Imperial College London
Department of Biomolecular Medicine

Exploring Correlation Structures of Metabolomics data for Quality Control and Biomarker Discovery

H. Paul Benton

December 11, 2013

Supervised by
Timothy MD Ebbels
Jeremy Nicholson

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Biochemistry of Imperial College London
and the Diploma of Imperial College London

Abstract

Metabolomics is a technology which allows us to probe a wide array of interactions between metabolites. These interactions can be revealed by statistical correlations between metabolite levels that may arise via a range of mechanisms. To measure metabolite levels, two main techniques are used: Liquid Chromatography Mass Spectrometry (LC-MS) and Nuclear Magnetic Resonance (NMR). For the measurement of correlation structure high analytical reproducibility of the assays is required. While NMR has previously been shown to be reproducible, LC-MS, has not been similarly assessed. To assess the reproducibility of LC-MS for urinary metabolomics, a multi-laboratory study was devised. We find that the technology is highly reproducible, both within and between laboratories with CVs of $< 17\%, < 5s$ drift and under 10% ppm between labs.

In LC-MS, ionisation of a single compound can lead to multiple charged species such as isotopologues, adducts etc. These multiple signals have a high mutual correlation and we show that this allows them to be identified with high sensitivity and specificity. The inferred statistical interactions between different metabolites can also be affected by analytical errors. An algorithm was designed to remove statistical metabolite links that could have been caused by the analytical technique. Using this method, a higher confidence can be placed on the remaining interactions, suggesting that they are potential biological interactions. Finally, most biological interactions are dynamic in nature, leading to correlations through time between metabolite levels. To explore these dynamic links, two temporal approaches were developed. These methods are designed to discover temporal correlations between metabolites and to test whether they vary between biological conditions. We successfully demonstrate the methods in both LC-MS and NMR datasets.

Overall, this thesis shows that correlation structure in metabolic profiling data is reliable, can be successfully filtered to improve quality and can be interrogated to reveal a new kind of dynamic metabolic biomarker.

Copyright Declaration

Declaration of Originality

I declare that the contents of this thesis are my work and efforts except where acknowledgment have been made to others.

Don’t Panic

First we would like to thank all of the people who have collected the data, in its many forms. This has allowed me to developed the methods herein. Specifically, Florian Geier who let me use his worm samples for data collection, all who were involved in COMET for development of the methods and collection of the data, Peter Dixon and Elizabeth Want who collected sample and data for the pregnancy dataset. Panagiotis Vorkas who collected the data for the arterial plaques. Finally, we would like to thank the Medical Research Council for the studentship, which made this work possible.

I would like to thank many people who have guided me and helped me along the way to the completion of the thesis and PhD. Each and every step has been filled with many people who encouraged me and taught me the skills I needed at the time. Francesco Falciani who got me interested in Bioinformatics and who always encouraged me to do more. Roman Mylonas who was a great friend in cold Geneva. Gary Siuzdak, who introduced me into the world of Metabolomics. Seeing my skills in computational work even when I didn’t. He allowed me to play with some of the best tools in the field and remains a great mentor. Anders Nömstrom gave me continuous help with Metabolomics and mass spectrometry when I didn’t know anything and would ask amazingly stupid questions. He has continued to allow me to ask stupid questions and has been a good friend. Timothy Ebbels who has been very patient supervisor. He has spent a lot of time helping me with my writing and helped me to develop my statistical skills leading me on the path to the PhD. I have greatly enjoyed coming up with new complex idea and methods during our meetings.

A big thank you to those who have read my thesis, especially Estitxu, my sister and my dad.

My friends in London, my past flatmates, all of you have been there for me a different times and have helped me on the roller coaster of science. Being able to celebrate Thanksgiving and the 4th in style. Friends who have already left London, thank you. A cup of Tea always helps. For always making London fun, whether always having friends over, making the tube journey more interesting or even waiting for the bus and of course ”BBC reporting”. Lunches at SAF with no more chocolate. Surfing dude! The introduction to Greek frappe and Greek death plant. The long Polish nights. A fantastic American road trip. Random wednesday nights. ReMaking burgers at 2:00am. From the midnight McCafe science talks, the east end exploring, the German road trip to flying planes in SAF. Thank you!

My family who have have always been there ready to read any weird crazy science and let me know if it makes any sense. My father who long before anyone was talking about Bioinformatics told me that it’s what I should do. You have always kept me up to date on the happenings in science.

Thank you to everyone.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Metabolite Profiling and its Background

Metabolites are the basic building blocks of any biochemical reaction. To understand these reactions, the kinetics, the products of the reactions and the effect would be to obtain a broad understanding of the biological system. This is to say they are a phenotype of the system. These metabolites are readily available from any biofluid or tissue. Collection from biofluids, such as blood or urine, eases ethical constraints, makes collection of samples simpler and makes medical diagnosis faster as there is less extraction work to do. For these reasons and more, metabolite profiling has become a promising new field and has been rapidly developing in the last 15 years [1–4]. Similar techniques and principals were described in 1966 by Dalgliesh [5]. In this publication Dalgliesh described how using a high through put methods such as Gas Chromatography Mass Spectrometry (GC-MS), would allow for the profiling of many metabolites to observe their changes in reaction to a stimuli. Later, Pauling *et al* [6] also published work in 1971 where they could observed around 250-280 metabolites using GC-MS analysis on breath and urine vapour samples. Using ^{31}P Nuclear Mass Resonance Spectroscopy (NMR) Hoult *et al*, published in 1974 a method of to look at metabolites in tissue [7]. He described a complex experiment where using NMR, energy metabolism and its dynamics where observed by using Phosphorus NMR containing molecules such as ATP. While the ideas of profiling the metabolites in order to understand the biological system and find unique metabolites to classify diseases were known and being practiced before they not defined in a publication until 1998-9 [2,8,9]. The paper in which the ideas of the field of metabolomics were described was part of a system wide understanding of yeast (*S.cerevisiae*). A publication by Oliver *et al* in 1998 [8,10] was the first to refer to the metabolome. While the metabolome was not strictly defined in the publication it was understood to be 'the collection of small molecules produced by cells' [3]. During this starting period of the field another term, rooted in greek origin, metabonomics was defined by Professor Jeremy Nicholson. It was defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification", meaning that "metabonomics deals with detecting, identifying, quantitating and cataloguing the history of time-related metabolic changes in an integrated biological system" [2]. The accepted difference between the two terms being mainly that the metabolome is an all encompassing term to profile all of the metabolites, where as metabonomics looks to profile metabolites in a sophisticated dynamic manner.

Many biomarkers have been discovered from metabolite profiling such as Taurine and Malic acid for myocardial injury [11], Oleamide for Hepatocellular Carcinoma [12] and even dietary biomarkers such as Dihydroferulic acid for coffee [13], proline betaine for citrus fruits [14] and many more [15,16]. While biomarkers for different disease and or changes are helpful and potentially life saving, understanding the metabolome has also been an area of intense research for metabolomics. In plant metabolomics it has helped to find and identify novel secondary metabolites [17,18] and to database plant metabolites [19,20]. Metabolite profiling has also started to address difficult and complex medical questions showing a strong link between gut microbiota and human diseases [21,22], a molecular understanding of the plasticity of stem cells [23], molecular models for the early detection of preeclampsia during pregnancy [24,25] and much more [26,27].

1.2 Analytical Platforms for Metabolite Profiling

While there are many ways to detect and analyse metabolites from a biological system there are two main detection systems. One is Nuclear Magnetic Resonance and the other is a hyphened technique of Liquid/Gas Chromatography-Mass Spectrometry. Both of these techniques allow for a high throughput analysis and can analyse tens to thousands of metabolites in a single run.

1.2.1 Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance (NMR) was developed and discovered by Professor Rabi in 1939 [28, 29]. The idea is to use the physical properties of certain atoms' nuclei that have a property called spin. Only certain nuclei have the property, some of the common ones are ^{1}H, ^{13}C, and ^{31}P. The nuclei act as little magnets, these magnetics can be aligned or mis-aligned with a given magnetic field. The nuclei that are aligned to the magnetic field are at a lower, more stable energy state than the mis-aligned nuclei. The nuclei will absorb different radio frequencies (RF) depending on which nuclei (ie ^{1}H, ^{13}C etc) and the magnetic field strength of the local magnetic field. Each nucleus on a particular atom has a slightly different environment. These different environments have different densities of electrons around them, this will affect the field local strength. This altering is due to the electrons on the neighbouring atoms that induce their own magnetic field. This can either add to or subtract from the external field. The nuclei will absorb energy at $E = h\nu$. Where E is the energy that they absorb h is Planck's constant and ν is the frequency of the radio frequency (RF). This absorbed energy will make the nuclei flip to the higher energy state. However, as the lower energy state is more stable the nuclei will flip back to the lower state releasing the same RF as they absorbed. Since a wide range of RF were pulsed at the sample and only certain RF were absorbed only those RF are resonated back. This allows detection of environments of the nuclei of each atom. However, since every magnet field strength is different the resonance frequencies will be different on each spectrometer. These frequencies can 'normalised' to a measured compound such as tetramethylsilane (TMS). By normalising the spectra to TMS there is now a base line which is made by the first signal of the TMS at roughly $100MHz$ on 9.4 Tesla NMR. The frequencies are measured in parts per million (ppm) shifts less from what the TMS baseline needed to produce a signal. This can now be plotted with the intensity of the resonated frequencies that were received by the NMR detector to make a spectrum. By knowing the the different ppm shifts a molecule can be described and structurally characterised.

The two main NMR experiments $^1H\ NMR$ and $^{13}C\ NMR$ require different pulse frequencies and need to be performed in separate experiments. The ^{13}C experiments take a much longer time than the ^1H experiments. This is due to the lower natural abundance of ^{13}C as opposed to ^1H. $^1H\ NMR$ experiments also have their own problems. Most samples need to be in a solvent however, most solvents have ^1H. If a solvent with ^1H was used then the signals from the solvent would dominate the analyte signals. Therefore, if a ^1H NMR experiment is being done all of the solvents used need to be deuterium ^2H solvents. However, if this is unavoidable specific signals such as water can be suppressed with negatives pulses. ^2H will also absorb RF and have spin however, they will absorb at a different wavelength than the ^1H.

NMR spectroscopy is very useful in metabolomics since it allows for a quick understanding of what compounds are in the mixture and can yield a lot of information quickly. NMR also has the advantage that it is relatively easy to use and the reproducibility of the technique is very high. This means that samples can be compared from different labs and still have the same results which is essential in the larger studies that are becoming more frequent.

1.2.2 Liquid Chromatography Mass Spectrometry

Liquid Chromatography

There are many different types of chromatography of which liquid chromatography is one. Liquid Chromatography (LC) is one of the most commonly used chemical separation techniques. In this technique, the chromatographic separation takes place by a competition between a tube packed with small solid particles, called the chromatographic column (stationary phase) and a liquid mobile phase. Nowadays a lot of chromatographic columns can be found in the market depending on the chemistry of the small particles they contain. These are normally porous silica particles however, newer technology have allowed polymeric particles to be used as well.

In metabolomics both gas chromatography (GC) and LC are used frequently [30, 31]. In LC a commonly used type of method is the reversed phase (RP) chromatography. In RP chromatography [32], the most widely used chromatographic columns are packed with porous silica particles that are linked to C_8 or C_{18} alkyl chains. These alkyl chains have a hydrophobic interaction with the analytes.

The analyte is initially dissolved in the mobile phase and once it reaches the chromatographic column it interacts with both phases at the same time. In RP chromatography the mobile phase is normally a mixture of water and organic solvent, that is substantially more polar than the stationary phase. As the mobile phase is flushed through the chromatographic column, analytes elutes faster or slower, depending on their partition coefficient and the non-polar analytes bind more strongly to the stationary phase particles than polar ones. By increasing the percentage of non-polar solvent (normally acetonitrile or methanol) the mobile phase turns less polar and thus elutes the retained analytes more quickly. This change is known as a gradient elution and the length and ramp of the gradient depends upon factors such as the amount of separation desired, complexity of the sample, the pressure used in the LC system and the solvents used. High Pressure LC (HPLC) [33] (50-350 bar) allows a good separation of complex samples and the typical run time is around 30-45 mins for a metabolite profiling experiment. However, if the particle size of the stationary phase is significantly decreased (from $3 - 5\mu m$ to $1.7 - 1.8\mu m$) then the system pressure is increased to 600-1200 bar and the same separation can be achieved within less run time. This means that the LC chromatographic peaks are narrower, which in-turn can increase the analytical sensitivity. These systems are known as ultra performance LC (UPLC) systems [34]. LC systems are only separation techniques, thus they need a detection system, such as UV detectors or inline mass spectrometry.

Mass Spectrometry

Mass spectrometry (MS) is a frequently used and popular technique in analytical chemistry and metabolomics. It works on the principal that chemicals can be charged/ionised and then manipulated. Once the chemicals are ions they can be introduced into the MS and its mass, ion charge ratio and relative amount can be measured. Once the molecule/chemical is an ion it can be manipulated by electromagnetic fields and guided through the instrument, filtered or separated and finally measured. The mass spectrometer records two measurements, the mass-to-charge ratio (*m/z*) and the number of ions hitting the detector (counts).

The process of ionisation is a complex and not fully understood process. There are many ways to make ions. The main decision is whether to use a solid support system, a liquid based ionisation or gas based. While there are many different ionisation systems [35–39] they can be broken up into the following main ionisation techniques:

- Matrix Assisted Laser Desorption/Ionisation (MALDI) - This is a solid support technique. A small amount of the analyte is deposited on a stainless steel plate with an equal amount of matrix. Matrix is a crystalline structure that absorbs energy from the laser. As this energy is absorbed the matrix desorbes from the plate and ionises in a plume. In this plume the matrix releases the energy given to it by the laser and gives it to the analyte.

Thus, the analyte is now ionised. The full mechanism of the MALDI process is not fully elucidated [40]. A drawback from this ionisation method is that both the analyte and matrix are seen in the final measurement. This co-ionisation can interfere with low molecular weight molecules such as metabolites.

- Electron Impact and Chemical Ionisation (EI & CI) - Electron impact and chemical ionisation are older techniques of ionisation. EI works by hitting the analyte, which is normally in the gas phase, with an electron beam. This method of ionisation is a high energy ionisation and so the molecule can fragment when this happens. Also, the method does not produce a $[M + H]^{+}$ or $[M - H]^{-}$ ion as in other ionisation methods. The simplest ion that can be formed is the addition of an electron or the removal of an electron $[M + .]^{-}$ or $[M - .]^{+}$. Chemical ionisation works in a similar fashion as EI and is also used on samples that can be analysed in the gas phase. With CI a primer gas such as ammonia or methane are ionised with an EI source. Then the charged gas is pressurised into the source cell with the analyte. The charge is then transferred between the gas ions and the analyte. CI is useful due to a softer ionisation and consequently lower fragmentation.

- Electrospray Ionisation (ESI) - This is the most common ionisation method in metabolite profiling. ESI is a low energy ionisation technique and so allows for the analysis of biomolecules with little fragmentation. ESI works by charging the injection needle with a very high voltage. The analyte liquid is then passed through this needle. Due to the high voltage of the needle the liquid is sprayed out of the needle in a cone shape. The needle and orifice of the MS are inside a chamber. This chamber is heated and is kept dry with continuously blowing nitrogen gas. This cone, referred to as a Taylor cone, then beads off droplets which undergo evaporation and Coulomb fission. Coulomb fission is an explosion of the droplets due to repulsion forces from the same charged ions. This process happens repetitively until the single charged molecule/chemical (ion) is left with no solvent. This process was popularised by John Fenn *et al* when they understood how multiple charged ions were being formed during the electrospray process [37]. These ions are attracted towards the instrument by opposite charges. This is shown in figure 1.1

All ionisation techniques produce more than one charged ion species per molecule. In ESI & MALDI the simplest and most sought after ion is the molecular ion or the molecule with or without a hydrogen atom ($[M+H]^+$ and $[M-H]^-$). However, ions can also be formed with any other atom that is the same charge as the charge given in the source. For example in positive mode another common ion is the sodium adduct $[M + Na]^+$ or more complex ions such as a water loss in negative mode $[M - H_2O - H]^-$. These other possible ions are collectively called adducts. These adducts can complicate detection, quantitation and identification of interesting results.

To separate out the different ions a separation chamber, mass analyser is needed. There are a lot of different options each with their own benefits and drawbacks. A few of the most popular ones are :

- Quadrupole - This is an older, low resolution separation technique. The ions are separated by four hyperbolic rods which have opposite charges. When these charges are applied, each of the opposite pairs of rods have a different charge, i.e. negative or positive. The ions are then attracted towards the rods and repelled from the opposite pair of rods. The rods then swap charges and so the ions have a forward cylindrical motion. Ions that are not stable in this environment due to the applied electric field will hit the rod. However, stable ions will reach a detector at the end of the quadrupole. By constantly scanning different RF's a range of *m/z*s can be filtered.
- Ion Traps - Ion traps (Paul ion traps) work on a similar concept as quadrupoles. However, the ions are trapped and are ejected from the trap to the detector when the RF are changed

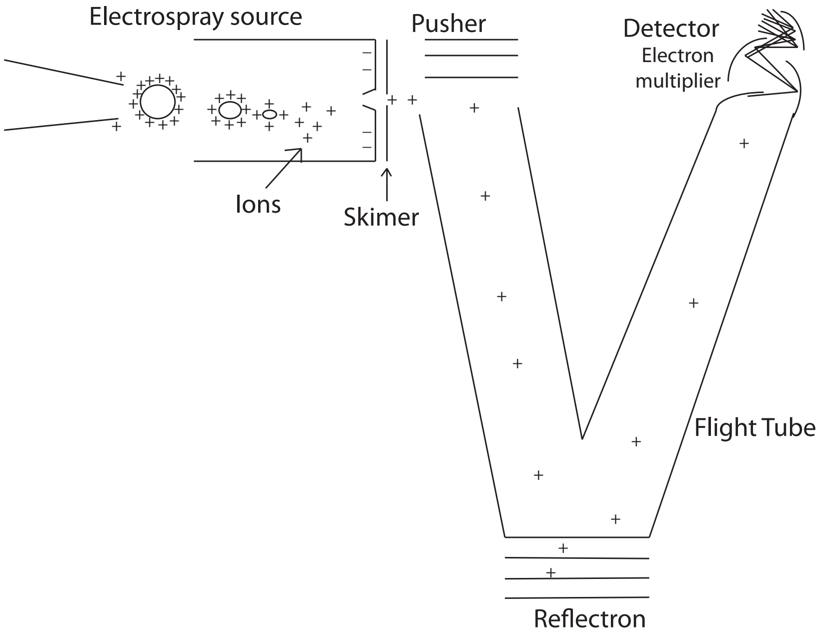


Figure 1.1: An overview of chemical detection using electrospray ionisation and the time of flight mass spectrometer detection with an electron multiplier. As droplets leave the high voltage needle they enter into a high temperature chamber and gather a charge from the needle. The droplets undergo coulomb fission until a single ion is left. The ions enter into the flight tube and are accelerated by the pusher down the flight tube and around the reflectron until they hit the detector. In the detector the signal is amplified by a electron multiplier.

so that the ions are no longer stable in the cell. Ion traps suffer from a low resolution and low dynamic range due to limited capacity to fill the cell with a wide range of ions.

- Fourier Transform Ion Cyclotron Resonance (FT-ICR) - This is by far the highest resolution mass spectrometer filter. However, it is also the most complex instrument to run and has high running costs. This instrument uses a high strength electro-magnet which allows for the trapping of the ions in a Penning trap. Once trapped, the ions are excited to their resonant cyclotron frequencies. As the ion packets oscillate they induce a charge.

All of the ions produce this charge in unison and so a Fourier transformation is required to transform the signal from the frequency domain to the time domain.

- Time Of Flight (TOF) - This is the most common mass analyser in metabolite profiling, first described in 1949 by [41]. The analyser is considered to have a high resolution and consequently can achieve a high mass accuracy. However, this is all dependant upon the size of the flight tube. TOF systems work on the simple principle that larger ions will travel slower than smaller ions and that ions that have more charge, than other ion will travel faster. The ions are timed from a pusher where the ions are perpendicularly accelerated down the flight tube to the detector as shown in figure 1.1. The ions are kept close to the centre of the flight tube by a charged guide wire until they reach the detector. The detector measures the time that it takes for ion packets to leave the pusher and arrive at the detector. Therefore, the longer then flight tube, the better resolution. [42]

When LC is combined with MS an ultimate metabolite profiling platform is made. However, this high throughput method comes at the price of complexity. The data has 3 measurements:

- Mass-to-charge ratio(*m/z*): The measured ratio of the mass and the charge of the ions.
- Retention time (RT): The time that the molecules eluted from the column.
- Intensity: The number of ions hitting the detector over a summed set of transients.

The resulting dataset has three different measurements. This complex 3D dataset makes data analysis difficult and consequently it requires specialised software.

1.3 Data Preprocessing

There are many different tools that have been designed to extract and interpret data from metabolomics experiments. Both NMR and LC-MS data have their own specialised software and scripts to allow successful data extraction and analysis. Data extraction can be difficult due to variance between samples within the dataset. This variance can cause a variable (a LC-MS or NMR peak) to shift in its location on the spectral scale. Consequently, matching the same variable between samples can be difficult. However, the end result of these data extractors is a list of variables with metadata (in NMR, ppm values and in MS *m/z* & RT values) and intensities for each sample in the dataset.

Typically, the experiments are setup to discover and detect a statistically significant metabolite that differentiates a set of class groups. One particular example is the FaahKO data where a set of mice had a gene knockout (KO) for the enzyme fatty acid amide hydrolase [43]. The experiment compared the KO to the wild-type (WT) mice brain and spinal cord tissue for each metabolite that was detected and assess if the metabolite concentration is different. For this purpose the term class will refer to a predefined similar set of samples. Once these softwares have a list of variables and their intensity values, a set of statistical test can be performed. The tests assesses which variable is important in differentiating the class or classes.

1.3.1 Data Preprocessing Software Platforms - LC-MS

The area of mass spectrometry pre-process software has seen many different solutions to the same problems. The main problem in LC-MS datasets is that the RT of the peaks shift from sample to sample. This is due to many different reasons, such as column degradation, non-efficient mixing of the solvents over the gradient and any outside factors that effect the column. This RT shifting causes an extra layer of problems in trying to match peaks between samples [44].

The software that extracts the data deals with these problems in a variety of different ways. However, before the peaks can be aligned they must first be extracted from the raw data and sorted into potential groups of peaks. While it is not always true that the software follows this sequence, it must complete the steps below at some point during its processing.

- Peak Detection - An algorithm that detects the signal from the noise. These algorithms use a filter function to assess if a possible peak fits to the specified function. These functions are often Gaussian or (mexican hat) wavelet functions. Using the fitted function the peak is integrated to find the area of the peak. The peak area has been shown to be more reliable than maximum peak intensity [45].

- Grouping - An algorithm that collects peaks within the same RT and *m/z*. Once the peaks are grouped they are referred to as ’features’. Grouping is often difficult and can lead to overlapping features being grouped together. RT alignment can help to reduce this problem along with choosing the correct *m/z* & RT windows.

- Retention time alignment - This is probably the biggest area of divergence in the field. The role of these algorithms is to correct for shifts in RT between samples for each set of features.

- Back filling - Not all software offer this option. The software goes back to the raw data if a peak was not detected in the sample for that feature. This is to say that in the classic KO vs WT experiment, if there are no peaks detected in the KO class for a feature then the software will go back to the KO samples and integrate whatever signal or baseline was present at that RT and *m/z*

As previously, stated all of these software platforms have difference between them. Below is a list highlighting some of the differences between the softwares and how they cope with these problems.

- Mass Profiler Pro - Commercial software from Agilent. This software is specifically tailored to the Agilent platform but does posses the ability to work with other instruments. It relies heavily on the ability to detect isotope patterns, fragments and adducts. With this information it uses these peaks to make a compound list. With the compound list it can then generate a 'true retention' time alignment and grouping is a manually assisted but uses the compound detection to aid in correct grouping.

- Progenesis CoMet - A commercial piece of software from Nonlinear Dynamics. Usable on all LC-MS platforms. This platform takes care of the alignment at the start of the process. Using image based software to align all of the LC-MS datasets to one global alignment. From this global alignment the peaks are then detected using added information about the isotope pattern and possible adducts. Since all of the data has been aligned to a global image the intensities for each of the peak can then be read independently from each file and compared between class groups. CoMet also makes use of the compound detection which is possible with the isotopes and adducts but groups the peaks that have not been detected to be a compound. Afterwards, a wide array of different multivariate and univariate analysis can be performed.

- mzMine - This is a well used open source software (OSS) platform that has been extended and continuously developed [46, 47]. Relying heavily on the graphical user interface, mzMine has incorporated some preprocessing sets such as SavitzkyGolay smoothing filter. For peak detection, there are many options, from simple detection of peaks above noise to a wavelet based algorithm [48] originally developed in xcms [49]. The RT alignment uses the RANdom SAmple Consensus (RANSAC) algorithm. This algorithm tries to group and align in one set. By using all of the features in the dataset, it makes a model for each set of features. The samples where peaks are outside of the model are then aligned to that RT.

- xcms - This is one of the most well used data platforms in the field. It is extendable OSS which has had many extensions published for it [48, 50–53].

  XCMS has two main peak detection methods. The first, matched Filter, uses a binned matrix. A user defined amount of bins are then combined and a second-derivative Gaussian filter is then fitted to the bins. The bin is a slice of *m/z* and RT so that each filter is being applied to an defined area of the chromatograph, or an extracted ion chromatogram (EIC). The other option is a high resolution peak detection algorithm called centWave [48]. This algorithm works by automatically detecting which area to apply the filter to. It does this by using a separate algorithm called region-of-interest (ROI). The ROI detection is a tracker algorithm based on the Kalman filter [54]. After an area of the chromatogram has been found to be a possible peak a mexican hat wavelet is used as the mother wavelet. The peak is fitted using a scaling Continuous Wavelet Transform. The final scale that the wavelet uses can be used for integrating the peak which is robust to noise or the original peak can be integrated which is more accurate but less robust.

Next, the peaks are grouped together between the samples using a grouping algorithm. Again xcms allows an option for this. The first is a fast kernel density estimation algorithm applied to slices of *m/z* to group the peaks close in retention time. Any areas that have a high density are selected as groups/features. The other algorithm is called 'nearest'. This algorithm is a nearest neighbour algorithm. By looking at the whole LC-MS landscape across all files/samples any peaks which are close to other peaks in other files are neighbours. Using a Euclidean distance the peaks are determined to be associated with the group or not. Figure 1.2 shows how different grouping algorithms can help to better define the groups. The figure demonstrates how erroneous peaks can be incorporated into the feature. Both figure 1.2 A&B show that more than the maximum number of samples (21 in this dataset) are in the feature. However, in C using 'nearest' as the grouping algorithm a correct feature group is made.

Once these groups have been determined, RT alignment can be performed. XCMS has a non-linear RT alignment algorithm. It works by finding 'well-behaved peak groups'. These well behaved peak groups act as anchors in a local regression model (LOESS). This model

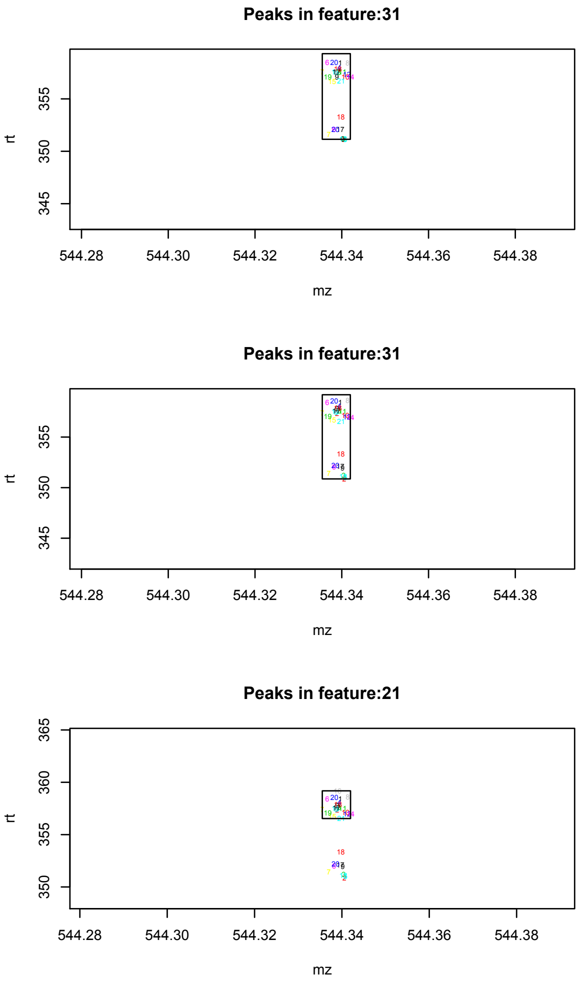


Figure 1.2: The figure demonstrates poor grouping. Each plot is a plot of RT and *m/z* for a set of group of peaks. Each peak is shown as a coloured number, where the colour and number demonstrates the file/sample which the peak came from. The box around the peaks shows the peaks that have been determined to be the same, a feature. In A, the peaks have are unaligned and grouped. The number of peaks is more than the total number of samples (21 in this data). In B, the peaks have been aligned and grouped using the kernel density algorithm. After aligning, the feature is still incorporating extra peaks. Finally, in C by using a different grouping algorithm, nearest a correct feature has been defined.

allows for a non-linear/segmented fitting and outlier detection which makes the model robust. Now that the data has been aligned the back filling or 'fillPeaks' algorithm can be run to integrate any background or peak that were not identified as previously stated. Finally, EIC, box plots and a Welches t-test are done to help the user find the most important feature in the dataset.

1.3.2 Data Preprocessing Software Platforms - NMR

Data extraction software for NMR can be much simpler [1]. Since the data is normally only 2 dimensions (ppm and intensity) either binning or signal processing techniques can be used. Also helping ease the analysis of NMR data is that NMR has a very high analytical reproducibility [55–58]. This high reproducibility reduces variable shifts between samples, allows data to be taken from different labs and or instruments and reduces false positives (FP) due to analytical variation being interpreted as biological variation.

The simplest way to extract the NMR data is to bin (sometimes called bucketing) each spectrum into thin ppm slices. For each bin, the intensity from the ppm values that are within the bin must be average in some fashion to produce a single value for the single bin.While some peaks shift a bit due to pH differences this can actually be information that is kept and can be informative about the sample set. Once all the spectra are binned a multivariate analysis can be performed on the data.

One of the advantages of NMR is that identification of metabolites is much simpler than in LC-MS datasets. With this in mind metabolites can be fitted based on their chemical shifts, multiplicities and J-couplings to a reference list of compounds. Once all of the peaks that can be determined are fitted a list of compounds is generated. This list can be used to integrated the

[1] A limited overview of data preprocessing will be discussed however, this report does not include any NMR preprocessing and so is simply given as background.

peaks to find their intensities. One of the popular software platforms to do this is Chenomix NMR suite [59]. However, this has to be performed manually and is a user intensive task. Recently a publication by Hao *et al* used a Bayesian model coupled with a Markov Chain Monte Carlo (MCMC) algorithm to automate identification and quantification of 1D NMR spectra [60, 61]. The software uses a lookup list of resonance patterns to identify the compounds. The final output of the software is a fully identified list of compounds with intensities for each sample. However, due to the high complexity and many possibilities the computational time is high.

Both of these methods have problems. Binning while fast, lowers the resolution of the spectra. This means that potentially two or more peaks can occupy the same bin. With binning the metabolites have not been identified before analysis and so the multiple peaks for each molecule can lead to redundant data. However, this can also be viewed as a confirmation that the analysis is correct. The fitting methods normally take a long time and the wrong metabolite can be fitted to a peak. However, once the fitting is performed and if done correctly then fitting the data is reduced and metabolites are already identified making direct analysis into pathway analysis possible. Fitting also removes the possibility of finding novel metabolites in the NMR dataset.

1.4 Quality Control

Variation can come from many different sources. These sources of variation can be summed into two categories, biological and analytical. Biological variation is seen in any collection biological samples and can be caused from small or big changes in the local environments of the animal/patient. Analytical variation however, is caused by changes from sample to sample in the local environment of the instrument. This is due to many reasons such as poorer ionisation efficiency caused by the sample coating the electrospray needle and skimmer [62], fluctuations in column chemistry (can be heavily reduced by QC sample equilibrating of the column) [63], changes in temperature and fluctuations in nitrogen flow between samples during the run will change the ionisation efficiency. Variation in *m/z* measurements can derive from can also come from temperature changes (in TOF systems), electronic noise and the number of ions hitting the detector.

In any experiment analytical variation needs to be minimised. While every effort is taken to do so it is not always possible to completely remove it. Quality Control samples, termed QC samples [63–65] allow a way to measure this variation These samples are normally injected every $10-20$ samples and at least 5 QC samples should be injected to equilibrate the column [63]. There are many types of QC samples that are used in LC-MS experiments. Some of these are:

- Pooled Samples: The QC sample is a linear combination of the biological samples. The assumption therefore is that the Biological variation has been removed. By using a linear combination no dilution effect should be seen for any one sample [62, 64].
- Standard Samples: These are where certain standards that are consistent with the profiled matrix are used. The standards are used in many ways to help with the normalisation of the signals and with identification of similar compounds [66, 67].

While typically used with the pooled QC samples the coefficient of variation (CV) value of the intensity measurement can be calculated for each peak or feature using the QC samples. When used with the pooled QC sample the CV value can be used as a filter, to remove any feature that might be too variable in the analytical measurement. If the instrument cannot measure the feature reliably in the absence of biological variation then the feature is deemed unreliable. Normally the CV filter is set to remove $> 20\%$ CV. The median CV in some datasets can 30% or more. A 30% CV value is still accepted in LC-MS datasets for untargeted profiling [65, 68]. The median CV value can be used as a measure of how reproducible the dataset is in terms of analytical variation. Other methods have been proposed where an internal standard is injected into all of the samples. However, this can cause ion suppression and consequently remove some of the analytes which are being seen. Also when untargeted profiling is being used finding the correct internal standard is difficult especially when the dealing with unknowns [69, 70].

1.5 Statistical Data Analysis

When all of the data has been extracted and only the robust variables are left the data can be analysed. There are many different ways to do this and depending upon the needs of the experiment this can range from a highly complex model to a simple statistical test.

1.5.1 Univariate Statistics

Univariate statistics is defined as assessing one variable/feature at a time independent of the any other variable/feature. In metabolite profiling this can be particularity useful if there is a known drug compound which has been metabolised and single metabolites need to be assessed to see if there were effected by this drug. Univariate statistics can also be very useful as they do not test the effects of other features only the selected feature. Consequently, the test is not altered by these other features, in a statistical fashion. Many clinical biomarker tests are performed with univariate tests. One of the tests which is often used is the Welches t-test. This test is an adaptation of the Students t-test. They, test the null hypothesis of whether or not the two distributions means are equal. If the p-value is below the chosen significance level (alpha or α) then the null hypothesis is rejected.

While this is the most widely used univariate test there are others. Univariate tests can be split into two main categories:

- Parametric or non-parametric

- paired or unpaired

These two main categories of univariate tests are the main assumptions when choosing a test. The first is simply defining if the data can be assumed to have a defined distribution,parametric or if there is no defined distribution, non-parametric. The next group is the assumption of how the data was sampled. If the data was sampled all from the same individual a paired test can be used whereas if independent sampling were used then an unpaired test should be chosen.

Biological datasets have many hundreds to thousands of variables that need to be tested. This means that the same test with the same assumptions is being run many times. With any test there is the possibility of getting false positive (FP) results. A FP result is a result below the cutoff limit, below α, in the test but truly the null hypothesis should have been accepted. When the cutoff level of 0.05 (alpha) is chosen the probability of a FP in one test is 0.05. However, this is in a single test. When many tests are performed, the probability of a FP result is much more than 0.05. Therefore some sort of multiple testing correction needs to be done. The most well known and rigorous correction is the Bonferroni.

$$\alpha' = \frac{\alpha}{n} \tag{1.1}$$

The correction is as shown in equation 1.1, where the n is the number of tests and α is the chosen level of significance. However, this correction method is very strict and increases the number of false negatives (FN). Another multiple testing correction termed false discovery rate (FDR) can reduce the number of FN. The FDR is the expected proportion of false positives among all discoveries (rejected null hypotheses). Finding the FDR is simple. One of the most popular FDR approaches is the Benjamini and Hochberg [42, 71] approach. In this approach a step down procedure is used. If the p-values are ranked so that $p_1 \leq p_2 \; ... \leq p_i$ for a total of m tests and an

FDR rate, δ is chosen then the equation 1.2 can be used to find the significant p-values.

$$p(i) \leq \delta \times \frac{i}{m} \tag{1.2}$$

This area is a well research area of statics and there have been newer approaches to the technique. A procedure by Storey [72] corrects for the false non-discovery rate or positive FDR. The method uses the distribution of the p-values and how the distribution of real data is different to that of null tests.

One of the other univariate methods that are often used in metabolomics are the Wilcoxon rank test. This test uses a ranking of the variables to come up with the test statistic. This test is used when the data is non-normally distributed and is non-parametric. Equation 1.3 shows how the test statistic is calculated.

$$W = \sum_{i=0}^{N} sgn(y_i - x_i) \cdot R_i \tag{1.3}$$

$$z = \frac{W - \mu_w}{\sigma_w} \tag{1.4}$$

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\sigma_1 + \sigma_2}} \tag{1.5}$$

Here, x and y are the same variables from different classes of sample i and R is the ranks of the differences of x and y. The statistic W is made by taking the sum of the signed ranks, excluding any pair of variables where $x - y = 0$. Once W has been calculated the test statistic can be transformed to a z statistic. To do so equation 1.4 is used, where σ_w is the standard deviation of W and μ_w is the mean of W. Once the z-score has been found a p-value can be easily calculated from a normal distribution.

Another common test univariate test used in metabolite profiling is the Welches t-test as seen in equation 1.5. Here, x_1 and x_2 are the same variables from two different classes and σ is the standard deviation from the respective classes. This equation produces a t-statistic like the Students t-test. This statistic can then be looked up on tables to find the p-value given the degrees of freedom. The Welches t-test is used when the two classes have unequal variances, seen by the two class σ values rather than the pooled σ of the Students t-test.

Sometimes there are multiple classes that need to be compared at the same time. It is possible to use a t-test or Wilcoxon rank test on each pair of classes. However, this is time consuming and increases the number of tests therefore increasing the number of false positives, even though this can be control. Another way is to use a univariate test that compares multiple classes at once. The parametric method of doing this is Analysis of Variance (ANOVA). This test statistic is very popular and has been used classically for design of experiments. It has also been extended to a multivariate form. With multiple class comparisons the test statistic does not describe which of the mean levels of the classes is different only that the set of classes are different. To find out which of the classes mean levels is statistically different a post-hoc test would need to be performed. These post-hoc tests are normally the two-way comparison tests. With Kruskall Wallis, another multi-class comparison test that is the non-parameteric cousin of ANOVA, post-hoc test are done with the Wilcoxon rank test. Kruskall-Wallis is an extension of the Wilcoxon rank test applied to a multiple class scenario. If the test showed that a particular variable was significant across the multiple classes then a Wilcoxon rank test could be done on all pairs to see which class was different.

1.5.2 Multivariate Statistics

Multivariate statistics can be very powerful in helping to define a class group. This is because they use all of the variables, and show how the variables are related to each other. One of the most well known and common multivariate techniques in metabolomics is principal component analysis (PCA). Some other multivariate analysis are multiple linear regression, Partial Least Squares regression and Multivariate analysis of variance[2].

[2] Only PCA will be described in detail as this is the only multivariate method which is used in this report.

PCA finds linear combinations of variables which explain the highest variation. The model is made up of multiple principal components. Each component is a description of the variance. Each component gets an R^2 value describing how much of the variance has been explained. The first principal component describes the most variance. The second component is orthogonal to the first. The maximum number of principal components is smaller than or equal to the number of samples -1 or until all 100% of the variance has been explained. Each component is orthogonal to the other.

To make sure that the model is predictive and is not simply being driven by a set of outliers, a predictive error value called the Q^2 value can be made. This value can be made by many different ways however, the most common is by leaving out a set of data or leaving one sample out (LOO). This is done many times until either a certain number of iterations has been satisfied or there are no more unique sets of data. For each iteration the left out data is predicted and the compared to the real data. A value called the predicted residual sum of squares (PRESS) can be made for the prediction. The value is an error value telling how close the predicted data is to the original. PRESS can be converted to a Q^2 value so that it is on the same scale as the R^2 value.

Scaling methods are often used to transform the data matrix into other spaces. This can help to explain the data and improve the accuracy of the model. PCA is sensitive to scaling methods. Methods such as univariate scaling give an equal weight to low intensity peaks and high intensity peaks however, this also has the effect of giving a lot of weight to noise peak in NMR data. Another popular scaling method is Pareto scaling [73]. This method is able to give low intensity peaks more weight without greatly enhancing baseline peaks. Scaling can also reduce or remove the Heteroscedastic noise in the data. Heteroscedastic noise is where each variable has a different amount of noise associated with it in a multivariate dataset. By changing the scaling of the original dataset this noise can be reduced [74].

1.6 Characterisation and Metabolite Databases

Characterisation can be an ambiguous term in metabolite profiling. Here, we use it to further define both the spectra; to annotate peaks and describe the relationship between peaks and the processed features in question; characterising these to metabolites. Spectral peaks have many relationships which can be used to understand them and describe the relationship to a set of rules. Once processed these peaks become features and need to be explained in terms of biological metabolites. We look at a few databases which help to translate these peaks into known metabolites.

1.6.1 Spectral Characterisation

In untargeted metabolite profiling, the most time consuming aspect is to identify metabolites [50]. There are many different paths to getting more information to help to characterise the metabolites. With more information the easier the database lookup becomes. The databases have also gained more information and have increased the ways to help identify the metabolites with online fragmentation databases and ppm shifts. Figure 1.3 shows a general workflow of finding a metabolite with the LC-MS platform and the data preprocessing workflow. It should be noted that this workflow is not always the answer, such methods as ultra high resolution mass spectrometry and retention time mapping can also characterise a feature to a metabolite.

