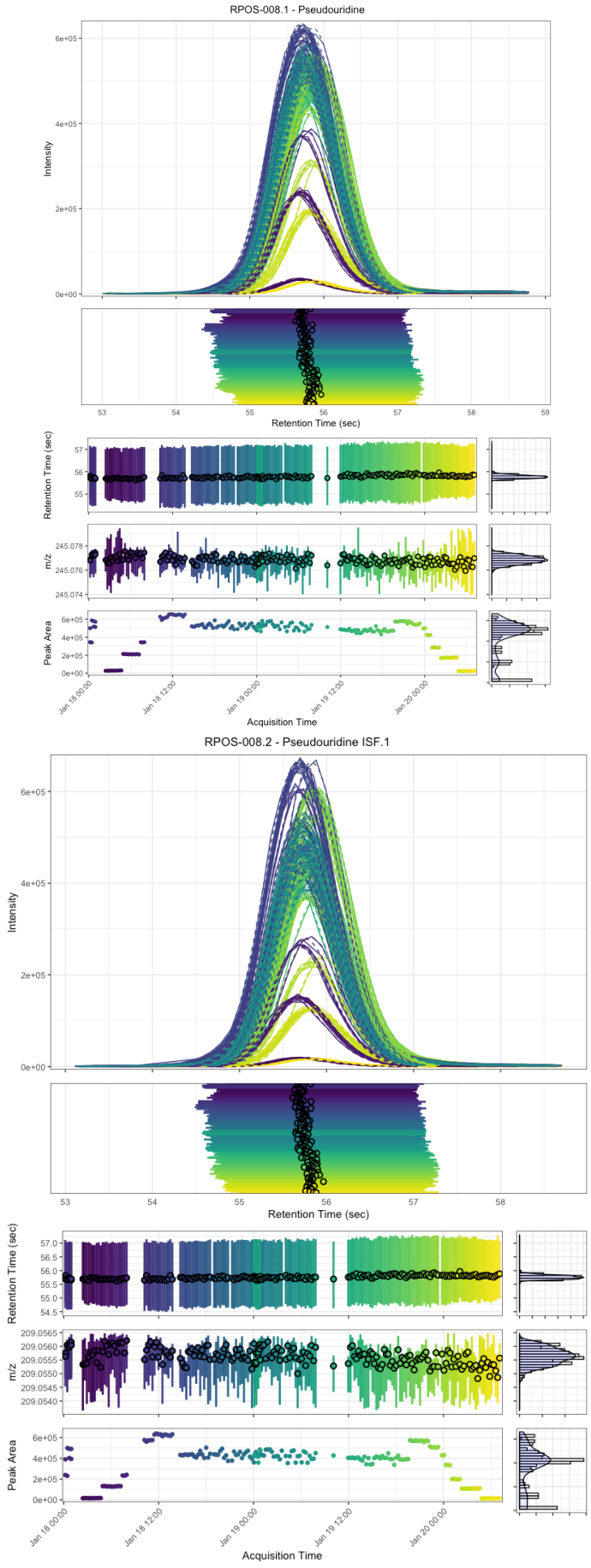


FIGURE B.3: Detection and integration of pseudouridine main ion and its its in-source fragment.

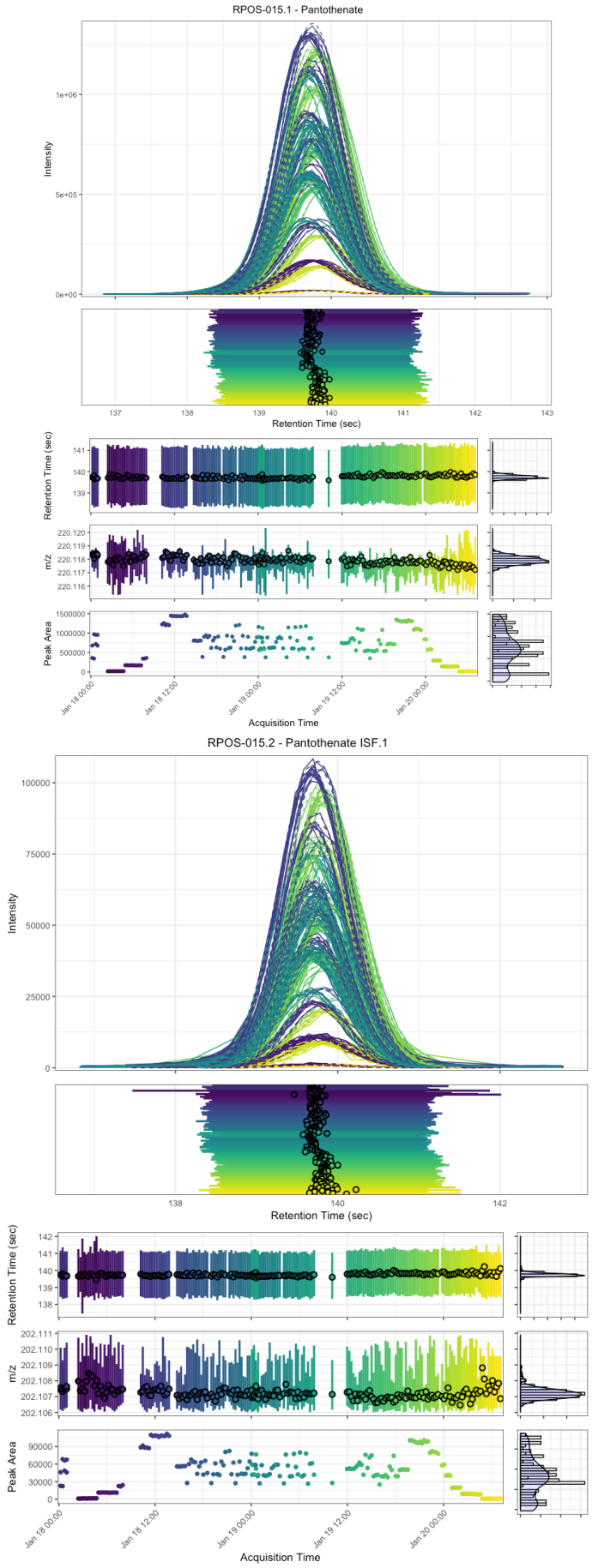


FIGURE B.4: Detection and integration of pantothenate main ion and its its in-source fragment.

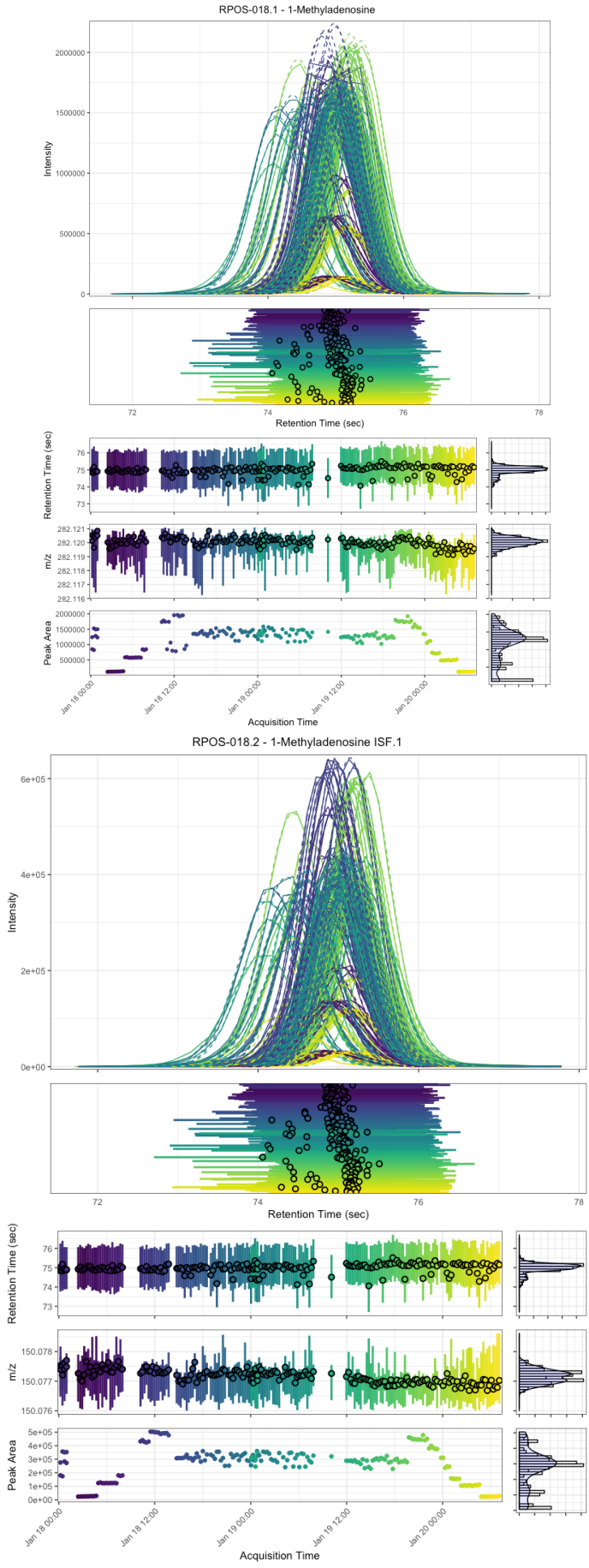


FIGURE B.5: Detection and integration of 2-Methyladenosine main ion and its its in-source fragment.

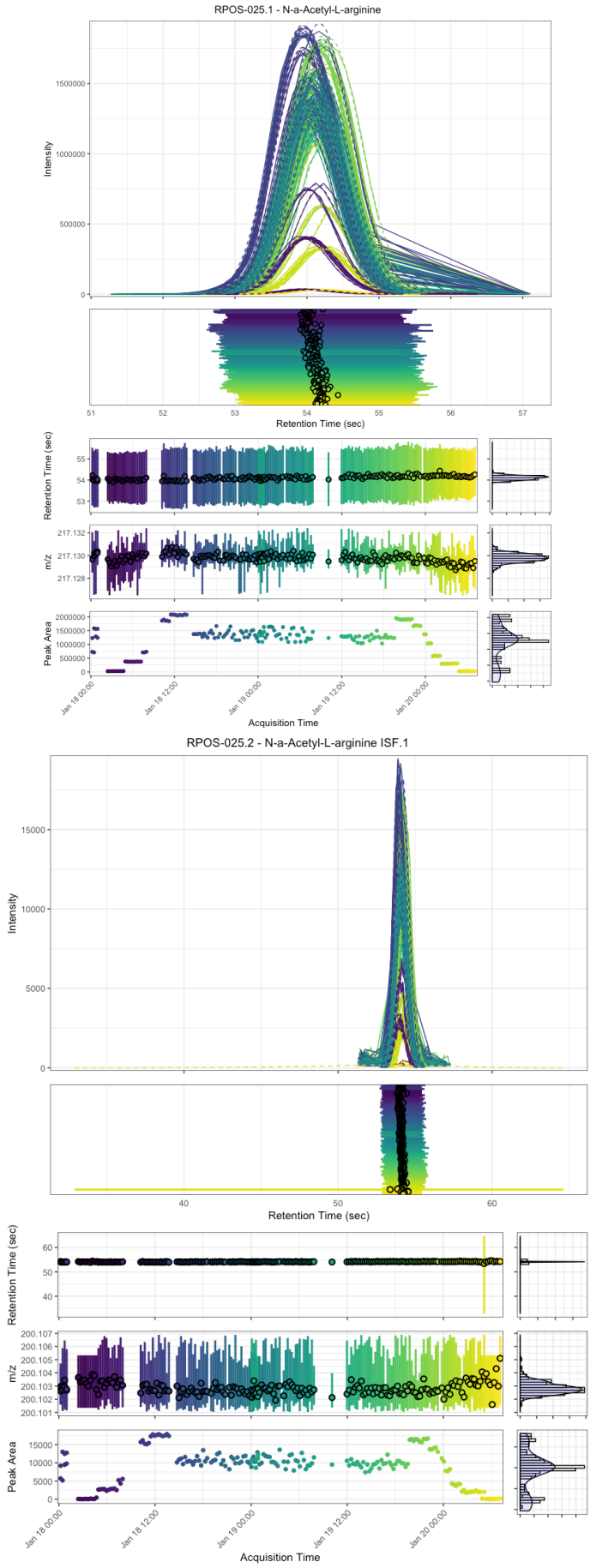


FIGURE B.6: Detection and integration of N-a-Acetyl-L-arginine main ion and its its in-source fragment.

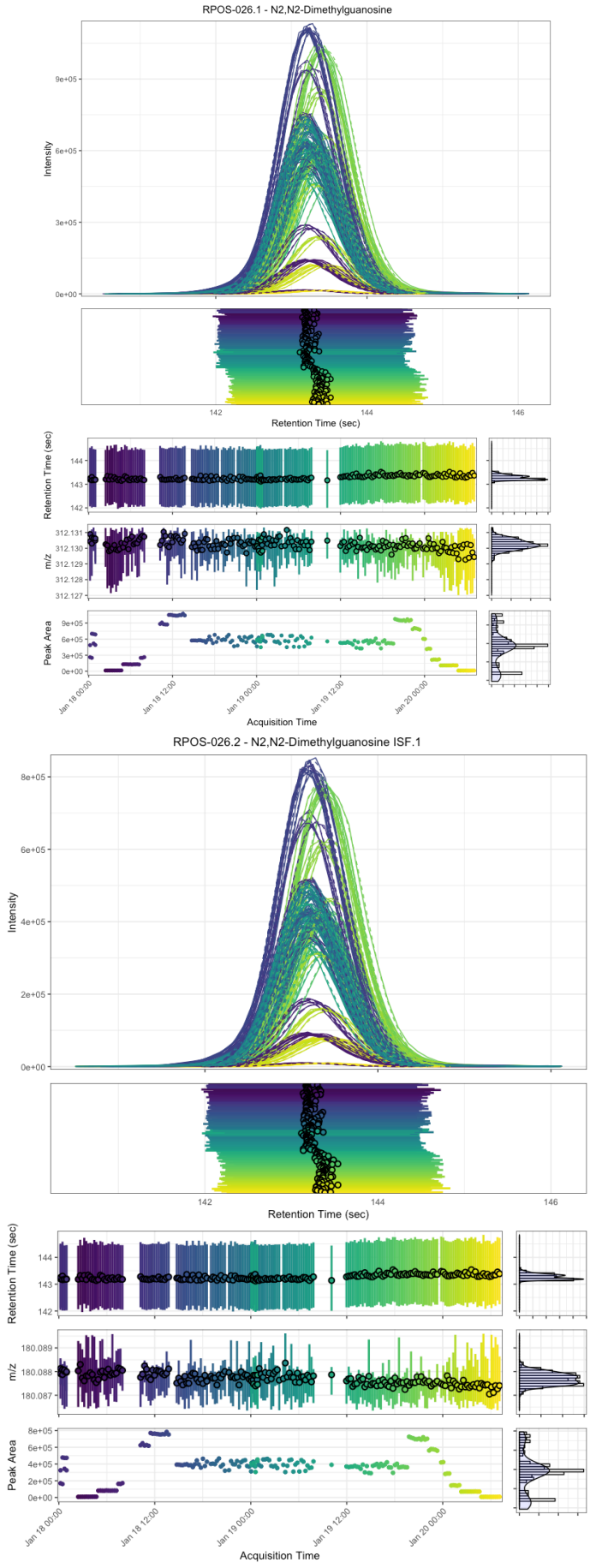


FIGURE B.7: Detection and integration of N2,N2-Dimethylguanosine main ion and its its in-source fragment.

