Characterisation of xenometabolome signatures in complex biomatrices for enhanced human population phenotyping

Mark John David

Division of Systems Medicine

Department of Metabolism, Digestion and Reproduction

Imperial College London

Supervised by

Dr Toby Athersuch

Dr Matthew R Lewis

Thesis submitted for the degree of Doctor of Philosophy

2020

Abstract

Metabolic phenotyping facilitates the analysis of low molecular weight compounds in complex biological samples, with resulting metabolite profiles providing a window on endogenous processes and xenobiotic exposures. Accurate characterisation of the xenobiotic component of the metabolome (the xenometabolome) is particularly valuable when metabolic phenotyping is used for epidemiological and clinical population studies where exposure of participants to xenobiotics is unknown or difficult to control/estimate. Additionally, as metabolic phenotyping has increasingly been incorporated into toxicology and drug metabolism research, phenotyping datasets may be exploited to study xenobiotic metabolism at the population level. This thesis describes novel analytical and data-driven strategies for broadening xenometabolome coverage to allow effective partitioning of endogenous and xenobiotic metabolome signatures.

The data driven strategy was multi-faceted, involving the generation of a reference database and the application of statistical methodologies. The database contains over 100 common xenobiotics profiles - generated using established liquid chromatography-mass-spectrometry methods – and provided the basis for an empirically derived screen for human urine and blood samples. The prevalence of these xenobiotics was explored in an exemplar phenotyping dataset (ALZ; n = 650; urine), with 31 xenobiotics detected in an initial screen. Statistical based methods were tailored to extract xenobiotic-related signatures and evaluated using drugs with well-characterised human metabolism.

To complement the data-driven strategies for xenometabolome coverage, a more analytical based strategy was additionally developed. A dispersive solid phase extraction sample preparation protocol for blood products was optimised, permitting efficient removal of lipids and proteins, with minimal effect on low molecular weight metabolites. The suitability and reproducibility of this method was evaluated in two independent blood sample sets (AZstudy12; n=171, MARS; n=285).

Finally, these analytical and statistical strategies were applied to two existing large-scale phenotyping study datasets: AIRWAVE (n = 3000 urine, n=3000 plasma samples) and ALZ (n= 650 urine, n= 449 serum) and used to explore both xenobiotic and endogenous responses to triclosan and polyethylene glycol exposure. Exposure to triclosan highlighted affected pathways relating to sulfation, whilst exposure to PEG highlighted a possible perturbation in the glutathione cycle.

The analytical and statistical strategies described in this thesis allow for a more comprehensive xenometabolome characterisation and have been used to uncover previously unreported relationships between xenobiotic and endogenous metabolism.

Statement of Originality

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

Mark David

November 2020

Copyright Declaration

Acknowledgements

First, I would like to express my sincere thanks and gratitude to my supervisors, Dr Toby Athersuch and Dr Matthew Lewis. I truly appreciate the time, support and guidance you have given me, and always encouraged me to see the bigger picture (at points when I was too bogged down in detail to see it myself). It was your critical thinking that laid the foundation for this work and has inspired me to become a better researcher.

In addition, my sincerest gratitude to the collaborators and clinicians who provided their expertise and datasets for the analytical and statistical developments throughout my PhD, particularly Prof Jeremy Nicholson, Prof Ian Wilson, Prof Paul Elliot (AIRWAVE cohort), Prof Simon Lovestone (ALZ cohort), Dr James Kinross for providing the materials that allowed us to generate the xenobiotic database, and Dr Timothy Schulz-Utermoehl from Sygnature Discovery for the triclosan *in vitro* work.

I am indebted to Dr Joram M. Posma, for sharing his vast breadth of expertise in relation to statistical analyses.

I wish to acknowledge the MRC-NIHR National Phenome Centre (NPC) for their financial support of this project and access to project data. With that, I wish to acknowledge my amazing colleagues at the NPC. It has been a great honour to be part of a team of such truly talented people. The work you do has truly inspired me over the years and has helped me become the scientist I am today. There is no way to express how much it has meant to me to be part of this NPC team, which such a diverse range of background and intellect. Without such a team behind me, I would not be in this place today and for that I am truly grateful.

I would like to especially thank Ms. Katie Chappell, Dr Verena Horneffer-van der Sluis, Dr Maria Gomez Romero, Dr Goncalo Dos Santos Correia, Dr Caroline Sands, Dr Elene Chekmeneva, Mr Stephane S M Camuzeaux, and Mr Ben Cooper. Although they are no longer part of the Imperial College team, I would like to also acknowledge Dr Dave Berry, and Dr Luke Whiley, for their support, practical help and advice along the way.

To my managers at the NPC, Matt and Maria, thank you so much for giving me the time off to write this thesis these last few weeks. It would have been near impossible to have made it here without your unwavering support.

The enzymatic hydrolysis method mentioned in this thesis would not have been possible without the efforts from Dr Elena Chekmeneva and her MRes student, Miss Ziyue Wang, to whom all credit should be directed for experimental design, sample preparation and instrument analysis.

Specific thanks to my colleague and friend Ada Armstrong, for being my partner in crime in the preparation and LC-MS profiling of biological samples. Her intellect and ability to think quickly, has gotten me through many a tough situation. More importantly, she never hesitated to pick up the slack at work when I was too busy having fun with PhD related activities, and for that, I am truly grateful.

Thank you to my sister-in-law, Priya, who provided hand-made and hand-delivered cakes to my front door in moments of dire need. And, to my brother-in-law Avi, whose brain I picked on so many occasions. To my family at home in Sydney – John (Appa) and Jaya (Amma) David, Joshua and Sarah. Being away from home has really made me appreciate you all even more and I miss all of you dearly. Amma and Appa, thank you for your countless sacrifices. I will always appreciate your difficult decision to migrate us to Australia and give me, Josh and Sarah the opportunities and education that you did not have. I would not be the husband, father, and more importantly, the man I am today without the support you have given, and for that, I am eternally grateful. To Josh and Sarah, you are amazing siblings, thank you for your love and encouragement despite the long distance between us.

Finally, I thank my wife Resh. You have been my best friend, who has loved, supported, and encouraged me during this difficult time and has stood by me through all my unravels frustration and impatience. Thank you so much. Along with her, I wish to acknowledge my son Casper, your smile truly lights up my day, and a relief from long day’s work of writing.

Table of Contents

List of figures

List of tables

Abbreviations

AUC	Area under the curve
BD	Bligh and Dyer
C18	Octadecyl carbon chain
Da	Dalton
DOE	Design of experiment
DG	Diglyceride
DSPE	Dispersive solid phase extraction
EIC	Extracted ion chromatogram
EN	Elastic net
ESI	Electrospray ionisation
FDR	False discovery rate
FWER	Family-wise error rate
GMM	Gaussian mixture models
HILIC	Hydrophilic interaction chromatography
HILIC-UPLC-MS	Hydrophilic interaction chromatography - ultra performance liquid chromatography - mass spectrometry
HPLC	High performance liquid chromatography
ID	Identification
LASSO	Least absolute shrinkage selector
LC-MS	Liquid chromatography – mass spectrometry
LIPID-UPLC-MS	Lipid-ultra performance liquid chromatography - mass spectrometry
LLE	Liquid-liquid extraction
LOESS	Locally weighted scatter-plot smoother
LPC	Lysophosphatidylcholine
LTR	Long term reference
m/z	Mass-to-charge (number) ratio
MeCN	Acetonitrile
MeOH	Methanol

MG	Monoglycerides
MR	Method reference
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NMR	Nuclear magnetic resonance
NPC	MRC-NIHR National Phenome Centre
OPLS	Orthogonal projection to latent structures
OPLS-DA	Orthogonal projection to latent structures discriminant analysis
PC	Phosphatidylcholines
PCA	Principal components analysis
PDF	Probability density function
PEG	Polyethylene glycol
PFOS	Perfluoroctanesulfonic acid
PG	phosphoglycerols
PGA	Pyroglutamic acid
PLS	Partial least square
PLS-DA	Partial Least Squares-Discriminant Analysis
PP	Protein precipitation
ppm	Parts per million
PQN	Probabilistic quotient normalisation
Q-TOF	Quadrupole time-of-flight
ROC	Receiver operating characteristic curve
RPC	Reversed phase chromatography
RPC-UPLC-MS	Reversed phase - ultra performance liquid chromatography - mass spectrometry
RSD	Relative standard deviation
RT	Retention time
S/N	Signal to noise ratio
SPE	Solid phase extraction
SR	Study reference
SM	sphingomyelins

TIC	Total ion current
TCS	Triclosan
TCS-Gluc	Triclosan glucuronide
TCS-SO4	Triclosan sulfate
TG	Triglycerides
UPLC	Ultra performance liquid chromatography
VIP	Variable importance for the projection

Chapter 1

Introduction

1.1 General Introduction

Top-down systems biology approaches that rely on data-driven hypothesis generation have helped progress understanding of biological phenomena across a wide variety of research areas. This has been particularly apparent in human health research where applications range from personalized medicine and patient stratification to population-level molecular epidemiological studies (Nicholson, 2006, Henderson et al., 2014, Naylor and Chen, 2010, Alyass et al., 2015).

The ability to characterise complex *molecular phenotypes*, i.e. the phenotypic endpoints resulting from interactions gene-environment interactions (e.g., dietary, lifestyle, environment, gut microbial, and genetic factors), has revolutionised areas in clinical care, epidemiology, and toxicology (Dumas et al., 2006, Nicholson et al., 2005, Sabeti et al., 2007, Holmes et al., 2008).

Molecular phenotypes complement traditional chemistry, physiological, environmental, and lifestyle measurements to help provide a more holistic view of the individual, with methods that can be deployed across large clinical and epidemiological cohorts. The metabolome – the complement of low molecular weight compounds in a biological system/tissue/fluid/compartment – is an important component of the molecular phenotyping picture, given the ubiquity of metabolism in biochemical processes.

The *metabolic phenotype* of an individual may therefore serve as a useful and objective multiparametric measure of prior environmental/xenobiotic exposures (drugs, environmental pollutants, food additives and toxicants) as well as endogenous responses, all played out over the genetic background of an individual (**Figure 1-1**).

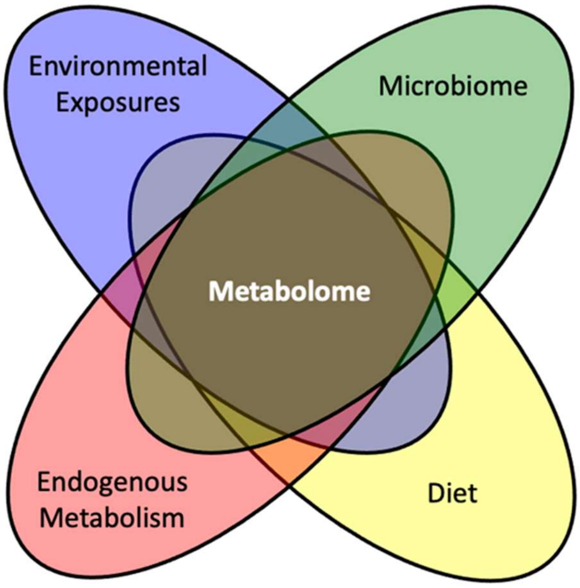


Figure 1-1 The human metabolome is a nexus for the internal and external environment, and a major component of how gene-environment interactions occur. Adapted from (Athersuch and Keun, 2015).

Methods for characterising the metabolome – now interchangeably called metabolic phenotyping / metabolomics / metabonomics / metabolic profiling – can be used to report on the complex chemical composition of the metabolome in biological fluids and tissues. This is a vibrant area of research, with continual improvements to methods being reported on a regular basis; recent developments have been reviewed in a number of different publications (Zhang et al., 2015, González-Riano et al., 2020, Trivedi et al., 2017)

Mass spectrometry (MS) has emerged as one of the most commonly used platforms for metabolic phenotyping, on account of the exquisite resolution and sensitivity that modern instrumentation can deliver; hyphenation with chromatographic separations such as liquid chromatography or gas chromatography provide an additional resolution dimension and richer metabolome dataset for analysis. Consequently, mass spectrometry based metabolic phenotyping has become a central pillar in studies across the biological sciences, and rapid profiling methods have facilitated their application to large cohort studies in clinical and epidemiological studies. In several countries (e.g., UK, Australia, Singapore, China), regional and/or national facilities have been established to deliver high-throughput, analytical services that can share best practice, develop protocols for wider adoption, and anchor efforts to ensure high data quality, and adherence to community guidelines for reporting and data sharing. The National Phenome Centre (NPC) was the first of these centres to be created and provides the context for the work presented in this thesis (NPC: https://phenomecentre.org/).

Metabolic phenotypes derived from MS analysis simultaneously capture (responses in) endogenous metabolism, as well as those arising from xenobiotics and their metabolites (collectively referred to as the *xenometabolome* (Holmes et al., 2007)). Linkage of these two components of the metabolome has the potential to help elucidate underlying mechanisms and identify key environmental determinants of disease (Niedzwiecki et al., 2019); the ability to distinguish components of the metabolome that are directly related to external chemical exposures, and those that reflect endogenous responses, can help reduce the confounding influence of xenobiotics on metabolome-wide analyses, and aid analyses focused on understanding endogenous metabolic regulation (e.g., understanding disease etiology, evaluating therapeutic responses, deriving clinically-relevant biomarkers, etc.).

Xenobiotic signatures in biofluids and tissues are often different to those of endogenous metabolites, including their spatial distribution and temporal variation. Detailed examination of these signatures highlights other distinguishing features of the species that comprise these signatures, including specific physicochemical properties (e.g., halogen substitution), characteristic metabolic fates (e.g., oxidative and conjugative metabolism), and that many xenobiotic exposures also occur as mixtures (e.g., co-administration of a drug and formulation-specific excipients), that often demonstrate highly correlated excretion kinetics.

In vitro and *in vivo* toxicological studies of metabolome responses to xenobiotics commonly partition components of the metabolic phenotype that are related to the test compound and its metabolites (commonly achieved by direct spectral comparison of pre-dose and post-dose samples to positively identify all drug-related compounds) so that they can be excluded from subsequent analysis. In uncontrolled population studies where participants may – knowingly or unknowingly – be exposed to a wide range of xenobiotics, no such comparator exists, yet these exposures may be of importance to data interpretation in these studies.

Additionally, annotation of xenobiotic signatures can report objectively on individual compliance with study protocols, identify outliers, and/or provide population level exposure data. A good example of this is given by Loo *et al.* who developed a xenometabolome screen to identify spectra from NMR metabolic phenotyping data, associated with xenobiotics (Loo et al., 2012). Through statistical exposure models, they were able to identify signal associated with Acetaminophen and ibuprofen in human urine specimens. The results from this demonstrated that the approach was feasible in validating self-reported use of these xenobiotic in the urine samples. Studying xenobiotic metabolism through molecular phenotyping can potentially provide insight into the inter-individual differences in xenobiotic responses (Holmes et al., 2007) which ultimately, can guide individualised drug therapies and drive advances in personalised medicine.

1.2 Scope of the Thesis

The overall aim of the work described in this thesis was to enhance common metabolic phenotyping assays and the analysis conducted on the datasets they generate by significantly increasing the number of positively identifiable and annotatable metabolic profile features relating to xenobiotics.

To this end, several different elements were brought together, including rational selection of prioritised xenobiotics using available literature data, generation of database-ready mass spectral data for authentic chemical standards, development of statistical analysis tools, and interrogation of existing large-scale population study data and to validate a platform for the untargeted profiling of blood products in large-scale molecular phenotyping studies.

Five key objectives were identified to address the overall aim of this work; each is briefly outlined below, and described in detail in subsequent chapters, as indicated:

1. Creation of spectral database of common xenobiotic profiles. *Chapter 3.*
   Pharmaceuticals that are commonly used by the general population – and others compounds of specific relevance to the exemplar populations in this work – were identified using publicly available/literature resources. These were then analysed using established metabolic phenotyping methods to generate a core spectral resource.

2. Identification of predicted xenobiotic metabolite features. *Chapter 3.*
   Extensive literature searching, and application of software-based methods for plausible metabolite prediction were employed to flag and putatively annotate potential xenobiotic-derived compounds in existing metabolic phenotype data.

3. Development of additional statistically based methods. *Chapter 3.*
   Discriminatory chemical properties and statistical relationships of xenobiotics were used to develop tools for enhanced extraction of xenometabolome signatures in existing metabolic phenotyping datasets.

4. Improved analytical methods for xenometabolome coverage. *Chapter 4.*
   An enhanced blood preparation protocol was developed and optimized for the removal of lipids and protein with minimal effect on other low molecular weight metabolites.

5. Application to real phenotyping studies. *Chapter 5.*
   Exploration of detectable exemplar xenobiotic exposures in selected large-scale epidemiological datasets using existing and the developed RPC methods:

   a. Detection of novel and direct xenobiotic metabolites
   b. Cross-correlation of xenobiotic and endogenous metabolic profiles
   c. Annotation and identification of associated metabolites of exposure and xenobiotic co-exposures
   d. Examination of endogenous metabolism in relation to xenobiotic exposure

Figure 1-2 illustrates the connectivity between the main element of this thesis, showing how the empirically-derived, data driven, and laboratory assay development have been separately developed and brought together for application on exemplar NPC epidemiological studies.

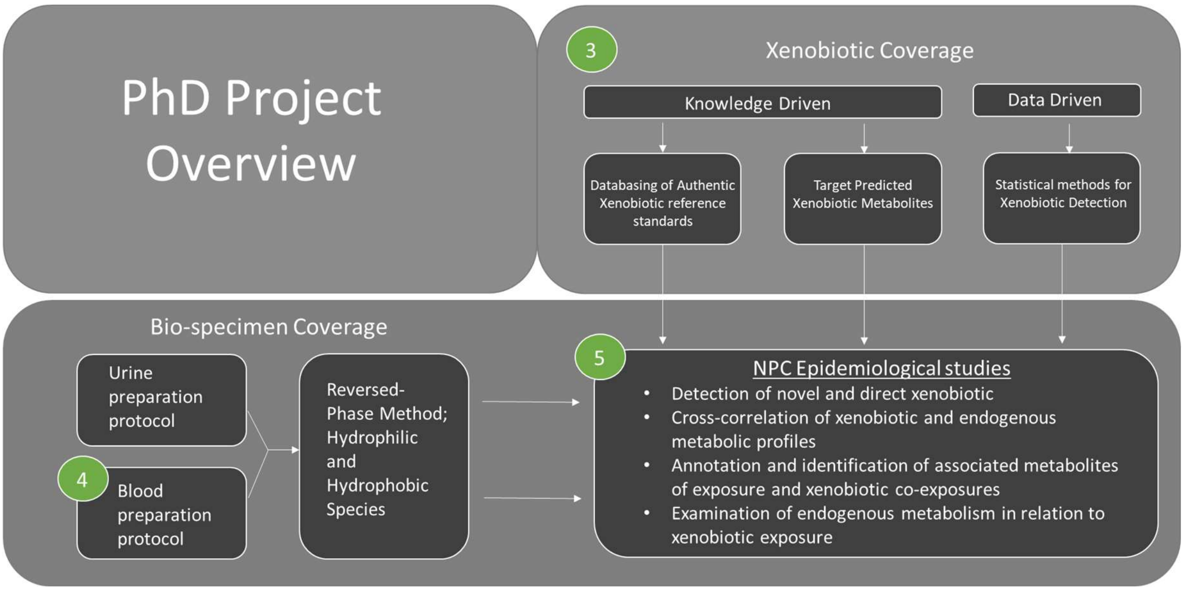

PhD Project Overview
3
Xenobiotic Coverage
Knowledge Driven
Data Driven
Databasing of Authentic Xenobiotic reference standards
Target Predicted Xenobiotic Metabolites
Statistical methods for Xenobiotic Detection
Bio-specimen Coverage
Urine preparation protocol
4
Blood preparation protocol
Reversed-Phase Method; Hydrophilic and Hydrophobic Species
5
NPC Epidemiological studies
• Detection of novel and direct xenobiotic
• Cross-correlation of xenobiotic and endogenous metabolic profiles
• Annotation and identification of associated metabolites of exposure and xenobiotic co-exposures
• Examination of endogenous metabolism in relation to xenobiotic exposure

Figure 1-2. A schematic indicating how the various elements of knowledge-based, analytical, and statistical method development will integrated and applied to a human molecular epidemiological study during the

course of the project. Highlighted by the green circles, Chapter 3 relates to the Identification and prioritization for the acquisition of known xenometabolome components, using authentic chemical standards to provide an empirically-derived spectral database, use of literature and software prediction to search for xenobiotic metabolites in existing metabolic phenotype datasets, and the development and integration of statistical approaches to further identify putative xenobiotic signatures. Chapter 4 relates to the development of a lipid removal method to enable analysis of blood samples using an existing and well-characterised reversed-phase assay for moderately hydrophobic biospecimen components, and Chapter 5, relates to the application of all available strategies to characterise and broaden the coverage of the xenometabaolome in key NPC-relevant sample sets.

1.3 Thesis Structure

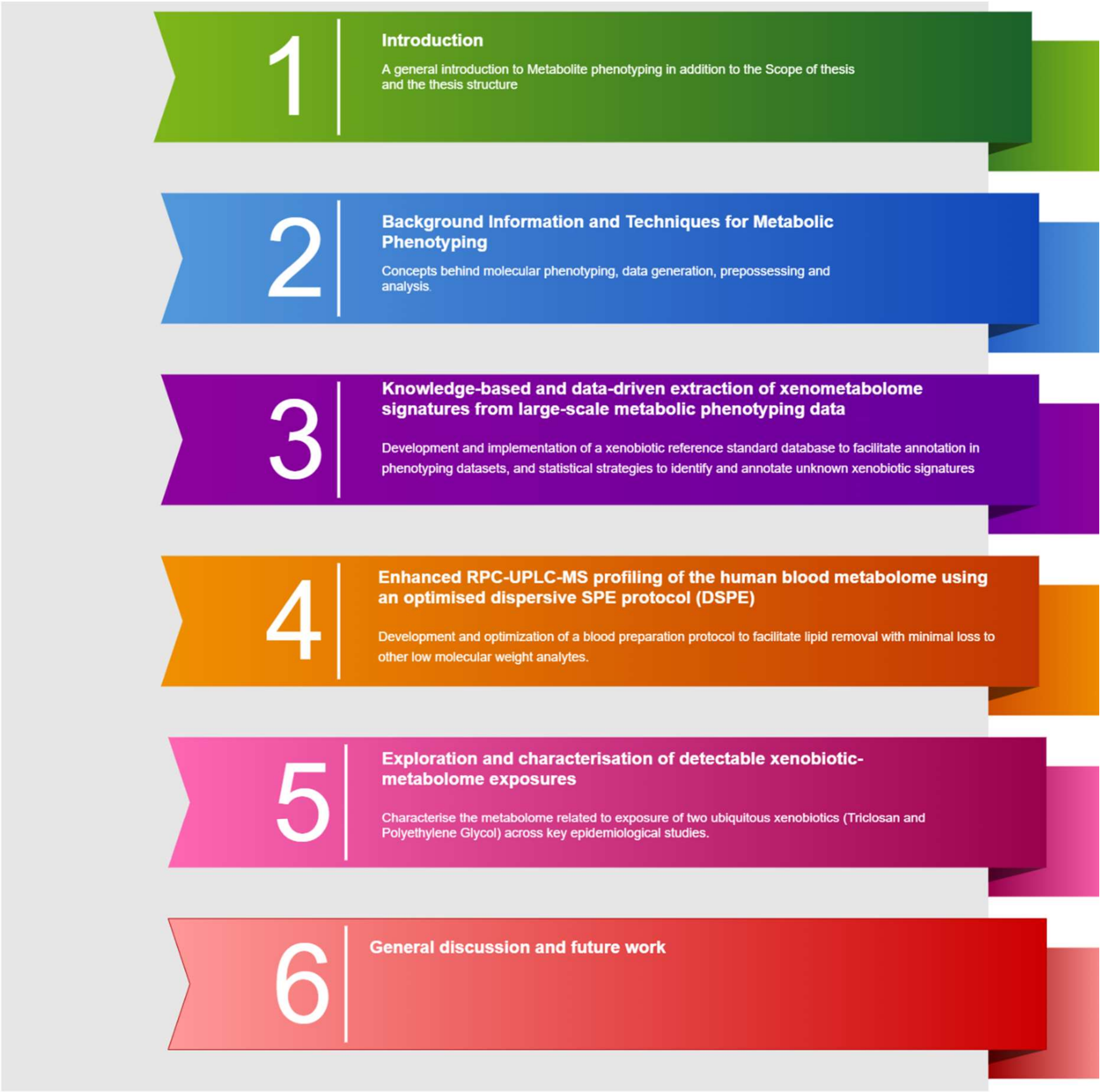

1
Introduction
A general introduction to Metabolite phenotyping in addition to the Scope of thesis and the thesis structure
2
Background Information and Techniques for Metabolic Phenotyping
Concepts behind molecular phenotyping, data generation, prepossessing and analysis.
3
Knowledge-based and data-driven extraction of xenometabolome signatures from large-scale metabolic phenotyping data
Development and implementation of a xenobiotic reference standard database to facilitate annotation in phenotyping datasets, and statistical strategies to identify and annotate unknown xenobiotic signatures
4
Enhanced RPC-UPLC-MS profiling of the human blood metabolome using an optimised dispersive SPE protocol (DSPE)
Development and optimization of a blood preparation protocol to facilitate lipid removal with minimal loss to other low molecular weight analytes.
5
Exploration and characterisation of detectable xenobiotic-metabolome exposures
Characterise the metabolome related to exposure of two ubiquitous xenobiotics (Triclosan and Polyethylene Glycol) across key epidemiological studies.
6
General discussion and future work

Figure 1-3. Schematic of the thesis structure.

Chapter 2

Background Information and Techniques for Metabolic Phenotyping

Summary

This chapter describes the theoretical background and analytical techniques/methods used to perform the work described in this thesis.

The main topics are:

1. Metabolic phenotyping, and its use in epidemiological studies.
2. Xenobiotic metabolism and the contribution of xenobiotics on the human metabolome,
3. Common biological sample types and biological samples
4. Simultaneous capture of endogenous and xenobiotic signatures and the perturbation in endogenous metabolism due to these external exposures is discussed.
5. Design of experiments to support untargeted method development and analytical platforms to characterize the metabolome are also discussed.
6. An overview of data pre-processing methods; these are of critical importance in the processing of MS based datasets. This section also includes a summary of data-driven methods that can be used in metabolite annotation and identification in/using untargeted phenotyping data.
7. Chapter concludes with a brief summary on metabolite identification.

2.1 Epidemiology and Molecular Phenotyping

In recent years, epidemiological studies have regularly incorporated biomarker measurements to augment questionnaire data, clinical records, external exposure assessment, and exposure models to help understand the relationship between gene, the environment, and disease risk. Such measurements can help reduce exposure misclassification and provide individual-level data that can help identify and quantify confounding factors that are not well captured by other means (i.e., in addition to age, sex, socioeconomic factors, and lifestyle behaviours).

Genome-wide association studies (GWAS) can often exhibit low explanatory power as a consequence of environmental (i.e. non-genetic) factors being responsible for much of the attributable disease risk, through initiation and mediation of disease pathways (i.e. gene-environment interactions) (Adamski, 2012). GWAS typically require very large sample sets and can be both expensive and labour intensive to conduct, and ultimately provide information about underlying predisposition for disease, with no incorporation of how genotypes are manifest in a real, complex environment.

Phenotypic responses may prove in explore such relations, due to the necessary mediation of most biochemical processes by low molecular weight metabolites. The measurement of low molecular weight metabolites from complex biofluids and the ubiquitous role metabolism plays in biochemical processes is collectively known as metabolic phenotyping.

Metabolic profiling, metabolomics, and metabonomics are all now variously and interchangeably and has been defined as ”the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification” (Nicholson et al., 1999).

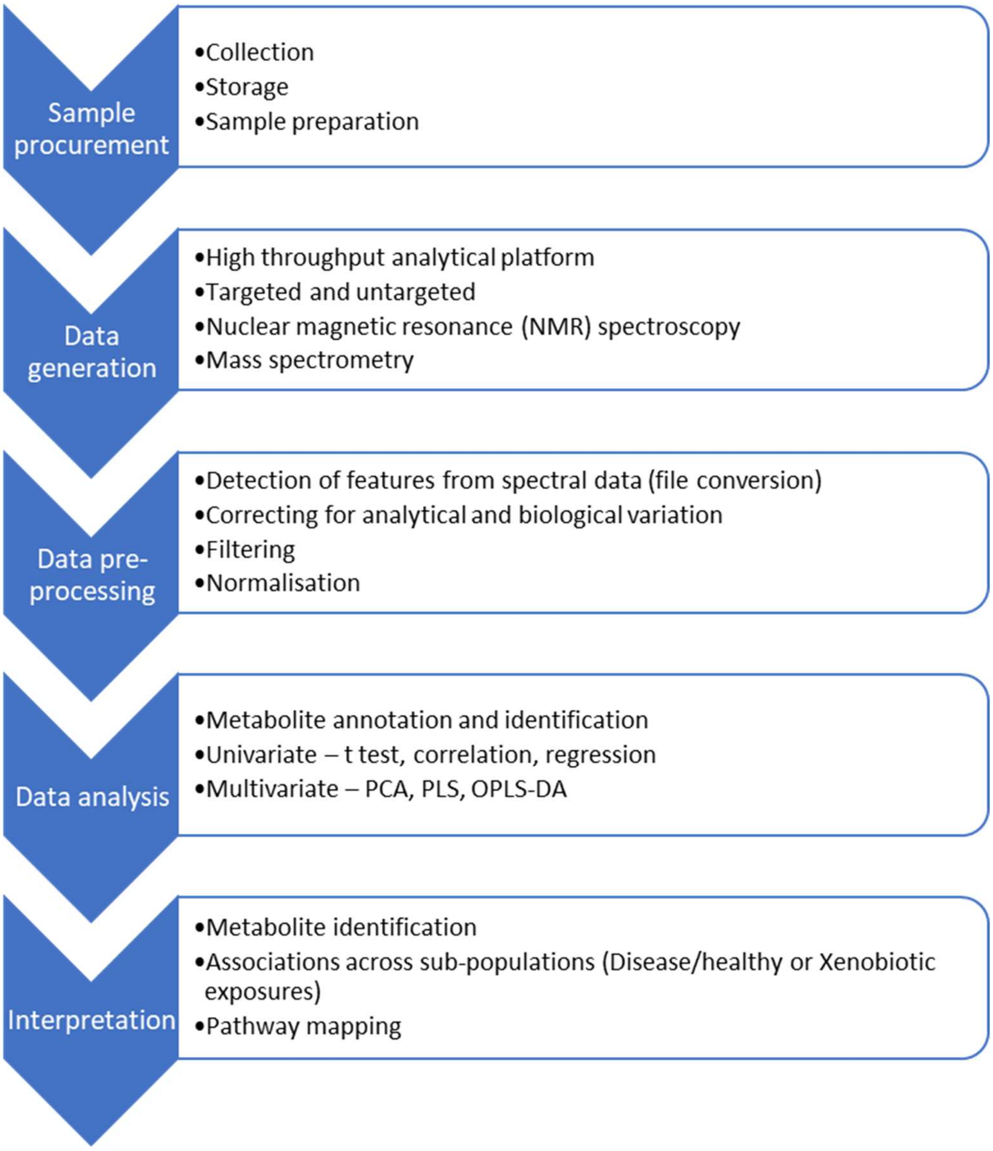
The total complement of these metabolites (typically < 2000 Da) is termed the metabolome, and is a reflection of the sum of the metabolic processes in a cell, organ, system etc. (Tweeddale et al., 1998). The metabolome is a critical component of the wider concept of the human phenome, which relates to all outwardly observable characteristics, and is complementary in nature to the genome. The field of metabolic profiling has been used to report on the complex chemical compositions of not only biological fluids and tissues in mammals, but also application to microbial communities (Donia and Fischbach, 2015), plant (Fiehn et al., 2008) and environmental systems (Bundy et al., 2008). In humans, its definition has expanded and encompasses not just metabolites related to core intracellular and extracellular metabolic processes, but those modified from external exposures such as diet (Holmes et al., 2008), xenobiotics (Marotta et al., 2006) , synthetic chemicals (Wishart et al.,

2012) and microbiome (Nicholson et al., 2005). Exposure of complex mixtures of xenobiotics from various sources detected in metabolomic studies is referred to as the xenometabolome (Holmes et al., 2007) , and collectively the wider exposome, which is a cumulative measure of all environmental exposures associated with biological responses, including lifestyle factors (Wild, 2005, Miller and Jones, 2014).

Typically, metabolomic studies utilise a top-down systems biology approach by describing and modelling biochemical networks that are a result of down-stream responses to perturbations in biofluids. As interactions and metabolite associations are not known *a priori*, the depiction of the metabolome and the observed phenotype can potentially highlight mechanistic pathways (Nicholson et al., 2012). The ability to monitor biochemical changes non-invasively, through typical biofluids measured in molecular phenotyping, makes investigations of this nature a highly valuable approach. The combination of urine, plasma and tissue metabolic profiling can reveal changes in metabolites to different compartments of the body and allows the visualisation of metabolic processes (Waters et al., 2001). Furthermore, the collection at specific compartments and time series will give insight into disease progression and the longitudinal metabolic trajectory of endogenous or exogenous molecules.

As a result, epidemiologists turn to molecular phenotyping as means for broad phenotypic insight as sample cohorts from large scale epidemiological studies generate unprecedented power in describing and modelling the biochemical networks that occur in organisms and an understanding of factors that underlie population disease demography.

To meet this need, there is a requirement for analytical techniques and platforms for broad metabolome coverage, producing high quality data with fast analysis times. Historically, nuclear magnetic resonance spectroscopy (NMR) has been the platform of choice, however there are recognised limitations arising from low sensitivity and chemical specificity (Dona et al., 2014). Mass spectrometry has emerged as a complimentary analytical platform for conducting metabolic phenotyping, and recently, radical advances in both software and hardware components all for the routine collection of high resolution, multidimensional data, with high sensitivity and precision across large sample datasets (Lewis et al., 2016). The diversity of chemical species present in biofluid samples is a substantial challenge; at present, a multitude of different techniques are required to capture metabolites across the vast chemical space of the metabolome (Athersuch, 2012). In addition, the deconvolution of chemical signals from spectral data is also vital and requires advanced analytics and data pre-processing approaches (Lindon et al., 2004). The challenges involved in identifying patterns in metabolites and biochemical pathways, as a result of disease or environmental factors, drive advances in the different key stages in a molecular phenotyping study which is summarised in **Figure 2-1** and discussed in detail throughout this chapter.


Sample procurement

- Collection
- Storage
- Sample preparation

Data generation

- High throughput analytical platform
- Targeted and untargeted
- Nuclear magnetic resonance (NMR) spectroscopy
- Mass spectrometry

Data pre-processing

- Detection of features from spectral data (file conversion)
- Correcting for analytical and biological variation
- Filtering
- Normalisation

Data analysis

- Metabolite annotation and identification
- Univariate – t test, correlation, regression
- Multivariate – PCA, PLS, OPLS-DA

Interpretation

- Metabolite identification
- Associations across sub-populations (Disease/healthy or Xenobiotic exposures)
- Pathway mapping

Figure 2-1. A typical metabolic phenotyping workflow.

2.2 Xenobiotic Metabolism

The term xenobiotic refers to any chemical substance that is not of an endogenous nature, and typically used to refer to drugs, environmental pollutants, contaminants, and other agents. Often, xenobiotic exposures result in toxicity, but these adverse effects are often mitigated by biotransformations that detoxify and eliminate these compounds. The human body has the ability to metabolize and detoxify a wide range of xenobiotic compounds, with the majority of metabolism taking place in the liver (hepatic metabolism) and small intestine (extrahepatic metabolism); a combination of functionalization and conjugation reactions, typically result in an increase in the water solubility (hydrophilicity) or molecular weight/size which facilitates efficient urinary or biliary excretion of xenobiotic metabolites, respectively.

Reactions that introduce or interconvert functional groups within a molecule are commonly known as Phase I reactions and are catalysed by the cytochrome P450 enzymes. The liver is the principal site of drug metabolism, with P450 3A4 and P4502C9 metabolism at this site accounting for ~70% of the metabolism of pharmaceutical drugs (Furge and Guengerich, 2006). Phase I reactions enzymatically modify compounds through hydroxylation, *N*- and *O*-dealkylation, epoxidation, and heteroatom oxygenation *via* biochemical reactions such as oxidation, reduction, and hydrolysis.

The functional groups introduced or uncovered by phase I reactions commonly provide a suitable moiety for conjugation reactions, commonly termed phase II reactions. Conjugation reactions such as these play an important role in the excretion of drug metabolites in that they radically alter the physicochemical properties of the drug (typically through increases in hydrophilicity or molecular weight); conjugation of a drug or its metabolite with hydrophilic endogenous molecules, (e.g., sulphate or glucuronic acid) is catalysed by phase II drug metabolising enzymes including UPD-transferase, glutathione-*S*-transferase, sulfotransferases and *N*-acetyltransferases. UPD-glucoronosyltransferase are used in glucuronidation reactions and approximately 40--70% of all clinical drugs are subjugated to these reactions (Wells et al., 2004). A third detoxification phase has also been documented which involves the transport and elimination of the final xenobiotic metabolite. Membrane transporters carry these metabolites across cellular membranes to the kidneys where they pass through the renal tubular membranes (which acts as a filter), ultimately being excreted through the urine (Kinne-Saffran and Kinne, 1999).

As such, the human body has a highly developed ability to handle xenobiotics, which are subjected to a variety of biotransformation reactions, resulting in elimination through excretion, or to an active or inactive metabolite(s). Elimination by metabolism (to active or inactive metabolites/s), can affect individuals differently, leading to large differences in xenobiotic and metabolite concentrations in biofluids. Currently, the most widely used techniques for studying xenobiotic metabolism and affected pathways, are largely based on *in vitro* incubation assays and *in vivo* radio-tracing experiments (Chen et al., 2007, Mortishire-Smith et al., 2005, Liu and Jia, 2007). *In vitro* assays use cells (primarily liver) from animals or cell lines, which have an infinite lifespan, making it a reliable, simple and relatively cheap procedure. However, the metabolites observed may not always translate *in vivo* due to lack of gene and protein regulation, adsorption, distribution, and elimination (ADME) mechanisms, and also the high concentration of xenobiotic exposure to the *in vitro* assay which is not always reflective of the therapeutic dosed amount. Radio-tracing offers a solution, to observations *in vivo* as techniques involves studying the metabolism route of a radiolabelled xenobiotic compound. Limitations however are cost, and complexity involved in synthesis, high purity needed of the radiolabelled compounds, and containment facilities for use.

The metabolic fate of drugs can vary widely within population groups depending upon an individual's pathophenotype at the time of administration due to both genetic and environmental factors (Maynert, 1961). Phase 1 Reactions that introduce or interconvert functional groups within a molecule (e.g., hydroxylation or carboxylation) and are catalysed by the cytochrome P450 enzymes. Slight variability in the P450 enzymes, i.e., different isoforms of P450, influence how they interact with certain xenobiotics and thus how they are metabolised in the body. The completion of the human genome project brought about the importance of individual variation (polymorphism) observed with xenobiotic metabolism amongst individuals due to the different isoforms of P450 (Gardiner and Begg, 2006).

The gut microflora or microbiome of an individual represents yet another source of variation observed between individuals in the population. Dihydroxylation, decarboxylation, dealkylation, dehalogenation, and deamination xenobiotic biotransformation reactions have been reported as gut microflora mediated reactions. Endogenous compounds such as p-cresol sulfate and indoxyl sulfate have also connection to gut microbes and given that sulfation is a fundamental reaction in Phase II drug metabolism, this also has implications for xenobiotic metabolism (Johnson et al., 2012).

The application of metabolic phenotyping to toxicology and pharmaceutical industries is increasingly being incorporated into drug metabolism research, providing an alternative approach to detect new and uncommon drug metabolites (Chen et al., 2007, Steuer et al., 2019). For example, the Consortium on Metabonomic Toxicology (COMET) generated toxicological databases for the screening of candidate drugs and implemented an expert system for toxicity prediction (Lindon et al., 2003), which included the need to identify drug related compounds (DRCs).

Metabolic phenotyping offers an unbiased approach for metabolite identification relating to drug exposure and thus may augment the study of xenobiotic absorption, distribution, metabolism, excretion and toxicity (ADMET). Without authentic reference materials for xenobiotic metabolites, measurements remain semiquantitative in nature (i.e. relative abundance with reference to parent compound), but integration of metabolite profiling early in the design stages of clinical drug development is sufficiently useful and has been encouraged by government authorities such as the FDA (Ufer et al., 2017).

2.2.1 *The xenometabolome*

The influence of xenobiotics on the metabolome, sometimes referred to as the xenometabolome, (Holmes et al., 2007b), is linked directly to xenobiotic signatures (i.e., xenobiotics and their metabolites) and has not been well characterised, particularly in epidemiological studies.

In large scale population study cohorts, estimates for exposures of individuals to xenobiotics have heavily relied upon meta-data provided through questionnaires. As a result, exposure misclassification (intentional or unintentional), underreporting and failure to remove outlier samples can lead to increased bias and a reduced data set in epidemiological studies. The ability of untargeted metabolic phenotyping platforms to capture xenobiotic signatures alongside endogenous signatures ascribed as responses to these exposures, potentially allows for biomarker-based exposure information to be captured and highlights metabolic pathways that may not have otherwise been seen through targeted biochemical methodologies. Ultra performance liquid chromatography-mass spectrometry (UPLC-MS)-based metabolomics in particular, have successfully been applied to numerous xenobiotic studies and have revealed novel metabolites and pathways (Johnson et al., 2012). In this thesis, both chemical and statistical methods for the efficient detection of xenobiotics and related metabolites, in untargeted metabolome profile data have been developed and evaluated in the context of exemplar studies relevant to the MRC-NIHR NPC.

2.3 Biological Samples

Urine and blood products (typically serum or plasma) are frequently the preferred sample type(s) in metabolomic research for a number of practical reasons, not least because they can be collected in minimally-invasive manner and sample volumes are sufficient to accommodate multiple types of analysis and archiving of sample aliquots. Patient compliance for urine and blood sample collection is high, and both sample type are sufficiently stable to enable self-collection (urine) or; Taken together, these characteristics mean urine and blood products minimal ethical constraints to study directors and bio-archive curators (Bouatra et al., 2013, González-Domínguez et al., 2020, Zikuan et al., 2019).

Metabolites detected in these fluids are subject to fluctuations, which can originate from pathophysiological conditions or external exposures to xenobiotics, diet, and environment.

By utilising knowledge on the biological function of each biofluid and what it pertains to, it is then possible to decide on the more appropriate sample type for a particular metabolic investigation.

On account of their prevalence in large-scale clinical and epidemiological study sample archives, urine and blood products (plasma and serum) are the matrices of choice used for methods development and application of the analytical strategies described; they are the focus of this thesis. It should be noted that the same methodology could be adapted and used for other sample types if required.

2.3.1 *Urine*

Urine is a sterile fluid generated by the kidneys and is comprised of mainly water and low molecular metabolites. The kidney's major role is the regulation of bodily fluids and excretion of water-soluble waste products which are a result of metabolic flux and external. In addition, they filter blood of toxins and undesirable products of metabolism. As urine is primarily water (approximately 95%), the remaining composition belongs to salts and waste products. This is a mixture of urea from amino acid metabolism, inorganic salts (sodium and potassium), creatinine from uric acid and muscle metabolism, ammonia (which give urine its odour) and water-soluble toxins such as xenobiotics (drugs, pesticides and food additives) (Lau et al., 2018). Subsequently, the reabsorption of important metabolites back into blood circulation are vital for certain metabolic functions. Influence from water consumption, nutrient intake and environmental factors (temperature and physical activity), can impact the physical characteristics of urine, such as such as colour, odour, density and pH (Bouatra et al., 2013). Imbalances of any of these characteristics may indicate disease, reflects health status and contribute to the metabolite presence in urine. Colour of urine is usually a mild yellow but can vary due to diet and water consumption. A red colour in urine is an indication of kidney damage or sexually transmitted diseases, due to the presence of red blood cells. Turbidity in urine is a result of crystalline particulates or protein, usually indicating infections or proteinuria. A sweet smell in urine is potentially due to ketones and glucose, which is usually observed with diabetics. As one of the main functions of kidneys, is the internal acid-base regulation, the pH of urine is heavily influenced, with typical readings of healthy urine in the range of 4.6-8. Urine pH can impact xenobiotic metabolism. Ionised substances (acidic or alkaline xenobiotics) will readily dissolve in urine for excretion depending on the pH.

Urine as a by-product of kidney function, and an end-product of pathological cellular processes, provides a highly rich matrix which reflect renal status and more importantly, the biochemical dynamics within a mammalian system. As urine is formed outside the body's circulatory system in the bladder, consequently, metabolite concentrations can build over a long period providing a cumulative longitudinal sample, which reflects the on-going catabolic processes occurring between sampling times. In addition, it is readily available, and least invasive, as it can be collected quite easily, therefore making it well suited for urinalysis for medical diagnosis, mapping metabolic networks and biomarker discovery in metabolomics (Fernández-Peralbo and Luque de Castro, 2012). 24hr samples if feasible, should be conducted to reduce variability in metabolic profiles as a result of change in microbiota activities. Also, first void and mid-stream urine collection is highly recommended to reduce bacterial contaminants. Other analytical advantages urine has over other biofluids include is it is mostly free of protein (in human, but high in rodents and mice), and higher thermodynamic stability of urinary peptides. Storage of urine samples in polypropylene containers are not endowed with special properties or preservatives as seen with blood products. Although special surfactants and antibacterial agents such has borate has been reported. Storage of urine at -80°C immediately after collection and has also been reported to be stable for up to 18 months. Collectively, urine biofluids require minimal pre-treatment and sample extraction for any subsequent urinalysis.

2.3.2 *Blood products*

Blood is a complex mixture of cells, enzymes, proteins and inorganic substances, accounting for approximately 7-8% of human body weight. The liquid portion of blood is plasma, which makes up 55% of total whole blood volume. It comprises of primarily water, red blood cells (erythrocytes), white blood cells (leukocytes) and platelets (involved in clotting). In addition to transporting metabolites, mediators and hormones, plasma is responsible for the transport of oxygen from the lungs to bodily tissues, bringing with it, nourishment to maintain cell life and in turn removing carbon dioxide and waste products of metabolism which inevitably is excreted by the kidney through urine. Blood, once sampled, will clot almost immediately. It clots when a protein in the plasma, known as fibrin, traps and enmeshes the red blood cells. Serum is the fluid component of clotted blood with the clotting process further releasing proteins like proinflammatory cytokines (Schnabel et al., 2010) and low molecular weight metabolites like sphingosine-1-posphate (Yatomi et al., 1997) back into the serum. The composition of both plasma and serum is unique in that it contains a variety of different low molecular metabolites from biochemical processes; be they, energetic substrates, signalling molecules, proteins peptides and lipids with relative differences observed between the two matrices dependent upon factors such as sample collection (Yu et al., 2011) and incubation(Liu et al., 2010), but due primarily to the clotting process present in plasma (Beheshti et al., 1994).

Blood products require a certain degree of sample pre-treatment in order to maintain and preserve the integrity of the measurement of low molecular weight compounds and the analytical platform used. Plasma is obtained by centrifugation of whole blood, and the procedure should be harmonised for high precision metabolomic investigations, especially if collected at different times or sampling locations. Differences in centrifugation protocols can also lead to differences in metabolic profiles in plasma (Lesche et al., 2016). Once centrifuged, it is best sealed in an airtight container after addition of an anticoagulant and a preservative. Serum can contain either a coagulation enhancer or no additive at all. Plasma requires both an anticoagulant and a preservative. The addition of an anticoagulants, such as ethylenediaminetetraacetic (EDTA), citrate, oxalate or Heparin, prevents clotting whereas a preservative such as sodium fluoride, inhibits the growth of bacterial microorganisms. However, these additives can potentially, interfere with certain analytical techniques and prevent measurement of certain endogenous metabolites, for example, certain heparin salts have been known to interfere with its endogenous counterpart and EDTA and citrates can form inactive complexes with calcium (Barton et al., 2010). In addition to the presence of additives for metabolite stability, the other significant factors which can cause fluctuations in metabolite concentrations, are storage temperature and the time of storage. Collection of blood at room temperature and at ice has shown to be susceptible to variation in the observed metabolic profile. It has also been established that plasma is stable in long term storage at -80°C for up to 7 years (Wagner-Golbs et al., 2019).

Being a truly systemic sample (due to homeostasis), plasma or serum (which can sometimes be used as a proxy for blood itself), can provide a snapshot of global metabolism, at the point of collection and so are frequently used in metabolomics (Suarez-Diez et al., 2017). However, studies have reported that the choice of blood product can influence the metabolic profile. Certain lipids (LPS, LPS,), amino acids (arginine, tryptophan, valine, serine and phenylalanine) and glucose have been reported to be higher in serum than plasma, whilst levels of citrate, pyruvate, urate, and lyso-phosphatidylinositol are higher in plasma (Liu et al., 2018).

Comprehensive coverage of chemically diverse metabolites present in human blood products benefits from the use of multiple extraction methods, each oriented toward a small molecule subset generally segregated by polarity and hydrophobicity. Whilst recent developments in LC-MS profiling methodologies have delivered numerous solutions for the analysis of polar molecules (e.g., *via* HILIC-MS) and complex lipids, the analysis of moderately hydrophobic and amphipathic molecules in blood products (including much of the xenometabolome) by RPC methodology, is complicated by the suppressive effects of lipids and proteins on the ionisation of low molecular weight (LMW) metabolites. Efficient and inexpensive sample preparation methods, such as Protein precipitation, liquid-liquid extraction and solid phase extraction have been developed for the separation of small molecules from the remaining sample matrix fit for large scale and high throughput applications.

2.3.3 *Other matrices*

Urine and blood products are typical sample used in untargeted metabolomics, as it offers a non-invasive approach with regard to sampling. Although there are other more relevant matrices which is more informative and offer greater insight on the study at hand.

The central nervous system (CNS) comprises essentially of the brain and spinal cord. The fluid that encompasses and protects the CNS is the cerebrospinal fluid (CSF). CSF protects the CNS by regulating of the intracranial pressure together with cerebral blood flow and excretion of toxic products as a result of cerebral metabolism (Di Terlizzi and Platt, 2006). CSF mainly comprises of water, inorganic salts, proteins and low molecular weight organic compounds. Although it's a highly invasive technique, *via* a lumbar puncture, CSF samples are utilised in metabolomic investigation involving neurodegenerative diseases due to its close proximity to the brain. (Willkommen et al., 2018). Sample collection, storage and even preparation are similar to that of blood products, however studies show that an initial centrifugation step and specific storage conditions, immediately after collection is vital for a cleaner matrix, due to contamination of blood or white blood cells (Rosenling et al., 2011).

Faecal samples offer metabolites which are a final product of cellular and microbial metabolism. It is an ongoing process which occurs in the gut or intestinal tracts of mammals. Faecal samples comprise of metabolites as a result of microbiota bacteria and unwanted waste products from the digestion process and so is largely affected by diet. Studies involving faecal samples have therefore given insight into the interactions between diet, human metabolism and host-gut microbiota status in relation to health and disease (Jain et al., 2019). As faecal sample are a complex matrices, preparation usually involves or a combination of freeze drying, sonication, filtration , solvent extraction and derivatisation (Deda et al., 2017).

The use of tissue samples in studies, although highly invasive, can often be used, as the origin of many disease often stems from the ongoing cellular processes in a tissue sample (Naz et al., 2014). Histological comparisons of metabolic profiles between diseased and non-diseased tissue gives has the potential to highlight the mechanistic pathways associated with disease which may have not otherwise been seen on a systemic level (Wu et al., 2008). Analytical challenges associated with sample prep often lies with tissue inhomogeneity, collection, sample quenching and extraction. Tissue biopsies are a composition of cells in which the intracellular or extracellular contents can be characterised. An example is hepatocytes or liver cells, which *in vivo* – is the primary site for metabolism of drugs. In vitro experiments in metabolomic investigations can be conducted on the S9, cytosolic and microsome fraction each with different enzymes and metabolites, and particularly used in drug metabolism evaluations (Bale et al., 2016).

Finally, cells which can be cultured from three sources; Primary cultures, cell lines and Stem cells, are particularly useful in characterising the biochemical changes in the intracellular metabolome and best indicator of an organism phenotype (Nomura et al., 2011). Quenching, homogenisation and metabolite extraction are typical sample preparation procedures which involves stopping enzymatic processes and releasing intracellular contents for analysis. The processes are now automated however the reagents used in quenching can potentially introduce contaminant metabolites and interfere with subsequent analysis (Zhang et al., 2013).

2.4 Sample Preparation Methods

A primary aim in metabolomic investigations, is the detection of as many metabolites as possible. Initial steps in any investigation firstly involves representative portion of sampling, then the use of cryogenics for storage (store in -80 °Cor freeze drying), buffering and finally metabolite extraction. To span the breadth of chemical diversity present in biofluids, comprehensive metabolome coverage can benefit from the use of multiple extraction solutions, each optimised to target a specific subset of metabolites, segregated by polarity and hydrophobicity. However, the extracted sample should closely represent the levels from the original specimen, thereby ensuring chemical stability and minimising the amount of chemical reactivity which can occur as a result of metabolite extraction. In addition, the type of sample preparation method needs to also be compatible with the analytical platform used for measurement.

2.4.1 *Filtration*

Used with both urine and blood samples, filtration is used as a means to filter suspended particles and cellular components, often immediately after some kind of chemical pre-treatment. Filtration uses a cellulose membrane (0.2-0.45 µm) accompanied with a sodium azide stabiliser which is primarily for preventing bacterial growth during storage. Ultrafiltration utilises special filters of various molecular weight – 3kDa,10kDa and 30kDa are commercially available. These filters are particularly useful for the removal of proteins and other larger macromolecules (Fernández-Peralbo and Luque de Castro, 2012).

2.4.2 *Enzymatic hydrolysis*

Metabolism is the enzymatic conversion of one chemical compound into another. With certain metabolites (endogenous or xenobiotics), can sometimes undergo a chemical change to prevent toxicity. Endogenous compounds include naturally compounds such as steroids and hormones, which are present in urine and blood as glucuronide and sulphate conjugates (Schiffer et al., 2019). Examples of xenobiotics are food additives introduced through diet, and exposure to drugs and pesticides. All which can be present as sulphate and glucuronide conjugates in urine and blood. The majority of xenobiotic metabolism takes place in the liver (hepatic metabolism) and small intestine (extra hepatic metabolism) and is referred to as biotransformation. Enzymes in the liver break down drugs to more polar metabolites through oxidation-reduction reactions (Phase 1) and/or conjugation reactions (Phase 2). This result in metabolites becoming more polar and easily being excreted through the urine. Enzymes such as β-glucuronidase and/or sulfatase are commonly used to hydrolyse glucuronide or sulphated compounds back to the native parent drug or metabolite This offers a solution where conjugated forms may not be commercially available which can hinder metabolic identification efforts. Another reason why hydrolysis is necessary is, conjugated forms may not be retained on RP chromatographic systems, which will impact measurement ability of the analytical method.

2.4.3 *Solvent extraction; protein precipitation (PP) and liquid-liquid extraction (LLE)*

Protein precipitation (PP) followed by centrifugation is the minimum and most often sample pre-treatment method used for either quenching, or the efficient extraction of metabolites, whilst also protein precipitating unwanted protein to biofluids prior to LC-MS analysis. Adequate deproteinization with polar organic solvents include, methanol (MeOH), ethanol (EtOH) and isopropanol (ISP), and acetonitrile (MeCN) have been used to extract mostly polar metabolites whilst non polar solvents, such as hexane, chloroform or methyl tertiary butyl ether are used for mostly non-polar metabolites (Raterink et al., 2014). A combination of both polar and non-polar solvents has been reported for lipids, amphipathic and moderately hydrophobic metabolites. Occasionally, acids can be added to extractions solvents to enhance and stabilise specific compounds such as acyl-Coenzyme A compounds (Basu and Blair, 2011), phosphoric acids for Triglycerides (Izzi-Engbeaya et al., 2018). A major drawback however is that acids can potentially result in an overall reduction in sensitivity and ion suppression. Other protein precipitation conditions which need to be considered are mixing times and temperature parameters. Temperature at 4°C and mixing for 15min minimum are sufficient enough to prevent any biological degradation (Bruce et al., 2009). So, although may not be suitable to capture low-level metabolites, PP is a simple and fast sample preparation technique, which can be beneficial for high throughput offering reasonable metabolome coverage.

The extraction of samples using a biphasic mixture of water, methanol, chloroform or methyl *tert*-butyl ether (MTBE), and the subsequent fractionation to concentrate metabolites into polar and lipophilic fractions, have been utilised to enhance metabolite signals. The three commonly used two-phase liquid extraction for lipid measurements are the Folch, Bligh and Dyer and Matyash methods (Folch et al., 1957, Bligh and Dyer, 1959, Matyash et al., 2008). Folch and Bligh and Dyer extractions result in an aqueous upper layer containing hydrophilic/polar metabolites and a lower organic layer containing mostly of lipophilic species. If the lipid rich organic phase is of interest, this can cause sample contamination issues, as retrieval would mean penetrating through the upper aqueous layer. The Matyash or MTBE method essentially reverses the aqueous and organic phases, thus eliminating this issue, however, has reported to have poorer recoveries for the more polar lipid species (Löfgren et al., 2012).

Irrespective of the extraction method used, complete phase separation is desired to allow efficient partitioning and recovery of metabolites. This is usually achieved by conducting sample preparation procedures at in-vessel temperatures below 0°C and/or the addition of a high concentration of salt to the organic phase (salt-assisted liquid-liquid extraction). The extra steps in the sample preparation process for solvent extraction using LLE mean it is not very compatible with high throughput workflows, and therefor infrequently employed in metabolic phenotyping investigations.

Moreover, LLE preparations often exhibit larger sample volumes, poor selectivity, matrix effects, require the use of glassware (expensive and/or labour intensive) to accommodate solvent choice, and lower analyte recoveries for polar analytes, when compared to PP and SPE protocols (Fiehn et al., 2000) (Kole et al., 2011).

2.4.4 *Solid phase extraction (SPE)*

The use of a combination of PP with SPE, in the form of a disk, cartridge or 96-well plate, has been increasingly used for the preparation of samples in metabolic phenotyping studies. Commonly described as off-line SPE, these formats allow for simultaneous protein precipitation and filtration, whilst also separating target analytes from interfering biological matrix components and enhancing their relative concentrations. In brief, biological samples are loaded onto a sorbent where analytes are then retained based on the affinity to the sorbent, usually *via* van der Waals interactions, or dipole-dipole interactions, hydrogen bonding, or electrostatic forces (Dettmer et al., 2007).

Common sorbents include carbon- or silica-based sorbents (C18 or polymeric silica) and ion-exchange resins. After retention, analytes are then eluted using solvents of sufficient elution strength. SPE can also be specifically used as a clean-up technique for biological extracts. Examples include phospholipid removal SPE plates, solid phase microextraction (SPME) and dispersive SPE (DSPE).

Commercial SPE plates such as OSTRO (Waters Corporation), ISOLUTE (Biotage) and PHREE (Phenomenex) can be purchased for phospholipid removal; these have fast sample preparation and analysis times, are suitable for high through put and in some instances, and often demonstrate higher reproducibility than manual PP (Walter et al., 2001). SPME can be used without the use of solvent and the fibre component can immediately be used on-line and coupled to a separation instrument such as GC or LC (Silva et al., 2011). DSPE is performed by addition of the sample to a sorbent material suspended in a liquid solvent, for extraction isolation or clean-up of specific analytes from complex matrices. DPSE differs from conventional SPE in that the time-based component of the extraction mechanism is removed with interactions occurring at a chemical specific on or of rate (i.e. at a certain point in time, a percentage of the analyte is bound to the sorbent).

Typically, a compromise between comprehensive metabolite coverage and selectivity is required when selecting a sorbent material and eluting solvent. Although SPE is used quite often for targeted metabolome analysis, it is not particularly suitable for untargeted global profiling as there is the potential to remove analytes of interest and introduce contaminants (Simón-Manso et al., 2013, Armirotti et al., 2014). Recently, mixed-mode cation-exchange materials and polymeric resins with weak/strong cation-exchange or anion-exchange sites, in the same sorbent material and online sample extraction have been introduced; these allow for higher analyte recovery on account of multiple retention mechanisms and thus may permit broader metabolome coverage in the extracted sample.(Mitra, 2003).

2.5 Design of Experiments

Appropriate experimental design is critical in any scientific endeavour, including the rational optimisation of assay conditions. Design of experiment (DoE) methodologies utilise a statistical approach to dealing with the complexity involved in the planning and conducting of analytical experiments; DoE is an applied statistical tool that can be used to study and measure the responses and interactions from a number of experimental factors simultaneously, resulting in an optimal and reliable outcome which is both cost efficient and time saving (Jacyna et al., 2018).

DoE provides an alternative to traditional One-Variable-At-Time (OVAT) approach (Barrentine, 2014) that can be used to establish the relationship between two factors; the levels (independent variables) of one factor will vary whilst measuring the changes or responses of the other (dependent variable). If multiple factors are involved, isolating those responsible for the observed effect would be harder to deduce using this approach. Additionally, the independent variables must cover a considerable range to be able to examine the desired effect, resulting in a large number of experiments that must be performed. With respect to optimisation of analytical protocols in which there are a large number of optimisable parameters, OVAT has clear limitations.

Initial steps for using a DoE approach is the specification of the factors which are to be studied and the levels which make up the factor (e.g., particle concentrations or volume of solvent added); preliminary screening tests should be conducted to determine which factors are worth considering and the experimental range of the levels which may influence the desired response.

Once these are set, minimum, target and maximum values are defined across the range in which the responses can be / are likely to be observed. The response must reflect the experimental observation that is to be studied (e.g., the signal intensity of a metabolite before and after a specific intervention), so careful consideration must be taken in determining accurate response parameters. Once the two factors have been defined, a model or design is proposed that can potentially solve the experimental problem (e.g., optimisation of the experimental factors).

If the factors are not well known, screening designs (**Figure 2-2**) can be implemented to determine those that are important, whilst minimising the number of experiments to be conducted (Cavazzuti, 2014). The most common screening design is the full factorial design (FFD), which can be 2 level or 3 level. This allows the entire experimental domain to be explored (as specified during the specification stages) and incorporates the possibility of interaction between factors. Another method is the D-optimal design and is particularly useful when there are unexplored regions in the experimental domain. This happens when experimental parameters are beyond the scope of exploration or unfeasible for other reasons. D-optimal designs can be implemented when optimisation is to be implemented rather than screening. When factors and responses are fully defined with confidence in the experimental ranges specified, optimisation designs (**Figure 2-2**) would be more suited. Another example of an optimisation design is the Central Composite Design (CCD). CCD are especially useful in sequential experimentation (statistical design methods to improve processes when many factors are be studied). It builds on a 2-level full fractional design by adding centre and star (axial) points. This means factors can be set outside their centre settings extending their ranges.

Finally, there is the implementation of statistical models and diagnostics to evaluate the performance of the designs in order to retrieve as much information from the obtained data for high-quality analysis. The data collected by the experimental design are used to estimate the coefficients of the model. The model represents the relationship between the response and the factors. PLS analysis is widely used for data analysis for omics data (Xu et al., 2016) and when the investigation has multiple responses.

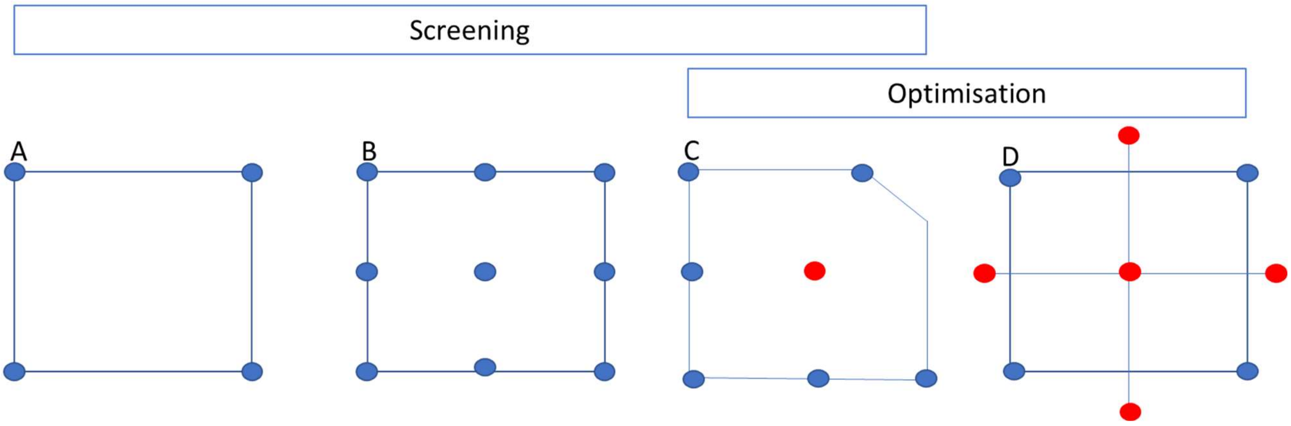


Figure 2-2. Common screening and optimisation experimental designs within a DOE protocol. Screening methods include A – full factorial 2 variables by 2 levels, denoted as 22, B – full factorial 2 variables by 3 levels denoted as 32, C – D-optimal design (2 variables), D – central composite design.

2.6 Data Generation

The ensemble of low molecular weight metabolites present in biofluids reflect the state of important life processes and respond to / are influenced by a variety of different stimuli such as underlying genetics and environmental factors (age, gender, and socioeconomic status).

Key metabolites include energy substrates, signalling molecules, amino acids, nucleotides, sugars, fatty acids, bile acids, proteins peptides and lipid species, and many others.

Collectively, the metabolite in a biological sample (the metabolome) represent chemically diverse range of compounds with a huge range of physicochemical (e.g. size, shape, pH, stability, hydrophobicity, solubility, etc.) (Dettmer et al., 2007). Metabolite concentrations in biofluids and tissues often exhibit a large dynamic range that can span several orders of magnitude (Cao et al., 2020). Furthermore, the combinatorial nature of metabolism means that characterising the metabolome therefore represents a substantial analytical challenge, as no single profiling method can be comprehensive, and it is not feasible to conduct individual assays for all individual known components in such complex biofluids. Consequently, spectroscopic techniques that are well-suited to providing a large number of parallel analytical measurements for biofluids have emerged and in combination offer broad (but not comprehensive) metabolome coverage. As instrumental performance increases so too does the resolution and sensitivity of the metabolome measurements that can be made.

2.6.1 *Analytical techniques in metabolic phenotyping*

The analytical platforms most commonly used in metabolic phenotyping are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS); these provide complementary metabolome coverage and are amenable to high throughput analysis. Furthermore, they both provide complementary structural information, which is valuable in the annotation of spectra and identification of unknown sample components. If applied appropriately, NMR spectroscopy and MS can also provide relative and/or absolute quantitation of metabolites in biological samples. The work conducted is in this thesis is specifically UPLC-MS driven; NMR techniques are not included in detail.

2.6.1.1 Mass spectrometry

Mass spectrometry is an analytical technique involving the filtering and separation of individual components within a sample in an ionised form, thereby measuring its abundance and mass to charge ratio. A mass spectrometer can fundamentally be divided into three basic components: ion source, mass analyser and ion detector.

The ion formation process used to convert sample molecules to a charged or ionised form has evolved throughout the years and is the first step in MS. Common ionisation techniques in which the sample needs to be in the gas phase are electron impact (EI) and Chemical ionisation (CI). Essentially, vaporised sample is exposed to a region of high ionised reagent gas. The gas is ionised by a similar process to electron ionisation where there is the transfer of a proton from the reagent gas to the analyte resulting in the M+1 molecular species. It is considered a soft ionisation technique as it produces very little to no fragmentation information. EI is considered a hard ionisation technique producing molecular and fragment ions. It is the oldest mode of ionisation which works by evaporating a solid deposit on a platinum filament and bombarding with a beam of electrons from a hot wire filament, thereby ionising volatile molecules, a technique which is still utilised today (Allison S K 1952). The need for ionisation sources that would transfer large non-volatile molecules into the gas phase without thermally degrading them was essential and the next logical step. This was the development of Fast Atom/Ion Bombardment (FAB) and Thermospray ionisation which was eventually superseded by electrospray ionisation (ESI) and Matrix assisted laser desorption/ionisation (MALDI). It was achieved by the transfer of analyte species, which is ionised in the condensed phase, to the gas phase. Once formed, analytes can be electrostatically directed to a mass analyser. This meant ionisation can occur with samples coming from an aqueous nature, thereby adding a new dimension to this analytical technique, making it almost routine in many scientific disciplines. ESI is commonly employed to LC as it is compatible with fast flow rates whilst achieving high sensitivity. Sample (i.e. sample and solvent) is pushed through the orifice of a very charged metal capillary. The capillary is held at either a positive or negative charge, thus charging the ionised sample at atmospheric pressure. A carrier gas such as Nitrogen, often referred to as a Curtain gas, flows over the liquid to help nebulise the sample and evaporate any neutral solvent. The sample components emerge in as a Taylor cone, formed by the elongation of the solution at the tip of the capillary consisting of droplets with a charge of between 50 to 70% of the Rayleigh limit, which is the maximum charge a spherical droplet can hold before columbic repulsion overcomes surface tension. As the solvent evaporates, the ions within the initial droplet move closer together and approaches this limit (Rayleigh, 1882). Uneven fission occurs which ejects offspring droplets.

Ultimately, solvent free ions are produced at are passed through to the mass analyser from atmospheric pressure into a vacuum, *via* the MS aperture, which can be polarised with either a positive or negative charge. This can be beneficial for metabolites that have a greater affinity to ionise in a specific polarity. Using polarity switching during or after an analytical run can increase metabolome coverage. More recently, desorption electrospray ionisation, allows for formation of ions at ambient environment outside the MS without the need for any sample preparation (Cooks et al., 2006). A somewhat similar approach is used by the DART (direct analysis in real time) method (Cody et al., 2005).

Mass analysers disperses ions based on mass to charge (*m/z*) ratio by focusing all ions to a single focal point thereby maximising its transmission. For global profiling applications, mass analysers need to be able to; 1) acquire data rapidly over a broad range of *m/z* and operate at 2) high accuracy and 3) high resolution. The first point satisfies most modern UPLC/MS/MS separations. Increases in acquired MS data points which permits better defined LC peak shapes, are observed with most modern UPLC/MS/MS systems with accurate quantitation of a chromatographic peak requiring at least 10 data points per peak as standard over the spectral range. Time-of-flight (TOF) mass analysers can perform up to 500 data points per second over a mass range of 1000 Daltons, and even faster when in tandem with a quadrupole. When coupled to UPLC, peaks can be generated with widths of less than a second. This means as the number of data point increases, so too does the resolving power and therefore the resolution of an instrument. Resolving power and resolution are terms used interchangeably in mass spectrometry, although differ when relating to performance. With TOF mass analysers, the resolving power is fixed throughout the *m/z* range and adopts the valley definition which describes the ability of an instrument to distinguish two adjacent ions of similar mass taken at 10% of the height of the peaks. Resolution (R)adopts the peak-width definition. This is where the resolution of the instrument can be expressed as a function of peak width at a percentage of the maximum peak height at a given *m/z* of a single spectral peak (Urban et al., 2014). Mathematically denoted as the mass m and the mass difference dm resolvable for separation at m:

$$\boldsymbol{R} = \frac{\mathbf{m}}{|\mathbf{dm}|} \qquad \textbf{\textit{(2. 1)}}$$

This is critical when analysing complex biofluids which contains thousands of metabolites of similar mass.

Finally, the third point, is mass accuracy, reported in ppm and is calculated by comparing the theoretical ion (exact mass) to an experimentally measured ion (accurate mass). The quotient is further multiplied by a constant value of 1000000, providing a more convenient integer to report when dealing with instrumentation that can measure at high accuracy. Mass accuracy is therefore reported in units of parts per million (ppm):

$$\mathbf{ppm} = \left(\frac{\textbf{measured ion} - \textbf{theoretical ion}}{\textbf{theoretical ion}}\right)\mathbf{10^6} \tag{2. 2}$$

The exact mass provides additional confirmation on the empirical formula of an analyte whereas the accurate mass is a measurement of an ions mass to within a specified error. High mass accuracy measurement is usually reported to within 5 ppm. Other options for mass analysers include the quadrupole and Orbitrap analysers.

A quadrupole mass analyser comprises of four symmetrically hyperbolic rods arranged in parallel. Ion sorting is based on applying, simultaneously, a constant DC voltage and radio frequency (rf) electric fields between a pair of rods. Scanning is then accomplished by changing magnitude of the rf and DC voltages but keeping ratio constant which results in in the ions undergoing a forward motion. As the ratio of voltages are interchanged, specific *m/z* will be stable and proceed through to the detector. When used as a single *m/z* filter, its offer less resolution than TOF and Orbitrap, and longer acquisition times, as measurement involves scanning a specific *m/z* window over the entire *m/z* range. TOF is most suited for untargeted global profiling and metabolic identification due to their higher mass resolving power, and rapid scanning capabilities which result in covering a wider mass range with higher mass accuracy. In a TOF, packets of ions are extracted into a flight tube, in short ionisation bursts or packets, and subjected to an accelerating voltage (Wolff and Stephens, 1953). Orthogonal acceleration is one of many options to create packets of ions, with ion optics producing either a transverse (90°) or oblique (3° and 10°) drift trajectory (Guilhaus, 1994). Ions are accelerated across identical distances (d) which result in them having the same kinetic energy (E), but velocity (v) is dependent on kinetic energy and mass (mass m is a reflection of *m/z*), and so lighter ions will have a greater velocity than heavier ions and therefore a shorter flight time (t). To increase mass resolution, a reflectron is also added to the TOF analyser. The reflectron is a series of lens which are held at contact electrostatic fields that reflects the ions back to the detector. The more kinetic energy ions have, the deeper they penetrate into the reflectron, therefore taking a longer path to the detector. Ions which are multiply charged will also travel faster than less charged ones. The equation below describes the formula for kinetic energy.

$$\mathbf{E = \frac{1}{2mv^2}} \qquad \textbf{\textit{(2. 3)}}$$

Rearranged,

$$\mathbf{v = \sqrt{\frac{2E}{m}}} \qquad \textbf{\textit{(2. 4)}}$$

The detector measures the time delay between the formation of the packets to when they strike the detector at the end of the flight tube. Combining the very known $v = d/t$ with equation X (above), yields;

$$\mathbf{m = \left(\frac{2E}{d^2}\right)t^2} \qquad \textbf{\textit{(2. 5)}}$$

Accurate measurement of the flight time (which include time for electronic interfaces), with calibration measurements, will correspond to an accurate mass value. Depending on the mass range the analyst wishes to observe, the simultaneous observation of acceleration and detection can be repeated tens-of-thousands of times per second, which can be sufficiently wide enough to transmit and capture low molecular weight compounds. Mass resolution, higher to that observed with TOF's can be achieved with the Orbitrap mass analysers. In an Orbitrap, ions are injected tangentially into an electric field generated by electrodes and trapped, essentially taking the ions off-line. Ions cycle around and along the electrodes which allows ions of a specific *m/z* to oscillate at a frequency which is inversely proportional to the square root of the *m/z*. TOF and Orbitrap mass analysers are compatible with UPLC, however with an Orbitrap, resolution is inversely proportional to the data acquisition rate. This is a factor to be considered for high throughput analysis. Furthermore, to having higher data acquisition speeds, developments in TOF analysers have demonstrated higher sensitivities and are superior in accurately establishing isotopic abundance patterns, which are vital for metabolite identification (Kind and Fiehn, 2006, Rousu et al., 2010). As the combination of high resolution and sensitivity is generally required in metabolic phenotyping, mass analyser operating in tandem is commonly used in place of single filters. A single Quadrupole's as a filter, will offer low resolution and therefore less suited for untargeted profiling. However, when in tandem, offer greater selectivity and sensitivity, which can be significant in both qualitative and quantitative applications. Common configurations are the triple quadrupole (TQ), and the quadrupole-TOF mass analysers (Q-TOF). A common method for fragmenting molecules involves the use of collision induced dissociation (CID). The ions are accelerated from one analyser to the collision cell where it collides with a neutral gas (often argon or helium). The kinetic energy on impact equals the internal energy of the covalent bonds, resulting in vibrations and ultimately cleavage of the bonds which produced characteristic molecular fragments in a reproducible manner. Whether it is a TQ or Q-TOF, the first quadrupole in the series of analysers, provides pre-selection of molecular ions, that is to be passed on for fragmentation. The TQ allows for multiple scanning modes, precursor ion scan, neutral loss scan, product ion scan and multiple reaction monitoring. MRM boasts the highest sensitivity and wide dynamic range of detection. Here, quadruple 1 (Q1) is fixed to filter a specific precursor ion, whilst the second quadrupole (Q2) is used as the collision cell to produce product or daughter ions specifically selected by quadrupole 3 (Q3). This is particularly useful in metabolite identification and quantitation applications. For routine profiling applications in this thesis, a Q-TOF was used, wherein the quadrupole and collision cell were selectively disabled allowing for ions to pass through to the TOF mass analyser for measurement.

2.6.1.2 Liquid chromatography

A chromatographic system consists of a moving phase and a stationary phase in contact with each other, where a specific sample can be separated into individual parts. Changing the nature of these phases can create different forms of chromatography. One form that is finding increasing utility in metabolic phenotyping investigations, is liquid chromatography. Its moving phase is a liquid that is pumped through a column filled with fine solid particles. The surfaces of these particles are chemically treated with an adsorbent material. Silica gel and Alumina are amongst the most popular adsorbents used. As the liquid moving phase (mobile phase) is pumped through the column, different components of the sample are separated and retained, based on molecular structure and/or polarities, to various degrees, depending on their affinity with the stationary phase or mobile phase. This leads to a separation of different components making up the sample mixture. The time taken for individual components to elute from a column is known as the retention time.

Liquid chromatography can be utilised in a number of different ways depending on an analytes' physico-chemical properties. The two main types of liquid chromatography which can found and applied to all fields are RPC and Normal phase. Both modes of chromatography are based on polarities of the sample and its interaction with the mobile and stationary phases. In RPC mode, the mobile phase is either a pure polar solvent or a mixture of various polar solvents while the stationary phase is a nonpolar solid or a liquid (Aygun and Ozcimder, 1996). Polar compounds elute first followed less polar compounds. In normal phase, a polar mobile phase is chosen, such as n-hexane. The stationary phase can either be silica; alumina or modified silica which is generally called bonded phases, such as Si-CN or Si-NH2 (Abbott, 1980). The opposite is observed here where, polar compounds will elute later (longer retention times). Hydrophilic interaction liquid chromatography (HILIC) is similar to normal phase with the main difference being the use of water immiscible organic solvents as the mobile phase. Polar molecules are well retained and elute in order of increasing hydrophilicity (Li and Huang, 2006). Normal phase or HILIC can be run complimentary to RP as polar molecules on a RP system will have very fast retention times resulting in co-elution of many peaks, running normal phase or HILIC will enable high resolution and separation of these types of compounds. Other forms of liquid chromatography include ion exchange and size exclusion chromatography. In exchange chromatography consists of two types, cation exchange, in which the stationary phase carries a negative charge, and anion exchange in which the stationary phase carries a positive charge. Charged molecules in the liquid phase pass through the column until a binding site in the stationary phase appears. The molecule will not elute from the column until a solution of varying pH or ionic strength is passed through it. Ion exchange chromatography is commonly used in the purification of biological materials (Cummins et al., 2011). Size exclusion chromatography, does not involve any adsorption and is extremely fast. The technique uses porous gel which allows separation of larger molecules which cannot penetrate the pores to elute first. This method is common in protein separation and purification (Irvine, 2001). Another separation technique used in liquid chromatography can be achieved by switching from an isocratic elution (same eluent throughout) to a gradient elution. This is done by mixing two or more different eluents, such that the mobile phase composition changes over time. This gradient elution offers the most complete separation of the peaks, without taking an inordinate amount of time. A sample containing compounds of a wide range of polarities can be separated by a gradient elution in a shorter time period without a loss of resolution.

It is this selectivity, fast analysis times, low solvent consumption and increased sensitivity that liquid chromatography has come a long way from its early days as a form of partition chromatography to what it is today. The introduction of ultra-performance liquid chromatography (UPLC), which permitted the use of smaller column particles in combination with high flow rates to provide fast and efficient separations, has been particularly beneficial in metabolome analysis, and affords vastly superior results in a fraction of the analysis time. (Wilson et al., 2005).

2.6.1.3 Hyphenated and parallel analytical platforms

The high sensitivity (ng, pg levels) of MS detection makes it an important method for measuring metabolites in complex biofluids. Mass spectrometry for profiling applications typically uses a high-resolution measurement system, which aids in the high specificity of the technique where the accurate molecular mass is well distributed across the detectable range. A high-resolution measurement system that enables accurate mass measurements allows for the elemental composition of many chemical compounds to be determined and the ability to distinguish between isobaric compounds.

Although the ionisation process required to form ions for mass spectrometry can be susceptible to ion suppression or enhancement, these effects can be minimised *via* the use of chromatographic techniques. When hyphenated to a chromatographic technique, the ability to acquire data in three-dimensions, i.e. retention time, mass to charge and intensity, results in a highly sensitive and selective technique which reduces the influence of co-eluting matrix components and therefore makes it a complementary platform to NMR.

Combing strengths from multiple analytical techniques, can aid in structural elucidation, and therefore the identification of unknown analytes which creates an opportunity to broaden metabolomic research (Whiley et al., 2019, Bhinderwala et al., 2018). The work in this thesis focuses on data derived only from UPLC-TOFMS systems (**Figure 2-3**) which have been set up for phenotyping applications at the NPC.

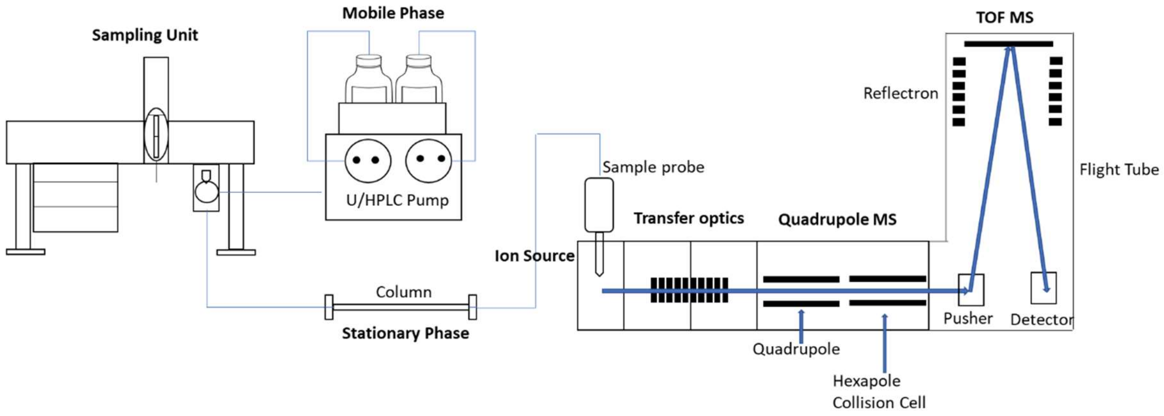


Figure 2-3. Typical schematic of a U/HPLC system coupled to Q-TOF mass spectrometer.

2.7 Pre-processing of Spectroscopic Data for Metabolic Phenotyping

2.7.1 Sources of variation

Advances in LC-MS allows for the separation and detection of thousands of metabolites in biofluids. The variation in metabolite concentrations and the ubiquitous role metabolism plays in biochemical processes are crucial in metabolomics.

The performance of all analytical devices are prone to the effects of environmental and sample conditions, and it is common practice to ensure any drift in performance is characterised so that assay quality can be determined, and, where appropriate, adjustment made to ensure data are comparable. Compared to other analytical devices, LC-MS instrumentation often exhibits considerable systematic over the course of sequential analyses within a batch, and inter-batch variability. This variation predominantly affects retention time (drift) and signal intensity (sensitivity), discussed below.

2.7.2 Batch effects

Often with metabolomic datasets, studies are split into batches to preserve the integrity of the components that make up an LC-MS system, thereby reducing the analytical variation observed with the instrumentation. Whilst cleaning, conditioning, calibration, randomisation of samples and limiting the number of samples within an analytical batch can mitigate some of the analytical variation observed, the analysis of larger sample batches, will not account for all of it. Analytical variation can arise from a variety of different sources primarily due to sample interaction with the analytics of the instrumentation. Such sources include, variation and gradual contamination of the stationary phase, variability in mobile phase preparation, e.g. pH, fluctuations in electrospray ionisation and gas flows, aging detector, changes in the ambient temperature and random imprecision in measurement, all of which can result in nonlinear retention time drift, and drift in detector response (Podwojski et al., 2009, Watrous et al., 2017). Therefore, there are ways to minimise and reduce variation specific to LC-MS.

2.7.2.1 Retention time drift

To account for retention time drift, several software packages exist, and have been implemented for pre-processing of untargeted metabolomic LC-MS datasets (Libiseller et al., 2015, Smith et al., 2006, Katajamaa and Orešič, 2005, Pluskal et al., 2010, Scheltema et al., 2011, Lommen, 2009). Proprietary software like Progenesis QI or open source software like XCMS (Tautenhahn et al., 2008, Smith et al., 2006), are such packages. Fundamentally, the software is a collection of algorithms that expedite the tedious but highly intuitive process of detecting, grouping, and aligning of LC-MS signals (herein referred to as *features* or *variables,* comprising of retention time, *m/z* and signal intensity) across multiple samples. Simply, peak detection involves locating spectral features that appear to exhibit a Gaussian distribution. Peaks over the *m/z* range are then searched across the chromatographic retention time range. All detected spectral features from a selected sample, are used as an alignment reference to which all corresponding features from all samples are corrected based on similarity in *m/z* and retention time.

2.7.2.2 Signal drift

Drift in signal intensities require different strategies for correction. This can be addressed at various stages of the metabolomic workflow, ranging from how samples are prepared, to during the sample acquisition process, and finally on the already acquired data. The spiking of reference or internal standards and then normalising to the intensity observed, is common practice in analytical experiments and can be used to correct for variability due to analyte loss in sample storage or extraction. Internal standards for relative retention time calculations can also reduce the uncertainty due to retention time drift and variation. There are limitations however to this method. Standards need to not interfere with metabolites and preferably be naturally occurring to be somewhat representative to the class of metabolites which care being studied. Usually, stable isotopes or structural analogues are used. However, an untargeted metabolomics experiment can result in the detection of thousands of metabolites, the type and number of internal standards used may not be representative and spiking many standards to account for this can be expensive and labour intensive (Ejigu et al., 2013).

Recently, the application of a detector gain control which increase gain response based on instrument performance have proven successful in achieving robust and reproducible measurements on the raw data without the need for post normalization or informatic correction (Lewis et al., 2016). The magnitude of the voltage that is applied to the detector in a waters UPLC-MS, is adjusted accordingly to account for any changes in the background chemical noise acquired for each sample in an analytical run. This therefore provides a means to adjust and minimise drift or fluctuations which have been affected due to analytical sources of variation. In addition, there is the added benefit of correcting signal intensity during sample acquisition which would have otherwise been loss and no amount of correction post analysis would have been able to restore.

Post-acquisition correction for signal drift utilises methods which are feature specific. Such a method is the quality control based robust local regression (LOESS) signal correction (Dunn et al., 2011). LOESS regression improves the overall precision in the data by eliminating the longitudinal drift observed within an analytical batch. LOESS regression applied in this thesis is an adapted version of the LOWESS approach proposed by Dunn et al. At each point in the data, a polynomial is fitted to a subset of the data using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. The smoothing parameter then determines how much of the data is used to fit each polynomial, in this case, the default window for the LOWESS smoother is 11 QC samples. So, for each MS feature, the LOESS estimator is fitted every 11th QC sample within the analytical batch to avoid overfitting. In this implementation, the LOWESS estimator is a cubic spline function which is fitted to the SR samples. Next, the value for each feature in a sample is corrected by dividing the original intensity value to that of the interpolated value given by the cubic spline fit at its position in the run order. The same procedure is undertaken with the QC samples and then finally the drift in the signal is calculated by taking the ratio between the study and the closest QC sample in the analytical sequence. To ensure features detected in the study sample are present in the QC sample, a pool of all study samples is usually recommended for this analysis thus reflecting the average metabolite concentration and is representative of the entire dataset. Pooling of samples from multiple analytical batches may also serve this purpose and can be done in-house or commercially bought. Commercially available QC samples can be found for matrices like urine and blood but can prove difficult for other biofluids. The application of this LOESS regression method and identical QC samples acquired regularly, will not only correct for signal drift within batches, but also between batches, allowing multiple batches to be merged, which can often be the case for large MS based metabolomic investigations.

2.7.3 *Biological variation*

Fluctuations in signal intensity can also stem from the complexity of metabolites associated with biofluid. Most matrices, such as plasma, serum, and cerebral spinal fluid, are physiologically controlled (Knepper et al., 1989). The diversity of urine however can exhibit large biological variation and is primarily a result of differences in concentration of metabolites which can be due to several factors. Natural differences in abundance between metabolites are common. Signal metabolites and metabolites as a result of central metabolism are generally lower or remain constant like Creatinine, whereas metabolites such as ATP or P-Cresol are generally larger in magnitude under identical experimental conditions. But not only that, certain metabolites can exhibit large fold changes between individuals based on physiology, genetics and environment. In urine biofluid, it has been reported that there can be a 20-fold change in water dilution (provide there is no renal impairment) which would significantly impact the concentration of metabolites (Yamamoto et al., 2019). In addition, there are also disease and pathophysiological factors and exposure confounders such as nutrition, medication and diet, that can influence the concentration of metabolites (Tsuchiya et al., 2003), further complicating interpretation and comparison of samples in an metabolomic investigation. The important thing to note however, is that high concentration metabolites are not any more intuitive than low concentration metabolites.

2.7.3.1 Creatinine and osmolality normalisation

The most reliable way to account for differences in metabolite concentrations in urine, is the collection of a 24hr sample which is planned in the study design (Warrack et al., 2009). This however can be a tedious process for both patients and researcher, so one collection submitted to the metabolomic study is often the case. As a result, signal variation in the acquired data tends to be heteroscedastic in nature, arising from both analytical and biological sources, and so normalisation methods can be applied post sample correction, to correct for signal intensity differences and stabilise the variance observed (van den Berg et al., 2006). One such way is normalising to the endogenous metabolite Creatinine (Alberice et al., 2013). Creatinine in urine takes on the assumption that there is a constant excretion of the metabolite and if a quantitative measurement is available, one can normalise by dividing the intensity for each feature to either the intensity of the ion relative to Creatinine or to the concentration quantified in NMR. Limitations of this method is that the assumption holds through provided there is no kidney impairment associated with the study, and creatinine in urine has been shown to have a minor age dependence (Gu et al., 2009).

Osmolality is another correction used to normalise urine measurements, and is largely unaffected by age, gender, diet and general health (Chadha et al., 2001). It is a measure of the concentration of solutes in biofluids and so is representative of the total metabolite output observed in urine matrices. Often used as the golden standard of estimating urinary concentration, osmolality determination is not always available in clinical laboratories, and so other clinical tests, such as freezing point depression or specific gravity, are used as an estimate (Chadha et al., 2001). These two clinical tests highly correlate to urinary osmolality, and readings are used and normalised in the same way as Creatinine measurements.

2.7.3.2 Statistical normalisation

In addition to these methods, are exclusive statistical approaches like MS total useful signal (MSTUS), median fold change (MFC) and probabilistic quotient normalisation (PQN). MSTUS forces all samples to have an equal total intensity, dividing the intensity of a feature to the sum of all features detected in a given sample (Warrack et al., 2009). The methodology is similar to that observed in proton NMR-based metabolomic analyses (Craig et al., 2006). Features considered for this normalisation must be present in all samples and although works well with the majority of stable features, metabolites which are large and variable in intensity, as in the case xenobiotics (which are the result of external exposures), can compromise the normalisation. PQN operates under the assumption that changes in metabolite concentration, as a result of urinary dilution, affects the entire profile whereas biological changes only affect parts of the profile. It was originally proposed by Dieterle *et al* (Dieterle et al., 2006) and applied to NMR data. Briefly, an integral or total area normalisation is firstly conducted in order to scale the data to approximately the same magnitude. After which, a quotient for each variable, between the reference sample and all study samples are calculated. The reference sample is the gold standard reference spectrum, either from a database or mean spectrum of all spectra in a study. Next, the median of the quotients is calculated, and finally all variables are divided by its specific median. A similar principle was applied to MS data but labelled as MFC (Veselkov et al., 2011) and a similar assumption made where peak intensities are directly proportional to metabolite concentrations so any changes in the overall profile of a sample, either by urinary dilution or fluctuations in the analytics of the instrumentation, would result in the same change and is linearly proportional to individual spectral features. For this normalisation, the data is rescaled by adjusting the median of the log fold change in peak intensities. As a result, all variables are distributed around zero. The major differences however are; a) there is no prior total area normalisation; and b) the reference sample can be a random sample selected from the study or a calculated median value from all samples for every given feature. Normalising this way, does not address or remove analytical drift which is observed with LC-MS acquisition. This can be an issue with large analytical batches and so a correction prior to MFC would be required. For all data analysis involving urine datasets in this thesis, LOESS regression for analytical drift correction is firstly applied, then a filtering protocol for the selection of high quality features which are shown to be measure accurately with respect to intensity, i.e. scale with dilution and reproducibility of features in the QC sample (measure by relative standard deviation) (Sands et al., 2019), and finally PQN to address biological variation.

2.7.4 *Feature quality (filtering)*

A deficiency of the pre-processing software packages is its inability to perform any advanced feature filtering. Filtering aims to reduce the spectral complexity observed with LC-MS data by reducing and removing non-relevant features and are usually applied post normalisation to ensure data quality. Features which are reported need to be considered as "real", i.e. features which are of low abundance against features which are a result of chemical noise and artefacts from the analytical system, or artefacts from the feature extraction process. Open source software such as XCMS incorporate basic filtering algorithms such as Minimum fraction filters to look for valid features present in a minimum number of samples within a sample group. The downside to this is that xenobiotics or metabolites detected in only a handful of samples, maybe filtered out prematurely. The Minfrac setting used for all project data in this thesis is set at 0.4. Not all software packages utilise a similar algorithm such as Minfrac, and so prospective filtering techniques are needed, i.e., the measurement of relative standard deviation (RSD) on each feature *via* the use of pooled QC metrics at repeated injections throughout the analytical run and in addition, a dilution series based on these QC samples, as a measurement of correlation to dilution. This method for filtering is applied for all datasets used in this thesis (Lewis et al., 2016).

2.8 Analysis of Metabolic Phenotyping Data

Metabolic phenotyping platforms ability to measure complex biofluids containing of thousands of different metabolites, would typically result in datasets where there are more variables than samples (Posma, 2019). As such, computational approaches are required to handle such large datasets, but also the correct use of statistical analyses is required to extract meaningful information and interpret the results in a biological context. When finding associations between spectral features (variables) among all samples (observations), classical univariate tests, such as t-tests or correlation analyses can be used. However, with such an elevated number of features typical of LC-MS based analyses, the repetition of any of these univariate tests increases the chance of false positives (FP). Although, multiple testing techniques can be used to reduce such issues, multivariate approaches can expose shared variable associations and are well suited in molecular phenotyping. The application of multivariate analysis accounts for reduction in spectral complexity, compensate for multicollinearity and to help visualise and identify patterns and similarities (clustering) between observations.

2.8.1 Univariate data analysis

2.8.1.1 Two sample t-tests

The two-sample t-test is a bivariate analysis with the ability to differentiate the means between two groups of sample data. It is different to a one sample t-test which compares the mean of the entire population to that of a theoretical value. The two-sample t-test can be either paired or unpaired. An unpaired test involves comparing means from two independent samples sets, whilst a paired t-test compares means of two related groups of sample sets. The dataset will dictate the appropriate t-test method, as certain methods come with certain assumptions. Traditional two sample t-test assumes data is continuous, normally distributed and with equal variances. Welch's t-test is appropriate for unequal variance but still assumes a normal distribution. The Wilcoxon-Mann-Whitney ranks the data prior to calculation of the t-test. It therefor allows for non-normally distributed datasets. The absolute value of the calculated t-test statistic using either of these methods, can then be used to determine if the difference is significant. If the level of significance or p-value used for the t-test is smaller than the cut off value, then it fulfils the hypothesis that there is no significant difference between the specified groups.

2.8.1.2 Correlation

Correlation is a bivariate analysis and a measure of the strength and direction of the relationship between two continuous variables. It is a widely used inferential statistical procedure used in multiple disciplines. Correlation is a measurement of the covariance. Covariance relates to how variables change with each other and uses only the sign to indicate the direction of the relationship. If the covariance is positive in a bivariate analysis, it means both variables increase together. If negative, it means as one variable increases, the other decreases. Correlation gives more information than covariance by also describing the strength of the association by the magnitude of the correlation. A high positive correlation indicates a strong relationship between two variables. A high negative correlation indicates a scenario where the variables move in opposite direction and so the increase in variable 1, is associated with a strong decrease in variable 2. The number associated with the correlation statistic is referred to as the correlation coefficient and is a result of specific correlation method. Popular methods are the Pearson and Spearman methods (Mukaka, 2012). The correct usage of correlation coefficient type depends on the types of variables being studied.

Pearson correlation coefficient measure the linear relationship between two variables and assumes that both variables follow a normal distribution. For a correlation between variables x and y, the correlation coefficient, denoted by r, has the following formula for a Pearson correlation:

$$\mathbf{r_{Pearson}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i} - \mathbf{x})(\mathbf{y_i} - \mathbf{y})}{\sqrt{\left[\sum_{\mathbf{i=1}}^{\mathbf{n}}\left(\mathbf{x_i} - \overline{\mathbf{x}}\right)^{2}\right]\left[\sum_{\mathbf{i=1}}^{\mathbf{n}}\left(\mathbf{y_i} - \overline{\mathbf{y}}\right)^{2}\right]}} \qquad \textbf{\textit{(2. 6)}}$$

where x_i and y_i are the values of x and y for the ith individual. The method is not robust to outlier samples and can falsely give a high correlation. A spearman rank correlation, however, does not carry any assumptions on the normality of data and is more robust to outlier samples. The method is used when changes between two variables occur at the same relative direction, but are not necessarily linear, i.e. when variables have a monotonic relationship. Unlike Pearson, the Spearman correlation coefficient is calculated based on ranked values of the variables rather than on the raw data and can be used for both continuous and ordinal variables. Briefly, every sample within the variable is given a rank score (e.g. 1 is the highest rank and 10 is the lowest rank based on a 10-sample dataset), next, the difference between each rank for that observation is subtracted and squared. The addition of these squared values, denoted by $\sum_{i=1}^{n} d_i^2$, is calculation for the Spearman correlation coefficient in the formula:

$$\mathbf{r_{Spearman}} = \mathbf{1} - \frac{\mathbf{6\sum_{i=1}^{n} d_i^2}}{\mathbf{n(n^2-1)}} \qquad \textbf{\textit{(2. 7)}}$$

Relationships using correlation coefficients determines associations not causal relationships. In the context of metabolomics, high pair-wise metabolite correlations may reflect metabolites that potentially belong to the same pathway of a metabolic network (Fiehn et al., 2000). Chemometric methods have been developed, such as statistical total correlation spectroscopy (STOCSY), which takes advantage of the multicollinearity and high correlation that maybe observed between spectral features arising from the same pathway, molecules with a common structure, and xenobiotics and direct metabolites (Cloarec et al., 2005, Crockford et al., 2008, Holmes et al., 2007). Statistical heterospectroscopy (SHY) extends this approach to provide correlation linkage across datasets produced by two different analytical spectroscopic platforms (Crockford et al., 2008).

2.8.1.3 Linear regression

A simple linear regression model assumes a linear relationship between one predictor input variable (independent) to a specific response or outcome (dependent), with both being a continuous numeric value. When there is more than one input variable, linear regression then models the relationship between the multiple variables to their outcomes. This is known as multiple linear regression (MLR).

The fundamental equation of a standard generalised linear model with one variable is denoted by the following equation;

$$\mathbf{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathbf{x} \qquad \textbf{\textit{(2. 8)}}$$

And with multiple variables;

$$\mathbf{y_n} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{x_1} + \boldsymbol{\beta}_2\mathbf{x_2} + \dots + \boldsymbol{\beta}_n\mathbf{x_n} \tag{2. 9}$$

Where y is the response variable and x_n are the predictor variables. The equation assigns regression coefficients, denoted by the Greek letter beta β, and represents the strength and influence of each variable to the response variable. In order to produce (or train) an equation and fit a regression line to the data, the most common method used is ordinary least squares or least squares regression. The method works by implementing a line (linear model) or regression plane (multiple regression model) that best fits the data. The algorithm is mathematically described as the minimum or least sum of squared residuals and as a result, fits a line that constitutes the smallest vertical distance, also referred to as errors or residuals, between each observation to the regression line or plane. Now that a model is constructed based on the data, we can use this model to make predictions on the response variable based on values of the predictor variables with new observations. In addition to predicator variables, are confounders. These are variables that share an association with both the response and predictor variables. Failure to identify these variables can distort the influence of the other predictor variables to the model. Age and gender are frequent confounders.

2.8.1.4 Logistic regression

Logistic regression (LR) (Cramer, 2002) is another regression type that utilises discriminant algorithms i.e. two classes or groups. LR utilises a classification algorithm by predicting the probability of a particular outcome where the response variable is binary or dichotomous, given a set of predictor variables. It therefore differs from linear models which assumes that the response Y is continuous, and predictions of the Y output is numeric ranging from negative to positive infinity. With logistic regression, the probability of an outcome is confined to values between zero and one and fits a line that best separates the two classes. However, if a linear model was projected onto a dataset with a binary outcome, it would be difficult to interpret. In addition, outlying samples can significantly skew regression lines. A way around this is the use of a transformation logistic (or logit) function that coverts Y to lie on the interval -infinity to infinity, also called the log(Odds):

$$\log(\mathrm{Odds}) = \log(\frac{p(y)}{1 - p(y)}) \qquad \textit{(2. 10)}$$

p is the expected probability that an outcome is present and so can be read as the ratio of the probability that an observation falls into a particular class, p*(y),* divided by the probability that the same observation is not a member of the class, $1 - p$*(y)*. Log transformation of the Odds, therefore, turns the Y variable from binary to continuous and is modelled to resemble a multiple linear regression equation:

$$\log(\frac{p(y)}{1 - p(y)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad \textit{(2. 11)}$$

x_1 through to x_n are the predictor variable and β_0 through to β_n are the regression coefficients. As the y-axis now ranges from +infinity to -infinity, the use of residuals and least squares, as seen in linear models, cannot possibly apply. Instead, LR uses maximum likelihood. The process that starts off with a candidate log(Odds) line, where each observation has a specific log(Odds) value. Transformation back to probabilities, occurs *via* the following equation:

$$p(y) = \frac{e^{\log(\mathrm{Odds})}}{1 + e^{\log(\mathrm{Odds})}} \qquad \textit{(2. 12)}$$

Graphically, the line takes a sigmoid shape. The likelihood (y-axis) for each observation can now be extrapolated from the sigmoid line, log transformed and added (likelihood = p*(y),* if it belongs to the class, and likelihood = $1 - p$*(y)*, if it doesn't). The sum is the overall log-likelihood of the original candidate log(Odds) line. This is an iterative process where the algorithm projects another log(Odds) line, but pivoted in a way where the loglikelihood is higher in magnitude. The optimal logistic regression line is one where the loglikelihood value is the highest, i.e. the line with the maximum likelihood estimation and therefore the best fit.

2.8.2 *Statistical significance*

Statistical significance tests the likelihood that a measurement or result is not attributed to chance. Utilizing the concept of a "null hypothesis", if true, then no difference or association is observed in the desired statistic. In the context of a correlation analysis between two variables, a correlation coefficient gives an indication of the relationship between two variables (direction and strength) and whether the relationship observed in the sample data is strong enough to model in the larger population. Therefore, the reliability of the model is dependent upon not only the correlation coefficient, but also the number paired data points in the sample. Testing the significance of the correlation coefficient takes into account both points. So, although the relative correlation coefficient maybe low in magnitude, a larger sample size may result in a correlation to still be statistically significant. Similarly, in regression models, especially with multiple predictor variables statistical significance can evaluate if the coefficients of certain predictor variables are statistically significant in relation to the response variable. It can therefore be used as a means of variable selection (Balding, 2006, Sterne and Davey Smith, 2001).

Whether its correlation or regression, statistical significance can be undertaken by estimating the p-values or the probability values of the coefficients. The p-value tests the null hypothesis that a coefficient is significantly different from zero. If a p-value is low, i.e. less than the stated significance level (α) e.g. 0.05, it indicates that there is enough evidence to reject the null hypothesis and the measured variable has an effect. Conversely, a larger p-value for a variable has no effect. In a correlation analysis, if p-value is less than or equal to α, the correlation is different from zero and is statistically significant. If the p-value is more than α, then one cannot conclude that the correlation is different from zero and is therefore not statistically significant. Using p-values for regression coefficients implies all variables in a model to be treated individually (univariate) in relation to the outcome, and only the significant variables are included into the final regression model. Indeed, as the number of variables increase in a univariate analysis (correlation and regression) or multivariate analysis (regression), the higher the chance that certain variables maybe deemed significant by chance, resulting in false positives. The significance of the p-value in this instance is therefore representative of the Type 1 error, which is a false rejection of the null hypothesis, and therefore equates to the probability of a wrongful significant correlation or wrongful inclusion of variables into the regression model, when in actual fact it's true.

To account for this, p-values can be "adjusted" with multiple testing correction methods, including the Bonferroni correction (Dunn, 1961, Armstrong et al., 2011) and the Benjamini correction (Benjamini and Hochberg, 1995). The probability of making at least one Type 1 error is referred to as the familywise error rate (FWER). Bonferroni correction attempts to compensate for this. There are two methods which can be employed for Bonferroni correction. Assuming, $\alpha = 0.05$, Bonferroni correction can firstly divide α by the number of tests (n) being performed, thereby ensuring FWER never exceeds 0.05, or secondly, multiplying the p-value by n, thereby reducing the number of variables which will be statistically significant. Both methods within Bonferroni are extremely stringent when compared to Benjamini-Hochberg (BH) correction. BH attempts to decrease the false discovery rate (FDR), and its analog the q-value (Storey, 2002), by avoiding all Type I errors. It works by firstly sorting the p-values in descending order of magnitude. The largest value remains unchanged, while the second largest is adjusted by multiplying by $n/(n\text{-}1)$, the third largest by $n/(n\text{-}2)$ and so on. In summary, corrections based on FWER provide a strong control of the number of false positives but are not really adapted for a high number of tests as they then have low statistical power. Corrections based on FDR are more powerful but offer a weaker control of the number of false positives.

Another method to test if an estimate of correlation or regression coefficient is robust and reliable is to set up a confidence interval. This can be used as an alternative to testing a null hypothesis, as described above. Confidence intervals gives an indication of the range a true value for a given statistic would lie, within a certain degree of probability which is defined (du Prel et al., 2009). A confidence interval of 95%, which is commonly cited, indicates the true value a measurement lies when performed 95 out of 100 times. As the confidence interval relates to the number of observations and standard deviation in a study, a narrow interval would imply a robust measurement with low variation. An interval range that does not include zero would therefore mean more confidence in the measurement. Finally, a higher probability confidence interval covers a larger range, e.g. 99% confidence interval is wider than a 95% confidence interval. Estimation of confidence intervals by bootstrapping has shown to have an number of advantages over conventional methods (Wood, 2004) without having to make assumptions on normality and other parametric tests.

2.8.3 *Multivariate data analysis*

Generally, split into unsupervised and supervised approaches, an unsupervised multivariate method provides an overview of data without any *a priori* information on the samples. A widely used method is Principal Components Analysis (PCA). In contrast, a supervised method aims to predict a specific outcome based on class information which is available or defined by a set of input variables. Partial

Least Squares (PLS) (Wold et al., 2001) analysis and orthogonal projection to latent structures (OPLS)(Trygg and Wold, 2002) are typical supervised approaches used for exploratory purposes (**Figure 2-5**). However, for any of these approaches to be fruitful, standardization of the data is generally required prior to any multivariate based analyses.

2.8.3.1 Multivariate standardisation

Pre-processing metabolomic data before any multivariate analysis is integral as the analysis involves the combination of variables of widely different magnitudes and feature to feature variation. There therefore are several pre-processing feature-wise normalisation methods to transform the numerical intensities to some common scale, so that the comparisons are easily interpretable and more meaningful. Firstly, phenotyping data is presented as a data matrix with the dimension's *m/z*, retention time and intensity. High resolution spectra will display the *m/z* values into equidistant intervals depending on the resolution of the mass spectrometer. This is referred to as binning and is all-purpose to allow for fast data pre-processing and processing (Tautenhahn et al., 2008).

Mean centring is as pre-treatment method whereby the mean of each variable is subtracted from all samples in the data, thus removing large offsets when investigating both low- and high-level metabolites. This results in the data to be oriented around zero and not the actual mean of the metabolite levels, thereby focusing on the actual fluctuating part of the data (Bro and Smilde, 2003). When there is emphasis on the moderate to low level metabolites, centring is often combined with a scaling pre-processing method. Unit variance and Pareto scaling (van den Berg et al., 2006) are such methods which focus on these metabolites by correcting for metabolic variances using the standard deviations of peak intensities. Unit variance is where every peak intensity is divided by the standard deviation of all intensities for that given feature. The combination of mean-centering and variance scaling is termed auto-scaling and is recommended if measured variables in a dataset are of differing units. Auto-scaling however has the potential to enhance variation associated with noise variables. Pareto scaling is less stringent than unit variance and is therefore more popular in metabolomics. Peak intensities are divided by the square root of the standard deviation and so is less susceptible to over manipulation of the data. Another option is transformation of the data by reducing the relative distances between feature intensities to be more equal and at the same time, correct for heteroscedasticity (Kvalheim et al., 1994). Taking the square root (power transformation) or the logarithm (log transformation) of peak intensities, will transform data which are heavily skewed to be more gaussian or follow a normal distribution. This can then allow the use of parametric method s which are frequently used and well understood to be applied even when the data is not normally distributed pre-transformation. The log transformation however is unable to handle zero values.

2.8.3.2 Hierarchical cluster analysis (HCA)

Hierarchical cluster analysis is one of a number of clustering algorithms (others include k-means and mean-shift), that is used to highlight similarity (typically pair-wise) between subsets within a larger population (Xu and Wunsch, 2010). HCA is an unsupervised technique that can be applied in two ways: agglomerative and divisive (Tong et al., 2003, Fukusaki and Kobayashi, 2005). Divisive analysis utilises a top-down approach, where all groups originate from one cluster, and then iteratively splits into smaller clusters. Agglomerative, is a bottom-up approach, where all observations start in its own cluster, and iteratively merges pairs of nearby observations into clusters, until all clusters have been merged into a single cluster. The end result from both approaches is usually depicted as a hierarchical tree or dendrogram, illustrating the relationship of clusters based on similarity (**Figure 2-4)**. A dendrogram, consists of stacked branches, called “clades”, that break down into smaller branches. At the end of a clade, are “leaves”. In the agglomerative HCA, working from the bottom-up, the arrangement of the clades tells which leaves are the most similar and the height of the clade, indicates how similar or different each clade is to one another. Thus, the greater the height, the greater the difference. Defining the distance between observations and the distance between clusters is necessary in HCA. Linking pairs of observations that are in close proximity, is based on the linkage criterion used. Common methods are Euclidean distance or Manhattan distance. Common cluster linkages include, Single linkage, Average linkage Complete linkage and Ward’s linkage. In this thesis, Euclidian distances were used as a measure of distances between observations and Ward’s linkage as a measure of distances between clusters. The Ward method is based on a sum of squared errors rationale that only works for Euclidean distance between observations.

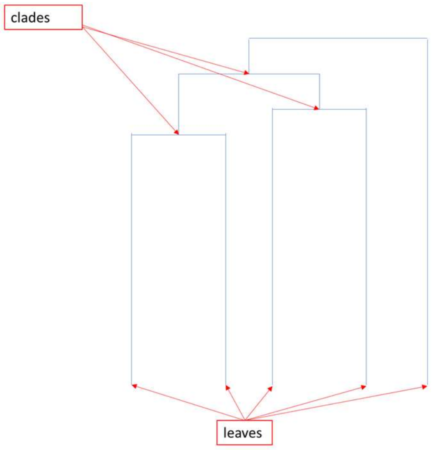


Figure 2-4. Typical example of a dendrogram used for Hierarchical clustering analysis.

2.8.3.3 Principal component analysis (PCA)

PCA is considered is an unsupervised multivariate technique that is widely used in metabolic phenotyping (Worley and Powers, 2013). It is used explore any general trends, clustering and outliers. Multiple principal components make up the model, and so the number of components reflect the greatest sources of variance present in the dataset with no consideration to any response variables. Divided into multiple principal components (PC) or sometimes called latent variables, the first PC explains the greatest variation in the data and each successive PC is independent and orthogonal to the one prior, explaining a different source of variation. A PCA scores plot can therefore be used to highlight patterns or relationships between different observations. In addition, a loadings plot can be viewed in which the loadings relate to the weights of each original variable (or spectral features) in the PC, by explaining the variables responsible for the observed variance in the scores plot.

2.8.3.4 Partial least squares (PLS) and orthogonal projection to latent structures (OPLS) analysis

PLS, an extension of PCA, and aims to model and predict relationships between an X data matrix consisting of independent variables (or predictors) and Y matrix of response dependent variables thereby linking the points. So, whilst PCA finds a subspace that explains the maximum variation in X, PLS captures the variance in X whilst also maximising the covariance between the scores in X and Y by creating a subspace which is a good representation of the relationship between X and Y. It is therefore both exploratory and predictive in nature. Classification with PLS is known as PLS-DA (PLS-

Discriminant analysis) explores the maximum co-variance between the X and Y variable, where Y can be categorical. OPLS-DA is a variant of PLS-DA which seeks to maximize, and capture confounding co-variance observed in X but is independent of Y. It uses an orthogonal signal correction filter (Wold et al., 1998) for maximum separation between the two groups, guided by known class information. Systematic variation that may otherwise confound the interpretation of the resulting PLS model is therefore removed, allowing OPLS to be an efficient tool with handling datasets with strongly collinear X predictors (or multicollinearity). As not all systematic variation in X is related to Y, an OPLS model is effective in separating the systematic variation into two parts; a.) the predictive or shared variation between X and Y and, b.) the orthogonal variation Y-uncorrelated variation in X and conversely the X-uncorrelated variation in Y. Simply, OPLS aims to condense all the predictive variance into the first component, and any subsequent components explain orthogonal (unrelated) variance. So, in theory, the addition of many components in the model is possible (as you would for regular PLS), but interpretation of the first component is the most important. Prediction power between PLS and OPLS have been shown to be similar (Kemsley and Tapp, 2009, Trygg and Wold, 2002). Generated from OPLS-DA models are the S-plots, which have been frequently used in metabolomic applications for biomarker identification (Wu et al., 2018, Liu et al., 2020, Banoei et al., 2019, Madala et al., 2012). As proposed by Wiklund *et al. (Wiklund et al., 2008),* S-plots were used in several areas of this work to identify discriminating features from OPLS-DA models. The S-plots are a scatter plot of the loadings that models the covariance (p) and correlation (p_{corr}) between the metabolite features and their modelled class designation. The covariance and correlation are plotted on opposite axis, resulting in features forming an “S” shape, thereby sending the most discriminant features to opposite quadrants of the plot.

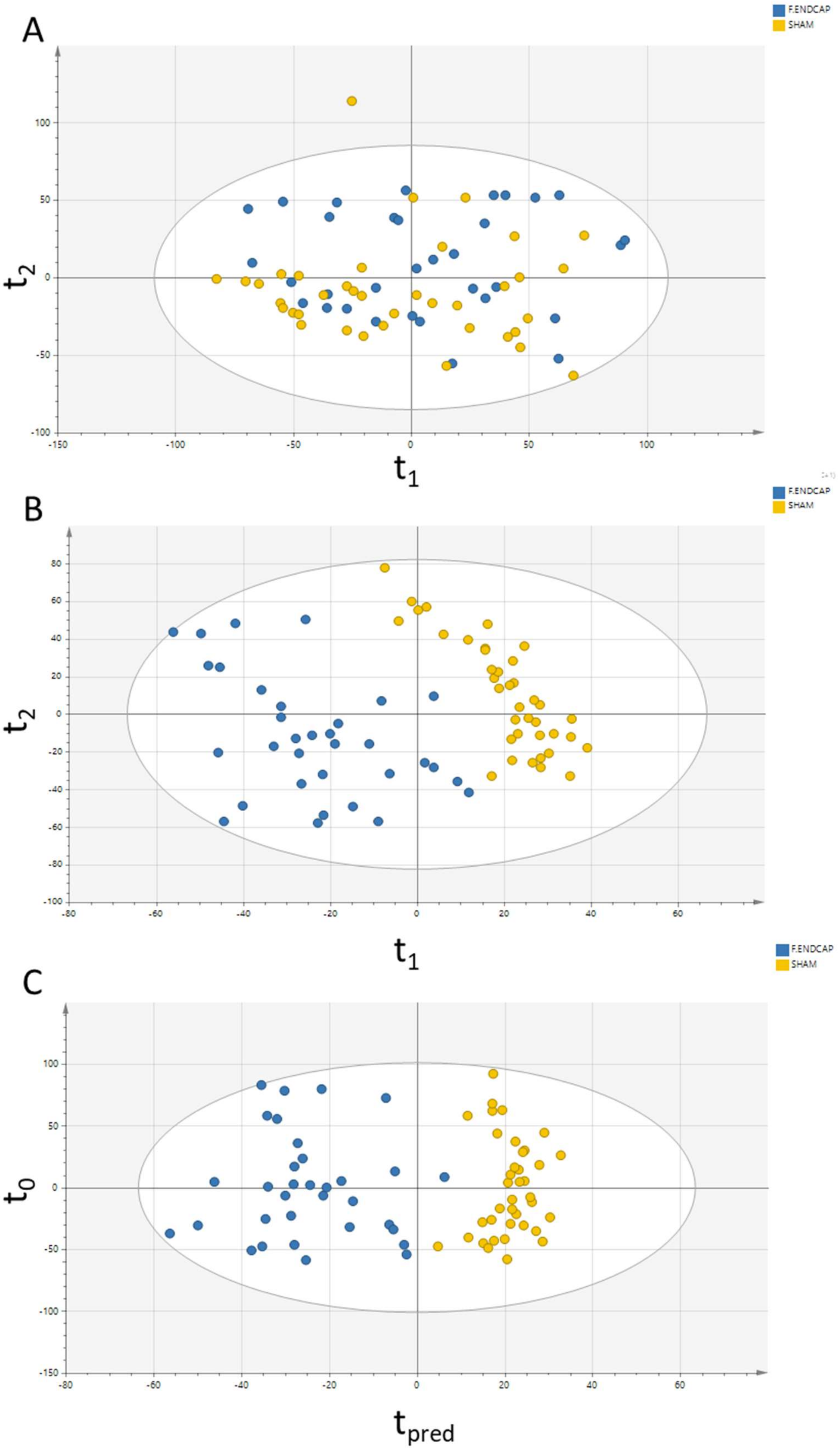


Figure 2-5. Comparison of the 2-component scores plots produced by PCA (A), PLS-DA(B) and OPLS-DA(C) using an exemplar RPC profiling experiment.

