University of London
Imperial College London
Biomolecular Medicine

Novel Computational Approaches to Characterising Metabolic Responses to Toxicity *via* an NMR-Based Metabonomic Database

Jake Thomas Midwinter Pearce

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computational Biology of the University of London
and the Diploma of Imperial College, February 2010

Abstract

This thesis develops and assesses computational approaches for handling large quantities of metabolic-profile data, with the aim of producing techniques that maximise data-extraction, while minimising subjective intervention. Metabolic-profiling studies of toxicity have thus far concentrated on single studies, often data preparation and processing in these studies has required many subjective judgements by the analyst. One of the key objectives of this work is the removal of this subjectivity, and thus potential bias. Here, this is exemplified by a search in serum for metabolic biomarkers of exposure to liver toxins.

Data were taken from COMET, a collaboration between pharmaceutical companies and Imperial College. COMET generated an array of 146 toxicological studies in the rat and mouse. This thesis deals with ^{1}H-NMR spectroscopic profiles of rat serum, consisting of 3,780 samples, in 126 studies, in combination with clinical chemistry and associated metadata.

Initial work consisted of designing and implementing a database to store these metadata, allowing the auditing and sorting of samples across the project. This was accompanied by the development of novel approaches to data standardisation, that removed subjectivity where possible. To ensure shift alignment of serum NMR spectra, a method of chemical-shift referencing to glucose was developed, based on the derivatisation of spectra; assessment of existing techniques for intensity normalisation lead to the development of a novel method of apportionment-entropy based normalisation; to correct for variations between profiles not rectified by shift referencing or global intensity normalisation, an orthogonal-filtering technique was developed, that attempted to identify and remove systematic variation, overlapping a subset of signals in a collection of spectra.

These data were then modelled as a whole, to assess metabolite-profile based estimates of exposure to liver-toxins. Several permutations of discriminant PLS models were generated and compared to each other, and to the current standard in serum-based clinical-chemistry, the ALT level. The strongest of these models was found to exceed the sensitivity of ALT to exposure, and is seen to model decreases in serum lipids, which likely represent a generic perturbation to lipoprotein metabolism resulting from toxic insult.

Acknowledgements

I wish to thank my supervisors, Dr Hector Keun and Professor Jeremy Nicholson for their support during the preparation of this thesis.

My thanks must also be extended to all those involved in the COMET project, and in particular, Olaf Beckonert, Mary Bollard and Tim Ebbels, without whom's hard work, I would have had no data!

My friends Mat, Tasha, Dan & Laura have given me great support, as have my colleagues at Imperial, particularly Alex, Jimmy, Rachel & Orla from the HK group, as well as Olivier, Tsz, Richard, Volker & Claire. Toby deserves especial thanks for having to deal with me both in and out of work!

Finally, my family have supported me fantastically throughout my various academic wanderings, Mum, Dad & 'O, Thank you!

Contents

List of Tables 12

List of Figures 14

Abbreviations 15

1 Introduction 17

1.1 Aims and Objectives . . . 17
1.2 Toxicological Screening in Safety Assessment . . . 17
1.3 Metabolic-Profiling . . . 19
1.3.1 Metabolite profiling of serum . . . 20
1.3.2 Metabolite profiling in preclinical toxicological screening . . . 21
1.4 Nuclear Magnetic Resonance Spectroscopy . . . 22
1.5 Chemometric Methods for Data Analysis . . . 24
1.5.1 Chemometrics in metabolite profiling . . . 25
1.5.2 Principal component analysis . . . 27
1.5.3 Regression models . . . 30
1.5.4 Partial least-squares regression . . . 31
1.5.5 Orthogonal filtering and O2-PLS . . . 32
1.5.6 Validation of PLS models . . . 34
1.6 The Receiver-Operating Characteristic . . . 34
1.7 The COMET Project and Dataset . . . 35
1.7.1 Study design . . . 36
1.7.2 Sample collection . . . 37

2 The COMET Database 39

2.1 Aims for the Database . . . 39
2.2 Database Concepts . . . 40
2.2.1 Database terminology . . . 40
2.2.2 Introduction to structured query language . . . 41
2.2.3 Selecting a DBMS . . . 41
2.3 The COMET Project . . . 41
2.3.1 Extent of the COMET project . . . 41
2.3.2 Design of the COMET data tables . . . 42

2.4 Example Queries on the COMET Database . . . 44
2.4.1 Single table queries . . . 44
2.4.2 Multiple table queries . . . 47

3 Automated Calibration of Serum NMR 49
3.1 NMR Calibration . . . 49
3.1.1 Factors effecting chemical shift . . . 49
3.1.2 Difficulties with referencing to TSP in serum spectra . . . 50
3.1.3 Calibration to glucose . . . 50
3.2 Methods and Algorithms . . . 51
3.3 Results and Discussion . . . 54
3.3.1 Comparison of calibration methods . . . 54
3.3.2 Effect of signal-to-noise ratio . . . 55
3.4 Conclusion . . . 57

4 Normalisation 59
4.1 Aims of Normalisation . . . 59
4.1.1 Established methods . . . 60
4.1.2 Entropy and statistical models . . . 63
4.1.3 Apportionment-entropy based normalisation . . . 64
4.2 Comparison of Normalisation Methods . . . 65
4.2.1 Simulated NMR spectra . . . 65
4.2.2 Experimental NMR spectra . . . 68
4.3 Apportionment-Entropy Profiles . . . 70
4.3.1 Apportionment-entropy can act as an indicator of spectral similarity . . . 72
4.4 Discussion . . . 73

5 Orthogonal Filtering for Relative Quantification 75
5.1 NMR Quantification in Biofluids . . . 75
5.1.1 Existing methods of quantification and their disadvantages . . . 75
5.2 Orthogonal Filtering . . . 77
5.2.1 Spectral post-processing . . . 77
5.2.2 Application to metabolic profiling data-sets . . . 78
5.2.3 Orthogonal filtering for spectral quantification . . . 78
5.2.4 Implementation of OFSQ . . . 79
5.2.5 Generation of simulated data . . . 80
5.3 Testing and Validation of OFSQ . . . 81
5.3.1 Performance on simulated data . . . 81
5.3.2 The effect of correlated variation on OFSQ . . . 82
5.3.3 Selection of the number of orthogonal components in OFSQ models . . . 84

5.3.4 The diagnostic measures are affected by data scaling . . . 86
5.3.5 Application of OFSQ to experimental data . . . 86
5.3.6 Examination of the orthogonal loadings can indicate the source of interference . . . 87
5.4 Utility of Orthogonal Filtering . . . 87

6 Detecting the Effect of Liver Toxins by Serum NMR 91
6.1 Estimation of Liver Injury from Serum Sampling . . . 91
6.1.1 Rationale for a metabolite profiling approach . . . 93
6.2 Data Analysis Methods . . . 95
6.2.1 Standardisation of NMR spectra . . . 95
6.2.2 Collation of clinical chemistry data . . . 96
6.2.3 Variable-removal strategies . . . 97
6.2.4 Predictive PLS modelling . . . 99
6.3 Analysis . . . 104
6.3.1 Assessment of predictive models . . . 104
6.3.2 The most predictive models make use of the entire spectrum . . . 105
6.3.3 Performance of the strongest model on the entire COMET dataset . . . 106
6.3.4 MPLE scores provides complementary information to ALT . . . 109
6.3.5 Effect of dietary manipulation on MPLE . . . 111
6.3.6 Metabolite identification . . . 112
6.4 Discussion . . . 114
6.4.1 MPLE scores provide information on liver function . . . 114
6.4.2 MPLE scores are not directly related to lipid clinical-chemistry . . . 116
6.4.3 Conclusion . . . 117

7 Discussion 119
7.1 Concluding Remarks . . . 119
7.1.1 The importance of data standardisation . . . 119
7.1.2 Mining the COMET data . . . 121
7.2 Wider Scope . . . 122
7.3 Conclusion . . . 124

A COMET Study Details 125

B Experimental Methods 133
B.1 COMET metadata collection . . . 133
B.2 Acquisition of COMET spectra . . . 133
B.3 Human blood-serum spectrum acquisition . . . 134
B.4 Human urine spectrum acquisition . . . 134

C Database Implementation 137
C.1 Database implementation . . . 137
C.2 MySQL data types . . . 137
C.3 Core table definitions . . . 138
C.4 Supplementary table definitions . . . 141
C.5 Queries . . . 141

D Data Tables 143

Bibliography 151

List of Tables

2.1 Listing of all records from the study table . . . 45
2.2 Listing of selected records from the study table . . . 45
2.3 Using text wildcards . . . 46
2.4 Example of attribute selection . . . 46
2.5 Example of the `distinct` operator . . . 47
2.6 Combining attribute and record selection . . . 47
2.7 Export of model species and strains by company . . . 48

4.1 Statistics for PLS regression of COMET data to glucose . . . 68
4.2 Statistics for PLS regression of urine spectra to age . . . 69

5.1 OFSQ applied to serum spectra . . . 86

6.1 MFC-normalised CPMG PLS-DA model statistics . . . 100
6.2 Entropy-normalised CPMG PLS-DA model statistics . . . 100
6.3 NOESYPR1D PLS-DA model statistics . . . 100

A.1 COMET study details . . . 126

C.1 Study table definition . . . 138
C.2 Animal table definition . . . 138
C.3 Histopathology table definition . . . 139
C.4 PR project table definition . . . 139
C.5 PR model table definition . . . 139
C.6 Urine sample table definition . . . 139
C.7 Serum sample table definition . . . 140
C.8 Tissue sample table definition . . . 140
C.9 Outliers sample table definition . . . 141
C.10 Company table definition . . . 141
C.11 Person table definition . . . 141

D.1 MFC-normalised CPMG variable sets . . . 143
D.2 Entropy-normalised CPMG variable sets . . . 143
D.3 NOESYPR1D variable sets . . . 144
D.4 ROC statistics, CPMG, liver vs controls . . . 145

D.5 ROC statistics, CPMG, liver vs other 146

List of Figures

1.1 Schematic outline of this thesis . . . 17
1.2 Energy levels in NMR . . . 22
1.3 Glossary of an NMR spectrum . . . 23
1.4 Acquisition of an NMR spectrum . . . 23
1.5 Resolution reduction of an NMR spectrum . . . 26
1.6 Geometric overview of PCA modelling . . . 28
1.7 Cross-validation of data . . . 29
1.8 O2-PLS modelling . . . 33
1.9 Outline of the O2-PLS model . . . 33
1.10 The sensitivity-specificity trade-off . . . 34
1.11 ROC basics . . . 35
1.12 Standard COMET sampling time-course . . . 37

2.1 Overview of a database relation . . . 40
2.2 An example ERD . . . 40
2.3 COMET database ERD . . . 43

3.1 Subset of unaligned spectra . . . 50
3.2 Use of spectral derivatisation to magnify sharp features . . . 52
3.3 Comparison of calibration method correlations . . . 54
3.4 Comparison of calibration methods . . . 55
3.5 Effect of signal-to-noise on derivative spectra . . . 56

4.1 Comparison of normalisations . . . 60
4.2 Schematic of normalisation by apportionment-entropy . . . 64
4.3 Simulated spectra . . . 66
4.4 Normalisation of simulated spectra . . . 67
4.5 Simple simulated spectra . . . 71
4.6 Entropy profile of simulated spectra . . . 71
4.7 Entropy profile of NMR spectra . . . 72
4.8 Distinguishing NMR pulse-sequences by apportionment-entropy . . . 72

5.1 Comparison of CPMG and NOESYPR1D spectra . . . 76
5.2 Generation of a simulated spectrum . . . 80

5.3 Example simulated spectra . . . 81
5.4 Improvement in modeling following OFSQ . . . 81
5.5 Example simulated data after OFSQ . . . 81
5.6 Simulated data peak-correlation comparison (UV) . . . 83
5.7 Simulated data peak-correlation comparison (MC) . . . 83
5.8 Effect of intensity of overlapping peaks . . . 84
5.9 OFSQ diagnostics . . . 85
5.10 Score on the orthogonal loadings for real data . . . 88

6.1 Distribution of serum ALT . . . 92
6.2 Diagram of the analytical process . . . 94
6.3 Determination of noise level . . . 98
6.4 ROCs of MFC-normalised CPMG models . . . 101
6.5 ROCs of entropy-normalised CPMG models . . . 102
6.6 ROCs of NOESYPR1D models . . . 103
6.7 Best ROC curves distinguishing dosed from controls . . . 104
6.8 Best ROC curves distinguishing liver dosed from others . . . 105
6.9 Study-wise breakdown of classifier scores . . . 107
6.10 Tally of metric performance . . . 108
6.11 Scores and ALT for isothiocyanate compounds . . . 109
6.12 Bilirubin response to isothiocyanates . . . 109
6.13 Scores and ALT for compounds sensitive to ALT . . . 110
6.14 Comparison of histopathology to measured metrics . . . 111
6.15 Scores and ALT for dietary modifications . . . 112
6.16 Assignments of NMR resonances . . . 113
6.17 Lipoprotein metabolism . . . 115
6.18 Lipid levels in D2O . . . 116

D.1 Breakdown of classifier scores for liver toxins . . . 147
D.2 Breakdown of classifier scores for kidney & testes toxins . . . 148
D.3 Breakdown of classifier scores for other stressors . . . 149

Abbreviations & Notations

Common abbreviations and notations used throughout this thesis. Where relevant, additional terms are defined in the text.

Abbreviations:

ALT	Alanine transaminase.	NOESYPR1D	One-dimensional water-suppressed pulse-sequence, see § B.2.
AUC	Area under the curve.	NMR	Nuclear magnetic resonance.
COMET	Consortium for metabonomic toxicology.	OFSQ	Orthogonal filtering for spectral quantification.
CPMG	Carr-Purcell-Meiboom-Gill pulse-sequence, see § B.2.	OSC	Orthogonal signal correction.
DBMS	Database management system.	PCA	Principal component analysis.
δ_n	Chemical shift for nucleus n.	PLS	Partial least-squares.
ERD	Entity-relationship diagram.	PPM	Parts-per-million.
f	Figure.	PR	Pattern recognition.
FDR	False discovery rate.	ROC	Receiver-operating characteristic.
FID	Free induction decay.	RMSE	Root-mean-squared error.
HDL	High-density lipoprotein.	§	Section.
LDL	Low-density lipoprotein.	SQL	Structured query language.
MC	Mean centring.	TSP	3-(trimethylsilyl)-propionic acid-d_4.
MFC	Median fold-change.	UV	Unit-variance.
MPLE	Metabolite profile of liver exposure.	VLDL	Very-low-density lipoprotein.
MS	Mass-spectrometry.		

Notations:

Matrices are set UPPERCASE, vectors lowercase, and scalar values *italic*.

X	Matrix of descriptor variables $[n \times j]$.	$\dot{\mathsf{X}}$	X mean centred $[n \times j]$.
Y	Matrix of response variables $[n \times k]$.	a	Number of components [scalar].
X^T	Matrix transpose of X $[j \times n]$.	T	Matrix of scores $[n \times a]$.
E	Matrix of residuals on X $[n \times j]$.	P	Matrix of loadings $[j \times a]$.
F	Matrix of residuals on Y $[n \times k]$.	W	Matrix of weights $[j \times a]$.
$\tilde{\mathsf{X}}$	Median of X $[1 \times j]$.		

Chapter 1

Introduction

1.1 Aims and Objectives

Metabolic-profiling techniques have proved an invaluable approach to scientists seeking to understand the metabolic status of an organism, and to use this information to comprehend and predict the consequences of any intervention. To date, the statistical-analysis conducted on these types of data have tended to focus on individual studies, rather than large agglomerations of data.

With the advent of large repositories of metabolite data, such as the COMET dataset discussed here, techniques for the analysis of large data sets become more pertinent, and present significant challenges.

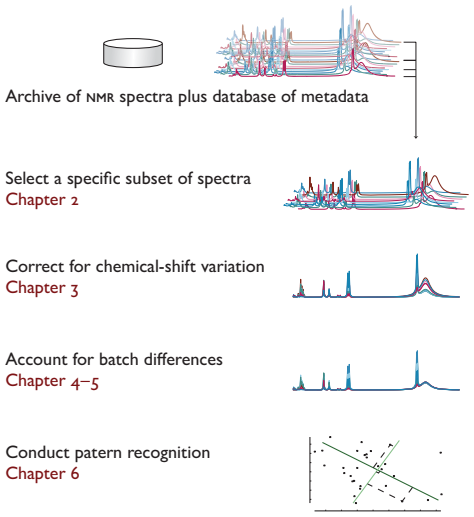

Figure 1.1: The analysis of a large metabolic-profiling data-set can be divided into a number of conceptual steps. Each step attempts to ensure the greatest possible uniformity in the data, while requiring the minimum subjective input.

There are many reasons for this, resulting from complicating factors that include; batch effects from separate samplings, whether the separation is spatial or temporal; Stages of data preparation are still subjective, and as the size of data-set increases beyond that easily manageable by one individual, this subjectivity can negatively affect analysis of data processed by more than one individual; The computational load of dealing with large volumes of data has restricted the simultaneous modelling of these data, limiting the potential analyses.

This work will examine the techniques, their limitations and potential, that pertain to the cross-comparison of a large collection of metabolic-profiling studies.

1.2 Toxicological Screening in Safety Assessment

Prior to the introduction of any novel chemical, whether a medical treatment, food additive, or industrial, agricultural or consumer product, toxicological screening is required to assess

the potential for adverse side-effects. As an attempt to understand and predict the effect of any putative compound on the human body as a whole, the heart of any toxicological study remains the use of animal models. Comparative inter-species studies in drug development have thus become a tenet of regulatory acceptance, both in Europe and the United States (Doull *et al.*, 1980).

The traditional format for studies in toxicological screening, the absorption, distribution, metabolism, and excretion (ADME) study, has evolved to extract the maximum information from the fewest animals over the shortest period of time. A typical ADME study doses the putative compound as a single acute dose, then observing the study animals for adverse affects, both physiological, behavioural and defined clinical-chemical markers (Hayes, 1989).

Acute *in vivo* toxicological studies have the advantage of allowing tight control over the experimental conditions, and allow a wide range of outcomes to be measured and assessed. However, they have the disadvantage of being expensive to conduct, and often dose compounds at far higher levels than would be expected of exposure in the wild, with implications for the validity of the results outside the laboratory. Additionally, as new modes of toxicity have been identified, further complications have been added to the screening process, such as the need to study the offspring of exposed mothers.

While such *in vivo* inter-species comparisons provide the current best method of assessing toxicity in other species, care must be taken to avoid making simple assumptions, a common example being that a basic scaling factor will allow the transfer of dosing levels from one species to the next (Jolyon West *et al.*, 1962). Indeed, the naivety of such an approach is widely acknowledged (Harwood, 1963; Boxenbaum, 1984) and any attempt to transfer dosing information from one species to another will inevitably require a range-finding exercise to locate the level at which the pharmacological effect is comparable.

As an attempt to control these factors, modern toxicological screening is beginning to take advantage of a far wider range of techniques. In their 2007 report, 'Toxicity Testing in the 21st Century: A Vision and a Strategy', the National Research Council of the USA, laid out a vision for the future of toxicological screening, using modern technology to both quicken and reduce the cost of screening, while enhancing the quality of data extracted.

The National Research Council present four options for the future development of screening techniques, ranging from a continuation of the *status quo*, to a complete re-assessment of the nature of such studies (Gibb, 2008; Krewski *et al.*, 2009). The report suggests the tiered introduction of high throughput, *in vitro* and *in silico* methodologies (Aardema & MacGregor, 2002) to rapidly screen compounds for harmful interactions with a number of 'toxicity pathways'. Such methods have the potential to assess toxicity not only in terms of overt injury, resulting from exposure to high levels of a compound, but also predict the affect of low-levels of exposure, by responding to modulations in gene-expression and metabolism of key pathways involved in toxic response.

A similar movement in the European Union has resulted in COST Action B15 as reviewed in

Gundert-Remy *et al.* (2005). This work posits the expansion and further systemisation of the uses and validation of biomarkers in clinical diagnosis, drug development, and toxicology.

These integrated approaches are often banded together under the name 'systems-biology'. A holistic approach that integrates a wide range of data, systems biology understands biology in terms of the interactions and fluxes within biological networks, taking the physiological, genetic, transcriptomic, proteomic and metabolite profile of the organism into account when assessing the effect of a xenobiotic (Kitano, 2002; Nicholson & Lindon, 2008).