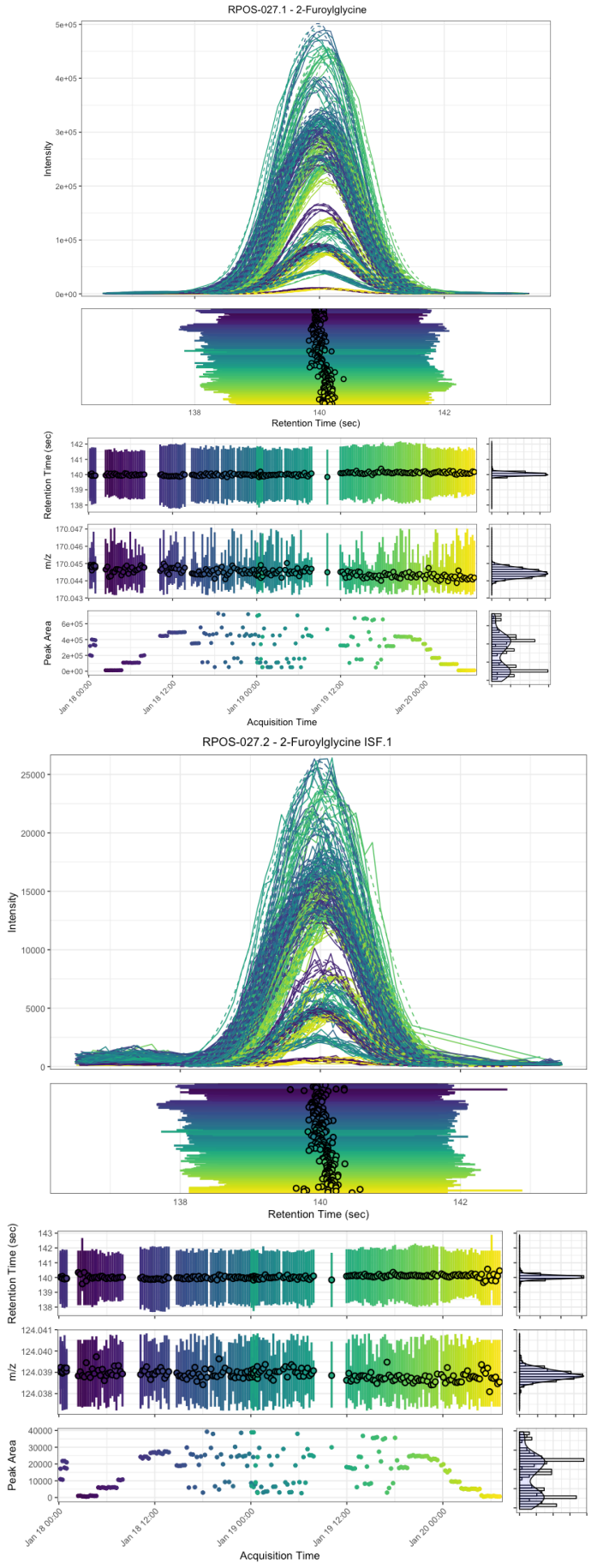


FIGURE B.8: Detection and integration of 2-Furoylglycine main ion and its its in-source fragment.

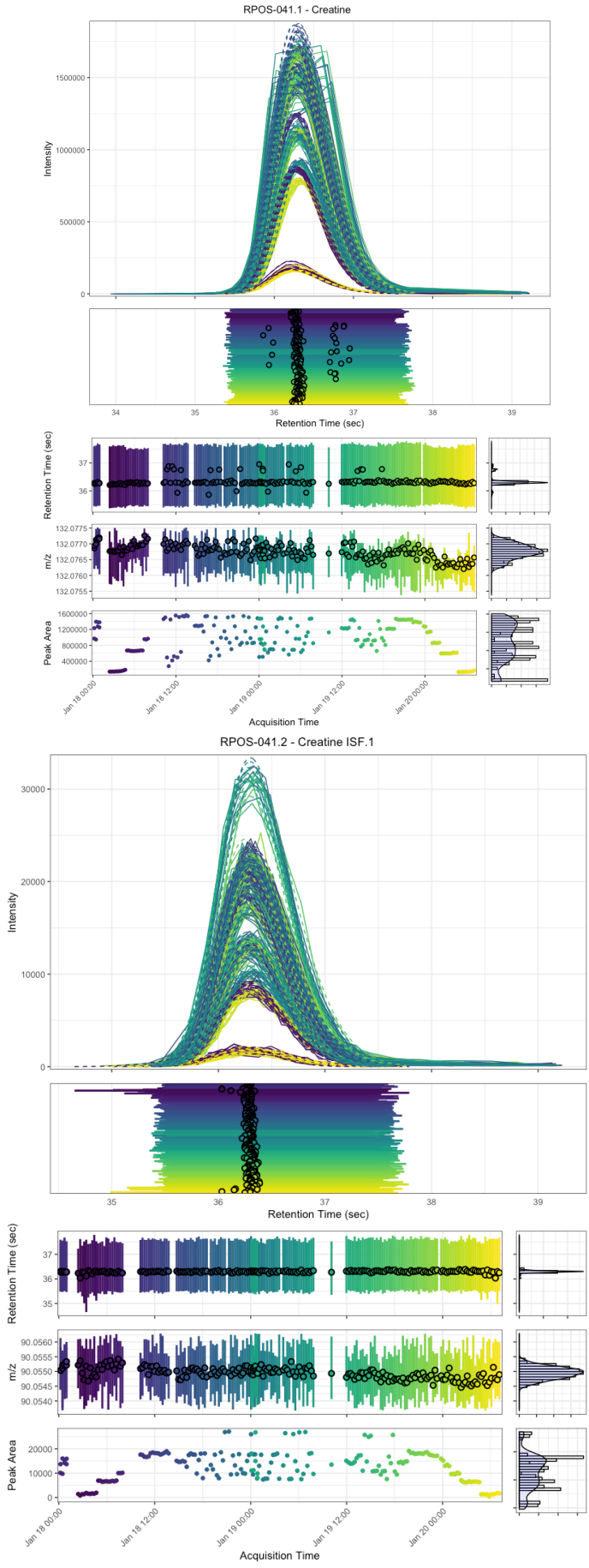


FIGURE B.9: Detection and integration of creatine main ion and its its in-source fragment.

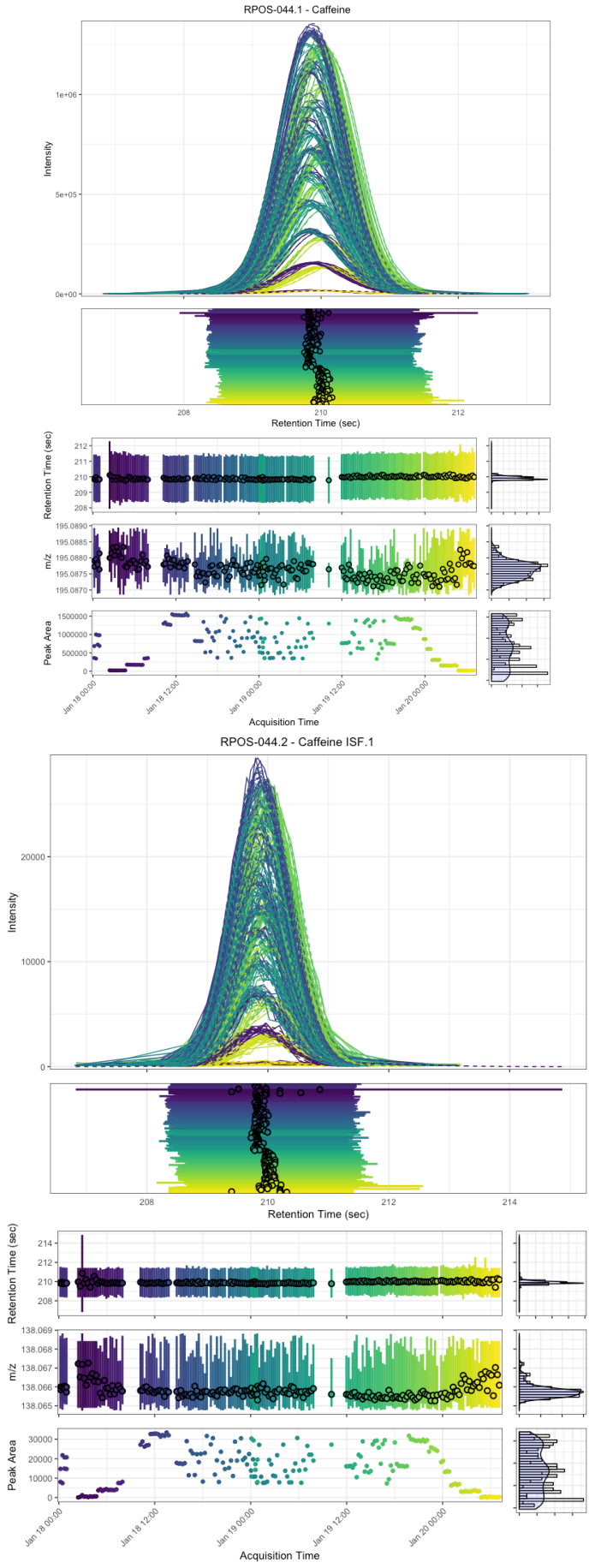


FIGURE B.10: Detection and integration of caffeine main ion and its its in-source fragment.

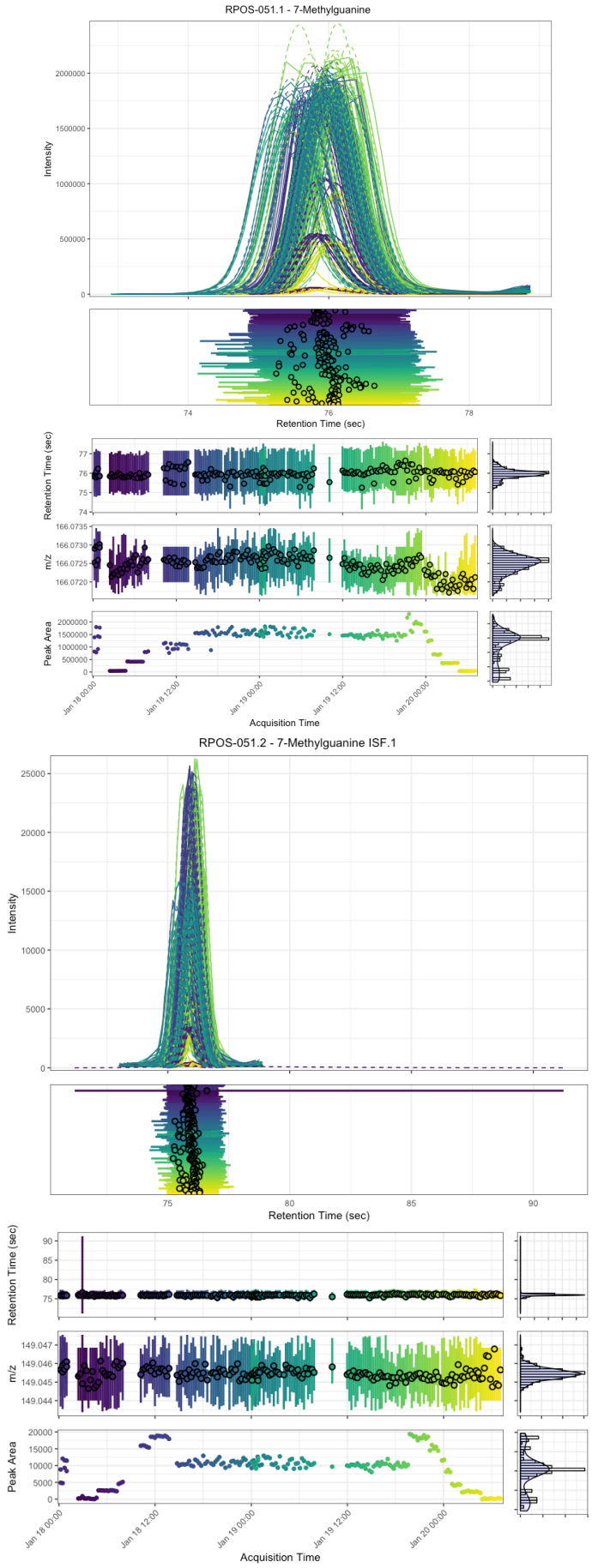


FIGURE B.11: Detection and integration of 7-Methylguanine main ion and its its in-source fragment.

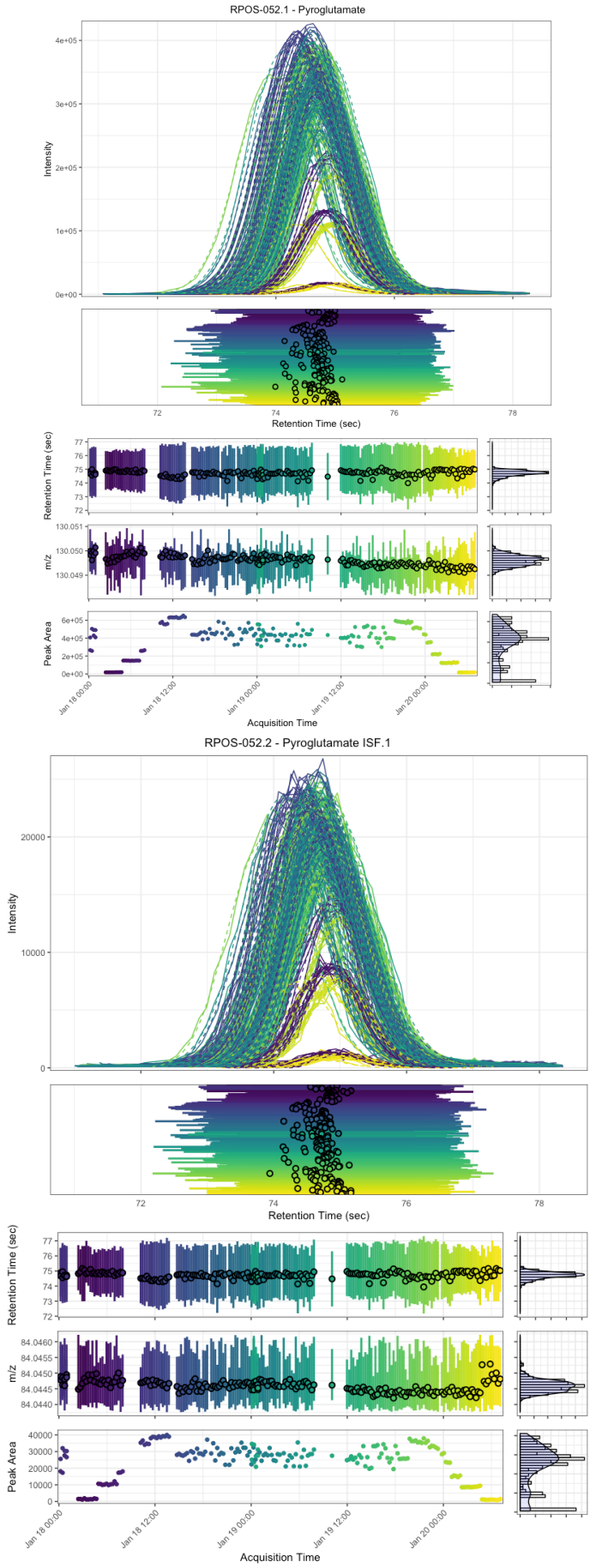


FIGURE B.12: Detection and integration of pyroglutamate main ion and its its in-source fragment.

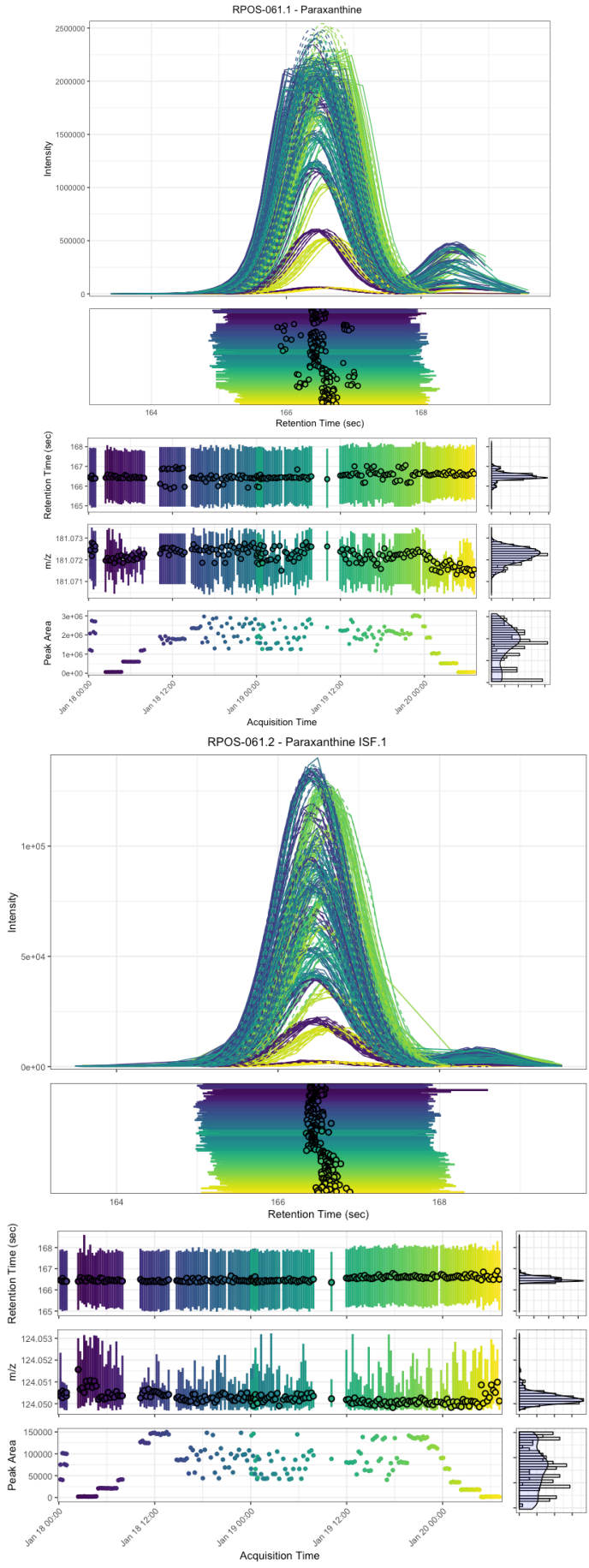


FIGURE B.13: Detection and integration of paraxanthine main ion and its its in-source fragment.