2.8.4 *Resampling and regularisation*

Classification problems potentially arise in phenotyping data as there are often in most cases more variables as there is samples. Poor modelling can result in scenarios where a model classifies the training data well but poorly with future data, resulting in what is called overfitting (Hawkins, 2004). Overfitting occurs when a training model incorporates all the data, which includes noise, random fluctuations, and outliers. The result however is a model which fits the data too well, incorporating all but then failing with new data. Multicollinearity is a common contributor to overfitting. The estimation of coefficients from regression models, heavily relies on the independence of the predictor variables. As the number of variables increase, there often is a chance of multicollinearity between variables occurring. Multicollinearity between variables can lead to incorrect interpretation of coefficients (Vatcheva et al., 2016) as regression lines becomes highly sensitive to deviations in the residuals resulting in large variance and a poor regression estimate. Resampling and regularisation are therefore important steps in addressing and minimising overfitting and thereby improving model performance.

2.8.4.1 Resampling

Resampling are methods in statistics which repeatedly draw samples from a population in order to estimate the precision of a specific statistic. One such method is cross validation. Cross validation (CV) is a powerful way prevent against overfitting (Vatcheva et al., 2016) whilst simultaneously optimising tuning parameters associated with the training model. The simplest and often most used CV method is randomly splitting the initial dataset into a training (or calibration) and test (or validation) set. This will give some idea if overfitting has occurred and determine how robust the model is and better approximate the ability of the model to perform on new data. K-fold cross validation is another method that splits the data into blocks or folds and depending on the number of folds (e.g. 4 folds or 10 folds), a minimum of one fold is left out as a test set and the remaining used to train the algorithm. Multivariate models such as PLS and OPLS are prone to overfitting, by separating classes even though there is no real difference between them (Westerhuis et al., 2008). Cross validation is therefore critical in ensuring model reliability and quality. In this thesis, SIMCA is used for all PLS-DA models. Quality assessment are measured through statistical parameters; R2Y and Q2Y. R2Y, although not a cross validation parameter, is reported as a measure of how well the model fits the data, i.e. the explained variation (Wold et al., 2001). It increases with the number of components in the model, eventually approaching 1, where an R2 of 1 perfectly describes the variation in the model. To guard against overfitting, the Q2Y is determined. The Q2 statistic is a measure of the predictive power of the model and is estimated through cross validation. The data is essentially divided into 7 parts, where 1/7th of the data is randomly selected and used as a testing set. Generally, the larger the Q2 the more confidence a model will be able to predict new data.

Permutation testing is another resampling technique used to ensure the validity of classification models. Random permutations in the Y response are generated to which individuals are randomly assigned to different classes and modelled. The theory is that now that they have been incorrectly classified, the result should be a poor model and an ineffective class prediction. Repetition of the permutation is carried out resulting in a distribution of Q2 values and since the groups are selected randomly, the assumption is that no difference exists between them. When plotted, a line of best fit is regressed from the Q2 of the original model through the distribution of permuted Q2 values where it interacts with the Y intercept to give the mean of the distribution. A reliable model (samples not used in permutation) should lie outside the 95% confidence interval ($p<0.05$) of such a distribution and the difference is statistically significant to the randomly permuted class labels indicating high validity of the model.

Like permutation testing, bootstrapping is a resampling method that builds a bootstrap distribution by resampling the observed data. It is a technique that independently draws sub-samples of the same sample size, with replacement, from the original dataset and then makes an inference on the measured statistic. As this is also a form of inferential statistics, the result is a calculation of the distribution of estimated values that would be expected if drawn from the original population, resulting in a confidence interval (Efron and Tibshirani, 1993).

2.8.4.2 Regularisation

Regularisation deals with multicollinearity by intentionally introducing some bias and thus reducing variance. Regularisation methods also has the added benefit of filtering noise variables and prevent therefore preventing overfitting. Both linear regression and LR allows for easy regularization to prevent and is another preventative measure against overfitting. It works by penalising the magnitude of the logistic/linear regression coefficients as well as minimising the error between predicted and actual observations. The outcome is to shrink the beta coefficient towards zero for unimportant variables thus being removed from model and reducing model complexity. There are three methods, Ridge (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996) and Elastic net (Zou and Hastie, 2005), and the difference lie in the application of the penalty to the coefficients. Ridge uses a

L2 regularization which adds a penalty term, Lambda (λ) which controls the importance of the regularization parameter.

$$\min\ (||\mathrm{y}-\mathrm{x}(\beta)||_2^2 +\ \lambda||\beta||_2^2) \tag{2. 13}$$

In the ridge function λ is denoted by alpha (α) and this controls the magnitude of the penalty on the coefficients, i.e. the higher the value of alpha, the bigger the penalty, and so the smaller the coefficients. Ridge regression won't remove any variables but minimises or shrinks the beta coefficient, thus all features remain in the model thereby reducing its complexity. This has added benefit of prevention of over fitting and works well with highly correlated features.

Least Absolute Shrinkage and Selection operator (LASSO), utilise similar concepts to that of Ridge but instead adds a L1 penalty term equivalent to the absolute value of the magnitude of the coefficients.

$$\min\ (||\mathrm{y}-\mathrm{x}(\beta)||_2^2 +\ \lambda||\beta||_1) \tag{2. 14}$$

Similarly, λ is equal to α in LASSO. L1 regularisation will shrink certain coefficients to zero, thereby removing variables entirely.

Elastic net (EN) regression incorporates penalties from both L1 and L2 regularisation.

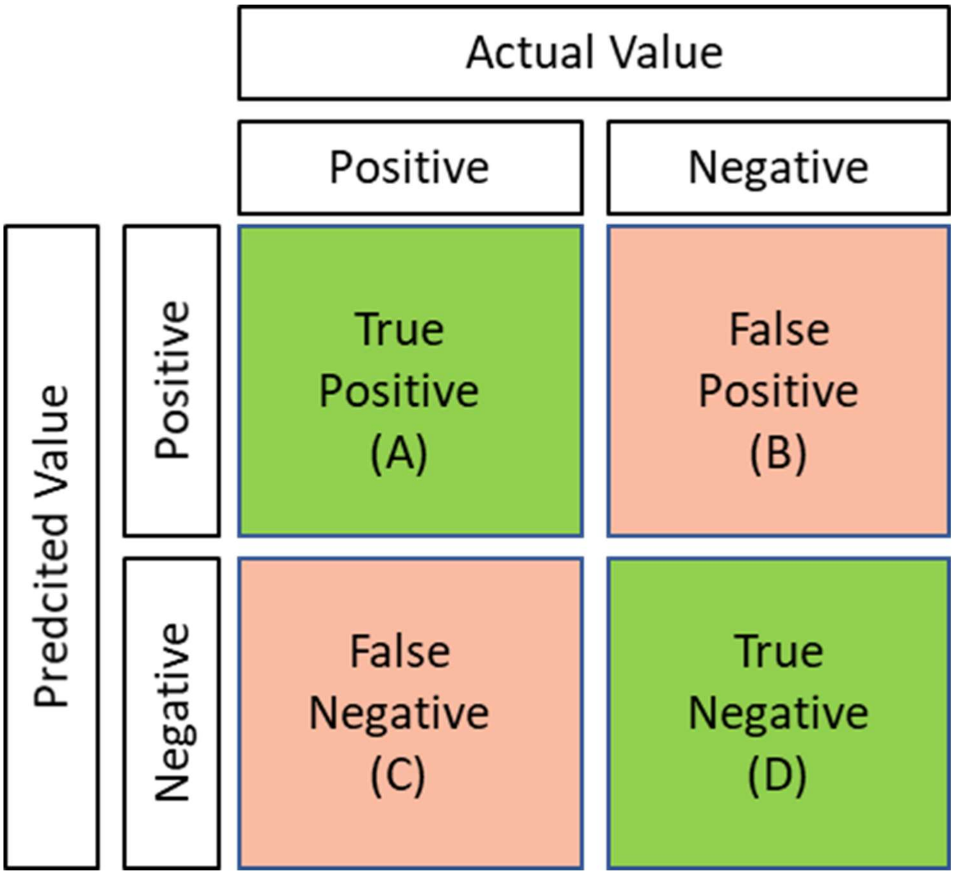
$$\min\ (||\mathrm{y}-\mathrm{x}(\beta)||_2^2 +\ \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2\ ) \tag{2. 15}$$

If α is set to 0, the penalty assimilates the Ridge L1 term and if α is set to 1, the penalty will be the LASSO L2 term. Therefore, for EN, the α and λ parameter must be optimised, usually *via* methods like cross validation. It's especially useful when multiple variables correlate with one another. EN will group the strongly correlated variables and if any one variable has a strong association with the dependent variable, all will be included in the model. In this scenario, LASSO will include only the one variable, whereas EN will likely include all. Regularization using LASSO or EN has therefore "built-in" variable selection, whereas Ridge would require other means like resampling and/or significance tests for variable selection.

2.8.5 *Performance assessment of linear and classification models*

In linear models, the optimal fitted line is one that minimises the total variance and can be expressed by the goodness-of-fit statistic, R^2. However, if the model has too many variables, it can be heavily influenced by the random noise and be incorporated into the model, thus leading to overfitting. The R squared adjusted is then a better measure of the performance of the model as it considers the number of predictor variables used in the model. The adjusted R2 will only increase if a new variable improves the model more by random chance. Root mean square error (RMSE) is also another statistic used to valuate model fit and is the square root of the standard deviation of the residuals (Heinze et al., 2018). RMSE is a measure of the spread of the data points surrounding the regression line and therefore how close observed values are to the predicted values. A lower RMSE indicates a better fit.

Typical performance parameters for binary classifier measurements such as LR, are assessed by different methods to that of linear models. Popular methods are the use of a confusion matrix (**Figure 2-6**) and Receiver Operating Characteristic (ROC) curves (**Figure 2-7**) (Tharwat, 2018). The confusion matrix (or contingency table) assesses the model's accuracy and misclassification error. It is also another means to detect and avoid overfitting. **Figure 2-6** is an example of a 2 x 2 confusion matrix.



		Actual Value	
		Positive	Negative
Predcited Value	Positive	True Positive (A)	False Positive (B)
	Negative	False Negative (C)	True Negative (D)

Figure 2-6. Example of a confusion matrix used in classification models.

And the accuracy of the model is calculated with the following formula:

$$\text{Accuracy} = \frac{(\text{True Positives} + \text{True Negatives})}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \qquad \textit{(2. 16)}$$

To evaluate the effectiveness of the LR model, probabilities from the model are converted into classifications by setting a threshold, e.g. if a threshold of 0.5 is set, then any observation with a probability higher than 0.5, then belongs to one group and conversely any observation below 0.5, belongs to the other group. Using this model and with new data, if an observation is classified as positive, and in fact it is, then its labelled as a true positive. If classified incorrectly, then it's a false negative and therefore a type II error. The inverse being, a negative observation that's classified as negative will be a true negative, but if classified as positive, then is a false positive and therefore a type I error. As different threshold is set, so too are different confusion matrices constructed. To decide the optimum threshold value, a model's sensitivity and specificity are generated, and when plotted is also referred to as a ROC curve.

The Receiver Operating Characteristic (ROC) summarises the performance and the discrimination ability of a logistic model, by optimising the trade-off between the true positive rate (TPR), sensitivity, and false positive rate (FPR), 1-specificity.

$$\mathbf{True\ Positive\ Rate = Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}} \qquad \textbf{\textit{(2. 17)}}$$

$$\mathbf{False\ Positive\ Rate = 1 - Specificity = \frac{False\ Positives}{False\ Positives + True\ Negatives}} \qquad \textbf{\textit{(2. 18)}}$$

Sensitivity is defined as the number of observations with a fitted predictive probability above a specific threshold and specificity is number of observations with a fitted predictive probability below the threshold. The direction of the threshold, either higher or lower, will ultimately determine if new data will fall in either the true-positive or false-positive distributions, thereby influencing the sensitivity and or specificity. Therein lies the trade of in choosing the optimum threshold. However, if translated to a ROC curve, the optimum threshold produces no false positives. The x-axis of the ROC curve is the specificity and the y-axis the sensitivity. Construction of the ROC curve is undertaken by plotting the sensitivity and specificity points for a given threshold from the confusion matrices. The area under the ROC curve (AUC) can then be used to determine the overall accuracy of the model. The higher the AUC, this is observed by how close the curve is to the left top left-hand corner of the graph, the better the prediction power of the model.

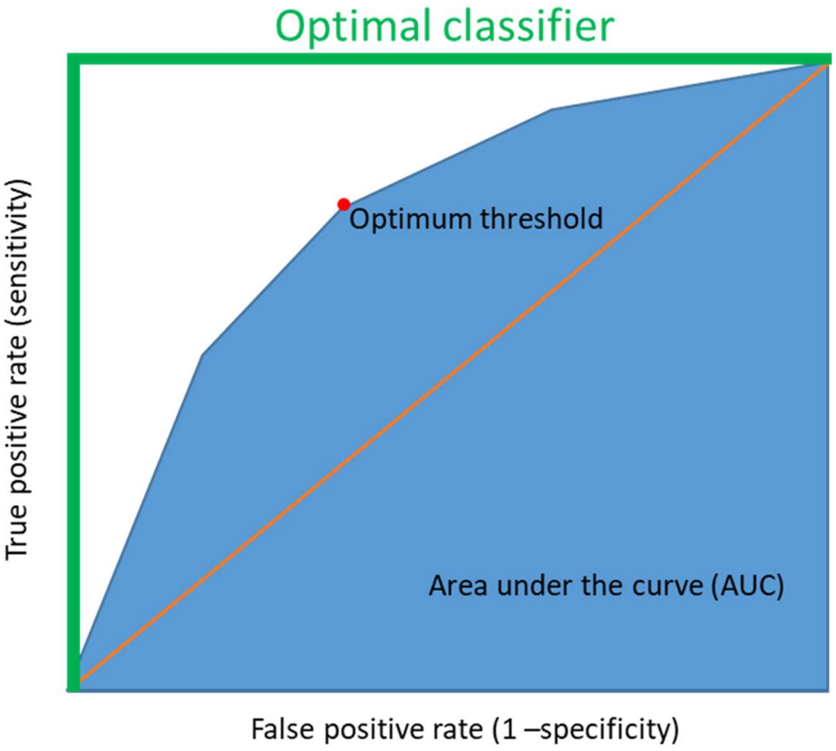

Optimal classifier
Optimum threshold
True positive rate (sensitivity)
Area under the curve (AUC)
False positive rate (1 –specificity)

Figure 2-7. Typical example of the area under the ROC curve. The top left corner of the ROC figure is the point where sensitivity and specfcity is 100% i.e. True positve and true negative are 100% . The closer the blue curved line or ROC curve is to the top left corner, the better the model is at distinguishing between two classes. The optimum threshold (red point), is the point that best discriminates between the two classes. The closer the curve is to the center diagonal orange line, the worst the model is at distinguishing between two classes.

Assessment metrics, such as accuracy from the confusion matrices, are sensitive to class imbalance. ROC curves however are insensitive and robust to any class imbalance, as ROC depends on the true and false positive rates irrespective of the actual classes in the data. However, for a better estimate of the accuracy, a threshold is determined which can be computed in several ways. One way is the point on the curve which is closest to the TPR of 1 and FPR of 0. Another option, which is what is used in the R software for ROC curves produced in this thesis, is the average of the "minimum value that gives the least number of false positives" and "maximum value that gives the least amount of false negatives". If there is a class imbalance observed in the population, this can result in inequivalent false positive and false negative predictions. To address this, weight can be added in the code to correct for this inconsistency. This can often be the case, when interrogating xenobiotic exposure from an untargeted investigation as case control sample numbers can differ. The threshold from the ROC curves can subsequently be inputted back into the confusion matrix. This serves as a correction for any class imbalance that may exist in the data, and the accuracy calculated from the confusion matrix should equate to the accuracy calculated from the ROC curve.

2.9 Metabolite Identification

Mass spectrometry has been widely used in untargeted metabolomics for a long period but metabolite identification continues to represent one of the most significant challenges in the workflow (Wishart, 2011). Putative annotations of spectral features of interest can be obtained by comparing these data to on-line spectral databases such as PubChem (Kim et al., 2018), HMDB (Wishart et al., 2017) and Metlin (Guijas et al., 2018). Despite providing highly accurate measurements (to within a few ppm), high-resolution MS measurements, may result in multiple putative candidates based on *m/z* alone. Additionally, isomers and isobaric compounds exhibit identical/very similar MS profiles (Creek et al., 2014) and are therefore impossible/challenging to distinguish.

Furthermore, as multiple ionisation products are commonly observed in MS based analyses (multimers, adducts, multiply charged ions, isotopes, and fragments), the formation of the molecular ion, may not necessarily be observed. The molecular ion peak corresponds to any molecule that has not undergone fragmentation. It can be considered reflective of the molecular weight of the compound but with a charge, i.e. M+H in positive mode and M-H in negative mode. Note, it is different to the base peak, which is the largest peak in the spectrum. In an MS/MS experiment, a fragment could potentially be the base peak. The molecular ion is therefore used as a reference point in identifying neutral losses and fragment ions. Thus, an understanding of the mass spectrum and the various ion types produced, in relation to the candidate metabolite, is therefore generally required.

Complementary analytical techniques such as NMR, or liquid chromatography, can also aid metabolite identification efforts, notwithstanding the additional resources that may be required (instrumentation, analyst, sample). Additional measurements can provide evidence for the true identity and annotation of an unknown target analyte.

Common practice for metabolite identification in LC-MS based metabolomics, requires acquisition of reference standards ascertained from the putative candidates and analysed under the identical experimental conditions (Kind and Fiehn, 2010). It is however important to define the distinction between putative “annotation” and “identification” of metabolites.

The annotation procedure, as defined by Metabolomics Standards Initiative (Sumner et al., 2007), relies on a positive comparison of two or more physicochemical properties of a given feature to spectral databases; whereas a true identification procedure involves comparative analysis of each detected feature with an authentic chemical standard measured on the same analytical platform. The latter however is clearly unfeasible for routine high-throughput analysis due to availability and cost of chemical standards but is still a necessary requirement for biomarker annotation validation.

The confidence of the metabolite identification described in this thesis refers to the identification levels as stated by the metabolomics standards initiative (MSI). The levels are as follows;

- **MSI level 1**
  Identification based on comparison of experimental spectral data to authentic chemical standard analysed with the same analytical conditions;
- **MSI level 2**
  Corresponds to putative annotation based on spectral similarity with public/commercial spectral libraries;
- **MSI level 3**
  Assignment to a class of metabolites;
- **MSI level 4**
  An unknown metabolite which could still be determined based on *in silico* fragmentation.

Chapter 3

Knowledge-based and data-driven extraction of xenometabolome signatures from large-scale metabolic phenotyping data

Summary

The efficiency and coverage of xenometabolome annotation in human metabolic phenotyping datasets has historically been relatively limited. This chapter describes the development of two complementary approaches aimed at improving routine spectral assignment of xenobiotics in biofluid spectra. A workflow was developed to generate an extensible reference standard database for xenobiotics, to support rapid annotation in metabolic phenotyping datasets analysed by RPC-UPLC-MS (ESI +/-). To accomplish this, a literature search and analysis of key public data was conducted to prioritise those compounds known to prevalent in the UK population, thus making use of what is already known about common, deliberate xenobiotic exposures. This chapter also describes the refinement of univariate and multivariate statistical methods for the annotation of xenobiotics and related metabolites, making explicit use of the key chemical and biochemical features and relationships that commonly exist within and between these compounds.

The statistical methods employed were:

1. Correlation analysis between the intensity pattern of the MS feature corresponding to APAP, to the intensity pattern of all other features in the dataset.
2. Correlation analysis using a newly developed i-STOCSY tool to annotate metabolites of memantine
3. Logistic regression (univariate and multivariate) to annotate metabolites of donepezil
4. PLS models to annotate metabolites of amlodipine

These approaches were then used to explore exemplar population study data.

Aim and Objectives

The central aim of this work was to develop strategies to identify xenometabolome signatures from large scale metabolic phenotyping datasets. To address this central aim, two specific objectives were identified and form the focus of the work described in this chapter.

1. Databasing of standards. Identification of commonly used pharmaceuticals in the general population, and others of specific relevance to the exemplar populations associated with this study, and generation of a database for supporting rapid annotation in NPC assays
2. Utilizing statistically based methods to identify xenobiotic signatures, consistent with exposure, in existing metabolic phenotype data.

3.1 Introduction

In the United Kingdom (UK), pharmaceutical medicines are legally classified into three categories: i) over the counter (OTC); ii) pharmacy only (P); iii) prescription only medicines (POMs). Pharmacy only (P) medication can only be found in a pharmacy under the supervision of a pharmacist. This means, medications "behind the counter", and are not available freely without the approval of a pharmacist depending. The creation of medical institutions such as NHS have led individuals to be become reliant upon them for medical advice and was not until the early 2000 that government bodies encouraged self-care. This was carried out by creating NHS walk in centres but more importantly, expanding medications which were once POMs to P (Rutter, 2012). The Medicines and Healthcare Regulatory Agency (MHRA) facilitated this process, by establishing proper protocols and consultation periods. The first of which being ibuprofen, which was classified as P status in 1983. POMs, need to be prescribed by a doctor or a qualified healthcare professional and can only be collected at pharmacies.

Since 1983, in the UK, 150 medications have been made available OTC, no longer requiring a prescription. OTC medication or General sales list medicines (GSL) are used to treat minor ailments and so does not require the supervision of a pharmacist and are subject to less regulation. They can be found in pharmacies or in local supermarkets. OTC drugs are available freely to individuals allowing for self-management and treatment of symptoms to minor ailments and illnesses. The Proprietary Association of Great Britain (PAGB), which are a UK trade association representing all OTC medicines and food ailments, quotes the value of the UK OTC market to be at £2.5 billion. 964

million packs of OTC medication were sold to individuals for self-care and treatment (www.pagb.co.uk/about/otcmarket). In comparison to the 1 billion in prescription medications, nine out of ten general practitioners believe that self-care is based on a continuum ranging from minor ailments, based on an individual's discretion, to the more series illnesses which requires the aid of healthcare professionals (www.pagb.co.uk/selfcare/home). Popular categories of OTC medication include painkillers, cough/cold and skin treatments (www.pagb.co.uk/publications/directory. Kidney Research UK additionally included the categories; non-steroidal anti-inflammatory medications, heartburn/acid reflux and antihistamines.

A statistical survey conducted by the Health and Social Care Information Centre (HSCIC) in 2013, revealed that £15 billion was the total drug cost in hospitals and the pharmaceutical community in the England (Croft, 2014). The survey was conducted between 2012 and 2013 and reflected both individuals within the healthcare (NHS) system, but of that of the general population. More recently in 2018, the Prescription cost analysis England 2018 report stated that the bill for prescriptions dispensed amounted to £8.8 billion (https://files.digital.nhs.uk/E5/A014A5/PCA-eng-2018-report.pdf).

A core drug list of the most commonly prescribed drug groups in England was published, ranking proton pump inhibitors (omeprazole, lansoprazole), statins (simvastatin, atorvastatin, pravastatin), paracetamol, beta-blockers (bisoprolol, atenolol, propranolol) and calcium-channel blockers (amlodipine, felodipine, diltiazem, nifedipine, lercanidipine) in the top five categories (Audi et al., 2018).

Accordingly, the HSCIC survey responses largely reflect population usage of OTC pharmaceuticals, and therefore and therefore represents a useful guide for prioritising efforts to characterise xenometabolome exposures according to expected prevalence.

The work presented in this chapter focused on increasing the number of positively identified and confidently annotated features related to xenobiotics in metabolic phenotyping data. To provide a solid basis for this work, rational selection of prioritised therapeutic medications was conducted using the HSCIC survey data, allowing prioritisation of compounds to for the development of a mass spectrometry-based authentic reference standards database, and attendant workflow.

3.2 Hypothesis

This chapter aims to broaden substantially the coverage of the xenometabolome through the construction of a xenobiotic reference standard database (knowledge-driven strategy) and the development of statical based methodologies (data-driven strategy) to further reveal additional xenobiotic metabolites using targets from the database or, from compliance/questionnaire metadata. These methods and workflows will be applied to existing human cohort study data to establish the prevalence and variation of these xenobiotic metabolites. Combined, these strategies will provide novel methodologies for high throughput xenometabolome analysis in population (epidemiological/clinical) sample sets, and provide greater coverage of xenobiotics for existing studies, permitting improved detection of non-compliant participants, and better confounder data.

3.3 Methods

A list of therapeutic drugs which are commonly administered in the United Kingdom was compiled. In brief, xenobiotics commonly administered in the UK were identified using the 2013 HSCIC survey and various other literature sources. The top 25 medications were used to initially populate the database (

Table 3-1). A secondary approach to increase xenobiotic annotations, was to integrate statistical based methods to extract unknown xenobiotic signatures. Both univariate and multivariate methods were explored and applied using an exemplar large-scale population study.

3.3.1 Materials

LC/MS grade water was obtained from fisher scientific (Fisher scientific, USA). 96-well plates were from Eppendorf (Hamburg, Germany) and well plate cap mats were from VWR (Leicestershire, UK). 5mL Cyrotube were purchased from Brook life sciences (Fluid X., Brook Life sciences, UK)

Table 3-1. **Priority drugs that are commonly prescribed, prevalent, or otherwise in use within the UK.** The list was compiled based on the comprehensive literature search described in the introduction. This list includes known/predicted major metabolites, where available from literature. Acquisition of the top 25 chemical reference standards for un-metabolised xenobiotic was initially used to populate the database, as indicated in the table below.

Xenobiotic	Number of known reported metabolites	Source	Use	Reference
Caffeine	14	Food/OTC	Supplement used for energy	(Andrews et al., 2007).
Nicotine/Cotinine	8	OTC	Widely used stimulant in smoking	(Baskin et al., 1998)
Aspirin	6	OTC	Reduce pain, anti-inflammatory	(Hutt et al., 1986)
Theobromine	5	Prescription/Food	Stimulant and metabolite of caffeine	(de Sena et al., 2011)
Theophylline	5	Prescription/Food	phosphodiesterase inhibiting drug used in therapy for respiratory diseases. Also, metabolite of caffeine	(Dubuis et al., 2014)
Simvastatin	5	Prescription	Statin used to lower cholesterol	(Prueksaritanont et al., 1997)

Xenobiotic	Number of known reported metabolites	Source	Use	Reference
Levothyroxine	4	Prescription	Used to treat thyroid hormone deficiency,	(Ianiro et al., 2014)
Acetaminophen	10	Prescription/OTC	Pain relief	(Johnson and Plumb, 2005)
Ibuprofen	4	Prescription/OTC	Non-steroidal anti-inflammatory	(Clayton et al., 1998)
Omeprazole	3	Prescription	Treatment for indigestion, heartburn and acid reflux	(Kobayashi et al., 1992)
Ramipril	5	Prescription	Lowers blood pressure	(Verho et al., 1995)
Amlodipine	7	Prescription	Lowers blood pressure	(Zhu et al., 2014)
Salbutamol	5	Prescription	Relieve symptoms of asthma and chronic obstructive pulmonary disease (COPD)	(Dominguez-Romero et al., 2013)
Lansoprazole	5	Prescription	Treat indigestion, heartburn, acid reflux and gastroesophageal-reflux-disease (GORD). Also stomach ulcers	(Song et al., 2008)

Xenobiotic	Number of known reported metabolites	Source	Use	Reference
Atorvastatin	5	Prescription	Statin used to lower cholesterol	(Macwan et al., 2011)
Metformin	-	Prescription	lowers blood sugar levels (Type II diabetics)	(Liu and Coleman, 2009)
Cholecalciferol	2	Prescription	Fat-soluble vitamin (d3) that helps your body absorb calcium and phosphorus	(Holick et al., 1972)
Bendroflumethiazide	-	Prescription	Diuretic, used to treat high blood pressure	(Beermann et al., 1977)
Bisoprolol	4	Prescription	Treat high blood pressure	(Horikiri et al., 1998)
Citalopram	7	Prescription	Antidepressant	(Dalgaard and Larsen, 1999)
Codeine	6	Prescription/OTC	Opiate used to treat pain	(Frost et al., 2015)
Amoxicillin	3	Prescription	Antibiotic used to treat bacterial infection	(Haginaka and Wakai, 1987)
Furosemide	2	Prescription	Diuretic	(Baranowska et al., 2010)
Amitriptyline	7	Prescription	Antidepressant	(Breyer-Pfaff, 2004)
Warfarin	4	Prescription	Anticoagulant/blood thinner	(Locatelli et al., 2005)

3.3.2 Acquisition and analysis of authentic xenobiotic reference standards

Reference standards and actual pharmaceutical formulations were acquired from various vendors (refer to Appendix 1). Preparation, acquisition and data extraction are in conjunction with ongoing protocols set in place for reference standard acquisition at the NPC and is described below.

Reference standards of xenobiotics were made in a qualitative manner by taking a fixed equal amount (if solid) or droplet (if liquid) into a clean 5mL cryotube and made to volume with ultrapure water. To handle the high volume of reference standards, certain measures were undertaken to account for differences in concentration between the reference standards. A 1000 µL aliquot was transferred to a 96-well deep well plate, and a further dilution of 1:10 (100 µL standard: 900 µL water), 1:100 (10 µL standard: 990 µL water) and 1:1000 (1 µL standard: 999 µL water) (v/v) was then subsequently prepared into additional 96 well plates.

An aliquot of a reference standard mixture was added to each well as an internal standard for chromatographic retention time(s). The reference standards correspond to a working concentration of the method reference mixture (MR) used for reversed phase profiling urine studies conducted at the NPC (Lewis et al., 2016). All individual standards and standard mixtures were frozen at -80°C. The 1:100 dilution plate was firstly acquired using reversed phase (RPC) separation on an Acquity UPLC system hyphenated with Xevo G2-S Q-TOF (Waters Corp., Milford MA, USA) -RPC-UPLC-MS, for the high-resolution detection of all observable chemical species. RPC was performed using an ACQUITY UPLC HSS T3 1.8µm, 2.1 x 150mm column with chromatographic conditions identical to that stated by Lewis *et al*. (Lewis et al., 2016). Mass spectral data was acquired under both electrospray positive ion and negative ion conditions, in continuum mode with two injections per well. One injection was at a low collision energy (4V) and the other utilising a collision energy ramp (10-30V). Within each injection, three interleaved full MS scans (0.05 second scan rate) were acquired for the *m/z* range between 50 and 1200 Da.

3.3.3 Population study used for xenobiotic exploration

An exemplar study were used for xenobiotic exploration; the Alzheimer’s Disease Multimodal Biomarkers study (ALZ) (Lovestone et al., 2009). ALZ is a nested case-control study of Alzheimer’s disease consisting of 650 urine samples. Urine samples were prepared and analysed (by myself) according to established protocols for urine phenotyping (Lewis et al., 2016) as previously described, by RPC- UPLC-MS in both positive and negative ion mode. The dataset acquired from RPC- UPLC-MS (positive ion mode) was used for statistical based exploration of xenobiotics.

Drug compliance metadata (records or known therapeutic drug use) were available for the ALZ dataset. With this information, several different statistical methods were used to identify signals relating to exposure from the xenobiotics reported, i.e. i-STOCSY and logistic regression. Prior to any statistical evaluation, a csv was constructed, indicating the samples with a known exposure, denoted by “1”, and the remaining samples in the dataset, denoted by “0”, for all reported medications.

3.3.4 *Pre-processing ALZ study data and reference standard data*

Reference standard data was firstly denoised using a compression/archival tool, and centroided for peak detection using Waters software (MassLynx™, Water Corporation). Within the archival tool, the processing options involved enabling noise reduction with the following parameters specified for RPC based analysis; threshold = 15, MS Resolution = set at the instrument resolution setting, Low Drift FWHM = 2, High Drift FWHM = 10, and Chromatographic Peak width = 0.02. Centroiding of the standard data was undertaken using the “Accurate Mass measure” tool located within the Masslynx software. Peak detection was set automatically with default values. Finally, all denoised and centroided data files were converted to the open NetCDF format.

There was no denoising or centroiding of the ALZ study data as all samples were already acquired in centroid mode. Mass spectral data files in .RAW format (Waters Corporation, USA) were converted to the open mzML format using the ProteoWizard msconvert tool (Chambers et al., 2012). During this conversion, all signals with an absolute intensity of less than 100 counts were removed.

Mass spectral peak picking (centwave), integration and grouping were performed on the converted NetCDF reference standards data and mzML ALZ study data using XCMS. The XCMS parameters for project data are part of the NPC workflow for project data acquisition and were processed by NPC informaticians. The XCMS parameters for reference standard acquisition were selected based of these parameters, with slight modifications on those which would have the biggest impact to feature selection. Both project and reference standard parameters are summarised in **Table 3-2.** Of the centwave peak picking parameters, “ppm” and “snthreshold” were kept identical. The “peakwidth” and “prefilter” parameters were altered to a slightly wider range for standards due to the qualitative way in which the standards were prepared. This preparation may result in broad saturated peaks, so the wider range would result in the integration of these peaks. The same qualitative approach may also result in lower signals for standards, so to account for this, the “noise” parameter was set at a slightly lower setting.

Table 3-2. XCMS parameters for NPC project data and reference standard RPC-UPLC-MS acquisitions (positive and negative ion mode).

Parameter	Project study data	Reference standard
ppm	25	25
peakwidth	1.5 to 5	2 to 8
snthreshold	10	10
noise	600	300
prefilter	4-1000	8 to 3000
mzdiff	0.001	0.001
mzCenterFun	wMean	wMean
integrate	2	2
minSamples	10	10
bw	3	3
binSize	0.01	0.01

The grouping function for both the reference standards and ALZ study data acquisitions were performed using the density method with a retention time window of 2 seconds and mass to charge (*m/z*) window of 0.001 Da. This grouping function was used as it, is a particularly faster process than its counterpart the “nearest” function, which matches features between samples one at a time and is therefore a slower approach, especially with regards to larger study sizes. The density function also can incorporate a minfrac filtering capability (Forsberg et al., 2018), which is necessary parameter for filtering purposes for project data (Refer to **section 2.7.4**). The sample groups argument was setup so that selection of features had to be present in either 40% of the SR or 40 %

of all remaining study samples (minFrac setting of 0.4). No retention time correction was performed, and the final “fillChromPeaks” method was applied with default parameters.

For the ALZ dataset, additional noise filtering protocols were implemented *via* the use of a pooled QC sample or Study Reference (SR), prepared by pooling together an aliquot of all study samples that represents the physical average. The quality control processes including the preparation of the SR sample and implementation of a dilution series for filtering purposes has already been previously described and established (Lewis et al., 2016, Dona et al., 2014). To account for the observed variable dilution with urine biofluids, probabilistic quotient normalization (PQN) was applied to the filtered dataset. The final result is feature matrix in the form of a .csv file, that summarizes all distinct signals (i.e. retention time, mass to charge (*m/z*) and intensities after normalization for all samples) captured by RPC-UPLC-MS in both polarities and meeting the filtering and QC criteria.

3.3.5 *Construction of a reference standard database to facilitate xenobiotic identification*

Electrospray ionisation mass spectra of individual compound typically contain multiple ionisation products that relate to a single molecule, including isotopes, adducts, multimers, neutral losses and in-source fragments; empirically-derived reference spectra obtained using a given analytical protocol may better reflect the features recorded for these compounds in real biological samples, compared with spectra predicted from *in silico* models.

An in-house script (written in R) was used to automate the following series of processes: i) features considered unique to the selected reference standard were extracted by comparing the standards against the standard acquired before and the standard acquired after in the analytical run; lower and upper signal intensity thresholds were set at 1 x 10^5 and 2.3 x 10^7, respectively; ii) features that fell within this range a minimum of two out of three times (from the three MS scans) were recorded; iii) features meeting these criteria were output to an RDA file (R specific encoded information e.g. the objects, variables etc) and used to create two-dimensional (2D) plots of retention time vs *m/z* for the unique features. Both low and high energy collision acquisitions were incorporated into this workflow.

Retention time, *m/z* for the molecular ion, one adduct, and a minimum of one fragment ion (in-source fragment) were manually curated and tabulated, alongside descriptor columns to link compounds/data to their physical/digital location, as well as to relevant entries in on-line spectral databases.

Identifications of xenobiotics using the database was exemplified in ALZ. Features corresponding to the spectral information from the database (molecular ion, adduct and in-source fragment) were targeted by implementing a peak fitting algorithm (*peakPanthR* - https://www.bioconductor.org/packages/release/bioc/html/peakPantheR.html) to the raw data for peak integration and identification. Detection was based on peaks above a certain threshold and with a minimum signal to noise ratio (S/NR) ≥ 5, which is a conservative criterion for defining the limit of detection (Armbruster and Pry, 2008). The analytical specificity was also considered and intended as a method specific measure of observed interferences and confirmation that the correct peak was integrated. Of the xenobiotics detected in ALZ, population prevalence was calculated as the proportion of samples exhibiting the xenobiotic as a percentage of the total study samples.

3.3.6 *Developments of data driven statistical methods to highlight xenobiotic exposure in ALZ*

3.3.6.1 Outlier detection

Certain outlier samples from previous studies revealed a pattern of xenobiotic presence within the larger dataset that was considered as the foundation for a possible strategy in identifying other xenobiotics. This pattern may be described as the presence of a minority of samples within the dataset that demonstrates an intensity for a given feature, that greatly differs from its main distribution. A script was written in the R language that assisted with the detection of features with this pattern. The function of the script is outlined below.

For each feature group in the pre-processed dataset:

1. Median and standard deviation feature intensity values were calculated for all samples within the experiment.
2. The maximum feature intensity was calculated.
3. The difference in feature intensity between the observed maximum and median value was expressed in number of standard deviations (SD).
4. All features within the dataset were then ranked by the calculated difference in descending order, highlighting the features responsible for having the greatest difference to lowest difference in the number of SD's between the median and max values.

5. A plot for each feature is generated and colour coded. It displays the feature rank and sample that is responsible for the greatest difference between the median and study samples (**Figure 3-1**).

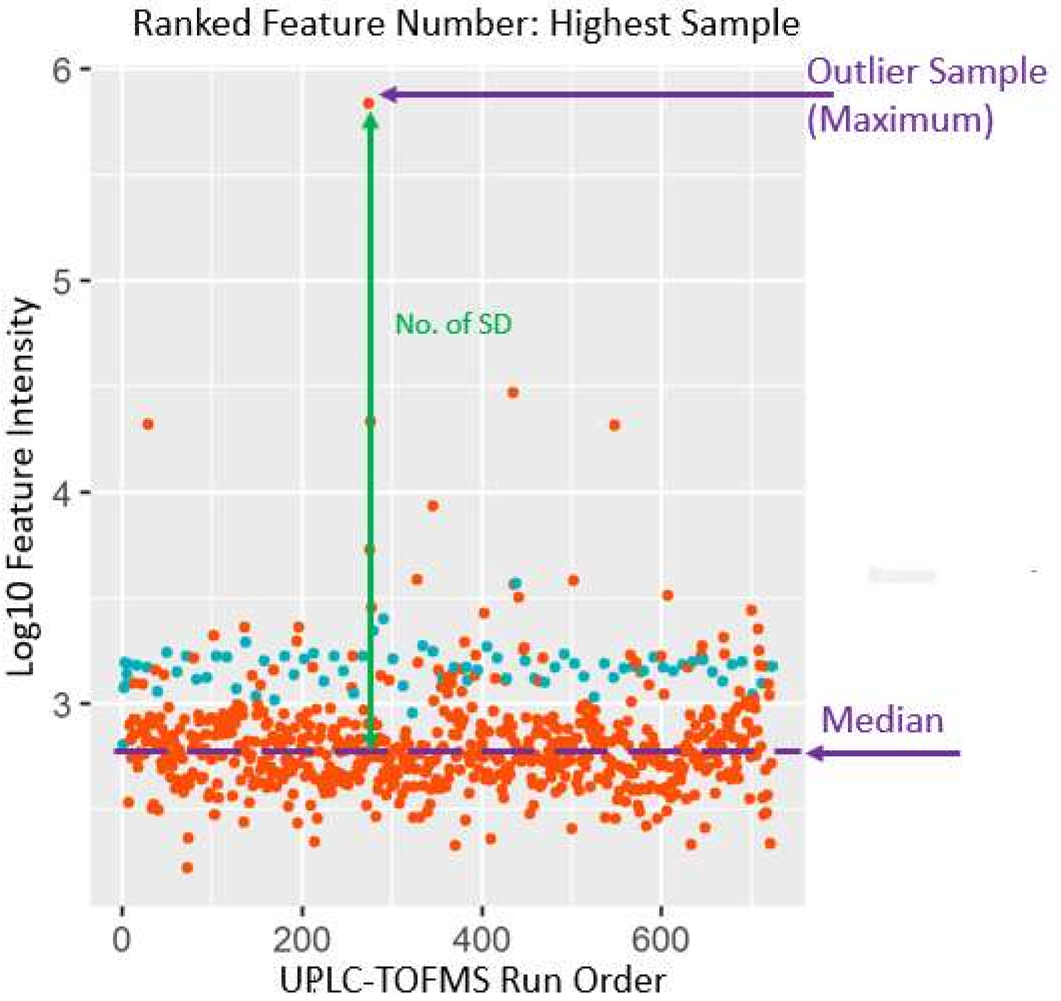


Figure 3-1. Output from the outlier sample script on an exemplar dataset. The figure is illustrating an outlier study sample for a single spectral feature with a significantly elevated feature intensity when compared to the median feature intensity across all study sample values (orange scatter points). In this instance, the outlying feature measurement was validated as biochemically relevant (not an artefact of data pre-processing or technical error) due to the elevated levels observed in the study reference samples (represented by the cyan coloured scatter points).

3.3.6.2 Intrasample and intersample correlation of mass spectral features

A correlation-based approach similar to STOCSY (Chapter 2) was used to identify chemicals that are metabolically and structurally associated with features of interest (herein referred to as “driver” feature, a term taken from STOCSY).

The correlation is carried out in two parts and highlights two aspects in relation to the driver:

- Structural correlates, from an intrasample correlation
- Biological correlates, from an intersample correlation

An intrasample correlation highlights only correlated features above an empirical correlation coefficient threshold of 0.8 or higher and restricted to a retention time window within 0.02 minutes of the driver. The analysis is undertaken using a sample with the highest measured signal for the driver (provided signal saturation was not observed). The Pearson method was used as its more suitable for identifying linear relationships between features within a sample. A two-dimension (2D) pseudo spectral peak is subsequently produced *via* an in-house R script, that allows peak shape to be further examined. The intersample correlation analysis (using spearman correlation as the default) involves all study samples from ALZ, where correlations are undertaken between the intensity pattern of the driver feature, to the intensity pattern of all remaining features. Embedded in the R script used to carry out the intersample correlation analysis, is the ability to evaluate feature distribution and intensity threshold settings. If the feature of interest exhibits a multimodal distribution in the data, multi-component Gaussian mixture models (GMMs) can be specified, placing clusters across the distributions. Once fitted, conversion of the distributions to probability distribution functions (PDF’s), can be obtained, and any sample with a probability (pr_n) of more than 0.90, assumes the classification for a specific distribution. The modality of the distribution observed in the dataset for a feature, can be an indication of exposure, which is further explored in chapter. Another feature of the code is multiple testing correction. Multiple testing correction (False Discovery Rate – Benjamini and Hochberg procedure) was applied, with a significance level cut-off of p_{adj}≤0.05, highlighting only statistically significant features.

The utility of this two-correlation analysis was exemplified using the driver feature that correspond to the molecular parent ion (M+H) of a prevalent xenobiotic, Acetaminophen (APAP). The metabolic fate for APAP has been extensively studied and well documented. The four main APAP metabolites (**Figure 3-2**) commonly observed in urine include, glucuronide, sulfate, cysteine and N-acetyl cysteine conjugates (Johnson and Plumb, 2005). Reference standards for APAP and the four metabolites have been prepared and analysed according to the databasing protocol stated in this chapter.

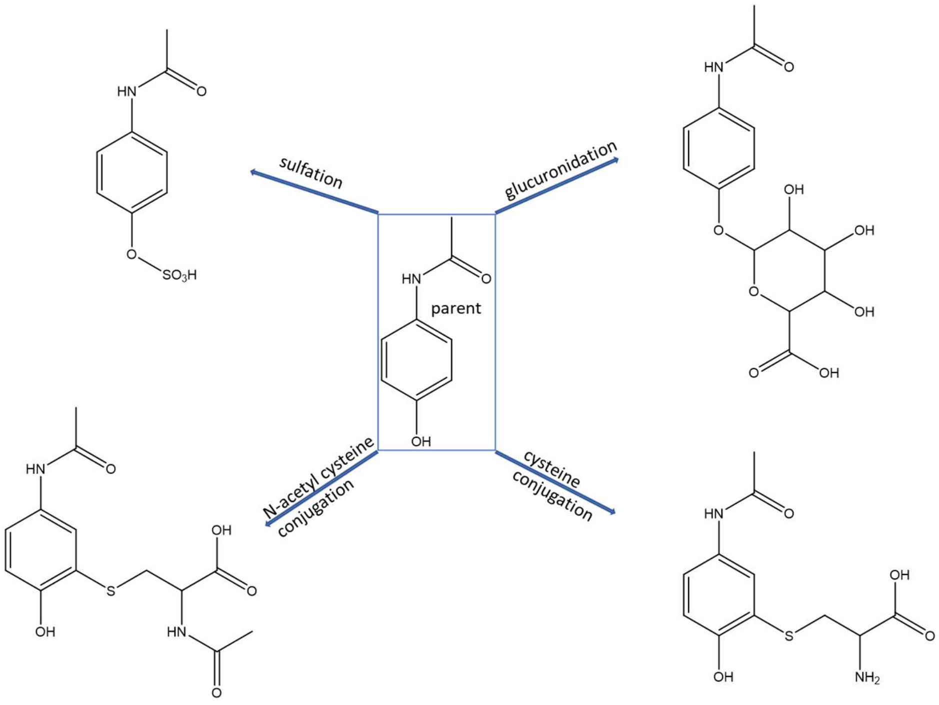


Figure 3-2. Known metabolites of acetaminophen observed in human urine (Johnson and Plumb, 2005).

3.3.6.3 *i-STOCSY*

Another correlation-based analysis was explored using a newly developed i-STOCSY tool (Zenodo DOI: https://doi.org/10.5281/zenodo.3886468, available from: https://github.com/phenomecentre/ISTOCSY). Briefly, it functions in the exact same way as the intersample correlation analysis, where a threshold value can also be inputted, but does not examine feature distribution or apply multiple testing. This tool can be accessed *via* a graphical interface that allows the driver feature to be selected interactively, yielding a rapid display of related correlation plots.

The compliance dataset and the ALZ profiling dataset were together uploaded, and the i-STOCSY tool was used to find correlations between the variables (MS features) from the profiling dataset, and variables (which are the medications) from the compliance dataset, based on matching samples (**Figure 3-3**). All figures produced are interactive and correlated features are coloured by the strength of the correlation. As this was an ALZ study, there were many patients on the ALZ drug memantine (MEM) (n=78), and therefore this medication was used to exemplify the i-STOCSY application. A correlation coefficient above 0.7 (Spearman) was set and the driver was the variable corresponding to MEM.

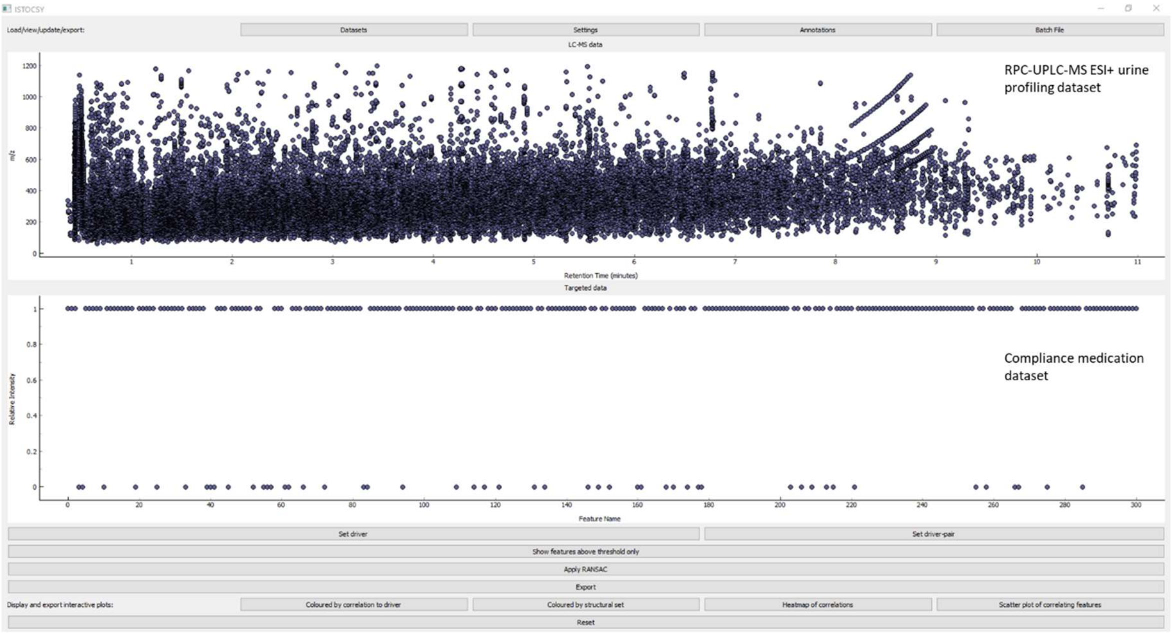


Figure 3-3. Typical example of the initial graphical interface of the i-STOCSY tool. The bottom figure is a scatter plot representing all drugs reported from patients (compliance) in the ALZ study. The tool automatically assigns each drug as either 1 or 0 (y -axis). The x-axis is the drug ID number. The top figure is a *m/z* (y-axis) vs retention time (x-axis) output from the ALZ profiling dataset. Correlations can be carried out by essentially clicking on the scatter points in either the top (RPC-UPLC-MS ESI+ urine profiling dataset) or bottom (compliance medication dataset) figures.

MEM has been reported to undergo metabolism *via* hydroxylation, N-oxidation and glucuronide conjugation (P S et al., 2014) as illustrated in **Figure 3-4**. Unfortunately, no reference standard was available for purchase.

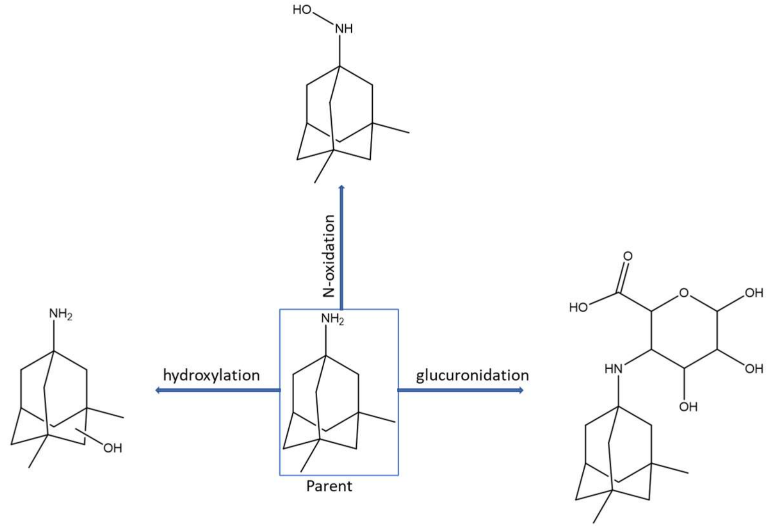


Figure 3-4. Known metabolites of memantine observed in human urine (P S et al., 2014).

3.3.6.4 Logistic regression

The medication donepezil was used to exemplify the logistic regression application. A total of 68 subjects reported use of the ALZ drug, donepezil hydrochloride (DNP – marketed under the brand name Aricept). DNP (chemical name 2-[((1-benzylpiperidin-4-yl)methyl)]-5,6-dimethoxy-2,3-dihydoinden-1-one monohydrochloride), is an acetylcholinesterase inhibitor used in the management of dementia in Alzheimer's disease and its metabolic fate has been extensively studied in humans (Sugimoto et al., 1990). Typical metabolic products of DNP found in urine are summarised in **Figure 3-5**.The 68 samples with a known exposure to DNP (based on the compliance meta-data) was assigned as the "high" exposure group (defined as 1 – case) and a random subset of an equal number of samples (n=68) was assigned as the "low" exposure group (defined as 0 – control).

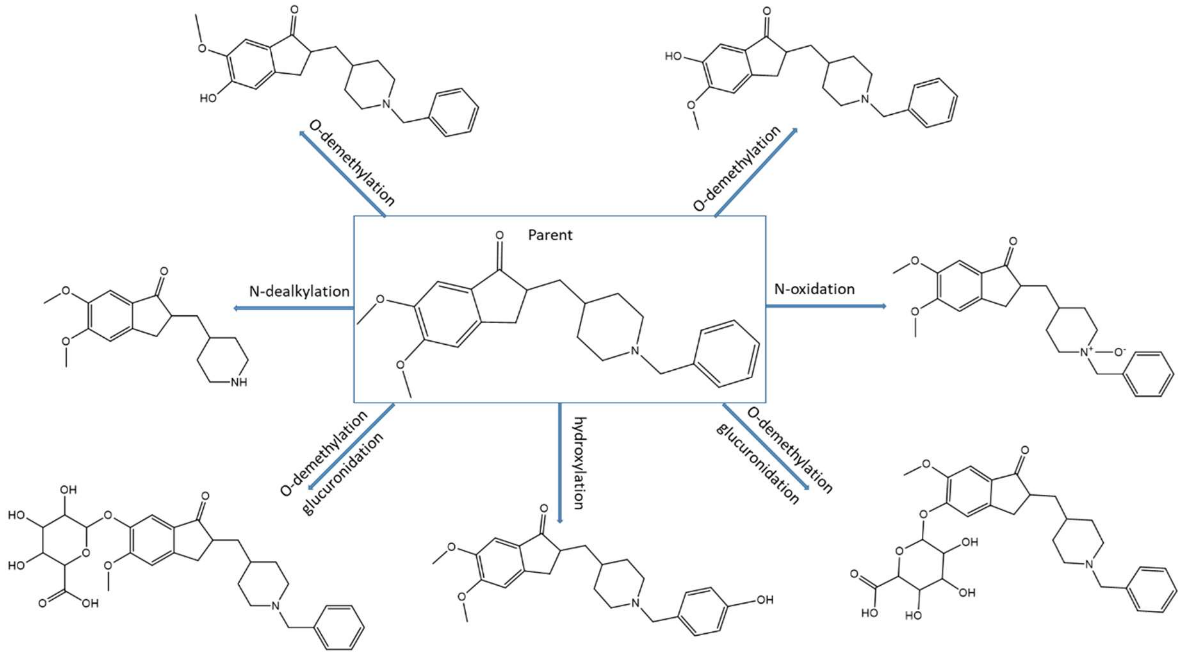


Figure 3-5. Known metabolites of donepezil observed in urine (human and animals) (Matsui et al., 1999).

Logistic regression (LogReg) was used to identify other spectral features (model variables) with the strongest associations to DNP. As summarised in Chapter 2, LogReg is a prediction mathematical modelling approach that is used to describe the relationship of several predictor variables X_1, X_2, …, X_n to a dichotomous dependent variable Y. The two exposure groups were further partitioned into training and test sets where selection of discriminant variables was conducted on the training set and the performance validated on the test set. From the zero group, 80% of the samples were assigned to a training set. Similarly, 80% of samples from the high group were selected and assigned to the same training set to maintain the same ratio between zero and high groups in the training and test sets. The remaining 20% of samples from each group were combined and assigned to the test set. The 80:20 split, incorporated a Euclidean distance metric and was undertaken using the DUPLEX algorithm in the "*prospectr*" package (Ramirez-Lopez, 2020) (version 0.2.0) in R.

Both univariate and multivariate LogReg models were calculated on the training set to predict case/control status. Multivariate models, notably Ridge, LASSO and Elastic Net (EN), were investigated to see how metabolites together relate to DNP exposure, whereas univariate models described the contribution for each feature individually. In the ridge model, the most important regression coefficients were found by bootstrapping and resampling the data with 500 iterations. This produced confidence intervals for the regression coefficient of each feature, and feature selection was based on intervals which did not include zero. LASSO and EN models have implicit variable selection as part of the regularization, therefore the important variables are those with non-zero coefficients as these can be directly derived from the model. The regularization parameters alpha and lambda were either set at a default value or tuned depending on the regularization model. In Ridge models, alpha is set as zero whereas for LASSO, alpha is 1. Lambda for both models are tuned using 10-fold cross-validation. With EN models, both alpha and lambda were tuned using the *"caret"* package (Kuhn,M. 2008) in R. This involved a grid of lambda (0.1 to 10, with 100 intervals) and alpha (0 to 1, 10 intervals) for which the optimal model was estimated using 10-fold cross-validation. As these penalised regression methods are multivariate, data was centred, and unit variance scaled prior to regression.

For the univariate logistic regression models, False Discovery Rate (Benjamini and Hochberg procedure) was used to adjust for multiple testing and applied on p-values (p_{ad}) derived from the model. Feature selection was then based on a cut-off, (p_{adj})<0.05, defining statistical significance. Of these significant features, three different training set models were produced:

- Univariate Model 1: A model that uses all significant features;

- Univariate Model 2: A model that includes only the significant features that correspond to the [M+H]+ ion of a reported metabolite (within 3ppm);

- Univariate Model 3: A model where all significant features were subjected to a backward elimination procedure. Backwards elimination iteratively removes one variable at a time and recalculates the model, it stops when no increase in the performance is observed following the removal of more variables. The Akaike information criterion (AIC) was used to select the optimal model, where a lower AIC is optimal.

In total there are 3 univariate models and 3 multivariate models. The accuracy, calculated from a confusion matrix and the AUC (area under the curve) from Receiver operating characteristic (ROC), was used to assess the predictive ability from each of the six models on a new set of data, i.e. the test set. A ROC curve was generated by plotting the true positive rate (TPR; sensitivity) against the false positive rate (FPR: 1-specificity) at various default threshold settings. The accuracy gives an indication of how much a model is capable of distinguishing between classes. A model with good predictive ability should have an accuracy and AUC closer to 1. A reference standard for DNP has been prepared and analysed according to the databasing protocol stated in this chapter.

3.3.6.5 PLS regression and discriminant models

The medication amlodipine was used to exemplify the PLS applications. A reference standard has been acquired for amlodipine and is part of the xenobiotic database, as such, retention time and main ion type are known. Based on this information, the distribution of the feature corresponding to amlodipine observed in the ALZ dataset, was assessed using the method described for the intersample correlation. PLS-R modelling was used to identify statistically significant covariation between a set (X) of independent variables (MS features) and the corresponding (Y) response (feature corresponding to the molecular ion of amlodipine). All study samples were used for PLS-R. For the PLS-DA model, samples were split into two exposure groups based on the feature distribution assessed for amlodipine. Two PLS-DA models were evaluated, one where exposure groups were unbalanced (therefore using all study samples in the dataset), and a second model where exposure groups were balanced. The models were performed using SIMCA (Version 15 Sartorius Stedim Biotech, Malmö, Sweden). After mean centering and Pareto scaling of the variables, the quality of the OPLS-DA models were validated by a seven-fold internal cross validation, assessment of the variance (R2Y) and predictive ability (Q2Y) of the model, and permutation tests (n=999). The appropriate number of components were selected for each model in order to optimise model quality without over-fitting. Discriminant features were evaluated based on variable importance for the projection (VIP) values greater than 2.

In a recent publication on amlodipine metabolism analysed by LC-MS/MS, metabolites observed in urine included a ketone metabolite, oxidised metabolite, oxidised metabolite which has undergone glucuronidation and, a carboxylic acid metabolite, as illustrated in **Figure 3-6** (van der Hooft et al., 2016).

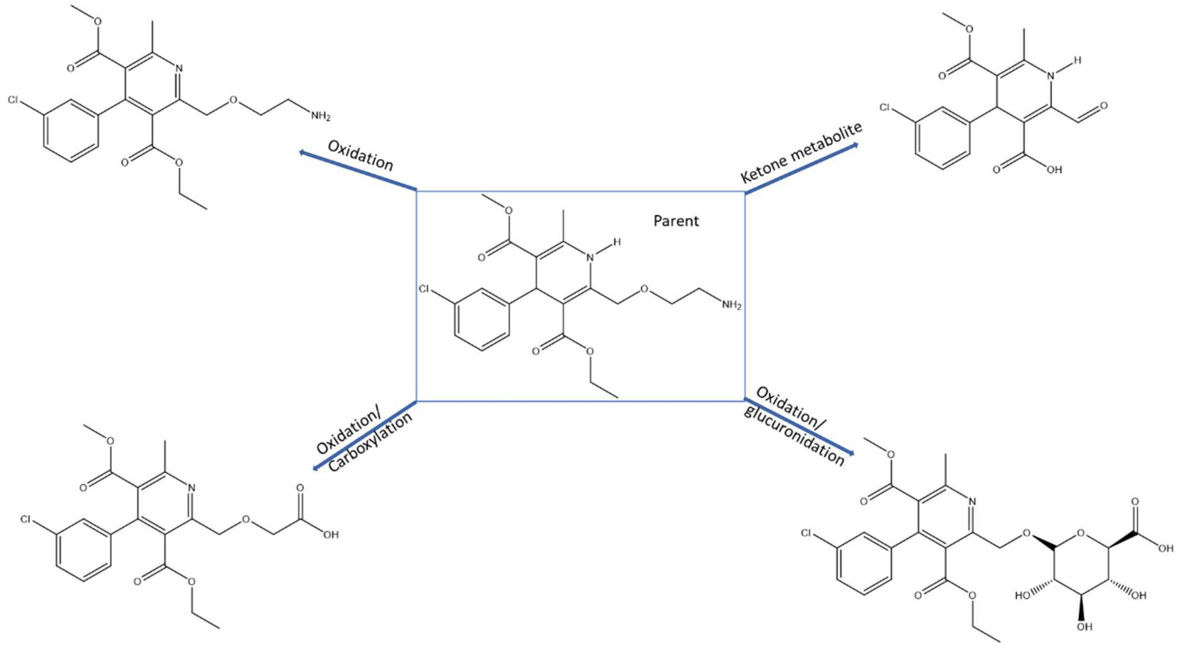


Figure 3-6. Known metabolites of amlodipine observed in human urine (van der Hooft et al., 2016).

3.4 Results and Discussion

3.4.1 *Application of the xenobiotic database for rapid annotation in ALZ*

A total of 25 chemical reference standards was initially acquired to populate the xenobiotic database since the start of this thesis. There are currently 41 reference standards and 57 pharmaceutical medications have undergone the acquisition and processing workflow stated in the methods. 35 of the reference standards showed evidence of a peak, and a molecular ion corresponding to the xenobiotic. The xenobiotic database was constructed as illustrated in the screenshot displayed in Figure 3-8. Examples of the 2D plots is illustrated in **Figure 3-7** using the reference standards APAP, Lansoprazole and Escitalopram. From the database, a total of 31 xenobiotics were detected in the ALZ dataset, based on comparison of retention time and spectral data, and their approximate population prevalence estimated (**Figure 3-9**). Detected xenobiotics from the ALZ dataset were based on peaks targeted by peakpantheR, and above an arbitrary intensity threshold. The threshold was specific to each compound and had to demonstrate a minimum signal to noise (S/N) > 5, and an elution time of within 15 seconds (as defined by peakPantheR) when compared to the reference standard. Of the xenobiotics detected in ALZ, population prevalence was calculated as the proportion of samples exhibiting the xenobiotic as a percentage of the total study samples in ALZ. Of the 35 refence standards detected by RPC-UPLC-MS methods, 31 xenobiotics were detected in the ALZ urine dataset.

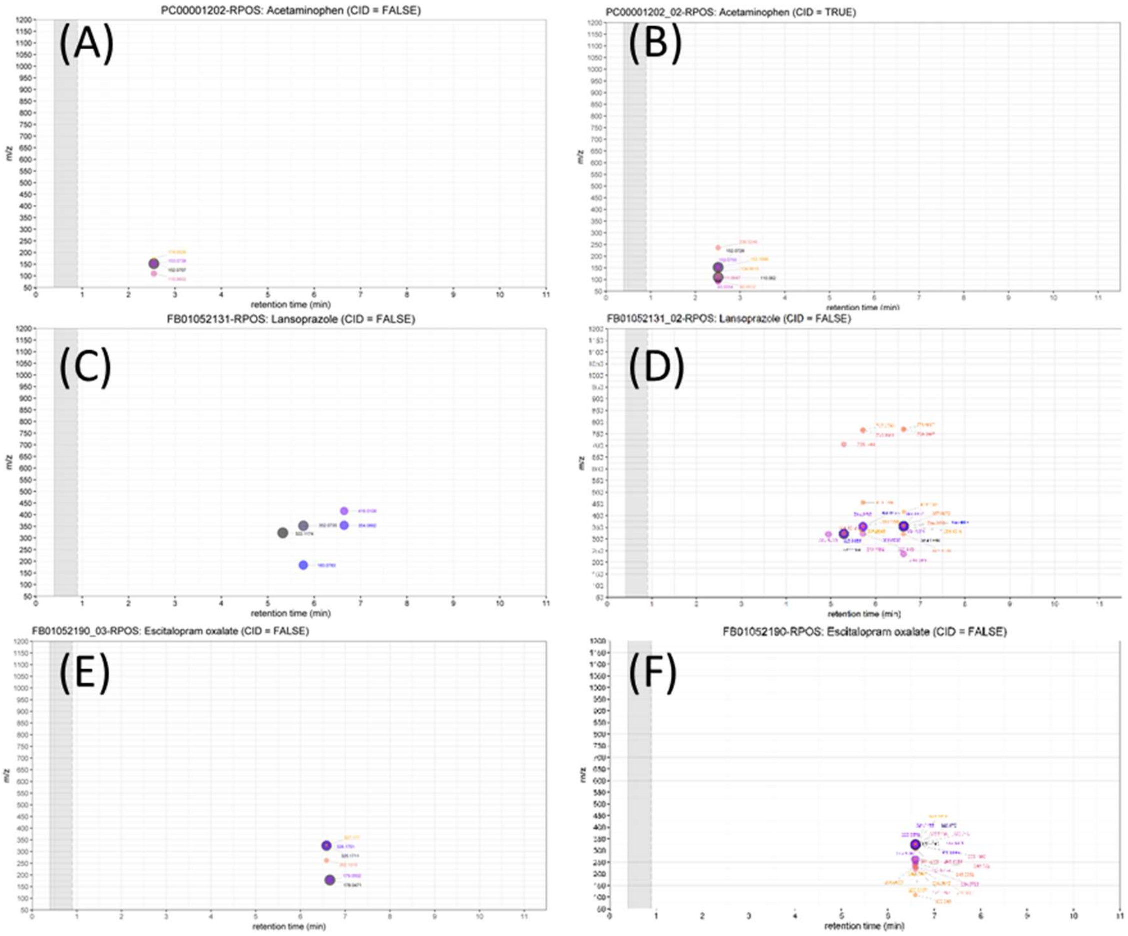


Figure 3-7. Example of the 2D outputs obtained for unique features detected by the reference standard workflow for databasing described in section 3.3.5. The x-axis on these 2D outputs represents the retention time (min) and the y-axis represents the *m/z.* The scatter points are the unique features detected by the workflow and are coloured by intensity (yellow – low intensity to purple – high intensity).

(A) acetaminophen at a low collision energy;

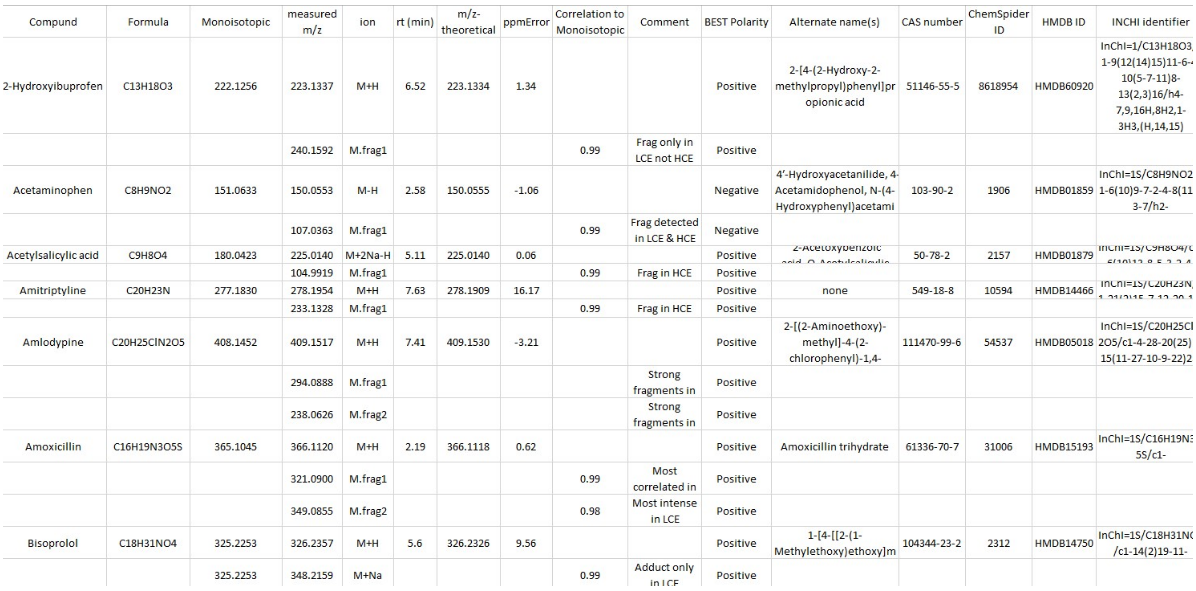
(B) acetaminophen at a high collision energy;

(C) first acquisition instance of lansoprazole was too dilute, and so needed re-acquisition;

(D) second acquisition of lansoprazole at the higher concentration (1:10 dilution);

(E) first acquisition of escitalopram at a concentration above the threshold range;

(F) second acquisition of escitalopram at a lower concentration (1:1000 dilution).



Compund	Formula	Monoisotopic	measured m/z	ion	rt (min)	m/z-theoretical	ppmError	Correlation to Monoisotopic	Comment	BEST Polarity	Alternate name(s)	CAS number	ChemSpider ID	HMDB ID	INCHI identifier
2-Hydroxyibuprofen	C13H18O3	222.1256	223.1337	M+H	6.52	223.1334	1.34			Positive	2-[4-(2-Hydroxy-2-methylpropyl)phenyl]propionic acid	51146-55-5	8618954	HMDB60920	InChI=1/C13H18O3/ 1-9(12(14)15)11-6-4-10(5-7-11)8-13(2,3)16/h4-7,9,16H,8H2,1-3H3,(H,14,15)
			240.1592	M.frag1				0.99	Frag only in LCE not HCE	Positive					
Acetaminophen	C8H9NO2	151.0633	150.0553	M-H	2.58	150.0555	-1.06			Negative	4′-Hydroxyacetanilide, 4-Acetamidophenol, N-(4-Hydroxyphenyl)acetami	103-90-2	1906	HMDB01859	InChI=1S/C8H9NO2 1-6(10)9-7-2-4-8(11 3-7/h2-
			107.0363	M.frag1				0.99	Frag detected in LCE & HCE	Negative					
Acetylsalicylic acid	C9H8O4	180.0423	225.0140	M+2Na-H	5.11	225.0140	0.06			Positive	2-Acetoxybenzoic acid, O-Acetylsalicylic	50-78-2	2157	HMDB01879	InChI=1S/C9H8O4/c 6(10)13-8-5-3-2-4-
			104.9919	M.frag1				0.99	Frag in HCE	Positive					
Amitriptyline	C20H23N	277.1830	278.1954	M+H	7.63	278.1909	16.17			Positive	none	549-18-8	10594	HMDB14466	InChI=1S/C20H23N/ 1-21(2)15-7-13-20-1
			233.1328	M.frag1				0.99	Frag in HCE	Positive					
Amlodypine	C20H25ClN2O5	408.1452	409.1517	M+H	7.41	409.1530	-3.21			Positive	2-[(2-Aminoethoxy)-methyl]-4-(2-chlorophenyl)-1,4-	111470-99-6	54537	HMDB05018	InChI=1S/C20H25Cl 2O5/c1-4-28-20(25) 15(11-27-10-9-22)2
			294.0888	M.frag1					Strong fragments in	Positive					
			238.0626	M.frag2					Strong fragments in	Positive					
Amoxicillin	C16H19N3O5S	365.1045	366.1120	M+H	2.19	366.1118	0.62			Positive	Amoxicillin trihydrate	61336-70-7	31006	HMDB15193	InChI=1S/C16H19N3 5S/c1-
			321.0900	M.frag1				0.99	Most correlated in	Positive					
			349.0855	M.frag2				0.98	Most intense in LCE	Positive					
Bisoprolol	C18H31NO4	325.2253	326.2357	M+H	5.6	326.2326	9.56			Positive	1-[4-[[2-(1-Methylethoxy)ethoxy]m	104344-23-2	2312	HMDB14750	InChI=1S/C18H31N(/c1-14(2)19-11-
		325.2253	348.2159	M+Na				0.99	Adduct only in LCE	Positive					

Figure 3-8. Screenshot of the constructed xenobiotic reference standard database.

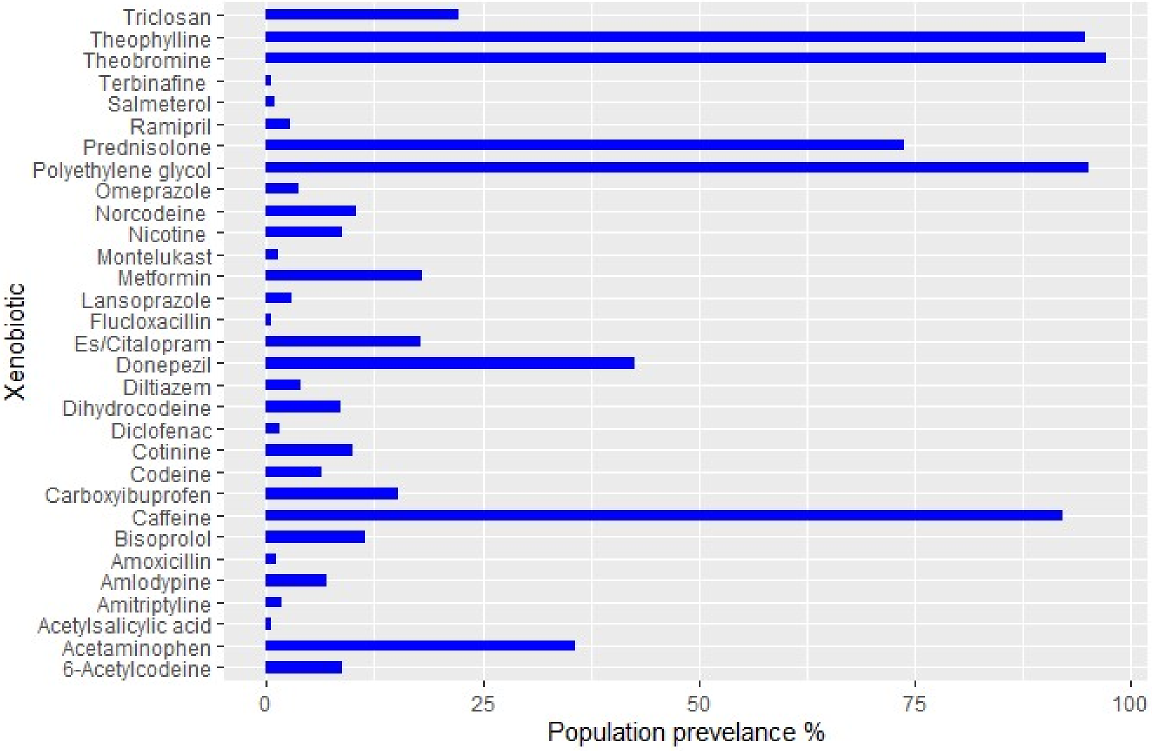


Figure 3-9. A horizontal bar chart representing the reported prevalence of xenobiotics detected in the ALZ cohort from data acquired by RPC-UPLC-MS (positive and negative ion mode). Detected xenobiotics from the ALZ dataset were based on peaks targeted by peakpantheR, and above an arbitrary intensity threshold. The threshold was specific to each compound and had to demonstrate a minimum signal to noise (S/N) > 5, and an elution time of within 15 seconds (as defined by peakPantheR) when compared to the reference standard. Of

the xenobiotics detected in ALZ, population prevalence was calculated as the proportion of samples exhibiting the xenobiotic as a percentage of the total study samples in ALZ. 31 xenobiotics were detected.

3.4.2 *Statistically based methods to identify xenobiotic signatures*

3.4.2.1 Outlier detection

An approach was undertaken to identify features with characteristic distributions across the sample set that are likely to reflect exposure in a minority of individuals (i.e. a potential exposure pattern for xenobiotics in a small proportion of volunteers, or non-compliant study participants). The first five ranked features (in the ALZ cohort) from this approach related to xenobiotics which were then later identified with reference standards.

Of the first five features, four different features indicated one specific outlier sample at the same retention time. Examining the spectrum for this sample, revealed a polymeric pattern, which is easily recognisable in UPLC-MS as polyethylene glycol (PEG), appearing as an envelope of repeating signals separated by approximately 44.026 Da (**Figure 3-10.A**). The PEG form was later identified as PEG3350 *via* a reference standard (As part of the investigation in Chapter 5, PEG forms which are commercially available were purchased and analysed by RPC-UPLC-MS, positive and negative ion mode). PEG3350 is frequently used as an excipient in many liquid and solid medicinal formulations (Stone et al., 2019).

The second ranked feature (**Figure 3-10.B**), after the PEG features (the 5th ranked feature), was initially putatively annotated (MSI level 2) as flucloxacillin (FLX). The putative annotation was undertaken using on-line spectral libraries, such as PubChem (Kim et al., 2018), HMDB (Wishart et al., 2017) and Metlin (Guijas et al., 2018). In this instance, the HMDB database indicated FLX as a potential suspect based on matches (within 3ppm) of this particular feature, to the theoretical molecular ion of FLX (454Da) and two fragment ions (160Da and 295Da) under positive ionisation conditions (Larmene-Beld et al., 2014). Additional evidence further confirming a positive assignment to FLX, was the presence in the mass spectrum of an isotopic distribution associated with halogen atoms, specifically chlorine. A reference standard was later purchased, and subjected to the protocol of refence standard acquisition, confirming its identity.

The presence of the outlying features associated with these two xenobiotics, was corroborated by an observed effect on the average pooled sample (SR), resulting in its elevation from the median of the study sample distribution. The calculation utilising the median is more effective than using the average (mean), as the median is robust to the presence of these outlier samples and thus better represents the amount observed in the remainder of the population. That being said, as samples with the exposure feature increases in number, the median becomes less representative of the other samples in the study, and so we are less able to use the resulting pattern to identify potential drug exposures. Therefore, this works best for less common xenobiotics and affected samples being in the minority. In addition, these datasets here were processed with XCMS which incorporates algorithms such as minimum fraction filters to look for valid features present in a minimum number of samples within a sample group. With ALZ, a minfrac setting of 0.4 was set, as per NPC protocol for phenotyping data. This means the feature must be present in at least 40% of samples to be included for the final dataset. Ideally a dataset with no in-built filtering function would work best for this scenario, however if a setting like this was applied to a typical metabolomic dataset, the end result could computationally be problematic.

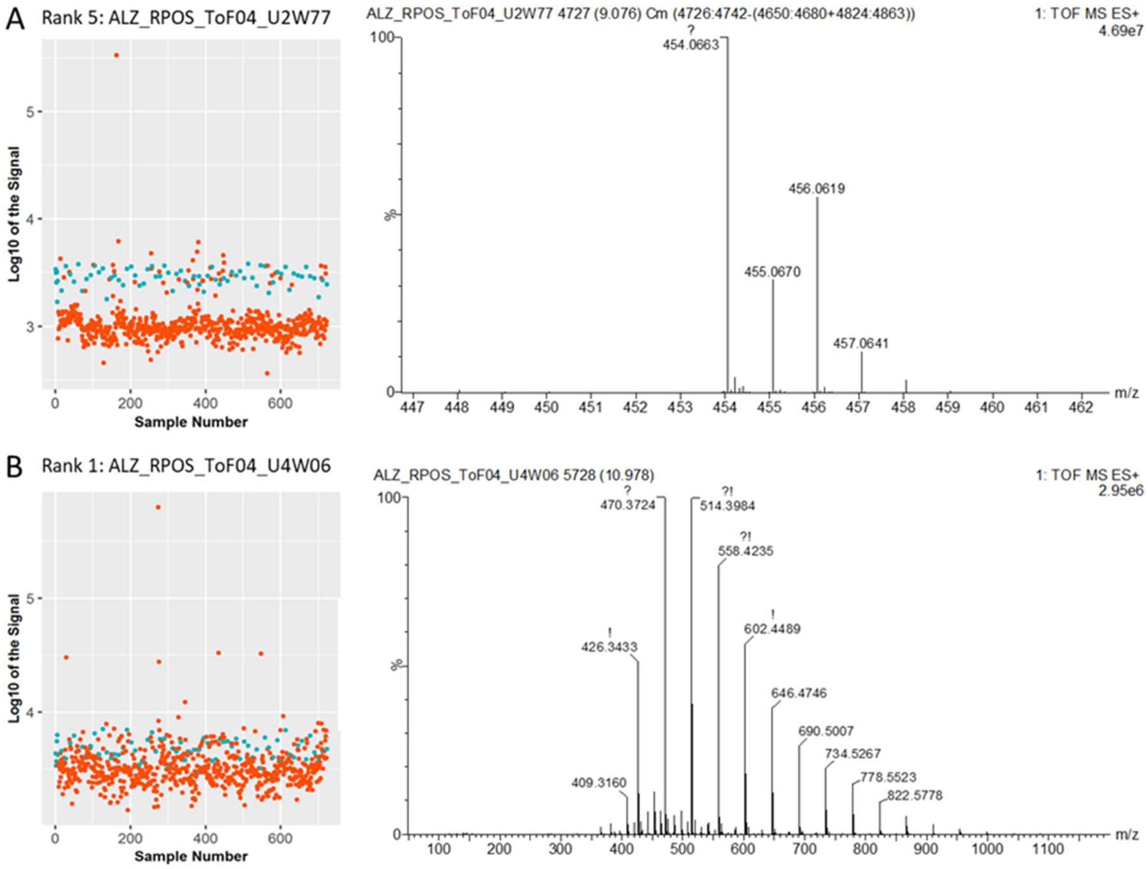


Figure 3-10. Two examples of samples (from urine ALZ RPC-UPLC-MS positive ion mode data) flagged from the outlier script. The two samples have an elevated amount of a feature and are ranked based on the greatest difference (number of standard deviations) between median value of all study samples and a sample with the highest recorded intensity. **A**. Feature is ranked number five and corresponds to the xenobiotic flucloxacillin. **B**. Feature is ranked number one and corresponds to PEG3350.

3.4.2.2 *Intrasample and intersample correlation analysis;*

A two-correlation analysis was undertaken to find features that correlate to the molecular ion $[M+H]^+$ corresponding to APAP-unmetabolised. The retention time of the ion was confirmed by comparison to the reference standard. The intrasample correlation successfully highlighted correlates which were structural, i.e. isotopes, adducts or in-source fragments (**Figure 3-11.C**). For the intersample correlation analysis, when all study samples were used for the correlation, a high number of features were observed to be statistically significant, making interpretation extremely difficult. The high multicollinearity between features, which are typical of UPLC-MS datasets is the most likely reason for this. Assessment of the APAP-unmetabolised molecular ion in the ALZ dataset revealed a bimodal distribution as illustrated in **Figure 3-11.B**.

The intersample correlation was therefore carried out on the samples that occupied Distribution 2 (pr_2), i.e. any sample with $pr_2 > 0.90$, as indicated by the green vertical dotted line in **Figure 3-11.B.** The samples with $pr_2 > 0.90$, exhibited a detectable level of APAP-unmetabolised, i.e. a minimum S/NR ≥ 10 in the correct elution region. The LOD was not evaluated for APAP-unmetabolised, however it does seem that the samples within distribution 1 (pr_1-red) exhibited negligible levels. The intersample correlation revealed several statistically significant correlated feature clusters as illustrated in **Figure 3-11.A**. The blue scatter points represent all statistically significant features ($p_{adj} \leq 0.05$), whereas the green represent features which are statistically significant and have correlation coefficients greater than 0.9. The predicted $[M+H]^+$ for reported metabolites of APAP were detected with statistically significant correlations (spearman, $p_{adj} < 0.05$) to the APAP-unmetabolised driver (**Table 3-3**). A hydroxyl sulfate conjugate was additionally annotated based on a predicted molecular ion match to within 5 ppm.

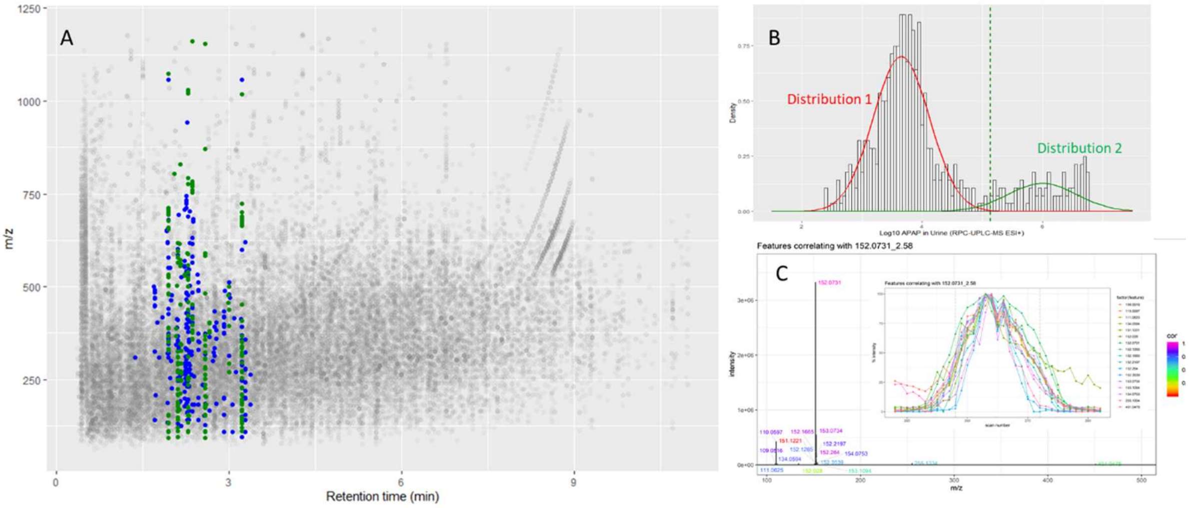


Figure 3-11. Two- dimensional representation of the intersample corelation analysis (A), feature distribution assessment (B), and intrasample correlation analysis (C). (A) A retention time (min) by *m/z* plot representing a Correlation analysis of the intensity pattern for the molecular ion corresponding to APAP-unmetabolised across all samples from distribution 2, to the intensity pattern observed for all other features in the dataset. Statistically significant Spearman correlations (p_{adj}<0.05) was coloured by the blue scatter points. The green scatter points represent padj<0.05, and correlation coefficient > 0.9. The greyed-out scatter points represent all features detected in the ALZ dataset. (B) Gaussian mixture models (GMM's) were fitted to the log(base10) MS intensity distribution of APAP-unmetabolised in ALZ urine data, and the PDF's for each gaussians (distribution) were obtained. Any sample with pr > 0.90, or a log10 signal greater than the green dotted vertical line, was used for the intersample correlation analysis. (C) For the intrasample correlation analysis, an empirically derived value of >0.8 correlation was used to select only strongly correlated features restricted to 0.02 minutes of a specified retention time. Within (C), is a representation of the pseudo spectrum of all features with a correlation coefficient >0.8, and the extracted ion chromatograms for the same correlated features at the specified retention time (noted as scan number).

Table 3-3. APAP and metabolites, based on authentic reference standards, reported metabolites from literature, and prediction from the intersample correlation analysis.

Compound	RT (min) measured	$[M+H]^+$ (measured)	Intersample correlation	Identification/ Putative annotation	Reference standard RT (min)	[M+H]+ (theoretical)	ppm error
APAP (driver)	2.58	152.0702	1*	APAP-unmetabolised	2.58	152.0712	6.58
M1	2.3	232.0276	0.89*	APAP Sulfate	2.31	232.028	1.72
M2	2.13	168.0661	0.72*	APAP hydroxyl Sulfate	-	168.0655	3.57
M3	1.95	328.1023	0.87*	APAP Glucuronide	1.95	328.1032	2.74
M4	2.25	271.0747	0.69*	APAP Cysteine	2.29	271.0753	2.21
M5	3.23	313.0849	0.68*	APAP Mercapturic Acid	3.23	313.0858	2.87

padj<0.05

3.4.2.3 Correlation analysis using i-STOCSY

Another correlation-based analysis was explored using i-STOCSY, where the correlation was undertaken between samples from the compliance dataset, and matching samples from the RPC-UPLC-MS (positive ion mode) profiling dataset. The driver used, was selected from the compliance dataset, and corresponded to the xenobiotic MEM. Correlated features, as indicated in **Figure 3-12.A1**, corresponded to the accurate theoretical [M+H]+ of reported metabolites to within 3ppm. Features at *m/z* 196.1692, match the accurate mass of the M+H ion for N-oxidation and hydroxylation metabolites. There are many retention time clusters present, each corresponding to different isomeric formations of the hydroxylated/N-oxidation metabolite. Other features include an ion at *m/z* 356.2065 which potentially corresponds to the glucuronide metabolite of MEM, and two RT clusters with a *m/z* of 373.2002, which corresponds to a hydroxylated glucuronide of MEM. The heat map (**Figure 3-12.A2**) was particularly useful for identifying structural correlates within a feature cluster. Although hydroxylation and glucuronidation are common metabolic products of xenobiotic metabolism, the discovery of a hydroxylated glucuronide using i-STOCSY, to my knowledge has never been reported in human urine. Interestingly, Meantime has been reported to undergo very little metabolism, with the majority of an administered does being excreted unchanged in urine. MEM-unmetabolised was not observed in this dataset, indicating either other metabolites would be a better exposure marker than the unemtabolised compound, or it may have been filtered during pre-processing.

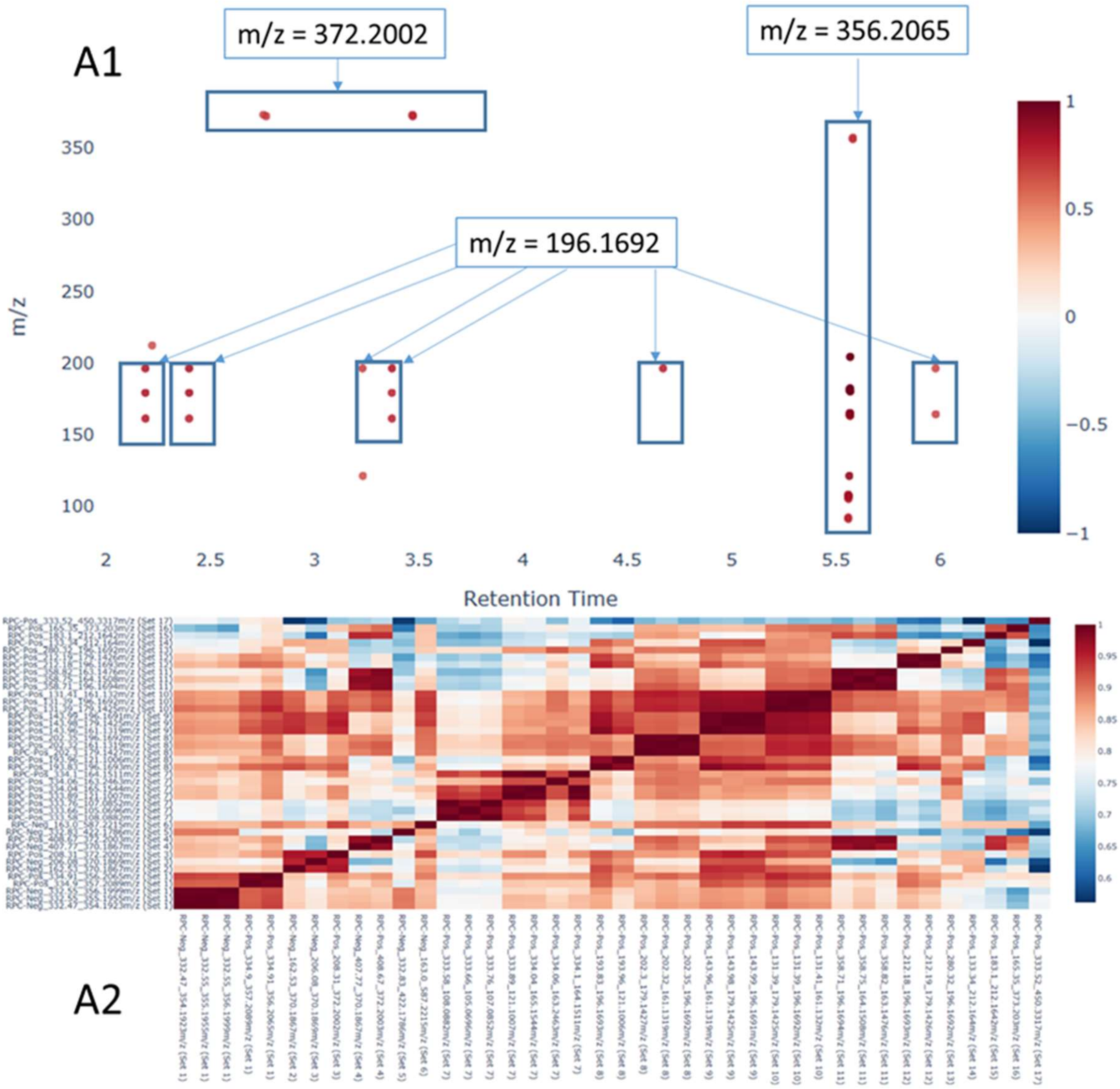


Figure 3-12. **i-STOCSY outputs expressed as a scatter plot and heatmap for amlodipine**.(A1*)* Represents all features in ALZ that correlate to the driver MEM. An empirically derived value of >0.8 correlation was used to select only strongly correlated features. Putative annotations were made from these features. Features at *m/z* 196.1692, correspond to N-oxidation and hydroxylation metabolites, a feature at *m/z* 356.2065 corresponds to the glucuronide metabolite of MEM, and two RT clusters with a *m/z* of 373.2002, corresponds to a hydroxylated glucuronide of MEM. (A2) is heatmap of all features above the 0.6 empirical correlation value, and in addition to the driver feature, represents the correlations observed to one another. This was particularly useful to identify structural correlates.

3.4.2.4 Logistic regression

The number of significant features and accuracy for all models are summarised in **Table 3-4**. ROC curves can be found in Appendix 1. A total of 189 features were statistically significant ($p_{adj} < 0.05$)

from univariate LogReg. From these features, the molecular ion for DNP-unmetabolised, and all reported metabolites, could be accounted for based on the theoretical $[M+H]^+$ ion that would be observed (**Table 3-5**). As no reference standard for the DNP metabolites are available for purchase, only a MSI level 3 putative annotation could be made (due to comparison of certain spectral ions to online databases and literature). The training set was used to fit three different univariate models (Univariate 1, Univariate 2 and Univariate 3) which was then implemented on the test set. The use of an automated algorithm to decide which features to include in the model (Univariate 3), demonstrated the poorest accuracy. The selection of all variables (Univariate 1) or selection of variables based on knowledge of DNP metabolism (Univariate 2) produced better models for DNP prediction with the latter being the most accurate. In multivariate LogReg regression, the results from the Ridge model, even after bootstrapping the regression coefficients, produced a large number of variables (n=11106), which complicated interpretation and feature selection. The LASSO model produced only two features, one was identified as the molecular ion of the DNP unmetabolised form, and the other putatively annotated as an in-source fragment ion of an O-demethylation product of DNP. In the EN model, 26 out of the 42 features corresponded to an ion associated with a metabolite of DNP. EN produced the best accuracy from the three multivariate models.

Table 3-4. Summarises the number of features that are statistically significant from both univariate and multivariate logistic regression models. The accuracy of these discriminant features to predict with new data was validated on the test set. The accuracy was determined using a confusion matrix and with ROC curve (AUC).

Model	Number of Features	Sensitivity	Specificity	Accuracy: AUC	Accuracy: Confusion Matrix
Univariate 1	189	0.77	0.69	0.73	0.73
Univariate 2	13	0.77	0.85	0.81	0.81
Univariate 3	16	0.69	0.62	0.65	0.65
Ridge	11106	0.61	0.77	0.67	0.65
LASSO	2	0.69	0.85	0.78	0.77
EN	43	0.69	0.92	0.79	0.81

Table 3-5. Donepezil and annotated metabolites which were significant from the univariate and multivariate LogReg models. The observed ions are all within 7ppm of the theoretical ion.

Proposed Annotation	Formula	$[M+H]^+$ (theoretical)	$[M+H]^+$ (measured)	ppm error
DNP-unmetabolised	$C_{24}H_{29}NO_3$	380.2226	380.2214	3.15
O-demethylation (isomer 1 and 2)	$C_{23}H_{27}NO_3$	366.2069	366.2043	7.10
O-demethylation glucuronide (isomer 1 and 2)	$C_{29}H_{35}NO_9$	542.2390	542.2385	0.92
N-oxidation	$C_{24}H_{29}NO_4$	396.2175	396.2173	0.50
N-dealkylation	$C_{17}H_{23}NO_3$	290.1756	290.1749	6.89
Mono hydroxylated metabolite	$C_{24}H_{29}NO_4$	396.2175	396.2173	0.50

The application of LogReg, was therefore successful in identifying features that best predict xenobiotic exposure (DNP), most of which corresponded to known metabolites from literature. The univariate LogReg model that used substantive knowledge of DNP metabolism produced the best model, however, is only applicable if metabolites are known. If the aim is to uncover new metabolites or even metabolites which are affected from endogenous pathways, then all features need to be examined. In addition to seeing how features together relate to an exposure, multivariate models can be investigated to mitigate the problem of multicollinearity and decrease model complexity, i.e., by exerting a penalty, as in the case of Ridge, LASSO and EN. Ridge will keep all variables (features) in the model, however the variables with a minor contribution to the outcome will have their coefficients closer to zero. Bootstrapping and resampling for an estimation of the regression coefficients did indeed reduced the number of variables in the model, however, was still difficult to interpret. LASSO and EN both have implicit variable selection, with the latter being less stringent. The EN model was a good compromise between Ridge and LASSO. The EN model predicted the best out of the three multivariate models, having the highest accuracy as estimated from the ROC curves. LogReg was successful in not only identifying the parent xenobiotic, but also known direct metabolites which are supported by the literature. Features not related to direct metabolism of DNP were also observed to be significant, suggesting that there is scope for this strategy to highlight other affected metabolites.

3.4.2.5 PLS regression and discriminant models

PLS regression and classification models were used to identify features relating to amlodipine exposure. A PLS-R model was firstly built and is based on the Y variable or output being continuous, i.e., the Y variable corresponds to the intensity profile for the molecular ion of amlodipine (amlodipine-unmetabolised) measured from all study samples. By contrast, PLS-DA models were also evaluated. As this is a discriminant analysis requiring groups or classes of samples, samples were classified into exposure groups based on the feature distribution assessment (as highlighted from the intersample correlation analysis). amlodipine-unmetabolised, like APAP, also exhibited a bimodal distribution in the data as illustrated in **Figure 3-13**. GMMs were implemented again, and samples occupying the green distribution, i.e., $pr_2 > 0.90$ or log10 signal greater than the vertical green dotted line, exhibited a feature corresponding to amlodipine-unmetabolised, with a minimum S/NR ≥ 10 eluting at an identical retention time to the reference standard from the database. A PLS-DA model was built, where samples were classified into two exposure groups, i.e., distribution 1 is the zero-low exposure group, and distribution 2 is the high exposure group. However, there were only 52 samples that occupied this high exposure group and 509 samples that occupied the low zero-low exposure group, so there could be an issue of bias related to unbalanced groups in a PLS-DA model. Therefore, a random equivalent number of samples were selected from the zero-low exposure group ($pr_1 > 0.90$ or less than the red vertical dotted line), and a secondary PLS-DA model with balanced groups was constructed. PLS loading plots were used to illustrate the findings and loadings highlighted in blue, exhibited VIP values ≥ 2 (**Figure 3-14**). The same features validated across all three PLS models. The R^2Y and Q^2Y values obtained with six calculated components were, 0.95 and 0.84 for the PLS-R model, 0.94 and 0.78 for the PLS-DA model (unbalanced), and 0.99 and 0.86 for the PLS-DA model (balanced), respectively. Permutation testing indicated low variability and an excellent predictive ability. The predicted $[M+H]^+$ for the reported metabolites of Amlodipine (which are represented by the red coloured circles), as well as ionisation products for these features, were the main features with VIP values ≥ 2 from the PLS loading plots. amlodipine-unmetabolised would be considered a MSI level 1 identification as there is reference standard, the other four metabolites would only be considered a MSI level 4 annotation. However, the mass spectrum of Amlodipine clearly shows two peaks indicative of an isotopic distribution pattern for a compound containing a single chlorine atom. This isotopic pattern was also observed for all the annotated metabolites which further adds confidence to their assignments. The oxidised-carboxylic acid metabolite was the most intense of all the annotated metabolites and was the most discriminant of all features (furthest away in the loadings plot). This agrees with the findings of Van der hooft *et al.* (van der Hooft et al., 2016), as they reported the same observation, indicating that it may be a better marker for exposure than the parent compound.

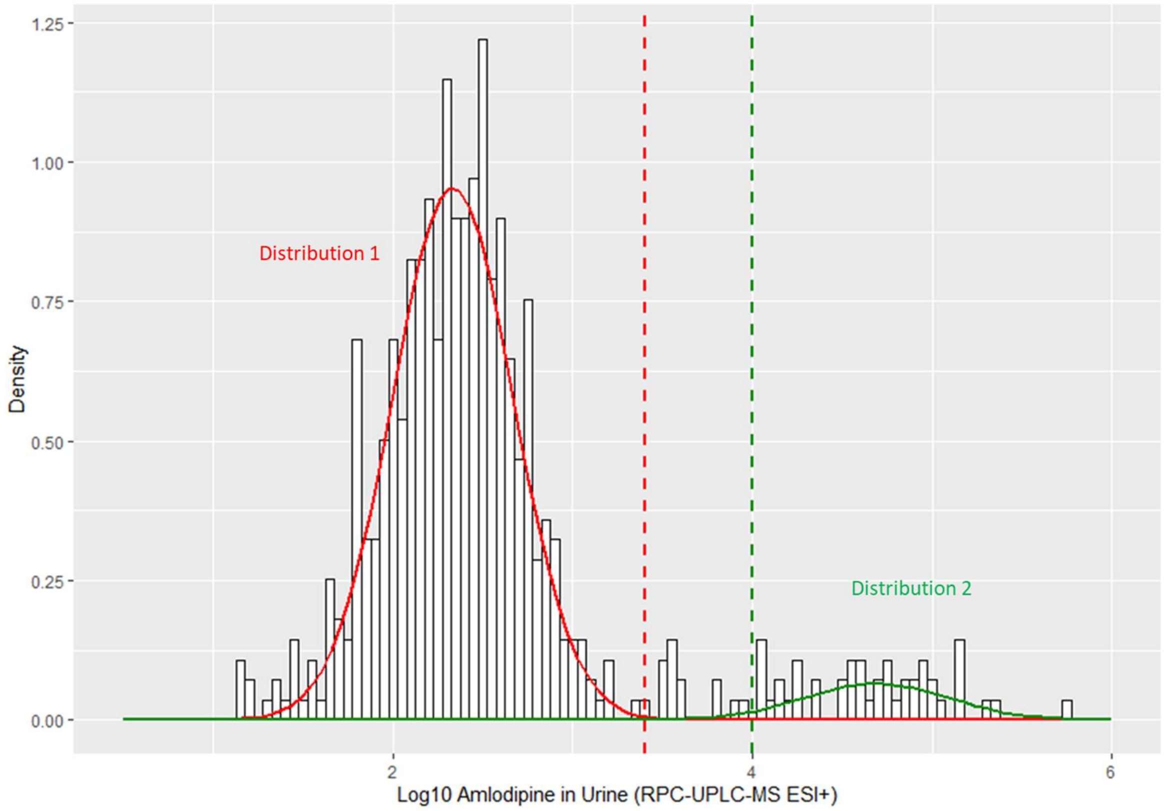


Figure 3-13. Gaussian mixture models (GMM's) fitted to the MS intensity distribution density plot of amlodipine-unmetabolised from ALZ urine data (RPC-UPLC-MS in positive ion mode). GMM's were fitted, and the PDF's for each gaussians were obtained, dividing the data to a High exposure group (Distribution 2, pr2 – green), and a Zero exposure group, (Distribution 1, pr1 -red). Any sample with pr1>0.90, or a log10 signal less than the red dotted line, assumed the classification of zero exposure, and any sample with pr2>0.90, or log10 signal more than the dotted green line, assumed the classification of high exposure.

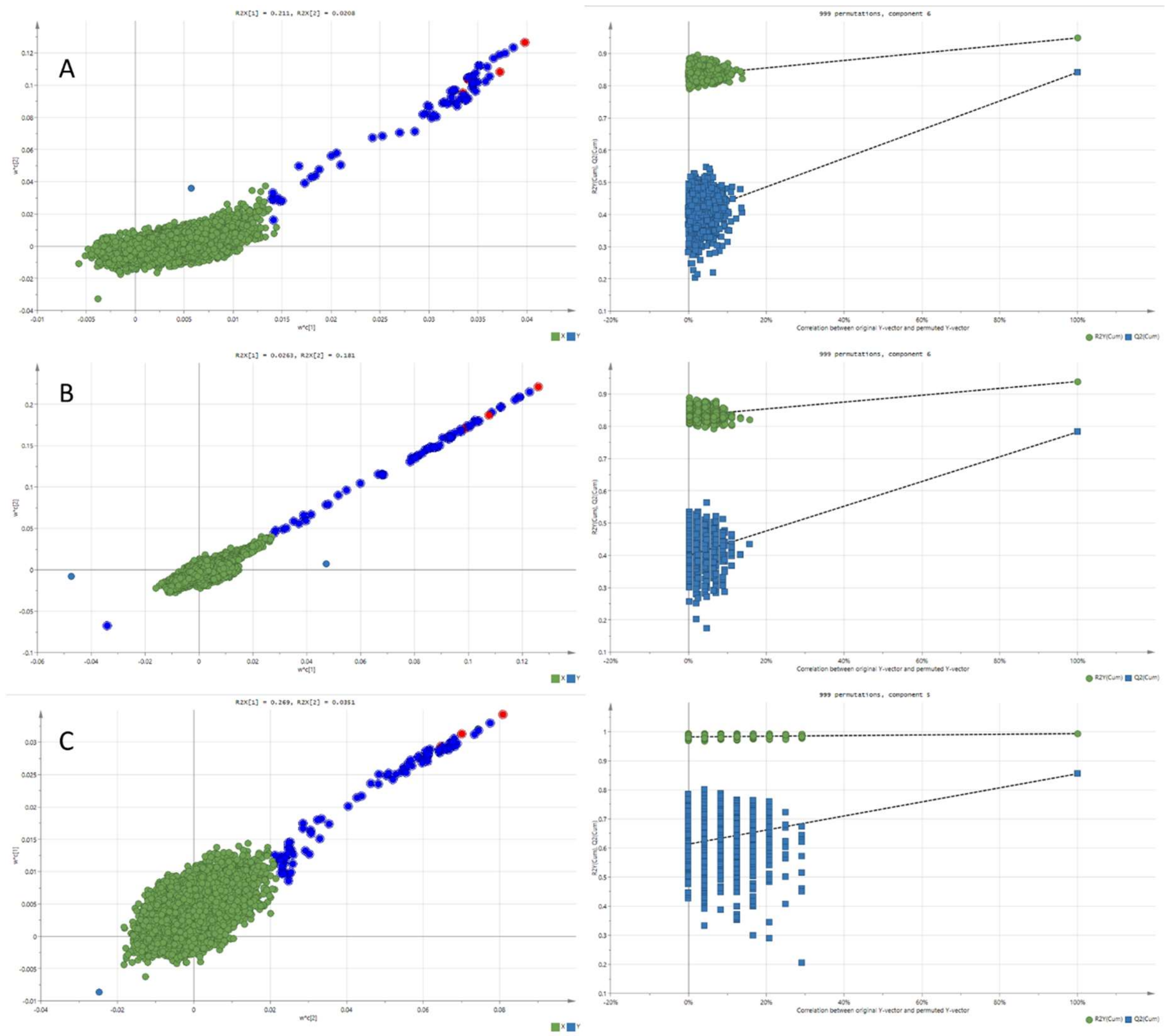


Figure 3-14. Six-component PLS loading plots showing the separation and discriminating features from the RPC-UPLC-MS (positive ion mode) profile observed between the high and zero-low exposure groups (with loadings coloured in blue exhibiting a VIP > 2) relating to the xenobiotic amlodipine. Loadings coloured in red are the molecular ions for reported metabolites of amlodipine, and loadings coloured in green are all remaining features in the dataset with VIP values < 2. Validation plots displaying 999 permutation tests are alongside the corresponding PLS models.

(A) The explained variance (R^2Y) was 0.95 and predictive ability was 0.84 for PLS-R model, where the Y variable is continuous, and is the intensity profile of the molecular ion corresponding to amlodipine-unmetabolised;

(B) The explained variance (R^2Y) was 0.94 and predictive ability was 0.78 for PLS-DA model (unbalanced), where the Y variable is categorical (i.e., 0 for control and 1 for case). Sample groups were classified and assessed based on the distribution of the molecular ion corresponding to amlodipine-unmetabolised. The control group consisted of 509 samples from the zero-low exposure group, and case group consisted of 52 samples from the high exposure group;

(C) The explained variance (R^2Y) was 0.99 and predictive ability was 0.86 for PLS-DA model (balanced), where the Y variable is categorical (i.e., 0 for control and 1 for case). Sample groups were classified and assessed based on the distribution of the molecular ion corresponding to amlodipine-unmetabolised. The control group consisted of 52 samples randomly selected from the low-zero exposure group, and case group consisted of all 52 samples from the high exposure group.

All three models highlighted approximately the same statically significant features (VIP > 2) relating to amlodipine exposure.

3.5 Results Summary

The aim of the work described in this chapter was to develop and apply strategies that could increase the number of identified and annotated xenobiotic-related features in UPLC-MS based metabolic phenotyping datasets.

This main aim was successfully addressed in two ways; i) the construction of a reference standard database for xenobiotic spectra, based on knowledge of those prevalent in the UK; ii) extraction of xenobiotic signatures using statistical methods that made use of characteristic chemical and biochemical relationships between metabolites.

For the knowledge driven strategy, in place, is now a reference standard database and workflow for reference standard acquisition.

Originally 25 xenobiotics were used to populate the database (**Table 3-1**). Also listed in the table were reported metabolites observed in human biofluids, and if available, were additionally purchased as they may potentially provide a better marker for exposure. There are currently 41 reference standards which have undergone the acquisition and processing workflow. There were also 57 pharmaceutical medications which were profiled and subjected to the same workflow. Profiling these medications highlighted common excipients used in these formulations, which should also be regarded as a separate exposure, like the excipient polyethylene glycol (PEG) 400. This excipient along with 30 xenobiotics were identified in the ALZ urine study and their population prevalence estimated. Caffeine, PEG-400 and Prednisolone were observed to be highly prevalent within this population.

In addition to the knowledge driven strategy were data-driven strategies using three main statistical methods; correlation (inter/intra correlation and i-STOCSY), regression (PLS and logistic), and a method exploring unknown outlying signals.

For each method, an exemplar xenobiotic was used to explore the feasibility of each statistical method to discover additional metabolic feature associations. Xenobiotics were chosen based on prevalence in the population (APAP), known reported metabolites detected in urine from the literature, and medications specific to the ALZ study (amlodipine, DNP and MEM).

Workflows to carry out inter/intra correlation were implemented using custom scripts written in R. This correlation method was successfully exemplified using the xenobiotic APAP, in which all known reported metabolites commonly observed in urine demonstrated statically significant correlations (padj < 0.05). Another correlation-based tool, i-STOCSY, was additionally explored using the xenobiotic Memantine. A hydroxylated glucuronide was putatively annotated and was a novel finding. Also, unmetabolised Memantine was not detected in the urine and therefore would not be the best marker for exposure when screening for this xenobiotic in future urine studies.

Stemming from this analysis, were workflows to examine feature distribution and exposure groups using Gaussian mixture models (GMMs) and Probability density functions (PDFs). This permitted the population distribution of a xenobiotic exposure to be evaluated in more detail, thereby allowing more discriminant-based methods to be explored, such as logistic regression and PLS-DA models.

The logistic regression method was exemplified using the xenobiotic DNP of which six different models (a mixture of univariate and multivariate) were explored. The univariate model that incorporated all statically significant variables, and the EN multivariate model, produced the best classifiers based on ROC curve metrics. All known reported metabolites were putatively accounted for, as well others. This suggests that these models are useful for uncovering unknown xenobiotic metabolites, or endogenous responses.

Another multivariate statical approach which was also investigated was PLS (which included regression and discriminant analysis) and exemplified using the xenobiotic amlodipine. Like memantine, the amlodipine carboxylic acid metabolite was the more appropriate marker for exposure than the unmetabolised form in the urine. This metabolite was therefore targeted in the ALZ study and used as a proxy for exposure for prevalence estimation.

Finally, an approach linking outlier samples to xenobiotic exposures was explored. The script was successful in enhancing xenobiotic coverage, with MSI level one identifications of PEG 3350, which is a common laxative, and the antibiotic flucloxacillin.

3.6 Significance of Findings

The work presented in this chapter is useful in three main ways:

1. The construction of a xenobiotic database which can aid in metabolic identification efforts.
2. Spectral features relating to identified xenobiotic exposures can be readily partitioned from endogenous features, to better inform interpretation and reduce confounding in metabolomic investigations. This will also help with exposure misclassification, so that stratification of samples groups can be undertaken more accurately.
3. Aids population-level exploration of exposure to sufficiently abundant xenobiotic compounds, and exploration of the variation in the metabolism of these compounds between individuals.

3.7 General Discussion

The Identification of commonly used pharmaceuticals and xenobiotics in the general population, allowed the generation of a database to supporting rapid annotation in NPC assays. Often is the case, running a reference standard on the instrumentation used to acquire biological samples, is the simplest approach and precludes the need for any statistical based approaches. It reflects almost identical conditions in instrumentation used in the profiling analysis with method specific retention times and assay specific adducts and isotopic distributions. However, many prescription medications are considered a controlled substance and are not available for research purposes. In addition, reference materials for metabolites of common xenobiotics, are not usually supplied by typical vendors and may require chemical synthesis. Synthesis of reference materials can often be time consuming and expensive. If compliance/questionnaire metadata for the study was not available, the construction of the database was therefore inevitable, as UPLC-MS signals for xenobiotics from the database were used as a starting point for any statistical based exploration in the exemplar ALZ dataset, to uncover metabolites related to exposure.

There are other means which can be utilised to annotate xenobiotic metabolites. In some instances, elemental composition can be derived directly from the MS data, and if MSMS is available, fragmentation can aid in structural elucidation efforts. Matching the MS or MSMS profile with online spectral databases is yet another means to annotate metabolites. Advances have been made with structural elucidators from online databases which can generate a more refined list of in-silico chemical structures (Blaženović et al., 2018). However, these approaches can still result in many possible candidates, and are limited to what is present in the database, meaning novel metabolites may be missed. Drug screening using high resolution mass spectrometry instrumentation in toxicological applications have utilised an alternative technique known as suspect screening. In these applications, xenobiotics are identified based on predicted or intrinsic properties, such as accurate mass, isotopic distribution and a theoretical fragmentation profile. These groups have demonstrated high accuracy in xenobiotic detection and applicability to xenobiotics where reference standards are not available, such as synthetic drugs of abuse. Lastly, the structure elucidation of an unknown xenobiotic feature can be greatly facilitated by combining or even direct linking of two analytical platforms, such as NMR and LC-MS (Zani and Carroll, 2017, Wolfender et al., 2019, Koehn and Carter, 2005, Posma et al., 2017). LC-NMR-MS instruments are now available that combines the unique information from NMR (e.g. chemical shifts) with MS (e.g. accurate mass and fragmentation data). Workflows for the purification and desalting of a urine matrix used as a proxy for systemic human and gut microbial metabolism have enabled both NMR and MS profiles to be captured resulting in greater confidence for a particular metabolic assignment (Whiley et al., 2019). However, this avenue of metabolite ID may also fall short, due to the large volumes needed for the pipeline. Although it has proven highly applicable to endogenous metabolites, less prevalent xenobiotics may not be in a sufficient concentration to be detected by less sensitive techniques such as NMR.In this chapter, prior to any statistical approach, an efficient way to classify samples into exposure groups was needed. Either through known information, such as compliance data, or through measurement of an exposure marker. Assessment of the MS feature corresponding to a xenobiotic, highlighted patterns in the data, which could be linked to exposure. Xenobiotic related feature often exhibited a multimodal distribution allowing samples to be classified into different exposure group (for instance, a low and high exposure group). The application of GMMs provided an efficient way to classify samples based on their distribution in the data. This classification method, therefore allowed statistical techniques such as correlation, logistic regression and PLS to be employed, which successfully identified additional features which were subsequently connected to known reported metabolites. There were other significant features from each statistical method which were not annotated, which could potentially reflect new xenobiotic related metabolites, co-exposures, or even affected endogenous metabolites, which demonstrates the strength untargeted metabolic phenotyping platforms to capture xenobiotic signatures alongside endogenous signatures.