1.3 Metabolic-Profiling

The technique of metabonomics has been defined by Nicholson *et al.* (1999) as:

> 'The quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification'

This definition is closely related to that of metabonomics, and both may be more generically referred to as *metabolite-profiling*, the term I will be using in this thesis.

Metabolite profiling observes the effect of a pharmacologically active xenobiotic, or any other stimulus, on an organism's metabolism by monitoring the changing concentrations of small metabolites in the organism's biofluids. The source of these biofluids may be any biological matrix, but is commonly urine, blood-plasma or -serum, whole tissues → or tissue extracts, depending on the analytical method used and the information desired.

Via magic-angle spinning NMR (MAS NMR).

Many studies have shown that the dynamics of the metabolic perturbations following an intervention, provides a valuable insight into the progression and mechanism of the perturbation (Holmes *et al.*, 1995; Bollard *et al.*, 2009) →.

See also CLOUDS and geometric trajectory analysis in § 1.7.

Metabolic profiling has an advantage in inter-species comparisons, as each metabolite is a clearly defined chemical entity, meaning observations can be more readily transferred between species than other —omics, where by definition, genomic and therefore transcriptomic and proteomic profiles will vary. Metabolite profiling also gains advantage as changes in metabolism are a phenotypic, and often functional, endpoint. Additionally, much of the interaction of an organism with its environment comes in the form of small molecules, most notably from the diet.

Typically, metabolite profiles have been generated *via* Nuclear Magnetic Resonance (NMR, see § 1.4) or Mass Spectrometry (MS). Both techniques are amenable to undirected profiling of a broad spectrum of molecules, generating quantitative measurements across a wide dynamic range. Both are also capable of resolving isotopologues where isotopes have been used to label metabolites to allow the tracking of their metabolic fate.

Of the two techniques, NMR has the advantage of providing an unbiased overview of all the small molecules in a solution, with minimal sample preparation, non-destructively and relatively quickly. This allows NMR to return to a sample and perform more complicated ex-

periments, such as 2D acquisitions, that can be used to identify unknown resonances. NMR is also more reproducible across platforms (Keun *et al.*, 2002b).

By comparison, MS has the disadvantage of destroying the sample as it is run, and typically requires a chromatographic separation, such as High- or Ultra-Performance Liquid Chromatography and Gas Chromatography (HPLC / UPLC / GC) to separate molecules as they enter the MS, to prevent unpredictable interactions between molecules affecting the levels of ionisation, that may in turn bias quantification, and some molecules may be resistant to ionisation entirely. However, MS is more sensitive than NMR by several orders of magnitude, and thus requires smaller sample sizes, typically in the 50 µL range for a rat urine sample ←. Modern mass-spectrometers are also capable of sophisticated analytical techniques, such as MS-MS which allows fragmentation information to be collected in concert with mass-to-charge ratios for whole ions, enhancing the ability to identify molecules in situations were the exact-mass and retention-time are insufficient.

NMR would usually require sample volumes of at least 200 µL for a urine sample, and even then could not match the sensitivity of MS.

1.3.1 Metabolite profiling of serum

The use of serum in general multiparametric metabolite profiling has been less well developed than urine, due to several complications related to the nature of the biofluid. Compared to urine, blood is a far more complex mixture, and even following removal of the cellular component by clotting and centrifugation, there remain many chemical species that have a strong effect on the spectra. Most notably in NMR, lipids-aggregates and proteins produce a large number of broad resonances that can dominate the spectrum ←. More fundamentally, whereas urine serves to export polar metabolites from the body, and thus differs markedly in composition depending on metabolic circumstance, the composition of blood is under tight homeostatic control, and therefore any metabolic perturbations are accordingly subtle.

See § 5.1.1 for further discussion of these effects.

However, useful information can be extracted from blood-serum by NMR. Sampling blood provides a useful counterpoint to urine sampling, as urine is typically pooled across a time period, while blood is (effectively) sampled instantaneously. This means while urine profiles represent an average of metabolite excretion over the pooled period, blood profiles represent the metabolic state of the organism at the instant of sampling.

The first blood-serum analysis by NMR was conducted by Bock (1982), as an extension of clinical-chemistry measurements. The initial rationale for NMR analysis of serum was to allow the simultaneous and non-destructive quantification of several compounds. Bock also demonstrated the ability to make novel diagnoses from NMR spectra, allowing the source of acidosis in a patient to be attributed to an excess of D-lactate, a stereoisomer that could not be detected by the enzymatic assay for L-lactate. Further use of NMR spectra of serum followed this approach (Nicholson *et al.*, 1983, 1984), using NMR to quantify small numbers of markers of interest. However, early use of serum-NMR of lipid levels in the diagnosis of cancer by Fossel *et al.* (1986), was found to be based on falsified data (Spratlin *et al.*, 2009), resulting in a significant setback to the acceptance of NMR profiling in cancer assessment.

Metabolic-profiling approaches to serum NMR were greatly facilitated by the application of advanced pulse-sequences (Brindle *et al.*, 1979; Nicholson *et al.*, 1983, 1984; Waters *et al.*, 2001; Beckwith-Hall *et al.*, 2003) that could take advantage of the differing properties of large and small molecules to exclude the resonances attributable to macromolecules, improving the detection of resonances from small molecules →. This combined with the multivariate statistical techniques that had previously been introduced to urine-based metabolic profiling (Antti *et al.*, 2004) has established serum-based metabolic profiling as an equally valuable adjutant to urinary profiles (Lindon *et al.*, 2007), with applications varying from toxicological studies, to clinical use such as cancer screening (Teahan, 2009; Keun *et al.*, 2009).

These sequences are further discussed in § 5.1.1.

1.3.2 Metabolite profiling in preclinical toxicological screening

Since inception, the development of metabolite profiling techniques has been tied to their use in toxicological studies. Metabolic profiling can provide a fresh perspective on toxicological assessment (Lindon *et al.*, 2007; Nicholson *et al.*, 2002), and its incorporation into such studies is facilitated, as typically it requires no more sampling than that available from the ADME studies that are generally run as part of any toxicological assessment process.

In addition to providing biomarkers of toxicity, including the ability to report otherwise silent lesions, metabolite profiling can provide an insight into the mechanistic processes resulting in toxicity. Recent work has also demonstrated that pre-dose metabolic profiles can predict the extent of toxic response (Clayton *et al.*, 2009), opening the possibility of using profiles in risk assessment to predict individual variability in toxicity, in a technique known as *pharmacometabonomics*.

Some of the earliest work by Nicholson *et al.* (1985) demonstrated the sensitivity of NMR profiles of urinary metabolites to renal damage. Further work by Holmes *et al.* (1992) and Anthony *et al.* (1994), demonstrated that model kidney toxins, known to cause injury by differing mechanisms could be distinguished by pattern recognition of the metabolite profiles. Amongst others, these studies demonstrated that metabolic-profiles not only respond to a toxic insult, but that this response can be unique to the location and mechanism of toxicity. As reported by Holmes *et al.* (2000), these observations supported the hypothesis that distinct mechanisms of toxicity generated distinct metabolic profiles, and served as the impetus for the COMET project, which is discussed fully in § 1.7.

The pioneering studies applying metabolite profiling to toxicological assessment concentrated on the use of urine samples, due both to the ease of sampling, and the ability to collect a timecourse of samples from a single animal →. Analysis of serum samples is now common, and has complimented many of the observations made in urine, including applications to hepatotoxicity by Beckwith-Hall *et al.* (2003), although due to the lower availability of samples, studies based around serum typically have far lower numbers of samples than those based around urine and therefore serum profiles are typically used as a compliment to urine data.

The volume of blood required for NMR typically rules out repeated sampling of serum from small experimental animals such as rats and mice, while repeated tissue sampling has negative implications for the ongoing wellbeing and health of experimental animals.

1.4 Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance spectroscopy provides a uniquely powerful, non-destructive method of examining the composition of a liquid or solid sample. Initially developed in the 1940s and much improved in the subsequent years, NMR takes advantage of the intrinsic magnetic moment of nuclei with a non-zero spin ← (Hore, 1995). Spin being a fundamental property of atomic nuclei, related to the number of protons and neutrons in the nucleus.

Such nuclei include: ^{1}H, ^{2}H, ^{3}H, ^{13}C, ^{10}B, ^{11}B, ^{14}N, ^{15}N, ^{17}O, ^{19}F, ^{23}Na, ^{29}Si, ^{31}P, ^{35}Cl, ^{37}Cl, ^{113}Cd & ^{195}Pt.

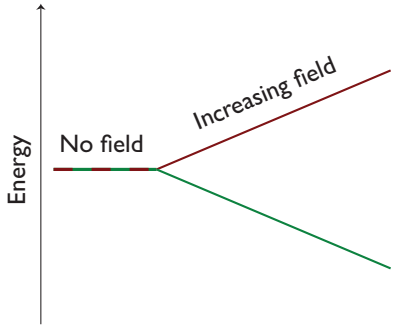

Figure 1.2: As an external magnetic field is applied to a spin-½ nucleus the gap between the two possible energy states widens.

Under normal conditions, a nucleus with a spin of i will randomly take one of $2i+1$ possible orientations. For a spin-½ nucleus such as ^{1}H, this means there will be two possible orientations. When subjected to an external magnetic field, the energy levels of these orientations will split, such that the population of nuclei will now exist in two states, one slightly higher in energy than the other (Figure 1.2). This difference in energy level between the two states is proportional to the external magnetic field, and the distribution of the population between each state is determined by this energy gap and the Boltzmann distribution. Thus, there will be a greater proportion of nuclei in the low-energy state, and this proportion will increases as the energy gap between the states widens.

As the sensitivity of an NMR spectrometer is driven by the difference in size between the high- and low-energy populations, this widening drives, in part, the use of higher field magnets. However, even at the highest fields in use today, the difference in the size of the populations is in the order of 1 part in 10^{4} (Harwood & Claridge, 1997), and this miniscule difference in population sizes is responsible for the low sensitivity of NMR spectroscopy.

At the most basic, spin-½ nuclei are often visualised in terms of an atomic-scale magnetic dipole. On application of an external magnetic field, the dipoles in a sample line up, either with the external field (the low-energy state), or opposing it. The rotation of the nucleus combined with its magnetic moment, will then cause the nucleus to precess about the applied field. The period of this precession is known as the *Larmor frequency* and is directly related to the energy gap between the high- and low-energy states.

Approximately 9.2 Tesla for a 400 MHz magnet and 14.1 Tesla for a 600 MHz instrument.

At the typical field-strength of modern superconducting NMR magnets ← the Larmor frequency, and thus the energy gap between states, falls in the radio-frequency (RF) range of the electromagnetic spectrum. By exciting a sample with a pulse of radio-frequency radiation, we can cause some of the nuclei to resonate at their Larmor frequency, alternating between the high and low energy states, as they absorb energy from the pulse. Following this excitation, the nuclei in the sample will return to thermal equilibrium, gradually returning to the original energy and phase distributions in a process known as *relaxation*.

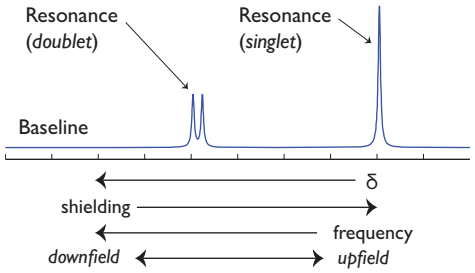

Figure 1.3: Glossary of terms used when referring to an NMR spectrum. The frequency scale of an NMR spectrum is plotted right to left for historical reasons.

Relaxation primarily take place by two mechanisms →, spin-lattice, and spin-spin relaxation. Spin-lattice, or T_1 relaxation occurs because an individual nucleus does not exist in isolation, rather it is part of a sample or *lattice*. The combined vibrational and rotational motion of all the nuclei within the lattice generates a complex magnetic field, and interactions between this field and nuclei in the high-energy state can cause the nuclei to return to the low-energy state, imparting the energy to the lattice, where it manifests as a small increase in temperature. Spin-spin, or T_2 relaxation involves the transfer of energy between nuclei with identical Larmor frequencies. Where one nucleus is in the excited state, and another is in the low-energy state, the two nuclei can exchange quantum states, the nuclei in the high-energy state moving to low-energy and vice-versa. While this process leaves the total distribution of nuclei between energy states unchanged, because, on average individual nuclei are in the excited state for a shorter period of time, spin-spin relaxation can cause an increase in the line-width of the acquired spectrum.

The third, direct emission of EM radiation at the wavelength of the energy gap, occurs with a probability that varies with the cube of the frequency. At the radio-frequency range observed in NMR this is minimal.

The utility of NMR for spectroscopy arises because the magnetic field at each nucleus in a sample is not identical to the applied field, rather, the electrons orbiting each nucleus will have a shielding effect, modulating the field experienced by the nucleus. This difference between the applied field and the field felt at the nucleus is known as *nuclear shielding*, and is observed as a shift in the resonant frequency of the nucleus. Because the precise electronic environment is strongly dependent on the chemical structure of the molecule, the frequency shift can be related to chemical structure, and therefore used to identify resonances belonging to a specific molecule.

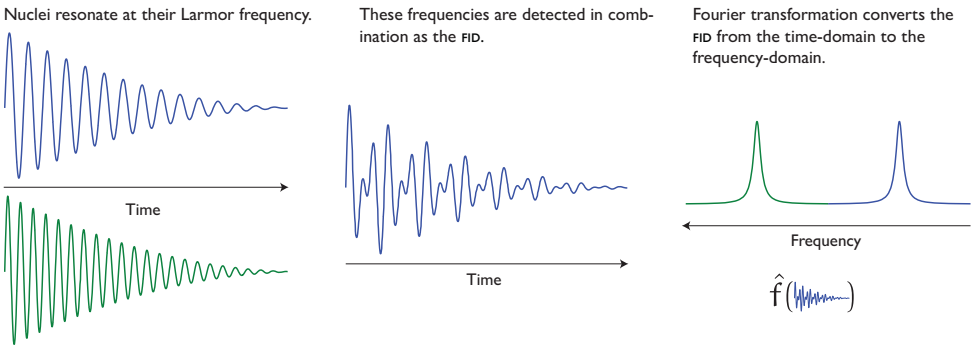

Figure 1.4: Following excitation, nuclei resonate at their Larmor frequency, while the population returns to thermal equilibrium. The combination of all these frequencies in a sample is detected as the FID. The Fourier-transform can then express the time-domain FID in terms of its constituent frequencies.

To acquire an NMR spectrum, a brief pulse of RF energy is applied to a sample, causing a subset of the nuclei present to resonate at their Larmor frequencies. The precessions of these nuclei can then be observed as a decaying oscillation, fading as the sample relaxes to the equilibrium state. This signal, known as the free-induction-decay (FID) consists of a combination of signals, each at the Larmor frequency of a nucleus under specific conditions, with an amplitude directly related to the number of nuclei precessing at that frequency. By means of the Fourier transform, the FID can then be transformed from repeating signal, decaying over time, to a frequency-domain representation of resonances spread across a frequency scale (see Figure 1.4), giving rise to the familiar NMR spectrum (Hausser & Kalbitzer, 1991).

NMR is a strongly quantitative experimental method (Evilia, 2001; Šárka Mierisová & Ala-Korpela, 2001), in which the area of a resonance is directly related to the number of nuclei generating the signal. Therefore the area of a resonance attributed to a known standard, either spiked into the sample, or an endogenous component quantified by other means, can be directly related to the chemical formula of that compound, and from there to the formula and signal-area of unquantified compounds. This assumes both molecules are allowed to fully relax during the acquisition of the NMR spectrum. In situations where equal relaxation cannot be assumed, quantification must be directed by the addition of an internal standard for each molecule to be calibrated (Sarpal *et al.*, 1997). In controlled situations, components of a solution can be quantified with an error of below 1%, down to components comprising less than 0.1% of the solution (Maniara *et al.*, 1998).

Preparation of biofluid samples for NMR requires minimal work, typically only pH buffering, as some resonances can be strongly pH sensitive, and the addition of a referencing agent. This coupled with the ability to re-use samples for more complex experiments, and the near-universality of hydrogen and carbon nuclei in metabolically relevant molecules, makes NMR especially suited to metabolic-profiling studies.

In metabolic profiling, there is a problem with overlapping resonances, particularly in ^{1}H spectra, where resonances with similar Larmor frequencies obscure each other, thus complicating identification and quantification. However, pulse sequences such as the Carr-Purcell-Meiboom-Gill (CPMG) sequence can mitigate this, by suppressing resonances associated with large molecules with short T_2 relaxation times (Carr & Purcell, 1954; Meiboom & Gill, 1958). Further work on fast 2D experiments show promise for further reducing this problem, without editing of the acquired resonances (Viant, 2003).

1.5 Chemometric Methods for Data Analysis

The name *chemometrics* describes a family of mathematical techniques (Geladi, 2003; Geladi *et al.*, 2004) that combine multivariate statistics and pattern recognition, and are primarily concerned with the use of various forms of linear algebra to describe complex chemical systems. Although the name chemometrics was only introduced by Herman Wold in the early 1970s (Esbensen &

Geladi, 1990; Geladi & Esbensen, 1990), many of the techniques were already in common use (Kowalski & Bender, 1972).

Chemometric techniques are directed toward systems in which there are many more variables than observations →, systems where traditional statistical regression models break down (see § 1.5.3). The primary tenet of all chemometric methods is the assumption that, in any system with a large number of observed-variables, there will be a lesser number of latent-variables. Each of these latent-variables represent a combination of the observed-variables in the data, some mixture of which act in concert to generate the observed variation. By mapping the samples onto these variables, the structure within multivariate data can be easily visualised and understood.

Often referred to as multivariate, megavariate or massively-multivariate systems.

These latent variables may depend on some experimental factor of interest, or they could describe an interference. The classic example of this are the data-points in an NMR spectrum. The intensities of all the variables that make up one resonance, and all the resonances from one compound are controlled by one latent-variable, the concentration of the compound in the sample.

This picture is complicated in most real world situations, where one observed-variable may be influenced to varying degrees by several latent-variables. In NMR this is seen where resonances overlap, and where the concentrations of several molecules show a strong correlation. Thus the latent variables calculated from a set of NMR spectra of a biofluid often represent some combinations of the resonances in several co-varying molecules.

1.5.1 Chemometrics in metabolite profiling

Advances in metabolite profiling have been intimately tied to the progress of the chemometric techniques used to interpret the multivariate data generated by the analytical instruments.

The very earliest work in the field of NMR assessment of toxicity, that eventually developed into metabonomics, eschewed pattern-recognition techniques, examining NMR spectra with univariate statistics, applied to individual resonances selected by eye, after searching for differences between groups (Nicholson *et al.*, 1985). However, the need to easily summarise and understand the relationships between hundreds of variables quickly drove the adoption of pattern-recognition techniques. One of the earliest data-reduction methods used was non-linear maps also known as Sammon mapping (Sammon Jr., 1969). This method distorts a set of observations in an n-dimensional space into two- or three-dimensions, while attempting to retain the inter-point distances, and was applied to the classification of toxicity (Gartland *et al.*, 1989, 1990). Further chemometric techniques were being applied to biological data by 1990 (Esbensen & Geladi, 1990), including Principal Components Analysis (PCA) to metabolite profiles (see § 1.5.2) (Holmes *et al.*, 1992).

The desire to discover biomarkers distinguishing classes in otherwise homogenous groups drove the adoption of supervised methods of pattern recognition. Supervised methods will take *a priori* knowledge of structure in data into account when constructing the model, allowing them

to 'force' a distinction where unsupervised methods may see no difference between the groups. This supervision may take the form of a regression to a continuous variable of interest, or to a class identifier. Many such techniques have been applied to metabolite profiles, including factor-analysis and neural-networks (El-Deredy, 1997). Recently, the most popular methods used in metabolite profiling have been the Partial Least-Squares (PLS)-based ← methods. Despite being a well known method of spectral analysis (Haaland & Thomas, 1988), PLS (see § 1.5.4) was not widely used in metabolic-profiling studies prior to 2000 (Gavaghan *et al.*, 2000).

PLS is sometimes also explained as projection to latent structures.

Extensive use of chemometrics in NMR based profiling first became computationally feasible with the introduction of data-reduction methods such as the integration of spectral variables into a limited number of regions (Spraul *et al.*, 1994), a technique that both reduces the computational time required to calculate models, and helps mitigate the effect of shifting resonances in NMR (Figure 1.5).