Even though metabolite characterisation and identification is easier with NMR datasets (multiple dimensional experiments with ^{13}C, ^{1}H ^{31}P, ^{15}N etc allow good coverage of the practically any biological molecule) it can still be difficult with a complex sample, such as those used in metabolomics. One of the methods that helps to identify related peaks is statistical correlation [75]. The peaks from each compound will shift in a coherent way between samples, in both intensity and ppm. While this ppm shift is not always linear it can be detected with Pearson correlation. These correlation can then identify the relate peaks which in turn helps to identify the compound.

In LC-MS the same correlation technique can also be use to detect isotopologues, adducts and fragments. Much like NMR, LC-MS produces multiple peaks for the same compound. The intensity of the isotopologue and adducts will have a linear change with regards to the molecular ion. This unified change happens due to the various forms of the compound (isotopologues, adducts and fragments) being formed during ionisation. Therefore, there is a linear relationship between the intensity of the isotopologue and the adducts[3]. Using this knowledge the molecular ion, can be found by relating all of the adducts together. The doubly charged molecular ion, $[M + 2H]^{2+}$ ions will have a linear relationship with the doubly charged adducts eg $[M + 2Na]^{2+}$ and $[M + 2K]^{2+}$. However, the intensities of the doubly charged adducts will not be linear with respect to the single charge molecular ion. The other single charged adducts can be worked out by using a simple lookup table of mass differences.

Once all of the adducts are found and the molecular ion has been identified this *m/z* can be used to find the metabolites from a database lookup. Any fragments that have been found in the spectrum will provide clues to how the molecule is made up and what are the possible atomic bonds.

1.6.2 Using Databases

Metabolite databases are an essential part of metabolomics. Identification is arguably the most difficult part of the process. There are many aspects to any database, the ways in which it holds the relevant information and what the database is customised for, some of these aspects are discussed below. As the field has grown and identification has remained one of the big challenges and there has been a need for standard terminology. This has lead to the standards initiative [76–

[3] Note that due to the non-linearity of the instrument in certain high or low intensities, the relationship may be non-linear if outside of the linear range of the instrument. Evidence is shown that the relationship can be detected using a linear relationship in chapter 4.

78]. Sumner *et al* defined the minimum reporting requirements for metabolomics studies. This has helped defined the various levels of identification. They are defined as:

- Identified compound - This is where a compound has been identified with a reference standard
- Putatively annotated compounds - Where a compound has been identified with a reference library but not side by side with a standard
- Putatively characterised compound classes - Compound has shown a high similarity to a class of compounds via spectral similarity or chemical properties
- Unknown compounds No known ID to the compound, however it can be verified as a true signal coming from a chemical entity.

To be able to get to each of these different levels of identification, the databases need to contain the information to leverage the research. The information can be roughly broken up into some of the following areas.

Accurate mass

For LC-MS experiments this is essential. Simply having the formula of the molecule allows the generation of the mono-isotopic mass. The *m/z* of an unknown can be compared to this mono-isotopic mass to find out the ppm difference. If the difference is very large and outside of the specified ppm range for the instrument, then this is suggestive that the *m/z* does not match to the compound in the database.

Spectral data

1. Tandem MS - In LC-MS experiments a simple MS spectrum of the compound gives little extra detail beyond the formula. However, fragmentation spectra from specialised tandem MS experiments yield not only what atoms are in the molecule but

also how they are connected. Tandem MS experiments require specialised MS instruments that have the ability to isolate a single *m/z* ion packet and accelerate it into inert gas, normally Helium or Nitrogen. These spectra need to be collected at several different collision energies and can be difficult to compare between instrument vendors. Figure 1.3 shows how a single statistically significant *m/z* from an EIC is selected and a tandem MS experiment produces a MS^2 spectrum. This spectrum is then search against a library of MS^2 spectra and the corresponding compound can be found. Some of the databases that have tandem MS data are shown in table 1.1

2. NMR - Most databases that hold NMR spectra have $1D$ spectra. However, some such as HMDB have $2D$ spectra. As table 1.1 shows there are fewer metabolite specific databases based on NMR. The spectra which are in the databases can be difficult to match up to the experimental spectra due to the experimental spectra having biological matrix peaks that overlap or hide the spectral peaks of the database match. Also pH shifts in the experimental spectra add another layer of complexity. Much like the MS databases where multiple fragmentation energies need to be stored, NMR databases need to have the ^1H spectra and the other NMR suitable nuclei. Most databases also try to have the ^{13}C spectra of each entry.

Annotated experimental spectra

Annotation of the spectra in the database can be difficult, but yields a high volume of information on each spectrum. In MS databases this means that the specific fragment is listed for each peak. In NMR databases this means that the chemical moiety that makes that chemical shift is listed. Sometimes this can be done automatically by working with specialised software. This has been used in many MS based databases where the compound is fragmented in-silco to generate every possible fragment and is then match back to the spectra [79].

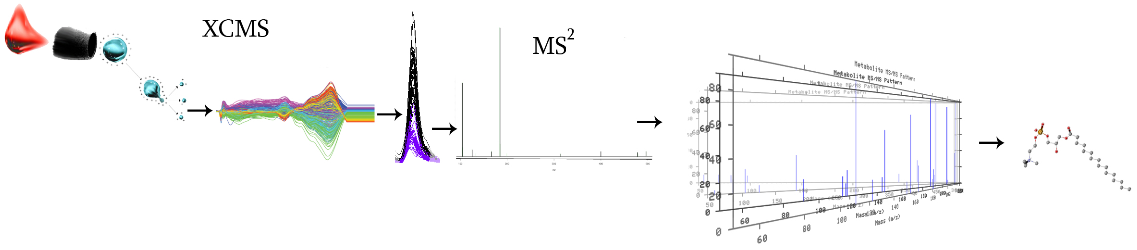


Figure 1.3: A general workflow for mass spectrometry based metabolite profiling. The figure shows the sample being injected into the mass spectrometer and then xcms being used for the data extraction. A feature is detected that differentiates the classes. Since the compound is unknown accurate tandem MS is performed to get the fragmentation pattern of the compound. This spectrum is then searched against a database and a characterised compound is identified. Figure altered from Benton *et al* [50]

Metabolite characterisation

Metabolite characterisation has become very important in the past few years of the field as the biology behind each metabolite has become an essential part of the research. A description of which pathway the metabolite is formed or is used in can help in the understanding of the underlying biology and consequently the disease.

The table 1.1 shows some of the most popular databases in metabolomics. Using the resources from these databases and the tools to help characterise the spectra from the raw data, metabolite identification has become easier. With more tools [50,80], algorithms [81] and new experimental methods identification will become easier. As metabolite ID becomes a simple single step process other more complex experiments will yield more information about the underlying biology happening in the studied system. Some of these experiments such as temporal metabolic profiling have already started to become more popular. The latest tools, such as complex correlation networks and temporal profiling help to find the metabolites that explain the biology in the studied system.

Database	Speciality	Interface	Accurate Mass	Outside Links	MS/MS Data	Spectral NMR	Entries
Metlin	Endogenous Metabolites	web	x	x	x		76,064
HMDB	Metabolites	web	x	x	x	x	40,437
KEGG	Pathways	web	x	x			16,980
ChemSpider	All Chemical	web	x	x	x		28M (all entries)
KnapSack	Plant Metabo- lites	Java app.	x	x			50,897
MassBank	Metabolites	web	x	x	x		39,407
LipidMaps	Lipids	web	x	x	x		37,127
NMRShiftDB	NMR data	web		x		x	41,921

Table 1.1: Metabolite databases and some of the relevant information [82–88].

1.7 Correlation Networks

Networks are always an interesting area in biology, especially with metabolites. Networks have the ability to infer a relationship between two variables or nodes. When metabolites are connected in the network this pair may have a relationship. The connection can be from any metric, linear correlation [89–91], mutual information [92–94] or even complex bayesian estimators [95]. The inferred relationship between metabolites can be a pathway, signalling event or any process that has affected the metabolites with similar dynamics. However, often more data is needed than simple correlations to make a biologically relevant network [89]. Yet, these simple correlations networks can give a large amount of information in understanding the apparent and underlying relationship between metabolites that is not seen when only considering mean levels. Also, correlation networks allow for a 'fingerprint' description of the state of the system which is different from 'fingerprint' analysis at the mean level.

In networks the relationship between the two variables(nodes) is called an edge. Any statistic can be used to make the edge. One of the most popular statistics is correlation. Linear correlations are computationally fast and intuitive. Unlike other metrics, especially non linear metrics, they are able to fit the data with fewer data points than many other statistics. However, correlation can be effected by outliers without the use of a robust correlation method. As previously stated correlations can be very useful in characterising the raw spectra and relating peak or spectral shifts. When the networks are made there needs to be a way to describe the network. Thus, there is a special area of statistics dedicated to describing the network.

1.7.1 Network Statistics

Networks can be difficult to describe. For this reason there are special statistics that help to describe how a network functions, what it looks like and even how the network can react when it is altered. These statistics mean that different networks can be compared against each other using these metrics to understand the differences between the networks. Some of the popular network statistics are:

- Degree - Defined as the number of edges connecting a specific node. This is one of the simplest forms of network statistics.

  $$2E = \sum_{v \in V} deg(v) \tag{1.6}$$

  In equation 1.6, E is the total number of edges in the network, $deg(v)$ is the degree of a single node[4]. The sum of all degrees in the network is double the total number of edges in the network. The degree of the network can be either a local or global statistic when averaged among all nodes. The global degree average can be very informative about a network. A network with a high global degree average means that the network is highly connected. The distribution of degree frequency can also be useful in understanding the structure of the network.

[4] nodes can also be called a vertex

- Shortest Path Length - This is the length of the shortest path between two specified nodes. There can also be a mean shortest path length for the network, which is the average of the shortest path between any two nodes in the network.

    1. Mean Path length - This is the average number of edges between any two nodes. For an undirected network it can be defined as seen in equation 1.7. Where v_i and v_j denote different nodes within a network and n is the number of nodes within the network.

$$l_g = \frac{1}{n \times (n-1)} \times \sum_{i \neq j} deg(v_i, v_j) \tag{1.7}$$

- Betweenness Centrally - is a centrality measurement that quantifies how many times a node is used in connecting two nodes along the shortest path. Explained in another way nodes that are nodes that are used in a shortest path between any two random nodes in a network will have a high betweenness. This can be shown as in equation 1.8 where $\sigma_{v_{i,j}}$ is the total number of shorest paths between all pairs of nodes in the network and $\sigma_{v_{i,j}}(v)$ is the number of shortest paths that go through node v. The sum is taken where node i cannot be the node in question.

$$C_B(v) = \sum_{v_i \neq v_j \neq t \in V} \frac{\sigma_{v_{i,j}}(v)}{\sigma_{v_{i,j}}} \tag{1.8}$$

Many of these statistics are altered when the graph is considered as a directed network. Directed networks state a direction of flow or which node affects which in the connection. Directionality is often seen in temporal networks. One of the statistics that is affected is the betweenness centrality. If there is a direction to the edges then the shortest path may need more edges to connect the nodes or it may not be possible to connect the nodes. Betweenness is more than just a measure of the connectivity of the network. It is a descriptor of how well the network is connected and can adjust or divert flow. If a highly central node is removed then the node betweenness will increase substantially. The statistics can be defined in either a directed or undirected network, here we have defined them in an undirected network.

1.8 Temporal Analysis

Temporal data is starting to be more common in metabolomics. This data has the ability to explain the dynamics of the underlying biology of the studied system. The aim of temporal data profiling is to be able to capture the events that are altering or affecting the system. Slowing the production of temporal metabolite profiling data has been a list of different issues as outlined below.

1. Difficulty in generating long (> 10 time points) time series

2. Unknown required sampling rate. Prior to recording the experimental data the required sampling frequency to observe the desired biological effect is unknown. However, this can normally be guessed on the magnitude of the scale which is needed However, this generates a large problem of either under-sampling or over-sampling. Under-sampling results in missing the effect, where as over sampling is expensive.

3. Variation between individuals. The variation in the experiment between individuals due to the speed of the response or timing of metabolite events, may not always be due to the desired effect. Consequently, this increases the amount of variation and dilutes the desired effect or may even make the effect undetectable.

However, once these issues have been over come and the data is collected analysis of the data can also be problematic. The data can be analysed for many different aims such as; temporal relationship between metabolites, prediction of future time points and the dynamics of a single metabolite.

With the extra level of complexity in the data it requires special software/algorithms. Temporal data algorithms use the temporal order to understand the dynamics. These algorithms make a model of the time series. To date few temporal metabolite profiling algorithms have been published [96]. Some of the above reasons, with the difficulty to generate the data, has hindered the development of these algorithm but also the high dependancy of the data. Each time point cannot simply be taken as an independent analysis, as each time point is dependant on the previous time point, the memory of the system. Added to this is the fact that many of the variables are correlated to each other. Some of the classical time series analysis methods are able to explain the memory of the system. However, as they are univariate methods they only consider one temporal profile at a time and consequently do not use or explain the dependant nature of temporal multivariate datasets.

1.8.1 Temporal Algorithms

Classic Algorithms

Classically, time series has been a univariable analysis trying to understand how the variable tracks through time. There has been a lot of development in financial data with tracking stocks and describing the patterns. These methods can describe if a time series is cyclic, if the time series has a repeating pattern, how long the memory of the time series is. Some of the models that can answer these questions are:

- Moving Average Models (MA) - This model is essentially a smoothing filter for the time series. It assume that the function is based around a centre mean. The model updates the average as each time point is assed.

- AutoRegressive Models (AR) - These models use the previous time points for the output variable. The model uses regression on itself to predict future time points, therefore the output is a linearly dependent upon the previous time points.

- AutoRegressive Moving Average Models (ARMA) - Incorporates both a moving average and an auto regressive portion to the model. These models require stationary data ie data that has a uniform mean.
- AutoRegressive Integrated Moving Average Model (ARIMA) - These models add an extra parameter in (*d*) to allow for time series that are non-stationary.
- Vector forms - The above models can be generalised to multivariate models, such as Vector Autoregression Moving-Average models. However, these models are less widely used. The vector forms require all the variables to be co-stationary which is rarely the case in metabolite data and require many time points to have a satisfactory model.

All these methods have the draw back that they require stationary data (the mean of the data is roughly stable) and are linear methods. Models such as the ARMA model, which is very popular, can produce useful autocorrelaiton function plots (correlograms) that give information about the memory in the data. They have been used for the basis of many specific biological methods. Due to the auto-regressive nature of these methods they require many time points to be able to generate a stable model. This means that they are not well suited for the median to low time point datasets of temporal metabolite profiling.

Temporal Specific Metabolite Profiling Algorithms

All of these algorithms use the order of the temporal data to understand the dynamics. Some of the simplest are the SME models [97]. This model will be discussed in more detail later in chapter 5. However, briefly this method uses spline fitting to fit the time series for one metabolite at a time in each set of classes. A median fit is then found for each metabolite and class. The median fits can then be compared between the classes using a modified t-test. Again this method compares only one metabolite at a time.

When a multivariate method is needed, such methods as PCA and PLS are often used. PLS models can be helpful in describing temporal data and even predicting future states. Wold *et al.* described a method where batches of chemical pharmaceutical processes were defined as either good or bad and the method could describe the next batch. In this model a 3D matrix was setup where the time points where columns and the batches were rows [98]. The multivariate models can also be used where the output of one model makes the input for the next. For instance a simple PCA can be performed on the data as normal. Then using the top components from the PCA model the scores for each component can be extracted. Using these scores they are then lagged against each other [99]. This method allows an exploration of the latent structures in the time series and the states in which the time series moves.

One of the areas of research which is gaining in publications is to build networks and then analyse the differences between the networks or the differences between the states [100–102]. However, there is a lot of difficulty in building an accurate network which describes the non-linearity and dynamics of metabolite networks. Many of these analysis try to describe the non-linearity of the metabolite networks to accurately describe and hopefully discover new pathways and signally networks. The effort is also seen directed towards flux balance analysis models [103–105].

Many of the current methods do describe a state or identify a new pathway. As discussed the common practice of metabolite profiling is to find a metabolite which is either responsible for the the change or is its concentration level is changing in response to the stimuli. New methods are needed to be able to look at differences in temporal relationships within and between metabolites. Network analysis has the advantage that it can find the relationships between the metabolites and they are flexible to be able to do this in both a temporal fashion and compare between classes.

1.9 Analysis Framework

Here, we propose a possible framework for the analysis of high throughput omics data. Figure 1.4 shows this framework and explains the different levels data analysis. We discuss this framework in terms of metabolite profiling. Level 1 is a comparison between the classes using a mean or median concentration. In figure 1.4 this is seen as two box plots that have different median levels the red class having a lower median level that the blue class. This is the conventional methodology and many studies do not go past this level.

The second level can be a comparison between classes as well but uses the covariance between metabolites to understand the data at a deeper level. In the figure, this is given as a simple no class comparison but shows a correlation network to describe the connections between the nodes. Some nodes are not connected and so have no relationship to other nodes at this level. These methods do not necessarily involve network depictions but involve relating one feature (variable) to another. Methods at this level would be PCA, correlation networks and factor analysis to name a few. Some methods such as PCA can be used in a level 1 capacity where they are simply used to find the mean level difference between two groups. However, they can also be used to understand the relationship between the features from the loading plots. As already stated correlation structures can yield a lot of information more than just pathways but also interactions beyond simple neighbouring metabolites [106]. In general term correlations do not have to be Pearson correlations or other correlations but can also be non-linear metrics such as mutual information. However, any metric that can determine a relationship between two variables is at this second level analysis. There has been a lot of interest in this area such as Opgen-Rhein *et al.* work. that looked at genetic networks using correlation structures and then reducing them using partial correlations [90]. This level is extendable to other areas outside of metabolite profiling and data for genetic studies can be very different, not only its origin but also in the relationships between the pairs. This level analysis will be used to help understand higher orders of repro-

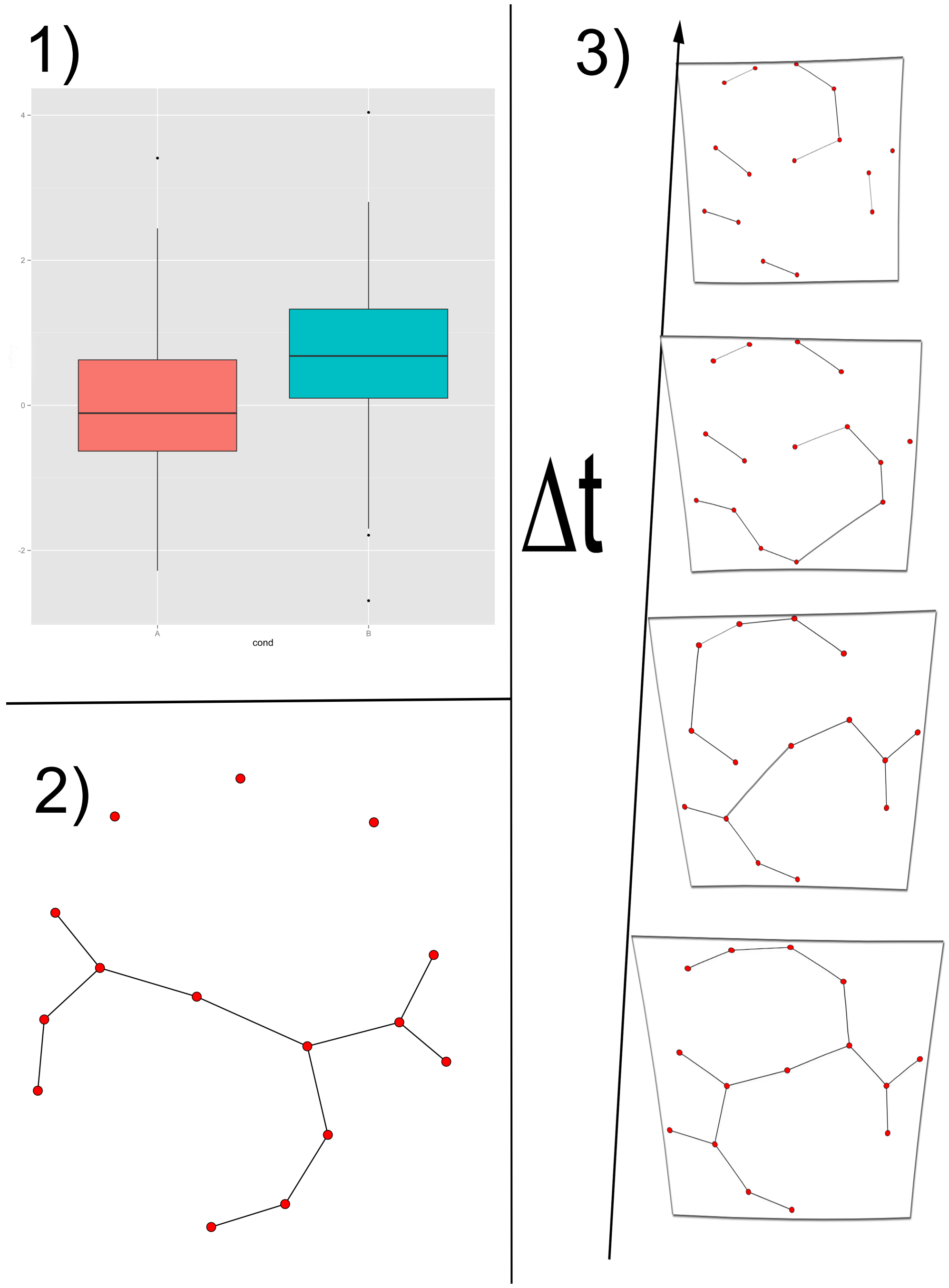


Figure 1.4: Proposed analysis Framework. This figure shows three levels of high throughput analysis. In 1 is the mean/median intensity level class comparison, between red and blue. The second level shows a relationship between the variables, depicted here by a network analysis of various nodes. Some of the nodes are connected and so share a relationship. This second level is only with static data and does not include any dynamics. Finally, the third level (3) is a demonstration of the change of the relationships between the variables. Each new network is a new lag (Δt). The relationship between the variable is lagged by a different amount of time for each network.

ducibility discussed later.

Finally, the third level is a dynamic level of temporal analysis. This is a deeper level of analysis with the hope of deeper understanding of the variable. In figure 1.4-3, this is demonstrated by the networks at different time lags. Each network has been independently generated using a metric such as correlation. The relationship between the variables at level 3 is performed by using the same metric as in level 2 but lagging the variable by a specific time lag. If the two variables are connected in the network at this level then there is a pattern that is seen to happen at this lag. The pattern is determined by the metric used. Level 3 can be performed in multiple ways when different classes are used. Depicted here is a simple single class analysis. As will be discussed later in chapter 5 the lagged networks can be made independently for each class and later compared. Level 3 does not strictly require network depictions but can be helped by the use of these structures in showing the flow between one state and the next as the time lag progresses.

1.10 Aim of Thesis

In this chapter, we have outlined some of the concepts and background that will be used in the following chapters. The aims of the thesis are to

1. Develop a computational method to improve quality of UPLC-MS data in metabolite profiling studies.
2. Assess the analytical reproducibility of UPLC-TOF-MS for urinary metabolomics both within and between laboratories.
3. Determine the ability of statistical correlation to identify structurally related signals in LC-MS metabolomics data.

4. Show that the analytical procedures introduce their own correlations and to develop a method to filter out these artefacts.
5. Develop new methods to find temporal metabolic biomarkers using differences in through-time correlaitons

We show that these methods and research, are novel and important to the field of metabolite profiling. The wider area of high throughput techniques and mass spectrometry based sciences are impacted by this research.

2 Correction of Mass Calibration Gaps in Liquid Chromatography – Mass Spectrometry Metabolomics Data

2.1 Introduction

Liquid chromatography-mass spectrometry (LC-MS) has become a widely used methodology in metabolomics. Profiles generated by this method are highly complex and many software packages have been designed to analyse this data (e.g. mzMine, XCMS [46, 49]). These typically employ a peak-picking algorithm for selection of peaks and removal of noise. Peaks are then matched between samples, correcting small drifts in retention time (RT) and the final result is a table of the integrated intensities of each peak in each sample.

An important goal in LC-MS experiments is the accurate measurement of the mass to charge ratio (*m/z*). Usually a mass calibration method is used throughout the experiment to keep the *m/z* deviation within the desired range. MS manufacturers employ different methods to calibrate their instruments. Some manufacturers use a switching method, which regularly changes the MS input between the analyte and a standard reference [107]. Switching the flow in this way can avoid the problem of ionisation suppression encountered with continuous flow methods. However, changing the flow to the lock mass spray temporarily stops the analyte from being detected, resulting in gaps in analyte data. These lock mass gaps can cause a single peak to be split into several smaller peaks (figure 2.1), confusing the process of peak matching. Even if the peak is detected correctly, the true intensity will be underestimated due to the gap. The situation is further exacerbated by RT drifts across samples, which will cause the position of the gap within the peak to shift. The peak detection, multiplicity and loss in intensity will thus vary depending on where the gap is in the peak. This variability can lead to both false positives and false negatives in peak detection, and inaccurate intensity estimation. While most vendor software is presumed to take account of such elements of MS design, this is not true of widely used and community supported open source software for LC-MS data processing. We have developed a method which recovers the lost intensity by filling in the gap. The method greatly reduces errors in peak detection, matching and intensity estimation and is freely available as a built-in function stitch [108], for the XCMS package.

2.2 Methods

2.2.1 Analytical Data and Processing

Caenorhabditis elegans were grown under standard conditions and snap frozen upon harvesting. The pellets were extracted in 80 % ice cold MeOH using a FastPrep bead beater (Precellys 24) for 40 seconds at a speed of 6.5 m/s. After centrifuging the supernatant was collected for analysis, the pellet re-extracted and the joined extracts dried prior to analysis.

10μl of Caenorhabditis elegans extract was separated on a Waters ACQUITY UPLC system using an HSS T3 column (2.1 x 100mm, 1.7μm) (Waters Corporation, Milford, MA). A 30 min water to acetronitrile linear gradient was used with a flow rate of 500 $\mu l.min^{-1}$. All solvents were HPLC grade (ACN: Fluka, H20: Romil Ltd). Detection was performed on a Waters LCT Premier ToF mass spectrometer. The sample was run twice, once with a lock mass interval (LMI) of 50 scans and once with an LMI of 1000 scans. The scan range was set to 50-1000 *m/z* and a scan time of 0.2 s and an inter-scan delay of 0.01 s. The majority of peak widths were in the range 2-20 s. Data was converted to mzXML using MassWolf (version 4.3.1). Each dataset was processed using XCMS (version 1.23.1). A third dataset was generated from the LMI50, by correcting the lock mass gaps using the new stitch method. These were then processed using the centWave [48] peak picker on all three datasets. Parameters were peakwidth of $2 - 20$ seconds and snthresh of 5. Data was then grouped using mzwid = 0.05; other parameters were set at default values.

2.2.2 Gap Filling Algorithm

The basic idea of the algorithm is to use the nearest available analyte scan to fill the gap. More precisely, if the location of the gap is from scan n to scan n+k, then the algorithm uses scan $n - 1$ to fill the first half of the gap and scan n+k+1 to fill the second half of the gap. If the gap length is odd, the middle scan is filled using scan $n - 1$. The algorithm requires three input parameters which are all readily available in the vendor software. These are: the scan number of the first gap, the lock mass interval (LMI, the time between gaps) and the length of each gap (default 2 scans). Despite its simplicity, this algorithm is very effective at correcting lock mass gaps and requires minimal user input.

2.3 Results

To test the method, data was acquired for a single sample at two different LMIs. A first acquisition at an LMI of 1000 scans (giving a total of 3 lock mass scans over the entire acquisition) gave a data set where few peaks would be affected by lock mass gaps. This approximates data unaffected by the lock mass gap problem. A second acquisition with an LMI of 50 scans (71 lock mass scans in total) was chosen to correspond to a typical metabolomics protocol. We then compared the XCMS output for the LMI 50 data with and without correction with the LMI 1000 data without correction.

2.3.1 Corrected Intensity of Single Peaks

To illustrate the success of the method we extracted two example peaks from the data (figure 2.1). The first peak (*m/z*=307.08, RT=126 s) demonstrates how the correction algorithm improves peak detection. The bottom panel of the figure, plotted in black, clearly shows two scans missing due to lock mass acquisition; these are corrected (shown in red) resulting in a peak with very similar shape and intensity to that in the LMI 1000 data (top panel). This peak is detected by the software in the LMI 1000 and LMI 50 corrected data, but not in the uncorrected LMI 50 data. In the second example the peak (*m/z*=308.08, RT=42s) is detected in all three datasets. However, in the uncorrected LMI50 the peak is much shorter, 3s compared to 5s in the LMI1000 and the corrected LMI50. The gap effectively leads to a shorter peak and this reduces the intensity estimate by 21% compared to the undamaged LMI1000 data (LMI50 intensity = 3146, LMI1000 intensity = 3987). Applying gap correction results in an intensity estimate just 1% higher than the LMI1000 data (LMI50 corrected intensity = 4030).

Other peaks were also checked to confirm that the algorithm was working correctly all of the peaks. Figure 2.2 shows a few example of extracted ion chromatograms (EICs) from the raw data for each peak group. These can be used to view the raw data and confirm that the algorithm has worked as expected. For the vast majority of the EICs it can be seen that where there is a gap, the correction method has successfully corrected for the lost intensity and reformed the peak shape.

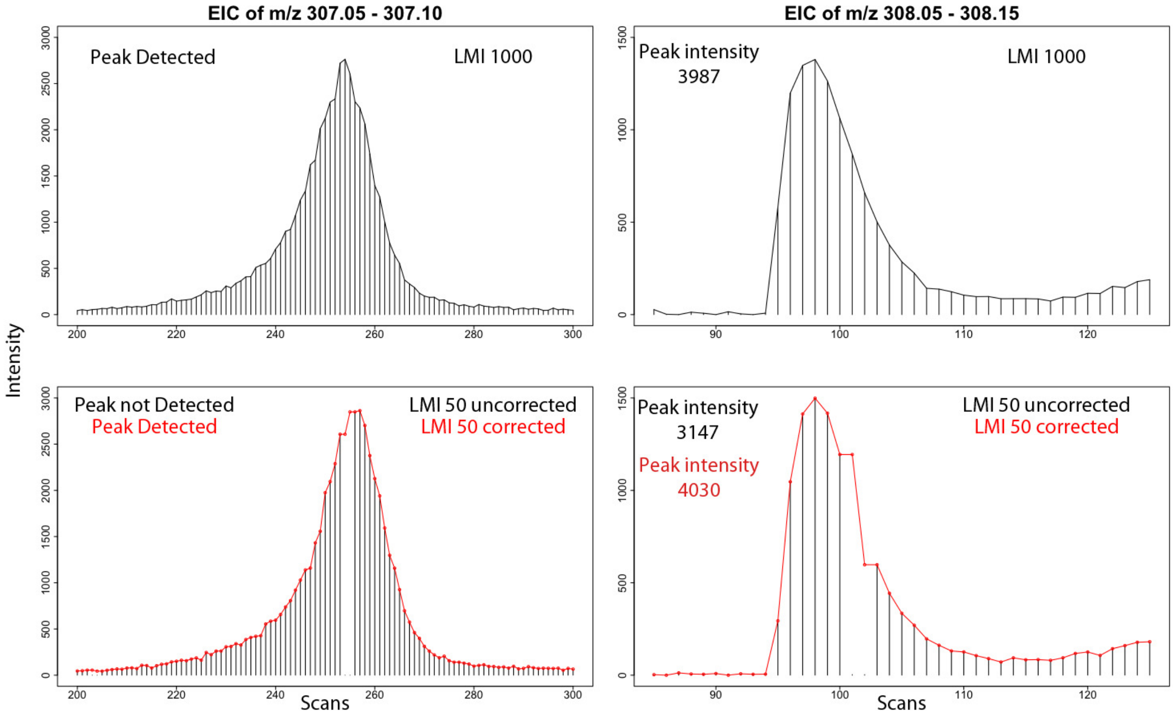
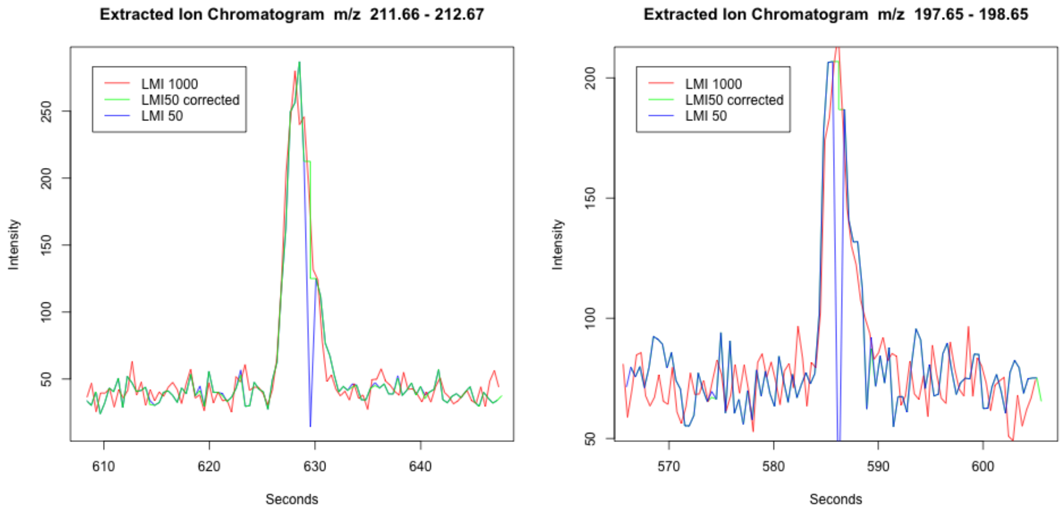


Figure 2.1: EICs of peak without lock mass gaps and with lock mass gaps. The red line demonstrates the filled in lock mass gap area. The EICs in the left hand column with gaps were originally not detected by the peak detector. The EICs on the right hand column were detected but with decreased intensity. The new integrated peak area is also shown in red.

2.3.2 Corrected Global Intensity of Detected Peaks

Many peak detection algorithms use the shape of the peak to help identify it. With the correction method, this shape is recovered, and more peaks are found. Using the same parameter settings 2784 peaks were detected in the LMI1000 data, while only 2625 peaks were detected in the LMI50 data, a decrease of 6%. In the LMI50 corrected data 2869 peaks were found, a gain of 3% over the LMI1000 data. After grouping, 2625 peak groups were matched between the three datasets. Extracted ion chromatograms were visually inspected, confirming successful gap correction. The correction method brought the global intensity distribution closer to the undamaged LMI1000 data. The median intensity dropped from 772 in the LMI1000 data to 700 in the LMI50 data (a 9% reduction) but was recovered with the correction method to 775 in the



(a) EIC

(b) EIC

Figure 2.2: Both of the peaks are affected by the gaps. Each of the three peaks groups are shown where the LMI 1000 is not affected and then the LMI 50 uncorrected and corrected. These peaks are examples of where the correction algorithm would correct for lost intensity

corrected LMI50 data (0.4% difference from the LMI1000 data).

To examine the difference between the LMI 50 corrected and uncorrected data sets as compared to the LMI 1000 dataset a subgroup of peaks that were eluting at the same time as lock mass acquisition was needed. This was necessary as the LMI 1000, although having very few lock mass gaps, was a separate run and therefore had slightly different intensities and peak integration ranges to the LMI50 data. To identify peaks, where a gap was present in the peak integration range, EICs were manually inspected, two such examples of the identified peaks can be seen in figure 2.2. The 300 peaks with the lowest *m/z* were evaluated, resulting in 78 peaks where correction had occurred. The fractional difference in intensity between the pairs of data sets for each group was calculated using equation 2.1 & 2.2. The fractional difference in intensity between the LMI50 and LMI1000 data, was computed for both uncorrected (f_c) and corrected (f_c) data for both gapped and non-gapped peaks. Gap correction resulted in a distribution for gapped peaks much closer to that for non-gapped peaks than for uncorrected data (median $f_u = -6.73$ & 3.26 and $f_c = 4.80$ & 3.13 for gapped & non-gapped peaks respectively).

$$f_c = \frac{(LMI50corrected - LMI1000)}{LMI1000} * 100 \tag{2.1}$$

$$f_u = \frac{(LMI50uncorrected - LMI1000)}{LMI1000} * 100 \tag{2.2}$$

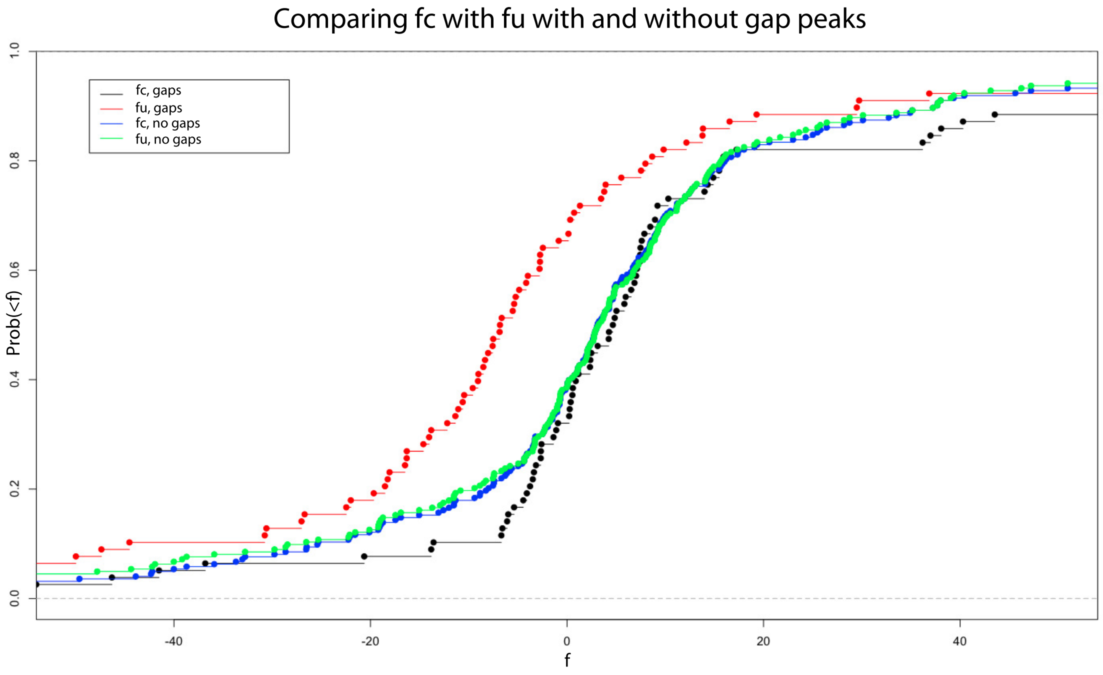


Figure 2.3: A cumulative distribution plot of the fractional corrected f_c and uncorrected f_u with and without gaps. The f_c no gaps in blue and the f_u no gaps in green are not affected by the gaps and are the expected fractional distribution. However, the red f_u gaps are the integrated peaks that are affected by the gaps but have not been corrected. The black line is the f_c gaps which has been corrected and is much closer to the expected fractional distribution.

Figure 2.3 shows the cumulative distribution of f_u and f_c for each subgroup: peaks with gaps and peaks without gaps. The corrected and uncorrected distributions (blue and green respectively) for non-gapped peaks are virtually identical as expected. However, the corrected gapped peaks distribution (black) is very different from the uncorrected gapped peaks distribution and is much closer to the non-gapped peaks distribution. This demonstrates both the major damage caused by the lock mass gaps and the considerable improvement obtained by the correction algorithm.

2.3.3 Other Software Comparison

The software mzMine [46, 47] is a widely used LC-MS data processing package. We were not able to find any mention of gap correction methods in the documentation for mzMine and therefore assume that the software is not explicitly designed to account for such defects. As a comparison, the LMI 50 uncorrected and LMI 1000 data were run in mzMine to investigate how other widely used software performs with lock mass gap damaged data. Both datasets were peak picked using the same parameter settings.

Parameter settings were as follows: Peak picking: centroid, noise level: 10, minimum time span: 5 seconds. Peak deconvolution using baseline, minimum duration 5 seconds. All other parameters were kept at default values.

Using mzMine in this way, 2545 peaks were detected in the LMI1000 data but only 1453 peaks were detected in the LMI50 data. It seems that the lock mass gaps have altered the raw data so much that 43% of the peaks are lost. Although it was not possible for us to exhaustively optimize peak detection parameters it would be expected that the same sample would have similar numbers of detected peaks in two consecutive runs. It is therefore clear that the lock mass acquisition is strongly affecting the analyte data in a way that, as expected, is not compensated for by the mzMine software.

2.4 Conclusion

The stitch function allows peak detection algorithms to accurately find and integrate each peak. We note that the method is generic and can be applied to data derived from any LC-MS experiment, for example in metabolomics or proteomics. We have implemented the algorithm within XCMS [49], one of the most widely used freely available LC-MS data processing packages.

The stitch function allows more reliable data and a great confidence that the result is TP. The stitch function has been used for all of the following chapters. This means that the peak detection is more reliable to find the peaks thereby removing FN in peak detection. In summary, our gap correction method allows LC-MS data acquired with lock mass gaps to be analysed with a higher degree of accuracy and confidence than currently possible.

3 Inter and Intra Laboratory Reproducibility of Liquid Chromatography-Mass Spectrometry a COMET Study

3.1 Introduction

Metabolic profiling has become a key component of the modern toolbox at the disposal of scientists investigating the integrated function of biological systems and gene environmental interactions [109]. Metabolic profiles characterise the global metabolic state of the system by assaying, without pre-selection, the levels of small molecule metabolites in a biological sample, for example a biofluid or tissue extract. The approach has yielded new insights across fields as diverse as functional genomics, toxicology, molecular epidemiology and clinical studies. The increasing impact of metabolic profiling has reinforced the need for stringent quality control (QC) procedures and measurements of reproducibility. This information is of interest both to research scientists and regulatory agencies across the world.