Multivariate unsupervised techniques, such as PCA or supervised techniques such as PLS-DA or orthogonal PLS (and variants such as OPLS), are widely used to analyse metabolomic data (Worley and Powers, 2013, Alonso et al., 2015, Ren et al., 2015). Outliers are normally detected using multivariate analysis such as PCA and can be the main cause of the underlying variance described by the first principal component (Ulaszewska et al., 2019, Tzoulaki et al., 2014). Removing these outlying samples from further analysis is often the case, as they can be mainly attributed to analytical issues.

However, the work presented in this chapter demonstrates that outlier samples can be linked to xenobiotic exposure and are not necessarily a result of analytical issues. If exposure related, or even a result of biological variation, consideration is needed before removing samples from a study as they can underline issues in the experimental design and potentially be of value. The putative annotation of Flucloxacillin using this outlier method, highlighted the unique isotopic pattern observed in MS data associated with chlorinated compounds. Mass defect and isotope filtering algorithms have been successfully implemented to identify xenobiotic and metabolites exhibiting this pattern, presenting another avenue of research which can be explored (Zhang et al., 2009, Rathahao-Paris et al., 2014).

These “outlying” signals can also be removed during sample pre-processing. As part of the NPC pre-processing (XCMS) workflow of UPLC-MS data, minimum fraction (minfrac) filtering can result in data that is biased towards the consensus metabolome. The parameter is usually set at 0.4 for all NPC project data. This means, that during the feature grouping stage of the workflow, for a feature to be qualified and included in the final dataset, it needs to be detected in at least 40% of all samples. There is therefore a good chance that less prevalent xenobiotic signals maybe filtered out, thereby demonstrating a limitation of applying any statistical method in extracting out signals relating to xenobiotics from these metabolomic studies. The minfrac setting can be adapted to a smaller setting but only to a certain extent. The smaller the setting, the more computational processing is required, which can be an issue with large (>1000 samples) studies. Another limitation which can construed from this line of thought is related to sample size in general. If the exposure feature is not detected in many samples, further annotations relating to metabolites will not be possible, and putative annotations will heavily rely on literature and online spectral databases. There also needs to be some level of diversity of the xenobiotic intensity measurements. Correlation and regression will not work if the measurement of a specific feature’s intensity does not fluctuate in a significant number of samples. The xenobiotic therefore needs to be moderately prevalent.

Compared to traditional univariate statistical methods, multivariate techniques can be better suited to handle high dimensional data, especially when there are strong inter-correlations (multicollinearity) between features, which care typical of UPLC-MS datasets. Regularised methods, such as Ridge, LASSO and EN, are alternative multivariate methods that can also deal with multicollinearity by imposing a penalty to regression coefficients (Full details in Chapter 2).

Multivariate Logistic regression using regularised methods, have also been implemented in metabolomic investigations (Yang et al., 2018, Yun et al., 2019, Goutman et al., 2020), however to my knowledge has not been implemented to specifically study xenobiotic metabolism. Both PLS and logistic regression (using regularisation) multivariate methods were explored in this chapter to study the metabolism of exemplar xenobiotics. Both were successful in highlighting associated features corresponding to reported metabolites.

A univariate application of logistic regression and correlation analyses (with multiple correction) was also successful in identifying a number of xenobiotic metabolites consistent with the literature. Regardless of the type of statistical method, a key observation was that metabolites in some cases maybe a better marker for exposure than the unmetabolised xenobiotic (as observed with MEM and amlodipine).

These methods and modelling approaches were specifically chosen due to ease of interpretation. When the xenometabolome was initially defined by Elaine Holmes, that work used STOCSY to identify signals directly linked to xenobiotic signatures from a given exposure using NMR. Correlation therefore formed the basis of the statistical approaches investigated throughout Chapter 3. Correlation can be considered a central part of statistics and quite a simple way to determine how two variables co-relate when measurements are continuous. The classification of samples into different exposure groups then invited other methods to be explored where variables can be categorical or binary. These included logistic regression and/or PLS-DA models. With these regression models, we can do more elegant statistics where different strata or groupings of people, factors of influence, for example gender or age, smokers, can be taken into account. These methods allow one to incorporate and accommodate these influences into the model and therefore control these potentially confounding variables. If we can account for these by stripping them away and removing their contribution, focus can be placed more on the biology.

The classification models discussed in the thesis use a form of regression to build a model, where if a new sample was introduced, we can confidently classify it into a particular group. Now this classification can be implemented using those same input variables used to make the model (in our case the features that best predict exposure) in a number of different ways. In a recent study, PLS-DA was compared with artificial neural networks (ANN) (Mendez et al., 2020). ANN takes into account the non-linear latent structure observed when handling biological data. The study demonstrated that it was possible to transfer the PLS-DA workflow to this ANN's, highlighting the same significant metabolites. The limitation of the ANN method is that they can be hard to interpret, which is why the methods investigated in this chapter were specifically chosen. With PLS models, loadings (weights) or p_{corr} (weights scaled as correlation coefficients) can be used to describe the contribution a variable may have to the model; Beta coefficient in regression models (linear and logistic) represents the strength and influence each variable has to the response variable; the use of regularised methods such as Ridge, LASSO and EN, imposes a penalty to regression coefficients, thereby highlighting the important variables in the model; and finally the use of p-values and multiple testing to account for false positives in univariate testing, thereby making feature selection simple. The relative ease of interpreting these models where why these specific statistical methods were selected to explore xenobiotic metabolism in profiling datasets.

Looking forward, making a note of the xenobiotic metabolites which are prevalent in specific biofluids, can be beneficial when targeting specific exposure markers in future studies. As demonstrated in this chapter (MEM and amlodipine in urine), the unmetabolised xenobiotic may not necessarily be the best marker for exposure.

Furthermore, as instrument conditions have remained the same since the very first project conducted at the NPC, the application of the database using the peakpantheR tool, could also be implemented to annotate xenobiotics on past project data. Due to the success of annotating metabolites from the database, the workflow for xenobiotic reference standard acquisition has now been implemented to the other profiling assays conducted at the NPC (HILIC systems for polar metabolites and RPC specifically geared to lipophilic species.). Similarly, the statistical approaches to further annotate xenobiotic metabolites can also extend to these profiling assays if ever the need requires it. The use of i-STOCSY in conjunction with compliance meta data provided with the ALZ cohort was successful in identifying metabolic features relating to exposure without the need of a specific driver feature MS feature. This success has now eventuated into a MRes project to further increase xenobiotic annotations in profiling cohorts.

Returning to the original hypothesis, construction of a reference standard database (analysis by RPC-UPLC-MS), and the development of statistical based methods, collectively increased coverage and characterisation of xenometabolome components in existing phenotyping datasets.

3.8 Conclusion

The developments described here have the potential to provide a richer set of annotated features with an exogenous origin, for existing and future studies.

The novelty of this chapter is therefore due to the strategies developed to partition signals relating to xenobiotics. There is novelty in the reference standard databasing workflow which permits the analysis and processing of a high number of standards with little manual assessment. There is also novelty in the strategies involved in partitioning xenobiotic signals statistically from endogenous signals (which includes the evaluation of feature distribution, and classification into exposure groups). Instead of dedicating a study to xenobiotic metabolism which can require quite a substantial investment (method development, time, cost etc), there is some novelty in how metabolomic datasets were retrospectively interrogated thereby demonstrating that these dedicated studies many not be necessary as the responses from exposures may already exist from these datasets.

For the purposes of most phenotyping studies, being able to separate out contributions / components of the metabolome that are directly related to (i.e. biomarkers of) chemical exposures can help limit their confounding influence on metabolome-wide analyses, and increase the efficiency of analyses focused on understanding endogenous metabolic regulation. Additionally, annotation of xenobiotic signatures can report objectively on individual compliance with study protocols, identify outliers, provide population-level exposure data and lead to discovery of new metabolites. Looking forward, these insights into the metabolism of xenobiotics have influenced how xenobiotics are being discovered, developed and administered (Wishart, 2016).

Chapter 4

Enhanced RPC-UPLC-MS profiling of the human blood metabolome using an optimised dispersive SPE protocol

Summary

Comprehensive coverage of chemically diverse metabolites present in human blood products benefits from the use of multiple methods, each oriented toward a small molecule subset generally segregated by polarity and hydrophobicity. Whilst recent developments in LC-MS profiling methodologies have delivered numerous solutions for the analysis of polar molecules (e.g. *via* HILIC-MS) and complex lipids, the analysis of moderately hydrophobic and amphipathic molecules in blood products (which includes much of the xenometabolome) is better suited to RPC methodology. The approach, however, is complicated by the suppressive effects of lipids on the ionisation of small molecule metabolites. A dispersive solid phase extraction (DSPE) protocol was developed to specifically remove lipids and protein efficiently, with minimal effect on other low molecular weight metabolites. The protocol therefore enables RPC-UPLC-MS blood profiling of the xenometabolome, with the added benefit of measuring a broader range of moderately hydrophobic endogenous metabolites. This was approached in three stages: optimisation, validation and application. Optimisation involved careful assessment and evaluation on the different components involved in DSPE (sorbent, solvent, and sorbent-solvent volume and concentration). Validation then assessed the reproducibility and recovery of measured small molecule metabolites using the final optimised sample extraction procedure and compares the method to conventional LLE methods and SPE protocols for lipid removal. Finally, an application was conducted to evaluate the performance of the protocol using two exemplar profiling studies. The DSPE method provided a straightforward and reproducible approach which enabled the use of uncompromised RPC-UPLC-MS to complement the coverage provided by HILIC and lipid analyses. Additional advantages include reduced cost and increased robustness when compared to liquid-liquid extraction (LLE) methods and conventional commercially available SPE sample clean-up protocols.

Aim and Objectives

The central aim of the work presented in this chapter was to develop an analytical strategy, to measure xenometabolome components in blood products. The strategy was undertaken with the development and optimisation of a DSPE blood preparation protocol to specifically removing lipids and protein efficiently and inexpensively, with minimal effect on other LMW metabolites.

Development of the protocol was divided into three stages.

1. Optimisation
   - Sorbent and solvent specification
   - Optimisation of sorbent-solvent conditions using Design-of-experiment (DOE)
2. Validation
   - Method reproducibility and precision
   - Assessment of recovery after DSPE; small molecule profile and targeted xenobiotics
   - Comparison of LLE methods and the DSPE protocol performance
   - Comparison of commercially available SPE phospholipid clean-up plates to DSPE protocol
3. Application
   - Method performance evaluated on plasma profiling study (MARS cohort)
   - Method performance evaluated on serum profiling study (AZ Study 12 cohort)

4.1 Introduction

Metabolic phenotyping or metabolomics has been used to report on the complex chemical compositions of biological fluids and tissues in mammals, as well as in microbial communities (Donia and Fischbach, 2015), plants (Fiehn et al., 2008) and environmental systems (Bundy et al., 2008). In humans, its definition has expanded to now encompass metabolites (the metabolome) reflecting intracellular and extracellular processes, as well as metabolites introduced and modified from external exposures such as diet (Holmes et al., 2008), xenobiotics (Marotta et al., 2006) , synthetic chemicals (Wishart et al., 2012) and the microbiome (Nicholson et al., 2005). Being minimally invasive in terms of accessibility, urine, and blood biofluids are frequently used in metabolomic research as they reflect import life processes and responses to environmental factors. Low molecular weight (LMW) metabolites represent a broad continuum of physiochemical and exposure-based metabolites with physiological concentrations spanning several orders of magnitude which help shape the phenotype of an individual. Plasma and serum, both of which are frequently used in metabolomics (Suarez-Diez et al., 2017) and as a proxy for blood itself, generally require sample pre-treatment in order to maintain and preserve the integrity of the analytical platform used. Both plasma and serum contain a unique variety of different LMW metabolites (e.g. energetic substrates and signalling molecules, proteins, peptides, lipids and lipoproteins. Relative differences are observed between the two matrices dependent upon factors such as sample collection (Yu et al., 2011), incubation(Liu et al., 2010), and due primarily to the clotting process present in plasma (Beheshti et al., 1994). Being a truly systemic sample (under homeostatic control), blood products can provide a snapshot of global metabolism at the point of collection, making them widely used in metabolomic studies. However, the diverse range of metabolites, represents an analytical challenge, as no single profiling method is yet comprehensive and so measurement of blood metabolites requires multiple complementary analytical techniques using multi technological platforms.

Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are two common profiling platforms used in metabolomic studies (Johnson and Gonzalez, 2012). NMR benefits from high analytical reproducibility and robustness, with biological samples making minimal contact with operational components due to the instrumental configuration, thereby minimising any risk of sample contamination or carry-over. The technique however is limited in that all signals from complex blood mixtures, are observed in the one spectrum and so high quantities of a particular analyte could potentially mask low level metabolites. MS for profiling applications offers a complementary approach to account for the poor sensitivity. The use of high-resolution measurement systems (e.g., time-of-flight mass analyser) allows high specificity of the technique where the accurate molecular mass is well distributed across the detectable range. Gas chromatography mass spectrometry (GC-MS) is a well-established platform in metabolomics, particularly in the analysis of volatile and non-polar compounds. Fragmentation of these species following the ionisation process, eliminates the need for further tandem MS (Papadimitropoulos et al., 2018) and as a result, spectra attained is highly reproduceable and independent of matrix effects leading to the creation of large spectral libraries (NIST11) to be available for metabolite annotation (Stein, 2012). In addition, head space GC-MS sampling allows for cleaner extract, free of interference from polar metabolites and has been successfully applied to a number of metabolomic applications (Silva et al., 2011). However, high throughput applications of blood products require an efficient extraction procedure which is both time and cost effective with minimum sample preparation, and as very few metabolites are truly volatile, a derivatisation step is usually required in GC-MS based analysis to make these metabolites thermally stable, introducing an extra step in sample preparation process which is generally not a concern for liquid chromatography (LC) MS applications. LC-MS boasts a broad metabolite coverage, utilising different column chemistries (Haggarty and Burgess, 2017) and ionisation sources (Lei et al., 2011). Developments in the retention mechanism in columns combined with the use of smaller column particles and high flow rates (ultra-performance liquid chromatography) provide fast and efficient separations, particularly beneficial in metabolome analysis, and affords vastly superior results in a fraction of the analysis time when compared to conventional HPLC systems (Wilson et al., 2005). Recently, we have developed a refined UPLC-MS platform capable of achieving robust and reproducible measurements in the raw data with minimal need for post normalization or informatic correction (Lewis et al., 2016).

Reversed-phase chromatography (RPC) with separation based on the hydrophobicity and length of fatty acyl chain moieties (Ovcacikova et al., 2016), is well suited to the analysis of serum, consisting of more than 70% non-polar lipid classes (Ovcacikova et al., 2016), making it a popular separation technique in lipidomics (Psychogios et al., 2011). Polar and ionic metabolites (some amino acids and organic acids) requires a different separation mechanism for detection, notably hydrophilic interaction liquid chromatography (HILIC) (Cai and Li, 2016), and ion-pairing chromatography (Coulier et al., 2006). To better accommodate lipids in global profiling analyses, the established standard is to design methods which measure smaller molecules and cleanly elute lipids, either allowing their additional measurement (Dunn et al., 2011), or simply preventing their problematic accumulation (Sarafian et al., 2015). However, where the target metabolite range is broad, the use of polar solvents to extract and solubilise polar metabolites can lead to poor recoveries of lipophilic species (Dunn et al., 2011). To span the breadth of chemical diversity present in blood products, comprehensive metabolome coverage can benefit from the use of multiple extraction procedures, each optimised to target a specific subset of metabolites. Within our laboratory, we have adopted such an approach, with methodologies for blood profiling that target polar metabolites using HILIC, and lipids using RPC (Izzi-Engbeaya et al., 2018). However, a major gap exists in the analysis of moderately hydrophobic and amphipathic molecules in blood products which can encompass much of the xenometabalome and the wider exposome (metabolites introduced and modified from external exposures). As molecules that fall within this category, exhibits a certain degree of hydrophobicity, RPC methodologies are better suited (Ordóñez et al., 2018, Lundgren and DePierre, 1990, Holcapek et al., 2008). RPC has been frequently used in metabolomic based analysis of biofluids due to a high level of analytical performance it can attain, i.e. superior peak shape, stable retention times and high speed of equilibration, across a broad range of analytes (Psychogios et al., 2011). RPC is therefore well suited for the analysis of moderately hydrophobic and amphipathic molecules which encompasses much of the metabolome and xenometabolome. With RPC however, comes chromatographic challenges, complicated by the suppressive effects of the existing lipid species present, on the ionisation of other LMW metabolites and the unpredictable manner in which lipid accumulate and elute from the column (Want et al., 2006, Dunn et al., 2011, Michopoulos et al., 2009, Rico et al., 2014).

As the focus of this work is on the analysis of moderately hydrophobic and amphipathic small molecules, the removal of lipids may be better suited during sample preparation. Protein precipitation (PP) followed by centrifugation, is the minimum and most often sample pre-treatment method used to blood products prior to LC-MS. Adequate blood deproteinization with organic solvents such as methanol (MeOH), ethanol (EtOH) and isopropanol (ISP), and acetonitrile (MeCN) have been reported to be the most effective (Raterink et al., 2014). Extraction selectivity of metabolites will differ based on the solvent used, and the overall extraction efficiency can be assessed using performance criteria such as instrumental stability (column lifetime and retention time), relative standard deviation for intensities (reproducibility) of all detected metabolite features and total number of extracted features (Polson et al., 2003, Want et al., 2006, Bruce et al., 2009). Biphasic extractions, or liquid-liquid extractions (LLE) are popular techniques used in the analysis of lipids on blood products (Li et al., 2014). Originally proposed by Folch (Folch et al., 1957), and later revised by Bligh-Dyer (Bligh and Dyer, 1959), the procedure predominantly uses a specific composition of methanol, water and chloroform. The Matyash method (Matyash et al., 2008) uses methyl tert-butyl ether (MTBE) in place of chloroform, which offers a safer alternative. Additionally, the extraction results in the organic layer forming above the aqueous hydrophilic layer (as opposed to forming on the bottom, as observed with Folch and Bligh-Dyer), which decreases risk of cross contamination between the organic and aqueous phases during sample preparation. Furthermore, integration of robotics and automated systems allows for high-throughput analysis (Patterson et al., 2015).

Whatever the LLE method, the use of an aqueous and an immiscible organic solvent, allows measurement of lipids from the organic fraction. The analysis of the hydrophilic fractions can additionally be used to measure polar metabolites however, this rarely appears to be the case. Rather, solid phase extraction (SPE) plates are often used to remove lipids. Examples which are commercially available include, Ostro™ (Waters), ISOLUTE ™ (Biotage) and Phree™ (Phenomenex). The extraction mechanism utilises a combination of PP and extraction on a C_{18} sorbent. SPE to remove phospholipids has observed to be better suited for targeted analyses, and unfavourable in untargeted metabolomics, with the potential risk of removing wanted metabolic features and potentially introducing contaminants (Armirotti et al., 2014, Simón-Manso et al., 2013). Both LLE and SPE have demonstrated less measurement variation of metabolic features but require additional steps during sample preparation (for instance, washing of sorbent with multiple solvents) which can be disadvantageous for global metabolic profiling (Yang et al., 2013).

In this chapter, a novel method using dispersive-SPE (DSPE) on blood products, was developed, validated, and applied as an alternative to LLE and SPE for phospholipid removal, for small molecule analysis. Originally designed as a clean-up technique for pesticide residue in produce (Anastassiades et al., 2003), the technique was implemented in a similar fashion, but instead, was used as a means to remove lipophilic species from blood products. In a DSPE preparation, sample is added to a sorbent material which is suspended in a liquid solvent. Based on the properties of the sorbent and solvent, certain components in the sample will therefore have an affinity for either one. C_{18} packing material was used as the sorbent in this application of DSPE to remove nonpolar interferences (lipids). The interaction between sample and components of the DSPE, occurs at a chemical specific on or of rate, i.e. at a certain point in time a percentage is bound to the sorbent and percentage is remains in the solvent. Equilibrium is reached quickly, reducing time significantly when compared to conventional SPE and LLE protocols (Islas et al., 2017).

Development of the DSPE protocol was split into three design stages: 1) optimisation, 2.) validation and 3.) application. Optimisation involved assessment of the sorbent and solvent, followed by the implementation of DOE to provide a more systematic approach to optimise the sorbent-solvent, or slurry, conditions. Validation then involved using the optimised protocol to determine method reproducibility, recovery, and comparison to other sample preparation procedures such as SPE and LLE. Finally, the application of the DSPE protocol to two exemplar profiling studies, will demonstrate the methods suitability to explore the metabolome and xenometabolome, with both endogenous metabolites and known xenobiotics being annotated.

4.2 Hypothesis

The hypothesis of this chapter is that DSPE will remove lipids and protein efficiently and inexpensively from blood products, with minimal effect on a specific subset of small molecule metabolites to enable large scale (i.e. fast, simple and reliable) RPC-UPLC-MS profiling. The DSPE blood sample preparation technique will provide a better alternative for the simultaneous capture of small molecules and removal of lipophilic species, than commercial SPE lipid removal plates and conventional LLE methods.

4.3 Methods

4.3.1 *Representative biological samples for development and testing*

Plasma was used as a proxy for all blood products and for the entirety of the optimisation experiments. EDTA anticoagulated human plasma, consisting of six individual donors, was purchased from Sera laboratories international (West Sussex, U.K) and was subsequently used in all parts of the development. The samples were sub aliquoted in 1 mL tubes and stored at -80°C. These samples were labelled as Development set samples or Devset-plasma. A pool of urine samples was also required for development and testing. It consisted of six individual donors of mixed genders and collected at multiple collection time points. No screening of contaminants was undertaken prior to pooling. A total of 200mL was collected and pooled as described by Lewis *et al.* These samples were labelled as Devset-urine. The volume of plasma and urine needed for any part of the development was calculated prior to the experiment and the appropriate number of aliquots thawed when needed for DSPE optimisation and validation.

4.3.2 *Reagents*

Organic solvents used for extraction and optimisation were HPLC grade and purchased from Sigma-Aldrich (Dorset, U.K.). All Reference standards used for various parts of the optimisation were purchased from Sigma-Aldrich (Dorset, U.K.), Avanti Polar Lipids (Alabaster, Alabama) and Qmx (Essex, U.K.). For instrumentation, LC/MS graded solvents and formic acid was purchased from Sigma-Aldrich (Dorset, U.K.). Solvents and acids used for the various extractions, MECN, IPA, MeOH, Formic acid, chloroform and MTBE were obtained from Honeywell solvents (Seelze, Germany).

4.3.3 *DSPE materials*

DSPE materials for optimisation, validation and application stages were obtained from numerous suppliers. High strength silica-bound C_{18} “HSS-T3” was obtained from Waters Corp. (Manchester, UK). Fully endcapped C_{18} (ENDCAPPED), non-endcapped C_{18} (NON.ENDCAPPED), and a commercial Lipid removal agent (LRA) were obtained from Sigma-Aldrich (Dorset, U.K.). Sepra C_{18} was obtained from Phenomenex (Torrance, USA). These materials were assessed as sorbents for DSPE, and its properties summarised in **Table 4-1.**

Before use, each material was washed to remove unwanted contaminants coating the sorbent and condition the sorbent. Washing of the DSPE sorbent with mixtures of H2O and IPA, firstly at 4:1 (v/v), then at 1:4 (v/v) were initially undertaken to minimise contaminants that maybe introduced from the outer coating of the particle. The sorbent was then further washed with a mixture 1:1 (v/v) mixture of I and MeOH, with the supernatant (after centrifugation) being dispensed into a 300 mL glass PYREX dish and left overnight for complete evaporation of solvent. Dry sorbent is then placed in an airtight, moisture free glass container ready for use. Prior to any extraction in subsequent experiments, the appropriate amount of dry sorbent is weighed, given the number of samples, and equilibrated at least once with the organic solvent.

Table 4-1. Physical properties of the sorbents, HSST3, C_{18} Fully endcapped (ENDCAPPED), C_{18} Fully non-endcapped (NON.ENDCAPPED), a generic Lipid removal agent (LRA), and Sepra C_{18}, material, evaluated for the DSPE protocol.

Property	HSST3	LRA	ENDCAPPED/ NON.ENDCAPPED	Sepra
Surface Area (m^2/g)	230	400	400	500
Particle Size (µm)	1.8	40	63	50
Pore Diameter (Å)	100	90	90	65
Carbon load (%)	11	11	15	17

4.3.4 *Sample preparation*

Sample extraction in all three design stages, unless stated otherwise, follows the DSPE procedure summarised in **Figure 4-1**. As both slurry concentration and volume had yet to be optimised, a maximum slurry concentration at 20 mg/mL of sorbent was implemented, as this was practical in terms of handling and sample preparation. Any concentration above 20 mg/mL produced a dense mixture that did not homogenise well. For slurry volume, a 1:3 proportion sample:solvent ratio was used, as this sample to solvent ratio has been reported to be sufficient for protein removal (Sarafian et al., 2014, Polson et al., 2003). Solvents used in the extraction were kept at -20°C. To ensure a homogenous mixture of slurry, the container holding the slurry (a 50mL centrifuge tube), was vortexed at every instance prior to sample addition. For example, experiments conducted within the optimisation phase involved the slurry to be made 20mL at a time in a 50mL centrifuge tube, so a 20mg/mL slurry concentration equates to 400mg of sorbent in 20mL of solvent. Next, the sample-slurry mixture is vortexed and incubated for 2 hours at 4°C. The sample is then centrifuged for 10 minutes at 3214 x g and 50% of the total supernatant is collected and dried under a gentle flow of Nitrogen. After drying, the sample was then resuspended with water, and in half the sample volume to account for the fact that only 50% of supernatant was recovered. This ensures no dilution of the sample is undertaken (*Method 1)*. Certain experiments also warranted lipid analysis (*Method 2*) and/or a SHAM sample (*Method 3*) to be prepared. A SHAM sample is when there is no DSPE treatment, i.e. sorbent free solvent is added to sample. In many experiments of the development, it was used as a control to test a particular treatment against.

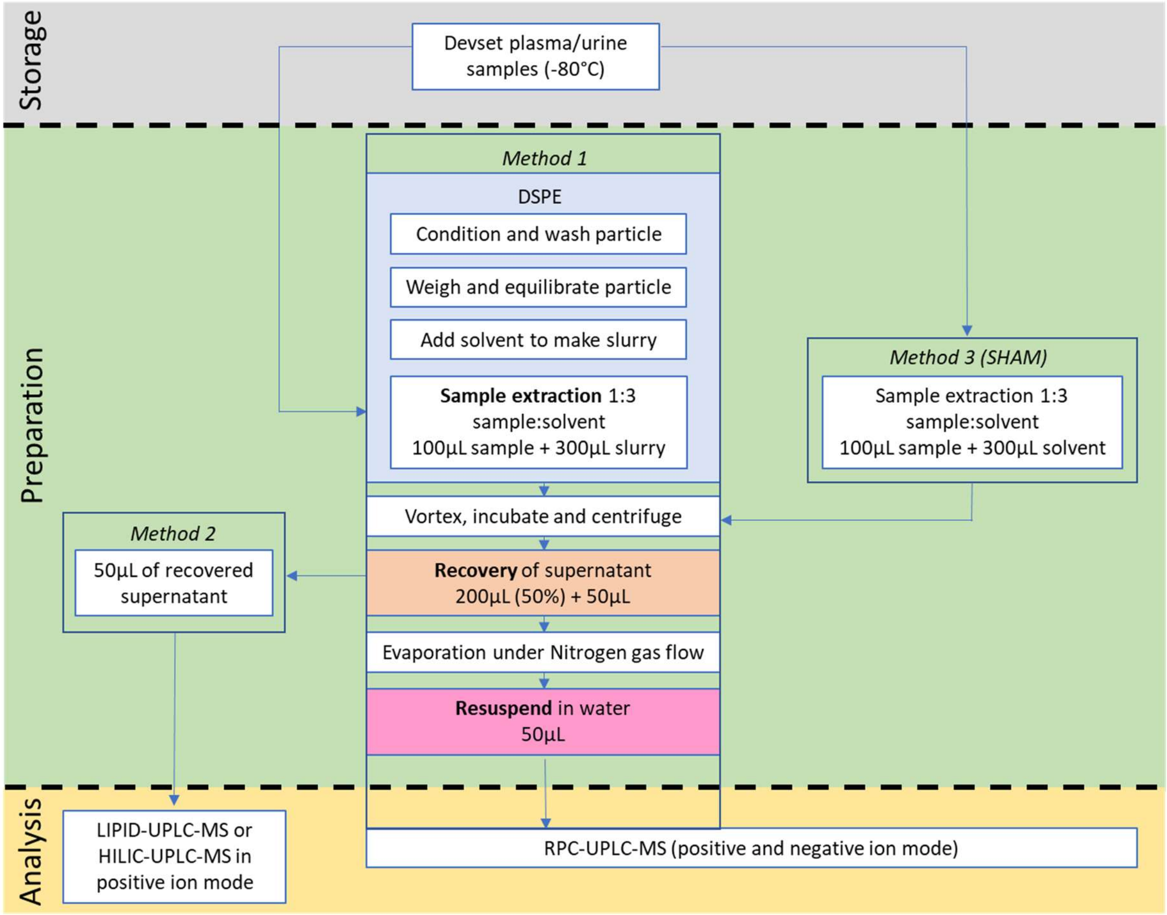

Storage
Devset plasma/urine samples (-80°C)
Preparation
Method 1
DSPE
Condition and wash particle
Weigh and equilibrate particle
Add solvent to make slurry
Sample extraction 1:3 sample:solvent 100µL sample + 300µL slurry
Method 3 (SHAM)
Sample extraction 1:3 sample:solvent 100µL sample + 300µL solvent
Vortex, incubate and centrifuge
Method 2
50µL of recovered supernatant
Recovery of supernatant 200µL (50%) + 50µL
Evaporation under Nitrogen gas flow
Resuspend in water 50µL
Analysis
LIPID-UPLC-MS or HILIC-UPLC-MS in positive ion mode
RPC-UPLC-MS (positive and negative ion mode)

Figure 4-1. Summary of the DSPE sample preparation protocol used in the development and validation stages of the experimental design.

4.3.5 *UPLC-MS*

The method for lipid removal from blood products was developed to enable reversed phase chromatography using a hyphenated ultra-performance liquid chromatography, together with a high-resolution orthogonal acceleration time-of-flight mass spectrometry system (RPC-UPLC-MS). This analytical method complements the already existing urine reversed phase method, thereby adopting identical instrumental conditions. The stationary phase and mobile phase conditions are methods previously adapted from those described by Want *et al.* (Want et al., 2010) and Wong *et al.* (Wong et al., 2008), and then optimised for large scale application by Lewis et al. (Lewis et al., 2016). Mass spectral acquisitions were measured in the range of 50 to 1200 *m/z* and collected in both electrospray positive and negative ion modes. Analysis of polar small molecules was performed using HILIC methodology previously described by Lewis *et al. 2016* (HILIC-UPLC-MS) in positive ion mode, and the analysis of lipids was performed using a lipid specific RPC method (LIPID-UPLC-MS) in positive ion mode as described (Izzi-Engbeaya et al., 2018).

4.3.6 *Data pre-processing*

Raw UPLC-MS data was firstly converted into *mzNLD* format for pre-processing using Nonlinear Dynamics pre-processing software, Progenesis QI 2.1 (Waters Corp., Manchester, UK), for peak detection, alignment and grouping. The minimum chromatographic peak width was set at 0.02 min in accordance with peak shapes observed using RPC methodology. Next, features from a selected sample in the dataset were used as the alignment reference to which all corresponding features from all samples are corrected. Experiments conducted in the optimisation and validation stages used a randomly selected sample for the alignment reference.

For the studies in the application stage, a randomly selected study reference (SR) sample is used as the alignment reference. The SR is a pool of all samples in the study and is the primary QC sample used for data quality purposes in NPC projects. A peak picking algorithm then generates and implements an aggregate map from the aligned runs of each sample, across the dataset. Finally, integration of the features in each acquisition was generated to create a data table output.

A series of QC measures were put in place, to all data succeeding pre-processing by Progenesis, ensuring high data quality and reduced bias, i.e. run order correction, feature filtering, and study sample randomisation. The studies exemplified in the application stage of the design incorporates all three QC measures. Run order correction by the LOWESS approach proposed by Dunn *et al* (Dunn et al., 2011) was applied using replicate SR samples injected throughout the analytical acquisition. Feature filtering is based on two criteria, 1.) the precision of individual features (%RSD < 30) detected in replicate SR and 2.) the correlation to dilution (Pearson correlation with r= 0.7 threshold) of features, *via* serial dilution (Dilution series) of the SR.

The preparation of the dilution series was conducted differently to that stated in Lewis *et al.*, as early experiments indicated a disruption to protein precipitation when sample was diluted prior to sample extraction and especially solvent compositions involving methanol. Full details explaining this disruption to protein precipitation is in **section 4.3.8.4.** Nevertheless, the outcome was that serial dilution was undertaken on the sample post-extraction, rather than pre-extraction. These measures are implemented using code written in Python (Sands et al., 2019), and in accordance to NPC protocols for project population studies (Lewis et al., 2016).

Data quality measures conducted in the optimisation and validation stages, adopted a similar approach, including the use of replicate samples for precision measurements and run order correction purposes (where the number of replicates is specified for each experiment), dilution series, and experimental sample randomisation. In addition to these QC measures, a third filtering method (blank filtering), was implemented, by substituting water for a given experimental condition (Broadhurst et al., 2018). Here, the mean intensity for each feature detected in the blank (n=3) is calculated ($\text{mean}_{\text{blank}}$). Then, the median intensity for all features from all samples is calculated ($\text{median}_{\text{all}}$). Any feature where $\text{mean}_{\text{blank}}$ is greater than 5% of $\text{median}_{\text{all}}$, is subsequently removed. Several different experiments within this chapter involves comparison between different extraction conditions (e.g., different solvents or sorbents). For most experimental conditions tested, unless stated otherwise, Individual dilution series and blank samples were prepared. Where the total number of features for a given experimental condition was evaluated, individual dilution series and blank samples were used for feature filtering (alongside %RSD < 30 calculated between replicates). When comparison between different experimental conditions were undertaken using multivariate methods, features included in the final dataset had to pass all three filtering criteria for at least one of the conditions. This resulted in one feature matrix which could then be inputted into the multivariate software. Run order correction and all feature filtering QC measures were implemented using custom scripts written in the R language. Normalisation using Probabilistic quotient normalisation (PQN) was undertaken when specified.

4.3.7 Multivariate analysis (optimisation and validation)

For multivariate data analysis, all feature matrices were exported to SIMCA (Version 15 Sartorius Stedim Biotech, Malmö, Sweden). Principal component analysis (PCA) and partial least squares-discriminant analysis (OPLS-DA) *via* orthogonal projection to latent structures were carried out on the filtered features. The quality of the OPLS-DA models were validated by a seven-fold internal cross validation, assessment of the variance (R2Y) and predictive ability (Q2Y) of the model, and permutation tests (n=999). The appropriate number of components were selected for each model in order to optimise model quality without over-fitting. Discriminant features were evaluated based on variable importance for the projection (VIP) values greater than 1.5. Log transformation and pareto scaling were carried out on the filtered features prior to multivariate analysis.

4.3.8 Optimisation

4.3.8.1 Sorbent assessment

HSST3 (Waters Corp., Manchester, UK) and three different sorbent materials from Sigma-Aldrich (Dorset, U.K.); C_{18} Fully endcapped (ENDCAPPED), C_{18} Fully non-endcapped (NON.ENDCAPPED), and a proprietary Lipid removal agent (LRA), were assessed as sorbents for DSPE. The experiment was undertaken in two parts, utilising both Devset-urine and Devset-plasma.

In the first experiment, a 100 µL Devset-plasma sample was added to a 300 µL slurry of a 20mg/mL sorbent made in MeOH and prepared in three replicates. This was repeated for all sorbents. These samples were subjected to *Method 1* followed by *Method 2* and acquired by the NPC lipid profiling method (LIPID-UPLC-MS) in positive ion mode. This was to understand the capacity of lipid removal using these sorbents. Assessment was based on a thorough evaluation of the total ion chromatogram (TIC) produced from each sorbent extraction.

The second experiment utilised Devset-urine samples and subjected to the same experimental conditions as previously but acquired by RPC-UPLC-MS in positive ion mode (*Method 1* only). Urine has been extensively studied at the NPC, with approximately 200 metabolite annotations. From the literature, approximately 4500 metabolites have been documented in urine, many of which are end products originating from metabolised nutrients and drugs. At this stage of the development, urine was therefore a better matrix than blood, to assess the effects DSPE treatment could have on a greater range of metabolites.

SHAM samples (*Method 3,* followed by *Method 1 and 2*) were included for both the first and second experiment and used as a reference to compare the effect of DSPE treatment. Feature filtering was undertaken as described in **section 4.3.6**, for data being examined by multivariate methods. PCA and OPLS-DA models were used in this second experiment, to highlight the differences between SHAM and DSPE treated samples.

4.3.8.2 Solvent assessment

To assess the solvents which would be needed as part of the slurry, five organic solvents were initially compared. The solvents were IPA, MeCN, MeOH, Acetone (Acet) and EtOH. In separate 50mL centrifuge tubes, 20 mL of each solvent was added to 400mg of DSPE sorbent (using the final sorbent chosen from section 4.3.8.1), making a slurry concentration of 20 mg/mL. Sample extractions were prepared in triplicate using Deveset-plasma and follows *Method 1* then *Method 2* i.e. sample extracts were only run by LIPID-UPLC-MS in positive ion mode. No data pre-processing was required for this part of the experiment, as evaluation was undertaken by a visual comparison of the TIC, for each solvent extraction.

The results from this experiment than warranted a secondary investigation with only three of the solvents. These solvents were tested individually and in combination with one another resulting in seven different extraction conditions; EtOH, MeOH:EtOH (1:1), MeOH:MeCN (1:1), MeCN:EtOH (1:1), MeOH, MeCN, and MeOH:MeCN:EtOH (1:1:1). The Devest-plasma was extracted identically to the previous experiment, in replicates of six and follows *Method 1*. The extracts were run by RPC in both positive and negative ion mode. The total number of features were evaluated for each extraction condition, based on features that passed the dilution series and blank filter for each extraction condition. The TIC's and the %RSD calculated from replicates for each extraction conditions were evaluated.

4.3.8.3 Sorbent and solvent (slurry) optimisation

Umetrics MODDE PRO12 (Sartorius Stedim Biotech, Malmö, Sweden), was implemented for the generation and evaluation of statistical experimental designs for a design-of-experiment (DOE) based analysis. DOE allowed for two variables (or factors) to be studied and optimised at the same time whilst also taking into account the interaction between the two factors (refer to Chapter 2 section 2.5 for full details on the DOE process). Initial steps in a DOE design involved specification of the factors and their ideal response measurements. The two factors investigated were the slurry concentration and slurry volume. Keeping the maximum concentration set at 20mg/mL and a sample to solvent ratio at 1:3, the input slurry concentration ranged from 2 mg/mL to 20 mg/mL and the slurry volume, ranged from 600 µL to 1000 µL for a sample volume of 200 µL. Eppendorf tubes with a wider diameter were used for sample preparation, to avoid risk of aspirating the protein pellet, a higher sample volume was used. The sample extraction using the sorbent and solvent that were optmised from the previous sections (**4.3.8.1** and **4.3.8.2**) were used. The sample extraction is similar to what is depicted in **Figure 4-1**, however volumes were scaled to account for the higher sample volume, and slurry conditions differed depending on the proposed conditions by the DOE design. The MODDE software recommended a quantitative central composite face design composing of eight different slurry concentration/solvent volume conditions with three replicated centre points for a total of 11 experiments. The conditions were 2mg/mL – 600 µL, 20mg/mL – 600 µL, 2mg/mL – 1000 µL, 20mg/mL – 1000 µL, 2mg/mL – 800 µL, 20mg/mL – 800 µL, 11mg/mL – 600 µL, 11mg/mL – 1000 µL and, three instances of 11mg/mL – 800 µL. Devset-plasma samples were prepared for each condition, with the inclusion of SHAM samples, and analysed by RPC-UPLC-MS (positive and negative ion mode) and LIPID-UPLC-MS (positive ion mode). Although MODDE only recommended 11 experiments, a dilution series and blank samples were prepared for each extraction condition for feature filtering purposes. Apart from the three instances of 11mg/mL – 800 µL, no replicates were prepared for the other extraction conditions. In hindsight, this should not have been conducted as the quality of features in the final dataset may be affected.

The responses required for DOE, was measured *via* the following steps:

1. Data was split into 1-minute RT bins, equalling a total of 11 bins which make up the chromatographic range;

2. Summation of peak intensities at each RT bin (up to 11 minutes) were calculated for each optimisation condition;

3. A ratio taken with its RT bin counterpart in the SHAM treated samples and calculated as a percentage. This percentage is herein referred to as the recovery. A recovery of 100% would equate to zero difference between DSPE treatment and without (SHAM).

The recoveries were input to MODDE as response measurements. Arbitrary response recoveries were selected. Ideal targets for the response recoveries were recorded as a range between 90-130% for RPC-UPLC-MS analyses and 20-40% for LIPID-UPLC-MS analysis. Performance indicators plots were used to evaluate and access the quality of the model. The first was a Summary of Fit plot. The model was fitted with partial least squares regression (PLS) and the validity of the model was verified by the summary of fit. The resulting models were evaluated using both R2 and Q2 metrics. R2 values reported the total amount of variance explained by the model in the data. Q2 reported the model accuracy and was calculated by cross-validation. For this investigation Q2/R2 ratios of greater than 0.5 was used as a measure of cross validation reproducibility and therefore model validity (Moltu et al., 2014). The second plots depict the regression coefficients of the models and their confidence intervals (or uncertainties) in the form of a bar graph. Any coefficients whose uncertainties exceeds their actual values have no significant contribution to the model and removed. This step also improves the R2/Q2. The final two plots were used for diagnostics, i.e. the Sweet spot contour plot and the design space plot. The Sweet spot plot highlights the areas where all responses are within the user specified range and is colour coded to how many specifications are fulfilled. The Design space plot represents how stable the model is by combining uncertainties from the specified factors and highlighting probabilities of failure. The plot is a result of disturbance-based Monte-Carlo simulations which estimates how sensitive the responses are to small fluctuations in the factors and thus how sensitive overall result is to different sources of uncertainties.

4.3.8.4 Protein quantification

Devset-plasma samples were prepared in triplicate and extracted with no dilution, with a five times dilution pre-extraction, and a five times dilution post-extraction. The DSPE parameters utilised the optimised conditions from the previous optimisation sections.

Proteins were quantified using BCA protein assay (Smith et al., 1985) and analysed by a ClarioSTAR Plate Reader from BMG LABTECH with UV/vis detection. A standard curve was developed using a series of Bovine Serum Albumin (BSA) standards in concentrations ranging 150 µg/mL to 2000 µg/mL. The absorbance of each sample was measured at 595 nm and plotted against the concentration of the BSA. The resulting line was fit by the linear least squared method. Relative albumin concentrations were calculated from the measured absorbance of each sample, together with the equation from the calibration line generated in the BSA standard curve (150 µg/mL to 200 µg/mL).

4.3.9 *Validation*

4.3.9.1 Method precision

The precision of the method was explored in a 96 well format and split into two parts: (1) the precision of the sorbent quantity added to each well and (2) the precision of the metabolic profile from repeated extractions of the Devset-plasma. For the first part of the experiment, to ensure the homogeneity of slurry addition, conditioned and washed sorbent is firstly weighed in a 300mL PYREX flat bottom glass flask. Cold solvent is then added to the flask which is then placed on a magnetic plate with a stirrer that constantly rotates. This constant rotation allows the mixture to remain in a homogenised state. Slurry was aspirated from the PYREX flask, and dispensed into special 96-well PCR tube racks, one column at a time, using an 8-multichannel automatic pipette. This type of rack has the same dimensions as a typical 96-well plate and allows individual tubes to be detached. Each tube was weighed and recorded prior to addition of slurry. Once the slurry was added, racks were allowed to rest for 10 minutes and the remaining solvent dried under a gentle flow of Nitrogen. Individual tubes were then reweighed and recorded. The difference in weight for each tube was calculated and the % RSD reported.

The second part of the experiment utilised the slurry conditions (optimised in **section 4.3.8**). Devset-plasma samples were distributed to three 96-well plates and subjected to the updated DSPE protocol, i.e. optimised sorbent, solvent and sorbent/solvent (slurry) slurry conditions. The precision of the metabolic profile was evaluated using PCA, and the relative standard deviation (RSD) calculated for the measured signal intensities of individual molecular species passing the dilution series filter and blank filter. Sample observations in PCA score plots were produced to highlight any general trends in variation, and trends that could potentially arise from the addition of slurry (slurry added to sample column-wise in the 96-well plate), and from resuspension (water added during row-wise in the 96-well plate).

4.3.9.2 Assessment of recovery after DSPE

To determine if the lipid removal conditions had any significant effect on the small molecule profile, two experiments were conducted utilising urine. The first experiment was a more targeted approach, evaluating a mixture of xenobiotics which were spiked into the urine and assessing intensity levels between SHAM (i.e. sorbent free solvent is added to sample) and DSPE treated samples. The second experiment evaluates feature intensity between SHAM and DSPE treated samples on a more global scale.

Devset-urine was initially screened *via* the RPC-UPLC-MS assay to make sure the sample were free of the selected xenobiotics. A working mixture was prepared by quantitatively spiking individual stock standards of six xenobiotics (selected from the in-house refence standard database), and two labelled standards (N-Benzoyl-D_5-Glycine and L-Phenylalanine-$^{13}C_9$,^{15}N), to make a final concentration of 10 µg/mL in MeOH/MeCN 1:1.

Table 4-2. Reference standards of xenobiotics and lipid species spiked at different concentrations into urine-SHAM and urine-DSPE treated, to assess if varying concentration of lipids will affect the recovery of xenobiotics.

	XENO Mixture			LIPID Mixture	
	RT (min)	[M+H]+		RT (min)	$[M+H]^+$
Amoxicillin	2.25	366.1118	Cer(d18:1/22:0)	8.457	604.6036
Acetaminophen	2.59	152.0706	Cer(d18:1/24:0)	9.003	632.6346
Diclofenac	10.66	296.0245	Cer(d18:1/24:1)	8.446	630.6191
Amitriptyline	7.5	278.1903	LPC(16:0/0:0)	1.868	496.3407
Lansoprazole	5.29	322.1164	LPC(16:1/0:0)	1.499	494.3249
Terbinafine	8.4	292.2060	LPC(18:0/0:0)	2.415	524.372
Ibuprofen	10.75	251.1036	PC(16:0/18:1)	6.55	760.5859
N-Benzoyl-D_5-Glycine	3.57	185.0975	PC(18:0/18:2)	6.725	786.6018
L-Phenylalanine-$^{13}C_9$,^{15}N	2.1	176.1140	PC(16:0/18:2)	6.031	758.5698
			PE(16:0/18:2)	6.188	716.525
			PE(16:0/20:4)	6.174	740.5222
			PE(18:0/20:4)	6.86	768.5567
			SM(d18:1/16:0)	5.683	703.5752
			SM(d18:1/18:0)	6.409	731.6065
			SM(d18:1/24:0)	8.401	815.7004

A serial dilution from the working mixture, using MeOH/MeCN 1:1 as a diluent, was made four times, resulting in a total of four xenobiotic mixtures; XENO1 (10 times dilution), XENO2 (twenty-five times dilution), XENO3 (fifty times dilution) and XENO4 (seventy-five times dilution). Each mix is made at a total volume of 500 mL and was used as the extraction solvent for DSPE and SHAM treatments. A secondary artificial mixture, comprising of fifteen individual lipids was also prepared.

This lipid mixture was prepared by pooling together 200 µL of stock standards, each at 1mg/mL and diluting to a total volume of 3000 µL with water, for a final concentration of 67 µg/mL. This was labelled as LIPID1. A further two dilutions from LIPID1 were also made in water, i.e. LIPID2 (2 times dilution) and LIPID3 (4 times dilution). The compounds which make up both the xenobiotic mixture and the lipid mixture are summarised in **Table 4-2**. In addition to these two mixtures, an albumin stock solution of 6000mg/mL was made and spiked into the urine sample to give a final concentration of 375mg/mL. Combinations of the lipid mixtures and xenobiotic mixtures resulted in a total of 12 different experimental conditions. Each condition was spiked into urine samples that will undergo DSPE treatment and SHAM treatment. Therefore 12 experimental conditions, with a DSPE and SHAM treatment, prepared in six replicates, totalled 144 individual extractions spanning across two 96-well plate (**Figure 4-2**). 50 µL of Devset-urine samples were transferred to a 96-well preparation plate with 25 µL of the Lipid mixture, and 25 µL of the albumin stock solution.

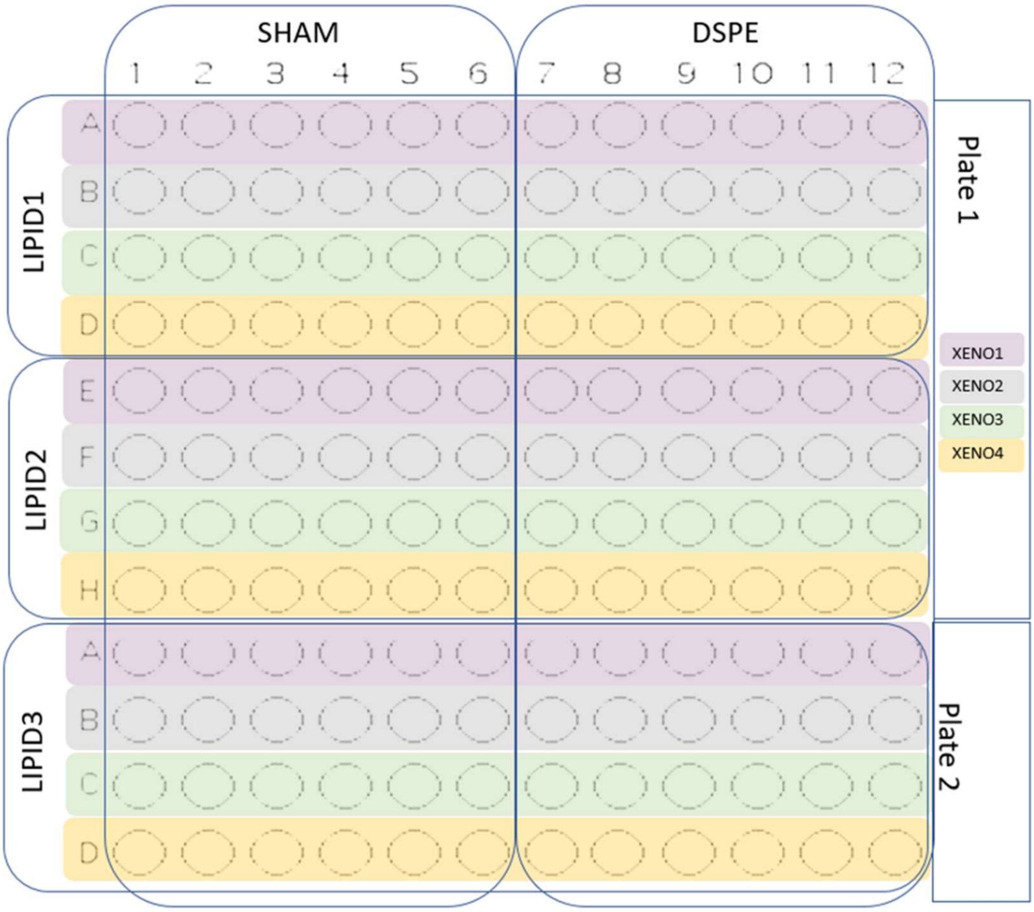


Figure 4-2. Plate layout of all extraction conditions for the xenobiotic/small molecule recovery validation experiments. Combinations of the lipid mixtures and xenobiotic mixtures resulted in a total of 12 different experimental conditions. Each condition was spiked into urine samples that will undergo DSPE treatment, and SHAM treatment. Therefore 12 experimental conditions, with a DSPE and 12 experiments with a SHAM treatment, prepared in six replicates, totalled 144 individual extractions spanning across two 96-well plates.

4.3.9.2.1 Recovery of target xenobiotics

Xenobiotics spiked into urine were identified by matching the observed retention time and accurate mass data from the Progenesis dataset to the in-house reference standard database. Within a xenobiotic concentration level, box plots and Kruskal-Wallis ANOVA test was used to determine if any statistically significant differences were found between the mean intensities measured for each xenobiotic in the DSPE and SHAM treated samples, regardless of the concentration of the lipid mixture. This was to demonstrate if varying concentration of lipids can also influence the signal intensity of the measured xenobiotics. The box plots and ANOVA tests were repeated for each concentration level of the xenobiotic mixture. The results from this analysis was acquired by RPC-UPLC-MS in positive ion mode only, as the xenobiotics from the XENO mixes do not ionise in negative ion mode.

4.3.9.2.2 Global small molecule recovery

In the second experiment, to complement this targeted assessment of signal intensities to a more global investigation, the recoveries of all filtered features were assessed from data acquired by RPC-UPLC-MS (both ionisation modes), for all experimental conditions associated with XENO3. Recoveries were calculated by taking the quotient of the average intensity (from replicate measurements) measured from the DSPE treated samples, with the average intensity observed in the corresponding SHAM treated samples, for all features passing the dilution series and blank filter. Filtered features were based on the dilution series and blanks prepared in only SHAM treated samples containing XENO3 and LIPID1 mixtures.

4.3.9.3 Comparison between DSPE and liquid-liquid extraction (LLE)

A series of two-phase LLE methods were applied to human plasma samples and compared to the DSPE protocol. Folch method (Folch), Bligh-Dyer (BD) and Matyash (Matyash) extractions were carried out in accordance with adapted protocols stated by Gil and co-workers (Gil et al., 2018). The metrics used to compare the different extractions were also adapted from Gil *et al.* For each extraction, samples were independently prepared in 10 mL centrifuge glass conical tubes in six replicates. The addition of plasma, water and MeOH, were all handled using a variable volume pipette, and any mixture involving chloroform or MTBE, was handled using a graduate glass pipette.

An incubation time of one hour and an incubation temperature of 20 °C were kept constant for each LLE method. The hydrophilic fractions were analysed by RPC-UPLC-MS (positive and negative ion mode) and HILIC-UPLC-MS (positive ion mode), and the organic fractions for the LLE methods were analysed by LIPID-UPLC-MS (positive ion mode), as illustrated in **Figure 4-3.** The details of the extractions are described below.

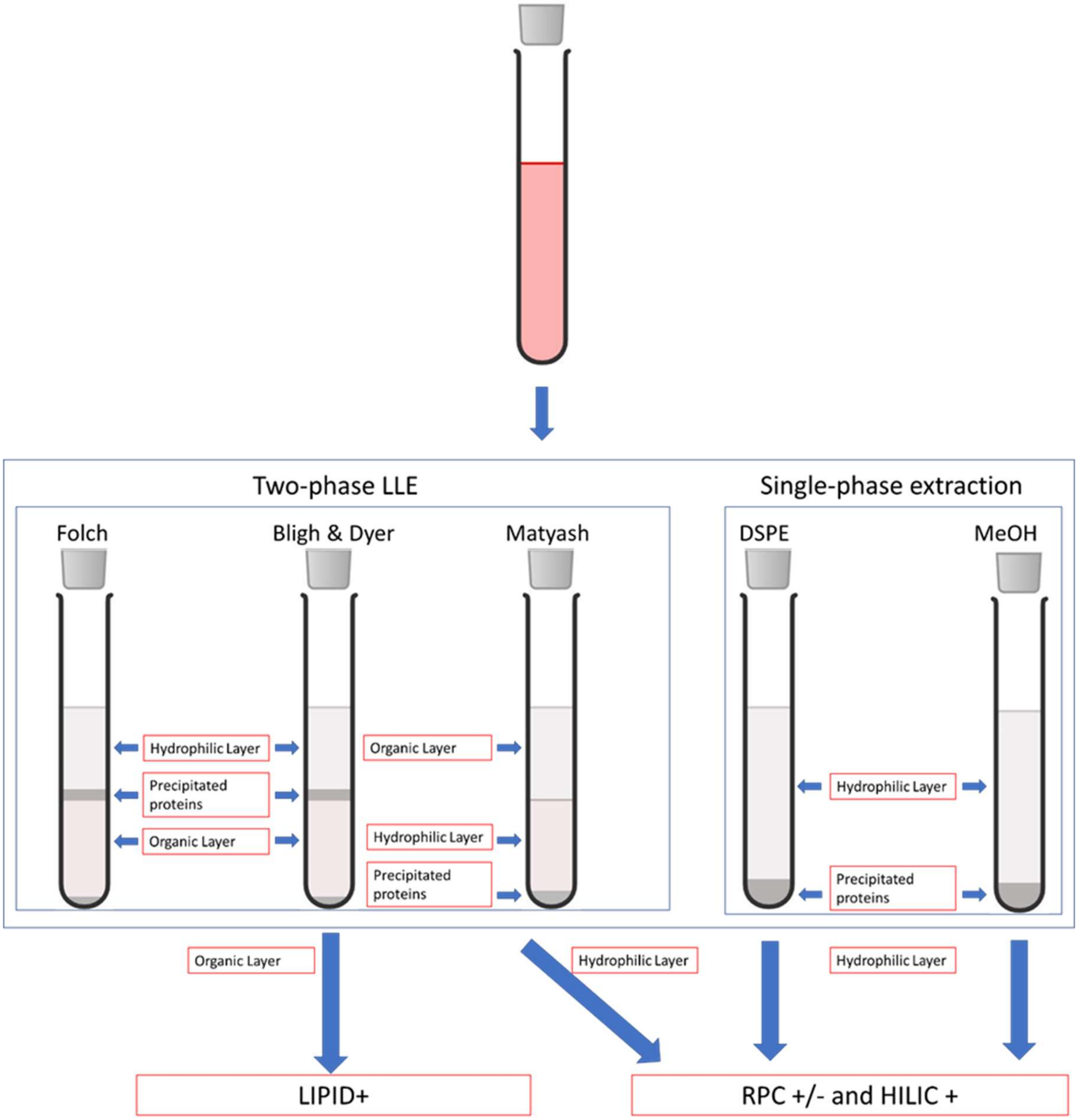


Figure 4-3. **Schematic of the five different extraction methods comparing the DSPE protocol to liquid-liquid extraction protocols commonly undertaken for lipidomics**. In all five methods, the hydrophilic layer was acquired using RPC-UPLC-MS (positive and negative ion mode) and HILIC-UPLC-MS (positive ion mode). The organic layer for the LLE protocols were acquired by LIPID-UPLC-MS (positive ion mode).

4.3.9.3.1 Folch method

Seventy-five microliters of plasma were mixed with 187.4 µL of MeOH (ice cold) and 375 µL chloroform – $CHCl_3$ (ice cold). The mixture was vortexed for 20 seconds and then incubated for 1 hour on a shaker. Phase separation was induced with the addition of 156.2 µL of water, and the mixture incubated for an additional 10 minutes. Using a glass bulb pipette, the maximum volume of the lower organic phase was aspirated into a clean 10mL glass tube and dried under nitrogen. The upper methanolic phase was re-extracted with 250 µL of a $CHCl_3$ /MeOH/H_2O 86:14:1 (v/v/v) mixture. The maximum volume of the upper hydrophilic methanolic phase was aspirated into a clean 10mL glass tube and dried under nitrogen. Once dried, both the lower organic phase and upper hydrophilic phase tubes were stored at − 20 °C.

4.3.9.3.2 Bligh-Dyer method

Seventy-five microliters of plasma were mixed with 562 µL of ice cold MeOH/ $CHCl_3$ 2:1 (v/v). The mixture was vortexed for 20 seconds and then incubated for 1 hour on a shaker. Phase separation was induced with the addition of 156.2 µL of water, and the mixture incubated for an additional 10 minutes. Using a glass bulb pipette, the maximum volume of the lower organic phase was aspirated into a clean 10mL glass tube and dried under nitrogen. The upper methanolic phase was re-extracted with 250 µL of a $CHCl_3$ /MeOH 2:1 (v/v) mixture. The maximum volume of the upper methanolic phase was aspirated into a clean glass tube and dried under nitrogen Once dried, both the lower organic phase and upper hydrophilic phase tubes were stored at − 20 °C.

4.3.9.3.3 Matyash method

Seventy-five microliters of plasma were mixed with 187.4 µL of ice cold MeOH. The mixture was vortexed for 20 seconds and added to 625 µL MTBE (room temperature) and then incubated for 1 hour on a shaker. Phase separation was induced with the addition of 156.2 µL of water and the mixture incubated for an additional 10 minutes. Using a glass bulb pipette, the maximum volume of the upper organic phase was aspirated into a clean 10mL glass tube and dried under nitrogen. The lower methanolic phase was re-extracted with 250 µL of a MTBE/MeOH/H_2O 10:3:2.5 (v/v/v) mixture. The maximum volume of the lower methanolic phase was aspirated into a clean glass tube and dried under nitrogen. Once dried, both the upper organic phase and lower hydrophilic phase tubes were stored at − 20 °C.

4.3.9.3.4 Additional sample preparations

Three additional experiments were further carried out in conjunction with the LLE prepared samples. A DSPE protocol (DSPE), a single-phase methanol extraction (MeOH) with no lipid removal, and a preparation of a pooled extract (Pool). The creation of a pooled extract to generate a reference sample is common technique used in phenotyping studies (Lewis et al., 2016). As it incorporates all extractions, it represents the average profile and therefore used as a reference to test against the other extraction procedures. To keep the sample volume (75 µL) consistent between the methods, the DSPE protocol was carried out with this revised volume and all parts of the procedure scaled accordingly. All extracts were resuspended with 60 µL of water. A volume of 10 µL were taken from the same replicate number of each extraction and combined to make a pool. For instance, 10 µL were taken from Folch replicate 1, BD replicate 1, Matyash replicate 1, DSPE replicate 1 and MeOH replicate 1, and combined to form Pool replicate 1. All replicates would therefore have a total volume of 50 µL, which was then further diluted to a final volume of 100 µL with water. In addition to being a reference sample, the pool replicates were injected at regular intervals throughout the analytical sequence and used as part of the LOWESS regression for run order correction.

4.3.9.3.5 Profiling data pre-processing and analysis

Individual blank samples and dilution series for all six extractions (Folch, BD, Matyash, DSPE, MeOH and Pool) were also prepared, and used for feature filtering as explained previously in **section 4.3.6**. Briefly, features are filtered out if there are found in the blank sample, have RSD > 30% between replicates (n=6), and a correlation to dilution of <0.7. The total number of features for each protocol were based of these metrics. As the pool represents the average of all extraction protocols, Venn diagrams, were produced to illustrate and compare the number of shared and unique features from each protocol, against the pool. Features included for any multivariate model had to pass the filtering criteria for only the pooled extracts. This allowed a comparison between shared features from all extraction protocols. Probabilistic quotient normalisation (PQN), log transformation and pareto scaling were carried out on these features prior to multivariate analysis. PQN was used to assess the relative differences between the samples extracted in different ways so that all samples are brought into a comparable range for comparison. PCA and Hierarchical clustering analysis – HCA (using Euclidean distance) were used to explore the similarities and trends between the different extraction methods. OPLS-DA models were also implemented to examine the number of discriminating features between the pool, and each extraction method.

4.3.9.3.6 Metabolite annotation and targeted data analysis

Metabolite annotation was made by applying a custom peak fitting algorithm (peakPantheR: https://www.bioconductor.org/packages/release/bioc/html/peakPantheR.html) that utilises the panels of metabolites, pre-annotated and confirmed, in specific analytical assays (RPC-UPLC-MS, HILIC-UPLC-MS and LIPID-UPLC-MS) using UPLC-MS data acquired from an in-house reference standard database. The targeted data for annotated metabolites were subjected to the same feature quality (filtering) process as used for the profiling data and normalised by PQN. Method-induced losses of metabolites identified by peakPantheR, were compared for each extraction as suggested in the work originally published by Klont et al. (Klont et al., 2018). The average intensity for each annotated metabolite was firstly taken from each replicate within each extraction followed by the calculation of the percentage of each average intensity against the most abundant condition. Statistically significant differences ($p < 0.05$, Newman-Keuls multiple comparison test) was performed on the absolute average levels between all extraction comparisons.

4.3.9.4 Solid phase extraction (SPE) comparison

Using the optimised conditions from previous experiments, the DSPE technique was prepared in conjunction with known 96-well SPE extraction plates designed for phospholipid removal. These include, Water's OSTRO, Biotage's ISOLUTE and Phenomenex's PHREE. In addition to these extractions, Phenomenex also packed the Sepra C_{18} material in a 96-well SPE format (Sepra-SPE), in the amount equivalent to the optimised dry weight. The purchased SPE plates were conditioned, equilibrated, and run, according to its own protocols for maximum efficiency in phospholipid removal (details for each extraction method are found in Appendix 2). The Sepra-SPE plate was conditioned and equilibrated in an identical manner as the DSPE sorbent. All extractions were prepared using Devset-plasma plasma, in six replicates. Solid phase extraction was carried out using a vacuum manifold with 10" Hg of vacuum for five to 10 minutes. Neat plasma samples (NEAT) were also included, i.e., samples were aliquoted into a UPLC-MS acquisition plate, with no protein precipitation or solid phase extraction, and used as a control. Ultimately these samples will cause major analytical issues as protein as the sample has not been treated in any way to remove protein, so as a result samples were acquired at the very end of the analytical run. The idea is to use these samples as a control. Samples were analysed by RPC-UPLC-MS (positive and negative ion mode). The total number of features were firstly evaluated using individual dilution series, blanks and %RSD prepared for each extraction condition. Then, a PCA was used to explore any general trends, clustering and outliers for each extraction condition, using only the features that meet the filtering criteria for multivariate analysis as described in **section 4.3.6.**

4.3.10 *Application – biological materials for application to cohort studies*

4.3.10.1 Serum

Serum samples were collected from female patients in an ovarian cancer population (Kaye et al., 2012), that were randomly assigned to one of three different treatment groups; DrugA (n=32), DrugB (n=33) and DrugC (n=33). A subset of these samples (AZ-Study12), consisting of 55 patients at up to four different time points (n=171), were then provided for metabolic phenotyping. Serum was collected from patients in a 10 mL Vacutainer™ tube. Tubes were held at room temperature for a minimum of 30 min but no longer than one hour for sample clotting. The tube was then centrifuged, and the serum supernatant transferred to cryotubes and stored at -80°C until shipment to the NPC under dry ice, where it was stored in a -80°C freezer.

4.3.10.2 Plasma

Plasma samples were collected from male patients from a cross sectional study of a cohort relating to prostate cancer and pelvic radiotherapy (Reis Ferreira et al., 2019). A subset of these samples (MARS), consisting of 285 plasma samples, were provided for phenotyping. Plasma was collected *via* venepuncture into 6mL heparinized vacutainers, immediately placed in ice, then later centrifuged to generate plasma. Plasma supernatant (0.5mL) drawn above the white blood cell layer, were aliquoted into Eppendorf tubes and stored at -80°C. Plasma was stored for a minimum of 8 months prior to being shipped to the NPC, which similarly was in dry ice and stored in a -80°C freezer until analysis.

4.3.10.3 Final sample preparation and analysis

A final sample preparation procedure was carried out on these two exemplar population studies using the optimised DSPE protocol. Sample recording using an in-house laboratory information management system (LIMS), randomisation, allocation, aliquoting and quality control sample preparation are all in accordance with guidelines set from Lewis et al (2016). Plasma samples are removed from the -80°C freezer and allowed to thaw at 4°C for approximately two hours prior to extraction. The samples are then prepared by transferring 100 µL to a 96-well preparation plate. A two-point internal standard (IS) solution, 0.05 µM of L-Phenylalanine-$^{13}C_9$,^{15}N and 0.04 µM of N-Benzoyl-D_5-Glycine, was spiked into the extraction solvent prior to the addition of sorbent. Sorbent was weighed and equilibrated using the protocol described in the sorbent conditioning section. For every 192 samples (two 96-well preparation plates), a slurry of 100 mL solvent (containing IS) at a 16 mg/mL concentration was made. A volume of 325 µL of slurry was added to each sample. Sample and slurry were vortexed thoroughly and left to incubate for 2 hours at 4°C. After centrifugation for 10 minutes at 3214 x g, 212.5 µL of the supernatant was collected and dried under gentle flow of Nitrogen. Resuspension was undertaken in 75 µL of water, containing a mixture of reference standards (Method Reference or MR). The composition of the MR and the concentration is identical to that used in RPC urinalysis (Lewis et al., 2016). The feasibility and robustness of the extraction methodology was then explored using data acquired from RPC-UPLC-MS (positive and negative ion mode), for high resolution detection of chemical species and data processing pipelines. Using the same metric for data quality of global profiling data as stated by Lewis *et al.,* the distribution of the RSD values (remaining features after dilution series), was calculated for the SR samples, and the median RSD values reported. To evaluate the quality of the raw data (uncorrected), the accuracy and precision of the method was examined. The mean RT and peak area % RSD were calculated by targeting the IS and MR from repeated injections of the two QC samples, i.e. SR and the long term reference (LTR). Targeted integration was undertaken using TargetLynx (MassLynx 4.2 SCN 982). Finally, all metabolite annotations were undertaken using peakPantheR and the RPC panel of metabolites. The metabolites annotated where then combined with metabolites annotated from the other NPC blood profiling extraction methods (illustrated as Venn diagrams) for polar metabolites (using HILIC-UPLC-MS) and lipid metabolites (using LIPID-UPLC-MS).

4.4 Results and Discussion

4.4.1 *Optimisation*

4.4.1.1 Sorbent selection

The depletion of lipid species was firstly evaluated by comparing the DSPE plasma profiles acquired by LIPID-UPLC-MS using HSST3, LRA, ENDCAPPED and NON.ENDCAPPED sorbents, against SHAM plasma profiles. The capacity of lipid removal was evaluated in the three most relevant retention time regions of the LIPID profiling method (positive ion mode). The 0-4 minutes (Region one) corresponds to Lysophosphatidylcholine (LPC), monoglycerides (MG), Lysophosphatidylethanolamine (LPE), lipophilic endogenous metabolites (acylcarnitines and bile acids) and xenobiotics; the 4-9 minutes (Region two) corresponds to phospholipids [phosphoglycerols (PG), phosphatidylcholines (PC), phosphatidylethanolamines (PE), phosphatidylserines (PS], sphingomyelins (SM), ceramides (Cer) and Diglycerides (DG); and 9 minutes and onwards (Region three) correspond to only Triglycerides (TG).

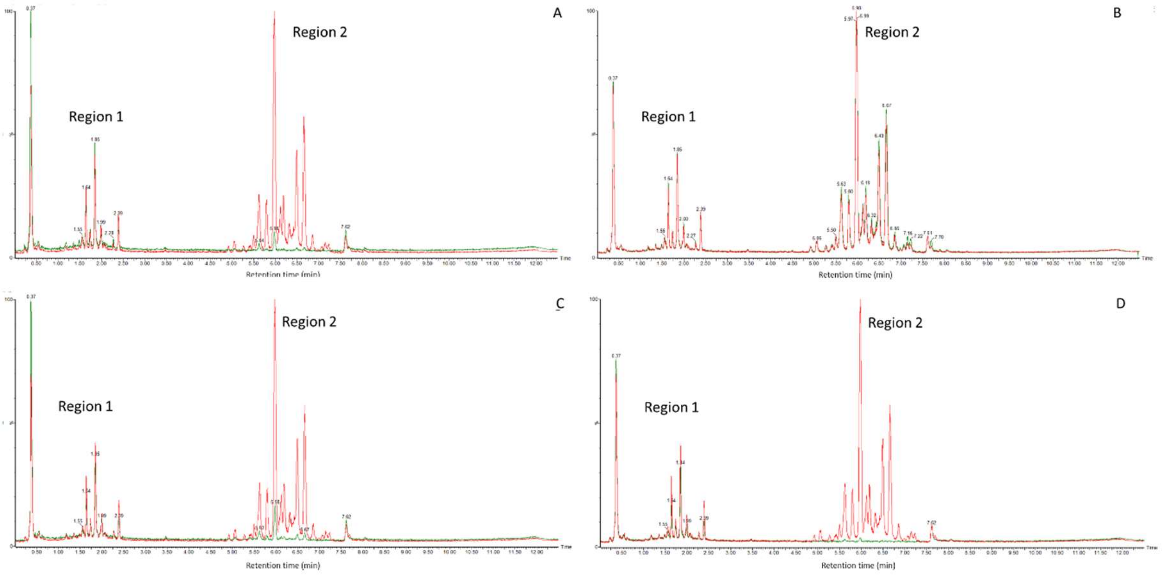


Figure 4-4. A comparison of the lipid profiles acquired by LIPID-UPLC-MS in positive ion mode, from plasma extracts using different DSPE sorbents (green) against the SHAM lipid profiles (red).

(A) Lipid profile of a plasma sample subjected to DSPE, using Waters HSST3 as the sorbent;

(B) Lipid profile of a plasma sample subjected to DSPE, using Sigma NON.ENDCAPPED as the sorbent;

(C) Lipid profile of a plasma sample subjected to DSPE, using Sigma ENDCAPPED as the sorbent;

(D) Lipid profile of a plasma sample subjected to DSPE, using Sigma LRA as the sorbent;

NON.ENDCAPPED sorbent had the poorest capacity to remove lipids, whereas HSST3, ENDCAPPED and LRA had a similar capacity of lipid removal, particularly in Region two of the lipid profile.

From the total ion chromatogram (TIC) illustrated in **Figure 4-4**, the NON.ENDCAPPED sorbent demonstrated very little lipid removal. The other three sorbents showed a similar lipid removal performance, mainly removing lipids associated with Region two. The TG's in Region three were not detected in both SHAM and plasma extractions due to the poor selectivity of MeOH. PCA was then used to display any general trends and compare the different sorbents to the SHAM, in urine samples acquired by RPC (**Figure 4-5**). This is to understand if any perturbation exists in the small molecule profile. From the PCA score plot, the LRA samples clearly clustered away from the SHAM samples and is responsible for the maximum source of variance in the data. A more global effect was observed with the LRA sorbent, which was observed with a significantly lower total signal (TIC) in comparison to the SHAM, when the chromatogram was evaluated. A similarity of the RPC-UPLC-MS profiles were observed between the SHAM samples, to the HSST3 and ENDCAPPED/NON.ENDCAPPED samples, as indicated by the four different groups and their close proximity to one another in the PCA.

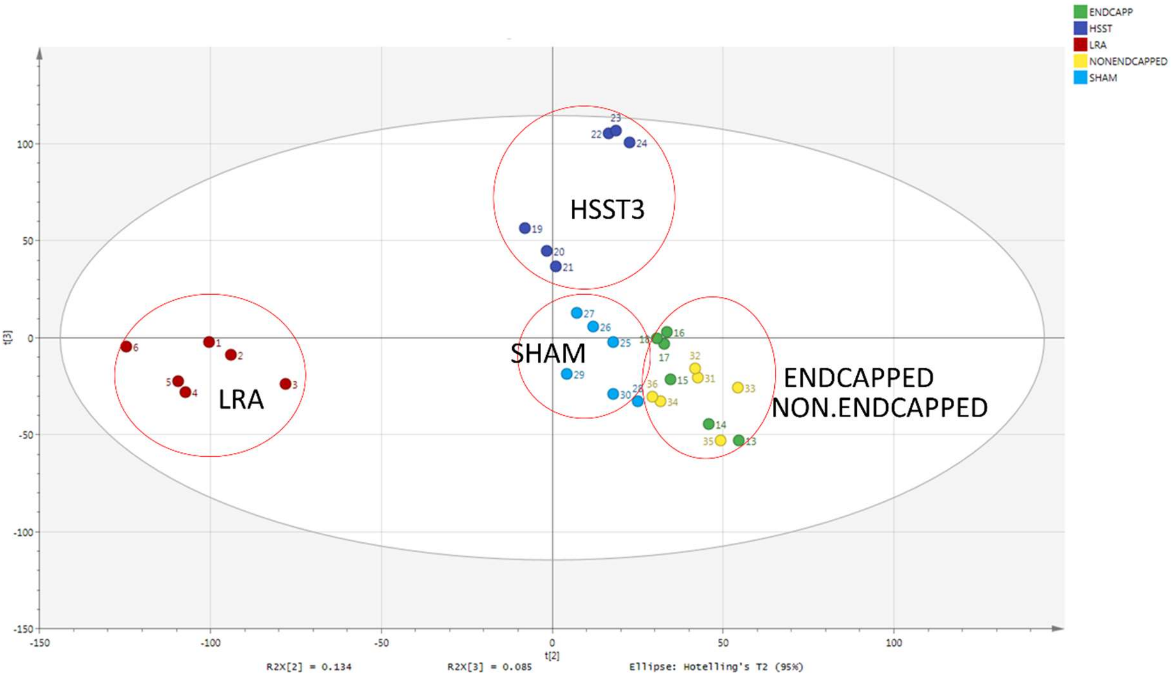


Figure 4-5. A PCA score plot comparing different sorbents used for DSPE in urine samples acquired by RPC-UPLC-MS (positive ion mode), alongside a SHAM sample as reference. The PCA showed clear clustering of samples based on the sorbent used in the DSPE treatment. LRA (brown) clustered furthest away from the other sorbents and was the main source of variance observed in the data. SHAM (light blue), HSST3 (dark blue), ENDCAPPED (green) and NONENDCAPPED (yellow) shared a similar RPC-UPLC-MS profile as indicated by their close proximity in the PCA.

However, separation of clusters indicates differences between the sorbent groups. As a consequence of the poor lipid removal efficiency of the NON.ENDCAPPED sorbent, and a lower total signal associated with the LRA samples, these two sorbents were no longer considered *via*ble options for our intended purpose of lipid removal from blood products. OPLS-DA models were then used to identify the features responsible for driving the separation between the SHAM samples, and the remaining HSST3 and ENDCAPPED samples. The R2Y and Q2Y values obtained with a single calculated component were 0.9 and 0.85 for HSST3 (**Figure 4-6.A**), and 0.9 and 0.73 for ENDCAPPED (**Figure 4-6.B**) and permutation testing indicated low variability and an excellent predictive ability. Features were coloured by "*w**", which represent the rotated weights, and how each variable/feature individually corelates with Y. Colouring by this parameter, highlights the features (X variables) which best associate with Y, and in addition, get a sense of direction with the latent variable. The scatter points closer to blue in colour therefore represents a decrease in signal associated with DSPE treatment. So, although the HSST3 and ENDCAPPED sorbents demonstrated lipid removal, particularly in Region two of the lipid profile (from the plasma extracts acquired by LIPID-UPLC-MS in positive ion mode), moderately hydrophobic small molecules (as indicated by features eluting after six minutes in the urine RPC-UPLC-MS profile) were also affected to a certain degree. A thorough examination into the specifications of each sorbent provided an explanation as to what was being observed experimentally.

Both sorbents are endcapped; endcapping is a process by which residue silanol groups which are void of C_{18} attachment, have substituted functional groups attached. Due to steric blocking, particles can have many "unexposed" silanol groups which are highly polar and can interact with the polar and moderately hydrophobic properties of small molecule metabolites. The substitution of a specific compound to any remaining unexposed silanol group heads, is therefore referred to as the process of endcapping. The ENDCAPPED sorbent utilises trimethyl silane (TMS) for endcapping, whereas the HSST3 uses a proprietary trifunctional C_{18} alkyl phase. It seems that the degree of endcapping for each sorbent, did not completely eliminate the interaction with moderately hydrophobic small molecules. To significantly reduce the amount of interaction with the endcapped silanol groups, the sorbent therefore needed to exhibit a smaller pore diameter and larger surface area. This results in an increase in the ligand density, or the number of fatty acid chain (C_{18}) attachment, reducing the number sites needed for endcapping and therefore greater interaction with only lipophilic species. The ligand density for these two sorbents are proprietary, however relates to the carbon load which is a reported property. The results from this investigation therefore required the procurement of a material with structural properties better suited for lipid removal. At the time of this experiment, the Sepra C_{18} sorbent from Phenomenex, was the only product commercially available to purchase in a bulk powder form that fulfilled these criteria, with specifications far more favourable than the HSST3 or ENCAPPED sorbents. It was therefore the sorbent of choice for subsequent experiments.

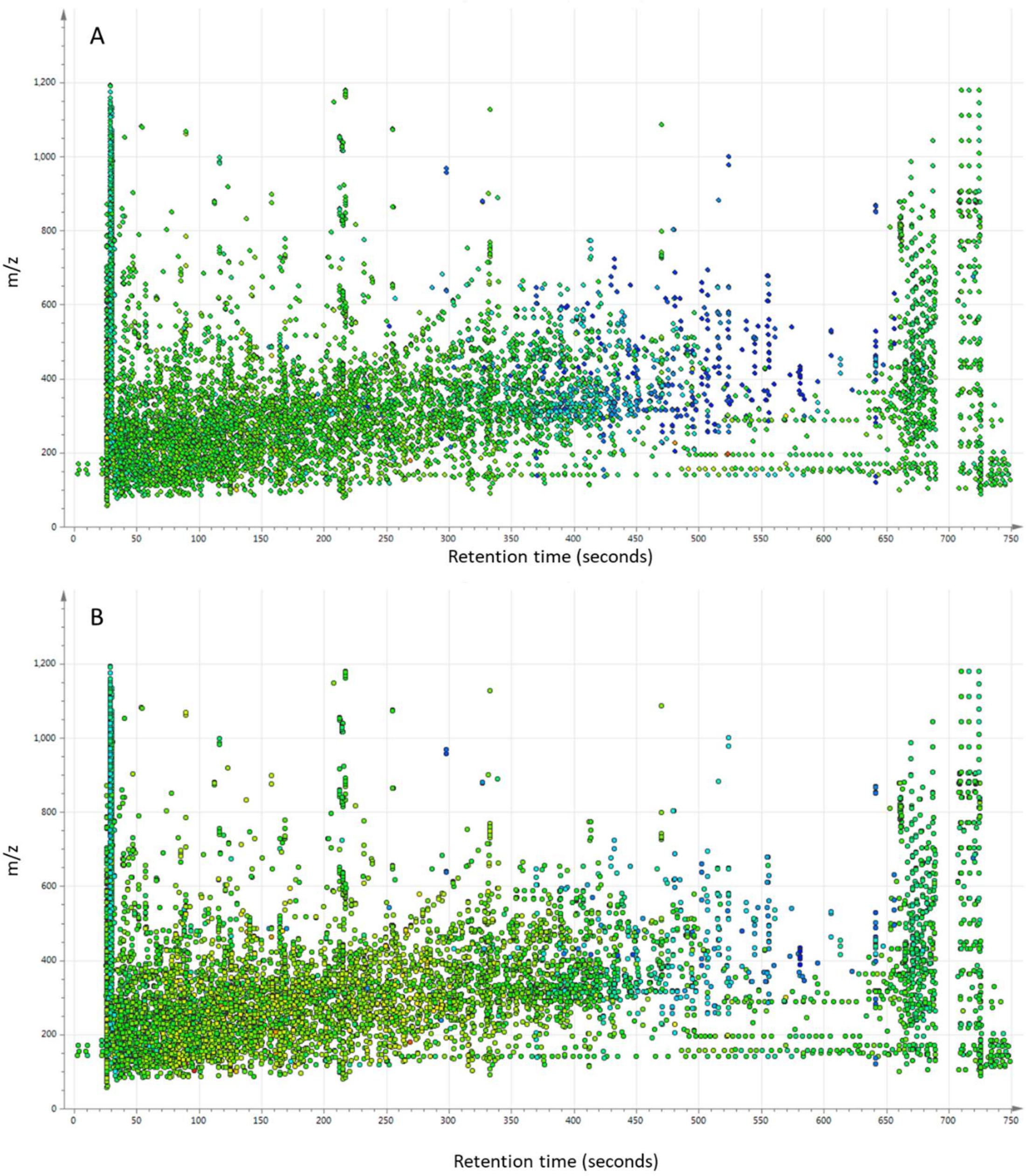


Figure 4-6. OPLS-DA loadings plots, comparing the SHAM reference samples to DSPE treatment samples (acquired by RPC-UPLC-MS in positive ion mode) using the HSST3 sorbent and ENCAPPED sorbent.

(A) OPLS-DA loading plot comparing the SHAM treated samples, to the HSST3 sorbent DSPE treated samples. The R2Y and Q2Y values obtained with a single calculated component were 0.9 and 0.85

(B) OPLS-DA loading plot comparing the SHAM treated samples, to the ENCAPPED sorbent DSPE treated samples. The R2Y and Q2Y values obtained with a single calculated component were 0.9 and 0.73

The loadings are coloured *w** which represents and how each variable/feature individually corelates with Y. The scatter points closer to blue in colour therefore represents a decrease in signal associated with DSPE treatment. So, although the HSST3 and ENDCAPPED sorbents provide a certain level of lipid removal (as demonstrated from the plasma extracts acquire by LIPID-UPLC-MS in positive ion mode), moderately hydrophobic small molecules (as indicated by features eluting after six minutes in RPC-UPLC-MS positive ion mode) were also being affected in the urine by both DSPE treatments.

Finally, volume of sorbent displacement was also evaluated for HSST3, ENDCAPPED and Sepra $C_{18.}$ A 5 mL volumetric flask was filled to volume with MeOH and weighed. A random amount of sorbent was added, and the flask reweighed. Using a syringe, a volume was drawn to bring the meniscus back to the 5mL mark on the flask. This drawn volume was then weighed and from this, the displacement volume can be extrapolated. As the displacement volume corresponds to the original amount of sorbent added, the volume which equates to 1mg can be calculated. The experiment was replicated in triplicate for all three sorbents, with random amounts of sorbent added to the 5mL flask each time. All three sorbents had approximately 0.7 µL of volume displaced for every 1 mg of sorbent and therefore regarded as negligible.

4.4.1.2 *Solvent optimisation*

The depletion of lipid species was firstly evaluated by their total ion chromatograms as depicted in **Figure 4-7**, in the three most relevant retention time regions of the lipid profiling method (LIPID-UPLC-MS positive ion mode) as specified in **section 4.4.1.1**. IPA and Acet, being less polar than the others, demonstrated to be the poorest solvent of choice, as lipids had a greater affinity to remain in the solvent and not bind to the sorbent. MeOH, EtOH and MeCN had efficient lipid removal in Regions two and Region three, but with little effect in Region one. Mainly, polar lysophospholipids (LPC) elute in Region one, but so to do other small molecules, such as lipophilic xenobiotics and acylcarnitines. It is these classes of compounds we want to remain unaffected by the DSPE method, and so a compromise was made to have metabolites within Region one remain in solution.

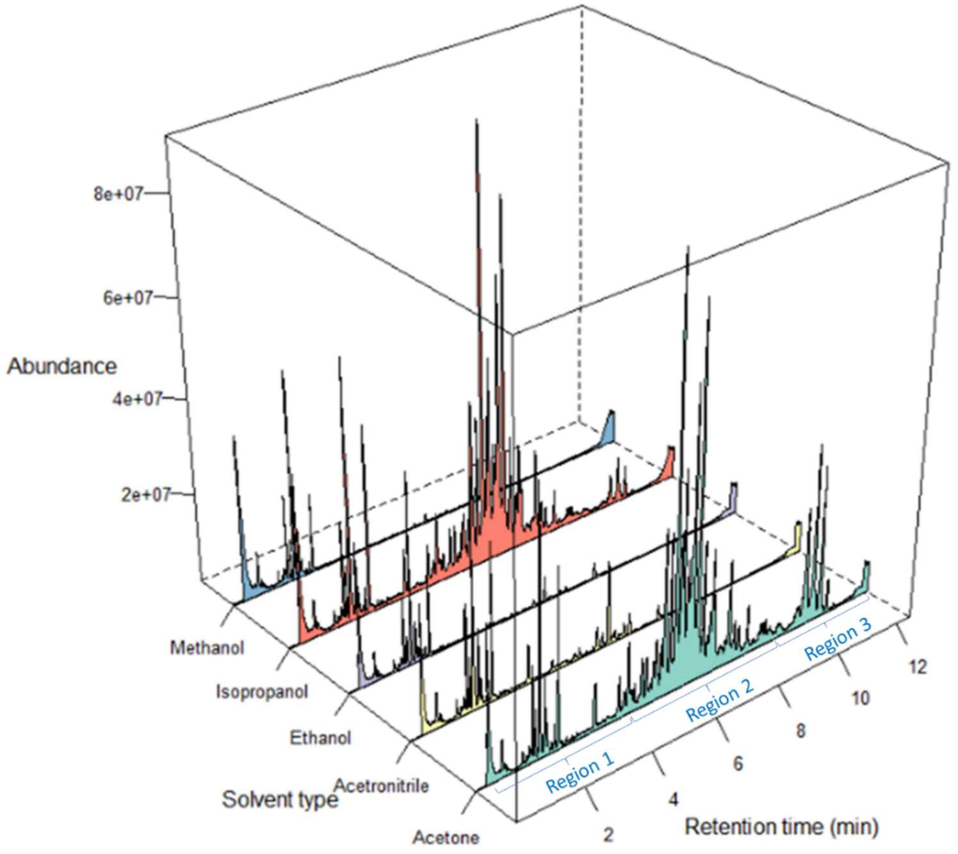


Figure 4-7. The total ion chromatogram comparing the effect five different extraction solvents (Methanol-MeOH, Isopropanol-ISP, Ethanol-EtOH, Acetonitrile-MeCN and Acetone-Acet), used in DSPE from plasma samples analysed by LIPID-UPLC-MS in positive ion mode. The three retention time regions relevant to LIPID analysis were examined. The 0-4 minute (Region one), the 5-9minute (Region two), and the 9min and onwards (Region three). MeOH, EtOH and MeCN produced a “cleaner” TIC spectrum which represents adequate removal of lipid species, especially in regions 1 and 2. Whereas ISP and Acet demonstrated denser TIC in all three regions, highlighting poor lipid removal. Both ISP and Acet were eliminated as viable options.

A secondary investigation was undertaken using solvent extraction conditions involving only MeOH, EtOH and MeCN. The total number of features detected by RPC-UPLC-MS in positive ion mode followed the order: EtOH > MeOH:EtOH > MeOH:MeCN > MeCN:EtOH > MeOH > MeCN > MeOH:MeCN:EtOH (6408, 6368, 6253, 5889, 5091, 4680 and 4248, respectively). In negative ion mode, the same order was observed (2416, 2398, 2281, 2281, 2239, 1974 and 1934, respectively). Overall, all solvents showed a similar extraction performance in negative ion mode, i.e. the number of features, median RSD levels and evaluation of the chromatography (**Figure 4-8.B**). Differences however were observed in positive mode. Although the extraction with only EtOH produced the highest number of features, the median RSD for these features was 68% (**Figure 4-8.A**). MeOH:EtOH was the next highest in the number of features detected, and was far more reproduceable, sharing more than 80% of features to the EtOH samples. However, a higher baseline, and large asymmetrical peaks, composing of many *m/z* values, were observed after seven minutes in the chromatogram (**Figure 4-8.A**). Any solvent composition with EtOH, produced these features, and is the underlying reason for the poor reproducibility associated with these samples. The MeOH samples also produced these features, and although lower in intensity and more reproduceable than EtOH extracts, still resulted in carry-over issues during analytical acquisition which eventuated into an increase in overall system pressure. The MeCN extraction has a lower baseline and a significantly lower signal of these large asymmetrical peaks, after the eight-minute region of the chromatogram. However, the MeCN only extracted samples did produce the lowest number of features and was more variable when compared to the other extraction solvents. Addition of MeCN to high salt samples has known to result in inconsistent biphasic partitioning of metabolites, leading to poor reproducibility (Watts and McDonald, 1990). However, in many cases, MeCN has reported to be superior to other solvents in terms of protein precipitation (Polson et al., 2003), therefore the incorporation of MeCN may be necessary for adequate protein removal. The MeOH:MeCN extracted samples produced the third highest number of features, and the lowest median %RSD at approximately 12%. The box plots in both modes also demonstrated this composition to be the least scattered and dispersed than the other solvents. To address the possible partitioning of metabolites affiliated with MeCN extractions, an experiment was conducted to test the effect of MeOH addition, to MeCN and a strong salt solution. One part of a 500mM sodium chloride solution, was mixed with three parts of four different solvent compositions, i.e., 100:0 MeCN:MeOH, 72:25 MeCN:MEOH, 50:50 MeCN:MeOH, and 25:75 MeCN:MeOH. As predicted, the 100:0 MeCN:MeOH produced two distinct layers. The addition of at least 25% of MeOH to a strong salt solution however, negated the biphasic behaviour (**Figure 4-9**). Therefore a 1:1 MeOH:MeCN mixture should have no issue handling high salt samples and metabolite partitioning during extraction.

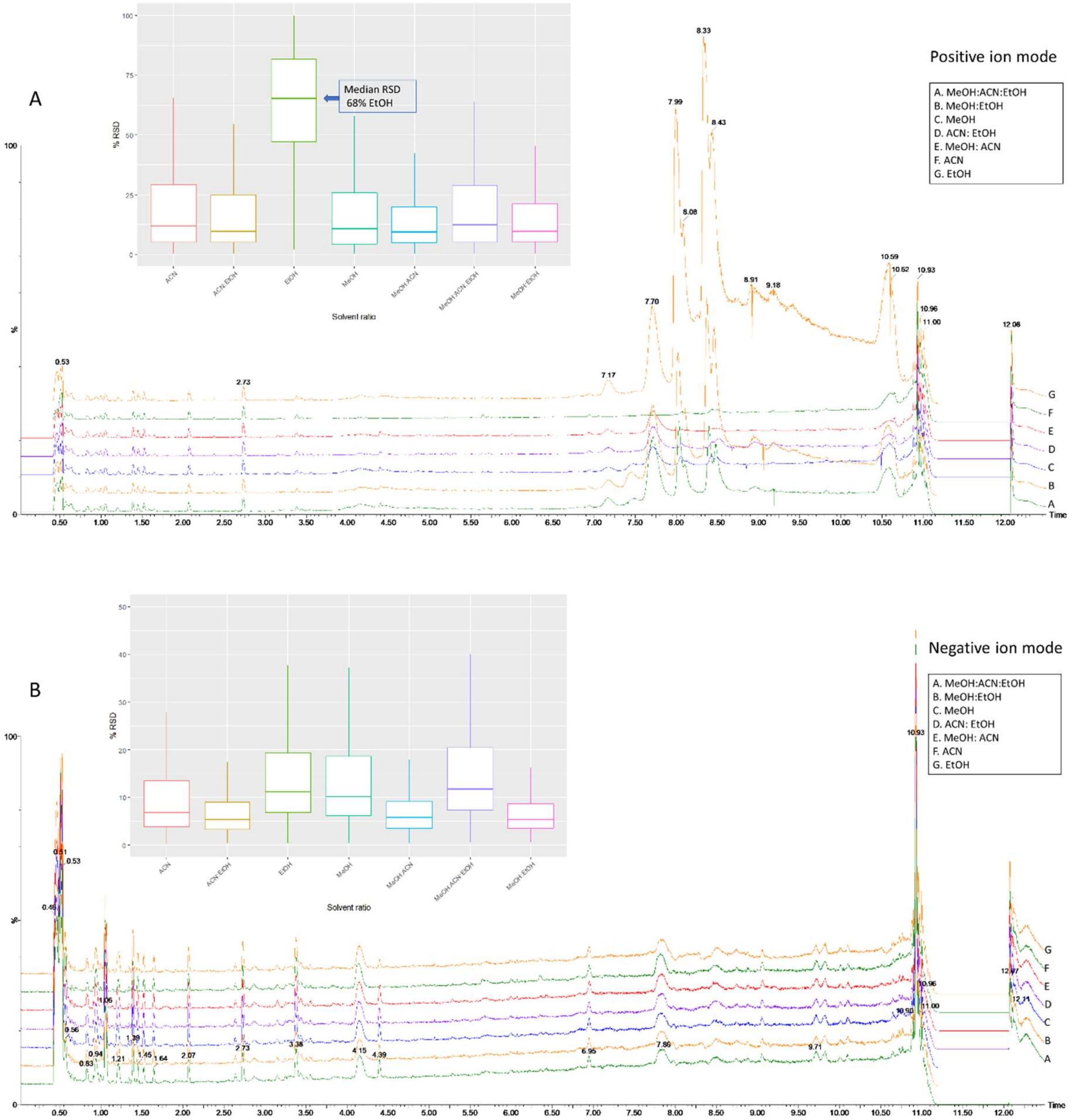


Figure 4-8. Total ion chromatograms (TIC) and %RSD values for features detected using the solvent compositions, EtOH, MeOH:EtOH (1:1), MeOH:MeCN (1:1), MeCN:EtOH (1:1), MeOH, MeCN, and MeOH:MECN:EtOH (1:1:1), for DSPE in plasma analysed by RPC-UPLC-MS (positive and negative ion mode).

(A) TIC and % RSD values for features detected in all extraction solvent conditions analysed by RPC-UPLC-MS (positive ion mode);

(B) TIC and % RSD values for features detected in all extraction solvent conditions analysed by RPC-UPLC-MS (negative ion mode);

A similar extraction performance was observed for all solvent compositions in RPC-UPLC-MS (negative ion mode). In RPC-UPLC-MS (positive ion mode), EtOH produced the highest number of features, however, was most variable and exhibited a higher baseline, and large asymmetrical peaks, composing of many *m/z* values.

MeOH also demonstrated these characteristics, but to a lesser extent. Ultimately, MeOH:MeCN composition produced the most features which were least variable, and incorporated both a methanol and acetonitrile component for efficient protein precipitation, thereby eliminating any likelihood of partitioning between sample and solvent.

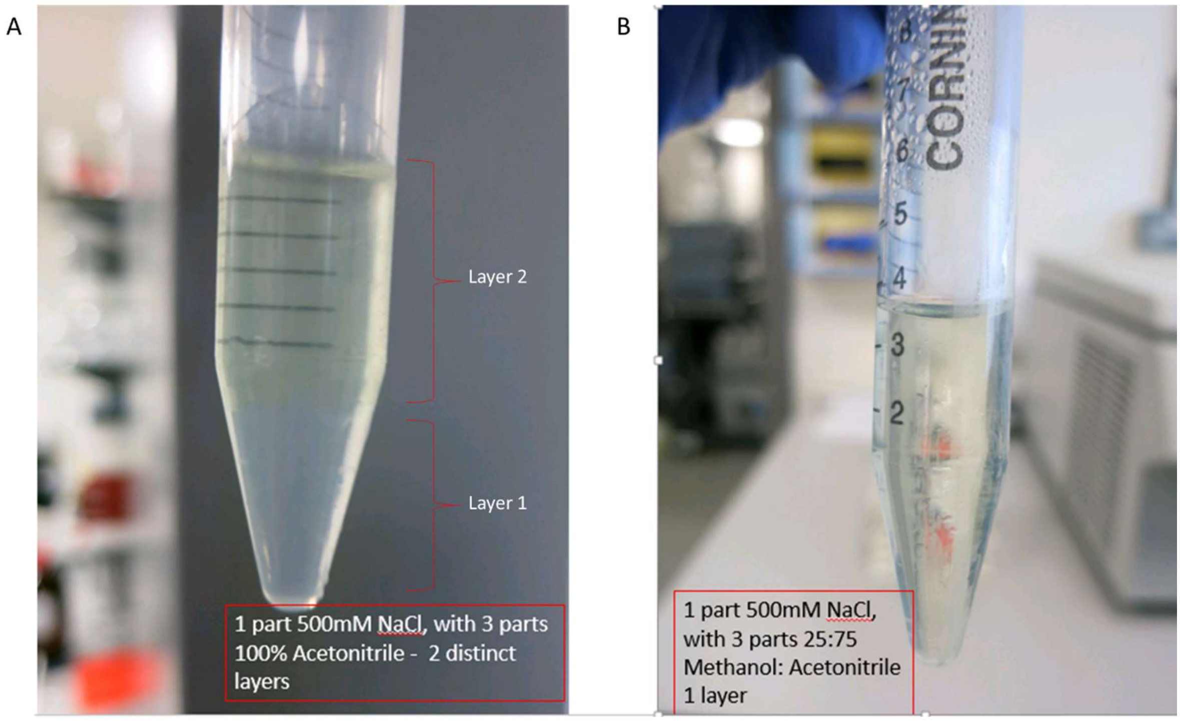

A
Layer 2
Layer 1
1 part 500mM NaCl, with 3 parts 100% Acetonitrile - 2 distinct layers
B
CORNING
1 part 500mM NaCl, with 3 parts 25:75 Methanol: Acetonitrile 1 layer

Figure 4-9. A comparison of a high concentrated salt sample (with yellow dye) mixed with acetonitrile, and mixed with a 1:1 MeOH:MeCN mixture.

(A) 1part 500mm NaCl mixture mixed with 3 parts MeCN. ;

(B) 1part 500mm NaCl mixture mixed with 3 parts MeOH:MeCN (1:1);

A clear biphasic separation is observed in the sample extracted purely with acetonitrile. The salt solution was dyed yellow to emphasis the clear separation between the salt solution and acetonitrile. A solvent composition incorporating as little as 25:75 MeOH:MeCN however, negated any biphasic separation between sample and solvent.

In summary, the selection of 1:1 MeOH:MeCN solvent mixture demonstrated a high number of reproduceable features, lower baseline, a significantly reduced signal of the large asymmetric peaks, and eliminated any likelihood of partitioning between sample and solvent. As a compromise was chosen for partial lipid removal in the earlier experiment, chromatographic performance parameters, such as pressure build up, carry-over and retention time drift were all assessed and considered negligible.

4.4.1.3 Slurry optimisation-DOE

The complexity involved in slurry optimisation was handled more efficiently and systematically using the DOE approach. The goal of this optimisation is to determine what slurry condition is best for lipid removal whilst maintaining maximum recovery of the remaining small molecules. A total of 11 experiments were undertaken as proposed by MODDE. If the experiment were to be undertaken in the conventional sense, i.e. testing all experimental conditions, the final experiment would have resulted 55 different conditions. This immediately highlights the benefit of the DOE approach to speed up the design process, thereby reducing cost and time. The recoveries from RPC-UPLC-MS (positive and negative ion mode), and LIPID-UPLC-MS (positive ion mode) were combined and tabulated (**Table 4-3**). The RT bins used on the data acquired from LIPID-UPLC-MS were based on regions one and two. Region three was not explored as TG's and DG's which elute in this region, would not extract due to poor selectivity of the MeOH/MeCN solvent. The recoveries beyond 10 minutes on RPC-UPLC-MS (positive ion mode), were substantially high (>130%). An explanation for this is that the washing phase of the chromatographic gradient (high solvent flow through the column) for this method is during this time, i.e. ranges from 10 minutes to 11.5 minutes. Chemicals which are very hydrophobic and lipophilic elute at this retention time range and can potentially affect the small molecule profile and precision of the data. Both RPC-UPLC-MS and LIPID-UPLC-MS are reversed phase methods, each having a limited useable range, with the former geared to features of lesser hydrophobicity. This means there is an overlap in metabolite retention between the two profiling methods. A simple experiment was then conducted to determine whether metabolites which elute past 10min by the RPC-UPLC-MS (positive ion mode) method, are still captured, with suitable retention, by the LIPID-UPLC-MS (positive ion mode) method. This involved acquiring different classes of analytical reference standards by the two methods. The classes included a selection of two antibiotics, three carnitines, nine xenobiotics (medicinal drugs), two LPC's and 45 Bile acids. Illustrated in **Figure 4-10**, the horizontal red dotted line represents the RT at 10minutes for RPC-UPLC-MS, and the vertical red dotted line, represents the RT of two times the solvent front (t_0), in the LIPID-UPLC-MS assay, i.e. the t_0 at 0.6 minutes. This essentially splits the figure into four quadrants and it is clear that none of the compounds fall in quadrant 1, meaning that all compounds which elute after 10 minutes by RPC-UPLC-MS, are captured in the LIPID-UPLC-MS assay and elute after the $t_{0.}$ We can therefore conclude that a cut-off RT at 10 min for RPC-UPLC-MS is acceptable and was kept consistent between the two polarities. The recoveries from RPC-UPLC-MS (positive and negative ion modes), with the updated RT ranges, and LIPID-UPLC-MS (positive ion mode) were combined into one DOE model and inputted into MODDE.

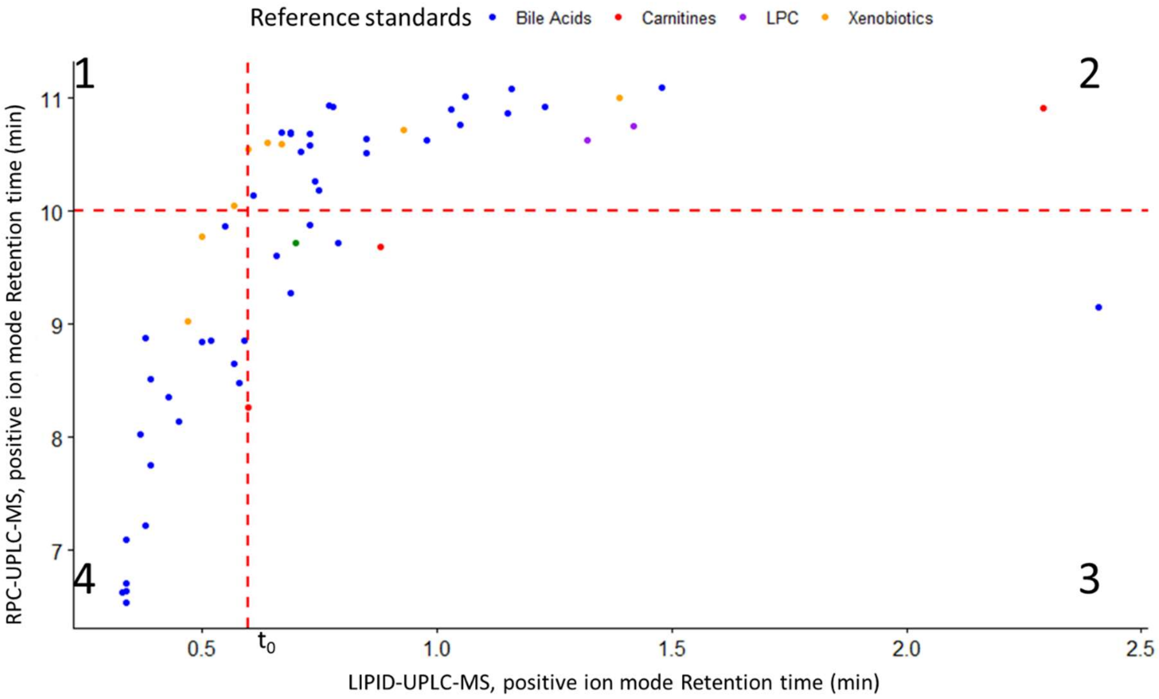


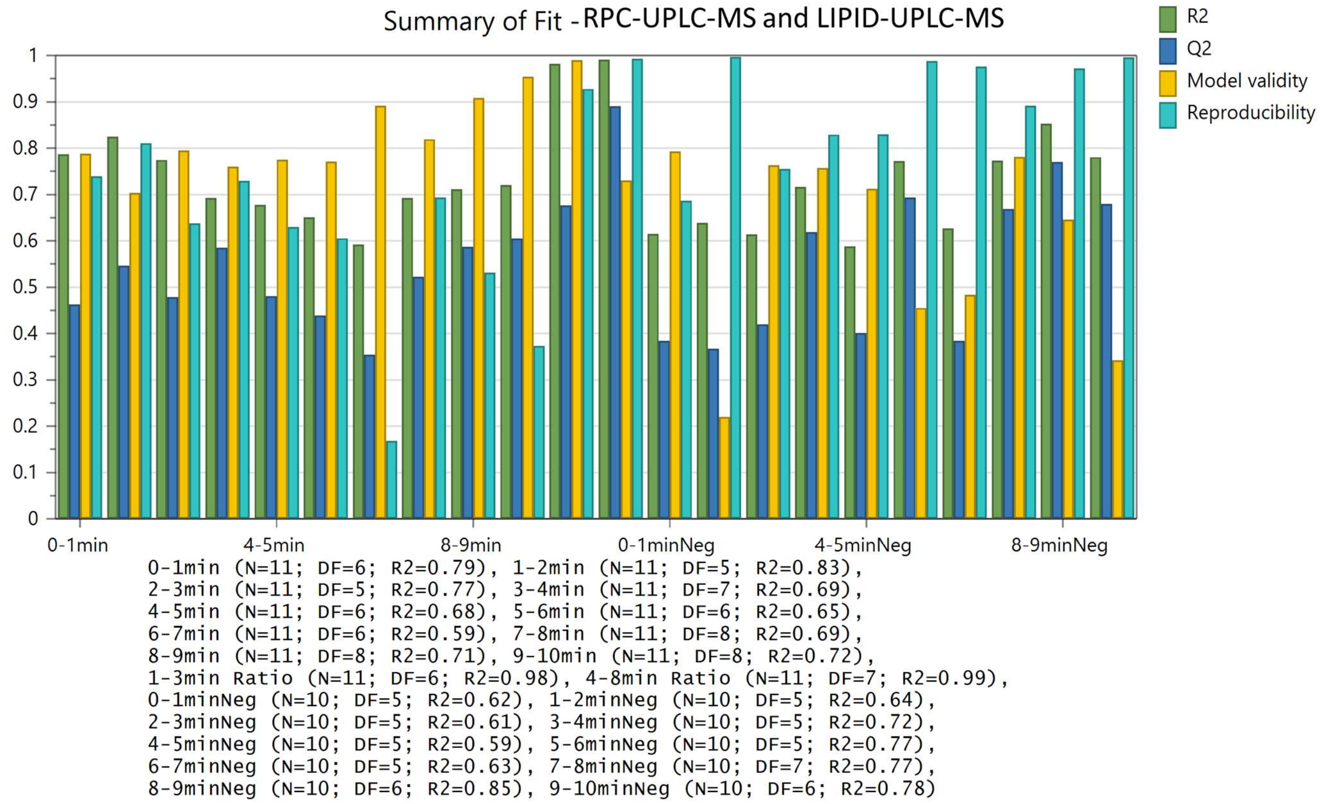
Figure 4-10. A mixture of lipophilic reference standards acquired by both LIPID-UPLC-MS and RPC-UPLC-MS profiling methods (both in positive ion mode), plotted on different axis to demonstrate the overlap in retention time between these two methods. The reference standards include two antibiotics (green), three carnitines (red scatter points), nine xenobiotics (yellow scatter points), two LPC's (purple scatter points) and 45 Bile acids (blue). The vertical red dotted line represents the t_0 of the LIPID-UPLC-MS method, and the horizontal red lone represents the retention time at 10 minutes for the RPC-UPLC-MS method. This splits the figure into four quadrants (1-4). Reference standards eluting after 10 minutes by RPC-UPLC-MS, eluted (with acceptable retention) well after the t_0 at 0.6 minutes in the LIPID-UPLC-MS method (quadrant 2). Therefore, the useable retention time range for the RPC-UPLC-MS method of blood extracts will stop at 10 minutes for all subsequent experiments and in both polarities.

Table 4-3. The response recoveries inputted into the DOE model, calculated as a percentage of the measured TIC signal (within a retention time window – or bin) observed in the DSPE treatment against the SHAM treatment. Ideal targets for the response recoveries were recorded as a range between 90-130% for RPC-UPLC-MS analyses (positive and negative ion mode) and 20-40% for LIPID-UPLC-MS (positive ion mode) analysis. One of the Replicate points in RPC-UPLC-MS (negative ion mode) was removed as a result of a mis-injection.

		Slurry conditions: Concentration and Volume										
MODDE	RT bins	2mg/mL	20mg/mL	2mg/mL	20mg/mL	2mg/mL	20mg/mL	11mg/mL	11mg/mL	11mg/mL	11mg/mL	11mg/mL
	(min)	600 µL	600 µL	1000 µL	1000 µL	800 µL	800 µL	600 µL	1000 µL	800 µL	800 µL	800 µL
RPC-UPLC-MS (positive ion mode)	0-1	99.00%	103.54%	105.54%	106.49%	94.58%	102.97%	100.93%	107.61%	101.18%	102.17%	104.84%
	1-2	99.54%	106.87%	107.98%	110.46%	93.44%	107.50%	102.89%	114.46%	103.70%	105.18%	108.49%
	2-3	98.75%	105.27%	107.59%	109.32%	93.36%	107.03%	101.55%	113.11%	103.52%	104.17%	109.58%
	3-4	95.89%	111.10%	108.28%	111.42%	89.99%	110.34%	101.33%	113.20%	107.93%	104.66%	112.40%
	4-5	95.85%	109.67%	107.07%	107.17%	88.04%	110.86%	98.52%	110.25%	104.19%	99.96%	108.79%
	5-6	93.44%	107.87%	102.90%	111.04%	89.40%	122.51%	97.80%	130.26%	104.06%	106.64%	118.63%
	6-7	93.44%	106.46%	106.65%	104.74%	83.79%	103.68%	96.72%	108.82%	94.99%	99.60%	109.05%
	7-8	97.34%	110.45%	112.48%	122.68%	85.97%	116.55%	100.69%	125.18%	105.61%	109.03%	118.04%
	8-9	94.98%	111.77%	109.01%	132.75%	80.48%	118.34%	100.98%	128.89%	101.83%	112.99%	122.88%
	9-10	93.78%	113.37%	116.86%	146.41%	76.96%	124.21%	95.44%	127.58%	96.96%	120.86%	126.63%
RPC-UPLC-MS (negative ion mode)	0-1	96.89%	102.55%	98.63%	100.61%	93.72%	101.42%	101.51%	97.79%	103.13%	105.92%	
	1-2	100.67%	106.69%	98.54%	100.38%	92.14%	99.47%	103.51%	101.03%	105.46%	105.78%	
	2-3	99.53%	105.49%	98.50%	99.49%	91.58%	103.97%	102.04%	100.69%	105.82%	109.26%	
	3-4	100.24%	107.54%	98.69%	103.64%	89.86%	106.24%	105.38%	104.21%	108.20%	111.78%	
	4-5	100.34%	106.23%	99.73%	102.21%	89.65%	103.80%	102.92%	101.98%	105.21%	108.17%	

		Slurry conditions: Concentration and Volume										
MODDE	RT bins	2mg/mL	20mg/mL	2mg/mL	20mg/mL	2mg/mL	20mg/mL	11mg/mL	11mg/mL	11mg/mL	11mg/mL	11mg/mL
	(min)	600 µL	600 µL	1000 µL	1000 µL	800 µL	800 µL	600 µL	1000 µL	800 µL	800 µL	800 µL
	5-6	101.97%	114.00%	103.66%	108.50%	88.78%	114.06%	112.02%	114.15%	114.11%	115.40%	
	6-7	105.78%	115.40%	104.52%	105.55%	90.67%	109.31%	110.12%	107.90%	112.28%	113.81%	
	7-8	108.35%	129.75%	108.32%	128.14%	100.57%	137.57%	122.78%	123.55%	137.33%	144.00%	
	8-9	111.75%	147.60%	115.85%	156.89%	103.47%	164.57%	138.26%	151.04%	165.95%	171.72%	
	9-10	120.49%	150.60%	121.40%	163.54%	104.98%	172.54%	144.31%	150.72%	177.34%	179.53%	
LIPID-UPLC-MS (positive ion mode)	1-3.5	101.47%	69.92%	106.38%	89.94%	103.34%	81.31%	82.50%	94.11%	83.61%	89.13%	87.96%
	4-8	84.75%	3.41%	86.13%	15.01%	78.39%	6.63%	7.50%	21.86%	12.14%	17.02%	16.86%

Models for each response demonstrated a Q2/R2 > 0.5 as observed with in **Figure 4-11**. In addition to these metrics, model validity of >0.25, indicating no lack of fit, and reproducibility of >0.5 was observed, all of which demonstrates good experimental control and robustness of the model (Jänsch et al., 2019).



```
0-1min (N=11; DF=6; R2=0.79), 1-2min (N=11; DF=5; R2=0.83),
2-3min (N=11; DF=5; R2=0.77), 3-4min (N=11; DF=7; R2=0.69),
4-5min (N=11; DF=6; R2=0.68), 5-6min (N=11; DF=6; R2=0.65),
6-7min (N=11; DF=6; R2=0.59), 7-8min (N=11; DF=8; R2=0.69),
8-9min (N=11; DF=8; R2=0.71), 9-10min (N=11; DF=8; R2=0.72),
1-3min Ratio (N=11; DF=6; R2=0.98), 4-8min Ratio (N=11; DF=7; R2=0.99),
0-1minNeg (N=10; DF=5; R2=0.62), 1-2minNeg (N=10; DF=5; R2=0.64),
2-3minNeg (N=10; DF=5; R2=0.61), 3-4minNeg (N=10; DF=5; R2=0.72),
4-5minNeg (N=10; DF=5; R2=0.59), 5-6minNeg (N=10; DF=5; R2=0.77),
6-7minNeg (N=10; DF=5; R2=0.63), 7-8minNeg (N=10; DF=7; R2=0.77),
8-9minNeg (N=10; DF=6; R2=0.85), 9-10minNeg (N=10; DF=6; R2=0.78)
```

Figure 4-11. A summary of fit plot from the DOE model. Summary of fit mode displaying four model performance indicators; the R^2 – model fit (green), Q^2 – estimate of the model to predict new data (blue), Model validity (yellow) and Reproducibility (cyan) for the DOE model incorporating all factor and response analyses from the RPC-UPLC-MS (positive and negative ion mode) and LIPID-UPLC-MS (positive ion mode) data. A good model is characterised by both R^2 and Q^2 close to 1, model validity of >0.25, and reproducibility of >0.5. The model has been fitted using PLS. The output summary statistic indicated the model to be sufficient for this optimisation.