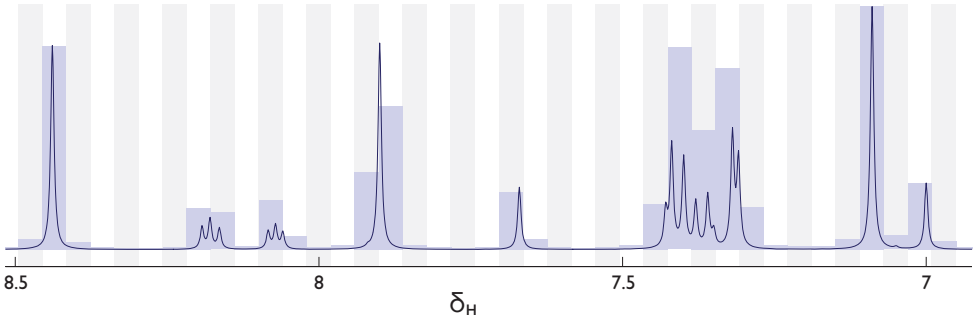

Figure 1.5: Here a spectrum has been reduced to a resolution of δ_H = 0.04 by summing the spectral intensities within defined regions. Reducing the resolution greatly reduces the complexity of the data, and compensates for the shifting of resonances to a degree. On the negative side, many resonances are split between integrated regions, and this may complicate interpretation of the data.

As computational power has increased and visualisation tools have improved, analysis of NMR spectra has moved towards using data at, or close to, the full spectral resolution (Cloarec *et al.*, 2005b). This approach, combined with developments in statistical reconstruction of spectra with correlation based methods (Cloarec *et al.*, 2005a), has been directed towards integrating the processes of modelling and biomarker identification.

More advanced unsupervised models have also been applied to metabolite profiles, such as the Tucker3 multiway model (Tucker, 1966). The Tucker3 model can decompose a dataset in higher dimensions than PCA, maintaining more of the structure inherent in the data than PCA, at the expense of an increase in model complexity. Dyrby *et al.* (2005a) used the Tucker3 model to explicitly model time as an axis in the development of a toxicological response. Further techniques such as multivariate curve resolution (MCR, Tauler *et al.*, 2002), have also been applied to NMR spectra. MCR is similar to PCA, but has the desirable attribute of imposing a non-negativity constraint on each constituent profile. In simple systems (Winning *et al.*, 2008),

this can result in components that are much closer to real-world chemical profiles of pure components than PCs →.

Which may be negative in parts, an attribute that is clearly not possible for the true profiles making up an NMR spectrum.

Correlation based methods take advantage of the internal correlations found in large sets of data, identifying variables that co-vary. The strongest correlations in s set of NMR spectra will be seen between resonances belonging to the same molecule, then, at lower levels, correlations will indicate molecules that co-vary to some degree. Therefore, by adjusting a cut-off on the correlation level, these methods allow the analyst to focus on specific relationships between variables, such as identifying resonances likely arising from the same molecule, or from molecules closely related in metabolism, such as those adjacent within a pathway (Couto Alves *et al.*, 2009).

Correlation based methods are also being directed towards integrating data acquired from divergent sources, both in terms of combining metabolite data between analytical platforms (Crockford *et al.*, 2006), and integrating metabolite profiles with other —omics data (Rantalainen *et al.*, 2006; Dumas *et al.*, 2007). However, these methods are susceptible to correlations driven by outliers in the data, and this must be accounted for with appropriate significance testing.

1.5.2 Principal component analysis

Principal component analysis is one of the longest established multivariate analysis techniques (Wold *et al.*, 1987), dating back to the turn of the 20th century, where it was described by Pearson (1901). PCA allows a set of multivariate data to be visualised in terms of the observations or variables that change in concert.

Taking no information not present in the original data matrix into account, PCA is an unsupervised method of data-reduction. This lack of bias makes PCA highly flexible, and has allowed it to find applications in a wide range of fields. PCA-type methods were separately developed in several disciplines, where the same process may be referred to as the Karhunen–Loève transform (in reference to stochastic processes and electrical engineering, referenced in Wold *et al.*, 1987), the Hotelling transform (in image analysis, Hotelling, 1933), proper-orthogonal decomposition (POD) or singular-value decomposition (SVD, Mandel, 1982).

PCA acts by generating combinations of variables known as Principal Components (PCs), each of which describe one of the underlying latent-variables that define the variation in these data. Each PC consists of a set of loadings, which describe the linear combination of variables making up the component, and a set of scores, that describe the contribution this loading makes to each observation:

$$X = TP^T + E \tag{1.1}$$

Thus the data matrix X may be decomposed into scores T on loadings P. Variance not captured in the model forms the residuals E.

PCs are extracted from the data in order of explained variance so that the first PC models the largest possible amount of variance. PCs are orthogonal to each other, meaning the scores on one component are uncorrelated with those for any other component. As components are generated in order of explained variance, examination of the observations in the axes defined by the first PCs (see Figure 1.6) will capture more of the structure in the data than any two individual variables, thus providing a quick, high-level overview of the data.

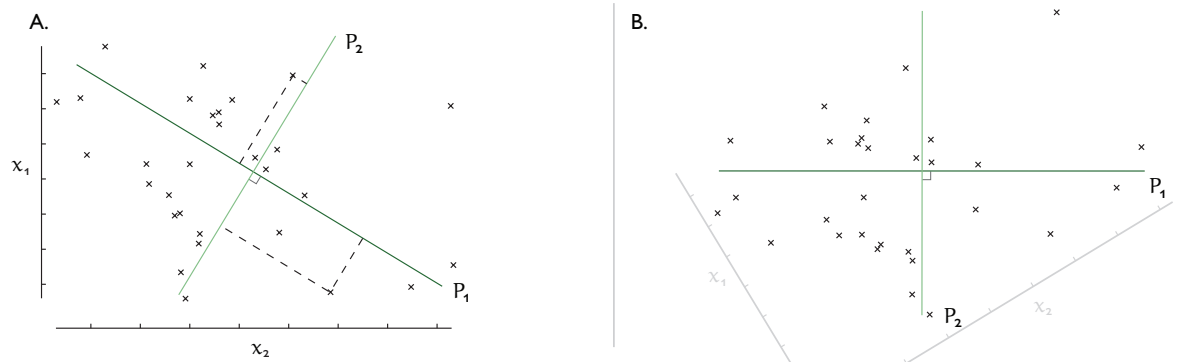

Figure 1.6: Geometric overview of PCA modelling. **A.** For a simple two-dimensional set of data, defined by the variables x_1 & x_2, PCs are generated by projecting (dashed lines) each observation onto orthogonal vectors, or *loadings* (in green), of maximum variation in the data, generating a score on each. **B.** The original data points can now be visualised in terms of their position (the score) on the new axes defined by the loadings, each of which represents a combination of the x_1 & x_2 variables.

As each successive PC is generated, the variation it models is removed from the data, until only noise remains. Even noise can be modelled, by calculating PCs up to the algebraic rank of the data matrix. Such a model, that explicitly includes data down to the noise, will inevitably be specific only to the current set of data. A model that is too specific to a single data-set is described as over-fitted. Any PCA model usually needs a balance between the number of PCs required to accurately model the structure of data in question, while remaining generalisable to novel observations. However, over-fitting is not necessarily a problem in PCA, as it merely means the model can only be considered relevant to data in question.

In chemometric use, PCA is generally calculated by an adaptation of the nonlinear iterative partial least-squares (NIPALS) algorithm described by Herman Wold (1974). For each PC, a:

1. $\mathbf{t}_a$ = the column in $\mathbf{X}$ with the maximum range.
2. $\mathbf{p}_a = \frac{\mathbf{X}^T\mathbf{t}_a}{|\mathbf{X}^T\mathbf{t}_a|}$
3. $\mathbf{t}_a = \mathbf{X}\mathbf{p}_a$
4. If $\mathbf{t}_a$ is unchanged, continue, else return to step 2.
5. $\mathbf{X} = \mathbf{X} - \mathbf{t}_a\mathbf{p}_a^T$

These steps can be repeated as many times as necessary to generate the desired number of components.

The NIPALS approach has an advantage over techniques such as SVD, which decompose the X-matrix in its entirety, as it calculates PCs iteratively, in order of their explained variance, starting with the largest. This makes the NIPALS approach faster in situations, such as metabolite profiling, were only the first few components are generally required. If PCs up to the full rank of X are desired, it is faster and more precise to use SVD, and decompose the matrix in its entirety, as this avoids the cumulative rounding errors inherent in the stepwise NIPALS approach (Varmuza & Filzmoser, 2009, p73).

The total amount of variation modelled in a PCA model is expressed in terms of goodness-of-fit or R^2. This is a measure of the proportion of the total variance in the data explained by the model:

$$R^2 = 1 - \sum \frac{(X - TP^T)^2}{X^2} \tag{1.2}$$

As more PCs are added to a PCA model, the R^2 will converge to 1 when all variation in the data has been modelled.

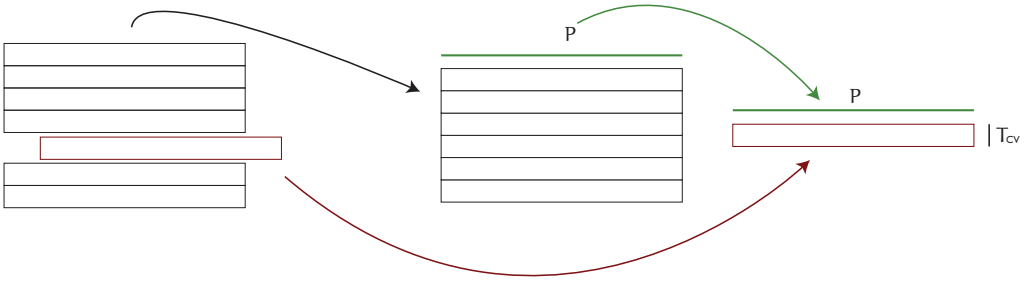

Figure 1.7: In cross-validation, data are split into several blocks (here seven). One block (in red) is then set aside, and a set of loadings P (in green) are generated from the remaining data. These loadings are then used to generate a set of cross-validated scores (T_{CV}) on the block left aside. This process is repeated for each block of data, and the T_{CV}s are joined to form a vector with dimensions matching T.

Over-fitting of a model is diagnosed with the cross-validated goodness-of-fit or Q^2. In cross-validation (Efron & Gong, 1983; Krzanowski, 1987; Kohavi, 1995), data are divided into several blocks of samples, chosen such that the cross-validation blocks are not coincident with any pre-existing structure in the data. Each of these blocks is then sequentially set aside, and a PCA model is generated from the remaining data. Scores are then predicted on the omitted block, using the loadings generated from the remainder of the data. The predicted scores for each block are then collated to form the cross-validated scores. The difference between the model generated with the cross-validated general scores and the original data then provides an estimate of how well the model generalises to novel data:

$$Q^2 = 1 - \sum \frac{(T_{CV}P^T)^2}{X^2} \tag{1.3}$$

Where T_{CV} are the cross-validated scores, and Q^2 is a value at maximum equal to the R^2 for this model, for a perfectly generalisable model, through to large negative numbers for models which are so over-fitted they offer no predictive value as to the structure of novel data.

Both R^2 and Q^2 can be calculated on a component-by-component basis, or simultaneously for every component in the model, which matches the sum of the component-wise values.

1.5.3 Regression models

Often it is desirable to relate a set of descriptive data, such as NMR spectra, to some response, such as levels of toxicity, or concentrations of the constituents of each sample. This may be achieved by the use of regression models, techniques that relate a set of descriptor variables, the X variables, to one or more response variables, known as Y variables. Such models are typically generated on one set of data, the *training* set, where both X and Y are known, which are used to generate a set of estimated regression coefficients $\leftarrow$, which can then used to estimate Y for novel Xs.

The estimated coefficients are typically denoted with $\hat{}$, so the estimate of b is $\hat{b}$.

Multiple Linear Regression (MLR) is the simplest extension of basic linear-regression to situations, such as NMR, where the predictor variable is a vector rather than a scalar value. In MLR, Y is modelled as the weighted sum of the X variables.

$$\begin{aligned} \hat{b} &= (X^T X)^{-1} X^T Y \\ Y &= 1c + Xb + E \end{aligned} \tag{1.4}$$

X is the matrix of variables, b is a vector of regression coefficient for the variables, c represents the intercept of the model and E represents random noise, that is assumed to be independent between each X variable, but with a common standard deviation.

The calculation of MLR requires the calculation of a non-singular least-squares solution to $\dot{X}^T\dot{X}$, and this imposes several restrictions on the use of MLR. Most notably, the number of X variables must be smaller than the number of samples, and there must be no collinearities in $\dot{X}$.

Essentially, MLR requires that the number of variables measured be smaller than the number of samples they are observed on, and that all the variables should exhibit independent variation. This is clearly not practical for NMR data, where spectra typically consist of tens of thousands of variables, many of which are tightly correlated.

An alternative methodology that addresses some of the limitations of MLR is provided by Classical Least-Squares (CLS). In CLS we reverse the previous assumption that we are modelling y in terms of x, and express the x variables as a function of y. In this, CLS is based on an adaption of the Beer-Lambert law of light absorption, taking the form:

$$\begin{aligned} \hat{A} &= (Y^T Y)^{-1} Y^T X \\ X &= 1\alpha + YA + E \end{aligned} \tag{1.5}$$

Where A is matrix of pure profiles, α is a row vector of offsets and E is a matrix of random noise.

CLS models X as a varying combination of the pure profiles in A (Haaland & Thomas, 1988). CLS has an advantage over MLR in that collinearities in X are allowed and X may contain more variables than samples. CLS is limited by the assumption that A should take every constituent profile into account, that is, if not all variation in X can be explained by parameters in Y, the model will not be optimally fitted, a problem in metabolic-profiling studies, where X often contains many unknown profiles.

Principal Components Regression (PCR) is a hybrid of PCA and MLR (Hawkins, 1973). Regression is carried out on the components of a PCA model of X, rather than X itself:

$$\begin{aligned} \hat{C} &= (T^T T)^{-1} T^T Y \\ Y &= TC + E \end{aligned} \tag{1.6}$$

Where T are the scores from a PCA model of X.

Due to the orthogonality of PCs, this has the advantage of obscuring the collinearities in X from the regression. Additionally, while PCR still retains the requirement that the number of variables be less than the number of samples, because in this case the PCs represent the variables, this is less of a problem.

However, because the variation modelled by PCA is determined by the magnitude of the variance, if the contribution of Y to X is small, it may not be effectively captured by the PCA model, and hence PCR will perform poorly.

Beyond these linear models, other regression techniques such as ridge-regression → have been applied to metabolite profiles, but none have achieved the success of partial-least-squares based methods (Höskuldsson, 1988).

Also known as Tikhonov regularisation.

1.5.4 Partial least-squares regression

PLS (Wold *et al.*, 2001) overcomes the limitations of the above regression models by directly modelling X, generating latent-variables that are maximally associated with Y:

$$\begin{aligned} \hat{C} &= (T^T T)^{-1} T^T Y \\ Y &= TC + E \end{aligned} \tag{1.7}$$

Note the similarity in the generation of the regression vector $\hat{C}$, to PCR. The important difference lies in the generation of the scores T. In PLS, these scores are generated from PLS components rather than PCs. The PLS loadings, P & Q are calculated *via* the NIPALS algorithm, to maximise the covariance of the scores between the X and Y matrices. For each PLS component a:

1. u_a = the column in Y with the maximum range.
2. $w_a = \frac{X^T u_a}{u_a^T u_a}$
3. $w_a = w_a^T \sqrt{w_a^T w_a}$
4. $t_a = X w_a$
5. $c_a = \frac{Y^T t_a}{t_a^T t_a}$
6. $c_a = c_a^T \sqrt{c_a^T c_a}$
7. $u_a = \frac{Y_a^T}{c_a^T c_a}$
8. If u_a is unchanged, continue, else return to step 2.
9. $p_a = \frac{X^T t_a}{t_a^T t_a}$
10. $q_a = \frac{Y^T u_a}{u_a^T u_a}$
11. $\hat{b}_a = \frac{u_a^T t_a}{t_a^T t_a}$
12. Set $X = X - t_a p_a^T$ and $Y = Y - \hat{b}_a t_a c_a^T$

As described above in § 1.5, the Y matrix in PLS can be any set of numerical values. This could be clinical chemistry measures, allowing the relation of NMR spectra to concentrations of compounds in the samples. Alternatively, by generating a set of dummy variables that indicate class membership, PLS can be used as a classification tool, in a process known as discriminant-analysis (PLS-DA). Examination of the weights following a PLS-DA regression to class membership is a common method of identifying variables that distinguish classes which appear similar in PCA space (Geladi, 1988; Wold *et al.*, 2001).

1.5.5 Orthogonal filtering and O2-PLS

For a PLS model with a single Y variable, the best relation between the X and Y matrices should be achieved with a single PLS component. However, if X contains a large amount of systematic variation unrelated to Y, known as orthogonal variation, PLS has to model this variation in concert with the correlated variation of interest (Figure 1.8). In turn, this may confuse the interpretation of the model, as the variation of interest becomes spread across several components. In response to this, Orthogonal-PLS (O-PLS) and its successor O2-PLS were developed by Johan Trygg (2002) (Trygg & Wold, 2003). O-PLS and O2-PLS incorporate an orthogonal signal correction (OSC) filter into the PLS modelling process.

The OSC filter models and removes the orthogonal variation from the data, prior to the modelling of variation correlated between X and Y. O2-PLS differs from O-PLS by the addition of the capability to model orthogonal variation in the Y matrix as well as X. By removing this orthogonal variation, both techniques concentrate the maximum information regarding covariation between the X and Y matrices, in the fewest components.

It is important to note that O-PLS does not provide better prediction than classical PLS, rather it improves the interpretability of the model, by separating related and confounding variation into correlated and orthogonal scores and weights. However, by modelling and removing orthogonal variation from the Y matrix, O2-PLS can offer an improvement in some situations.

The O2-PLS model is more complex than the previous regression models, as it simultane-

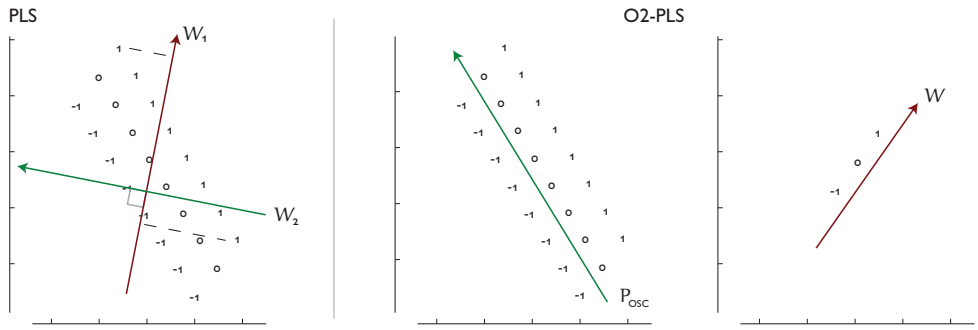

Figure 1.8: When modelling data with systematic variation, observations with similar Y values (indicated by the value of each numeral), will produce different scores on the first PLS component W_1. In PLS this must be corrected for with a second component, W_2. In an O- or O2-PLS model, the systematic uncorrelated variation is modelled and removed by one or more orthogonal components, P_{OSC}, allowing the weights, W, to directly model the correlated variation.
Adapted from Trygg (2002).

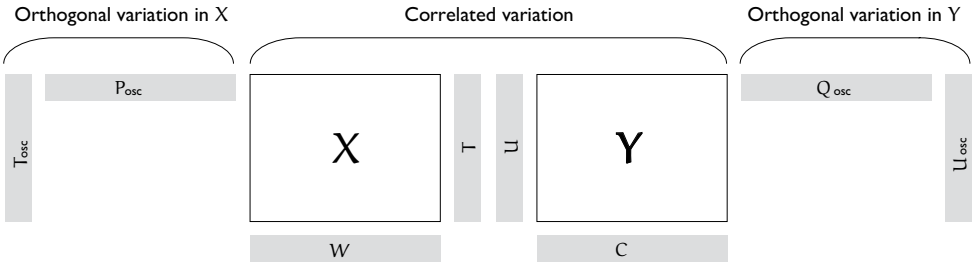

Figure 1.9: O2-PLS models a set of data in three parts (Equation 1.8), the correlated variation between the X and Y matrices, and the orthogonal variation found in each matrix.

ously models and predicts both X and Y.

$$\begin{aligned}
\hat{B_U} &= (U^T U)^{-1} U^T T \\
\hat{B_T} &= (T^T T)^{-1} T^T U \\
Y &= T B_T C^T \\
X &= U B_U W^T \\
X &= T W^T + T_{osc} P_{osc}^T + E \\
Y &= U C^T + U_{osc} Q_{osc}^T + F
\end{aligned} \tag{1.8}$$

Where the orthogonal scores and loadings, denoted $[]_{osc}$, capture the orthogonal variation in the X and Y matrices.