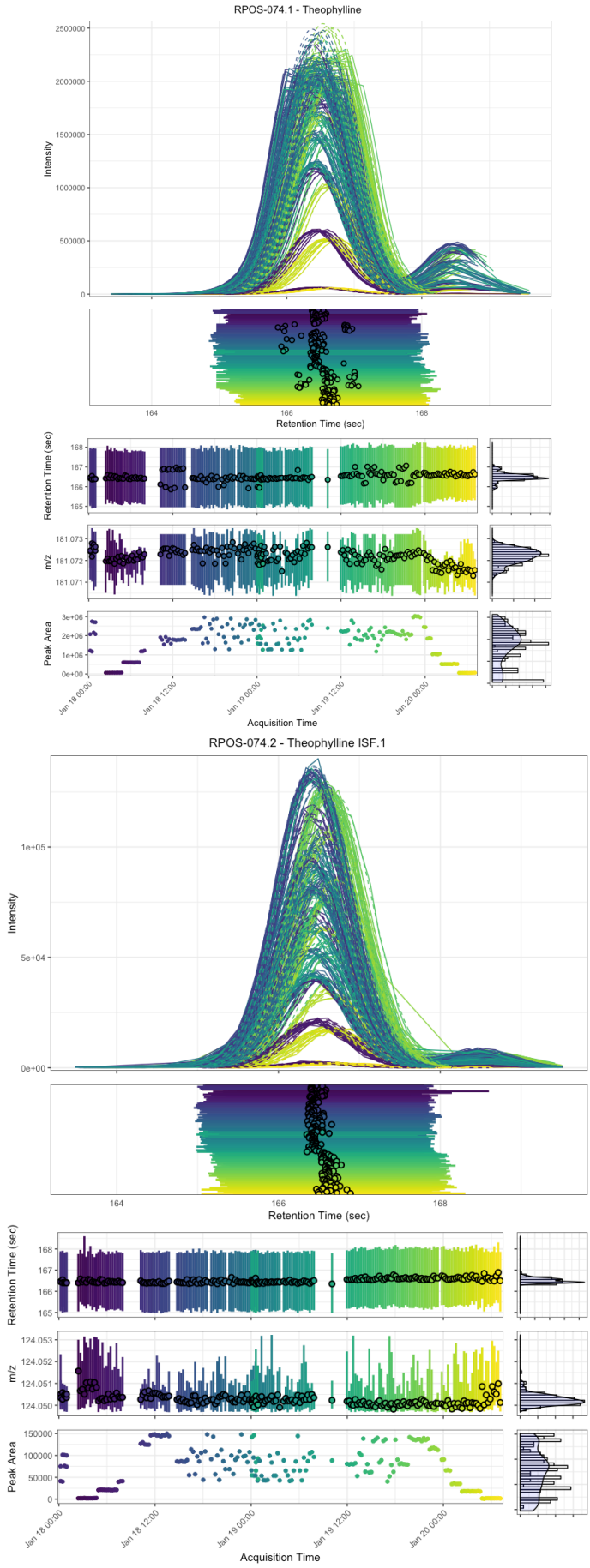


FIGURE B.14: Detection and integration of theophylline main ion and its its in-source fragment.

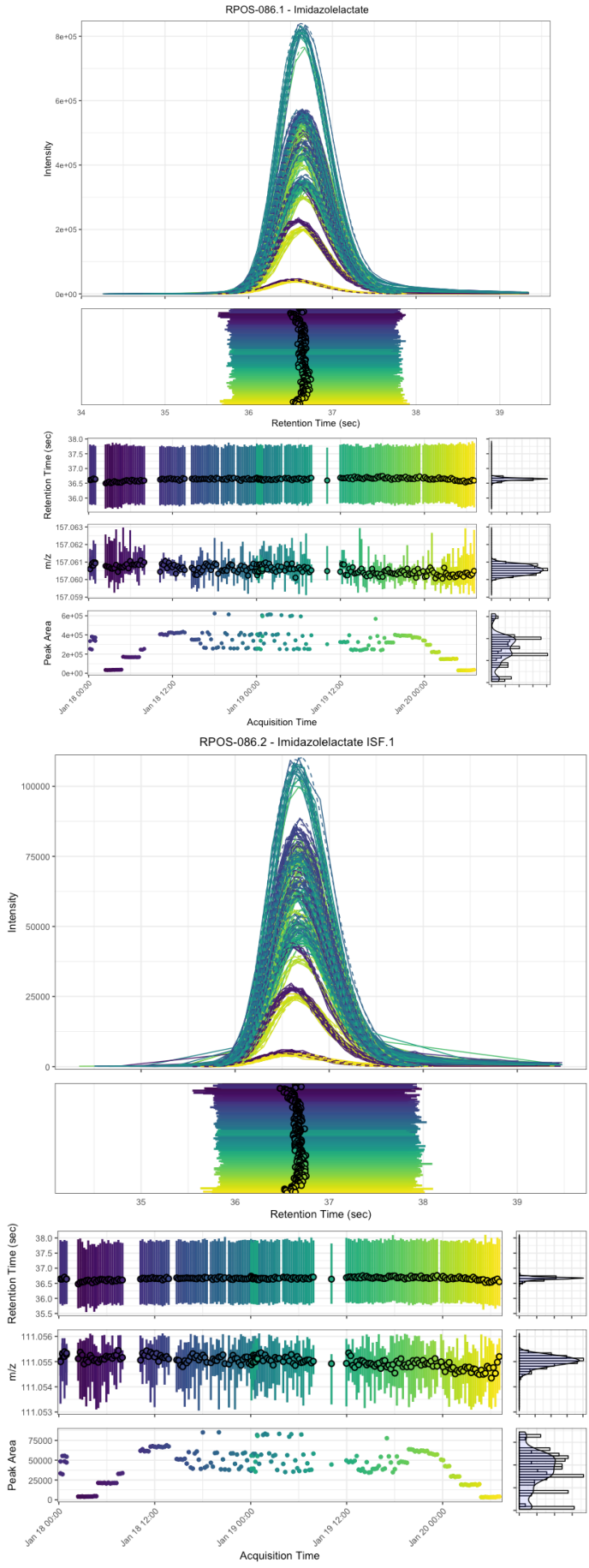


FIGURE B.15: Detection and integration of imidazolelactate main ion and its its in-source fragment.

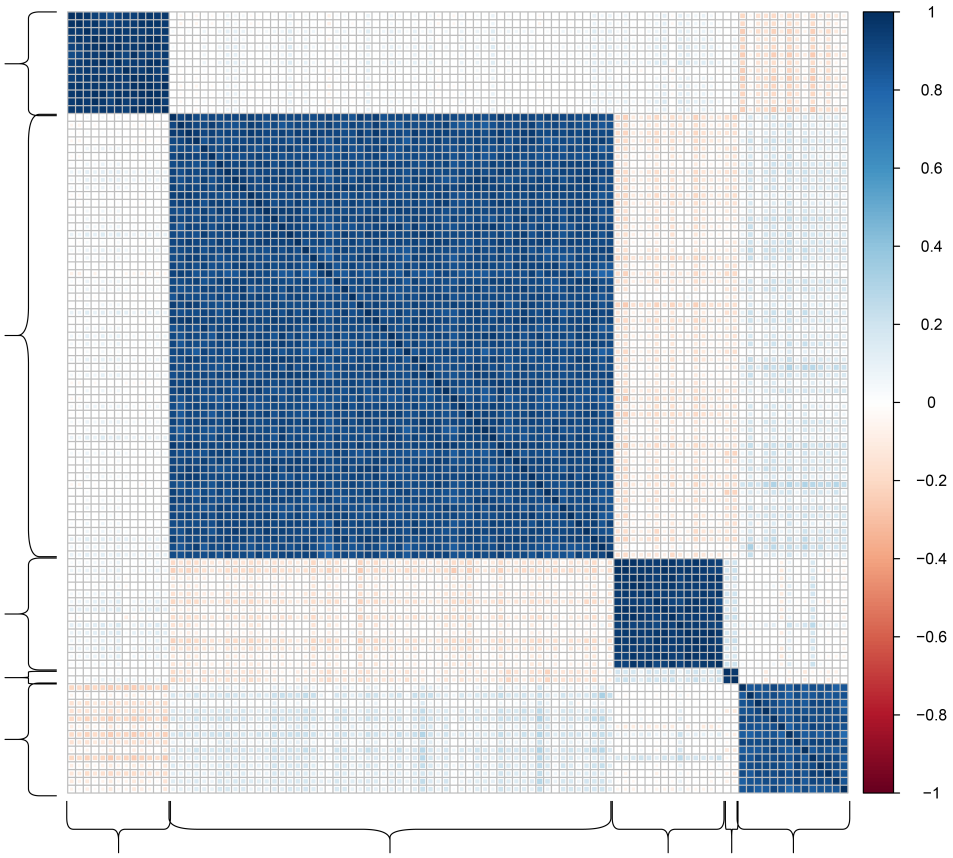


FIGURE B.16: Synthetic datasets were used to assess the performance of the developed feature alignment algorithm. Intensity correlation coefficients obtained between features of five randomly selected pseudo chemical spectra (PCS) in one of the synthetic datasets are visualised as a correlation matrix, with colours corresponding to Pearson correlation coefficients. Lines underneath and on the left-hand side of the matrix indicate feature grouping into the five PSC.

Appendix C

Spectral similarity sensitivity to scaling

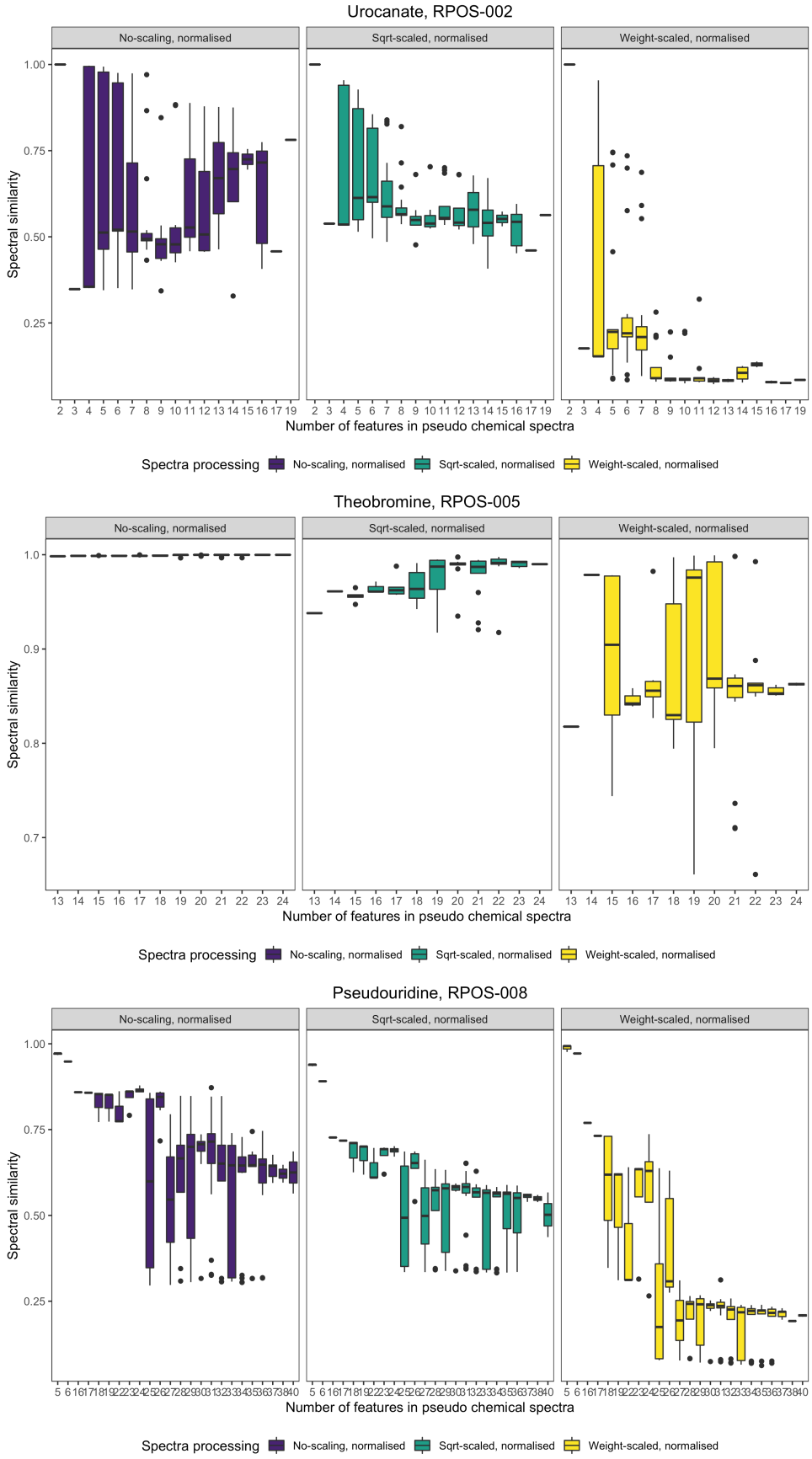


FIGURE C.1: Spectral similarity variation for pseudo chemical spectra containing different metabolite adducts.

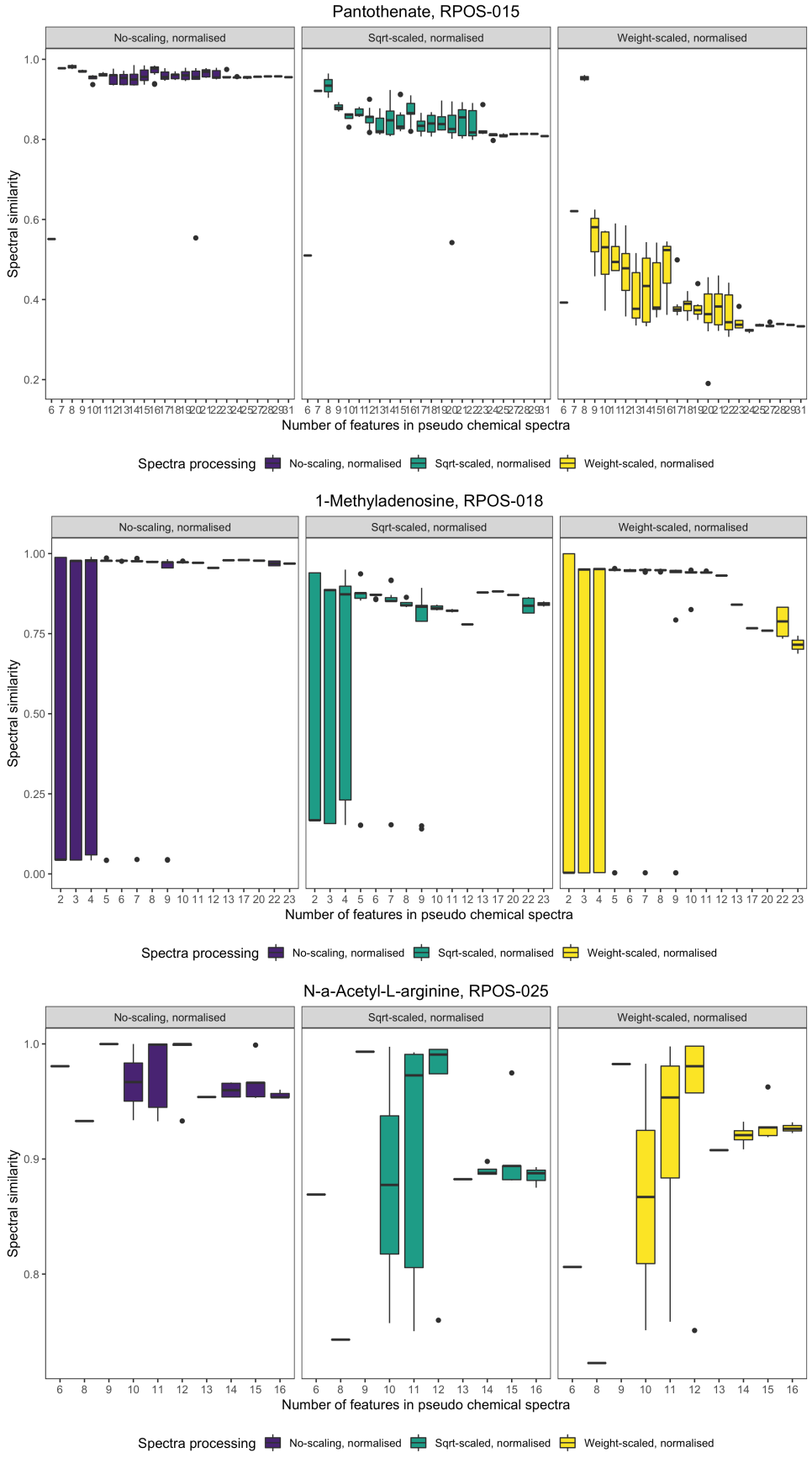


FIGURE C.2: Spectral similarity variation for pseudo chemical spectra containing different metabolite adducts.