Regression coefficients and their confidence intervals or uncertainties are shown in **Figure 4-12** with slurry concentration labelled as “C18”, and slurry volume labelled as “Vol”. In RPC-UPLC-MS (positive and negative ion mode), both the slurry concentration and slurry volume behaved similarly with no significant impact in the retention time range spanning 0-8minutes. This is to be expected as there should be no change in the small molecule profile when plasma is subjected to DSPE. A factor is only considered significant, if the factor effect is greater than its error, and in all RT bins associated with

RPC-UPLC-MS (positive and negative ion mode), this was observed. One can argue however, that a small influence may be observed, in the RT bin corresponding to 8-10 minutes in both ionisation modes and in the positive direction. This means, that as either factor increases, the higher the response (away from the ideal target) is observed. This could be attributed to the fact that lipophilic features elute later in this assay and so an increase in slurry concentration can result in less suppression on other small molecules that elute at this time. In the LIPID-UPLC-MS profiling assay, the regression coefficient for the slurry concentration (C18) is in the negative direction and significant to the model in regions one (1-3 min) and two (4-8 min). What this represents is what is already known, and that is the higher the sorbent concentration, the more lipid is removed, resulting in lower recoveries.

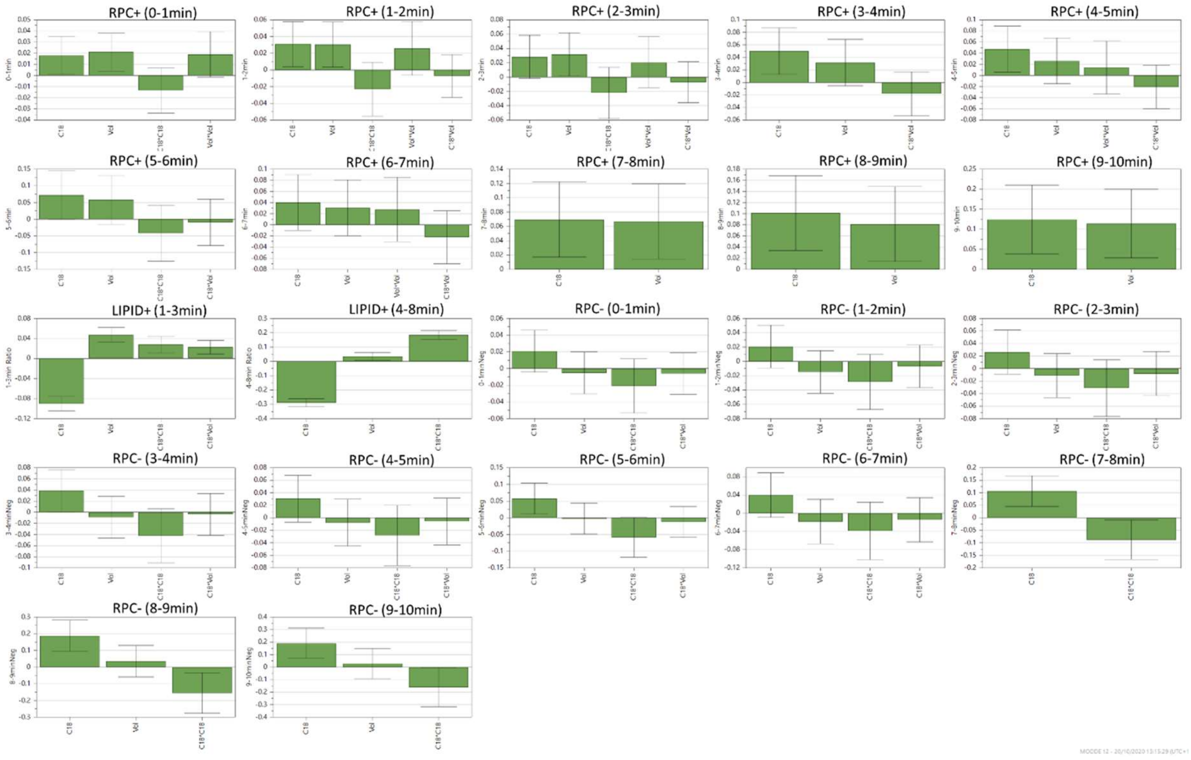
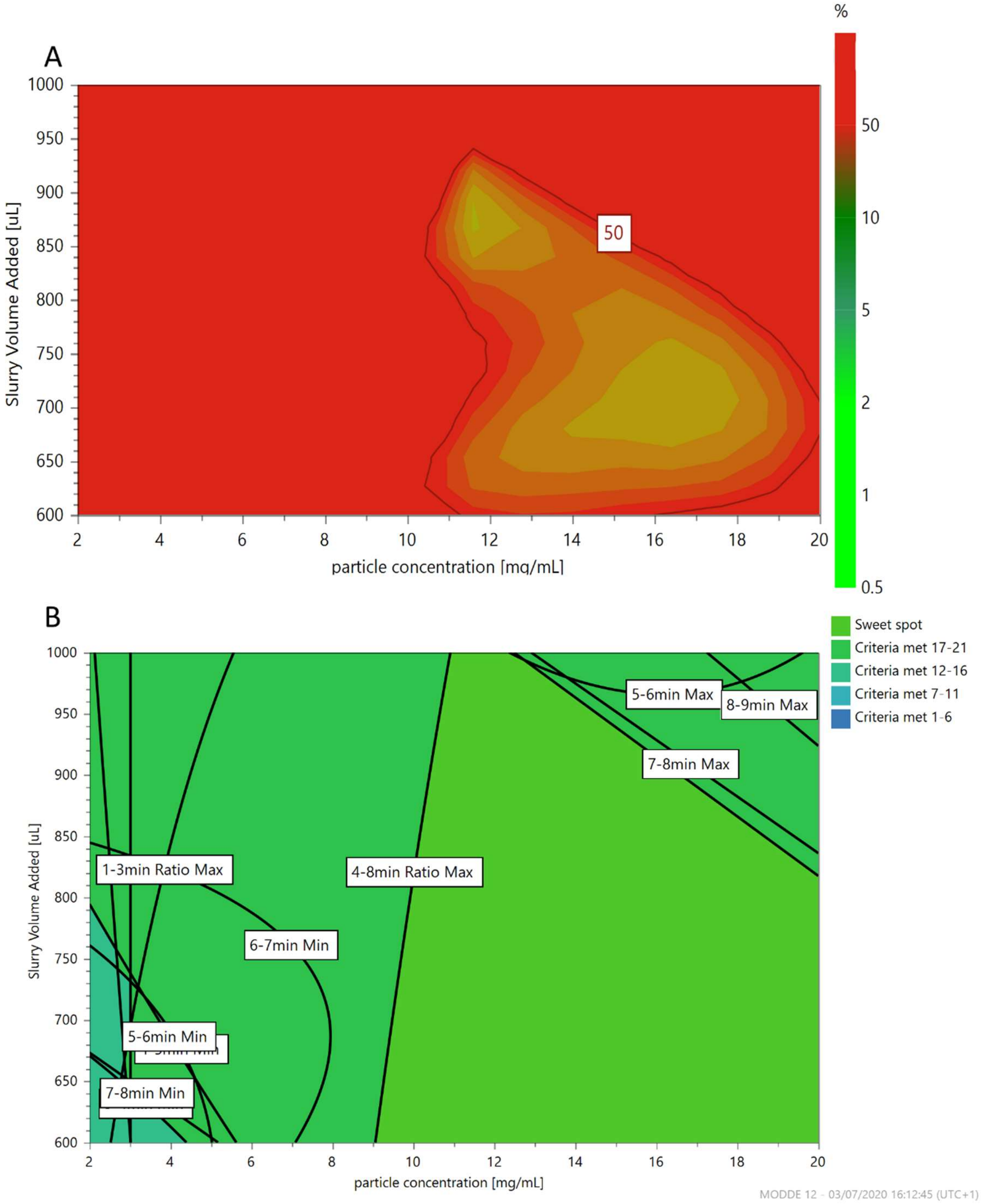


Figure 4-12. The scaled and centered regression factors calculated from the DOE optimisation for RPC-UPLC-MS (RPC+ and RPC-) and LIPID-UPLC-MS (LIPID+). The regression factors were slurry concentration (C18), slurry volume (Vol) and interaction factors; C18*C18, Vol*Vol and C18*Vol. In most retention time bins, certain factors were removed to improve the Q2/R2 ratio. Large regression coefficients represent factors with large contributions to the response. If a factor's regression coefficient is smaller than the associated errors bars, then the factor is not significant. A slight positive influence was observed in the retention time bin associated with 8-10 minutes in RPC-UPLC-MS (both ionisation modes; RPC + and RPC -), indicating higher recoveries were observed. The opposite was in fact observed in LPID-UPLC-MS (LIPID +), demonstrating that the C18 factor had a diminishing effect on the response.

Finally, based on Sweet spot (**Figure 4-13.A**) and design space (**Figure 4-13.B**) plots, the optimum conditions with a low % uncertainty was estimated to be to be 14-18 mg/mL for the slurry concentration and 650 µL for slurry volume. From this point forward in the development, all samples subjected to DSPE, utilised the Sepra C_{18} sorbent, the 1:1 MeOH:MeCN solvent mixture, and a slurry concentration at 16 mg /mL at a slurry volume of 325 µL for any 100 µL sample.



MODDE 12 - 03/07/2020 16:12:45 (UTC+1)

Figure 4-13. Contour (sweet spot) and design space plot showing the interaction between both slurry volume and concentration factors. Both plots are used for diagnostics and interpretation and indicate combinations of the factors where all responses are within the target ranges (i.e. fulfil all criteria).

(A) Design space plot illustrates model stability by combing uncertainties from the specified factors. The areas in the lighter green, indicate combinations of the factors where the risk of failure to meet the criteria, is less than 5%. As the risk increases, indicated by the contour lines, the colour transitions from green to red;

(B) Contour plot highlights the areas where all criteria are met based on the user specified target ranges. The “sweet spot” is represented by the lightest green colour;

The optimised condition was therefore estimated to be 14-18 mg/mL for the slurry concentration and 650 µL for slurry volume.

4.4.1.4 *Protein presence and dilution of sample*

Current NPC blood profiling methodologies (polar metabolites acquired by HILIC-UPLC-MS and lipids acquired by LIPID-UPLC-MS), for sample preparation of the dilution series QC samples, requires a dilution of the blood product prior to extraction with solvent. It became apparent however that this is not applicable for the small molecule RPC based assays. As the QC sample increased in dilution, the extracts exhibited a higher baseline and large asymmetrical peaks in the TIC comprising of many *m/z* values eluting from 7min onwards (**Figure 4-14**). These are the same features that were detected from the MeOH and EtOH compositions in the solvent optimisation experiment (**section 4.4.1.2**)

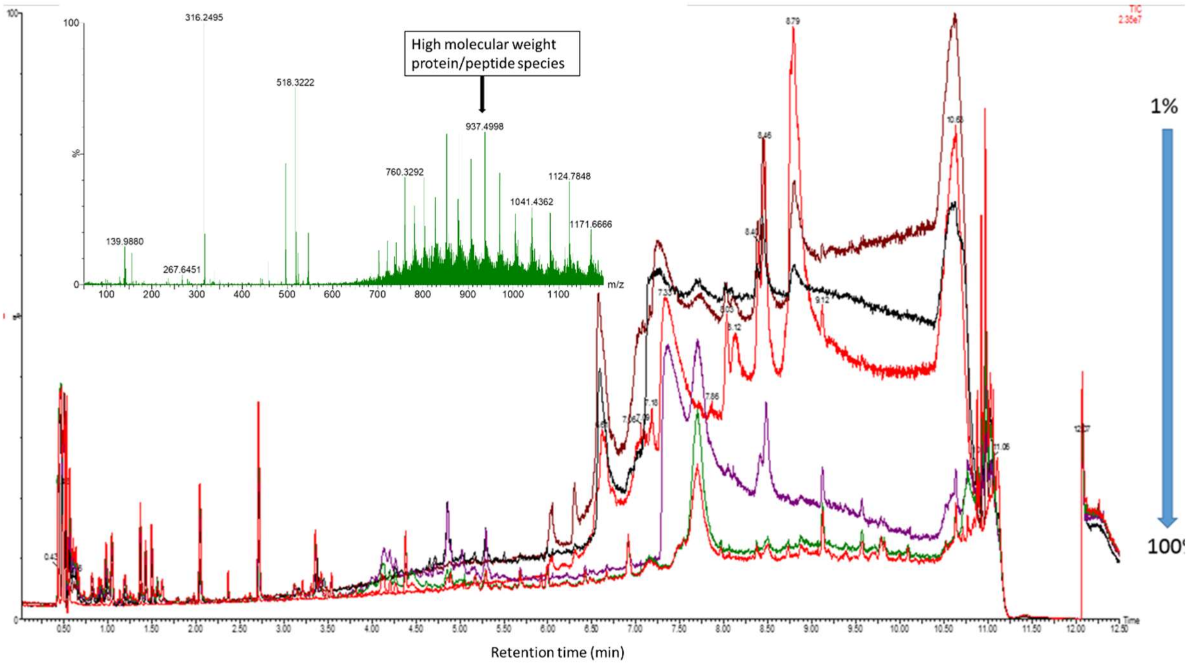


Figure 4-14. The TIC of dilution series plasma samples that were subjected to the DSPE protocol, where plasma was diluted post-DSPE treatment. As the sample concentration decreases (from 100% to 1%), the intensity of high molecular weight, multiply charged species increased. These species exhibited protein/peptide spectral characteristics.

The spectral features from these large peaks were multiply charged hinting that peptides or proteins remained in solution even after PP. To test this, the albumin content for samples dilute pre-extraction and post-extraction were measured and summarised in **Table 4-4.** As predicted, samples which were diluted pre-extraction exhibited large levels of albumin, exceeding the highest point on the calibration curve (> 3500 µg/mL). Samples diluted post-extraction, were significantly lower in albumin content. So, although the exact peptides have not been properly identified, these features are related to albumin still in solution.

A possible reason for this is that proteins contain both hydrophobic/hydrophilic parts in its structure, and therefore in solution may generally exhibit an overall net charge which can be either positive or negative. The presence of water molecules can cause an interaction with protein therefore resulting in protein molecules to dissociate itself from others and remain soluble. This solubility depends on the chemical environment and is a measure of the dielectric constant (Frigerio and Hettinger, 1962). If protein molecules are soluble, this implies a large dielectric constant. The dielectric constant of aqueous solutions can be lowered by adding water-soluble organic solvents, such as methanol and acetonitrile. The combination of water and alcohols, such as ethanol and methanol, was not sufficient in reducing the dielectric constant, therefore resulting in proteins to remain in solution as demonstrated in this experiment. The decision was therefore made to dilute the samples post extraction for the dilution series samples to be diluted post-extraction.

Table 4-4. A summary of the concentration of albumin in plasma samples diluted pre and post DSPE. Samples where the plasma sample was diluted pre-DSPE exhibited significant levels of detectable albumin (>3500 µg/mL), thereby demonstrating that plasma samples diluted pre-DSPE is unsuitable for this assay

	Concentration (µg/mL)	Replicate 1	Replicate 2	Replicate 3	Mean
Calibration	2000	2158	2165	2164	2162
	1500	1616	1573	1623	1604
	1200	1302	1260	1254	1272
	1000	967	760	1105	944
	600	584	570	596	583
	400	389	384	382	385
	200	215	200	198	204
LOQ	150	177	170	170	172
Sample (no dilution)		525	500	485	
Sample (5X DIL -Before DSPE)		>3500	>3500	>3500	
Sample (5X DIL -After DSPE)		231	233	244	

4.4.2 *Validation*

4.4.2.1 Precision

The precision of the sorbent weight (in a 96 well format suitable for high throughput) and metabolic profiles were assessed. Each tube of the PCR rack was weighed before and after the addition of slurry, and the difference calculated (Appendix 2). An RSD of less than 10% sorbent weight was measured across all wells of a 96-well plate. This not only confirms that a uniform amount of slurry is added to the sample, but also, larger amounts of the slurry can be made which can then be easily distributed to 96-well sample plates in an easy manner, therefore making the procedure applicable for high throughput profiling projects.

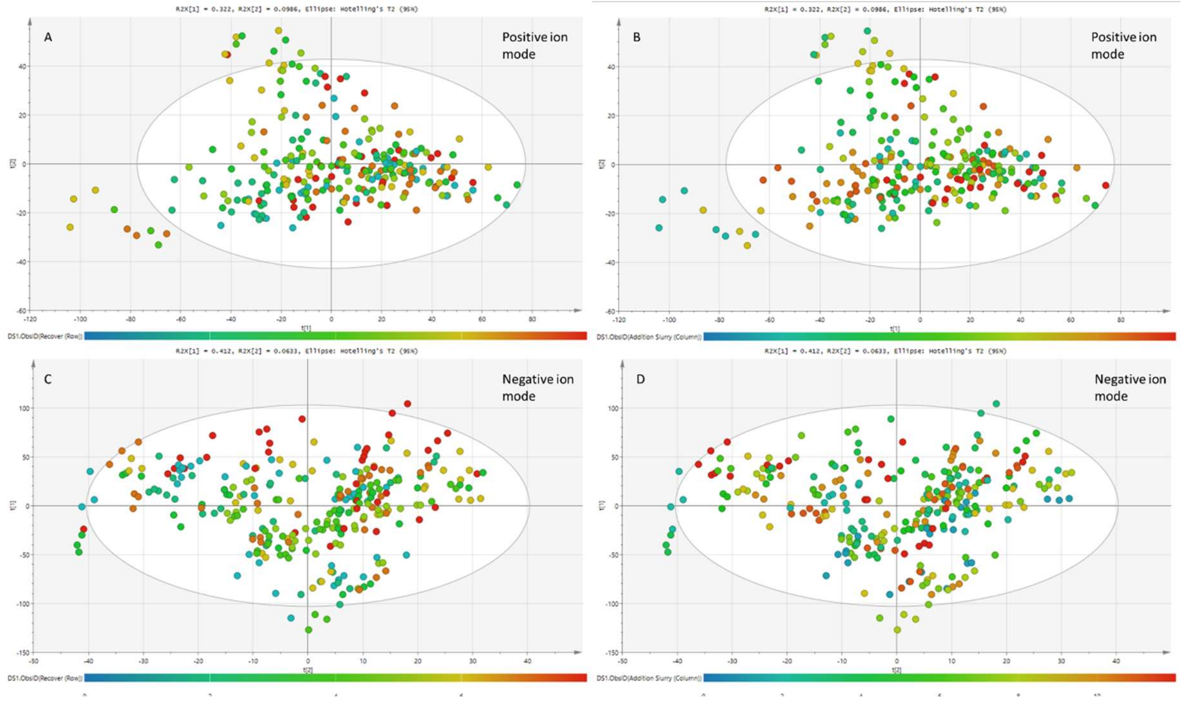


Figure 4-15. PCA score plot of 288 (three 96-well plate) samples prepared using the updated DSPE protocol to assess method reproducibility.

(A) PCA score plot of all sample observations coloured by order of slurry addition (column-wise using an 8-channel multi-pipette) from RPC-UPLC-MS (positive ion mode). No trend observed;

(B) PCA score plot of all sample observations coloured by order of water resuspension (row-wise, using a 12-channel multi-pipette) from RPC-UPLC-MS (positive ion mode). No trend observed;

(C) PCA score plot of all sample observations coloured by order of slurry addition (column-wise using an 8-channel multi-pipette) from RPC-UPLC-MS (negative ion mode). No trend observed;

(D) PCA score plot of all sample observations coloured by order of water resuspension (row-wise, using a 12-channel multi-pipette) from RPC-UPLC-MS (negative ion mode). No trend observed;

Overall, no trends observed, indicating high method reproducibility.

For the second part of the experiment, plasma samples from all three plates were randomised and acquired by RPC-UPLC-MS (positive and negative ion mode). PCA scores plots demonstrated no outlying samples and no trend in variation between the samples when coloured by addition of slurry (**Figure 4-15.A** and **Figure 4-15.C**) or addition of the water during resuspension (**Figure 4-15.B** and **Figure 4-15.D**). From the data acquired in both modes, the majority of features had an %RSD< 30 (**Figure 4-16**). In RPC-UPLC-MS (positive ion mode), a total of 3055 features were detected. Of these features, only 22 features had an %RSD more than the 30% threshold. Similarly, in RPC-UPLC-MS (negative ion mode), a total of 4236 features were detected, with 106 over this threshold. The median %RSD measured was 4.5% in positive ion mode and 4.2% in negative ion mode. Chromatographic performance was also assessed with no observable pressure spikes throughout the analytical run, low background, negligible drift in RT and a stable baseline. Overall, no significant global variation was observed across all detected chemical features. Finally, as the DSPE method described is subject to variances in the precision and accuracy of measurements, evaluation of data quality was accounted for by careful deliberation in various steps during the sample preparation. One such step is during the recovery stage of the sample preparation, although manual recovery of the supernatant is possible, the risk of disturbing the pellet after centrifugation is high and was mitigated using a multichannel automated fluid pipetting robot.

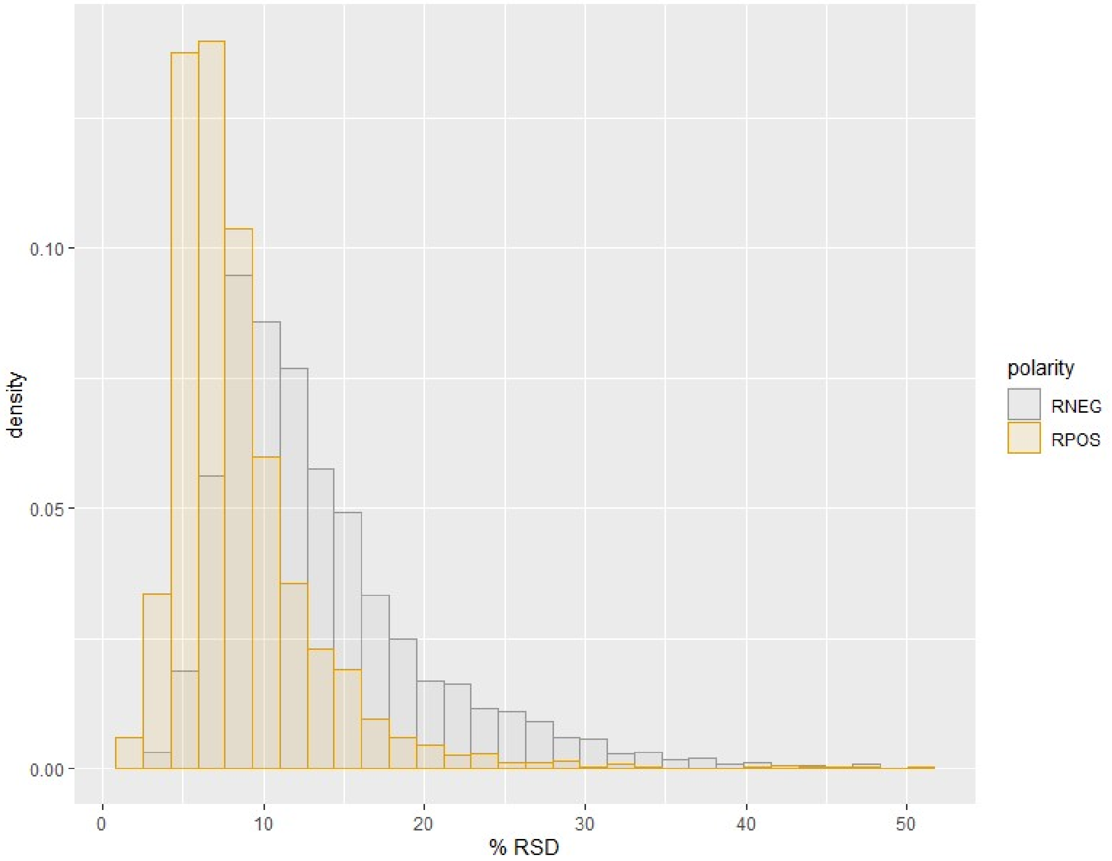


Figure 4-16. The relative standard deviation (RSD) for feature groups passing the dilution series and blank filtering from all samples acquired in both RPC-UPLC-MS positive ion mode (yellow) and RPC-UPLC-MS negative ion mode (grey). The median %RSD measured was 4.5% and 4.2%, respectively.

4.4.2.2 Performance of lipid removal on xenobiotics and endogenous small molecules

A mixture of prevalent xenobiotics was spiked into urine at varying concentrations and subjected to DSPE. A decision was made to spike into urine a variety of xenobiotics, which are known to elute throughout the chromatographic range, with an emphasis on lipophilic xenobiotics (lansoprazole, amitriptyline ,terbinafine and diclofenac), as these would have the potential, like lipids, to be removed by the DSPE procedure. The spiking of LIPID mixes at different concentrations, added a further layer to the experiment to explore whether varying levels of lipids can have a suppressive effect on the intensity of the spiked xenobiotics. The lipids selected for the lipid mixture are based on lipid classes which are prevalent in normal human blood and commonly detected by the LIPID-UPLC-MS profiling method. The approximate final concentration of the LIPID mix (held at three concentrations; 985 ng/mL, 1970 ng/mL, and 3941 ng/mL) in the urine extracts were designed to span average lipid levels observed in human plasma (Bowden. J *et al.* 2017). Finally, albumin was spiked into urine at a concentration consistent to reported levels found in circulating plasma of healthy individuals (Varacallo, 2020). Although the composition between urine and blood products are vastly different, the aim here was to try and create an artificial mixture resembling blood. This was undertaken by spiking lipids and albumin into a urine matrix. Urine profiling by RPC-UPLC-MS is extensively studied at the NPC and exhibits a greater number of features in comparison to blood products, therefore a more global evaluation can be undertaken if such a significant effect between SHAM and DSPE treatment were to be observed.

4.4.2.2.1 Recovery of xenobiotics

Differences between DSPE and SHAM samples were firstly examined by targeting the spiked xenobiotics from the xenobiotic mixtures and comparing signal intensities. Xenobiotics were spiked into urine at four different concentrations. Within each concentration exist six different preparations, i.e., three different concentrations of lipids each with a SHAM and DSPE treatment. The results are illustrated as box plots (**Figure 4-17**), and an ANOVA test was used to determine if any statistically significant differences exist between the mean intensities observed in the DSPE treated samples, and the mean intensity measure from the SHAM treated samples within a specific concentration of the xenobiotic mixture. For some xenobiotic’s, there were p<0.05 within a XENO group (highlighted in red in **Table 4-5**), however no real pattern was observed to explain these p-value results. Ibuprofen demonstrated significant difference in intensity between SHAM and DSPE treated samples, at different concentrations (XENO1-4). This agrees with the observations made during the DOE analysis **(section 4.4.1.3**), as this compound elutes after 10 minutes and measurement of features eluting after this time may be unstable. Diclofenac also elutes after 10 minutes, but apart from its highest concentration (XENO1), no statistically significant differences were observed between groups at lower concentrations. Overall, regardless of the concentration of the lipid mixture, SHAM and DSPE treated samples demonstrated little significant difference in mean signal intensities for this exemplar set of xenobiotics.

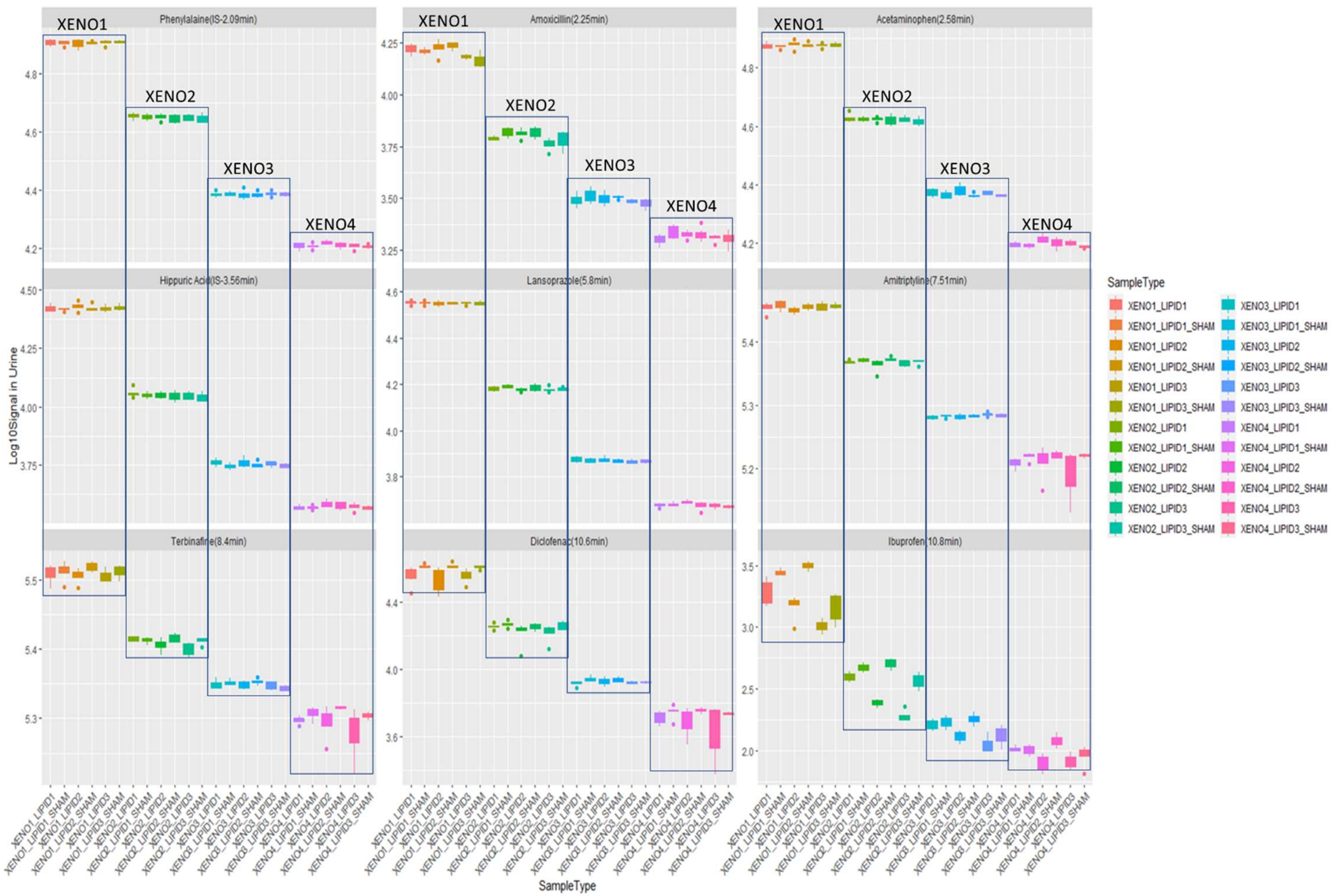


Figure 4-17. Box plots depicting the distribution of signal intesity from an exemplar set of xenobiotics and internal standards, between DSPE and SHAM treated samples. Overalll (within each XENO concentration level), regardless of the concentration of the lipid mixture (LIPID1-3), SHAM and DSPE treated samples demonstrated little statistically significant difference ($p < 0.05$ Kruskal-Wallis ANOVA) in mean signal intensities of these xenobiotics.

Table 4-5. p-values from Kruskal-Wallis ANOVA comparing the mean intensities of an exemplar set of xenobiotics between DSPE and SHAM treated samples.

Xenobiotic/IS	ANOVA p-value			
	XENO1	XENO2	XENO3	XENO4
L-Phenylalanine-$^{13}C_9$,^{15}N	0.93	0.75	0.99	0.12
Amoxicillin	0.01	0.01	0.14	0.21
Acetaminophen	0.70	0.83	0.06	0.01
N-Benzoyl-D_5-Glycine	0.66	0.67	0.16	0.12
Lansoprazole	0.95	0.20	0.33	0.01
Amitriptyline	0.24	0.10	0.11	0.11
Terbinafine	0.23	0.01	0.06	0.06
Diclofenac	0.01	0.12	0.06	0.12
Ibuprofen	<0.001	<0.001	<0.001	<0.001

4.4.2.2.2 Global recovery

A secondary analysis was subsequently undertaken to explore any significant differences from a more global interrogation of the data. The distribution of recovery values was calculated between the DSPE and SHAM treated samples acquired by RPC-UPLC-MS, in positive ion mode (**Figure 4-18.A-C**) and negative ion mode (**Figure 4-18.D**-F) at the three different lipid concentrations. For clarity the x-axis has been truncated to between 50% and 200% recovery, and features were further divided into retention time bins of 2minutes and coloured accordingly. The median recovery, as indicated by the purple dotted line in the figures (**A-F**), were approximately 102-104% in both positive and negative ion mode. The results of this experiment demonstrate that the DSPE treatment has had minimal global effect on small molecule recovery.

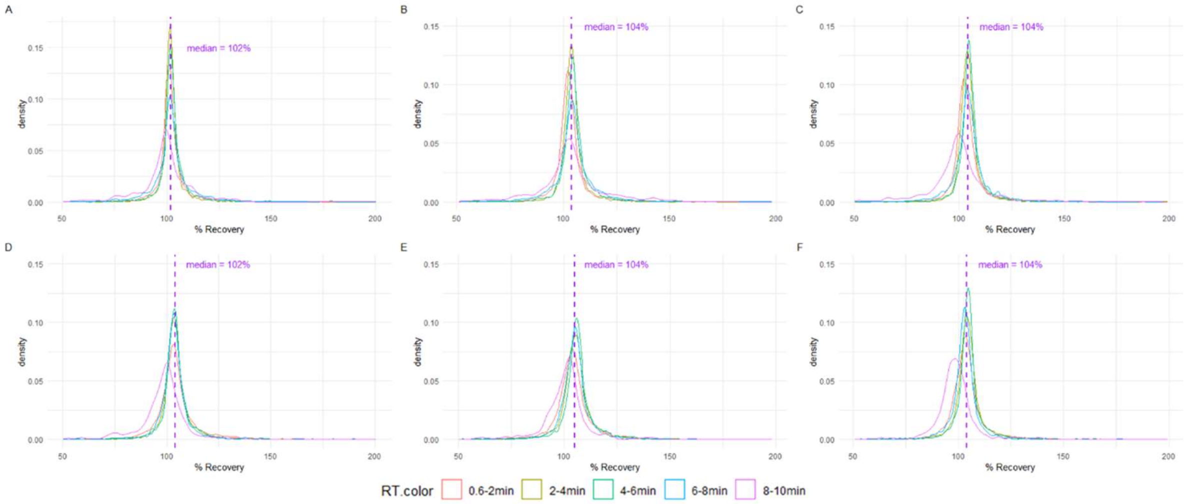


Figure 4-18. Recovery density plots for feature groups passing the dilution series and blank filtering, calculated between DSPE and SHAM treated samples analysed by both RPC-UPLC-MS (positive and negative ion mode). Feature groups are segregated by 2-minute retention time bins; 0.2min (pink), 2-4min (yellow), 4-6min (green), 6-8min (blue) and 8-10min (purple).

(A) Density plot measuring the recovery (between SHAM and DSPE treated samples) of filtered features spiked at the first lipid concentration of lipids (LIPID1) in positive ion mode. Median recovery calculated at 102% (vertical purple dotted line);

(B) Density plot measuring the recovery (between SHAM and DSPE treated samples) of filtered features spiked at the second lipid concentration of lipids (LIPID2) in positive ion mode. Median recovery calculated at 104% (vertical purple dotted line);

(C) Density plot measuring the recovery (between SHAM and DSPE treated samples) of filtered features spiked at the third lipid concentration of lipids (LIPID3) in positive ion mode. Median recovery calculated at 104% (vertical purple dotted line);

(D) Density plot measuring the recovery (between SHAM and DSPE treated samples) of filtered features spiked at the first lipid concentration of lipids (LIPID1) in negative ion mode. Median recovery calculated at 102% (vertical purple dotted line);

(E) Density plot measuring the recovery (between SHAM and DSPE treated samples) of filtered features spiked at the second lipid concentration of lipids (LIPID2) in negative ion mode. Median recovery calculated at 104% (vertical purple dotted line);

(F) Density plot measuring the recovery (between SHAM and DSPE treated samples) of filtered features spiked at the second lipid concentration of lipids (LIPID3) in negative ion mode. Median recovery calculated at 104% (vertical purple dotted line).

4.4.2.3 Liquid-Liquid extraction (LLE) comparison

4.4.2.3.1 Residual protein

Within 20 injections of the hydrophilic fraction acquisitions *via* the NPC RPC-UPLC-MS profiling method assay, the instrument stopped due to the pressure exceeding its maximum limit (15000 psi). Viewing the spectra from these injections indicated evidence of protein that accumulated on the column resulting in a blockage that caused the overpressure stoppage. As a result, all extracts were then subjected to an additional protein precipitation step using cold acetonitrile in the 1:3 proportion sample:acetonitrile. During sample preparation for both the Folch and BD extractions, the protein precipitate forms in between the organic and aqueous phases. It was evident from these extractions, that precipitate also formed on the walls of the glass tubes and its aspiration with the upper hydrophilic aqueous phase was therefore inevitable. In the Matyash extraction, the precipitate formed at the bottom of the tube. Visually, there seemed to be no evidence that precipitate was present on the walls of tube, however the spectra did show evidence of protein, possibly be due to carry-over from the Folch and BD extractions.

From the literature, lipid analysis using LLE has never indicated the presence of protein. Analysis of the organic extracts *via* the NPC LIPID-UPLC-MS profiling assay (Izzi-Engbeaya et al., 2018) requires resuspension of the dried organic phase in a method appropriate solution of 1:4 water:IPA. Apart from the selectivity of this mixture to extract lipids, this step also ensures protein precipitation. The organic fractions were subsequently analysed and showed no evidence of protein in the spectra. The presence of protein in the hydrophilic fractions when acquired by RPC methodologies was a novel finding. LLE of blood products is primarily for lipid analysis, and the analysis of the aqueous phase in most cases is analysed (if analysed at all) by HILIC methodologies (Fei et al., 2014, Lange and Fedorova, 2020, Fauland et al., 2011, Schwaiger et al., 2018). Both the organic and aqueous phase will have some level of protein precipitation associated with it in its preparation. The analysis of the aqueous phases by RPC-UPLC-MS evidently can not be done so easily as observed with protein on the walls of the sample extraction tubes and its inevitable contamination during aspiration.

4.4.2.3.2 Unsupervised multivariate comparison of the extraction systems

All extracts and the pooled sample were analysed by the RPC-UPLC-MS profiling method. The total number of features for each extraction in positive ion mode had the following order: DSPE > Pool > BD > MeOH > Folch > Matyash (2655, 2535 ,2362, 2230, 2103, 2086). In negative ion mode, the order was: DSPE > BD > Pool > MeOH > Matyash > Folch (3681, 3481, 3290 ,3235, 2964, 2921). In both modes, the DSPE RPC data demonstrated the highest number of high-quality features, and the highest number of common features, when compared to the pooled extract, as illustrated by the Venn diagrams (**Figure 4-19**). The pooled sample will be more diluted when compared to the other extracts, as it is a mixture of all the samples, which could explain why a smaller number of features were detected.

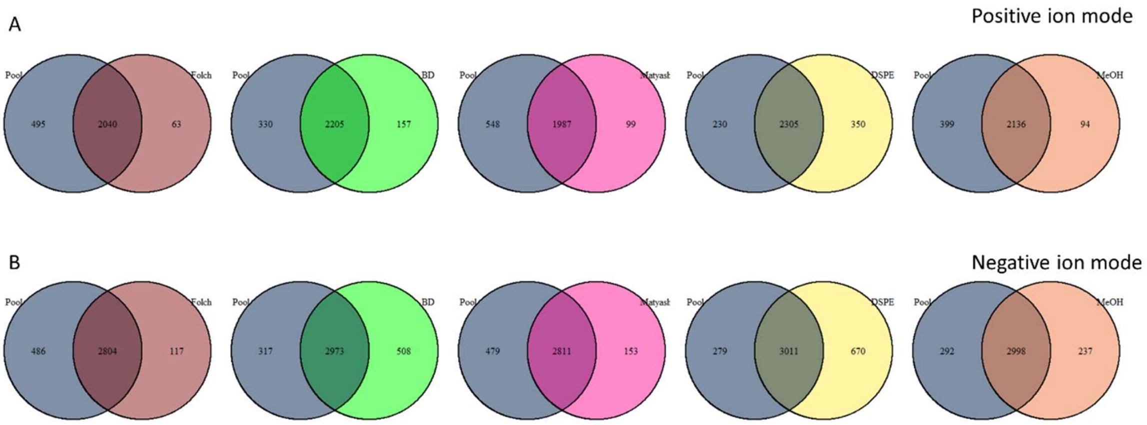


Figure 4-19. Venn diagrams representing the number of common features between the pool samples (grey circle) , to the different extraction methods analysed by the RPC-UPLC-MS method. These are Folch (green), Mataysh (pink), DSPE (yellow) and MeOH (orange).

(A) are Pool samples compared to the Folch, BD, Matyash, DSPE and MeOH in positive ion mode;

(B) are Pool samples compared to the Folch, BD, Matyash, DSPE and MeOH in negative ion mode.The DSPE treated samples exhibited the highest number of features when compared to the other extraction protocols, and shared the highest number of common features to the pooled samples.

Using only the features that passed all filtering criteria (Blank filtering, RSD<30% and correlation to dilution of 0.7) for the pooled extract, the RPC-UPLC-MS data from all extractions were exported to SIMCA software for multivariate analysis. Grouping, trends and outliers were examined from PCA scores plots shown in **Figure 4-20.A** and **Figure 4-20.C**. In both ion modes, a PCA score plot illustrated little method variability associated with all extraction methods, as the samples were relatively clustered together. However, the DSPE, MeOH and pool samples clustered more tightly than the LLE samples. One Matyash replicate clustered with the Folch extraction, highlighting a possible contamination. This sample was then removed from further analysis. A dendrogram from HCA, was also plotted (**Figure 4-20.B** and **Figure 4-20.D**) to visually and quantitively demonstrate the similarity between the RPC profiles from each extraction (where the height of the link that joins two groups are the smallest). The profiles of the Matyash method in both ionisation modes were the most similar to the pool sample profiles as it is also observed from the PCA score plots. In positive ion mode, the order of proximity is as follows: Pool > Mataysh > Folch > MeOH > BD > DSPE, while in negative ion mode, the order is Pool > Matyash > Folch > BD > DSPE > MeOH. The DSPE method seems to differ significantly in the profile from the pool, even though the DSPE and pool samples share the most common features (Venn diagrams above). The Venn diagrams produced in this instance relate to only the number of features and whether it is present in each extraction. Both PCA and HCA can illustrate the similarity of general profiles, taking into account the relative abundances of the features.

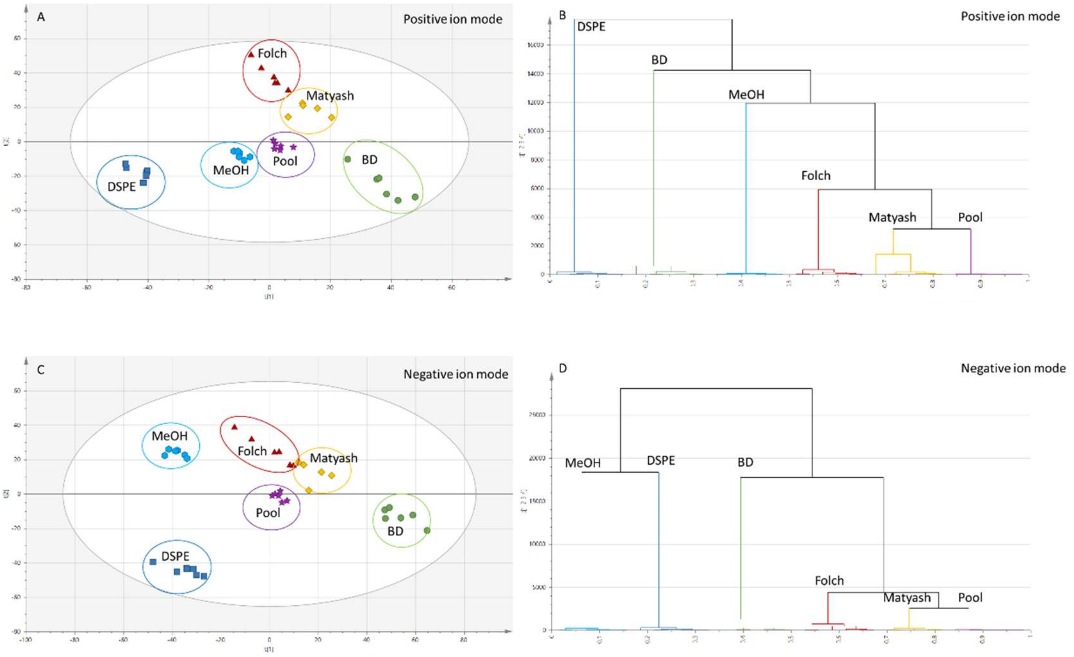


Figure 4-20. PCA score plots and HCA dendrograms depicting the relationships of the different extraction methods of the hydrophilic fractions acquired by RPC-UPLC-MS, to one another. The PCA score plot will depict how the different extraction methods relate to each other by identifying clusters, whilst also highlighting outliers and time-based patterns. The dendrogram quantitatively highlights similar groups based on the height at which any two objects are joined together. The most similar group have the smallest link that joins them are the smallest.

(A) PCA score plot of the Pool, Folch, BD, Matyash, DSPE and MeOH profiles in positive ion mode;

(B) HCA dendrogram comparing Pool, Folch, BD, Matyash, DSPE and MeOH profiles in positive ion mode;

(C) PCA score plot of the Pool, Folch, BD, Matyash, DSPE and MeOH profiles in negative ion mode;

(D) HCA dendrogram comparing Pool, Folch, BD, Matyash, DSPE and MeOH profiles in negative ion mode.

4.4.2.3.3 Supervised multivariate comparison of the extraction systems

Even though the RPC-UPLC-MS profiles obtained for Matyash extracts proved to be more similar to the pool sample, the PCA did demonstrate different clustering groups, thereby suggesting different profiles. A series of OPLS-DA models were therefore produced between the pool and all other extraction methods to highlight the ratio of the number of discriminant features to common features. Discriminant features were based on VIP values greater than 1.5. S plots produced from the OPLS-DA models were used to visualise the features that were driving the separation between two extraction groups in positive ion mode (**Figure 4-21**) and negative ion mode (**Figure 4-22**). The results show that for all OPLS-DA models in both positive and negative ion modes, the R2 (> 0.995) and Q2 (> 0.945) values of the original models were above the permutated models, indicating low variability and a high predictive ability. From these models, in both ionisation modes, the DSPE method has the highest ratio of discriminant to common features, therefore indicating that a higher number of features are being affected by this method. As to be expected, the number a lower ratio of discriminant to common features was observed for the Mataysh samples, as indicated by their close proximity to the pool in the PCA and HCA.

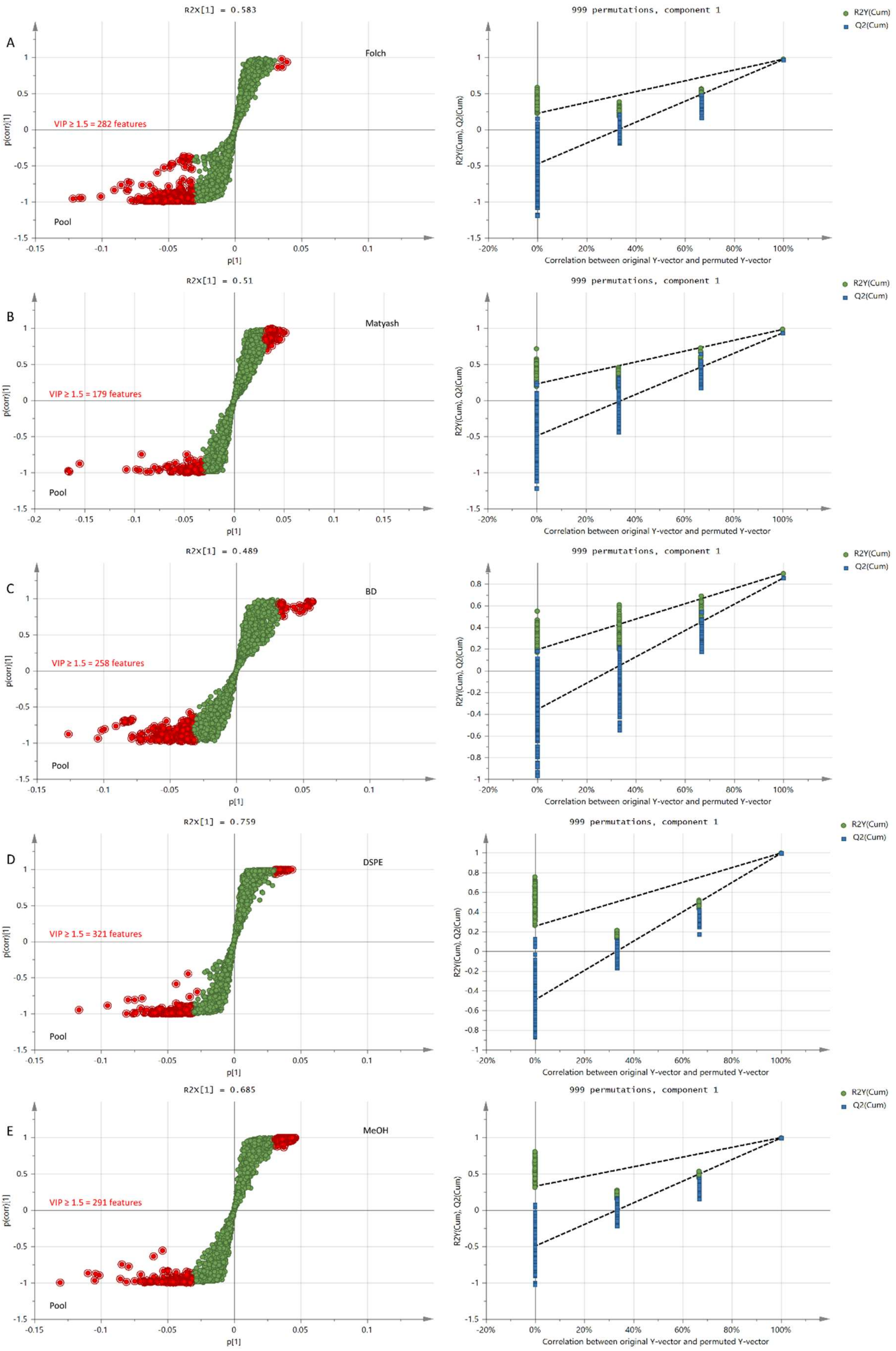

A
R2X[1] = 0.583
Folch
VIP ≥ 1.5 = 282 features
Pool
p[corr][1]
p[1]
999 permutations, component 1
R2Y(Cum)
Q2(Cum)
R2Y(Cum), Q2(Cum)
Correlation between original Y-vector and permuted Y-vector
B
R2X[1] = 0.51
Matyash
VIP ≥ 1.5 = 179 features
Pool
p[corr][1]
p[1]
999 permutations, component 1
R2Y(Cum)
Q2(Cum)
R2Y(Cum), Q2(Cum)
Correlation between original Y-vector and permuted Y-vector
C
R2X[1] = 0.489
BD
VIP ≥ 1.5 = 258 features
Pool
p[corr][1]
p[1]
999 permutations, component 1
R2Y(Cum)
Q2(Cum)
R2Y(Cum), Q2(Cum)
Correlation between original Y-vector and permuted Y-vector
D
R2X[1] = 0.759
DSPE
VIP ≥ 1.5 = 321 features
Pool
p[corr][1]
p[1]
999 permutations, component 1
R2Y(Cum)
Q2(Cum)
R2Y(Cum), Q2(Cum)
Correlation between original Y-vector and permuted Y-vector
E
R2X[1] = 0.685
MeOH
VIP ≥ 1.5 = 291 features
Pool
p[corr][1]
p[1]
999 permutations, component 1
R2Y(Cum)
Q2(Cum)
R2Y(Cum), Q2(Cum)
Correlation between original Y-vector and permuted Y-vector
-20%
0%
20%
40%
60%
80%
100%

Figure 4-21. One-component OPLS-DA S-Plots showing the separation and discriminating features (respectively) between the RPC-UPLC-MS profiles of the Pool samples (with loadings coloured in red exhibiting a VIP > 1.5), and all other extraction methods in positive ion mode. Validation plots displaying 999 permutation tests are alongside the corresponding OPLS-DA model.

(A) The explained variance (R2Y) was 0.972 and predictive ability was 0.968 for “Pool vs Folch”;

(B) The explained variance (R2Y) was 0.979 and predictive ability was 0.930 for “Pool vs BD”;

(C) The explained variance (R2Y) was 0.898 and predictive ability was 0.854 for “Pool vs Matyash”;

(D) The explained variance (R2Y) was 0.997 and predictive ability was 0.994 for “Pool vs DSPE”;

(E) The explained variance (R2Y) was 0.995 and predictive ability was 0.990 for “Pool vs MeOH”.

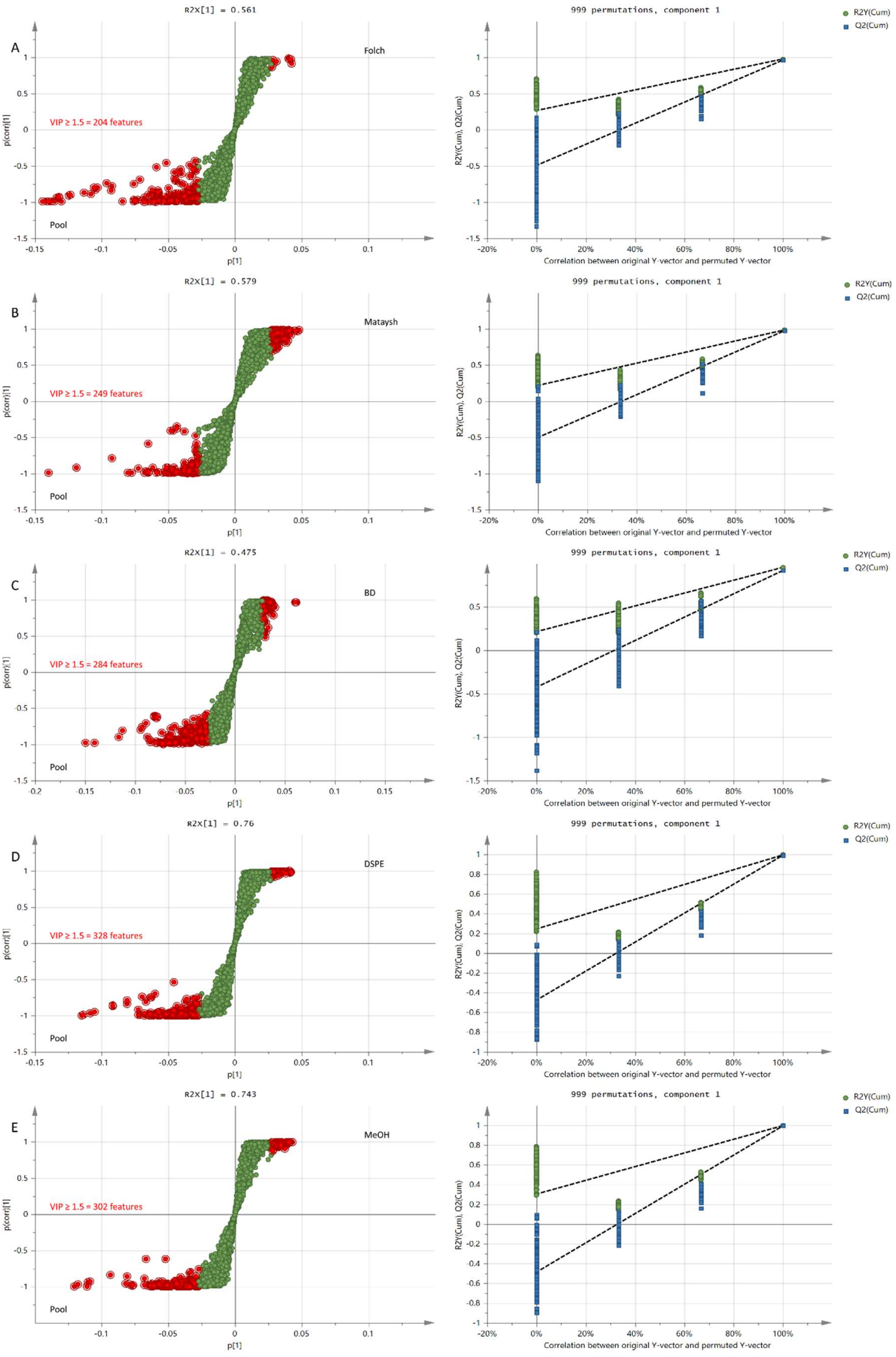

A
R2X[1] = 0.561
Folch
VIP ≥ 1.5 = 204 features
Pool
p(corr)[1]
p[1]
999 permutations, component 1
R2Y(Cum)
Q2(Cum)
R2Y(Cum), Q2(Cum)
Correlation between original Y-vector and permuted Y-vector
B
R2X[1] = 0.579
Mataysh
VIP ≥ 1.5 = 249 features
Pool
C
R2X[1] = 0.475
BD
VIP ≥ 1.5 = 284 features
Pool
D
R2X[1] = 0.76
DSPE
VIP ≥ 1.5 = 328 features
Pool
E
R2X[1] = 0.743
MeOH
VIP ≥ 1.5 = 302 features
Pool

Figure 4-22. One-component OPLS-DA S-Plots showing the separation and discriminating features (respectively) between the RPC-UPLC-MS profiles of the Pool samples (with loadings coloured in red exhibiting a VIP > 1.5), and all other extraction methods in negative ion mode. Validation plots displaying 999 permutation tests are alongside the corresponding OPLS-DA model.

(A) The explained variance (R^2Y) was 0.977 and predictive ability was 0.962 for “Pool vs Folch”;

(B) The explained variance (R^2Y) was 0.983 and predictive ability was 0.972 for “Pool vs BD”;

(C) The explained variance (R^2Y) was 0.955 and predictive ability was 0.20 for “Pool vs Matyash”;

(D) The explained variance (R^2Y) was 0.995 and predictive ability was 0.992 for “Pool vs DSPE”;

(E) The explained variance (R^2Y) was 0.996 and predictive ability was 0.994 for “Pool vs MeOH”.

4.4.2.3.4 Method-induced losses of annotated small molecules

The identification of all unknown discriminant features was not feasible due to their high number. However, relative method-induced losses for RPC-UPLC-MS metabolites annotated by peakpantheR, were evaluated for each extraction. This method was implemented, as the conventional way to calculate recoveries could not be conducted due to improper use of internal standards (i.e. internal standard were spiked pre-extraction but a number of sample were not subjected to a post-extraction spike of the same standards). A total of 105 metabolites were annotated in both ionisation modes for RPC-UPLC-MS data from the different extraction methods (Appendix 2). The method-induced losses are also listed in Appendix 2 and illustrated in the jitter plots (**Figure 4-23**). Overall, all extraction procedures demonstrated a similar performance with average metabolite losses less than 10% for both ionisation modes (as indicated by the red horizontal line in **Figure 4-23**). There were no statistically significant average metabolite losses observed in negative mode, however in positive mode, only the DSPE samples demonstrated statistically significant differences to all other extraction methods (apart from BD). Both BD and DSPE yielded the highest levels for pre-annotated metabolites measured by peakPantheR, with the lowest average loss. The DSPE method in particular, gave the highest levels for xenobiotics (Cortisone, Caffeine, Theobromine, Theophylline, Cotinine and Prednisolone) and carnitine metabolites (L-acetylcarnitine, Carnitine, 2-Octenoylcarnintine and Propionylcarntinine).

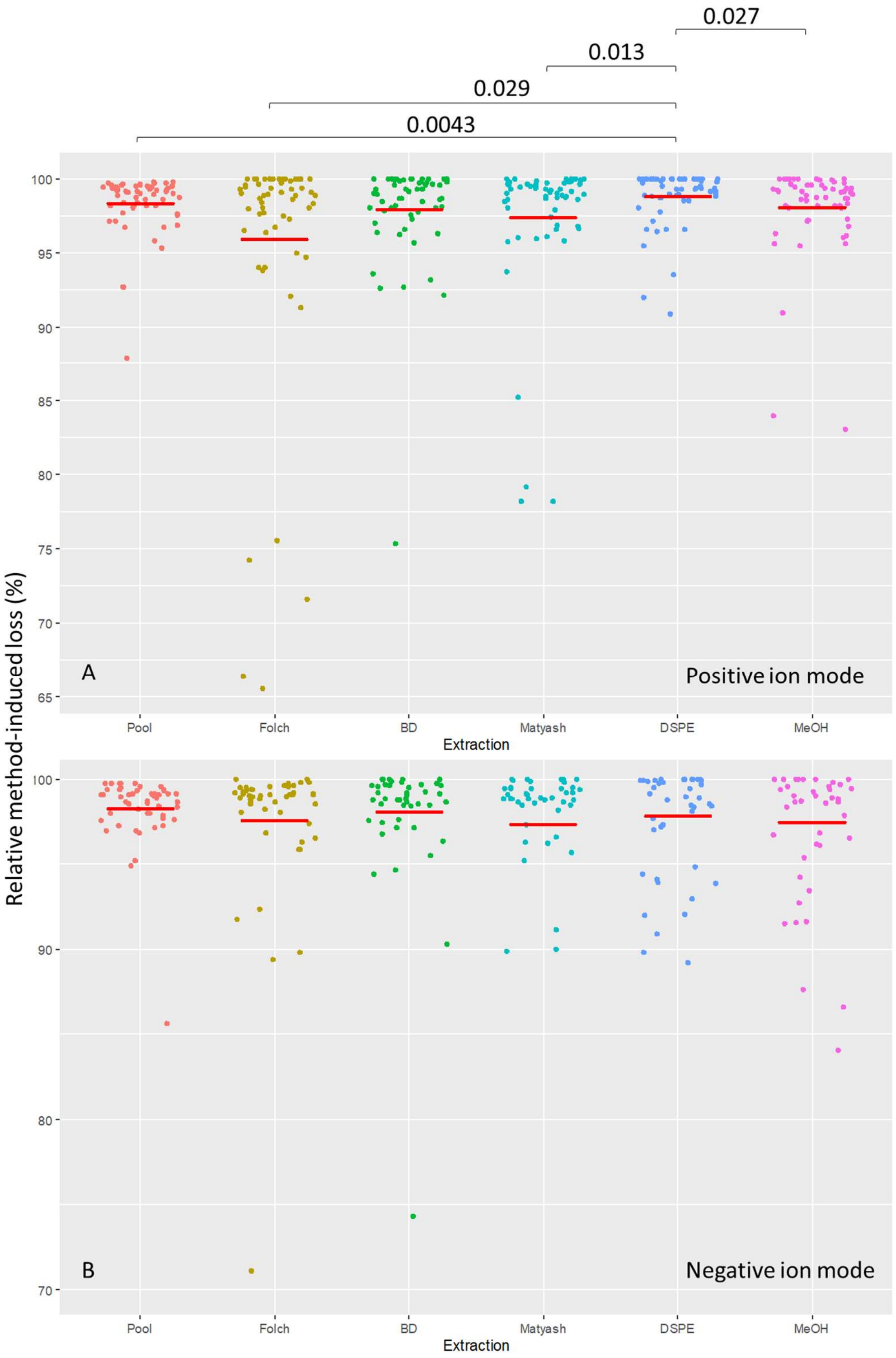


Figure 4-23. Method-induced losses for metabolites identified from RPC-UPLC-MS peakPantheR on all extraction approaches. (Pool - red, Folch - yellow, BD – green, Matyash – aqua, DSPE – blue, MeOH – purple**).** The average losses were calculated for each method, as indicated by the red horizontal line. Statistically significant differences, p < 0.05, Newman-Keuls multiple comparison test, were performed on the absolute average levels.

(A) In positive ion mode, statistically significant differences were observed between DSPE and all other extraction methods apart from BD. Both BD and DSPE yielded the highest average level of targeted metabolites. The DSPE method in particular yielded highest levels for carnitine and related metabolites and xenobiotics.

(B) In negative ion mode, no statistically significant differences were observed between the different extraction groups

4.4.2.3.5 Carnitine presence from the different extraction systems

The high yield associated with carnitine metabolites in the DSPE extracts corroborates the work in **section 4.4.1.2**. As stated in this section, the DSPE protocol is not a method that enables complete lipid removal, and a compromise was made that allowed metabolites within region one of the lipid profile, to remain, i.e. polar lipids such as LPC, acylcarnitines and lipophilic endogenous and exogenous metabolites. To explore this further, when the samples were treated with acetonitrile to remove the residual protein (1:3 sample:acetonitrile), an aliquot was taken and run *via* the NPC HILIC-UPLC-MS assay which provides a wider range of detected and pre-annotated acylcarnitine species. A total of 76 metabolites were annotated and measured using peakPantheR (list provided in Appendix 2). A PCA biplot was used to summarise the findings. The biplot illustrates the features/metabolites which associate best to a particular extraction method. It models similarities and dissimilarities between sample clusters and the relationship of metabolites to the clusters simultaneously. As highlighted in **Figure 4-24,** most of the extracted carnitines present, as well as lipophilic xenobiotics such as warfarin, are strongly associated to the DSPE method (blue circles). This also explains the higher number of features observed in the Venn diagrams, indicating that the DSPE method covers a wider range of metabolites amendable to RPC small molecule analysis when compared to the other LLE methods.

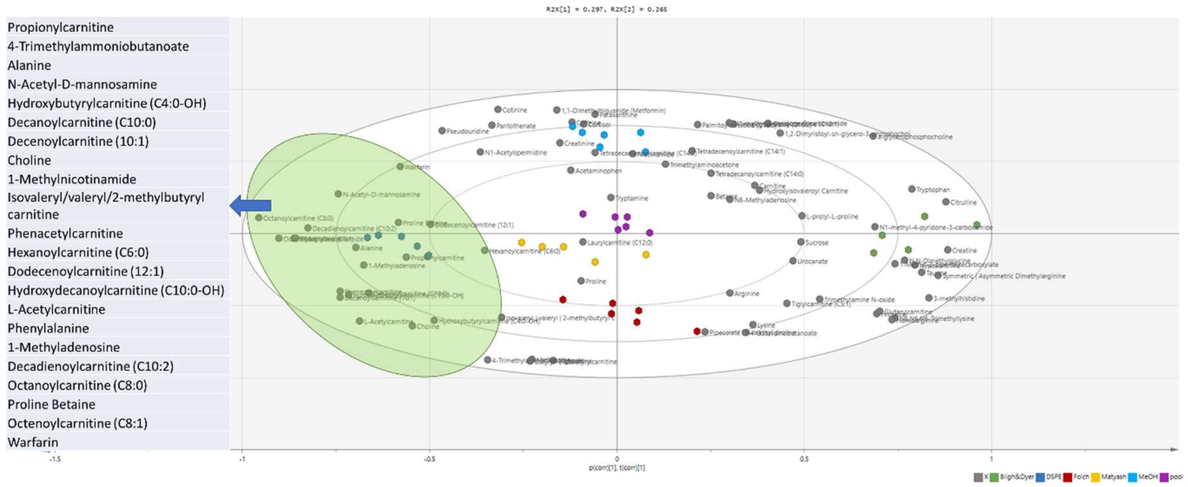


Figure 4-24. A PCA biplot comparing the levels of targeted metabolites measured using peakpantheR in hydrophilic fractions of samples from the different extractions acquired by HILIC-UPLC-MS (positive ion mode). The grey circles represent the targeted metabolites (loadings), and all other coloured cirlces represent the different extrcation methods; BD (green), DSPE (dark blue), Folch (maroon), Matyash (yellow), MeOH (light blue) and pool (purple). The biplot demonstrates a strong association for acylcarnitine affiliated metabolites to the DSPE method as indicated by the light green eclipse and the accompanying table of targeted metabolites. The Matyash and Pool samples are closest to the origin and therefore do not significantly contribute to the formation of the observed clusters as described by PC1 and PC2.

4.4.2.3.6 Selectivity of the LLE methods for different lipid species

Gil *et al*. found that metabolites associated with Region one (specifically LPC) in the Matyash and BD methods, had a lower overall signal in the organic fractions when compared to the Folch method. The recovery of these LPC affiliated metabolites were not lost during the sample reparation procedure but rather were detected in higher quantities than the Folch method in the hydrophilic/aqueous fractions. The organic extracts from all LLE extractions obtained in this work were acquired by the NPC LIPID-UPLC-MS assay in positive mode only, to examine whether their findings were consistent. In hindsight, the DSPE sorbent should have been re-extracted with the IPA mixture, so that comparison of lipid profiles could be made, however the collection plate was discarded at the time of preparation. A total of 243 lipid species were annotated in the organic extracts using a LIPID panel of metabolites. Apart from lipid species, the panel also included a mixture of short, medium, and long chain acylcarnitine's, most of which elute in Region one of the LIPID assay. Repeatability of Folch and Matyash methods were similar, with approximately the same number of metabolites demonstrating a %RSD < 30 (208 and 206, respectively). The BD method was more variable, and only 113 metabolites had a %RSD < 30%. Method-induced losses and a PCA-biplot were again used to examine the results, and metabolites were divided into the three retention time regions**. Figure 4-25.B** shows that the Folch method produced the highest mean lipid yield, with the lowest losses associated with metabolites from region one. This can also be observed in the biplot, where the majority of these metabolites (as indicated by the green circles in the PCA-Biplot in **Figure 4-25.A**) is related to the Folch method. The biplot also demonstrates that the Matyash samples are closest to the origin and therefore do not significantly contribute to the formation of the observed clusters as described by PC1 and PC2. A clustering of TG's (red circles), and the majority of later eluting metabolites affiliated with region two (blue circles), are observed, relating to samples from the BD method. The BD extraction method resulted in the highest yields for metabolites associated with region's two and three (**Figure 4-25.C-D**), specifically, phospholipids, SM, CER, DG and TG. However, the BD method was the most variable, with only 113 metabolites with a CV < 30%, thus demonstrating a greater overall loss between the three methods. These results are consistent with Gil *et al*., who found that the Folch method was the best of the three LLE methods in terms of lipid coverage along with precision of lipid measurements. A summary of the annotated lipid metabolites and method induced losses, are summarised in Appendix 2.

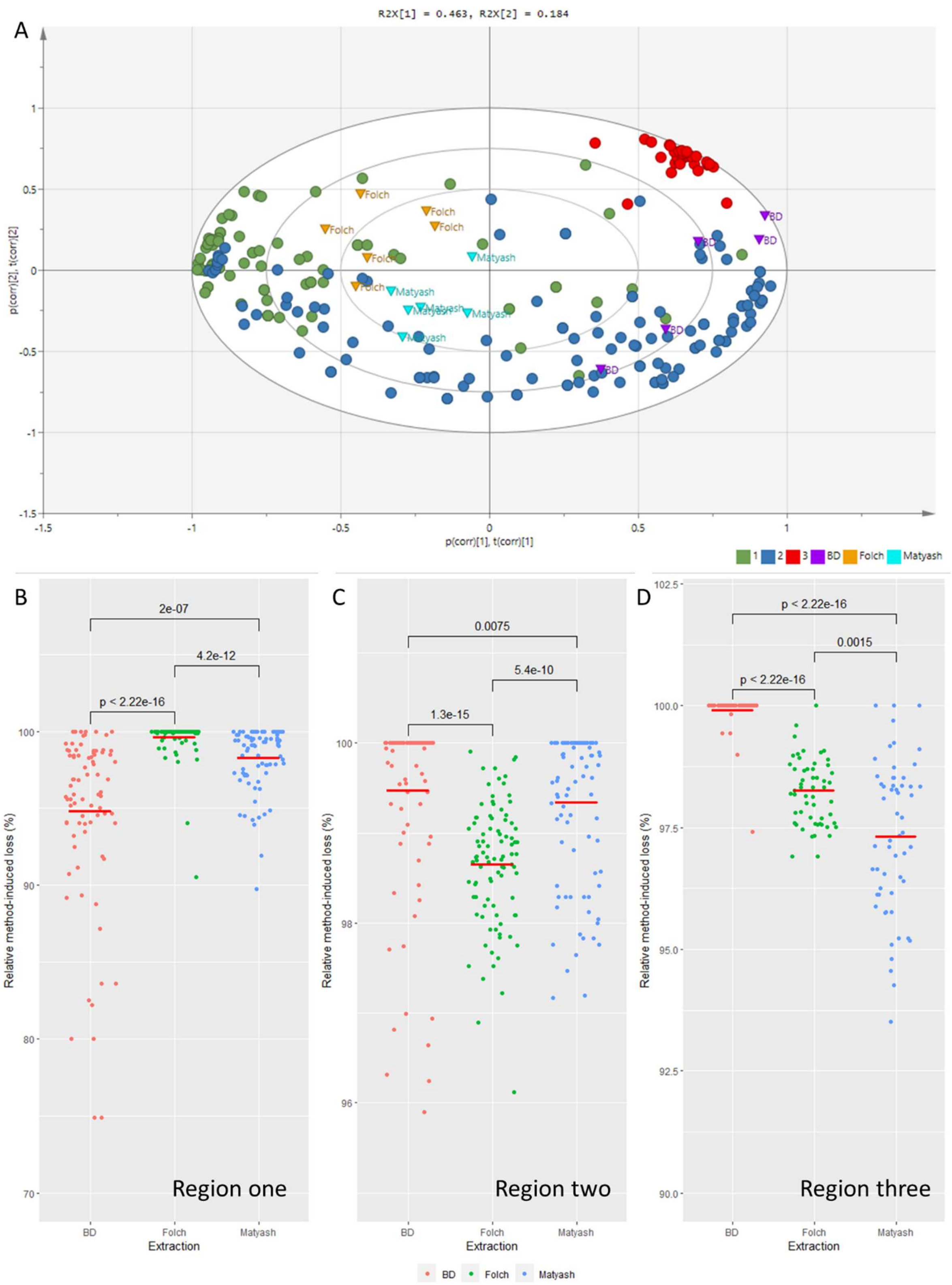


Figure 4-25. A PCA-Biplot (A) and method-induced jitter plots (B-D) comparing the organic fractions of the LLE methods (Folch, BD and Mataysh) acquired by the LIPID-UPLC-MS (positive ion mode). (A) PCA-Biplot of the different LLE methods (highlighted by the coloured triangles, Folch (orange), BD (purple) and Matyash (aqua)) overlayed with the metabolites annotated by LIPID (ESI positive) peakpantheR. Metabolites are colour coded by region: Region one (Blue circles), region two (green circles), region three (red circles). (B), (C) and (D)

are plots depicting the method induced losses observed between the three LLE methods for metabolites identified by LIPID peakpantheR. The coloured circles represent lipid metabolites extracted by Folch (red), BD (green) and Matyash (blue). The average method induced losses for each LLE method is represented by the red horizontal line. Statistically significant differences ($p < 0.05$, Newman-Keuls multiple comparison test) was performed on the absolute average levels between all extraction comparisons. (B) Method induced losses for lipid metabolites associated with Region one. The Folch method demonstrated the highest average signal in this region. (C) Method induced losses for lipid metabolites associated with Region two. The BD method demonstrated the highest average signal but also was the most variable. (D) Method induced losses for lipid metabolites associated with Region three. The BD method demonstrated the highest average signal for TG's. Overall, the Folch method demonstrated the highest average lipid yield, particularly with metabolites associated with Region one. The BD extraction method resulted in the highest yields for metabolites associated with region's two and three, specifically, phospholipids, SM, CER, DG and TG. However, was also the most variable, with only 133 metabolites with a CV<30%. Both Folch and Matyash had 206 and 208 metabolites with CV<30%, respectively.

4.4.2.3.7 The detection of LPC in the hydrophilic fraction

To determine if the losses associated with LPC in the organic fractions is due to its partitioning into the hydrophilic fraction during sample extraction, an endogenous LPC (14:0/0) and LPC (15:0/0) were identified from the extracts acquired by the RPC-UPLC_MS (positive ion mode) and compared. The extracted ion chromatogram (EIC) from one replicate of each extraction method were overlayed for these two compounds to illustrate the difference in peak area (**Figure 4-26**). The highest signals observed for both LPC species were detected in BD samples, followed by Matyash, MeOH, DSPE and very minor levels in Folch. The order of decreasing intensity associated with LPC in the LLE methods (BD > Matyash > Folch), correspond exactly to the findings by Gil *et al.* Regarding the DSPE protocol, its lipid removal capacity is greater than the BD and Matayash methods but less that the Folch. However, the higher yield associated with lipophilic small molecule metabolites (acylcarnitines and other lipophilic xenobiotics), overweighs the presence of the small amount of lipids in the DSPE extraction compared to the Folch extraction, which yielded the lowest amount of lipids in the hydrophilic fractions but an overall lower yield of small molecule metabolites.

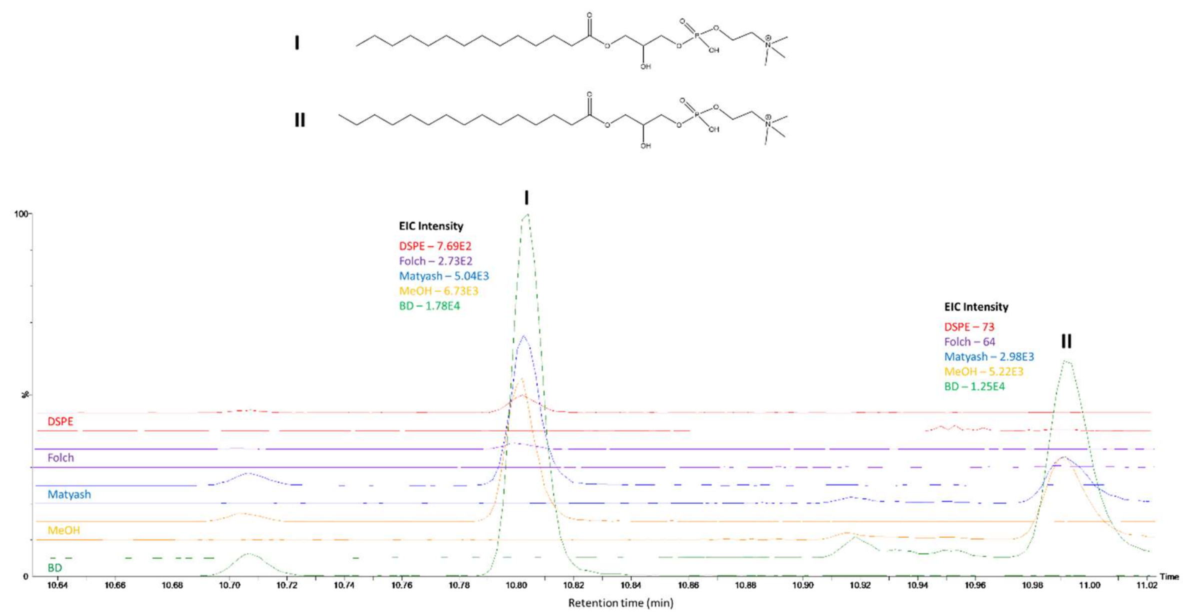


Figure 4-26. Comparison of the relative abundance of a representative set of LPC species , 14 (I) and 15 (II) carbon saturated LPC, present in the hydrophilic fractions of the five different LLE extraction methods, Folch (purple), BD (green), Matyash (blue), DSPE (red) and MeOH (yellow), analysed by RPC-UPLC-MS (positive ion mode). The BD method yielded the highest levels of the LPC species, followed in decreasing order by MeOH, Matyash, DSPE and Folch.

4.4.2.3.8 Liquid-liquid extraction (LLE) comparison summary

The analysis of the hydrophilic/aqueous fractions by RPC-UPLC-MS (positive and negative ion mode) demonstrated that the DSPE method produced the highest number of features between all extraction methods. The method also had the highest number of common features when compared to the pooled extract, which is representative of all the extraction methods and used as a reference to compare the performance of an extraction method. The DSPE samples were the least similar to the pooled extract, with OPLS-DA models indicating that the DSPE method had the highest ratio of discriminant to common features. Targeted analysis (peakpantheR) and acquisition of the extracts *via* the HILIC-UPLC-MS assay, highlighted DSPE, to be the best method for the measurement of short, medium, and long chain acylcarnitine in the hydrophilic fractions, providing a possible explanation as to the difference and higher number of discriminant features observed between the DSPE and pooled extract. These fractions also revealed that the DSPE method was better at removing LPC affiliated metabolites, than the BD and Matyash methods. The organic fractions analysed by the LIPID-UPLC-MS assay (positive ion mode) revealed that the Folch method is the better of the three LLE methods in terms of lipid coverage and overall precision. The BD method was best for TG extractions but demonstrated the greatest overall loss and variability. Lastly, the presence of protein in the hydrophilic fraction of the LLE methods is therefore a limitation for the analysis of polar metabolites by RPC methodologies, especially for high throughput applications.

4.4.2.4 SPE VS DSPE

The optimised DSPE protocol was compared to known SPE methods/plates for lipid removal. As DSPE and SPE are essentially different extraction techniques, comparison between the two were undertaken by creating an SPE plate using the DSPE parameters. In conjunction with Phenomenex, the optimised sorbent weight was packed into cartridges in a 96-well format to replicate a typical 96-well SPE plate. Sample extraction using this plate (Sepra-SPE) was carried out in a similar manner to Sepra-DSPE. This included making sure that sample volume, conditioning/washing of the sorbent, solvent, recovery, and reconstitution volumes were all identical. A comparison between the different protocols were then conducted using PCA, and by reporting the total number of reproducible features detected. For PCA, a score plot was generated to illustrate the variance within the dataset (**Figure 4-27.A**). In both positive and negative modes, both Sepra-DSPE and Sepra-SPE samples clustered together highlighting a similarity in their profiles. This highlights the importance of sorbent and solvent optimisation and its impact on small molecule measurement.

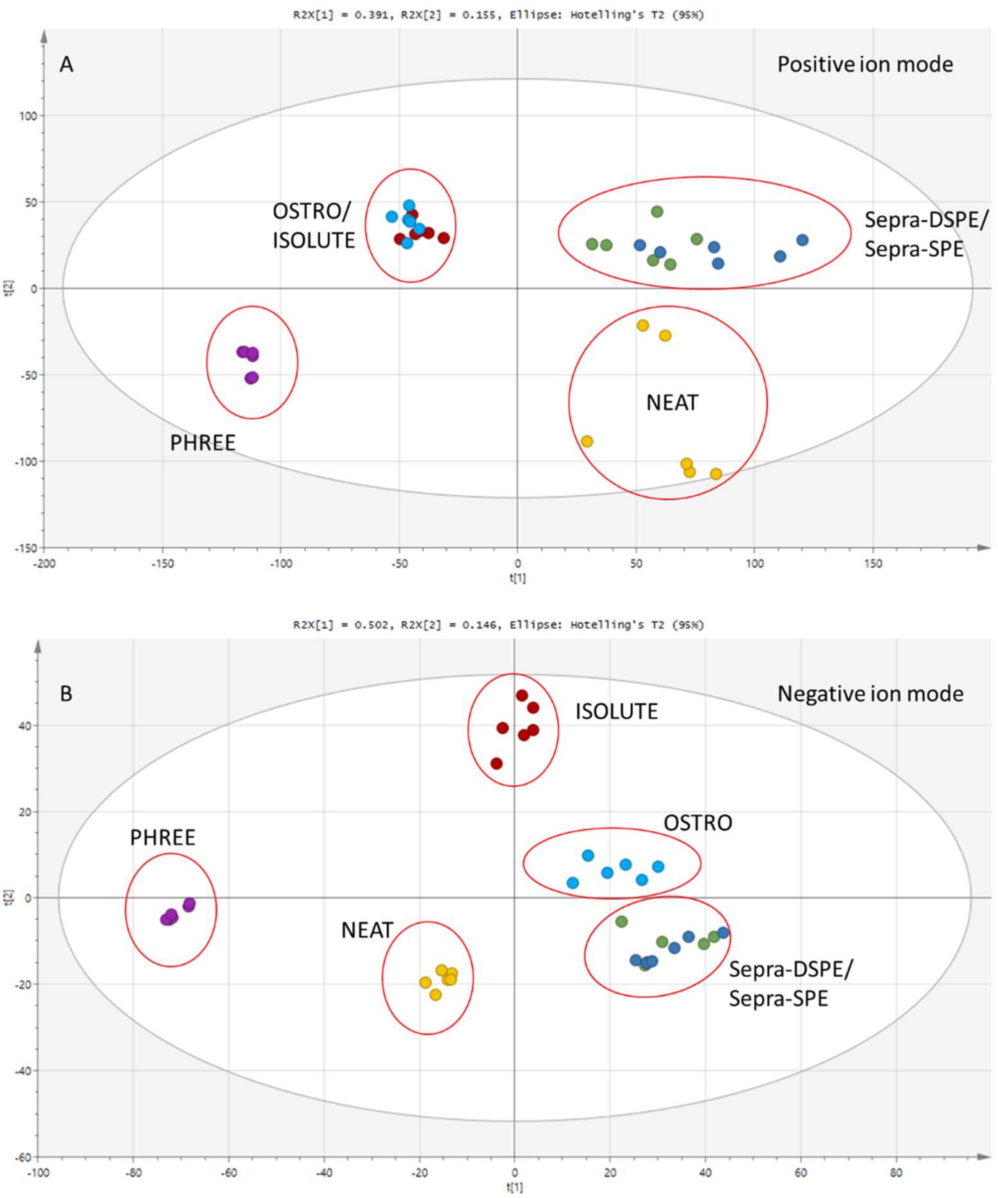


Figure 4-27. A PCA score plot comparing different phospholipid removal SPE protocols in plasma samples acquired by RPC-UPLC-MS (positive and negative ion mode)-, alongside a DSPE treated samples, the DSPE sorbent packed into a SPE format, and Neat plasma samples with no treatment.

(A) PCA score plot comparing Sepra-DSPE (green), Sepra-SPE (blue), Neat (yellow), OSTRO (light blue), ISOLUTE (brown) and PHREE (purple), which were analysed by RPC in positive ion mode

(B) PCA score plot comparing Sepra-DSPE (green), Sepra-SPE (blue), Neat (yellow), OSTRO (light blue), ISOLUTE (brown) and PHREE (purple), which were analysed by RPC in negative ion mode

The PCA showed clear clustering of samples based on their treatment. The Sepra samples clustered together indicating highly similar profiles, thereby allowing comparison to SPE lipid removal protocols. OSTRO, ISOLUTE and PHREE plates all incorporated acetonitrile and therefore was the main driver of variance, as a lower number of features were detected with these protocols. The PHREE plate was the least favourable exhibiting species related to protein (like the NEAT), causing major instrument issues.

The total number of features detected in the positive mode followed the order: Sepra-DSPE > Sepra-SPE > Neat > OSTRO > ISOLUTE > PHREE (2566, 2312, 1986, 1580, 1405 and 469, respectively). In negative ion mode, the same order was observed (828, 810, 679, 549 and 176, respectively). In both modes, the PHREE samples clustered on its own and had the lowest number of features detected. This is in part due to MeCN as the protocol's extraction solvent. We have demonstrated in **section 4.4.1.2**, that MeCN has a lower extraction efficiency and selectivity of small molecules in comparison to MeOH. A similar RPC-UPLC-MS profile was observed between the PHREE and NEAT samples, with the detection of high molecular weight multiply charged species, eluting between seven and 11 which we have now attributed to residual protein. It seems that during sample extraction with the PHREE plates, the precipitate from the PP, did not remain within the cartridge, and made it through to the collected filtrate. This later becomes apparent in the analytical run, as both the NEAT and PHREE samples resulted in stoppages of the LC-MS due to over overpressure caused by an accumulation of protein in the LC column. The presence of these species may potentially affect the ionisation of other LMW ions, resulting in a fewer number of features detected, as observed with the PHREE samples. The OSTRO and ISOLUTE plates also used MeCN in the extraction but was void of any protein/peptide species. As a result, a higher number of features, when compared to the PHREE plate, were observed, and explains the clustering in the top-left quadrant of the PCA, away from the PHREE samples in positive ion mode. In negative ion mode, the OSTRO plate clustered with the Sepra samples, indicating a similarity in the blood profile, however, was lower in the number of features detected. It is the fact that MeCN was used as the extraction solvent for ISOLUTE, PHREE, and OSTRO which explains the greatest source of variation in the data as described by PC1 (45%). PC2 explained approximately 20% of the variation and was attributed to the presence of the protein/peptide species. This is corroborated by the fact that both PHREE and NEAT samples are in line along PC2 and contain significant levels of these species. OPLS-DA models were used to identify the features that contribute to the greatest influence between Sepra-DSPE and NEAT samples, and further highlighted the above point. Significant features were based on their VIP scores (VIP > 1.5). The results show that for single component OPLS-DA models, in both positive and negative ion modes, the R2Y (> 0.991) and Q2Y (> 0.855), and permutation testing indicated low variability and an excellent predictive ability. A loadings plot (with retention time on the x-axis and *m/z* on the y-axis) of the two OPLS-DA models, show that the majority of discriminating features eluted between 7 and 11 minutes, which are representative of the protein/peptide species (**Figure 4-28.B**). There were also features eluting after 11 minutes which were attributed to LPC affiliated metabolites, as annotated in **section 4.4.2.3.7**. Other than these features, the remaining profile between the Sera-DSPE and NEAT samples are similar (between zero and seven minutes).

The use of NEAT plasma was not a good sample to use as a control. Urine should have been used in this experiment (as demonstrated in previous sections), as we could then evaluate the effect of SPE treatment on small molecules. However, the use of blood samples did highlight a similarity between Phree and Neat samples suggesting an inefficiency of protein removal during extraction. Overall, the Sepra samples boasted the highest number of detected features in comparison to the other lipid removal lipid removal SPE plates. The experiment demonstrated that both the sorbent and solvent can affect the selectivity of small molecules, and the parameters use for the DSPE protocol was the better of all protocols, detecting a greater number of high-quality features whilst also delivering a cleaner extract.

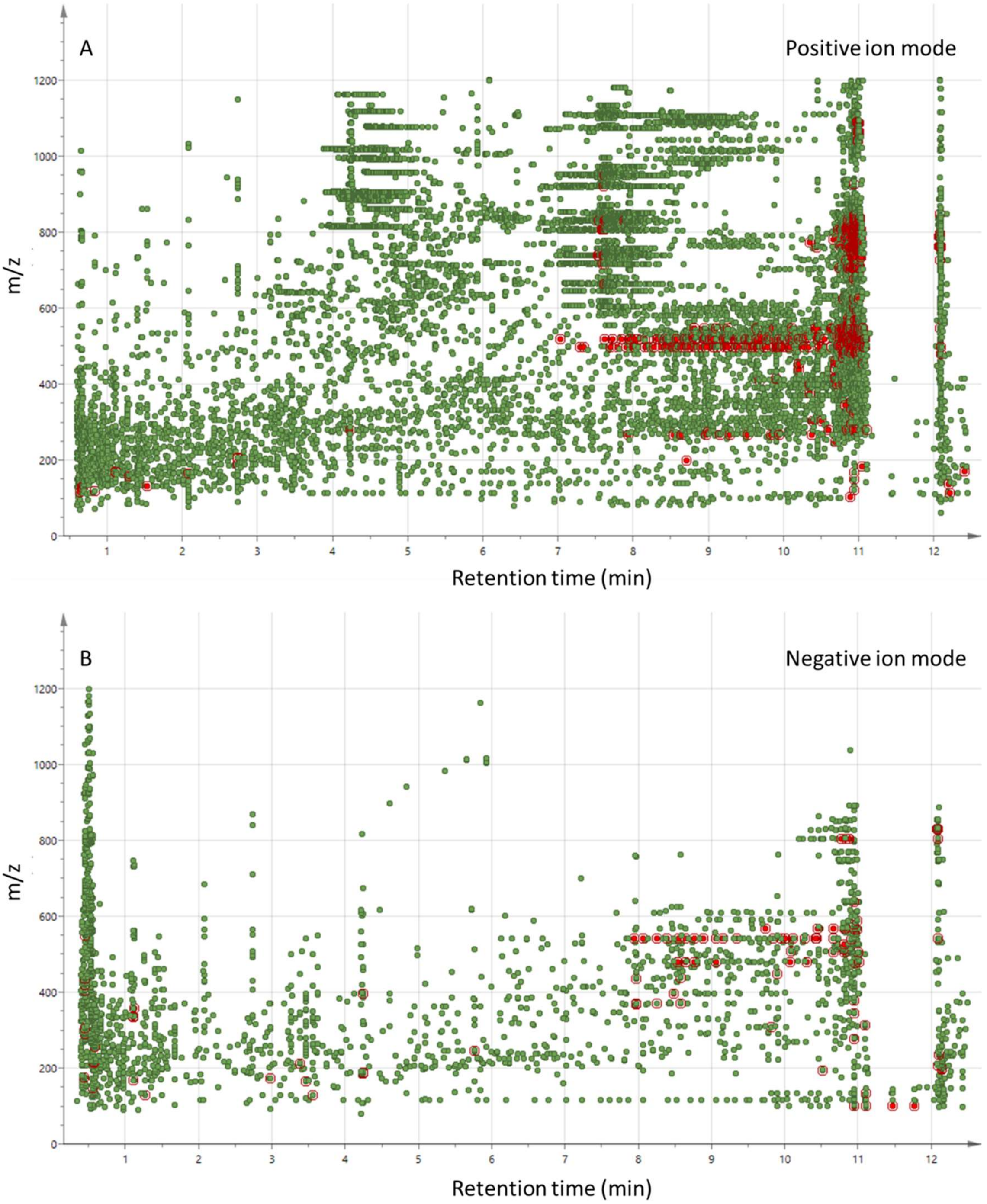


Figure 4-28. One-component OPLS-DA loadings plots showing the separation and discriminating features (respectively) between the RPC profiles of the NEAT samples (with loadings coloured in red exhibiting a VIP > 1.5), to the DSPE treated samples.

(A) The explained variance (R2Y) was 0.994 and predictive ability was 0.812 for Neat samples vs DSPE treated samples in positive ion mode;

(B) The explained variance (R2Y) was 0.974 and predictive ability was 0.964 for Neat samples vs DSPE treated samples in negative ion mode;

The loadings plot (with retention time on the x-axis and *m/z* on the y-axis) of the two OPLS-DA models, highlight most discriminating features to elute between 7 and 11 minutes in both modes. These discriminating features are representative of the protein/peptide species identified in previous sections.

4.4.3 *Application*

4.4.3.1 Analysis of plasma and serum samples

The acquired data were pre-processed within ongoing operations of the NPC according to established QC protocols for metabolic phenotyping (Lewis et al., 2016, Sands et al., 2019) thereby ensuring high data quality. No obvious drifts or outliers were observed in TIC for both the SR and LTR QC samples in both ionisation modes, and for both studies (**Figure 4-29.A-B** and **Figure 4-30.A-B**). The distribution of the % RSD in relation to the feature intensity in the SR samples was also assessed (**Figure 4-29.C-D** and **Figure 4-30.C-D**). The distribution is further divided into a lower quartile range (green), interquartile range (blue) and upper quartile range (green), highlighting the precision of features based on their measured signals. The median RSD values in the MARS and AZ Study 12 studies were 8.4% and 5.9% respectively in positive ion mode and 7.6% and 8.3% in negative ion mode.

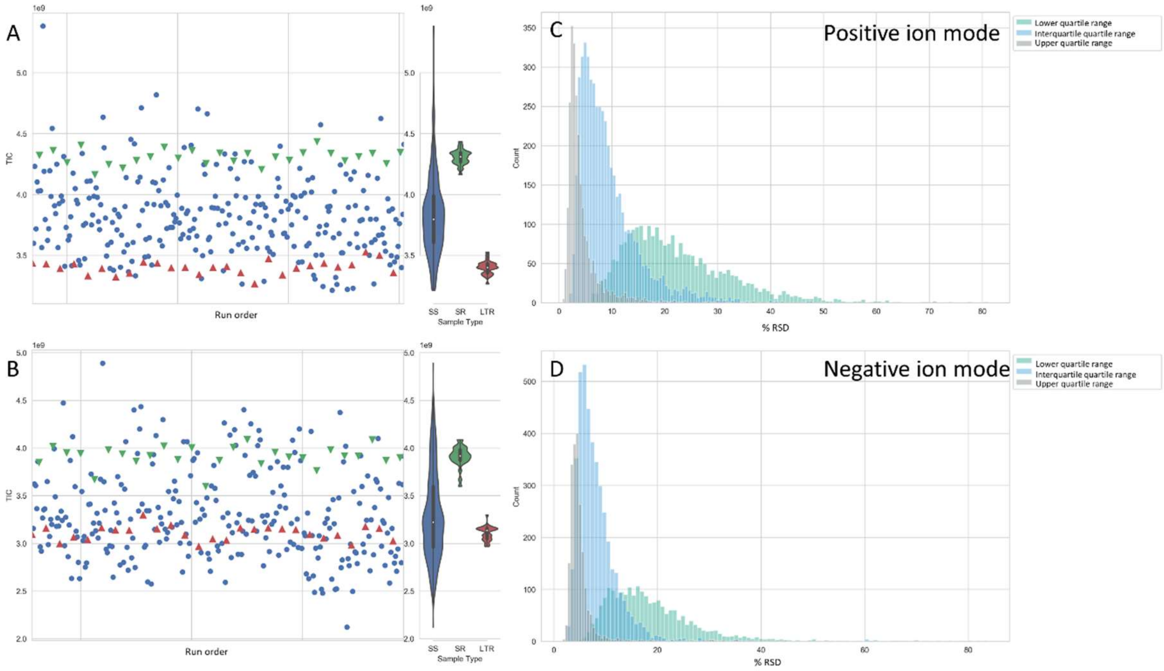


Figure 4-29. RPC-UPLC-MS TIC of all plasma samples (Study samples – blue, SR – green, and LTR – red) in MARS (A-B), and the % RSD distribution for all features passing the dilution series filter (C-D). Alongside the TIC scatter plots are violin plots exhibiting the TIC density for each sample type. The distribution plots are %RSD segmented by mean feature intensity into quartiles.

(A) TIC of all samples in MARS (positive ion mode) against the run order;

(B) TIC of all samples in MARS (negative ion mode) against the run order;

(C) % RSD distribution in positive ion mode. Median RSD value was 8.4%;

(D) % RSD distribution in negative ion mode. Median RSD value was 7.6%.

Data from the MARS plasma study (n=285) resulted in 2837 detected metabolite features in positive ion mode and 1523 in negative ion mode. Repeated observation of the pooled sampled (SR) throughout the analytical batch demonstrated high precision, with the majority of features that occupied the interquartile and upper quartile intensity range having an RSD of less than 30%. The TIC plots exhibited no major outliers or trends with respect to the QC samples.

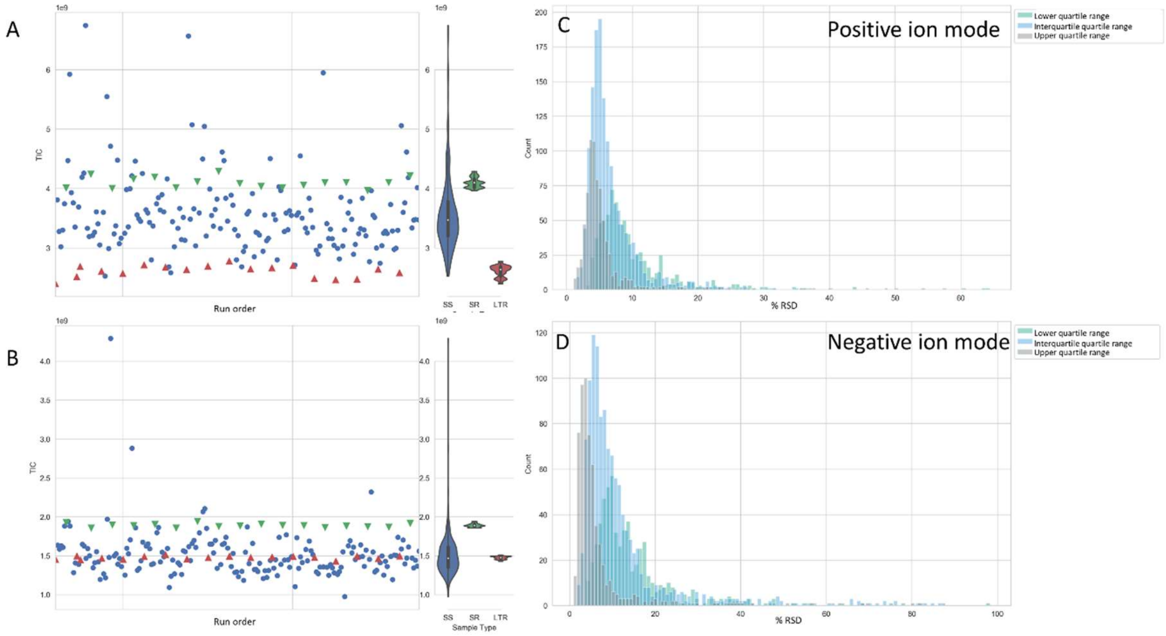


Figure 4-30. RPC-UPLC-MS TIC of all serum samples (Study samples – blue, SR – green, and LTR – red) in AZ Study 12 (A-B), and the % RSD distribution for all features passing the dilution series filter (C-D). Alongside the TIC scatter plots are violin plots exhibiting the TIC density for each sample type. The distribution plots are %RSD segmented by mean feature intensity into quartiles.

(A) TIC of all samples in MARS (positive ion mode) against the run order.

(B) TIC of all samples in MARS (negative ion mode) against the run order.

(C) % RSD distribution in positive ion mode. Median RSD value was 5.9%;

(D) % RSD distribution in negative ion mode. Median RSD value was 8.3%.

Data from the AZ Study 12 serum study (n=169) resulted in 2936 detected metabolite features in positive ion mode and 2342 in negative ion mode. Repeated observation of the pooled sampled (SR) throughout the analytical batch demonstrated high precision, with the majority of features that occupied the interquartile and upper quartile intensity range having an RSD of less than 30%. The TIC plots exhibited no major outliers or trends with respect to the QC samples.

Repeated observation of reference features from the pooled QC samples (SR) throughout the two profiling studies, demonstrated high precision with both mean peak area and retention time RSD < 10% (**Table 4-6** and **Table 4-7**).

Table 4-6. Retention time and Peak area precision of reference standards within the MARS plasma project acquired by RPC-UPLC-MS in positive and negative ion mode. Repeated observations of reference features from the pooled QC samples throughout the analytical batch demonstrated high precision with mean retention time RSD < 1% and mean peak area RSD <10% with no post batch correction required.

	%RSD							
NPC Project	MARS							
Polarity	RPC-UPLC-MS (positive ion mode)				RPC-UPLC-MS (negative ion mode)			
QC Sample	SR		LTR		SR		LTR	
	RT	Peak Area	RT	Peak Area	RT	Peak Area	RT	Peak Area
L-Glutamine-$^{13}C_5$	0.0	6.3	0.0	7.0	0.0	5.0	0.0	3.8
L-Glutamic acid-$^{13}C_5$	0.0	6.7	0.0	7.5	0.5	5.4	0.5	4.8
Creatinine-(methyl-d_3)	0.6	3.4	0.6	3.6	0.5	4.6	0.4	4.3
L-Isoleucine-13C6,15N	0.3	7.8	0.3	7.6	0.3	9.3	0.3	6.5
L-Leucine-$^{13}C_6$	0.3	8.1	0.3	8.4	0.3	7.7	0.2	9.2
L-Tryptophan-$^{13}C_{11}$, $^{15}N_2$	0.4	8.9	0.3	10.3	0.1	5.3	0.1	4.1
Cytidine-5,6- d_2	0.0	5.3	0.0	5.4	0.1	4.2	0.1	3.9
L-Phenylalanine-$^{13}C_9$, ^{15}N	0.1	5.8	0.1	6.6	0.0	2.0	0.0	1.7
N-Benzoyl- d_5-glycine	0.0	6.3	0.0	7.0	0.0	5.0	0.0	3.8

Table 4-7. Retention time and Peak area precision of reference standards within the AZ Study12 Serum project acquired by RPC-UPLC-MS in positive and negative ion mode. Repeated observations of reference features from the pooled QC samples throughout the analytical batch demonstrated high precision with mean retention time RSD < 1% and mean peak area RSD <10% with no post batch correction required.

	%RSD							
NPC Project	AZ Study 12							
Polarity	RPC-UPLC-MS (positive ion mode)				RPC-UPLC-MS (negative ion mode)			
QC Sample	SR		LTR		SR		LTR	
	RT	Peak Area	RT	Peak Area	RT	Peak Area	RT	Peak Area
L-Glutamine-$^{13}C_5$	0.0	5.2	0.6	4.0	0.0	4.1	0.0	3.1
L-Glutamic acid-$^{13}C_5$	0.0	2.7	0.0	4.1	0.0	5.4	0.8	3.5
Creatinine-(methyl-d_3)	0.0	3.8	0.0	4.4	0.0	2.1	0.0	1.7
L-Isoleucine-13C6,15N	0.3	2.2	0.3	3.6	0.0	6.5	0.2	7.4
L-Leucine-$^{13}C_6$	0.3	2.1	0.3	4.1	0.2	6.8	0.3	8.2
L-Tryptophan-$^{13}C_{11}$, $^{15}N_2$	0.0	4.4	0.0	5.7	0.2	2.5	0.2	3.5
Cytidine-5,6- d_2	0.3	1.3	0.3	2.4	0.1	6.6	0.1	6.5
L-Phenylalanine-$^{13}C_9$, ^{15}N	0.2	2.7	0.0	2.2	0.1	4.0	0.0	2.4
N-Benzoyl- d_5-glycine	0.1	2.7	0.1	3.3	0.0	1.6	0.0	1.1

4.4.3.2 *Targeted annotations*

Of the features which passed QC protocols, 144 metabolites were annotated in AZ-Study12 (**Figure 4-31.B**) and 141 metabolites in MARS (**Figure 4-31.A**). In conjunction with the two other blood profiling assays, polar metabolites by HILIC-UPLC-MS and lipid metabolites by LIPID-UPLC-MS, an additional 129 metabolites were reported in AZ-Study12 and 111 in MARS. Of these metabolites for both studies, 23 were xenobiotics, eluted well within acceptable RT confines of the RPC-UPLC-MS assay. Selectivity of all compounds that are moderately hydrophobic and/or amphipathic, are poor for LIPID-UPLC-MS and HILIC-UPLC-MS analysis. If detected, would not retain, or would elute too close to the t_0 of the chromatographic method. As the LIPID assay is a method specifically designed to target lipid classes, this leaves only the HILIC assay as the only other blood profiling assay for xenobiotic detection at the NPC. Only four xenobiotics, cotinine, caffeine, acetaminophen and cortisol passed QC and was detected *via* the HILIC assay, but all these xenobiotics were close to the t_0. So, although detected in HILIC, caution should be taken using this data for further analysis.

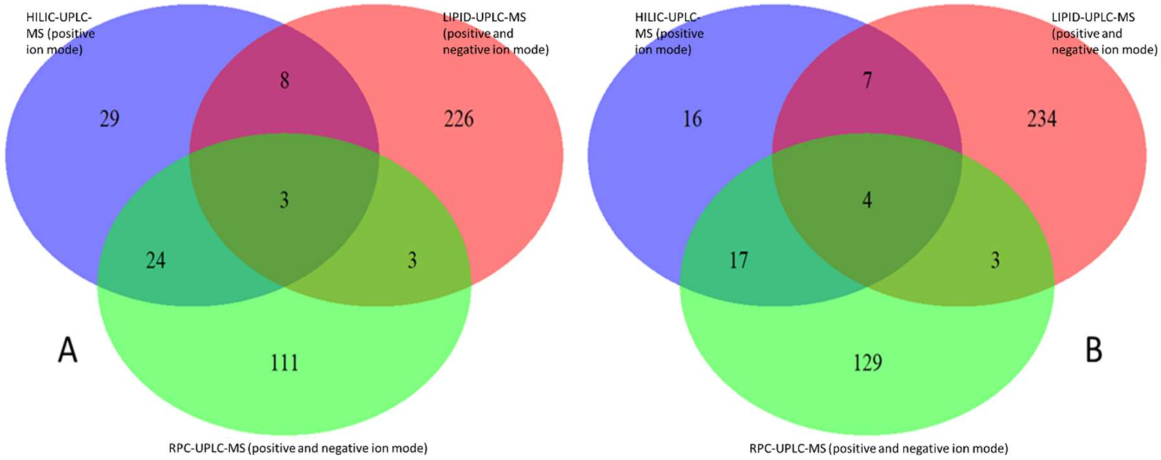


Figure 4-31. Venn diagram for the number of annotated metabolites detected in MARS and AZ STUDY12 studies using all blood NPC profiling assays, **i.e. polar metabolites by HILIC-UPLC-MS (positive ion mode- purple), lipids by LIPID-UPLC-MS (positive and negative ion mode - red) and the developed DSPE method by RCP-UPLC-MS (positive and negative ion mode – light green)**

(A) Venn diagram for the number of annotated metabolites detected in MARS (plasma)

(B) Venn diagram for the number of annotated metabolites detected in AZ Study12 (serum)

An additional 111 (MARS) and 129 (AZ Study12) metabolites (endogenous and xenobiotics) were annotated from these two exemplar studies.

4.5 Results Summary

The aim of the work described in this chapter was to develop an analytical strategy to improve xenometabolome coverage in blood products.

Comprehensive effort was therefore undertaken to construct a method for the depletion of lipids from blood which rendered the sample type fit for the analysis of amphipathic and moderately hydrophobic metabolites, encompassing many xenobiotics, by RPC. The approach is accomplished using DSPE with C18 sorbent, thereby enabling large scale high throughput LC-MS (RPC) profiling. To address this main aim, development of this DSPE sample preparation method for blood products was split into three design stages; optimisation, validation and application.

The first stage involved optimising the components that make up DSPE, i.e., the C18 sorbent use to target lipophilic species, and the solvent used to extract the metabolite subset of interest.

Properties of the Sepra C18 sorbent from Phenomenx was most favourable in comparison to other C18 sorbent material, due to small particle size, and greater surface area. As a result of these properties, a greater number of fatty acid or C18 chains will attach to the surface of the underlying support material, reducing the number of polar endcapped sites, and thereby ensuring that only highly lipophilic metabolites in blood are targeted by the sorbent.

A methanol:acetronitrile (1:1) mixture provided the best solvent composition for extracting this subset of metabolites, producing the greatest number of highly reproducible metabolic features, minor signals of residual protein, and minimal evidence of biphasic partitioning.

The combination of the sorbent and solvent variables make a slurry, and it is both the concentration and volume of the slurry which were optimised simultaneously using a DOE protocol. Rather than a full factorial design which would have incorporated all the possible combinations in the design model, a central composite face design was efficiently used, providing much of the necessary information on variable effects and overall experimental error, in a minimum number of experimental analyses. The MODDE software highlighted several performance indicators plots which were used to evaluate and assess the quality of the PLS model produced, this included;

1. summary of fit, in which R2/Q2 ratios greater than 0.5 were evaluated and used as a measure of cross validation reproducibility and model validity.
2. a plot depicting regression coefficients for slurry concentration and slurry volume factors at each retention time bin. Coefficients whose uncertainties exceeds their actual values have no significant contribution to the model and removed. Slurry volume and concentration had

no significant impact within each retention time bin for the RPC-UPLC-MS based assays. For the LIPID-UPLC-MS assay, the slurry concentration was significant in the negative direction, implying as more sorbent is added, more lipid is removed.

3. Finally, the sweet spot contour plots and design space plots, both of which highlight regions in the design space, where all responses are within the specified range, and most sensitive to small fluctuations in the factors.

Collectively, utilising these performance indicator plots, both the slurry concentration and slurry volume were optimised; 325 uL of a 16mg/mL slurry solution is added to every 100 uL of blood sample.

The second stage involved testing method reproducibility, recovery, and effectiveness of the protocol at lipid removal and small molecule measurement (comparison to SPE and LLE).

Reproducibility addressed whether the sorbent can be reproducibly dispensed into a 96-well format, and the precision of the molecular profile of detected metabolic features from repeated measures of the same sample. An RSD of less than 10% sorbent weight was measured across all wells of a 96-well plate and the median RSD of less than 5% was measured for all detected features passing the dilution series filter.

Recovery was addressed by evaluating the signals of specific xenobiotic targets and metabolic features passing the dilution series filter, between SHAM and DSPE treated samples. Overall, SHAM and DSPE treated samples demonstrated little significant difference in mean signal intensities for the exemplar set of xenobiotics, eluting before 10 minutes for data acquired by RPC-UPLC-MS. DSPE treatment has also had minimal global effect on small molecule recovery.

To address whether the DSPE procedure is effective at lipid removal and small molecule measurement, blood profiles were compared to other sample preparation methods, namely SPE lipid removal plates, and LLE. The DSPE method was the better of all protocols, detecting a greater number of reproducible metabolic features, whilst also delivering a cleaner extract free of any residual protein. The solvent composition of methanol:acetonitrile for DSPE was crucial in detecting more features, as most lipid removal SPE plates utilise acetonitrile, which has poorer selectivity. Targeted analysis of DSPE samples demonstrated higher yields for short, medium and long chain acyl carnitine metabolites, and lipophilic drugs. Finally, although a compromise was made to leave metabolites associated with Region 1 of the lipid profile intact (LPC's), the DSPE method demonstrated lower yields of these LPC metabolites in comparison to BD and Matyash LLE methods.

The final stage were applications of the DSPE method to the exemplar MARS and AZ Study 12 population studies.

In both studies, no obvious outliers were observed in the TIC for the QC samples (SR and LTR), the median RSD values for all detected features were less than 10% in both ionisation modes, and repeated observation of the reference standards from the SR samples for both intensity and retention time had RSD values of < 10%. Overall, an additional 129 metabolites were reported in AZ-Study12 and 111 in MARS. Approximately 23 metabolites were identified as xenobiotics, which would not have been annotated by the other profiling studies offered at the NPC. This therefore demonstrates that xenobiotic coverage was significantly enhanced for both studies via this analytical strategy.

4.6 Significance of Findings

The work presented in this chapter is useful in three main ways:

1. An additional profiling method for blood products has been added to the NPC profiling portfolio; this method is optimized for the subset of metabolites that fall in-between those suited to the existing HILIC or lipid methods and encompasses many xenobiotics of interest.
2. The method combines an efficient sample preparation technique (DSPE) to a robust chromatographic technique (RPC).
3. This method supports an analytical strategy to separate and distinguish between xenobiotic and endogenous signatures potentially increasing xenobiotic coverage in population studies.

4.7 General Discussion

Whilst the development of the analytical strategy features a substantial amount of trial and error and process of elimination in the scoping of materials and chemistry conditions, it also relies on the principles of DOE. The approach benefitted the work greatly by allowing assessment of a broad parameter space with less overall experiments physically performed. For example, when optimising the slurry composition, if the experiment set was to be undertaken in the conventional sense, i.e. testing all experimental conditions, the final experiment would have resulted 55 different conditions to analyse. Using DOE, only 11 experiments were required to assess the same parameter space. This immediately highlights the benefit of the DOE approach to speed up the design process, thereby reducing cost and time. The final method is fit for application to blood product analysis, bringing the benefits of RPC as a high performing and dependable separation technique into focus for less hydrophobic metabolites in an otherwise complex sample type.

Historically, RPC has been the benchmark for LC/MS profiling in untargeted metabolomic studies (Zelena et al., 2009, Wilson et al., 2005, Plumb et al., 2006, Dunn et al., 2011). Uniform peak shape across the separation, stable retention times and high speed of equilibration make RPC a popular analytical separation technique for examining the chemically diverse range of metabolites present in biofluids (Psychogios et al., 2011). Lipids constitute a large portion of the metabolic content present in biofluids, with more than 60% annotated of all metabolites listed in the human metabolome database (Wishart et al., 2013).

As such, RPC has been most widely used for the analysis of complex lipids and therefore well established (Cajka and Fiehn, 2014). However, highly polar, and ionic metabolites suffer retention on the RPC systems, making measurement unsuitable. Ion-paring agents for highly polar metabolites have been added to RPC mobile phase buffers to improve measurement although these additives can result in heavy contamination of the entire chromatographic system (Tulipani et al., 2015, Yanes et al., 2011). HILIC has increasingly been incorporated into metabolomic research, as an alternative to RPC for the analysis of highly polar metabolites (Cubbon et al., 2010, Spagou et al., 2010).

The presence of endogenous phospholipids in significant levels are however considered to be a real problem in HILIC (Tsakelidou et al., 2017). The analysis of polar metabolites and lipids using RPC based methods for large scale profiling has also been reported, however the authors have stated that that phospholipids can interfere with retention time of all endogenous compounds (Dunn et al., 2011).

Sample preparation plays an important role in blood-based metabolomics and lipidomic analysis, as methods need to be reproducible and efficiently denature and precipitate proteins, whilst also ensuring maximum extraction recoveries for metabolites. Therefore, the most popular practice for large scale metabolome studies is to focus on different sample extraction and separation (chromatographic) procedures. Studies that have reported comprehensive coverage of metabolites have analysed polar extracted fractions using HILIC and non-polar lipid extracted fractions using RPC with C_8 or C_{18} columns (Dunn et al., 2011, García-Cañaveras et al., 2011, Urayama et al., 2010). This approach of complementary analytical methods each covering a specific subset of metabolites based on hydrophobicity (i.e. the HILIC and LIPID analysis) helps to build more complete metabolome coverage.

To limit time, cost and sample consumption, simultaneous analysis of both the metabolome and lipidome has also been undertaken (Cai and Li, 2016). Cai *et al.* prepared one extraction for the analysis of both polar metabolites and lipids in plasma using HILIC-MS. The solvent composition of MEOH/MeCN/Acetone used in the plasma extraction demonstrated the highest efficiency for the extraction of both polar metabolites and lipids. In this chapter, MeOH:MeCN was assessed to be the most suitable extraction solvent. Acetone was eliminated as a candidate extraction solvent due to an affinity of lipids to remain in solution rather than bind to the DSPE sorbent. It is therefore not surprising that the combination incorporating all three yielded highest recoveries for both polar metabolites and lipids.