1.5.6 Validation of PLS models

Much like PCA, PLS models, whether O2-, O-, or standard PLS may be judged in terms of the explained variance in the model, and generalisability to novel data. As PLS takes two matrices into account, there are two explained variance statistics R^2X and R^2Y, representing the X & Y matrices respectively. As PLS models are typically generated with the aim of using X to predict Y, the explained variance in Y is generally the more reported statistic.

Rather than deriving from the predicted scores of the excluded data, the Q^2 of a PLS model is based on the ability of a model to predict the Y values of the set aside data. As PLS models are generated to relate two matrices, the Q^2 is much more important than in PCA, indicating whether the observed relationship is likely to be a general one, or is as a result of over-fitting the model.

Predictive models are often also validated with more general permutation tests. By repeatedly randomising the Y matrix and generating new models, an estimate of the likely distribution of R^2[X or Y] and Q^2 in the current data can be made, under the assumption that there is no true relationship between X and Y. By comparing this distribution to the true values, it is possible to diagnose cases where a large value of Q^2 could arise by chance, or conversely, a small value could still represent a model that is performing better than chance (Lindgren *et al.*, 1996).

1.6 The Receiver-Operating Characteristic

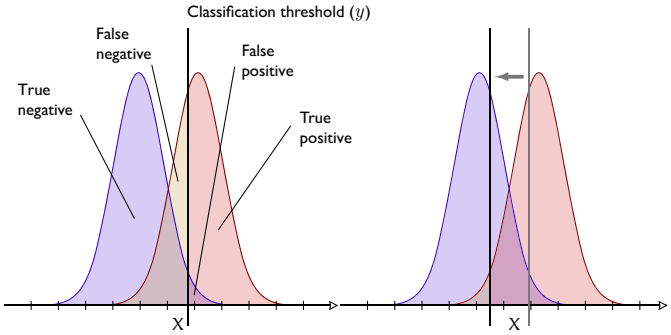

Figure 1.10: Two populations (red & blue) exhibit two separate distributions of the measure x. As the classification threshold, y, is lowered, the proportion of false-negatives falls (sensitivity increases) but the number of false-positives rises (specificity decreases). The extent of these changes will depend on the distribution of samples in the positive and negative groups.

The Receiver-Operating Characteristic (ROC) is a long-established means of describing the sensitivity and specificity of a binary classifier over the complete range of potential classification thresholds.

The use of the ROC was pioneered in the 1940s as part of the field of signal-detection

theory, applied to the analysis of radar signals. As reported by Zweig & Campbell (1993), by the late 1950s the utility of the ROC saw its use quickly spread to medicine, becoming a common validation tool in many fields by the present day.

The ROC is intended to judge the quality of a metric used to separate two groups in a population (Fawcett, 2004). In such a population, any continuous (numeric) observation, x, may show a significantly different distribution between the groups. The question then arises, knowing only the value of x, how accurately can you assign samples to the correct group? Measuring x for an observation, a binary classifier will assign the observation to a group based upon a threshold value y. Observations where $x < y$ will be assigned to one group and the remainder to the other. This can be seen graphically in Figure 1.10.

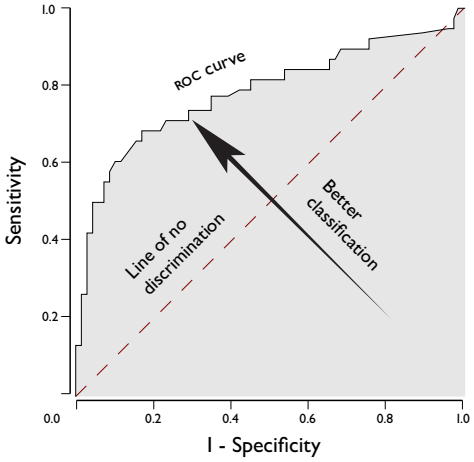

Figure 1.11: Basic outline of an ROC curve. Each point on the curve represents one value of the classification threshold y. The area under the curve is shaded; the dashed diagonal indicates the line of no discrimination, where prediction is no better than chance.

In this situation, the classification of any observation may be sorted into one of four categories, those correctly assigned to their group, known as true-positives and -negatives, and those erroneously assigned to the incorrect group, known as false-positives and -negatives. Depending on the threshold value and the overlap and shape of the distributions, the proportion of these errors will shift.

These categories are often summarised as →; sensitivity, the proportion of true positives to true positives plus false negatives; and specificity, the proportion of true negatives to true negatives plus false positives. The ROC can be used to summarises this intrinsic trade-off between sensitivity and specificity by plotting the sensitivity vs. 1-specificity → for the entire range of possible threshold values. An ROC curve that runs along the diagonal of the plot essentially classifies samples randomly, while any curve that extends into the upper left half improves on this, as seen in Figure 1.11. The ROC is often summarised in terms of the area under the curve (AUC), simply the integral of the ROC curve →, though this value can be misleading if the ROC curve is not a simple parabola.

As proportions both these metrics potential values fall between 0 and 1.

Purely by convention, to provide a plot that rises to the top-right.

Also a proportion, of the maximum possible area, with 1 representing perfect discrimination.

1.7 The COMET Project and Dataset

The COnsortium for MEtabonomic Toxicology (COMET) project (Lindon *et al.*, 2003) was a large-scale collaboration between six pharmaceutical companies; Bristol-Myers-Squib, Eli Lilly, Novo-Nordisk, Hoffman-La Roche, Pfizer, Pharmacia (since acquired by Pfizer) and Imperial

College, London. The consortium aimed to generate a database of metabolic-profiles of toxic insult, as an opening into the use of metabolic profiling techniques in preclinical toxicological screening (Lindon *et al.*, 2005a).

COMET ran 146 studies, mainly concentrating on acute dosing of model liver and kidney toxins, but also including chronic studies of toxicity and physiological studies, such as the effect of dietary restriction or tissue regeneration following surgical removal of a kidney or lobe of the liver. Concomitant to these *in vivo* studies, was the acquisition of ^{1}H-NMR spectra of urine and serum collected from the studies to allow the generation of multivariate statistical models and an expert system based on the acquired data, designed to classify the class of toxicity in novel studies.

In addition to the sheer quantity of data generated in COMET, the project advanced the field of metabolite profiling, both in experimental and analytical terms. One of the earliest achievements in COMET was a demonstration of the reproducibility of NMR-based metabolite profiling studies between sites and instruments (Keun *et al.*, 2002b), as well as an in-depth analysis of the differences in hydrazine toxicity between the rat and mouse (Bollard *et al.*, 2005). Some of the first cryo-NMR ← data on biofluids, taking advantage of the extra sensitivity of cryo-NMR to acquire ^{13}C-NMR, were acquired from COMET samples (Keun *et al.*, 2002a).

NMR in which the receiver-coil and preamplifier are cooled to cryogenic temperatures to reduce thermal noise (Styles *et al.*, 1984).

On the data analysis side, COMET introduced variable-stability (VAST) scaling, a variable scaling technique that weights individual variables according to their stability across a data-set (Keun *et al.*, 2003). Higher-level pattern recognition techniques included the CLassification Of Unknowns by Density Superposition (CLOUDS) method, used to classify the target organ and mechanism of action of unknown toxins, by positioning dosed animals in a multivariate 'metabolic space' based on a model of the metabolic state of control animals, and comparing their overlap with known toxins (Ebbels *et al.*, 2003, 2007). Related to CLOUDS, geometric-trajectory analysis considers a toxic insult in terms of a deviation from normal in metabolic space, with similar mechanisms of toxicity causing characteristic deviations from the control space. These deviations may be scaled according to dose or the extent of toxic effect (Keun *et al.*, 2004), but conserve the characteristic trajectory.

1.7.1 Study design

To the greatest possible extent, all COMET studies were conducted to a common protocol. The typical COMET study was run using the Sprague-Dawley rat (strain Crl:CD (SD) IGS BR) as the model animal. Twenty comparative studies were also run in the mouse (strain B6C3F1), and one in the Hans-Wistar rat (strain HanBrl:WIST(SPF)), to provide the opportunity to explore the variation in metabolic response between species and strains. Each study was conducted with 30 rats or 24 mice. Animals were all male, and between 5 and 8 weeks old ($5-10$ for the mice). Prior to the commencement of each study, animals were placed into metabolism cages and allowed approximately one week to acclimatise to their new conditions. Animals were fed and watered *ad libitum* for the duration of each study.

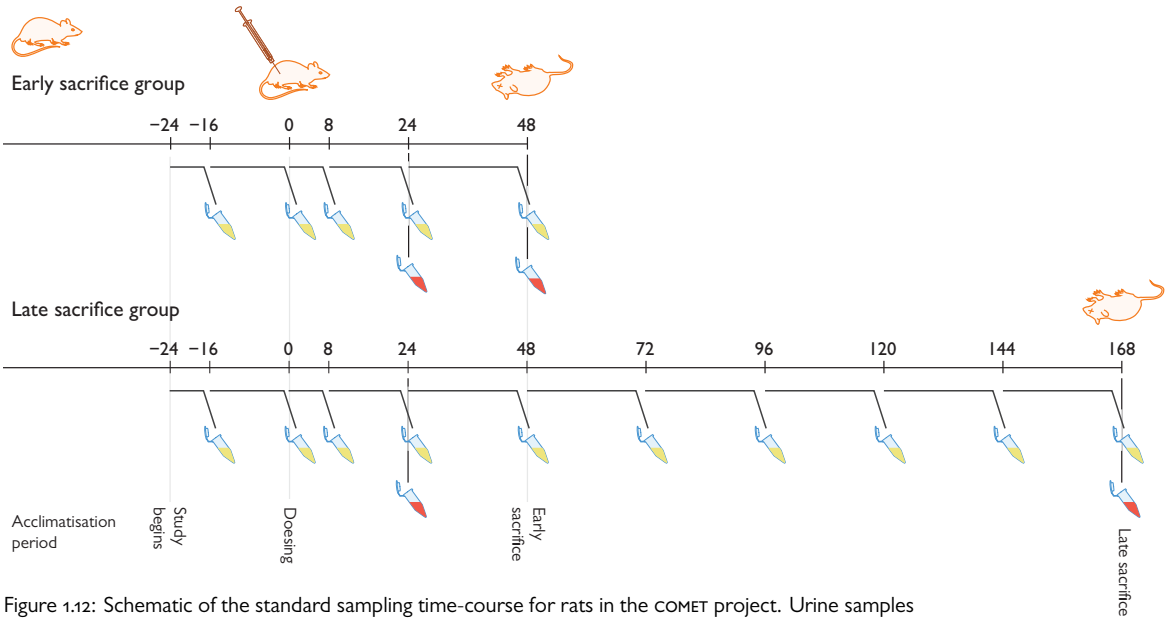

Figure 1.12: Schematic of the standard sampling time-course for rats in the COMET project. Urine samples (yellow) were pooled across each time-period, while serum (red) was sampled directly at the timepoints indicated. Note that 24h serum was only analysed by clinical-chemistry, due to limited volumes of sample.

Within a study, the animals were randomly assigned into one of three groups; controls, that were dosed only with the dosing vehicle; low-dose animals, dosed at a level intended to elicit a metabolic response but no overt toxicity; high-dose animals, dosed at a level intended to produce toxic lesions visible during histopathological examination. In addition to dose-groups, the animals were separated into an early and late sacrifice group, such that half the animals from each dose-group were in each sacrifice group. Animals in the early sacrifice group were euthanised after 48 hours, while those in the late group were euthanised after 168 hours to allow comparative histopathological tissue analysis from the two timepoints.

1.7.2 Sample collection

COMET sampled urine, serum and tissue from each animal. Samples were collected for a period of eight days including one day of pre-dose sampling to establish a metabolic baseline for each animal, Figure 1.12 shows a summary of the sampling timeline. Urine was pooled over the course of each sampling period, being collected into refrigerated containers with sodium azide to prevent bacterial growth. Blood was collected into microcentrifuge tubes *via* puncture of the tail vein, and serum separated from the plasma. Liver and kidney, in addition to any other tissue considered relevant for the study, were collected at each animals designated sacrifice point. All samples were maintained below -40°C during shipment to, and storage at, Imperial College. NMR spectra of serum were acquired with both a standard 1D water-suppressed pulse sequence (referred to as NOESYPR1D spectra) and a Carr-Purcell-Meiboom-Gill pulse sequence (referred

to as CPMG spectra), to suppress the resonances associated with macromolecules in solution. Technical details of sample preparation and data acquisition may be found in Appendix B.

Chapter 2

The Comet Database

2.1 Aims for the Database

In a large study such as COMET, there exists a large amount of descriptive data above and beyond the metabolite profiles that were the primary data collected. These data referring to other data are known as *metadata*, and capture high-level information such as study design and dosing regimes, all the way down to basic data like additional measurements made on the samples, for instance clinical-chemistry or histopathology.

Due to the large numbers of studies and samples, any in-depth analysis conducted on the COMET data as a whole would require an ability to quickly and simply, sort, select and match samples according to various criteria on their metadata. To simplify these processes, a relational database was designed and implemented to hold the metadata associated with the COMET samples and studies. This would both serve as a central, audited repository for these data, and also a research tool, by allowing sophisticated queries of these data, such as those exemplified in § 2.4. In turn, by simplifying the extraction and annotation of subsets of the samples, comparison of these groups would be accelerated.

This chapter outlines the concepts and techniques utilised in the construction of the COMET database. An introduction to the basic concepts and technology involved in relational databases can be found in § 2.2, while examples of the applications of database queries to the COMET data are given in § 2.4. Full implementation details along with specifics of the queries used to extract the data used in this thesis can be found in Appendix C.

2.2 Database Concepts

2.2.1 Database terminology

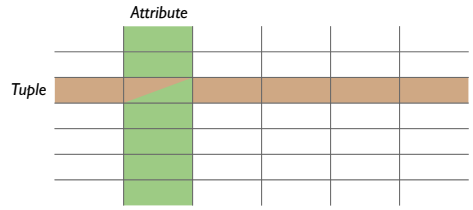

Figure 2.1: Components of a database relation. Each relation is defined by a number of *attributes* that determine the data the relation is capable of storing. *Tuples* refer to individual records.

Relational databases encapsulate systems for modelling and storing data, capturing them as a hierarchy of relationships or connections between distinct units within the data, known as *relations*. A relation represents a specific aspect of the data, and they are usually described in terms of tables of data (Figure 2.1). Each table consists of a set of records, or *tuples* in database terminology, capturing a specific set of fields or *attributes*. Typically *tuples* are the rows of the data table and *attributes* the columns, but this arrangement is purely convention.

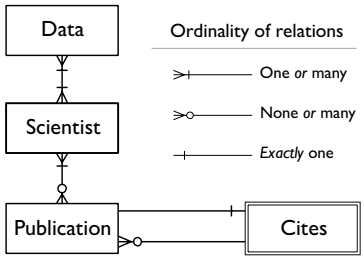

Figure 2.2: The relationship between tables is often visualised as an entity-relationship diagram. Each box represents a table, and the connections between entities represents their relationship. Connections have a specific ordinality, arising from the structure of the data, for instance, a publication record needs to refer to several citations, while each citation must only refer to a single publication.

Any set of data may be modelled by one, but more usually, many tables. The structure of the overall data then defines the relation of these tables to each other. The relationships between tables are specified by connections between specific attributes in each table. An example of this could be a database intended to keep track of a group of scientists and the data and publications they generate. The organisation of relations in a database is often visualised as an entity-relationship diagram (ERD), Figure 2.2 shows a simple ERD for the above example. Such a database could be structured as a set of three tables, one listing scientists, one data, and one publications. Each *scientist* record (one line in the *scientist* table) may be related to one or more records in the *data* table and may also related to any number of *publication* records. Each *publication* record must be related to at least one *scientist* record, and by the addition of a *cites* table, may also be related to any number of other publications.

When implemented, the relationships in a database take the form of *keys*, attributes that directly relate one record in a table to records in another table. Taking our above example, each *scientist* record would have a unique ID, and any *data* or *publication* record that wished to refer to it would have a record of the ID.

The implementation and optimisation of relational-databases is a well-established field (Date, 1986), with its own terminology that sometimes clashes with terms in other fields. In particular the verb *normalisation*, commonly used in reference to standardising metabolic profiles (Chapter 4), refers to a specific form of progressive table optimisation in the database world. Any database may be said to be in one of several, cumulatively more rigorous normal-forms depending on the degree of redundancy present. Normalisation serves to protect the database from corruption by limiting the situations in which the same data stored in two separate tables may be in an inconstant state. By reducing duplication, a well normalised database is also more storage efficient, although this is usually only of peripheral concern.

2.2.2 Introduction to structured query language

Structured Query Language (SQL) is a database management language →, initially developed in the 1970s by Donald D. Chamberlin and Raymond F. Boyce (Date, 1986). SQL was intended to provide a unified, interactive method of accessing a database, regardless of the specifics of the underlying storage implementation. SQL is implemented as the query language in a number of the leading Database Management Systems (DBMS – the software running a database) from several vendors including IBM, Oracle and Microsoft. While the standard fragmented in the 30 years following its creation, the query languages of most popular current DBMS descend from the SQL specification →, with the basic syntax remaining mostly constant and each DBMS adding different custom extensions to the language.

Standard: ISO/IEC 9075 SQL

This can be seen in the names of current DBMS, including: MySQL, Oracle, PostgreSQL, T-SQL & SQL PL. (list from http://en.wikipedia.org/wiki/SQL#Procedural_extensions).

2.2.3 Selecting a DBMS

Many DBMS, such as Oracle are targeted at large organisations needing to store many millions of records, and have licensing fees of many thousands of U.S. Dollars, depending on the specific configuration. Despite the proven capability and reliability of Oracle and its ilk, such expense is obviously unacceptable for a small-scale academic project. Luckily there are several open-source implementations of SQL, each of which provides a high quality DBMS, free for download. Chief amongst the open-source implementations of SQL is MySQL from Sun Co. MySQL is a technically highly capable implementations of SQL, favoured for use in web-based database applications →(Suehring, 2002). With minimal set-up costs and an active community available to provide support and tools, MySQL was an obvious choice for use as the database engine for the COMET project.

MySQL is used as the DBMS for such popular websites as Google, Wikipedia and Wordpress, the popular blogging system. It is also expanding into the enterprise business market following its acquisition by Sun Co.

2.3 The COMET Project

2.3.1 Extent of the COMET project

As outlined in § 1.7 the COMET project was a large scale application of metabolic profiling methods to *in vivo* toxicology experimentation. A total of 146 studies were run by six compa-

nies, ranging from the dosing of model toxins, to dietary and surgical interventions. In total over 30,000 urine samples and 7,000 serum samples were generated, each with associated NMR spectra and clinical chemistry measurements. Liver and kidney tissue were also collected and assessed histopathologically, while gross morphological observations were recorded for all animals.

The intention of the COMET database was to store metadata from the project, that is, contextual information for the samples and data acquired along with the main thrust of NMR data-generation in the project. These metadata include numerical values, such as clinical chemistry measurements, housekeeping information such as sample tracking codes, and subjective observations such as comments from scientists participating in relevant sections of the project.

While consisting of many diverse data types, the COMET data-set is bound in a strict hierarchy of relations. At the root of the hierarchy, the project itself gives rise to a set of studies, each of which then contains a group of experimental animals, that in turn produce a set of samples of various types.

Excepting NMR spectra, all the COMET metadata were generated at the participating companies or their designated experimental houses. These data were then reported to Imperial College as Microsoft Office Excel spreadsheets, with the relevant values filled in by the experimental technicians. A major aspect of this work then consisted of the standardisation of these sheets, as despite the experimental protocol, data were not always reported correctly. For instance the units of measurements may have been non-standard, or simply not reported. These corrections were in addition to an examination for gross mistakes.

Once the metadata had been checked and standardised, the Excel sheets were exported into comma-separated text files using a custom macro written in Microsoft visual basic. These data files were then imported into a local mirror of the database and checked for consistency. The local database was then synced to the main database driving a website, once a week.