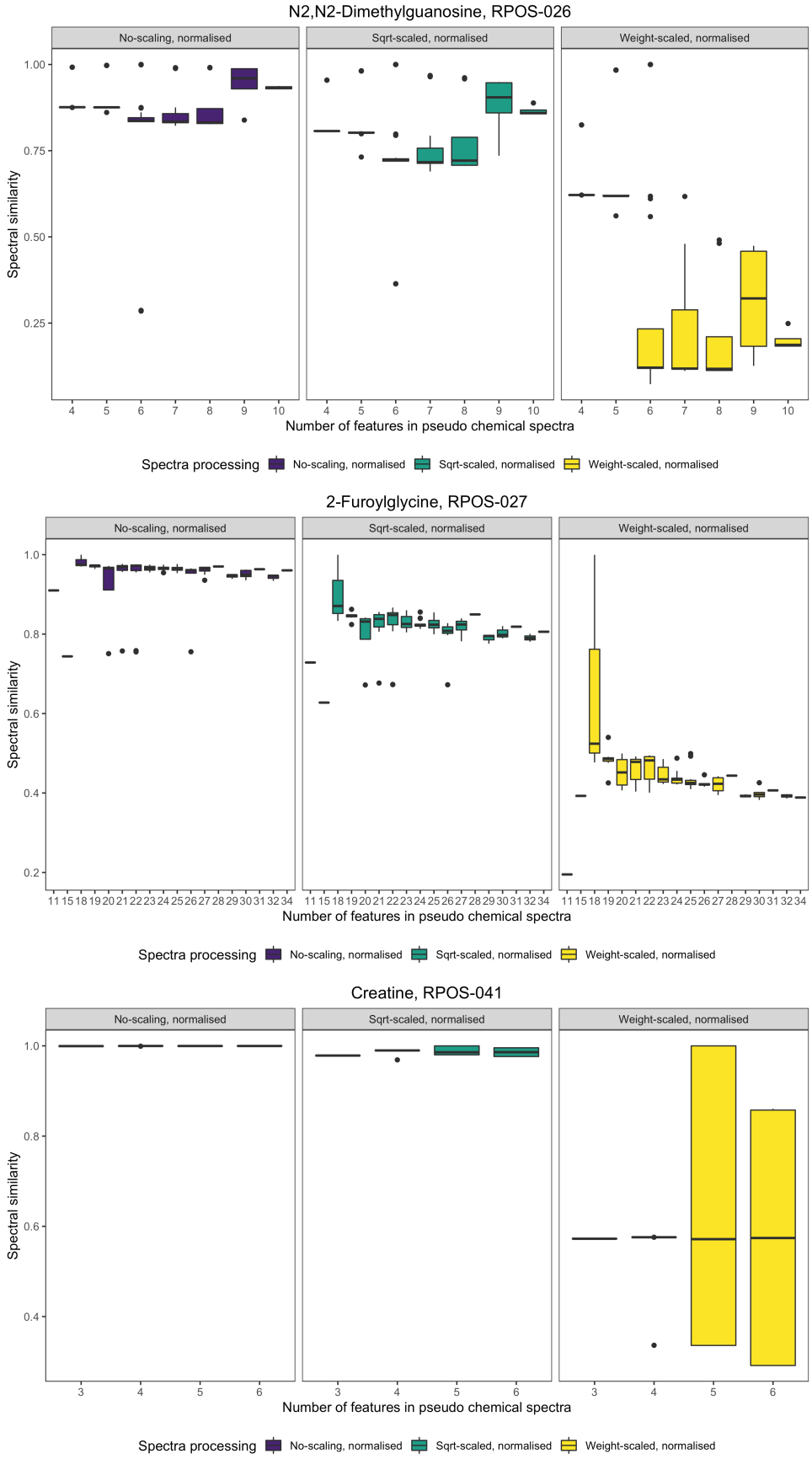


FIGURE C.3: Spectral similarity variation for pseudo chemical spectra containing different metabolite adducts.

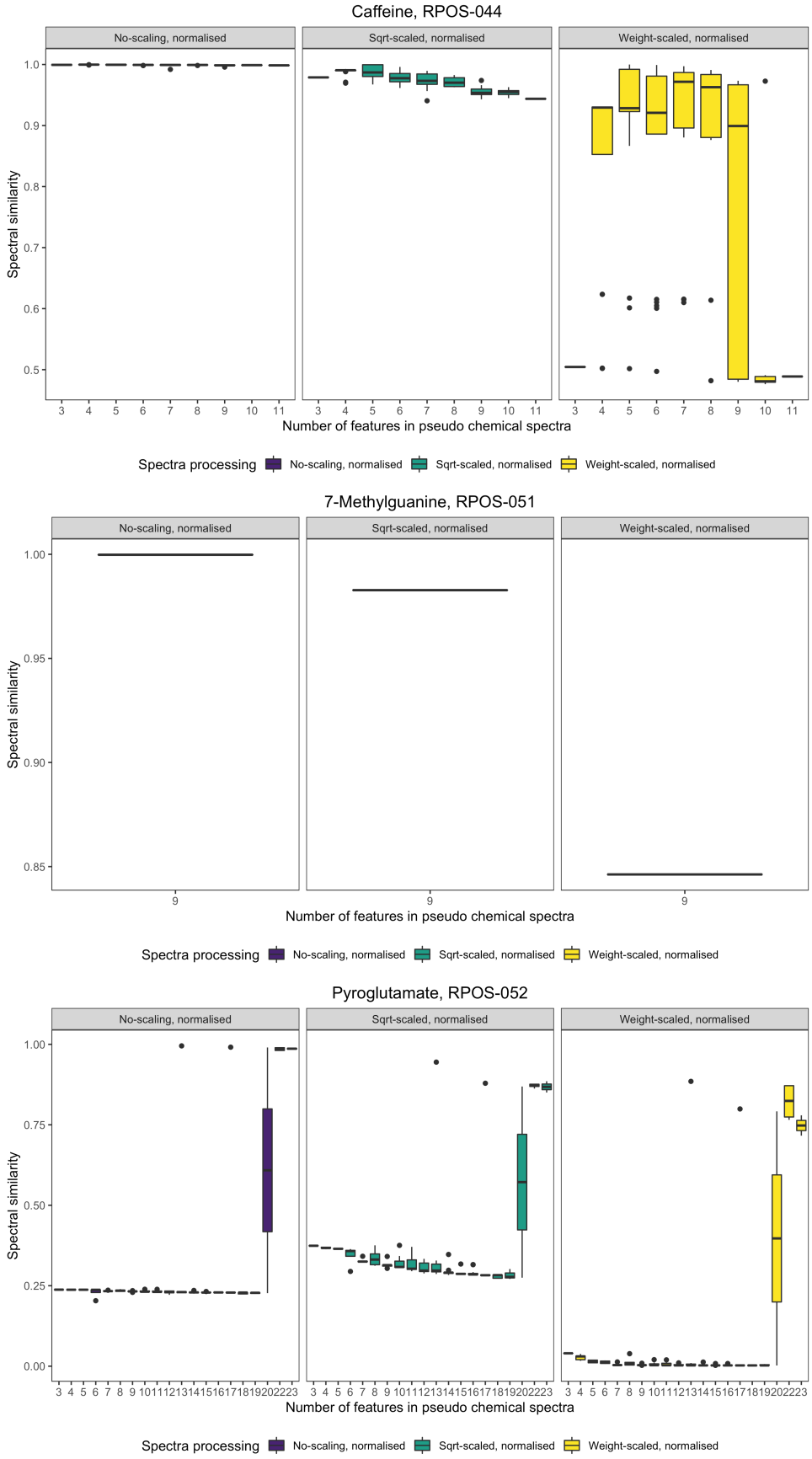


FIGURE C.4: Spectral similarity variation for pseudo chemical spectra containing different metabolite adducts.

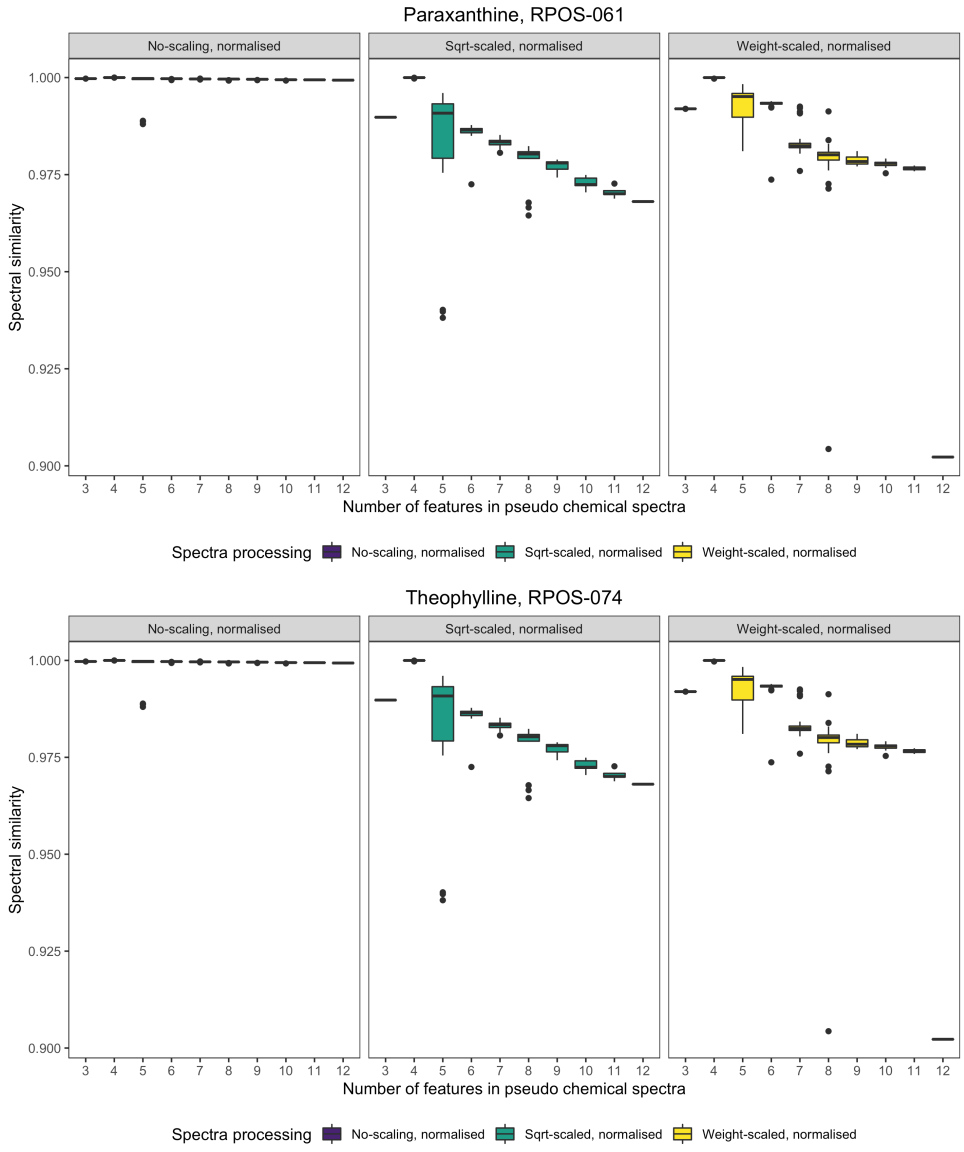


FIGURE C.5: Spectral similarity variation for pseudo chemical spectra containing different metabolite adducts.

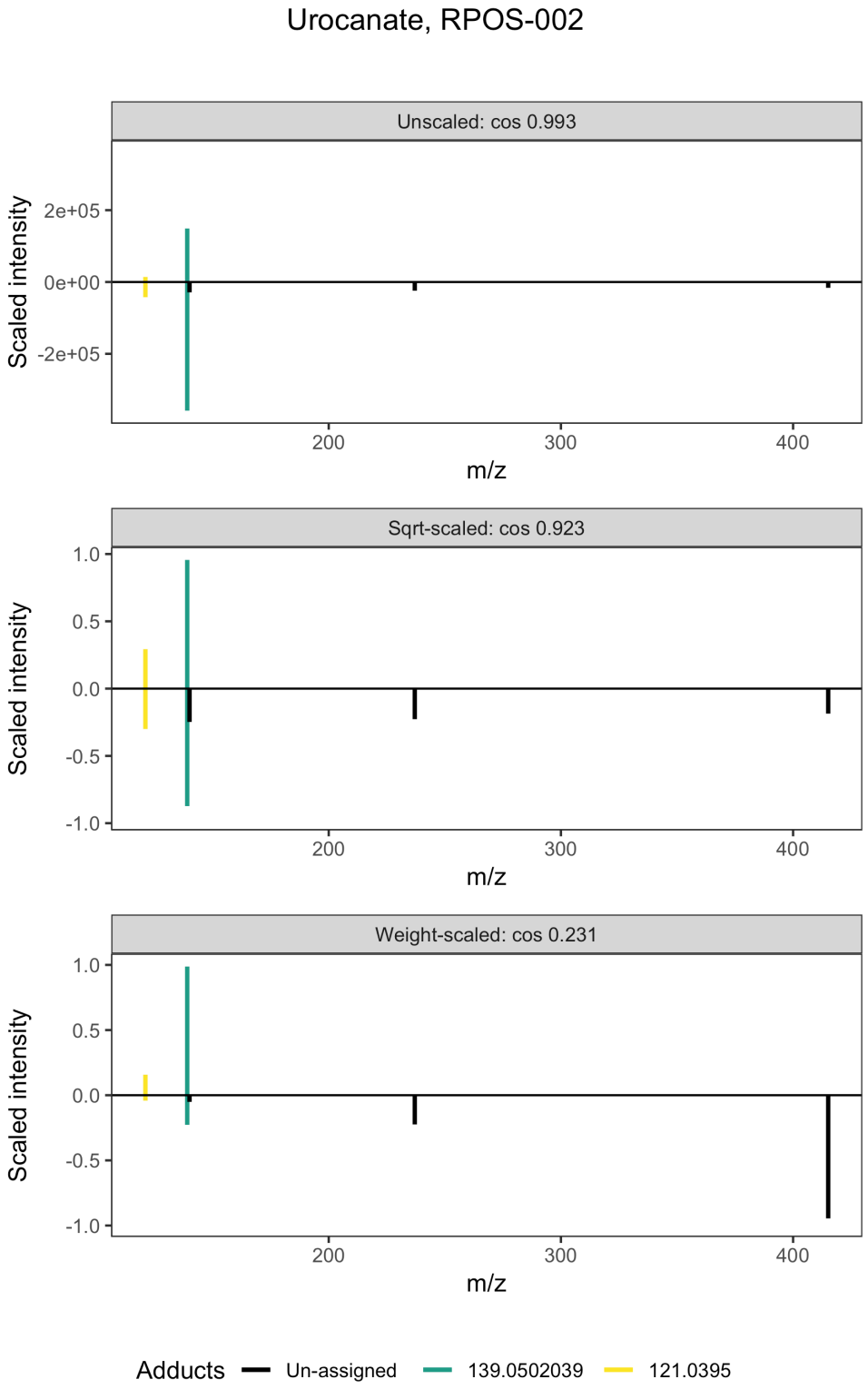


FIGURE C.6: Spectral similarity score between PCS containing two Urocanate ions varies depending on spectra scaling method.