In addition to analytical issues, the extraction procedure involved for the simultaneous analysis of both polar metabolites and lipid species, are tailored specifically for these types of metabolites. However, these protocols provide poor coverage for the subset of metabolites that fall in between these two chemistries, i.e., moderately hydrophobic, and amphipathic, which encompasses many xenobiotics. RPC is still the better choice of the two separation chemistries, to analyse this specific subset of metabolites, however as discussed in the introduction, the presence of complex and neutral lipid species in blood products can accumulate and elute unpredictably from the column which could thereby exhibit suppressive effects on the ionisation of LMW metabolites. Nasar *et al*. describes the metabolites that fall within this polar-lipid range as “semi-polar”(Naser et al., 2018). These semi-polar metabolites were classified by their inherent lipophilicity calculated by using the logarithm of the partition coefficient (logP). Semi-polar metabolites exhibited logP values from approximately −2 to 1.5. This group also found that for comprehensive metabolome coverage, analysis had to be split into a HILIC method for polar metabolites and two RPC methods for semi-polar (using a C_{18} column) and lipid metabolites (using a C8 column), which agrees with the approach adopted by the NPC. Another study classified xenobiotics (drugs and steroids), as having a logP between 0-5 (Drouin et al., 2018), and its measurement in blood samples required sample clean-up methods, such as LLE and SPE, to remove most of the lipophilic interference. The group also describes a variant of DSPE called dispersive liquid–liquid microextraction (DLLME) which can be used as an alternative sample extraction method to LLE and SPE. Rather than using a solid sorbent component as with DSPE, in DLLME, a nonpolar water immiscible high-density solvent (such as chloroform) is added to the sample-solvent mixture and acts as extraction phase. This method reports high yields for drugs with faster analysis time when compared to SPE, however has not been implemented for metabolomic applications.

The DSPE method was optimised for high recoveries for the subset of metabolites of interest. It also has the added benefit of being reproducible, inexpensive, and effective for high through-put analyses. However, a limitation of the method, is that the procedure does not eliminate lipids completely, especially those with single fatty acyl chains. For example, lysophospholipids were heavily retained in solution by the method in order to ensure more hydrophobic metabolites were not depleted. If a more complete reduction of these less hydrophobic lipids were required, the solvent ratio used in the slurry could be adapted (e.g., making the solvent component more polar), but some sacrifice to the yield of those small molecules would certainly be made. Alternatively, a higher concentration of sorbent could be used to target lipids, as demonstrated in the DOE experiments. However higher slurry concentrations would also mean higher slurry volumes and ultimately higher sample volume, which maybe an issue if sample volume is limited. The approach is therefore a compromise, intended to deplete the most highly retained lipid species and thereby reduce their accumulation and carryover on column, but with a bias towards complete retention of small molecules in solution (at the expense of more complete lipid removal).

A range of different SPE sorbents, packed into 96-well plates for high throughput analyses, has now been introduced that remove proteins and phospholipids from biofluids in a single step, with each exhibiting different degrees of extraction efficiency (Tulipani et al., 2013, Neville et al., 2012, Patterson et al., 2015). However, these plates appear to be more appropriate for targeted applications (sample cleaning, compound enrichment) than for global profiling, with some groups reporting loss of recovery of small molecules and introduction of contaminants (Armirotti et al., 2014, Simón-Manso et al., 2013). SPE sorbent materials tested in this chapter exhibited a lower number of metabolic features and were found to retain hydrophilic analytes together with endogenous interferences (protein/peptides). This agrees with the findings by Tsakelidou *et al.*, where the SPE phospholipid removal plate required additional elution steps to remove these interferences (Tsakelidou et al., 2017). The same species were also observed with the initial analysis of the hydrophilic fractions of the LLE extractions by the RPC profiling method demonstrated in this chapter. As a result, all extracts were subjected to an additional protein precipitation step. Compared to the DSPE method, the hydrophilic fraction from LLE methods also exhibited significant levels of LPC affiliated metabolites, which were highest in the BD and Matyash extraction methods. Additionally, small molecule metabolites such acylcarnitine’s, have been reported to be lost in biphasic extraction, and was also observed in this investigation (Patterson et al., 2015). The additional steps and effort needed to analyse the hydrophilic fraction from LLE methods for small molecules analysis, and loss of small molecule metabolites makes LLE as means for lipid removal, unfavourable. Returning to the original hypothesis, where we stated that DSPE would provide a better alternative to LLE and SPE phospholipid removal plates for metabolome and xenometabolome measurement, the results support this hypothesis.

The application of DSPE for lipid removal in metabolomic investigations involving blood products is novel, in that it has not been explored in any great level of detail. For blood sample preparations in a metabolomic investigation, Tsakelidou *et al.* used a QuECHERS (quick, easy, cheap, effective, rugged and safe) kit that utilise DSPE. It is a form of liquid-liquid extraction with implements a solvent and a high salt buffer where analytes of interest will partition into the organic phase. The authors used a specific QuECHERS clean-up step which involves an initial protein precipitation, followed by a DSPE that utilises a sorbent mixture of MgSO4 and a proprietary primary secondary amine. They did report however lower recoveries for 24 out of the 53 metabolites when compared to conventional protein precipitation extraction method (using methanol). A high recovery was observed for xenobiotics from this list of metabolites and recommended that a C18 sorbent should be investigated, which was exactly what was investigated for this work. At the time of this work, no C18 material as the sorbent was available in the QuECHERS kits, but kits are now available with C18 sorbents from various vendors (Waters, Sigma, Thermo Fisher Scientific and Agilent) which represents another avenue of research which can potentially be explored.

Configurations of instrumentation to partition lipids, either by directing to waste or measurement of the lipid profile, can also be undertaken via 2-dimensional LC methods (Li et al., 2015, Schwaiger et al., 2019, Broeckling and Prenni, 2018, Chalcraft and McCarry, 2013). Commonly used in proteomics, eluent from one column is collected in a sample loop and then injected onto a second column exhibiting orthogonal properties. In these studies, dual HILIC and RPC were combined in serial or parallel to high-resolution mass spectrometry. Injection stacking was proposed as an alternative to the reported conventional based 2-D methods, significantly reducing cost and hardware. Broeckling *et al.* utilised a Mataysh extraction on whole human blood, whereby the injection of the organic (lipophilic) extract onto a RPC column under isocratic conditions, followed immediately by injection of the aqueous (hydrophilic) extract onto the same column (stacking). The authors did state that this platform does not claim to cover comprehensive profiling and should only be considered a viable option if time and resources are limited.

A Q-TOF-MS was used for this work as it can offer a higher sensitivity, high mass resolving power, rapid scanning capabilities, and cover a wider mass range with high mass accuracy than other profiling platforms used in metabolomics (GCMS or NMR). As such, the use of Q-TOF-MS instrumentation has also gained considerable attention in forensic and clinical toxicological applications due to its high sensitivity and specificity in xenobiotic detection (Allen and McWhinney, 2019, Ranasinghe et al., 2012, Maurer and Meyer, 2016, Bidny et al., 2016, Dalsgaard et al., 2012, Dresen et al., 2010). The unique MS scanning capabilities can allow MS/MS fragmentation data to be acquired without any prior knowledge of the parent ion through acquisition modes such as data dependent (DDA) and data independent (DIA) acquisition (Roemmelt et al., 2014). The simultaneous acquisitions of both parent and fragment ions are formed in one analytical run, with the added benefit of matching MS signatures alongside spectral databases for putative annotations to be made in real time, which is something which is to be considered for future analyses. Furthermore, unknown xenobiotic metabolites which cannot be matched to spectral libraries will often exhibit common mass fragments with their parent compound (Rathahao-Paris et al., 2014, Sauvage and Marquet, 2012). Therefore, putative matches can be potentially made if spectra of the parent xenobiotic are known. The limitation of these techniques however is the need for high performance computing (memory, storage, computer processing capabilities etc), and less sensitivity. In addition to this limitation, given the diverse range of metabolites detected in blood products, with > 30 000 small molecules present in human serum (Ivanisevic et al., 2013), the Q-TOF-MS may not have the desired selectivity to resolve such a vast array of metabolites, especially with regards to biologically relevant isomer pairs. The higher resolution attained by Orbitrap mass spectrometers can characterise the isomeric composition of samples without sacrificing sensitivity. As a result, the Orbitrap is increasingly being used in metabolomics investigations, thereby enabling scientist to carry out quantitation similar to the Q-TOF-MS with its DIA capabilities (Barbier Saint Hilaire et al., 2020, Zhou et al., 2017, Bonner and Hopfgartner, 2019). Separation methods, such as the use of UPLC, can be hyphenated to high-resolution mass spectrometers providing fast separation of sample components. The high-speed acquisition to enable the necessary number of data points across a chromatographic peak is well suited to Q-TOF-MS instrumentation and if optimised and developed properly, can be used to address, and resolve isomeric metabolites. Furthermore, as mentioned in Chapter 2 of this thesis, developments in TOF analysers have demonstrated higher sensitivities than Orbitraps and are superior in accurately establishing isotopic abundance patterns, which are vital for metabolite identification (Kind and Fiehn, 2006, Rousu et al., 2010). As the NPC mass spectrometry facility incorporates UPLC-Q-TOF-MS for high throughput metabolic phenotyping, the use of this analytical platform was therefore suitable for this piece of work as it can attain the high sensitivity and resolution required for metabolic measurements. However, in a publication by Rappaport *et al*. on the human blood exposome, untargeted analyses using high resolution mass spectrometry (Q-TOF-MS or Orbitrap) may miss about 90% of environmental pollutants, with concentration levels approximately 1000 times lower than endogenous metabolites (Rappaport et al., 2014). Concentrating the sample can increase the concentration levels of these metabolites marginally, as do larger sample volumes, although such approaches would not enable the required levels of these low-level metabolites to be detected. Therefore, the combination of both untargeted and targeted methods maybe needed to detect these metabolites, such as the use of triple quadrupole (TQ) instruments which can attain the necessary level of sensitivity required.

4.8 Conclusion

The untargeted nature of metabolomics allows measurement of biofluid chemistry related to both endogenous metabolism and host-environment exposures. Comprehensive coverage of chemically diverse metabolites present in human blood products benefits from the use of multiple methods, each oriented toward a small molecule subset generally segregated by polarity and hydrophobicity. Whilst recent developments in LC-MS profiling methodologies have delivered numerous solutions for the analysis of polar molecules (e.g., *via* HILIC-MS) and complex lipids, the analysis of moderately hydrophobic and amphipathic molecules in blood products by RPC methodology, is complicated by the suppressive effects of lipids on the ionisation of LMW metabolites. SPE techniques and LLE methodologies can offer a solution to remove lipophilic species, but can often be expensive, time consuming, effect recoveries on the other small molecules, and introduce contamination. Lipids can also chromatographically be separated from other small molecules, however, have been reported to accumulate on the column and elute in an unpredictable manner. DSPE therefore provides an alternative to these approaches for lipid removal from blood products, which is where the novelty lies. As demonstrated in this study, a high throughput and highly precise DSPE sample preparation technique provided a way to efficiently removing highly lipophilic species from the sample, but with minimal effect on moderately hydrophobic and amphipathic LMW compounds. This offers a solution for one of the major remaining gaps in end-to-end comprehensive metabolome coverage. As no single methodology is yet able to capture the entire plasma metabolome, our approach complements two established analytical assays (HILIC for polar metabolites and specific lipidomic assays) commonly used in the field. The approach enables the use of RPC methodology for metabolome measurement and has both the advantages of being cheaper and more robust than conventional SPE and LLE methodologies, making it a highly suitable way to study the human metabolome and xenometabolome.

Chapter 5

Exploration and characterisation of detectable xenobiotic-metabolome exposures

Summary

The novel analytical and statistical strategies for xenometabolome exploration described in Chapters 3 and 4 of this thesis were applied to two exemplar phenotyping studies; these are discussed in turn in this chapter.

The first application focused on the xenobiotic triclosan (TCS). Exploration is undertaken from molecular phenotyping data acquired from key epidemiological studies (RPC-UPLC-MS) which have been previously acquired. As exposure prevalence was relevant in this study, large datasets with matching urine-blood pairs were needed. The AIRWAVE (n=3000; plasma and urine) cohort were therefore used for this investigation, and the distribution of a feature marker relating to exposure was explored. A semi-quantitative analysis was undertaken using this marker to evaluate approximate exposure levels observed in the population. *In vitro* models were also explored to evaluate TCS metabolism. Finally, logistic regression models were used to explore features (metabolites) that best predict exposure, highlighting direct metabolites, co-exposures, and endogenous metabolites. The second application involves the xenobiotic polyethylene glycol (PEG), and utilises an intersample correlation analysis, and PLS models to find feature associations where the response (outcome) is continuous. Exploration was undertaken on the ALZ cohort using previously acquired urine samples (by RPC-UPLC-MS), and serum samples prepared using the blood lipid removal protocol developed in Chapter 4. The PEG application exploits both data-driven and analytical strategies to enhance coverage of the xenometabolome.

Aim and Objectives

The central aim of the work described in this chapter was to explore metabolism related to two exemplar xenobiotics using strategies developed in Chapter 3 and 4 of this thesis. The chapter is divided into two applications.

The first application involves the xenobiotic Triclosan (TCS). The following was undertaken:

- Acquisition of reference standards for TCS and known metabolites by current NPC profiling methodologies;
- Retrospective interrogation of broad profiling data for the large-scale assessment of TCS exposure prevalence and distribution within the UK population – Airwave (AW) cohort (urine and plasma);
- Semi-quantitative assessment of a TCS exposure marker in AW cohort;
- Logistic regression analysis to explore features (metabolites) that best predict TCS exposure in AW cohort;
- Investigate the products of TCS metabolism in an *in vitro* investigation and how it translates *in vivo* across exemplar population studies.

The second application involves the xenobiotic polyethylene glycol (PEG). The following was undertaken:

- Acquisition of the different PEG forms by RPC-UPLC-MS (positive ion mode);
- Interrogation of urine and serum (Blood lipid removal protocol developed in Chapter 4) metabolite phenotyping data for PEG signatures;
- Distinguishing between PEG-contamination and PEG-excipient, as surrogate for drug exposure in serum;
- Examining metabolism in relation to PEG exposure within and between biofluids using intersample correlation (urine and serum) and PLS models (serum only).

5.1 Introduction

Erroneous outlier signals consistent with a polymeric signature and chlorinated compounds were highlighted in Chapter 3. The presence of these signals in biofluids can reflect contamination during sample handling or a legitimate external exposure from a variety of different xenobiotics including drugs, cosmetics, environmental pollutants, and food additives. The absorption, metabolism, and excretion of xenobiotics produce metabolites that can potentially have a negative impact on human health producing unintended toxic effects. Considerable effort has been placed in drug discovery, testing and development into the disposition of individual xenobiotics on the human body, however, can often be a time-consuming and costly process (Wishart, 2016, Dickson and Gagnon, 2004, Vishwakarma and Patel, 2010). The untargeted nature of a molecular phenotyping workflows has the potential to provide direct metabolism data as well as capturing information about downstream metabolic perturbations in population-level studies. As molecular phenotyping approaches are increasingly applied to the study of large populations, it is increasingly likely that such data already exists for exploitation. The measurement of metabolites using phenotyping approaches can improve our understanding of how factors such as xenobiotic exposure, can influence the phenotype. This unique advantage of accessing population studies provides an opportunity to survey the metabolome without *a priori* analyte selection, demonstrating the power and value of untargeted molecular phenotyping in studying the xenometabolome.

5.2 Hypothesis

Using the strategies developed in Chapter 3 and 4, xenobiotice signatures can be partitioned and distinguished from endogenous signals in population studies. Strategies are knowledge driven (xenobiotic reference standard databasing), data driven (application of methodologies to statistically identify relationships between exposures and metabolic responses) and analytical driven (some physiochemical property that distinguishes between xenobiotics and endogenous metabolites). The result of these strategies will effectively increase coverage and enhance large scale human population phenotyping by providing novel insight into population level exposure and metabolism of xenobiotics. The provision of xenometabolome data will augment the established profiling methods that exist for these sample sets (which are largely focused on endogenous metabolites).

5.3 Methods

5.3.1 *Sample metabolic phenotyping*

The cohorts of interest for these two applications were the Airwave Health Monitoring study (AW) and Alzheimer's Disease Multimodal Biomarkers study (ALZ) (Elliott et al., 2014, Lovestone et al., 2009). For both studies, urine and blood samples were collected from participants in the same visit.

The AW study is an observational cohort study on British police officers (various age and sex), intended to evaluate possible health risks associated with Terrestrial Trunked Radio use. It consists of 3000 plasma and 3000 urine samples. Urine and blood products were prepared and analysed according to established protocols for blood (LIPID) and urine (RPC) phenotyping (Izzi-Engbeaya et al., 2018, Lewis et al., 2016). ALZ is a nested case-control study of Alzheimer's disease consisting of 650 urine samples and 650 serum samples.

AW and ALZ Urine samples and AW plasma samples were prepared and analysed according to established protocols for blood (LIPID) and urine (RPC) phenotyping (Izzi-Engbeaya et al., 2018, Lewis et al., 2016). The ALZ serum samples however were prepared using the RPC lipid removal protocol for blood products described in Chapter 4. The ALZ study for both biofluids were acquired as one continuous analytical batch. The data acquired for serum under negative ionisation conditions were unsuitable due to unforeseen instrument stoppages and limited sample volume. The serum samples were prepared for the second application of this chapter involving PEG, however PEG signals do not generate very stable negative ions through deprotonation, so the loss of this dataset was not a major concern. For AW however, due to the large sample size (n=3000), samples for both biofluids were acquired as three separate analytical batches. Each analytical batch consisted of 1000 study samples (excluding QC samples). Batch sizes are capped at 1000 samples as per NPC protocol for large scale phenotyping applications with certain consumables, such as columns, sample loops and syringes, kept consistent between batches (Lewis et al., 2016).

5.3.2 *Data pre-processing*

Mass spectral data files in .RAW format (Waters Corporation, USA) were converted to the open mzML format using the ProteoWizard msconvert tool (Chambers et al., 2012). During this conversion, all signals with an absolute intensity of less than 100 counts were removed.

All datasets, i.e. AW plasma (LIPID-UPLC-MS, negative ion mode) and urine (RPC-UPLC-MS, negative ion mode), ALZ urine and serum (RPC-UPLC-MS, positive and negative ion mode), were then pre-processed using the R, version 3.6.1 (R Development Core Team, 2019), package XCMS, version 3.6.1 (Smith et al., 2006). Peak detection was performed using the centwave algorithm with identical parameters as stated in Chapter 3 for the RPC ALZ study data. For the blood LIPID assay, apart from peakwidth = 3 to 12, all other parameters for this assay were identical. The overall dataset produced from the three analytical batches of the AW urine study were too large to be pre-processed together using XCMS on available computational resources. Therefore, AW urine data pre-processing was conducted in four individual processed cohorts; the first analytical batch was able to be processed as one cohort, i.e. AW-B1; the second analytical batch had to be processed as two smaller cohorts due to its size, i.e. AW-B2.1 and AW-B2.2, and the third analytical batch was processed as the fourth cohort, i.e. AW-B3.

The size of the spectral datasets for AW plasma, ALZ urine and ALZ serum were sufficiently small to be pre-processed as single batches. Following the pre-processing steps in XCMS, data matrices were imported into and filtered using the nPYc-Toolbox (Sands et al., 2019), running in the Python environment (Python Software Foundation. Python Language Reference, version 3.5 and above. Available at http://www.python.org).

A pooled QC samples, or Study reference (SR), was used for noise filtering purposes and signal drift correction, as previously mentioned in Chapter 4. Briefly, filtering was based on features which had passed several criteria. Firstly, all features which are included must not have a relative standard deviation (RSD) exceeding 30% from repeated extractions of the SR, which were systematically injected throughout the analytical run. The features passing this initial step, whose Pearson correlation with dilution of the SR, as estimated from the serial dilution series, was less than 0.7 or, having a residual standard deviation (RSD) in the study samples smaller than 1.1 times the value estimated from the SR, were subsequently removed. These SR were also systematically injected throughout the analytical run, and LOESS regression applied, to correct for any signal drift. The SR sample was prepared by pooling together an aliquot of all samples in a study and is a representation of the physical average. The quality control processes described in Lewis *et al.*, including the preparation of the SR sample and implementation of a dilution series for filtering purposes has already been previously described and established (Lewis et al., 2016, Dona et al., 2014). A further normalization step was implemented on the urine samples, using PQN to adjust for urinary sample dilution on global sample intensity (Dieterle et al., 2006). Both blood and urine datasets were also log transformed (log base 10) to reduce the impact of outliers causing highly skewed feature distributions, thereby converting data to be more normally distributed.

5.3.3 Metabolite Identification (MetID)

Initial steps for MetID efforts required an intrasample and intersample correlation analysis (Chapter 4) to be undertaken for the unknown feature from a representative sample, i.e. a sample exhibiting the highest signal of the unknown feature in the study. This allowed the molecular ion to be identified and can reveal additional information on structurally and biologically related metabolites.

The representative sample underwent a series of MS/MS experiments, where the target molecular ion was subjected to five different static collision energies (5V, 10V, 20V, 30V and 40V). The acquisitions at the five collision energies were undertaken as means to obtain the accurate mass of fragment ions, and to identify the correct fragment ions, i.e. as the collision energy increases, the intensity of the parent molecular ion will decrease, and the intensity of the correct fragment ions will increase. Putative annotations were therefore made by comparison of the accurate mass of the unknown feature (molecular ion and fragment ions), to in-house and online spectral databases such as PubChem (Kim et al., 2018), HMDB (Wishart et al., 2017) and Metlin (Guijas et al., 2018). Identification of metabolites were confirmed based on matching candidate metabolites from the putative annotations to purchased referenced standards (if commercially available) or to the chromatographic and spectral data (MS) from an in-house reference standard database using a narrow *m/z* window (1-5mDA) and retention time range (0.02 minutes). Exact definitions for putative annotation and identification can be found in Chapter 2 in the Metabolite identification section. Additional MS/MS on the reference standard were further carried out and compared to the fragmentation pattern of the unknown feature. For a successful identification (MSI level 1), both the retention time value and the fragmentation data acquired for the unknown feature and the reference standard must be identical. When reference standards were not available, two other techniques were implemented to synthesise the desired standard, i.e. a sulfation protocol (for sulfate conjugates), which was adapted from the protocol described by Sarafian *et al* for bile acids (Sarafian et al., 2015), and enzymatic hydrolysis (for glucuronide conjugates). For both techniques, the candidate unmetabolised compound was purchased and then subjected to either the sulfation protocol or analysed in parallel with the unconjugated metabolite obtained upon the enzymatic hydrolysis of the glucuronide conjugated metabolites.

5.3.3.1 Sulfation

A 20 µL of a 1 mg/mL reference solution was firstly added to a 500 µL mixture of sulfur trioxide – pyridine complex (50mg) and sodium sulfate (5mg) in 10 mL chloroform. The solution was evaporated to dryness by using nitrogen flow at 55 °C and then resuspended in 100 µL of water for RPC-UPLC-MS analysis.

5.3.3.2 Enzymatic hydrolysis

This protocol combines an enzymatic hydrolysis sample preparation method, and a neutral loss (NL) MS acquisition. Prepared samples were acquired using RPC-UPLC-MS chromatographic conditions, but with NL MS acquisition. The protocol used the enzyme β-glucuronidase from different biological sources to cleave the glucuronide moiety from the unknown conjugated metabolite, leaving only the unconjugated parent in the solution, the structure of which could then be then ascertained by confirmation with an analytical reference standard. Briefly, a 90 µL representative sample was subjected to a hydrolysis incubation followed by protein precipitation (**Figure 5-1**). Reference standards for the unconjugated forms were available for purchase and acquired by reversed phase profiling (RPC-UPLC-MS) (Lewis et al., 2016) using the method developed for reference standard acquisition described in detail in Chapter 3.

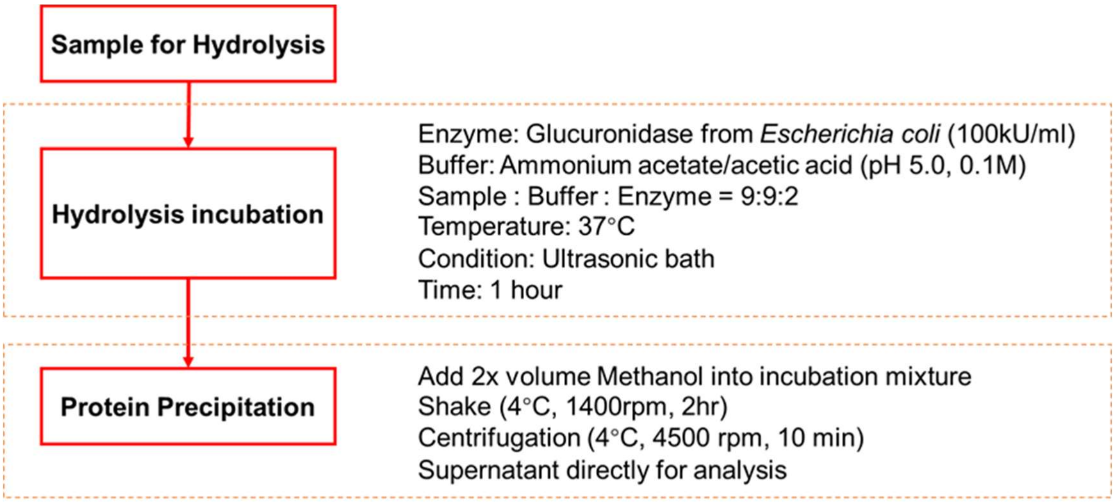

Sample for Hydrolysis
Hydrolysis incubation
Enzyme: Glucuronidase from *Escherichia coli* (100kU/ml)
Buffer: Ammonium acetate/acetic acid (pH 5.0, 0.1M)
Sample : Buffer : Enzyme = 9:9:2
Temperature: 37°C
Condition: Ultrasonic bath
Time: 1 hour
Protein Precipitation
Add 2x volume Methanol into incubation mixture
Shake (4°C, 1400rpm, 2hr)
Centrifugation (4°C, 4500 rpm, 10 min)
Supernatant directly for analysis

Figure 5-1. Sample preparation protocol for enzymatic hydrolysis.

5.4 Application 1: Triclosan

5.4.1 Introduction

Endogenous halogenated compounds are rare in humans and its presence in biofluids can be good indicators of environmental exposures (Gribble, 2003). The incorporation of halogens into pharmaceutical formulations is common, often as a means to improve and enhance certain properties of the medication, i.e. facilitate the active drug molecules across biological barriers, fill hydrophobic pockets within protein targets, prolong medication lifetime and allow for easy adsorption (Wagner et al., 2009, Rahman et al., 2016). Importantly, the halogen atoms have distinctive patterns of isotopic abundance and therefore the analysis of halogenated compounds by mass spectrometry can make use of the resulting spectral signatures that can be readily identify on account of these their unique and characteristic isotopic distributions. For example, the antibiotic flucloxacillin that was identified in Chapter 3, contains both a fluorine and a chlorine atom. While fluorine-19 is 100% abundant, chlorine has two stable isotopes: ^{35}Cl and ^{37}Cl, with relative abundances in the ratio 3:1. The distinctive isotope pattern of flucloxacillin and its chlorine containing metabolites can be readily identified in mass spectra alongside endogenous metabolites.

Triclosan (TCS), is a chlorinated anti-microbial chemical. Originally confined to use in hospital environments for pre-operative skin preparation and post-surgical sutures (Leaper et al., 2011, Leaper et al., 2010), TCS was soon after introduced into US and European populations *via* incorporation into antiseptic over the counter consumer products (Bhargava and Leonard, 1996, Jones et al., 2000) with concentrations in the range of 0.1-0.3% (w/w) (Pannek and Vestweber, 2011, Dhillon et al., 2015). As a consequence of its widespread use, TCS has now been shown to contaminate water supplies, furthering population-level exposure to the chemical (Lindström et al., 2002, Lopez-Avila and Hites, 1980, McAvoy et al., 2002, Okumura and Nishikawa, 1996, Singer et al., 2002, Lehutso et al., 2017, Dhillon et al., 2015).

Phenolic xenobiotics generally undergo rapid metabolism and detoxification to glucuronide and sulfate conjugates by phase II metabolic enzyme. To date, phase II conjugates and oxidative metabolites of TCS has been reported in a diverse number of species, *in vitro* (Hanioka et al., 1996, Moss et al., 2000) and *in vivo* (Moss et al., 2000), with the major excretion product being the glucuronide in urine samples (Fang et al., 2016, Fang et al., 2010). More recently a new metabolism route was discovered, reporting a TCS dimer in microsomal extracts, but only in the absence of phase II enzymes (Ashrap et al., 2017). These reported metabolites are summarised in **Figure 5-2**. Exposure to TCS in can occur *via* absorption through the oral cavities (Lin, 2000), gastrointestinal tract (Sandborgh-Englund et al., 2006) and dermal exposure (Moss et al., 2000). Its occurrence has been reported in plasma, urine and breast milk (Allmyr et al., 2006, Li et al., 2013). TCS is estimated to have a half-life of approximately 8 hours with the majority being eliminated through the urine in the course of 24hrs (Sandborgh-Englund et al., 2006). Baseline levels are reached within 8 days of exposure with evidence that daily exposure has bioacculumative effects (Mustafa et al., 2003).

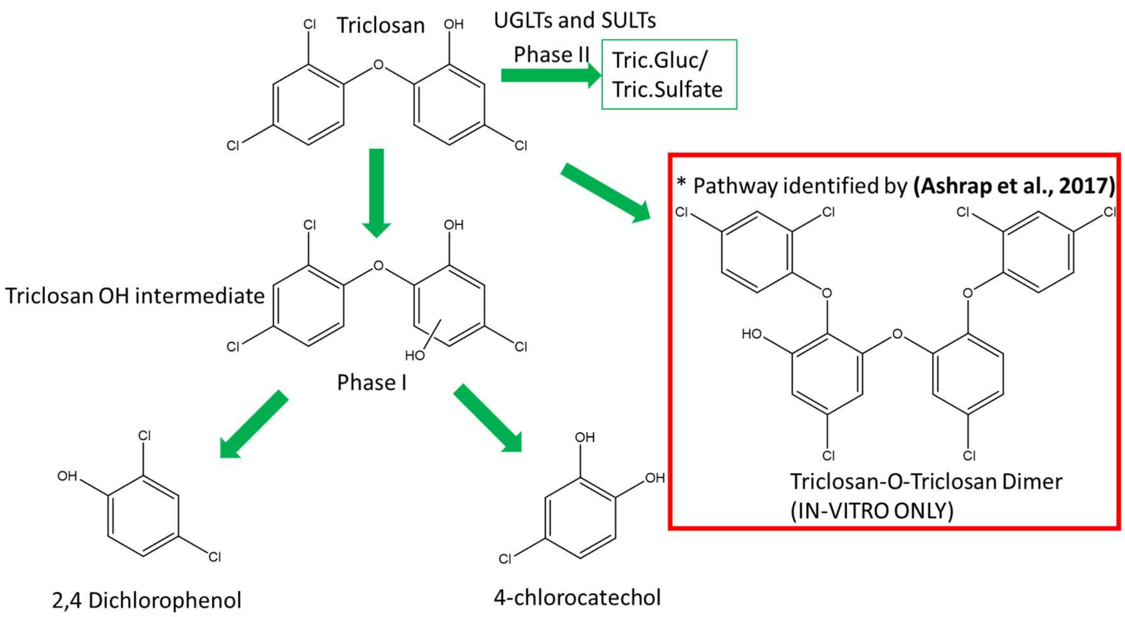


Figure 5-2. ***in vivo*** **and** ***in vitro*** **phase I and phase II metabolism of TCS** (Fang et al., 2010).

In the recent decade, the efficacy of TCS has been repeatedly brought into question (Fang et al., 2010), and in September 2016, the US food and drug administration (FDA), investigated and banned the use of antibacterial soaps containing TCS and similar agents citing a failure to produce evidence of effectiveness and health benefit (Aiello et al., 2007). This decision has emerged due to increasing concerns regarding anti-microbial resistance (Carey and McNamara, 2015, Braoudaki and Hilton, 2004), TCS-induced disruption to the endocrine system responsible for reproductive and developmental functions in mammalian systems (Chen et al., 2008), shifts in the microbiome in animal studies (Gaulke et al., 2016, Lawrence et al., 2015) and mitochondrial uncoupling in living organisms (Shim et al., 2016).

Notwithstanding the mounting concern related to use of TCS in consumer products, such products remain in great abundance, with 450 tonnes still used for domestic uses within the EU annually (Halden et al., 2017). At the time of writing, TCS has not been investigated by the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK. Furthermore, despite its ubiquity as an environmental exposure, no population-level study on human samples has been performed to assess TCS exposure in UK cohorts and/or access the effect that low-level TCS exposure may have on endogenous metabolism and/or disease risk.

TCS was therefore selected as an exemplar to assess the prevalence in UK cohorts, and to implement a statistical strategy as discussed in Chapter 3, to characterise the xenometabolome in relation to exposure.

5.4.2 *Materials and additional methods*

5.4.2.1 Reference standard preparation and preliminary assessment

All TCS-related metabolites that have been previously reported in the literature and that were commercially available, were purchased. The chemical standards: TCS, TCS O-β-D-Glucuronide Sodium Salt (TCS-Gluc) and TCS O-Sulfate (TCS-SO4) were purchased from Toronto Research Chemicals Inc. 2, 4 Dichlorophenol and 4-chlorocaetchol purchased from SIGMA-ALDRICH. Compounds were subjected to the acquisition workflow for reference standards as described in Chapter 3, and the 1:100 dilution standard was analysed by RPC-UPLC-MS (in both positive and negative ion mode). The preparation of standards undertaken in Chapter 3 are however, for RPC based assays, and therefore made to volume in water. A 50 µL aliquot of each standard were individually mixed with 200 µL of isopropanol to make them compatible with the LIPID assay. Solutions were then acquired using the RPC and LIPID profiling methods. For each standard (as discussed in Chapter 3), data was acquired under both electrospray positive (ESI+) and negative (ESI-) ionisation conditions, in continuum mode with 2 injections per well. One injection was at a low collision energy (4V) and the other utilising a collision energy ramp (10-30V). Within each injection, three interleaved full MS scans (0.05 second scan rate) were acquired for the *m/z* range between 50 and 1200 Da.

5.4.2.2 In vitro incubations of TCS with human hepatocytes

Metabolic profiling was undertaken on human liver samples prepared by Sygnature Discovery and analysed by UPLC-MS (negative ion mode) using in-house methodologies. Samples were prepared from an incubation of TCS (10 µM final concentration) with cryopreserved human hepatocytes (0.5 million cells/mL). Samples were taken at 0, 10 and 60 minutes and were compared with blank samples of hepatocytes to identify compound related peaks. All detectable metabolites were identified by UPLC-MS (negative ion mode). A TCS dimer has been reported in the literature (Ashrap, Zheng et al. 2017) and so in addition to the above analysis, a further incubation in human liver microsomes was carried out to see if the dimer was formed in the absence of phase II metabolism. The incubation was carried out at both 10 µM and 50 µM to see if an increased concentration resulted in the formation of dimer. Sample extracts were analysed under the same UPLC-MS conditions. Assignment of TCS and metabolites were based by matching the theoretical accurate mass of the molecular ion, to peaks consistent with the EIC observed in the samples. There was no further spectral information provided. Reports from both analyses detailing the results were kindly provided by Sygnature Discovery. Further details on the experimental details and instrument conditions can be found in Appendix 3.

Finally, the hepatocyte and microsomal extracts analysed by Sygnature Discovery were sampled in 200 µL UPLC-MS vials and shipped to the phenome centre. These samples were then acquired using RPC-UPLC-MS (negative ion mode) and LIPID-UPLC-MS (negative ion mode) profiling methodologies for method specific retention times and spectral data.

5.4.2.3 Exposure prevalence and distributions of TCS in urine phenotyping data (AW)

5.4.2.3.1 Exposure prevalence

The *m/z* and retention time of TCS and metabolites, recorded from the reference standards, for each profiling assay, were then screened for in the corresponding profiling data for urine AW. Prevalence in the population was assessed using a specific metabolite as a proxy for TCS exposure. The LOD reported in **5.4.2.4** below, is not appropriate to apply to retrospective data as calibration curves were not run with the original AW profiling data. Prevalence of TCS exposure in AW was therefore evaluated differently. For a sample to be classified as detected, or “exposed”, the following four criteria had to be fulfilled:

1.) A retention time and molecular ion [M-H] match against a reference standard. The *m/z* of the ion must be within 3ppm and the retention time of the chromatographic peak within 0.02 minutes between the reference standard and that observed in phenotyping data;
2.) A minimum Signal to noise ratio (S/NR) ≥ 5 in the correct elution region. This equated to an approximate peak area ≥ 25000 arbitrary units (or approximately 4.3 arbitrary units on a log10 scale);
3.) The detection of the unique isotopic distribution in the mass spectrum, which is characteristic of chlorinated compounds. A molecule containing a chlorine atom will produce two molecular ion peaks in a mass spectrum, i.e. M^+ and M+2. The chlorine atom can exist as

two major isotopes, 35Cl and 37Cl with relative abundances 3:1 respectively. TCS-Gluc contains 3 three chlorine atoms resulting in 4 isotopic combinations in the ratio 27:27:9:1. This isotopic pattern had to be present, and the two main isotopes had to be approximately the same in relative intensity;

4.) In addition to the M$^+$ ion, an in-source fragment ion of *m/z* 286.9425, which represents the loss of a conjugation moiety, had to also be present with a similar isotopic cluster as stated in 3.).

5.4.2.3.2 Implementation of Gaussian mixture models

As multimodal distributions of TCS exposure exists in the data, multi-component Gaussian mixture models (GMMs) were specified, placing clusters across the distributions (as explained in Chapter 3). Once fitted, conversion of the distributions to probability distribution functions (PDF's), were calculated, dividing the data into three main exposure groups; a zero, low-mid range and high group. The PDF's for each of these Gaussians were obtained and any sample with a probability (pr_n) of more than 0.90, assumed the classification for that group. The samples that occupied the "zero" group demonstrated no evidence of TCS, based on the criteria for exposure prevalence listed previously.

5.4.2.4 Semi-quantitative screening of TCS-Gluc

The approximate concentration range observed in the AW urine dataset was estimated. Calibration curves were firstly prepared in triplicate to access the limit of detection (LOD), limit of quantitation (LOQ) and linear range. Two separate curves were constructed by spiking TCS-Gluc into water and blank urine (i.e. urine free of TCS-Gluc). The concentration ranged from 0.1 µg/mL to 5000 µg/mL. Like the study samples, a dilution of 1:1 (v/v) was undertaken where the diluent was water with a spiked concentration of an isotope-labelled internal standard (IS), L-Phenylalanine-$^{13}C_9$,^{15}N at 50 µg/mL. For each curve, the ratio of the TCS-Gluc area to the IS area is plotted against the measured concentration with no weighting factor applied. Standard concentrations were then back extrapolated from the calibration curves. The LOQ was established as the concentration five times the LOD. The concentration of the AIRWAVE SR sample and a secondary developmental set urine, Devset-urine (same set used in Chapter 3), was measured using both calibration curves and prepared in six replicates to assess precision. Precision is defined as the closeness of six individual preparations of the QC samples, and the percentage relative standard deviation (RSD) reported. It is precise if the RSD is less than 15% and reported levels are above the LOQ. An additional curve was also prepared *via* a standard addition method using the AIRWAVE SR. The standard addition method allowed the quantitative measurement of existing TCS-Gluc present in the SR, and at the same time, minimise the matrix effects that would have otherwise interfered with analyte measurement signals. This allowed a comparison of the SR measured from to the two calibration curves, to the standard addition measurement. The standard addition curve was prepared by aliquoting equal volumes of TCS-Gluc reference standards held at different concentrations ranging from 20 µg/mL to 0.2 µg/mL, into an aqueous diluent, and into the aqueous diluent, whist keeping the sample volume, overall volume and IS concentration constant. Finally, a random subset of samples selected from AW, which were classified as either zero (n=10), mid-range (n=5) or high (n=10), as described in the exposure prevalence section above, were additionally acquired and approximate concentration ranges for each group reported. All developmental, QC and study samples were prepared with the same dilution factor as the original AW study.

5.4.2.5 Metabolite associations from TCS exposure

To highlight the features with the strongest associations to TCS exposure, LogReg models were individually applied to each cohort. LogReg models binary response variables by fitting a regression curve between two groups of samples, in this case, the zero and high TCS exposure groups. The application of LogReg is applied in an almost identical manner to that observed in Chapter 3. Briefly, the zero-exposure group was defined as 0 (control) and the high exposure group defined as 1 (case). The data was then split into a training and test set, where selection of the statistically significant features was conducted on the training set and validated in the test set. It should be noted that prior to the splitting of the data, all TCS related variables (adduct, isotopes and in-source fragments) were removed. These variables would be predicted perfectly as it is essentially predicting itself.

Two Multivariate regularised LogReg models, Ridge and Elastic Net (EN), were firstly explored to see how features together relate to TCS exposure. LASSO was not explored as it is considered too stringent in feature selection, as observed in Chapter 3.

A univariate method was used to estimate the individual contribution of each feature. The exact details of the procedure used for data partitioning, tuning parameters for the multivariate models and, and feature selection (based on statistical significance and bootstrapping of the regression coefficients), can be found in the LogReg application in Chapter 3.

LogReg was carried out on each of the four cohorts of the AW urine study. . A decision was made to also include a fifth cohort, ALZ urine dataset (RPC-UPLC-MS, negative ion mode). Features selected for the final model had to be significant from at least two out of the four cohorts for AW and be significant in the ALZ cohort. This way there was no issue with the cohorts being unbalanced. EN has implicit feature selection and imposes a penalty similar to both Ridge and LASSO by effectively shrinking regression coefficients (Ridge) or sets them as zero (LASSO). In the univariate model, feature selection was based on a cut-off of p_{adj}<0.05. As there were instances where features were statistically significant in some cohorts but not all, in order to assess all cohorts together and therefore achieve a higher statistical power, a meta-analysis was performed to combine p-values (p_{adj}). Fisher's combination test (Fisher, 1992), Stouffer's method (Stouffer et al., 1949) and Stouffer's method with weights (Zaykin, 2011) are common approaches for combining p -values (Vaitsiakhovich et al., 2014). This Stouffer method with weights has been implemented in several different metabolomic applications (Kaever et al., 2014, Laíns et al., 2019, Whitlock, 2005) and is applicable in this circumstance as all cohorts share the same experimental design and the combined p-values are from multiple tests of the same hypothesis. It is summarised by the following three equations:

$$w_j = \sqrt{n_j}, j = 1, \ldots k^{th} \tag{5. 1}$$

$$\mathrm{Z} = \sum_{j=1}^{k} w_j \varphi^{-1}(1 - p_j) / \sqrt{\sum_{j=1}^{k} w_j^2} \tag{5. 2}$$

$$p_{\mathrm{META}} = 1 - \varphi(\mathrm{Z}) \tag{5. 3}$$

Where:

n is the number of observations for the j^{th} cohort to the k^{th} cohort.

w_j is cohort specific weight

Z is the z-score

φ *and* φ^{-1} is normal standard cumulative distribution and its inverse

p_j cohort specific p-value

The p_{adj}-values from all cohorts were converted to z-scores which are weighted based on their sample sizes. In any of the cohorts, if a feature was statistically significant (FDR; $p_{adj} \leq 0.05$), the corresponding features were found in the remaining datasets, and its p_{adj}-value was used in the Stouffer method equation to produce a combined p_{adj}-value (p_{adj}meta). Features with a p_{adj}meta <0.05 are included in the final model. As these p_{adj}-values are combined for the meta-analysis, the samples used to train the models for the five cohorts were also combined. Application of the combined training set model, using only the statistically significant features from the meta-analysis, was then applied to the combined test set. The calculation of the AUC (area under the curve) from a ROC curve, which is a typical performance parameter for binary classifier measurements (Chapter 2), was assessed. The optimum cut-off point was defined as that which maximized the AUC value, as indicated by the minimum distance to the top-left corner of a ROC curve plot. A summary of the steps carried out for LogReg is illustrated in **Figure 5-3**.

5.4.2.6 Blood samples

The evaluation of TCS exposure in blood products was not as comprehensive as urine. Exposure was examined using only plasma samples from the AW study. Blood samples from the ALZ dataset were not available as data quality was compromised. Nevertheless, the distribution of a TCS exposure marker was firstly evaluated, then both univariate and multivariate LogReg models were subsequently explored using samples from the AW cohort. Furthermore, AW had matching urine and plasma samples collected during the same visit. Assessment of the correlation between the intensity profile of the molecular ion of a TCS metabolite in urine and in blood, was therefore undertaken to determine if a significant relationship existed between circulating plasma and excreted urine.

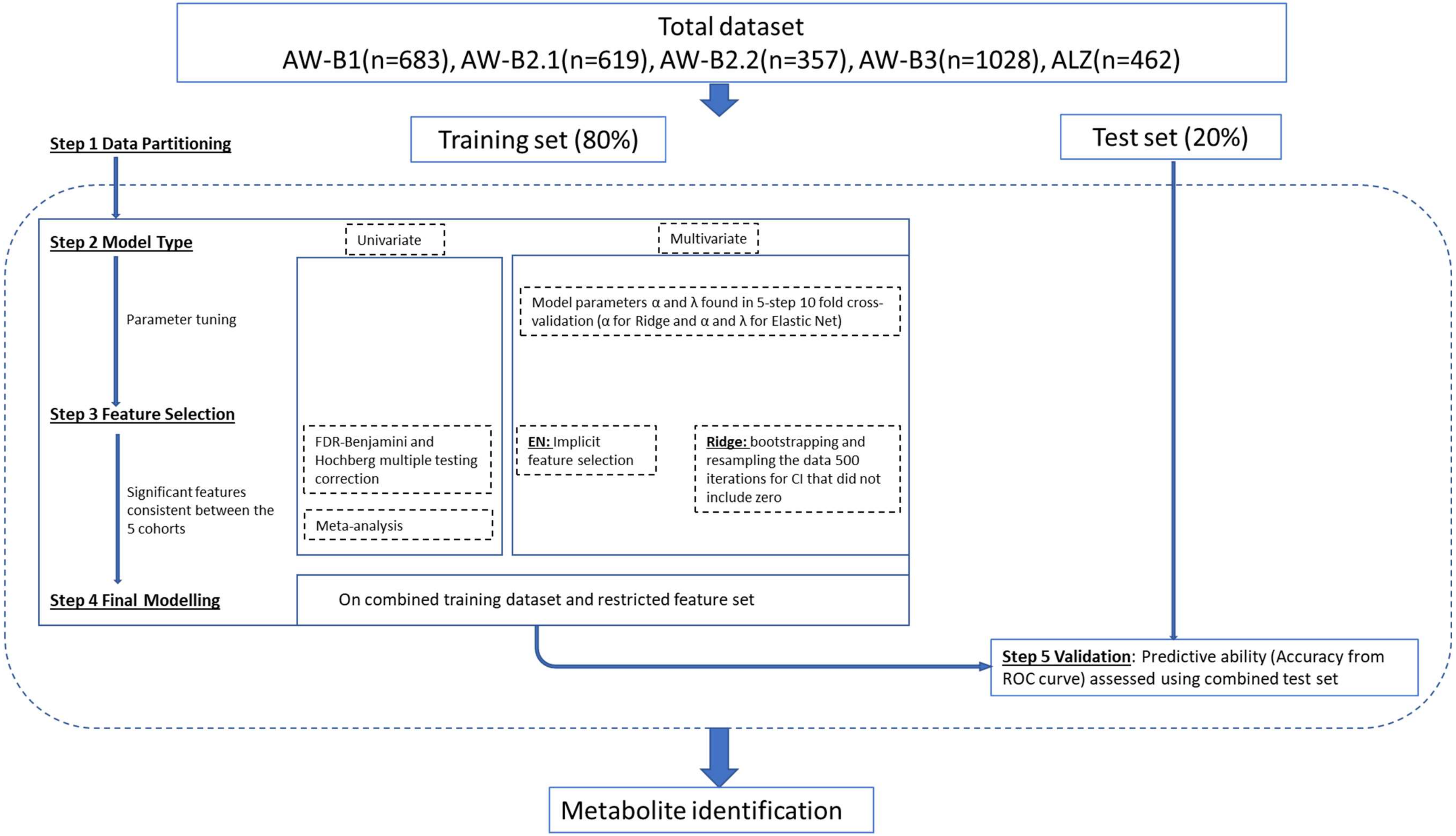

Total dataset
AW-B1(n=683), AW-B2.1(n=619), AW-B2.2(n=357), AW-B3(n=1028), ALZ(n=462)
Step 1 Data Partitioning
Training set (80%)
Test set (20%)
Step 2 Model Type
Univariate
Multivariate
Parameter tuning
Model parameters α and λ found in 5-step 10 fold cross-validation (α for Ridge and α and λ for Elastic Net)
Step 3 Feature Selection
FDR-Benjamini and Hochberg multiple testing correction
EN: Implicit feature selection
Ridge: bootstrapping and resampling the data 500 iterations for CI that did not include zero
Significant features consistent between the 5 cohorts
Meta-analysis
Step 4 Final Modelling
On combined training dataset and restricted feature set
Step 5 Validation: Predictive ability (Accuracy from ROC curve) assessed using combined test set
Metabolite identification

Figure 5-3. A summary of the steps involved in univariate and multivariate (Ridge and Elastic net regularisation) logistic regression models to assess metabolite associations in relation to TCS exposure in RPC-UPLC-MS (negative ion mode) datasets from AW and ALZ studies

5.4.3 *Results and discussion*

5.4.3.1 Preliminary assessment

Reference standards for TCS, TCS-Gluc, TCS-SO4, 2, 4 dichlorophenol and 4-chlorocatechol were only identified by RPC-UPLC-MS and LIPID-UPLC-MS instrumentation, in negative ion mode. The retention time and mass spectrum for the standard acquired using a collision energy ramp (as discussed in Chapter 3 for reference standard acquisitions), are summarised in **Figure 5-4**. Screening of these compounds in XCMS outputs for AW, revealed detection of TCS-Gluc in urine, and TCS-SO4 in blood. The molecular ion corresponding to TCS-Gluc (*m/z* 462.9754) was therefore used as a proxy for TCS exposure for all subsequent analyses involving urine and similarly, the molecular ion corresponding to TCS-SO4 (*m/z* 366.9002) was used as a proxy for TCS exposure for all analyses involving plasma.

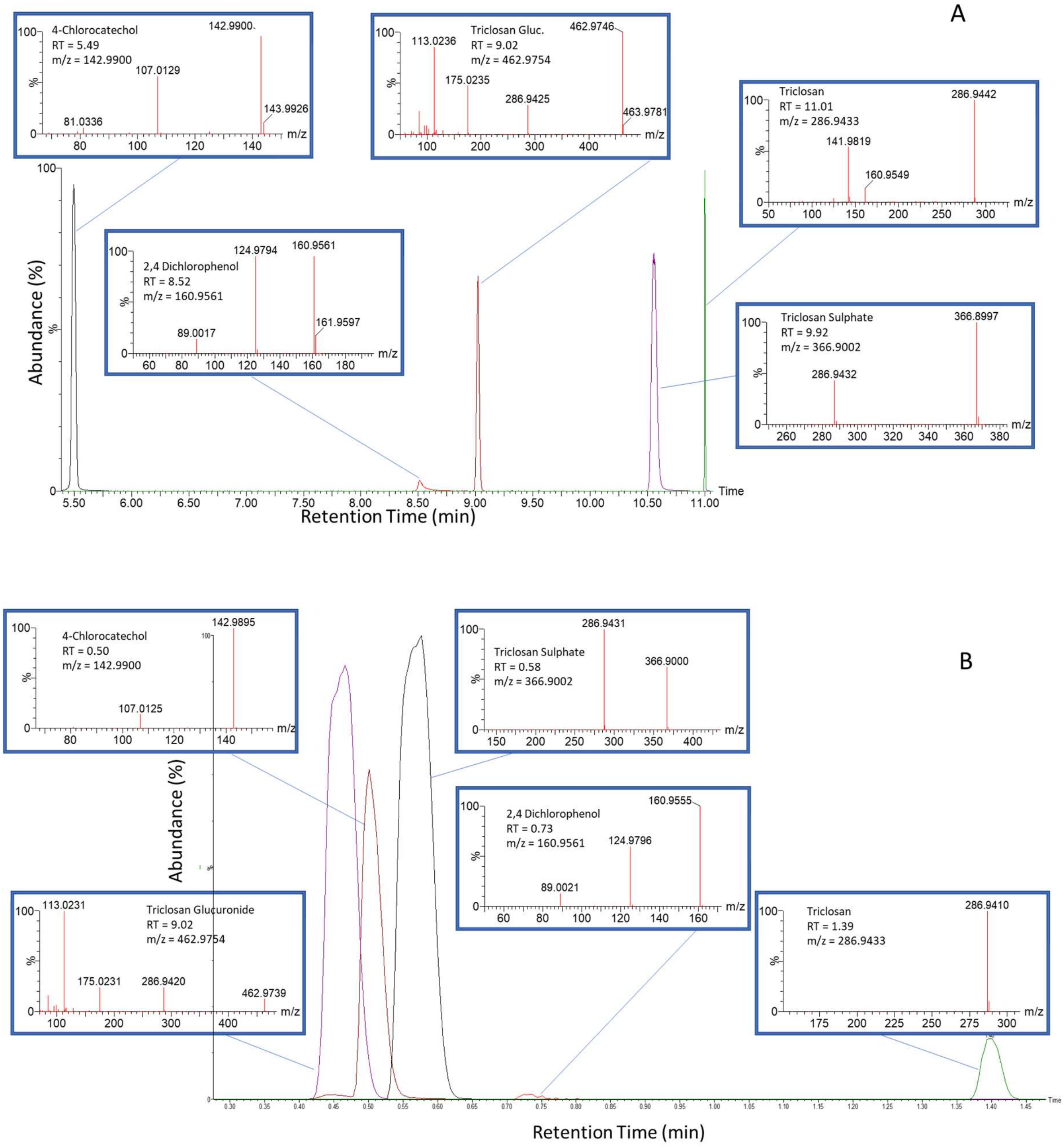


Figure 5-4. Retention time and mass spectrum for TCS and reported metabolites acquired by RPC-UPLC-MS (negative ion mode) and LIPID-UPLC-MS (negative ion mode). (A) Retention time and mass spectrum of TCS and metabolites (TCS-Gluc, TCS-SO4, 4-Chlorocatechol and 2,4Dichlorphenol) acquired by RPC-UPLC-MS (negative ion mode) using a collision energy ramp; (B) Retention time and mass spectrum of TCS and metabolites (TCS-Gluc, TCS-SO4, 4-Chlorocatechol and 2,4Dichlorphenol) acquired by LIPID-UPLC-MS (negative ion mode) using a collision energy ramp; The RPC and LIPID profiling methods were successful in detecting TCS and metabolites, which allowed for easy screening of the same signatures in the AW dataset. TCS-Gluc was detected in AW urine RPC-UPLC-MS (negative ion mode) and TCS-SO4 was detected in AW blood LIPID-UPLC-MS (negative ion mode).

5.4.3.2 Assessing exposure prevalence and distributions of TCS-Gluc

Using the criterion stated in the methods, the prevalence of exposure was studied. As there are 4 pre-processed cohorts within AW, the prevalence was measured 4 times and summed, resulting in approximately 28% TCS-Gluc exposure. Using this information, the distribution of TCS-Gluc was evaluated in all four cohorts. It was clear that a multimodal distribution exists in the data. The PDF's for each gaussian were obtained, dividing the data into a High, Low-Mid and Zero exposure groups as illustrated in the density plot for AW-B3 in **Figure 5-5**. The distribution was also evaluated for the ALZ cohort. Density plots for each cohort (AW-B1, AW-B2.1, AW-B2.3, and ALZ) can be found in Appendix 3. The samples occupying the high and zero groups for all cohorts (four from AW and 1 from ALZ), were then used for logistic regression.

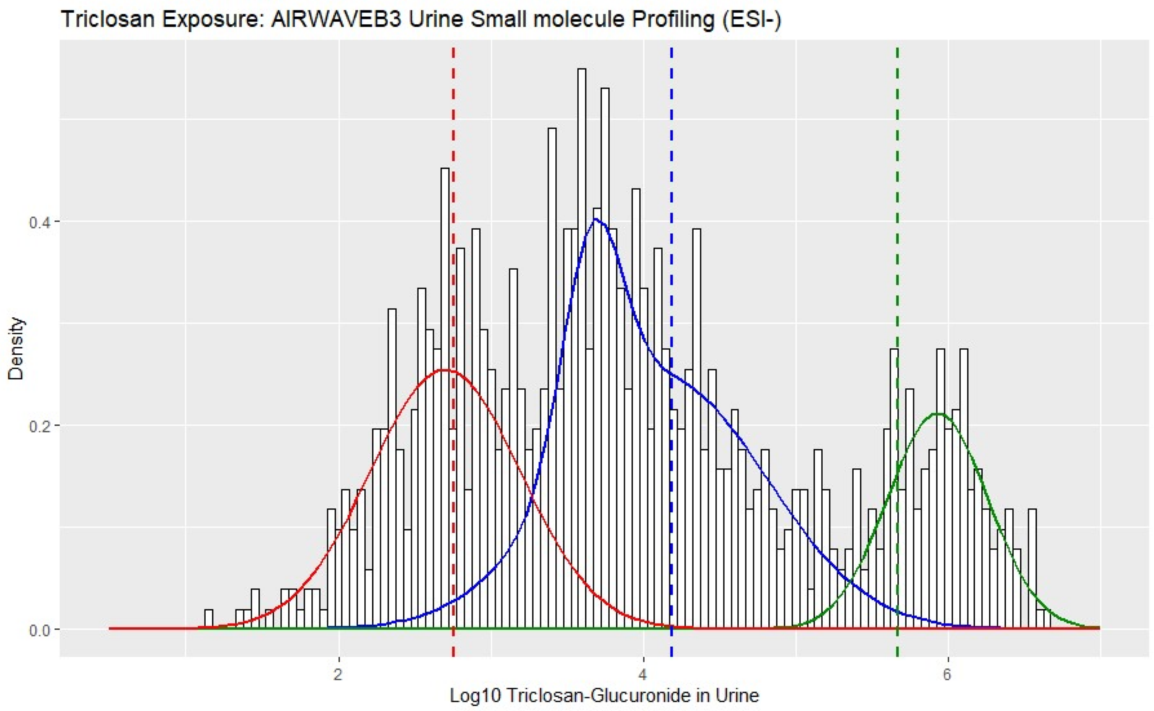


Figure 5-5. Gaussian mixture models (GMM's) fitted to the MS intensity distribution of TCS-Gluc from AW-B3 urine RPC-UPLC-MS (negative ion mode) data. GMM's were fitted to AW-B3, and the PDF's for each gaussians were obtained, dividing the data to a High exposure group (Distribution 2, pr_3 – green), Low-Mid exposure group (Distribution 2, pr_2 – blue) and a Zero exposure group, (Distribution 1, pr_1 -red). Any sample with pr_1>0.90, or a log10 signal less than the red dotted line, assumed the classification of zero exposure, and any sample with pr_3>0.90, or log10 signal more than the dotted green line, assumed the classification of high exposure. The blue dotted line is equivalent to a signal that fulfils all three criteria as stated in **sectio**n **5.4.2.3.1.**

5.4.3.3 *In vitro incubations of TCS with human hepatocytes*

The observed ions for TCS and metabolites reported by Sygnature Discovery, were then assessed by analysing the same extracts by RPC-UPLC-MS (negative ion mode) and LIPID-UPLC-MS (negative ion mode). The observed metabolites at 60 minutes incubations, are summarised in **Table 5-1.**

Table 5-1. Observed metabolites from the incubation of TCS with human hepatocytes analysed by RPC-UPLC-MS (negative ion mode) and LIPID-UPLC-MS (negative ion mode).

Name	RPC-UPLC-MS (RT min)	LIPID-UPLC-MS (RT min)	Observed ion $[M-H]^-$	Formula $[M-H]^-$
TCS-unmetabolised	11.01	1.37	286.9433	$C_{12}H_6Cl_3O_2^-$
TCS -Oxidised Glucuronide	6.79 & 6.81	Not detected	478.9703	$C_{18}H_{14}Cl_3O_9^-$
TCS – Oxidised Sulfate	7.78 & 7.95	0.42	382.8951	$C_{12}H_6Cl_3O_6S^-$
TCS-Gluc	9.04	0.47	462.9754	$C_{18}H_{14}Cl_3O_8^-$
TCS-SO_4	10.48	0.58	366.9002	$C_{12}H_6Cl_3O_5S^-$

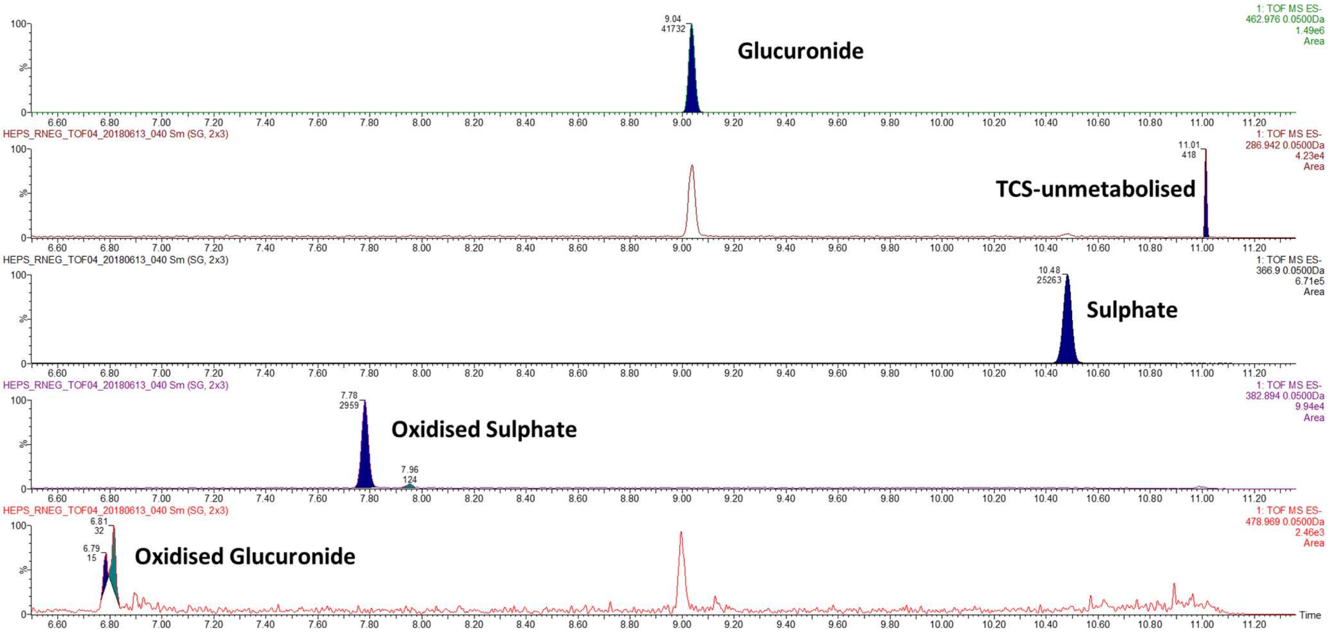


Figure 5-6. Extracted mass chromatograms (EIC) for TCS-unmetabolised, and metabolites, from the incubation of TCS with human hepatocytes at 60 minutes acquired by RPC-UPLC-MS (negative ion mode).

Metabolites observed in the RPC assay (**Figure 5-6**) were predominantly sulfate and glucuronide conjugates, with lower levels of oxidised sulfate, and oxidised glucuronide conjugates. However, the dimer reported by Asphrap *et al*. (2017), was not detected. According to the publication, the dimer was observed in microsomal extracts. Microsomal incubations provide by Sygnature Discovery were analysed by LIPID-UPLC-MS (negative ion mode), due to the lipophilic nature of the dimer. The EIC of the observed ions provided by Sygnature Discovery in their report, matched two peaks corresponding to hydroxylated metabolites of TCS in the LIPID-UPLC-MS profile, and five peaks corresponding to the dimer (summarised in **Table 5-2** and **Figure 5-7**), highlighting that multiple different isomeric forms exist for these two metabolites. TCS, TCS-Gluc and TCS-SO4 assigned in the in vitro extracts are considered identifications, as reference standards acquired by the sample analytical platform were conducted for these compounds. Although comparison of spectral data is not made with a public library, comparison of the molecular ion of all other TCS metabolites, to the assignments made by Sygnature Discovery would be sufficient to constitute a level two putative annotation (MSI level 2). The complex isotopic pattern observed in the mass spectrum for each annotated metabolite is consistent with the presence of multiple chlorine atoms, which also gives additional confidence to their assignment.

Table 5-2. Observed metabolites from microsomal incubations of TCS analysed by LIPID-UPLC-MS (negative ion mode).

Name	LIPID-UPLC-MS (min)	Observed ion $[M-H]^-$	Formula $[M-H]^-$
TCS-unmetabolised	1.37	286.9433	$C_{12}H_6Cl_3O_2^-$
TCS Hydroxylated isomer 1	0.71	302.9383	$C_{12}H_6Cl_3O_3^-$
TCS Hydroxylated isomer 2	1.05	302.9383	$C_{12}H_6Cl_3O_3^-$
TCS Dimer 1	2.07	572.8789	$C_{24}H_{11}Cl_6O_4^-$
TCS Dimer 2	2.44	572.8789	$C_{24}H_{11}Cl_6O_4^-$
TCS Dimer 3	2.67	572.8789	$C_{24}H_{11}Cl_6O_4^-$
TCS Dimer 4	2.88	572.8789	$C_{24}H_{11}Cl_6O_4^-$
TCS Dimer 5	3.17	572.8789	$C_{24}H_{11}Cl_6O_4^-$
TCS Dimer 6	Not detected	572.8789	$C_{24}H_{11}Cl_6O_4^-$

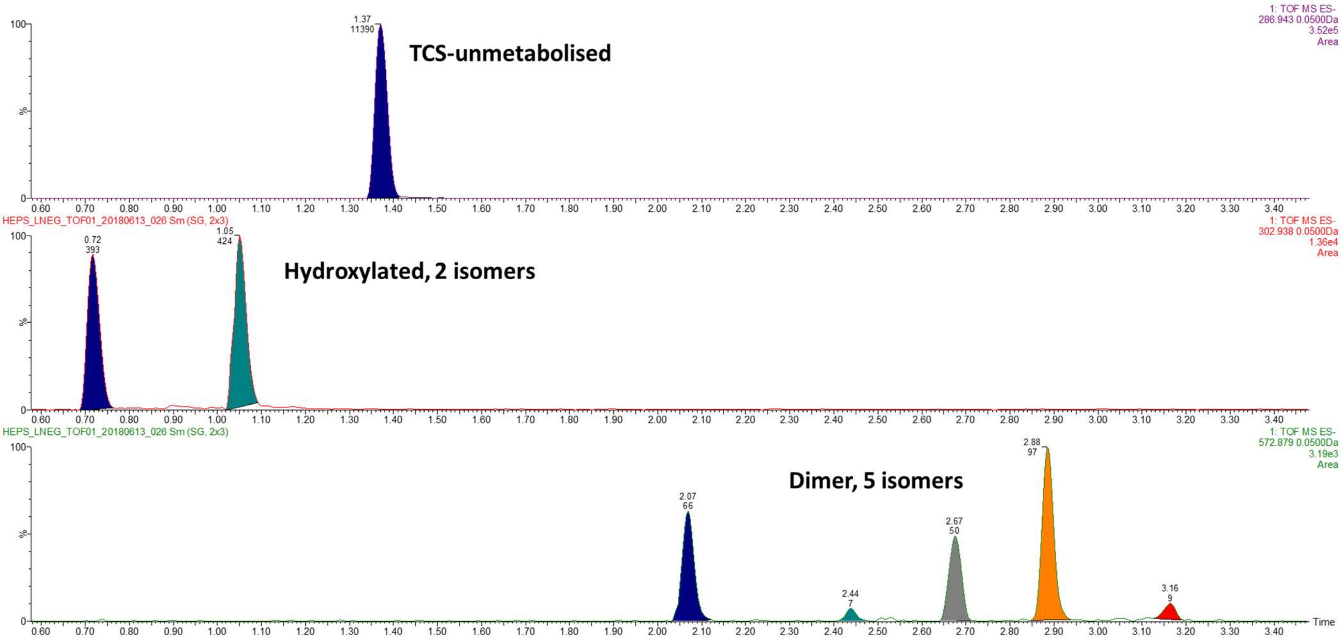


Figure 5-7. Extracted mass chromatograms for TCS unmetabolised, and metabolites from the 50µM incubation of TCS with human liver microsomes at Time = 60 minutes analysed by LIPID-UPLC-MS (negative ion mode). The dimerised TCS metabolites as identified by Rang *et al,* were present, but only in the absence of phase II enzymes. These metabolites were not observed in the AW and ALZ plasma/serum datasets.

5.4.3.4 LOD, LOQ and linear range

Calibration curves, made in water and blank urine, were constructed to approximate the LOD, LOQ, and linearity range for TCS-Gluc. A linear range was measured to be between 0.1 µg/mL and 500 µg/mL. The curve began to plateau at levels greater than 500 µg/mL. A LOD of 0.5 µg/mL and a LOQ of 2 µg/mL were estimated from these curves. Both water and urine calibration curves reported mean concentrations at 36 µg/mL and 34 µg/mL for SR and 26 µg/mL and 24 µg/mL for the Devset-urine, respectively. The precision was well within 5% for both QC samples. A two-sample t-test was subsequently used with a p-value < 0.05, demonstrating no statistically significant deference between the measured means of the SR and Devset-urine using either water or urine calibration curves. As there is no difference between the water and urine curves, only the urine curve was used for subsequent measurements.

A standard addition method using the SR samples from AW, calculated the SR to be 40 µg/mL (n=3), similar to the levels measured using the calibration curves. As a dilution series, is acquired with project data, it was also prepared and acquired with this experiment. The concentration of the SR from standard addition, corresponds to the intensity measured for the 100%SR sample, and each dilution series level is then calculated as a percentage of the 100% SR, i.e. 80%,60%,40%,20%,10% and 1%. This curve was plotted, and for ease of comparison, the dilution series intensities were additionally projected onto the urine calibration curve and the concentration back-calculated (**Figure 5-8**). The standard addition curve (red), is much steeper than the urine (blue) or water (green) curve, demonstrating a certain level of matrix effects. This can be explained by the fact that the SR sample is a complex pool of approximately 3000 samples, each with various levels of diversity. Although the gradient is steeper, a similar concentration of TCS-Gluc in the SR was observed. The goodness of fit R2, for all three curves was superior to 0.9955.

Finally, a subset of samples from the three exposure groups were run and their TCS-Gluc signal measured using the urine calibration curve. All samples in the zero group had levels less than the LOD, in the Low-mid range group, concentrations ranged from 3-10 µg/mL, and in the high group, concentrations ranged from 70 µg/mL to concentrations exceeding 500 µg/mL.

To my knowledge, Provencher et al. is the only piece of work that has quantitative data on TCS-Gluc in human urine samples using a LC-MS/MS platform (Provencher et al., 2014). They established an LOD at 0.089 µg/mL, a LOQ at 0.30 µg/mL and a calibration range from 0.3 µg/mL to 100 µg/mL. Overall, a similar LOD, LOQ and linear range was observed in this study. The method presented by Provencher et al., is a fully validated method meeting laboratory accreditation standards, which explains the lower LOD achieved and the fact they were also able to detect the TCS unmetabolised form and sulfate conjugate in urine. They stated that 97.7% of the total TCS detected in the human urine samples measured, were attributed to TCS-Gluc. The platform and analysis used for urine phenotyping in this study was still able to capture the best biomarker for evaluating human exposure to TCS, whilst also achieving similar quantitative metrics.

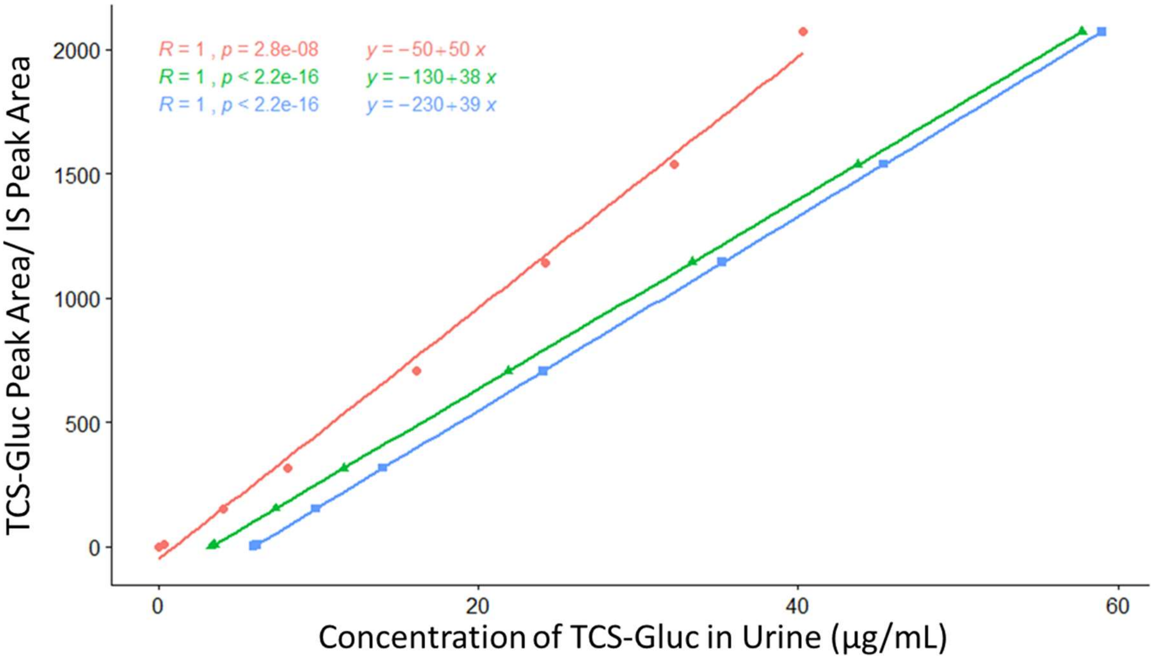


Figure 5-8. Calibration curves for TCS-Gluc made in solvent and urine, superimposed with a standard addition curve using AW SR analysed by RPC-UPLC-MS (negative ion mode). The curve associated with the standard addition method (red) presents a much steeper gradient, which is potentially due to the SR sample being a much more complex matrix then the urine (blue) and water (green) curves. However, all three curves estimated a similar concentration of TCS-Gluc in the SR. The urine calibration curve was used to estimate the approximate concentration ranges observed for samples occupying the High, Low-Mid, and the Zero exposure groups.

5.4.3.5 *TCS-Gluc associations using LogReg models*

A penalised logistic regression model was firstly employed to find TCS related associations. The results from Ridge regression however, even after bootstrapping the regression coefficients, produced too many variables which made interpretation and selection too complicated. EN however had the opposite issue, where the number of significant variables were less than 10 for each AW cohort. Only two features replicated between the cohorts of AW. This was immediately identified as the oxidised sulfate and the oxidised glucuronide of TCS, which was observed from the *in vitro* work. These metabolites however were not present in the ALZ dataset. This is possibly due to the smaller sample size or, these features could have been filtered due to the minfrac setting during XCMS pre-processing. The data was re-processed to include these features, however, was not significant from the models. EN was therefore successful in identifying direct TCS metabolites, however, to find more feature associations, other than direct xenobiotic related correlates, a univariate approach which incorporated meta-analysis (p_{meta}), was implemented. The outcome from the meta-analysis across the five cohorts demonstrated an association of 16 features ($p_{meta} \leq 0.05$) (**Table 5-3**). The five cohorts were combined (using all data allocated to the training set), to create a model using only these 16 features. This model was then applied to the combined test set data. The ROC curve estimated an AUC of 0.96 as illustrated in **Figure 5-9**. ROC analyses were used to further characterize the predictive value of these individual metabolites independently.

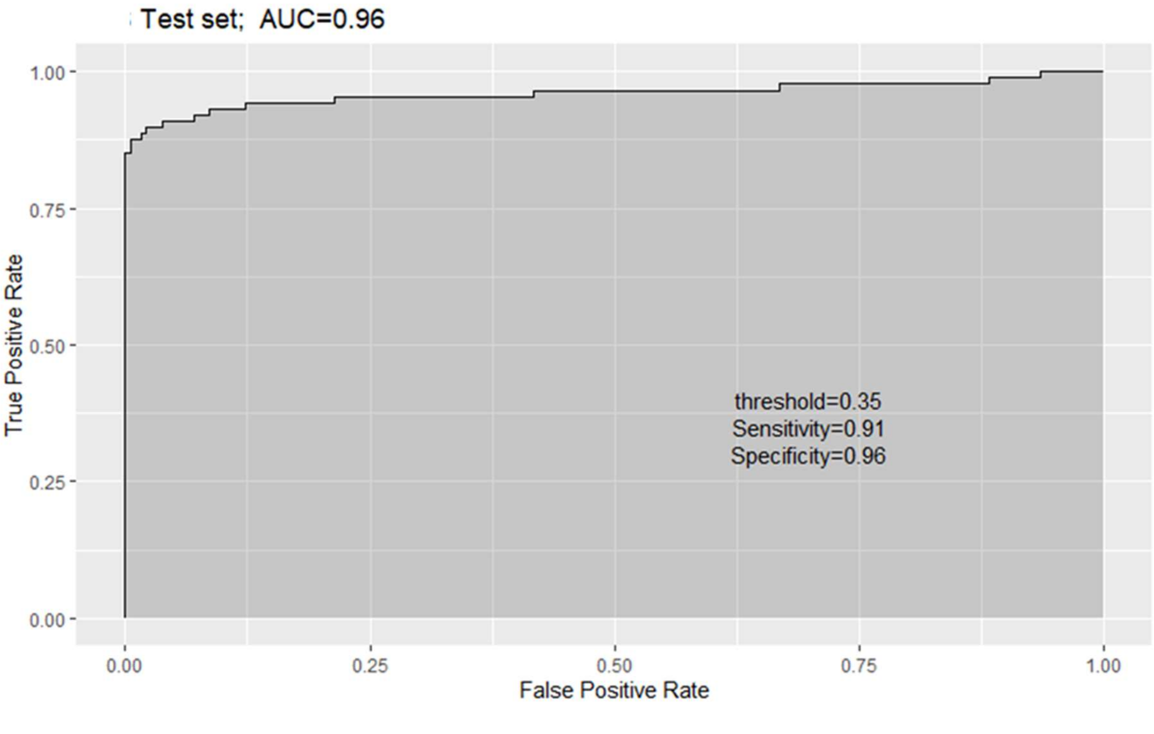


Figure 5-9. The ROC analysis from a binary logistic regression model used to characterise the metabolites associated with TCS exposure in AW urine RPC-UPLC-MS (negative ion mode), from thetest set data. Exposure groups (zero and High) were firstly partitioned into training and testing sets where selection of discriminant variables were conducted on the training set and the performance validated on the test set. From the zero group, 80% of the samples were assigned to a training set. Similarly, 80% of samples from the high group were selected and assigned to the same training set to maintain the same ratio between zero and high

groups in the training and test sets. The remaining 20% of samples from each cohort were combined and assigned to the testing set. The 80:20 split, incorporated a Euclidean distance metric. A logistic regression model was constructed on the training sample sets (from AW and ALZ totalling five cohorts) to design the best metabolite combination using features selected from the (p_{meta}) analysis. A ROC curve was used to evaluate the accuracy of this model in the combined independent test set. The performance of the model was evaluated using the AUC and the determination of sensitivity and specificity at the optimal cut-off point defined by the minimum distance to the top-left corner. The optimised model resulted in an AUC of 0.96 demonstrating an excellent ability of the model to distinguish between the high and zero TCS exposure groups.

Table 5-3. padj -values and regression coefficients of 16 features with the most significant association to TCS exposure following logistic regression applied to urine AW RPC-UPLC-MS (negative ion mode) data. A meta-analysis was used to combine results from five different cohorts. The meta-analysis of these results revealed that 16 metabolites differed significantly (padj-value < 0.05) between samples occupying the low and high exposure groups.

Metabolite	AW-B1		AW-B1.2		AW-B2.2		AW-B3		ALZ		Combined p-value (p_{meta})
	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	
Metabolite 1	0.35	7.83×10^{-1}	2.07	4.03×10^{-2}	2.98	1.45×10^{-4}	1.85	4.28×10^{-4}	2.35	9.47×10^{-4}	3.88×10^{-7}
Metabolite 2	0.53	7.81×10^{-1}	1.75	5.53×10^{-2}	2.65	8.95×10^{-5}	1.57	1.02×10^{-3}	2.83	6.62×10^{-4}	6.01×10^{-7}
Metabolite 3	0.82	4.58×10^{-1}	1.14	1.68×10^{-1}	2.24	8.95×10^{-5}	1.53	3.43×10^{-4}	2.27	2.28×10^{-4}	7.02×10^{-8}
Metabolite 4	0.95	7.81×10^{-1}	1.82	4.03×10^{-2}	2.94	8.95×10^{-5}	1.07	1.29×10^{-2}	2.00	5.94×10^{-3}	1.40×10^{-5}
Metabolite 5	0.38	8.20×10^{-1}	1.89	4.03×10^{-2}	0.91	5.24×10^{-1}	1.39	2.61×10^{-3}	1.58	2.56×10^{-3}	1.66×10^{-3}
Metabolite 6	0.67	7.81×10^{-1}	2.00	4.03×10^{-2}	3.02	9.69×10^{-5}	1.48	1.64×10^{-3}	1.96	3.82×10^{-3}	2.13×10^{-6}
Metabolite 7	0.25	9.33×10^{-1}	0.64	5.21×10^{-2}	0.60	7.18×10^{-1}	4.13	4.90×10^{-9}	1.90	5.31×10^{-4}	3.29×10^{-5}
Metabolite 8	0.34	2.20×10^{-1}	0.75	2.20×10^{-1}	0.90	1.30×10^{-2}	1.21	1.47×10^{-4}	1.57	2.12×10^{-3}	1.53×10^{-6}
Metabolite 9	0.090	8.79×10^{-1}	1.72	1.18×10^{-1}	3.19	1.65×10^{-4}	2.18	3.43×10^{-4}	3.07	2.28×10^{-4}	1.17×10^{-7}
Metabolite 10	0.43	1.08×10^{-1}	0.62	4.59×10^{-2}	1.24	8.51×10^{-4}	1.04	6.58×10^{-4}	1.25	5.63×10^{-3}	5.75×10^{-8}
Metabolite 11	0.57	7.88×10^{-1}	1.09	4.21×10^{-1}	1.24	4.03×10^{-1}	2.06	3.66×10^{-3}	3.74	5.30×10^{-4}	4.47×10^{-3}
Metabolite 12	1.00	6.17×10^{-1}	1.80	4.03×10^{-2}	2.48	8.95×10^{-5}	1.56	4.90×10^{-2}	2.90	1.95×10^{-4}	2.32×10^{-6}

Metabolite	AW-B1		AW-B1.2		AW-B2.2		AW-B3		ALZ		Combined p-value (p_{meta})
	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	Coefficient	p_{adj} -value	
Metabolite 13	0.39	6.82×10^{-2}	0.41	1.63×10^{-1}	0.55	4.96×10^{-2}	0.71	3.63×10^{-3}	1.69	3.09×10^{-3}	7.71×10^{-6}
Metabolite 14	0.59	7.81×10^{-1}	1.58	4.32×10^{-2}	2.40	8.95×10^{-5}	1.16	8.40×10^{-4}	2.23	2.28×10^{-4}	2.03×10^{-7}
Metabolite 15	9.31	1.01×10^{-2}	11.7	9.25×10^{-2}	18.9	9.25×10^{-5}	12.52	2.12×10^{-2}	0.16	8.60×10^{-1}	3.73×10^{-3}
Metabolite 16	4.39	4.00×10^{-10}	8.06	3.16×10^{-3}	8.69	5.35×10^{-3}	7.66	3.67×10^{-4}	0.19	7.70×10^{-1}	1.00×10^{-11}

5.4.3.6 Blood products

Initial screening for TCS revealed only the sulfate conjugate (TCS-SO4) present in phenotyping data. Interestingly, 20% prevalence was observed, which was less than the reported prevalence in urine (28%). Both univariate and multivariate LogReg models revealed four consistent metabolites with an association to exposure in the AW study (**Table 5-4**). The ROC curve (**Figure 5-10**) estimated an AUC of 0.675 (from an independent test set not used in training the model). This AUC value falls somewhere between poor and acceptable discrimination; therefore, caution should be taken in the biological interpretation of metabolites identified from this model.

Table 5-4. p_{adj} -values and regression coefficients of four features with the most significant association to TCS exposure following logistic regression applied to the AW plasma cohort acquired by LIPID-UPLC-MS (negative ion mode).

Metabolite	AW-plasma	
	Coefficient	p_{adj} -value
Metabolite 17	2.15	7.14×10^{-3}
Metabolite 18	0.83	3.97×10^{-2}
Metabolite 19	1.14	9.64×10^{-4}
Metabolite 20	0.58	8.73×10^{-3}

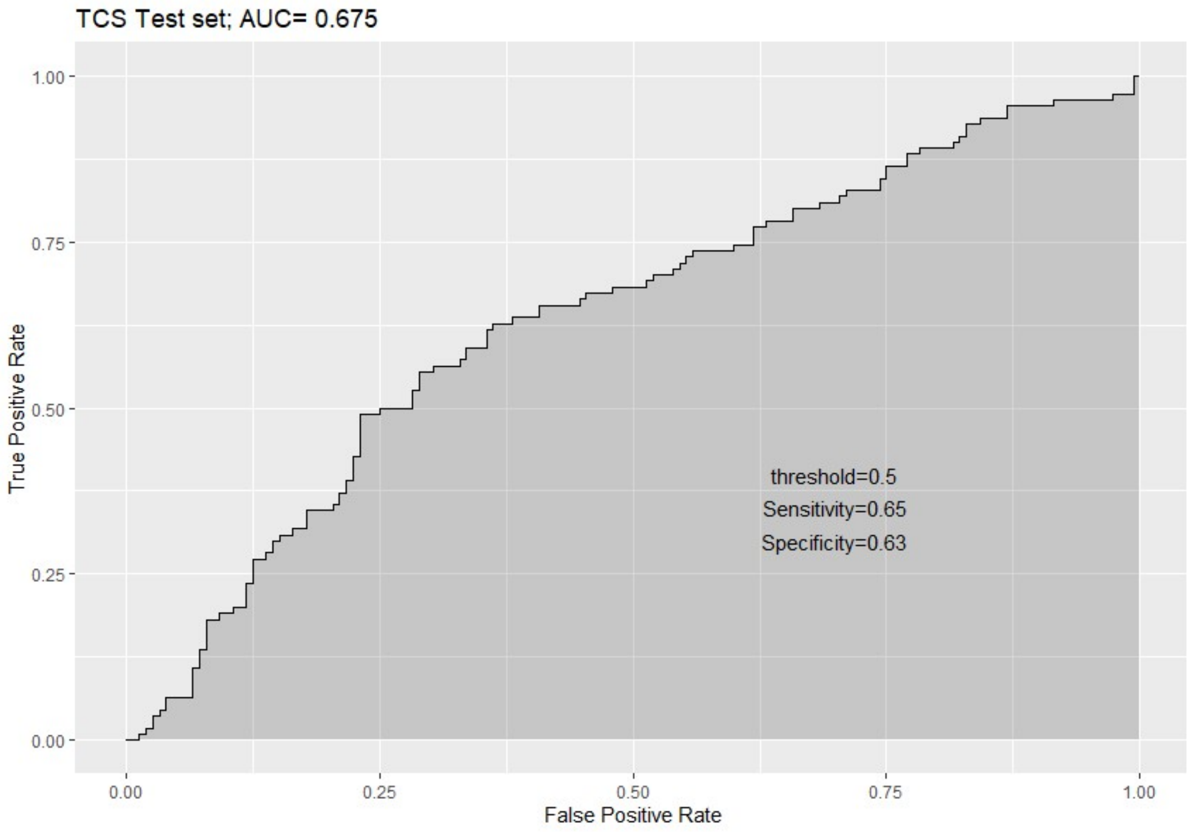


Figure 5-10. The ROC analysis from a binary logistic regression model used to characterise the metabolites associated with TCS exposure in AW plasma LIPID-UPLC-MS (negative ion mode), from the test set data. Exposure groups (zero and High) were firstly partitioned into training and testing sets where selection of discriminant variables were conducted on the training set and the performance validated on the test set. From the zero group, 80% of the samples were assigned to a training set. Similarly, 80% of samples from the high group were selected and assigned to the same training set to maintain the same ratio between zero and high groups in the training and test sets. The remaining 20% of samples from each group were combined and assigned to the testing set. The 80:20 split, incorporated a Euclidean distance metric. A logistic regression model was constructed on the training sample sets (AW plasma – one cohort) to design the best metabolite combination using features which had p_{adj} – value ≤ 0.05 (FDR multiple testing correction). A ROC curve was used to evaluate the accuracy of this model in the combined independent test set. The performance of the model was evaluated using the AUC and the determination of sensitivity and specificity at the optimal cut-off point defined by the minimum distance to the top-left corner. The optimised model resulted in an AUC of 0.675.

5.4.3.7 Metabolite identification

Of the 20 metabolites from both urine and blood bio-fluids, there were four metabolic identifications; 2,4-dichlorophenol glucuronide (Metabolite 7 – **Figure 5-14**), γ-hydroxybutyric acid sulfate (Metabolite 9 – **Figure 5-13**), menthol glucuronide (Metabolite 13 – **Figure 5-12**) and perfluorooctanesulfonic acid (Metabolite 18 – **Figure 5-11**). There were seven putative annotations, and the remaining features could not be deduced from the MS/MS acquisitions. The results of metabolite identification efforts are summarised in **Table 5-5**.

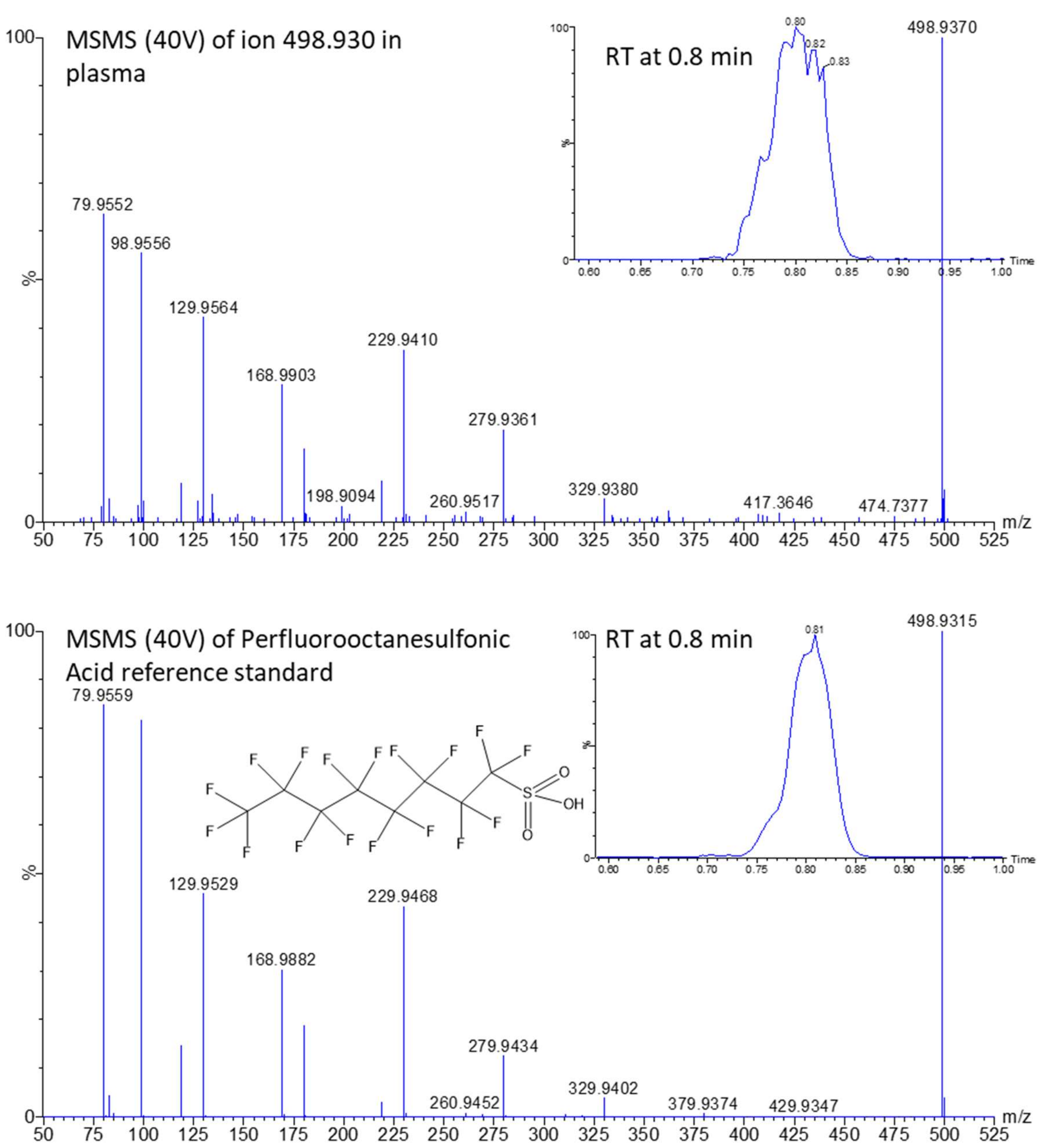


Figure 5-11. Metabolite identification efforts to compare retention time and MS/MS for Metabolite 18, to the MS/MS of the candidate reference compound perfluorooctanesulfonic acid analysed by LIPID-UPLC-MS (negative ion mode) chromatographic conditions. The MS/MS fragmentation spectra and retention time of Metabolite 18 observed in the profiling data was a match to the candidate reference standard. Metabolite 18 has therefore been identified as perfluorooctanesulfonic acid.

5.4.3.7.1 Intersample correlation

Using the molecular ion associated with Metabolite 12 (putatively annotated as lauric acid sulfate) as the driver feature, statistically significant correlations (Spearman; $p_{adj}\leq0.05$) were observed with, Metabolite 1 (r = 0.80), Metabolite 2 (r = 0.82), Metabolite 3 (r = 0.69), Metabolite 6 (r = 0.82), and Metabolite 14 (r = 0.86), respectively. Metabolite 6 was putatively annotated as hydroxy lauric acid sulfate, so is not surprising that they correlate strongly to one another. Unfortunately, the unknown metabolites could not be deduced from this correlation study, however their statistically significant correlation to Metabolite 12, strongly suggest that they are metabolically connected.

Metabolite 13 was identified as menthol glucuronide by matching the retention time and fragmentation spectra to a reference standard (**Figure 5-12**). When the molecular ion was used as a driver feature, statistically significant correlations were observed to the metabolites putatively annotated as menthol correlates, i.e. Metabolite 8 (menthol bound with two glucuronic acid molecules linked together, r = 0.71) and Metabolite 10 (hydroxy menthol glucuronide of *p*-menthane-3,7-diol, r = 0.61).

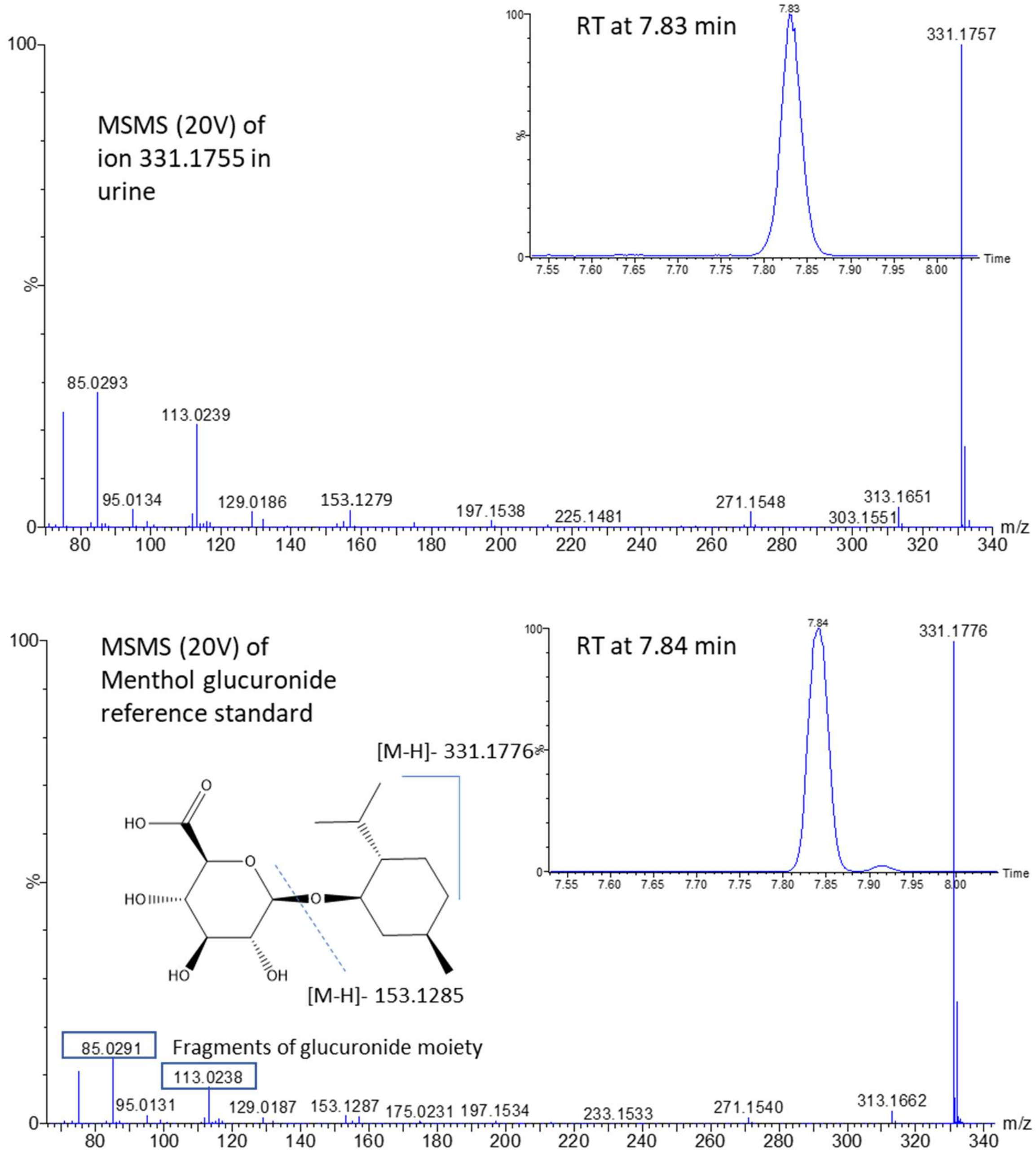


Figure 5-12. Metabolite identification by comparison of retention time and MS/MS spectra for Metabolite 13, to the MS/MS of the candidate reference compound menthol glucuronide analysed by RPC-UPLC-MS (negative ion mode) chromatographic conditions. The MS/MS fragmentation spectra and retention time of Metabolite 13 observed in the profiling data was a match to the candidate reference standard. These data confirm the identity of Metabolite 13 as menthol glucuronide.

5.4.3.7.2 Metabolite sulfation

No peak was present at the highest concentration for both unmetabolised and sulfate conjugates for the candidate metabolites suspected for Metabolite 2, Metabolite 6 and Metabolite 12. MassLynx elemental composition software predicted a chemical formula $C_4H_7O_6S$ for Metabolite 9, which was subsequently inputted into the online database HMDB. Hydroxybutyric acid Sulfate was identified. Three isomeric unconjugated forms have been reported in human urine; γ-hydroxybutyric acid, β-hydroxybutyric acid and α-hydroxybutyric acid (Petersen et al., 2013, Gall et al., 2010, Stojanovic and Ihle, 2011). Sulfated versions are not commercially available which presents a bottleneck in identification and structural separation of the three metabolites. The unconjugated forms were purchased, and the sulfation protocol were implemented on all three compounds. The unconjugated and sulfate conjugated reference standards were acquired (MS scan and MS/MS) alongside a study sample positive for Metabolite 9, confirming the identity to be γ-hydroxybutyric acid sulfate *via* matching of retention time and fragmentation pattern (**Figure 5-13**). Interestingly, although γ-hydroxybutyric acid has two hydroxyl functional groups, sulphation of the metabolite did not result in two isomeric peaks.

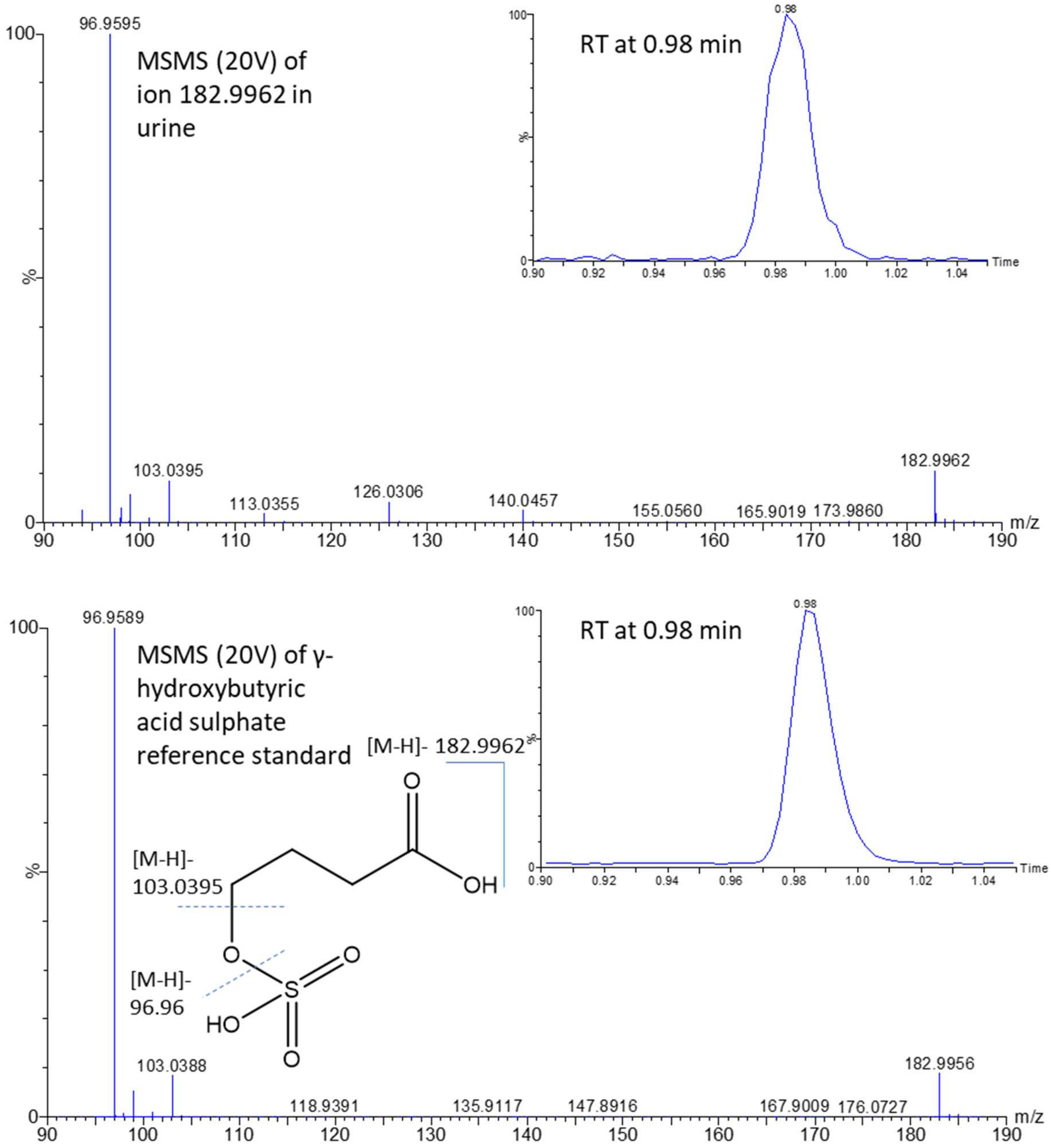


Figure 5-13. Metabolite identification by comparison of retention time and MS/MS spectra for Metabolite 9, to the MS/MS of the candidate reference standard compound γ-hydroxybutyric acid sulfate analysed by RPC-UPLC-MS (negative ion mode) chromatographic conditions. The MS/MS fragmentation spectra and retention time of Metabolite 9 observed in the profiling data was a match to the candidate reference standard. These data conform the identity of Metabolite 9 to be γ-hydroxybutyric acid sulfate.

5.4.3.7.3 Enzymatic hydrolysis

Metabolite 7 was putatively annotated as 2,4-dichlorophenol glucuronide (2,4 DCP-Gluc); 2,4-dichlorophenol (2,4 DCP) is a phase 1 metabolite of TCS and its glucuronidated conjugate is a reported metabolite in urine (Somani and Khalique, 1982). A search of the molecular ion 336.9881 revealed an additional isomer which can also be present in urine, 2,5-dichlorophenol glucuronide (2,5 DCP-Gluc) (Park and Kim, 2018). An MS/MS experiment was acquired on this feature (using RPC-UPLC-MS chromatographic conditions), however fragmentation patterns revealed nothing useful for differentiating between the two glucuronide isomers. Reference standards for the unconjugated forms were available for purchase and acquired by RPC using the method for reference standard acquisition, described in Chapter 3. 2,4 DCP, was detected at retention time 8.70 minutes while 2,5 DCP was detected at retention time of 8.66 minutes. Using identical chromatography as the profiling methods, a NL acquisition was conducted on the enzymatic hydrolysed study sample. The retention time of Metabolite 7 was at 5.78 minutes. After enzymatic hydrolysis, the peak at 5.78 minutes disappeared, and the emergence of a new peak at retention time 8.70 minutes was observed in the NL acquisition (**Figure 5-14**). This confirms that the unknown metabolite at retention time 5.78 min is 2,4-dichlorophenol glucuronide.

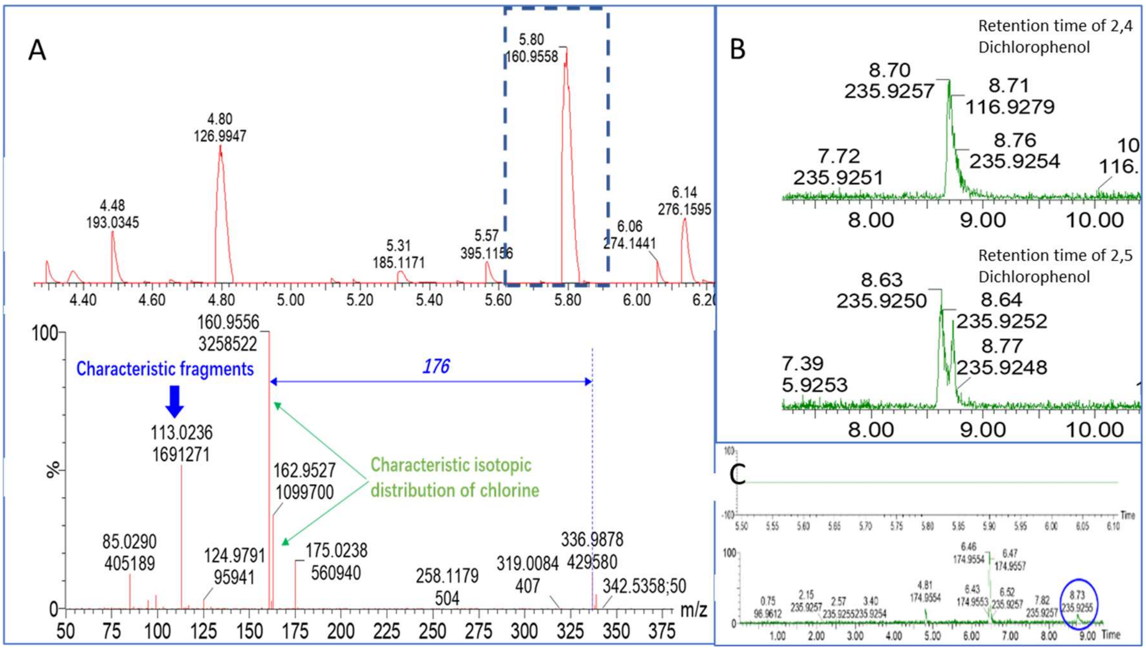


Figure 5-14. Metabolite identification efforts to identify Metabolite 7, by enzymatic hydrolysis.

(A) The top figure is the TIC of the NL acquisition at 176 Da (glucuronide moiety) in a urine sample analysed by RPC-UPLC-MS chromatographic conditions. The retention time at 5.8 minutes corresponds to the retention time of Metabolite 7. The bottom figure is an additional MSMS acquisition of the unknown feature at 5.8 minutes, which clearly depicts the loss of 176 to give the unconjugated M_ H ion at 160.9556. Fragments at *m/z* 113.0236 and *m/z* 85.0290 are typical fragment ions associated with the glucuronide moiety;

(B) The retention time of 2,4 dichlorophenol (top) and 2,5 dichlorophenol (bottom) reference standards analysed by RPC-UPLC-MS (negative ion mode);

(C) TIC of the urine sample (RPC-UPLC-MS) after enzymatic hydrolysis. The top image represents the TIC of the NL acquisition at 176, demonstrating that the moiety has been cleaved. The bottom image is the EIC of the ion 160.9561, which is the molecular ion of the unconjugated form of dichlorophenol (either isomer). The peak at 5.8minutes is no longer detected, but a peak at 8.73 minutes has appeared. This matches the retention time to the reference standard of 2,4 dichlorophenol.

All images for this analysis have been provided by Miss Ziyue Wang (Imperial).

Table 5-5. MS/MS summary of all metabolites with a significant association to TCS exposure, possible putative annotations (MSI levels), and confirmation with a reference standard if available.

Metabolite	Biofluid	Assay (Polarity)	Experiment	RT (min)	[M-H]$^-$ Molecular ion	*m/z* -Fragment	Structural Elucidation	Candidate Molecular Formula	Candidate Compound	Confirmation	Identification/ Putative Annotation
Metabolite 1	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS	5.82	297.1383	96.96	Fragment at *m/z* 96.96 – sulfate. Many possible candidates, but should have -OH group or two	$C_{12}H_{26}O_6S$	unknown		No match (MSI level 4)
Metabolite 2	Urine	RPC-UPLC-MS (negative ion mode)	Metabolite Sulfation, MS/MS and reference standard	7.42	281.143	96.96	Fragment at *m/z* 96.96 – sulfate.	$C_{12}H_{26}O5S$	1,2-dodecanediol sulfate	No peak present for both unmetabolised and sulfate	Putative annotation (MSI level 4)
Metabolite 3	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS	6.41	394.2267	335.153, 96.96	Fragment *m/z* 335.153 – loss of 59 Da, -N(CH3)3 betaine group. Fragment *m/z* 96.96 – sulfate.	$C_{18}H_{37}NO_6S$	unknown		No match (MSI level 4)
Metabolite 4	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS	6.51	281.1064	201.15, 155, 96.96	Fragment at 96.96 – sulfate. Fragment at *m/z* 201.15 – loss of 80 Da from sulfate. MS/MS at 30 V shows the fragment at *m/z* 155 – loss of 46 Da from *m/z* 201, formic acid CH2O2. Alpha hydroxy acids typically dissociate in tandem mass spectrometric experiments to produce product ions	$C_{11}H_{22}O_6S$	2-hydroxyundecanoic acid sulfate	No reference standard commercially available	Putative Annotation (MSI level 3)

Metabolite	Biofluid	Assay (Polarity)	Experiment	RT (min)	$[M-H]^-$ Molecular ion	*m/z* -Fragment	Structural Elucidation	Candidate Molecular Formula	Candidate Compound	Confirmation	Identification/ Putative Annotation
							representing a neutral loss of 46 Da (CH2O2) in negative ion mode. (i.e. 201->155)				
Metabolite 5	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS	6.21	295.1216	96.96	Fragment at 96.96 – sulfate	$C_{12}H_{24}O_6S$	unknown		No match (MSI level 4)
Metabolite 6	Urine	RPC-UPLC-MS (negative ion mode)	Metabolite Sulfation, MS/MS and reference standard	7.33	295.1222	215.1647, 169.17, 96.96	Fragment at 96.96 – sulfate. Fragment at *m/z* 215.1647 – loss of 80 Da from sulfate. MS/MS at 30 V shows the fragment at *m/z* 169.17 – loss of 46 Da from *m/z* 215, formic acid CH2O2.	$C_{12}H_{24}O_6S$	Hydroxydodecanoic acid sulfate or hydroxy lauric acid sulfate. Possible metabolite of lauric acid	No peak present for both unmetabolised and sulfate	Putative annotation (MSI level 3)
Metabolite 7	Urine	RPC-UPLC-MS (negative ion mode)	Enzymatic Hydrolysis	5.65	336.9876	160.95, 96.96	Fragment at 160.9 – loss of glucuronide. 113 and 85 are typical fragment ions of the glucuronide moiety	$C_{12}H_{12}Cl_2O_7$	2,4 Dichlorophenol glucuronide	Confirmed by enzymatic hydrolysis	Identification (MSI level 1)
Metabolite 8	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS	6.32	507.2081	351.12, 113.02, 85.03	Fragment at *m/z* 351 – loss of menthol, which means that two glucuronic acid moieties are bound together.	$C_{22}H_{36}O_{13}$	Menthol bound with two glucuronic acid molecules linked together	No reference standard commercially available	Putative annotation (MSI level 3)
Metabolite 9	Urine	RPC-UPLC-MS	Metabolite Sulfation, MS/MS and	1.01	182.9976	96.96	Fragment at 96.96 – sulfate	$C_4H_8O_6S$	γ-hydroxybutyric acid sulfate	Reference standard *via*	Identification (MSI level 1)

Metabolite	**Biofluid**	**Assay (Polarity)**	**Experiment**	**RT (min)**	**[M-H]$^-$ Molecular ion**	***m/z* -Fragment**	**Structural Elucidation**	**Candidate Molecular Formula**	**Candidate Compound**	**Confirmation**	**Identification/ Putative Annotation**
		(negative ion mode)	reference standard							sulfation match	
Metabolite 10	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS	3.15	347.1711	113.02, 85.03	Typical fragment ions of the glucuronide moiety	$C_{16}H_{28}O_8$	Hydroxy Menthol glucuronide of p-menthane-3,7-diol	No reference standard commercially available	Putative annotation (MSI level 3)
Metabolite 11	Urine	RPC-UPLC-MS (negative ion mode)	Metabolite Sulfation, MS/MS and reference standard	0.75	168.9827	96.96	Fragment at 96.96 – sulfate	$C_3H_6O_6S$	Sulfonatolactate	Reference standard does not match	No match (MSI level 4)
Metabolite 12	Urine	RPC-UPLC-MS (negative ion mode)	Metabolite Sulfation, MS/MS and reference standard	7.76	279.1275	96.96	Fragment at 96.96 – sulfate	$C_{12}H_{24}O_5S$	Lauric acid sulfate	No peak present for both unmetabolised and sulfate	Putative annotation (MSI level 4)
Metabolite 13	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS and reference standard	7.82	331.1774	153.128, 113.02, 85.03	Fragment at 153 – loss of glucuronide. 113 and 85 are typical fragment ions of the glucuronide moiety	$C_{16}H_{28}O_7$	Menthol glucuronide	Reference standard match	Identification
Metabolite 14	Urine	RPC-UPLC-MS (negative ion mode)	MS/MS	8.17	295.1225	235.11, 96.96	Fragment at 96.96 – sulfate. MS/MS at 20 V shows the fragment at *m/z* 235.11 – loss of 60 Da from *m/z* 295, acetic acid CH2O2. Ethyl ester moiety in the structure.	$C_{12}H_{24}O_6S$	unknown		No match (MSI level 3)

Metabolite	**Biofluid**	**Assay (Polarity)**	**Experiment**	**RT (min)**	**$[M-H]^-$ Molecular ion**	***m/z* -Fragment**	**Structural Elucidation**	**Candidate Molecular Formula**	**Candidate Compound**	**Confirmation**	**Identification/ Putative Annotation**
Metabolite 15	Urine	RPC-UPLC-MS (negative ion mode)	*in vitro* comparison	6.79	478.969			C_{18}H15Cl3O9	TCS Oxidised Glucuronide	Match with *in vitro*	Putative annotation (MSI level 2)
Metabolite 16	Urine	RPC-UPLC-MS (negative ion mode)	*in vitro* comparison	7.94	382.8943			$C_{12}H_{7}Cl_{3}O_{6}S$	TCS Oxidised Sulfate	Match with *in vitro*	Putative annotation (MSI level 2)
Metabolite 17	Plasma	LIPID-UPLC-MS (negative ion mode)	Internal Reference LIPID Database	4.42	657.4961		Ion type $[M-CH_3]$-. Also present was 731.5327, ion type $[M+CH_3COO]$-	$C_{37}H_{73}N_{2}O_{6}P$	Sphingomyelin (d18:2/14:0)	Reference standard match	Putative annotation (MSI level 2)
Metabolite 18	Plasma	LIPID-UPLC-MS (negative ion mode)	MS/MS and reference standard	0.79	498.9291	80, 98.95, 168.98, 279.95	Retention time and MS/MS match with reference standard	$C_{8}HF_{17}O_{3}S$	Perfluorooctanesulfonic Acid	Reference standard match	Identification (MSI level 1)
Metabolite 19	Plasma	LIPID-UPLC-MS (negative ion mode)	MS/MS	2.18	473.2802	457.25, 205.16	Fragment at 457.25 – loss of water		unknown		No match (MSI level 4)
Metabolite 20	Plasma	LIPID-UPLC-MS (negative ion mode)	MS/MS	1.07	660.8435	126.90, 216.91, 243.04, 279.03, 315.01, 344.82	315.01 loses HCl to 279.03 which loses HCl to 243.04, thus compound contains two chlorine atoms. 344.82 to 216.91, neutral loss of 126.905 which is iodine	$C_{22}H_{14}Cl_{2}I_{2}N_{2}O$ 2	Closantel?		No match (MSI level 3)

5.4.3.8 Correlation between matching biofluids

Using the Pearson method, statistically significant ($p_{adj} \leq 0.05$) positive correlations were observed between TCS-SO4 in plasma to TCS-Gluc (r = 0.89), TCS-oxidised glucuronide (r = 0.75), and TCS-oxidised sulfate (r = 0.79) in urine, suggesting, a significant relationship between circulating plasma TCS and excreted TCS. This is illustrated in **Figure 5-15** where the intersection of the two purple dotted lines represent levels of TCS which are detected in accordance with the detection criteria stated in **section 5.4.2.3.1**.