Liquid Chromatography hyphenated to Mass Spectrometry (LC-MS) has become a key technology for metabolic profiling, because of its ability to detect hundreds to thousands of metabolites with wide ranging chemistries from a single sample in a few minutes with minimal sample preparation. The power of the method has been augmented by the advent of ultra performance systems (UPLC-MS) [110], which achieve a higher resolution and sensitivity than conventional chromatography. While the other key technology for metabolic profiling, Nuclear Magnetic Resonance (NMR) spectroscopy has been shown to be highly stable and reproducible [56–58], this has been harder to achieve in practice using LC-MS. Traditional QC protocols for LC-MS, such as monitoring stable isotope labeled analogues of the analytes of interest, cannot be used in untargeted metabolic profiling since the analytes are unknown a priori [111]. An alternative approach is to use a biological QC sample, comprised of a representative pool of the analytical samples, which is injected frequently throughout the analytical run and monitored for changes in spectrometer response. Several studies have examined within-lab repeatability of the technology using this approach and have found excellent repeatability of mass, retention time and intensities [63, 111–113]. Notably, for a large proportion of metabolites, these within-lab figures complied with FDA guidelines on acceptable reproducibility for regulatory acceptance. However, to date there have been no studies examining the between-day and between-lab reproducibility of LC-MS for urinary metabolic profiling.

To address this need, we have conducted a large inter-laboratory reproducibility study for UPLC-MS metabolic profiling. The study was conducted within the framework of the second Consortium for Metabonomic Toxicology (COMET-2) an academic collaboration between Imperial College London, four pharmaceutical companies and two instrument manufacturers. We aim to assess within-run, within-lab (between-day) and between-lab reproducibility of UPLC-MS for metabolic profiling of urine. To assess reproducibility of both known compounds and the much larger set of unknowns typically assayed in metabolic profile experiments, the design incorporated a set of stable isotope labeled standard compounds spiked into normal human urine samples. To gauge linearity of response within a complex urinary background matrix, the samples were subject to a standard dilution series. A more detailed description of the experimental design is given in the methods section. Analysis of the unknown signals is left to later work; here we focus on the standard compounds and how they inform about within and between lab reproducibility of UPLC-MS for urinary metabolic profiling.

3.2 Methods

3.2.1 Experimental Design and Sample Preparation

The experimental design comprised a Standards Dilution Series (SDS) for system suitability assessment and a Standards in Urine Dilution Series (SUDS) to assess reproducibility. The SDS samples consisted of standard compounds in water, whereas the SUDS comprised the same standards spiked into a pooled urine sample collected from 3 healthy volunteers and mixed in equal proportions. Fourteen metabolites commonly observed in human urine were chosen as standard compounds and obtained in stable isotope (^2H or ^{13}C) labeled forms. A full list of the compounds chosen is seen in table 3.1. We note that deuterium exchange can complicate quantitation of ^2H labeled compounds. However, we found no systematic difference between the reproducibility of ^2H or ^{13}C labeled compounds, suggesting this was not a problem in our study. All samples were prepared in a single batch at Imperial College London.

All reagents including LC-MS grade water, acetonitrile with pre-added 0.1% formic acid, and standard compounds were obtained from Sigma Aldrich (Gillingham, UK). Columns and maximum recovery vials were donated by Waters Corporation (Milford, USA). The urine samples were treated with sodium azide (at 0.1 g per 100 ml of urine) and filtered at 0.2 micron to remove particulates and cells. Stock mixtures of the standards (1mg/ml) were prepared, and diluted to a working concentration of $10\mu g/ml$ per standard. Similarly the standard mixture was spiked into the pooled urine at a similar concentration. These stock solutions were subject to four two-fold dilutions resulting in dilutions of 1, 1/2, 1/4, 1/8, 1/16. These dilutions are identified as SDS1 or SUDS1 for the starting dilution then sequentially as SDS2, SUDS2 until SDS5 and SUDS5 for the most diluted sample. The SDS samples were employed as a system suitability check to determine whether a given run was to be accepted to the study (see below). Each participating laboratory received two identical sample batches to accommodate positive and negative mode analysis. Each batch was further divided into two phases to assess within-lab reproducibility. In phase 1, a full SUDS was run. In phase 2, run at least one week after phase 1, replicate aliquots of SUDS 1, 3 and 5 were run. To facilitate handling and storage the four batch / phase combinations were packed in 4 separate boxes and shipped on dry ice to the participating laboratories. Each laboratory recieved the same solvents and column.

3.2.2 UP-LC-MS Analysis

In each lab, samples were analysed using a Waters®(UPLC) chromatography system connected to a LCT Premier TOF mass spectrometer (Waters MicroMass, UK) with electrospray ionisation (ESI). The same batch of C18 Acquity (2.1 x 100mm 1.7μm) columns (Waters) was distributed to each laboratory. The same chromatographic gradient was used in all laboratories, over 14 min, with mobile phases A (water, 0.1% formic acid) and B (acetonitrile, 0.1% formic acid). The gradient is summarised as follows: 0 min 100% A, 0% B; 4 min 80% A, 20% B; 9 min 0% A, 100% B; 11 min 0% A 100% B; 12 min 100% A, 0% B; 14 min 100% A, 0 % B. The flow rate was 0.5ml/min and an injection volume of 5μl was used. The mass spectrometer was set to scan from 85-1000 in mass to charge ratio (*m/z*), calibrated with sodium formate. Sample temperature was maintained at 4°C and column temperature at 40°C. To allow for differing MS response, each lab optimised the individual MS conditions separately (e.g. cone and capillary voltage, desolvation temperature and gas flows, see table 7.2). Technical replication was achieved by repeated injections from the same vial and this was performed 10 times for SUDS 1, 3 and 5 and 3 times for SUDS 2 and 4. A randomised run order was chosen, ensuring that it was orthogonal to the dilution factors of both SUDS and SDS, to avoid systematic bias resulting from instrumental drift. The same run order was used for each mode and in each laboratory. Data were sent back to Imperial College in raw format.

3.2.3 Data Analysis

These data were converted to mzXML format using massWolf (SPC tools, Institute for Systems Biology, Seattle, WA, USA). The files were grouped according to dilution series (SUDS/SDS), ionisation mode (+/-) and laboratory (A-C). XCMS [49] (version 1.23.3) was used to process each dataset separately. Files were processed with gap filling [108] and the following parameters: peak picking method = centWave, peakwidth 2-20 sec; grouping bw/bandwidth of 30 sec, mzwid= 0.07; retcor missing $\frac{n}{2}$ extra maximum of 3, span maximum 0.7. The retention time alignment was iteratively processed until a global drift below 2 s was observed using the loess method. The second grouping was dependent on 'retcor' and was taken as the global retention time drift. The standard compound signals were identified in the XCMS output table using an in-house adduct finder script written in R. This script used estimated retention times based on a pilot study using an identical experimental set-up (data not shown). The adduct finder used a comprehensive list of regularly observed adduct forms, chosen from the literature [114] and the R package CAMERA [51]. The list is seen in table 7.1. The masses of each adduct was added or subtracted from the expected mass of each compound in positive and negative mode, respectively. Mass windows of 20 ppm and retention time (RT) windows of ±20 s, around the values estimated from a pilot study seen in table 3.1, were used to find the standard compound adducts. The pilot study confirmed that, as expected all 14 compounds ionised in positive mode, and 10 ionised in negative mode. The adduct finder detections were inspected manually using extracted ion chromatograms (EIC) to confirm the presence of a peak. We analysed the performance of the adduct finder to assess the likelihood of false identifications of the standard compounds.

Standard	Formula	Accurate MW	RT	POS ION	NEG ION
dimethylglycine-d_6	$C_4H_9NO_2$	109.1010	0.5200	Y	N
L-DOPA-ring-d_3	$C_9H_{11}NO_4$	200.0877	1.0400	Y	Y
DL-methionine- ^{13}C	$C5H_{11}NO_2S$	150.0545	1.0400	Y	Y
acetylcarnitine-d_3	$C_9H_{17}NO_4$	206.1347	1.0900	Y	N
nicotinamide-d_4	$C_6H_6N_2O$	126.0731	1.2400	Y	N
dopamine-d_4	$C_8H_{11}NO_2$	157.1041	1.3100	Y	Y
Succinic acid-d_4	$C_4H_6O_4$	122.0517	1.4500	Y	Y
DL-leucine-d_3	$C_6H_{13}NO_2$	134.1135	1.6600	Y	Y
Glutaric acid-d_4	$C_5H_8O_4$	136.0674	1.9000	Y	Y
DL-phenylalanine-13C	$C_9H_{11}NO_2$	166.0824	2.2500	Y	Y
Adipic acid-d_8	$C_6H_{10}O_4$	154.1082	2.5300	Y	Y
tryptamine-d_4	$C_{10}H_{12}N_2$	164.1252	3.1900	Y	N
Heptanedioic acid-d_4	$C_7H_{12}O_4$	164.0987	3.4000	Y	Y
Hippuric acid-d_2	$C_9H_9NO_3$	181.0708	3.4000	Y	Y

Table 3.1: The isotopically labeled standards ordered by retention times with accurate masses, expected retention times (calculated from the pilot study) and whether they should ionise in positive or negative mode.

3.2.4 Performance Analysis of the Adduct Finder

The reproducibility analysis depends heavily on the ability to automatically identify the standard compounds and their adducts in the XCMS output. With low numbers of samples, this can be performed manually. However, this study incorporated 81 samples analysed in two modes across three labs, with 14 labeled standards and a possible 23 different adducts. This detection of the resulting 156492 possible unique ions could not therefore be verified manually and thus an automated adduct finder script was developed for this purpose. To quantify the false positive rate of the adduct finder, three datasets were tested in which none of the labeled standard compounds used in this study should be present. The datasets were as follows: an HPLC-TOF-MS human plasma dataset, with 2368 metabolite features and an UPLC-TOF-MS human urine dataset that reported 788 features. Both sets of data had been processed with XCMS similarly to the main reproducibility study. Additionally, a synthetic dataset was made by modelling the distribution of RT and *m/z* from the SUDS data of lab B. The RT distribution was modelled by a mixture of two Normal distributions with means $\mu 1 = 110$ & $\mu 2 = 425$ s and standard deviations $\sigma 1 = 60$ & $\sigma 2 = 140$ s and truncated at 0 minutes. The *m/z* distribution was modelled by a Normal distribution, with $\mu = 200$ Da and $\sigma = 100$ Da. Using these parameters, we simulated data sets of up to 1,000,000 features. No false positives were found in either of the two biological data sets. For the synthetic data set a maximum false positive rate of 0.07% was recorded. These tests indicate that it is very unlikely that any of the detections reported in this study are false.

3.2.5 System Suitability Assessment

The SDS samples were used to assess system suitability, i.e. to determine whether each analytical run was of acceptable quality to include in the study. To pass this assessment, we required a minimum of 40% of the standards to be detected by the adduct finder as molecular ions (e.g. $[M + H]^+$ or $[M - H]^-$ ions). The peak shape of the detected ion was also required have a good signal to noise ratio (minimum 5) as well as being approximately Gaussian, e.g. the peak could not be flat topped due to detector saturation. The EICs of the molecular ion ($[M + H]^+$ and $[M - H]^-$) of each standard compound detected in the SDS are shown in appendix figure 7.2. These plots illustrate that for lab C negative mode, no molecular ions corresponding to the standard compounds were observed. Additionally,figure 7.2b shows that few adducts of the standard compounds were detected in lab C negative mode. Consequently, the lab C negative mode data was removed from the reproducibility analysis. All other lab, mode and phase combinations passed the system suitability assessment.

Phase	A+	A-	B+	B-	C+
1	78	70	100	70	71
2	78	70	100	60	71

(a) All adducts

Phase	A+	A-	B+	B-	C+
1	78	70	100	50	71
2	78	70	85	40	64

(b) Molecular ion only

Table 3.2: The percentage of detected compounds as either molecular ion or adduct. Each lab is shown in both positive and negative mode.

3.2.6 Reproducibility Measures

In this study [68] we report the reproducibility of all three recorded parameters: *m/z*, RT and intensity. For mass, we report the mean mass measurement accuracy [115] (hereafter referred to as mass accuracy). For retention time, we report retention time drift corresponding to the retention times before correction by XCMS. For intensity, we report two measures. For within-day repeatability, we report the coefficient of variation (CV) within each replicate group. For between-day and between-lab reproducibility we report the R^2 of a linear regression between the mean intensities of ions detected in both conditions.

3.3 Results / Discussion

3.3.1 Detection of Spiked-In Standard Compounds

Raw extracted ion chromatograms (EICs) were inspected for each adduct of each standard compound. Manual inspection of the peak shape and intensity confirmed the authenticity of each detected peak. Figure 3.1 illustrates a typical set of EICs from the dilution series for hippuric acid$-d_2$, showing a well-defined peak in each case. The signal is seen to decrease in intensity according to the two-fold dilution factors. As expected, we observed differences between the three labs of both the mean and spread of retention times, with median retention times for labs A, B and C of 190 s, 196 s and 193 s respectively. Lab C had the highest spread of retention times and poorest peak shapes which could be due to changes in the platform setup such as change of column temperature[1], fluctuating column pressures or flow rates and changes in ionisation efficiency. The other main difference is that of peak intensity with Lab B having by far the highest intensity, while labs A and C record similar intensities. Different peak intensities are to be expected when comparing MS results from different instruments and laboratories

[1] Under normal running conditions column age could also effect the retention times however, here all labs received the same column therefore all columns were the same age.

since source settings are usually optimised for each individual experiment (as in this case, see appendix table 7.2 for details of source parameters). It should also be noted that Figure 3.1 shows pre-aligned data and that the post-alignment peaks are all within 2 seconds drift.

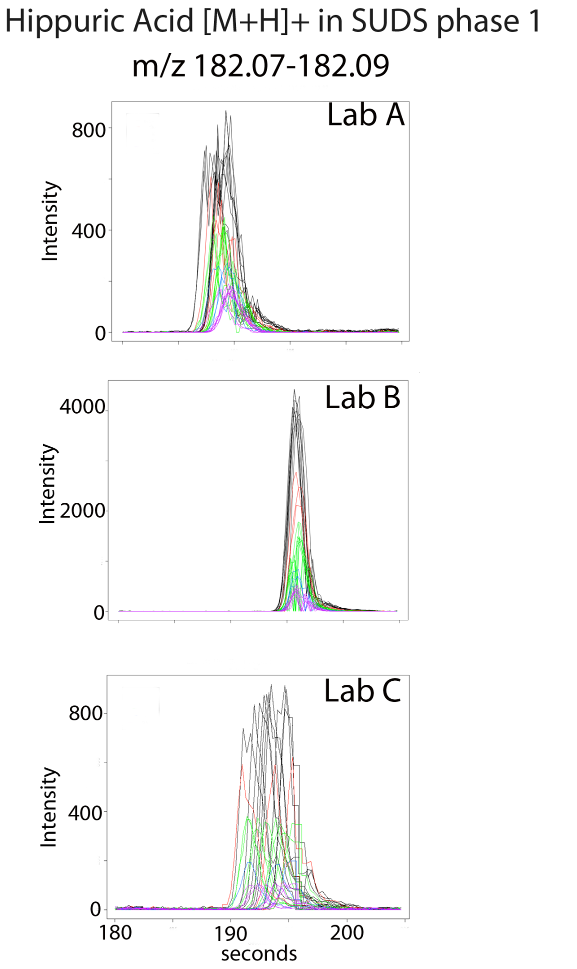


Figure 3.1: Extracted ion Chromatography of raw pre-aligned data from each laboratory for Hippuric acid-d_3 $[M+H]^+$

The percentage of standard compounds detected in each lab, phase and mode is given in table 3.2. When considering all adduct forms, the majority of the standard compounds were detected in all data sets, with a minimum of 60% detected in lab B, negative mode. The number of standards detected was similar between the two phases, indicating good within-lab reproducibility. Detection rates were always lower in negative than positive mode, though the only by 8% in lab A. Lab B, however, showed a greatly increased detection rate in positive mode, detecting all standards, while labs A and C detected 78% and 71% respectively. Interestingly, labs A and C showed little difference in detection rates when the analysis was restricted to molecular ions, while the detection rate in lab B dropped by up to 20%, indicating a significantly increased amount of adduct formation demonstrated later in the chapter 3.3.4. To check that these differences were not simply a result of the peak detection algorithm settings, we reprocessed the data using a lower signal-to-noise threshold with almost identical results (data not shown). Ionisation efficiency can be lower in negative mode [116]. This effect could have resulted in the corresponding lower detection rate the standard compounds. Conversely, Nordstrom *et al* [117] found that metabolite coverage was similar between ESI- and ESI+ modes in human blood serum extracts and it may be that the number of standard compounds in our study is too small to draw a general conclusion in this area. For the remainder of the analysis, molecular ion data is presented in the main figures, with equivalent results including all adduct forms available in the appendix.

3.3.2 Within Run Repeatability of Intensity, Mass Accuracy and Retention Time

For each standard compound, we calculated reproducibility measures for intensity, mass to charge ratio and retention time for each set of analytical replicates across dilutions, phases, labs and ionisation modes. Figure 3.2 reports this information for molecular ions ($[M + H]^+$ and $[M - H]^-$); results for the full range of adducts are reported in appendix figure 7.1.

Overall intensity reproducibility was good, with median CVs of 12% and 17% for positive and negative modes across all labs, phases and dilutions. The majority of CVs were low, with 70% and 74% of features having CVs below 20% in positive and negative mode respectively. Across labs, median CVs were 11%, 12% and 9% for labs A, B and C in positive mode and 17% and 15% for labs A and B in negative mode. The positive mode exhibited more high CV outliers, particularly nicotinamide (162% and 62% CV in lab B phase 2, dilutions 1 & 2 respectively), acetylcarnitine (58% and 27% CV in lab A, phase 1 & 2, dilution 1 respectively). These high values are primarily due to the untargeted nature of global metabolite profiling, where more than one unique peak can be assigned to a given feature by automated processing software, depending on the *m/z* and RT tolerances set. Note that, although such errors could be corrected manually, they are representative of error that would occur in any automated global profiling strategy, and

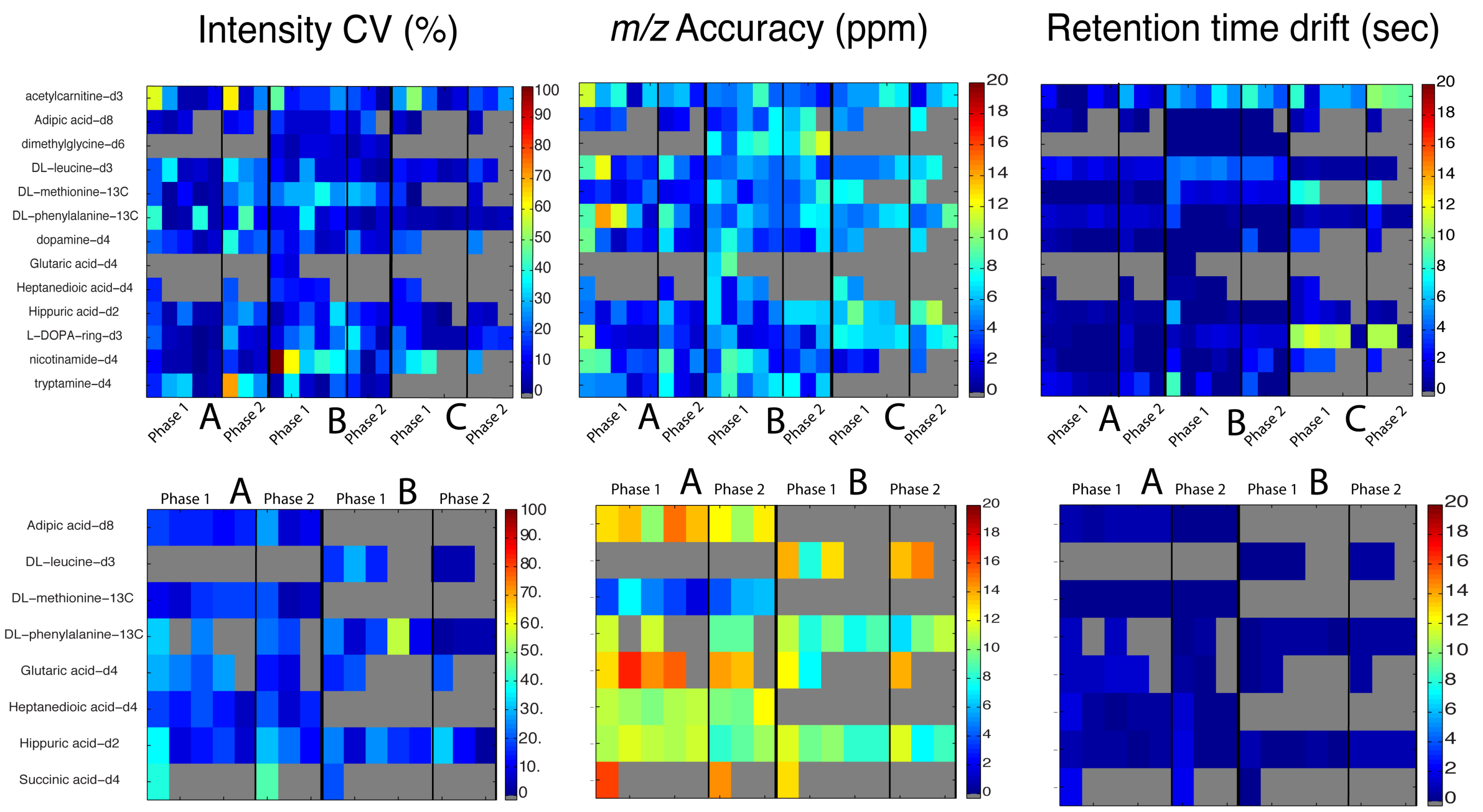


Figure 3.2: Reproducibility of intensity, mass accuracy and retention time for the standard compounds. Heat maps show from left to right the coefficient of variation (CV) of intensity, mean mass measurement accuracy (ppm) and retention time drift (s). Upper panels show positive mode, lower panels show negative mode. Each heat map cell reports data for a one dilution of a given phase (1&2) and lab (A, B & C). Dilution increases left to right. Lab C negative mode data did not meet required QC criteria and was therefore not included in the analysis. Grey cells indicate the standard compounds that were not detected.

have therefore not been corrected here. Tryptamine also shows an unusually high CV (71% CV in lab A, phase 2, dilution 1). This was due to the fact that the corresponding peak was not detected in all samples, an issue that affects the higher dilutions more severely, as a non-detected peak will contrast more strongly with the high counts for detected analytes. Many other MS based metabolic profiling studies have reported typical CVs of approximately 20%. Using an HPLC-ESI-linear ion trap for urine, our study has shown very similar values to Gika *et al* [65] for intra-day reproducibility. Our values show that 73% and 64% of peaks have a CV below 20% in positive and negative modes respectively. A study on human cell extracts analysed by LC-MS found inter-day CVs below 25% for 6 spiked-in compounds. Crews *et al* analysed serum and cerebral spinal fluid (CSF) by HPLC-TOF-MS to show the overlap between the two biological matrices [118]. They also demonstrated that the analytical variation in intensities was 15%. However, the biological variation between the samples was high, > 50% and this is usually the case irrespective of the analytical method [56,57]. Our work is consistent with Masson *et al* [119] work where an optimised extraction protocol for liver recorded analytical CV of approximately 20% using UPLC-TOF-MS.

We observed a median mass accuracy of 4.4 ppm for positive mode but a substantially poorer value of 11.8 ppm in negative mode. The mass accuracy was consistent both within laboratories and across phases with median values of 3.5, 5.2 & 6.2 ppm for labs A, B and C in positive mode and 11.0 & 9.9 ppm for labs A and B in negative mode. As with intensity, the lower mass accuracy in negative mode is to be expected due to lower ionisation efficiency. In positive mode, the compounds with the lowest mass accuracy were phenylalanine and acetylcarnitine. These outliers could be attributed to saturation at high concentrations (phenylalanine) and the multiple peak problem described in the above paragraph (3.3.2.3) for acetylcarnitine but also in more detail in figure 1.2.

Drift in chromatographic retention times is a well-known complication in metabolic profiling studies, complicating the process of matching peaks between samples. However, with UPLC the drift problem is much reduced, confirmed in our study. The median RT drift was remarkably good at 0.73s in positive mode and 0.38 s in negative mode. 75% of RT drifts were below 2.12s in positive mode and 0.67s in negative mode. These figures can be compared to typical RT peak widths of 2-20s. As illustrated in Figure 3.1, RT drift varied across the labs, with median values of 0.88, 1.61 and 3.55s for Labs A, B and C respectively in positive mode and 0.56 and 0.40 s for Labs A and B in negative mode. It is important to bear in mind that the chromatographic conditions were identical between modes and so ionisation mode cannot affect RT drift. Our results are consistent with Lange *et al* found that RT drifts were below 5 seconds for the majority of peaks in an untargeted UPLC analysis [44].

3.3.3 Linearity of Intensity with Dilution

Linearity in response of any analytical method is crucial. However, there have been no studies investigating the linearity of response for multiple metabolites in complex biological mixtures with TOF-MS. Table 3.3 reports the R^2 for the regression of intensity against dilution for the standard compounds in each lab, phase and mode (results including non-molecular ions are given in appendix table 7.3 & 7.4). Overall, highly linear responses were obtained across the 16-fold dilution gradient with median R^2 values of 0.95 (compounds detected in $n = 63$ lab/phase combinations) in positive mode (3.3a) and 0.93 ($n = 15$) in negative mode (3.3b). Across the labs, median R^2 values of 0.89 ($n = 22$), 0.95 ($n = 22$) and 0.98 ($n = 19$) were obtained for labs A, B and C in positive mode and 0.93 ($n = 12$)& 0.89 ($n = 5$) for labs A and B in negative mode respectively. This indicates minimal variation in linearity between the labs. Median R^2 values for phases 1 and 2 were 0.95 ($n = 32$)& 0.95 ($n = 31$) in positive mode, with 0.89 ($n = 9$) and 0.93 ($n = 8$) in negative mode, again indicating minimal variation between phases in linearity. The poorest linearity was observed for tryptamine with an R^2 of 0.41 for lab A in positive mode phase 2, due to the signal falling below the limit of detection at lower dilutions. Hippurate in negative mode lab B phase 2 also showed poor linearity ($R^2 = 0.49$). This was due to a single

Compound	A Pos	A2 Pos	B Pos	B2 Pos	C Pos	C2 Pos
Acetylcarnitine$-d_3$	0.69	0.57	0.79	0.97	0.53	0.91
Adipic acid-d_8	0.99	0.98	0.97	0.97	0.99	0.99
Dimethylglycine-d_6			0.96	0.97		
DL-leucine$-d_3$	0.91	0.75	0.94	0.99	0.99	0.98
DL-methionine-^{13}C	0.96	0.90	0.89	0.92	0.97	0.98
DL-phenylalanine-^{13}C	0.68	0.74		1.00	1.00	1.00
Dopamine$-d_4$	0.93	0.82	0.95	0.93	0.95	0.95
Glutaric acid$-d_4$			0.93			
Heptanedioic acid$-d_4$	0.94	0.63			0.96	
Hippuric acid$-d_2$	0.89	0.85	0.96	0.97	0.98	0.99
L-DOPA-ring$-d_3$	0.91	0.84	0.95	0.93	0.97	0.98
Nicotinamide$-d_4$	0.96	0.94	0.87	0.91	0.95	0.97
Tryptamine$-d_4$	0.89	0.41	0.95	0.98		

(a) Positive mode

Compound	A Neg	A2 Neg	B Neg	B2 Neg
Adipic acid-d_8	0.95	0.93		
DL-leucine$-d_3$			0.95	0.98
DL-phenylalanine-^{13}C	0.80	0.93		
Glutaric acid$-d_4$	0.83	0.98	0.89	
Heptanedioic acid$-d_4$	0.95	0.96		
Hippuric acid$-d_2$	0.90	0.89	0.70	0.49
Succinic acid$-d_4$	0.87	0.93	0.69	

(b) Negative mode

Table 3.3: Dilution series R^2 values for each lab, phase and compound. Blank entries are where the compound was not detected in more than 50% of the samples for one dilution series.

replicate in the most concentrated group having a much lower intensity than the majority of the samples. If this sample is removed the R^2 increases to 0.94. DL-Phenylalanine showed a markedly lower linearity in lab A than the other two labs. This was due to higher variability of intensity at dilutions 1 and 4 in lab A as seen in figure 3.1.

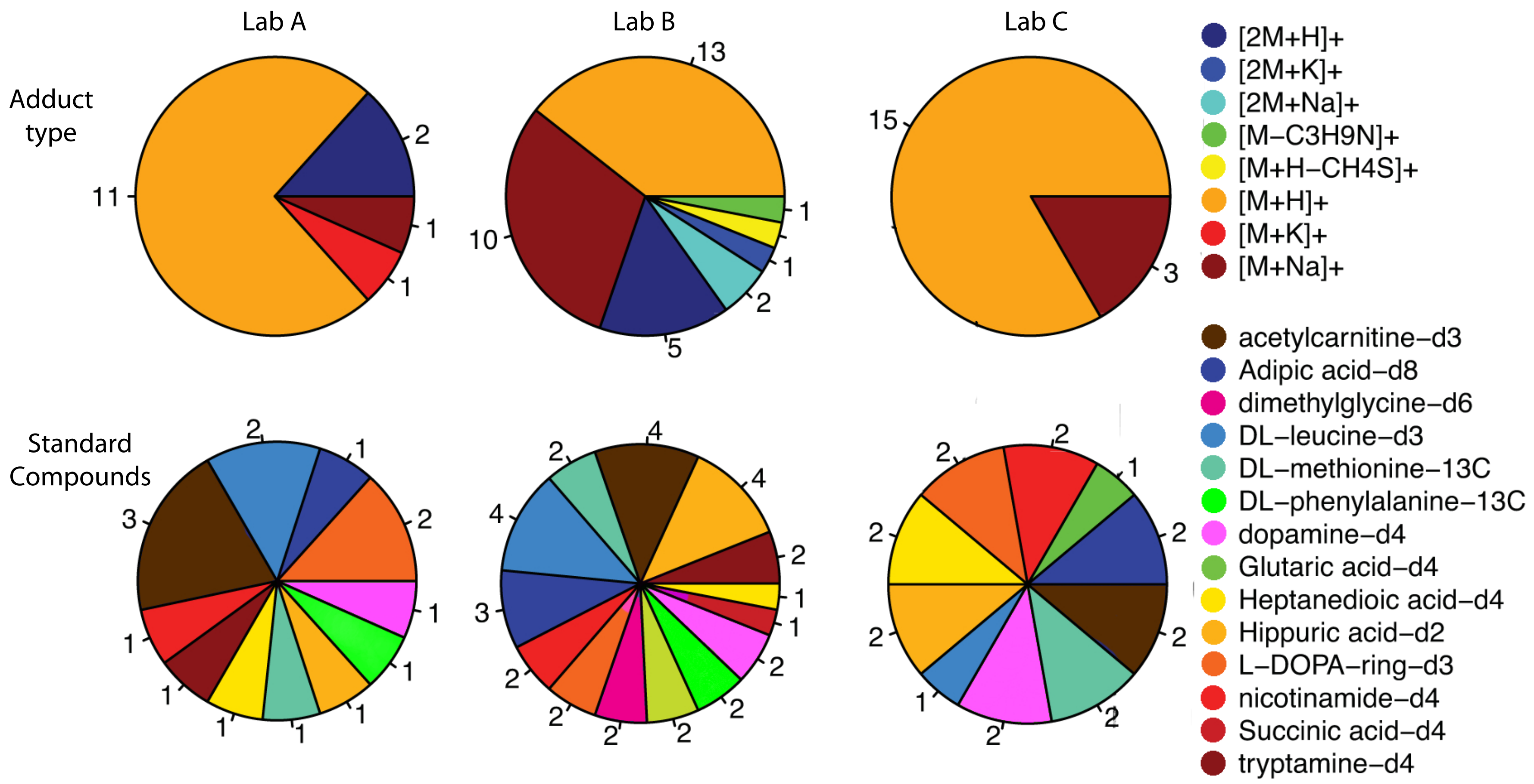
3.3.4 Adduct and Dimer Formation

To facilitate a straightforward comparison between the phases and laboratories, the results reported thus far have been confined to molecular ions. Nonetheless, many adducts and dimers of the standard compounds were detected, a total of 35 and 20 distinct ions (including molecular ions) in positive and negative modes respectively, illustrated in Figure 3.3 & 3.4.

In both modes, the molecular ion ($[M + H]^+$ or $[M - H]^-$) was the most commonly identified adduct form. Clear differences between laboratories were seen in the amount and types of adducts found. For labs B and C in positive mode, the sodium adduct, $[M + Na]^+$ was the next most commonly detected form; however this was not seen in lab A. Overall, lab B detected the most adducts, with 12/14 standard compounds detected in more than one adduct form in positive mode (4/7 in negative mode). These differences may be due to differing optimal MS source settings, such a source voltage and source temperature. Only one compound, acetylcarnitine, showed a propensity for consistent adduct formation in all labs (positive mode), while three others were detected as adducts in two or more labs (phenylalanine and hippuric acid in negative mode and L-DOPA in positive mode).

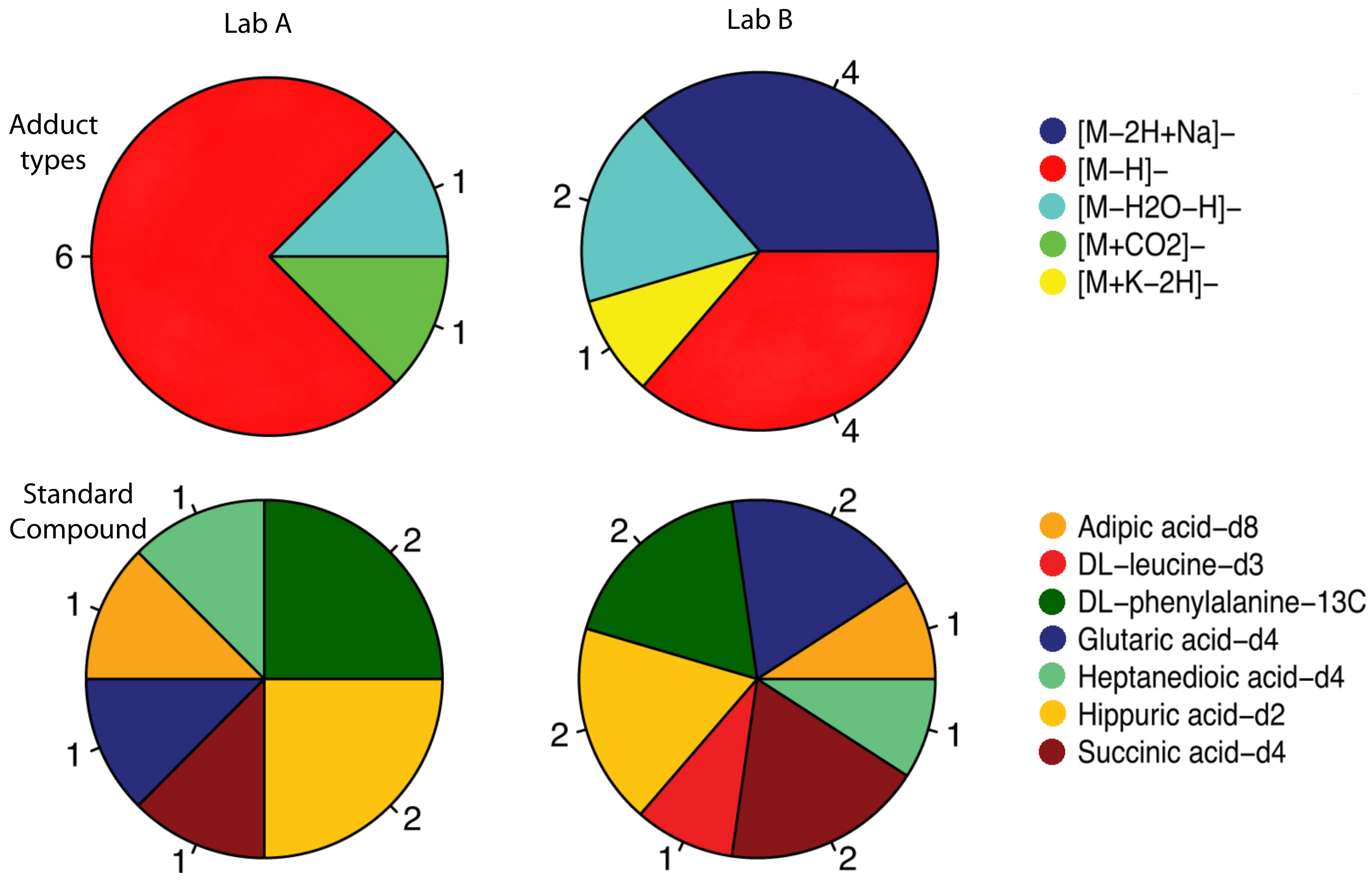
3.3.5 Between-Day Reproducibility

We estimated inter-day reproducibility using a subset of identical samples run in phase 2, at least one week after phase 1. Spectrometer intensity response can vary over time, especially between before and after cleaning the instrument. Therefore in this instance the comparison of absolute intensity measurements does not demonstrate the normal reproducibility. Instead we assessed reproducibility by examining quality of the linear relationship between intensities in the two runs, as shown in Figure 3.5. R^2 values from the regression of phase 2 against phase 1 intensities were above 0.84 in all cases, with a mean of 0.93, indicating excellent reproducibility. Two comparisons showed lower R^2 values: lab A negative mode, and lab B positive mode. In



(a)

Figure 3.3: Adduct formation for each lab and mode in phase one in positive mode. In each panel the upper row of pie charts report the adduct forms observed, while the lower rows report the numbers of adducts observed for each standard compound.



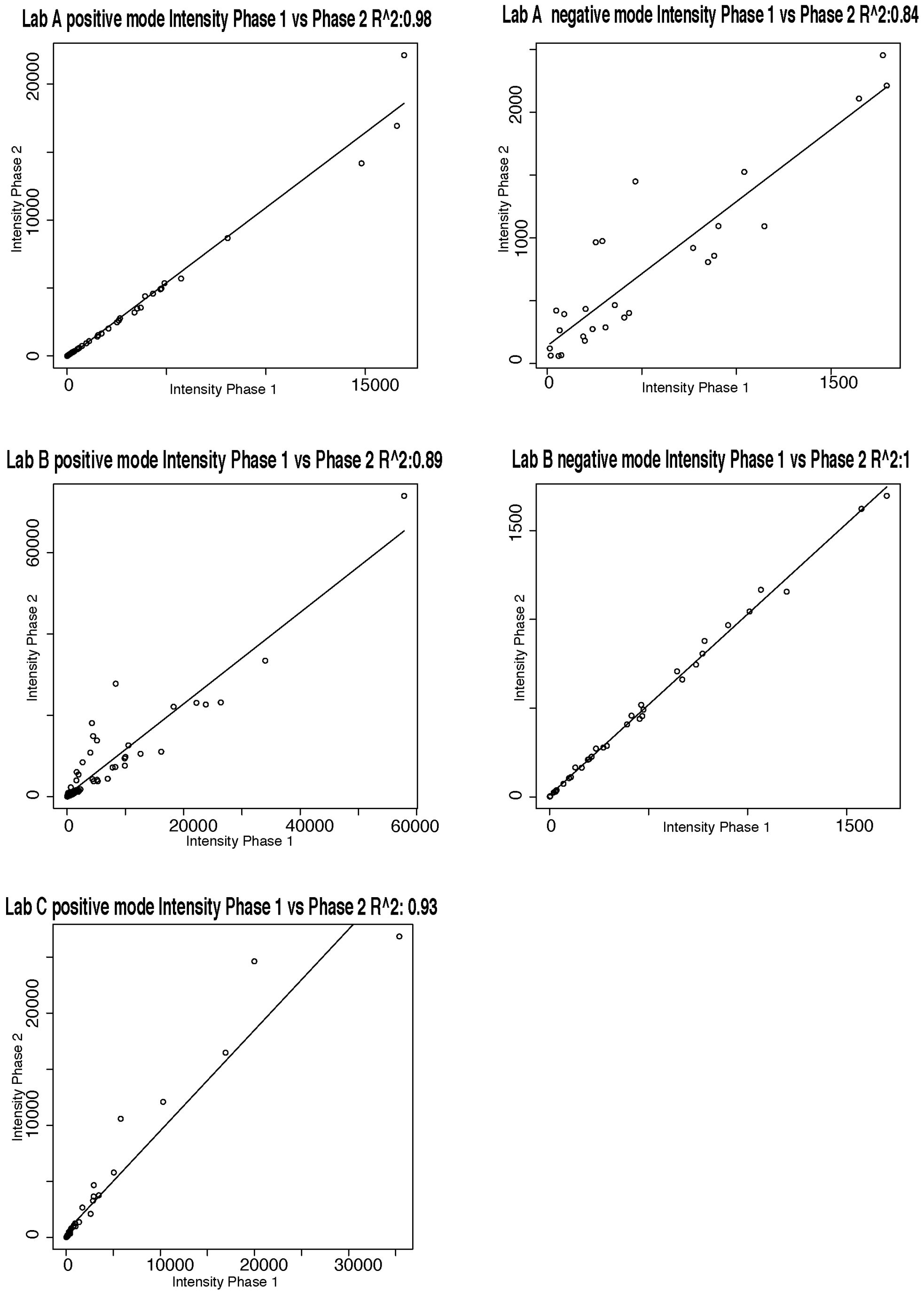
(a)

Figure 3.4: Adduct formation for each lab and mode in phase one in negative mode. In each panel the upper row of pie charts report the adduct forms observed, while the lower rows report the numbers of adducts observed for each standard compound.

neither case could the lower reproducibility be attributed to individual (or groups of) standard compounds, adduct types, or dilutions. To our knowledge, no other studies have investigated inter-day reproducibility of LC-MS based metabolic profiles of urine.