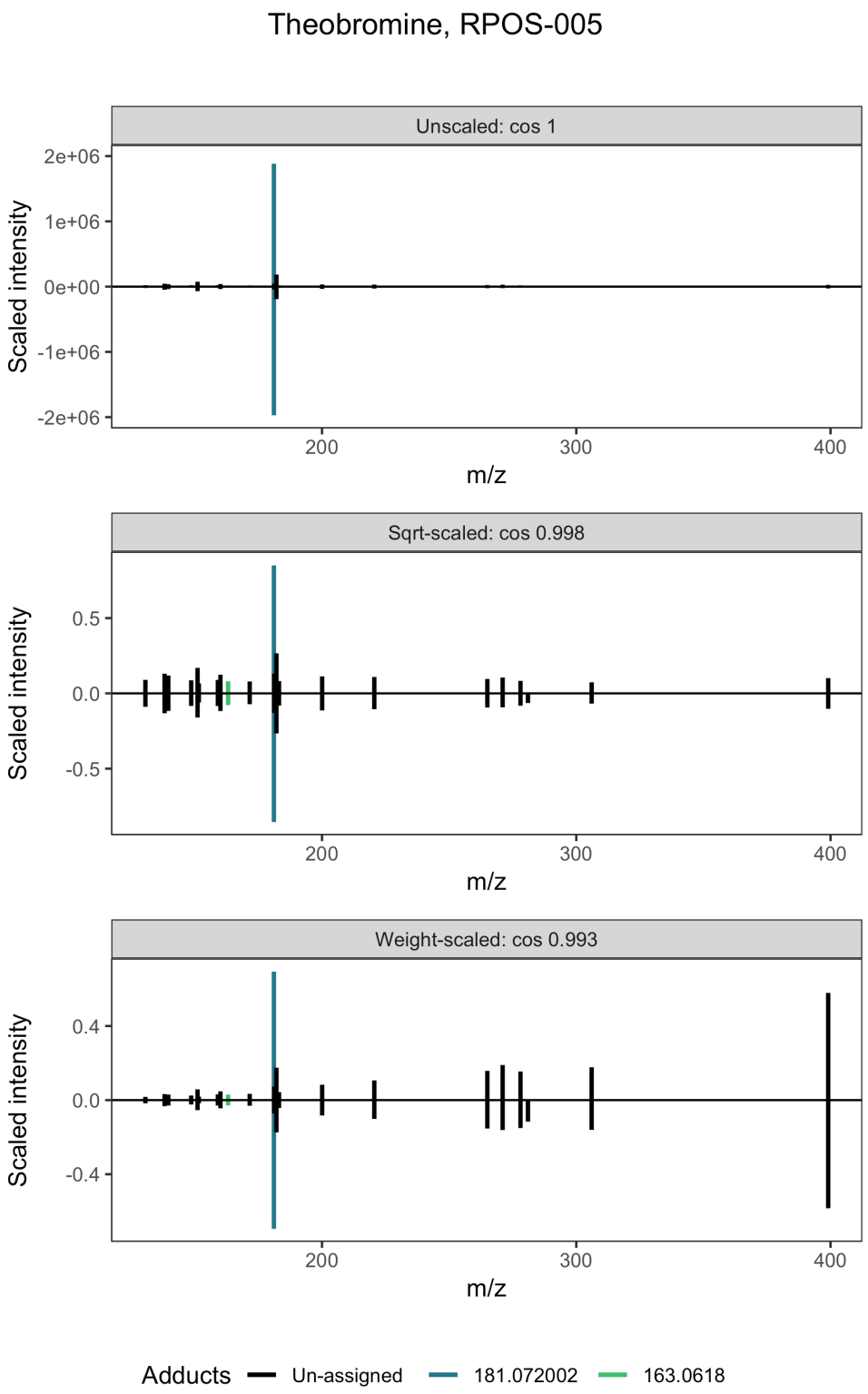


FIGURE C.7: Spectral similarity score between PCS containing two theobromine ions varies depending on spectra scaling method.

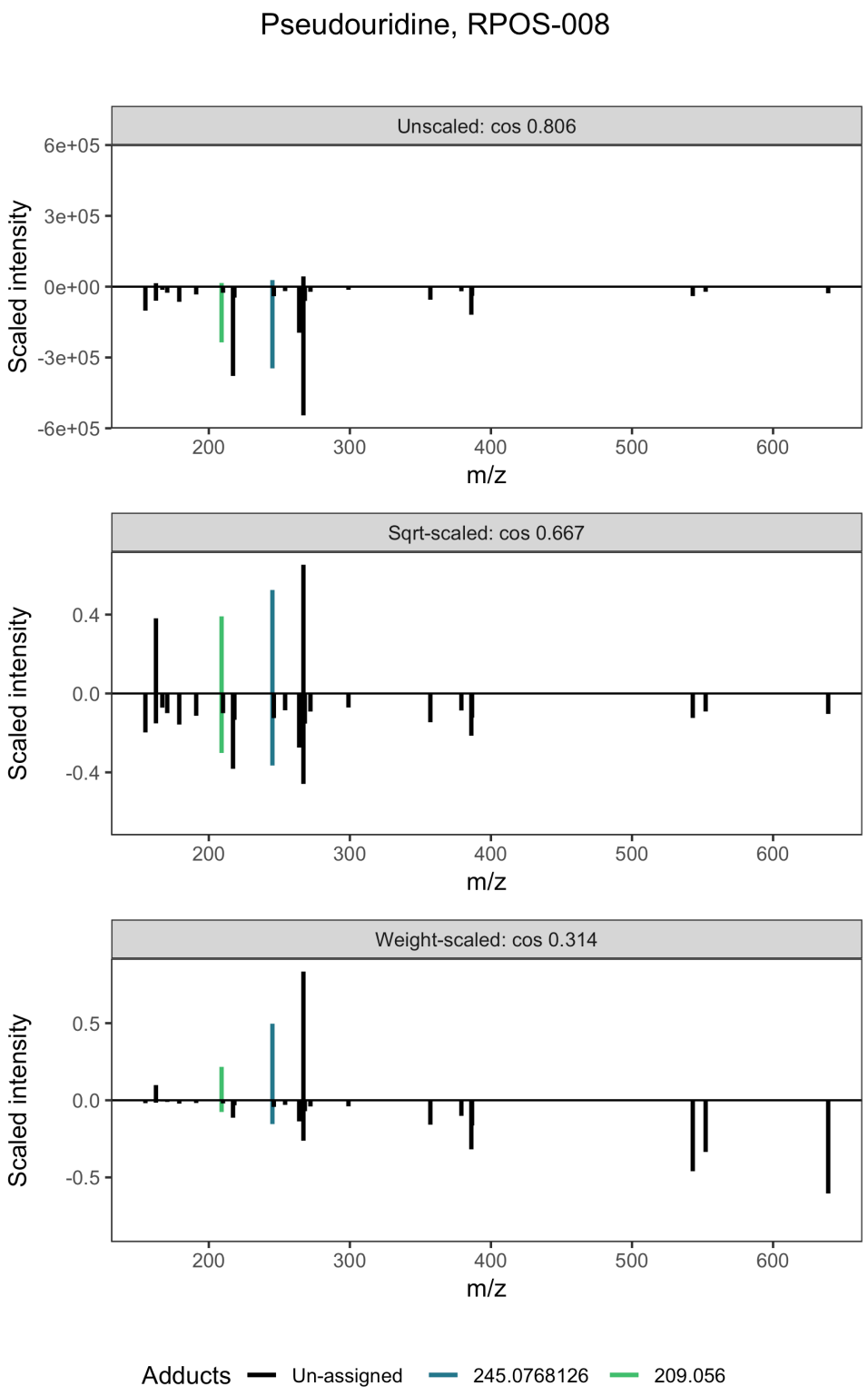


FIGURE C.8: Spectral similarity score between PCS containing two pseudouridine ions varies depending on spectra scaling method.

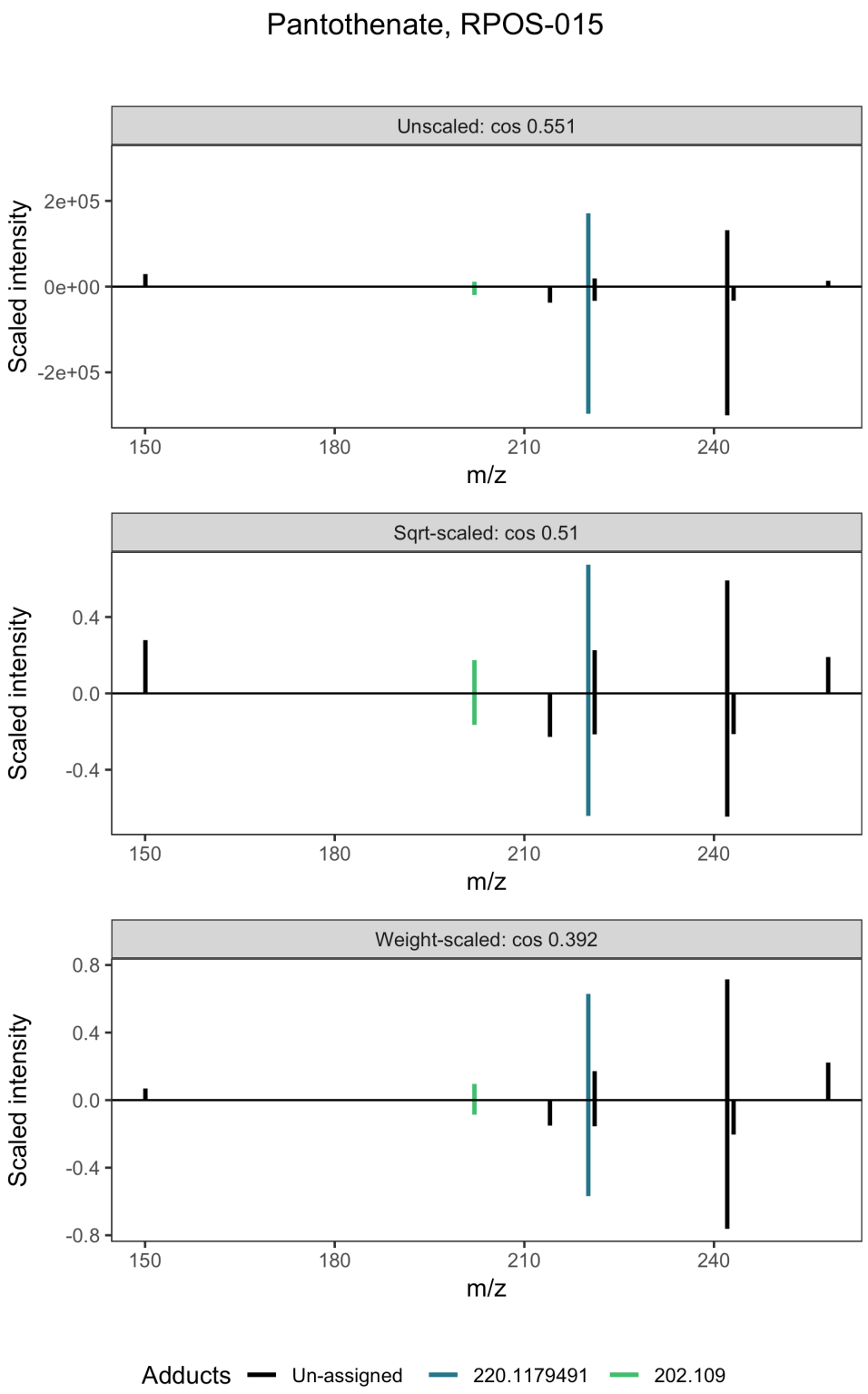


FIGURE C.9: Spectral similarity score between PCS containing two pantothenate ions varies depending on spectra scaling method.

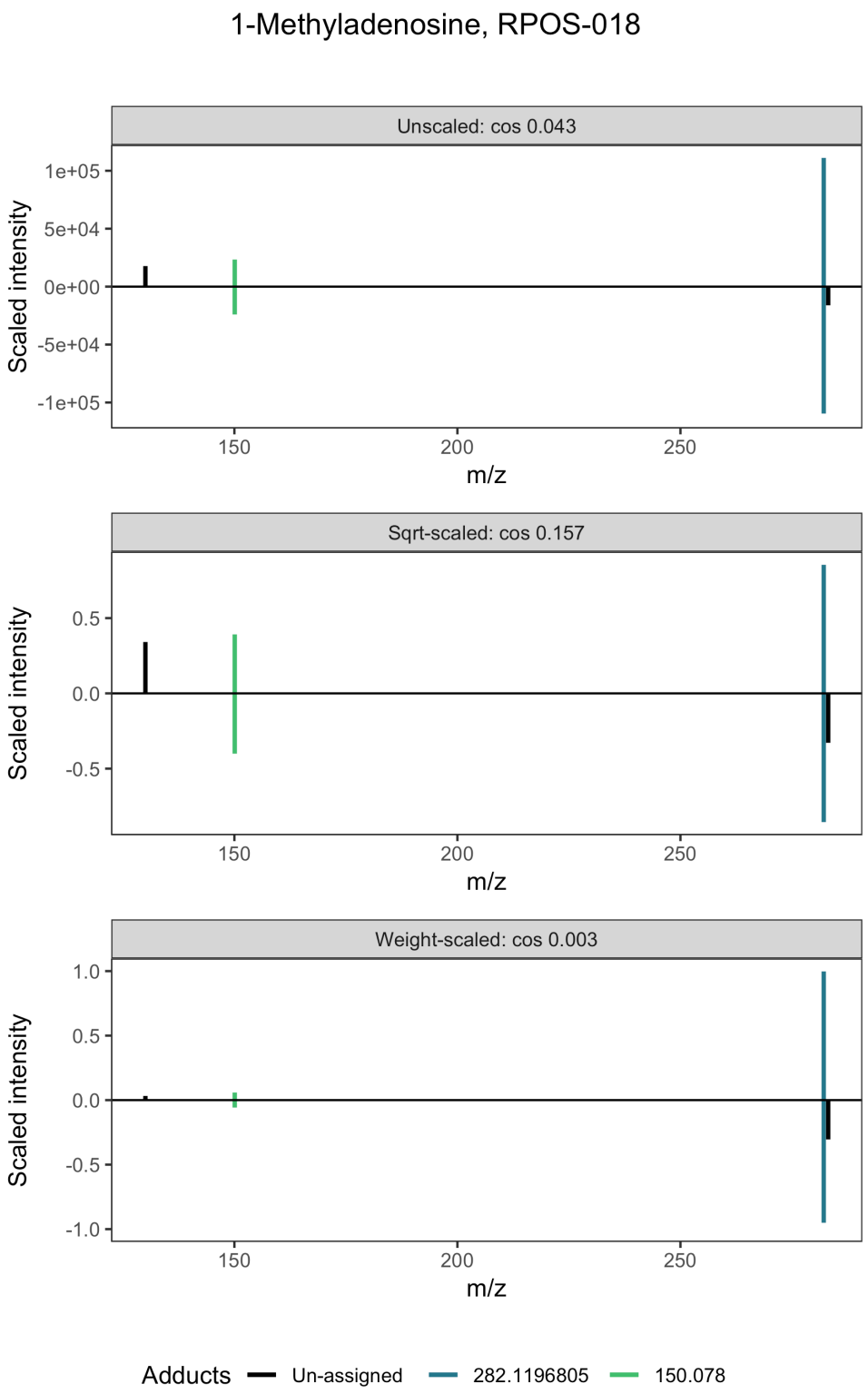


FIGURE C.10: Spectral similarity score between PCS containing two 1-Methyladenosine ions varies depending on spectra scaling method.

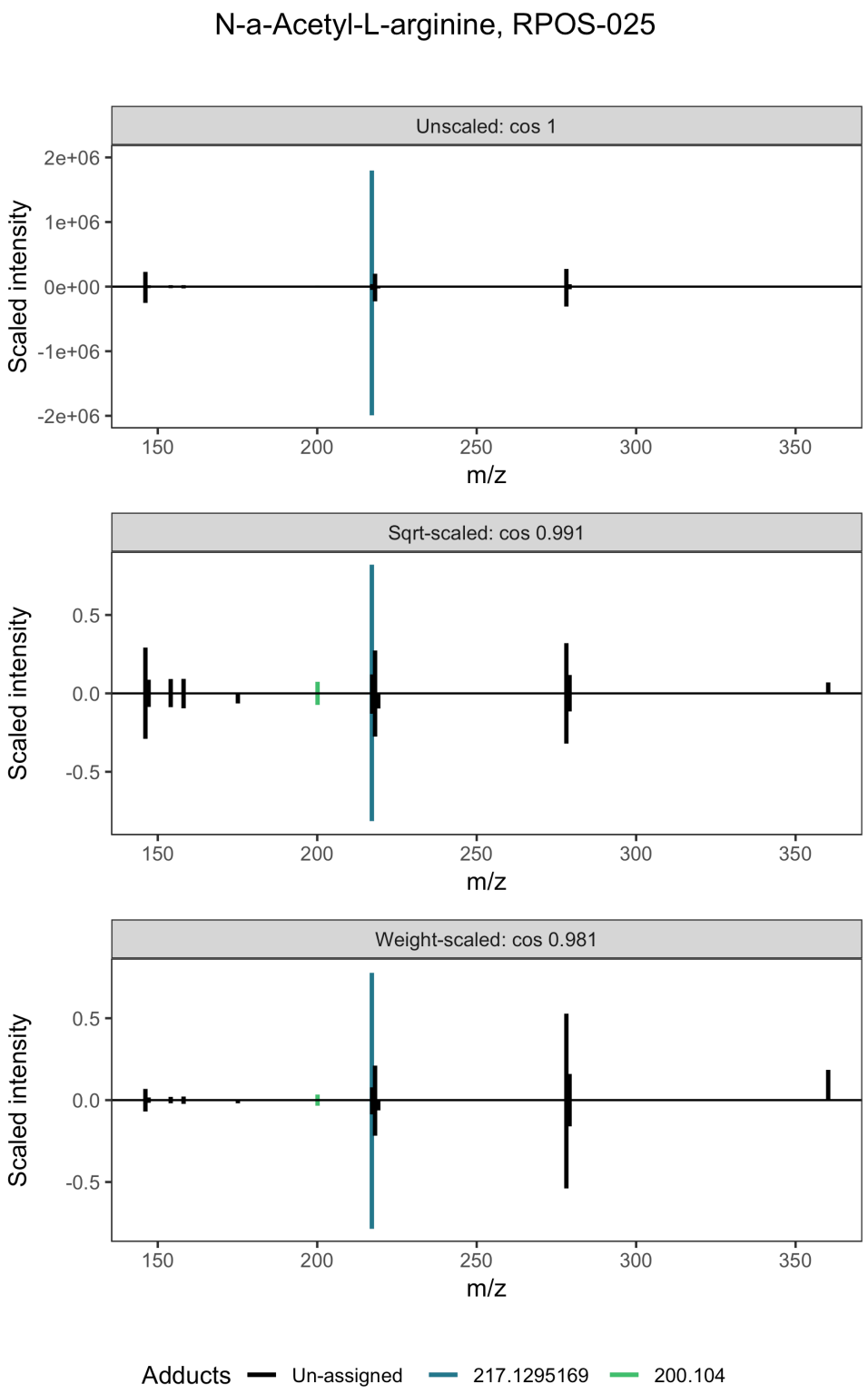


FIGURE C.11: Spectral similarity score between PCS containing two N-a-Acetyl-L-arginine ions varies depending on spectra scaling method.