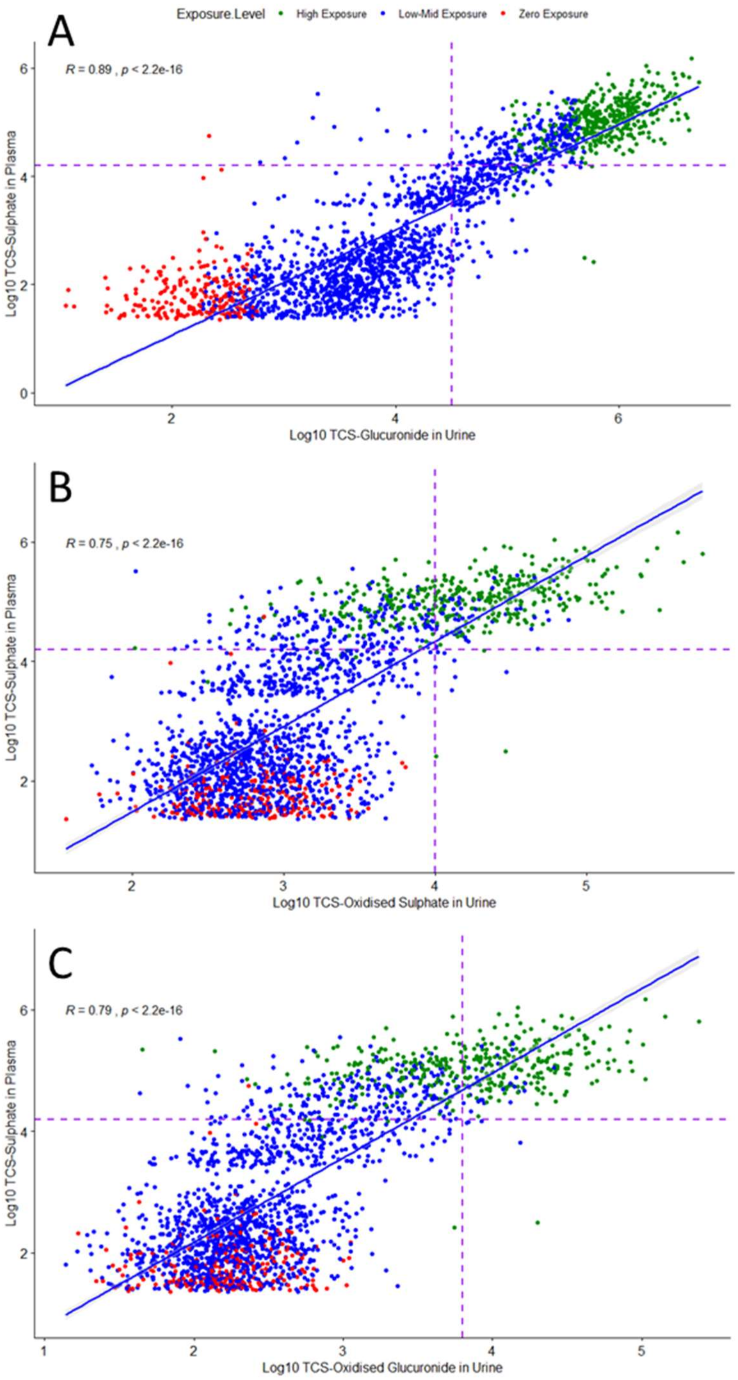


Figure 5-15. Correlation of the TCS signals detected in plasma (LIPID-UPLC-MS negative ion mode), to TCS signals in urine (RPC-UPLC-MS negative ion mode).

(A) A scatter plot demonstrating the correlation between the intensity level of TCS-Sulfate in plasma and TCS-Gluc in urine;

(B) A scatter plot demonstrating the correlation between the intensity level of TCS-Sulfate in plasma and TCS oxidised sulfate in urine;

(C) A scatter plot demonstrating the correlation between the intensity level of TCS-Sulfate in plasma and TCS oxidised glucuronide in urine.

A liner regression line fitted to the scatter plot which visually illustrates a positive statistically significant correlation between TCS-Sulfate in plasma to TCS-Gluc (r = 0.89), TCS-oxidised glucuronide (r = 0.75), and TCS-oxidised sulfate (r = 0.79) in urine, thereby suggesting a significant relationship between circulating plasma TCS metabolite to excreted urinal TCS metabolites. The points are coloured by their exposure group (red = Zero to low, blue = Low-Mid, and green = High exposure) and the intersection of the two purple dotted lines represent levels of TCS which are detected in accordance with the detection criteria stated in prevalence section of this study.

5.4.4 *General Discussion*

The chemical complexity of urine samples combined with such a large study size was problematic in terms of sample pre-processing. Due to lack of resources and computational power, the urine AW dataset could not be pre-processed as one cohort but rather split into four. As a result of this limitation, a meta-analysis was necessary to combine cohorts to enable subsequent statistical exploration. Meta-analyses take into account the heterogeneity observed with the different cohorts which can stem from various sources, such as differences in participants, or study design. In this application, the Stouffer method with weights meta-analysis was specifically useful, as it applied a weighting algorithm which accounts for sample size. This therefore enabled combining p-values from the logistic regression models conducted on the different cohorts which ultimately led to the discovery of TCS exposure related metabolites. This all could have been avoided if the datasets were able to be pre-processed and combined as one cohort. The complexity involved with such a task can be difficult and therefore it is common practice to concentrate on the statistically significant metabolites, whether it be for metabolite identification, or for classifying new data. As demonstrated in this application, once the statistically significant variables were determined using logistic regression on the training dataset, only then were the five test set cohorts stitched together and the model validated using this combined cohort. However, there are tools now available for combining datasets for untargeted metabolomic investigations prior to any statistical testing. The package “MSCombine” is such an example, which uses a series of algorithms to essentially match molecular features from different datasets with provision for mass accuracy and retention time (Calderón-Santiago et al., 2016). Another way to reduce the computational power needed to handle the complexity observed with MS datasets, is to remove redundant metabolic features relating to ionisation products, such as adducts or fragments. A group from the Broad institute has introduced a computational suite named “netome”, that not only can remove these redundancies, but also align features between separately acquired datasets (Rahnavard et al., 2018). Both tools have been successfully applied to metabolomic studies and represents a solution to merge datasets, which can then be statistically analysed.

The main metabolites reported from TCS exposure, are the glucuronide and sulfate conjugates (Rodricks et al., 2010, Wu et al., 2010, Ranganathan et al., 2015, Ye et al., 2007). Rodricks *et al*. noted that the glucuronide metabolite predominates in humans while the sulfate conjugate is the dominant metabolite in mice. The sulfate was not observed in AW urine, however the oxidised sulfate was. Other direct metabolites were oxidised glucuronide and 2,4 dichlorophenol glucuronide. To my knowledge these three metabolites are the first be reported *in vivo*. Wu *et al*, although did not report the presence of these metabolites, did report additional metabolites in rat urine in the form of glucosidated, cysteine and mercapturic conjugates. They did however also report hydroxylated TCS, which is interesting because from the in vitro model evaluated in this chapter, the hydroxylated metabolites were only present in the absence of phase II conjugation enzymes (microsomes). The Logistic regression models from both AW urine and plasma highlighted other feature associations due to TCS exposure. Menthol was another xenobiotic that was identified (MSI level 1) as highlight associated, from the LogReg models. A possible source of exposure, could arise from topical compositions used in cosmetics, where Menthol and TCS are co-ingredients (Samani et al., 2004). The putative annotation of Lauric acid is yet another ingredient used in these cosmetic applications.

The application of MS/MS to features with an association to TCS exposure, did reveal most features to contain a sulfate moiety. A sulfation method was implemented to putative unconjugated candidates, however sulfation can occur on a molecule hydroxyl or amine group, and if the candidate compound has multiple groups present, this can result in the formation of multiple different isomeric forms. Thus, care should therefore be taken when using this method for the preparation of the reference standards to avoid incorrect metabolite annotations. Nevertheless, these highly associated features raised an interesting point. As mentioned in Chapter 2, section 2.2, metabolism of xenobiotics in mammalian systems occurs in two phases, Phases I and II. Both involve the biotransformation of molecules to increase water solubility and facilitate excretion (more details in Chapter 2). Phase 2 reactions, or conjugation reactions, occurs when phase I is insufficient to clear a compound from circulation, or if phase I generates a reactive metabolite. It consists of the conjugation of a drug or its metabolite, with hydrophilic endogenous molecules such as glycine, sulfate or glucuronic acid. Sulfotransferases (SULT), are enzymes that catalyse phase II conjugation by transferring a sulfuryl group, donated by 3-phospoadenosine5’-phosphosulfate (PAPS), to the hydroxy or amine group of a molecule often referred to as sulfonation. There are at least 11 different SULT isoforms in the human body that catalyse sulfate conjugation of endogenous metabolites and xenobiotics. In addition to being a detoxifying step in xenobiotic metabolism, SULTs are involved in many endogenous processes, which include hormone regulation, transport of steroids and modulation of neurotransmitters. SULT activity however has been known to be inhibited by certain xenobiotic exposures, which includes therapeutic drugs, dietary additives, and environmental pollutants. Hydroxylated polychlorinated biphenyls such as Triclosan has been reported to inhibit the sulfonation of Bisphenol A, Acetaminophen, 4-Nitrophenol, all of which occur ubiquitously in the environment. In addition, sulfonation is regarded as a low capacity, high affinity conjugation system (Leung et al., 2016). What this means, is although it is used in a number of different metabolic processes, the capacity to sulfonate can be inhibited due to competitive sulfonation of endogenous metabolites and xenobiotics (Clayton et al., 2009). Whether it is an inhibition of a particular SULT isoform or competition with another metabolite, a disturbance in the capacity to sulfonate could potentially result in prolonged exposure to xenobiotics and therefore toxicity. As, the majority of chemicals with the strongest associations with TCS, were all sulphated, collectively, this could place a biochemical strain on the sulfonation potential needed in endogenous and exogenous metabolism. Another group also made such a claim, highlighting an effect on the sulfonation pathway on the rat metabolome when subjected to human level exposure of TCS (Houten et al., 2016).

Another identified metabolite with a strong association to TCS exposure, was γ- hydroxybutyric acid sulfate (GHB-SO4). The unconjugated parent, GHB, can be used as an anaesthetic drug, enhancing supplement in body building and a substance of abuse used in drug facilitated sexual assault. It is however also generated endogenously in human brain, urine and blood as an *in vivo* metabolic product and short fatty acid derivative of gamma-aminobutyric acid (GABA). GABA is a principal inhibitory neurotransmitter that plays a major role in the central nervous system (CNS) and together with glutamate, has excitatory effects on nerve cells, thereby decreasing the brains overall level of excitation. Much research has been conducted linking the microbiome to the CNS with levels of GABA being associated with depression and mood disorders (Ma et al., 2019, Sharon et al., 2016, Wang and Kasper, 2014). TCS has been reported to change the gut microbiota in fish and rats, with modest effects observed in humans (Gálvez-Ontiveros et al., 2020). However, to date, there is no definitive study that demonstrates a disruption in the microbial community due to TCS exposure. As GHB-SO4 is a direct metabolite of GHB, which in turn is a metabolite of GABA, a link does seem to exist between TCS exposure and a known biomarker associated with the microbiome, however as the parent GHB was not significant from this investigation, care must be taken in its interpretation.

In plasma samples, the association of SM to TCS observed in this study, highlighted a perturbation in lipid metabolism. SM, a class of sphingolipids, together with ceramides and glycosphingolipids, are integral in cell signalling pathways and cell apoptosis in humans (Slotte, 2013). Varying levels of SM are associated with diseases such as multiple sclerosis and Abetalipoproteinemia (Jana and Pahan, 2010, Cooper et al., 1977). An association of SM to TCS has been reported in Daphnia magna (water fleas) (Sengupta et al., 2017). This stemmed from an original metabolomic study, which successfully examined Daphnia and the metabolomic dysregulation due to exposure from the pollutant propranolol (Jeong and Simpson, 2019). The study served as an indicator for water quality and the basis for studying Daphnia in relation to environmental toxicants. Sengupta *et* al. reported an effect of SM metabolism specifically from TCS exposure in new-borns of Daphnia. The changes occurred on a molecular and genetic level, which could potentially result in a cascading effect throughout the food web. Perfluorooctanesulfonic acid (PFOS), like TCS, is a chemical which is widely found in the environment and used in a number of different applications including industrial surfactants and in the manufacture of textiles, cleaning agents, paints and polishes. Due to its ubiquitous nature, its environment fate has been thoroughly examined, with an estimated half-life of 5.4 years and a high potential for bioaccumulation (Sznajder-Katarzyńska et al., 2019). As such, PFOS has been detected in human serum and in organs such as spleen kidneys and brain (Zeng et al., 2019). Studies suggest long term exposure can impact estrogenic activity, result in liver toxicity and alter endocrine functions (Chaparro-Ortega et al., 2018, Karzi et al., 2018, Henry and Fair, 2013). Recently, co-exposures to both TCS and PFOS has increased toxicity in freshwater organisms (González-Doncel et al., 2020). The detection of both these chemicals in human biofluids is therefore a concern.

5.5 Application 2: Polyethylene Glycol

5.5.1 Introduction

Polyethylene glycol (PEG) is a semi-crystalline polyether, composed of repeating sub-units of ethylene oxide. It is one of the most commonly encountered contaminants in biochemical and molecular biological research. It can derive from various sources including, sample collection containers, plastic labware, plastic tubing, LC column manufacture processes, detergents and common analytical reagents (Hodge et al., 2013, Waters, 2003, Mihailova et al., 2006, Weaver and Riley, 2006, Michopoulos et al., 2010, Forcisi et al., 2013). PEG signatures are expressed as an envelope of repeating signals separated by 44.0262 Da, making it easily recognizable in UPLC-MS measurements. These patterns can cause signal loss by supressing other low molecular weight compounds, thereby representing a risk to the validity of data generated from PEG-containing biofluids (Weaver and Riley, 2006, Mortier et al., 2002, Antignac et al., 2005). As a result, metabolomic workflows often remove these samples from the study. However, PEG (<600) is popular water soluble vehicle, widely present in a number of common over-the-counter and prescription drug formulations (D'souza and Shegokar, 2016, Gullapalli and Mazzitelli, 2015), food products (FDA, 2014, EFSA, 2015), supplements (EFSA, 2007), infant diapers (Goodpaster et al., 2011) and cosmetics (Fruijtier-Pölloth, 2005). Therefore, its presence in biofluids may not necessarily derive from a contamination, but rather form a legitimate exposure to any of the aforementioned sources. Commercially, PEG exists as a mixture of varying polymeric lengths with each mixture being reflective of its average molecular weight, for example, PEG400 refers to a distribution of PEG polymers with an average of approximately 400 Da. Its naming convention can also be defined by the number of ethylene oxide monomers present, e.g., PEG(n8) equates to a polymer of 8 ethylene oxide unit repeats. Low molecular weight PEG forms, such as PEG 400 is notorious within analytical chemistry as a ubiquitous contaminant but is also a common vehicle attached used in pharmacological formulations. PEG as an excipient, provides better solubility for poorly water-soluble active ingredients, thereby playing a major role in formulating a dosage form to facilitate drug absorption. The metabolic fate of PEG (specifically PEG 400) has been well studied in mammalian systems due to its presence in a variety of different pharmaceutical formulations. PEG 400 remains in the blood for a minimum of 4hrs prior to oral administration (Tong et al., 2002). Where absorbed, *in vivo* oxidation to diacid and hydroxy acid metabolites is known to occur prior to renal excretion (Boyd Shaffer et al., 1950, Fruijtier-Pölloth, 2005, Prentice and Majeed, 1978, Hunt et al., 1982, Herold et al., 1982, Herold et al., 1989, Friman et al., 1993, Baumann et al., 2014) (**Figure 5-16**).

In this work, commercially available forms of PEG were selected and analysed by RPC-UPLC-MS (positive ion mode). These signatures were then assessed for in the ALZ urine UPLC-MS (positive ion mode) dataset. This project was specially chosen as many samples were flagged by NMR as being positive for PEG. It provided an opportunity to acquire blood samples with the newly developed lipid removal protocol (Chapter 4) and explore the potential perturbation in the blood metabolome due to PEG exposure. Using a combination of different statistical methods and data acquired from the serum samples, PEG signals that are due to contamination, and those that are a result of an external exposure, maybe differentiated. It is the former that should be removed from metabolomic workflows and the latter which should be considered as part of the everyday metabolome, when investigating health status.

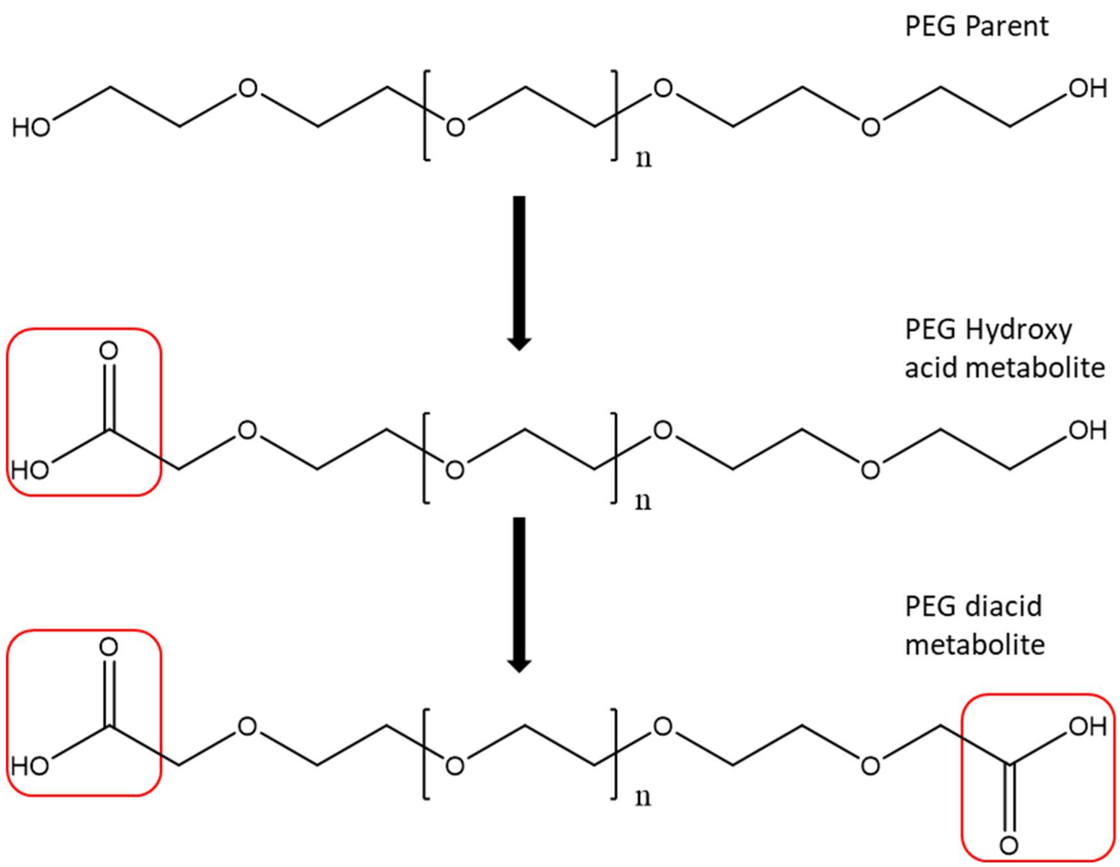


Figure 5-16. Reported PEG metabolism products (Webster et al., 2007). PEG metabolism *in vivo*, undergoes oxidation to its two major metabolites, a hydroxy acid and a diacid form. Both have been reported in urine and blood products.

5.5.2 Materials and additional methods

5.5.2.1 Materials

Reference standards for PEG 400, 600, 1000, 2000, 3350, 4000, 6000 and 8000 were purchased from Toronto Research Chemicals Inc., and SIGMA-ALDRICH. Reference standards were prepared as discussed in Chapter 3. Mass spectral analysis was performed using the 1:100 dilution in RPC-UPLC-MS (positive ion mode only).

5.5.2.2 Validation study

A study to observe PEG metabolism was undertaken. A urine sample was collected from three volunteers and their metabolic profile obtained using phenotyping protocols as described by Lewis *et al.* This step was to ensure that the urine samples were free of PEG signals. The volunteers were then subjected to consumption of a PEG containing OTC gel capsule. The capsule contained PEG 400. Metabolic profiles were then obtained for three sets of samples:

1. An aqueous extraction of a typical over-the-counter pharmaceutical gel capsule containing PEG

2. Urine collected prior to consumption of the gel capsule (control)

3. Urine collected within 24 hours of consumption of the gel capsule.

5.5.2.3 Data analysis

5.5.2.3.1 Univariate correlation analysis

The details for all univariate correlation analysis implemented for this investigation can be found in Chapter 3, namely, intersample correlation (Spearman), intrasample correlation (Pearson) and i-STOCSY (Pearson). Bonferroni multiple testing correction is applied to the p-values associated with each intersample correlation test, defining statistical significance (p_{adj}-value ≤ 0.05). The i-STOCSY tool was also used to explore correlations between the profiling data and the same compliance medication data that accompanied the ALZ dataset, used in the example application in Chapter 3.

5.5.2.3.2 Multivariate analysis

Multivariate analysis of phenotyping data was performed SIMCA-P v.14.1 (Umetrics, Umeå, Sweden). Principal component analysis (PCA) was used for the detection of outliers and to display any general trends and clustering observed in the data. Partial-least-squares regression (PLS-R) modelling was used to identify statistically significant covariation between a set (X) of independent variables (MS features) and the corresponding (Y) response (PEG). PCA and PLS models were performed on mean-centered and unit variance-scaled (MC-UV) data. The quality of the PLS models were validated by a seven-fold internal cross-validation, and assessment of the variance explained (R^2Y) and predictive ability (Q^2Y) of the model. The number of components for each model was selected to optimise for model quality and avoid over-fitting. Finally, variable selection for metabolic identification were based on variable importance for the projection (VIP) values, summarising the contribution the variables make to the overall model. Only variables with VIP values ≥ 1.5 were considered for metabolite identification.

5.5.3 *Results and discussion*

5.5.3.1 Acquisition of reference standards and initial observations in urine data

Reference solutions for various PEG mixtures were made according to the protocol set in Chapter 4 and acquired on RPC-UPLC-MS (positive ion mode). PEG 400, 600, 1000 exhibited highly resolved chromatographic peaks with some overlap in RT. The mass range acquired for typical profiling studies are between 50 and 1200 Da, so as a result, PEG mixtures which exceeded 1000Da suffered in chromatographic resolution resulting in large broad peaks. The TIC for each PEG form is summarised in **Figure 5-17**.

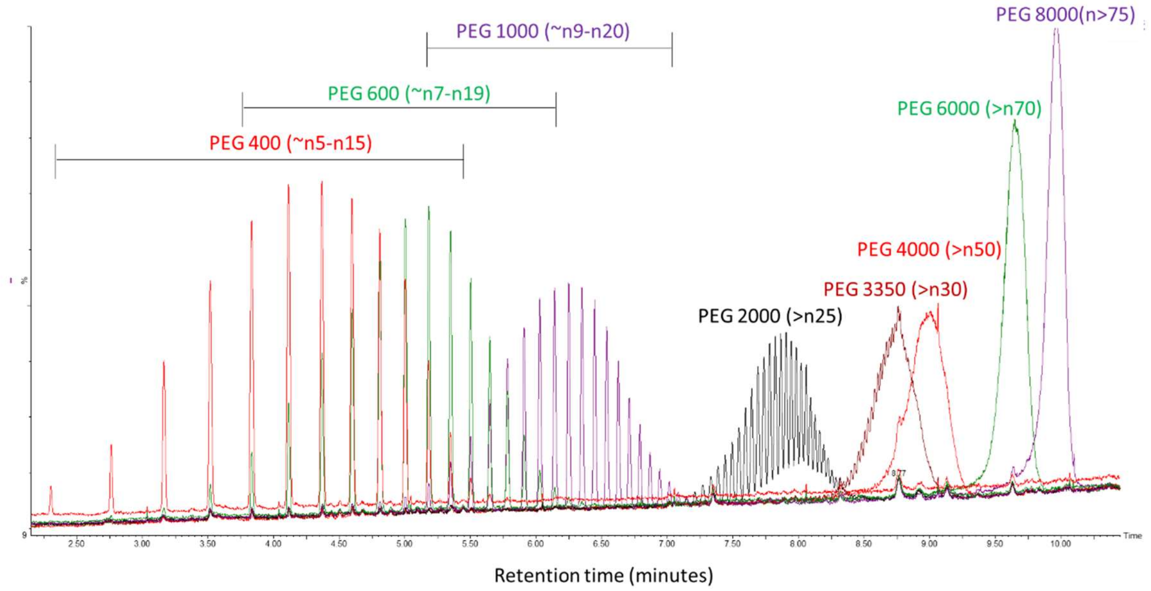


Figure 5-17. Reference standard of the different PEG forms acquired by RPC-UPLC-MS (positive ion mode). The reference standards for the PEG forms; PEG 400, 600, 1000, 2000, 3350, 4000, 6000 and 8000 were acquired using the same profiling methods as sample acquisitions. PEG forms of a lower polymeric length exhibited highly

resolved peaks. As the polymeric length increased (starting at PEG2000), resolution became poorer resulting in large broad peaks.

Part of the workflow for the analysis of urine samples at the NPC involves NMR sample acquisition prior to LC-MS. This serves two purposes, 1.) a complementary acquisition by a secondary analytical platform and 2.) to screen for species like PEG and protein, which would otherwise cause ionisation interferences for MS based acquisition. The samples which are flagged are subsequently removed and in addition, removed in the making of the QC study pool, which for MS based analysis is crucial as it is used for filtering and batch correction purposes. However, as NMR is less sensitive analytically to MS, PEG containing samples are not entirely removed, as was the case in the ALZ urine RPC dataset. Multiple metabolic features were observed in the ALZ urine RPC dataset, that exhibited a repeating polymer pattern that differed by 44 Daltons. These features were identified as PEG with a varying polymer length of n5-n16, matching the LC-MS signature observed from the PEG400-600 reference standards. Further, polymer patterns alongside the PEG parent were also observed and the accurate mass of these features matched perfectly to the oxidised and di-oxidised form of PEG. These forms of oxidised PEG have previously been reported as the major metabolites in mammalian studies (Fruijtier-Pölloth, 2005) (Boyd Shaffer et al., 1950, Prentice and Majeed, 1978, Baumann et al., 2014, Friman et al., 1993, Herold et al., 1989, Herold et al., 1982, Hunt et al., 1982). As there are many spectral features corresponding to PEG, one representative ion was which did not exhibit saturation was chosen for all subsequent statistical analysis. The molecular ion of a PEG form with eight ethylene glycol monomers (PEG(n8)-unmetabolised) was selected, i.e. *m/z* 388.2540. This feature was used as the driver feature in a intersample correlation analysis (Chapter 3), and statistically significant correlates included features corresponding to ionisation products (i.e. isotopes, adducts and in-source fragments) and to multiple isotopes and adducts of the acid (PEG(n8)-COOH) and diacid (PEG(n8)-2xCOOH) metabolites (accurate mass of within 3 ppm), as illustrated in **Figure 5-18**. For example, the correlation coefficient observed with the molecular ion of PEG(n8)-COOH and PEG(n8)-2xCOOH, were 0.93 and 0.90, respectively. This correlation suggests a strong relationship between the concentration of PEG(n8) in the study samples, to its two putatively annotated metabolic products. The intensity of the acid metabolite was also generally higher than the diacid metabolite. There were no other correlated features present in the urine data.

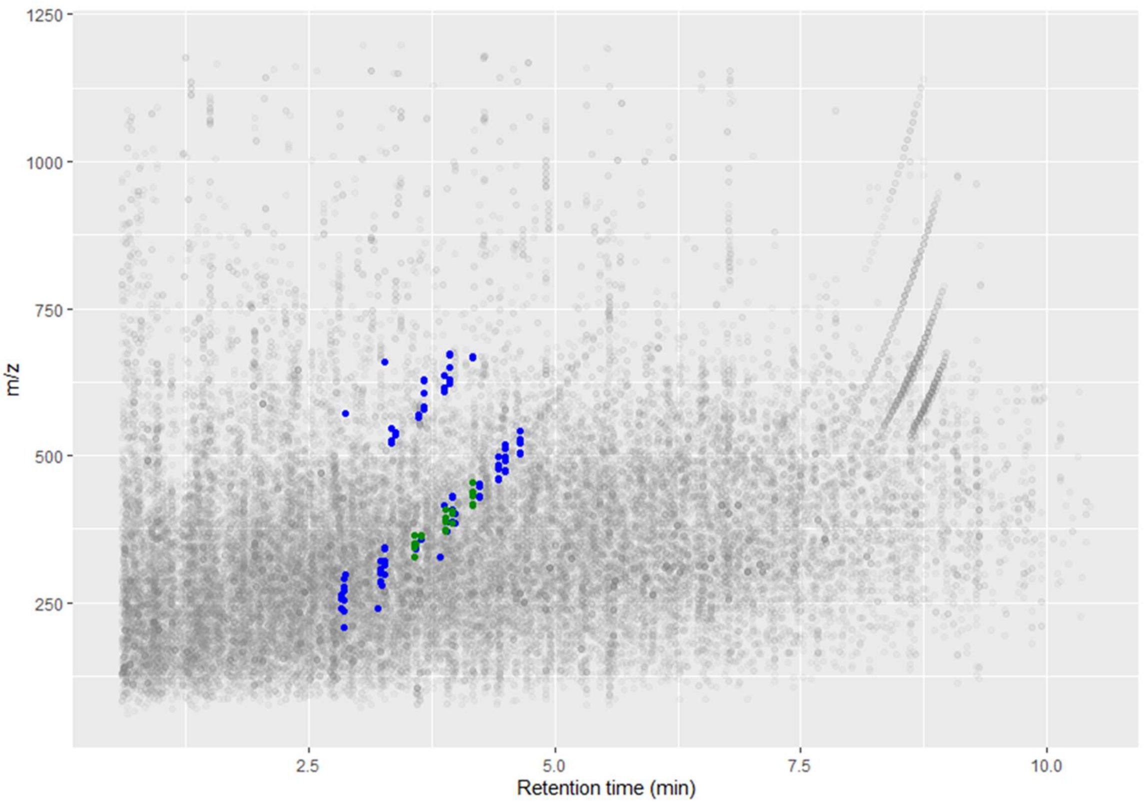


Figure 5-18. Intersample correlation using PEG(n8)-unmetabolised as the driver in ALZ urine analysed by RPC-UPLC-MS (positive ion mode). Results from the intersample correlation were presented as a retention time (RT) vs *m/z* plot, where features that correlated to the driver feature (PEG(n8)-unmetabolised), were coloured by statistical significance ($p_{adj} \leq 0.05$)- blue, and a statistically significant correlation coefficient greater than 0.7 (i.e. Spearman: r >0.7 and $p_{adj} \leq 0.05$) as green. All other detected features are coloured grey. Correlated features included ionisation products of PEG(n8)-unmetabolised, and features corresponding to acid (PEG(n8)-COOH) and diacid (PEG(n8)-2xCOOH) metabolites.

5.5.3.2 PEG metabolism and renal excretion validation study

In this study, volunteers were exposed to PEG 400. As 400 represents an average molecular weight, the number of ethylene oxide units lies in the range between n8 to n13. All three volunteers produced the same result. Extracted ion chromatograms for features corresponding to PEG n8 – n13 of the three different sets of samples (specified in the method) for one of the volunteers is illustrated in **Figure 5-19**. The gel capsule profile only showed evidence of the molecular ion associated with the unmetabolised form of PEG. The mass spectrometry profile for the control urine samples contained low, background traces of PEG(n8-n13)-unmetabolised, PEG(n8-n13)-COOH and PEG(n8-n13)-2xCOOH. In the urine sample collected following the self-consumption of an over-the-counter gel capsule, metabolite profiles contained PEG and the two metabolites at high concentrations, suggesting absorption, metabolism, and renal excretion of PEG compounds. The findings suggest that consumption of the PEG containing gel capsule results in renal excretion of both PEG unmetabolised and PEG-COOH and PEG-2xCOOH metabolites, thereby supporting the findings in the ALZ urine cohort above.

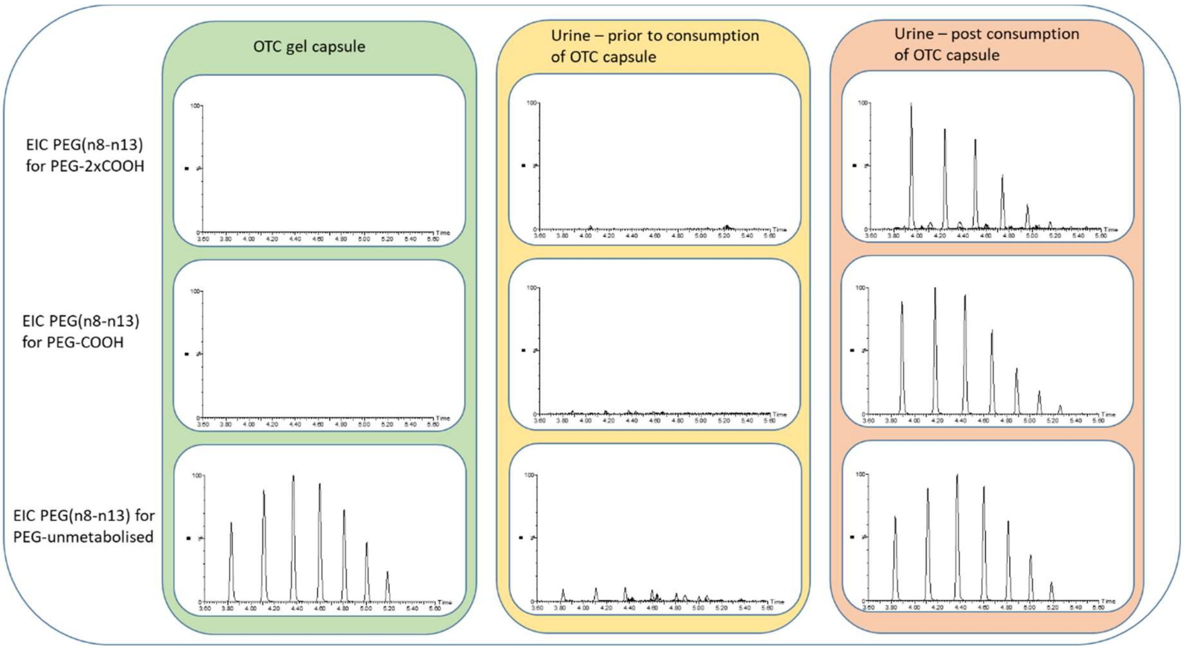


Figure 5-19. EIC of PEG (n8-n13) unmetabolised, PEG (n8-n13)-COOH and PEG (n8-n13)-2xCOOH in a urine sample of a volunteer, pre/post consumption of PEG containing OTC gel capsule analysed by RPC-UPLC-MS (positive ion mode). EIC profiles of PEG(n8-n13) in a gel capsule extraction (green), a first pass morning urine sample taken as a control (yellow), and a urine sample taken within 24 hours following consumption of a PEG containing gel capsule (orange). All chromatograms are scaled to the same intensity on the Y-axis. Chromatographic features corresponding to the PEG polymer (n8-n13) can be observed in an example gel capsule extract, and a urine sample collected post consumption of a gel capsule. Chromatographic features for PEG oxidised acid metabolite PEG (n8-n13)-COOH and diacid metabolite PEG (n8-n13)-2xCOOH can be observed only in urine collected post consumption of a gel capsule. First pass urine only contains trace levels of both parent PEG and PEG-COOH.

5.5.3.3 Investigation of PEG metabolism in ALZ serum

5.5.3.3.1 Serum profiling by lipid removal protocol

At the time of writing, blood samples analysed at the NPC are not subjected to a preliminary screening process for PEG. A HILIC LC-MS assay employed at the NPC was previously run on these samples highlighting the presence of large PEG signatures, however PEG metabolites were not present in this assay. Using the developed lipid removal protocol, all serum samples were prepared and acquired by RPC-UPLC-MS (positive ion mode). Data was pre-processed using XCMS software in accordance with NPC workflows, resulting in 2499 detected metabolite features. Repeated observation of specific reference features from pooled QC samples throughout an analytical batch demonstrated mean retention time RSD <0.3% and mean peak area RSD <10% with no post normalisation, except for the labelled creatinine (**Table 5-6**). No obvious drifts or outliers were observed in TIC for both the SR and LTR samples (**Figure 5-20.A**). The distribution of the % RSD in relation to the feature intensity in the SR samples, is illustrated in **Figure 5-20.B**. The distribution is further divided into a lower quartile range (green), interquartile range (blue) and upper quartile range (green), highlighting the precision of features based on their measured signals. The median RSD values for the ALZ study was 20.3%.

Table 5-6. Retention time and peak area precision of reference standards within the ALZ serum project analysed by RPC-UPLC-MS (positive ion mode). Repeated observations of reference features from the pooled QC samples throughout the analytical batch demonstrated high precision with mean retention time RSD < 1% and mean peak area RSD <20% with no post batch correction required.

NPC Project	ALZ Serum RPC-UPLC-MS (+)			
Metric	RT %RSD		Area %RSD	
QC Sample	SR	LTR	SR	LTR
L-Glutamine-$^{13}C_5$	0.00	0.70	7.98	7.48
L-Glutamic acid-$^{13}C_5$	0.00	0.00	9.13	8.24
Creatinine-(methyl-d_3)	0.00	0.33	18.91	21.45
L-Isoleucine-13C6,15N	0.39	0.39	7.67	7.78
L-Leucine-$^{13}C_6$	0.41	0.37	7.64	7.89
L-Tryptophan-$^{13}C_{11}$, $^{15}N_2$	0.33	0.35	6.54	6.86
Cytidine-5,6- d_2	0.40	0.61	5.25	6.78
L-Phenylalanine-$^{13}C_9$, ^{15}N	0.33	0.33	7.63	7.21
N-Benzoyl- d_5-glycine	0.16	0.17	7.08	7.69

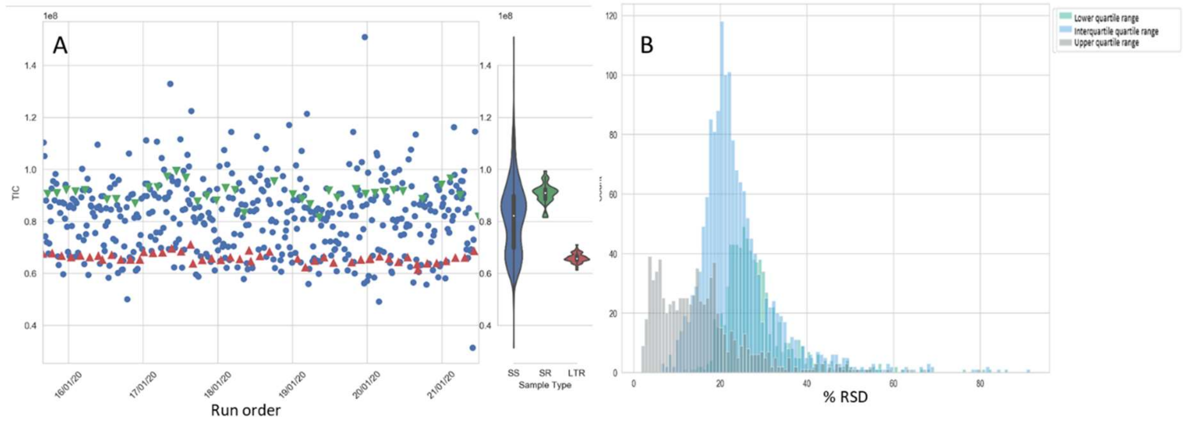


Figure 5-20. RPC-UPLC-MS (positive ion mode) TIC of all plasma samples (Study samples – blue, SR – green, and LTR – red) in ALZ (A), and the % RSD distribution for all features passing the dilution series filter (B). Alongside the TIC scatter plots are violin plots exhibiting the TIC density for each sample type. The distribution plots are %RSD segmented by mean feature intensity into quartiles.

(A) TIC of all samples in ALZ (positive ion mode) against the run order;

(B) % RSD distribution in positive ion mode. Median RSD value was 20%;

Data from the ALZ serum study (n=449) resulted in 2499 detected metabolite features. Repeated observation of the pooled sampled (SR) throughout the analytical batch demonstrated high precision, with the majority of features that occupied the interquartile and upper quartile intensity range having an RSD of less than 30%. The TIC plots exhibited no major outliers or trends with respect to the QC samples.

5.5.3.3.2 PEG exploration in serum samples

Intersample correlation

The presence of PEG in biofluid samples can result from both an environmental exposure (for example, from medicinal use), or from contamination during sample handling. To evaluate the potential biochemical impact of PEG exposure, samples contaminated with PEG had to be first removed from the dataset to avoid skewing the findings. Most of the urine samples positive for PEG were originally excluded by the NMR screening process, and of the samples remaining, differentiating between an exposure or a contamination was not particularly obvious. As no samples were excluded from serum prior to profiling, exploration of PEG exposure/contamination was undertaken by firstly exploring the distribution of PEG associated signals in the serum dataset. The density plot (as illustrated in **Figure 5-21**) represents the distribution of PEG(n8)-unmetabolised signal intensities detected in all serum samples. A clear bimodal distribution exists in the data.

PEG(n8)-unmetabolised was detected at basal levels in all samples, however the relative intensity was significantly higher in the group on the right. The two groups from the density plot was therefore labelled as PEG-Low (left) and PEG-High (right).

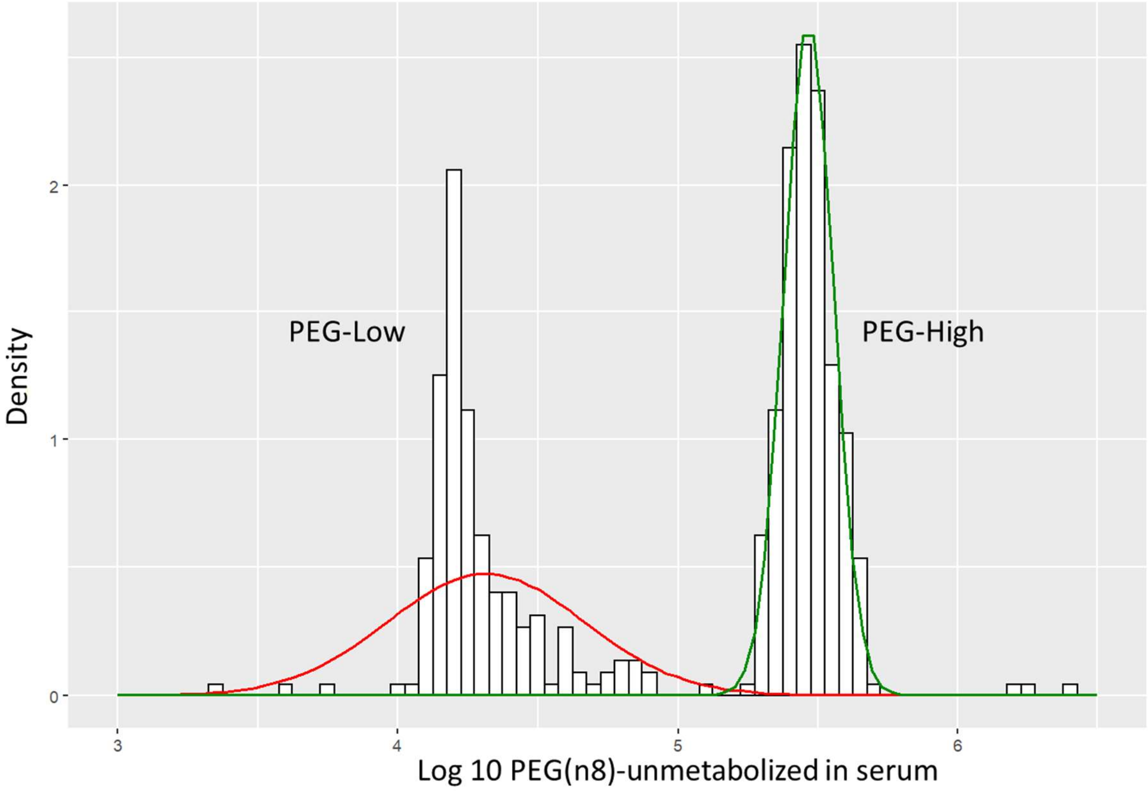


Figure 5-21. The distribution range of signal intensity associated with PEG(n8)-unmetabolised, observed in the ALZ serum analysed by RPC-UPLC-MS (positive ion mode).

A bimodal distribution PEG(n8)-unmetabolised was observed in ALZ, **and Gaussian mixture models (GMM's) were fitted to the MS intensity distribution** essentially splitting the data into a PEG-Low and PEG-High group. PEG(n8)-COOH and PEG(n8)-2xCOOH were largely absent from samples in the PEG-High group. These samples were subsequently removed to form a new "amended" dataset.

The features associated with PEG(n8)-COOH and PEG(n8)-2xCOOH were largely absent from samples in the PEG-High group, as such, these samples can be attributed to a significant PEG contamination and were therefore flagged for exclusion from subsequent data analysis. The exclusion of these samples required an approximate LOD for PEG(n8)-COOH to be estimated. The LOD of only the acid metabolite was evaluated, because like urine, the serum exhibited generally higher signals than the diacid metabolite. The LOD involved approximating an arbitrary intensity threshold, that corresponded to an intensity greater than five times the signal to noise. All samples within the PEG-

High group which showed no evidence of PEG(n8)-COOH, i.e. less than the threshold, were removed. This resulted in a new “amended” dataset, which contained approximately 232 samples (from 449), 45 of which were positive for PEG(n8)-COOH. An intersample correlation was then undertaken using PEG(n8)-unmetabolised as the driver (**Figure 5-22**). Significant correlates included the two oxidised metabolites, PEG(n8)-COOH (Spearman, r = 0.76, p <2.2e^{-16}) and PEG(n8)-2xCOOH (Spearman, r = 0.57, p <2.2e^{-16}). Three other feature groups (Feature Group 1, Feature Group 2 and Feature Group 3 – as in **Figure 5-22**) were also identified with statistically significant correlations to the driver, and to one another. To further demonstrate that the eliminated samples from the analysis are products of PEG contamination and have no relation to exposure, the samples from the PEG-High group which were removed, was also subjected to an intersample correlation analysis (Figure can be found in Appendix 3). As suspected, the only correlates to the PEG(n8)-unmetabolised driver, were features related to the ionisation products of the PEG(n8) molecule (i.e. isotopes, adducts and in-source fragments).

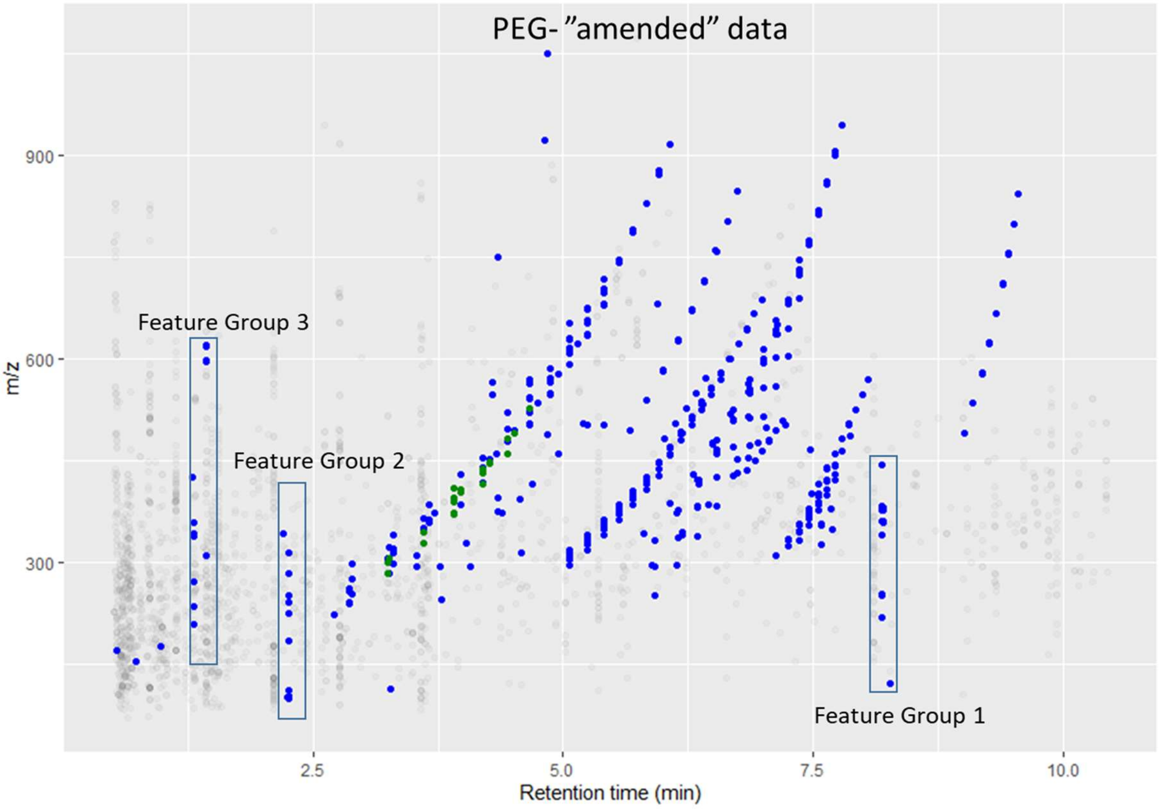


Figure 5-22. Intersample correlation using PEG(n8)-unmetabolised as the driver, observed in the ALZ serum analysed by RPC-UPLC-MS (positive ion mode). Results from the intersample correlation were presented as a retention time (RT) vs *m/z* plot, where features that correlated to the driver feature (PEG(n8)-unmetabolised), were coloured by statistical significance ($p_{adj} \leq 0.05$)- blue, and a statistically significant correlation coefficient

greater than 0.7 (i.e. Spearman: r >0.7 and $p_{adj} \leq 0.05$) as green. All other detected features are coloured grey. Correlated features included ionisation products of PEG(n8)-unmetabolised, features corresponding to acid (PEG(n8)-COOH) and diacid (PEG(n8)-2xCOOH) metabolites and three unknown feature groups, which will be the subject of metabolic identification efforts.

Multivariate analysis using PCA and PLS-R

Multivariate models were used to establish if the same feature groups highlighted from the correlation analyses, together relate to PEG exposure. The heavy presence of PEG in ALZ serum was immediately observed with two tightly clustered groups in the PCA model. PC1 and PC2 are shown in the score plot (**Figure 5-23.A**), accounting for 25% and 6% of the total dataset variance (R^2X). The loadings (coloured by the weighting values scaled as correlation coefficients – p_{corr}) highlighted the polymeric pattern associated with PEG as the major source of variance between these two clusters. The samples within these two clusters are consistent with the samples from the PEG-Low and PEG-High groups observed in the density plots above. Feature group 1 and Feature group 2 were also observed as underlying drivers in the data (**Figure 5-23.B**). PLS modelling was then used on the amended dataset to identify statistically significant covariation between MS features and the corresponding Y response. Four PLS models were made using different Y response variables; PEG(n8)-unmetabolised, and features representing the molecular ion (deduced from the intrasample correlation) from each of the feature groups. The same feature groups, as observed from the intersample correlation above, replicated across the four models as illustrated in the PLS loadings plot (**Figure 5-24.A-D),** where features with VIP values ≥ 1.5 was used as a cut-off point for feature selection. There were other feature groups with VIP values > 1.5, a decision was made to concentrate on only the consistent feature groups between the univariate and multivariate analyses.

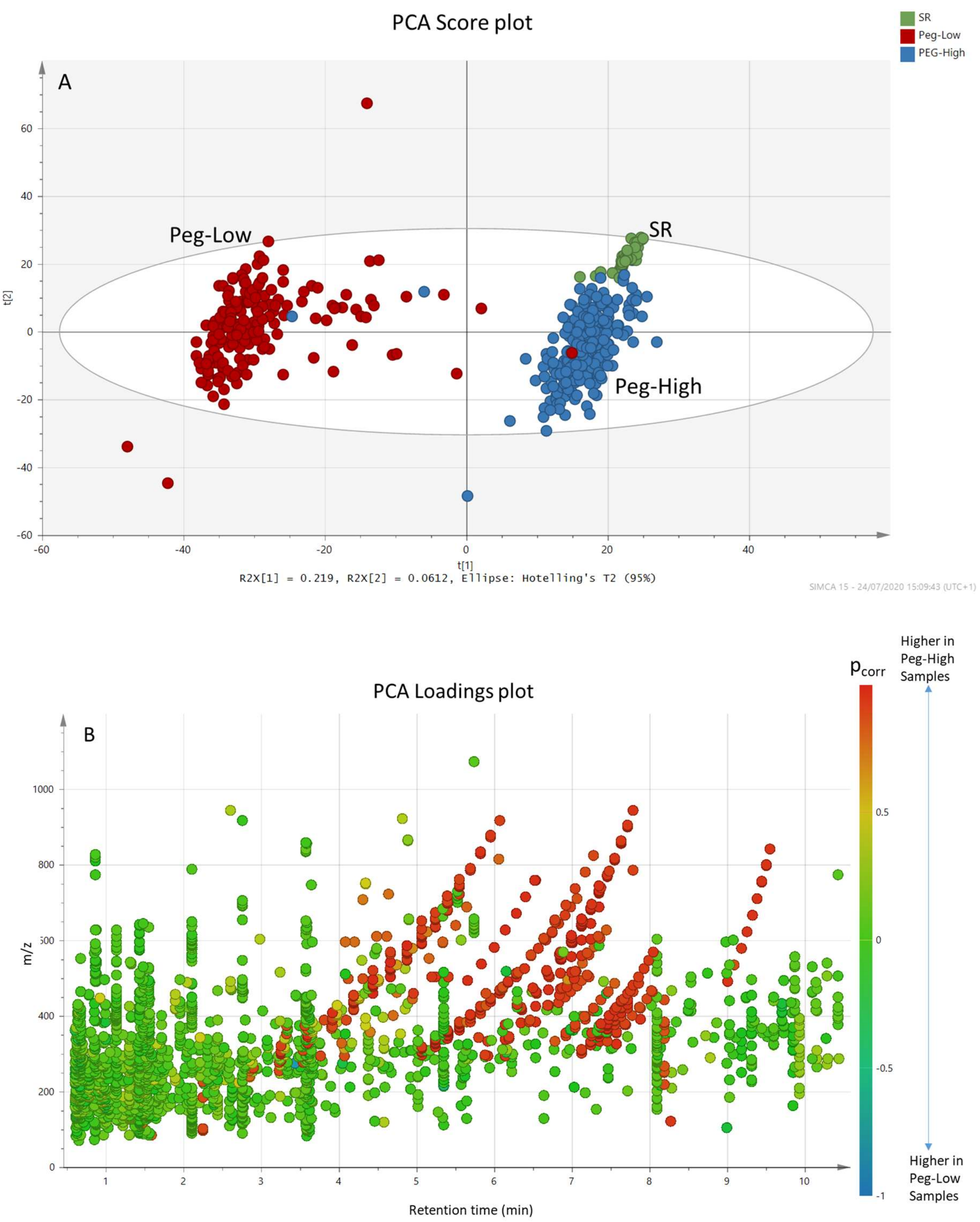


Figure 5-23. PCA Score and loadings plot on samples from the amended dataset for ALZ serum. (A) A score scatter plot, coloured by sample type, PEG-low (red), PEG-High (blue) and SR (green); (B) Loadings plot, coloured by p_{corr}.

An unsupervised PCA model demonstrates two clear groups along PC1. The loadings indicate that the features, associated with PEG, are the main source of variance driving the separation between these two groups (PEG-Low and PEG-High). Two unknown clusters of features not related to PEG were also observed in the loadings plot (at 2.25 minutes and 8.18 minutes).

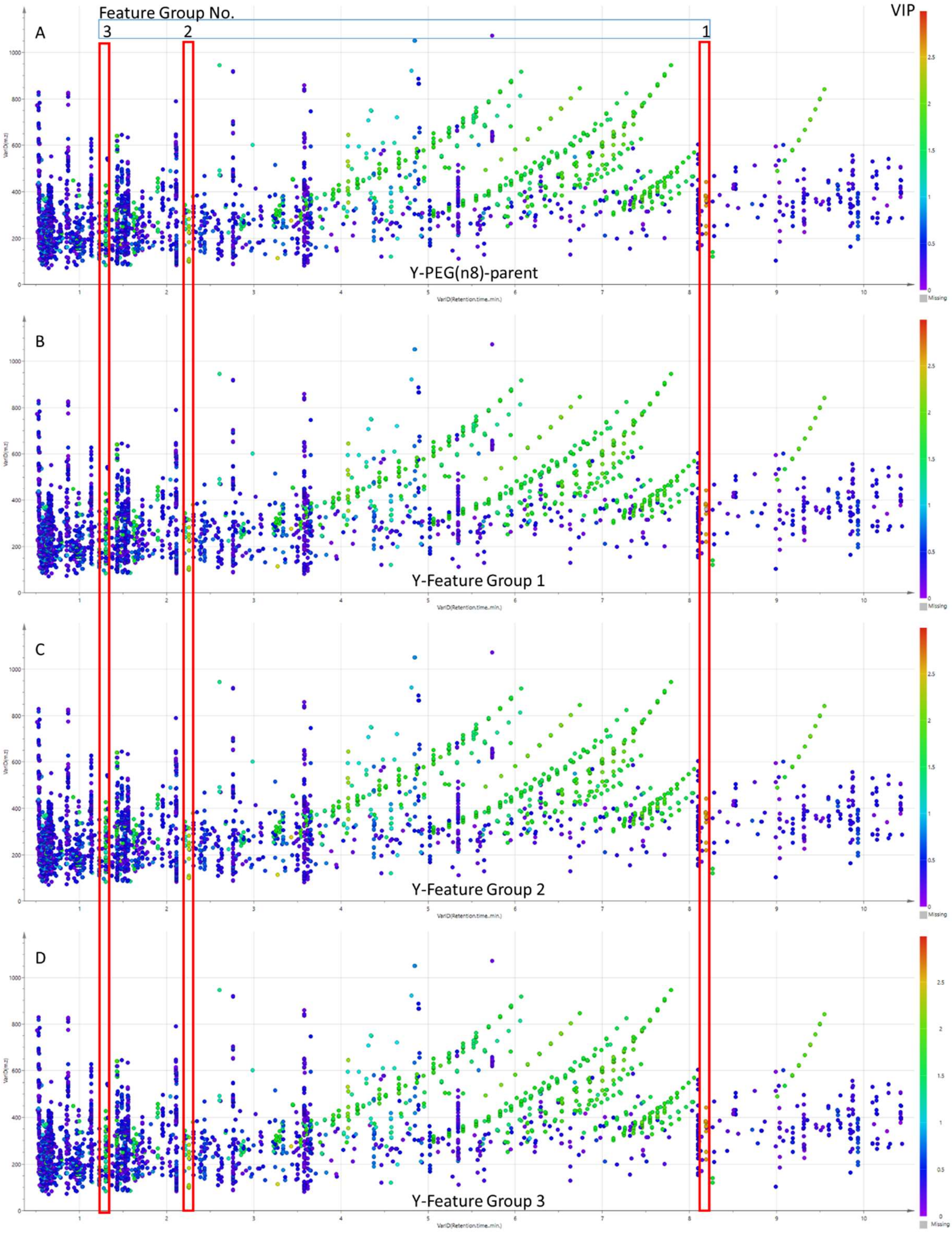


Figure 5-24. PLS regression (PLS-R) models (loading plots) from the amended ALZ serum dataset, coloured by VIP using a single y-response.

(A) PLS model where the Y response variable is the molecular ion associated with PEG(n8)-unmetabolised;

(B) PLS model where the Y response variable is the molecular ion associated with Feature Group 1;

(C) PLS model where the Y response variable is the molecular ion associated with Feature Group 2;

(D) PLS model where the Y response variable is the molecular ion associated with Feature Group 3;

PLS-R captured the relationship between additional phenotypic information (PEG exposure) and metabolite concentrations (MS features). The three feature groups (coloured by the light green and correspond to a VIP > 1.5) are consistent between the four models thereby validating the findings from the univariate correlation analyses.

i-STOCSY

To explore potential sources for PEG exposure, i-STOCSY was implemented using the profiling dataset and reported medication, as demonstrated in Chapter 3. Three drugs were flagged (**Figure 5-25**) using the amended dataset and PEG(n8)-unmetabolised as the driver feature. The strongest correlation observed, were to the drugs buflomedil, nifedepine and sulpride (correlation value of 0.44). Although without any calculation of p-values accompanying these correlation coefficients, the statistical significance of these values cannot be determined. The i-STOCSY tool for this investigation should only be used as an approximate indicator to correlates that maybe observed in the profiling data. For a more comprehensive calculation on correlation, an intersample correlation analysis should be carried out once the features corresponding to a drug has been assigned. Furthermore, the three drugs highlighted from i-STOCSY, does not necessarily mean that each drug has a relationship to PEG. The i-STOCSY tool not only looks for correlations between the medication and profiling datasets, but also highlights the correlations within each dataset separately. Buflomedil and nifedipine are vasoactive drugs which can be both used to treat high blood pressure and peripheral arterial disease (Chacón-Quevedo et al., 1994). It is therefore not surprising that they correlate to one another. According to the European Medicines Agency (https://www.ema.europa.eu/en) common formulations for all three-drugs mention PEG as part of their film coating, however as we are not privy to the exact brand that patients were administered, there is no way of knowing the exact PEG form.

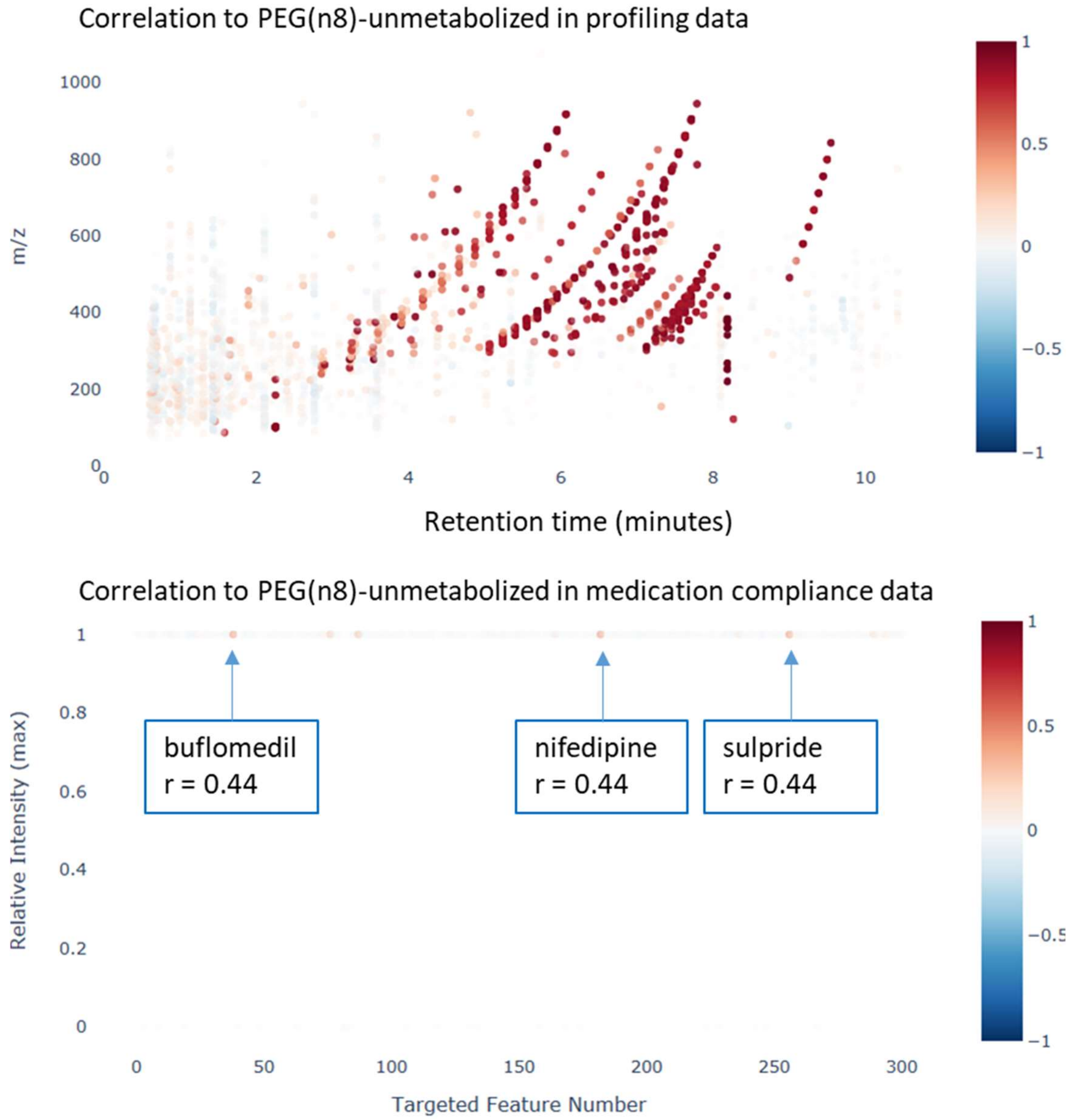


Figure 5-25. i-STOCSY output plots highlighting correlations between the PEG(n8)-unmetabolised feature, to all other features within the profiling dataset (ALZ serum RPC-UPLC-MS in positive ion mode) , and within the medication compliance dataset. The results are presented as two interactive plots in which driver feature (PEG(n8)-unmetabolised) was selected and correlated features coloured by the strength of the correlation. The dark red features indicate a correlation coeffecent close to 1 and therfore strongly correlated to the driver. Features which are whitw and greyed out, represrent no correaltion associationm, and finally features which are dark blue inidcate a strong negative correlation. In the profiling data (top plot), strongly correlated features indicate the ionisartion products of PEG(n8)-unmetabolised, acid (PEG(n8)-COOH) and diacid (PEG(n8)-2xCOOH) metabolites, and the Feature group 1 and Feature Group 2. In the medication data (bottom plot), the strongest correlation were to the drugs buflomedil, nifedepine and sulpride (r = 0.44).

5.5.3.3.3 Correlation between biofluids

Correlations between serum samples, and its corresponding urine pair were evaluated, i.e. PEG(n8)-unmetabolised in serum, against both PEG(n8)-COOH and PEG(n8)-2xCOOH in urine. The correlation was undertaken using the log transformed intensities of these PEG compounds, from samples in the amended dataset. A positive statistically significant correlation (spearman) was observed between PEG(n8)-unmetabolised in serum and the two metabolites in urine (**Figure 5-26**), suggesting a significant relationship between circulating plasma PEG and excreted urinal PEG metabolite.

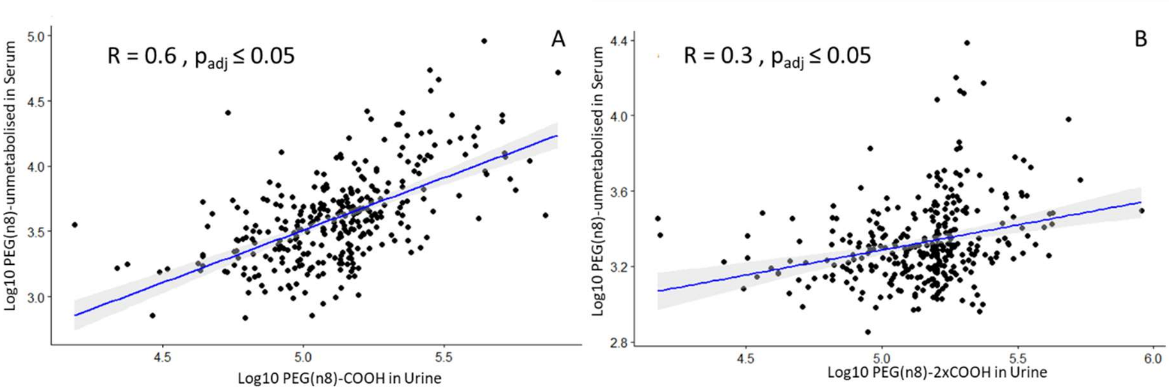


Figure 5-26. Correlation of the PEG signals detected in ALZ serum (RPC-UPLC-MS positive ion mode), to PEG signals in ALZ urine (RPC-UPLC-MS positive ion mode). A. Scatter plot demonstrating the correlation between the intensity level of PEG(n8)-unmetabolised in serum and PEG(n8)-COOH in urine; **B**. Scatter plot demonstrating the correlation between the intensity level of PEG(n8)-unmetabolised in serum and PEG(n8)-2xCOOH in urine.

A linear regression line fitted to the scatter plot that illustrates a positive statistically significant correlation between PEG(n8)-unmetabolised in serum to PEG(n8)-COOH (r= 0.6, $p_{adj} \leq 0.05$), and PEG(n8)-COOH (r= 0.3, $p_{adj} \leq 0.05$), thereby suggesting a significant relationship between circulating serum PEG and excreted urinal PEG metabolite.

5.5.3.4 *Metabolite identification in serum samples*

A total of three feature groups were identified from PLS and the intersample correlation analysis. This molecular ion from each feature group (now labelled Metabolite 1, Metabolite 2, and Metabolite 3) was used as the target ion for MS/MS acquisitions. A description of the MetID efforts for each metabolite are provided below.

Metabolite 1 (8.18 minutes, *m/z* 359.1495)

Fragments corresponding to a toluene ion (*m/z* 91.055) and a benzaldehyde ion (*m/z* 105.03) were present from the MS/MS. These fragments suggest that a Nitrogen atom may not be present in the overall molecule. MassLynx™ elemental composition software predicted a chemical formula C20H22O6, which was subsequently inputted into the online database Pubchem. There are two possible candidates based on the fragmentation data and molecular formula: triethylene glycol dibenzoate (TGD) and dibenzylidene D-sorbitol (DBS). Out of the two, only a reference standard for DBS was commercially available. Acquisition of TGD by RPC-UPLC-MS revealed a chromatographic peak at 10 minutes (not 8.18 minutes) and a different fragmentation pattern (**Figure 5-27**) thereby excluding it as a potential candidate. DBS is therefore only a putative annotation (MSI level 4), as no reference standard was available to confirm its identity. DBS has been used in a variety of different applications including as an active excipient for transdermal pharmaceutical compositions that enables high drug release (Okesola et al., 2015).

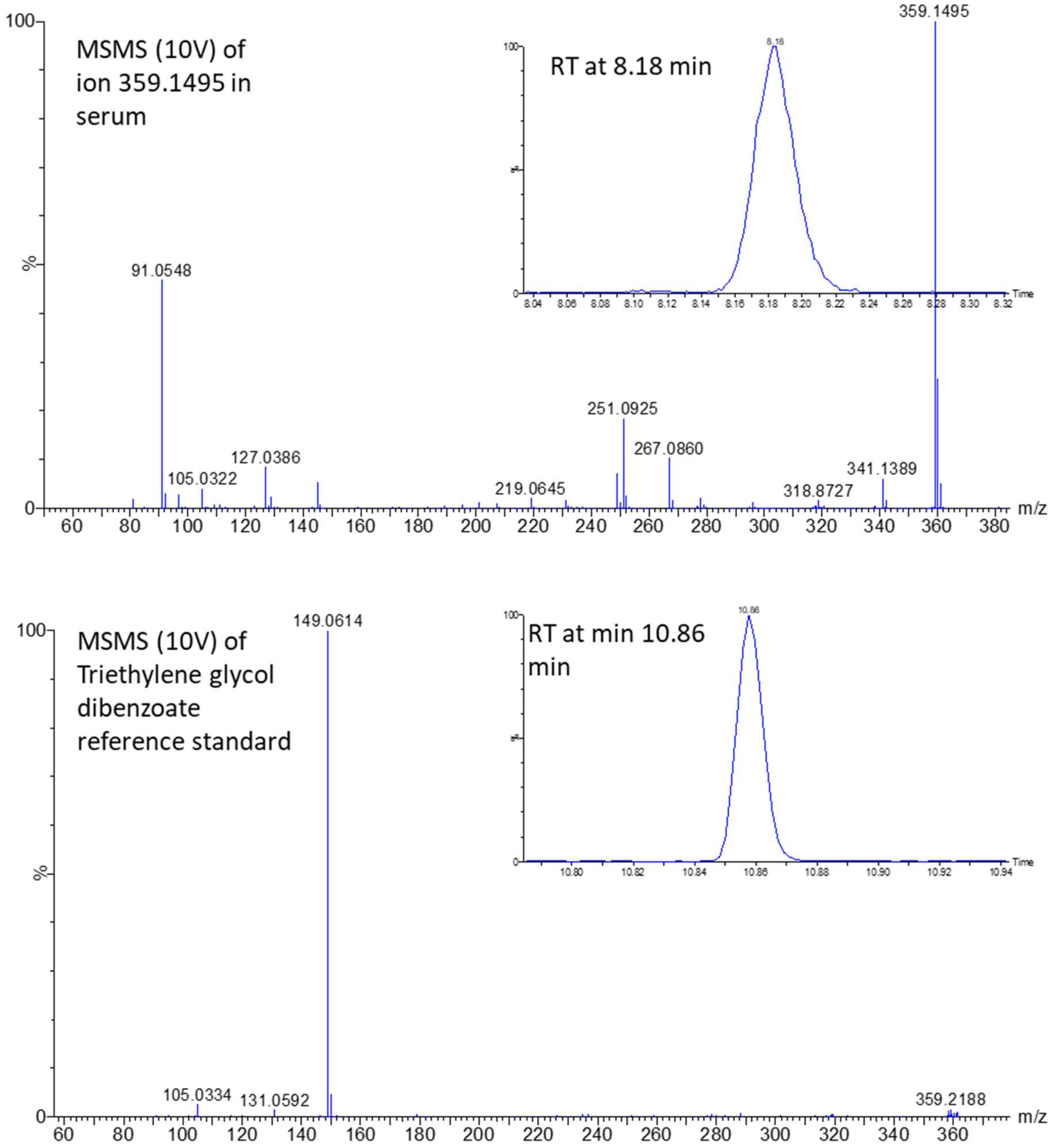


Figure 5-27. Metabolite identification efforts to compare Retention time and MS/MS of Metabolite 1, to the MS/MS of a candidate reference standard triethylene glycol dibenzoate analysed by RPC-UPLC-MS (positive ion mode) chromatographic conditions. The MS/MS fragmentation spectra and retention time of Metabolite 1 observed in the profiling data did not match the candidate reference standard of triethylene glycol dibenzoate. No identification could be made for this feature.

Metabolite 2 (2.25 minutes, *m/z* 100.077)

The molecular ion for Metabolite 2 was inputted into the online database HMDB. The first hit was piperidone. There are two isomeric forms available, 2-piperidone and 4-piperidone, and both are used in the manufacture of pharmaceuticals such as Fentanyl (Valdez et al., 2014). From the literature, both these compounds undergo hydroxylation metabolism (Cheng et al., 2013). An intersample correlation using this feature as the driver, highlighted a statistically significant correlate corresponding to the monoisotopic mass (within 1 ppm) of the hydroxyl metabolite of piperidone (Spearman, r = 0.60, $p_{adj} \leq 0.001$). Upon closer inspection, the PCA and PLS models also highlighted the significance of the hydroxyl metabolite. An MS/MS experiment was conducted for the standard 2-piperidone by Cheng *et al,* revealing a fragment at *m/z* 82.07, matching the fragment observed in the MS/MS acquisition for this feature in ALZ serum (**Figure 5-28**). Unfortunately, no reference standards for both compounds were commercially available, so retention time could not be confirmed. Therefore piperidone (either isomer) is only a putative annotation (MSI level 2).

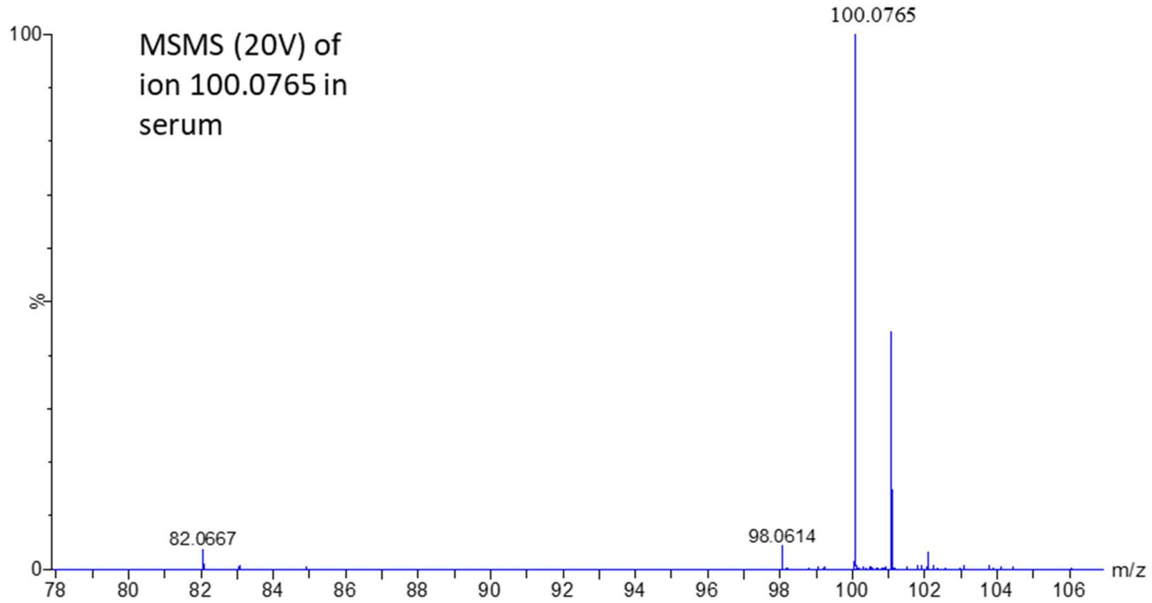


Figure 5-28. Metabolite identification efforts to compare only the MS/MS of Metabolite 2 analysed by RPC-UPLC-MS (positive ion mode) chromatographic conditions, to the MS/MS of a candidate reference compound observed in the literature. The MS/MS fragmentation spectra of metabolite 2 observed in the profiling data matched the candidate reference standard of 2-piperidone from literature. No standard was commercially available; therefore Metabolite 2 is putatively annotated (MSI level 2) as piperidone (the 2 or 4 isomer).

Metabolite 3 (1.3 minutes, *m/z* 130.0491)

Was i**dentified** as pyroglutamic acid (PGA), by comparison of the molecular ion, in-source fragment, and retention time, to an already acquired reference standard from an in-house standard database (**Figure 5-29**). The intersample correlation using PGA as the driver, revealed statistically significant correlations in the serum, to features identified from the database as glutamic acid (Spearman, r = 0.40, p_{adj}≤0.001), glutamyl threonine (Spearman, r = 0.60, p_{adj}≤0.001) and glutamine (Spearman, r = -0.30, p_{adj}≤0.001). PGA, glutamic acid, glutamyl threonine and glutamine are all involved in the glutathione cycle or gamma-glutamyl cycle, so it is not surprising that they correlate to one another (Bachhawat and Yadav, 2018).

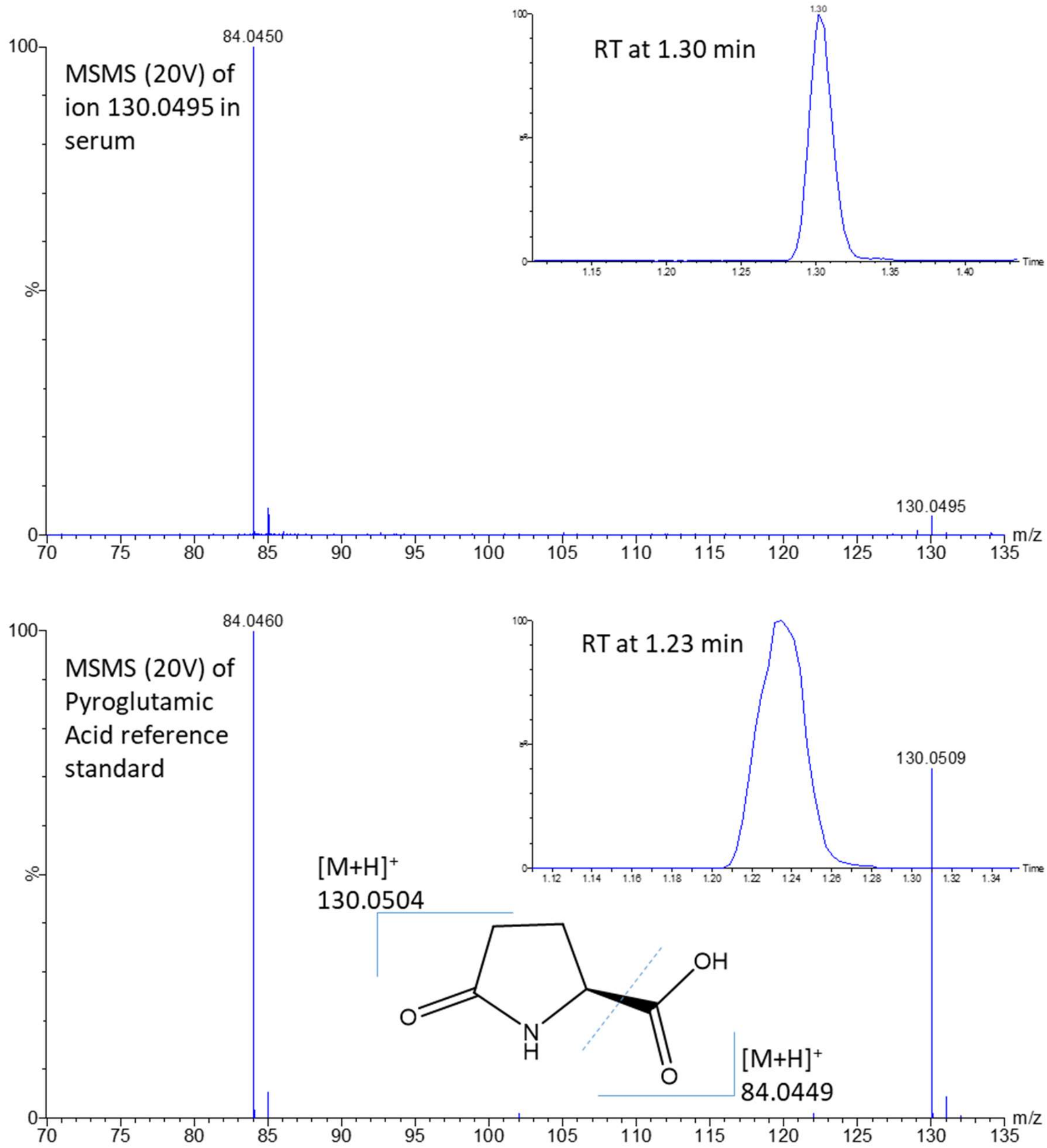


Figure 5-29. Metabolite identification efforts to compare Retention time and MS/MS for Metabolite 3, to the MS/MS of a candidate reference standard pyroglutamic acid analysed by RPC-UPLC-MS (positive ion mode) chromatographic conditions. The MS/MS fragmentation spectra and retention time of Metabolite 3 observed in the profiling data was a match to the candidate reference standard. Metabolite 3 has therefore been identified as pyroglutamic acid (PGA).

5.5.4 General discussion

5.5.4.1 PEG as a sample contaminant vs PEG as a true xenometabolome signature

Analytical laboratories have historically viewed PEG as a contaminant that enters the biological sample as a by-product of laboratory equipment and LC solvents (Hodge et al., 2013, Mihailova et al., 2006, Michopoulos et al., 2010, Weaver and Riley, 2006, Waters, 2003).

PEG presence in UPLC-MS profiling experiments is an analytical concern, as signals can potentially compromise the analytical system causing carry-over contamination issues and ion suppression effects (Weaver and Riley, 2006). One approach is for samples to be screened for PEG prior to mass spectral analysis using another technique; the NPC utilises data acquired from NMR to identify samples containing appreciable (µM – mM) PEG that allows these samples to be removed prior to MS acquisition.

Metabolomic workflows have implemented strategies to account for PEG containing samples, and computational methods such as the Kendrick mass filter has been successfully applied to dataset to filter PEG (da Silva et al., 2019). Sample preparation techniques such as SPE (Kamleh et al., 2008) have also been proposed to reduce PEG content in biological samples. but SPE (and similar) methods are not well suited for integration with high-throughput metabolic phenotyping assays.

This study suggests that not all PEG signals can be attributed to contamination. It is evident that PEG undergoes significant metabolism in humans. Evidence to support this include the correlation of urinary PEG(n8) signals with those of its primary oxidised metabolite PEG(n8)-COOH (Spearman, r = 0.93, $p<2.2e^{-16}$) in 900 urinary samples analysed by RPC-MS. Both oxidised metabolites were detected in corresponding serum samples (n=449) and similarly, a statistically significant correlation was observed between PEG(n8)-unmetabolised and PEG(n8)-COOH (Spearman, r = 0.76, $p<2.2e^{-16}$) in serum. Statistically significant correlations were also observed between PEG(n8)-unmetabolised in serum, to both metabolites in urine. This presence of PEG metabolites in different biofluids is perhaps not surprising, especially since PEG has long been used as an excipient in a range of over the prescription drug formulations (D’souza and Shegokar, 2016, Gullapalli and Mazzitelli, 2015), food products (FDA, 2014, EFSA, 2015), supplements (EFSA, 2007), infant diapers (Goodpaster et al., 2011) and cosmetics (Fruijtier-Pölloth, 2005). As a result of its ubiquitous nature, PEG products has mostly been reported to be safe (Jang et al., 2015) however despite its perceived safety, there are instances of toxicity in both human and animals (Pellegrini et al., 2013, Descamps et al., 2000, Biondi et al., 2002, Fruijtier-Pölloth, 2005, Erickson et al., 1996).

As metabolic phenotyping is now frequently applied to large clinical and epidemiology sample cohorts, eliminating PEG contaminated samples, designated for metabolite phenotyping would ensure greater data quality and applicability for retrospective data mining. As sample collection protocols are developed that are specific for metabolic phenotyping the implementation of the restriction or substitution of products that are known to contain PEG for biofluid study participants in the lead up to sample collection may be beneficial to study design. An example that could be implemented is the substitution of over–the-counter pharmaceuticals in gel capsules could be replaced with powder form alternatives that do not contain PEG. Exposure to PEG, be that a pharmaceutical excipient or elsewhere, can potentially have a negative impact on the metabolome and confound phenotyping studies. However, what is apparent is that this exposure to PEG, in the presence of PEG metabolites, should no-longer simply be disregarded as a laboratory contamination and should be considered as part of the human metabolome.

5.5.4.2 Potential biological implications

The association of PGA to PEG and the putative annotated metabolites DBS (Metabolite 1) and piperidone (Metabolite 2) were highlighted from both multivariate (PCA and PLS) and univariate approaches (correlation). When PGA is used as a driver feature in the correlation analysis, a statistically significant correlation is observed to PEG(n8)-unmetabolised (Spearman, r = 0.47, p_{adj}≤0.001), TGD (Spearman, r = 0.4, p_{adj}≤0.001), and piperidone (Spearman, r = 0.42, p_{adj}≤0.001). PGA also demonstrated statistically significant covariation with these metabolites from the PLS models.

PGA, also known as 5-oxoproline, is an intermediate organic acid formed in the production and recycling of glutathione. Elevated levels can be a marker for glutathione deficiency, which can be dangerous in humans as glutathione is a crucial antioxidant needed in ridding the body of toxins and in amino acid transport. A defect in the γ-glutamyl cycle, can cause an up-regulation of γ-glutamylcysteine synthtase resulting in a build-up of this enzyme. Conversion to glutathione is rate-limiting. As levels of this enzyme increase, it can then enter a secondary pathway in which PGA is produced. Accumulation of PGA in serum can lead to a condition known as pyroglutamic acidosis and can be an underlying cause for high anion gap metabolic acidosis (HAGMA). The major causes of HAGMA have been commonly attributed to the accumulation of lactate, ketones, urea and ingestion of toxins, however more recently, elevated levels of PGA have been linked to this condition (Spector et al., 2019). Chronic acetaminophen use and antibiotic therapies have been reported to deplete glutathione reserves, resulting in pyroglutamic acidosis (Spector et al., 2019, Croal et al., 1998, Wardell et al., 2012).

Another well documented cause for HAGMA is ethylene glycol (EG) poisoning (Latus et al., 2012). EG has been extensively studied and its metabolism similar to PEG, with the formation of the oxidised hydroxy acid (glycolic acid) and diacid metabolites (Singh et al., 2016). EG is responsible for increased osmolality after exposure, resulting in a higher concentration of EG metabolites that accumulate in serum and produces the HAGMA associated in EG poisoning. They have also been numerous reports indicating PEG as a likely culprit in HAGMA observed in patients, due to exposure from topical and intravenous use of drugs, even as far as stating the PEG form as PEG400 (Bruns et al., 1982, Laine et al., 1995). These reports do not suggest that PEG depolymerises to EG and that it is EG toxicity that is the cause of the HAGMA, but rather, the oxidised PEG metabolites react similarly to the oxidised EG metabolites, resulting in the same hyperosmolar state of the serum, and an accumulation of the oxidised acid metabolites. The exact mechanism causing the acidosis from PEG exposure is still unknown.

The results from this investigation suggest that PGA has a linear relationship to PEG, DBS and piperidone. Drug compliance metadata and i-STOCSY revealed the presence and detection of many drugs in the ALZ serum dataset. As PEG, DBS and piperidone are common excipients used in these drugs, it is therefore possible that increased levels of PEG metabolites, combined with numerous medication intake, could result in a synergistic interaction that increases PGA levels. The metabolic handling of many drugs, consumes glycine, which is integral in glutathione production (Jackson et al., 1997). A limitation in its availability may also increase PGA levels. Biologically, this could be of significance, as a rise in PGA levels may induce pyroglutamic acidosis and even potentially provide an explanation for the HAGMA that was observed from patients with PEG exposure in the studies referenced above. To my knowledge this is the first reported case of PGA association with PEG exposure. However, care should be taken in this interpretation as this observation has not been validated in other studies and therefore a limitation of this discovery. As PEG is easily detected by UPLC-MS, in future studies we can identify PEG/metabolites and evaluate if any associations truly exist with PGA (using the strategies developed in this thesis) and the glutathione cycle.

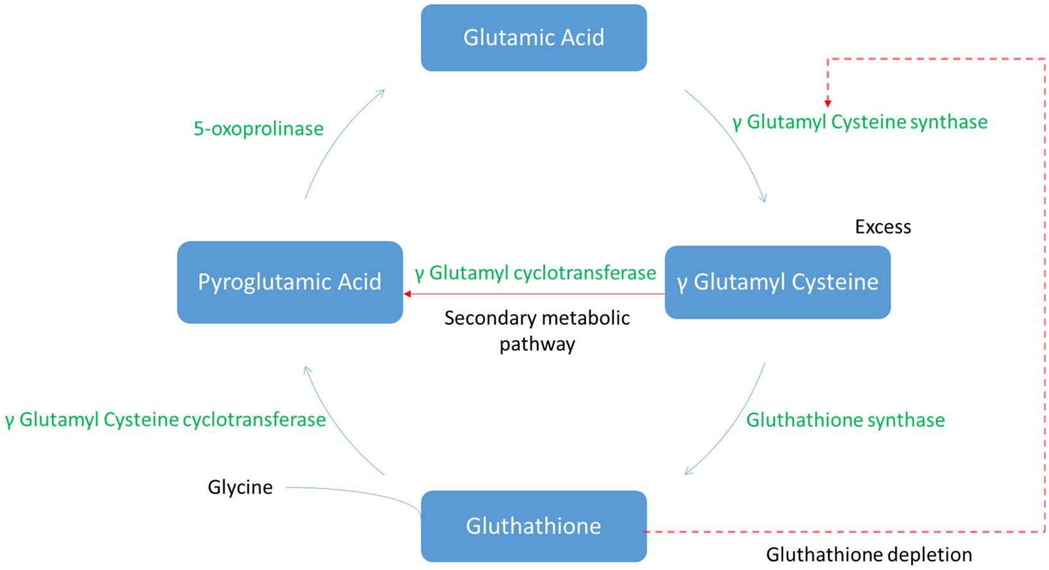


Figure 5-30. A depiction of the gamma-glutamyl cycle: an essential pathway for cells in the human body to regulate intracellular glutathione (GSH) levels (Bachhawat and Yadav, 2018).

5.6 Results Summary

The aim of the work described in this chapter was to apply the strategies developed in the two previous chapters to explore xenobiotic metabolism (xenometabolome).

The first application involved the retrospective examination of urine and blood MS datasets on the xenobiotic Triclosan. *In vitro* models with TCS liver incubations were undertaken and profiled to discover the main TCS metabolites, which was then explored in both urine and plasma AW phenotyping datasets. Once the best marker for TCS exposure was identified in these datasets, subsequent analyses such as measurement of its population distribution, semi-quantitation and finally the application of logistic regression was used to identify feature (metabolite) associations. The second application was on the xenobiotic PEG 400, which is a known excipient used in pharmaceutical formulations. Similarly, the best marker for exposure and its distribution was evaluated in the ALZ phenotyping datasets. In this application however, both data-driven and analytical strategies were implemented**.** Correlation analyses and PLS models were applied to datasets acquired using the developed small molecule (DSPE lipid removal) method for blood products.

Triclosan

Molecular phenotyping techniques has successfully detected TCS metabolites in the bio-fluids of human volunteers enabling a large-scale assessment of TCS exposure prevalence and metabolism within the UK population (AW). An *in vitro* incubation of TCS with human hepatocytes revealed several metabolites, predominately sulfate and glucuronide conjugates. A reported prevalence of approximately 20-30% was observed *in vivo*, confirmed by the presence of the conjugates TCS-Gluc in urine, and TCS-SO4 in blood respectively. Multimodal distributions for TCS-Gluc in urine and TCS-SO4 in plasma essentially divided the data into exposure groups of varying levels, i.e. zero, Low-mid and high. Logistic regression analysis was performed between the zero and high exposure groups to identify metabolites with significant differences in concentration using a measured TCS exposure marker (TCS-Gluc in urine and TCS-SO4 in plasma) as a dependent variable and each detected metabolite as an explanatory variable. Significant metabolites (univariate, $p_{adj} \leq 0.05$ or EN) in urine, highlighted strong associations to direct drug metabolites, i.e. the oxidised glucuronide and sulfate forms as well as the detection of phase1 metabolite 2,4 dichlorophenol glucuronide. To my knowledge this is the first reported case of oxidised phase 2 conjugates of TCS, and a glucuronidated phase 1 metabolite of TCS, to be reported *in vivo.* An endogenous metabolite, GHB-SO4, was also identified with a significant association to TCS, suggesting a possible link with the microbiome. Other significant metabolites which were identified or annotated were exogenous in nature (surfactants like the identification of Menthol and putative annotation of lauric acid derivatives), indicating that they possibly derived from the same source, such as diet or personal care products. In addition, the majority of associated metabolites were sulfated which could potentially result in a reduced sulfonation capacity. In plasma samples, the detection of PFOS as a co-exposure, and the perturbation in lipid metabolism, specifically SM, were observed with known reported toxic health implications. In closing, the studies and data presented here are part of an effort to demonstrate the ability of molecular phenotyping for xenobiotic metabolism applications.

Polyethylene glycol

Due to the common usage of PEG in food products, cosmetics and as an excipient in pharmaceutical formulations, PEG metabolism has been well documented in mammalian systems. Urine excretory metabolites include the diacid (PEG-2xCOOH) and hydroxy acid (PEG-COOH) metabolites. Interrogation of MS data from the ALZ cohort demonstrated a significant correlation between the PEG unmetabolised form, to these metabolites in both urine and serum biofluids. The correlation coefficients for PEG(n8)-COOH and PEG(n8)-2xCOOH, were 0.93 and 0.90 in urine and 0.76 and 0.57

in serum respectively. Additionally, matched urine and serum collected at the same ALZ participant visit, also demonstrated a significant correlation (r = 0.6 and r = 0.3). Together, these findings suggest that PEG encountered in a subset of samples has undergone metabolism *in vivo* and can therefore originate as part of an environmental exposure and should not be immediately disregarded when phenotyping the population.

The detection of highly intense PEG signals in the absence of PEG metabolites, highlighted the samples that have been contaminated through sample handling or instrumentation. By differentiating the signals due to exposure, exploration on associations of PEG to the endogenous profile were possible. Through the application of univariate correlation models and multivariate PLS models (regression and discriminant analysis), the endogenous metabolite PGA was identified, and two exogenous metabolites, DBS and piperidone were putatively annotated. DBS and piperidone are excipients used in pharmaceutical formulations, and although have only putatively been annotated, if true, further provides evidence of PEG as an external exposure due to medications and not contamination.

A linear relationship was also observed between excipient profiles and metabolites involved in the γ-glutamyl cycle. A biological implication of this is that Increases in medical use, compounded by PEG used as an excipient in the same medical formulations, can potentially result in an accumulation of PGA, leading to pyroglutamic acidosis, which is one of many causes of HAGMA. PEG intoxication can also result in HAGMA. It therefore seems that careful consideration is perhaps needed not only on the number of medications being administered to an individual patient, but also the excipients used in the formulation.

An epidemiological implication of this work is how the presence of PEG in biofluids can reduce the quality of a dataset and can negatively impact the statistical power and ultimately the economic value of a study if contaminated samples are not removed from the dataset. For this reason, PEG contamination in population studies may be preventable. Steps can be taken in metabolic phenotyping sample collection protocols to reduce polymer presence in biological samples and ensuring mass spectrometry data acquisition is not compromised by ion suppression effects. Steps such as the substitution of medicines or any supplements which are in gel form (known to contain PEG), to perhaps a powdered alternative, in the lead up to sample collection.

The findings of this study strongly suggest that excipients such as PEG should be regarded as a separate exposure as they too can influence metabolic systems. Analytical considerations aside, samples that are found to contain PEG (and metabolite) should not necessarily be excluded on the basis of post-sampling contamination and be considered for inclusion.

5.7 Significance of Findings

The examination of the two exemplar xenobiotics highlighted the increased importance of environmental chemicals, and their potential influence on health status.

Data driven strategies using statistical based analyses such as correlation, logistic regression and PLS models were implemented (from strategies explored in Chapter 3) resulting in additional feature associations to be identified from the target exposure marker. This was exemplified in the TCS application, which allowed the prevalence and metabolism of exposure to be explored at the population level (*in vivo*). In addition to the data-driven strategies, the analytical lipid removal method (Chapter 4) provided an additional strategy to broaden the coverage of the xenometabolome in blood products which was exemplified in the PEG application. Both xenobiotics have therefore been profiled in the population and as a result, associated metabolites have been identified and added to the xenobiotic database.

Thus, the applications studied in this chapter utilised both data-driven and analytical strategies highlighting additional xenobiotic derived annotations related to direct metabolism and co-exposures, whilst also highlighting perturbations in endogenous metabolism leading to the discovery of novel metabolites and potential affected metabolic pathways. Ultimately both applications demonstrated the effectiveness of metabolic phenotyping to study drug metabolism, and to generate new hypotheses.

5.8 Conclusion

Establishing a prospective and dedicated investigation into individual ubiquitous chemicals such as PEG and TCS, as an environmental risk factor would require substantial investment, thereby presenting a barrier to such studies. Comprehensive work on the metabolism kinetics of xenobiotics in human subjects uses targeted methodology with mass spectrometry offering higher sensitivity and lower limits of detection (Calafat et al., 2008, Allmyr et al., 2006, Gonzalez-Marino et al., 2009, Vijaya Bhaskar et al., 2013, Gong et al., 2014, Su et al., 2019).

While these provide an extremely rigorous measurement, they are limited to parent xenobiotic and its metabolized forms. The application of molecular phenotyping to epidemiological studies is increasingly being incorporated into disease research and the data produced from these studies may already contain the metabolic responses from external xenobiotic exposures, potentially offering a far more cost-effective and productive route. Although not always quantitative in an absolute sense, it can provide insight into direct metabolism, associated metabolites (endogenous and exogenous co-exposures) and affected metabolic pathways.

Also, from both applications, metabolic profiling of multiple biofluids collected simultaneously enabled a greater understanding of the metabolic associations related to exposure from these two exemplar xenobiotics. Urine and plasma/serum were collected at the same participant study visit and analysed using the same analytical method (RPC-UPLC-MS). Samples are however processed independently, with biofluid specific sample preparation protocols, randomised run orders and sample acquisition using different mass spectrometers. Nevertheless, the analysis of both urine and serum biofluids enabled a multicompartment snapshot of metabolic status, with statistically significant correlation indicative of a relationship between circulating plasma metabolite(s) and those present in the urine of the same individual.

To conclude, the results from this chapter demonstrated that the strategies developed from previous Chapters 3 and 4, and the phenotyping workflow was successful in capturing the metabolic imprint of TCS and PEG exposures, thereby supporting the hypothesis for this chapter. As a result of these applications, the novelty of this chapter came about from new knowledge discovered relating to biology. Profiling the metabolism of exemplar xenobiotics over an entire population of people, and evaluating the proportion exposed, led to the discovery of novel xenobiotic metabolites, with possible implications affecting major metabolic pathways such as the γ-glutamyl cycle and sulfonation pathways.

Chapter 6

General discussion and future work

Outside of targeted analysis of specific compounds in toxicological, pharmaceutical, and environmental studies, the xenometabolome has been largely uninvestigated; the majority of metabolic phenotyping studies to date focusing on biomarkers which are endogenous.

The high analytical sensitivity from phenotyping platforms (such as mass spectrometry), will not only improve detection of these endogenous metabolites, but also xenobiotics. Estimates for exposures of individuals to xenobiotics are commonly poorly characterised in population studies and relies on participant recall or be estimated from exposure models.

As a result, exposure misclassification, inadequate adjustment for confounders, and failure to remove outlier samples may lead to reduced study power and increased bias in metabolic phenotyping analyses. The work presented in this thesis has sought to explore and deliver strategies to enable the annotation of xenobiotic derived signatures that augments existing metabolome profiles in large-scale molecular epidemiological analysis. The strategies have additionally provided novel insight into xenobiotic metabolism at the population level.

Throughout this work, the focus has been to substantially broaden the coverage of the xenometabolome through the development of new laboratory and data-driven statistical methods to achieve this goal. Combined, this provided novel methods for xenobiotic annotations and identifications in common metabolic phenotyping assays of urine and blood products and a workflow for prospective xenobiotic annotation in current and future studies. Two strategies were therefore developed to broaden the coverage of the xenometablome.

The first strategy (Chapter 3) was a multi-faceted data-driven approach which initially involved the generation of a xenobiotic database, in which reference standards were acquired by RPC-UPLC-MS. Next, statistical based methods were used to further increase xenobiotic related annotations present in urine and blood phenotyping datasets, from a target xenobiotic.

Finally, a method was developed to identify outlying signals with the potential to be affiliated with xenobiotic exposure. The second strategy (Chapter 4) was an analytical based approach to characterise xenometabolome components. This involved the development of a sample preparation protocol for the analysis of moderately hydrophobic and amphipathic metabolites, which includes much of the xenometabolome, in blood products, thereby enabling RPC-UPLC-MS profiling which would have otherwise been compromised by lipophilic species. The two strategies were then applied to existing human cohort studies, to enhance and characterise the xenometabolome, relating to two exemplar xenobiotics (Chapter 5).

6.1 Knowledge-based and data-driven strategies

6.1.1 *Reference standard database*

Metabolite identification of unknown metabolic features obtained from UPLC-MS data, remains a bottleneck in untargeted metabolomics. Metabolites often reported are by spectral matching with online mass spectrometry databases. Therefore, the simplest approach to increase xenobiotic annotations was the analysis of authentic standards of xenobiotics and xenobiotic metabolites using the existing NPC RPC platforms (for method specific retention times) to generate a library of common xenobiotic signatures. Xenobiotics selected for the database were chosen based on comprehensive knowledge through literature search and collation of priority therapeutic drugs, excipients, and additives, that are commonly prevalent (through prescription or otherwise) in the United Kingdom. The pharmacokinetic/pharmacodynamic (PK/PD) and metabolism of many pharmaceutical agents are very well characterized during regulatory steps and was the basis for any statistical-based survey to complement the laboratory-based analyses. A total of 25 chemical reference standards were initially acquired to populate the xenobiotic database. There are currently 41 reference standards and 57 pharmaceutical medications that have undergone the acquisition workflow. This number is increasing as more studies are being conducted at the NPC. Using targeting software such as peakpantheR, has allowed for many xenobiotics to be annotated in existing datasets. The workflow necessary to acquire, process and examine xenobiotics have been set in place for future procurement of xenobiotic refence standards. There is also scope to run these standards *via* other NPC profiling platforms (HILIC, LIPID and NMR), as xenobiotics can exhibit a range of different molecular properties, and so may be better measured by these methodologies.

6.1.2 *Statistical methods*

The acquisition of reference standards may not always be feasible due to time and cost restraints. Reference standards available commercially are usually related to the unconjugated parent from, and metabolites often require synthesis which can also be quite costly. Sample preparation protocols, such as enzymatic hydrolysis and sulfation protocols discussed in Chapter 5 can also be used to further increase annotations analytically, but still require the necessary reagents and further analytical experiments (MS/MS) for annotations.

This does raise a limitation for exposure exploration of this nature. Although we were able to extract out features/metabolites related to exposure, identification in some instances were difficult due to lack reference materials. Xenobiotics are often expensive or not commercially available, requiring expensive synthesis. As mentioned in the discussion in Chapter 3 (**section 3.6**), metabolic identification represents a bottleneck in phenotyping studies. As a result, interpreting the biology from putative annotations should be approached with caution. Suspect screening, protocols to purify and concentrate signals in biofluids such as urine to enable additional analyses and combining or linking analytical platforms such as MS and NMR, are some examples of what other groups in the scientific community are implementing to annotate unknown metabolites in biofluids.

An alternative approach for xenobiotic annotations, are statistical based methods for extracting xenobiotic related MS signals from datasets, which were exemplified in Chapter 3. Human exposure to xenobiotics may result in elimination of the xenobiotic unchanged, however the vast majority undergo endogenous metabolism such as conjugation and enzymatic functionalisation. Existing metabolomics datasets, which are an unbiased measurement of all downstream low molecular weight compounds, may already contain data related to xenobiotic metabolism data which can be exploited. Annotations of MS features related to xenobiotic exposure, can be further validated by known metabolites of the xenobiotic (literature or from databases), e, g. conjugation moieties, if detected by the analytical platform, as these will in most cases have some relationship statistically to the unconjugated form.

In this thesis, the statistical methods explored were, correlation (intersample and i-STOCSY), logistic regression and PLS models. Univariate methods used in metabolomics are often used and simplest to implement. The application of multiple testing correction for statistical significance in univariates tests, therefore, makes interpretation easy. Scripts for both the intersample correlation and logistic regression univariate methods were written in R and successfully implemented on existing datasets allowing for further annotations relating to xenobiotics. Embedded in the code was the use of multiple testing correction. FDR was used as the default parameters for the intersample correlation analysis, although can be changed easily in the code if needed. The graphical interface of i-STOCSY makes carrying out correlation-based analyses efficient, as driver features can be selected at any point. However, unlike i-STOCSY, added features to the intersample correlation developed in this thesis includes multiple testing correction. i-STOCSY can be used to give an indication of correlated features and used with the intersample correlation analysis for a more vigorous and accurate measurement.

As with any univariate method, relationships are between pairs of features, however instances may arise where feature associations observed in a univariate sense, may not exhibit the same association when other features are taken into account. Alternately, the inverse could be observed, where features together may indicate an association, but is exhibits no association individually. Multivariate models (PCA, PLS, OPLS) may also be of value for xenobiotic exposure related investigations. Furthermore, correlations between features within a dataset may not always be biological, but can be the result of instrumentation artefacts, resulting in many intercorrelated features. To account for this, dimension reduction using PCA or PLS models, have been widely used in the metabolomics, however, care should be taken as these models can be prone to overfitting. In the PEG application, PLS-R (continuous variables) was implemented and for feature selection, variable importance in projection (VIP) was used as a means to highlight the features (loadings) which have a strong contribution to exposure. Apart from PLS based methods, alternative multivariate approaches which can provide implicit feature selection together with model development relates to regularization-based methods where a penalty is imposed on regression coefficients, (Ridge, LASSO and Elastic net) as demonstrated in TCS application. Both LASSO and EN has implicit variable selection so as part of the regularisation, important variables are those with non-zero coefficients, which was directly derived from the model. For models, that utilised ridge regression, the most important regression coefficient, were found by bootstrapping and resampling the data (500 iterations). This produced confidence intervals in which statistical significance could be derived. The multivariate based approaches can be and was, additionally utilized to further complement the univariate analysis, highlighting additional xenobiotic feature associations.

The identification of exposure groups was vital when exploring xenobiotic feature associations using both univariate and multivariate methods. The classification of samples into exposure groups was undertaken in one of two ways; either the xenobiotic of interest is known through compliance patient meta data, or a spectral feature which was representative of exposure to be specified (through identification from a reference standard), and its population distribution in the data to be evaluated. Assessment of the distribution was vital for each statistical method. Non-zero metabolite features in metabolomic datasets often presents as right-skewed distributions, and any modality observed in the data may potentially be a result a measurement artefact (such as batch effects). Some variation of transformation (log transformation used in this thesis) can usually be carried out to correct for any skewed distribution thereby reducing the impact of outliers causing the skewness. Multimodal distributions were observed for xenobiotic related features in phenotyping data (as seen in the PEG and TCS). Often is the case with metabolomic investigations involving a healthy and a disease group, exposure to particular xenobiotics will mostly only present in the disease as a result of medicinal intake. If a multimodal distribution exists, code was written in R, where multi-component gaussian mixture models (GMMs) were specified, placing clusters across the distributions. Once fitted, conversion of the distributions to probability distribution functions (PDF's), were calculated. This resulted in a more accurate assessment of samples belonging to a particular distribution. From this, any univariate or multivariate analyses could be carried out, which were exemplified in the TCS and PEG applications, and in Chapter 3 with Amlodipine.

The statical methods explored to identify xenobiotic metabolites, resulted in models which were generally easy to interpret. The evaluation of beta coefficients observed in regression models, statistical significance through the calculation of p-values or the use of resampling and regularization, all are approaches which were utilised to highlight metabolic features relating to xenobiotic exposure in profiling datasets. In addition to interpretation, different methods are more appropriate depending on whether the measurement for exposure is continuous or categorical. Where measurements are continuous, correlation or covariation in PLS regression models are more suitable methods for highlighting metabolic features associated with exposure. However, a lack of correlation does not necessarily mean there is no associations, as metabolites could exhibit non-linear relationships. Therefore, evaluation of the feature distribution, and classification into exposure groups then allowed for other methodologies to be implemented whereby variables relating to exposure could be categorical. Logistic regression and PLS-DA are such examples of methodologies more suited to identify the most predictive or discriminative metabolic features in the classification of sample groups. Also, regression can be considered a more detailed analysis, as it can accommodate confounders, such as age or gender. As a result, regression models which incorporate confounders are increasingly being used in epidemiology in place of stratification methods (McNamee, 2005).

6.1.3 Outlier samples

Finally, a more untargeted statistical approach for identifying xenobiotic signatures was developed, based on distinguishing outliers caused by erroneous signals (e.g., poor peak integrations or system contaminants) from those caused by the legitimate presence of a feature in an unusually high concentration highlighted a pattern in the data which could be the result of a xenobiotic exposure. The feature in an outlying sample is corroborated by an observed effect on the average (pooled) sample, resulting in its elevation from the mean distribution of the study sample population. This approach was successful in identifying signature relating to xenobiotics (Flucloxacillin and PEG3350). As with correlation and logistic regression, a script was written in the R language to automate xenobiotic annotations from existing and future metabolomic datasets. However, the current pre-processing tool for MS profiling datasets at the NPC, uses XCMS which incorporates algorithms such as minimum fraction filters (Minfrac) to look for valid features present in a minimum number of samples within a sample group. Outlying features may therefore be removed at this stage of the workflow, therefore this method for outlier detection may be better suited using a lower Minfrac setting. A lower setting does however come with risks, such as the inclusion of features relating to artefactual noise. The additional filtering protocol adopted by the NPC, i.e. dilution series and RSD under a threshold, will account for some of the noise features but only to a certain extent, as such, the setting used for LC-MS based analyses by the NPC is currently fixed at 0.4. The detection of the chlorinated compounds from this work, highlighted an avenue of research where metabolites can be identified based on their specific isotopic distribution. Synthetic compounds such as Flucloxacillin and TCS, and other xenobiotics (pharmaceutical drugs and pesticides), are more likely to contain halogens, thereby exhibiting a unique isotopic pattern in the mass spectrum (Hernandes et al., 2010, Jeschke, 2010). The MassLynx™ software suite (Waters Inc., USA) includes OpenLynx™ , an application that can perform isotopic cluster analysis; inputting the isotopic distribution (*m/z* peak ratios based on peak area) and target retention time range allows identification of matching features and feature extraction across sample sets. Note: this analysis is computationally intensive as it is based on a scan-by-scan. Pre-processing software used for MS data, such as Progenesis QI (Nonlinear Dynamics, Newcastle, UK) measures isotopic abundance profiles for each spectral feature during the peak picking process. This was one of the main reasons why Progenesis QI was the pre-processing software of choice for the first few initial projects conducted by the NPC. However, incorrect assignment of isotopic profiles was consistently observed. Furthermore, feature grouping parameters such as the Minfrac setting, is not available within the Progenesis peak picking method which can result in larger less manageable datasets comprising of more noise features. Thus, a move from Progenesis QI to XCMS was eventually adopted by the NPC for the pre-processing of MS datasets.

More recently, mass defect and isotope filtering algorithms have been developed that identifies oxidative metabolites with mass defects similar to or significantly different from those of the parent drugs, and isotopic distributions indicative of a chlorine atom presence in the molecule (Zhu et al., 2006, Rathahao-Paris et al., 2014). The algorithms may not cover all biotransformation products of a xenobiotic *in vivo* but can be useful to identify xenobiotics and metabolites from LC-MS datasets. This suggests that halogen identification may lie in a more informatic based approach and warrants further investigation.

6.2 Analytical driven strategy

To complement the data-driven strategies for xenometabolome coverage, a more analytical based strategy was developed. The majority of xenobiotics exhibit both a certain degree of hydrophobicity and amphipathic properties (semi-polar). As such, the analysis of molecules which fit this specificity, are better suited with reversed phased methodologies due to binding of the non-polar properties of the molecule, to the hydrophobic stationary phase in RP columns and the excellent analytical performance (uniform peak shape, stable retention times and quick equilibration times) offered by RPC based systems. A urine RPC profiling assay was previously developed and is part of the NPC profiling portfolio (Lewis et al., 2016), so a complementary blood protocol to measure the xenometabolome using RPC was developed, with the added benefit of being utilised to measure a broader range of moderately hydrophobic metabolites, which includes many endogenous molecules. Significant challenges in the analysis of blood extracts using RPC platforms are the presence of highly lipophilic compounds (e.g., lysophospholipids, phospholipids, triglycerides, etc.) which can result in significant methodological issues for high-throughput analysis. The DSPE sample preparation protocol developed in this thesis efficiently removes lipids from blood products, but with minimal effect on other low-molecular weight metabolites. Briefly, the development was split into three stages, optimisation, validation and application. Optimizing the components of a DSPE protocol was firstly undertaken, i.e. the sorbent, solvent, and slurry (sorbent-solvent concentration). Selection of sorbent and solvent were based on metrics which produced the greatest number of features that were highly precise. The optimum sorbent-solvent (slurry) concentration was undertaken using a design-of-experiment (DOE) protocol, allowing both variables (sorbent and solvent) to be optimised simultaneously by maximising response recoveries associated with the small molecule profile and lipid profile, i.e. high recoveries for the former, and low recoveries for the latter. The validation demonstrated high reproducibility when implemented in a 96-well format, and therefore is applicable for high through-put. When the small molecule profile was compared to samples without DSPE treatment, higher recoveries were observed thereby demonstrating that the protocol was able to efficiently remove the lipids, without any significant change to the remaining profile. The validation stage also compared other lipid removal protocols (SPE) and LLE methods. The DSPE delivered extracts free of large concentrations of peptides and protein, were highly reproducible, less tedious in terms of sample preparation, and greater metabolite coverage. The final stage of the development was the application of the DSPE method to two exemplar profiling studies. In both cases, high precision, and annotation of a greater set of metabolites (both endogenous and xenobiotic) were observed, thereby offering greater xenometabolome and metabolome coverage for blood products. The analysis of these projects also demonstrated that the DSPE method was also applicable with serum biofluids. A limitation of the DSPE protocol was that only plasma was used in the development. However, the findings from the application of DSPE to serum samples associated with AZ Study 12 population study, suggest that this may not necessarily be an issue. Although a comparison between serum and plasma was not undertaken in this development, literature suggests that either matrix will generate similar metabolite profiles if sample preparation protocols are identical (Yu et al., 2011, Liu et al., 2018). Another weakness of the method that was uncovered during the development involved preparation of the dilution series. Sample dilution prior to extraction is common procedure for all profiling methods at the NPC. As such, any contaminant introduced as part of the sample extraction procedure, in theory will not correlate to dilution, and therefore be removed as part of the filtering process. However, in this DSPE protocol, diluting the sample prior to a solvent extraction highlighted a disruption in protein precipitation, resulting in protein presence in the final extract. Thus, contaminants introduced during sample preparation, will also serially dilute. This was accounted for in the development, by utilising blank samples, that underwent the extraction procedure and therefore used as a third criteria for filtering. It would mean, for future project samples, an extra extraction using water blanks is a necessary requirement for filtering purposes. There is however promise for this protocol to be implemented on blood products analysed for HILIC based assays. Tsakelidou *et al.* highlighted that endogenous phospholipids contained in blood products in significant levels are considered to be a real problem in HILIC based assays (Tsakelidou et al., 2017). Southam *et al* demonstrated that a 50 : 50, methanol : acetonitrile solvent composition delivered the greatest number of putatively-identified polar metabolites with high reproducibility, in blood analysed by HILIC (Southam et al., 2020). Both points have been addressed with the developed protocol. Furthermore, the DSPE extracts when acquired *via* the NPC HILIC method, also produced substantial coverage in comparison to other samples preparation protocols (i.e. LLE methods and monophasic extraction methods) of polar small molecules especially in relation to acylcarnitine's. Overall, this protocol, provided an analytical option to measure xenometabolome and metabolome components in blood products, which was exemplified in a number of different profiling studies, adding another a dataset which not only measured endogenous metabolites, but also a range of xenobiotics of this specificity.