3.3.6 Between-Lab Reproducibility

The lab-to-lab studies measure one of the most important aspect of analytical reproducibility. As for the within-lab measurement, it is not meaningful to compare absolute responses and thus we look for a linear relationship between intensities in pairs of labs. Further, given that each lab optimised its own spectrometer settings, ionisation efficiency would be expected to be compound specific and thus each compound was modelled separately. Figure 3.6 shows a separate regression line for each standard compound adduct which was detected in all dilutions for each lab pair and illustrates that the response is highly linear between each pair of labs in both positive and negative modes. In positive mode 24 compound-lab pair matches were made with a median R^2 of 0.96, and 79% of values above 0.90 indicating excellent reproducibility. In negative mode, only four compounds could be found in both labs, yet the median R^2 was 0.98, and three of the four had values of 0.96 or higher. R^2 values for each standard can be found in appendix table 7.5 & 7.6 . Some compounds showed lower R^2 values. For example, acetylcarnitine gave an anomalous response in lab B at high intensities, but a highly linear relationship was seen for this compound when comparing labs A and C ($R^2 = 0.96$). L-DOPA appeared to saturate the detector in lab A, but again this effect was not seen in labs B and C ($R^2 = 1.00$). To our knowledge there have been no studies investigating inter-lab reproducibility of LC-MS for metabolic profiling of urine. For NMR, Keun *et al* [56] have shown excellent inter-laboratory reproducibility with R^2 values above 0.93 for three metabolites in rat urine measured at 600 Mhz.


Lab A positive mode Intensity Phase 1 vs Phase 2 R^2:0.98

Lab A negative mode Intensity Phase 1 vs Phase 2 R^2:0.84

Lab B positive mode Intensity Phase 1 vs Phase 2 R^2:0.89

Lab B negative mode Intensity Phase 1 vs Phase 2 R^2:1

Lab C positive mode Intensity Phase 1 vs Phase 2 R^2: 0.93

Figure 3.5: Between-day reproducibility of intensity for each laboratory and mode. Each data point corresponds to a single standard compound adduct averaged across all replicates at a given dilution.

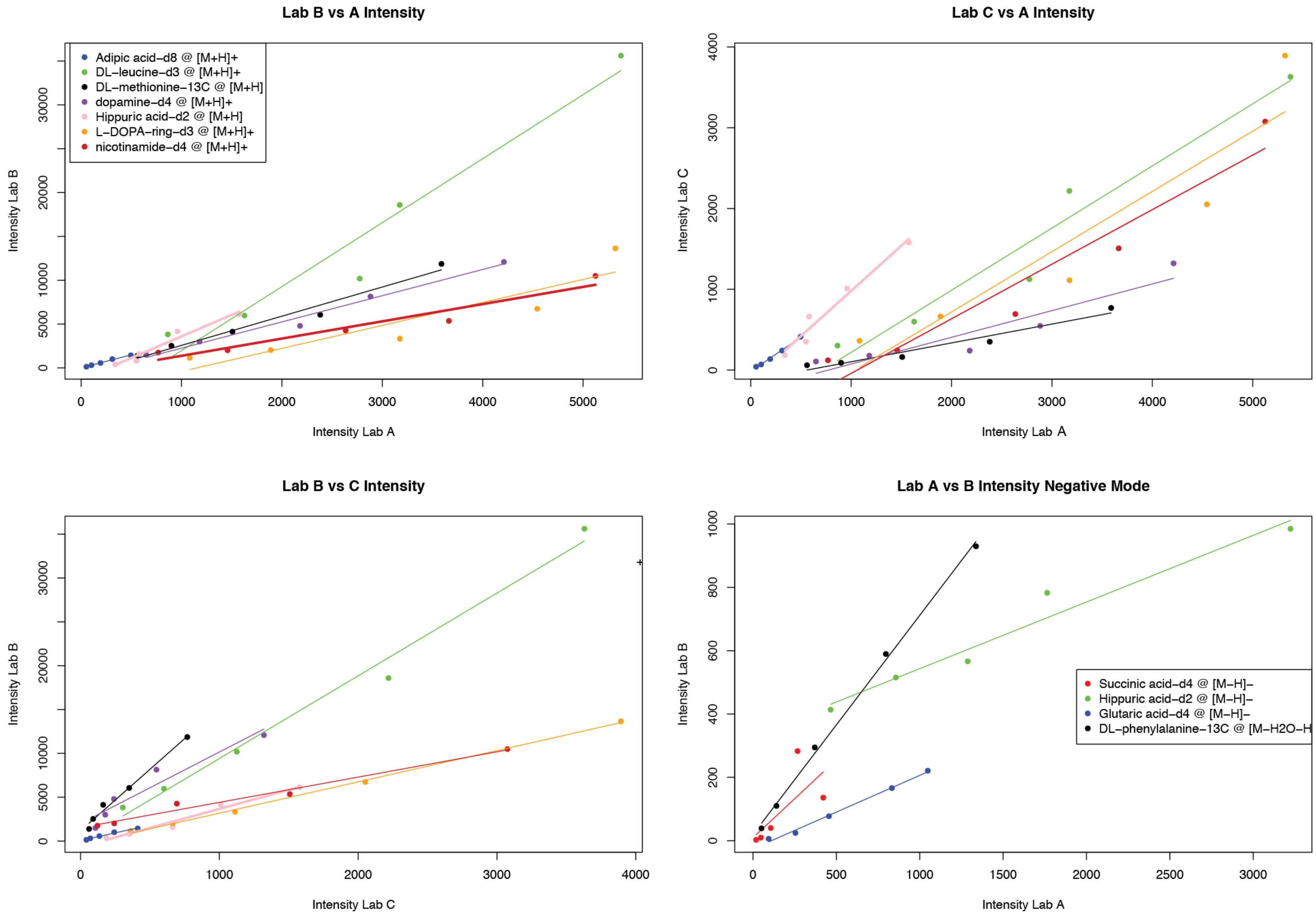


Figure 3.6: Between-laboratory reproducibility for pairs of laboratories in positive and negative modes. Each data point corresponds to a single standard compound adduct averaged across all replicates at a single dilution. Each standard is indicated by a different colour and a separate linear regression line. Key for positive mode: green, DL-leucine; black, DL-methionine; purple, dopamine; pink, hippuric acid; orange, L-DOPA; light green, nicotinamide. Key for negative mode: red, succinic acid; green, hippuric acid; blue, glutaric acid; black, DL-phenylalanine. Acetyl carnitine obtained much higher counts than the other standards, obscuring the response from the other compounds. A version of this figure including acetyl carnitine can be found in figure 7.3 in the appendix.

3.4 Conclusion

The study presented here is the first one to examine reproducibility of UPLC-MS for urinary metabolic profiling comprehensively across three levels: within-day, between-day and between labs. By spiking stable isotope labeled standard compounds into a complex biological matrix, and using standard global profiling software tools, we have been able to estimate reproducibility of untargeted metabolic profiling using identified compounds, thus increasing confidence in our estimates. We have shown that all three aspects of the experiment are reproducible, with median CVs for intensity of less than 18%, median mass accuracy below 12 ppm and median retention time drift less than 0.73 s. Within each run, the intensity response is highly linear for most compounds, with a median R^2 of 0.95 in positive and 0.93 in negative modes, despite the complex background of diverse urinary constituents. Between-day reproducibility was excellent (even with runs separated by at least a week in which instruments were used for other projects), with a mean phase 1 to phase 2 correlation R^2 of 0.93 across the labs and modes. All compounds were found to form adducts and/or dimers to an extent which varied widely between the labs. Finally, we have shown that intensity responses are also highly reproducible both within a lab and between labs with median inter-lab R^2 of 0.96 in positive mode and 0.98 in negative mode respectively. This implies that the intensity ratio for a given metabolite across different samples (for example the fold change between two biological conditions) will be highly reproducible. While we have not shown in detail we suggest the need for strict system suitability and quality control criteria. Overall, we can conclude that UPLC-MS is a highly valuable technique for global metabolic profiling, yielding measurements of thousands of features which are highly reproducible both within, and between different labs.

4 Characterising the Static Correlation Structure in LC-MS Data for Structural Identification and Removal of Analytical Artifacts

4.1 Introduction

Mass spectrometry is a commonly used method for exploring biological samples and has become widely used in metabolomics. Samples are analysed according to their mass to charge values (*m/z*), which is helpful in metabolomics with the limited mass range. Used in conjunction with mass spectrometry to separate chemicals according to chemistry is liquid chromatography (LC). A final 3 dimensional dataset is produced for each experiment with *m/z*, retention time (rt) and intensities. As discussed in section 1.3.1 many software platforms have been designed to analyze this data. Among these are openMS [120] a popular proteomic software platform, MZmine2 [47] which places an emphasis on graphical output and display and XCMS. XCMS [49] is a popular metabolomics platform written in R and is highly modular. Commonly, these platforms find 1000s of *m/z*-rt pairs or features across a dataset. However, these features are not unique to a particular molecule and any one molecule can have multiple ions. These ions are adducts, isotoplogues and fragments. The intensity of the variables from MS are highly correlated. These correlations come about by different processes taking place. These processes can be :

- Structural process Structurally related ions such as isotopes and adducts.

- Biological processes chemicals are biologically related in a pathway or are being controlled by the same biological event.

- Analytical process A non-biological or structural chemical change in the way that each metabolite is related to each other. This could happen due to column changes with certain chemistries as the alky chains gain more or less affinity for the chemistries. Another way would be due to poor extraction efficiency, where some metabolites were extracted at a lower or higher efficiency between samples. This change can be a non-linear change however, here we have looked at a linear change.

There have been many publication on the identification of structural pairs within LC-MS datasets. They have shown that these structural pairs can be identified a number of ways including intensity ratios, retention time (RT) windows and correlation [106, 121, 122]. However, correlation is a common area of research for identifying structural links especially in automation with programs [51, 123, 124]. There are two main ways in which the data can be correlated find adducts, fragments and isotopes in these large metabolomics datasets. The difference between them is the data that they use in the correlation. The two main ways are:

- Within file correlation (WFC) - This is where the peak height, intensity data from a centroided peak (not peak integration) is used for correlation. The correlation can only happen between peaks that are within a selected RT window. This is the method that Kuhl *et al* [51] uses and utilises a density window to allow for a non statically defined RT window. To define the window the peak with the largest maximum peak intensity is chosen. Then the median RT of that integrated peak is used to make the RT window. All peaks that are within the window are chosen for pairwise correlation. This is outlined in figure 4.1. Using a similar method Scheltema *et al* reduces the LC-MS datasets and eases identification [125]. This method was shown to reduce the dataset by 60% on test data. A normal range of structural pairs in general is between 50-75% of the data. Brown *et al.* [80, 122] uses within file Pearson correlation for identification of the frequency of adducts and isotopes in a dataset. No RT window was set allowing a wide analysis of the accuracy of the required RT window. The most frequent structural identification was ^{13}C. A graphical user interface for xcms [49], IDEOM [124] uses the same method of WFC for structural pair identification.

- Across file correlation (AFC) - AFC is an easier technique to implement, where the integrated intensity of the peak from each file for any pair of features are correlated. This is the method that Alonso *et al* [123] uses in the publication. The RT window is defined in a stricter sense where a simple RT window is implemented around the median RT of the first selected peak in the pair by *m/z*.

After the correlation has shown that the peaks are related, a lookup table of mass differences, can be used to try and identify what the adducts are. Both of these methods are able to successfully identify related structural pairs in the dataset which is normally around 50%-75% of the peaks. In the Kuhl publication a score is made where both WFC and AFC are used to allow for a high confidence in structural pairs. As long as the peak is able to be separated and there are no other peaks that are co-eluting the WFC method is more robust as inter sample noise can lower the correlation coefficient of the AFC method.

There are many other correlations that are not structural. These correlations do not have overlapping RT windows and are defined as biological correlations with the underlying processes causing them outlined above. Biological correlations can only be found using the AFC method with no needed RT window. The correlations also need to be in samples that represent biological

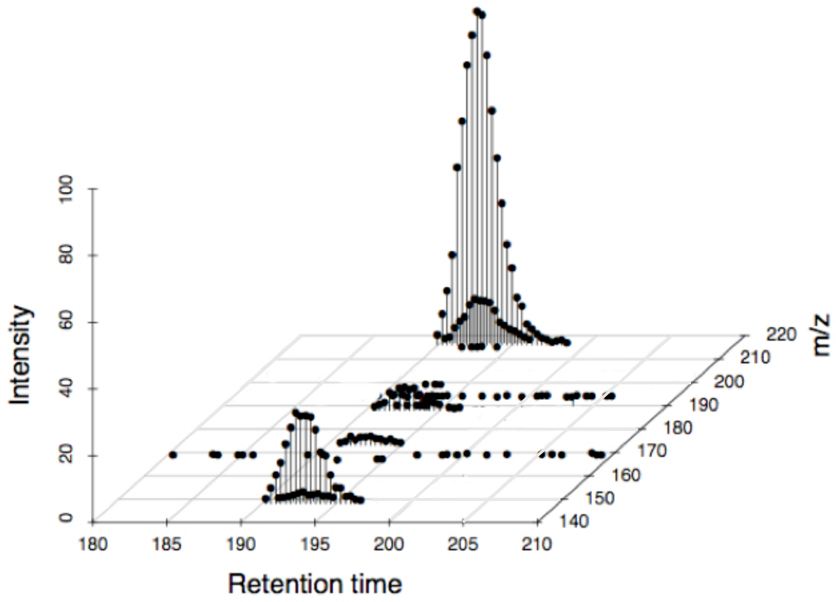


Figure 4.1: The figure demonstrates the intensity data that is correlated in the WFC method. It can be seen how each point on the peak is correlated between the other peaks. Only peaks within the RT window will be correlated.

change and not QC or QC like samples.

Before the biological processes behind the biological correlations can be explained, good data must first be extracted. The pooled biological quality control samples, deemed QC samples [126] as explained in section 1.4, can help to identify features with high analytical variation. These QC samples are used to measure the analytical variation and show the variation in a repeated measurement. Poorly performing features can then be excluded using coefficient of variation (CV) above a certain value, commonly 30%. Research teams can then use the resulting data to form networks with the correlations. These correlations networks are increasing in popularity within the metabolomics field [89] and the wider omic field [127]. Many network based analyses have helped explain the biological experimental hypothesis [91, 128, 129] and many newer network based algorithms have been published and used in to demonstrate the usefulness of networks [130, 131].

Finally, analytical correlations can be induced for example, by changes in the column chemistry for different functional groups over the run, delayed quenching between samples collection or changing extraction efficiencies. To our knowledge these correlation have been largely ignored, with the exception of Karakach *et al* [132] that showed that these correlations or covariances affected PCA analysis for NMR data. These analytical correlations can cause many problems when trying to understand the data. First, the analytical correlation could be mistaken as a biological correlation, leading to erroneous results. PCA is a very popular multivariate analysis technique used in metabolomics. As previously described the technique uses a covariance matrix. While not a statistical assumption there is generally an assumption with PCA analysis that the covariance of the features has been caused from biological variation. Again, the analytical correlations could cause erroneous results.

Here we demonstrate that correlation can significantly improve the identification of structurally related signals over simple small retention time windows. We present global correlation distributions for the biological & QC samples, showing that these distributions are relatively stable between datasets and LC methods. While it is very difficult to separate the structural and analytical correlations from the biological one, as each one is a function of the other, we present a method for removing correlations that come from the analytical process and show that the resulting network is statistically different than a network that has had nodes randomly deleted. This final network is a highly improved correlation network with greater confidence that each edge has a non-analytical process behind it [1]. It is believed that this method will have a wide range of uses where ever correlations or covariances are used in analytical data.

[1] This is most likely to be a biological process as all other known processes have been accounted for.

4.2 Methods

4.2.1 Sample Preparation & Data Acquisition

Two different datasets were obtained. The first was a urine dataset collected from a in-vitro fertilisation (IVF) pregnancy study, dataset A. The study had two biologically relevant groups: patients that remained pregnant and those that did not. The second is a tissue dataset from artery plaques in humans, dataset B. Below is a brief description of the methods used herein, the full details of the analytical methods will be published at a later date. The objective of the chapter is show the use of the data in the development of the informatics methods.

Dataset A

Urine samples were collected from 14 different patients who were on IVF treatment. The samples were collected each day over the course of the first month of IVF treatment. The sample collection was a spot collection. The urine samples were treated with sodium azide (at 0.1 g per 100 mL of urine) and filtered at 0.2 μm to remove particulates and cells. Dataset A data was collected on a LCT Premier TOF mass spectrometer (Waters MicroMass, U.K.) with electrospray ionisation (ESI) source. Using a C18 Acquity (2.1 mm x100 mm, 1.7 μm) columns (Waters) the chromatography gradient was over 14 min, with mobile phases A (water, 0.1% formic acid) and B (acetonitrile,0.1% formic acid). The gradient is summarised as follows: 0 min 100% A, 0% B; 4min 80%A, 20%B; 9min 0%A,100%B; 11min 0%A, 100% B; 12min 100%A, 0%B; 14min 100%A, 0%B.The flow rate was 0.5 mL/min, and an injection volume of 5μL was used. The mass spectrometer was set to scan from 85 to 1000 in mass to charge ratio (*m/z*), calibrated with sodium formate. Sample temperature was maintained at 4 $^\circ C$ and column temperature at 40 $^\circ C$. Quality control (QC) samples was used for the UPLC-MS analysis. Briefly, a pooled sample of the reconstituted extracts was prepared. This sample was re-injected 10 times before initiating the run to condition the column. Then the sample was re-injected once at the beginning, every

10 injections of samples, and at the end of the run[2].

Dataset B

Parts of plaque tissue, and intact tissue adjacent to plaque, were dissected and forwarded to metabolite extraction. Pre-chilled methanol:water solution (1:1) (methanol HPLC gradient grade, Fisher; water LC-MS grade, Fluka), was added to the tissue samples. Samples were centrifuged (Eppendorf) at 13,000 rcf for 20 min, at $40\,^{\circ}C$, followed by aliquoting at $250\mu L$ of supernatant into Eppendorf tubes. Samples were then spun to concentrate the samples. Reconstitution of the samples was done in $200\mu L$ of solvent mixture of H2O/acetonitrile (5:95) (LC-MS grade, Fisher Scientific, USA and centrifugation for 20min at 13000 rcf and $4\,^{\circ}C$. An Acquity UPLC BEH HILIC 2.1x100mm, $1.8\mu m$, column (Waters, Milford, USA) was used. Injection volume of $10\mu L$ was used for positive and negative ionisation modes. Mobile phase A consisted of acetonitrile/water (95:5) and mobile phase B acetonitrile / water (50:50). The chromatographic gradient was 23 min summarised as follows: 0 min 99%A, 1%B; 9min 1%A, 99%B; 23 min 1%A, 99%B. Detection of eluting UPLC fractions was achieved using a Premier Q-TOF (Waters Corp, USA). Data was collected with a mass range of 50-1500 *m/z*. A set of QC samples were made up in a similar fashion as in the method for dataset A[3].

4.2.2 Peak Detection and Alignment

The data was converted into netCDF using Waters MicroMass databridge$^{\text{TM}}$software. The files were then sorted into their different classes, with QC samples having their own class. For dataset A this was P (pregnant), NP (non-pregnant) and the QC samples. Peak detection and alignment was performed by XCMS (version 1.30.1). The 'centWave' peak detection algorithm was used for high mass accuracy, using peak widths of 3-20 sec, lock mass correction was used [108].

[2] We would like to thank Elizabeth Want and Peter Dixon for data and sample collection

[3] Thanks is given to Panagiotis Vorkas who collected and developed the analytical methods for the data

Grouping was performed with the 'nearest' algorithm, and alignment used the following parameters, missing=2, extra=3. The 'nearest' grouping algorithm was used due to many closely eluting peaks between samples and a better performance was observed with this algorithm over the default density algorithm. Both datasets were normalised with a median fold change normalisation [74]. A minFrac of 50% was imposed on the data, which removes any peaks that are not detected in < 50% of samples for any one class, including the QC class.

A CV of cutoff of 27% was used for dataset A and a 26% CV filter was used for dataset B. This corresponded to the median CV in both datasets. The median CV is a useful measure as previously stated, and this allows for removal of 50% of the data. These CV's were calculated from the across the QC samples, any feature that had a CV above this value was subsequently removed. The resulting dataset A had 2321 features across 9 samples. The number of samples was reduced due to the collection of data being outside of the specified 1st month of pregnancy. Dataset B had 1044 features across 146 samples. Again a CV filter was used on the QC samples to return only reproducible data.

4.2.3 Manual Detection of Adducts & Isotopes

To be able to assess the ability of correlation to find adducts and isotopes a list of known compounds in urine, as shown in table 7.7 and previously identified compounds in dataset B shown in table 7.8 were used to search the data matrix. An in-house R script was made to automatically find adducts and isotopes from named compounds within the data matrix by using retention time windows and explained *m/z* differences. This script was described in section 3.2.3. Briefly, the script takes a list of *m/z*s from a targeted list and calculates all possible adducts, isotopes and fragments using a list of possible adduct forms. The output is a list of any feature that matches the adduct, isotopes or fragments that is within the RT and *m/z* range specified. For both datasets a range of 25ppm was used and a range of 10s was used for the RT window. A list of the calculated adducts can be seen in table 7.1

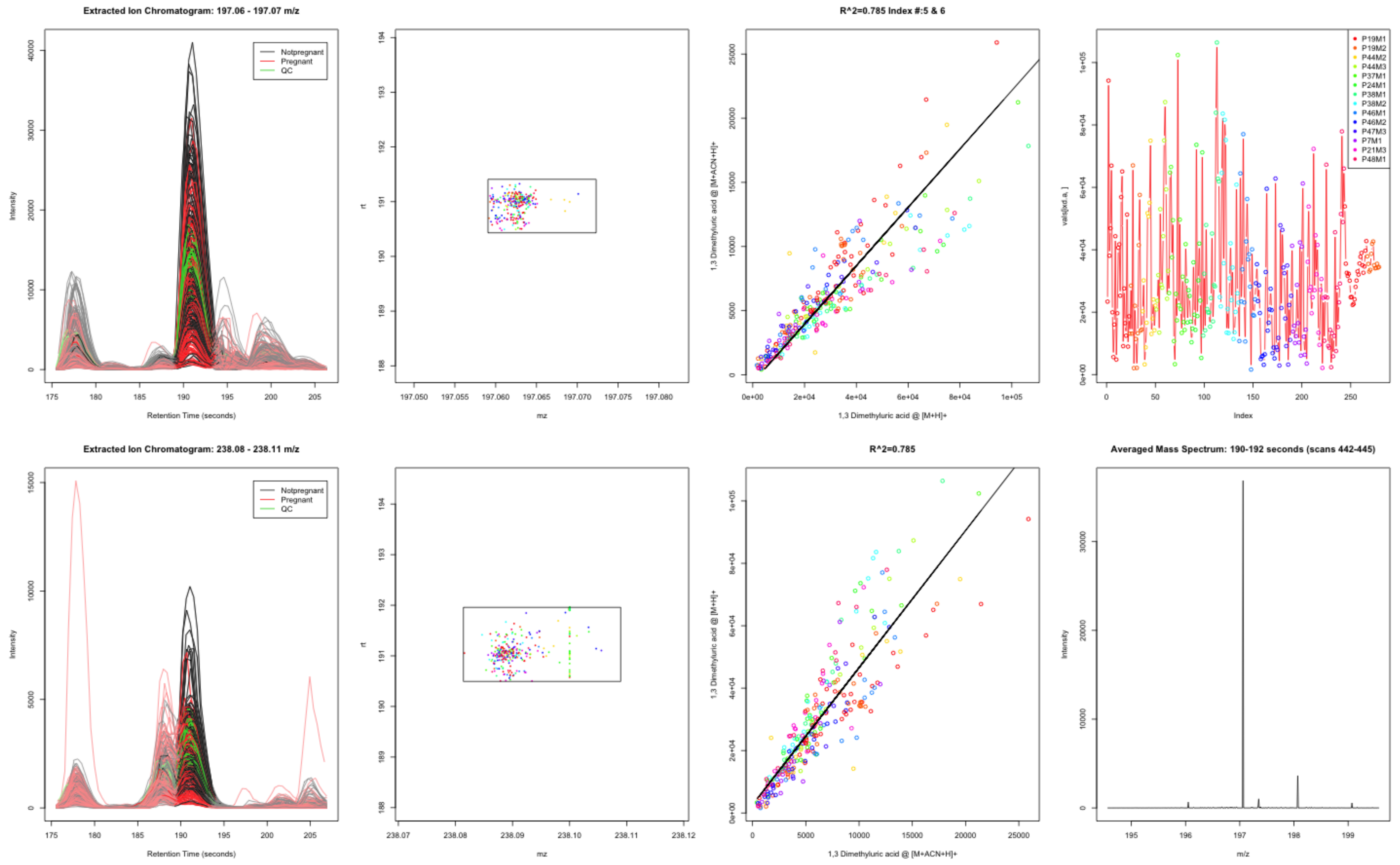
The results from the script were then verified by manual inspection. Conformation of isotopologues and adducts pairs were done by manual inspection of linked EICs. Figure 4.2 shows an example plot. These plots were generated for possible structural pair. These complex plots were directly used for the confirmation of each pair and contained the following:

- Extracted Ion Chromatographs of each feature in the pair - The EIC's are very informative as they show if the peak is real. The criteria is being roughly Gaussian and intensity above local noise. While all peaks should be above a signal to noise of 5, some peaks were close to or difficult to discriminate from the local noise by eye. In these case the peak was excluded

- A feature definition plot - These indicate where each peak from each sample was detected in terms of *m/z* and RT. These are very useful for defining if there are extra peaks not belonging to the feature are present. For example if a sample had two peaks within the feature box.

- A Mass Spectrum - This plot helped to understand where the peak was coming from and if it was actually an isotopologue or if it was the molecular ion. Although there is no precise method for identification of the isotopologues, the overall pattern needs to be an isotopic pattern that is normally seen with metabolite. The expected pattern is a majority of ^{12}C and then a smaller peak with 1 ^{13}C etc. It is also helpful to check that the isotope is +1 *m/z*, unless the identified feature is believed to be multiply charged.

- Integrated Intensity plots - these plots show the integrated intensity across all samples. This can indicate if there was a falling intensity across the run. If there is a falling intensity normalisation can normally help.



(a) Good Pair

Figure 4.2: Manual verification of putative structural pairs in dataset A. From left to right in the first column are extracted ion chromatograms of the selected feature. In black is the non pregnant, red pregnant and green the QC samples. The next column shows the feature definition plots. These plots indicate where the peak was detected in terms of *m/z* and rt for each sample using the same colour scheme. The 3rd column is the Pearson correlation of the two features. Each colour represents different patients in the study. The 4th column top box is an intensity plot of the top feature across all samples. This can indicate if there was a falling intensity across the run. Lastly the bottom right box shows a mass spectra of the integrated retention time areas. This figure represents data that would pass the checks.

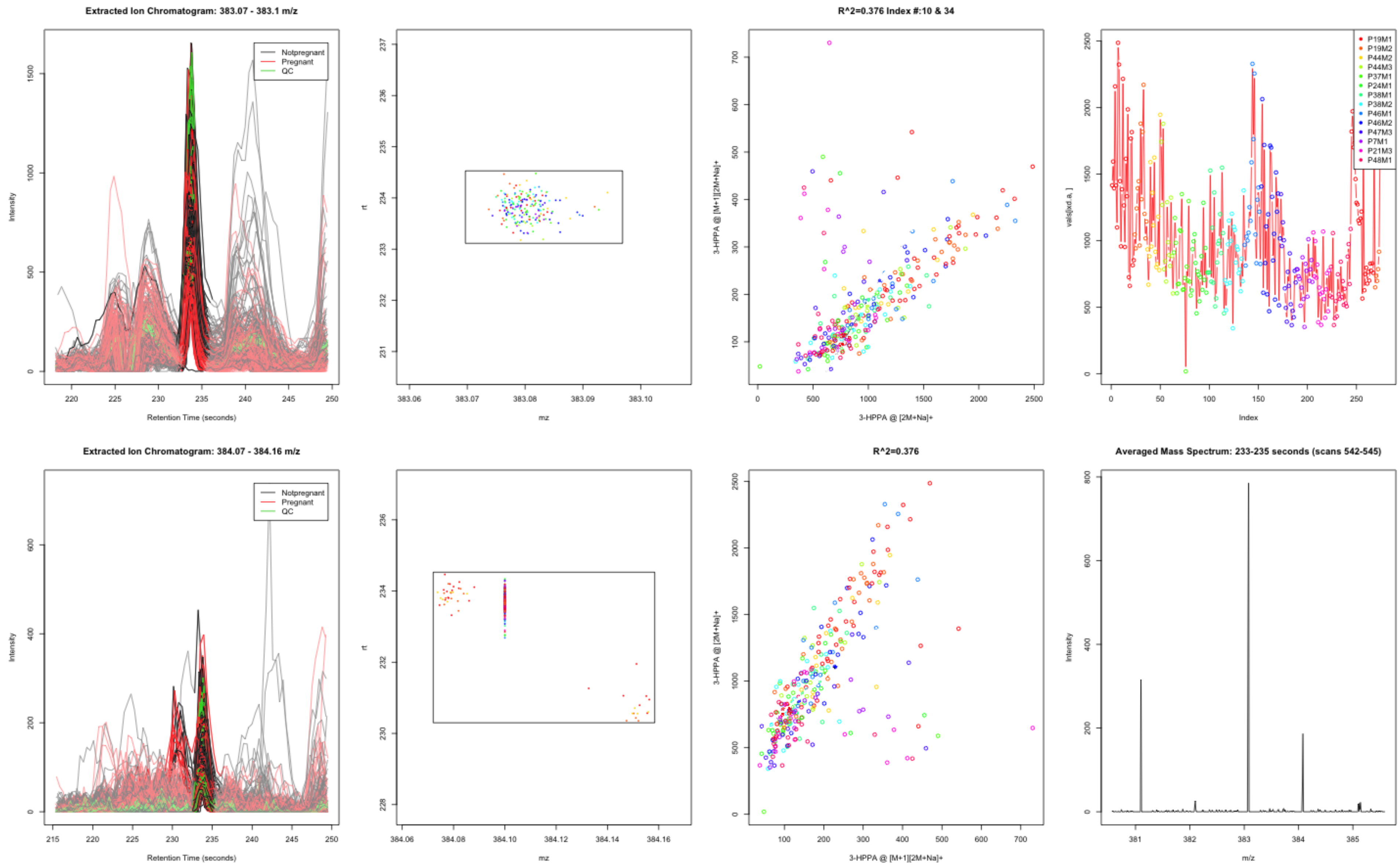


Figure 4.3: Manual verification of putative structural pairs in dataset A. Figure is same as figure 4.2 however, this pair was rejected due. The rejected pair was due to the definition plot having poor grouping of peaks with the feature box.

Two sets of the example figures have been shown; one for an accepted structural pair 4.2a and one for a rejected structural pair 4.3. After all pairs had been verified a list of accepted structural pairs was made. This set was a set of true structural pairs. A set of false pairs was any feature that did not have an overlapping RT and therefore could not be structural pairs. A 5 second RT window was used for the exclusion of structural pairs. Using ρ as a threshold to declare pairs as structural (positive) or non-structural (negatives). A Receiver Operator Characteristic curve (ROC) was calculated by varying the ρ threshold. Plotting of the (ROC) was done using the pCor package. The following definitions were used for the confusion matrix:

- True Positives (TP) - These were confirmed structural pairs via the structural pair plots that had a $\rho >$ the chosen threshold.
- False Positives (FP) - These were confirmed non-structural pairs that had a higher ρ than the chosen threshold.
- True Negatives (TN) - Confirmed structural pairs that had a $\rho <$ the chosen threshold.
- False Negatives (FN) - Confirmed non-structural pairs that had a $\rho <$ the chosen threshold

Using these pairs a calculation of sensitivity and specificity was done using the Pearson correlation of all feature pairs to produce the ROC curve.

4.2.4 Statistical Correlation Dependency Network

For both datasets the samples were combined into two categories, biological samples and QC samples. The QC samples should have no detectable biological process to influence the correlation and consequently need to be in a separate category.

The intensities of the features in each category were then correlated using Pearson correlation.

The resulting correlation matrix was then used for network construction of the 'biological network' using the biological samples and the QC sample for the 'QC network'. Each network was made using the igraph [133] package, any edge having a $\rho \geq +0.75$ or $\rho \leq -0.75$. A value of ±0.75 was chosen as weak spurious correlations would be removed and only stronger correlations would remain. However, as this is a proof of principal the final chosen value is not essential. The networks were made for both datasets, independently. In the networks the nodes are the features and the edges or connections are the correlations between the feature pairs.

Filtered Network

An improved 'filtered network' was made from the biological and QC networks. The QC network, which has analytical and structural correlations can be used to remove edges which are statistically similar in their ρ to the biological sample network. These edges should be analytical edges however, depending upon the dataset the edges removed could also be structural.

To do this the QC network was bootstrapped, a method of resampling with replacement. The bootstrapping was performed by bootstrapping the QC samples 5000 times and remaking the correlation network for each bootstrap. The bootstrapping gave a distribution of ρ for each pair of features. From this distribution a mean, rho_{qcBoot} and standard deviation, σ_{qcBoot} was used in the Z transformation, equation 4.1. The ρ_{bio} was used to transform in the Z domain. The Z transformation allows for a higher resolution p-value. The p-value is the probability the correlation is, or more extreme than ρ_{bio} could be observed in the QC network. Therefore, any edge with a low $Z_{filtered}$ value i.e. high p-value was removed since correlation may have resulted from analytical procedure alone. The chosen threshold p-value was > 0.05 post Bonferroni correction. This made the new network, the 'filtered network'.

$$\frac{\rho_{bio} - \overline{\rho_{qcBoot}}}{\sigma_{qcBoot}} = Z_{filtered} \tag{4.1}$$

Testing the Filtered Network

The new 'filtered network' needs to be tested that it was statistically different to a network that can be generated by randomly deleting edges from the biological network. This was performed by randomly deleting edges from the biological network. This process was repeated 1000 times deleting the 578 edges each time. This number corresponded to the difference between the biological and filtered network edge number. For each iteration, the shortest average path length, betweenness centrality and degree statistic were recorded. Both the degree and betweenness centrality statistic were also recorded as a rank statistic. This was done by calculating the rank difference between the original biological network and the iterated random deletion network.

Another way to test the new filtered network is to calculated how different it is from the biological network. While simple statistics on the biological network give an answer it is unknown what error is associated with these measurements. Consequently, a method is needed to generate a network that is no different from the biological network except by permutation of samples. This network is a null network. This null network describes the error boundaries of the original biological network. It is important to know how different the null network is to the filtered network. If the filtered network is within the boundaries of the null network then it could be no different than a permutation of the biological network. A null model network was defined using equation 4.2. The null model is made in a similar fashion as the filtered network.

$$\frac{\rho_{bio} - \overline{\rho_{bioBoot}}}{\sigma_{bioBoot}} = Z_{null} \tag{4.2}$$

Using equation 4.2 the biological samples are bootstrapped and for each bootstrap a new correlation matrix is made. The distribution of each pair of features is then used to find the mean ρ, $\overline{\rho_{bioBoot}}$ and the standard deviation $\sigma_{bioBoot}$. These are then used in a transformation to the Z domain by subtracting the ρ_{bio} for that particular pair of features. Again, these Z values were converted to p-values where, after Bonferroni correction any p-value that was ≤ 0.05 had the corresponding feature pair/edge removed. Using the same cutoff ≤ -0.75 & $\geq +0.75$ a network was made in a similar fashion as before. 5000 unique bootstraps were performed calculating a new network each time. The rank betweenness was also performed on the new null network, again ranking against the biological sample network.

In total there are now 5 different networks they are as follows:

- Biological Sample Network - This is the original network from the correlation matrix of the biological sample category
- QC Sample Network - This is the network from the correlation matrix of the QC samples category.
- Filtered Network - This is the network that uses the bootstrapped QC samples to remove analytical correlations. This network was made using equation 4.1
- Random Edge Delete Network - This is the network that was derived from the original Biological Sample Network but has had edges deleted at random.
- Null network - This is the null model network to examine the variability in the biological network. This network was made using equation 4.2

4.2.5 Network Visualisation

Using VANTED [134] the network was first visualised using the grid layout. In this mode singlelet nodes were deleted. The layout was then reselected to circular and finally to a force directed layout. This allowed the same nodes to be in similar locations between the original network and the filtered network.

4.3 Results and Discussion

Both datasets showed that correlation is a useful technique to distinguish structurally related compounds within a defined retention time (RT) window. Using a combination of automatic analysis using a previously published script [68] and manual inspection to identify adducts, isotopes and fragments, 65 structurally related features with 72 pairs, were detected for dataset A. In the B dataset, 32 structurally related ions were detected with 36 pairs. After manual conformation, based on peak shape, peaks that had overlapping RT windows, mass spectra plots and feature definition plots, as demonstrated in figure 4.2, 25 adduct and 8 isotope pairs were identified for the dataset A. Dataset B resulted in 11 adduct and 7 isotope pairs. Dimers were also detected during this search with 3 monomer-dimer pairs in dataset A and only 1 pair in dataset B.

4.3.1 Adducts & Isotopes ROC

Figure 4.4 shows performance statistics for detecting structurally related pairs using a correlation threshold and retention time windows. As previously discussed in section 4.2.3, a confusion matrix was setup for the adducts and isotopes. The positives examples were adduct pairs that were deemed to have to correct mass difference and were within the retention time window that was specified. These positives were manually confirmed via confirmation plots. Negatives were defined as any non-structural pair outside of the RT window. The graphs show that the false positive (FP) rate is very low at high cutoffs or correlation coefficients (1.5% at $\rho = 0.75$ Dataset A, adducts). The isotopes have a similar false positive value for the same cutoff as the adducts, 6% at $\rho = 0.75$. Dataset B had higher FP rates than dataset A, at 5.5% ($\rho = 0.75$) for adducts and 6.7% ($\rho = 0.75$) for isotopes. However, this value is affected by the low resolution of the curve, due to a lower number of confirmed structural pairs. Yet the positive predictive value (PPV) shows that dataset B is more accurate in predicting both isotopes and adducts. PPV shows the proportion of pairs predicted to be structural that are correctly identified at the corresponding cutoff.

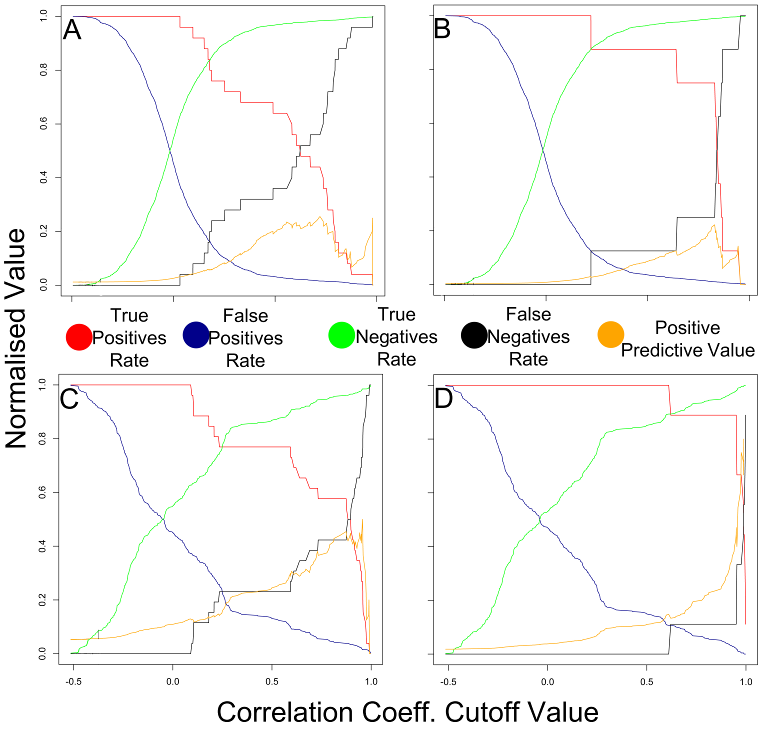


Figure 4.4: Performance statistic for the ability of a single correlation threshold to identify structurally related ions. A - adducts & B -isotopes dataset A. C adducts & D - isotopes dataset B.

The PPV is a maximum at the following cutoffs 25% PPV at $\rho = 0.72$ in dataset A and 50% at $\rho = 0.95$ for dataset B with the adducts and 25% at $\rho = 0.82$ in dataset A and 80% at 0.99 ρ in dataset B for the isotopes. The PPV value shows that a single threshold is more predictive for isotopologues pairs than adduct pairs, which is to be expected given their higher correlation coefficients. The PPV is also heavily influenced by negatives. This is particularly important when trying to compare between datasets. The PPV does not have the same prevalence between datasets, that is to say the number of confirmed structural & non structural pairs is different between the different datasets. This therefore, arbitrarily increases the PPV value for dataset B.