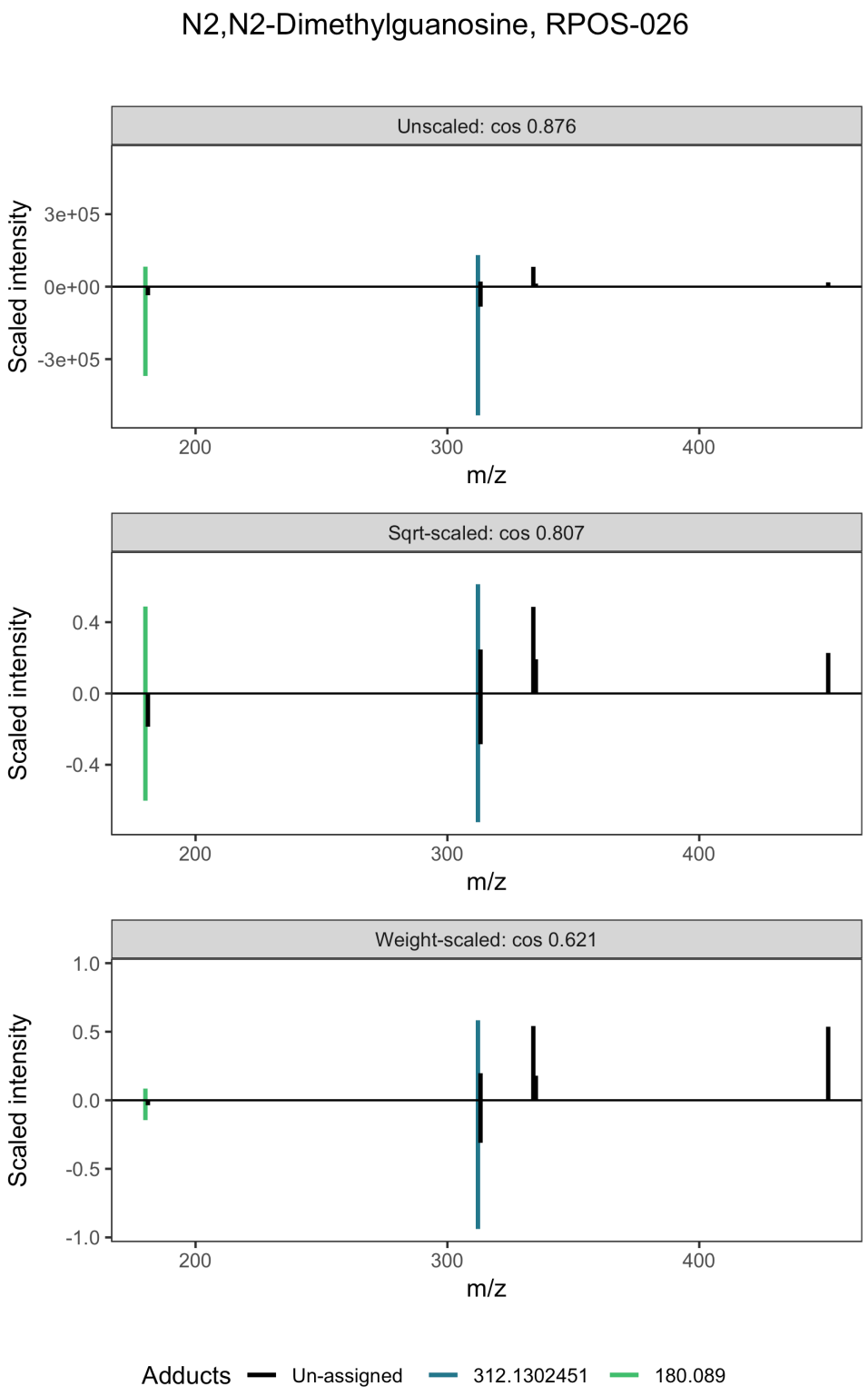


FIGURE C.12: Spectral similarity score between PCS containing two N2,N2-Dimethylguanosine ions varies depending on spectra scaling method.

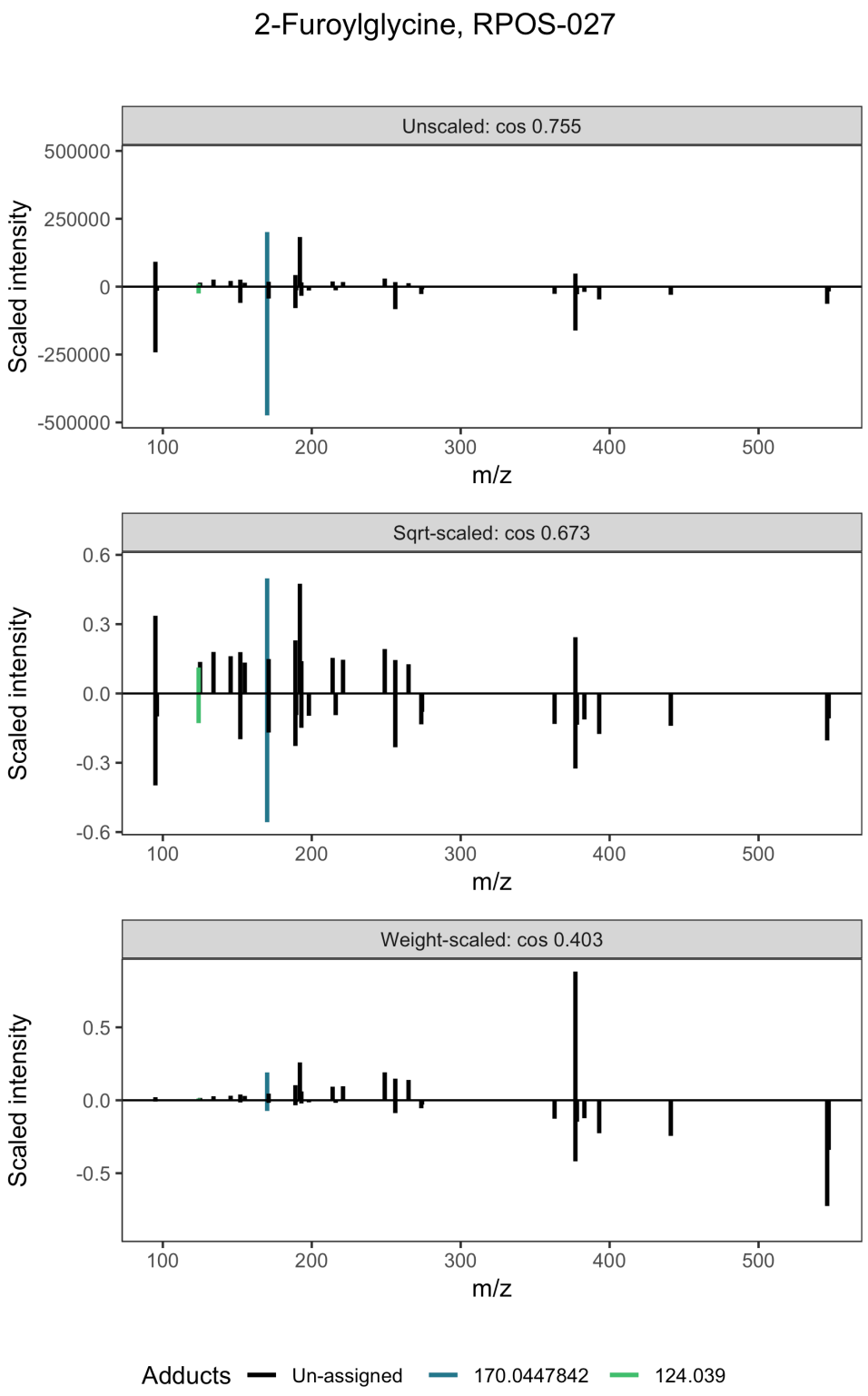


FIGURE C.13: Spectral similarity score between PCS containing two 2-Furoylglycine ions varies depending on spectra scaling method.

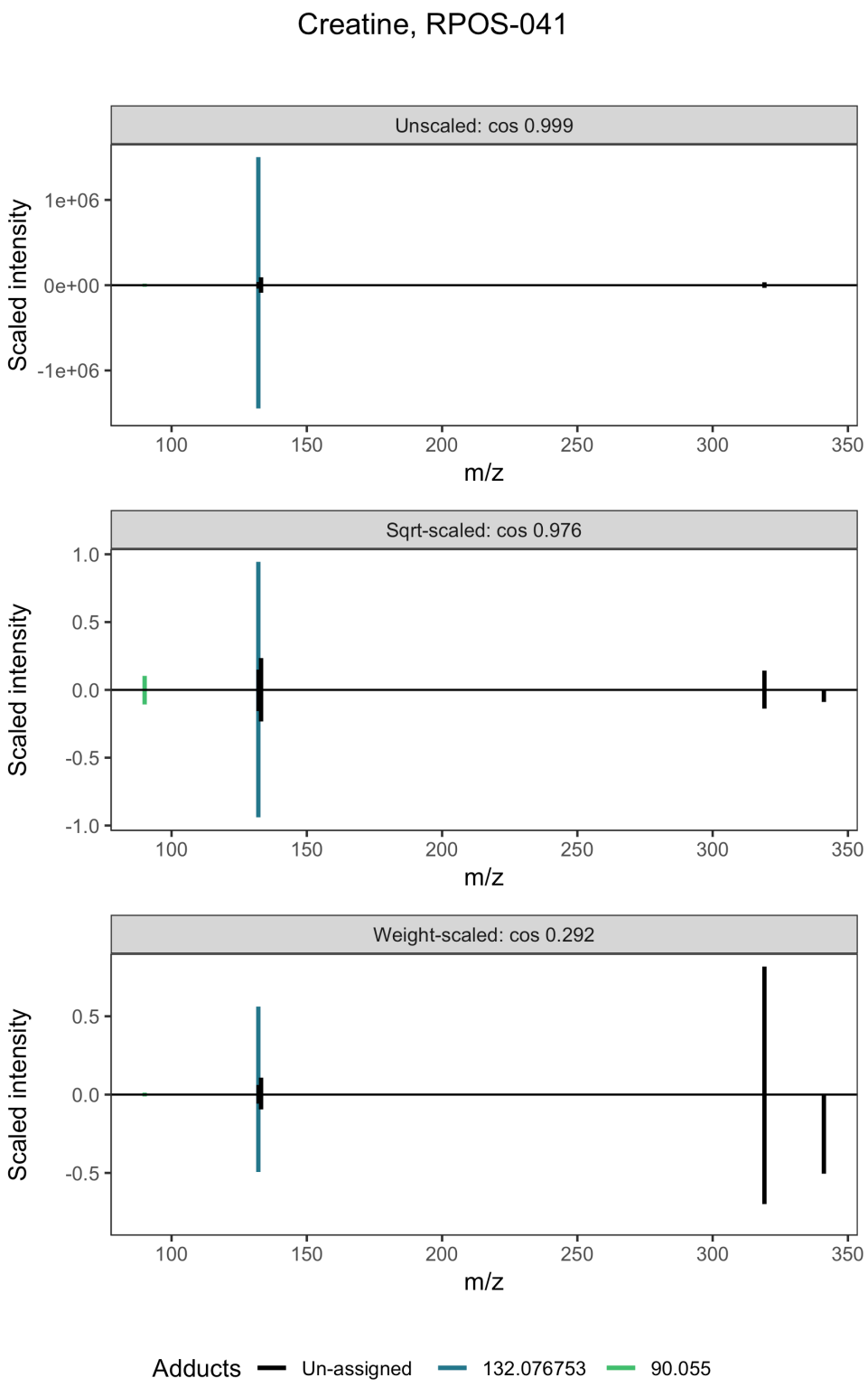


FIGURE C.14: Spectral similarity score between PCS containing two Creatine ions varies depending on spectra scaling method.

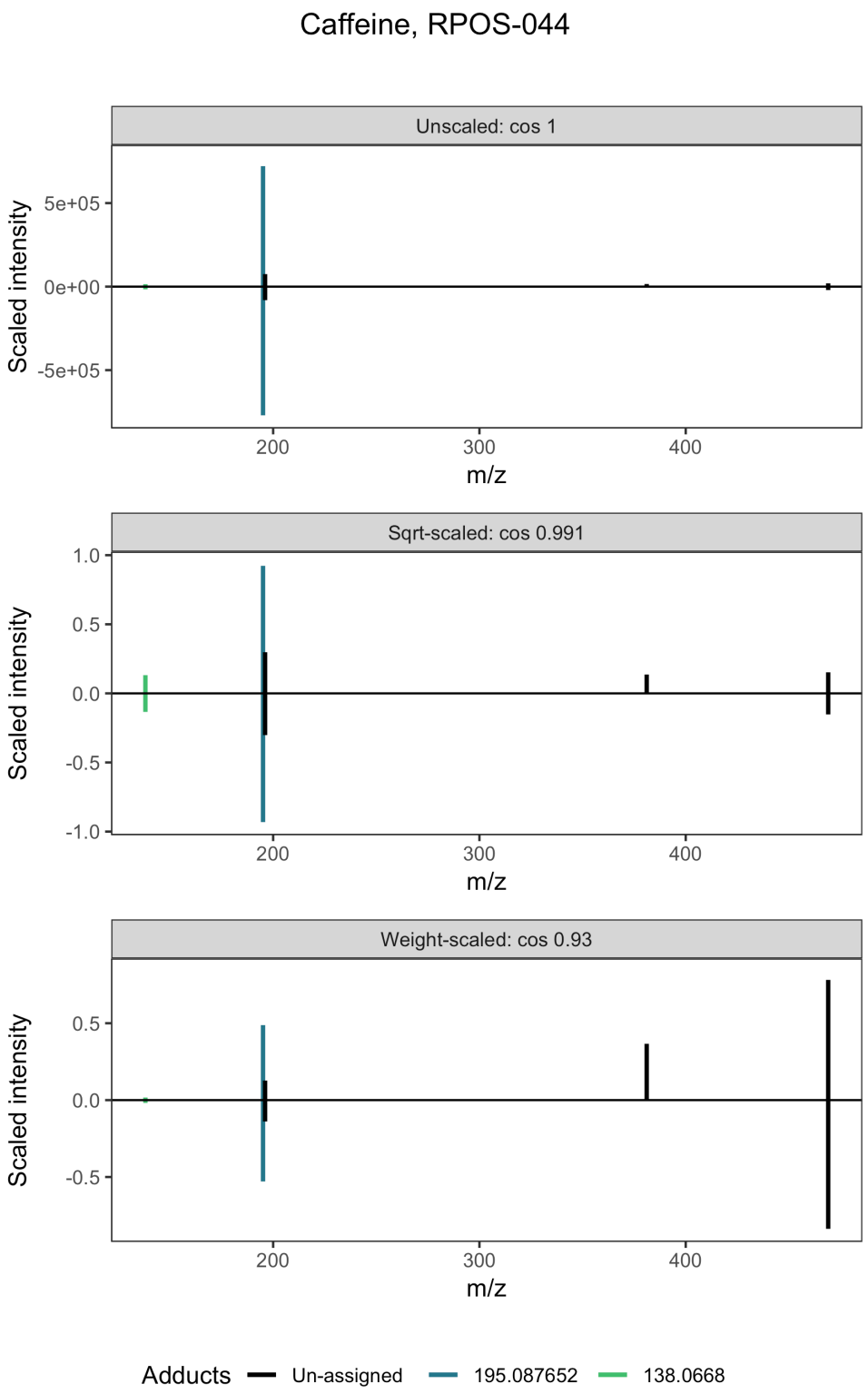


FIGURE C.15: Spectral similarity score between PCS containing two Caffeine ions varies depending on spectra scaling method.