6.3 Data-driven and analytical application

Complementary analytical and statistical strategies were developed in this work described in this thesis; both identified signatures pertaining to xenobiotics and/or xenobiotic-induced endogenous changes in the metabolome when applied to biofluid samples within epidemiological studies. These strategies were exemplified in Chapter 5 by the exploration of the xenometabolome and metabolome in relation to the xenobiotics, TCS and PEG. The workflow identified several aspects in relation to exposure evaluation. Firstly, metabolism related to these two xenobiotics could be assessed at the population level. For instance, known metabolites of TCS were reported, in the form of the glucuronide and the sulfate. The oxidised sulfate and oxidised glucuronide metabolites were additionally detected, and to my knowledge are novel and has never been reported in human biofluids *in vivo.* The workflow also highlighted affected endogenous pathways. In both examples, pathways relating to sulfation, microbiome, lipid and glutathione metabolism were highlighted. There were also toxicity concerns that could potentially arise due to co-exposures (TCS and PFOS). Furthermore, changes to study design can also be inferred from this evaluation. In the PEG example, PEG should not always be regarded as a laboratory contamination and in the presence of PEG metabolites, should be considered as part of the human metabolome and the wider exposome. Environmental chemicals will be of influence on our health status and are becoming part of the human metabolite phenotype. The substitution of medicinal products, from gel caps (known to contain PEG) to powder alternatives, in the lead up to sample collection may be a possible solution in reducing PEG exposure and be beneficial to attaining analytical data that is better in quality and would be more appropriate for population phenotyping and retrospective data mining. However, as some of the annotations were putative, care must be taken in some of the interpretations made. Identification of less prevalent xenobiotics and their metabolites will always be a bottleneck for these investigations and requires additional analysis (such as MS/MS) for additional structural information. If sample is limited, in future studies, it may be helpful to run analyses with MSE mode for fragmentation data, with the added benefit of enabling real time putative annotations using spectral databases. It was evident from the PEG and TCS investigations, that validity and confidence in metabolic feature associations related to exposure, lies in robust exposure models (accurate exposure classification, cross validation and/or training and independent test sets), significant features replicating across multiple datasets, (as seen with the significant features relating to TCS exposure validating across both AW and ALZ datasets), and confirmation of putatively annotated metabolites with reference standards. The observations made from the TCS and PEG examples demonstrate the strength of phenotyping approaches in studying xenobiotic metabolism, highlighting metabolites which are better markers for exposure, and identifying affected endogenous pathways.

6.4 Final note

The statistical and analytical strategies described can therefore be useful in two possible real-world applications. The first application is useful in metabolic phenotyping investigations. In untargeted metabolic phenotyping studies, MS features associated with xenobiotic exposure can potentially bias the interpretation of such studies. However, the use of these strategies allows such exposures to identified and partitioned, thereby permitting improved detection of both outliers and non-compliant participants, and better confounder data. The second application is more suited to toxicology and even pharmaceutical industries, where these strategies can be implemented on samples sets produced from epidemiological studies, in an effort to study the metabolism of xenobiotics on a population level. The benefit of studying xenobiotic metabolism using untargeted phenotyping approaches *in vivo* rather than the conventional *in vitro* or stable isotope route (the latter of which is commonly used in drug metabolism studies), is due to its potential to capture both xenobiotic metabolism, and changes in endogenous metabolism, as demonstrated by the TCS and PEG examples. This insight has led to an increased interest in metabolic phenotyping approaches to study how drugs are being developed and dosed.

Looking forward, strategies to identify and partition samples in large population studies with exposure to xenobiotics will not only be important to help improve the quality of phenotyping datasets but can be particularly important with respect to personalised healthcare. The characterisation of the xenometabolome in population studies may reveal unique biochemical signatures that are a result of xenobiotic exposure and aid in explaining the variation observed with its metabolism in the human body. As a result, adverse drug reactions could be potentially avoided, and drug efficacy could be enhanced.

Lastly, the workflow developed for identifying xenobiotics, *via* the statistical or analytical strategies, has many applications in the area of metabolic phenotyping, with the latter (blood lipid removal protocol) now employed in a large range of MS based studies at the NPC. The statistical methods for annotating xenobiotics, can be applied to existing human cohort study as needed to characterize priority xenometabolome components and partition prior to subsequent data analysis. Furthermore, the generation of the reference standard database has also proved very useful in further annotating xenobiotics for current and future metabolic phenotyping studies, with the addition of new xenobiotic compounds to the database being driven by metabolite identification efforts on studies conducted at the NPC.

References

Dispersive Solid-Phase Extraction. *Analytical Separation Science.*

ABBOTT, S. R. 1980. PRACTICAL ASPECTS OF NORMAL-PHASE CHROMATOGRAPHY. *Journal of Chromatographic Science,* 18**,** 540-550.

ADAMSKI, J. 2012. Genome-wide association studies with metabolomics. *Genome medicine,* 4**,** 34-34.

AIELLO, A. E., LARSON, E. L. & LEVY, S. B. 2007. Consumer Antibacterial Soaps: Effective or Just Risky? *Clinical Infectious Diseases,* 45**,** S137-S147.

ALBERICE, J. V., AMARAL, A. F., ARMITAGE, E. G., LORENTE, J. A., ALGABA, F., CARRILHO, E., MARQUEZ, M., GARCIA, A., MALATS, N. & BARBAS, C. 2013. Searching for urine biomarkers of bladder cancer recurrence using a liquid chromatography-mass spectrometry and capillary electrophoresis-mass spectrometry metabolomics approach. *J Chromatogr A,* 1318**,** 163-70.

ALLEN, D. R. & MCWHINNEY, B. C. 2019. Quadrupole Time-of-Flight Mass Spectrometry: A Paradigm Shift in Toxicology Screening Applications. *The Clinical biochemist. Reviews,* 40**,** 135-146.

ALLMYR, M., ADOLFSSON-ERICI, M., MCLACHLAN, M. & SANDBORGH-ENGLUND, G. 2006. Triclosan in plasma and milk from Swedish nursing mothers and their exposure via personal care products. *Science of the Total Environment,* 372**,** 87-93.

ALLMYR, M., ADOLFSSON-ERICI, M., MCLACHLAN, M. S. & SANDBORGH-ENGLUND, G. 2006. Triclosan in plasma and milk from Swedish nursing mothers and their exposure via personal care products. *Sci Total Environ,* 372**,** 87-93.

ALONSO, A., MARSAL, S. & JULIÀ, A. 2015. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology,* 3**,** 23-23.

ALYASS, A., TURCOTTE, M. & MEYRE, D. 2015. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics,* 8**,** 33.

ANASTASSIADES, M., LEHOTAY, S. J., STAJNBAHER, D. & SCHENCK, F. J. 2003. Fast and easy multiresidue method employing acetonitrile extraction/partitioning and "dispersive solid-phase extraction" for the determination of pesticide residues in produce. *J AOAC Int,* 86**,** 412-31.

ANDREWS, K. W., SCHWEITZER, A., ZHAO, C., HOLDEN, J. M., ROSELAND, J. M., BRANDT, M., DWYER, J. T., PICCIANO, M. F., SALDANHA, L. G., FISHER, K. D., YETLEY, E., BETZ, J. M. & DOUGLASS, L.

2007. The caffeine contents of dietary supplements commonly purchased in the US: analysis of 53 products with caffeine-containing ingredients. *Anal Bioanal Chem,* 389**,** 231-9.

ANTIGNAC, J.-P., DE WASCH, K., MONTEAU, F., DE BRABANDER, H., ANDRE, F. & LE BIZEC, B. 2005. The ion suppression phenomenon in liquid chromatography–mass spectrometry and its consequences in the field of residue analysis. *Analytica Chimica Acta,* 529**,** 129-136.

ARMBRUSTER, D. A. & PRY, T. 2008. Limit of blank, limit of detection and limit of quantitation. *The Clinical biochemist. Reviews,* 29 Suppl 1**,** S49-S52.

ARMIROTTI, A., BASIT, A., REALINI, N., CALTAGIRONE, C., BOSSU, P., SPALLETTA, G. & PIOMELLI, D. 2014. Sample preparation and orthogonal chromatography for broad polarity range plasma metabolomics: application to human subjects with neurodegenerative dementia. *Anal Biochem,* 455**,** 48-54.

ARMSTRONG, R. A., DAVIES, L. N., DUNNE, M. C. & GILMARTIN, B. 2011. Statistical guidelines for clinical studies of human vision. *Ophthalmic Physiol Opt,* 31**,** 123-36.

ASHRAP, P., ZHENG, G., WAN, Y., LI, T., HU, W., LI, W., ZHANG, H., ZHANG, Z. & HU, J. 2017. Discovery of a widespread metabolic pathway within and among phenolic xenobiotics. *Proceedings of the National Academy of Sciences,* 114**,** 6062-6067.

ATHERSUCH, T. J. 2012. The role of metabolomics in characterizing the human exposome. *Bioanalysis,* 4**,** 2207-12.

ATHERSUCH, T. J. & KEUN, H. C. 2015. Metabolic profiling in human exposome studies. *Mutagenesis,* 30**,** 755-762.

AUDI, S., BURRAGE, D. R., LONSDALE, D. O., PONTEFRACT, S., COLEMAN, J. J., HITCHINGS, A. W. & BAKER, E. H. 2018. The ‘top 100’ drugs and classes in England: an updated ‘starter formulary’ for trainee prescribers. *British Journal of Clinical Pharmacology,* 84**,** 2562-2571.

AYGUN, S. F. & OZCIMDER, M. 1996. A comparison of normal (-CN) and reversed (C-18) phase chromatographic behaviour of polycyclic aromatic hydrocarbons. *Turkish Journal of Chemistry,* 20**,** 269-275.

BACHHAWAT, A. K. & YADAV, S. 2018. The glutathione cycle: Glutathione metabolism beyond the γ-glutamyl cycle. *IUBMB Life,* 70**,** 585-592.

BALDING, D. J. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics,* 7**,** 781-791.

BALE, S. S., MOORE, L., YARMUSH, M. & JINDAL, R. 2016. Emerging In Vitro Liver Technologies for Drug Metabolism and Inter-Organ Interactions. *Tissue engineering. Part B, Reviews,* 22**,** 383-394.

BANOEI, M. M., IUPE, I., BAZAZ, R. D., CAMPOS, M., VOGEL, H. J., WINSTON, B. W. & MIRSAEIDI, M. 2019. Metabolomic and metallomic profile differences between Veterans and Civilians with Pulmonary Sarcoidosis. *Scientific Reports,* 9**,** 19584.

BARANOWSKA, I., WILCZEK, A. & BARANOWSKI, J. 2010. Rapid UHPLC method for simultaneous determination of vancomycin, terbinafine, spironolactone, furosemide and their metabolites: application to human plasma and urine. *Anal Sci,* 26**,** 755-9.

BARBIER SAINT HILAIRE, P., ROUSSEAU, K., SEYER, A., DECHAUMET, S., DAMONT, A., JUNOT, C. & FENAILLE, F. 2020. Comparative Evaluation of Data Dependent and Data Independent Acquisition Workflows Implemented on an Orbitrap Fusion for Untargeted Metabolomics. *Metabolites,* 10.

BARRENTINE, L. B. 2014. *Introduction to Design of Experiments : a Simplified Approach,* Milwaukee, ASQ Quality Press.

BARTON, R. H., WATERMAN, D., BONNER, F. W., HOLMES, E., CLARKE, R., NICHOLSON, J. K. & LINDON, J. C. 2010. The influence of EDTA and citrate anticoagulant addition to human plasma on information recovery from NMR-based metabolic profiling studies. *Mol Biosyst,* 6**,** 215-24.

BASKIN, L. B., ANDERSON, R. W., CHARLSON, J. R., HURT, R. D. & LAWSON, G. M. 1998. A solid phase extraction method for determination of nicotine in serum and urine by isotope dilution gas chromatography/mass spectrometry with selected ion monitoring. *Ann Clin Biochem,* 35 (Pt 4)**,** 522-7.

BASU, S. S. & BLAIR, I. A. 2011. SILEC: a protocol for generating and using isotopically labeled coenzyme A mass spectrometry standards. *Nat Protoc,* 7**,** 1-12.

BAUMANN, A., TUERCK, D., PRABHU, S., DICKMANN, L. & SIMS, J. 2014. Pharmacokinetics, metabolism and distribution of PEGs and PEGylated proteins: quo vadis? *Drug Discovery Today,* 19**,** 1623-1631.

BEERMANN, B., GROSCHINSKY-GRIND, M. & LINDSTROM, B. 1977. Pharmacokinetics of bendroflumethiazide. *Clin Pharmacol Ther,* 22**,** 385-8.

BEHESHTI, I., WESSELS, L. M. & ECKFELDT, J. H. 1994. EDTA-plasma vs serum differences in cholesterol, high-density-lipoprotein cholesterol, and triglyceride as measured by several methods. *Clinical Chemistry,* 40**,** 2088-2092.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological),* 57**,** 289-300.

BHARGAVA, H. N. & LEONARD, P. A. 1996. Triclosan: Applications and safety. *American Journal of Infection Control,* 24**,** 209-218.

BHINDERWALA, F., WASE, N., DIRUSSO, C. & POWERS, R. 2018. Combining Mass Spectrometry and NMR Improves Metabolite Detection and Annotation. *Journal of Proteome Research,* 17**,** 4017-4022.

BIDNY, S., GAGO, K., CHUNG, P., ALBERTYN, D. & PASIN, D. 2016. Simultaneous Screening and Quantification of Basic, Neutral and Acidic Drugs in Blood Using UPLC-QTOF-MS. *Journal of Analytical Toxicology,* 41**,** 181-195.

BIONDI, O., MOTTA, S. & MOSESSO, P. 2002. Low molecular weight polyethylene glycol induces chromosome aberrations in Chinese hamster cells cultured in vitro. *Mutagenesis,* 17**,** 261-264.

BLAŽENOVIĆ, I., KIND, T., JI, J. & FIEHN, O. 2018. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites,* 8**,** 31.

BLIGH, E. G. & DYER, W. J. 1959. A RAPID METHOD OF TOTAL LIPID EXTRACTION AND PURIFICATION. *Canadian Journal of Biochemistry and Physiology,* 37**,** 911-917.

BONNER, R. & HOPFGARTNER, G. 2019. SWATH data independent acquisition mass spectrometry for metabolomics. *TrAC Trends in Analytical Chemistry,* 120**,** 115278.

BOUATRA, S., AZIAT, F., MANDAL, R., GUO, A. C., WILSON, M. R., KNOX, C., BJORNDAHL, T. C., KRISHNAMURTHY, R., SALEEM, F., LIU, P., DAME, Z. T., POELZER, J., HUYNH, J., YALLOU, F. S., PSYCHOGIOS, N., DONG, E., BOGUMIL, R., ROEHRING, C. & WISHART, D. S. 2013. The human urine metabolome. *PLoS One,* 8**,** e73076.

BOYD SHAFFER, C., CRITCHFIELD, H. & NAIR, J. H. 1950. The Absorption and Excretion of a Liquid Polyethylene Glycol. *Journal of the American Pharmaceutical Association (Scientific ed.),* 39**,** 340-344.

BRAOUDAKI, M. & HILTON, A. C. 2004. Adaptive resistance to biocides in Salmonella enterica and Escherichia coli O157 and cross-resistance to antimicrobial agents. *J Clin Microbiol,* 42**,** 73-8.

BREYER-PFAFF, U. 2004. The metabolic fate of amitriptyline, nortriptyline and amitriptylinoxide in man. *Drug Metab Rev,* 36**,** 723-46.

BRO, R. & SMILDE, A. K. 2003. Centering and scaling in component analysis. *Journal of Chemometrics,* 17**,** 16-33.

BROADHURST, D., GOODACRE, R., REINKE, S. N., KULIGOWSKI, J., WILSON, I. D., LEWIS, M. R. & DUNN, W. B. 2018. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics,* 14**,** 72.

BROECKLING, C. D. & PRENNI, J. E. 2018. Stacked Injections of Biphasic Extractions for Improved Metabolomic Coverage and Sample Throughput. *Analytical Chemistry,* 90**,** 1147-1153.

BRUCE, S. J., TAVAZZI, I., PARISOD, V., REZZI, S., KOCHHAR, S. & GUY, P. A. 2009. Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. *Anal Chem,* 81**,** 3285-96.

BRUCE, S. J., TAVAZZI, I., PARISOD, V., REZZI, S., KOCHHAR, S. & GUY, P. A. 2009. Investigation of Human Blood Plasma Sample Preparation for Performing Metabolomics Using Ultrahigh Performance Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry,* 81**,** 3285-3296.

BRUNS, D. E., HEROLD, D. A., RODEHEAVER, G. T. & EDLICH, R. F. 1982. Polyethylene glycol intoxication in burn patients. *Burns,* 9**,** 49-52.

BUNDY, J. G., DAVEY, M. P. & VIANT, M. R. 2008. Environmental metabolomics: a critical review and future perspectives. *Metabolomics,* 5**,** 3.

CAI, X. & LI, R. 2016. Concurrent profiling of polar metabolites and lipids in human plasma using HILIC-FTMS. *Scientific reports,* 6**,** 36490-36490.

CAI, X. & LI, R. 2016. Concurrent profiling of polar metabolites and lipids in human plasma using HILIC-FTMS. *Scientific Reports,* 6**,** 36490.

CAJKA, T. & FIEHN, O. 2014. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. *Trends in analytical chemistry : TRAC,* 61**,** 192-206.

CALAFAT, A. M., YE, X., WONG, L. Y., REIDY, J. A. & NEEDHAM, L. L. 2008. Urinary concentrations of triclosan in the U.S. population: 2003-2004. *Environ Health Perspect,* 116**,** 303-7.

CALDERÓN-SANTIAGO, M., FERNÁNDEZ-PERALBO, M. A., PRIEGO-CAPOTE, F. & LUQUE DE CASTRO, M. D. 2016. MSCombine: a tool for merging untargeted metabolomic data from high-resolution mass spectrometry in the positive and negative ionization modes. *Metabolomics,* 12**,** 43.

CAO, G., SONG, Z., HONG, Y., YANG, Z., SONG, Y., CHEN, Z., CHEN, Z. & CAI, Z. 2020. Large-scale targeted metabolomics method for metabolite profiling of human samples. *Analytica Chimica Acta,* 1125**,** 144-151.

CAREY, D. E. & MCNAMARA, P. J. 2015. The impact of triclosan on the spread of antibiotic resistance in the environment. *Frontiers in Microbiology,* 5.

CAVAZZUTI, M. 2014. *Optimization Methods From Theory to Design Scientific and Technological Aspects in Mechanics,* Berlin, Springer Berlin.

CHACÓN-QUEVEDO, A., EGUARAS, M. G., CALLEJA, F., GARCIA, M. A., ROMAN, M., CASARES, J., MUÑOZ, I. & CONCHA, M. 1994. Comparative evaluation of pentoxifylline, buflomedil, and nifedipine in the treatment of intermittent claudication of the lower limbs. *Angiology,* 45**,** 647-53.

CHADHA, V., GARG, U. & ALON, U. 2001. Measurement of urinary concentration: A critical appraisal of methodologies. *Pediatric nephrology (Berlin, Germany),* 16**,** 374-82.

CHADHA, V., GARG, U. & ALON, U. S. 2001. Measurement of urinary concentration: A critical appraisal of methodologies. *Pediatric Nephrology,* 16**,** 374-382.

CHALCRAFT, K. R. & MCCARRY, B. E. 2013. Tandem LC columns for the simultaneous retention of polar and nonpolar molecules in comprehensive metabolomics analysis. *Journal of Separation Science,* 36**,** 3478-3485.

CHAMBERS, M. C., MACLEAN, B., BURKE, R., AMODEI, D., RUDERMAN, D. L., NEUMANN, S., GATTO, L., FISCHER, B., PRATT, B., EGERTSON, J., HOFF, K., KESSNER, D., TASMAN, N., SHULMAN, N., FREWEN, B., BAKER, T. A., BRUSNIAK, M. Y., PAULSE, C., CREASY, D., FLASHNER, L., KANI, K., MOULDING, C., SEYMOUR, S. L., NUWAYSIR, L. M., LEFEBVRE, B., KUHLMANN, F., ROARK, J., RAINER, P., DETLEV, S., HEMENWAY, T., HUHMER, A., LANGRIDGE, J., CONNOLLY, B., CHADICK, T., HOLLY, K., ECKELS, J., DEUTSCH, E. W., MORITZ, R. L., KATZ, J. E., AGUS, D. B., MACCOSS, M., TABB, D. L. & MALLICK, P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol,* 30**,** 918-20.

CHAPARRO-ORTEGA, A., BETANCOURT, M., ROSAS, P., VÁZQUEZ-CUEVAS, F. G., CHAVIRA, R., BONILLA, E., CASAS, E. & DUCOLOMB, Y. 2018. Endocrine disruptor effect of perfluorooctane sulfonic acid (PFOS) and perfluorooctanoic acid (PFOA) on porcine ovarian cell steroidogenesis. *Toxicology in Vitro,* 46**,** 86-93.

CHEN, C., GONZALEZ, F. J. & IDLE, J. R. 2007. LC-MS-Based Metabolomics in Drug Metabolism. *Drug Metabolism Reviews,* 39**,** 581-597.

CHEN, J., AHN, K. C., GEE, N. A., AHMED, M. I., DULEBA, A. J., ZHAO, L., GEE, S. J., HAMMOCK, B. D. & LASLEY, B. L. 2008. Triclocarban Enhances Testosterone Action: A New Type of Endocrine Disruptor? *Endocrinology,* 149**,** 1173-1179.

CHENG, J., CHEN, C., KRISTOPHER, K. W., MANNA, S. K., SCERBA, M., FRIEDMAN, F. K., LUECKE, H., IDLE, J. R. & GONZALEZ, F. J. 2013. Identification of 2-piperidone as a biomarker of CYP2E1 activity through metabolomic phenotyping. *Toxicol Sci,* 135**,** 37-47.

CLAYTON, E., TAYLOR, S., WRIGHT, B. & WILSON, I. D. 1998. The application of high performance liquid chromatography, coupled to nuclear magnetic resonance spectroscopy and mass spectrometry (HPLC-NMR-MS), to the characterisation of ibuprofen metabolites from human urine. *Chromatographia,* 47**,** 264-270.

CLAYTON, T. A., BAKER, D., LINDON, J. C., EVERETT, J. R. & NICHOLSON, J. K. 2009. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 14728-14733.

CLOAREC, O., DUMAS, M. E., CRAIG, A., BARTON, R. H., TRYGG, J., HUDSON, J., BLANCHER, C., GAUGUIER, D., LINDON, J. C., HOLMES, E. & NICHOLSON, J. 2005. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal Chem,* 77**,** 1282-9.

CODY, R. B., LARAMÉE, J. A. & DURST, H. D. 2005. Versatile New Ion Source for the Analysis of Materials in Open Air under Ambient Conditions. *Analytical Chemistry,* 77**,** 2297-2302.

COOKS, R. G., OUYANG, Z., TAKATS, Z. & WISEMAN, J. M. 2006. Ambient Mass Spectrometry. *Science,* 311**,** 1566-1570.

COOPER, R. A., DUROCHER, J. R. & LESLIE, M. H. 1977. Decreased fluidity of red cell membrane lipids in abetalipoproteinemia. *The Journal of clinical investigation,* 60**,** 115-121.

COULIER, L., BAS, R., JESPERSEN, S., VERHEIJ, E., VAN DER WERF, M. J. & HANKEMEIER, T. 2006. Simultaneous Quantitative Analysis of Metabolites Using Ion-Pair Liquid Chromatography−Electrospray Ionization Mass Spectrometry. *Analytical Chemistry,* 78**,** 6573-6582.

CRAIG, A., CLOAREC, O., HOLMES, E., NICHOLSON, J. K. & LINDON, J. C. 2006. Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets. *Analytical Chemistry,* 78**,** 2262-2267.

CRAMER, J. S. 2002. The Origins of Logistic Regression. *Tinbergen Institute, Tinbergen Institute Discussion Papers*.

CREEK, D. J., DUNN, W. B., FIEHN, O., GRIFFIN, J. L., HALL, R. D., LEI, Z., MISTRIK, R., NEUMANN, S., SCHYMANSKI, E. L., SUMNER, L. W., TRENGOVE, R. & WOLFENDER, J.-L. 2014. Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics,* 10**,** 350-353.

CROAL, B. L., GLEN, A. C., KELLY, C. J. & LOGAN, R. W. 1998. Transient 5-oxoprolinuria (pyroglutamic aciduria) with systemic acidosis in an adult receiving antibiotic therapy. *Clin Chem,* 44**,** 336-40.

CROCKFORD, D. J., MAHER, A. D., AHMADI, K. R., BARRETT, A., PLUMB, R. S., WILSON, I. D. & NICHOLSON, J. K. 2008. 1H NMR and UPLC-MSE Statistical Heterospectroscopy: Characterization of Drug Metabolites (Xenometabolome) in Epidemiological Studies. *Analytical Chemistry,* 80**,** 6835-6844.

CROFT, K. 2014. Prescriptions Dispensed in the community. *In:* CENTER, P. A. P. C. H. A. S. C. I. (ed.). England.

CUBBON, S., ANTONIO, C., WILSON, J. & THOMAS-OATES, J. 2010. Metabolomic applications of HILIC-LC-MS. *Mass Spectrom Rev,* 29**,** 671-84.

CUMMINS, P. M., DOWLING, O. & O'CONNOR, B. F. 2011. Ion-exchange chromatography: basic principles and application to the partial purification of soluble mammalian prolyl oligopeptidase. *Methods Mol Biol,* 681**,** 215-28.

D’SOUZA, A. A. & SHEGOKAR, R. 2016. Polyethylene glycol (PEG): a versatile polymer for pharmaceutical applications. *Expert Opinion on Drug Delivery,* 13**,** 1257-1275.

DA SILVA, R. R., VARGAS, F., ERNST, M., NGUYEN, N. H., BOLLEDDU, S., DEL ROSARIO, K. K., TSUNODA, S. M., DORRESTEIN, P. C. & JARMUSCH, A. K. 2019. Computational Removal of Undesired Mass Spectral Features Possessing Repeat Units via a Kendrick Mass Filter. *Journal of The American Society for Mass Spectrometry,* 30**,** 268-277.

DALGAARD, L. & LARSEN, C. 1999. Metabolism and excretion of citalopram in man: identification of O-acyl- and N-glucuronides. *Xenobiotica,* 29**,** 1033-41.

DALSGAARD, P. W., RASMUSSEN, B. S., MÜLLER, I. B. & LINNET, K. 2012. Toxicological screening of basic drugs in whole blood using UPLC-TOF-MS. *Drug Testing and Analysis,* 4**,** 313-319.

DE SENA, A. R., DE ASSIS, S. A. & BRANCO, A. 2011. Analysis of Theobromine and Related Compounds by Reversed Phase High-Performance Liquid Chromatography with Ultraviolet Detection: An Update (1992-2011). *Food Technology and Biotechnology,* 49**,** 413-423.

DEDA, O., CHATZIIOANNOU, A. C., FASOULA, S., PALACHANIS, D., RAIKOS, N., THEODORIDIS, G. A. & GIKA, H. G. 2017. Sample preparation optimization in fecal metabolic profiling. *Journal of Chromatography B,* 1047**,** 115-123.

DESCAMPS, C., CABRERA, G., DEPIERREUX, M. & DEVIERE, J. 2000. Acute renal insufficiency after colon cleansing. *Endoscopy,* 32**,** S11-S11.

DETTMER, K., ARONOV, P. A. & HAMMOCK, B. D. 2007. Mass spectrometry-based metabolomics. *Mass spectrometry reviews,* 26**,** 51-78.

DHILLON, G., KAUR, S., PULICHARLA, R., BRAR, S., CLEDON, M., VERMA, M. & SURAMPALLI, R. 2015. Triclosan: current status, occurrence, environmental risks and bioaccumulation potential. *International Journal of Environmental Research and Public Health,* 12**,** 5657-5684.

DHILLON, G. S., KAUR, S., PULICHARLA, R., BRAR, S. K., CLEDÓN, M., VERMA, M. & SURAMPALLI, R. Y. 2015. Triclosan: current status, occurrence, environmental risks and bioaccumulation potential. *International journal of environmental research and public health,* 12**,** 5657-5684.

DI TERLIZZI, R. & PLATT, S. 2006. The function, composition and analysis of cerebrospinal fluid in companion animals: Part I – Function and composition. *The Veterinary Journal,* 172**,** 422-431.

DICKSON, M. & GAGNON, J. P. 2004. The cost of new drug discovery and development. *Discov Med,* 4**,** 172-9.

DIETERLE, F., ROSS, A., SCHLOTTERBECK, G. & SENN, H. 2006. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Analytical Chemistry,* 78**,** 4281-4290.

DOMINGUEZ-ROMERO, J. C., GARCIA-REYES, J. F., MARTINEZ-ROMERO, R., MARTINEZ-LARA, E., DEL MORAL-LEAL, M. L. & MOLINA-DIAZ, A. 2013. Detection of main urinary metabolites of beta2-agonists clenbuterol, salbutamol and terbutaline by liquid chromatography high resolution mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci,* 923-924**,** 128-35.

DONA, A. C., JIMENEZ, B., SCHAFER, H., HUMPFER, E., SPRAUL, M., LEWIS, M. R., PEARCE, J. T., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2014. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem,* 86**,** 9887-94.

DONA, A. C., JIMÉNEZ, B., SCHÄFER, H., HUMPFER, E., SPRAUL, M., LEWIS, M. R., PEARCE, J. T. M., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2014. Precision High-Throughput Proton NMR Spectroscopy of Human Urine, Serum, and Plasma for Large-Scale Metabolic Phenotyping. *Analytical Chemistry,* 86**,** 9887-9894.

DONIA, M. S. & FISCHBACH, M. A. 2015. HUMAN MICROBIOTA. Small molecules from the human microbiota. *Science,* 349**,** 1254766.

DRESEN, S., FERREIRÓS, N., GNANN, H., ZIMMERMANN, R. & WEINMANN, W. 2010. Detection and identification of 700 drugs by multi-target screening with a 3200 Q TRAP® LC-MS/MS system and library searching. *Analytical and Bioanalytical Chemistry,* 396**,** 2425-2434.

DROUIN, N., RUDAZ, S. & SCHAPPLER, J. 2018. Sample preparation for polar metabolites in bioanalysis. *Analyst,* 143**,** 16-20.

DU PREL, J.-B., HOMMEL, G., RÖHRIG, B. & BLETTNER, M. 2009. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. *Deutsches Arzteblatt international,* 106**,** 335-339.

DUBUIS, E., WORTLEY, M. A., GRACE, M. S., MAHER, S. A., ADCOCK, J. J., BIRRELL, M. A. & BELVISI, M. G. 2014. Theophylline inhibits the cough reflex through a novel mechanism of action. *J Allergy Clin Immunol,* 133**,** 1588-98.

DUMAS, M.-E., MAIBAUM, E. C., TEAGUE, C., UESHIMA, H., ZHOU, B., LINDON, J. C., NICHOLSON, J. K., STAMLER, J., ELLIOTT, P., CHAN, Q. & HOLMES, E. 2006. Assessment of Analytical Reproducibility of 1H NMR Spectroscopy Based Metabonomics for Large-Scale Epidemiological Research: the INTERMAP Study. *Analytical Chemistry,* 78**,** 2199-2208.

DUNN, O. J. 1961. Multiple Comparisons among Means. *Journal of the American Statistical Association,* 56**,** 52-64.

DUNN, W. B., BROADHURST, D., BEGLEY, P., ZELENA, E., FRANCIS-MCINTYRE, S., ANDERSON, N., BROWN, M., KNOWLES, J. D., HALSALL, A., HASELDEN, J. N., NICHOLLS, A. W., WILSON, I. D., KELL, D. B. & GOODACRE, R. 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc,* 6**,** 1060-83.

DUNN, W. B., BROADHURST, D., BEGLEY, P., ZELENA, E., FRANCIS-MCINTYRE, S., ANDERSON, N., BROWN, M., KNOWLES, J. D., HALSALL, A., HASELDEN, J. N., NICHOLLS, A. W., WILSON, I. D., KELL, D. B., GOODACRE, R. & THE HUMAN SERUM METABOLOME, C. 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols,* 6**,** 1060-1083.

EFRON, B. & TIBSHIRANI, R. 1993. *An introduction to the bootstrap,* New York, Chapman and Hall.

EFSA 2007. Opinion of the Scientific Panel on food additives, flavourings, processing aids and materials in contact with food (AFC) related to an application on the use of polyethylene glycol (PEG) as a film coating agent for use in food supplement products. *EFSA Journal.* John Wiley & Sons, Ltd.

EFSA 2015. Current EU approved additives and their E Numbers.

EJIGU, B. A., VALKENBORG, D., BAGGERMAN, G., VANAERSCHOT, M., WITTERS, E., DUJARDIN, J.-C., BURZYKOWSKI, T. & BERG, M. 2013. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *Omics : a journal of integrative biology,* 17**,** 473-485.

ELLIOTT, P., VERGNAUD, A. C., SINGH, D., NEASHAM, D., SPEAR, J. & HEARD, A. 2014. The Airwave Health Monitoring Study of police officers and staff in Great Britain: rationale, design and methods. *Environ Res,* 134**,** 280-5.

ERICKSON, T. B., AKS, S. E., ZABANEH, R. & REID, R. 1996. Acute Renal Toxicity After Ingestion of Lava Light Liquid. *Annals of Emergency Medicine,* 27**,** 781-784.

FANG, J. L., STINGLEY, R. L., BELAND, F. A., HARROUK, W., LUMPKINS, D. L. & HOWARD, P. 2010. Occurrence, efficacy, metabolism, and toxicity of triclosan. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev,* 28**,** 147-71.

FANG, J. L., VANLANDINGHAM, M., DA COSTA, G. G. & BELAND, F. A. 2016. Absorption and metabolism of triclosan after application to the skin of B6C3F1 mice. *Environ Toxicol,* 31**,** 609-23.

FAULAND, A., KÖFELER, H., TRÖTZMÜLLER, M., KNOPF, A., HARTLER, J., EBERL, A., CHITRAJU, C., LANKMAYR, E. & SPENER, F. 2011. A comprehensive method for lipid profiling by liquid chromatography-ion cyclotron resonance mass spectrometry. *Journal of lipid research,* 52**,** 2314-2322.

FDA 2014. FDA. Food Additives & Ingredients - Food Additive Status List. .

FEI, F., BOWDISH, D. M. E. & MCCARRY, B. E. 2014. Comprehensive and simultaneous coverage of lipid and polar metabolites for endogenous cellular metabolomics using HILIC-TOF-MS. *Analytical and Bioanalytical Chemistry,* 406**,** 3723-3733.

FERNÁNDEZ-PERALBO, M. A. & LUQUE DE CASTRO, M. D. 2012. Preparation of urine samples prior to targeted or untargeted metabolomics mass-spectrometry analysis. *TrAC Trends in Analytical Chemistry,* 41**,** 75-85.

FIEHN, O., KOPKA, J., DORMANN, P., ALTMANN, T., TRETHEWEY, R. N. & WILLMITZER, L. 2000. Metabolite profiling for plant functional genomics. *Nat Biotechnol,* 18**,** 1157-61.

FIEHN, O., KOPKA, J., TRETHEWEY, R. N. & WILLMITZER, L. 2000. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal Chem,* 72**,** 3573-80.

FIEHN, O., WOHLGEMUTH, G., SCHOLZ, M., KIND, T., LEE, D. Y., LU, Y., MOON, S. & NIKOLAU, B. 2008. Quality control for plant metabolomics: reporting MSI-compliant studies. *Plant J,* 53**,** 691-704.

FISHER, R. A. 1992. Statistical Methods for Research Workers. *In:* KOTZ, S. & JOHNSON, N. L. (eds.) *Breakthroughs in Statistics: Methodology and Distribution.* New York, NY: Springer New York.

FOLCH, J., LEES, M. & SLOANE STANLEY, G. H. 1957. A simple method for the isolation and purification of total lipides from animal tissues. *J Biol Chem,* 226**,** 497-509.

FORCISI, S., MORITZ, F., KANAWATI, B., TZIOTIS, D., LEHMANN, R. & SCHMITT-KOPPLIN, P. 2013. Liquid chromatography–mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal of Chromatography A,* 1292**,** 51-65.

FORSBERG, E. M., HUAN, T., RINEHART, D., BENTON, H. P., WARTH, B., HILMERS, B. & SIUZDAK, G. 2018. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nature protocols,* 13**,** 633-651.

FRIGERIO, N. A. & HETTINGER, T. P. 1962. Protein solubility in solvent mixtures of low dielectric constant. *Biochimica et Biophysica Acta,* 59**,** 228-230.

FRIMAN, S., EGESTAD, B., SJÖVALL, J. & SVANVIK, J. 1993. Hepatic excretion and metabolism of polyethylene glycols and mannitol in the cat. *Journal of Hepatology,* 17**,** 48-55.

FROST, J., LOKKEN, T. N., BREDE, W. R., HEGSTAD, S., NORDRUM, I. S. & SLORDAL, L. 2015. A Validated Method for Simultaneous Determination of Codeine, Codeine-6-Glucuronide, Norcodeine, Morphine, Morphine-3-Glucuronide and Morphine-6-Glucuronide in Post-Mortem Blood, Vitreous Fluid, Muscle, Fat and Brain Tissue by LC-MS. *J Anal Toxicol,* 39**,** 203-12.

FRUIJTIER-PÖLLOTH, C. 2005. Safety assessment on polyethylene glycols (PEGs) and their derivatives as used in cosmetic products. *Toxicology,* 214**,** 1-38.

FUKUSAKI, E. & KOBAYASHI, A. 2005. Plant metabolomics: potential for practical operation. *Journal of Bioscience and Bioengineering,* 100**,** 347-354.

FURGE, L. L. & GUENGERICH, F. P. 2006. Cytochrome P450 enzymes in drug metabolism and chemical toxicology. *Biochemistry and Molecular Biology Education,* 34**,** 66-74.

GALL, W. E., BEEBE, K., LAWTON, K. A., ADAM, K.-P., MITCHELL, M. W., NAKHLE, P. J., RYALS, J. A., MILBURN, M. V., NANNIPIERI, M., CAMASTRA, S., NATALI, A., FERRANNINI, E. & FOR THE THE, R. S. G. 2010. α-Hydroxybutyrate Is an Early Biomarker of Insulin Resistance and Glucose Intolerance in a Nondiabetic Population. *PLOS ONE,* 5**,** e10883.

GÁLVEZ-ONTIVEROS, Y., PÁEZ, S., MONTEAGUDO, C. & RIVAS, A. 2020. Endocrine Disruptors in Food: Impact on Gut Microbiota and Metabolic Diseases. *Nutrients,* 12**,** 1158.

GARCÍA-CAÑAVERAS, J. C., DONATO, M. T., CASTELL, J. V. & LAHOZ, A. 2011. A Comprehensive Untargeted Metabonomic Analysis of Human Steatotic Liver Tissue by RP and HILIC Chromatography Coupled to Mass Spectrometry Reveals Important Metabolic Alterations. *Journal of Proteome Research,* 10**,** 4825-4834.

GARDINER, S. J. & BEGG, E. J. 2006. Pharmacogenetics, drug-metabolizing enzymes, and clinical practice. *Pharmacol Rev,* 58**,** 521-90.

GAULKE, C. A., BARTON, C. L., PROFFITT, S., TANGUAY, R. L. & SHARPTON, T. J. 2016. Triclosan Exposure Is Associated with Rapid Restructuring of the Microbiome in Adult Zebrafish. *PLOS ONE,* 11**,** e0154632.

GIL, A., ZHANG, W., WOLTERS, J. C., PERMENTIER, H., BOER, T., HORVATOVICH, P., HEINER-FOKKEMA, M. R., REIJNGOUD, D.-J. & BISCHOFF, R. 2018. One- vs two-phase extraction: re-evaluation of sample preparation procedures for untargeted lipidomics in plasma samples. *Analytical and Bioanalytical Chemistry,* 410**,** 5859-5870.

GONG, J., GU, X., ACHANZAR, W. E., CHADWICK, K. D., GAN, J., BROCK, B. J., KISHNANI, N. S., HUMPHREYS, W. G. & IYER, R. A. 2014. Quantitative analysis of polyethylene glycol (PEG) and PEGylated proteins in animal tissues by LC-MS/MS coupled with in-source CID. *Anal Chem,* 86**,** 7642-9.

GONZÁLEZ-DOMÍNGUEZ, R., GONZÁLEZ-DOMÍNGUEZ, Á., SAYAGO, A. & FERNÁNDEZ-RECAMALES, Á. 2020. Recommendations and Best Practices for Standardizing the Pre-Analytical Processing of Blood and Urine Samples in Metabolomics. *Metabolites,* 10**,** 229.

GONZÁLEZ-DONCEL, M., FERNÁNDEZ TORIJA, C., PABLOS, M. V., GARCÍA HORTIGÜELA, P., LÓPEZ ARÉVALO, M. & BELTRÁN, E. M. 2020. The role of PFOS on triclosan toxicity to two model freshwater organisms. *Environmental Pollution,* 263**,** 114604.

GONZALEZ-MARINO, I., QUINTANA, J. B., RODRIGUEZ, I. & CELA, R. 2009. Simultaneous determination of parabens, triclosan and triclocarban in water by liquid chromatography/electrospray ionisation tandem mass spectrometry. *Rapid Commun Mass Spectrom,* 23**,** 1756-66.

GONZÁLEZ-RIANO, C., DUDZIK, D., GARCIA, A., GIL-DE-LA-FUENTE, A., GRADILLAS, A., GODZIEN, J., LÓPEZ-GONZÁLVEZ, Á., REY-STOLLE, F., ROJO, D., RUPEREZ, F. J., SAIZ, J. & BARBAS, C. 2020. Recent Developments along the Analytical Process for Metabolomics Workflows. *Analytical Chemistry,* 92**,** 203-226.

GOODPASTER, A. M., RAMADAS, E. H. & KENNEDY, M. A. 2011. Potential Effect of Diaper and Cotton Ball Contamination on NMR- and LC/MS-Based Metabonomics Studies of Urine from Newborn Babies. *Analytical Chemistry,* 83**,** 896-902.

GOUTMAN, S., BOSS, J., GUO, K., ALAKWAA, F., PATTERSON, A., KIM, S., SAVELIEFF, M., HUR, J. & FELDMAN, E. 2020. Untargeted metabolomics yields insight into ALS disease mechanisms. *Journal of neurology, neurosurgery, and psychiatry*.

GRIBBLE, G. W. 2003. The diversity of naturally produced organohalogens. *Chemosphere,* 52**,** 289-97.

GU, H., PAN, Z., XI, B., HAINLINE, B. E., SHANAIAH, N., ASIAGO, V., GOWDA, G. A. & RAFTERY, D. 2009. 1H NMR metabolomics study of age profiling in children. *NMR Biomed,* 22**,** 826-33.

GUIJAS, C., MONTENEGRO-BURKE, J. R., DOMINGO-ALMENARA, X., PALERMO, A., WARTH, B., HERMANN, G., KOELLENSPERGER, G., HUAN, T., URITBOONTHAI, W., AISPORNA, A. E., WOLAN, D. W., SPILKER, M. E., BENTON, H. P. & SIUZDAK, G. 2018. METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Analytical chemistry,* 90**,** 3156-3164.

GUILHAUS, M. 1994. Spontaneous and deflected drift-trajectories in orthogonal acceleration time-of-flight mass spectrometry. *J Am Soc Mass Spectrom,* 5**,** 588-95.

GULLAPALLI, R. P. & MAZZITELLI, C. L. 2015. Polyethylene glycols in oral and parenteral formulations—A critical review. *International Journal of Pharmaceutics,* 496**,** 219-239.

HAGGARTY, J. & BURGESS, K. E. V. 2017. Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Current Opinion in Biotechnology,* 43**,** 77-85.

HAGINAKA, J. & WAKAI, J. 1987. Liquid chromatographic determination of amoxicillin and its metabolites in human urine by postcolumn degradation with sodium hypochlorite. *Journal of Chromatography B: Biomedical Sciences and Applications,* 413**,** 219-226.

HALDEN, R. U., LINDEMAN, A. E., AIELLO, A. E., ANDREWS, D., ARNOLD, W. A., FAIR, P., FUOCO, R. E., GEER, L. A., JOHNSON, P. I., LOHMANN, R., MCNEILL, K., SACKS, V. P., SCHETTLER, T., WEBER, R., ZOELLER, R. T. & BLUM, A. 2017. The Florence Statement on Triclosan and Triclocarban. *Environmental health perspectives,* 125**,** 064501-064501.

HANIOKA, N., OMAE, E., NISHIMURA, T., JINNO, H., ONODERA, S., YODA, R. & ANDO, M. 1996. Interaction of 2,4,4'-trichloro-2'-hydroxydiphenyl ether with microsomal cytochrome P450-dependent monooxygenases in rat liver. *Chemosphere,* 33**,** 265-76.

HAWKINS, D. M. 2004. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences,* 44**,** 1-12.

HEINZE, G., WALLISCH, C. & DUNKLER, D. 2018. Variable selection - A review and recommendations for the practicing statistician. *Biometrical journal. Biometrische Zeitschrift,* 60**,** 431-449.

HENDERSON, D., OGILVIE, L. A., HOYLE, N., KEILHOLZ, U., LANGE, B., LEHRACH, H. & CONSORTIUM, O. 2014. Personalized medicine approaches for colon cancer driven by genomics and systems biology: OncoTrack. *Biotechnology Journal,* 9**,** 1104-1114.

HENRY, N. D. & FAIR, P. A. 2013. Comparison of in vitro cytotoxicity, estrogenicity and anti-estrogenicity of triclosan, perfluorooctane sulfonate and perfluorooctanoic acid. *Journal of Applied Toxicology,* 33**,** 265-272.

HERNANDES, M. Z., CAVALCANTI, S. M., MOREIRA, D. R., DE AZEVEDO JUNIOR, W. F. & LEITE, A. C. 2010. Halogen atoms in the modern medicinal chemistry: hints for the drug design. *Curr Drug Targets,* 11**,** 303-14.

HEROLD, D. A., KEIL, K. & BRUNS, D. E. 1989. Oxidation of polyethylene glycols by alcohol dehydrogenase. *Biochemical Pharmacology,* 38**,** 73-76.

HEROLD, D. A., RODEHEAVER, G. T., BELLAMY, W. T., FITTON, L. A., BRUNS, D. E. & EDLICH, R. F. 1982. Toxicity of topical polyethylene glycol. *Toxicology and Applied Pharmacology,* 65**,** 329-335.

HODGE, K., HAVE, S. T., HUTTON, L. & LAMOND, A. I. 2013. Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS. *Journal of Proteomics,* 88**,** 92-103.

HOERL, A. E. & KENNARD, R. W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* 12**,** 55-67.

HOLCAPEK, M., KOLÁROVÁ, L. & NOBILIS, M. 2008. High-performance liquid chromatography-tandem mass spectrometry in the identification and determination of phase I and phase II drug metabolites. *Analytical and bioanalytical chemistry,* 391**,** 59-78.

HOLICK, M. F., DELUCA, H. F. & AVIOLI, L. V. 1972. Isolation and identification of 25-hydroxycholecalciferol from human plasma. *Arch Intern Med,* 129**,** 56-61.

HOLMES, E., LOO, R. L., CLOAREC, O., COEN, M., TANG, H., MAIBAUM, E., BRUCE, S., CHAN, Q., ELLIOTT, P., STAMLER, J., WILSON, I. D., LINDON, J. C. & NICHOLSON, J. K. 2007. Detection of urinary drug metabolite (xenometabolome) signatures in molecular epidemiology studies via statistical total correlation (NMR) spectroscopy. *Analytical chemistry,* 79**,** 2629-2640.

HOLMES, E., LOO, R. L., CLOAREC, O., COEN, M., TANG, H., MAIBAUM, E., BRUCE, S., CHAN, Q., ELLIOTT, P., STAMLER, J., WILSON, I. D., LINDON, J. C. & NICHOLSON, J. K. 2007. Detection of urinary drug metabolite (xenometabolome) signatures in molecular epidemiology studies via statistical total correlation (NMR) spectroscopy. *Anal Chem,* 79**,** 2629-40.

HOLMES, E., LOO, R. L., STAMLER, J., BICTASH, M., YAP, I. K., CHAN, Q., EBBELS, T., DE IORIO, M., BROWN, I. J., VESELKOV, K. A., DAVIGLUS, M. L., KESTELOOT, H., UESHIMA, H., ZHAO, L., NICHOLSON, J. K. & ELLIOTT, P. 2008. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature,* 453**,** 396-400.

HOLMES, E., LOO, R. L., STAMLER, J., BICTASH, M., YAP, I. K. S., CHAN, Q., EBBELS, T., DE IORIO, M., BROWN, I. J., VESELKOV, K. A., DAVIGLUS, M. L., KESTELOOT, H., UESHIMA, H., ZHAO, L., NICHOLSON, J. K. & ELLIOTT, P. 2008. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature,* 453**,** 396-400.

HORIKIRI, Y., SUZUKI, T. & MIZOBE, M. 1998. Pharmacokinetics and metabolism of bisoprolol enantiomers in humans. *J Pharm Sci,* 87**,** 289-94.

HOUTEN, S. M., CHEN, J., BELPOGGI, F., MANSERVISI, F., SÁNCHEZ-GUIJO, A., WUDY, S. A. & TEITELBAUM, S. L. 2016. Changes in the Metabolome in Response to Low-Dose Exposure to Environmental Chemicals Used in Personal Care Products during Different Windows of Susceptibility. *PLOS ONE,* 11**,** e0159919.

HUNT, D. F., GIORDANI, A. B., RHODES, G. & HEROLD, D. A. 1982. Mixture analysis by triple-quadrupole mass spectrometry: metabolic profiling of urinary carboxylic acids. *Clinical Chemistry,* 28**,** 2387.

HUTT, A. J., CALDWELL, J. & SMITH, R. L. 1986. The metabolism of aspirin in man: a population study. *Xenobiotica,* 16**,** 239-49.

IANIRO, G., MANGIOLA, F., DI RIENZO, T. A., BIBBO, S., FRANCESCHI, F., GRECO, A. V. & GASBARRINI, A. 2014. Levothyroxine absorption in health and disease, and new therapeutic perspectives. *Eur Rev Med Pharmacol Sci,* 18**,** 451-6.

IRVINE, G. B. 2001. Determination of molecular size by size-exclusion chromatography (gel filtration). *Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.],* Chapter 5**,** Unit 5.5-Unit 5.5.

ISLAS, G., IBARRA, I. S., HERNANDEZ, P., MIRANDA, J. M. & CEPEDA, A. 2017. Dispersive Solid Phase Extraction for the Analysis of Veterinary Drugs Applied to Food Samples: A Review. *International journal of analytical chemistry,* 2017**,** 8215271-8215271.

IVANISEVIC, J., ZHU, Z. J., PLATE, L., TAUTENHAHN, R., CHEN, S., O'BRIEN, P. J., JOHNSON, C. H., MARLETTA, M. A., PATTI, G. J. & SIUZDAK, G. 2013. Toward 'omic scale metabolite profiling: a dual separation-mass spectrometry approach for coverage of lipid and central carbon metabolism. *Anal Chem,* 85**,** 6876-84.

IZZI-ENGBEAYA, C., COMNINOS, A. N., CLARKE, S. A., JOMARD, A., YANG, L., JONES, S., ABBARA, A., NARAYANASWAMY, S., ENG, P. C., PAPADOPOULOU, D., PRAGUE, J. K., BECH, P., GODSLAND, I. F., BASSETT, P., SANDS, C., CAMUZEAUX, S., GOMEZ-ROMERO, M., PEARCE, J. T. M., LEWIS, M. R., HOLMES, E., NICHOLSON, J. K., TAN, T., RATNASABAPATHY, R., HU, M., CARRAT, G., PIEMONTI, L., BUGLIANI, M., MARCHETTI, P., JOHNSON, P. R., HUGHES, S. J., JAMES SHAPIRO, A. M., RUTTER, G. A. & DHILLO, W. S. 2018. The effects of kisspeptin on beta-cell function, serum metabolites and appetite in humans. *Diabetes Obes Metab,* 20**,** 2800-2810.

IZZI-ENGBEAYA, C., COMNINOS, A. N., CLARKE, S. A., JOMARD, A., YANG, L., JONES, S., ABBARA, A., NARAYANASWAMY, S., ENG, P. C., PAPADOPOULOU, D., PRAGUE, J. K., BECH, P., GODSLAND, I. F., BASSETT, P., SANDS, C., CAMUZEAUX, S., GOMEZ-ROMERO, M., PEARCE, J. T. M., LEWIS, M. R., HOLMES, E., NICHOLSON, J. K., TAN, T., RATNASABAPATHY, R., HU, M., CARRAT, G., PIEMONTI, L., BUGLIANI, M., MARCHETTI, P., JOHNSON, P. R., HUGHES, S. J., JAMES SHAPIRO, A. M., RUTTER, G. A. & DHILLO, W. S. 2018. The effects of kisspeptin on β-cell function, serum metabolites and appetite in humans. *Diabetes Obes Metab,* 20**,** 2800-2810.

JACKSON, A. A., PERSAUD, C., HALL, M., SMITH, S., EVANS, N. & RUTTER, N. 1997. Urinary excretion of 5-L-oxoproline (pyroglutamic acid) during early life in term and preterm infants. *Archives of Disease in Childhood - Fetal and Neonatal Edition,* 76**,** F152.

JACYNA, J., KORDALEWSKA, M. & MARKUSZEWSKI, M. 2018. Design of Experiments in metabolomics-related studies: An overview. *Journal of Pharmaceutical and Biomedical Analysis,* 164.

JAIN, A., LI, X. H. & CHEN, W. N. 2019. An untargeted fecal and urine metabolomics analysis of the interplay between the gut microbiome, diet and human metabolism in Indian and Chinese adults. *Scientific reports,* 9**,** 9191-9191.

JANA, A. & PAHAN, K. 2010. Sphingolipids in multiple sclerosis. *Neuromolecular medicine,* 12**,** 351-361.

JANG, H.-J., SHIN, C. Y. & KIM, K.-B. 2015. Safety Evaluation of Polyethylene Glycol (PEG) Compounds for Cosmetic Use. *Toxicological research,* 31**,** 105-136.

JÄNSCH, N., COLIN, F., SCHRÖDER, M. & MEYER-ALMES, F.-J. 2019. Using design of experiment to optimize enzyme activity assays. *ChemTexts,* 5**,** 20.

JEONG, T.-Y. & SIMPSON, M. J. 2019. Daphnia magna metabolic profiling as a promising water quality parameter for the biological early warning system. *Water Research,* 166**,** 115033.

JESCHKE, P. 2010. The unique role of halogen substituents in the design of modern agrochemicals. *Pest Manag Sci,* 66**,** 10-27.

JOHNSON, C. H. & GONZALEZ, F. J. 2012. Challenges and opportunities of metabolomics. *Journal of cellular physiology,* 227**,** 2975-2981.

JOHNSON, C. H., PATTERSON, A. D., IDLE, J. R. & GONZALEZ, F. J. 2012. Xenobiotic metabolomics: major impact on the metabolome. *Annual review of pharmacology and toxicology,* 52**,** 37-56.

JOHNSON, C. H., PATTERSON, A. D., IDLE, J. R. & GONZALEZ, F. J. 2012. Xenobiotic Metabolomics: Major Impact on the Metabolome. *Annual Review of Pharmacology and Toxicology,* 52**,** 37-56.

JOHNSON, K. A. & PLUMB, R. 2005. Investigating the human metabolism of acetaminophen using UPLC and exact mass oa-TOF MS. *J Pharm Biomed Anal,* 39**,** 805-10.

JONES, R. D., JAMPANI, H. B., NEWMAN, J. L. & LEE, A. S. 2000. Triclosan: A review of effectiveness and safety in health care settings. *American Journal of Infection Control,* 28**,** 184-196.

KAEVER, A., LANDESFEIND, M., FEUSSNER, K., MORGENSTERN, B., FEUSSNER, I. & MEINICKE, P. 2014. Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets. *PLOS ONE,* 9**,** e89297.

KAMLEH, M. A., DOW, J. A. T. & WATSON, D. G. 2008. Applications of mass spectrometry in metabolomic studies of animal model and invertebrate systems. *Briefings in Functional Genomics,* 8**,** 28-48.

KARZI, V., TZATZARAKIS, M., VAKONAKI, E., ALEGAKIS, T., KATSIKANTAMI, I., SIFAKIS, S., RIZOS, A. & TSATSAKIS, A. 2018. Biomonitoring of bisphenol A, triclosan and perfluorooctanoic acid in hair samples of children and adults: Biomonitoring of endocrine disruptors in hair samples. *Journal of Applied Toxicology,* 38.

KATAJAMAA, M. & OREŠIČ, M. 2005. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics,* 6**,** 179.

KAYE, S. B., LUBINSKI, J., MATULONIS, U., ANG, J. E., GOURLEY, C., KARLAN, B. Y., AMNON, A., BELL-MCGUINN, K. M., CHEN, L.-M., FRIEDLANDER, M., SAFRA, T., VERGOTE, I., WICKENS, M., LOWE, E. S., CARMICHAEL, J. & KAUFMAN, B. 2012. Phase II, Open-Label, Randomized, Multicenter Study Comparing the Efficacy and Safety of Olaparib, a Poly (ADP-Ribose) Polymerase Inhibitor, and Pegylated Liposomal Doxorubicin in Patients With BRCA1 or BRCA2 Mutations and Recurrent Ovarian Cancer. *Journal of Clinical Oncology,* 30**,** 372-379.

KEMSLEY, E. K. & TAPP, H. S. 2009. OPLS filtered data can be obtained directly from non-orthogonalized PLS1. *Journal of Chemometrics,* 23**,** 263-264.

KIM, S., CHEN, J., CHENG, T., GINDULYTE, A., HE, J., HE, S., LI, Q., SHOEMAKER, B. A., THIESSEN, P. A., YU, B., ZASLAVSKY, L., ZHANG, J. & BOLTON, E. E. 2018. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research,* 47**,** D1102-D1109.

KIND, T. & FIEHN, O. 2006. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics,* 7**,** 234.

KIND, T. & FIEHN, O. 2010. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews,* 2**,** 23-60.

KINNE-SAFFRAN, E. & KINNE, R. K. 1999. Vitalism and synthesis of urea. From Friedrich Wohler to Hans A. Krebs. *Am J Nephrol,* 19**,** 290-4.

KLONT, F., BRAS, L., WOLTERS, J. C., ONGAY, S., BISCHOFF, R., HALMOS, G. B. & HORVATOVICH, P. 2018. Assessment of Sample Preparation Bias in Mass Spectrometry-Based Proteomics. *Analytical Chemistry,* 90**,** 5405-5413.

KNEPPER, M. A., PACKER, R. & GOOD, D. W. 1989. Ammonium transport in the kidney. *Physiological Reviews,* 69**,** 179-249.

KOBAYASHI, K., CHIBA, K., SOHN, D. R., KATO, Y. & ISHIZAKI, T. 1992. Simultaneous determination of omeprazole and its metabolites in plasma and urine by reversed-phase high-performance liquid chromatography with an alkaline-resistant polymer-coated C18 column. *J Chromatogr,* 579**,** 299-305.

KOEHN, F. E. & CARTER, G. T. 2005. The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery,* 4**,** 206-220.

KOLE, P. L., VENKATESH, G., KOTECHA, J. & SHESHALA, R. 2011. Recent advances in sample preparation techniques for effective bioanalytical methods. *Biomedical Chromatography,* 25**,** 199-217.

KVALHEIM, O. M., BRAKSTAD, F. & LIANG, Y. 1994. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry,* 66**,** 43-51.

LAINE, G. A., HOSSAIN, S. M., SOLIS, R. T. & ADAMS, S. C. 1995. Polyethylene glycol nephrotoxicity secondary to prolonged high-dose intravenous lorazepam. *Ann Pharmacother,* 29**,** 1110-4.

LAÍNS, I., CHUNG, W., KELLY, R. S., GIL, J., MARQUES, M., BARRETO, P., MURTA, J. N., KIM, I. K., VAVVAS, D. G., MILLER, J. B., SILVA, R., LASKY-SU, J., LIANG, L., MILLER, J. W. & HUSAIN, D. 2019. Human Plasma Metabolomics in Age-Related Macular Degeneration: Meta-Analysis of Two Cohorts. *Metabolites,* 9.

LANGE, M. & FEDOROVA, M. 2020. Evaluation of lipid quantification accuracy using HILIC and RPLC MS on the example of NIST® SRM® 1950 metabolites in human plasma. *Analytical and Bioanalytical Chemistry,* 412**,** 3573-3584.

LARMENE-BELD, K. H., VRIES-KOENJER, M. T., TER HORST, P. G. & HOSPES, W. 2014. Development and validation of a liquid chromatography/tandem mass spectrometry method for the quantification of flucloxacillin and cloxacillin in microdialysis samples. *Biomed Chromatogr,* 28**,** 1096-101.

LATUS, J., KIMMEL, M., ALSCHER, M. D. & BRAUN, N. 2012. Ethylene glycol poisoning: a rare but life-threatening cause of metabolic acidosis-a single-centre experience. *Clinical kidney journal,* 5**,** 120-123.

LAU, C.-H. E., SISKOS, A. P., MAITRE, L., ROBINSON, O., ATHERSUCH, T. J., WANT, E. J., URQUIZA, J., CASAS, M., VAFEIADI, M., ROUMELIOTAKI, T., MCEACHAN, R. R. C., AZAD, R., HAUG, L. S., MELTZER, H. M., ANDRUSAITYTE, S., PETRAVICIENE, I., GRAZULEVICIENE, R., THOMSEN, C., WRIGHT, J., SLAMA, R., CHATZI, L., VRIJHEID, M., KEUN, H. C. & COEN, M. 2018. Determinants of the urinary and serum metabolome in children from six European populations. *BMC Medicine,* 16**,** 202.

LAWRENCE, J. R., TOPP, E., WAISER, M. J., TUMBER, V., ROY, J., SWERHONE, G. D. W., LEAVITT, P., PAULE, A. & KORBER, D. R. 2015. Resilience and recovery: The effect of triclosan exposure timing during development, on the structure and function of river biofilm communities. *Aquatic Toxicology,* 161**,** 253-266.

LEAPER, D., ASSADIAN, O., HUBNER, N. O., MCBAIN, A., BARBOLT, T., ROTHENBURGER, S. & WILSON, P. 2011. Antimicrobial sutures and prevention of surgical site infection: assessment of the safety of the antiseptic triclosan. *Int Wound J,* 8**,** 556-66.

LEAPER, D., MCBAIN, A. J., KRAMER, A., ASSADIAN, O., SANCHEZ, J. L., LUMIO, J. & KIERNAN, M. 2010. Healthcare associated infection: novel strategies and antimicrobial implants to prevent surgical site infection. *Ann R Coll Surg Engl,* 92**,** 453-8.

LEHUTSO, R. F., DASO, A. P. & OKONKWO, J. O. 2017. Occurrence and environmental levels of triclosan and triclocarban in selected wastewater treatment plants in Gauteng Province, South Africa. *Emerging Contaminants,* 3**,** 107-114.

LEI, Z., HUHMAN, D. V. & SUMNER, L. W. 2011. Mass spectrometry strategies in metabolomics. *The Journal of biological chemistry,* 286**,** 25435-25442.

LESCHE, D., GEYER, R., LIENHARD, D., NAKAS, C. T., DISERENS, G., VERMATHEN, P. & LEICHTLE, A. B. 2016. Does centrifugation matter? Centrifugal force and spinning time alter the plasma metabolome. *Metabolomics,* 12**,** 159.

LEUNG, A. W. Y., BACKSTROM, I. & BALLY, M. B. 2016. Sulfonation, an underexploited area: from skeletal development to infectious diseases and cancer. *Oncotarget,* 7**,** 55811-55827.

LEWIS, M. R., PEARCE, J. T., SPAGOU, K., GREEN, M., DONA, A. C., YUEN, A. H., DAVID, M., BERRY, D. J., CHAPPELL, K., HORNEFFER-VAN DER SLUIS, V., SHAW, R., LOVESTONE, S., ELLIOTT, P., SHOCKCOR, J., LINDON, J. C., CLOAREC, O., TAKATS, Z., HOLMES, E. & NICHOLSON, J. K. 2016. Development and Application of Ultra-Performance Liquid Chromatography-TOF MS for Precision Large Scale Urinary Metabolic Phenotyping. *Anal Chem,* 88**,** 9004-13.

LEWIS, M. R., PEARCE, J. T. M., SPAGOU, K., GREEN, M., DONA, A. C., YUEN, A. H. Y., DAVID, M., BERRY, D. J., CHAPPELL, K., HORNEFFER-VAN DER SLUIS, V., SHAW, R., LOVESTONE, S., ELLIOTT, P., SHOCKCOR, J., LINDON, J. C., CLOAREC, O., TAKATS, Z., HOLMES, E. & NICHOLSON, J. K. 2016. Development and Application of Ultra-Performance Liquid Chromatography-TOF MS for Precision Large Scale Urinary Metabolic Phenotyping. *Analytical Chemistry,* 88**,** 9004-9013.

LI, M., YANG, L., BAI, Y. & LIU, H. 2014. Analytical Methods in Lipidomics and Their Applications. *Analytical Chemistry,* 86**,** 161-175.

LI, R. & HUANG, J. 2006. Hydrophilic interaction chromatography and its applications in the separation of basic drugs. *Progress in Chemistry,* 18**,** 1508-1513.

LI, X., YING, G., ZHAO, J., CHEN, Z., LAI, H. & SU, H. 2013. 4-Nonylphenol, bisphenol-A and triclosan levels in human urine of children and students in China, and the effects of drinking these bottled materials on the levels. *Environment International,* 52**,** 81-86.

LI, Y., ZHANG, Z., LIU, X., LI, A., HOU, Z., WANG, Y. & ZHANG, Y. 2015. A novel approach to the simultaneous extraction and non-targeted analysis of the small molecules metabolome and lipidome using 96-well solid phase extraction plates with column-switching technology. *Journal of Chromatography A,* 1409**,** 277-281.

LIBISELLER, G., DVORZAK, M., KLEB, U., GANDER, E., EISENBERG, T., MADEO, F., NEUMANN, S., TRAUSINGER, G., SINNER, F., PIEBER, T. & MAGNES, C. 2015. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics,* 16**,** 118.

LIN, Y. J. 2000. Buccal absorption of triclosan following topical mouthrinse application. *Am J Dent,* 13**,** 215-7.

LINDON, J. C., HOLMES, E., BOLLARD, M. E., STANLEY, E. G. & NICHOLSON, J. K. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers,* 9**,** 1-31.

LINDON, J. C., NICHOLSON, J. K., HOLMES, E., ANTTI, H., BOLLARD, M. E., KEUN, H., BECKONERT, O., EBBELS, T. M., REILY, M. D., ROBERTSON, D., STEVENS, G. J., LUKE, P., BREAU, A. P., CANTOR, G. H., BIBLE, R. H., NIEDERHAUSER, U., SENN, H., SCHLOTTERBECK, G., SIDELMANN, U. G., LAURSEN, S. M., TYMIAK, A., CAR, B. D., LEHMAN-MCKEEMAN, L., COLET, J. M., LOUKACI, A. & THOMAS, C. 2003. Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol Appl Pharmacol,* 187**,** 137-46.

LINDSTRÖM, A., BUERGE, I. J., POIGER, T., BERGQVIST, P.-A., MÜLLER, M. D. & BUSER, H.-R. 2002. Occurrence and Environmental Behavior of the Bactericide Triclosan and Its Methyl Derivative in Surface Waters and in Wastewater. *Environmental Science & Technology,* 36**,** 2322-2329.

LIU, A. & COLEMAN, S. P. 2009. Determination of metformin in human plasma using hydrophilic interaction liquid chromatography-tandem mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci,* 877**,** 3695-700.

LIU, H., GARRETT, T. J., SU, Z., KHOO, C., ZHAO, S. & GU, L. 2020. Modifications of the urinary metabolome in young women after cranberry juice consumption were revealed using the UHPLC-Q-Orbitrap-HRMS-based metabolomics approach. *Food & Function,* 11**,** 2466-2476.

LIU, L., AA, J., WANG, G., YAN, B., ZHANG, Y., WANG, X., ZHAO, C., CAO, B., SHI, J., LI, M., ZHENG, T., ZHENG, Y., HAO, G., ZHOU, F., SUN, J. & WU, Z. 2010. Differences in metabolite profile between blood plasma and serum. *Analytical Biochemistry,* 406**,** 105-112.

LIU, X., HOENE, M., WANG, X., YIN, P., HÄRING, H.-U., XU, G. & LEHMANN, R. 2018. Serum or plasma, what is the difference? Investigations to facilitate the sample material selection decision making process for metabolomics studies and beyond. *Analytica Chimica Acta,* 1037**,** 293-300.

LIU, X. & JIA, L. 2007. The conduct of drug metabolism studies considered good practice (I): analytical systems and in vivo studies. *Current drug metabolism,* 8**,** 815-821.

LOCATELLI, I., KMETEC, V., MRHAR, A. & GRABNAR, I. 2005. Determination of warfarin enantiomers and hydroxylated metabolites in human blood plasma by liquid chromatography with achiral and chiral separation. *J Chromatogr B Analyt Technol Biomed Life Sci,* 818**,** 191-8.

LÖFGREN, L., STÅHLMAN, M., FORSBERG, G. B., SAARINEN, S., NILSSON, R. & HANSSON, G. I. 2012. The BUME method: a novel automated chloroform-free 96-well total lipid extraction method for blood plasma. *J Lipid Res,* 53**,** 1690-700.

LOMMEN, A. 2009. MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry,* 81**,** 3079-3086.

LOO, R. L., CHAN, Q., BROWN, I. J., ROBERTSON, C. E., STAMLER, J., NICHOLSON, J. K., HOLMES, E. & ELLIOTT, P. 2012. A comparison of self-reported analgesic use and detection of urinary ibuprofen and acetaminophen metabolites by means of metabonomics: the INTERMAP Study. *Am J Epidemiol,* 175**,** 348-58.

LOPEZ-AVILA, V. & HITES, R. A. 1980. Organic compounds in an industrial wastewater. Their transport into sediments. *Environmental Science & Technology,* 14**,** 1382-1390.

LOVESTONE, S., FRANCIS, P., KLOSZEWSKA, I., MECOCCI, P., SIMMONS, A., SOININEN, H., SPENGER, C., TSOLAKI, M., VELLAS, B., WAHLUND, L. O. & WARD, M. 2009. AddNeuroMed--the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Ann N Y Acad Sci,* 1180**,** 36-46.

LUNDGREN, B. & DEPIERRE, J. W. 1990. The metabolism of xenobiotics and its relationship to toxicity/genotoxicity: studies with human lymphocytes. *Acta Physiol Scand Suppl,* 592**,** 49-59.

MA, Q., XING, C., LONG, W., WANG, H. Y., LIU, Q. & WANG, R.-F. 2019. Impact of microbiota on central nervous system and neurological diseases: the gut-brain axis. *Journal of Neuroinflammation,* 16**,** 53.

MACWAN, J. S., IONITA, I. A., DOSTALEK, M. & AKHLAGHI, F. 2011. Development and validation of a sensitive, simple, and rapid method for simultaneous quantitation of atorvastatin and its acid and lactone metabolites by liquid chromatography-tandem mass spectrometry (LC-MS/MS). *Anal Bioanal Chem,* 400**,** 423-33.

MADALA, N., STEENKAMP, P., PIATER, L. & DUBERY, I. 2012. Collision energy alteration during mass spectrometric acquisition is essential to ensure unbiased metabolomic analysis. *Analytical and bioanalytical chemistry,* 404**,** 367-72.

MAROTTA, F., LECROIX, P., HARADA, M., MASULAIR, K., SAFRAN, P., LORENZETTI, A., ONO-NITA, S. K. & MARANDOLA, P. 2006. Liver exposure to xenobiotics: the aging factor and potentials for functional foods. *Rejuvenation Res,* 9**,** 338-41.

MATSUI, K., MISHIMA, M., NAGAI, Y., YUZURIHA, T. & YOSHIMURA, T. 1999. Absorption, Distribution, Metabolism, and Excretion of Donepezil (Aricept) after a Single Oral Administration to Rat. *Drug Metabolism and Disposition,* 27**,** 1406-1414.

MATYASH, V., LIEBISCH, G., KURZCHALIA, T. V., SHEVCHENKO, A. & SCHWUDKE, D. 2008. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of lipid research,* 49**,** 1137-1146.

MAURER, H. H. & MEYER, M. R. 2016. High-resolution mass spectrometry in toxicology: current status and future perspectives. *Archives of Toxicology,* 90**,** 2161-2172.

MAYNERT, E. W. 1961. Metabolic Fate of Drugs. *Annual Review of Pharmacology,* 1**,** 45-63.

MCAVOY, D. C., SCHATOWITZ, B., JACOB, M., HAUK, A. & ECKHOFF, W. S. 2002. Measurement of triclosan in wastewater treatment systems. *Environ Toxicol Chem,* 21**,** 1323-9.

MCNAMEE, R. 2005. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine,* 62**,** 500.

MENDEZ, K. M., BROADHURST, D. I. & REINKE, S. N. 2020. Migrating from partial least squares discriminant analysis to artificial neural networks: a comparison of functionally equivalent visualisation and feature contribution tools using jupyter notebooks. *Metabolomics,* 16**,** 17.

MICHOPOULOS, F., LAI, L., GIKA, H., THEODORIDIS, G. & WILSON, I. 2009. UPLC-MS-Based Analysis of Human Plasma for Metabonomics Using Solvent Precipitation or Solid Phase Extraction. *Journal of Proteome Research,* 8**,** 2114-2121.

MICHOPOULOS, F., THEODORIDIS, G., SMITH, C. J. & WILSON, I. D. 2010. Metabolite Profiles from Dried Biofluid Spots for Metabonomic Studies using UPLC Combined with oaToF-MS. *Journal of Proteome Research,* 9**,** 3328-3334.

MIHAILOVA, A., LUNDANES, E. & GREIBROKK, T. 2006. Determination and removal of impurities in 2-D LC-MS of peptides. *Journal of Separation Science,* 29**,** 576-581.

MILLER, G. W. & JONES, D. P. 2014. The nature of nurture: refining the definition of the exposome. *Toxicol Sci,* 137**,** 1-2.

MITRA, S. 2003. *Sample preparation techniques in analytical chemistry,* Hoboken [NJ], J. Wiley.

MOLTU, S. J., SACHSE, D., BLAKSTAD, E. W., STROMMEN, K., NAKSTAD, B., ALMAAS, A. N., WESTERBERG, A. C., RONNESTAD, A., BRAEKKE, K., VEIEROD, M. B., IVERSEN, P. O., RISE, F., BERG, J. P. & DREVON, C. A. 2014. Urinary metabolite profiles in premature infants show early postnatal metabolic adaptation and maturation. *Nutrients,* 6**,** 1913-30.

MORTIER, K. A., MAUDENS, K. E., LAMBERT, W. E., CLAUWAERT, K. M., VAN BOCXLAER, J. F., DEFORCE, D. L., VAN PETEGHEM, C. H. & DE LEENHEER, A. P. 2002. Simultaneous, quantitative determination of opiates, amphetamines, cocaine and benzoylecgonine in oral fluid by liquid chromatography quadrupole-time-of-flight mass spectrometry. *Journal of Chromatography B,* 779**,** 321-330.

MORTISHIRE-SMITH, R. J., O'CONNOR, D., CASTRO-PEREZ, J. M. & KIRBY, J. 2005. Accelerated throughput metabolic route screening in early drug discovery using high-resolution liquid chromatography/quadrupole time-of-flight mass spectrometry and automated data analysis. *Rapid Communications in Mass Spectrometry,* 19**,** 2659-2670.

MOSS, T., HOWES, D. & WILLIAMS, F. M. 2000. Percutaneous penetration and dermal metabolism of triclosan (2,4, 4'-trichloro-2'-hydroxydiphenyl ether). *Food Chem Toxicol,* 38**,** 361-70.

MUKAKA, M. M. 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi,* 24**,** 69-71.

MUSTAFA, M., WONDIMU, B., HULTENBY, K., YUCEL-LINDBERG, T. & MODEER, T. 2003. Uptake, distribution and release of 14C-triclosan in human gingival fibroblasts. *Journal of Pharmaceutical Sciences,* 92**,** 1648-1653.

NASER, F. J., MAHIEU, N. G., WANG, L., SPALDING, J. L., JOHNSON, S. L. & PATTI, G. J. 2018. Two complementary reversed-phase separations for comprehensive coverage of the semipolar and nonpolar metabolome. *Analytical and bioanalytical chemistry,* 410**,** 1287-1297.

NAYLOR, S. & CHEN, J. Y. 2010. Unraveling human complexity and disease with systems biology and personalized medicine. *Personalized medicine,* 7**,** 275-289.

NAZ, S., MOREIRA DOS SANTOS, D. C., GARCIA, A. & BARBAS, C. 2014. Analytical protocols based on LC-MS, GC-MS and CE-MS for nontargeted metabolomics of biological tissues. *Bioanalysis,* 6**,** 1657-77.

NEVILLE, D., HOUGHTON, R. & GARRETT, S. 2012. Efficacy of plasma phospholipid removal during sample preparation and subsequent retention under typical UHPLC conditions. *Bioanalysis,* 4**,** 795-807.

NICHOLSON, J. K. 2006. Global systems biology, personalized medicine and molecular epidemiology. *Molecular Systems Biology,* 2**,** 52-52.

NICHOLSON, J. K., HOLMES, E., KINROSS, J. M., DARZI, A. W., TAKATS, Z. & LINDON, J. C. 2012. Metabolic phenotyping in clinical and surgical environments. *Nature,* 491**,** 384-392.

NICHOLSON, J. K., HOLMES, E. & WILSON, I. D. 2005. Gut microorganisms, mammalian metabolism and personalized health care. *Nat Rev Microbiol,* 3**,** 431-8.

NICHOLSON, J. K., HOLMES, E. & WILSON, I. D. 2005. Gut microorganisms, mammalian metabolism and personalized health care. *Nature Reviews Microbiology,* 3**,** 431-438.

NICHOLSON, J. K., LINDON, J. C. & HOLMES, E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica,* 29**,** 1181-1189.

NIEDZWIECKI, M. M., WALKER, D. I., VERMEULEN, R., CHADEAU-HYAM, M., JONES, D. P. & MILLER, G. W. 2019. The Exposome: Molecules to Populations. *Annual Review of Pharmacology and Toxicology,* 59**,** 107-127.

NOMURA, D. K., MORRISON, B. E., BLANKMAN, J. L., LONG, J. Z., KINSEY, S. G., MARCONDES, M. C., WARD, A. M., HAHN, Y. K., LICHTMAN, A. H., CONTI, B. & CRAVATT, B. F. 2011. Endocannabinoid hydrolysis generates brain prostaglandins that promote neuroinflammation. *Science,* 334**,** 809-13.

OKESOLA, B. O., VIEIRA, V. M. P., CORNWELL, D. J., WHITELAW, N. K. & SMITH, D. K. 2015. 1,3:2,4-Dibenzylidene-d-sorbitol (DBS) and its derivatives – efficient, versatile and industrially-relevant low-molecular-weight gelators with over 100 years of history and a bright future. *Soft Matter,* 11**,** 4768-4787.

OKUMURA, T. & NISHIKAWA, Y. 1996. Gas chromatography—mass spectrometry determination of triclosans in water, sediment and fish samples via methylation with diazomethane. *Analytica Chimica Acta,* 325**,** 175-184.

ORDÓÑEZ, J. L., PEREIRA-CARO, G., LUDWIG, I., MUÑOZ-REDONDO, J. M., RUIZ-MORENO, M. J., CROZIER, A. & MORENO-ROJAS, J. M. 2018. A critical evaluation of the use of gas chromatography- and high performance liquid chromatography-mass spectrometry techniques for the analysis of microbial metabolites in human urine after consumption of orange juice. *Journal of Chromatography A,* 1575**,** 100-112.

OVCACIKOVA, M., LISA, M., CIFKOVA, E. & HOLCAPEK, M. 2016. Retention behavior of lipids in reversed-phase ultrahigh-performance liquid chromatography-electrospray ionization mass spectrometry. *J Chromatogr A,* 1450**,** 76-85.

P S, S., MULLANGI, R. & KUMAR S, S. 2014. Highly Sensitive LC-MS/MS Method for Determinationof Memantine in Rat Plasma: Application to Pharmacokinetic Studies in Rats. *Biomedical Chromatography,* 28.

PANNEK, J. & VESTWEBER, A. M. 2011. [Clinical utility of an antimicrobial blocking solution in patients with an indwelling catheter]. *Aktuelle Urol,* 42**,** 51-4.

PAPADIMITROPOULOS, M.-E. P., VASILOPOULOU, C. G., MAGA-NTEVE, C. & KLAPA, M. I. 2018. Untargeted GC-MS Metabolomics. *In:* THEODORIDIS, G. A., GIKA, H. G. & WILSON, I. D. (eds.) *Metabolic Profiling: Methods and Protocols.* New York, NY: Springer New York.

PARK, H. & KIM, K. 2018. Concentrations of 2,4-Dichlorophenol and 2,5-Dichlorophenol in Urine of Korean Adults. *International journal of environmental research and public health,* 15**,** 589.

PATTERSON, R. E., DUCROCQ, A. J., MCDOUGALL, D. J., GARRETT, T. J. & YOST, R. A. 2015. Comparison of blood plasma sample preparation methods for combined LC-MS lipidomics and metabolomics. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences,* 1002**,** 260-266.

PELLEGRINI, G., STARKEY LEWIS, P. J., PALMER, L., HETZEL, U., GOLDRING, C. E., PARK, B. K., KIPAR, A. & WILLIAMS, D. P. 2013. Intraperitoneal administration of high doses of polyethylene glycol (PEG) causes hepatic subcapsular necrosis and low-grade peritonitis with a rise in hepatic biomarkers. *Toxicology,* 314**,** 262-266.

PETERSEN, I. N., TORTZEN, C., KRISTENSEN, J. L., PEDERSEN, D. S. & BREINDAHL, T. 2013. Identification of a New Metabolite of GHB: Gamma-Hydroxybutyric Acid Glucuronide. *Journal of Analytical Toxicology,* 37**,** 291-297.

PLUMB, R. S., JOHNSON, K. A., RAINVILLE, P., SHOCKCOR, J. P., WILLIAMS, R., GRANGER, J. H. & WILSON, I. D. 2006. The detection of phenotypic differences in the metabolic plasma profile of three strains of Zucker rats at 20 weeks of age using ultra-performance liquid chromatography/orthogonal acceleration time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry,* 20**,** 2800-2806.

PLUSKAL, T., CASTILLO, S., VILLAR-BRIONES, A. & OREŠIČ, M. 2010. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics,* 11**,** 395.

PODWOJSKI, K., FRITSCH, A., CHAMRAD, D. C., PAUL, W., SITEK, B., STUHLER, K., MUTZEL, P., STEPHAN, C., MEYER, H. E., URFER, W., ICKSTADT, K. & RAHNENFUHRER, J. 2009. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics,* 25**,** 758-64.

POLSON, C., SARKAR, P., INCLEDON, B., RAGUVARAN, V. & GRANT, R. 2003. Optimization of protein precipitation based upon effectiveness of protein removal and ionization effect in liquid chromatography-tandem mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci,* 785**,** 263-75.

POLSON, C., SARKAR, P., INCLEDON, B., RAGUVARAN, V. & GRANT, R. 2003. Optimization of protein precipitation based upon effectiveness of protein removal and ionization effect in liquid chromatography–tandem mass spectrometry. *Journal of Chromatography B,* 785**,** 263-275.

POSMA, J. M. 2019. Chapter 9 - Multivariate Statistical Methods for Metabolic Phenotyping. *In:* LINDON, J. C., NICHOLSON, J. K. & HOLMES, E. (eds.) *The Handbook of Metabolic Phenotyping.* Elsevier.

POSMA, J. M., GARCIA-PEREZ, I., HEATON, J. C., BURDISSO, P., MATHERS, J. C., DRAPER, J., LEWIS, M., LINDON, J. C., FROST, G., HOLMES, E. & NICHOLSON, J. K. 2017. Integrated Analytical and Statistical Two-Dimensional Spectroscopy Strategy for Metabolite Identification: Application to Dietary Biomarkers. *Anal Chem,* 89**,** 3300-3309.

PRENTICE, D. E. & MAJEED, S. K. 1978. Oral toxicity of polyethylene glycol (PEG 200) in monkeys and rats. *Toxicology Letters,* 2**,** 119-122.

PROVENCHER, G., BÉRUBÉ, R., DUMAS, P., BIENVENU, J.-F., GAUDREAU, É., BÉLANGER, P. & AYOTTE, P. 2014. Determination of bisphenol A, triclosan and their metabolites in human urine using isotope-dilution liquid chromatography–tandem mass spectrometry. *Journal of Chromatography A,* 1348**,** 97-104.

PRUEKSARITANONT, T., GORHAM, L. M., MA, B., LIU, L., YU, X., ZHAO, J. J., SLAUGHTER, D. E., ARISON, B. H. & VYAS, K. P. 1997. In vitro metabolism of simvastatin in humans [SBT]identification of metabolizing enzymes and effect of the drug on hepatic P450s. *Drug Metab Dispos,* 25**,** 1191-9.

PSYCHOGIOS, N., HAU, D. D., PENG, J., GUO, A. C., MANDAL, R., BOUATRA, S., SINELNIKOV, I., KRISHNAMURTHY, R., EISNER, R., GAUTAM, B., YOUNG, N., XIA, J., KNOX, C., DONG, E., HUANG, P., HOLLANDER, Z., PEDERSEN, T. L., SMITH, S. R., BAMFORTH, F., GREINER, R., MCMANUS, B., NEWMAN, J. W., GOODFRIEND, T. & WISHART, D. S. 2011. The Human Serum Metabolome. *PLOS ONE,* 6**,** e16957.

R DEVELOPMENT CORE TEAM 2019. R: A language and environment for statistical computing. *R Foundation for Statistical computing*.

RAHMAN, A., ALI, M. T., SHAWAN, M. M. A. K., SARWAR, M. G., KHAN, M. A. K. & HALIM, M. A. 2016. Halogen-directed drug design for Alzheimer’s disease: a combined density functional and molecular docking study. *SpringerPlus,* 5**,** 1346.

RAHNAVARD, A., HITCHCOCK, D., PACHECO, J. A., DEIK, A., DENNIS, C., JEANFAVRE, S., PIERCE, K., BULLOCK, K., COSTLIOW, Z. & CLISH, C. B. 2018. netome: a computational framework for metabolite profiling and omics network analysis. *bioRxiv***,** 443903.

RAMIREZ-LOPEZ, A. S. A. L. 2020. An introduction to the prospectr package.

RANASINGHE, A., RAMANATHAN, R., JEMAL, M., D'ARIENZO, C. J., HUMPHREYS, W. G. & OLAH, T. V. 2012. Integrated quantitative and qualitative workflow for in vivo bioanalytical support in drug discovery using hybrid Q-TOF-MS. *Bioanalysis,* 4**,** 511-28.

RANGANATHAN, A., GEE, S. J. & HAMMOCK, B. D. 2015. An immunoassay for the detection of triclosan-O-glucuronide, a primary human urinary metabolite of triclosan. *Analytical and Bioanalytical Chemistry,* 407**,** 7263-7273.

RAPPAPORT, S. M., BARUPAL, D. K., WISHART, D., VINEIS, P. & SCALBERT, A. 2014. The blood exposome and its role in discovering causes of disease. *Environmental health perspectives,* 122**,** 769-774.

RATERINK, R.-J., LINDENBURG, P. W., VREEKEN, R. J., RAMAUTAR, R. & HANKEMEIER, T. 2014. Recent developments in sample-pretreatment techniques for mass spectrometry-based metabolomics. *TrAC Trends in Analytical Chemistry,* 61**,** 157-167.

RATHAHAO-PARIS, E., PARIS, A., BURSZTYKA, J., JAEG, J.-P., CRAVEDI, J.-P. & DEBRAUWER, L. 2014. Identification of xenobiotic metabolites from biological fluids using flow injection analysis high-resolution mass spectrometry and post-acquisition data filtering. *Rapid Communications in Mass Spectrometry,* 28**,** 2713-2722.

RAYLEIGH, L. 1882. XX. On the equilibrium of liquid conducting masses charged with electricity. *Philosophical Magazine Series 5,* 14**,** 184-186.

REIS FERREIRA, M., ANDREYEV, J., MOHAMMED, K., TRUELOVE, L., GOWAN, S. M., LI, J., GULLIFORD, S. L., MARCHESI, J. & DEARNALEY, D. P. 2019. Microbiota and radiotherapy-induced gastrointestinal side-effects (MARS) study: a large pilot study of the microbiome in acute and late radiation enteropathy. *Clinical Cancer Research***,** clincanres.0960.2019.

REN, S., HINZMAN, A. A., KANG, E. L., SZCZESNIAK, R. D. & LU, L. J. 2015. Computational and statistical analysis of metabolomics data. *Metabolomics,* 11**,** 1492-1513.

RICO, E., GONZALEZ, O., BLANCO, M. E. & ALONSO, R. M. 2014. Evaluation of human plasma sample preparation protocols for untargeted metabolic profiles analyzed by UHPLC-ESI-TOF-MS. *Anal Bioanal Chem,* 406**,** 7641-52.

RODRICKS, J. V., SWENBERG, J. A., BORZELLECA, J. F., MARONPOT, R. R. & SHIPP, A. M. 2010. Triclosan: a critical review of the experimental data and development of margins of safety for consumer products. *Crit Rev Toxicol,* 40**,** 422-84.

ROEMMELT, A. T., STEUER, A. E., POETZSCH, M. & KRAEMER, T. 2014. Liquid chromatography, in combination with a quadrupole time-of-flight instrument (LC QTOF), with sequential window acquisition of all theoretical fragment-ion spectra (SWATH) acquisition: systematic studies on its use for screenings in clinical and forensic toxicology and comparison with information-dependent acquisition (IDA). *Anal Chem,* 86**,** 11742-9.

ROSENLING, T., STOOP, M. P., SMOLINSKA, A., MUILWIJK, B., COULIER, L., SHI, S., DANE, A., CHRISTIN, C., SUITS, F., HORVATOVICH, P. L., WIJMENGA, S. S., BUYDENS, L. M., VREEKEN, R., HANKEMEIER, T., VAN GOOL, A. J., LUIDER, T. M. & BISCHOFF, R. 2011. The Impact of Delayed Storage on the Measured Proteome and Metabolome of Human Cerebrospinal Fluid. *Clinical Chemistry,* 57**,** 1703-1711.

ROUSU, T., HERTTUAINEN, J. & TOLONEN, A. 2010. Comparison of triple quadrupole, hybrid linear ion trap triple quadrupole, time-of-flight and LTQ-Orbitrap mass spectrometers in drug discovery phase metabolite screening and identification in vitro – amitriptyline and verapamil as model compounds. *Rapid Communications in Mass Spectrometry,* 24**,** 939-957.

RUTTER, P. M. 2012. Over-the-counter medicines: their place in self-care. *Br J Nurs,* 21**,** 806-10.

SABETI, P. C., VARILLY, P., FRY, B., LOHMUELLER, J., HOSTETTER, E., COTSAPAS, C., XIE, X., BYRNE, E. H., MCCARROLL, S. A., GAUDET, R., SCHAFFNER, S. F., LANDER, E. S., FRAZER, K. A., BALLINGER, D. G., COX, D. R., HINDS, D. A., STUVE, L. L., GIBBS, R. A., BELMONT, J. W., BOUDREAU, A., HARDENBOL, P., LEAL, S. M., PASTERNAK, S., WHEELER, D. A., WILLIS, T. D., YU, F., YANG, H., ZENG, C., GAO, Y., HU, H., HU, W., LI, C., LIN, W., LIU, S., PAN, H., TANG, X., WANG, J., WANG, W., YU, J., ZHANG, B., ZHANG, Q., ZHAO, H., ZHAO, H., ZHOU, J., GABRIEL, S. B., BARRY, R., BLUMENSTIEL, B., CAMARGO, A., DEFELICE, M., FAGGART, M., GOYETTE, M., GUPTA, S., MOORE, J., NGUYEN, H., ONOFRIO, R. C., PARKIN, M., ROY, J., STAHL, E., WINCHESTER, E., ZIAUGRA, L., ALTSHULER, D., SHEN, Y., YAO, Z., HUANG, W., CHU, X., HE, Y., JIN, L., LIU, Y., SHEN, Y., SUN, W., WANG, H., WANG, Y., WANG, Y., XIONG, X., XU, L., WAYE, M. M. Y., TSUI, S. K. W., XUE, H., TZE-FEI WONG, J., GALVER, L. M., FAN, J.-B., GUNDERSON, K., MURRAY, S. S., OLIPHANT, A. R., CHEE, M. S., MONTPETIT, A., CHAGNON, F., FERRETTI, V., LEBOEUF, M., OLIVIER, J.-F., PHILLIPS, M. S., ROUMY, S., SALLÉE, C., VERNER, A., HUDSON, T. J., KWOK, P.-Y., CAI, D., KOBOLDT, D. C., MILLER, R. D., PAWLIKOWSKA, L., et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature,* 449**,** 913-918.

SAMANI, S. M., MONTASERI, H. & BANANI, A. Formulation and evaluation of an alcohol free mouth-refreshing product. Proceedings World Automation Congress, 2004., 28 June-1 July 2004 2004. 9-14.

SANDBORGH-ENGLUND, G., ADOLFSSON-ERICI, M., ODHAM, G. & EKSTRAND, J. 2006. Pharmacokinetics of triclosan following oral ingestion in humans. *J Toxicol Environ Health A,* 69**,** 1861-73.

SANDS, C. J., WOLFER, A. M., CORREIA, G. D. S., SADAWI, N., AHMED, A., JIMÉNEZ, B., LEWIS, M. R., GLEN, R. C., NICHOLSON, J. K. & PEARCE, J. T. M. 2019. The nPYc-Toolbox, a Python module for the pre-processing, quality-control and analysis of metabolic profiling datasets. *Bioinformatics,* 35**,** 5359-5360.

SARAFIAN, M. H., GAUDIN, M., LEWIS, M. R., MARTIN, F.-P., HOLMES, E., NICHOLSON, J. K. & DUMAS, M.-E. 2014. Objective Set of Criteria for Optimization of Sample Preparation Procedures for Ultra-High Throughput Untargeted Blood Plasma Lipid Profiling by Ultra Performance Liquid Chromatography–Mass Spectrometry. *Analytical Chemistry,* 86**,** 5766-5774.

SARAFIAN, M. H., LEWIS, M. R., PECHLIVANIS, A., RALPHS, S., MCPHAIL, M. J., PATEL, V. C., DUMAS, M. E., HOLMES, E. & NICHOLSON, J. K. 2015. Bile acid profiling and quantification in biofluids using ultra-performance liquid chromatography tandem mass spectrometry. *Anal Chem,* 87**,** 9662-70.

SARAFIAN, M. H., LEWIS, M. R., PECHLIVANIS, A., RALPHS, S., MCPHAIL, M. J. W., PATEL, V. C., DUMAS, M.-E., HOLMES, E. & NICHOLSON, J. K. 2015. Bile Acid Profiling and Quantification in Biofluids Using Ultra-Performance Liquid Chromatography Tandem Mass Spectrometry. *Analytical Chemistry,* 87**,** 9662-9670.

SAUVAGE, F.-L. & MARQUET, P. 2012. LC-MS/MS Screen for Xenobiotics and Metabolites. *In:* LANGMAN, L. J. & SNOZEK, C. L. H. (eds.) *LC-MS in Drug Analysis: Methods and Protocols.* Totowa, NJ: Humana Press.

SCHELTEMA, R. A., JANKEVICS, A., JANSEN, R. C., SWERTZ, M. A. & BREITLING, R. 2011. PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis. *Analytical Chemistry,* 83**,** 2786-2793.

SCHIFFER, L., BARNARD, L., BARANOWSKI, E. S., GILLIGAN, L. C., TAYLOR, A. E., ARLT, W., SHACKLETON, C. H. L. & STORBECK, K.-H. 2019. Human steroid biosynthesis, metabolism and excretion are differentially reflected by serum and urine steroid metabolomes: A comprehensive review. *The Journal of steroid biochemistry and molecular biology,* 194**,** 105439-105439.

SCHNABEL, R. B., BAUMERT, J., BARBALIC, M., DUPUIS, J., ELLINOR, P. T., DURDA, P., DEHGHAN, A., BIS, J. C., ILLIG, T., MORRISON, A. C., JENNY, N. S., KEANEY, J. F., JR., GIEGER, C., TILLEY, C., YAMAMOTO, J. F., KHUSEYINOVA, N., HEISS, G., DOYLE, M., BLANKENBERG, S., HERDER, C., WALSTON, J. D., ZHU, Y., VASAN, R. S., KLOPP, N., BOERWINKLE, E., LARSON, M. G., PSATY, B. M., PETERS, A., BALLANTYNE, C. M., WITTEMAN, J. C., HOOGEVEEN, R. C., BENJAMIN, E. J., KOENIG, W. & TRACY, R. P. 2010. Duffy antigen receptor for chemokines (Darc) polymorphism regulates circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators. *Blood,* 115, 5289-99.

SCHWAIGER, M., SCHOENY, H., EL ABIEAD, Y., HERMANN, G., RAMPLER, E. & KOELLENSPERGER, G. 2018. Merging metabolomics and lipidomics into one analytical run. *Analyst,* 144, 220-229.

SCHWAIGER, M., SCHOENY, H., EL ABIEAD, Y., HERMANN, G., RAMPLER, E. & KOELLENSPERGER, G. 2019. Merging metabolomics and lipidomics into one analytical run. *Analyst,* 144, 220-229.

SENGUPTA, N., REARDON, D. C., GERARD, P. D. & BALDWIN, W. S. 2017. Exchange of polar lipids from adults to neonates in Daphnia magna: Perturbations in sphingomyelin allocation by dietary lipids and environmental toxicants. *PLOS ONE,* 12, e0178131.

SHARON, G., SAMPSON, T. R., GESCHWIND, D. H. & MAZMANIAN, S. K. 2016. The Central Nervous System and the Gut Microbiome. *Cell,* 167, 915-932.

SHIM, J., WEATHERLY, L. M., LUC, R. H., DORMAN, M. T., NEILSON, A., NG, R., KIM, C. H., MILLARD, P. J. & GOSSE, J. A. 2016. Triclosan is a mitochondrial uncoupler in live zebrafish. *J Appl Toxicol,* 36, 1662-1667.

SILVA, C. L., PASSOS, M. & CAMARA, J. S. 2011. Investigation of urinary volatile organic metabolites as potential cancer biomarkers by solid-phase microextraction in combination with gas chromatography-mass spectrometry. *Br J Cancer,* 105, 1894-904.

SILVA, C. L., PASSOS, M. & CMARA, J. S. 2011. Investigation of urinary volatile organic metabolites as potential cancer biomarkers by solid-phase microextraction in combination with gas chromatography-mass spectrometry. *British Journal of Cancer,* 105, 1894-1904.

SIMÓN-MANSO, Y., LOWENTHAL, M. S., KILPATRICK, L. E., SAMPSON, M. L., TELU, K. H., RUDNICK, P. A., MALLARD, W. G., BEARDEN, D. W., SCHOCK, T. B., TCHEKHOVSKOI, D. V., BLONDER, N., YAN, X., LIANG, Y., ZHENG, Y., WALLACE, W. E., NETA, P., PHINNEY, K. W., REMALEY, A. T. & STEIN, S. E. 2013. Metabolite Profiling of a NIST Standard Reference Material for Human Plasma (SRM 1950): GC-MS, LC-MS, NMR, and Clinical Laboratory Analyses, Libraries, and Web-Based Resources. *Analytical Chemistry,* 85, 11725-11731.

SINGER, H., MÜLLER, S., TIXIER, C. & PILLONEL, L. 2002. Triclosan: Occurrence and Fate of a Widely Used Biocide in the Aquatic Environment: Field Measurements in Wastewater Treatment Plants, Surface Waters, and Lake Sediments. *Environmental Science & Technology,* 36, 4998-5004.

SINGH, R., ARAIN, E., BUTH, A., KADO, J., SOUBANI, A. & IMRAN, N. 2016. Ethylene Glycol Poisoning: An Unusual Cause of Altered Mental Status and the Lessons Learned from Management of the Disease in the Acute Setting. *Case Reports in Critical Care,* 2016, 9157393.

SLOTTE, J. P. 2013. Biological functions of sphingomyelins. *Prog Lipid Res,* 52, 424-37.

SMITH, C. A., WANT, E. J., O'MAILLE, G., ABAGYAN, R. & SIUZDAK, G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem,* 78**,** 779-87.

SMITH, C. A., WANT, E. J., O'MAILLE, G., ABAGYAN, R. & SIUZDAK, G. 2006. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry,* 78**,** 779-787.

SMITH, P. K., KROHN, R. I., HERMANSON, G. T., MALLIA, A. K., GARTNER, F. H., PROVENZANO, M. D., FUJIMOTO, E. K., GOEKE, N. M., OLSON, B. J. & KLENK, D. C. 1985. Measurement of protein using bicinchoninic acid. *Anal Biochem,* 150**,** 76-85.

SOMANI, S. M. & KHALIQUE, A. 1982. Distribution and metabolism of 2,4-dichlorophenol in rats. *J Toxicol Environ Health,* 9**,** 889-97.

SONG, M., GAO, X., HANG, T. & WEN, A. 2008. Simultaneous determination of lansoprazole and its metabolites 5'-hydroxy lansoprazole and lansoprazole sulphone in human plasma by LC-MS/MS: application to a pharmacokinetic study in healthy volunteers. *J Pharm Biomed Anal,* 48**,** 1181-6.

SOUTHAM, A. D., HAGLINGTON, L. D., NAJDEKR, L., JANKEVICS, A., WEBER, R. J. M. & DUNN, W. B. 2020. Assessment of human plasma and urine sample preparation for reproducible and high-throughput UHPLC-MS clinical metabolic phenotyping. *Analyst,* 145**,** 6511-6523.

SPAGOU, K., TSOUKALI, H., RAIKOS, N., GIKA, H., WILSON, I. D. & THEODORIDIS, G. 2010. Hydrophilic interaction chromatography coupled to MS for metabonomic/metabolomic studies. *J Sep Sci,* 33**,** 716-27.

SPECTOR, S. R., MAYAN, H., LOEBSTEIN, R., MARKOVITS, N., PRIEL, E., MASSALHA, E., SHAFIR, Y. & GUETA, I. 2019. Pyroglutamic acidosis as a cause for high anion gap metabolic acidosis: a prospective study. *Scientific Reports,* 9**,** 3554.

STEIN, S. 2012. Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification. *Analytical Chemistry,* 84**,** 7274-7282.

STERNE, J. A. & DAVEY SMITH, G. 2001. Sifting the evidence-what's wrong with significance tests? *Bmj,* 322**,** 226-31.

STEUER, A. E., BROCKBALS, L. & KRAEMER, T. 2019. Metabolomic Strategies in Biomarker Research– New Approach for Indirect Identification of Drug Consumption and Sample Manipulation in Clinical and Forensic Toxicology? *Frontiers in Chemistry,* 7.

STOJANOVIC, V. & IHLE, S. 2011. Role of beta-hydroxybutyric acid in diabetic ketoacidosis: a review. *The Canadian veterinary journal = La revue veterinaire canadienne,* 52**,** 426-430.

STONE, C. A., JR., LIU, Y., RELLING, M. V., KRANTZ, M. S., PRATT, A. L., ABREO, A., HEMLER, J. A. & PHILLIPS, E. J. 2019. Immediate Hypersensitivity to Polyethylene Glycols and Polysorbates: More Common Than We Have Recognized. *The journal of allergy and clinical immunology. In practice,* 7**,** 1533-1540.e8.

STOREY, J. D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 64**,** 479-498.

STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. & WILLIAMS JR, R. M. 1949. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1,* Oxford, England, Princeton Univ. Press.

SU, C., LIU, Y., LI, R., WU, W., FAWCETT, J. P. & GU, J. 2019. Absorption, distribution, metabolism and excretion of the biomaterials used in Nanocarrier drug delivery systems. *Advanced Drug Delivery Reviews,* 143**,** 97-114.

SUAREZ-DIEZ, M., ADAM, J., ADAMSKI, J., CHASAPI, S. A., LUCHINAT, C., PETERS, A., PREHN, C., SANTUCCI, C., SPYRIDONIDIS, A., SPYROULIAS, G. A., TENORI, L., WANG-SATTLER, R. & SACCENTI, E. 2017. Plasma and Serum Metabolite Association Networks: Comparability within and between Studies Using NMR and MS Profiling. *Journal of proteome research,* 16**,** 2547-2559.

SUGIMOTO, H., TSUCHIYA, Y., SUGUMI, H., HIGURASHI, K., KARIBE, N., IIMURA, Y., SASAKI, A., KAWAKAMI, Y., NAKAMURA, T., ARAKI, S. & ET AL. 1990. Novel piperidine derivatives. Synthesis and anti-acetylcholinesterase activity of 1-benzyl-4-[2-(N-benzoylamino)ethyl]piperidine derivatives. *J Med Chem,* 33**,** 1880-7.

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. & VIANT, M. R. 2007. Proposed minimum reporting standards for chemical analysis. *Metabolomics,* 3**,** 211-221.

SZNAJDER-KATARZYŃSKA, K., SURMA, M. & CIEŚLIK, I. 2019. A Review of Perfluoroalkyl Acids (PFAAs) in terms of Sources, Applications, Human Exposure, Dietary Intake, Toxicity, Legal Regulation, and Methods of Determination. *Journal of Chemistry,* 2019**,** 2717528.

TAUTENHAHN, R., BÖTTCHER, C. & NEUMANN, S. 2008. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics,* 9**,** 504.

THARWAT, A. 2018. Classification assessment methods. *Applied Computing and Informatics*.

TIBSHIRANI, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 58**,** 267-288.

TONG, W., WELSH, W. J., SHI, L., FANG, H. & PERKINS, R. 2003. Structure-activity relationship approaches and applications. *Environ Toxicol Chem,* 22**,** 1680-95.

TONG, X. S., WANG, J., ZHENG, S., PIVNICHNY, J. V., GRIFFIN, P. R., SHEN, X., DONNELLY, M., VAKERICH, K., NUNES, C. & FENYK-MELODY, J. 2002. Effect of Signal Interference from Dosing Excipients on Pharmacokinetic Screening of Drug Candidates by Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry,* 74**,** 6305-6313.

TRIVEDI, D. K., HOLLYWOOD, K. A. & GOODACRE, R. 2017. Metabolomics for the masses: The future of metabolomics in a personalized world. *New Horizons in Translational Medicine,* 3**,** 294-305.

TRYGG, J. & WOLD, S. 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics,* 16**,** 119-128.

TSAKELIDOU, E., VIRGILIOU, C., VALIANOU, L., GIKA, H. G., RAIKOS, N. & THEODORIDIS, G. 2017. Sample Preparation Strategies for the Effective Quantitation of Hydrophilic Metabolites in Serum by Multi-Targeted HILIC-MS/MS. *Metabolites,* 7.

TSUCHIYA, Y., TAKAHASHI, Y., JINDO, T., FURUHAMA, K. & SUZUKI, K. T. 2003. Comprehensive evaluation of canine renal papillary necrosis induced by nefiracetam, a neurotransmission enhancer. *Eur J Pharmacol,* 475**,** 119-28.

TULIPANI, S., LLORACH, R., URPI-SARDA, M. & ANDRES-LACUEVA, C. 2013. Comparative Analysis of Sample Preparation Methods To Handle the Complexity of the Blood Fluid Metabolome: When Less Is More. *Analytical Chemistry,* 85**,** 341-348.

TULIPANI, S., MORA-CUBILLOS, X., JÁUREGUI, O., LLORACH, R., GARCÍA-FUENTES, E., TINAHONES, F. J. & ANDRES-LACUEVA, C. 2015. New and Vintage Solutions To Enhance the Plasma Metabolome Coverage by LC-ESI-MS Untargeted Metabolomics: The Not-So-Simple Process of Method Performance Evaluation. *Analytical Chemistry,* 87**,** 2639-2647.

TWEEDDALE, H., NOTLEY-MCROBB, L. & FERENCI, T. 1998. <Effect of Slow Growth on Metabolism of Escherichia coli, as revelaed by Global Metabolite Pool (Metabolome) analysis.pdf>. *Journal Of Bacteriology.* Australia: American Society for Microbiology.

TZOULAKI, I., EBBELS, T. M. D., VALDES, A., ELLIOTT, P. & IOANNIDIS, J. P. A. 2014. Design and Analysis of Metabolomics Studies in Epidemiologic Research: A Primer on -Omic Technologies. *American Journal of Epidemiology,* 180**,** 129-139.

UFER, M., JUIF, P.-E., BOOF, M.-L., MUEHLAN, C. & DINGEMANSE, J. 2017. Metabolite profiling in early clinical drug development: current status and future prospects. *Expert Opinion on Drug Metabolism & Toxicology,* 13**,** 803-806.

ULASZEWSKA, M. M., WEINERT, C. H., TRIMIGNO, A., PORTMANN, R., ANDRES LACUEVA, C., BADERTSCHER, R., BRENNAN, L., BRUNIUS, C., BUB, A., CAPOZZI, F., CIALIÈ ROSSO, M., CORDERO, C. E., DANIEL, H., DURAND, S., EGERT, B., FERRARIO, P. G., FESKENS, E. J. M., FRANCESCHI, P., GARCIA-ALOY, M., GIACOMONI, F., GIESBERTZ, P., GONZÁLEZ-DOMÍNGUEZ, R., HANHINEVA, K., HEMERYCK, L. Y., KOPKA, J., KULLING, S. E., LLORACH, R., MANACH, C., MATTIVI, F., MIGNÉ, C., MÜNGER, L. H., OTT, B., PICONE, G., PIMENTEL, G., PUJOS-GUILLOT, E., RICCADONNA, S., RIST, M. J., ROMBOUTS, C., RUBERT, J., SKURK, T., SRI HARSHA, P. S. C., VAN MEULEBROEK, L., VANHAECKE, L., VÁZQUEZ-FRESNO, R., WISHART, D. & VERGÈRES, G. 2019. Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. *Molecular Nutrition & Food Research,* 63**,** 1800384.

URAYAMA, S., ZOU, W., BROOKS, K. & TOLSTIKOV, V. 2010. Comprehensive mass spectrometry based metabolic profiling of blood plasma reveals potent discriminatory classifiers of pancreatic cancer. *Rapid Communications in Mass Spectrometry,* 24**,** 613-620.

URBAN, J., AFSETH, N. K. & ŠTYS, D. 2014. Fundamental definitions and confusions in mass spectrometry about mass assignment, centroiding and resolution. *TrAC Trends in Analytical Chemistry,* 53**,** 126-136.

VAITSIAKHOVICH, T., DRICHEL, D., HEROLD, C., LACOUR, A. & BECKER, T. 2014. METAINTER: meta-analysis of multiple regression models in genome-wide association studies. *Bioinformatics,* 31**,** 151-157.

VALDEZ, C. A., LEIF, R. N. & MAYER, B. P. 2014. An efficient, optimized synthesis of fentanyl and related analogs. *PloS one,* 9**,** e108250-e108250.

VAN DEN BERG, R. A., HOEFSLOOT, H. C. J., WESTERHUIS, J. A., SMILDE, A. K. & VAN DER WERF, M. J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics,* 7**,** 142.

VAN DEN BERG, R. A., HOEFSLOOT, H. C. J., WESTERHUIS, J. A., SMILDE, A. K. & VAN DER WERF, M. J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics,* 7**,** 142-142.

VAN DER HOOFT, J. J. J., PADMANABHAN, S., BURGESS, K. E. V. & BARRETT, M. P. 2016. Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. *Metabolomics : Official journal of the Metabolomic Society,* 12**,** 125-125.

VARACALLO, R. N. M. N. G. N. S. S. M. 2020. Physiology, Albumin.

VATCHEVA, K. P., LEE, M., MCCORMICK, J. B. & RAHBAR, M. H. 2016. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale, Calif.),* 6**,** 227.

VERHO, M., LUCK, C., STELTER, W. J., RANGOONWALA, B. & BENDER, N. 1995. Pharmacokinetics, metabolism and biliary and urinary excretion of oral ramipril in man. *Curr Med Res Opin,* 13**,** 264-73.

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., PELCZER, I., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2011. Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. *Analytical Chemistry,* 83**,** 5864-5872.

VIJAYA BHASKAR, V., MIDDHA, A., TIWARI, S. & SHIVAKUMAR, S. 2013. Liquid chromatography/tandem mass spectrometry method for quantitative estimation of polyethylene glycol 400 and its applications. *J Chromatogr B Analyt Technol Biomed Life Sci,* 926**,** 68-76.

VISHWAKARMA, D. N. & PATEL, D. 2010. Drug Discovery. *Journal of Antivirals & Antiretrovirals,* 2.

WAGNER-GOLBS, A., NEUBER, S., KAMLAGE, B., CHRISTIANSEN, N., BETHAN, B., RENNEFAHRT, U., SCHATZ, P. & LIND, L. 2019. Effects of Long-Term Storage at -80 degrees C on the Human Plasma Metabolome. *Metabolites,* 9.

WAGNER, C., EL OMARI, M. & KÖNIG, G. M. 2009. Biohalogenation: nature's way to synthesize halogenated metabolites. *J Nat Prod,* 72**,** 540-53.

WALTER, R. E., CRAMER, J. A. & TSE, F. L. 2001. Comparison of manual protein precipitation (PPT) versus a new small volume PPT 96-well filter plate to decrease sample preparation time. *J Pharm Biomed Anal,* 25**,** 331-7.

WANG, Y. & KASPER, L. H. 2014. The role of microbiome in central nervous system disorders. *Brain Behav Immun,* 38**,** 1-12.

WANT, E. J., O'MAILLE, G., SMITH, C. A., BRANDON, T. R., URITBOONTHAI, W., QIN, C., TRAUGER, S. A. & SIUZDAK, G. 2006. Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Anal Chem,* 78**,** 743-52.

WANT, E. J., SMITH, C. A., QIN, C., VAN HORNE, K. C. & SIUZDAK, G. J. M. 2006. Phospholipid capture combined with non-linear chromatographic correction for improved serum metabolite profiling. 2**,** 145-154.

WANT, E. J., WILSON, I. D., GIKA, H., THEODORIDIS, G., PLUMB, R. S., SHOCKCOR, J., HOLMES, E. & NICHOLSON, J. K. 2010. Global metabolic profiling procedures for urine using UPLC-MS. *Nat. Protocols,* 5**,** 1005-1018.

WARDELL, R. J., BURROWS, L. A., MYALL, K. & MARSH, A. 2012. An unusual cause of high anion gap metabolic acidosis: pyroglutamic acidosis. *Critical Care,* 16**,** P148.

WARRACK, B. M., HNATYSHYN, S., OTT, K.-H., REILY, M. D., SANDERS, M., ZHANG, H. & DREXLER, D. M. 2009. Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B,* 877**,** 547-552.

WATERS 2003. Controlling Contamination in Ultra Performance LC/MS and HPLC/MS Systems. 715001307, RevD. 2003:21. http://www.waters.com/. Accessed October 11, 2017.

WATERS, N. J., HOLMES, E., WILLIAMS, A., WATERFIELD, C. J., FARRANT, R. D. & NICHOLSON, J. K. 2001. NMR and pattern recognition studies on the time-related metabolic effects of alpha-naphthylisothiocyanate on liver, urine, and plasma in the rat: an integrative metabonomic approach. *Chem Res Toxicol,* 14**,** 1401-12.

WATROUS, J. D., HENGLIN, M., CLAGGETT, B., LEHMANN, K. A., LARSON, M. G., CHENG, S. & JAIN, M. 2017. Visualization, Quantification, and Alignment of Spectral Drift in Population Scale Untargeted Metabolomics Data. *Analytical chemistry,* 89**,** 1399-1404.

WATTS, M. T. & MCDONALD, O. L. 1990. The effect of sodium chloride concentration, water content, and protein on the gas chromatographic headspace analysis of ethanol in plasma. *Am J Clin Pathol,* 93**,** 357-62.

WEAVER, R. & RILEY, R. J. 2006. Identification and reduction of ion suppression effects on pharmacokinetic parameters by polyethylene glycol 400. *Rapid Communications in Mass Spectrometry,* 20**,** 2559-2564.

WEBSTER, R., DIDIER, E., HARRIS, P., SIEGEL, N., STADLER, J., TILBURY, L. & SMITH, D. 2007. PEGylated Proteins: Evaluation of Their Safety in the Absence of Definitive Metabolism Studies. *Drug metabolism and disposition: the biological fate of chemicals,* 35**,** 9-16.