The ROC curves from figure 4.5 demonstrate that the using Pearson correlation of intensity values with retention time windows is an effective way to identify structurally related pairs. The area under the curves (AUC) are above 89.8%. The isotopes show the same trend as the false

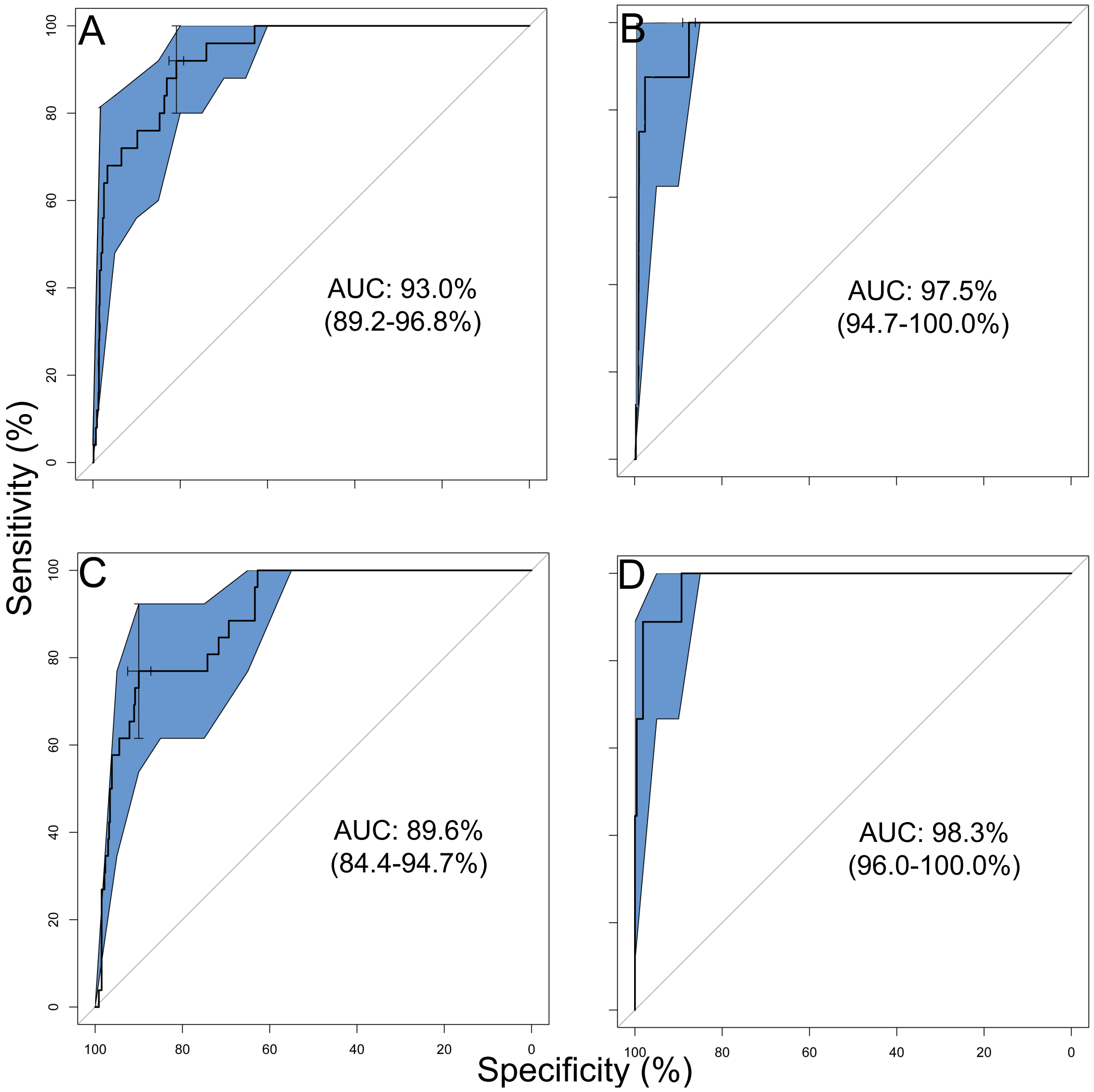


Figure 4.5: Receiver operator characteristic (ROC) curves for the performance of a simple correlation cut off in identifying structurally related pairs of features. These pairs were manually identified adducts and isotopes in both datasets. Graph A is the ROC curve for adducts of dataset A. graph B is the isotopes for dataset A, graph C are adducts for dataset B and graph D is the isotopes for dataset B. The isotopes ROC curves have a much higher area under the curve (AUC) than the adducts in both datasets. The error bars shown in blue are a 95% confidence interval calculated via 2000 bootstraps.

positive rates (97.5% (*C.I.* 94.7 – 100.0%) dataset A & 98.3% (*C.I.* 96.0 – 100.0%) dataset B) having a higher AUC than the adducts (93.0% (*C.I.* 89.2–96.8%) dataset A & 89.6% (*C.I.* 84.4–94.7%) dataset B). This increase in variation of correlation and reduced correlation coefficient may be due to the small changes in source conditions across samples that effect the ionisation of the adducts. These results are inline with what has been found using the scoring system published in Kuhl *et al* where a 84% AUC was found. However, Kuhl's [51,135] uses a complex scoring system that uses a binary score for the presence of isotope peaks, a WFC and a AFC. This means that the $84\% AUC$ is not directly comparable to the AUC described here. Further more the method by which the false positives are defined is also different. The paper defines the false positives by artificially moving randomly selected peaks to co-elute with the known molecular ion. This method is needed due to the use of the WFC being included in the score. With the complex scoring system and confirmation between the two structural pair correlation methods, WFC and AFC it would be expected that the AUC would be better. The number of FPs is probably higher than it should be due to more WFC causing a higher ρ. This would decrease the specificity and consequently the AUC.

4.3.2 Correlation in Relationship to Other Parameters

Using the known compounds in dataset A the adducts and isotopes were checked to see if there was any relationship with other chemical parameters. The tested parameters were: logP and Polarity. The frequency distributions plots of Pearson correlation, against these parameters showed no identifiable relationship with either of these parameters. It could be due to the low amount of identified compounds and a wide numerical range for both logP and Polarity both of which were referenced values. It would be expected that with specialised source setting that a relationship could be formed between these parameters and the number of adducts however, given that the source settings are generalised for all expected ions this is much less likely.

4.3.3 Non Structurally Related Correlation Pairs

Many of the high correlations are between non-structurally related pairs. Pairs that lie outside of the retention time windows and still have a high correlation are most likely not structural pairs. These pairs are most likely biological correlations or analytical correlations as described earlier. Seen in equation 4.3 there are substantially more non-structural pairs than structural.

$$No.\ of\ non - structural\ pairs = \frac{2321^2}{2} - (2321 * 3) = 2.7x10^6 \tag{4.3}$$

If it is assumed that each of the 2321 features (dataset A) will have 3 structural pairs then the non structural pairs will be 2.7×10^6 and 6963 structural pairs. A distribution of correlations for both sets of sample categories, biological samples and QC samples can be seen in figure 4.6. This distribution was again made by a pairwise correlation of intensity of the features. Pairs that were within 5 seconds of each other were removed. This was done to reduce or remove the structural correlations. Consequently, the distribution seen in 4.6 should only be a distribution of biological and analytical correlations.

As previously mentioned the biological samples have biological information. Whereas, the QC samples are the pooled biological samples, these samples give information about the repeatability of the method and instrumentation. Consequently, the distributions for the two sample sets should be different. As expected the mean (μ) of the distributions are close to 0 for both datasets (biological 0.01 dataset A & 0.01 dataset B QC samples 0.03 dataset A & 0.15 dataset B). However, the standard deviations (σ) are much larger for the QC samples in dataset B (0.19 Biological, 0.38 QC samples). This is suggestive that the QC distribution is closer to a random distribution and therefore many of the correlations are spurious/random correlation. Dataset A has little difference between the standard deviation of the classes (0.20 Biological & 0.28 QC samples). The biological samples distribution is a tighter distribution due to the biological variation which, on a mean level comparing CVs, is know to be quite large [62, 118]. The larger

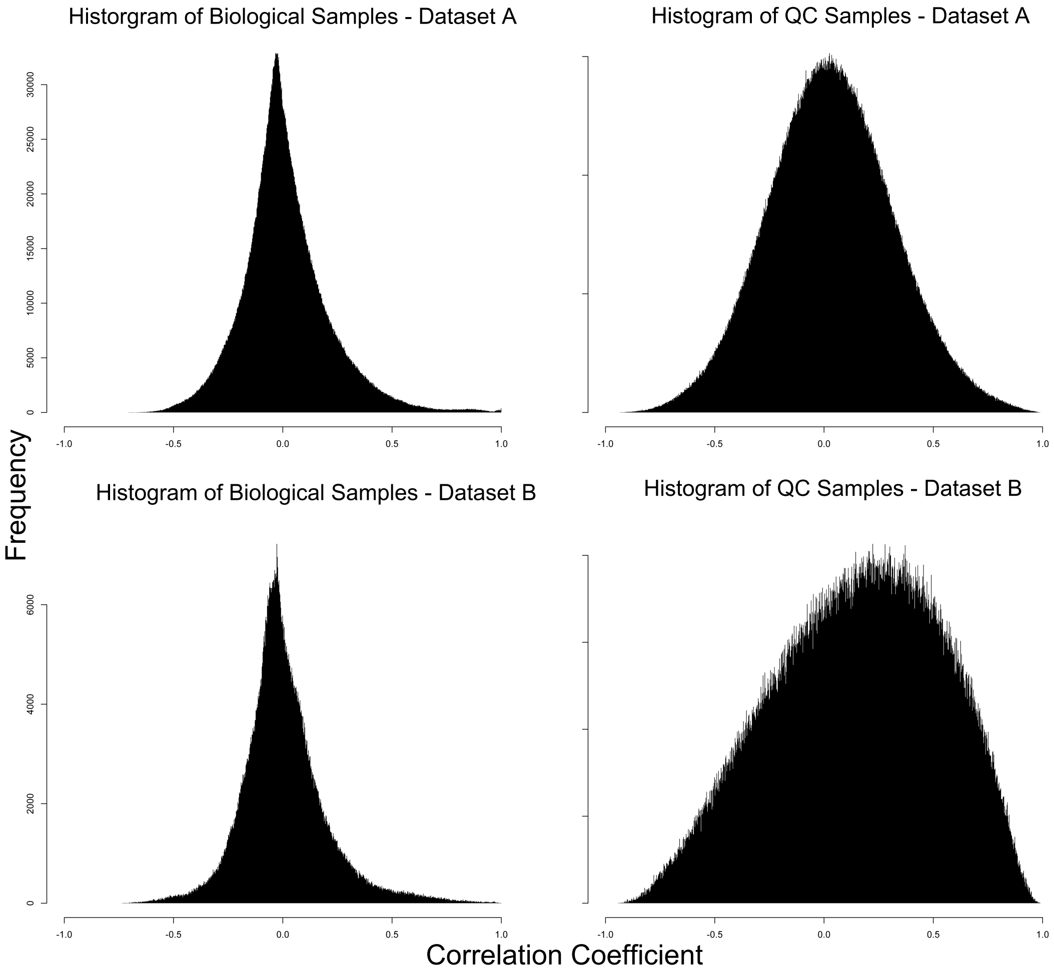


Figure 4.6: Non-structural correlation distributions from the Biological and QC samples. The diagonal of the correlation matrix was removed along with a 5second window around the diagonal. This removal had the effect of removing structural pair correlations. The distributions for the biological, analytical and structural distributions in the biological samples and the analytical and structural distributions for dataset A on top and dataset B below.

variation means that correlations are not due to random correlations as seen in the QC samples that should have a small variation. The larger variation of the biological samples cancels out the smaller variation of the analytical process known to be much smaller [68, 112]. Any real correlation from the QC samples cannot be biological correlation as the QC samples are a uniform mixture of the various biological samples thereby removing any biological effect. The strong correlation in figure 4.6 are not structural as they have been removed. These real correlations must therefore be analytical correlation.

4.3.4 Effects of Analytical Correlations on Inferred Statistical Dependency Networks

Commonly, inference networks are made to help in the analysis and understanding of omics data [89, 136]. These networks could be highly influenced by correlation induced purely by analytical process and thus not deriving from true biological relationships. Additionally the analytical process could reduce or eliminate biological correlations in some cases. We therefore aimed to investigate this by removing or reducing correlations that are from an analytical correlation origin. For this set of analysis only dataset A was used.

Using the integrated intensities of the features in the biological samples, pairwise correlation was used to make a correlation matrix. From this correlation matrix any correlation over $\rho \geq 0.75$ & $\rho \leq -0.75$ was used to make a feature correlation network. The structural correlations were expected to be revealed as tightly grouped and highly connected sub-networks where as the analytical and biological correlations structures were harder to define and identify. Ideally, the edges would be just biological correlations. It is not straightforward to separate the distributions into their component parts. Each distribution is a function of the other. Structural correlations are fairly easy to classify and most analytical correlations should be weak correlations. This can be tested and any correlation that is being driven by the analytical correlation can be removed, using the QC samples.

The QC samples were bootstrapped to form a ρ distribution for each pair of features. Using this distribution a z test could be performed using the same pair in the biological samples. The z values could then be transferred to p-values. If the p-value was not significant (> 0.05), that is to say that the ρ for the biological samples pair was within the correlation coefficient distribution of the QC samples, the edge was deleted in the filtered network. This filtered network is visually similar to the unfiltered network, figure 4.7. Both graphs were assembled in identical ways, thereby allowing a visual comparison between the locations of the different sub-graphs.

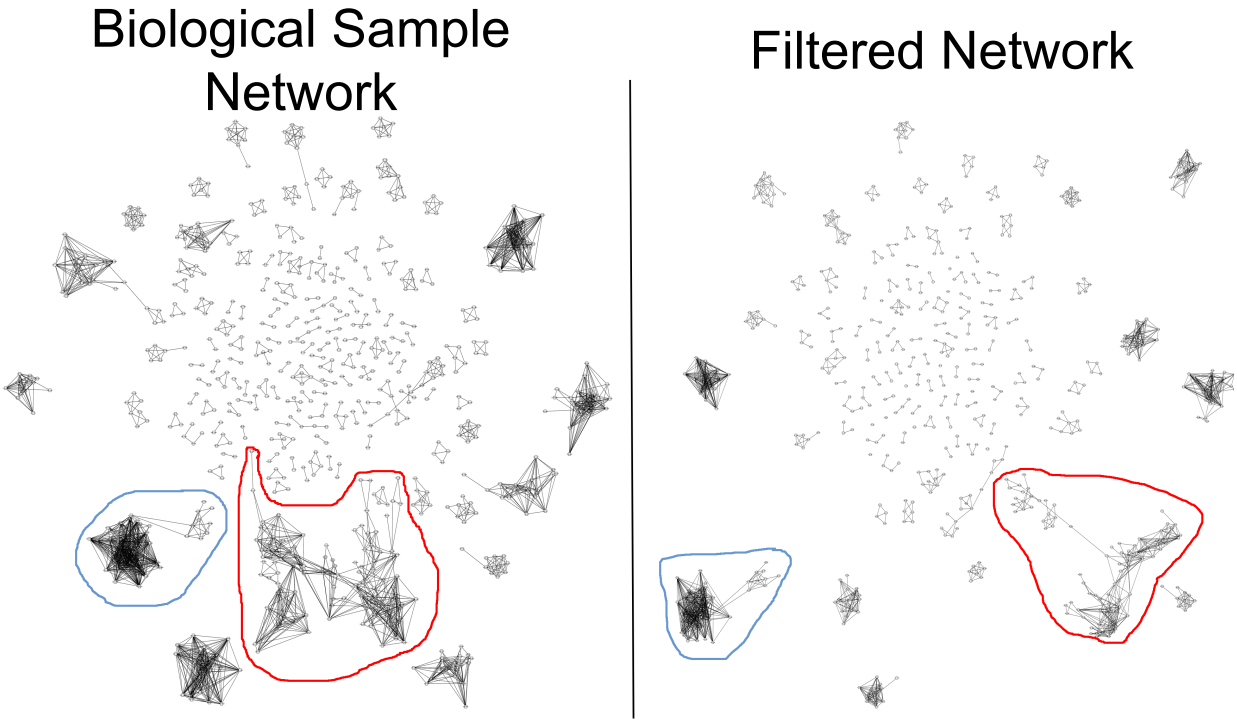


Figure 4.7: A force directed layout of both the original biological sample network and the filtered network. Many visual similarities can be seen, while some sub-graphs have been reduced in the number of edges and nodes. The circled sub-networks correspond to the same group of nodes in both networks. A visual difference can be seen between the circled sub-networks between the original biological sample network and the filtered network.

There are many highly connected sub-graphs in both networks however, the filtered network (7973 edges) has fewer connections than the biological sample network (8551 edges). This is also shown by the drop in the average degree between the two networks dropping from 7.37 in the biological sample network to 6.87 in the filtered network. Due to drop in the number of edges there is also a small rise in the number of singlet nodes in the filtered network (921 nodes in the biological sample network to 949 nodes in the filtered network).

To assess whether the filtered network was statistically different to deleting random edges from the biological sample network a series of randomly deleted edge networks were created. In each iteration, a randomly selected list of edges (578 edges each time), from the biological sample network were deleted. Performed many times (1000), network statistics could be used to identify if the filtered network was different to the mean random edge deletion network. Figure 4.8 shows the distribution of average path length from the deleted network. The average path length for the original network is 1.564 shown as the red line. The average path length for the filtered network is less, at 1.552 the blue line in figure 4.8. It can be seen from the graph that the filtered networks, average path length is different from the median 1.60 (σ = 0.02) of the random deletes average path length. The edges that are targeted in the filtered network are the analytical and the structural edges. The structural links are normally highly connected subgraphs and few analytical connects should be connecting between subgroups. This means that the overall network structure should not change too much. A good example of this has been highlighted in blue in figure 4.7. The decrease in the average path length between the original biological network and the filtered network is due to the unconnected nature of the network. If there are few connections between two subnetworks, seen highlighted in red of figure 4.7 (Biological network), and the edges which connect the two subnetworks are broken to form smaller networks then the average path length can decrease. There is a second smaller distribution in figure 4.8 which lies close to the both the filtered and biological sample network. This is not unexpected as the method randomly deletes edges and therefore could come very close to a network that is very similar to the filtered network. However, these similar randomly deleted edge networks do not happen the majority of the time. The degree distribution was also evaluated for the random edge deletion networks. This did not show any significant difference to the filtered network.

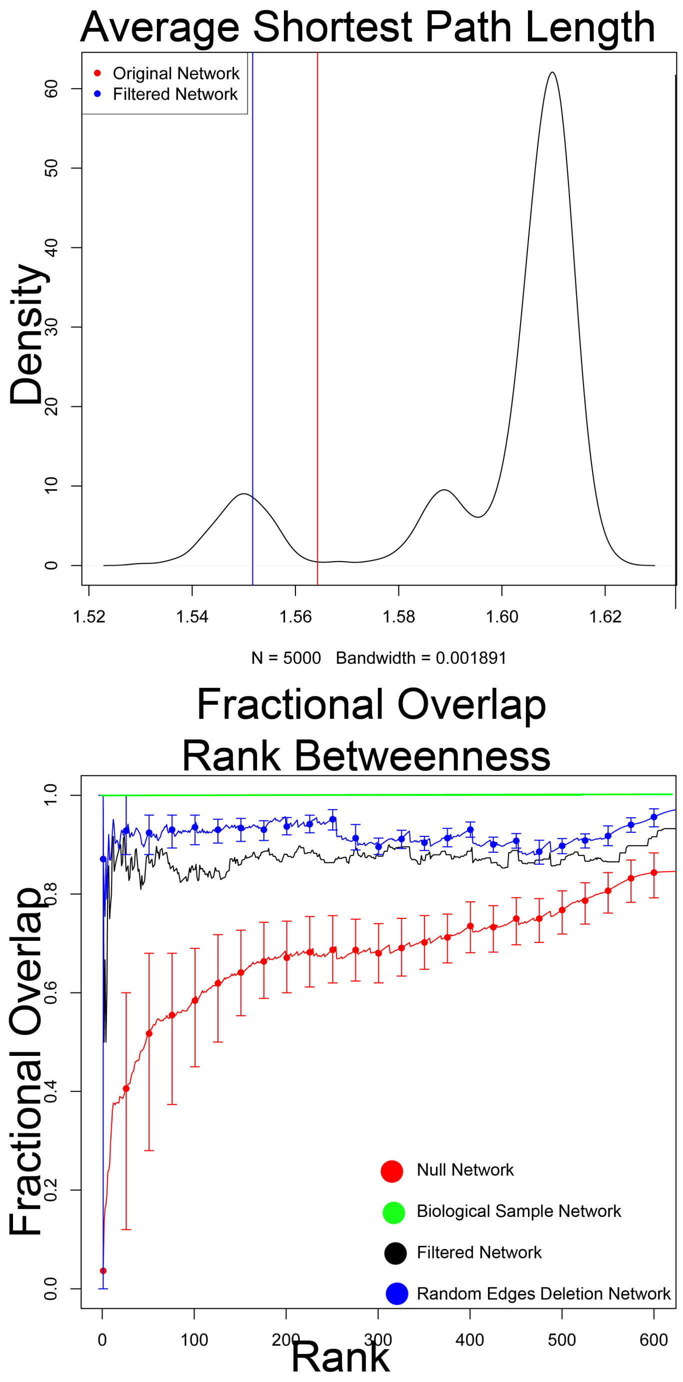
Betweenness is a commonly used network statistic and is a measure of how many paths between pairs of nodes, go through a given node. Figure 4.8 For each network the betweenness values were ranked. Then the overlap of the top n ranked nodes were calculated for $n = 1$ to 2321 (all of the nodes) e.g. if half of the top nodes were the same for the networks, then the overlap is 50%. The green line is the fractional overlap of the biological sample network ranked against itself. Therefore, it is always at 100%. The black line is the fractional overlap of the rank betweenness of the filtered network. It can be seen that the first 10 ranks there is an 80% rank overlap to the biological sample network. The blue line is the random edge deletion network. For the first 10 ranks there is a 89% overlap. Finally, the red line is the fractional overlap of rank betweenness for the null network. The null network was made in a similar fashion as the filtered network. However, where the filtered network used the QC samples category for bootstrapping the null model uses the biological samples for the bootstrapping. This network makes sure that the filtered network is not just a rearrangement of the biological sample network. This means that if the filtered network is very close to the null network, then the filtered network is just removing spurious correlations. The null network has a maximum overlap of 60% to the filtered network for the first 10 ranks. The null network rank betweenness is very different than the filtered network and stays outside of the 95% error bars for more than the first 600 ranks. The rank betweenness shows that the filtered network is significantly different than the biological sample network, the null network and the random edge deletion networks. When the edges are deleted in the random edge deletion network the betweenness stays close to filtered network. However, they are different and the random edge deletion network is closer to the biological sample network than the filtered network for the first 200 ranks. This shows that the filtered network does not resemble random edge deletion. Rather the filtered network has targeted edges to delete.This changes both the average path length and the betweenness score.

A side product of the filtering of the network for analytical correlation is that structural correlation may also be removed. However, if the experiment has been designed and carried out well then the variation in the QC samples will be less. There will be fewer analytical correlations and due to the lower variation the structural correlations will be weaker.

The filtering method reduces the analytical correlations which could lead to FP links. Those FP links could be interpreted as significant biological links. Multiple testing correction was also performed on the p-values of the filtered network. This lowers the p-value cut-off for significance, which in turn causes more edges to be removed. Consequently, the edges that remain will almost certainly be due to a biological correlation. As network based analysis has become more prominent in the field the risk of trying to identify a pair of metabolites, a lengthy process, of a non-biological relevant link would be a time consuming process for no reward and could lead to a false hypothesis.

4.4 Conclusion

In conclusion, there are many different types of correlations in liquid chromatography mass spectrometry data. The correlations that are structural have been explored by publications such as Kuhl and Alonso *et al.* We have further defined how well correlation alone can classify structurally related pairs. We examined ROC curves for adducts and isotopes independently which has not been previously investigated. It is also clear that a retention time window is needed for classification of structural pairs. Without retention time windows correlations just as strong that are either biological or analytical correlations may be found. By removing structural links from the global correlation distribution a set of correlations could be seen in both sets of sample categories, biological and QC samples. The QC sample correlation distribution should have only had



Average Shortest Path Length

Fractional Overlap
Rank Betweenness

Figure 4.8: Network statistics to show the difference between the unfiltered biological network and the filtered network. A) The overlap of the different network using betweenness as a measure and ranking score shows that the filtered network is very close to the random delete networks. Yet, the bootstrapping of the biological network is very different to the filtered network. B) The average path length was calculated for each random delete network. The filtered network (blue) is clearly substantially different to the main distribution of the random deletes, and different to unfiltered biological network (red)

structural correlation. However, the global correlations distributions are strikingly different from each other. This correlation distribution for the QC samples was made up of analytical correlation and the biological sample distribution was made up of biological and analytical correlations. To our knowledge no previous work has separated out these distributions or has studied the correlation at length. The analytical correlations are derived from analytical processes that have a linear effect, such as degradation of column chemistry, variable sample extraction efficiency etc. As inference and correlation networks become more popular in the field the risk of including analytical links in biological networks is prominent. We have developed a method to help remove or reduce the amount of analytical correlations and return a high confidence correlation network. We have shown that the network is statistically different to randomly deleting edges in the original biological correlation network. It is different on the level of connectivity between pairs of nodes for the original biological correlation network, random edge deletions and a null network that is an altered biological network.

This method presented here has many more far reaching effects than just correlation networks but also any process that relies of pairwise feature metric, such as PCA, MDS, nearest neighbour approaches and many more.

5 Using Temporal Correlations as Biomarker Signatures in Metabolite Profiling

5.1 Introduction

Time is an experimental parameter that is rarely analysed in great detail in modern high throughput experiments. However, temporal analysis can yield a better understanding of the dynamics of the system than simple ’snapshot data’. If used correctly with good quality data the change from one state to the next can be understood and even predicted.

Classically time series analysis has been preformed at a univariate level. These methods include auto-regressive (AR) models, moving average (MA) models, combination models such as autoregressive moving average (ARMA) models and autoregressive integrated moving average (ARIMA) models, as previously discussed in section 1.8.1. These methods give the ability to understand the dynamics of the data by using the all of the information available however, since they are univariate methods they do not consider variable relationship. In general, time series analysis methods have the potential to:

1. Predict future time points of the series being modelled

2. Give an understanding of the underlying dynamics (the progression of one state into the next)
3. Increase the statistical power to accurately classify samples

There is no universally agreed definition of what makes method a temporal method. However, the method should use the order of the time points to understand the dynamics of the data. This definition was used by Smilde *et al* [96]. There are many well received methods that use this definition such as the STEM method [137], TimeClust [138], NETGEM [101] and miniTuba [102]. Here, we define a temporal method as one which uses the order of the time points to model the data.

5.1.1 Metabolite Time Series

In many high throughput biological experiments there is an interest in finding the difference between two class states and identifying a statistically significant difference between the two. Commonly, this is done by evaluating the mean or median feature intensity levels between class states. Without a second dimension of time the understanding of how the class states interact or transform into each other is lost.

While traditional metabolite biochemistry experiments, which follow only one or two metabolites or a certain pathway commonly have a temporal aspect, this is still rare in high throughput metabolite profiling experiments. The cost of running multiple samples and the need for repeated sampling from the study individuals across time reduces reduces the frequency of these experiments. With little data being generated for these studies there is a lack of metabolomics specific methods for time series. Some methods such as PCA [56, 139, 140] have been used to show the temporal trajectories of each study individual. This method is very useful to show toxicology data, as was the case in the COMET study [139, 141, 142]. The study showed how mice and rats which were given doses of toxic chemicals moved away from their starting metabolic state and then corrected and returned to their previous state. However, this method does not give detailed information about the dynamics that are happening and it does not use the ordering of the temporal data in the model. Another way that temporal data is routinely analysed is to look at each time point separately and perform a univariate test such as a Wilcoxon rank t-test on the metabolites levels for each time point separately. Again, this method does not use the temporal ordering to gain more information about the system.

Existing Time Series Methods

Methods such as the smoothing splines mixed effects (SME) models [97] method have been specifically designed for temporal metabolite data and use the ordering of the temporal data to detect differences between experimental conditions (classes). This method makes a spline model of each class so that a median curve is modelled for each condition. Once these median curves are found a functional t-test can be used to find if the curves are significantly different from each other. The functional t-test integrates the area between the two curves. By modelling the individual deviations from the average curve the variance of the test statistic is estimated. Finally, a null model is estimated via bootstrapping. This method takes into account the temporal nature of the data by creating a model of the temporal profile. Some models uses aligned continuous curves of the two expression profiles. By doing so they are able to collapse the data down to simple chi-squared test. This method has been used to find differently expressed genes with few data points [143]. Other models also use the continuous curves of B-splines, such as the continuous hidden process model (CHPM) which is able to assign a biological function to a set of genes using previously known biological information [144].

Many methods look at the differences between the temporal states [101,145] using networks to describe these changes. Others look at the a genetic pathway approach again using networks for the description [92,146,147]. Correlation of the intensity or counts is an often used and powerful metric [89, 90, 137]. As previously stated one of the problems with time series data is that it can be difficult to fit the correct curve to the series. Often this is due to either short time series and or few biological replicates. A lot of biology tends not to be a linear process. However, nonlinear methods have a difficulty in being able to fit the model. Pearson correlation requires much less data to be able to fit the linear model than many nonlinear methods.

Developed Method

Often an interesting and useful way to look at temporal data is to use lags. Lags are used classically in AR models, to look at the cyclic nature and memory of time series among other values in the model. With multivariate time series we can compare different time series to each other using a selected metric. As previously stated Pearson correlation is a powerful metric and is quick to compute. To best of the authors knowledge there are no methods that use multivariate temporal lags in metabolite profiling.

Here we present two temporal methods for the discovery of temporal biomarkers between two or more sample classes. The first method uses metabolite correlation networks to identify similar processes within classes and differences between classes. The second method presented identifies differences in the rate of the processes between classes. For example if one class has a faster rate of metabolism than the other class, but the process stays the same, method 2 will identify this difference. These methods are demonstrated using two different temporal datasets, a pregnancy study (from chapter 4) and the COMET Hydrazine study [139, 148]. Using these datasets we are able to show that these methods produce statistically significant results that can help with the understanding of the temporal structures in these complex datasets.

5.2 Methods

5.2.1 LC-MS Pregnancy Study

The data was prepared in the same way as the previous chapter 4.2.1. Briefly, 14 women who were on the same in vitro fertilisation (IVF) trial gave daily urine samples. The 14 women were split into two classes of successful IVF treatment (P) and failed IVF treatment (NP). The samples were collected over the whole month. However, there were some missed days and not all of the women overlapped with the sample collection days of the rest of the group. After filtering the samples were analysed by an LCT Premier TOF mass spectrometer (Waters MicroMass, U.K.) with electrospray ionisation (ESI) source. A reversed-phase C18 chromatography method was run using an Acquity (2.1 mm x100 mm, 1.7 μm) columns (Waters) mobile phases A (water, 0.1% formic acid) and B (acetonitrile,0.1% formic acid). The method was a linear gradient of 14 minutes. Injection volume of $5\mu L$. Quality control (QC) samples were made using a combination of all the samples. These samples were analysed at the same time and were inter-dispersed between the samples every 10 samples.

Again the data was processed and extracted in the same manner as in section 4.2.2. The data files were converted to netCDF files using Waters MicroMass databridge software$^{\text{TM}}$ and data extracted using xcms [49] version 1.30.1. The files were separated into classes of P and NP. For peak detection the centWave algorithm [48] was used with 3-20second peak width and lock mass correction on [108]. Grouping was performed using the nearest method and retention time alignment was done with extra=3, missing=2. The final matrix output of intensities was normalised using the median fold change normalisation method [74]. Filters applied were a post-processing minFrac of 50% and a CV filter of 27% on the QC samples, as explained earlier 4.2.2. The final dataset had 2321 features.

5.2.2 NMR COMET Hydrazine Study

The historical COMET dataset [139, 141] for the Hydrazine toxicology studies was a suitable temporal effect study and had a good length time series with many biological replicates. All of the Hydrazine study data was retrieved from the COMET database [139]. Urine samples of Sprague Dawley rats were collected over 168 hours (h). An injection of Hydrazine was given to a population of rats at $30mg/kg$ and $90mg/kg$ dosage. This time point at the injection was time 0, two time points before this were collected. This gave the following time series $-16, \ -8, \ 0, \ 4, \ 8, \ 24$ hours and then every 24hrs after that up to 168hrs. A final population of 3 classes was made, no dose, low dose and high dose. There were the following number of replicates in each class respectively: 26, 28, 24. These samples were collected from 6 different laboratories.

The spectra were collected on 600MHz Bruker spectrometer ^1H NMR. This data was archived and retrieved on the COMET database retrieval system [139]. The intensity data was interpolated over 0.04 ppm wide bins. Following the binning of the spectrum hydrazine and its metabolites spectral resonates were removed. The water peak was also removed from the spectra from 6.00-4.48ppm. Finally, the citrate peaks at bins 2.7 & 2.54 were combined to be two bins such that the 2.7 bin was made up of 0.08 ppm units. The resulting matrix incorporated 200 spectral variables from 78 individuals across 8 time points.

5.2.3 Computational Method - Univariate Analysis

A Welch's two sample t-test was performed on each feature between P and NP class, for the LC-MS dataset across all the time points. For each feature a box plot and EIC were plotted and checked to make sure that the peaks were real. Bonferroni correction (as defined in eq 1.2) was performed due to the multiple tests, using the correction factor of 2321.

Next a time point Welch's t-test was performed. This was done so that each time point was tested separately. However, there were many missing data points in both datasets. For the LC-MS, IVF treatment dataset, a set of samples that had the fewest missing time points and had the most time points was found. Missing time points from this set of samples were interpolated over using linear interpolation. No more than 1 time point was interpolated over and no extrapolation was performed. The resulting matrix of 14-28 days for the LC-MS IVF treatment dataset and 0-168 equally space at 24hr intervals for the NMR hydrazine study was used for the rest of the analysis. This gave the LC-MS IVF treatment dataset 9 individuals and the NMR Hydrazine study remained with the same number (78) of rats as before. After the time point by time point t-test was performed Bonferroni correction was used, with the correction factor of 2321.

5.2.4 Computational Method - Multivariate Analysis

A principal component analysis (PCA) was performed on the interpolated data matrix. First a scaling method was chosen to compensate for the heteroscedastic noise. The following scaling types were tested; unscaled, univariate, Pareto and log base 10. Only the Pareto and UV scaling were chosen based on their results and a PCA model was made in R using the pcaMethods package [149]. For each model a Q^2 value was generated via the Krzanowski method [150]. The Q^2 value helped to decide the scaling method used for the model and number of principal components (PC) to examine. If the Q^2 did not increase from one component to another or started as a negative then the model was removed as a choice. The number of PC were chosen by the Q^2. If the Q^2 did not change from one component to another then the component in which the last change was seen was used. A scores plot of the resulting model was plotted and was coloured using class labels of class and individual, separately.

5.2.5 Computational Method - Intra-unit Temporal Correlations

To analyse the temporal data in a temporal fashion two different methods were designed. The first method has been designed to find processes that affect metabolites differently between classes and lags. These differences are between pairs of different metabolites that have a lagged effect within the individual but have a different lag between the classes.

Using the temporal profile for each feature a pair wise correlation matrix is produced for each individual. First a 0 delay correlation matrix was made. The equation for this can be seen below.

$$\rho_{ixy} = cor(S_{i_x}(t), S_{i_y}(t)) \tag{5.1}$$

where S_{i_x} is the intensity of feature x in the individual i. Here, x and y represent the different features. The correlation of all the features generates the correlation matrix ρ_i for one individual. Using the same structure the temporal profiles can then be lagged to reveal any common structures :

$$\rho_{ixy_\tau} = cor(S_{i_x}(t), S_{i_y}(t + \tau)) \tag{5.2}$$

$S(t)$ is the signal at time (t) and $S(t + \tau)$ is the signal lagged by τ. A lagged correlation matrix ρ_{ixy_τ} is made for each individual subject. Each matrix at lag τ is grouped with the lag matrices from other individuals at τ. Once all of the required lagged correlation matrices are made within each individual, each feature pair of correlation coefficients ρ is tested between classes. A Wilcoxon rank sum test is used to test whether the correlation $\rho_{xy}(t)$ are different between the classes.

Figure 5.1 shows the overall workflow of the method. In Step 1 a classical two class experimental design can be seen with samples being taken over time at equally spaced time points. These samples are then analysed using a high through-put technique such as NMR or MS. In the

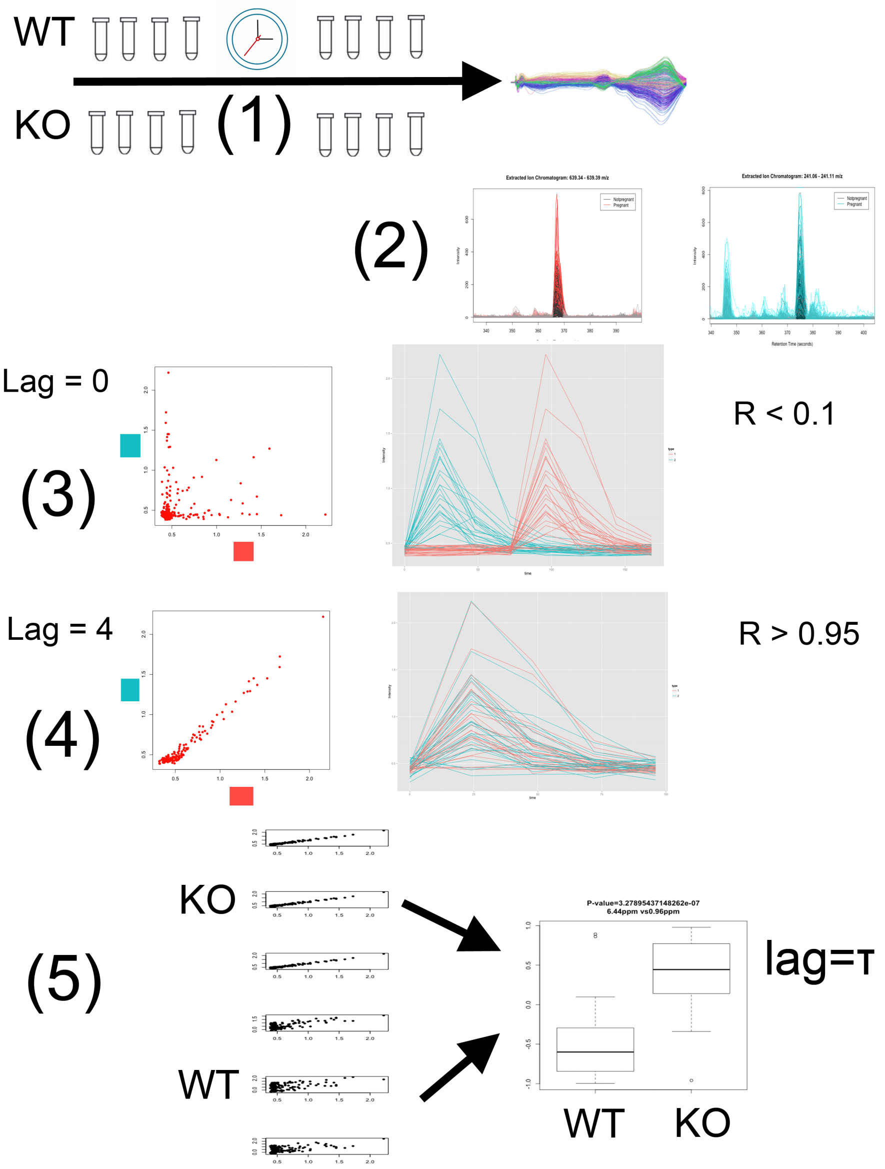

WT
KO
(1)
(2)
Extracted Ion Chromatogram: 639.34 - 639.39 m/z
Extracted Ion Chromatogram: 241.06 - 241.11 m/z
Retention Time (seconds)
Lag = 0
(3)
R < 0.1
Lag = 4
(4)
R > 0.95
(5)
KO
WT
P-value=3.27895437148262e-07
6.44ppm vs0.96ppm
lag=τ
WT
KO

Figure 5.1: The workflow for method 1. In this method samples are taken over time (1) from a two class experiment eg KO vs WT. These samples are then analysed with either NMR or MS. If MS is used then xcms extracts the data and (2) EIC's are checked for quality. The EIC's show two different feature in every class and sample. No obvious mean difference is apparent. Next each feature's time profile for one individual is plotted against the other feature of that individual. At 0 lag there may be no correlation (3) however, at higher delays there may be a high correlation (4). Finally, every feature pair ρ is compared between classes and each lag τ using a Wilcoxon t-test (5).

case of MS analysis a program such as xcms can be used to extract the data. Next, each EIC is inspected (2). The figure shows two different features one in red and the other blue. These features are from different retention times and are different masses. They are not structural pairs. In (3) the temporal profiles are plotted for both features against each other. The blue feature is on the y axis and the red feature is on the x axis. The correlation from these two features is very low. However, the temporal profiles of the two features appears to be very similar is shape but are delayed by a few time points. If we delay or lag the time profiles by 4 time points, in this case, the correlation is now very high (4). The temporal plot next to the scatter plot shows that there is almost a perfect overlap in the red and blue signals. Also since the profiles are now lagged the profile is also shorter by 4 time points. Once this pairwise correlation is performed within each individual subject, the two ρ values can be compared between the two classes. Using a Wilcoxon test, any significant results can be investigated further. A Bonferroni correction is performed to control the false positive rate.

5.2.6 Computational Method - Inter-unit Temporal Correlation

Method 2 is designed to capture differences in temporal events between classes. To do this each feature is correlated with itself between different individuals. Correlations between different features are not used due to the complexity of a biological meaning. The correlation is performed in a pairwise manner for within and between classes. Figure 5.2, shows the general workflow of method 2. A single feature is selected in both classes, eg KO and WT. The temporal profile is then correlated in a pairwise fashion between all individuals of that class and between all individuals of both classes. Each individual is shown as a red (KO) or black circle (WT) coming from the same feature depicted by the EIC. A line between the circles (individuals) depicts a correlation. A ρ is calculated for each given τ for each feature:

$$\rho_{ijx}(\tau) = cor(S_{i_x}(t+\tau), S_{j_x}(t)) \qquad (5.3)$$

$$\rho_{ijx}(-\tau) = cor(S_{i_x}(t-\tau), S_{j_x}(t)) \qquad (5.4)$$

As shown by equations 5.3 and 5.4 the correlation coefficient will be different depending on which individual, (S_i or S_j) for feature x is lagged. For lagged correlations the correlation is performed in both directions, as the correlation will be different. This is because $\rho_{ijx}(\tau) \neq \rho_{ijx}(-\tau)$. This is seen in figure 5.2 where the the two correlation plots have a different ρ value depending the direction of the lag. While this is also true for the within class correlations, the method is trying to identify a difference between the classes. Consequently, the within class correlations are not calculated in both directions. In addition to this, the within class temporal profiles should be similar and so a low variation between the ρ for both directions should be observed. The correlations in figure 5.2 are grouped so that all green correlation (intra-KO class correlation) are grouped, separate from blue correlations (intra-WT class correlations). This results in 4 different classes to compare between; KO-KO, WT-WT, WT-KO, KO-WT. It can also be seen that the same data lagged in different directions give a different ρ. A Kruskal Wallis test is performed to find metabolites that are significantly different between classes. The p-values are then corrected using Bonferroni correction to help reduce for the number of false-positives.