2.3.2 Design of the COMET data tables

The job of the COMET database is to successfully capture the complete data set in a manner that would allow overarching analysis and comparison of the information within. The COMET database was custom designed to capture the full structure of the COMET project. Full technical details of this the implementation can be found in Appendix C, while Appendix C.3 lists the complete range of data acquired and stored in the COMET database.

Figure 2.3 gives a summary of the structure of the tables (*relations*) used to model the COMET database. In this section we will see an overview of the table structure and relations in the COMET data-set, for complete implementation details for each table, see Appendix C.3.

The database is centred about the *study* table (Table C.1); records in this table represent an individual COMET study and hold information general to each study as a whole. The table is keyed to the unique study name used to identify a study, a three character code of the form *xyy*, where *x* is a single letter identifying the COMET participant responsible for the study, and

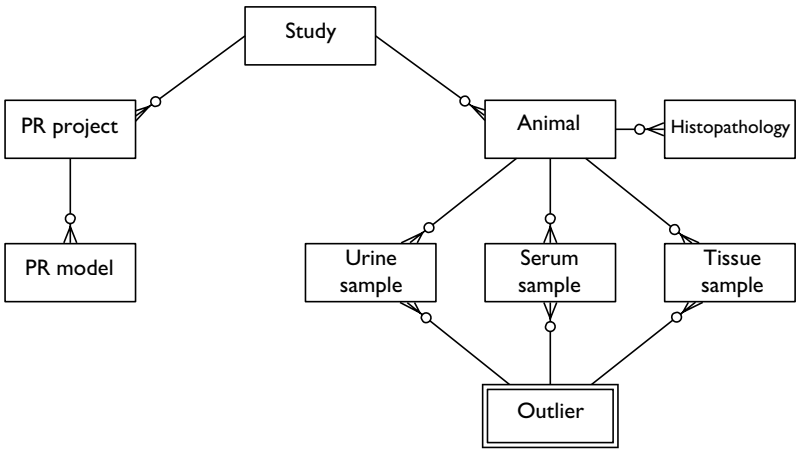

Figure 2.3: Entity-relationship diagram, showing the relationships between data tables in the comet database. Each box represents a table in the database, ordinality is represented as in *f* 2.2.

yy is the individual study number for that company, e.g. 101, the first study rum by Eli Lilly. Additional information includes: the identity of the compound dosed →; details of dosing, such as vehicle composition, dose-route and volume; the number of animals involved in the study and their supplier, species and strain; details of the personnel involved with running the study; Finally any general comments the study director may have had regarding the *in vivo* aspect of study were recorded along with the overall pattern recognition findings.

Or the nature of a dietary or surgical intervention.

Subordinate to the *study* table, the *animal* table (Table C.2) details individual animals. Each animal record is associated with a single, specific study by the study name. Animals are uniquely identified across the entire project by a numeric animal id, and within their study by an animal number. The additional information in the table can be broken into two approximate groups: data on animal health, such as: age; sex; observations on its pre- and post-dose condition; weight, and information regarding treatment, such as: dose-groups and levels; and sacrifice times.

Histopathological examinations of tissues from each animal are recorded in the *histopathology* table (Table C.3). Histopathological examination was carried out on tissues predicted to show damage as a result of the dosed compounds expected mode of action. Individual histopathologists scored the tissues using a simple controlled vocabulary in an attempt to standardise the results. However, considerable divergence in the use of the vocabulary was observed between individual histopathologists, many of whom took advantage of the ability to add their own terms to the reporting sheets. This significantly limits our ability to compare histopathology across studies, as similar lesions may be scored or described differently.

Each record in the three sample tables: *urine*, *serum* and *tissue* (Tables C.6, C.7 & C.8) is associated with a single animal and records information regarding the samples taken from this animal. Records in each of these tables reference the animal id of the animal the sample was

taken from. Each record contains additional metadata regarding the timepoint of collection, the sample volume or weight, various clinical chemistry measurements (depending on the sample type, see Appendix C), tracking information, comments and reduced-resolution NMR spectra, if these had been recorded or generated.

The sample tables show probably the greatest acquiescence to practicality at the expense of formal correctness of implementation in the COMET database. Conceptually there is no need to split the sample types into three separate tables, all sample information could be stored in one table, with the different clinical-chemistry data acquired accounted for by separate tables for each measure (each consisting of a measurement and a reference to a sample). However, by requiring multiple table-joins for both the loading of data into the database and the most anticipated form of data-extraction, extracting all the clinical chemistry, such a design would have significantly complicated the database implementation. Therefore each data-type was modelled individually, explicitly incorporating all the relevant clinical-chemistry measurements as attributes of the table.

The pattern recognition *project* and *model* tables (Tables C.4 & C.5) encapsulate the set of pattern recognition models made of a study. Each PR project represents one set of spectral data, with certain regions of the spectra removed, and certain regions combined. Typically two projects would be generated for each study, an initial project including the entire set of NMR data, and one with the signals attributed to drug-related compounds removed, to focus on endogenous metabolic changes. Each model represents a multivariate model constructed with the data in a specific project. Generally this would include an overview PCA model of the entire data set, followed by more specific models excluding outliers or focusing on a particular aspect of the data set. Attributes of the model table include; the model statistics used to assess the quality of multivariate models, such as the R^2 and the Q^2, and the identity and reason for any samples or NMR-regions excluded from the analysis.

2.4 Example Queries on the COMET Database

By placing the COMET metadata into a relational database we are able to considerably simplify the collation of data for further analyses of individual COMET studies and especially meta-analyses across many studies. In this section we shall see some examples ← of SQL queries to the database.

Code is set `monospaced`, with SQL commands set **UPPERCASE BOLD**.

2.4.1 Single table queries

The simplest queries are made by listing all the records and attributes from a single table. We can extract all data stored regarding every study record with the following query:

```
SELECT * FROM study
```

Table 2.1: An example of listing of all records from the study table. Only the first three lines and five attributes are shown for clarity.

study_id	compound	dose_route	vehicle	volume_administered	...
N15	Lead acetate	I.P.	sterile water	10.00	...
S16	atractyloside	I.P.	saline	10.00	...
N18	Acivicin	I.P.	saline	10.00	...
⋮	⋮	⋮	⋮	⋮	⋱

Generating the output we see in Table 2.1. By default the * argument to the **select** command lists all the attributes of the table. This can output a large amount of superfluous information, so to restrict the output we may name the attributes we wish returned. To list only; each study, the compound dosed, and the model species, we specify the names of these attributes when running the following query on the study table and so generate Table 2.2:

```
SELECT study_id, compound, strain FROM study
```

Table 2.2: Listing of selected records from the study table. Additional records have been trimmed for space.

study_id	compound	strain
N15	Lead acetate	Crl:CD (SD) IGS BR
S16	atractyloside	Crl:CD (SD) IGS BR
N18	Acivicin	Crl:CD (SD) IGS BRs
⋮	⋮	⋮

For tables with numeric attributes, we can limit the number of records returned, based on some property of the results, for instance to select all the urine samples from animals with more than a threshold level of glucose in the urine we could run the following query:

```
SELECT * FROM urine_sample WHERE glucose > 50
```

Text values may be queried on the basis of finding an exact match:

```
SELECT * FROM urine_sample WHERE present = 'Y'
```

Here `= 'Y'` will match records containing only a single capital character Y in the *present* attribute.

Alternatively we can use wildcards to search for a range of possible matches. Here, in combination with the **like** operator, the % sign indicates that the database should return all records that have a title beginning with R23:

```
SELECT title, timepoint, alt FROM serum_sample WHERE title LIKE 'R23%'
```

Table 2.3: Wildcards such as '%' combined with the **like** operator allow searches for partial text strings. Additional records have been trimmed for space.

title	timepoint	alt
R23sr01h+024	24	52.00
R23sr02h+024	24	55.00
R23sr03h+024	24	51.00
⋮	⋮	⋮

Generating a list of output such as that in Table 2.3. We can query more than one attribute at a time, based on simple Boolean logic (Boole, 1848). To restrict the results of the above query on the urine_sample table to samples that are present at Imperial College:

```
SELECT * FROM urine_sample WHERE glucose > 50 AND present = 'Y'
```

And to limit the output to only sample name, glucose concentration and reduced NMR spectra:

```
SELECT title, glucose, reduced_spectra FROM urine_sample WHERE glucose >
    50 AND present = 'Y'
```

Producing output as in Table 2.4.

Table 2.4: Selection of attributes from the urine_sample table, filtered by glucose concentration, and presence at Imperial College.

title	glucose	reduced_data
S16r21h+048	57.90	⟨ommited⟩
S16r35h+024	75.55	⟨ommited⟩
S16r30h+024	55.57	⟨ommited⟩
⋮	⋮	⋮

We can also query tables to return only the unique values of a certain attribute, for instance to find each of the different dosing levels for each group in study D19:

```
SELECT DISTINCT dose_group, dose_level FROM animal WHERE study_ref =
    'D19'
```

This filters out all the identical records returned when we select the values of dose group and dose level from study D19, leaving an overview of the unique doses used in the study (see Table 2.5).

Table 2.5: Use of the `distinct` operator to select the unique dose levels from study D19.

dose_group	dose_level
1	0.00
2	10.00
3	100.00

2.4.2 Multiple table queries

For more complicated queries, we must combine the information contained in several tables. In SQL this is achieved with an operation known as a *table join*. A join takes the records from one table and matches them to records in a second table, according to an attribute that relates the tables.

Here we use the study_id and study_ref attributes of the study and animal tables to join the tables:

```
SELECT * FROM study JOIN animal ON study.study_id = animal.study_ref
```

This works by combining the records from both tables in the database, using the attributes of both. Records are joined where the study_id in the study table matches the study_ref in the animal table. The output of a basic table join such as this is a new pseudo-table that exists only as the return from this query, with all the attributes of both tables. Records are only output when a match can be made between the selected attributes, meaning records that do not have a match in the other table are not returned. Conversely, where a record in one table can match several in the other, the single record will be replicated to join each of the possible matches. We can then use the same selection syntax as demonstrated above for a single table to extract only the records and attributes of interest. For instance to select only the animals dosed with hydrazine:

```
SELECT study.study_id, study.species, animal.study_animal_no FROM study
    JOIN animal ON study.study_id = animal.study_ref WHERE
    study.compound = 'hydrazine' ORDER BY study.study_id ASC
```

Above we also introduce the **`order by`** command, which, in association with the **`asc`** option sorts the output table by ascending study_id.

Table 2.6: Selection of attributes and records in a table join.

study_id	species	study_animal_no
D01	Rat	1
D01	Rat	2
⋮	⋮	⋮

Generating the output seen in Table 2.6. Note how the DO1 study record has been replicated to pair every matching animal record.

Sophisticated queries can be constructed from these elements, for example, to identify which companies worked with which species and strains, we can again use the **distinct** operator:

```
SELECT DISTINCT company.name, study.species, study.strain FROM study
    JOIN company ON study.company_ref = company.company_id
```

Which would generate output as in Table 2.7. Note that the Pfizer and Roche records exist only once in the companies table, but the combinations of the **join** and **distinct** operators causes them to be replicated in the output to account for all existing permutations of company, species and strain.

Table 2.7: Export of model species and strains by company

name	species	strain
Novo Nordisk	Rat	Crl:CD (SD) IGS BR
Pharmacia	Rat	Crl:CD (SD) IGS BR
Bristol Myers-Squibb	Rat	Crl:CD (SD) IGS BR
Lilly	Rat	Crl:CD (SD) IGS BR
Roche	Rat	Crl:CD (SD) IGS BR
Roche	Rat	HanBrl:WIST(SPF)
Pfizer	Mouse	B6C3F1
Pfizer	Rat	Crl:CD (SD) IGS BR

In conclusion, the implementation of the COMET database has greatly simplified the analysis conducted in this thesis, particular with respect to extracting the clinical chemistry used in Chapters 5 & 6. Examples of the queries used during the course of this thesis can be found in Appendix C.5.

Chapter 3

Robust Automated Calibration of 1D Serum ^{1}H-NMR

3.1 NMR Calibration

3.1.1 Factors effecting chemical shift

One of the most valuable features of ^{1}H-NMR spectroscopy in structural analysis is that the observed resonance frequencies are extremely sensitive to the local chemical environment of the nucleus (as described in § 1.4). It is this property that causes dispersion of the resonances from the various nuclei in a molecule across the spectrum, facilitating structural characterisation and the differentiation between various compounds in a mixture. However, this extreme shift sensitivity also means that resonance frequencies can be affected by even minor fluctuations in temperature, pH, and the external magnetic field. Hence precise comparison from one sample to another requires frequencies being presented as relative to some internal standard (Wishart *et al.*, 1995). In high resolution NMR spectroscopy-based metabolic-profiling studies such as COMET, hundreds or even thousands of individual spectra must be compared, and typically require automated sample handling and spectral acquisition (Spraul *et al.*, 1994; Keun & Athersuch, 2007) (see § B.2). While the first sample in a run may be manually calibrated, subtle variations in sample temperature often causes variation in the frequency of the lock solvent, which, in turn, introduces small errors in alignment for each individual spectrum (Figure 3.1). This effect must then be corrected for by spectral calibration. Calibration is distinct from spectral alignment procedures, as is seeks to correct the frequency scale for the entire spectrum and not individual resonances.

As outlined in § 1.5, these data are typically analysed by pattern recognition methods such as PCA. Since there is an increasing tendency to perform analyses at the maximum spectral resolution possible, it is crucial that shift errors are removed and the spectra accurately calibrated to prevent this variation influencing pattern recognition (Stoyanova *et al.*, 1995; Brown & Stoyanova, 1996). The number of samples involved lends some advantages to the use of automated algorithms for this process, as manual calibration may be time consuming. Additionally a deterministic algorithm has the advantage of absolute reproducibility that cannot be guaranteed by human calibration, reducing one element of subjectivity in the processing of spectra. This is of great advantage in a project such as COMET, were it is unfeasible for one individual to reference every spectrum acquired.

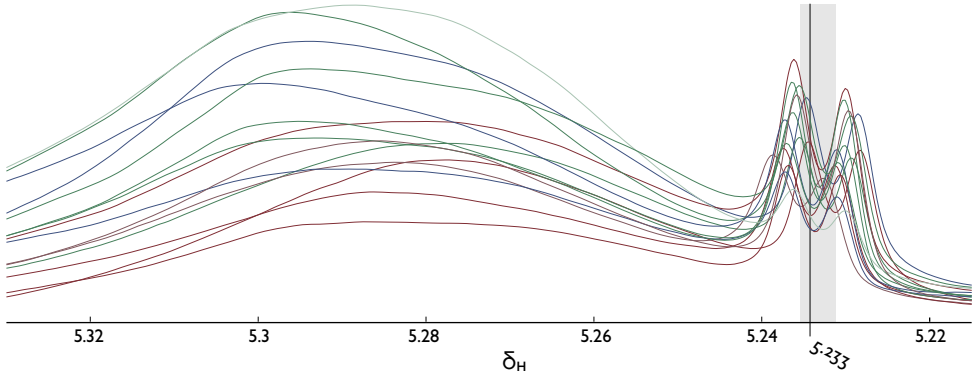

Figure 3.1: Subset of the unaligned spectra, focused upon the α-glucose doublet. Note the overlapping broad lipid-olefinic resonance. δ_H = 5.233 is marked, and the centre of the α-glucose doublet can be seen to shift over a δ_H = 0.004 range (greyed area).

3.1.2 Difficulties with referencing to TSP in serum spectra

TSP has a true shift of δ_H = -0.015 compared to DSS.

Proton NMR spectra of aqueous samples are typically referenced to either 3-(trimethylsilyl)-propionic acid-d_4 (TSP) or 2,2-Dimethyl-2-silapentane-5-sulfonate (DSS), with the Si-Me signal assigned to δ_H = 0 ←. As both TSP and DSS produce a singlet resonance that occupies a sparsely inhabited region of the ^{1}H spectrum, it is algorithmically trivial to locate and reference to either compound by searching for the greatest intensity within a small defined chemical shift range. However, in some routinely analysed biofluids such as blood plasma or serum, the chemical shift of the DSS or TSP resonance becomes highly variable and unpredictable, due to interaction with proteins in solution, preventing the use of these compounds as an internal reference.

Several techniques have been developed to overcome this problem, including the use of a referencing agent: 4,4-dimethyl-4-silapentane-1-ammonium trifluoroacetate (DSA) that is not susceptible to shifting in plasma (Alum *et al.*, 2008) or the use of a microcapillary containing a solution of TSP inserted into the sample tube during acquisition of a spectrum, providing a reference while remaining separate from the sample solution (Ala-Korpela *et al.*, 1994). The latter method in particular has significant drawbacks, such as an increased difficulty ensuring the homogeneity of the magnetic field, and all are limited to new acquisitions and cannot be applied to historical data sets.

3.1.3 Calibration to glucose

As an alternative, in these sample types the anomeric doublet resonance of glucose (typically the α-anomer at δ_H = 5.233) is often used as a substitute reference due to its abundance and positional stability (Nicholson *et al.*, 1995; Farrant *et al.*, 1994), and this was the reference chosen for the COMET serum samples. However, in normal mammalian blood-serum the α-anomeric glucose doublet at δ_H = 5.233 is partially overlapped by broad lipid olefinic resonances at δ_H ~ 5.27

(Figure 3.1). Depending on the precise composition of the sample, the lipid or α-glucose resonance may appear more intense, thus complicating the use of an automated search for the resonance of greatest intensity in locating the α-glucose doublet. Changes in the intensity of the lipid resonance may also result in an apparent shift in the glucose resonance due to contribution of confounding intensity from the lipid signal.

Considered in abstract terms, we are presented with a situation in which we must locate the correct pair of peaks out of a set of at least three, where the peaks of interest are also distinguished by their line-shape. Previous work (Crockford *et al.*, 2005; Weljie *et al.*, 2006) has demonstrated the value of line-shape fitting for identification and quantification in NMR. While this technique would undoubtedly be effective at distinguishing the α-glucose resonance from the lipid, due to the distinctive line-shapes of the resonances, it is computationally expensive and time-consuming.

Historically, techniques based on spectral derivatisation have proved successful as a method of resolution enhancement in various fields of spectroscopy, often being used to resolve overlapping signals (Butler & Hopkins, 1970; Giese & French, 1955). Therefore, we have developed novel, efficient methods of locating a multiplet resonance for calibration, using spectral derivatisation, with an emphasis on simplicity of implementation and speed, to allow mass calibration of archived spectra, as well as potential applicability to other resonances in distinct biofluids.

3.2 Methods and Algorithms

Full details of sample preparation and spectral acquisition may be found in Appendix B.3. Briefly, 77 standard water-suppressed ^{1}H-NMR spectra of human blood-serum were acquired as part of a separate study, then phased and baseline-corrected with an in-house routine, however the chemical shift drift was not corrected from the raw instrument files.

In order to allow an automatic, robust selection of the correct resonance for calibration, it is necessary to reliably identify the pair of resonances that comprise the glucose doublet. We have made use of spectral transforms that take advantage of the sharp signal of the glucose doublets, to distinguish the doublet from the ill-defined broad resonance of the interfering lipid resonances.