WELLS, P. G., MACKENZIE, P. I., CHOWDHURY, J. R., GUILLEMETTE, C., GREGORY, P. A., ISHII, Y., HANSEN, A. J., KESSLER, F. K., KIM, P. M., CHOWDHURY, N. R. & RITTER, J. K. 2004. Glucuronidation and the UDP-glucuronosyltransferases in health and disease. *Drug Metab Dispos,* 32**,** 281-90.

WESTERHUIS, J. A., HOEFSLOOT, H. C. J., SMIT, S., VIS, D. J., SMILDE, A. K., VAN VELZEN, E. J. J., VAN DUIJNHOVEN, J. P. M. & VAN DORSTEN, F. A. 2008. Assessment of PLSDA cross validation. *Metabolomics,* 4**,** 81-89.

WHILEY, L., CHEKMENEVA, E., BERRY, D. J., JIMÉNEZ, B., YUEN, A. H. Y., SALAM, A., HUSSAIN, H., WITT, M., TAKATS, Z., NICHOLSON, J. & LEWIS, M. R. 2019. Systematic Isolation and Structure Elucidation of Urinary Metabolites Optimized for the Analytical-Scale Molecular Profiling Laboratory. *Analytical Chemistry,* 91**,** 8873-8882.

WHITLOCK, M. 2005. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of evolutionary biology,* 18**,** 1368-73.

WIKLUND, S., JOHANSSON, E., SJÖSTRÖM, L., MELLEROWICZ, E. J., EDLUND, U., SHOCKCOR, J. P., GOTTFRIES, J., MORITZ, T. & TRYGG, J. 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal Chem,* 80**,** 115-22.

WILD, C. P. 2005. Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiology Biomarkers &amp; Prevention,* 14**,** 1847-1850.

WILLKOMMEN, D., LUCIO, M., MORITZ, F., FORCISI, S., KANAWATI, B., SMIRNOV, K. S., SCHROETER, M., SIGAROUDI, A., SCHMITT-KOPPLIN, P. & MICHALKE, B. 2018. Metabolomic investigations in cerebrospinal fluid of Parkinson's disease. *PloS one,* 13**,** e0208752-e0208752.

WILSON, I. D., NICHOLSON, J. K., CASTRO-PEREZ, J., GRANGER, J. H., JOHNSON, K. A., SMITH, B. W. & PLUMB, R. S. 2005. High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *J Proteome Res,* 4**,** 591-8.

WILSON, I. D., PLUMB, R., GRANGER, J., MAJOR, H., WILLIAMS, R. & LENZ, E. M. 2005. HPLC-MS-based methods for the study of metabonomics. *Journal of Chromatography B,* 817**,** 67-76.

WISHART, D. S. 2011. Advances in metabolite identification. *Bioanalysis,* 3**,** 1769-82.

WISHART, D. S. 2016. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery,* 15**,** 473-484.

WISHART, D. S., FEUNANG, Y. D., MARCU, A., GUO, A. C., LIANG, K., VÁZQUEZ-FRESNO, R., SAJED, T., JOHNSON, D., LI, C., KARU, N., SAYEEDA, Z., LO, E., ASSEMPOUR, N., BERJANSKII, M., SINGHAL, S., ARNDT, D., LIANG, Y., BADRAN, H., GRANT, J., SERRA-CAYUELA, A., LIU, Y., MANDAL, R., NEVEU, V., PON, A., KNOX, C., WILSON, M., MANACH, C. & SCALBERT, A. 2017. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research,* 46**,** D608-D617.

WISHART, D. S., JEWISON, T., GUO, A. C., WILSON, M., KNOX, C., LIU, Y., DJOUMBOU, Y., MANDAL, R., AZIAT, F., DONG, E., BOUATRA, S., SINELNIKOV, I., ARNDT, D., XIA, J., LIU, P., YALLOU, F., BJORNDAHL, T., PEREZ-PINEIRO, R., EISNER, R., ALLEN, F., NEVEU, V., GREINER, R. & SCALBERT, A. 2012. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research,* 41**,** D801-D807.

WISHART, D. S., JEWISON, T., GUO, A. C., WILSON, M., KNOX, C., LIU, Y., DJOUMBOU, Y., MANDAL, R., AZIAT, F., DONG, E., BOUATRA, S., SINELNIKOV, I., ARNDT, D., XIA, J., LIU, P., YALLOU, F., BJORNDAHL, T., PEREZ-PINEIRO, R., EISNER, R., ALLEN, F., NEVEU, V., GREINER, R. &

SCALBERT, A. 2013. HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res,* 41**,** D801-7.

WOLD, S., ANTTI, H., LINDGREN, F. & ÖHMAN, J. 1998. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems,* 44**,** 175-185.

WOLD, S., SJÖSTRÖM, M. & ERIKSSON, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 58**,** 109-130.

WOLFENDER, J.-L., NUZILLARD, J.-M., VAN DER HOOFT, J. J. J., RENAULT, J.-H. & BERTRAND, S. 2019. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography–High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics. *Analytical Chemistry,* 91**,** 704-742.

WOLFF, M. M. & STEPHENS, W. E. 1953. A Pulsed Mass Spectrometer with Time Dispersion. *Review of Scientific Instruments,* 24**,** 616.

WONG, M. C., LEE, W. T., WONG, J. S., FROST, G. & LODGE, J. 2008. An approach towards method development for untargeted urinary metabolite profiling in metabonomic research using UPLC/QToF MS. *J Chromatogr B Analyt Technol Biomed Life Sci,* 871**,** 341-8.

WOOD, M. 2004. Statistical inference using bootstrap confidence intervals. *Significance,* 1**,** 180-182.

WORLEY, B. & POWERS, R. 2013. Multivariate Analysis in Metabolomics. *Current Metabolomics,* 1**,** 92-107.

WU, H., CHEN, Y., LI, Z. & LIU, X. 2018. Untargeted metabolomics profiles delineate metabolic alterations in mouse plasma during lung carcinoma development using UPLC-QTOF/MS in MS(E) mode. *Royal Society open science,* 5**,** 181143-181143.

WU, H., SOUTHAM, A. D., HINES, A. & VIANT, M. R. 2008. High-throughput tissue extraction protocol for NMR- and MS-based metabolomics. *Anal Biochem,* 372**,** 204-12.

WU, J.-L., LIU, J. & CAI, Z. 2010. Determination of triclosan metabolites by using in-source fragmentation from high-performance liquid chromatography/negative atmospheric pressure chemical ionization ion trap mass spectrometry. *Rapid Communications in Mass Spectrometry,* 24**,** 1828-1834.

XU, R. & WUNSCH, D. C. 2010. Clustering Algorithms in Biomedical Research: A Review. *IEEE Reviews in Biomedical Engineering,* 3**,** 120-154.

XU, Y., MUHAMADALI, H., SAYQAL, A., DIXON, N. & GOODACRE, R. 2016. Partial Least Squares with Structured Output for Modelling the Metabolomics Data Obtained from Complex Experimental Designs: A Study into the Y-Block Coding. *Metabolites,* 6.

YAMAMOTO, M., PINTO-SANCHEZ, M. I., BERCIK, P. & BRITZ-MCKIBBIN, P. 2019. Metabolomics reveals elevated urinary excretion of collagen degradation and epithelial cell turnover products in irritable bowel syndrome patients. *Metabolomics,* 15**,** 82.

YANES, O., TAUTENHAHN, R., PATTI, G. J. & SIUZDAK, G. 2011. Expanding Coverage of the Metabolome for Global Metabolite Profiling. *Analytical Chemistry,* 83**,** 2152-2161.

YANG, W., MU, T., JIANG, J., SUN, Q., HOU, X., SUN, Y., ZHONG, L., WANG, C. & SUN, C. 2018. Identification of Potential Biomarkers and Metabolic Profiling of Serum in Ovarian Cancer Patients Using UPLC/Q-TOF MS. *Cellular Physiology and Biochemistry,* 51**,** 1134-1148.

YANG, Y., CRUICKSHANK, C., ARMSTRONG, M., MAHAFFEY, S., REISDORPH, R. & REISDORPH, N. 2013. New sample preparation approach for mass spectrometry-based profiling of plasma results in improved coverage of metabolome. *Journal of chromatography. A,* 1300**,** 217-226.

YATOMI, Y., IGARASHI, Y., YANG, L., HISANO, N., QI, R., ASAZUMA, N., SATOH, K., OZAKI, Y. & KUME, S. 1997. Sphingosine 1-phosphate, a bioactive sphingolipid abundantly stored in platelets, is a normal constituent of human plasma and serum. *J Biochem,* 121**,** 969-73.

YE, X., BISHOP, A. M., REIDY, J. A., NEEDHAM, L. L. & CALAFAT, A. M. 2007. Temporal stability of the conjugated species of bisphenol A, parabens, and other environmental phenols in human urine. *Journal of Exposure Science & Environmental Epidemiology,* 17**,** 567-572.

YU, Z., KASTENMÜLLER, G., HE, Y., BELCREDI, P., MÖLLER, G., PREHN, C., MENDES, J., WAHL, S., ROEMISCH-MARGL, W., CEGLAREK, U., POLONIKOV, A., DAHMEN, N., PROKISCH, H., XIE, L., LI, Y., WICHMANN, H. E., PETERS, A., KRONENBERG, F., SUHRE, K., ADAMSKI, J., ILLIG, T. & WANG-SATTLER, R. 2011. Differences between Human Plasma and Serum Metabolite Profiles. *PLOS ONE,* 6**,** e21230.

YU, Z., KASTENMÜLLER, G., HE, Y., BELCREDI, P., MÖLLER, G., PREHN, C., MENDES, J., WAHL, S., ROEMISCH-MARGL, W., CEGLAREK, U., POLONIKOV, A., DAHMEN, N., PROKISCH, H., XIE, L., LI, Y., WICHMANN, H. E., PETERS, A., KRONENBERG, F., SUHRE, K., ADAMSKI, J., ILLIG, T. & WANG-SATTLER, R. 2011. Differences between human plasma and serum metabolite profiles. *PloS one,* 6**,** e21230-e21230.

YUN, J. H., LEE, H.-S., YU, H.-Y., KIM, Y.-J., JEON, H. J., OH, T., KIM, B.-J., CHOI, H. J. & KIM, J.-M. 2019. Metabolomics profiles associated with HbA1c levels in patients with type 2 diabetes. *PLOS ONE,* 14**,** e0224274.

ZANI, C. L. & CARROLL, A. R. 2017. Database for Rapid Dereplication of Known Natural Products Using Data from MS and Fast NMR Experiments. *Journal of Natural Products,* 80**,** 1758-1766.

ZAYKIN, D. V. 2011. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J Evol Biol,* 24**,** 1836-41.

ZELENA, E., DUNN, W. B., BROADHURST, D., FRANCIS-MCINTYRE, S., CARROLL, K. M., BEGLEY, P., O'HAGAN, S., KNOWLES, J. D., HALSALL, A., WILSON, I. D. & KELL, D. B. 2009. Development of a Robust and Repeatable UPLC−MS Method for the Long-Term Metabolomic Study of Human Serum. *Analytical Chemistry,* 81**,** 1357-1364.

ZENG, Z., SONG, B., XIAO, R., ZENG, G., GONG, J., CHEN, M., XU, P., ZHANG, P., SHEN, M. & YI, H. 2019. Assessing the human health risks of perfluorooctane sulfonate by in vivo and in vitro studies. *Environment International,* 126**,** 598-610.

ZHANG, A., SUN, H., XU, H., QIU, S. & WANG, X. 2013. Cell metabolomics. *Omics : a journal of integrative biology,* 17**,** 495-501.

ZHANG, A., SUN, H., YAN, G., WANG, P. & WANG, X. 2015. Metabolomics for Biomarker Discovery: Moving to the Clinic. *BioMed Research International,* 2015**,** 354671.

ZHANG, H., ZHANG, D., RAY, K. & ZHU, M. 2009. Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. *Journal of Mass Spectrometry,* 44**,** 999-1016.

ZHOU, J., LI, Y., CHEN, X., ZHONG, L. & YIN, Y. 2017. Development of data-independent acquisition workflows for metabolomic analysis on a quadrupole-Orbitrap platform. *Talanta,* 164**,** 128-136.

ZHU, M., MA, L., ZHANG, D., RAY, K., ZHAO, W., HUMPHREYS, W. G., SKILES, G., SANDERS, M. & ZHANG, H. 2006. Detection and characterization of metabolites in biological matrices using mass defect filtering of liquid chromatography/high resolution mass spectrometry data. *Drug Metab Dispos,* 34**,** 1722-33.

ZHU, Y., WANG, F., LI, Q., ZHU, M., DU, A., TANG, W. & CHEN, W. 2014. Amlodipine metabolism in human liver microsomes and roles of CYP3A4/5 in the dihydropyridine dehydrogenation. *Drug Metab Dispos,* 42**,** 245-9.

ZIKUAN, S., HAOYU, W., XIAOTONG, Y., PENGCHI, D. & WEI, J. 2019. Application of NMR metabolomics to search for human disease biomarkers in blood. *Clinical Chemistry and Laboratory Medicine (CCLM),* 57**,** 417-441.

ZOU, H. & HASTIE, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67**,** 301-320.

Appendix 1 Chapter 3

Reference standards acquired by RPC-UPLC-MS (positive and negative ion mode) for the database (Chapter 3)

All standards purchased by Sigma Aldrich (Vienna, Austria).

1	Amoxicillin
2	Bisoprolol
3	Caffeine
4	Citalopram hydrobromide
5	Codeine-6-beta-D-glucuronide solution C-II
6	(-)-Cotinine
7	Escitalopram oxalate
8	Lansoprazole
9	Metformin hydrochloride
10	Norcodeine solution C-II
11	Omeprazole
12	Ramipril
13	Simvastatin
14	Acetaminophen
15	6-Acetylcodeine.HCL
16	Acetylsalicylic acid
17	Amitriptyline hydrochloride
18	Amlodypine besylate
19	Atorvastatin calcium salt trihydrate
20	Bendroflumethiazide

21	Carboxyibuprofen
22	Cholecalciferol
23	Codeine.HCL
24	Dihydrocodeine.HCL
25	(+)-cis-Diltiazem hydrochloride
26	Furosemide
27	2-Hydroxyibuprofen
28	Ibuprofen
29	Levothyroxine
30	Paracetamol sulfate potassium salt
31	Paracetamol β-D-glucuronide
32	Salicylic acid
33	Terbinafine hydrochloride
34	Theobromine
35	Theophylline
36	Warfarin
37	Naproxen
38	N-Acetylbenzoquinoneimine
39	Nicotine
40	Salbutamol
41	Cotinine

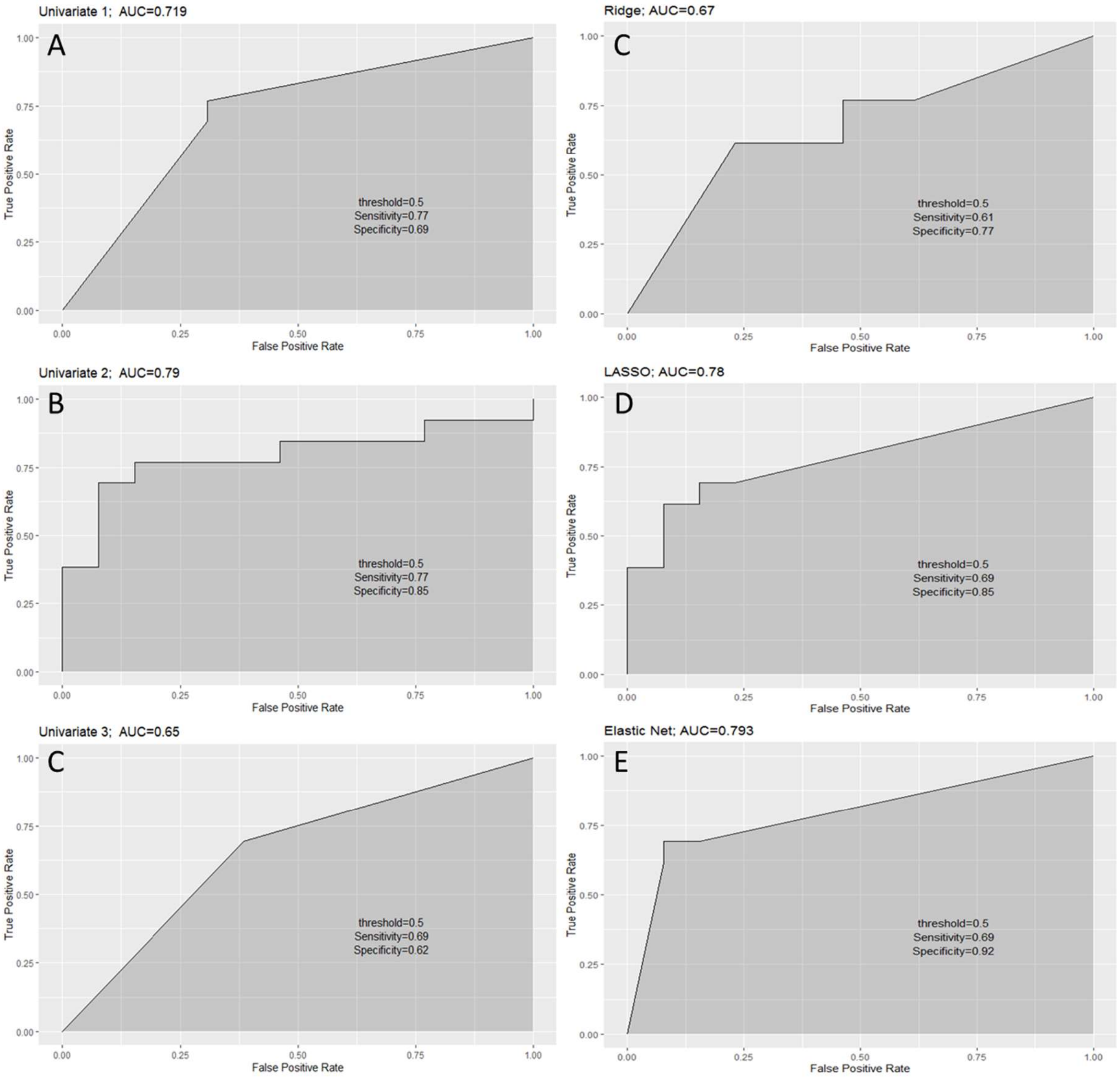
Pharmaceutical medications acquired by RPC-UPLC-MS (positive and negative ion mode) for the database (Chapter 3)

All medications were provided by Mr James Kinross

1	Paracetamol
2	Codeine
3	Morphine
4	Duhydrocodeine
5	Gabapentin
6	Pregabalin
7	Carbamazepine
8	Lamotrgine
9	Enoxaparin
10	Zopiclone
11	Citalopram
12	Amitriptyline
13	Sertraline
14	Fluoxetine
15	Mirtazapine
16	Venlafaxine
17	Paroxetne
18	Duloxetine
19	Simvastatin
20	Atorvastatin
21	Pravastatin
22	Ezetimibe
23	Quetiapine
24	Olanzapine
25	Risperidone
26	Naproxen
27	Allopurinol
28	Ibuprofen
29	Diclofenac

30	Omeprazole
31	Lansoprazole
32	Ranitidine
33	Esomeprazole
34	Trimethoprim
35	Metronidazole
36	Metformin
37	Gliclazide
38	Nitrazepam
39	Diazepam
40	Ramipril
41	Lisinopril
42	Losartan
43	Doxazosin
44	Candesartan
45	Perindopril
46	Amlodipine
47	Isosorbide
48	Felodipine
49	Diltiazem
50	Warfarin
51	Bendroflumethiazide
52	Furosemide
53	Indapamide
54	Bisoprolol
55	Atenolol
56	Propranolol
57	Metoprolol

ROC curves produced from the logistic regression univariate and multivariate models using the test set samples



ROC curves produced from the logistic regression univariate and multivariate models using the test set samples.

(A) **univariate Model 1; Sensitivity = 0.77, Specificity = 0.69, Accuracy (AUC) = 0.73;**

(B) **univariate Model 2; Sensitivity = 0.77, Specificity = 0.85, Accuracy (AUC) = 0.81;**

(C) **univariate Model 3; Sensitivity = 0.69, Specificity = 0.62, Accuracy (AUC) = 0.65;**

(D) **Ridge; Sensitivity = 0.61, Specificity = 0.77, Accuracy (AUC) = 0.67;**

(E) **LASSO; Sensitivity = 0.69, Specificity = 0.85, Accuracy (AUC) = 0.78;**

(F) **Elastic Net; Sensitivity = 0.69, Specificity = 0.92, Accuracy (AUC) = 0.79.**

Appendix 2: Chapter 4

Mass of 96-well plate prior to sorbent addition, post addition, and the difference (Mass of plate post sorbent addition – mass of plate) - Chapter 4

Table X Mass of plate prior to sorbent additon

BEFORE												
(mg)	1	2	3	4	5	6	7	8	9	10	11	12
A	609.91	607.82	610.1	607.89	610.41	608.41	610.09	608.09	610.48	607.79	610.33	608.33
B	604.83	610.31	609.67	610	609.7	610.58	609.56	610.63	609.7	609.12	609.38	608.68
C	605.92	612.16	605.8	612.49	605.77	612.5	605.77	612.61	605.84	612.52	605.71	612.57
D	619.1	603.9	619.07	604	619.04	603.89	617.5	603.91	618.75	603.87	618.69	603.93
E	609.95	611.64	610.56	611.8	610.56	611.8	610.52	611.54	610.52	611.83	610.51	611.82
F	605.25	612.07	605.61	612.19	605.67	612.03	605.52	612.16	605	612.04	605.52	612.1
G	609.9	605.28	609.98	605.22	609.88	605.18	609.85	605.22	609.84	605.16	609.84	605.2
H	613.22	607.6	613.45	607.55	613.15	607.45	613.33	609.64	613.14	605.57	613.17	607.46

Table X Mass of plate post sorbent addition

AFTER												
(mg)	1	2	3	4	5	6	7	8	9	10	11	12
A	618.51	616.3	617.83	616.48	618.83	617.13	618.61	615.94	618.48	615.99	618.33	617.09
B	614	619.93	617.62	619.12	618.95	620.14	618.83	618.74	618.82	617.74	618.84	619.01

C	616.23	622.35	614.34	622.23	615.12	622.67	616.13	621.52	615.24	622.42	615.41	622.99
D	628.97	613.07	627.43	613.86	628.08	613.29	626.57	611.99	627.71	613.54	627.93	612.95
E	619.89	620.39	618.77	621.07	619.6	621.64	619.91	619.33	619.05	621.15	619.4	621.04
F	614.56	620.97	613.54	620.9	613.97	621.22	614.33	620.08	613.42	621.24	613.68	620.92
G	617.84	613.78	617.57	613.69	618.04	613.18	618.15	613.09	617.81	614	618.3	613.27
H	621.32	615.75	621.23	615.61	621.19	615.49	621.21	617.44	621.11	615.77	621.72	616

Table X Mass of sorbent (Mass of plate post sorbent addition – mass of plate)

DIFFERENCE (mg)	1	2	3	4	5	6	7	8	9	10	11	12
A	8.6	8.48	7.73	8.59	8.42	8.72	8.52	7.85	8	8.2	8	8.76
B	9.17	9.62	7.95	9.12	9.25	9.56	9.27	8.11	9.12	8.62	9.46	10.33
C	10.31	10.19	8.54	9.74	9.35	10.17	10.36	8.91	9.4	9.9	9.7	10.42
D	9.87	9.17	8.36	9.86	9.04	9.4	9.07	8.08	8.96	9.67	9.24	9.02
E	9.94	8.75	8.21	9.27	9.04	9.84	9.39	7.79	8.53	9.32	8.89	9.22
F	9.31	8.9	7.93	8.71	8.3	9.19	8.81	7.92	8.42	9.2	8.16	8.82
G	7.94	8.5	7.59	8.47	8.16	8	8.3	7.87	7.97	8.84	8.46	8.07
H	8.1	8.15	7.78	8.06	8.04	8.04	7.88	7.8	7.97	10.2	8.55	8.54

List of identified metabolites (RPC positive mode peakpantheR) and relative method induced losses (%)

								Average intensity						Relative percentage					
Compound	**HMDBClass**	**HMDBSubClass**	**Retention time (min)**	**cpdMonoisot**	**ion**	**m/z**	**Pool-mean**	**Folch-mean**	**BD-mean**	**Matyash-mean**	**DSPE-mean**	**MeOH-mean**	**Max intensity**	**Pool**	**Folch**	**BD**	**Matyash**	**DSPE**	**MeOH**
Histidine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.53	155.0695	M+H	156.0768	4.84	4.88	4.65	4.88	4.58	5.05	5.05	95.84	96.68	92.15	96.62	90.84	100.00
Acetaminophen glucuronide	Organooxygen compounds	Carbohydrates and carbohydrate conjugates	1.95	327.0954	M+Na	350.0846	4.05	4.01	4.10	4.05	3.96	4.06	4.10	98.72	97.64	100.00	98.79	96.55	98.83
Symmetric dimethylarginine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.64	171.1002	M+H	172.1080	3.97	4.02	3.86	3.97	3.97	3.96	4.02	98.84	100.00	96.21	98.90	98.77	98.50
4-Guanidinobutanoate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.87	145.0851	M+H	146.0924	4.75	4.81	4.87	4.71	4.47	4.72	4.87	97.55	98.83	100.00	96.75	91.95	97.11
Riboflavin (vit B2)	Pteridines and derivatives	Alloxazines and isoalloxazines	3.86	376.1383	M+H	377.1456	3.29	2.83	2.82	2.96	3.74	3.14	3.74	87.86	75.61	75.37	79.23	100.00	84.00
Citrate	Carboxylic acids and derivatives	Tricarboxylic acids and derivatives	1.10	192.0270	M+Na	215.0162	5.85	5.49	5.96	5.78	5.79	5.86	5.96	98.13	92.02	100.00	96.82	97.13	98.19
Phenylacetylglutamine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	3.62	264.1110	M+H	265.1183	5.20	5.19	5.19	5.21	5.14	5.22	5.22	99.59	99.36	99.54	99.80	98.49	100.00
Guanidinosuccinate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.59	175.0593	M+H	176.0666	3.95	3.93	3.82	3.92	3.89	4.08	4.08	96.74	96.33	93.55	96.08	95.42	100.00
Paracetamol sulfate potassium salt	Organooxygen compounds	Carbohydrates and carbohydrate conjugates	2.20	231.0201	M+H	232.0274	3.93	3.91	3.96	3.92	3.88	3.92	3.96	99.12	98.59	100.00	98.84	98.01	99.06
N2,N2-Dimethylguanosine	Purine nucleosides		2.39	311.1230	M+H	312.1302	4.07	4.05	4.10	4.05	4.07	4.03	4.10	99.30	98.87	100.00	98.79	99.34	98.29
Imidazolelactate	Azoles	Imidazoles	0.61	156.0535	M+H	157.0608	4.69	4.71	4.71	4.69	4.64	4.62	4.71	99.51	100.00	99.92	99.48	98.48	98.00
S-Adenosylhomocysteine	5'-deoxyribonucleosides	5'-deoxy-5'-thionucleosides	1.40	384.1216	M+H	385.1289	3.73	3.69	3.78	3.74	3.74	3.63	3.78	98.70	97.66	100.00	99.00	98.96	95.99
1-Methyladenosine	Purine nucleosides		1.26	281.1124	M+H	282.1197	5.06	4.95	5.03	4.99	5.21	5.01	5.21	97.13	94.95	96.58	95.71	100.00	96.16
N-a-Acetyl-L-arginine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.92	216.1222	M+H	217.1295	4.51	4.48	4.54	4.49	4.45	4.55	4.55	99.02	98.30	99.65	98.57	97.74	100.00
Indole-3-acetate	Indoles and derivatives	Indolyl carboxylic acids and derivatives	5.59	175.0633	M+H	176.0706	4.80	4.87	4.91	4.92	4.75	4.09	4.92	97.67	99.00	99.86	100.00	96.59	83.08
Homocitrulline	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.65	189.1113	M+H	190.1186	4.31	4.28	4.33	4.29	4.31	4.30	4.33	99.55	98.88	100.00	99.20	99.72	99.34
7-Methylguanine	Imidazopyrimidines	Purines and purine derivatives	1.26	165.0651	M+H	166.0723	4.79	4.74	4.80	4.77	4.82	4.79	4.82	99.41	98.35	99.57	98.98	100.00	99.32
Prednisolone	Steroids and steroid derivatives	Hydroxysteroids	6.89	360.1937	M+H	361.2010	4.17	2.95	4.17	3.52	4.50	4.31	4.50	92.66	65.56	92.65	78.23	100.00	95.59
Acetaminophen	Phenols	1-hydroxy-2-unsubstituted benzenoids	2.58	151.0633	M+H	152.0706	4.25	4.26	4.34	3.71	4.35	4.26	4.35	97.62	97.93	99.70	85.25	100.00	97.99
1,1-Dimethylbiguanide (Metformin)	Organonitrogen compounds	Guanidines	0.67	129.1014	M+H	130.1087	5.69	5.70	5.64	5.71	5.68	5.73	5.73	99.41	99.57	98.41	99.69	99.25	100.00
Cortisone	Steroids and steroid derivatives	Hydroxysteroids	6.96	360.1937	M+H	361.2010	4.17	2.99	4.17	3.52	4.50	4.31	4.50	92.68	66.43	92.63	78.22	100.00	95.59
Pantothenate	Alcohols and polyols	Polyols	2.35	219.1107	M+H	220.1179	4.22	4.20	4.25	4.23	4.20	4.17	4.25	99.30	98.91	100.00	99.64	98.90	98.16
Disaccharides	Organooxygen compounds	Carbohydrates and carbohydrate conjugates	0.68	342.1162	M+Na	365.1054	5.20	5.26	5.25	5.26	5.07	5.09	5.26	98.94	100.00	99.91	99.97	96.54	96.79
Caffeine	Imidazopyrimidines	Purines and purine derivatives	3.48	194.0804	M+H	195.0877	5.96	4.41	5.93	5.91	6.16	6.10	6.16	96.84	71.62	96.33	95.96	100.00	99.17
Aminoadipate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.63	161.0688	M+H	162.0761	4.13	4.12	4.11	4.11	4.12	4.15	4.15	99.36	99.27	98.91	99.08	99.31	100.00
1-Methylurate	Imidazopyrimidines	Purines and purine derivatives	1.83	182.0440	M+H	183.0513	4.56	4.51	4.57	4.54	4.60	4.55	4.60	99.13	98.03	99.29	98.58	100.00	98.91
a-glycerophosphocholine	Glycerophospholipids	Glycerophosphocholines	0.56	257.1028	M+H	258.1101	6.21	6.02	6.27	6.17	6.01	6.42	6.42	96.69	93.78	97.56	95.98	93.51	100.00
(-)-Cotinine	Pyridines and derivatives	Pyrrolidinylpyridines	1.43	176.0950	M+H	177.1022	5.48	4.19	5.44	5.54	5.65	5.60	5.65	97.11	74.26	96.28	98.03	100.00	99.24
Kynurenine	Organooxygen compounds	Carbonyl compounds	2.02	208.0848	M+H	209.0921	5.25	5.22	5.13	5.17	5.51	5.01	5.51	95.32	94.71	93.14	93.73	100.00	90.97
Propionylcarnitine	Fatty Acyls	Fatty acid esters	1.64	217.1314	M+H	218.1387	5.50	5.52	5.50	5.55	5.54	5.30	5.55	99.04	99.35	99.03	100.00	99.82	95.44
Urate	Imidazopyrimidines	Purines and purine derivatives	1.13	168.0283	M+H	169.0356	6.81	6.87	6.72	6.84	6.78	6.87	6.87	99.21	99.99	97.82	99.60	98.76	100.00
Prolylhydroxyproline	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.77	228.1110	M+H	229.1183	5.21	5.16	5.19	5.17	5.31	5.16	5.31	98.15	97.24	97.72	97.38	100.00	97.24
Tetrahydropentoxyline	Harmala alkaloids		2.01	366.1427	M+H	367.1509	4.87	4.86	4.89	4.86	4.86	4.82	4.89	99.46	99.29	100.00	99.33	99.36	98.60
Pipecolate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.91	129.0790	M+H	130.0863	5.69	5.70	5.68	5.70	5.69	5.66	5.70	99.77	99.91	99.56	100.00	99.74	99.25
Theobromine	Imidazopyrimidines	Purines and purine derivatives	2.42	180.0647	M+H	181.0720	5.84	5.42	5.86	5.82	5.94	5.89	5.94	98.18	91.25	98.56	97.91	100.00	99.06
Niacinamide (vit B3)	Pyridines and derivatives	Pyridinecarboxylic acids and derivatives	1.12	122.0480	M+H	123.0553	4.54	4.46	4.56	4.43	4.62	4.54	4.62	98.14	96.52	98.62	95.83	100.00	98.19
1-Methyl-2-piperidinecarboxylate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.93	143.0946	M+H	144.1019	4.77	4.76	4.77	4.78	4.79	4.72	4.79	99.65	99.43	99.73	99.79	100.00	98.69
Tyramine ISF.2	Benzene and substituted derivativ	Phenylethylamines	1.46	102.0452		103.0530	4.50	4.50	4.50	4.48	4.52	4.48	4.52	99.64	99.61	99.72	99.28	100.00	99.11
Isobutyrylcarnitine	Fatty Acyls	Fatty acid esters	2.41	231.1471	M+H	232.1543	5.10	5.09	5.09	5.10	5.14	5.06	5.14	99.35	99.05	99.17	99.29	100.00	98.62
N-acetyl-DL-glutamate ISF.1	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	1.36	129.0422		130.0500	5.98	5.99	5.95	5.99	5.95	5.99	5.99	99.72	100.00	99.27	99.94	99.29	99.91
Pyroglutamate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	1.27	129.0426	M+H	130.0499	5.98	5.99	5.95	5.99	5.95	5.99	5.99	99.72	100.00	99.27	99.94	99.29	99.91
Pseudouridine ISF.1	Nucleoside and nucleotide analogues		0.95	208.0482		209.0560	4.79	4.80	4.81	4.78	4.79	4.72	4.81	99.51	99.71	100.00	99.40	99.52	98.15
Trigonelline			0.62	137.0477	M+H	138.0550	5.65	5.67	5.63	5.67	5.60	5.65	5.67	99.55	100.00	99.29	99.94	98.72	99.59
Paraxanthine	Imidazopyrimidines	Purines and purine derivatives	2.77	180.0647	M+H	181.0720	5.97	5.69	5.97	5.97	6.05	6.00	6.05	98.57	94.02	98.62	98.69	100.00	99.12
Theophylline	Imidazopyrimidines	Purines and purine derivatives	2.76	180.0647	M+H	181.0720	5.97	5.69	5.97	5.97	6.05	6.00	6.05	98.57	94.02	98.62	98.69	100.00	99.12
L-Acetylcarnitine	Fatty Acyls	Fatty acid esters	0.98	203.1158	M+H	204.1230	6.43	6.47	6.27	6.47	6.56	6.31	6.56	98.05	98.67	95.70	98.67	100.00	96.29
N6-Acetyl-L-lysine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.92	188.1161	M+H	189.1234	4.54	4.51	4.55	4.52	4.55	4.54	4.55	99.73	99.07	100.00	99.28	99.96	99.75
Hypoxanthine	Imidazopyrimidines	Purines and purine derivatives	1.21	136.0385	M+H	137.0458	5.65	5.66	5.65	5.64	5.65	5.59	5.66	99.72	100.00	99.79	99.58	99.86	98.75
Tryptophan	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	2.75	204.0899	M+H	205.0972	6.30	6.35	6.23	6.33	6.27	6.31	6.35	99.12	100.00	98.00	99.58	98.69	99.39
2-Octenoylcarnitine	Fatty Acyls	Fatty acid esters	5.65	285.1940	M+H	286.2013	5.66	5.64	5.60	5.67	5.76	5.59	5.76	98.34	98.01	97.29	98.45	100.00	97.17
dimethylarginine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.63	202.1430	M+H	203.1503	5.87	5.90	5.81	5.89	5.87	5.86	5.90	99.43	100.00	98.44	99.82	99.50	99.32
Carnitine	Organonitrogen compounds	Quaternary ammonium salts	0.54	161.1052	M+H	162.1125	6.54	6.59	6.41	6.55	6.55	6.60	6.60	99.00	99.86	97.02	99.15	99.15	100.00
Tyrosine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	1.40	181.0739	M+H	182.0812	6.37	6.41	6.29	6.39	6.36	6.39	6.41	99.29	100.00	98.08	99.61	99.16	99.70
Creatine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.61	131.0695	M+H	132.0768	6.32	6.37	6.25	6.36	6.30	6.35	6.37	99.24	100.00	98.00	99.77	98.88	99.65
Betaine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.56	117.0790	M+H	118.0863	6.40	6.45	6.32	6.43	6.37	6.42	6.45	99.23	100.00	98.02	99.77	98.89	99.52
Proline betaine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.68	143.0946	M+H	144.1019	6.52	6.57	6.45	6.55	6.50	6.54	6.57	99.25	100.00	98.15	99.69	98.92	99.53

List of identified metabolites (RPC negtaive mode peakpantheR) and relative method induced losses (%)

								Average intensity						Relative percentage					
Compound	**HMDBClass**	**HMDBSubClass**	**Retention time (min)**	**cpdMonoisotopic**	**ion**	**m/z**	**Pool-mean**	**Folch-mean**	**BD-mean**	**Matyash-mean**	**DSPE-mean**	**MeOH-mean**	**Max intensity**	**Pool**	**Folch**	**BD**	**Matyash**	**DSPE**	**MeOH**
1,3,7-Trimethylurate	Imidazopyrimidines	Purines and purine derivatives	2.95	210.0753	M-H	209.0680	3.75	3.68	3.69	3.77	3.79	3.80	3.80	98.58	96.83	97.16	99.11	99.84	100.00
1,7-Dimethylurate	Imidazopyrimidines	Purines and purine derivatives	2.57	196.0596	M-H	195.0524	4.35	4.32	4.34	4.34	4.39	4.37	4.39	99.16	98.51	98.79	98.80	100.00	99.57
2-Hydroxy-2-methylbutanoate	Fatty Acyls	Fatty acids and conjugates	2.45	118.0630	M-H	117.0557	3.98	3.90	4.01	4.00	4.00	3.95	4.01	99.30	97.40	100.00	99.78	99.89	98.61
Methylglutarate	Fatty Acyls	Fatty acids and conjugates	2.66	146.0579	M-H	145.0506	4.26	4.33	4.35	4.32	4.00	3.98	4.35	97.99	99.60	100.00	99.45	92.02	91.53
Adipate	Fatty Acyls	Fatty acids and conjugates	2.59	146.0579	M-H	145.0506	4.25	4.32	4.34	4.31	3.99	3.97	4.34	97.97	99.56	100.00	99.43	91.99	91.50
Azelaic Acid	Fatty Acyls	Fatty acids and conjugates	5.48	188.1049	M-H	187.0976	5.60	5.44	5.89	5.37	5.35	5.40	5.89	95.18	92.32	100.00	91.15	90.89	91.64
Cholate	Steroids and steroid derivatives	Bile acids, alcohols and derivatives	9.75	408.2876	M-H	407.2803	4.76	4.78	4.82	4.88	4.63	4.56	4.88	97.57	98.01	98.76	100.00	94.80	93.41
Indolelactate	Indoles and derivatives	Indolyl carboxylic acids and derivatives	4.86	205.0739	M-H	204.0666	5.18	5.18	5.22	5.23	5.09	5.15	5.23	99.13	99.07	99.80	100.00	97.32	98.58
4-Hydroxyphenyllactate	Phenylpropanoic acids		2.81	182.0579	M-H	181.0506	5.10	5.09	5.10	5.11	5.10	5.10	5.11	99.75	99.48	99.67	100.00	99.69	99.78
Histidine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.52	155.0695	M-H	154.0622	5.18	5.17	5.12	5.19	5.27	5.10	5.27	98.23	98.21	97.16	98.46	100.00	96.82
Hypoxanthine	Imidazopyrimidines	Purines and purine derivatives	1.20	136.0385	M-H	135.0312	4.89	4.89	4.88	4.88	4.94	4.88	4.94	99.09	98.95	98.78	98.80	100.00	98.73
Indoxyl sulfate	Organic sulfuric acids and derivatives	Arylsulfates	3.40	213.0096	M-H	212.0023	6.32	6.36	6.26	6.35	6.32	6.34	6.36	99.47	100.00	98.51	99.94	99.41	99.67
Tyrosine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	1.38	181.0739	M-H	180.0666	6.35	6.37	6.25	6.36	6.40	6.41	6.41	99.04	99.28	97.46	99.18	99.73	100.00
Acetaminophen Glucuronide	Organooxygen compounds	Carbohydrates and carbohydrate conjuga	1.92	327.0954	M-H	326.0881	5.04	5.03	5.04	5.03	5.00	5.08	5.08	99.13	99.04	99.16	99.05	98.31	100.00
Pantothenate	Alcohols and polyols	Polyols	2.31	219.1107	M-H	218.1034	4.52	4.51	4.52	4.52	4.53	4.50	4.53	99.75	99.49	99.74	99.87	100.00	99.44
Acetaminophen Sulfate	Organic sulfuric acids and derivatives	Arylsulfates	2.26	231.0201	M-H	230.0129	5.51	5.51	5.47	5.51	5.51	5.54	5.54	99.34	99.37	98.66	99.41	99.47	100.00
p-Cresol sulfate	Organic sulfuric acids and derivatives	Arylsulfates	4.20	188.0143	M-H	187.0071	6.95	7.00	6.82	6.97	6.97	7.05	7.05	98.63	99.31	96.76	98.86	98.87	100.00
Salicylate	Benzene and substituted derivatives	Benzoic acids and derivatives	5.48	138.0317	M-H	137.0244	4.69	4.69	4.70	4.67	4.68	4.74	4.74	99.14	99.00	99.20	98.67	98.78	100.00
Succinate	Carboxylic acids and derivatives	Dicarboxylic acids and derivatives	1.37	118.0266	M-H	117.0193	5.40	5.43	5.42	5.41	5.31	5.41	5.43	99.36	100.00	99.68	99.50	97.67	99.58
Theophylline	Imidazopyrimidines	Purines and purine derivatives	2.76	180.0647	M-H	179.0574	5.03	4.73	5.03	5.02	5.16	5.10	5.16	97.64	91.75	97.64	97.30	100.00	99.01
Urate	Imidazopyrimidines	Purines and purine derivatives	1.12	168.0283	M-H	167.0211	6.89	6.94	6.74	6.91	6.90	7.00	7.00	98.41	99.15	96.32	98.74	98.54	100.00
Uridine	Pyrimidine nucleosides		1.38	244.0695	M-H	243.0623	5.60	5.59	5.57	5.58	5.65	5.60	5.65	99.05	98.92	98.53	98.64	100.00	99.01
D-Gluconate	Organooxygen compounds	Carbohydrates and carbohydrate conjuga	0.58	196.0583	M-H	195.0505	5.82	5.48	5.54	5.51	6.13	6.05	6.13	94.88	89.40	90.31	89.89	100.00	98.63
Malate	Hydroxy acids and derivatives	Beta hydroxy acids and derivatives	0.75	134.0215	M-H	133.0137	4.75	4.76	4.83	4.76	4.68	4.66	4.83	98.50	98.63	100.00	98.61	97.03	96.50
Glutamate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	0.54	147.0532	M-H	146.0453	6.18	6.11	6.03	6.10	6.27	6.37	6.37	96.94	95.86	94.67	95.66	98.43	100.00
3-Hydroxyhippurate	Benzene and substituted derivatives	Benzoic acids and derivatives	2.73	195.0532	M-H	194.0453	4.17	4.17	4.17	4.15	4.15	4.22	4.22	98.97	99.00	98.84	98.49	98.45	100.00
N-acetyl-L-glutamate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	1.35	189.0637	M-H	188.0559	3.94	3.88	3.99	3.90	3.82	4.05	4.05	97.25	95.85	98.63	96.31	94.40	100.00
Phenylalanine	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	2.05	165.0790	M-H	164.0712	6.14	6.17	5.96	6.17	6.25	6.17	6.25	98.33	98.82	95.48	98.85	100.00	98.82
Glutarate	Carboxylic acids and derivatives	Dicarboxylic acids and derivatives	1.86	132.0423	M-H	131.0344	4.39	4.51	4.50	4.34	4.20	4.19	4.51	97.15	100.00	99.59	96.23	92.93	92.72
N-Acetylneuraminate	Organooxygen compounds	Carbohydrates and carbohydrate conjuga	0.62	309.1060	M-H	308.0987	6.17	6.27	6.24	6.41	5.24	5.49	6.41	96.24	97.79	97.24	100.00	81.78	85.59
Quinate	Organooxygen compounds	Alcohols and polyols	0.65	192.0634	M-H	191.0561	5.28	5.25	5.28	5.27	5.31	5.27	5.31	99.37	98.88	99.26	99.20	100.00	99.23
Xanthine	Imidazopyrimidines	Purines and purine derivatives	1.31	152.0334	M-H	151.0262	5.10	5.09	5.09	5.10	5.14	5.08	5.14	99.32	99.09	99.10	99.34	100.00	98.86
4-Hydroxyhippurate	Benzene and substituted derivatives	Benzoic acids and derivatives	2.45	195.0532	M-H	194.0459	4.27	4.27	4.28	4.27	4.25	4.29	4.29	99.57	99.56	99.62	99.50	99.11	100.00
Pyroglutamate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	1.25	129.0426	M-H	128.0353	6.35	6.36	6.26	6.36	6.35	6.42	6.42	98.86	99.13	97.58	99.07	98.96	100.00
Xanthosine	Purine nucleosides		1.83	284.0757	M-H	283.0684	3.92	3.82	3.95	3.90	3.97	3.84	3.97	98.66	96.27	99.59	98.16	100.00	96.73
N-Acetylaspartate	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	1.00	175.0481	M-H	174.0408	4.30	4.28	4.34	4.29	4.26	4.31	4.34	99.08	98.54	100.00	98.83	98.12	99.38
N-acetyl-L-carnosine	Peptidomimetics	Hybrid peptides	0.92	268.1172	M-H	267.1090	4.36	4.32	4.38	4.35	4.40	4.31	4.40	99.00	98.03	99.51	98.85	100.00	97.87
Riboflavin	Pteridines and derivatives	Alloxazines and isoalloxazines	3.82	376.1383	M-H	375.1310	2.88	2.39	2.50	1.58	3.36	2.83	3.36	85.63	71.08	74.29	46.97	100.00	84.03
Suberate	Fatty Acyls	Fatty acids and conjugates	4.46	174.0892	M-H	173.0819	4.70	4.79	4.81	4.65	4.52	4.59	4.81	97.70	99.60	100.00	96.60	93.91	95.39
Sucrose	Organooxygen compounds	Carbohydrates and carbohydrate conjuga	0.68	342.1162	M-H	341.1089	4.86	4.84	4.84	4.87	4.90	4.82	4.90	99.15	98.74	98.81	99.45	100.00	98.35
Phenyllactate	Phenylpropanoic acids		4.51	166.0630	M-H	165.0552	4.62	4.60	4.63	4.63	4.62	4.60	4.63	99.73	99.39	99.88	100.00	99.79	99.38
a-Hydroxyisobutanoate	Hydroxy acids and derivatives	Alpha hydroxy acids and derivatives	1.62	104.0473	M-H	103.0401	5.43	5.43	5.40	5.44	5.44	5.44	5.44	99.73	99.79	99.10	100.00	99.90	99.95
Indoxyl glucuronide	Organooxygen compounds	Carbohydrates and carbohydrate conjuga	3.39	309.0849	M-H	308.0776	3.05	3.11	3.12	3.14	2.82	2.72	3.14	97.23	99.18	99.53	100.00	89.81	86.62
Citrate	Carboxylic acids and derivatives	Tricarboxylic acids and derivatives	1.13	192.0270	M-H	191.0197	6.55	6.07	6.77	6.44	6.35	6.37	6.77	96.80	89.79	100.00	95.21	93.87	94.20
Pregnanediol-3-glucuronide	Steroids and steroid derivatives	Steroidal glycosides	9.23	496.3036	M-H	495.2953	5.09	5.16	5.12	5.20	4.89	5.00	5.20	97.95	99.20	98.48	100.00	94.13	96.11
p-Cresol glucuronide	Organooxygen compounds	Carbohydrates and carbohydrate conjuga	4.30	284.0896	M-H	283.0824	4.37	4.39	4.37	4.38	4.31	4.43	4.43	98.69	99.09	98.49	98.75	97.21	100.00
2-hydroxybutyric acid	Hydroxy acids and derivatives	Alpha hydroxy acids and derivatives	1.30	104.0473	M-H	103.0403	4.84	4.85	4.79	4.86	4.86	4.86	4.86	99.53	99.76	98.38	100.00	99.92	99.97
4-hydroxybutyric acid	Hydroxy acids and derivatives	Alpha hydroxy acids and derivatives	1.67	104.0473	M-H	103.0404	5.43	5.43	5.40	5.45	5.44	5.44	5.45	99.73	99.80	99.11	100.00	99.90	99.96
Epinephrine sulfate	Organic sulfuric acids and derivatives	Arylsulfates	0.89	263.0464	M-H	262.0425	4.65	4.59	4.49	4.28	4.75	4.57	4.75	97.86	96.54	94.42	89.99	100.00	96.18

List of annotated metabolites (HILIC positive mode peakpantheR)

Compound	HMDBSubClass	HMDB Class	cpdMonoisotopic	ion	Retention time (min)	m/z
1,2-Dimyristoyl-sn-glycero-3-phosphocholine	Organooxygen compounds	Carbohydrates and carbohydrate conjugates	677.4996	M+H	4.18	678.5068
Citrulline	Carboxylic acids and derivatives	Amino acids, peptides, and analogues		M+Na	5.74	198.0850
L-prolyl-L-proline			212.1161	M+H	5.40	213.1234
Sucrose			342.1162	M+Na	2.27	365.1053
Oleoylcarnitine (C18:1)				M+H	3.65	426.3578
a-glycerophosphocholine			257.1028	M+H	6.02	258.1101
Hypoxanthine	Glycerophospholipids	Glycerophosphocholines	136.0385	M+H	1.59	137.0458
Tryptamine			160.1000	M-NH3+H	2.79	144.0813
Propionylcarnitine	Fatty Acyls	Fatty acid esters		M+Na	4.82	240.1180
1-Methyl-2-piperidinecarboxylate			143.0946	M+H	4.59	144.1019
Histidine			155.0695	M+H	6.15	156.0768
Cotinine	Pyridines and derivatives	Pyridinecarboxylic acids and derivatives	176.0950	M+H	1.40	177.1022
Trigonelline			137.0477	M+H	4.85	138.0550
Lysine			146.1055	M+H	5.98	147.1128
Methionine			149.0510	M+H	4.11	150.0583
4-Trimethylammoniobutanoate			146.1181	M	4.91	146.1181
Palmitoylcarnitine (C16:0)			399.3349	M+H	3.78	400.3421
N6,N6,N6-Trimethyllysine			188.1525	M+H	6.16	189.1598
N,N-Dimethylglycine			103.0633	M+H	4.64	104.0712
Glutarylcarnitine			275.1369	M+H	5.10	276.1442
Alanine			89.0477	M+2Na-H	4.57	134.0188
Tetradecenoylcarnitine (C14:1)			369.2879	M+H	3.86	370.2952
Taurine			125.0147	M+H	2.61	126.0219
3-methylhistidine			169.0851	M+H	6.35	170.0924
Linoleoylcarnitine (C18:2)			423.3349	M+H	3.80	424.3421
N-Acetyl-D-mannosamine			221.0899	M+Na	1.81	244.0792
Acetaminophen			151.0633	M+H	0.71	152.0706
Trimethylaminoacetone			116.1075	M+	3.51	116.1075
Symmetric | Asymmetric Dimethylarginine			202.1430	M+H	5.91	203.1503
Tetradecadienoylcarnitine (C14:2)	Fatty Acyls	Fatty acid esters	367.2723	M+H	3.81	368.2795
N1-Acetylspermidine			187.1685	M+H	5.99	188.1757
Tetradecanoylcarnitine (C14:0)			371.3036	M+H	3.78	372.3109
Creatine			131.0695	M+H	5.06	132.0768
Hydroxybutyrylcarnitine (C4:0-OH)			247.1420	M+H	5.40	248.1492
Decanoylcarnitine (C10:0)			315.2410	M+H	3.92	316.2482
Tiglylcarnitine (C5:1)			243.1471	M+H	4.54	244.1543
Laurylcarnitine (C12:0)			343.2723	M+H	3.97	344.2795
Pipecolate | N-methyl proline			129.0790	M+H	4.48	130.0863
N6-Methyladenosine			281.1124	M+H	1.43	282.1197
Hydroxyisovaleroyl Carnitine	HMDBClass	HMDBSubClass	261.1576	M+H	5.22	262.1649
Butyryl- | isobutyrylcarnitine			231.1471	M+H	4.69	232.1543
Decenoylcarnitine (10:1)			313.2253	M+H	4.00	314.2326
Proline			115.0633	M+H	4.46	116.0706
N1-methyl-4-pyridone-3-carboxamide				M+H	1.50	153.0659
Cortisol			362.2093	M+H	0.71	363.2166
Urocanate			138.0429	M+H	1.30	139.0502
Choline			104.1075	M	3.97	104.1075
Homoarginine			188.1273	M+H	5.92	189.1346
4-Guanidinobutanoate			145.0851	M+H	3.84	146.0930
1-Methylnicotinamide			137.0715	M+	4.08	137.0713
Isovaleryl | valeryl | 2-methylbutyryl carnitine			245.1627	M+H	4.39	246.1700
Phenacetylcarnitine			280.1549	M	4.47	280.1549
Pantothenate			219.1107	M+H	1.13	220.1179
Hexanoylcarnitine (C6:0)			259.1784	M+H	4.34	260.1856
Arginine			174.1117	M+H	5.92	175.1190
Carnitine	Organonitrogen compounds	Quaternary ammonium salts	161.1052	M+H	5.32	162.1125
Betaine	Organooxygen compounds	Carbonyl compounds	117.0790	M+H	4.78	118.0863
Tryptophan			204.0899	M+H	3.82	205.0972
Creatinine			113.0589	M+Na	2.51	136.0481
Pseudouridine			244.0695	M+H	1.25	245.0768
Paraxanthine			180.0647	M+H	0.95	181.0720
Caffeine			194.0804	M+H	0.85	195.0877
Trimethylamine N-oxide			75.0684	M+H	4.18	76.0757
Dodecenoylcarnitine (12:1)			341.2566	M+H	3.89	342.2639
Niacinamide			122.0480	M+H	1.07	123.0553
Hydroxydecanoylcarnitine (C10:0-OH)			331.2359	M+H	4.49	332.2431
L-Acetylcarnitine			203.1158	M+H	5.03	204.1230
Phenylalanine			165.0790	M+H	3.78	166.0863
1-Methyladenosine			281.1124	M+H	1.43	282.1197
Decadienoylcarnitine (C10:2)			311.2097	M+H	4.11	312.2169
Octanoylcarnitine (C8:0)	Organoheterocyclic compounds	Pyridines and derivatives	287.2097	M+H	4.13	288.2169
Proline Betaine			143.0946	M+Na	4.91	166.0839
Octenoylcarnitine (C8:1)	Fatty Acyls	Fatty acid esters	285.1940	M+H	4.20	286.2013
1,1-Dimethylbiguanide (Metformin)			129.1014	M+H	3.42	130.1087
Warfarin	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	308.1049	M+H	0.67	309.1121
N1-methyl-2-pyridone-5-carboxamide				M+H	1.50	153.0659

List of annotated metabolites from region 1 (LIPID positive mode peakpantheR), and relative method induced losses (%)

										Average intensity							Relative percentage		
Compound	**HMDBSubClass**	**HMDBDirectParent**	**LMapsID**	**LMapsSubclass**	**cpdMonoisotopicMZ**	**ion**	**m/z**	**Retention time (min)**	**Region**	**BD-mean**	**BD-%CV**	**Folch-mean**	**Folch-%CV**	**Matyash-mean**	**Matyash-%CV**	**Max intensity**	**BD**	**Folch**	**Matyash**
CAR(8:1)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	285.1940	[M+H]+	286.2015	0.42	1	4.45	1.17	4.19	0.72	4.09	0.65	4.45	100.00	94.06	91.95
CAR(8:0)	Fatty acid esters	Acyl carnitines	LMFA07070095	Fatty acyl carnitines [FA0707]	287.2097	[M+H]+	288.2175	0.46	1	4.13	1.68	4.08	0.89	3.94	1.02	4.13	100.00	98.90	95.43
CAR(10:2)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	311.2097	[M+H]+	312.2168	0.47	1	3.79	6.11	3.74	2.15	3.60	2.33	3.79	100.00	98.76	94.88
CAR(10:0-OH)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	331.2359	[M+H]+	332.2440	0.49	1	3.56	2.78	3.22	2.63	3.44	5.20	3.56	100.00	90.53	96.73
CAR(10:1)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	313.2253	[M+H]+	314.2332	0.53	1	4.46	0.73	4.68	0.57	4.40	0.73	4.68	95.33	100.00	93.93
CAR(12:1-OH)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	357.2515	[M+H]+	358.2585	0.56	1	3.45	1.85	3.63	2.05	3.26	4.30	3.63	95.03	100.00	89.73
CAR(10:0)	Fatty acid esters	Acyl carnitines	LMFA07070059	Fatty acyl carnitines [FA0707]	315.2410	[M+H]+	316.2490	0.61	1	4.61	0.90	4.87	0.43	4.60	0.33	4.87	94.58	100.00	94.38
CAR(12:1)_1	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	341.2566	[M+H]+	342.2645	0.72	1	4.36	0.58	4.75	0.66	4.47	0.69	4.75	91.90	100.00	94.25
CAR(18:0-DC)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	457.3403	[M+H]+	458.3472	0.84	1	2.55	19.69	0.85	155.09	0.96	167.59	2.55	100.00	33.17	37.50
CAR(14:2)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	367.2723	[M+H]+	368.2803	0.85	1	4.25	1.46	4.69	0.66	4.43	0.50	4.69	90.72	100.00	94.58
CAR(12:0)	Fatty acid esters	Acyl carnitines	LMFA07070062	Fatty acyl carnitines [FA0707]	343.2723	[M+H]+	344.2804	0.88	1	4.40	0.49	4.76	0.58	4.50	0.61	4.76	92.47	100.00	94.52
CAR(14:0-OH)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	387.2985	[M+H]+	388.3057	0.97	1	3.35	2.18	3.77	1.62	3.67	1.74	3.77	88.77	100.00	97.27
CAR(14:1)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	369.2879	[M+H]+	370.2959	1.06	1	4.56	0.82	4.90	0.42	4.72	0.57	4.90	93.20	100.00	96.29
CAR(16:1-OH)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	413.3141	[M+H]+	414.3219	1.15	1	3.24	4.92	3.72	2.58	3.52	1.68	3.72	87.18	100.00	94.60
CAR(16:2)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	395.3036	[M+H]+	396.3109	1.16	1	3.61	3.29	4.04	1.38	3.89	0.95	4.04	89.34	100.00	96.28
CAR(14:0)	Fatty acid esters	Acyl carnitines	LMFA07070107	Fatty acyl carnitines [FA0707]	371.3036	[M+H]+	372.3116	1.26	1	4.10	1.89	4.47	0.94	4.32	0.45	4.47	91.70	100.00	96.63
LPC(0:0/14:0)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines	LMGP01050073	Monoacylglycerophosphocholines [GP0105]	467.3012	[M+H]+	468.3090	1.27	1	5.14	1.12	5.44	0.63	5.28	0.34	5.44	94.66	100.00	97.13
LPC(0:0/18:3)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	517.3168	[M+H]+	518.3236	1.28	1	4.47	1.60	4.76	0.65	4.62	0.48	4.76	94.01	100.00	97.20
LPC(0:0/20:5)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	541.3168	[M+H]+	542.3242	1.29	1	4.60	1.89	4.89	0.80	4.76	0.52	4.89	94.06	100.00	97.32
CAR(18:3)_2	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	421.3192	[M+H]+	422.3259	1.35	1	3.48	1.46	3.82	1.52	3.70	1.24	3.82	91.17	100.00	96.84
LPC(14:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050012	Monoacylglycerophosphocholines [GP0105]	467.3012	[M+H]+	468.3092	1.36	1	5.14	1.11	5.44	0.63	5.28	0.34	5.44	94.66	100.00	97.13
LPC(18:3/0:0)_1	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	517.3168	[M+H]+	518.3245	1.36	1	4.47	1.52	4.75	0.67	4.62	0.45	4.75	94.04	100.00	97.22
LPC(20:5/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	541.3168	[M+H]+	542.3246	1.37	1	4.60	1.89	4.89	0.80	4.76	0.52	4.89	94.06	100.00	97.32
LPA(18:2/0:0)	Glycerophosphates	1-acylglycerol-3-phosphates		Monoacylglycerophosphates [GP1005]	434.2433	[M+H-H2O]+	417.2409	1.39	1	3.47	2.31	2.31	49.48	3.67	1.41	3.67	94.55	63.08	100.00
LPA(20:4/0:0)	Glycerophosphates	1-acylglycerol-3-phosphates		Monoacylglycerophosphates [GP1005]	458.2433	[M+H-H2O]+	441.2401	1.39	1	2.95	6.25	1.54	78.98	3.18	2.35	3.18	92.77	48.40	100.00
LPC(0:0/16:1)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	493.3168	[M+H]+	494.3243	1.41	1	5.45	0.92	5.68	0.64	5.56	0.38	5.68	95.91	100.00	97.92
CAR(16:1)_2	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	397.3192	[M+H]+	398.3269	1.45	1	4.69	5.26	4.98	1.34	4.86	2.05	4.98	94.09	100.00	97.56
LPC(16:1/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	493.3168	[M+H]+	494.3249	1.50	1	5.45	0.92	5.68	0.64	5.56	0.38	5.68	95.91	100.00	97.92
LPC(0:0/15:0)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	481.3168	[M+H]+	482.3244	1.50	1	5.05	1.06	5.31	0.63	5.20	0.41	5.31	95.16	100.00	97.85
LPC(0:0/22:6)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	567.3325	[M+H]+	568.3403	1.52	1	5.28	1.74	5.52	0.64	5.44	0.40	5.52	95.60	100.00	98.40
CAR(18:2)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	423.3349	[M+H]+	424.3431	1.56	1	4.85	0.98	5.07	0.67	5.01	0.42	5.07	95.57	100.00	98.72
LPC(0:0/18:2)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	519.3325	[M+H]+	520.3407	1.56	1	6.69	0.70	6.77	0.59	6.71	0.47	6.77	98.81	100.00	99.05
LPC(0:0/20:4)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	543.3325	[M+H]+	544.3402	1.57	1	6.24	0.98	6.35	0.60	6.29	0.44	6.35	98.33	100.00	99.09
CAR(20:4)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	447.3349	[M+H]+	448.3429	1.58	1	3.56	3.28	4.00	1.09	3.92	1.14	4.00	89.20	100.00	98.15
LPC(22:6/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	567.3325	[M+H]+	568.3402	1.60	1	5.28	1.74	5.52	0.64	5.44	0.40	5.52	95.60	100.00	98.40
LPC(15:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050016	Monoacylglycerophosphocholines [GP0105]	481.3168	[M+H]+	482.3247	1.61	1	5.05	1.06	5.31	0.63	5.20	0.41	5.31	95.16	100.00	97.85
LPE(0:0/18:2)	Glycerophosphoethar	2-acyl-sn-glycero-3-phosphoethanolamines		Monoacylglycerophosphoethanolamines [GI	477.2855	[M+H]+	478.2939	1.63	1	3.50	14.36	4.67	0.61	4.57	1.34	4.67	74.89	100.00	97.80
LPC(20:4/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	543.3325	[M+H]+	544.3402	1.66	1	6.24	0.98	6.35	0.60	6.29	0.44	6.35	98.33	100.00	99.09
LPC(18:2/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	519.3325	[M+H]+	520.3406	1.66	1	6.69	0.70	6.77	0.59	6.71	0.47	6.77	98.81	100.00	99.05
LPC(0:0/22:5)_1	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	569.3481	[M+H]+	570.3550	1.66	1	4.95	1.47	5.11	0.82	5.07	0.61	5.11	96.86	100.00	99.11
LPE(18:2/0:0)	Glycerophosphoethar	1-acyl-sn-glycero-3-phosphoethanolamines		Monoacylglycerophosphoethanolamines [GI	477.2855	[M+H]+	478.2939	1.71	1	3.50	14.36	4.67	0.61	4.57	1.34	4.67	74.89	100.00	97.80
LPE(20:4/0:0)	Glycerophosphoethar	1-acyl-sn-glycero-3-phosphoethanolamines		Monoacylglycerophosphoethanolamines [GI	501.2855	[M+H]+	502.2944	1.71	1	3.92	8.21	4.77	0.77	4.66	1.04	4.77	82.22	100.00	97.74
LPA(18:1/0:0)	Glycerophosphates	1-acylglycerol-3-phosphates		Monoacylglycerophosphates [GP1005]	436.2590	[M+H-H2O]+	419.2559	1.71	1	3.57	3.68	3.73	2.56	3.70	2.09	3.73	95.51	100.00	99.11
LPC(22:5/0:0)_1	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	569.3481	[M+H]+	570.3556	1.74	1	4.95	1.47	5.11	0.82	5.07	0.61	5.11	96.86	100.00	99.11
LPC(17:1/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	507.3325	[M+H]+	508.3404	1.75	1	4.43	1.35	4.69	0.94	4.61	0.54	4.69	94.48	100.00	98.37
LPC(0:0/20:3)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	545.3481	[M+H]+	546.3558	1.75	1	5.24	4.36	5.57	4.54	5.53	3.27	5.57	94.09	100.00	99.26
CAR(16:0)	Fatty acid esters	Acyl carnitines	LMFA07070004	Fatty acyl carnitines [FA0707]	399.3349	[M+H]+	400.3429	1.75	1	4.95	1.03	5.16	0.66	5.11	0.59	5.16	95.80	100.00	98.90
CAR(20:3)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	449.3505	[M+H]+	450.3580	1.75	1	3.58	1.36	3.83	1.26	3.80	0.69	3.83	93.48	100.00	99.33
LPC(0:0/16:0)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines	LMGP01050074	Monoacylglycerophosphocholines [GP0105]	495.3325	[M+H]+	496.3407	1.76	1	7.43	0.87	7.45	0.59	7.41	0.55	7.45	99.77	100.00	99.39
LPE(0:0/16:0)	Glycerophosphoethar	2-acyl-sn-glycero-3-phosphoethanolar	LMGP02050036	Monoacylglycerophosphoethanolamines [GI	453.2855	[M+H]+	454.2935	1.82	1	3.65	8.09	4.57	0.67	4.56	1.03	4.57	79.98	100.00	99.75
LPC(20:3/0:0)_1	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	545.3481	[M+H]+	546.3561	1.84	1	5.62	1.20	5.71	0.64	5.68	0.38	5.71	98.49	100.00	99.51
LPC(16:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050018	Monoacylglycerophosphocholines [GP0105]	495.3325	[M+H]+	496.3407	1.87	1	7.43	0.87	7.45	0.59	7.41	0.55	7.45	99.77	100.00	99.39
LPC(20:3/0:0)_2	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	545.3481	[M+H]+	546.3559	1.91	1	5.62	1.20	5.71	0.64	5.68	0.38	5.71	98.49	100.00	99.51
CAR(18:1)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	425.3505	[M+H]+	426.3588	1.91	1	5.07	1.06	5.24	0.61	5.22	0.59	5.24	96.83	100.00	99.61
LPC(0:0/18:1)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	521.3481	[M+H]+	522.3563	1.91	1	6.65	0.78	6.73	0.50	6.69	0.51	6.73	98.76	100.00	99.43
LPE(16:0/0:0)	Glycerophosphoethar	1-acyl-sn-glycero-3-phosphoethanolar	LMGP02050002	Monoacylglycerophosphoethanolamines [GI	453.2855	[M+H]+	454.2935	1.93	1	3.66	8.09	4.57	0.67	4.56	1.03	4.57	80.02	100.00	99.75
LPC(18:1/0:0)_1	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	521.3481	[M+H]+	522.3562	2.01	1	6.65	0.78	6.73	0.50	6.69	0.51	6.73	98.76	100.00	99.43
LPC(0:0/20:2)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	547.3638	[M+H]+	548.3710	2.03	1	4.95	0.64	5.03	1.04	5.03	0.73	5.03	98.41	100.00	99.97
LPC(22:4/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	571.3638	[M+H]+	572.3713	2.05	1	4.80	1.32	4.85	0.82	4.84	0.59	4.85	98.79	100.00	99.75
LPC(18:1/0:0)_2	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	521.3481	[M+H]+	522.3565	2.08	1	6.65	0.78	6.73	0.50	6.69	0.51	6.73	98.76	100.00	99.43
LPC(O-16:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01060010	Monoalkylglycerophosphocholines [GP0106]	481.3532	[M+H]+	482.3610	2.09	1	5.89	0.65	5.97	0.95	5.93	0.34	5.97	98.66	100.00	99.29
LPC(P-16:0/0:0)	Glycerophosphocholir	1-(1Z-alkenyl)-glycero-3-phosphocholi	LMGP01070006	1Z-alkenylglycerophosphocholines [GP0107]	479.3376	[M+H]+	480.3446	2.09	1	5.66	1.13	5.81	1.02	5.85	0.46	5.85	96.86	99.41	100.00
LPC(20:2/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	547.3638	[M+H]+	548.3701	2.13	1	4.95	0.64	5.03	1.04	5.03	0.73	5.03	98.41	100.00	99.97
LPC(17:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050024	Monoacylglycerophosphocholines [GP0105]	509.3481	[M+H]+	510.3562	2.14	1	5.52	0.57	5.64	0.78	5.61	0.73	5.64	98.02	100.00	99.51
LPC(O-18:1/0:0)_1	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoalkylglycerophosphocholines [GP0106]	507.3689	[M+H]+	508.3768	2.22	1	5.69	0.85	5.74	0.95	5.71	0.91	5.74	99.21	100.00	99.57
CAR(18:0)	Fatty acid esters	Acyl carnitines	LMFA07070051	Fatty acyl carnitines [FA0707]	427.3662	[M+H]+	428.3743	2.30	1	4.59	0.71	4.72	0.90	4.70	0.52	4.72	97.20	100.00	99.54
LPC(0:0/18:0)	Glycerophosphocholir	2-acyl-sn-glycero-3-phosphocholines	LMGP01050076	Monoacylglycerophosphocholines [GP0105]	523.3638	[M+H]+	524.3721	2.30	1	7.17	0.71	7.16	0.68	7.15	0.52	7.17	100.00	99.87	99.68
LPC(O-18:1/0:0)_2	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoalkylglycerophosphocholines [GP0106]	507.3689	[M+H]+	508.3729	2.30	1	5.69	0.85	5.74	0.95	5.71	0.91	5.74	99.21	100.00	99.57
LPE(0:0/18:0)	Glycerophosphoethar	2-acyl-sn-glycero-3-phosphoethanolar	LMGP02050038	Monoacylglycerophosphoethanolamines [GI	481.3168	[M+H]+	482.3247	2.38	1	4.07	6.86	4.86	0.59	4.87	1.04	4.87	83.62	99.90	100.00
CAR(20:1)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	453.3818	[M+H]+	454.3891	2.40	1	3.76	1.24	3.83	0.55	3.83	1.12	3.83	98.27	100.00	100.00
LPC(18:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050026	Monoacylglycerophosphocholines [GP0105]	523.3638	[M+H]+	524.3720	2.42	1	7.17	0.71	7.16	0.68	7.15	0.52	7.17	100.00	99.87	99.68
LPE(18:0/0:0)	Glycerophosphoethar	1-acyl-sn-glycero-3-phosphoethanolar	LMGP02050001	Monoacylglycerophosphoethanolamines [GI	481.3168	[M+H]+	482.3247	2.48	1	4.07	6.86	4.86	0.59	4.87	1.04	4.87	83.62	99.90	100.00
LPC(20:1/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoacylglycerophosphocholines [GP0105]	549.3794	[M+H]+	550.3874	2.51	1	5.00	0.79	5.08	0.99	5.10	0.67	5.10	98.01	99.62	100.00
MG(16:0)	Monoradylglycerols	na		Monoacylglycerols [GL0101]	330.2770	[M+H-H2O]+	313.2740	2.52	1	4.01	2.57	4.11	3.24	4.13	2.37	4.13	97.19	99.41	100.00
MG(18:1)_1	Monoradylglycerols	na		Monoacylglycerols [GL0101]	356.2927	[M+H-H2O]+	339.2891	2.60	1	3.99	1.09	4.02	2.30	4.05	1.97	4.05	98.57	99.19	100.00
LPC(P-18:0/0:0)	Glycerophosphocholir	1-(1Z-alkenyl)-glycero-3-phosphocholi	LMGP01070009	1Z-alkenylglycerophosphocholines [GP0107]	507.3689	[M+H]+	508.3767	2.66	1	4.95	1.24	5.00	1.30	5.07	0.93	5.07	97.66	98.66	100.00
LPC(O-18:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01060014	Monoalkylglycerophosphocholines [GP0106]	509.3845	[M+H]+	510.3922	2.66	1	5.43	0.76	5.42	1.06	5.43	0.73	5.43	99.98	99.87	100.00
MG(18:1)_2	Monoradylglycerols	na		Monoacylglycerols [GL0101]	356.2927	[M+H-H2O]+	339.2896	2.66	1	3.99	1.23	4.02	2.28	4.05	1.95	4.05	98.72	99.26	100.00
LPC(19:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050041	Monoacylglycerophosphocholines [GP0105]	537.3794	[M+H]+	538.3874	2.70	1	4.52	0.92	4.62	1.22	4.62	1.00	4.62	97.86	100.00	99.97
LPE(P-18:0/0:0)	Glycerophosphoethar	1-(1Z-alkenyl)-glycero-3-phosphoethanolamines		Glycerophosphoethanolamines [GP02]	465.3219	[M+H]+	466.3293	2.75	1	3.86	6.55	4.63	0.59	4.68	0.98	4.68	82.51	98.90	100.00
CAR(20:0)	Fatty acid esters	Acyl carnitines	LMFA07070052	Fatty acyl carnitines [FA0707]	455.3975	[M+H]+	456.4047	2.88	1	3.53	1.92	3.66	1.73	3.68	2.53	3.68	95.75	99.53	100.00
LPC(20:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050045	Monoacylglycerophosphocholines [GP0105]	551.3951	[M+H]+	552.4030	3.02	1	4.71	0.59	4.76	1.30	4.79	0.97	4.79	98.41	99.55	100.00
LPC(O-20:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01060041	Monoalkylglycerophosphocholines [GP0106]	537.4158	[M+H]+	538.4231	3.32	1	4.65	0.95	4.62	1.05	4.69	0.99	4.69	99.16	98.49	100.00
CAR(24:1)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	509.4444	[M+H]+	510.4515	3.59	1	3.58	1.56	3.66	1.20	3.73	1.00	3.73	96.07	98.15	100.00
LPC(22:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050053	Monoacylglycerophosphocholines [GP0105]	579.4264	[M+H]+	580.4342	3.73	1	3.99	1.74	4.06	1.28	4.13	1.11	4.13	96.48	98.26	100.00

List of annotated metabolites from region 2 (LIPID positive mode peakpantheR), and relative method induced losses (%)

										Average intensity							Relative percentage		
Compound	**HMDBSubClass**	**HMDBDirectParent**	**LMapsID**	**LMapsSubclass**	**cpdMonoisotopicMZ**	**ion**	**m/z**	**Retention time (min)**	**Region**	**BD-mean**	**BD-%CV**	**Folch-mean**	**Folch-%CV**	**Matyash-mean**	**Matyash-%CV**	**Max intensity**	**BD**	**Folch**	**Matyash**
LPC(O-24:1/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoalkylglycerophosphocholines [GP0106]	591.4628	[M+H]+	592.4704	4.03	2	5.03	0.93	4.95	0.80	5.05	0.80	5.05	99.69	98.01	100.00
SM(d30:1); SM(d16:1/	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	646.5050	[M+H]+	647.5125	4.25	2	4.59	0.66	4.69	0.66	4.74	0.91	4.74	96.82	98.91	100.00
CAR(26:1)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	537.4757	[M+H]+	538.4831	4.33	2	4.11	1.04	4.16	1.04	4.20	1.22	4.20	97.74	99.05	100.00
CAR(24:0)	Fatty acid esters	Acyl carnitines		Fatty acyl carnitines [FA0707]	511.4601	[M+H]+	512.4675	4.34	2	4.44	0.56	4.47	0.56	4.50	1.01	4.50	98.70	99.35	100.00
SM(d18:2/14:0)	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	672.5206	[M+H]+	673.5280	4.39	2	4.91	0.49	4.96	0.49	5.00	0.59	5.00	98.08	99.20	100.00
LPC(24:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines	LMGP01050057	Monoacylglycerophosphocholines [GP0105]	607.4577	[M+H]+	608.4656	4.52	2	4.67	0.83	4.74	0.83	4.82	0.91	4.82	96.94	98.29	100.00
Cholesterol	Cholestane steroids	Cholesterols and derivatives	LMST01010001	Cholesterol and derivatives [ST0101]	386.3549	[M+H-H2O]+	369.3527	4.55	2	5.79	1.18	5.72	1.18	5.76	1.48	5.79	100.00	98.79	99.49
LPC(O-26:1/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoalkylglycerophosphocholines [GP0106]	619.4941	[M+H]+	620.5002	4.80	2	3.89	1.69	3.94	1.69	4.04	1.65	4.04	96.25	97.38	100.00
LPC(O-24:0/0:0)	Glycerophosphocholir	1-acyl-sn-glycero-3-phosphocholines		Monoalkylglycerophosphocholines [GP0106]	593.4784	[M+H]+	594.4853	4.84	2	4.69	0.86	4.71	0.86	4.77	0.97	4.77	98.33	98.87	100.00
SM(d32:1); SM(d16:1/	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	674.5363	[M+H]+	675.5438	4.97	2	6.16	0.35	6.17	0.35	6.22	0.59	6.22	98.96	99.16	100.00
SM(d18:2/16:0)_1	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	700.5519	[M+H]+	701.5596	5.11	2	6.47	0.37	6.45	0.37	6.50	0.52	6.50	99.46	99.17	100.00
PC(14:0/22:6)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	777.5309	[M+H]+	778.5376	5.14	2	4.40	0.88	4.49	0.88	4.53	1.11	4.53	96.99	99.05	100.00
CAR(26:0)	Fatty acid esters	Acyl carnitines	LMFA07070069	Fatty acyl carnitines [FA0707]	539.4914	[M+H]+	540.4990	5.15	2	4.71	0.68	4.69	0.68	4.74	0.49	4.74	99.46	98.99	100.00
PC(14:0/20:4)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	753.5309	[M+H]+	754.5378	5.32	2	5.15	1.09	5.18	1.09	5.24	0.33	5.24	98.42	98.94	100.00
PC(14:0/18:2)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	729.5309	[M+H]+	730.5377	5.32	2	5.75	0.89	5.71	0.89	5.77	0.30	5.77	99.59	98.90	100.00
SM(d17:1/16:0)	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	688.5519	[M+H]+	689.5595	5.33	2	5.95	0.64	5.93	0.64	6.01	0.31	6.01	99.01	98.71	100.00
PC(16:1/20:4)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	779.5465	[M+H]+	780.5534	5.46	2	6.06	0.81	6.07	0.81	6.13	0.29	6.13	98.88	99.12	100.00
PC(16:0/20:5)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	779.5465	[M+H]+	780.5539	5.55	2	6.06	0.81	6.07	0.81	6.13	0.29	6.13	98.88	99.12	100.00
PC(18:2/20:4)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	805.5622	[M+H]+	806.5690	5.63	2	6.87	0.72	6.82	0.72	6.87	0.34	6.87	99.95	99.32	100.00
PC(18:2/18:2)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	781.5622	[M+H]+	782.5694	5.64	2	6.60	0.76	6.53	0.76	6.58	0.51	6.60	100.00	98.90	99.63
PC(15:0/18:2)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	743.5465	[M+H]+	744.5510	5.67	2	5.64	0.64	5.56	0.64	5.59	0.48	5.64	100.00	98.45	99.12
SM(d18:1/16:0)	Phosphosphingolipids	Phosphosphingolipids	LMSP03010003	Ceramide phosphocholines (sphingomyelins) [S	702.5676	[M+H]+	703.5752	5.68	2	7.46	0.58	7.32	0.58	7.37	0.71	7.46	100.00	98.07	98.82
LacCer(d18:1/16:0)	Glycosphingolipids	Glycosyl-N-acylsphingosines	LMSP0501AB03	Simple Glc series [SP0501]	861.6177	[M+H-H2O]+	844.6120	5.75	2	5.01	1.94	5.02	1.94	5.22	0.63	5.22	95.90	96.12	100.00
PC(16:0/20:4)_1	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	781.5622	[M+H]+	782.5691	5.77	2	7.49	0.63	7.33	0.63	7.35	0.75	7.49	100.00	97.84	98.12
PC(16:0/22:6)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	805.5622	[M+H]+	806.5697	5.84	2	6.87	0.72	6.82	0.72	6.87	0.34	6.87	99.95	99.32	100.00
SM(d18:2/18:0)	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	728.5832	[M+H]+	729.5905	5.86	2	6.31	0.66	6.28	0.66	6.36	0.35	6.36	99.33	98.74	100.00
PC(32:1); PC(14:0/18:1	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	731.5465	[M+H]+	732.5541	5.87	2	6.39	0.81	6.32	0.81	6.39	0.29	6.39	100.00	98.85	99.95
SM(d18:0/16:0)	Phosphosphingolipids	Phosphosphingolipids	LMSP03010004	Ceramide phosphocholines (sphingomyelins) [S	704.5832	[M+H]+	705.5907	5.94	2	6.09	0.78	6.04	0.78	6.12	0.28	6.12	99.56	98.73	100.00
SM(d19:1/16:0)	Phosphosphingolipids	Phosphosphingolipids	LMSP03010045	Ceramide phosphocholines (sphingomyelins) [S	716.5832	[M+H]+	717.5902	5.95	2	5.77	0.77	5.67	0.77	5.74	0.87	5.77	100.00	98.27	99.57
SM(d18:1/18:1)	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	728.5832	[M+H]+	729.5898	5.97	2	6.31	0.66	6.28	0.66	6.36	0.35	6.36	99.33	98.74	100.00
PC(16:0/16:1)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	731.5465	[M+H]+	732.5536	5.97	2	6.39	0.81	6.32	0.81	6.39	0.29	6.39	100.00	98.85	99.95
HEXCer(d18:1/16:0)	Glycosphingolipids	Glycosyl-N-acylsphingosines		Simple Glc series [SP0501]	699.5649	[M+H-H2O]+	682.5614	5.98	2	4.99	1.69	4.91	1.69	4.98	0.37	4.99	100.00	98.43	99.85
PC(16:0/18:2)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	757.5622	[M+H]+	758.5698	6.03	2	7.99	0.65	7.77	0.65	7.80	0.81	7.99	100.00	97.22	97.65
SM(d35:1); SM(d17:1/	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	716.5832	[M+H]+	717.5906	6.07	2	5.77	0.81	5.67	0.81	5.74	0.88	5.77	100.00	98.29	99.56
PC(O-16:0/14:0)	Glycerophosphocholir	1-alkyl,2-acylglycero-3-phosphocholines	LMGP01020178	1-alkyl,2-acylglycerophosphocholines [GP0102]	691.5516	[M+H]+	692.5586	6.14	2	4.95	1.14	4.88	1.14	4.92	0.72	4.95	100.00	98.55	99.42
PE(16:0/20:4)	Glycerophosphoethar	Phosphatidylethanolamines		Diacylglycerophosphoethanolamines [GP0201]	739.5152	[M+H]+	740.5222	6.17	2	4.11	0.69	4.86	0.69	4.81	1.09	4.86	84.53	100.00	98.93
PE(16:0/18:2)	Glycerophosphoethar	Phosphatidylethanolamines		Diacylglycerophosphoethanolamines [GP0201]	715.5152	[M+H]+	716.5250	6.19	2	4.23	0.84	4.83	0.84	4.80	0.96	4.83	87.64	100.00	99.25
PC(33:1); PC(15:0/18:1	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	745.5622	[M+H]+	746.5694	6.21	2	5.72	0.38	5.66	0.38	5.70	0.38	5.72	100.00	98.96	99.74
PC(O-18:1/22:6) and/c	Glycerophosphocholir	na		na	817.5985	[M+H]+	818.6051	6.34	2	5.09	1.07	5.17	1.07	5.21	0.65	5.21	97.71	99.41	100.00
PC(18:1/20:3)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	809.5935	[M+H]+	810.5986	6.38	2	7.34	0.61	7.20	0.61	7.21	0.66	7.34	100.00	98.09	98.29
PC(16:0/16:0)	Glycerophosphocholir	Phosphatidylcholines	LMGP01010564	Diacylglycerophosphocholines [GP0101]	733.5622	[M+H]+	734.5700	6.38	2	6.63	0.41	6.53	0.41	6.56	0.33	6.63	100.00	98.49	98.96
SM(d18:1/18:0)	Phosphosphingolipids	Phosphosphingolipids	LMSP03010001	Ceramide phosphocholines (sphingomyelins) [S	730.5989	[M+H]+	731.6065	6.41	2	6.66	0.43	6.58	0.43	6.62	0.54	6.66	100.00	98.77	99.43
PC(O-16:0/22:5)	Glycerophosphocholir	1-alkyl,2-acylglycero-3-phosphocholines		1-alkyl,2-acylglycerophosphocholines [GP0102]	793.5985	[M+H]+	794.6060	6.43	2	6.38	0.48	6.32	0.48	6.34	0.53	6.38	100.00	99.03	99.29
PC(O-16:0/18:2)	Glycerophosphocholir	1-alkyl,2-acylglycero-3-phosphocholines		1-alkyl,2-acylglycerophosphocholines [GP0102]	743.5829	[M+H]+	744.5903	6.45	2	6.24	0.58	6.20	0.58	6.23	0.36	6.24	100.00	99.26	99.74
PC(O-16:0/15:0)	Glycerophosphocholir	1-alkyl,2-acylglycero-3-phosphocholines	LMGP01020180	1-alkyl,2-acylglycerophosphocholines [GP0102]	705.5672	[M+H]+	706.5738	6.47	2	4.28	1.26	4.31	1.26	4.35	0.75	4.35	98.25	99.17	100.00
PC(16:0/22:4)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	809.5935	[M+H]+	810.6008	6.47	2	7.34	0.61	7.20	0.61	7.21	0.66	7.34	100.00	98.09	98.29
PC(16:0/18:1)_1	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	759.5778	[M+H]+	760.5856	6.48	2	7.68	0.51	7.51	0.51	7.54	0.56	7.68	100.00	97.85	98.18
SM(d16:1/20:0)	Phosphosphingolipids	Phosphosphingolipids	LMSP03010052	Ceramide phosphocholines (sphingomyelins) [S	730.5989	[M+H]+	731.6063	6.50	2	6.66	0.43	6.58	0.43	6.62	0.54	6.66	100.00	98.77	99.43
Cer(d18:1/16:0)	Ceramides	Ceramides	LMSP02010004	N-acylsphingosines (ceramides) [SP0201]	537.5121	[M+H-H2O]+	520.5095	6.50	2	4.94	0.73	4.79	0.73	4.80	0.65	4.94	100.00	96.90	97.17
PC(18:0/22:6)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	833.5935	[M+H]+	834.6009	6.52	2	6.40	0.46	6.37	0.46	6.39	0.42	6.40	100.00	99.60	99.84
SM(d18:1/20:1)	Phosphosphingolipids	Phosphosphingolipids		Ceramide phosphocholines (sphingomyelins) [S	756.6145	[M+H]+	757.6222	6.60	2	6.09	0.38	6.06	0.38	6.11	0.18	6.11	99.66	99.15	100.00
PC(36:2); PC(16:0/20:2	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	785.5935	[M+H]+	786.6011	6.61	2	7.74	0.60	7.55	0.60	7.58	0.63	7.74	100.00	97.52	97.83
PC(18:0/20:4)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	809.5935	[M+H]+	810.6018	6.71	2	7.34	0.61	7.20	0.61	7.21	0.66	7.34	100.00	98.09	98.29
PC(18:0/18:2)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	785.5935	[M+H]+	786.6018	6.73	2	7.74	0.60	7.55	0.60	7.58	0.63	7.74	100.00	97.52	97.83
PC(O-16:0/16:0)	Glycerophosphocholir	1-alkyl,2-acylglycero-3-phosphocholines	LMGP01020029	1-alkyl,2-acylglycerophosphocholines [GP0102]	719.5829	[M+H]+	720.5906	6.81	2	5.90	0.68	5.84	0.68	5.89	0.34	5.90	100.00	98.99	99.87
DG(36:4)_1; DG(18:2/1	Diradylglycerols	na		Diacylglycerols [GL0201]	616.5067	[M+H]+	617.5140	6.86	2	4.20	2.27	4.23	2.27	4.23	2.03	4.23	99.27	99.91	100.00
PE(18:0/20:4)	Glycerophosphoethar	Phosphatidylethanolamines		Diacylglycerophosphoethanolamines [GP0201]	767.5465	[M+H]+	768.5567	6.86	2	4.72	0.85	5.33	0.85	5.29	1.02	5.33	88.48	100.00	99.24
PC(18:0/20:3)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	811.6091	[M+H]+	812.6173	6.91	2	6.87	0.43	6.77	0.43	6.79	0.23	6.87	100.00	98.56	98.90
PC(18:0/22:5)	Glycerophosphocholir	Phosphatidylcholines		Diacylglycerophosphocholines [GP0101]	835.6091	[M+H]+	836.6168	6.96	2	5.62	0.60	5.65	0.60	5.67	0.21	5.67	99.10	99.55	100.00
PC(O-16:0/18:1)	Glycerophosphocholir	1-alkyl,2-acylglycero-3-phosphocholines		1-alkyl,2-acylglycerophosphocholines [GP0102]	745.5985	[M+H]+	746.6063	6.97	2	6.19	0.60	6.14	0.60	6.18	0.22	6.19	100.00	99.19	99.87

List of annotated metabolites from region 3 (LIPID positive mode peakpantheR), and relative method induced losses (%)

										Average intensity							Relative percentage		
Compound	**HMDBSubClass**	**HMDBDirectParent**	**LMapsID**	**LMapsSubclass**	**cpdMonoisotopicMZ**	**ion**	**m/z**	**Retention time (min)**	**Region**	**BD-mean**	**BD-%CV**	**Folch-mean**	**Folch-%CV**	**Matyash-mean**	**Matyash-%CV**	**Max intensity**	**BD**	**Folch**	**Matyash**
TG(44:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	748.6581	[M+NH4]+	766.6908	9.87	3	4.24	5.49	4.18	1.79	4.17	2.67	4.24	100.00	98.52	98.16
TG(46:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	774.6737	[M+NH4]+	792.7068	9.91	3	4.45	7.10	4.39	2.51	4.34	3.07	4.45	100.00	98.69	97.70
TG(48:3)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	800.6894	[M+NH4]+	818.7225	9.96	3	4.61	7.37	4.53	2.08	4.50	2.86	4.61	100.00	98.33	97.78
TG(50:4)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	826.7050	[M+NH4]+	844.7384	10.00	3	4.78	7.18	4.73	1.58	4.65	2.91	4.78	100.00	98.92	97.39
TG(52:5)_1	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	852.7207	[M+NH4]+	870.7541	10.05	3	5.17	6.92	5.10	1.85	5.02	2.83	5.17	100.00	98.66	97.12
TG(54:6)_1	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	878.7363	[M+NH4]+	896.7699	10.10	3	4.94	6.76	4.88	1.99	4.79	2.73	4.94	100.00	98.90	96.97
TG(44:0)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	750.6737	[M+NH4]+	768.7061	10.17	3	4.12	6.45	4.01	3.43	4.00	2.87	4.12	100.00	97.32	97.10
TG(46:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	776.6894	[M+NH4]+	794.7227	10.21	3	4.75	6.52	4.64	2.08	4.60	2.76	4.75	100.00	97.82	96.93
TG(48:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	802.7050	[M+NH4]+	820.7384	10.25	3	5.15	6.90	5.06	2.22	4.97	2.76	5.15	100.00	98.26	96.48
TG(50:3)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	828.7207	[M+NH4]+	846.7541	10.29	3	5.60	7.01	5.51	2.20	5.40	2.74	5.60	100.00	98.39	96.40
TG(52:4)_1	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	854.7363	[M+NH4]+	872.7698	10.33	3	6.23	6.89	6.15	1.84	6.02	2.34	6.23	100.00	98.70	96.65
TG(54:5)_1	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	880.7520	[M+NH4]+	898.7852	10.37	3	5.63	7.13	5.57	2.02	5.44	2.48	5.63	100.00	98.82	96.55
TG(47:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	790.7050	[M+NH4]+	808.7356	10.37	3	3.69	8.84	3.62	2.91	3.63	3.77	3.69	100.00	98.04	98.39
TG(49:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	816.7207	[M+NH4]+	834.7535	10.41	3	4.11	9.14	4.12	3.21	3.97	4.18	4.12	99.82	100.00	96.25
TG(51:3)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	842.7363	[M+NH4]+	860.7700	10.44	3	4.47	7.52	4.40	2.70	4.30	2.91	4.47	100.00	98.46	96.16
TG(56:6)_1	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	906.7676	[M+NH4]+	924.8014	10.48	3	4.68	8.37	4.65	2.53	4.50	2.81	4.68	100.00	99.36	96.12
TG(53:4)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	868.7520	[M+NH4]+	886.7849	10.48	3	4.21	8.74	4.16	2.48	4.04	3.29	4.21	100.00	98.97	96.12
TG(46:0)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	778.7050	[M+NH4]+	796.7365	10.50	3	4.00	6.69	3.92	1.93	3.89	1.92	4.00	100.00	98.14	97.34
TG(48:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	804.7207	[M+NH4]+	822.7543	10.53	3	5.12	6.41	5.00	2.02	4.91	2.47	5.12	100.00	97.58	95.77
TG(50:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	830.7363	[M+NH4]+	848.7698	10.56	3	5.82	6.37	5.69	2.05	5.57	2.31	5.82	100.00	97.71	95.77
TG(52:3)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	856.7520	[M+NH4]+	874.7863	10.60	3	5.48	7.50	5.42	2.02	5.43	5.26	5.48	100.00	99.01	99.17
TG(54:4)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	882.7676	[M+NH4]+	900.8014	10.64	3	5.86	7.49	5.77	2.21	5.61	2.58	5.86	100.00	98.52	95.74
TG(49:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	818.7363	[M+NH4]+	836.7694	10.68	3	4.31	9.82	4.29	2.67	4.16	3.30	4.31	100.00	99.59	96.64
TG(51:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	844.7520	[M+NH4]+	862.7856	10.71	3	4.68	7.96	4.62	2.54	4.45	3.26	4.68	100.00	98.70	95.23
TG(53:3)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	870.7676	[M+NH4]+	888.8012	10.74	3	4.49	7.94	4.44	2.72	4.31	2.78	4.49	100.00	98.79	95.88
TG(48:0)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	806.7363	[M+NH4]+	824.7689	10.80	3	4.16	5.76	4.04	1.65	4.04	1.43	4.16	100.00	96.91	97.10
TG(50:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	832.7520	[M+NH4]+	850.7853	10.82	3	5.57	6.55	5.42	2.26	5.30	2.17	5.57	100.00	97.31	95.19
TG(52:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	858.7676	[M+NH4]+	876.8021	10.85	3	6.22	6.54	6.07	2.11	5.92	2.27	6.22	100.00	97.69	95.22
TG(54:3)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	884.7833	[M+NH4]+	902.8172	10.89	3	5.68	6.75	5.55	2.12	5.40	2.26	5.68	100.00	97.76	95.09
TG(53:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	872.7833	[M+NH4]+	890.8170	10.99	3	4.31	8.25	4.24	2.71	4.07	3.23	4.31	100.00	98.19	94.27
TG(52:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	860.7833	[M+NH4]+	878.8171	11.10	3	4.83	7.04	4.70	2.30	4.58	2.15	4.83	100.00	97.32	94.80
TG(54:2)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	886.7989	[M+NH4]+	904.8328	11.13	3	4.97	7.36	4.87	2.55	4.70	2.51	4.97	100.00	97.96	94.55
TG(54:1)	Triradylcglycerols	Triacylglycerols		Triacylglycerols [GL0301]	888.8146	[M+NH4]+	906.8483	11.36	3	4.05	14.04	4.16	3.25	3.89	7.13	4.16	97.41	100.00	93.52

Sample preparation method for phospholipid removal using 96-well SPE plates

96-well SPE extraction plates designed for phospholipid removal. These include, Water's OSTRO, Biotage's ISOLUTE and Phenomenex's PHREE. In addition to these extractions, Phenomenex also packed the Sepra C18 material in a 96-well SPE format (Sepra-SPE). The extraction conditions and sample volumes are summarised in the table below. All plates were washed with their specific extraction solvents three times. Then samples are loaded into the wells, followed by their solvents. A series of aspiration and dispensing was undertaken to mix the sample (total of three times). The extraction plate was then placed above a collection plate and placed on the vacuum manifold for 5-10 minutes at 15" Hg of vacuum.

COMPANY	NAME	Sample volume (uL)	Solvent	Solvent (uL)
Waters	OSTRO	100	MECN+0.1%FA	300
Phenomenex	PHREE	100	MECN	300
Biotage	ISOLUTE	100	MECN	300

Appendix 3: Chapter 5

Sygnature discovery experimental conditions

Incubations

Hepatocytes: Triclosan (10µM) was incubated with cryopreserved human hepatocytes at 0.5 million cells/ml in a total volume of 330 µl Leibovitz buffer at 37^{O}C for 60 minutes. Negative control incubations in buffer without hepatocytes, and in hepatocytes without compound were performed in parallel to eliminate peaks that are not compound related or to identify any transformations that are non-metabolic.

Microsomes: Triclosan (10µM and 50 µM) was incubated with human liver microsomes at 0.5 mg/ml with NADPH (1mM) in a total volume of 500 µl phosphate-buffered saline at 37^{O}C for 60 minutes. A negative control was carried out in parallel without NADPH.

Aliquots were taken from incubations in both hepatocytes and microsomes at 0 and 60 minutes and quenched 1:1 with 3% formic acid in acetonitrile.

Sample Analysis

Samples were analysed using a Waters Xevo-G2-XS-TOF mass spectrometer with Waters Acquity UPLC system. Full scan spectra were acquired using an MS^{E} method in negative ion mode.

Chromatography was performed on an ACQUITY UPLC HSS C18 Column, (1.8 µm, 2.1 mm X 100 mm) with a solvent flow of 0.45 ml min^{-1} and a column temperature of 40^{O}C. An ammonia (0.002%) / acetonitrile gradient with 0.1% formic acid was run (Table A2).

Table A1: MS Conditions

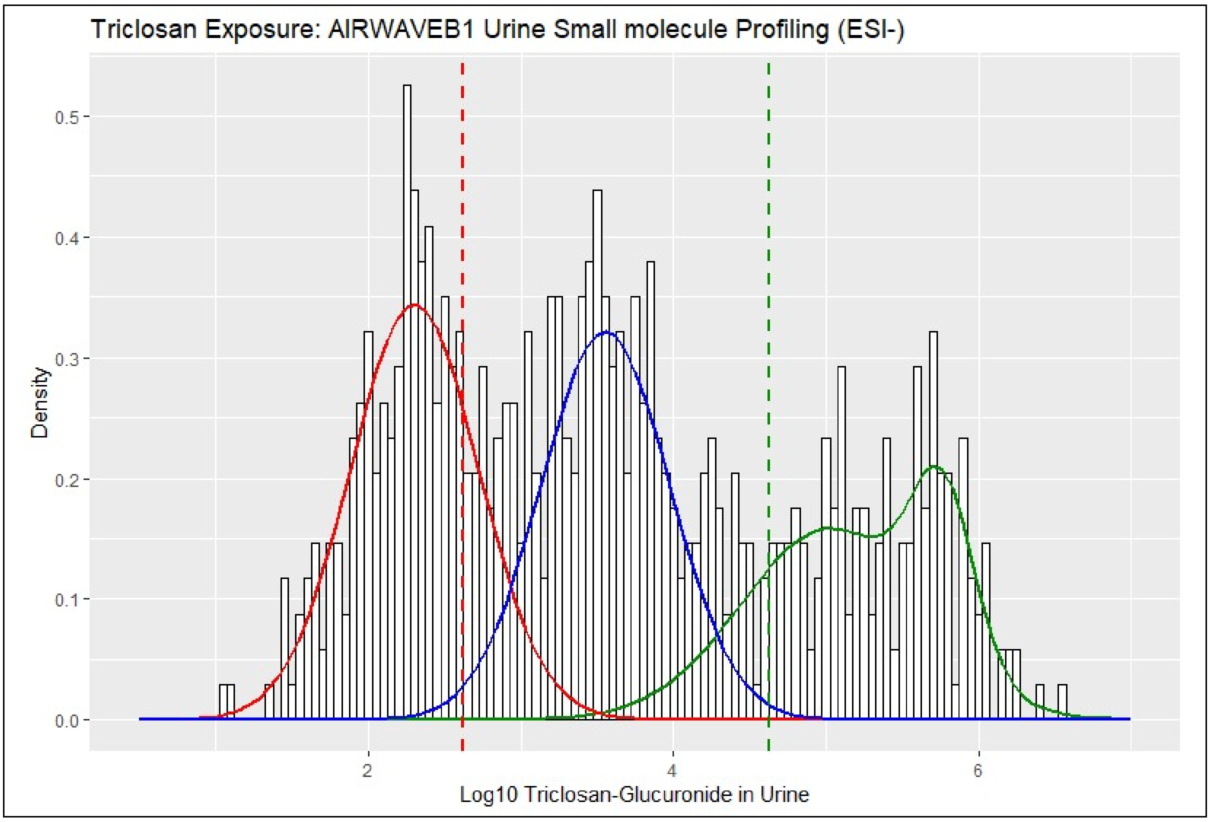
Capillary Voltage	2 kV
Cone Voltage	25 V
Collision Energy	20-50 V ramped (high energy MS^{E} mode)
Lock mass	Leucine Enkephalin (200 pg/mL @ 10 µL/min)

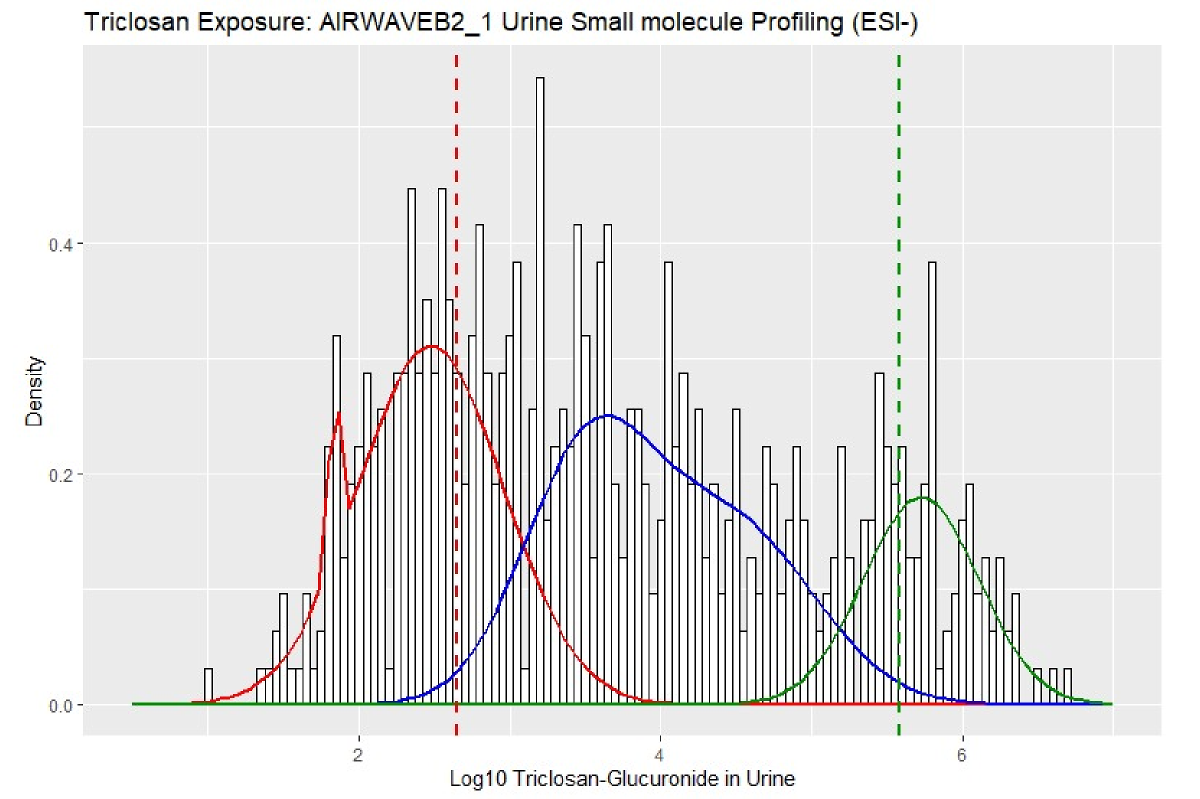
Table A2: LC Conditions

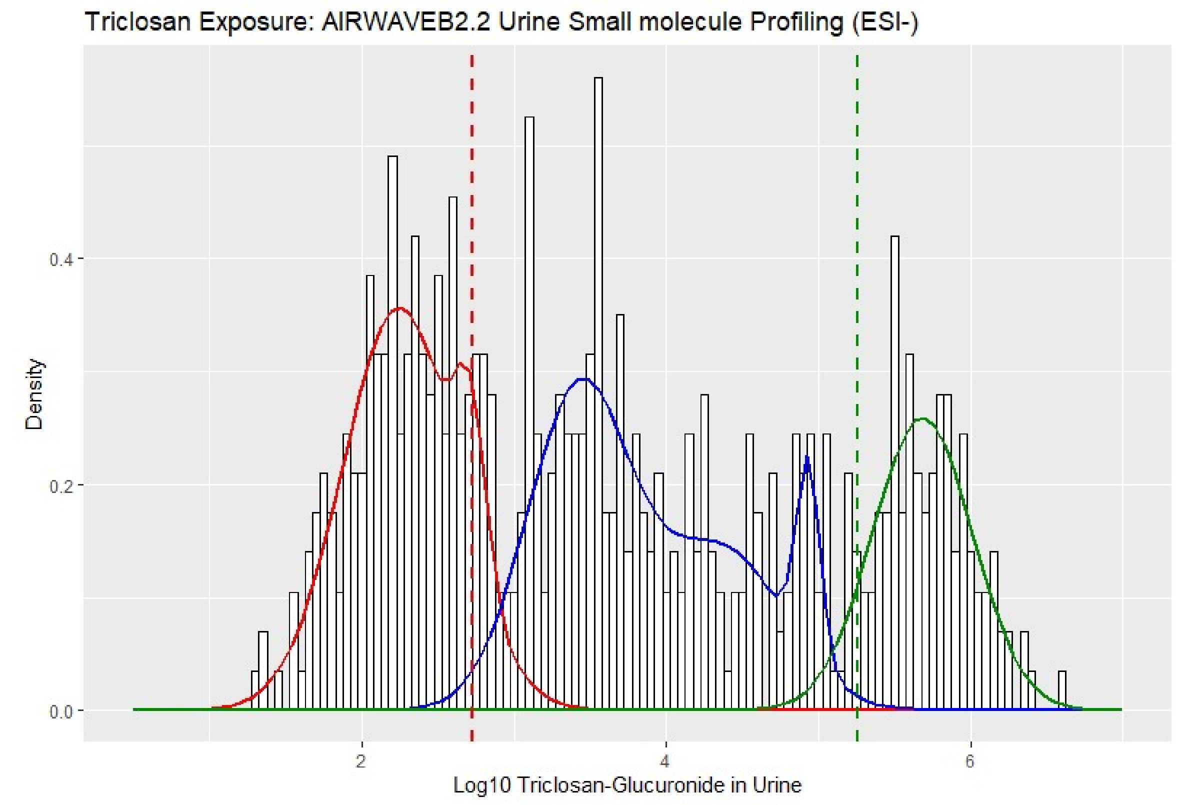
Time	% Acetonitrile
0.20	5

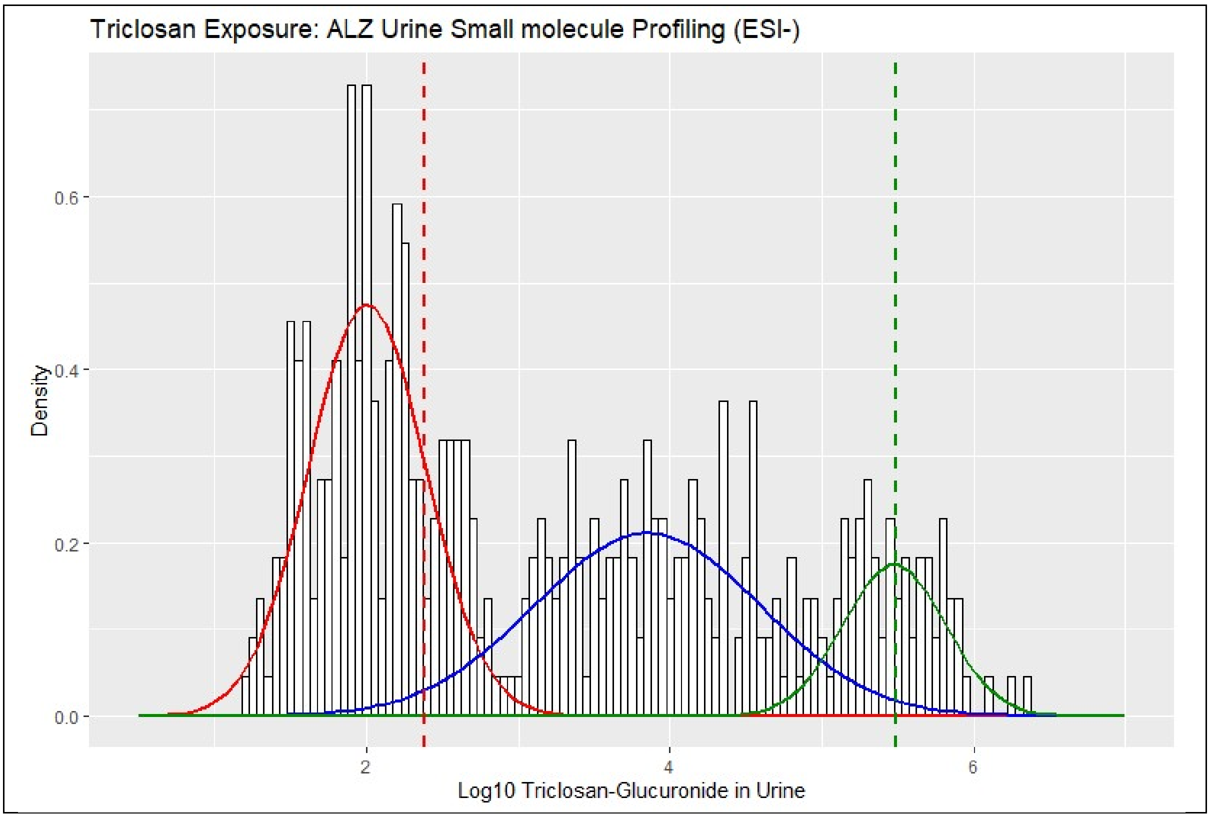
11.00	95
13.00	95
14.00	5
15.00	5

Gaussian mixture models (GMM's) fitted to the MS intensity distribution of TCS-Gluc from AW and ALZ cohorts



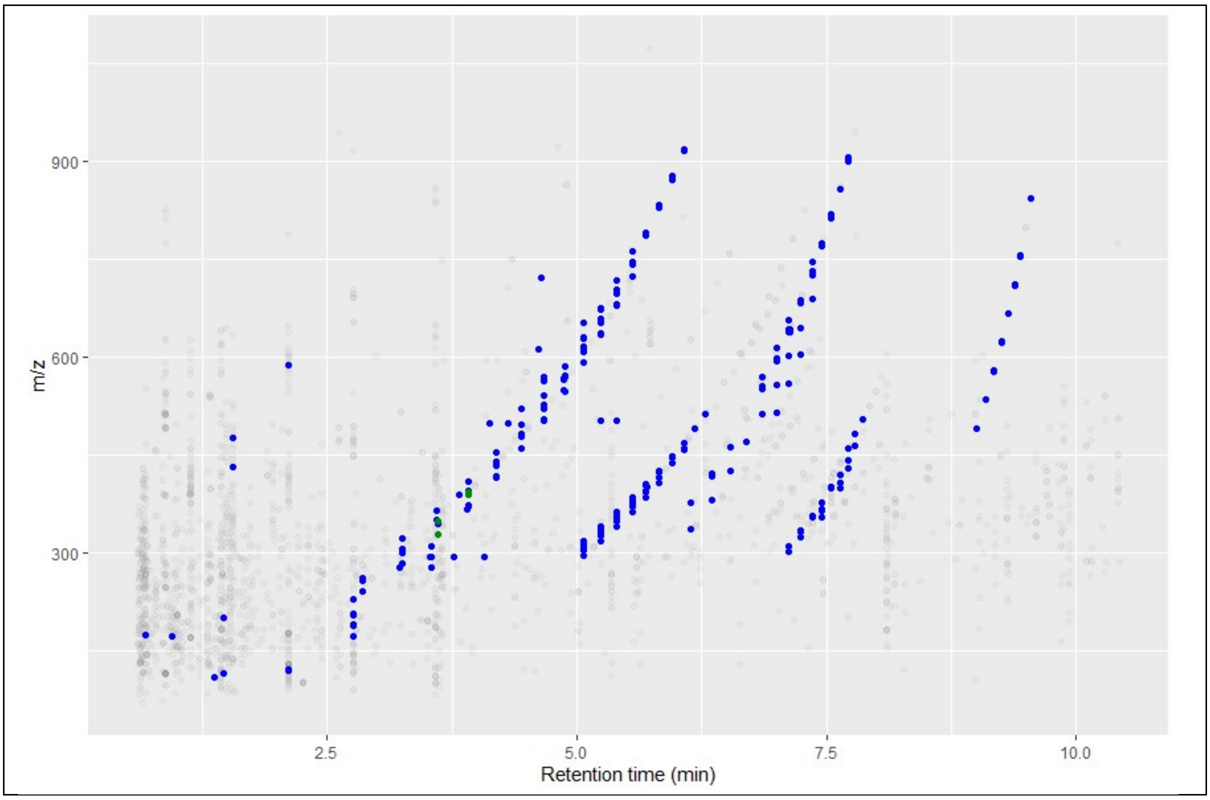

Triclosan Exposure: AIRWAVEB2_1 Urine Small molecule Profiling (ESI-)
Density
0.4
0.2
0.0
2
4
6
Log10 Triclosan-Glucuronide in Urine


Triclosan Exposure: AIRWAVEB2.2 Urine Small molecule Profiling (ESI-)
Density
0.4
0.2
0.0
2
4
6
Log10 Triclosan-Glucuronide in Urine



GMM’s were fitted to each cohort, and the PDF’s for each gaussians were obtained, dividing the data to a High exposure group (Distribution 2, pr3 - green), Low-Mid exposure group (Distribution 2, pr2 - blue) and a Zero exposure group, (Distribution 1, pr1 -red). Any sample with pr1>0.90, or a log10 signal less than the red dotted line, assumed the classification of zero exposure, and any sample with pr3>0.90, or log10 signal more than the dotted green line, assumed the classification of high exposure. The blue dotted line is equivalent to a signal that fulfils all three criteria as stated in the Prevalence section.

Between-sample correlation using PEG(n8)-unmetabolised as the driver, observed in the ALZ serum samples which were removed as PEG contaminated samples.



Results from the between-sample correlation were presented as a retention time (RT) vs *m/z* plot, where features that correlated to the driver feature (PEG(n8)-unmetabolised), were coloured by statistical significance ($p_{adj} \leq 0.05$)- blue, and a statistically significant correlation coefficient greater than 0.7 (i.e. Spearman: r >0.7 and $p_{adj} \leq 0.05$) as green. All other detected features are coloured grey. Correlated features included ionisation products of PEG(n8)-unmetabolised only, Features corresponding to acid (PEG(n8)-COOH) and diacid (PEG(n8)-2xCOOH) metabolites and three unknown feature groups, are not present.

Appendix 4: Publications and conference presentations

Publications

LEWIS, M., PEARCE, J. T. M., SPAGOU, K., GREEN, M., DONA, A., YUEN, A. H. Y., **DAVID, M**., BERRY, D. J., CHAPPELL, K., SLUIS, V., SHAW, R., LOVESTONE, S., ELLIOTT, P., SHOCKCOR, J., LINDON, J. C., CLOAREC, O., TAKATS, Z., HOLMES, E. & NICHOLSON, J. 2016a. Development and Application of Ultra-Performance Liquid Chromatography-TOF MS for Precision Large Scale Urinary Metabolic Phenotyping. Analytical Chemistry, 88.

GAFSON, A. R., SAVVA, C., THORNE, T., **DAVID, M.,** GOMEZ-ROMERO, M., LEWIS, M. R., NICHOLAS, R., HESLEGRAVE, A., ZETTERBERG, H. & MATTHEWS, P. M. 2019. Breaking the cycle: Reversal of flux in the tricarboxylic acid cycle by dimethyl fumarate. Neurol Neuroimmunol Neuroinflamm, 6, e562

BIDNY, S., GAGO, K., **DAVID, M**., DUONG, T., ALBERTYN, D. & GUNJA, N. 2015. A validated LC-MS-MS method for simultaneous identification and quantitation of rodenticides in blood. J Anal Toxicol, 39, 219-24.

Conference

DMG early career meeting 2018 – Oral presentation, "Drug metabolism in populations – profiling the human xenometbaolome". (**Chapter 4**)

MSACL 2020 poster presentation – "Enhanced mass spectrometric profiling of the human blood exposome using an optimised dispersive SPE protocol". (**Chapter 4**)