5.2.7 Power Analysis

A power analysis was needed to find the number of samples needed, at 80% power to find a significant result of a given effect size. The calculation was performed with a numerical simulation to test the effect of different sample sizes. Using this simulation allowed for the same amount of variance and overall distributions of the data to be preserved. The work was carried out in a similar fashion as Chadeau *et al* [151] work on power calculations. A logistic model was applied to to determine if a variable was going to be a case or control. The prevalence of the model

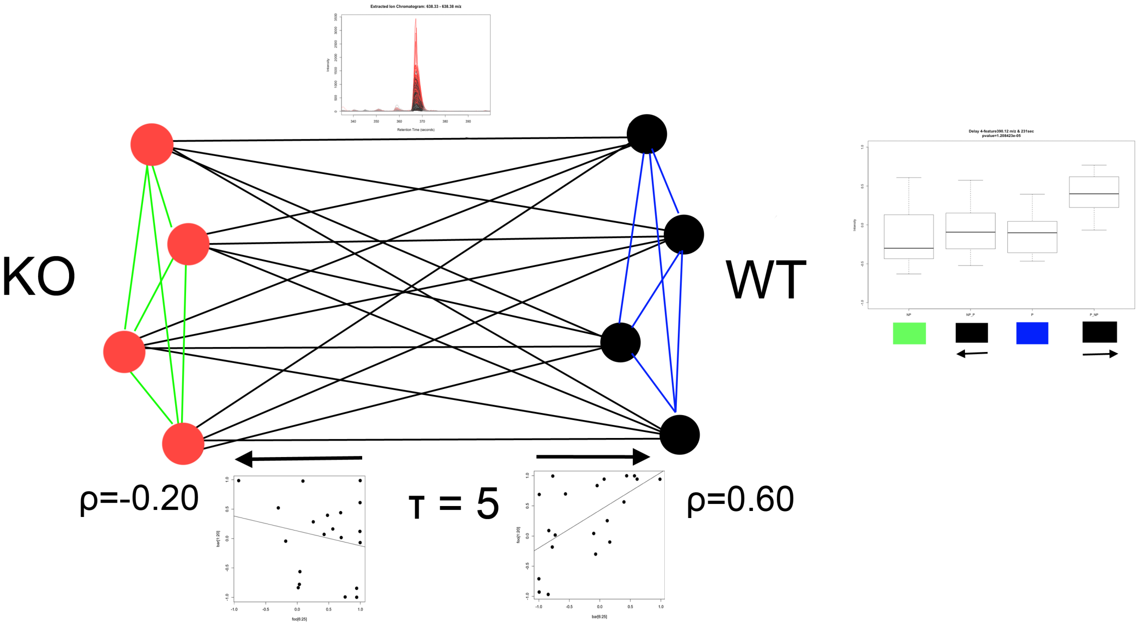


Figure 5.2: The workflow for method 2. In this method each feature's temporal profile is compared between and within each class. This comparison is then also performed at lags and consequently both forward and back directions are needed. Finally the ρ are compared between and within classes to find the differences between classes.

was set to 50%. This allowed both case and control samples classes to be the same size. The logistical model followed equation 5.5

$$P(case) = \frac{e^{B_0+(B_1*\rho)}}{1+e^{B_0+(B_1*\rho)}} \qquad (5.5)$$

$$d = \frac{\bar{x_1} - \bar{x_2}}{\sigma} \qquad (5.6)$$

where B_0 is the prevalence of cases. This decided how balanced the classes are going to be, or the centre point of the logistic function. Since the prevalence was set to 50% this means that in figure 5.3 this is roughly around the correlation coefficient ($\rho = 0$), as this was the median of the distribution. The B_1 value controls the effect size. From figure 5.3 a low effect size can be seen in the top left hand corner of the figure ($B_0 = 0$), whereas a high effect size is seen in the bottom right hand corner of the figure ($B_0 = 10$).

To run the power analysis on method 1 all of the classes were combined into one dataset. From this combined class dataset the B_0 was set to the median (LC-MS pregnancy dataset -0.016 and

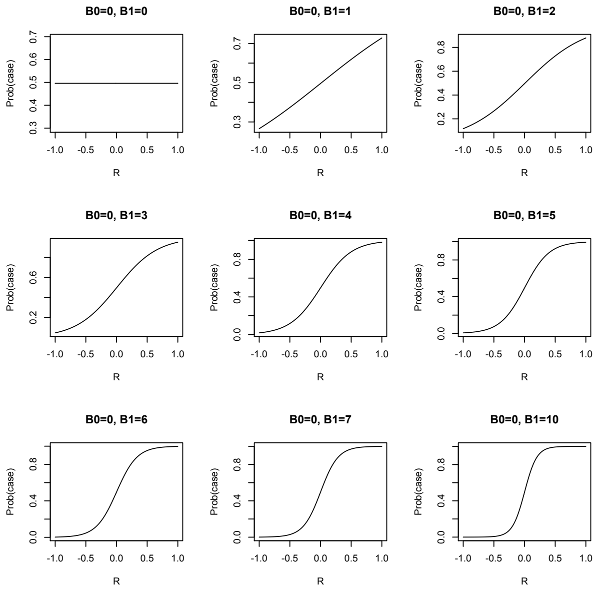


Figure 5.3: The figure shows the effect on the probability of a case changing B_1 while keeping B_0 static. When B_1 is close to 0 the value of the correlation coefficient makes no difference to deciding if the sample is a case or control sample. However, as B_1 increases then the class description of case or control becomes more likely that a high ρ value will result in the sample being a case.

Hydrazine NMR dataset −0.002) correlation coefficients. Using this combined class dataset the values were randomly chosen from the distribution. The probability of being a case was then calculated the sample was assigned to this class accordingly This was repeated until the number of samples desired was reached. The whole procedure was also repeated 500 times and results were averaged. The resulting matrix was an effect size given in Cohens d as shown in equation 5.6. Here, $\bar{x_1}$ and $\bar{x_2}$ are the mean of the two correlation distributions being tested and σ is the standard distribution of the two distributions.

For method 2 a power analysis was also performed however, as there were 4 groups an effect size was more difficult to calculate. The results needed to be comparable between the methods and so a Cohens d was also chosen to express the effect size. However, Cohens d is normally between two means as shown in equation 5.6. Since method 2 is trying to identify differences between the classes the means between the two groups were used. From figure 5.2 this would be the KO-KO and one direction of between class correlations KO-WT.

5.3 Results

5.3.1 Pregnancy Dataset

Conventional Data Analysis

The IVF dataset was collected using a regular 12 minute reverse phase LC-TOF-MS protocol. Each patient submitted urine samples over the course of a month, resulting in 279 samples with a time range spanning from 12 days to 40 days. There were 9 failed IVF pregnancies (NP) and 5 successful IVF pregnancies (P). The raw data was analysed with xcms. This processing resulted in 32,721 features. Features that were not present in either class of NP or P in at least 50% of the samples were removed. This makes the data more reliable and limits the possibility of false positives. The number of features after this removal was 4915. Using the QC samples the coefficient of variation (CV) could be calculated for each feature on the unfiltered data. This gives a measure of how much analytical variation there is. The median CV was 28% and the mean was 38% with the minimum being 7% and the maximum being 591%. These CVs are higher than expected for a LC-MS dataset. A typical value being around 15-20% median CV [63,68]. Using both the minFrac and CV filters (filter of median CV of 28%) the resulting dataset was 2321 features.

Normally with snapshot data i.e. non-temporal data, a simple t-test can be preformed to find the most significant result that differentiates the means of the two classes as shown in chapter 1.9 figure 1.4. Figure 5.4 shows the top two results from a mean level analysis using a Welches t-test. The p-values that are below the Bonferroni cut-off of 0.01 alpha are $1.46E^{-13}$ for M659T367 and $1.46E^{-13}$ for M223T375. Both of the results show a good peak, from the EIC plot. The box plots show that the QC samples have a low amount of variation. However, even though these features have passed all of the selection criteria the there is a high amount of variation in the up-regulated samples. This is due to the experimental design of the time course. A few time points have high intensity whereas some have a low intensity.

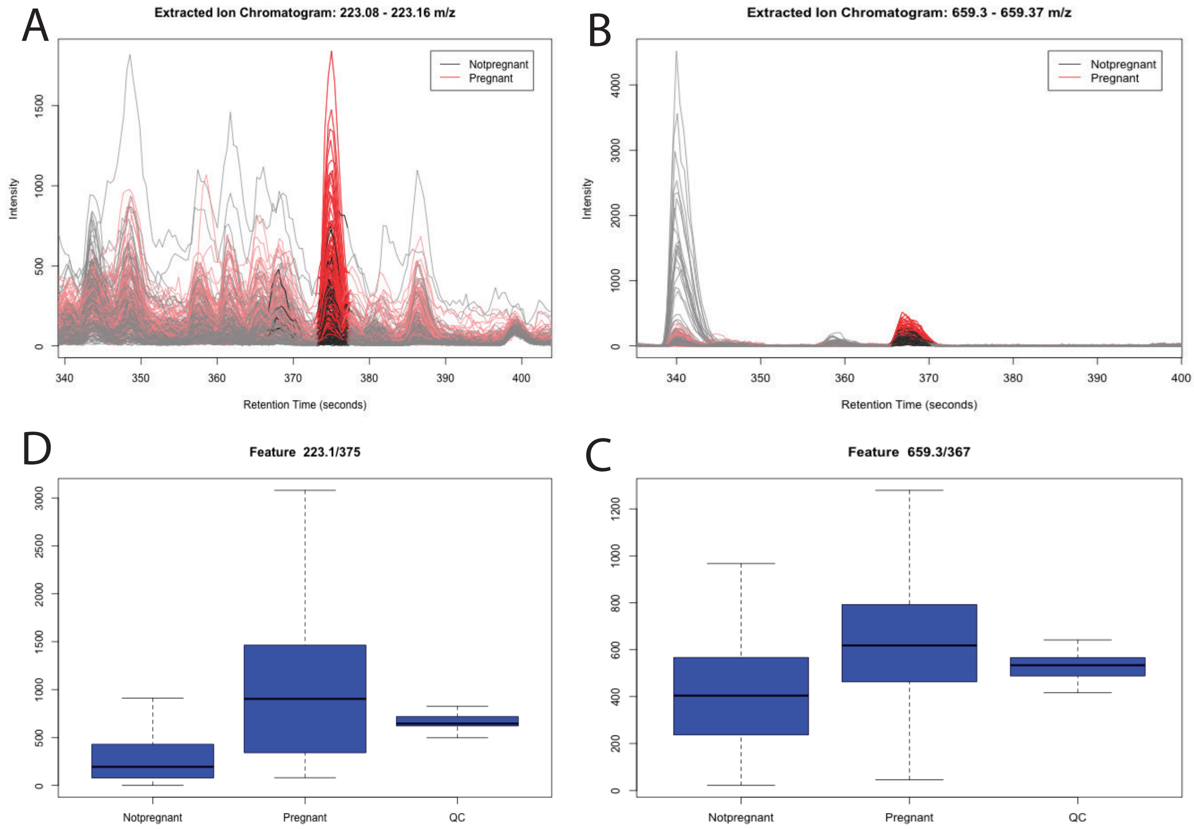


Figure 5.4: The figure shows two significant features from the mean level comparison via a t-test. Below the two EICs are the box plots showing the differences in median level intensity for the three classes of failed IVF (non-pregnant), successful IVF (pregnant) and Quality control samples (QC).

While this method produces some results it does not take into account the dynamic nature of the data. As previously stated the variability is very high. To understand the dynamics the temporal nature must be taken into account. A widely used method is to take each time point separately and perform a t-test on the features. This means that the number of samples will decrease as the full time course is not being used. The variation should also decrease as longitudinal variation has been removed. This is under the assumption that all of the individuals are metabolising at the same rate and that there is a biological event that has synchronised all of the individuals together.

The collected data has many gaps in the temporal profile. This is due to failed sample collection from the individuals. To allow a full temporal profiling for the each patient a simple linear interpolation was performed. This interpolation was then checked manually for each feature. In total no more than one separate time points were interpolated over, in each individual. After this operation there were 15 continuous time points from 14-28 days in 9 individuals. Many individuals had to be removed as they either had too many time points missing or did not overlap with the other individuals. The final result was a continuous time profile that overlaped on all of the patients therefore allowing for a comparison between the individuals and classes. The NP class had 5 individuals and the P class had 4.

Performing the T-test on each time point gave no significant results after Bonferroni correction. Bonferroni correction is a very conservative method of correcting for family-wise error rates. There is also an inherent problem associated with performing the analysis in this fashion. This form of analysis makes the assumption that the different classes are at the same metabolic rate. This is to say that the same metabolic activity is happening on day 14 in all individuals.

Another typical analysis with temporal data is to perform a multivariate analysis. The most common is principal component analysis (PCA). While PCA does not take into account the temporal nature of the data, temporal trajectories can be visualised, which can help in the analysis and understanding of the biological data. As previously stated, PCA can be highly influenced by different scaling methods. The scaling method needs to be chosen as to reduce the amount of variation in the dataset and need to be chosen so that equal weight is given to the variables. In multivariate datasets there can be a different amount and different distributions of variation on the variables. This quantity is known as heteroscedasticity. A good way to test which scaling method will reduce the heteroscedasticity is by plotting the rank mean of each feature intensity vs standard deviation of the intensity [74]. Figure 5.5 shows 4 different scaling factors using the above-described method. If a scaling method is removing or limiting the heteroscedasticity noise then the overall trend should be flat. The test does not work very well for univariate scaling as

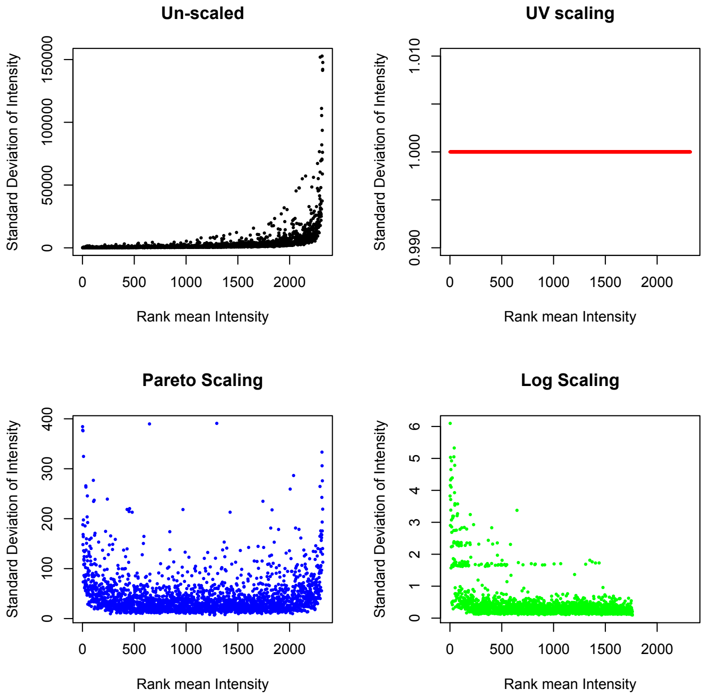


Figure 5.5: Four different plots showing the effects of heteroscedasticity after various scaling methods. In black is the unscaled data, red is univariate scaling, blue is pareto scaling and green is log_{10} scaling. Plots without a trend show a reduction in heteroscedastic noise.

all features have a standard deviation of 1 by definition. However, we can tell that the un-scaled data has heteroscedasticitic noise associated with it. The log scaling has also not been able to remove all of the noise, while the pareto scaling has done a better job there is still a trend of high standard deviations at high and low ranks.

A PCA model was made for the best two scaling methods, univariate and pareto scaling. The Q^2 value is a estimate of how predictive the model is. This value should increase with the number of component of the model. However, as seen in figure 5.6 the Q^2 value decreases in the pareto scaling model. This means that the 2nd component is not valid and so only the 1st component is valid. For this reason the pareto scaling was dropped for further analysis. The score plots in figure 5.7 show the 1st (20% R^2) and 2nd (8% R^2) principal components. There is no obvious temporal trend in the scores plots and there is a spread of the individuals.

Trying to get an accurate metabolic descriptor from this analysis can be difficult. The loadings

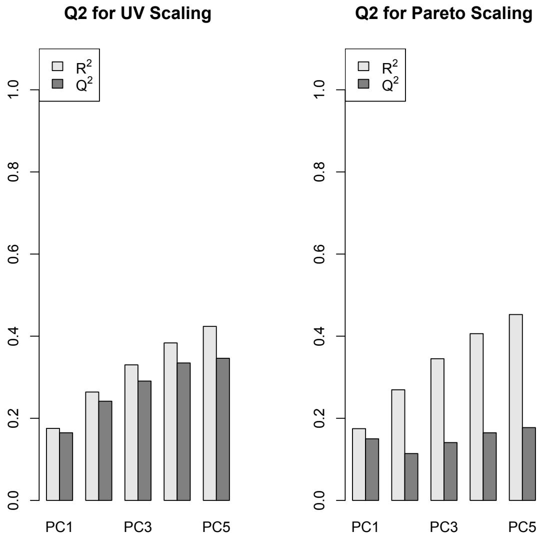
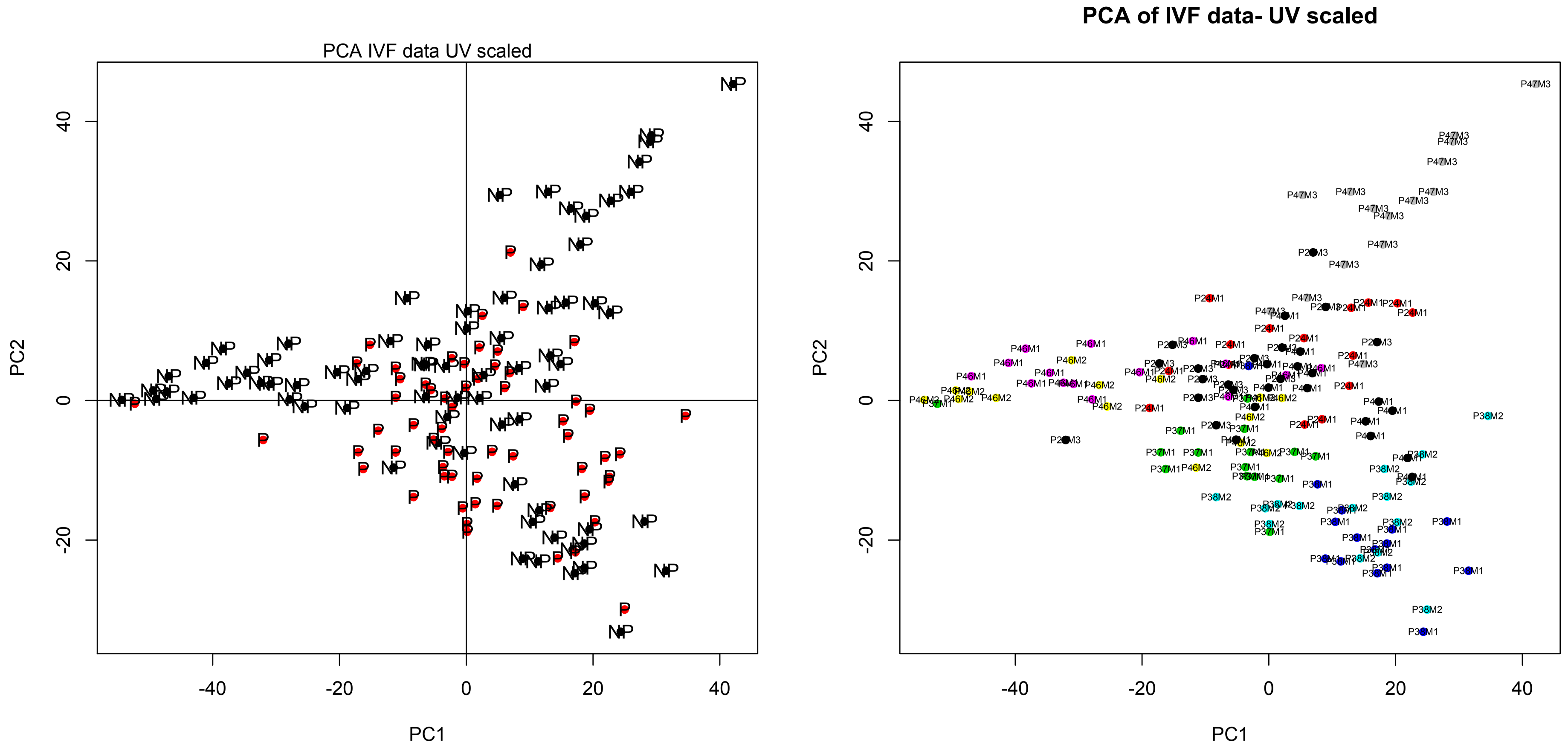


Figure 5.6: The Q^2 and R^2 of two different scaling methods, Pareto and univariate scaling. The Pareto scaling PCA models shows that the Q^2 values decreases after the 1st component. However, the univariate scaling roughly increases linearly until the 5th component.

plot will describe the features that move the samples away from the centre (away from the P class). But these descriptors will mainly be individual specific as demonstrated by figure 5.7b. To understand this data better more individuals would be needed.

A method is needed where the dynamic nature of data is taken into account. A temporal statistical t-test was developed and published called SME (Smoothing spline Mixed Effect models) [97]. This method uses smoothing splines to make an averaged temporal profile for the different classes and then uses a functional t-test to test for significance. The functional t-test integrates the area between the median temporal profiles for each class and then compares this to a bootstrapped null model. While this test is very applicable to this data, the written software was unable to run on the very high number of features in the dataset. A simple reduction of the number of features used was not applicable as the results would not be comparable between methods. Also, a suitable selection of features was not able to be found.



(a) Class

(b) Individual

Figure 5.7: PCA score using UV scaling. Plots are coloured by class (B) and by individual (B). There is grouping/clustering of the individuals and a moving of class NP away from the P class.

Method 1

After correction with Bonferroni there were no significant results. At first this result was surprising as we believed that there should be some differences over time between the two classes in some process between the metabolites. The most likely reason for this is a low statistical power. If we consider the effect size as in equation 5.6, then we can see from $\bar{x_1} - \bar{x_2}$ that the maximum would be ${}^{+}1 - {}^{-}1 = 2$. Consequently, the distributions would have to be very tight to have a significant p-value if the number of samples is kept small. With the small sample set and the limited maximum effect size both contribute to the limited power. To demonstrate that power was the reason for negative results a power calculation was done for the method using the existing dataset.

Power Analysis

The simulation used a specific lag to generate the correlation matrix used for the data input to the simulation. Correlation values were assigned a class using a logistic function. This process was repeated for increasing numbers of samples to a maximum of 100. Figure 5.8 shows the results of this simulation. Using the closest simulated sample size to the true sample size the median effect size was 0 with a 3rd quartile of 1.2 and a maximum of 15.3. However, even with this high effect size the p-value was only 0.0153. This p-value would satisfy an alpha of 0.05 but with false positive discovery correction or Bonferroni correction this p-value is no longer significant. With so few samples it is difficult to get a large enough effect size which will be significant after multiple testing correction. A minimum of 35 samples are needed to get 80% power at an effect size of 2.0. This effect size was not reached with the low number of samples in this dataset.

The power analysis highlights some of this issues with this method. The method requires a median to long time series (> 10 time points) so that the ρ value can be accurately estimated. This has not been shown here with the power test but non-accurate ρ values would increase the

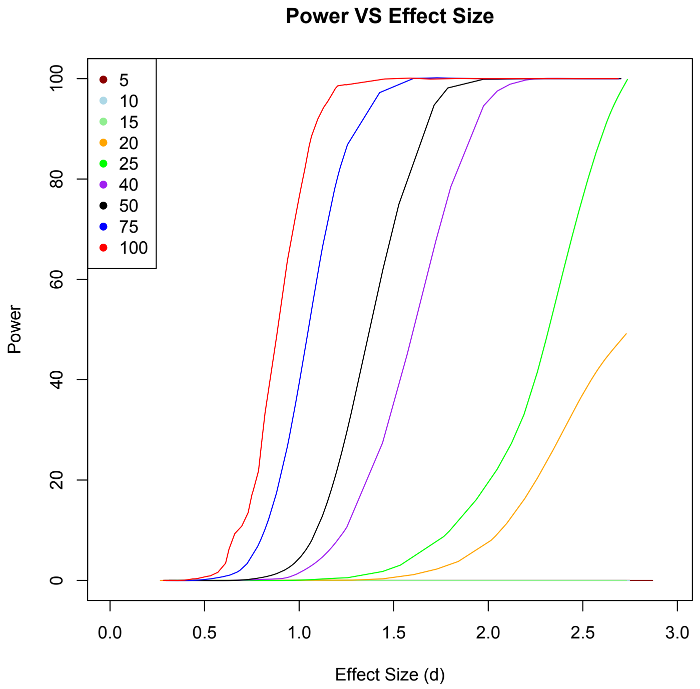


Figure 5.8: Mean power curves for different number of samples. The power is plotted against the Cohen's d effect size.

variation in the effect size. The power analysis shows that the method requires > 25 samples for a dataset with this variation. This requirement for a large sample size is partial due to the limited range of effect sizes. If the two classes have a maximum effect size i.e. one class $\rho = {}^{-}1$ and the other class was $\rho = {}^{+}1$, this is the maximum difference in means possible. This limit translates to a limit in the maximum effect size. The theoretical effect size is still infinite however, this depends upon the standard deviation being very small.

Method 2

Another analysis to perform using correlation is to look at the correlations of the intensities of the features (metabolite) between individuals. This method looks for a lagged onset of a signal from the same feature between classes. The ρ were tested in a 4 way Kruskall-Wallis test. The groups are the two within class correlations and the two between class correlations. The between class correlations are different at different lags as seen in figure 5.2. This method was tested and after using Bonferroni correction on the results a few features were still significant as shown in

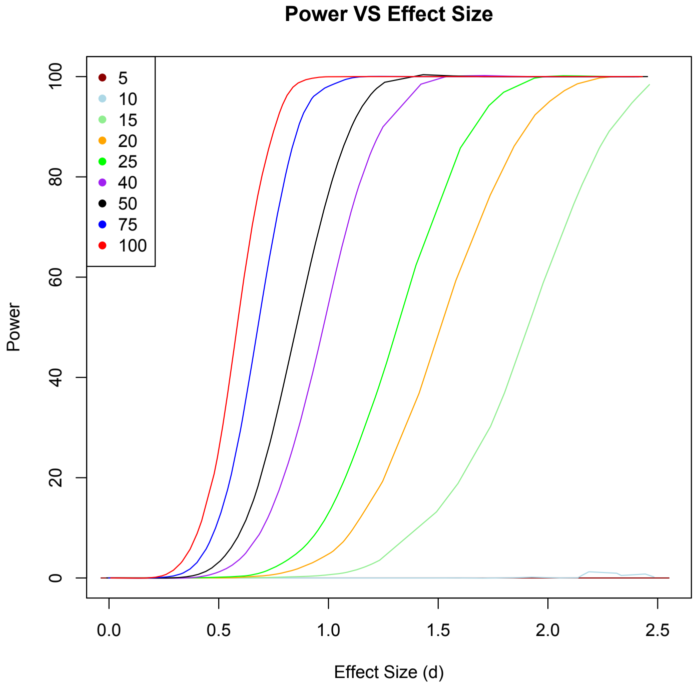


Figure 5.9: Each mean curve is a power simulation. The x axis shows Cohens d. Since both of the between class correlations are chosen as control classes the difference can be demonstrated using Cohens d. The Cohens d

table 5.1.

Power Analysis

Again a power analysis was performed to test how much power the method had. The same logistical model was used as in the previous example. For the method there were 4 classes. This meant that the two classes from the logistic model are repeated twice. These classes were: two cases and two control classes. The cases were the between class correlations and the controls were the within class correlations. A Cohens d was used between one case and one control. This allowed a comparison between the two methods in terms of power. For this analysis the number of observations is 20. While there are only 4 and 5 sample in the P and NP the number of pair observations is 20. Looking at the graph this means that there is sufficient power to be able to find a significant result. If 80% power is used then an effect size of 1.8 with 20 samples is possible.

Table 5.1 shows the features which were significant. To try and identify the metabolites the

CAMERA package was used [51]. This uses a correlation based algorithm to collect adducts, fragments & isotopes into groups. This analysis easies the identification process by characterising the feature and related features thereby allowing for identification of the molecular ion.

Name	p-value	Effect Size	Lag	*m/z* med	RT med	Pseudo ID	Δppm	Metlin ID	KEGG ID
M326T212	4.5×10^{-7}	1.7	1	326.08	212				
M328T243	4.0×10^{-7}	1.9	1	328.10	242	Als Cys Cys	4	18736	
M299T391	1.7×10^{-5}	1.5	1	299.23	390	NSC-76989	15	70609	C15117
M281T391	9.7×10^{-6}	1.3	1	281.22	390				
M539T409	6.1×10^{-7}	1.5	1	539.25	409				
M513T409	9.0×10^{-7}	1.8	1	513.31	409	PA(10:0/10:0)	9	40925	
M328T243	1.0×10^{-5}	1.6	2	328.10	242	Ala Cys Cys	4	18736	
M328T243	2.4×10^{-6}	1.4	3	328.10	242	Ala Cys Cys	4	18736	
M242T181	1.6×10^{-5}	1.4	4	242.09	180	Vitamin B_5	16	70034	
M446T405	8.0×10^{-6}	1.6	4	446.28	404	Trp Leu Lys	14	17082	
M172T201	2.7×10^{-6}	1.7	5	172.09	200				

Table 5.1: Results from method 2 of the Pregnancy data. These non-0 lag results were significant after Bonferroni correction ($\alpha = 0.05$, correction factor 2321). The metabolite ID's are still preliminary.

From this analysis the molecular ion and consequently the neutral mass could be found. After this a search was performed in Metlin, KEGG and HMDB [82,83,85]. If a result was present then the CAMERA [51] spectral plots were used to look at fragments of the correlated group. Using the characterised fragment masses from the CAMERA plots the MS/MS databases in Metlin and HMDB were used to tentatively identify the features. These results are still very preliminary however, such results as vitamin B_5 and the tri-peptides are encouraging. It is known that thyroid problems are a key factor during pregnancy and that the thyroid is responsible for many different signalling processes [152, 153]. Doctors recommend for women with thyroid problems to finish treating these problems before starting on IVF treatments [153]. The thyroid sends many messages to the body via small peptides, such as Thyrotropin-releasing hormone (TRH). It must be remembered that the identities of the metabolites are preliminary and so no conclusion can be draw from these pseudo ID's. To secure the identities of the metabolites which would be needed before any more analysis is done, MS/MS and retention time mapping against a pure standard would be required.

5.3.2 COMET Dataset

Many of the long time series with many biological replicates are toxicology experiments. The COMET project was is a very large study and investigated how the exposure of toxins effected a biological system. One of the larger studies was a hydrazine study. This study was carried out over many different labs. There was a total of of 78 animals in 3 different classes of low dose (n=28), high dose (n=24) and control (n=26). This study was chosen to increase the amount of power for methods 1 and 2. For the analysis the low dose vs control was used as these classes had the highest number of samples.

As a first analysis of the data and to see that the data was the same as the original publication, a PCA model was made [139, 142, 154]. This was carried out in the same fashion as before using no scaling, the scores plot of the first two principal components (54% and 15% R^2 for the first two components respectively) is shown in figure 5.10. It shows the same trend as before, where the higher dose group animals are shifted outside of the main group of control and low dose. As the time course continues the trajectories return to the original position. The data was treated with the same single time point interpolation as the pregnancy dataset. All time points were used for the interpolation, $^{-}16$ to 168 hours. However, so that the time points were equally spaced the pretreatment times were dropped from the analysis. Consequently, the remaining time points were every 24hr starting at $0hr - 168hr$.

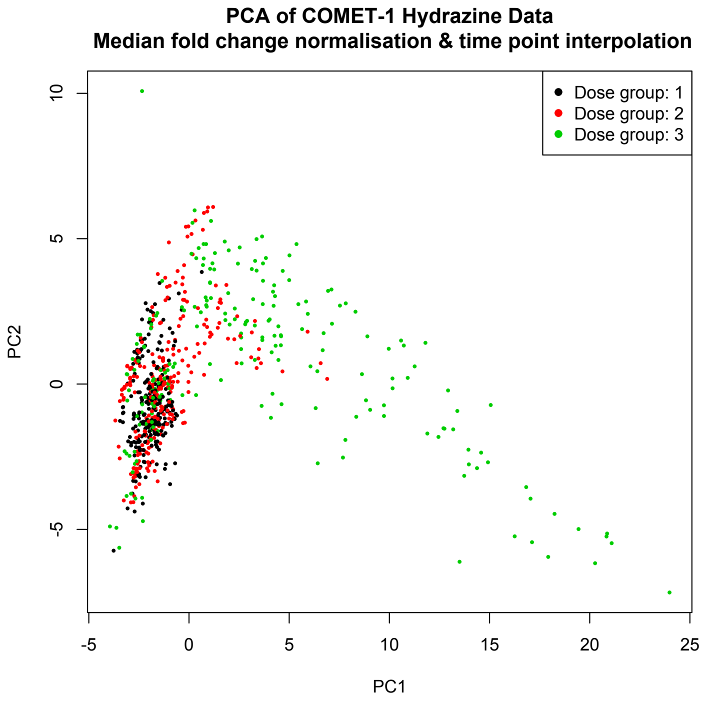


Figure 5.10: A PCA scores plot of the Hydrazine data. The scores plot shows a similar pattern as the published data.

Method 1

Method 1 was run on the COMET Hydrazine NMR data. Briefly, this method identifies feature pairs that have similar temporal patterns at a lag and where these correlations are statistically different from the other class. Using this method on the COMET Hydrazine NMR data a lot of significant features were seen at 0 lag. These significant results are similar to a STOCY analysis. The 0 lags do not use the temporal information of the data and a better assessment can be done using T-tests or multivariate methods. Removing the 0 delay results table 5.2 shows the results using the temporal information.

Due to the large ppm width of the original binning, identification of the spectral resonaces was complicated. Each of the results were checked to make sure that the spectral bins were not noise and contained real peaks. While this excludes that the results were erroneous it is more difficult to say which signal in the bin was the main contributor to the significant result. More work needs to be done with reprocessed high spectral resolution binning or with integrated identified

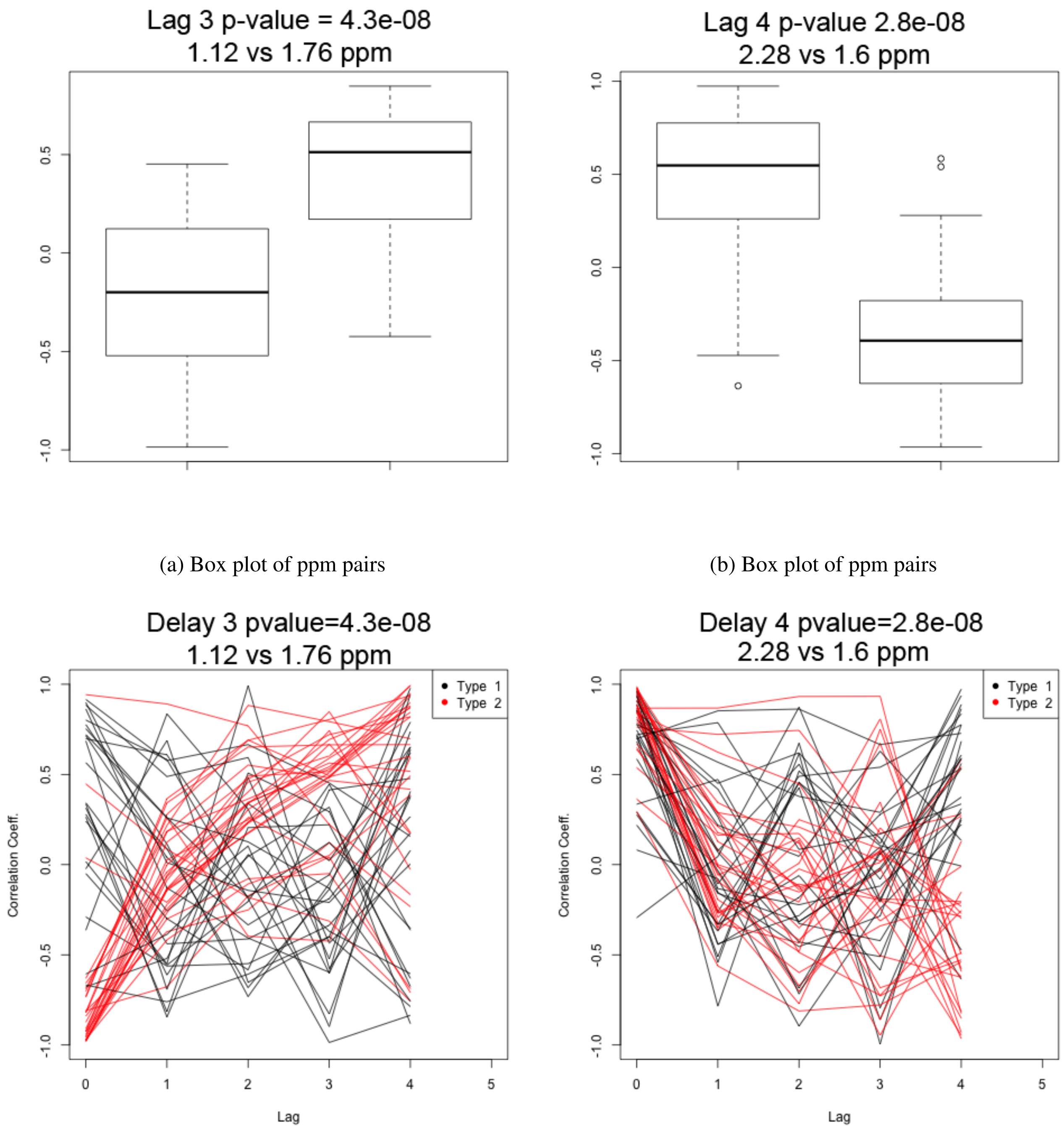
Bin1 (ppm)	Bin2 (ppm)	Lag	p-value	Effect Size
2.24	2.24	1	$8.33e^{-07}$	1.6
1.12	6.48	2	$5.58e^{-08}$	1.7
1.12	1.76	3	$4.26e^{-08}$	1.7
1.12	1.68	3	$5.91e^{-07}$	1.6
1.12	1.64	3	$1.04e^{-08}$	1.9
2.28	1.60	4	$2.82e^{-08}$	2.1
1.64	1.60	4	$9.32e^{-07}$	1.6
1.60	1.60	4	$3.28e^{-07}$	1.8
1.52	1.60	4	$1.04e^{-06}$	1.6
2.28	1.56	4	$6.38e^{-08}$	1.9
1.80	1.56	4	$1.17e^{-06}$	1.6
1.52	1.56	4	$5.26e^{-07}$	1.6
6.44	0.96	4	$3.28e^{-07}$	1.8

Table 5.2: Non zero delay results from Method 1 on the Hydrazine COMET data

compounds (e.g. using BATMAN [60]). The identified compounds would add to the confidence of the results, aiding a biological explanation. Figure 5.11 shows the results from method 1. Figure 5.11a & 5.11b show the differences in ρ between the control and low dose classes in the form of box plots. Both of the plots show a difference in the median ρ between the control and low dose classes resulting in highly significant p-values. These plots were chosen as they are some of the most significant p-values and have a clear correlogram plot.

Power Analysis

As with the pregnancy dataset, a power analysis was performed, figure 5.12. This analysis allowed confirmation that the results with the observed level of significance was viable and that with enough samples the method does have enough power to identify lagged temporal events with reasonable effect sizes. The power curve has been produced in the same fashion as the previous study. The highest effect size observed in this study was 2.05. The number of samples used was 26 samples, this gives a maximum effect sizes of 2.05 for > 80% power.