The 'power-derivative' transform takes the sum of the squares of the first derivative of the real and imaginary parts of the spectrum. Taking the first derivative of the spectrum has the effect of down-weighting the strength of signals that increase slowly, such as the broad lipid-resonance, as compared to the strength of sharper resonances such as the glucose doublet (Figure 3.2A). We have used a simple difference to approximate the derivative *i.e.* the value of the derivative at point x_i is the difference between this value and the subsequent data point x_{i+1}. This results in a well-defined glucose doublet that is significantly stronger than the adjacent lipid resonance. The centre of this doublet can then be located as the midpoint between the two highest points in the spectrum, using the following algorithm:

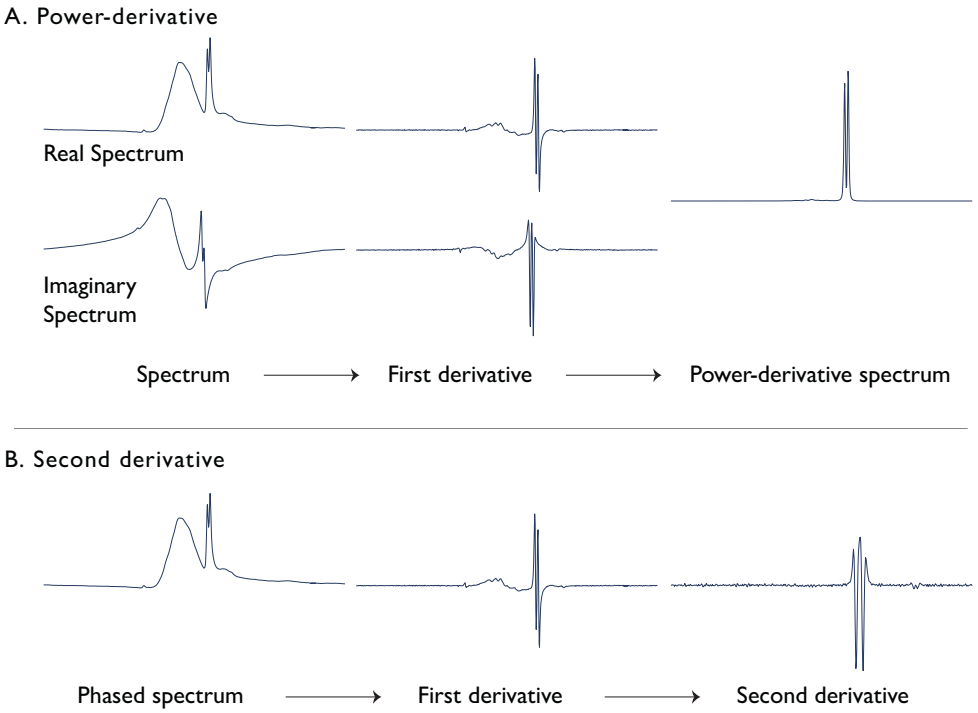

Figure 3.2: Application of spectral derivatisation to magnify sharp features.
A, The sum of squares of the approximate derivative of the real and imaginary parts of the spectrum gives the power-derivative spectrum. The broad signals are attenuated, while sharp resonances are boosted. B, First and second approximate derivatives of the real part of an NMR spectrum. Magnitude of the first derivative indicates the gradient of slope in the real spectrum, magnitude of the second derivative indicates the rate of change of the slope. (magnitude not to scale).

1. Select a region of the spectrum $\delta_H \pm 1$ (for the current study, can be defined by user) around the expected position of the α-glucose doublet at $\delta_H = 5.233$
2. Take the power-derivative of this region:
   $\text{PowerDerivative} = f'_{(\text{realSpectrum})}{}^2 + f'_{(\text{imaginarySpectrum})}{}^2$
3. Find the maximum peak in the region
4. Locate the second largest peak in the region
5. Set the midpoint of these two peaks to $\delta_H = 5.233$

While the 'power-derivative' based method allows for reliable detection of the glucose doublet, it still possesses drawbacks. The requirement for the imaginary part of the spectrum limits the use of this method to situations where these data are available. This would mandate the retention of the otherwise unused imaginary part of the spectrum in situations where calibration was expected to be repeated on archived data or require an additional computational

step to regenerate the imaginary spectrum using the Hilbert transform (Zolnai *et al.*, 1990; Shaw, 1976).

To eliminate these limitations, a second method of calibration was developed. By taking an approximate second derivative of the real part of the spectrum, it is possible to obtain a data vector where magnitude indicates the rate of change in gradient of the spectrum (Figure 3.2B). It is then simple to identify the two most negative peaks in the second derivative, which correspond to the tops of the two sharpest resonances in the original spectrum, using the following algorithm:

1. Select a region of the spectrum $\delta_H \pm 1$ around the expected position of the α-glucose doublet at $\delta_H = 5.233$

2. Take the second derivative of this region:
   $\text{SecondDerivaitve} = f''_{(\text{realSpectrum})}$

3. Find the minimum peak in the region

4. Locate the second smallest peak in the region

5. Set the midpoint of these two peaks to $\delta_H = 5.233$

Considered formally, Fourier-transformed NMR resonances have a Lorentzian distribution across frequency space (Claridge, 1999) (Equation 3.1):

$$y = \frac{y_0}{1 + (\frac{x - x_0}{w})^2} \tag{3.1}$$

where: y is the intensity at $\delta_H = x$; the resonance is at maximum intensity y_0 at $\delta_H = x_0$; and w represents the half-height line-width.

Taking the second derivative of this (Equation 3.2):

$$\frac{d^2y}{d(x - x_0)^2} = \frac{2y_0(3(\frac{x - x_0}{w})^2 - 1)}{w^2(1 + (\frac{x - x_0}{w})^2)^3} \tag{3.2}$$

We can then simplify to show the maximum intensity of a peak in the second derivative spectrum is inversely proportional to the square of the half-height line-width (Equation 3.3):

$$\frac{d^2y}{d(x - x_0)^2} = \frac{2y_0}{w^2} \tag{3.3}$$

This transform favours sharp resonances, and thus these are more pronounced in the derivative spectra compared to broad signals.

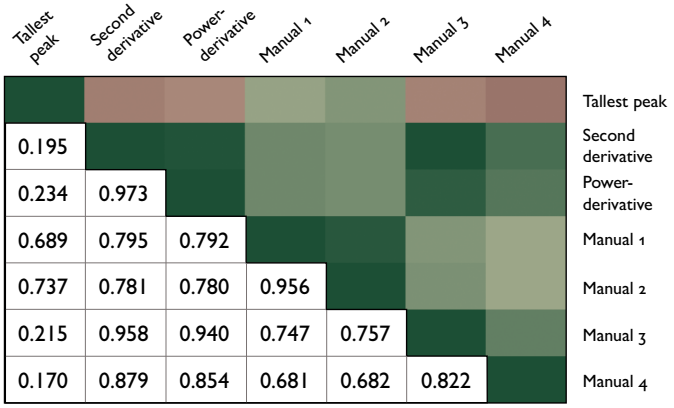

Figure 3.3: Heatmap of pairwise Pearson's correlations between the values of the corrective offset each method applied to calibrate the α-glucose doublet. Both computational methods based on both peaks of the doublet can be seen to be very consistent with each other, while manual calibration shows a greater degree of variation between spectroscopists. The window region for selecting the 'tallest peak' was δ_H = 0.006 compared to δ_H = 2 for the derivative-based methods.

3.3 Results and Discussion

3.3.1 Comparison of calibration methods

Following phasing and baseline correction, but prior to calibration, the centre of the α-glucose doublet at $\delta_H \sim 5.233$ was observed to vary over a range of δ_H = 0.004, with a standard deviation of $\delta_H = 1 \times 10^{-3}$ within a set of 77 ^{1}H-NMR spectra of human sera (see Figure 3.1). We then compared the accuracy and stability of several methods of calibration:

- manually using the software Xwin-NMR
- detection of the largest resonance in a given region ('tallest peak')
- the 'power-derivative'
- second derivative

Manual referencing was repeated from the unreferenced spectra by four different spectroscopists. The computational methods, producing deterministic results, were each run once on the test data. The 'tallest peak' method was compared with our algorithms because it represents the only widely used automated means of spectral calibration. The window region selected within which we searched for the tallest peak was chosen to include the entire visible α-anomeric doublet for all the uncalibrated spectra, but excluding the broad olefinic resonance (typically 0.06 PPM wide). This was narrower than the region used for the derivative approaches since the 'tallest' peak in this region is frequently the olefinic resonance, leading to unrealistically poor spectral calibration.

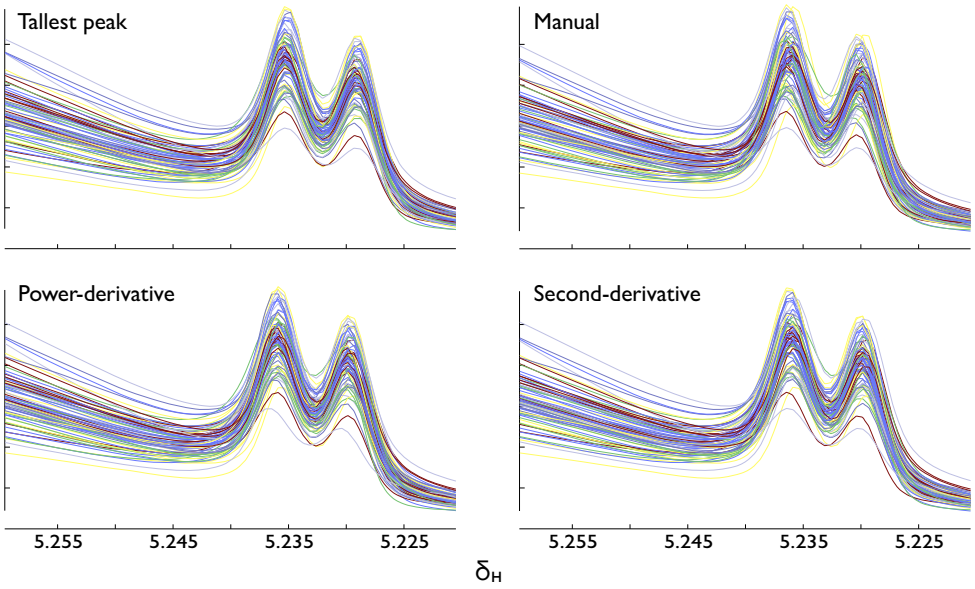

Figure 3.4: Comparison of the position of the glucose doublet in 77 ^{1}H-NMR serum spectra after referencing with the: tallest-peak; manual; power-derivative; or second-derivative methods.

Following each method of calibration, the offset applied to move each spectrum into register was calculated. This value was then used to generate a pair-wise Pearson's correlation matrix comparing the correction each method applied to each spectrum (see Figure 3.3). Examination of the correlation matrix shows the comparative reliability of each calibration method. The two derivative methods produce remarkably similar calibrations, however the greater robustness of the second derivative methods to differences in spectral composition makes it the best choice for use in the referencing of a large number of spectra. While each of the three computational methods tested produce completely reproducible results for each spectrum, the manual calibration can be seen to produce results that are variable, both by comparison to the computational methods and the calibration by other individuals (see Figure 3.4). Following manual calibration the α-glucose doublet is observed varying over a range of $\delta_H = 0.005$, with a standard deviation of $\delta_H = 5 \times 10^{-4}$ across all the manually calibrated datasets, a figure not appreciably more consistent than prior to calibration.

3.3.2 Effect of signal-to-noise ratio

One limitation of derivatisation-based methods is their sensitivity to noise. To estimate the effect of the signal-to-noise (S/N) ratio on the performance of the second-derivative algorithm, a simulated set of doublets of varying intensity were generated by the addition of two Lorentzian peaks. Instrument-noise recorded from an empty, high-field region of one of the previously acquired ^{1}H-NMR spectra was then added to this simulated data and the second derivative was

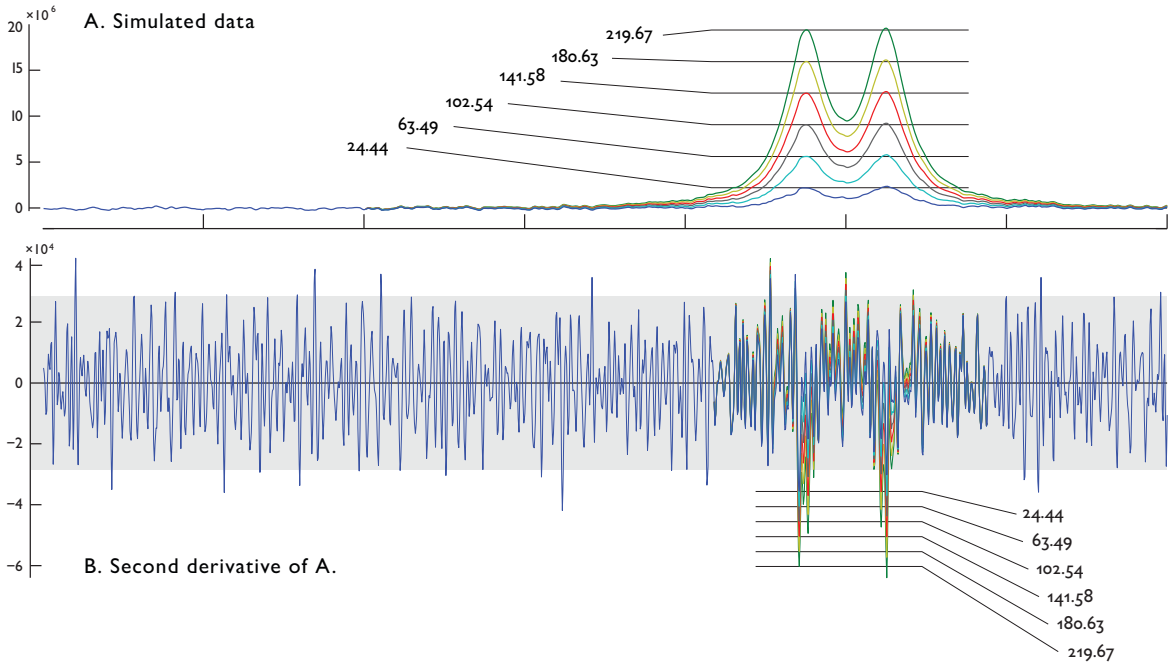

Figure 3.5: Effect of a low signal-to-noise ratio on the second-derivative of a simulated doublet. **A**, The simulated doublet with signal-to-noise ratio indicated. **B**, The second derivative of these simulated spectra. The greyed area indicates twice the standard deviation of the noise about the mean in the second derivative.

calculated (See Figure 3.5). From this dataset, the point at which the doublet in the second derivative could no longer be reliably identified was observed to lie between S/N ratios of 100 and 140 (S/N = max signal intensity ÷ standard deviation of noise). Typically with the sample types of interest, such as blood-serum, the signal-to-noise ratio for the α-glucose doublet is an order of magnitude greater than this, allowing good confidence that calibration performance will be sufficiently robust. Additional 'safety' checks could be easily added to the algorithm to provide confidence in borderline situations, for example confirming that the coupling constant for the doublet ← lies within the expected range.

The distance between peaks in a multiplet.

Given that the problem is largely one of resolving resonances based on line-width, appropriate apodisation would be a possible alternative in some cases. However, this approach requires the manipulation of the time domain data, which is inconvenient for archived resources, typically worsens signal-to-noise and can introduce artefacts into the spectrum confounding the detection of resonances. The 'tallest peak' approach, while clearly successful in some circumstances is always highly dependent on a subjective selection of the search window region, and is fundamentally flawed since a doublet resonance is in principle symmetric with no 'tallest' peak. The success of this approach for serum and plasma NMR spectra relies upon the consistent abundance of the overlapping olefinic resonance.

3.4 Conclusion

Accurate and automated calibration to the α-glucose doublet at $\delta_H = 5.233$ was demonstrated for 1D ^{1}H-NMR spectra of human sera, using two derivative-based methods. It was also shown that derivative-based methods were robust for typically observed signal-to-noise ratios and more consistent than manual calibration of 1D ^{1}H-NMR, particularly when using the second derivative. This makes this method ideal for referencing large collections of serum spectra such as COMET, and so this algorithm has been used in the calibration of all COMET spectra presented in this thesis.

These methods are also applicable to other biofluids in which glucose is present, such as cell and tissue extracts, and with minor alteration, to calibration using other well-defined resonances of any multiplicity. The second-derivative technique is also solely reliant on the real part of the Fourier-transformed spectra and may be simply applied to archived spectra, without the need for reprocessing of the samples.

The contents of this chapter have been adapted for publication in Analytical Chemistry (Pearce *et al.*, 2008).

Chapter 4

Normalisation of Metabolite-Profiles

4.1 Aims of Normalisation

Recorded metabolite profiles, measured from biofluid data, are often subject to systematic variations in intensity, across all measured variables in a sample. The perturbations that lead to these variations arise from a wide range of factors, including, but not limited to: changes in sample volume or concentration, such as the diuretic effect of a dosed compound on urine volume; changes in instrumentation, such as the gradual loss of sensitivity in MS due to contamination of the source; and the need to compare data-sets acquired on several instruments with differing sensitivity profiles. All these interferences may have a common effect, causing a consistent increase or decrease in signal intensity across an entire profile. This perturbation can then dominate the true metabolic changes, masking the biologically-relevant differences between profiles.

The normalisation of metabolite profiles is a process of row-wise standardisation, intended to remove this spurious between-sample variation from these data. This process is distinct from database normalisation (§ 2.2) and OSC, as discussed in Chapter 5, this is because the interference normalisation is intended to remove is assumed to be similar for all variables in each sample, while OSC techniques remove systematic variation that affects each variable to a different extent, which may leave some variables unaffected.

The common operation of all normalisation techniques is a row-wise scaling by a scalar normalisation factor f:

$$x_{\text{norm}} = \frac{x}{f} \tag{4.1}$$

for each metabolite-profile x, in an attempt to remove the interfering variation. Normalisation may be viewed as an attempt to remove the biases in sample acquisition, resulting in a profile that reflects the 'true' composition of the sample as closely as possible. In addition to using normalisation in an attempt to cause each metabolic-profile to represent the true experimental situation more accurately, for instance; adjusting for excretion rate or removing instrument effects on sample intensity. Normalisation can also be used to increase the interpretability of the data, removing variation that is not expected to be related to any change in metabolic state from the data-set.

One risk of any normalisation method is the possibility of introducing spurious correlations

to the data. Pearson (1886–87) first raised the possibility of the introduction of significant, yet spurious correlations between variables, arising due to variables being expressed as ratios with a common divisor. More recent work by Kim (1999) has allowed the precise extent of introduced correlation to be calculated, allowing an estimate of the extent of any problem.

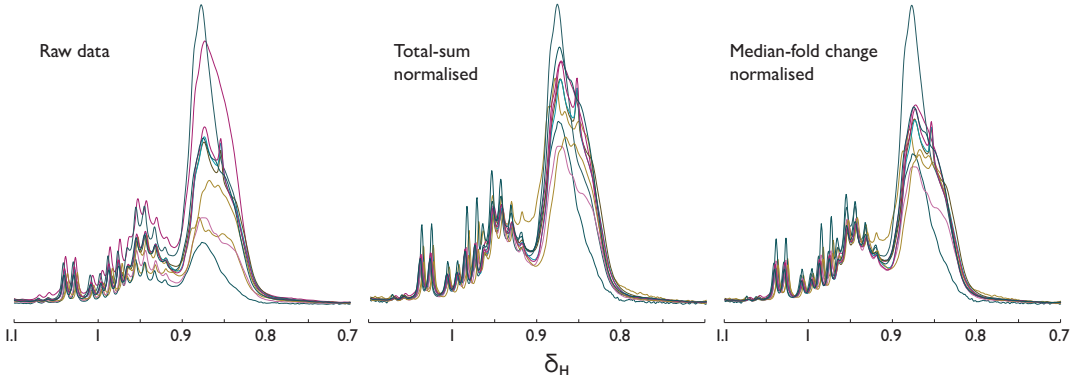

Figure 4.1: Comparison of the two most common normalisation methods, applied to 10 randomly selected CPMG spectra of serum. The large methyl resonance can be seen to dominate the total-sum normalisation, while MFC results in a more consistent set of intensities for the more numerous smaller resonances.

The chosen method of normalisation can have a decisive effect on the outcome of any further analysis, whether by automated pattern-recognition and statistical methods, or manual inspection of the data. This makes it imperative to consider both the nature of the data-set and the intended interpretation of the data, prior to selecting a method of normalisation. Figure 4.1 shows the effect of two common forms of normalisation on the intensities of 10 randomly selected CPMG spectra of serum from COMET.

4.1.1 Established methods

There are a great many established methods for normalising metabolite profiles. Every method of normalisation is based on assumptions as to the nature of the data, which may limit its applicability to certain types of data, or lead to drawbacks under certain conditions. Some of the most significant methods along with their operating assumptions are as follows:

Total area or total sum

Normalisation to total area, or its close counterpart total sum, are two of the longest established methods of normalisation in metabolite profiling. Both methods apply a basic closure constraint onto each individual profile, setting the integrated area under a spectrum $f(s)$, or the total sum of all data-points to a constant value for all samples:

$$x_{\text{norm}} = \frac{x}{\int_a^b f(s)ds} \tag{4.2}$$

for total area, over the spectral range $a - b$, or:

$$x_{\text{norm}} = \frac{x}{\sum_{i=1}^{j} x_i} \tag{4.3}$$

for total-sum of a profile of j variables. Typically area normalisation is used for spectral data, while total sum is applied to categorical data such as ion-intensities. As the sampling resolution of spectral data increases, the area method converges with the total sum calculation. Indeed with the dominance of digital sampling of spectra, there are now few truly continuous sets of data. This has resulted in the two terms often being used interchangeably, with total-area used to describe what is actually total-sum normalisation.