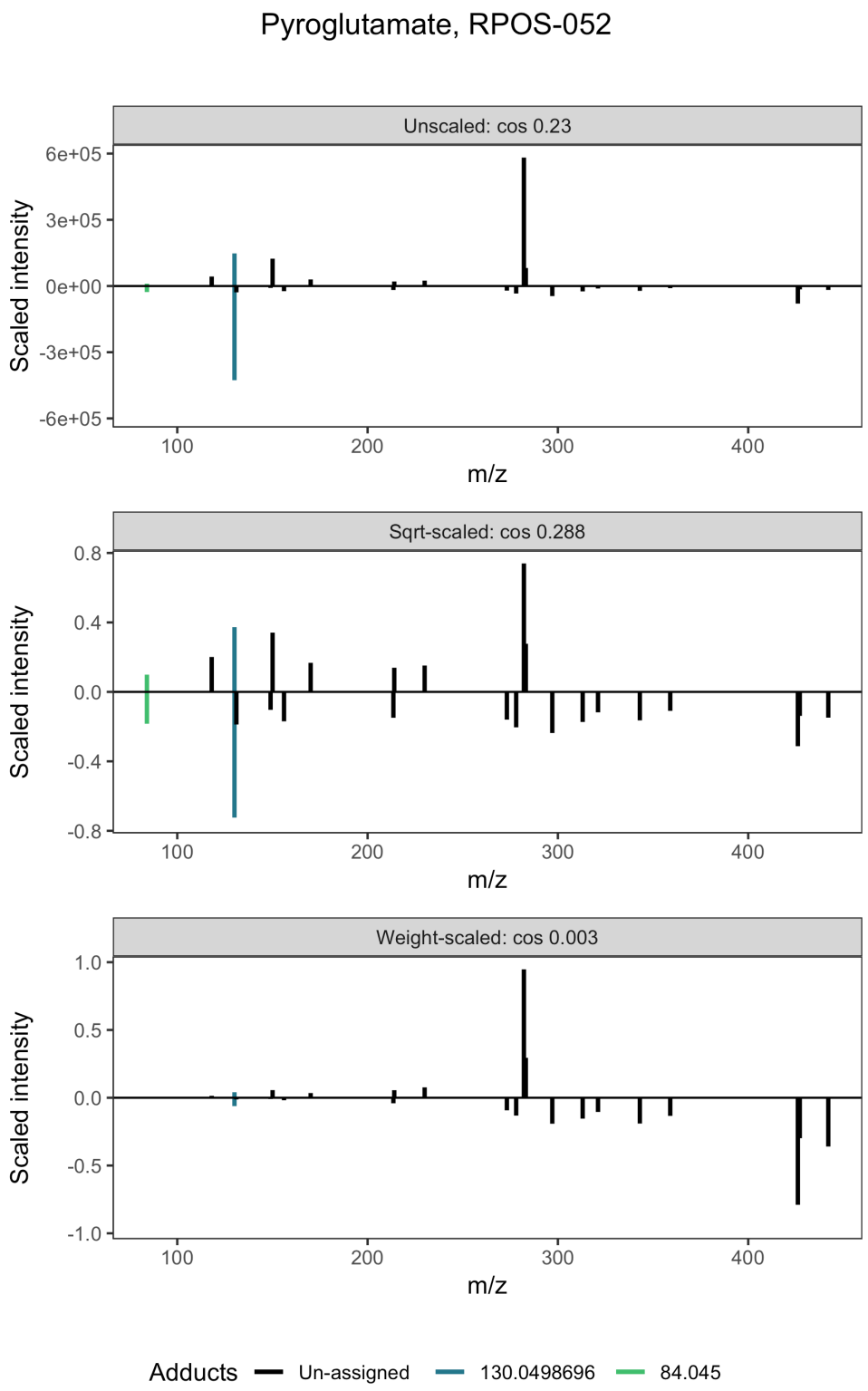


FIGURE C.16: Spectral similarity score between PCS containing two Pyroglutamate ions varies depending on spectra scaling method.

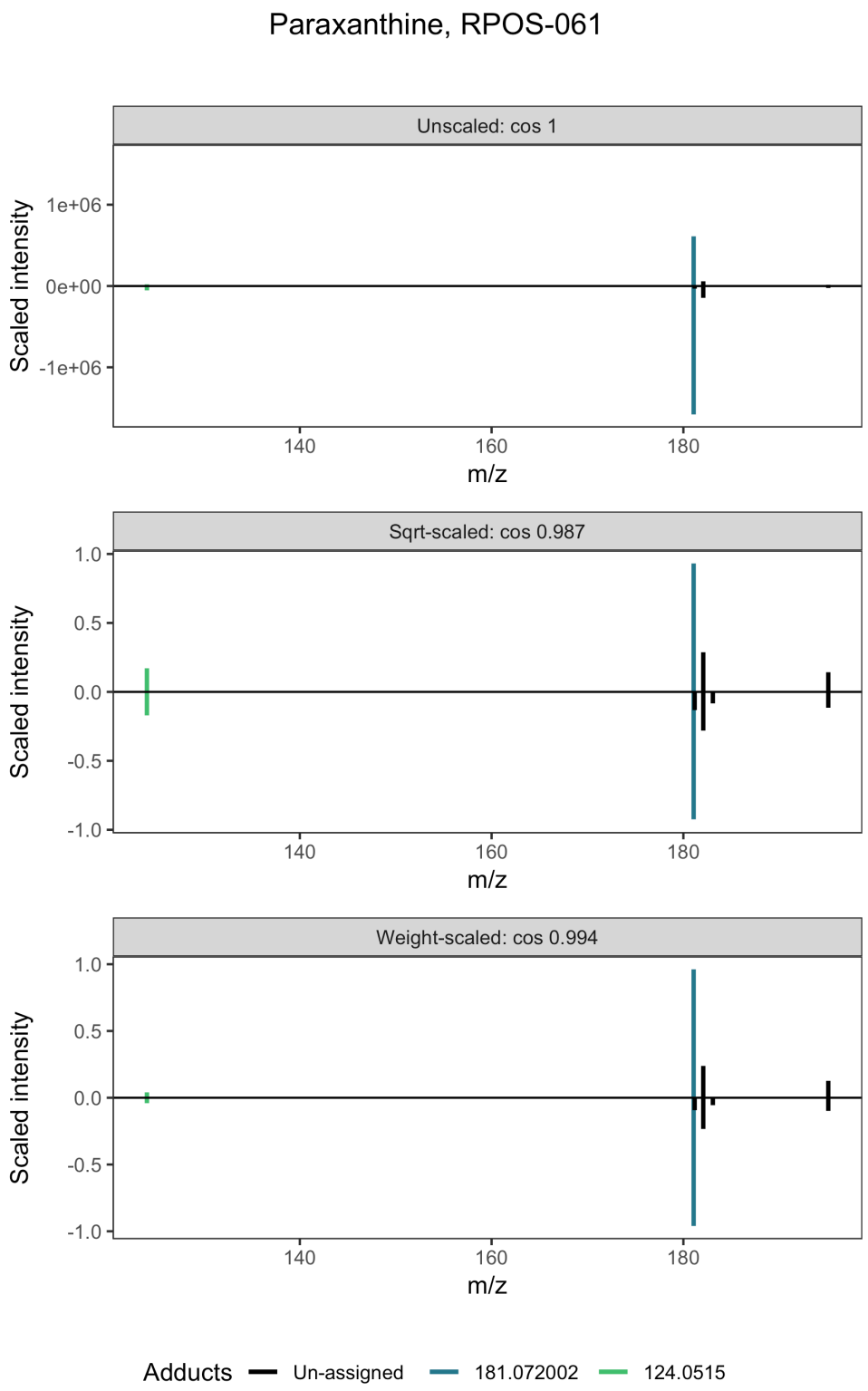


FIGURE C.17: Spectral similarity score between PCS containing two Paraxanthine ions varies depending on spectra scaling method.

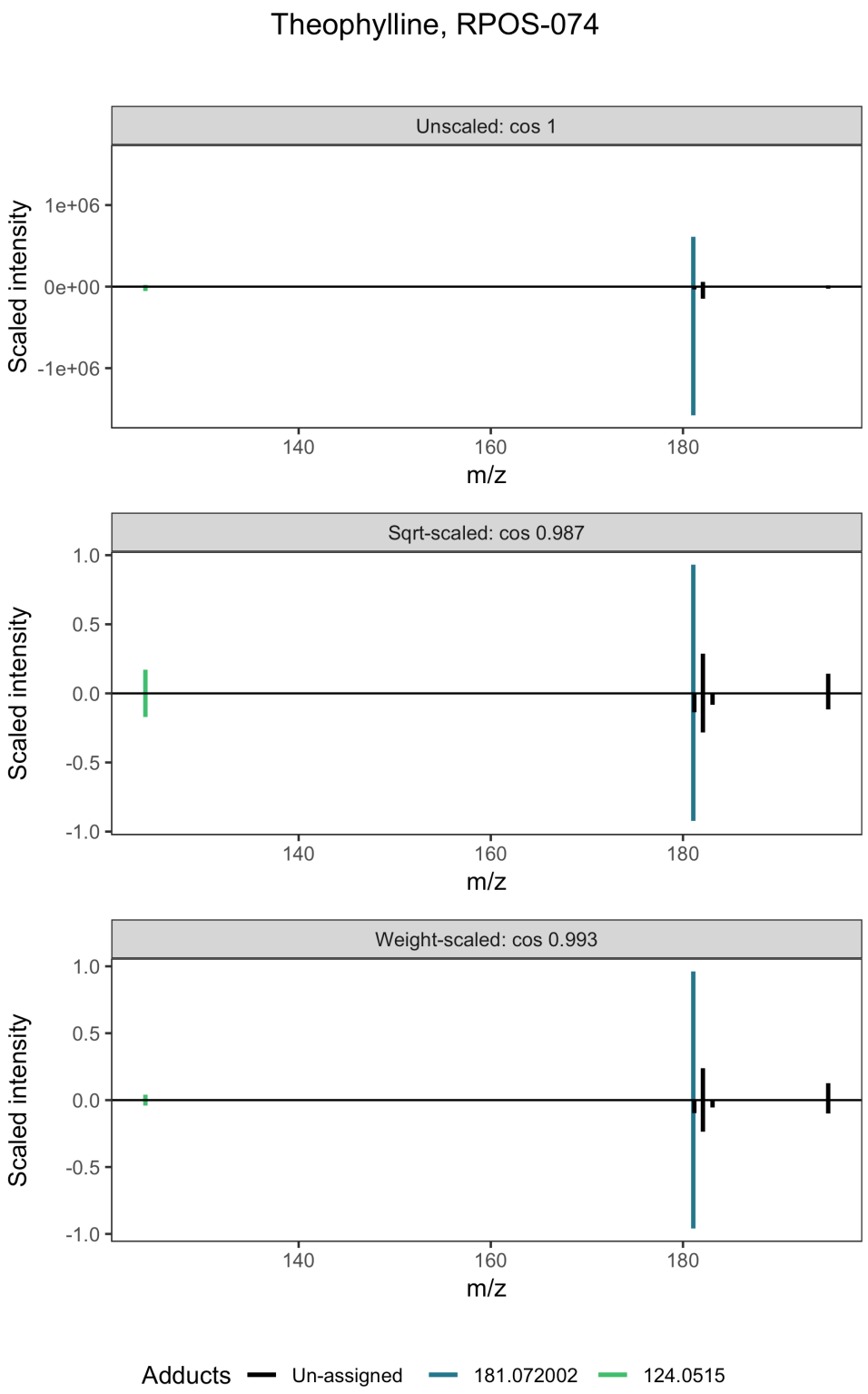


FIGURE C.18: Spectral similarity score between PCS containing twoTheophylline ions varies depending on spectra scaling method.

Appendix D

User guides to massFlowR

For more details, please refer to the GitHub repository `https://github.com/lauzikaite/massFlowR`.

LC-MS data processing with massFlowR

10 January 2020

Package: massFlowR (https://github.com/lauzikaite/massFlowR)
Authors: Elzbieta Lauzikaite
Date: Fri Jan 10 14:28:56 2020

1 Introduction

This documents provides an overview of the LC-MS data pre-processing with `massFlowR` using dataset faahKO.

2 Data import

LC-MS data in mzML/NetCDF format import is supported. Data import is implemented via mzR (http://bioconductor.org/packages/mzR) package.

In this document, functionality of the package will be demonstrated using data from faahKO (http://bioconductor.org/packages/faahKO) package. Raw LC-MS data files (in NetCDF format) are provided for spinal cords samples taken from six knock-out (KO) and six wild-type (WT) mice. Each datafile contains centroided data acquired in positive ionization mode, with data recorded at 200-600 m/z and 2500-4500 seconds.

Load the package and locate the raw CDF files within the `faahKO` package:

```
library(massFlowR)
## Get the full path to the CDF files
faahKO_files <- dir(system.file("cdf/WT", package = "faahKO"), full.names = TRUE)
```

/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt15.CDF
/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt16.CDF
/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt18.CDF
/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt19.CDF
/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt21.CDF
/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt22.CDF

3 Individual samples processing

The first stage in the pipeline is chromatographic peak detection and grouping via function `groupPEAKS` .

3.1 Chromatographic peak detection

Peaks are detected using the *centWave* algorithm from xcms (http://bioconductor.org/packages/xcms) package (see (Tautenhahn, Bottcher, and Neumann 2008)). Appropriate parameters for the LC-MS experiment must be selected. For advise on this, please see the official `xcms` manual (https://bioconductor.org/packages/release/bioc/vignettes/xcms/doc/xcms.html#3_initial_data_inspection). Selected parameters must be built into a *CentWaveParam* class object:

```
## xcms parameters for peak-picking
cwt_param <- xcms::CentWaveParam(ppm = 25,
                                 snthresh = 10,
                                 noise = 1000,
                                 prefilter =  c(3, 100),
                                 peakwidth = c(30, 80),
                                 mzdiff = 0)
```

3.2 Chromatographic peak grouping

Detected peaks are put into groups, which comprise peaks originating from the same chemical compound: adducts and isotopes. For each peak in a sample, function `groupPEAKS` :

- Finds co-eluting chromatographic peaks;
- Performs extracted ion chromatogram (EIC) correlation between all co-eluting peaks;
- Builds a network of peaks with high EIC correlation;
- Detects communities of peak within the correlation network (implemented by `igraph` package algorithm, see (Raghavan, Albert, and Kumara 2007)).

Peaks that group into a community form a **pseudo chemical spectra**. Only communities with more than one peak are retained for further processing.

3.3 Implementation

Function `groupPEAKS` processes every LC-MS datafile independently and thus can be implemented in parallel, or during sample acquisition on the machine linked to the LC-MS. A list of paths to LC-MS datafiles (in mzML/NetCDF format) is feeded to `groupPEAKS`, together with the *CentWaveParam* class object, path to output directory and parameters for parallelisation.

`groupPEAKS` writes a csv with detected and grouped peaks in the selected directory for each LC-MS sample separately. The filenames of the generated csv files will be needed for the next stage in the pipeline. The filename starts with the original raw LC-MS filename and ends with "peakgrs.csv".

`massFlowR` pipeline requires a metadata table with the following columns for each sample:

- `filename` specifies the basename of the raw LC-MS file.
- `run_order` specifies the acquisition order for the corresponding LC-MS sample.
- `raw_filepath` specifies the absolute path to the raw LC-MS file (netCDF/mzML).

```
## define where processed datafiles should be written
# out_directory <- "absolute_path_to_output_directory"

## create metadata table with required columns 'filename', 'is_sr', 'run_order' and 'raw_filepath'
metadata <-
  data.frame(
  filename = gsub(".CDF", "", basename(faahKO_files)),
  is_sr = rep(TRUE, length(faahKO_files)), # here we will assume that every file is SR
  run_order = seq(length(faahKO_files)),
  raw_filepath = faahKO_files,
  stringsAsFactors = FALSE
  )
write.csv(metadata, file = file.path(out_directory, "metadata.csv"), row.names = FALSE)
```

filename	is_sr	run_order	raw_filepath
wt15	TRUE	1	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt15.CDF
wt16	TRUE	2	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt16.CDF
wt18	TRUE	3	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt18.CDF
wt19	TRUE	4	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt19.CDF
wt21	TRUE	5	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt21.CDF
wt22	TRUE	6	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt22.CDF

```
## run peak detection and grouping for the listed faahKO files with two workers
groupPEAKS(file = file.path(out_directory, "metadata.csv"), out_dir = out_directory, cwt = cwt_param, ncores = 2)
#> 2 out of 6 files were processed.
#> 4 out of 6 files were processed.
#> 6 out of 6 files were processed.
#> Peak-groups for all files were succesfully generated.
```

4 Peak alignment

To align structurally-related peaks as a group across samples in LC-MS experiment, an alignment algorithm, which preserves the structural spectral information, is implemented.

Peaks are aligned by taking samples in the order of raw sample acquisition and matching them against a template. Template is list of all previously aligned peaks, which is updated with each sample by:

- adding new peaks
- averaging the *m/z* and *rt* values of matching peaks between the sample and the template.