(a) Box plot of ppm pairs

(b) Box plot of ppm pairs

(c) Correlogram of ppm pair

(d) Correlogram of ppm pair

Figure 5.11: Method 1 applied to COMET Hydrazine NMR data. (a) & (b) Two box plots for the control (type1) vs low dose (type 2) with the y-axis as the correlation coefficient ρ; Two correlograms (c) & (d) show ρ for each sample in each class across the different lags. It can be seen that (c) ρ is most different at delay 3 and that (d) is most different at delay 4. The effect sizes are 1.7 and 2.0 for the two results as seen in table 5.2

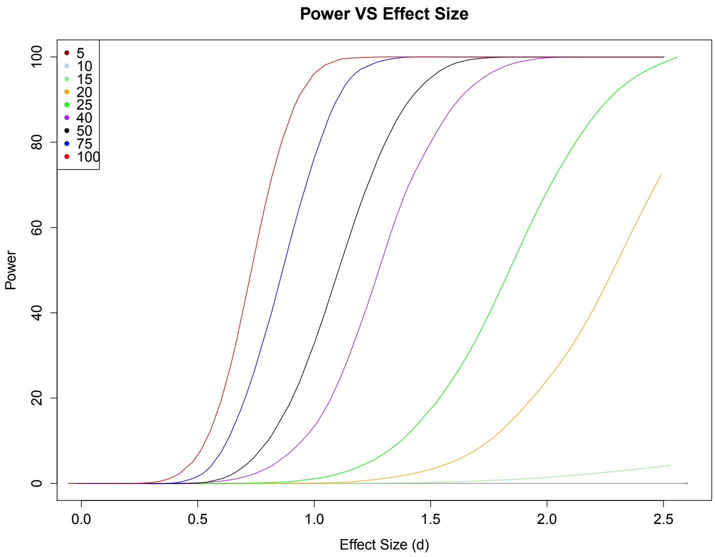
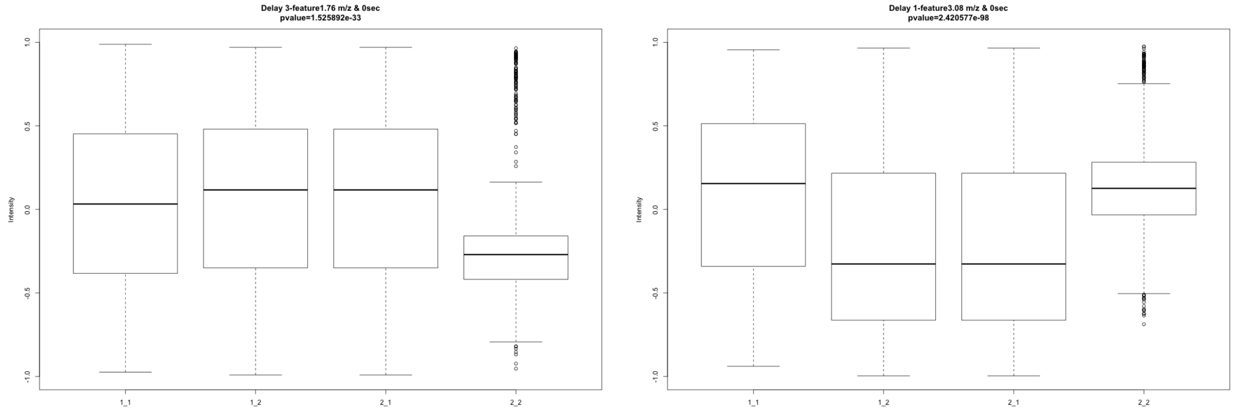


Figure 5.12: Power curves from NMR COMET Hydrazine temporal metabolite data. The power curves are from temporal analysis method 1.

In figure 5.12 where the green line represents 25 samples. As expected, with more samples there is more power, but there is also a sharp pick up in the amount of power after 10 samples. This is suggestive that with the variance in this dataset somewhere between 20 to just below 40 samples is optimum. With the observed effect sizes in this dataset a larger number of samples would be preferred as many of the effect sizes have a lower power than 80%. The mean effect size of the significant results is 1.7. This gives a power of 42%.

Method 2

Method 2 was also run with the Hydrazine data. With the larger sample size more power was expected and consequently more reliable results. In total there were 368 results from 0 lag, whereas the non-zero lags gave much lower number of significant results (38, 23, 29 results for lag 1, 2 and 3 respectively). Many of these results had p-value below 10^{-9} while some had p-values below 10^{-23}. These p-values are low due to the high number of pairs used to calculate the test statistics. The method uses each pair between and within class groups to calculate the p-value. This means that rather than the actual number of biological replicates for each temporal profile used (26 vs



(a) Box plot of 1.76 ppm pairs at lag 3

(b) Box plot of 3.08 ppm pairs at lag 1

Figure 5.13: Method 2 applied to COMET Hydrazine NMR data

28) the test actually uses all of the pairs of temporal profiles. In this larger study this come out to be a lot of pairs, 676, 728, 728 and 784 pairs for the within control group, control vs low dose, low dose vs control and within low dose respectively. As it can be seen these numbers are much higher than the numbers of individuals and the data in each class are not independent. Consequently, the p-values are more overly optimistic than they should be. It is difficult to correct for these effects and the analysis was performed assuming independence but using the conservative bonferroni multiple testing correction to limit false positives. Bonferroni correction lowers the p-value for significance and so was helpful in removing some FP hits. Below, in figure 5.13 are two of the most significant results from this analysis; ppm 1.76 at a lag of 3 and ppm 3.08 at a lag of 1.

In figure 5.13a a strange result can be seen where the median correlation within the the low dose groups has a negative median correlation. This shows poor within group similarity, along with the control group having a close to 0 median correlation. The amount of variance in the distribution of each box is very large. Again, due to the artificial increase in sample size the p-value remains highly significant at 1.52×10^{98}. Figure 5.13b shows a higher within class similarity and ρ. The between class comparisons have a negative ρ. This is very interesting and suggests that

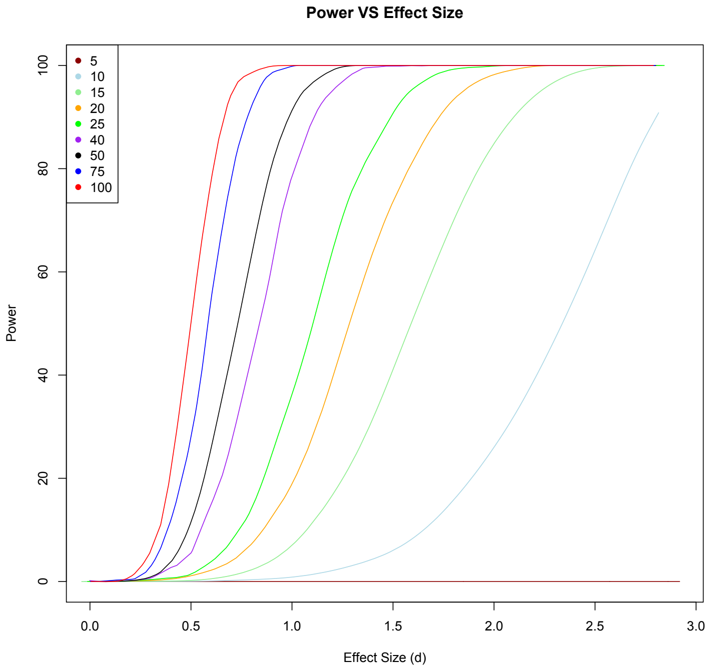


Figure 5.14: Power curves for COMET Hydrazine NMR temporal data, using temporal analysis method 2.

the features have a negative feedback between each other. The temporal profile plots for these features is not revealing as the amount of variation (even on a normalised scale) between the temporal profiles is quite large.

Power Analysis

To see how much power was possible with the increased number of samples a power analysis was done in the same way as with the previous study. Figure 5.14 shows the results from the power analysis. The number of biological replicates was 26 and 28. Since each feature is only being compared to itself across and within classes there should only be 1 comparison for each biological replicate. However, as previously shown (figure 5.2) the lags means that both directions of the comparison need to be considered. Therefore the smallest group has 676 pairs. This means that even small effect sizes have a lot of power. This can be in figure 5.14 where almost all of the power curves hit 100% power by an effect size of 2.0.

5.4 Discussion

Temporal metabolite profiling is still new and many methods are not available or the current coding does not allow a full scale analysis. Some methods, such as the SME algorithm [97] are limited because they are computationally intensive. A highly reduced dataset could have been produced by filtering however, these results would not comparable to the results discussed here as the methods here use the whole dataset. Other non-temporal metabolite profiling methods have been used such as PCA. These results were able to show some overall trends. However, these trends did not show temporal trajectories in the case of the pregnancy data but in the case of the Hydrazine NMR COMET data the PCA shows both smooth temporal trajectories and class differences. The t-tests that were done on mean levels at each time point also did not show any difference between the classes with multiple testing correction. As previously stated this highlights the need for specialised temporal methods.

It was particular noted that the pregnancy dataset had a higher variation than the NMR Hydrazine dataset. Some of the variation can be attributed to the known lower reproducibility in LC-MS datasets [63, 65, 68, 155]. In addition to this the pregnancy dataset was also a human dataset and so consequently had more biological variation than the highly controlled Hydrazine rat study. The increase in variation is seen in the difference between the power curves where a smaller effect size is needed in the NMR Hydrazine study data to achieve equivalent power. For example to achieve 80% power in the LC-MS pregnancy dataset 35 samples were needed with an effect size of 2.0, in the NMR Hydrazine dataset only 26 samples were needed to achieve the same power with a similar effect size. Also the variation can be seen in the PCA plots of both the NMR Hydrazine data and the LC-MS Pregnancy dataset. Due to this increased variation the pregnancy dataset needs more biological replicates to be able to obtain equivalent power as the NMR study. As shown in the previous chapter the amount of analytical variation in the Pregnancy dataset was enough to create some analytical correlations. However, these correlations were at 0 lag. The processes that causes the analytical correlation are highly unlikely to correlate different metabolites through their time profile. Therefore, a lagged temporal correlation will be unlikely to be caused by an analytical process. The analytical variation will however, decrease the ρ if there is a true biological correlation for both method 1 and method 2.

The question of variation in the biological dimension also raises the question of synchronisation of the individuals. Here, method 1 is unique in that it does not suffer from non synchronised individuals. This is because the correlation networks are made within the individual, there is no cross individual connections. Only the correlation coefficient is compared at the end between classes. Unfortunately method 2 is effected by synchronisation of the individuals as it correlates across individuals. This means that if the individuals are metabolising at a different rate or different starting points the method will not work as designed.

While the datasets were from different species they were both urine. Urine datasets can have more variance than plasma or tissue datasets. Unfortunately obtaining other fluids was not possible for both studies. In the pregnancy study it would be unethical to collect endometrial tissue and even daily plasma samples would be difficult. For the Hydrazine study the rats would not have enough blood for daily collection. This was unfortunate as tissue would have allowed for a greater understanding and more targeted analysis of the temporal data. With the body fluids, especially urine, it can be difficult to understand what the process is that is causing the correlation.

In method 2 the p-values in the Hydrazine study are abnormally low. While this means that they pass the Bonferroni correction and are very significant it is unexpected. The reason for the values being so low is due to the number of samples used in the test. Due to the way in which the test is uses pairs rather than single data points the number of test samples is an artificially increased number. Overall the amount of false positives could be increased. While the Bonferroni cutoff is used exactly for this purpose we believe that the correction may not remove the stated number of false positives.

Overall, both methods have been shown to yield useful results that would not have been detected if a temporal metabolite profiling method was not used. Identification of metabolites is still a time consuming and difficult venture however, a final validation of the methods is the biological understand and explanation of the results.
Others work in transcriptomics have used lags on small datasets of micorarrays [92]. Zoppoli *et al* showed that a biologically meaningful network could be reconstructed using a temporal extension the the ARACNE [156] algorithm. The algorithm used the mutual information between transcripts at different lags to build up the network. However, this work did not compare between class groups to find statistically significant transcripts between classes. Others have also looked at elucidating pathways [157]. Lecca *et al* also uses a lags Pearson correlation network to "inference ... pharmacokinetic network". The networks explain the response of the system to a drug or stimuli. Again, these networks do not analyse the difference between two biological class states. We have developed and demonstrated two temporal metabolite profiling analysis methods. The methods have the ability to detected lagged signals or events happening between classes without the need for temporal synchronising of biological replicates. Method 2 is able to detect significantly different lags between classes of the same process. The results show that the methods picks up different metabolites and with enough samples are able to find statistically significant results. We believe that with further developments that these methods and framework will help to promote new temporal algorithm for metabolite profiling.

5.5 Future Work

While both methods have shown that they are able to identify new connections in the data there are improvements that could be made.

- Pre selection filter on features - This could be done in numerous ways:

  Sub grouping:

  > A method could be used to subgroup the different temporal trends into a set number of profiles such as was used in the STEM [137] algorithm. After this method has placed the features into different groups they the most similar and most different could be compared between biological classes.

  Smooth trends:

  > Temporal profiles that are smooth trend may be less likely to be noise profiles

  Similarity metrics:

  > Within class similarity metric for removal of features with low within class similarity. A similarity identifier to find features that have a high similarity within the biological class would allow for a higher degree of confidence in the final results. This would also reduce the amount of computational time required for the comparison. This would be specifically for method 2, but could be extended to method 1 if the network level was used instead.

- Development of a better statistical test for method 2. This developed method would account for the artificially increased test samples and their non independent nature.

- To be able to run the tests in a multivariate fashion. For method 1 this could be done by fitting a smooth spline to all temporal profiles within the individual. Then common features between individuals but within class could be combined to find the median fit within the class. Knowing the variance between the median fit a correlation with the included variance could be used to assess the difference between the classes. This could be done in much the same way as in SME [97]

- Partial correlations. The use of partial correlations would shrink the ρ and reduce the network size removing correlations that were spurious and false positives.
- Use of a non-linear metric such as Mutual information (MI). As previously stated, most of metabolism is not linear. Using a non-linear metric such as MI may help to find biological relationships between the pairs. MI also does not have an upper bound and so the issue associated with using Pearson correlation of a maximum of $\rho = {}^{+}1$ and a minimum of $\rho = {}^{-}1$ would not be seen with MI.

The above list addresses some of issues which were met during the analysis of the datasets. However, the use of the two datasets with both method 1 and 2 is a demonstration that lagged structures in temporal metabolite profiling can help to explain temporal data and adds another dimension to the analysis.

These methods have shown that there is a wide area of research needed not only in temporal metabolite profiling but in this 3rd level analysis. We have simply shown a two class analysis with lags. With more classes and a well designed experiment the within classes lags could be tested as a predictive model for the other classes. Nevertheless, the temporal lags are a new and unexplored area in temporal metabolite profiling.

6 Conclusion

In conclusion we have shown that the aims of the thesis were realistic and have demonstrated methods to resolve the problems discussed. Some of these problems are inherent in the data as part of the profiling of metabolites, while others are specifically due to the platform that has been used.

6.1 Correction of Mass Calibration Gaps in LC-MS

The existence of mass calibration gaps in LC-MS data was a previously unknown and unsolved problem associated with the raw data. This problem could not simply be resolved using a smoothing filter on the peaks as not all of the peaks were accurately detected. The method described has been able to increase the confidence for all of the detected peaks. The integrated intensity of the detected peaks was restored along with a decrease in the false negatives of detected peaks. This software was publicly released and is part of the open source software suite, XCMS [49]. While no computation solution will be able to perfectly restore the lost intensity, the current method alters the expected peak shape due to its design. This solution was taken since it is fast and accurate. This method could be improved upon by using interpolation to return the expected peak shape. Another way would be to apply a smoothing filter to the raw data after the mass calibration gaps have been filled. It is debatable how much the accuracy would be improved by this approach in comparison to the extra computational time required. However, the filling method could also be used for recovering lost intensity when MS/MS experiments are being performed.

6.2 Inter- and Intra-Lab Reproducibility of LC-MS in Urine Metabolomics

The chapter shows that UPLC-MS-TOF is a reproducible technique to be used across multiple labs at different times. This study was essential in being able to understand what is the expected and accepted amount of analytical variability within the lab and how this changed between labs. The study demonstrated that across lab studies, if well designed, are capable of generating highly valuable information. One of the main problems in analysing the data was the change in source settings which meant that different compounds and different adducts were detected. To resolve this, an in-depth study which changes the source settings one parameter at a time and analyses the changes in adduct formation and intensity of those adducts is needed. This would allow for a better understanding of why and how the adducts were made and what, if any, source setting should be changed between instruments and between labs.

To be able to examine the data, all of the spiked-in compounds needed to found in their various adduct forms. The software that was made to find these adducts proved highly useful and versatile. However, this also meant that only the selected spike-ins were analysed. A larger study with a highly complex mixture of labeled compounds, needs to be analysed in this same manner. This would enable a wider range of compounds to be detected and a wider range of endogenous metabolites intensities could be studied. Such a study would also have the drawback of only generating data on one biofluid. However, it is the authors opinion that these studies would be extendable to other biological matrixes.

6.3 Characterising the Static Correlation Structure in LC-MS

We have shown that not only are labeled compounds detectable with accurate RT windows but that, with the use of Pearson correlation, adducts and isotopologues can be confirmed with a high degree of confidence. As more software platforms have started to use correlation for identification of structurally related compounds, this is an important and needed step. However, as previously discussed there are many other metabolite pairs that have a high correlation. These other correlated pairs are either biological correlations or analytical correlations. The analytical correlations may prove to be problematic for downstream analysis where metabolite to metabolite relationships are considered. This was shown using metabolite correlation networks however, methods such as PCA and other factor analysis will also be affected. It is important to note that such methods as PCA are widely used in the field. These analytical correlations could prove very time consuming to researcher following up false biological separations. An automated check in PCA methods to remove these correlations or limit them should prove useful to the field.

While this higher level of reproducibility or variably was shown using a linear metric it is expected that a non-linear metric would also show analytical relationships which may require removal. As these linear and non-linear analytical relationships are discovered the cause of them needs to be studied in greater detail so that this level of reproducibility can also be limited and tested for. Further computational research could be directed towards studying how to separate the analytical distribution without impacting the biological correlation distribution.

6.4 Lagged Correlations in Temporal Metabolite Profiling

Two new and different methods were developed for the 3rd level analysis of metabolite data. These lagged temporal analyses have been shown to be able to extract new and different information from the dataset compared to traditional methods. Like many temporal methods they suffer from low power which in our case prevented further biological analysis. Further to this is the fact that the current implementation of these methods uses a linear metric. Most metabolic reactions would not lead to simple linear relationship and so consequently some relationships may not be detectable. As discussed in the chapter, these methods are a development into the 3rd level analysis for temporal metabolite data processing. Both of the methods developed have shown that more work is needed in this third level analyses. This area of research promises to help produce more temporal data and to further develop temporal metabolite understanding through the use of the temporal lags.

6.5 Summary

Overall we have seen that there are many different levels of reproducibility in LC-MS datasets and that higher levels of reproducibility is based on the mean level (level 1) reproducibility. We have been able to develop new methods to help increase the reproducibility of the data, identify and remove feature pairs that are not reproducible. We have shown that the structural pairs can be identified with confidence. As proposed in the framework another level of analysis is the lagged temporal level. This level requires further analysis, however two novel methods have been proposed and are able to successfully identify new relationships in the data.

This thesis covers many different areas of research with a focus on reproducibility of data within the presented framework. This framework gives new levels for analysis of metabolite profiling data and omics data in a whole. Further work on extending this framework to 'lag' between the different omic technologies would help the understand of not only the metabolites but the biology in question.

Bibliography

[1] van derGreef, J. and Smilde, A. K. (2005) *Journal of Chemometrics* **19(57)**, 376–386.

[2] Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999) *Xenobiotica* **29(11)**, 1181–1189.

[3] Patti, G. J., Yanes, O., and Siuzdak, G. (2012) *Nature Reviews Molecular cell biology* **13(4)**, 263–269.

[4] Dettmer, K., Aronov, P. A., and Hammock, B. D. (2007) *Mass Spectrometry Reviews* **26(1)**, 51–78.

[5] Dalgliesh, C. E., Horning, E. C., Horning, M. G., Knox, K. L., and Yarger, K. (1966) *Biochemical Journal* **101**, 792–810.

[6] Pauling, L., Robinson, A. B., Teranishi, R., and Cary, P. (1971) *Proceedings of the National Academy of Sciences of the United States of America* **68(10)**, 2374–2376.

[7] Hoult, D. I., Busby, S. J., Gadian, D. G., Radda, G. K., Richards, R. E., and Seeley, P. J. (1974) *Nature* **252(5481)**, 285–287.

[8] Oliver, S., Winson, M., Kell, D., and Baganz, F. (1998) *Trends in Biotechnology* **16(9)**, 373–378.

[9] Tweeddale, H., Notley-McRobb, L., and Ferenci, T. (1998) *Journal of Bacteriology* **180(19)**, 5109–5116.

[10] Fiehn, O. (2002) *Plant Molecular Biology* **48(1/2)**, 155–171.

[11] Lewis, G. D., Wei, R., Liu, E., Yang, E., Shi, X., Martinovic, M., Farrell, L., Asnani, A., Cyrille, M., Ramanathan, A., Shaham, O., Berriz, G., Lowry, P. A., Palacios, I. F., Taşan, M., Roth, F. P., Min, J., Baumgartner, C., Keshishian, H., Addona, T., Mootha, V. K., Rosenzweig, A., Carr, S. A., Fifer, M. A., Sabatine, M. S., and Gerszten, R. E. (2008) *The Journal of clinical investigation* **118(10)**, 3503–3512.

[12] Chen, T., Xie, G., Wang, X., Fan, J., Qiu, Y., Zheng, X., Qi, X., Cao, Y., Su, M., Wang, X., Xu, L. X., Yen, Y., Liu, P., and Jia, W. (2011) *Molecular & Cellular Proteomics* **10(7)**, 4945.1–4945.13.

[13] Stalmach, A., Mullen, W., Barron, D., Uchida, K., Yokota, T., Cavin, C., Steiling, H., Williamson, G., and Crozier, A. (2009) *Drug Metabolism and Disposition* **37(8)**, 1749–1758.

[14] Heinzmann, S. S., Brown, I. J., Chan, Q., Bictash, M., Dumas, M. E., Kochhar, S., Stamler, J., Holmes, E., Elliott, P., and Nicholson, J. K. (2010) *American Journal of Clinical Nutrition* **92(2)**, 436–443.

[15] Wijeyesekera, A., Clarke, P. A., Bictash, M., Brown, I. J., Fidock, M., Ryckmans, T., Yap, I. K. S., Chan, Q., Stamler, J., Elliott, P., Holmes, E., and Nicholson, J. K. (2012) *Analytical Methods* **4(1)**, 65–72.

[16] Dyer, A. R., Elliott, P., Stamler, J., Chan, Q., Ueshima, H., and Zhou, B. F. (2003) *Journal of Human Hypertension* **17(9)**, 641–654.

[17] Ohta, D., Kanaya, S., and Suzuki, H. (2010) *Current opinion in biotechnology* **21(1)**, 35–44.

[18] Nakabayashi, R. and Saito, K. (2013) *Analytical and Bioanalytical Chemistry* **405(15)**, 5005–5011.

[19] Afendi, F. M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L. K., Saito, K., and Kanaya, S. (2012) *Plant & cell physiology* **53(2)**, 1–12.

[20] Sawada, Y., Nakabayashi, R., Yamada, Y., Suzuki, M., Sato, M., Sakata, A., Akiyama, K., Sakurai, T., Matsuda, F., Aoki, T., Hirai, M. Y., and Saito, K. (2012) *Phytochemistry* **82**, 38–45.

[21] Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C., and Siuzdak, G. (2009) *Proceedings of the National Academy of Sciences of the United States of America* **106(10)**, 3698–3703.

[22] Swann, J. R., Want, E. J., Geier, F. M., Spagou, K., Wilson, I. D., Sidaway, J. E., Nicholson, J. K., and Holmes, E. (2011) *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4523–4530.

[23] Yanes, O., Tautenhahn, R., Patti, G. J., and Siuzdak, G. (2011) *Analytical Chemistry* **83(6)**, 2152–2161.

[24] Kenny, L. C., Broadhurst, D. I., Dunn, W., Brown, M., North, R. A., McCowan, L., Roberts, C., Cooper, G. J. S., Kell, D. B., Baker, P. N., and Screening for Pregnancy Endpoints Consortium (2010) *Hypertension* **56(4)**, 741–749.

[25] Dunn, W. B., Brown, M., Worton, S. A., Davies, K., Jones, R. L., Kell, D. B., and Heazell, A. E. P. (2011) *Metabolomics* **8(4)**, 579–597.

[26] Aggio, R. B. M., Ruggiero, K., and Villas-Boas, S. G. (2010) *Bioinformatics (Oxford, England)* **26(23)**, 2969–2976.

[27] Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., and Fernie, A. R. (2006) *Nature Protocols* **1(1)**, 387–396.

[28] Rabi, I. (1927) *Physical Review* **29(1)**, 174–185.

[29] Rabi, I. I., Millman, S., Kusch, P., and Zacharias, J. R. (1939) *Physical Review* **55(6)**, 526.

[30] Want, E. J., Wilson, I. D., Gika, H., Theodoridis, G., Plumb, R. S., Shockcor, J., Holmes, E., and Nicholson, J. K. (2010) *Nature Protocols* **5(6)**, 1005–1018.

[31] Ivanisevic, J., Zhu, Z., Plate, L., Tautenhahn, R., Chen, S., O'Brien, P. J., Johnson, C. H., Marletta, M. A., Patti, G. J., and Siuzdak, G. E. (2013) *Analytical Chemistry* **85**, 6876–6994.

[32] Howard, G. A. and Martin, A. J. P. (1950) *Biochemical Journal* **46(5)**, 532–538.

[33] Horvath, C. G., Preiss, B. A., and Lipsky, S. R. (1967) *Analytical Chemistry* **39(12)**, 1422–1428.

[34] Plumb, R. S., Castro-Perez, J., Granger, J., Beattie, I., Joncour, K., and Wright, A. (2004) *Rapid Communications in Mass Spectrometry* **18(19)**, 2331–2337.

[35] Takáts, Z., Wiseman, J. M., and Cooks, R. G. (2005) *Journal of Mass Spectrometry* **40(10)**, 1261–1275.

[36] Carroll, D. I., Dzidic, I., Stillwell, R. N., Haegele, K. D., and Horning, E. C. (1975) *Analytical Chemistry* **47(14)**, 2369–2373.

[37] Fenn, J., Mann, M., Meng, C., Wong, S., and Whitehouse, C. (1989) *Science* **246(4926)**, 64–71.

[38] Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988) *Rapid Communications in Mass Spectrometry* **2(8)**, 151–153.

[39] Go, E. P., Shen, Z., Harris, K., and Siuzdak, G. (2003) *Analytical Chemistry* **75(20)**, 5475–5479.

[40] Karas, M. and Krüger, R. (2003) *Chemical reviews* **103(2)**, 427–440.

[41] Wolff, M. M. and Stephens, W. E. (1953) *Review of Scientific Instruments* **24**, 616–617.

[42] Adams, D. (1985) The Original Hitchhiker Radio Scripts, Random House Value Publishing, .

[43] Saghatelian, A., Trauger, S. A., Want, E. J., Hawkins, E. G., Siuzdak, G., and Cravatt, B. F. (2004) *Biochemistry* **43(45)**, 14332–14339.

[44] Lange, E., Tautenhahn, R., Neumann, S., and Gröpl, C. (2008) *BMC Bioinformatics* **9**, 375.

[45] Synovec, R. E. and Yeung, E. S. (1985) *Analytical Chemistry* **57(12)**, 2162–2167.

[46] Katajamaa, M., Miettinen, J., and Oresic, M. (2006) *Bioinformatics (Oxford, England)* **22(5)**, 634–636.

[47] Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010) *BMC Bioinformatics* **11(1)**, 395–395.

[48] Tautenhahn, R., Böttcher, C., and Neumann, S. (2008) *BMC Bioinformatics* **9**, 504.

[49] Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006) *Analytical Chemistry* **78(3)**, 779–787.

[50] Benton, H. P., Wong, D. M., Trauger, S. A., and Siuzdak, G. (2008) *Analytical Chemistry* **80(16)**, 6382–6389.

[51] Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2012) *Analytical Chemistry* **84(1)**, 283–289.

[52] Tautenhahn, R., Patti, G. J., Rinehart, D., and Siuzdak, G. (2012) *Analytical Chemistry* **84(11)**, 5035–5039.

[53] Tautenhahn, R., Patti, G. J., Kalisiak, E., Miyamoto, T., Schmidt, M., Lo, F. Y., McBee, J., Baliga, N. S., and Siuzdak, G. (2011) *Analytical Chemistry* **83(3)**, 696–700.

[54] Kalman, R. E. (1960) *Journal of basic Engineering* **82(1)**, 35–45.

[55] Bauer, M., Bertario, A., Boccardi, G., Fontaine, X., Rao, R., and Verrier, D. (1998) *Journal of Pharmaceutical and Biomedical Analysis* **17(3)**, 419–425.

[56] Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M., Beckonert, O. P., Schlotterbeck, G., Senn, H., Niederhauser, U., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2002) *Chemical Research Toxicology* **15(11)**, 1380–1386.

[57] Dumas, M.-E., Maibaum, E. C., Teague, C., Ueshima, H., Zhou, B., Lindon, J. C., Nicholson, J. K., Stamler, J., Elliott, P., Chan, Q., and Holmes, E. (2006) *Analytical Chemistry* **78(7)**, 2199–2208.

[58] Ward, J. L., Baker, J. M., Miller, S. J., Deborde, C., Maucourt, M., Biais, B., Rolin, D., Moing, A., Moco, S., Vervoort, J., Lommen, A., Schäfer, H., Humpfer, E., and Beale, M. H. (2010) *Metabolomics* **6(2)**, 263–273.

[59] Weljie, A. M., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. M. (2006) *Analytical Chemistry* **78(13)**, 4430–4442.

[60] Hao, J., Astle, W., deIorio, M., and Ebbels, T. M. D. (2012) *Bioinformatics (Oxford, England)* **28(15)**, 2088–2090.

[61] Liebeke, M., Hao, J., Ebbels, T. M. D., and Bundy, J. G. (2013) *Analytical Chemistry* **85(9)**, 4605–4612.

[62] Zelena, E., Dunn, W. B., Broadhurst, D., Francis-McIntyre, S., Carroll, K. M., Begley, P., O'Hagan, S., Knowles, J. D., Halsall, A., HUSERMET Consortium,, Wilson, I. D., and Kell, D. B. (2009) *Analytical Chemistry* **81(4)**, 1357–1364.

[63] Gika, H. G., Macpherson, E., Theodoridis, G. A., and Wilson, I. D. (2008) *Journal of Chromatography B* **871(2)**, 299–305.

[64] Sangster, T., Major, H., Plumb, R. S., Wilson, A. J., and Wilson, I. D. (2006) *The Analyst* **131(10)**, 1075–1078.

[65] Gika, H. G., Theodoridis, G. A., Wingate, J. E., and Wilson, I. D. (2007) *Journal of Proteome Research* **6(8)**, 3291–3303.

[66] Wang, S.-Y., Kuo, C.-H., and Tseng, Y. J. (2013) *Analytical Chemistry* **85(2)**, 1037–1046.

[67] Kamleh, M. A., Ebbels, T. M. D., Spagou, K., Masson, P., and Want, E. J. (2012) *Analytical Chemistry* **84(6)**, 2670–2677.

[68] Benton, H. P., Want, E., Keun, H. C., Amberg, A., Plumb, R. S., Goldfain-Blanc, F., Walther, B., Reily, M. D., Lindon, J. C., Holmes, E., Nicholson, J. K., and Ebbels, T. M. D. (2012) *Analytical Chemistry* **84(5)**, 2424–2432.

[69] Lei, Z., Huhman, D. V., and Sumner, L. W. (2011) *The Journal of biological chemistry* **286(29)**, 25435–25442.

[70] Sysi-Aho, M., Katajamaa, M., Yetukuri, L., and Oresic, M. (2007) *BMC Bioinformatics* **8**, 93.

[71] Benjamini, Y. and Hochberg, Y. (1995) *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

[72] Storey, J. D. (2002) *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64(3)**, 479–498.

[73] van denBerg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., and van derWerf, M. J. (2006) *BMC Genomics* **7**, 142.

[74] Veselkov, K. A., Vingara, L. K., Masson, P., Robinette, S. L., Want, E., Li, J. V., Barton, R. H., Boursier-Neyret, C., Walther, B., Ebbels, T. M. D., Pelczer, I., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2011) *Analytical Chemistry* **83(15)**, 5864–5872.

[75] Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., Holmes, E., and Nicholson, J. (2005) *Analytical Chemistry* **77(5)**, 1282–1289.

[76] Fiehn, O., Robertson, D., Griffin, J., van derWerf, M., Nikolau, B., Morrison, N., Sumner, L. W., Goodacre, R., Hardy, N. W., and Taylor, C. (2007) *Metabolomics* **3(3)**, 175–178.

[77] Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W. M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. (2007) *Metabolomics* **3(3)**, 211–221.

[78] Johnson, C. H. and Gonzalez, F. J. (2012) *Journal of Cellular Physiology* **227(8)**, 2975–2981.

[79] Wolf, S., Schmidt, S., Muller-Hannemann, M., and Neumann, S. (2010) *BMC Bioinformatics* **11(1)**, 148.

[80] Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., Mamas, M. A., Neyses, L., and Dunn, W. B. (2011) *Bioinformatics (Oxford, England)* **27(8)**, 1108–1112.

[81] Mylonas, R., Mauron, Y., Masselot, A., Binz, P.-A., Budin, N., Fathi, M., Viette, V., Hochstrasser, D. F., and Lisacek, F. (2009) *Analytical Chemistry* **81(18)**, 7604–7610.

[82] Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005) *Therapeutic drug monitoring* **27(6)**, 747–751.

[83] Wishart, D., Tzur, D., Knox, C., Eisner, R., Guo, A., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G., MacInnis, G., Weljie, A., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B., Vogel, H., and Querengesser, L. (2007) *Nucleic Acids Research* **35**, D521–D526.

[84] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) *Nucleic Acids Research* **27(1)**, 29–34.

[85] Pence, H. E. and Williams, A. (2010) *Journal of Chemical Education* **87(11)**, 1123–1124.

[86] Stein, S. (2012) *Analytical Chemistry* **84(17)**, 7274–7282.

[87] Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D., and Kanaya, S. (2006) KNApSAcK: A Comprehensive Species-Metabolite Relationship Database In Plant Metabolomics pp. 165–181 Springer-Verlag Berlin/Heidelberg.

[88] Fahy, E., Sud, M., Cotter, D., and Subramaniam, S. (2007) *Nucleic Acids Research* **35**, W606–12.

[89] Steuer, R., Kurths, J., FIEHN, O., and Weckwerth, W. (2003) *Bioinformatics (Oxford, England)* **19(8)**, 1019–1026.

[90] Opgen-Rhein, R. and Strimmer, K. (2007) *BMC Systems Biology* **1**, 37.

[91] Valcarcel, B., Wurtz, P., alBasatena, N.-K. S., Tukiainen, T., Kangas, A. J., Soininen, P., Jarvelin, M.-R., Ala-Korpela, M., Ebbels, T. M. D., and deIorio, M. (2011) *PLoS ONE* **6(9)**, e24702.

[92] Zoppoli, P., Morganella, S., and Ceccarelli, M. (2010) *BMC Bioinformatics* **11(1)**, 154.

[93] Cakir, T., Hendriks, M. M. W. B., Westerhuis, J. A., and Smilde, A. K. (2009) *Metabolomics* **5(3)**, 318–329.

[94] Meyer, P. E., Lafitte, F., and Bontempi, G. (2008) *BMC Bioinformatics* **9(1)**, 461.

[95] Mallick, B. K., Gold, D., and Baladandayuthapani, V. (2009) Bayesian Analysis of Gene Expression Data, John Wiley & Sons, .

[96] Smilde, A. K., Westerhuis, J. A., Hoefsloot, H. C. J., Bijlsma, S., Rubingh, C. M., Vis, D. J., Jellema, R. H., Pijl, H., Roelfsema, F., and Greef, J. (2009) *Metabolomics* **6(1)**, 3–17.

[97] Berk, M., Ebbels, T., and Montana, G. (2011) *Bioinformatics (Oxford, England)* **27(14)**, 1979–1985.

[98] Wold, S., Kettaneh, N., Fridén, H., and Holmberg, A. (1998) *Chemometrics and Intelligent Laboratory Systems* **44(1-2)**, 331–340.

[99] Wikström, C., Albano, C., Eriksson, L., Fridén, H., Johansson, E., Nordahl, Å., Rännar, S., Sandberg, M., Kettaneh-Wold, N., and Wold, S. (1998) *Chemometrics and Intelligent Laboratory Systems* **42(1-2)**, 233–240.

[100] Liang, Y. and Kelemen, A. (2007) *Biometrical Journal* **49(6)**, 801–814.

[101] Jethava, V., Bhattacharyya, C., Dubhashi, D., and Vemuri, G. N. (2011) *BMC Bioinformatics* **12(1)**, 327.

[102] Xiang, Z., Minter, R. M., Bi, X., Woolf, P. J., and He, Y. (2007) *Bioinformatics (Oxford, England)* **23(18)**, 2423–2432.

[103] Nöh, K., Grönke, K., Luo, B., Takors, R., Oldiges, M., and Wiechert, W. (2007) *Journal of Biotechnology* **129(2)**, 249–267.

[104] Sokol, S. S., Millard, P. P., and Portais, J.-C. J. (2012) *Bioinformatics (Oxford, England)* **28(5)**, 687–693.

[105] Hiller, K., Metallo, C., and Stephanopoulos, G. (2011) *Current Pharmaceutical Biotechnology* **12(7)**, 1075–1086.

[106] Camacho, D., de laFuente, A., and Mendes, P. (2005) *Metabolomics* **1(1)**, 53–63.

[107] Balogh, M. (2009) The Mass Spectrometry Primer, Milford Massachusetts: Waters Corporation, .

[108] Benton, H. P., Want, E. J., and Ebbels, T. M. D. (2010) *Bioinformatics (Oxford, England)* **26(19)**, 2488–2489.

[109] Holmes, E., Wilson, I. D., and Nicholson, J. K. (2008) *Cell* **134(5)**, 714–717.

[110] Wilson, I. D., Nicholson, J. K., Castro-Perez, J., Granger, J. H., Johnson, K. A., Smith, B. W., and Plumb, R. S. (2005) *Journal of Proteome Research* **4(2)**, 591–598.

[111] Sangster, T. P., Wingate, J. E., Burton, L., Teichert, F., and Wilson, I. D. (2007) *Rapid Communications in Mass Spectrometry* **21(18)**, 2965–2970.

[112] Gika, H. G., Theodoridis, G. A., Earll, M., Snyder, R. W., Sumner, S. J., and Wilson, I. D. (2010) *Analytical Chemistry* **82(19)**, 8226–8234.

[113] Geier, F. M., Want, E. J., Leroi, A. M., and Bundy, J. G. (2011) *Analytical Chemistry* **83(10)**, 3730–3736.

[114] Keller, B. O., Sui, J., Young, A. B., and Whittal, R. M. (2008) *Analytica Chimica Acta* **627(1)**, 71–81.

[115] Brenton, A. G. and Godfrey, A. R. (2010) *Journal of the American Society for Mass Spectrometry* **21(11)**, 1821–1835.

[116] Cech, N. B. and Enke, C. G. (2002) *Mass Spectrometry Reviews* **20(6)**, 362–387.

[117] Nordström, A., Want, E., Northen, T. R., Lehtiö, J., and Siuzdak, G. (2008) *Analytical Chemistry* **80(2)**, 421–429.

[118] Crews, B., Wikoff, W. R., Patti, G. J., Woo, H.-K., Kalisiak, E., Heideker, J., and Siuzdak, G. (2009) *Analytical Chemistry* **81(20)**, 8538–8544.

[119] Masson, P., Alves, A. C., Ebbels, T. M. D., Nicholson, J. K., and Want, E. J. (2010) *Analytical Chemistry* **82(18)**, 7779–7786.

[120] Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008) *BMC Bioinformatics* **9(1)**, 163.

[121] Werner, E., Heilier, J.-F., Ducruix, C., Ezan, E., Junot, C., and Tabet, J.-C. (2008) *Journal of Chromatography B* **871(2)**, 143–163.

[122] Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C. L., Francis-McIntyre, S., Begley, P., Carroll, K., Broadhurst, D., Tseng, A., Swainston, N., Spasic, I., Goodacre, R., and Kell, D. B. (2009) *The Analyst* **134(7)**, 1322–1332.

[123] Alonso, A., Julia, A., Beltran, A., Vinaixa, M., Diaz, M., Ibanez, L., Correig, X., and Marsal, S. (2011) *Bioinformatics (Oxford, England)* **27(9)**, 1339–1340.

[124] Creek, D. J., Jankevics, A., Burgess, K. E. V., Breitling, R., and Barrett, M. P. (2012) *Bioinformatics (Oxford, England)* **28(7)**, 1048–1049.

[125] Scheltema, R. A., Kamleh, A., Wildridge, D., Ebikeme, C., Watson, D. G., Barrett, M. P., Jansen, R. C., and Breitling, R. (2008) *Proteomics* **8(22)**, 4647–4656.

[126] Gika, H. G., Theodoridis, G. A., and Wilson, I. D. (2008) *Journal Of Chromatography A* **1189(1-2)**, 314–322.

[127] Wang, J., Zhang, Y., Marian, C., and Ressom, H. W. (2012) *Briefings in Bioinformatics* **13(4)**, 406–419.

[128] DiLeo, M. V., Strahan, G. D., denBakker, M., and Hoekenga, O. A. (2011) *PLoS ONE* **6(10)**, e26683.

[129] Waterfield, C. J., Delaney, J., Kerai, M. D. J., and Timbrell, J. A. (1997) *Toxicology in Vitro* **11(3)**, 217–227.

[130] Jourdan, F., Breitling, R., Barrett, M. P., and Gilbert, D. (2007) *Bioinformatics (Oxford, England)* **24(1)**, 143–145.

[131] Gong, P., Cui, N., Wu, L., Liang, Y., Hao, K., Xu, X., Tang, W., Wang, G., and Hao, H. (2012) *Analytical Chemistry* **84(6)**, 2995–3002.

[132] Karakach, T. K., Wentzell, P. D., and Walter, J. A. (2009) *Analytica Chimica Acta* **636(2)**, 163–174.

[133] Csardi, G. and Nepusz, T. (2006) *InterJournal, Complex Systems* **1695(5)**, 1–9.

[134] Junker, B. H., Klukas, C., and Schreiber, F. (2006) *BMC Bioinformatics* **7**, 109.

[135] Tautenhahn, R., Böttcher, C., and Neumann, S. (2007) *Bioinformatics Research and Development* pp. 371–380.

[136] Friedman, J. and Alm, E. J. (2012) *PLoS Computational Biology* **8(9)**, 1–11.

[137] Ernst, J. and Bar-Joseph, Z. (2006) *BMC Bioinformatics* **7(1)**, 191.

[138] Magni, P., Ferrazzi, F., Sacchi, L., and Bellazzi, R. (2008) *Bioinformatics (Oxford, England)* **24(3)**, 430–432.

[139] Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M., Keun, H. C., Beckonert, O. P., Ebbels, T. M. D., Reily, M., Robertson, D., Stevens, G., Luke, P., Breau, A., Cantor, G., Bible, R., Niederhauser, U., Senn, H., Schlotterbeck, G., Sidelmann, U., Laursen, S., Tymiak, A., Car, B., Lehman-McKeeman, L., Colet, J., Loukaci, A., and Thomas, C. (2003) *Toxicology and Applied Pharmacology* **187(3)**, 137–146.