The imposition of this closure constraint onto each profile has a significant drawback, in that it assumes that the total intensity of each profile is invariant. That is, individual variables may vary in intensity, but the aggregate intensity will remain constant (Johansson *et al.*, 1984). This is a risky assumption for profiles of biofluids, and is clearly wrong in situations, such as the dosing of NMR-visible compounds in toxicological studies, where some resonances may only be expected to be observed in a subset of the profiles. As a result, in situations were the overall intensity is varying, area or sum normalisation will introduce a degree of self-correlation between all variables within the data-set. While the strength of this correlation is inversely related to the number of variables in the sample, it can introduce significant artefacts to pattern recognition (Chayes & Trochimczyk, 1978; Rietjens, 1995).

Euclidian norm

Another closure-based method, the Euclidian- or Root Mean Squared- (RMS) norm, treats each j-variable profile as a vector in j-dimensional space, normalisation then sets this vector to unit length for each profile:

$$x_{\text{norm}} = \frac{x}{\sqrt{\sum_{i=1}^{j} x_i^2}} \tag{4.4}$$

Although this makes for a simple method of normalisation, the assumption that a constant Euclidian norm provides any form of meaningful normalisation for metabolite profiles has been shown to be false by Torgrip *et al.* (2008).

Probabilistic quotient / median fold-change

Probabilistic-quotient normalisation, described by Dieterle *et al.* (2006) and its closely related counterpart, median fold-change (MFC), are based on identifying a common invariant set of variables in the data. This is achieved by calculating the fold-change of each data-point in a profile to that in a reference profile:

$$d_i = \frac{x_i}{r_i}$$

$$x_{\text{norm}} = \frac{x}{\tilde{d}} \tag{4.5}$$

The reference r, is a manually selected 'golden' spectrum for probabilistic-quotient normalisation, while the median of the data, $r = \tilde{X}$, is used in median-fold-change. The median fold-change of each spectrum from this reference is then taken as the normalisation factor.

Both these methods work under the assumption that the dominant variation between profiles represents the systematic interference to be removed. While both methods are not subject to the closure problem of area normalisation, as median-based methods they are only applicable to situations in which the median of the data is in fact invariant. Beyond this threshold the median value will no longer provide a stable target for normalisation. Additionally, as both methods are dependent on the relationship between variables across the data, meaning they may be negatively affected in situations, such as shifting resonances in NMR, where signals may move and be detected in one of several variables in subsequent observations.

Reference feature

A feature present in all samples at a known or constant level may be used as a normalisation target. All other variables in the sample may be expressed relative to the intensity of this feature, which may be judged according to either the maximum value or its area. Often this may be a standard compound added to the sample in known concentrations, to allow for quantification of other components in the sample matrix. For NMR, the referencing standards TSP and DSS often provide this, along with chemical-shift referencing.

In some cases an internal standard, already present in the sample matrix, may be used. This is useful either where quantification is available from another source, such as clinical-chemistry, or a good understanding of the underlying biology allows the selection of a compound expected to be stable across the data-set. For example, as far back as Rose (1933), creatinine has been used as a reference compound in urine clinical-chemistry, under the assumption that it scales well with muscle mass. In situations where it is desirous to compare metabolite profiles obtained by NMR or MS to clinical-chemistry measures (Viau *et al.*, 2004), the profiles may be normalised to creatinine.

Normalisation to a reference feature is required when absolute quantifications for constituents of the sample are desired. This form of normalisation is limited to situations where a suitable external or internal standard exists, and may also add complexity to sample preparation. Also, as each variable is expressed as a ratio of the reference feature, the above remarks concerning spurious correlations are relevant, and should be taken into account.

Histogram matching

Described by Torgrip *et al.* (2008), histogram matching is a method specific to spectral-data adapted from the field of image-processing, where the technique is utilised to standardise the brightness of grayscale images. The technique attempts to match the histogram of intensities

in a spectrum as closely as possible to that of a reference spectrum, suggested to be the median spectrum of the data by the authors.

This method has the advantage of being less tied to the precise shift of resonances than the MFC derived methods. To its detriment, the inability to directly calculate the normalisation factor requires the use of either an exhaustive search, or a suitable search-algorithm (as used in the paper) to derive the optimal factor. The former is limited by the length of time required for the exhaustive search, which rises proportional to nf, where n is the number of samples and f is the number of dilution factors to test →, and the latter by the possibility that the search function may fail to converge to the true minima on a case-by-case basis.

In other words a doubling of the sample number results in a doubling in the time required for normalisation.

4.1.2 Entropy and statistical models

Given that a further goal of normalisation in metabolic profiling is often also to concentrate the variation of the data-set into the fewest number of variables, we have developed a novel method of normalisation, applicable to both spectral and categorical data. This development of normalisation takes its inspiration from the statistical-thermodynamic concept of entropy.

The concept of entropy was developed by Rudolf Clausius during the 1850s and '60s, in the field of classical thermodynamics, as a description of the energy state of a system. In the late 1870s Ludwig Boltzman and later Josiah W. Gibbs, developed the modern statistical-thermodynamic definition of entropy, and the field was further expanded by Shannon (1948) who took the concept beyond the context of heat and energy to describe the uncertainty in communication channels.

Entropy can be considered a measure of the uncertainty in a system, or in informational terms, the amount of information required to establish the content of a message unambiguously. Watanabe (1981) has described pattern recognition techniques in general as methods of minimising the entropy of a data-set. Specifically, Watanabe references his own previous work (Watanabe, 1965) describing the Karhunen–Loeve expansion and by extension, PCA in terms of a process of entropy-minimisation, where the entropy of the system is defined as the degree of participation of the j variables in representing a data-set of n points:

$$S = -\sum_{i=1}^{j} p_i \log(p_i)$$

where

$$p = \frac{1}{n}\sum_{\alpha=1}^{n}(x_i^{(\alpha)})^2 \qquad (4.6)$$

$$\sum_{i=1}^{n}(x_i^{(\alpha)})^2 = 1$$

We might therefore expect by minimising a suitable entropy measure with respect to a set of metabolite profiles, we may be able to generate more robust statistical models.

4.1.3 Apportionment-entropy based normalisation

Apportionment-entropy, as introduced by Kawachi *et al.* (2001) is a system for describing the uncertainty of a set of events falling within a defined interval. Initially used to describe the temporal variability of rainfall (Maruyama *et al.*, 2005; Mishraa *et al.*, 2009) in Japan, here we adapt the concept to metabolite profiles, interpreting the difference between two profiles in terms of the apportionment of the total difference into individual variables.

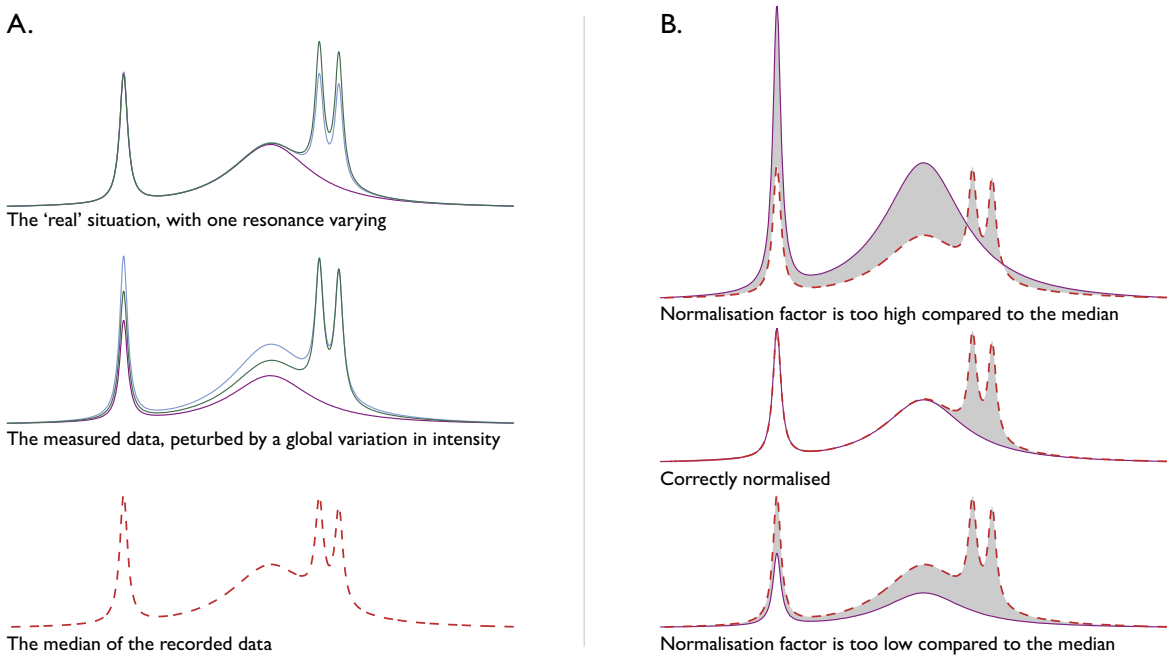

Figure 4.2: **A**, For any set of data which is acquired with random permutations in intensity, a median spectrum can be generated. **B**, The apportionment-entropy with respect to this median can then be calculated for each individual sample, for a range of normalisation factors. Here the apportionment-entropy is indicated by the area shaded grey, where this is lowest, the maximum number of variables are aligned.

To normalise an entire data-set X, each profile in X must have the apportionment-entropy, S, minimised with respect to some reference profile, effectively finding the minima in S space. S between two profiles of j variables is calculated according to an adaption of Equation 4.7:

$$S = -\sum_{i=1}^{j} d_i \log_2(d_i) \tag{4.7}$$

with:

$$d = \frac{|r_i - t_i|}{\sum_{i=1}^{j} |r_i - t_i|} \tag{4.8}$$

where r is the reference profile and t is the profile to be normalised. Apportionment-entropy is lowest when all variation between the reference and the profile to be normalised is contained in one variable. As the total intensity of one profile is adjusted with respect to the other, the value of S will vary as seen in Figure 4.2B, resulting in a distribution of values as seen in Figure 4.7. I have chosen to use the median profile of the data-set, $\tilde{X}$, as the reference profile, but a manually selected 'golden' profile may equally be used.

Figure 4.2 provides a schematic view of the process of normalisation by apportionment-entropy. The steps required are as follows:

1. Generate the reference profile, by calculating the median profile.
2. For each spectrum, estimate the normalisation factor that minimises S with respect to the reference.
3. Return the data set with each spectrum corrected by the calculated normalisation factor.

The minimisation of S may be achieved either by an exhaustive grid-search, calculating S for every possible dilution factor, or potentially *via* an appropriate minimisation function. In this work I have been unable to find a minimiser that will reliably find the global minima of S, so the grid-search method has been used.

Normalisation by apportionment-entropy aims to minimise the variation across all samples, for the greatest possible number of variables. In this, it is analogous in objective to the MFC method. However in real world samples, the difference in weighting of small differences in intensity mean the two methods usually produce close, but non-identical normalisation factors. This can be seen in comparisons in § 4.2.2 and the entropy profiles in Figure 4.7.

4.2 Comparison of Normalisation Methods

4.2.1 Simulated NMR spectra

To compare the performance of the four most common methods of normalising biofluid NMR in a well-characterised situation, simulated NOESYPR1D spectra of serum were generated in MATLAB →. A representative, experimentally acquired NOESYPR1D spectrum of rat serum was peak-picked by selection of all features with an intensity greater than ten-times the standard deviation of the noise. The position of these features was then used as the position of the simulated resonances, while the integral of each feature provided the default intensity. The intensity of each simulated resonance could be independently varied according to a log-normal distribution, with a variable base to the exponent. Peak intensity $i = cb^r$, for initial concentration c,

The Mathworks Inc., Natick, MA.

exponent base b and value r chosen randomly from a normal distribution. Simulated spectra were generated with 20,000 data points between δ_H = -1 & 10, according to the following algorithm:

1. Create an empty matrix X, of the required dimensions.
2. Generate a set of random intensities by $i = cb^r$.
3. Use these intensities to generate a set of Lorentzian peaks centred at $\delta_H = j$.
4. Add these peaks to X.
5. If further peaks are to be produced repeat steps $2-4$.
6. Multiply each simulated spectrum by a random dilution factor between 0.1 and 10.
7. Add random noise with a standard deviation of 1/500 of the mean intensity of the resonances used to define the data-set.

This resulted in spectra such as those seen in Figure 4.3. To estimate the effect of the degree of variation within a spectrum on various normalisation methods, each simulated resonance could be individually switched between a random log-normal distribution and an invariant intensity across the simulated data-set.

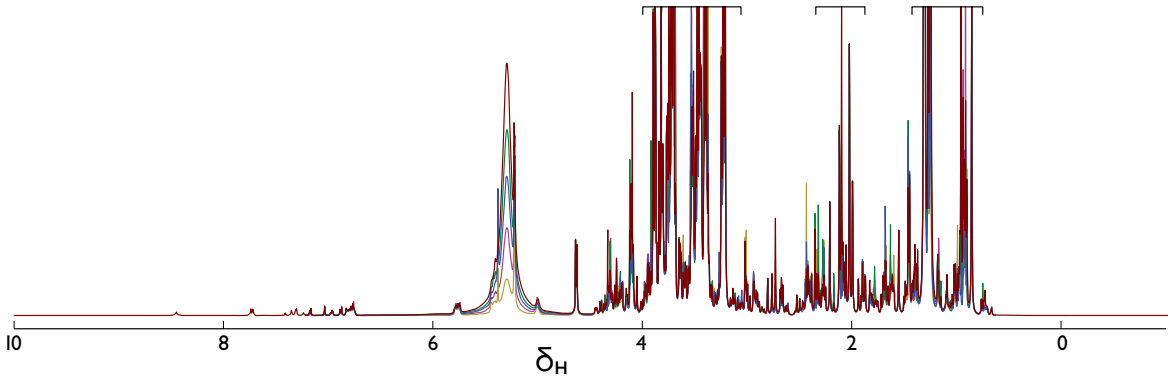

Figure 4.3: Simulated spectra of serum before multiplication by a random dilution factor. Here 1/3 of the resonances in each spectrum are varying with a log-normal distribution, while the remainder are constant.

These simulated data have one primary limitation, due to the random variation in intensity of the simulated resonances, spectral area and median are on average constant for any data-set. This can most clearly be seen in Figure 4.4 where spectra that are almost entirely random still produce a correlation of normalisation factor to dilution of 0.5.

Three of the most common methods of normalisation, area, MFC and histogram matching were compared to the new apportionment-entropy method by systematically increasing the

degree of variation in a simulated data-set and correlating the reciprocal of the normalisation factor of each method with the true dilution-factor applied. In this simulation two axes of variation in the data were gradually increased. First, the number of simulated-resonances varying within the data-set was increased in twelfths of the total number of simulated-resonances, between 1/12, where the majority of resonances were not changing in intensity, to 12/12 where every resonance in each spectrum was varying in intensity randomly. Second, the extent of the variation within each resonance was raised, by an incrementally increasing in the exponent base b, of the log-normal distribution used to generate the intensity values.

The simulation was repeated 50 times for each combination of number of resonances varying and exponent base, and the correlation between the true and predicted normalisation factor for each simulation was averaged to the mean, as illustrated in Figure 4.4. The results for histogram-matching are clearly dubious and in no way match the claims made for the method. This appeared to be due to an inability of the MATLAB built in function minimiser, `fminbnd()`, to converge on the correct minima of the matching function. As the implementation used here follows the algorithm as published as closely as possible, it was decided not to rerun the histogram matching using an exhaustive search to minimise the objective function and therefore the histogram matching results will be disregarded for the remainder of this analysis.

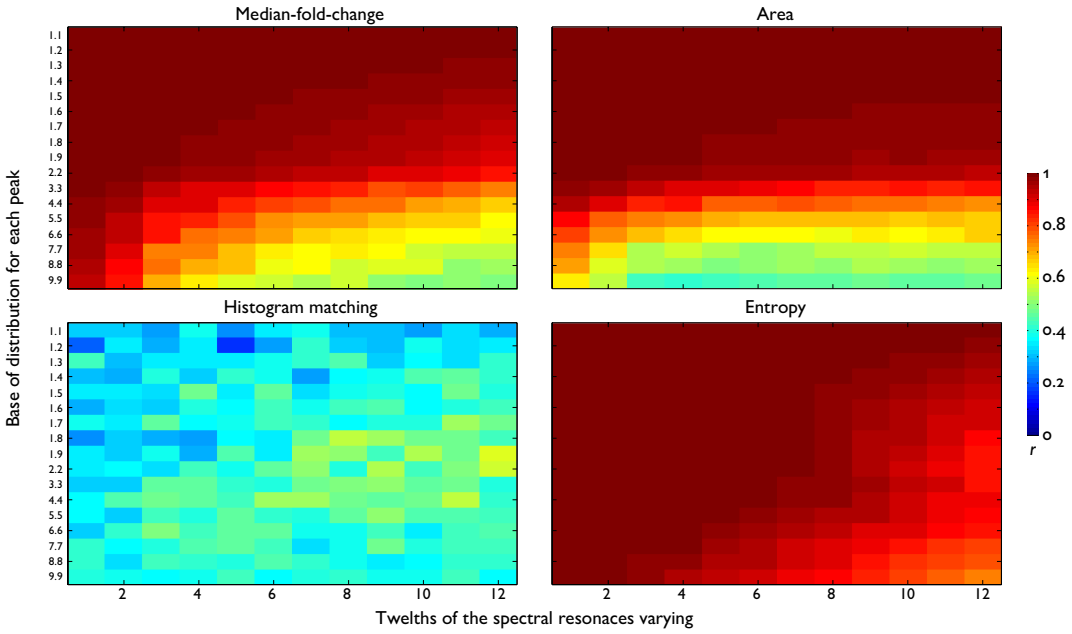

Figure 4.4: Grid searches illustrating the correlation of normalisation factors generated by four normalisation methods to the known true dilution-factor of simulated spectra, with both an increasing proportion of varying resonances and an increasing degree of variation.

The apportionment-entropy method can clearly be seen to perform better over a wider range of conditions than both the MFC and total-area based methods. Total-area normalisation

performs strongly in cases were the exponent base is lower than 1.6, but this may be expected due to limitations with the simulation outlined above. As the exponent base rises and so the total variability in spectral-area increases, area normalisation performs less well, however once again the nature of the simulated spectra means that the mean performance never falls below an r of 0.5.

Median-fold-change normalisation performs as expected, generating the correct normalisation factor as long as the majority of the spectrum is invariant, but losing effectiveness as the proportion of the spectrum varying increases. Additionally as the base of the log-normal distribution increases and individual resonances have a wider effect, due to the tails of the resonances, the effectiveness of MFC normalisation decreases markedly.

4.2.2 Experimental NMR spectra

To estimate the performance of the methods in real-world tasks, the effect of each method on the pattern recognition of two exemplar data sets was assessed in terms of the robustness of the generated models. PLS models of real-world data normalised by the established methods were compared to those generated from apportionment-entropy normalised data. In the absence of true dilution values for the real world data, each normalisation method was judged on the quality of model it produced, as estimated from the model Q^2, indicating the predictive ability of the training data. The maximum number of components taken for each model was chosen by; the number of components required to maximise the Q^2, on the condition that the cumulative Q^2 must always increase from one component to the next.

COMET serum spectra

In the first case the NOESYPR1D spectra of three randomly chosen COMET studies were regressed to the clinical-chemistry derived glucose level. Data preparation was as in § 6.2.1. Data were normalised in MATLAB to area, MFC and apportionment-entropy, then reduced to a resolution of $\delta_H = 0.0090$ *per* data point. PLS models were generated in SIMCA-P+ 11.5 ← using UV-scaled data.

Umetrics AB, Umeå.

Along with the outlined methods, this data was also used without normalisation, as serum is often considered to be exempt from dilution effects due to the homeostatic controls exerted by the body ←, and thus assuming no experimental error, any apparent dilution can be attributed to a fundamental metabolic basis.

This assumes spectra are acquired with a fixed receiver-gain, and the same probe-head, true in the case of these COMET serum spectra.

Table 4.1: Regression of COMET NOESYPR1D serum data to the clinical chemistry derived glucose concentration. All values are cumulative.