Therefore, template stores the **moving averages** of *m/z* and *rt* values.

For each peak in a sample, alignment algorithm:

1. Finds all template peaks within a *m/z* and *rt* window.
2. Identifies the true match by comparing the spectral similarity between the peak-group of the peak-of-interest and all matching template's peak-groups.
3. Merges the selected template's peak-group with the peak-group of the peak-of-interest. It updates template's *m/z* and *rt* values for the matching peaks across the template and the sample.

Spectral similarity is measured by obtaining the cosine of the angle between two 2D vectors, representing each PCSs *m/z* and *intensity* values.

4.1 Implementation

To enable peak alignment, previous metadata table has to contain additional column:

- `proc_filepath` specifies the absolute path to the csv files generated by the `groupPEAKS` function.

```
##  update previous metadata table and add paths to generated csv files
processed_files <- list.files(out_directory, "peakgrs.csv", full.names = TRUE)
metadata$proc_filepath <- processed_files
write.csv(metadata, file.path(out_directory, "metadata_grouped.csv"), row.names = FALSE)
```

filename	is_sr	run_order	raw_filepath	proc_filepath
wt15	TRUE	1	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt15.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt15_peakg
wt16	TRUE	2	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt16.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt16_peakg
wt18	TRUE	3	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt18.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt18_peakg

filename	is_sr	run_order	raw_filepath	proc_filepath
wt19	TRUE	4	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt19.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt19_peakg
wt21	TRUE	5	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt21.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt21_peakg
wt22	TRUE	6	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt22.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt22_peakg

To initiate peak alignment, use function `buildTMP`, which constructs a `massFlowTemplate` class object. `massFlowTemplate` object stores sample alignment and annotation data and is updated with every sample. Define the desired error window for *m/z* and *rt* (seconds) values, which will be used for the whole experiment. `mz_err = 0.01` and `rt_err = 2` are recommended for high-resolution UPLC-MS data. `mz_err = 0.01` and `rt_err = 10` are suitable for the `faahKO` package data.

```
## initiate template
template <- buildTMP(file = file.path(out_directory, "metadata_grouped.csv"), out_dir = out_directory, mz_err = 0.
#> Building template using sample: wt15  ...
```

To review the samples that are in the experiment, use slot `@samples`.

```
## review samples in the experiment using slot "samples"
template@samples
```

filename	is_sr	run_order	raw_filepath	proc_filepath
wt15	TRUE	1	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt15.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt15_peakg
wt16	TRUE	2	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt16.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt16_peakg
wt18	TRUE	3	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt18.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt18_peakg
wt19	TRUE	4	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt19.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt19_peakg
wt21	TRUE	5	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt21.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt21_peakg
wt22	TRUE	6	/Users/el1514/anaconda3/envs/r/lib/R/library/faahKO/cdf/WT/wt22.CDF	/Users/el1514/Documents/Scripts/massFlowR/vignettes/wt22_peakg

`massFlowTemplate` object stores the most up-to-date template in the `@tmp` slot. Function `buildTMP` creates a template using the first sample in the experiment.

```
## review gnerated template using slot "tmp"
head(template@tmp, 20)
```

peakid	mz	rt	into	peakgr
1	508.2	3515.468	53910493	1
2	496.2	3384.012	38200390	2
3	343.0	2678.218	26239739	3
4	524.2	3662.573	26223613	4
5	526.1	3168.048	25760336	5
6	522.2	3344.888	23482538	6
7	502.1	3157.093	20113369	7
8	522.2	3409.051	19327938	8
9	365.0	2679.783	15389162	3
10	509.2	3517.033	15123847	1
11	496.2	3316.719	13639593	9
12	497.2	3384.012	10219380	2
13	360.0	2684.478	10062820	10
14	464.2	3454.435	9807117	11
15	577.4	4122.670	9073006	12
16	482.2	3585.891	9047697	13
17	525.2	3662.573	7829455	4
18	527.1	3168.048	7615135	5
19	531.2	3344.888	7268278	6
20	523.2	3344.888	6909134	6

To align peaks across all samples in the study, apply method `alignPEAKS`. `alignPEAKS` updates the `massFlowTemplate` object:

1. Selects next sample to be aligned and checks whether it was already peak-picked and grouped (waits until the corresponding csv file is written).
2. Matches every peak in the sample against the template.
3. Selects best matches using spectral similarity comparison.
4. Updates template with sample's peaks: adds new and averages matching peaks.

Parameter `ncores` allows a quicker implementation using the parallel backend that is available on the user's machine (i.e. multicore on Unix/Mac and snow on Windows). Select the desired number of parallel workers.

```
## align peaks across all remaining samples
template <- alignPEAKS(template, out_dir = out_directory, ncores = 2)
#> 5 out of 6 samples left to align.
#> Aligning to sample: wt16 ...
#> 4 out of 6 samples left to align.
#> Aligning to sample: wt18 ...
#> 3 out of 6 samples left to align.
#> Aligning to sample: wt19 ...
#> 2 out of 6 samples left to align.
#> Aligning to sample: wt21 ...
#> 1 out of 6 samples left to align.
#> Aligning to sample: wt22 ...
#> Peaks were aligned across all samples.
```

Aligned sample's peak tables are stored within the `@data` slot, which lists tables for each sample separately.

```
## review alignment results for an individual sample, e.g. the second, using slot "data"
head(template@data[[2]])
```

peakid	mz	mzmin	mzmax	rt	rtmin	rtmax	into	intb	maxo	sn	egauss	mu	sigma	h	f	dppm	scale	s
1	508.2	508.2	508.2	3532.682	3498.253	3571.806	56426907	55783902	1336832	189	NA	NA	NA	NA	3887	0	10	
2	496.2	496.2	496.2	3407.486	3374.622	3443.480	34418258	33952491	1064960	190	NA	NA	NA	NA	3009	0	10	
3	524.2	524.2	524.2	3693.872	3659.443	3734.561	32311995	31172950	821056	40	NA	NA	NA	NA	3674	0	10	
4	526.1	526.1	526.1	3180.567	3146.138	3216.561	27029850	25682671	769536	29	NA	NA	NA	NA	1862	0	10	
5	530.2	530.2	530.2	3357.407	3326.108	3396.531	26175357	25052715	831104	50	NA	NA	NA	NA	2340	0	10	
6	522.2	522.2	522.2	3365.232	3324.543	3401.226	24818362	24519586	809088	153	NA	NA	NA	NA	3014	0	14	

5 Alignment validation

Once peaks are aligned across all samples, the obtained peak-groups are validated. Intensity values for each peak in a group are correlated across all samples. Correlation estimates are then used to build networks of peaks, that behave similarly across all samples. Peaks exhibiting a different pattern in their intensities are put into a new peak-group.

Each peak-group is a *pseudo chemical spectra* (PCS), which comprised peaks exhibiting consistent behaviour across samples.

To enable alignment validation, a metadata table and final template file that both were written by the `alignPEAKS` function in the selected directory, must be used. A `massFlowTemplate` class object is first created.

```
## get the absolute paths to the updated metadata file and the final template writen by "alignPEAKS"
m_file <- file.path(out_directory, "aligned.csv")
tmp_file <- file.path(out_directory, "template.csv")

## initiate validation by first loading aligned samples into a massFlowTemplate object
template <- loadALIGNED(file = m_file, template = tmp_file, rt_err = 10)
#> A 'massFlowTemplate' object was succesfully built with aligned samples.
```

Peak-group validation is enabled by applying the method `validPEAKS` on the `massFlowTemplate` class object. Validation can be implemented in parallel using `ncores` parameter.

`validPEAKS` will return a `massFlowTemplate` class object with validated *pseudo chemical spectra*, as well as write peak tables for the obtained PCS:

- intensity_data.csv (intensity values for every peak in PCS in every sample)
- peaks_data.csv
- sample_data.csv

```
## Start validation using a massFlowTemplate object
template <- validPEAKS(template, out_dir = out_directory, ncores = 2, cor_thr = 0.5)
#> value for 'min_samples_prop' not provided. Setting the minimum number of samples to 3!
#> All peak-groups were succesfully validated.
```

6 Peak filling

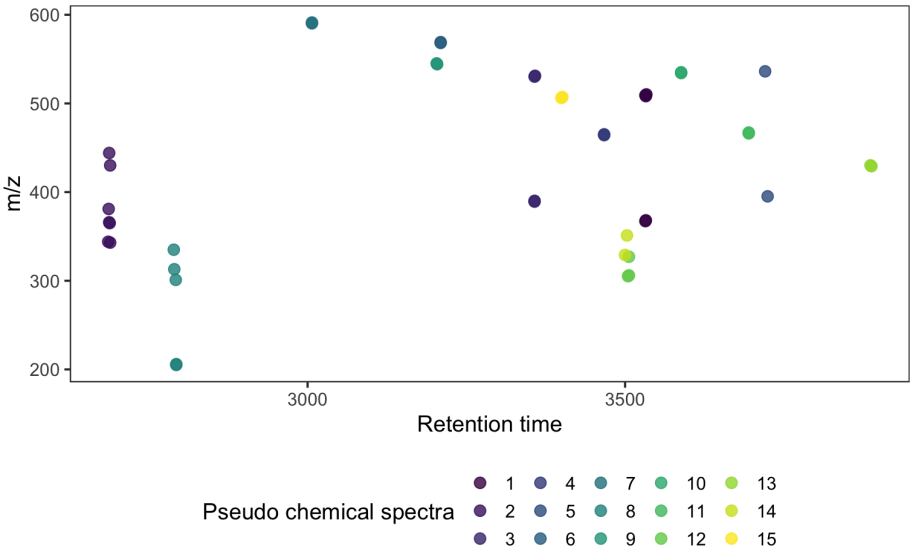
Final step in the pipeline is to re-integrate intensity values for peaks that were not detected by the *centWave* using raw LC-MS files. In contrast to `xcms` package, *m/z* and *rt* values for intensity integration are estimated for each sample separetely. *m/z* and *rt* values are modelled using a cubic smoothing spline.

```
## Fill peaks using validated massFlowTemplate object
template <- fillPEAKS(template, out_dir = out_directory, ncores = 2)
#> 6 samples left to fill.
#> Filling next 2 samples:
#> wt15
#> wt16
#>  ...
#> 4 samples left to fill.
#> Filling next 2 samples:
#> wt18
#> wt19
#>  ...
#> 2 samples left to fill.
#> Filling next 2 samples:
#> wt21
#> wt22
#>  ...
#> All peak-groups were succesfully filled
```

7 Final output

`fillPEAKS` writes file 'filled_intensity_data.csv', which contains features metadata (including pseudo chemical spectra number) and intensity values for each sample in the experiment.

```
final_dt <- read.csv(file.path(out_directory, "filled_intensity_data.csv"), header = TRUE, stringsAsFactors = FALS
```



8 Peak annotation

If in-house chemical reference database is available, PCS are annotated. For more details how to build a database file, see annotation using database (https://htmlpreview.github.io/?https://github.com/lauzikaite/massFlowR/blob/master/doc/annotation.html)).

9 See also

- massFlowR overview (https://htmlpreview.github.io/?https://github.com/lauzikaite/massFlowR/blob/master/doc/massFlowR.html)
- Automatic annotation (https://htmlpreview.github.io/?https://github.com/lauzikaite/massFlowR/blob/master/doc/annotation.html)

References

Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. 2007. "Near linear time algorithm to detect community structures in large-scale networks." *Physical Review E* 76 (3): 036106. doi:10.1103/PhysRevE.76.036106 (https://doi.org/10.1103/PhysRevE.76.036106).

Tautenhahn, Ralf, C Bottcher, and S Neumann. 2008. "Highly sensitive feature detection for high resolution LC/MS." *BMC Bioinformatics* 9: 16. doi:10.1186/1471-2105-9-504 (https://doi.org/10.1186/1471-2105-9-504).

Annotation using chemical reference database

10 January 2020

Package: massFlowR (https://github.com/lauzikaite/massFlowR)
Authors: Elzbieta Lauzikaite
Date: Fri Jan 10 14:28:53 2020

`massFlowR` performs automatic annotation of final feature table if LC-MS files for chemical reference compounds are available. Database table can be obtained from raw LC-MS files in two steps:

- Build pseudo chemical spectra (PCS) for each compound using the raw LC-MS file (in mzML/NetCDF format);
- Build database table.

1 Building database

Raw LC-MS files acquired for each chemical standards were processed by the National Phenome Centre. Each chemical standard was written as an rda file. Code for corresponding functionality will be added to `massFlowR` package for those that have acquired LC-MS data independently.

1.1 Building database table

Function `buildDB` can be used to build a database table from rda files. The generated table will have the following columns:

- `peakid` (unique peak number)
- `mz` (peak m/z)
- `rt` (peak retention time, sec)
- `into` (peak intensity)
- `peakgr` (unique peak-group number)
- `chemid` (unique database chemical number)
- `dbid` (compound identifier)
- `dbname` (compound chemical name)

```
# rda_dir <- "path to rda files"
# out_directory <- "path to output directory"
buildDB(rda_dir = rda_dir, out_dir = out_directory)
db_table <- read.csv(file.path(out_directory, "database.csv"))
```

peakid	mz	rt	into	chemid	dbid	dbname
1	664.1180	74.913	303022.40	1	IROA_P01W01	NAD
2	332.5625	74.913	196370.35	1	IROA_P01W01	NAD
3	542.0694	74.913	84401.63	1	IROA_P01W01	NAD

peakid	mz	rt	into	chemid	dbid	dbname
4	665.1209	74.913	75311.62	1	IROA_P01W01	NAD
5	333.0638	74.913	51325.25	1	IROA_P01W01	NAD
6	686.0997	74.913	49965.38	1	IROA_P01W01	NAD
7	123.0552	74.913	47608.35	1	IROA_P01W01	NAD
8	524.0586	74.913	43952.75	1	IROA_P01W01	NAD
9	564.0509	74.913	33665.41	1	IROA_P01W01	NAD
10	580.0154	74.913	28148.40	1	IROA_P01W01	NAD
11	428.0368	74.913	27833.21	1	IROA_P01W01	NAD
12	666.1227	74.913	16549.82	1	IROA_P01W01	NAD
13	626.0221	74.913	16058.00	1	IROA_P01W01	NAD
14	543.0715	74.913	14569.47	1	IROA_P01W01	NAD
15	135.0664	31.095	202712.16	2	IROA_P01W02	L-GLUTAMINE
16	130.0495	31.095	101983.22	2	IROA_P01W02	L-GLUTAMINE
17	152.0924	31.095	61172.92	2	IROA_P01W02	L-GLUTAMINE
18	147.0759	31.095	40568.49	2	IROA_P01W02	L-GLUTAMINE
19	174.0744	31.095	22007.72	2	IROA_P01W02	L-GLUTAMINE
20	169.0575	31.095	17936.97	2	IROA_P01W02	L-GLUTAMINE
21	349.0541	60.951	168396.59	3	IROA_P01W04	INOSINE 5’-PHOSPHATE
22	371.0361	60.951	118149.89	3	IROA_P01W04	INOSINE 5’-PHOSPHATE
23	137.0455	60.951	113153.09	3	IROA_P01W04	INOSINE 5’-PHOSPHATE
24	697.1015	60.951	77696.85	3	IROA_P01W04	INOSINE 5’-PHOSPHATE
25	735.0478	60.951	21372.32	3	IROA_P01W04	INOSINE 5’-PHOSPHATE
26	350.0562	60.951	20850.84	3	IROA_P01W04	INOSINE 5’-PHOSPHATE
27	393.0174	60.951	15408.79	3	IROA_P01W04	INOSINE 5’-PHOSPHATE
28	387.0024	60.951	15040.19	3	IROA_P01W04	INOSINE 5’-PHOSPHATE

peakid	mz	rt	into	chemid	dbid	dbname
29	433.0062	60.951	13116.59	3	IROA_P01W04	INOSINE 5'-PHOSPHATE

2 Automatic annotation

To annotate a features table, you will need:

- path to the metadata file with columns 'filename' and 'run_order'.
- path to the intensity table generated by `fillPEAKS` function. Sample names in this table must correspond to 'filename' column in metadata.

First, `massFlowAnno` class object is created by `buildANNO` function.

```
# meta_file <- "path to metadata csv file"
# ds_file <-  "path to filled intensity data csv file"
# out_directory <- "path to output directory"
anno <- buildANNO(ds_file = ds_file, meta_file = meta_file, out_dir = out_direc
#> A 'massFlowAnno' object was succesfully built with 6 samples.
```

`massFlowAnno` class object can be annotated with different databases tables using function `annotateDS`.

```
anno <- annotateDS(object = anno, db_file = db_file, out_dir = out_directory, r
#> Annotating dataset...
#> Dataset was annotated succesfully.
```

3 See also

- Introduction to massFlowR (https://htmlpreview.github.io/?https://github.com/lauzikaite/massFlowR/blob/master/doc/massFlowR.html)
- Data processing with massFlowR (https://htmlpreview.github.io/?https://github.com/lauzikaite/massFlowR/blob/master/doc/processing.html)