[140] Kim, J. K., Bamba, T., Harada, K., Fukusaki, E., and Kobayashi, A. (2006) *Journal of Experimental Botany* **58(3)**, 415–424.

[141] Lindon, J. C., Keun, H. C., Ebbels, T. M. D., Pearce, J. M., Holmes, E., and Nicholson, J. K. (2005) *Pharmacogenomics* **6(7)**, 691–699.

[142] Ebbels, T. M. D., Keun, H. C., Beckonert, O. P., Bollard, M. E., Lindon, J. C., Holmes, E., and Nicholson, J. K. (2007) *Journal of Proteome Research* **6(11)**, 4407–4422.

[143] Bar-Joseph, Z. (2003) *Proceedings of the National Academy of Sciences of the United States of America* **100(18)**, 10146–10151.

[144] Shi, Y., Klustein, M., Simon, I., Mitchell, T., and Bar-Joseph, Z. (2007) *Bioinformatics (Oxford, England)* **23(13)**, i459–i467.

[145] Lèbre, S., Becq, J., Devaux, F., Stumpf, M. P., and Lelandais, G. (2010) *BMC Systems Biology* **4(1)**, 130.

[146] Steuer, R. (2006) *Briefings in Bioinformatics* **7(2)**, 151–158.

[147] Huang, T., Liu, L., Qian, Z., Tu, K., Li, Y., and Xie, L. (2010) *BMC research notes* **3(1)**, 142.

[148] Keun, H. C. (2006) *Pharmacology & therapeutics* **109(1)**, 92–106.

[149] Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007) *Bioinformatics (Oxford, England)* **23(9)**, 1164–1167.

[150] Krzanowski, W. J. (1987) *Biometrics* **43(3)**, 575–584.

[151] Chadeau-Hyam, M., Ebbels, T. M. D., Brown, I. J., Chan, Q., Stamler, J., Huang, C. C., Daviglus, M. L., Ueshima, H., Zhao, L., Holmes, E., Nicholson, J. K., Elliott, P., and deIorio, M. (2010) *Journal of Proteome Research* **9(9)**, 4620–4627.

[152] Lazarus, J. H. (2011) *British Medical Bulletin* **97(1)**, 137–148.

[153] Kwak-Kim, J., Han, A. R., Gilman-Sachs, A., Fishel, S., Leong, M., and Shoham, Z. (2013) *American journal of reproductive immunology (New York, N.Y. : 1989)* **69(1)**, 12–20.

[154] Ebbels, T. M. D., Keun, H. C., Beckonert, O. P., Antti, H., Bollard, M., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2003) *Analytica Chimica Acta* **490(1-2)**, 109–122.

[155] Pelikan, R., Bigbee, W. L., Malehorn, D., Lyons-Weiler, J., and Hauskrecht, M. (2007) *Bioinformatics (Oxford, England)* **23(22)**, 3065–3072.

[156] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006) *BMC Bioinformatics* **7**, S7.

[157] Lecca, P., Morpurgo, D., Fantaccini, G., Casagrande, A., and Priami, C. (2012) *BMC Systems Biology* **6(1)**, 51.

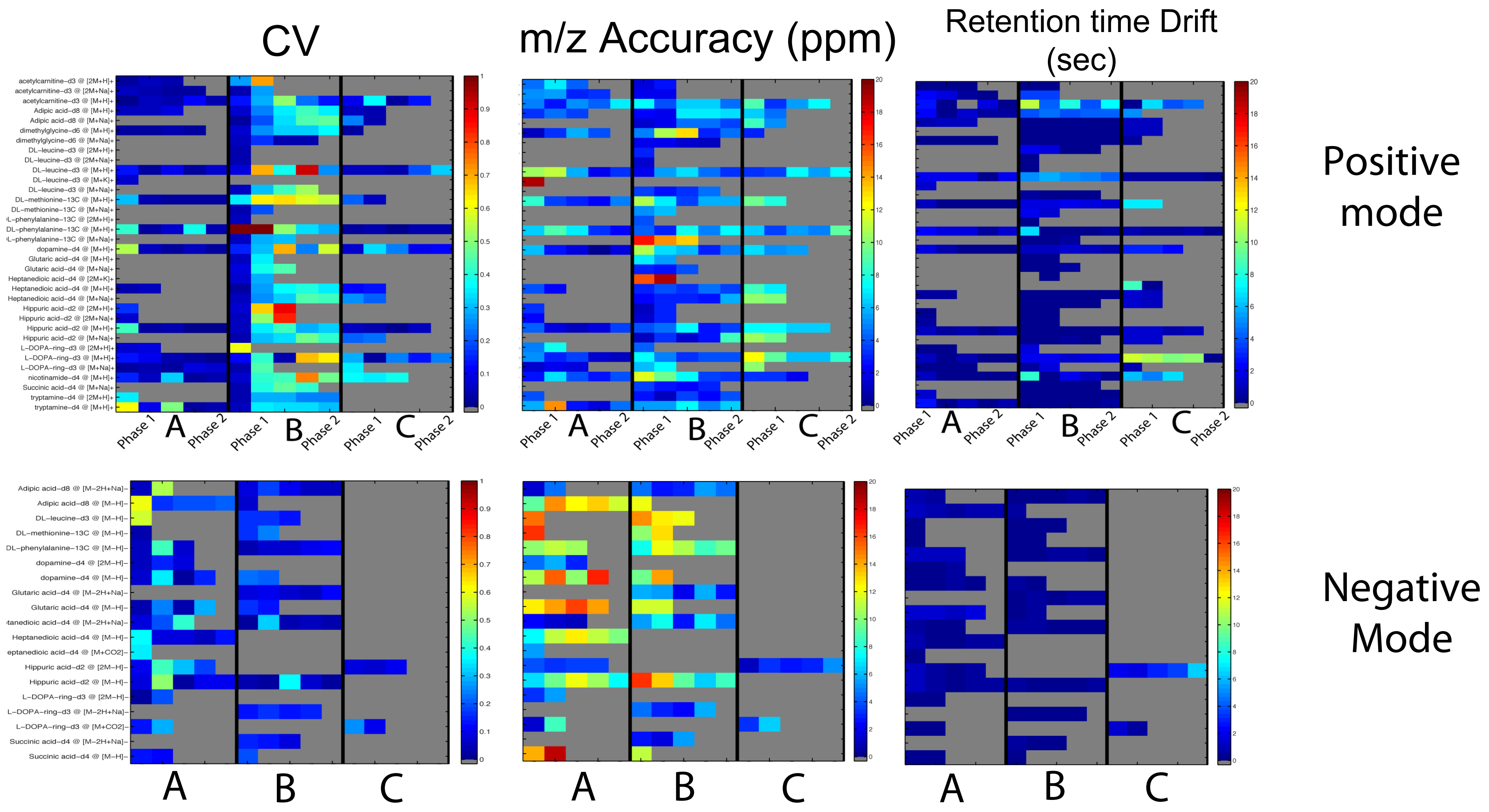
7 Appendix

Name	Charge	Mass Difference
$[M+H]^{+}$	+1	1.01
$[M+2H]^{2+}$	+2	2.01
$[M+3H]^{3+}$	+3	3.02
$[M+Na]^{+}$	+1	22.99
$[M+Na]^{2+}$	+2	22.99
$[M+2Na]^{2+}$	+2	45.98
$[M+K]^{+}$	+1	38.96
$[M+NH_4]^{+}$	+1	18.03
$[2M+H]^{+}$	+1	1.01
$[2M+Na]^{+}$	+1	22.99
$[2M+K]^{+}$	+1	38.96
$[M+2K]^{2+}$	+2	77.93
$[M+2K+H]^{+}$	+1	76.92
$[M+H+Na]^{2+}$	+2	24.00
$[M+H+K]^{2+}$	+2	39.97
$[M+Na+K]^{2+}$	+2	61.95
$[M-C_3H_9N]^{+}$	+1	-59.07
$[M+H-CH_4SO]^{+}$	+1	-62.99
$[M+H-CH_4S]^{+}$	+1	-47.00
$[M+H-C_6H_{10}O_5]^{+}$	+1	-161.04
$[M+H-C_6H_{10}O_4]^{+}$	+1	-145.05
$[M-H]^{-}$	-1	1.01
$[M-2H]^{-}$	-2	2.01
$[M-H2O-H]^{-}$	-1	19.02
[M-OH]-	-1	17.00
[M+Cl]-	-1	-34.97
[2M-H]-	-1	1.01
[M+CO2]-	-1	-43.99
[M-2H+Na]-	-1	-20.97
[M+K-2H]-	-1	-36.95

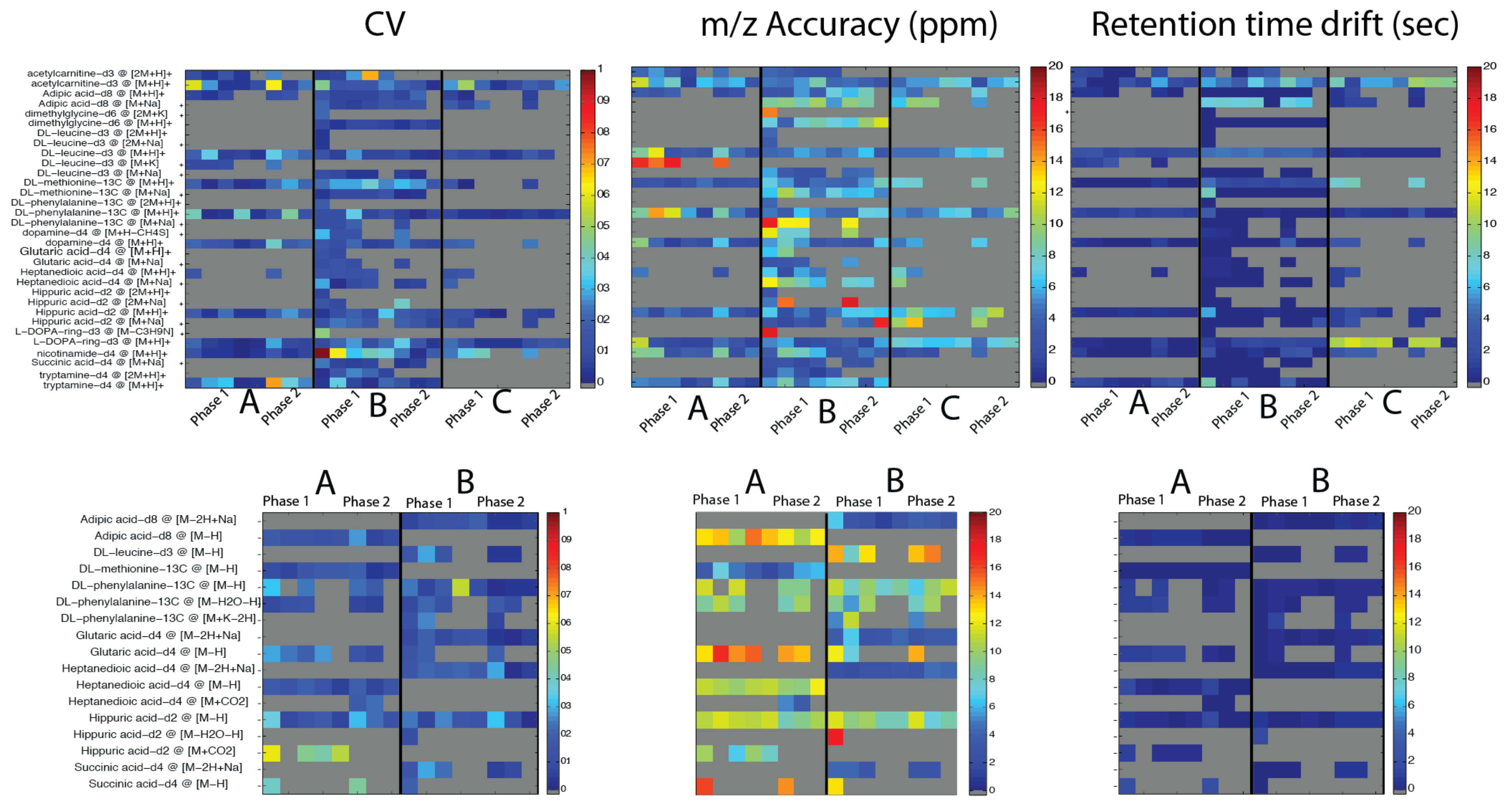
Table 7.1: A list of the possible adduct formations with the charge states and mass difference.

Setting	A Pos	A Neg	B Pos	B Neg	C Pos	C Neg
Ref Scan Freq	50	50	15	15	20	20
Ref Cone Voltage	40	40	50	50	30	30
TTP 4GHz	YES	YES	Yes	YES	Yes	Yes
Dynamic Range Enhancements	YES	YES	Yes	YES	Yes	Yes
mode	W	W	W	W	W	W
Capillary voltage	3000	2500	2700	2000	3000	2800
Sample Cone Voltage	40	40	50	50	30	30
Desolvation Temp	380	380	400	400	300	300
Source Temp	120	120	150	150	120	120
Cone Gas Flow	10	10	50	50	40	40
Reflectron Voltage	1800	1800	1800	1800	1800	1800
MCP detector	2800	2750	2000	2150	2100	2200
Trigger Voltage	600	600	600	650	600	600
Signal Threshold	20	20	90	90	60	60
m/z range	85-1000	85-1000	85-1000	85-1000	85-1000	85-1000
cycle time	0.11	0.11	0.11	0.11	0.16	0.16
Scan Time	0.1	0.1	0.1	0.1	0.15	0.15
inter scan time	0.01	0.01	0.01	0.01	0.01	0.01

Table 7.2: Instrument settings from the different labs

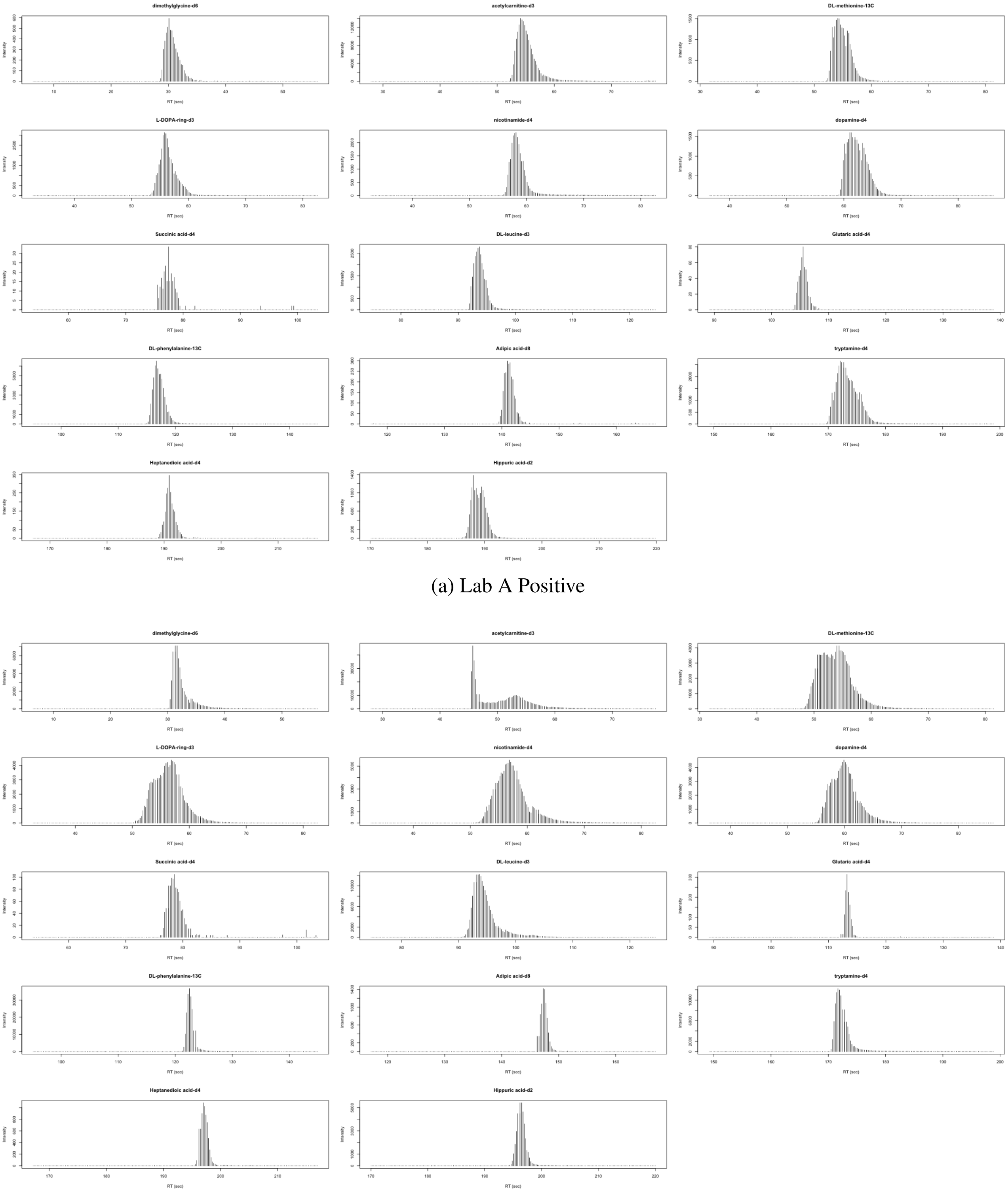

CV
m/z Accuracy (ppm)
Retention time Drift (sec)
Positive mode
Negative Mode
acetylcarnitine–d3 @ [2M+H]+
acetylcarnitine–d3 @ [2M+Na]+
acetylcarnitine–d3 @ [M+H]+
Adipic acid–d8 @ [M+H]+
Adipic acid–d8 @ [M+Na]+
dimethylglycine–d6 @ [M+H]+
dimethylglycine–d6 @ [M+Na]+
DL–leucine–d3 @ [2M+H]+
DL–leucine–d3 @ [2M+Na]+
DL–leucine–d3 @ [M+H]+
DL–leucine–d3 @ [M+K]+
DL–leucine–d3 @ [M+Na]+
DL–methionine–13C @ [M+H]+
DL–methionine–13C @ [M+Na]+
L–phenylalanine–13C @ [2M+H]+
DL–phenylalanine–13C @ [M+H]+
L–phenylalanine–13C @ [M+Na]+
dopamine–d4 @ [M+H]+
Glutaric acid–d4 @ [M+H]+
Glutaric acid–d4 @ [M+Na]+
Heptanedioic acid–d4 @ [2M+K]+
Heptanedioic acid–d4 @ [M+H]+
Heptanedioic acid–d4 @ [M+Na]+
Hippuric acid–d2 @ [2M+H]+
Hippuric acid–d2 @ [2M+Na]+
Hippuric acid–d2 @ [M+H]+
Hippuric acid–d2 @ [M+Na]+
L–DOPA–ring–d3 @ [2M+H]+
L–DOPA–ring–d3 @ [M+H]+
L–DOPA–ring–d3 @ [M+Na]+
nicotinamide–d4 @ [M+H]+
Succinic acid–d4 @ [M+Na]+
tryptamine–d4 @ [2M+H]+
tryptamine–d4 @ [M+H]+
Adipic acid–d8 @ [M–2H+Na]–
Adipic acid–d8 @ [M–H]–
DL–leucine–d3 @ [M–H]–
DL–methionine–13C @ [M–H]–
DL–phenylalanine–13C @ [M–H]–
dopamine–d4 @ [2M–H]–
dopamine–d4 @ [M–H]–
Glutaric acid–d4 @ [M–2H+Na]–
Glutaric acid–d4 @ [M–H]–
tanedioic acid–d4 @ [M–2H+Na]–
Heptanedioic acid–d4 @ [M–H]–
eptanedioic acid–d4 @ [M+CO2]–
Hippuric acid–d2 @ [2M–H]–
Hippuric acid–d2 @ [M–H]–
L–DOPA–ring–d3 @ [2M–H]–
L–DOPA–ring–d3 @ [M–2H+Na]–
L–DOPA–ring–d3 @ [M+CO2]–
Succinic acid–d4 @ [M–2H+Na]–
Succinic acid–d4 @ [M–H]–
Phase 1
A
Phase 2
Phase 1
B
Phase 2
Phase 1
C
Phase 2
A
B
C
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
18
16
14
12
10
8
6
4
2
0

(a) SDS



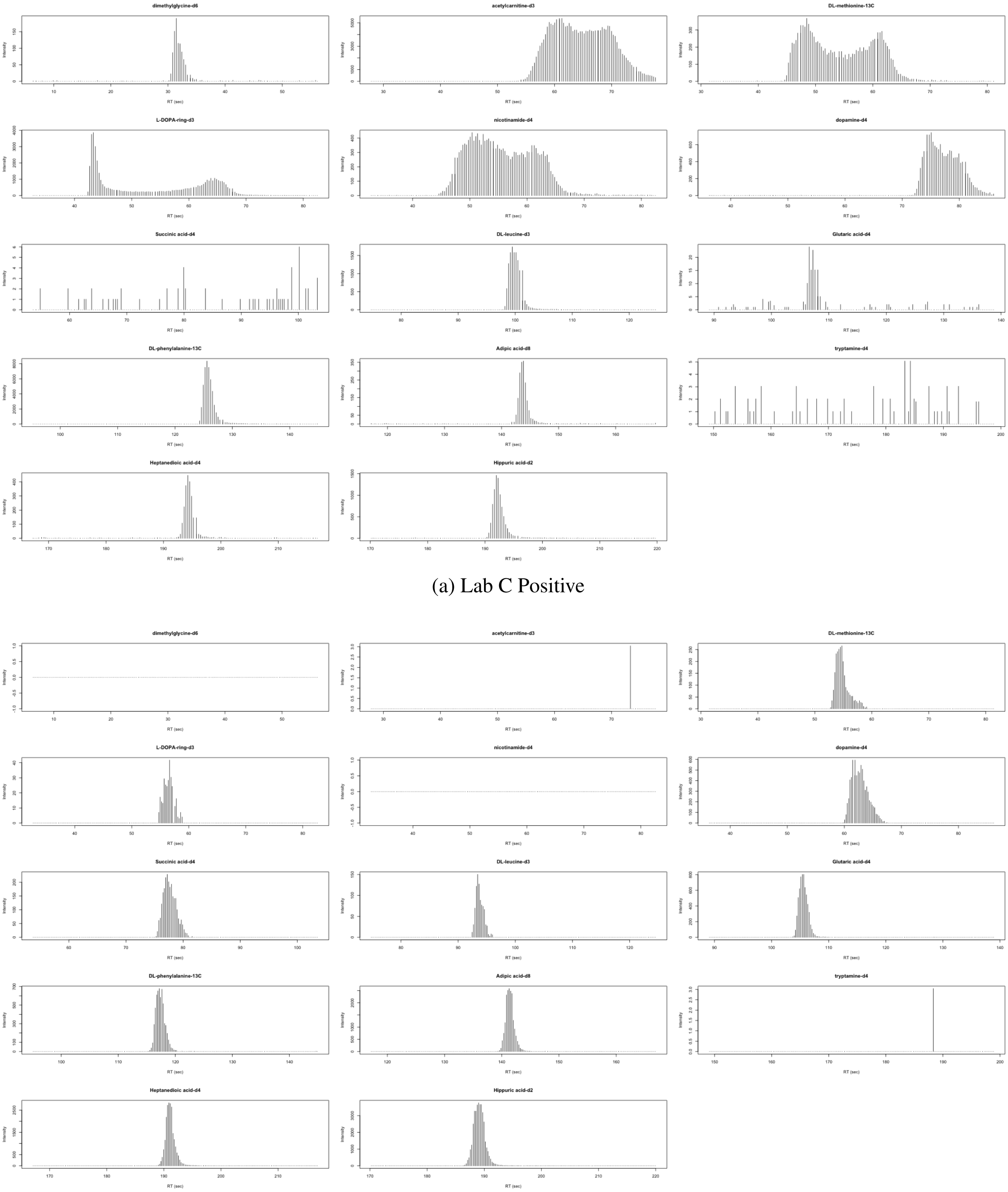
(a) SUDS

Figure 7.1: Reproducibility of intensity, mass accuracy and retention time for the standard compounds including all detected adducts. Heat maps show from left to right the coefficient of variation (CV) of intensity, mean mass measurement accuracy (ppm) and retention time drift (s). Upper panels show positive mode, lower panels show negative mode. Each heat map cell reports data for a one dilution of a given phase (1&2) and lab (A, B & C). Dilution increases left to right. 7.1a: Standard Dilution Series (SDS). 7.1a: Standards in Urine Dilution Series (SUDS). Note: these figures should be viewed at high magnification.



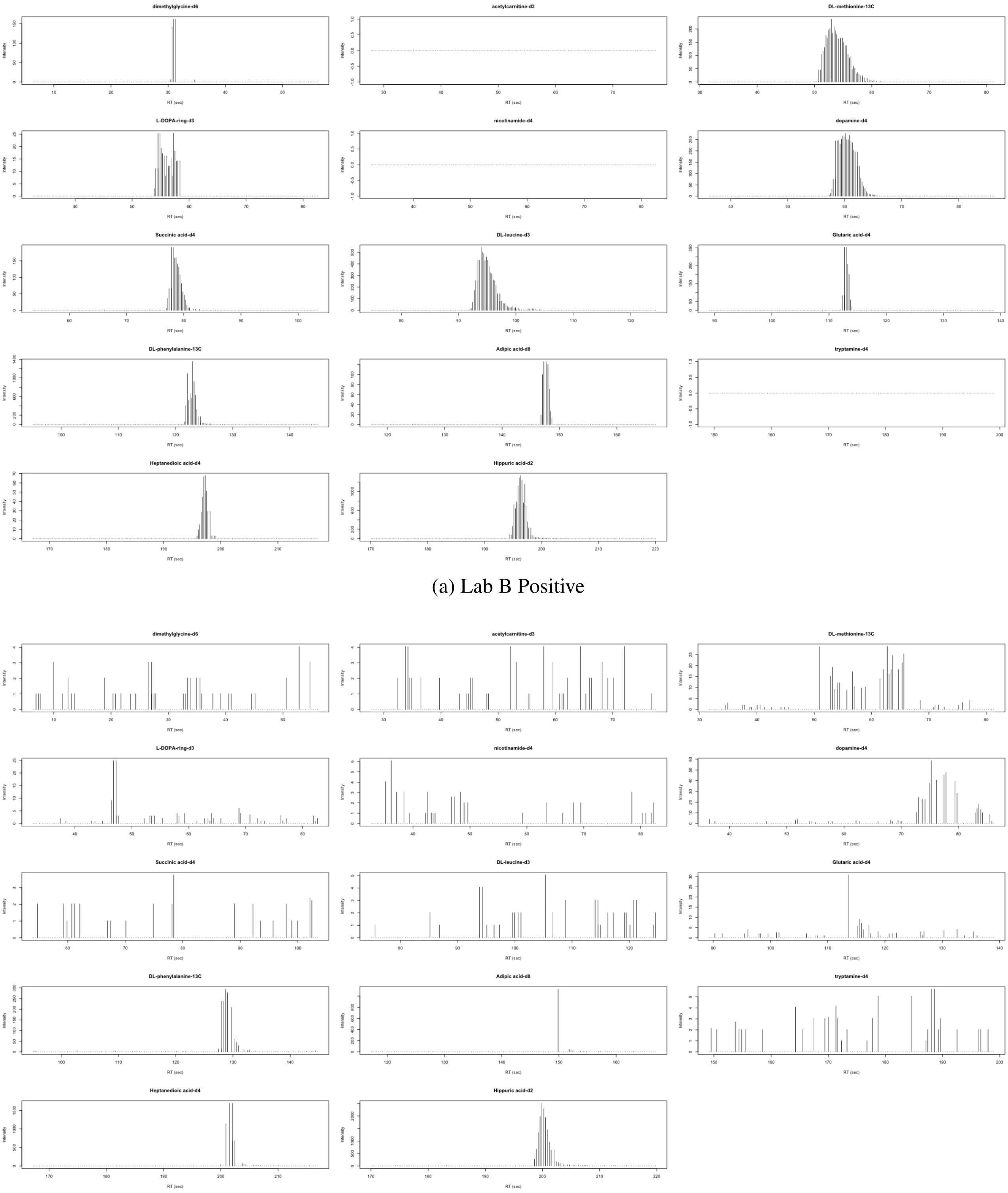
(a) Lab A Positive

(b) Lab B Positive



(a) Lab C Positive

(b) Lab A Positive



(a) Lab B Positive

(b) Lab C Positive

Figure 7.2: This EICs for each of the labelled compounds seen if table 3.1

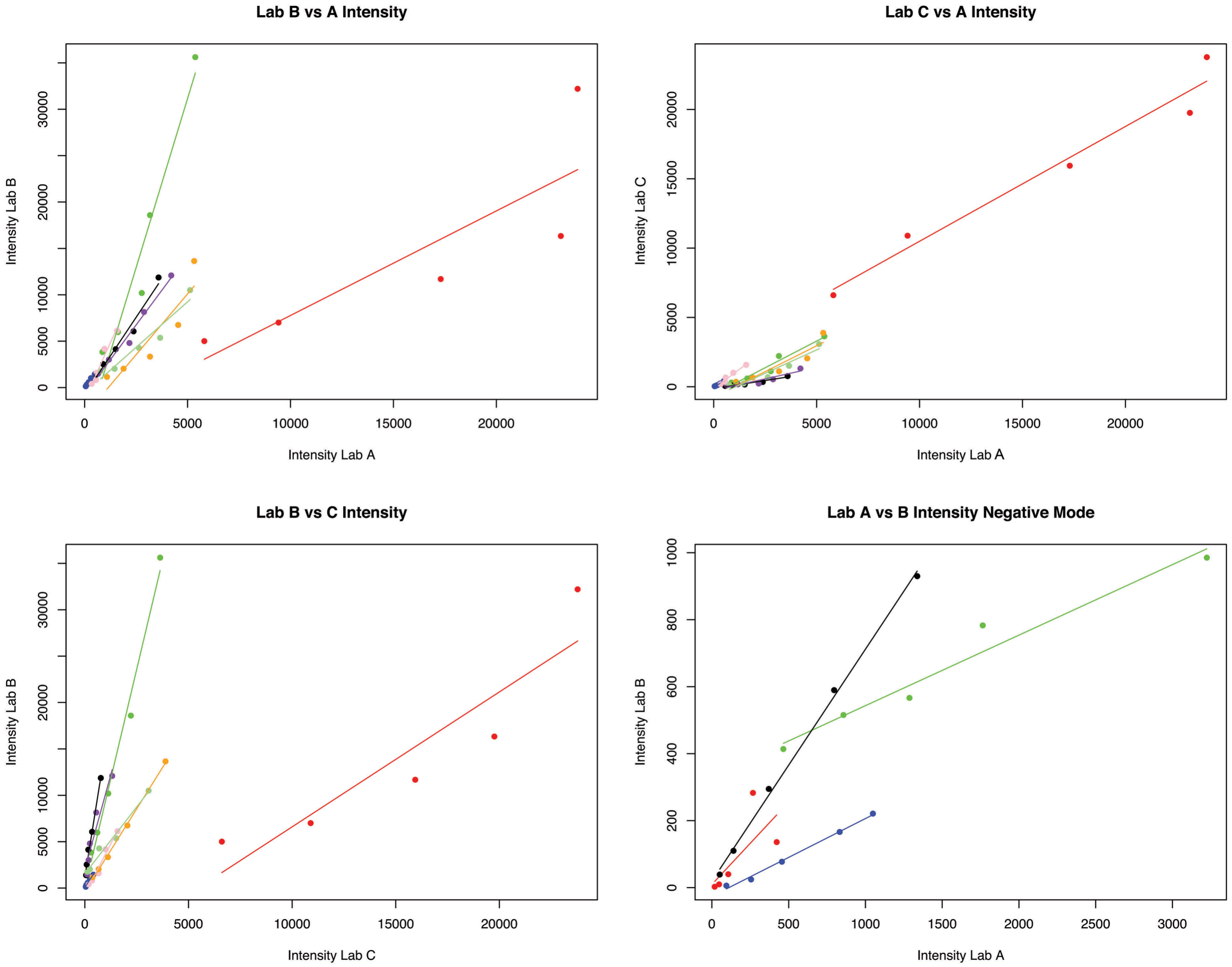


Figure 7.3: Between-lab reproducibility for pairs of labs in positive and negative mode, including acetyl carnitine. Each data point corresponds to a single standard compound adduct averaged across all replicates at a single dilution. Each standard is indicated by a different colour and a separate linear regression line. Key: positive mode: red acetyl carnitine, green - DL-leucine, black - DL-methionine, purple - dopamine, pink - Hippuric, orange - L-DOPA, light green - nicotinamide. Negative mode: red - Succinic acid, Green - Hippuric acid, blue - Glutaric acid, Black - CL-phenylalaine.

Ion	lab.A.phase.1	lab.A.phase.2	lab.B.phase.1	lab.B.phase.2	lab.C.phase.1	lab.C.phase.2
acetylcarnitine-d_3 @ [2M+H]+	0.341	0.564	0.849			
acetylcarnitine-d_3 @ $[M+H]^+$	0.689	0.570	0.791	0.967	0.531	0.910
acetylcarnitine-d_3 @ [M+Na]+	0.928		0.989	0.979		0.988
Adipic acid-d_8 @ $[M+H]^+$	0.986	0.978	0.966	0.968	0.993	0.989
Adipic acid-d_8 @ [M+Na]+			0.916	0.880	0.961	0.992
dimethylglycine-d6 @ [2M+K]+			0.854			
dimethylglycine-d6 @ $[M+H]^+$			0.963	0.967		
DL-leucine-d_3 @ $[M+H]^+$	0.912	0.754	0.943	0.991	0.988	0.977
DL-leucine-d_3 @ [M+K]+	0.974	0.920				
DL-leucine-d_3 @ [M+Na]+			0.977	0.987		
DL-methionine- ^{13}C @ $[M+H]^+$	0.960	0.899	0.890	0.917	0.974	0.977
DL-methionine- ^{13}C @ [M+Na]+			0.894	0.891		
DL-phenylalanine- ^{13}C @ [M]+			0.906		0.992	0.997
DL-phenylalanine- ^{13}C @ $[M+H]^+$	0.677	0.737		0.995	0.997	0.998
DL-phenylalanine- ^{13}C @ [M+Na]+			0.973	0.920		
dopamine-d_4 @ [M+H-CH4S]+			0.971	0.934		
dopamine-d_4 @ $[M+H]^+$	0.931	0.815	0.951	0.926	0.951	0.949
Glutaric acid-d_4 @ $[M+H]^+$			0.926			
Glutaric acid-d_4 @ [M+Na]+			0.951	0.952		
Heptanedioic acid-d_4 @ $[M+H]^+$	0.939	0.633			0.958	
Heptanedioic acid-d_4 @ [M+Na]+			0.495	0.985	0.884	
Hippuric acid-d_2 @ [2M+H]+			0.969			
Hippuric acid-d_2 @ [2M+Na]+			0.949	0.851		
Hippuric acid-d_2 @ $[M+H]^+$	0.888	0.847	0.959	0.971	0.976	0.991
Hippuric acid-d_2 @ [M+Na]+			0.963	0.942	0.967	0.993
L-DOPA-ring-d_3 @ [2M+H]+	0.957	0.958				
L-DOPA-ring-d_3 @ [M-C3H9N]+			0.789			
L-DOPA-ring-d_3 @ $[M+H]^+$	0.913	0.841	0.946	0.930	0.972	0.984
nicotinamide-d_4 @ $[M+H]^+$	0.964	0.940	0.865	0.914	0.949	0.975
Succinic acid-d_4 @ [M+Na]+			0.966	0.991		
tryptamine-d_4 @ [2M+H]+			0.917			
tryptamine-d_4 @ $[M+H]^+$	0.889	0.413	0.953	0.982		

Table 7.3: R^2 values from a linear regression between intensity and dilution for each standard compound adduct in each lab and phase in positive mode. NaN indicates that the ion was detected in less than three dilutions.

Ion	lab.A.phase.1	lab.A.phase.2	lab.B.phase.1	lab.B.phase.2	lab.C.phase.1	lab.C.phase.2
Adipic acid-d_8 @ [M-2H+Na]-			0.828	0.957		
Adipic acid-d_8 @ $[M-H]^-$	0.951	0.925				
DL-leucine-d_3 @ $[M-H]^-$			0.952	0.980		
DL-methionine- ^{13}C @ [2M-H]-					0.804	
DL-methionine- ^{13}C @ [M+Cl]-					0.974	
DL-phenylalanine- ^{13}C @ $[M-H]^-$	0.797	0.932				
DL-phenylalanine- ^{13}C @ [M-H2O-H]-	0.929	0.984	0.961	0.860		
DL-phenylalanine- ^{13}C @ [M+K-2H]-			0.948			
Glutaric acid-d_4 @ [M-2H+Na]-			0.898	0.970		
Glutaric acid-d_4 @ $[M-H]^-$	0.830	0.979	0.887			
Heptanedioic acid-d_4 @ [M-2H+Na]-			0.025	0.165		
Heptanedioic acid-d_4 @ $[M-H]^-$	0.947	0.962				
Heptanedioic acid-d_4 @ [M+CO2]-		0.935				
Hippuric acid-d_2 @ [2M-H]-					0.983	
Hippuric acid-d_2 @ $[M-H]^-$	0.901	0.889	0.699	0.492		
Hippuric acid-d_2 @ [M-H2O-H]-			0.952			
Hippuric acid-d_2 @ [M+CO2]-	0.661				0.828	
Succinic acid-d_4 @ [M-2H+Na]-			0.789	0.944		
Succinic acid-d_4 @ $[M-H]^-$	0.873	0.926				

Table 7.4: R^2 values from a linear regression between intensity and dilution for each standard compound adduct in each lab and phase in negative mode. NaN indicates that the ion was detected in less than three dilutions.

Compound	Comparison	R^2
acetylcarnitine-d_3 @ $[M + H]^+$	Lab C vs B	0.840
acetylcarnitine-d_3 @ $[M + H]^+$	Lab C vs A	0.960
acetylcarnitine-d_3 @ $[M + H]^+$	Lab B vs A	0.710
Adipic acid-d_8 @ $[M + H]^+$	Lab C vs B	0.980
Adipic acid-d_8 @ $[M + H]^+$	Lab C vs A	1.000
Adipic acid-d_8 @ $[M + H]^+$	Lab B vs A	0.990
DL-leucine-d_3 @ $[M + H]^+$	Lab C vs B	0.990
DL-leucine-d_3 @ $[M + H]^+$	Lab C vs A	0.950
DL-leucine-d_3 @ $[M + H]^+$	Lab B vs A	0.950
DL-methionine- ^{13}C @ $[M + H]^+$	Lab C vs B	0.990
DL-methionine- ^{13}C @ $[M + H]^+$	Lab C vs A	0.950
DL-methionine- ^{13}C @ $[M + H]^+$	Lab B vs A	0.970
dopamine-d_4 @ $[M + H]^+$	Lab C vs B	0.920
dopamine-d_4 @ $[M + H]^+$	Lab C vs A	0.870
dopamine-d_4 @ $[M + H]^+$	Lab B vs A	0.980
Hippuric acid-d_2 @ $[M + H]^+$	Lab C vs B	0.970
Hippuric acid-d_2 @ $[M + H]^+$	Lab C vs A	0.960
Hippuric acid-d_2 @ $[M + H]^+$	Lab B vs A	0.960
L-DOPA-ring-d_3 @ $[M + H]^+$	Lab C vs B	1.000
L-DOPA-ring-d_3 @ $[M + H]^+$	Lab C vs A	0.860
L-DOPA-ring-d_3 @ $[M + H]^+$	Lab B vs A	0.830
nicotinamide-d_4 @ $[M + H]^+$	Lab C vs B	0.980
nicotinamide-d_4 @ $[M + H]^+$	Lab C vs A	0.930
nicotinamide-d_4 @ $[M + H]^+$	Lab B vs A	0.930

Table 7.5: Between Lab Reproducibility, (Positive mode
)

Compound	Comparison	R^2
Succinic Acid-d_4 $[M - H]^-$	Lab A vs B	0.499
Hippuric Acid-d_2 $[M - H]^-$	Lab A vs B	0.960
Glutaric Acid-d_4 $[M - H]^-$	Lab A vs B	0.994
DL-phenylalanine- ^{13}C [M-H2O-H]-	Lab A vs B	0.998

Table 7.6: Between Lab Reproducibility, (Negative mode)

	Name	*m/z*	RT(sec)
1	Alanine	89.0480	20
2	Hypotaurine	109.0197	24
3	D-Proline	115.0633	40
4	TLC-A	152.0473	255
5	Uric Acid	168.0283	70
6	L-Citrulline	175.0957	40
7	Asorbic Acid	176.0321	176
8	Hippuric Acid	179.0580	232
9	3-HPPA	180.0423	235
10	Citric Acid	192.0270	73
11	caffeine	194.0804	264
12	L-Tryptophan	204.0900	200
13	Uridine	244.0700	125
14	Chenodeoxycholic acid	392.2930	550
15	ADP	427.0290	70
16	3-Hydroxybutaric acid	104.1073	100
17	1,3 Dimethyluric acid	196.0596	180
18	1-Methyladenine	146.0714	150
19	1-Methylnicotinamide	137.0715	230
20	1-Naphthol	144.0575	160
21	11-Dehydro-thromboxane	368.2199	270
22	11-Hydroxyandersterone	306.2195	300
23	2-Aminobenzoic acid	137.0477	150
24	2-Furoic acid	112.0160	175

Table 7.7: Identified Metabolites in Urine used for Dataset A

	Name	*m/z*	RT(min)
1	Adenosine	268.1032	2.0900
2	Glycerophocholine	258.1095	8.3300
3	Inosine	137.0453	3.0800
4	Iso-Butyl-carnitine	232.1533	6.4600
5	Lidocaine	543.3280	6.0000
6	CPA	415.2110	0.7190
7	L-PC	520.3396	6.0000
8	Hypoxanthine	137.0453	2.2100
9	Uridine	243.0613	1.4000

Table 7.8: Identified Metabolites in Arterial Plaques used for Dataset B