	One Component			Two Components			Three Components		
	R^2Y	Q^2	RMSE	R^2Y	Q^2	RMSE	R^2Y	Q^2	RMSE
Entropy	0.266	0.117	3.18	0.387	0.128	3.12	0.636	0.225	2.95
MFC	0.274	0.176	3.11	0.529	0.127	3.12	0.665	0.0767	3.20

Continued on next page

	One Component			Two Components			Three Components		
	R^2Y	Q^2	RMSE	R^2Y	Q^2	RMSE	R^2Y	Q^2	RMSE
Area	0.288	0.173	3.08	0.46	0.102	3.11	0.581	0.157	2.97
None	0.26	0.083	3.24	0.573	0.0174	3.36	0.66	0.111	3.26

Examining the model statistics in Table 4.1, entropy-normalised data shows the maximum Q^2 of 0.255 after three components. MFC-normalised data has a lower Q^2 overall, but a higher Q^2 for a one component model. This swiftly declines as further components are removed. Area-normalisation also performs well for the first component, but the Q^2 swiftly declines as further components are removed. The initial strong performance is likely due to the essential similarity of the spectra, which exhibit a stable collection of resonances, with no drug-related resonances appearing and affecting the sum in the dosed subset. Contrary to expectation, the un-normalised data performed worst of all methods, potentially due to a systematic difference between studies, such as those discussed in § 5.3.5.

Correlation of human urine spectra to subject age

Previous analysis of this data-set, consisting of human urine from a study of environmental cadmium exposure (Thomas *et al.*, 2009), had been shown to demonstrate a relationship between metabolic profile and age (Keun *et al.*, In Press).

Here we compare the models generated by regressing urine spectra, after various methods of normalisation, to the age of the subject. Spectra were acquired with a standard 1D-water-suppressed NOESY-presaturated pulse-sequence. Spectra were phased, baseline corrected, referenced to the TSP resonance at $\delta_H = 0$ and imported to MATLAB at full spectral-resolution using a set of in-house software. Full details of the data acquisition can be found in § B.4. Spectra were normalised to: total area, area of the TSP resonance, MFC, and apportionment-entropy, then reduced to a resolution of $\delta_H = 0.0090$ in MATLAB and exported for chemometric analysis in SIMCA-P+ 11.5. PLS regression to the age of each subject was carried out with UV scaled data, and the model statistics are reported in Table 4.2.

Table 4.2: PLS model statistics for a regression of ^{1}H-NMR of human urine to age. All values are cumulative.

	One Component			Two Components			Three Components		
	R^2Y	Q^2	RMSE	R^2Y	Q^2	RMSE	R^2Y	Q^2	RMSE
Entropy	0.0947	0.0446	16.91	0.369	0.263	15.93	0.513	0.346	16.11
MFC	0.269	0.167	15.79	0.476	0.257	15.50	0.703	0.274	17.19
Area	0.256	0.143	16.02	0.43	0.196	16.01	0.58	0.236	16.86
TSP	0.127	0.1	16.41	0.463	0.177	16.17	0.598	0.267	16.77

Table 4.2 repeats the findings of Table 4.1, as MFC-normalised data gives the highest Q^2 after one component, with entropy overtaking as further components are removed.

4.3 Apportionment-Entropy Profiles

By systematically varying the normalisation factor and recording the apportionment-entropy, an entropy-profile for each spectrum can be generated across this range of dilutions. Examination of these profiles can reveal some interesting aspects of the apportionment-entropy methodology. To explore this, a simple set of 30 simulated spectra were generated, each spectrum consisting of 100 resonances evenly distributed across the spectrum, with identical intensity distributions. Figure 4.6 shows the entropy profile of a single simulated spectrum as the number of simulated-resonances varying in the spectrum is increased from none, where the only variance is noise, to every resonance in the spectrum. The minimum entropy value remains at the point of true dilution until $^{96\pm2}/_{100}$ resonances are varying.

That is, the only difference between the spectrum to be normalised and reference spectrum is noise.
The darkest red profile.

Counterintuitively, in ∫ 4.6 we can see that when there is no variation in the data-set, ←, the S profile ← is at its maximum at the correct normalisation factor, and decreases symmetrically about this point. Clearly this presents a unique case, where apportionment-entropy will not converge to the correct solution.

Examination of the normalisation process reveals this is due to the scaling of the difference vector in Equation 4.8. When the only variation between each spectrum is noise, this results in a difference vector dominated by the disparity between the noise in the two spectra. This when scaled, results in the observed large values of S. This effect is also observed in simulations without noise, due to the inherent quantisation errors associated with binary representations of real numbers, which like the differences in noise values, are magnified by the scaling process to produce much the same effect.

Machine epsilon refers to the bound on the quantisation error resulting from a binary representation of a real number (Higham, 2002, p39).

These effects might be countered by setting to zero any value that lies below either, the noise level in the spectra, or machine epsilon ←. However, this simply results in a value of S that is constant for all possible dilution factors.

While this limits the use of apportionment-entropy for the normalisation of data with no biological variation, the real-world rarity of this situation, coupled to the ease of its diagnosis by examination of the entropy profiles, means the technique remains valid for use in real-world data-sets. Figure 4.6 illustrates how once $^1/_{100}$ of the resonances are varying, normalisation proceeds correctly.

A set of example profiles from spectra of urine can be seen in Figure 4.7, these profiles are less regular than those from the simulation seen in ∫ 4.6, due to the greater variation in the area of resonances above and below the reference spectrum. These spectra were previously normalised to median-fold-change, so the point of no change, a normalisation factor of 1, in the plot, represents the normalisation applied by this method. Typically we see that the apportionment-entropy value is close, but not identical to that calculated by MFC.

The tailing-off of the entropy value towards the highest and lowest normalisation factors is also due to the vector scaling described above. As the difference between the spectrum to be normalised and the reference grows larger, the contributions of individual resonances decreases, making the scaled difference vector increasingly homogenous.

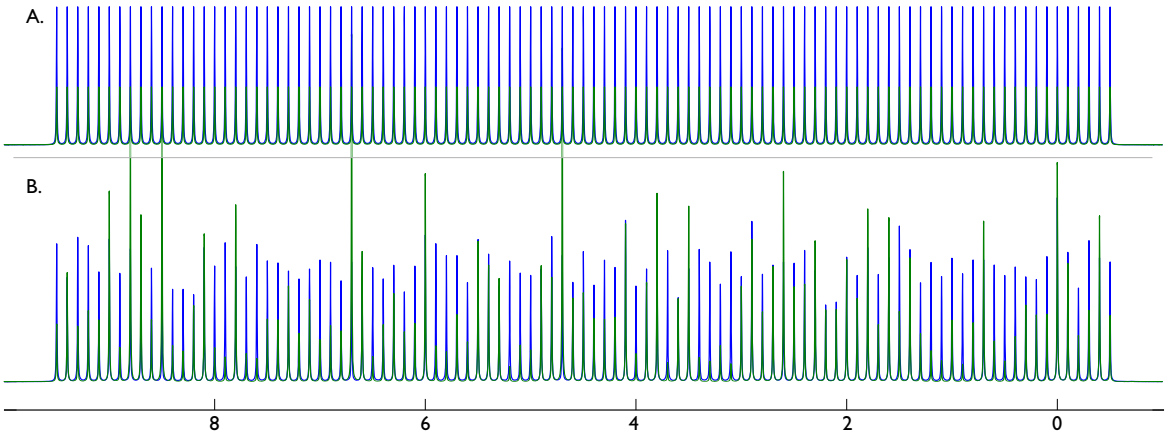

Figure 4.5: Basic simulated spectra with 100 identical resonances. The median spectrum of all 30 spectra is in blue, while the first spectrum in these data is in green. **A**. All resonances have a fixed intensity. **B**, Each resonance varies in intensity across the data-set according to a log-normal distribution.

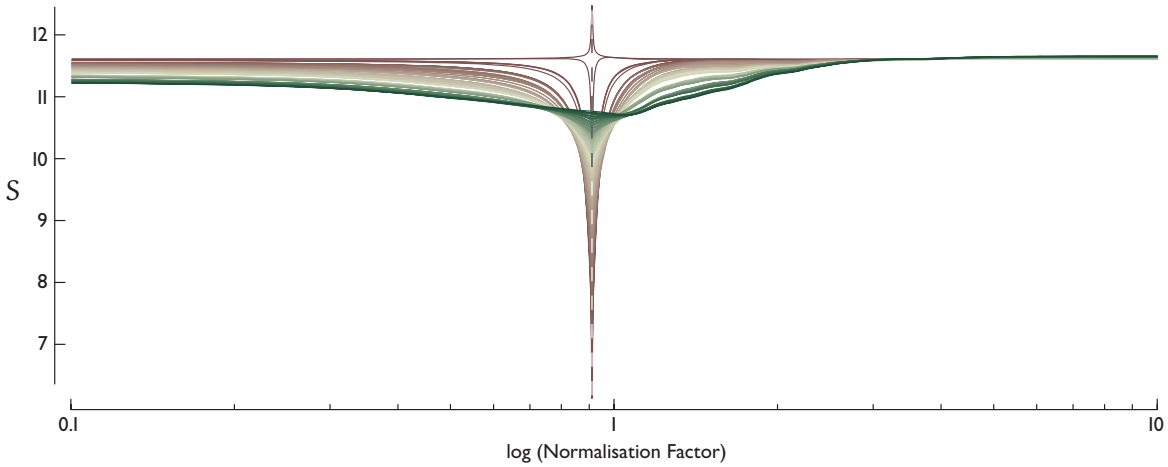

Figure 4.6: Entropy profiles of a single simulated 'spectrum', as seen in *f* 4.5, as the proportion of varying resonances is increased from $^{0}/_{100}$ (red) to $^{100}/_{100}$ (green), plotted against the log normalisation factor. The true dilution (0.913 ×) is marked by a dashed line. The minimum entropy remains at this point until greater than $^{95}/_{100}$ resonances are varying.

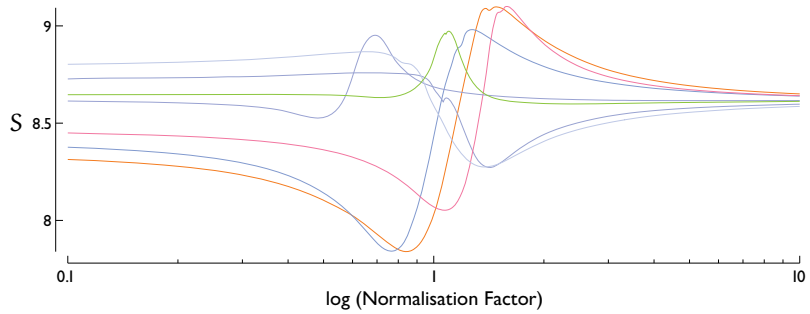

Figure 4.7: Entropy profiles of a random selection of experimental NMR spectra as the dilution factor is adjusted. These spectra were previously normalised to MFC, we can see the minimum S is close to, but not identical to this value.

4.3.1 Apportionment-entropy can act as an indicator of spectral similarity

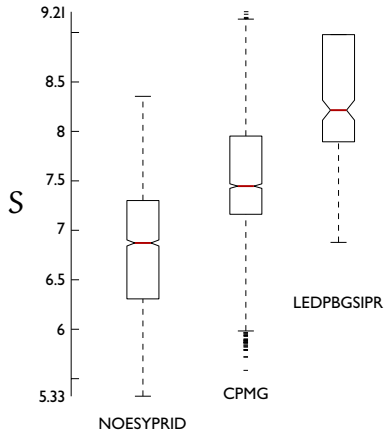

Figure 4.8: Apportionment-entropy with respect to a randomly selected NOESYPR1D spectrum. The NOESYPR1D spectra have significantly lower entropy values than spectra of the same samples acquired with a CPMG or LEDPBGS1SPR pulse-sequence. Whiskers extend to a maximum of 1.5 times the interquartile range, notches mark the comparison interval of the median at the 5% level.

Unlike many other normalisation methods, such as MFC or total-area, which standardise the metric of difference between spectra as part of the normalisation process, following apportionment-entropy normalisation, each profile will retain a value of S. This value could be considered a measure of the similarity of the spectrum to the reference.

To illustrate this, Figure 4.8 shows the minimum apportionment-entropy values of every COMET serum spectrum calculated with a randomly selected control NOESYPR1D spectrum as the reference. Each of the three pulse-sequences used to acquire data shows a separate distribution of entropy values, with the medians significantly different at $p < 0.05$. Interestingly the CPMG spectra (2,539 total), which edits out larger macromolecules, fall between the NOESYPR1D spectra (2,879), which suppresses only the water resonance and the LEDPBGS1SPR spectra (88) that focus on macromolecules and suppress smaller metabolites.

This could be understood in terms of the relative information content of the three types of spectra, with NOESYPR1D reporting all resonances in a sample, and naturally being most similar to the reference spectrum. The CPMG spectra suppress the broad resonances, that make up around 30% of the total area of a serum spectrum (see ∫ 5.1), while the LEDPBGS1SPR spectra focus on these resonances and suppress those associated with smaller molecules. Thus we

might expect a greater degree of information overlap between the NOESYPR1D and CPMG spectra than between NOESYPR1D and LEDPBGS1SPR.

This would appear to indicate apportionment-entropy has potential to act as a generalised measure of similarity between profiles. While there is a degree of overlap between NOESYPR1D and CPMG spectra, the difference in distribution of S between otherwise good spectra suggests a truly malformed spectrum could be expected to be further outlier, and this could be applied as part of an automated outlier-detection or quality-control system.

4.4 Discussion

Normalisation is strongly dependent on both the nature of the variation affecting the data and the methods intended to be used in analysis and interpretation. Apportionment-entropy provides a powerful, novel, technique for normalising metabolite profiles, with the specific aim of producing the most interpretable and robust models. In this, the above presented analysis of both simulated and experimentally acquired data has shown apportionment-entropy based normalisation to equal or improve upon the performance of the current standards, median-fold-change and total-area or -sum.

While the need for an exhaustive grid-search means that apportionment-entropy is not the fastest method, the fact that most data-sets are only normalised once means that in practice, this is not a major drawback.

Besides providing a strong basis for normalisation, the value of apportionment entropy between a reference and a set of NMR spectra shows potential as a generalised measure of spectral similarity. With further development this could aid in analysis or prove a powerful method of outlier detection.

While untested on a large set of data such as COMET, apportionment-entropy appears to work well enough to be used in parallel, and thus compared to median-fold change while conducting further analysis of the COMET data.

Chapter 5

Orthogonal Filtering for Relative Quantification

5.1 NMR Quantification in Biofluids

While the majority of statistical analysis of metabolic profiles is based on the the analysis of spectra-based profiles, this is a concession to the limitations of the analytic technologies, with the ultimate aim of the subject being to remove this layer of abstraction and deal with the metabolite concentrations directly. With this aim comes the desire to deconvolve the contribution of overlapping signals to a set of resonances of interest.

Often pure spectra of such compounds are available, or can be easily acquired. In such cases, it would be desirable to have a technique capable of using knowledge of the pure spectrum of the compound of interest to extract quantification information from a set of biofluid spectra, discarding the influence of the overlapping signals. Even if this data was available for only a subset of the known metabolites in a sample, and only in the form of relative values, it could prove a useful counterpart to purely spectrum-based pattern-recognition. For instance, this simplifies pattern-recognition in a manner similar to resolution reduction, but also has the advantage of separating co-variation caused by overlapping resonances from biological variation.

5.1.1 Existing methods of quantification and their disadvantages

Many methods exist to extract discrete values from NMR spectra. These vary between simple methods such as direct integration of the area under a resonance, to advanced techniques including; the generation of a PLS model from a representative training set (Haaland & Thomas, 1988); intelligent integration, using curve-fitting to maximise the area of a resonance captured (Crockford *et al.*, 2005); and many others (De Beer *et al.*, 1998), some utilising wavelet transforms (Mainardi *et al.*, 2002), modelling of the FID (Vanhamme *et al.*, 1997) or finite impulse response filters in the time domain (Vanhamme *et al.*, 2000).

Estimation of the area of a resonance in biofluid NMR is complicated by several factors. Following the Fourier-transform, NMR resonances have a Lorentzian line-shape (Equation 3.1 in § 3.2). As a result, to capture 99% of the area of a resonance, it must be integrated over a region of twenty times its half-height line-width (Claridge, 1999). This hampers accurate quantification of metabolites in biofluids by NMR →, due to the complexity of the chemical environment in these samples, small variations in which may affect the line-shape or observed

Including relative quantifications, relative meaning internally consistent within a set of data, but not necessarily related to real-world concentrations.

frequency of resonances, meaning that integration over the same height-proportional spectral-range may not capture a comparable proportion of a resonances area.

Biofluid NMR is further complicated at typical spectrometer frequencies, as many resonances from disparate compounds inhabit contiguous regions of the spectrum, and the cumulative contribution of many unresolved low-intensity resonances can have the effect of elevating the observed spectral baseline, negatively affecting our ability to quantify metabolites by the integration of resonances (Figure 5.1). Furthermore, variance between subsequent samples, such as pH, ion and protein content, which affects wide-ranging spectral features such as the baseline contribution, can negatively effect our ability to compare measurements between experimental runs.

Due to their typical composition, biofluids such as blood-serum and plasma- are especially affected by overlap when compared to urine. Large concentrations of proteins, lipids and other macromolecules in solution, generate many resonances in most ^{1}H-NMR pulse-sequences. These signals can contribute over 30% of the total area of a serum spectrum, much in the form of broad, unresolved resonances ($\int$ 5.1).

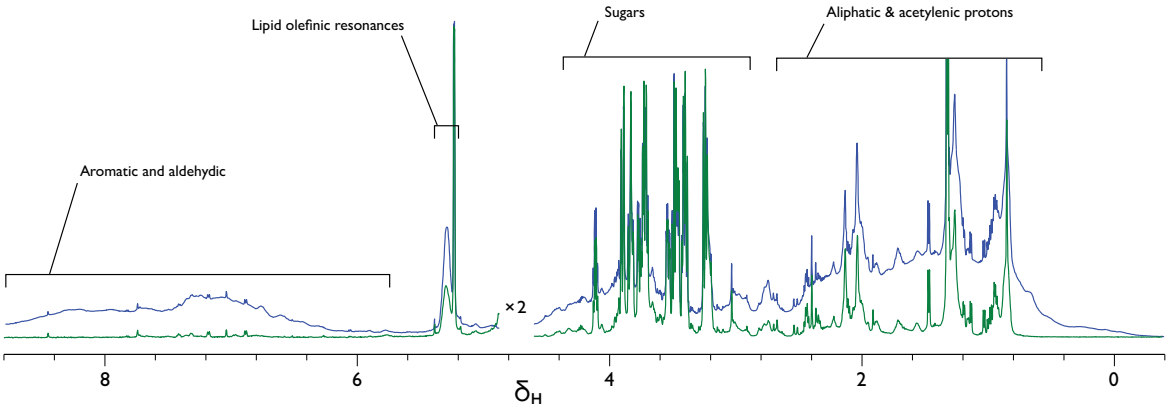

Figure 5.1: Comparison of two ^{1}H-NMR spectra of a control rat serum-sample, one acquired with a CPMG pulse-sequence and the other a NOESYPR1D sequence. Broad resonances attributed to macromolecules can be seen to contribute over 30% of the area of the spectrum. The CPMG pulse-sequence edits out the (mainly broad) signals with a short T_2, leaving behind those from small molecules.

To overcome these interferences, experimental methods such as the Carr-Purcell-Meiboom-Gill pulse sequence (CPMG) have been developed (Carr & Purcell, 1954). The CPMG takes advantage of the lengthy T_2 relaxation times of small molecules to allow the analyst to edit out signals from macromolecules with rapid T_2 relaxation times (see $\int$ 5.1 for an example of a CPMG spectrum compared to a standard 1D water-suppressed pulse-sequence). However CPMG spectra are susceptible to phase or baseline distortions, caused by magnetic field inhomogeneities (Song, 2002) and inaccuracies in the pulse-sequence, resulting from the complexity of the sequence (Meiboom & Gill, 1958; Lucas *et al.*, 2005). These distortions may be compounded by repeated scans, further complicating the quantification of low-concentration metabolites. Additionally,

as the area of a resonance in a CPMG spectrum is partially related to the T_2 relaxation time, the quantification and quantitative comparison between molecules with differing T_2s becomes complicated.

Furthermore, while the CPMG provides a reasonable method of suppressing the resonances from macromolecules, it requires a second acquisition on every sample for which this data is desired. This is a requirement that may not be realistic in situations where the spectra were not acquired firsthand or spectrometer time is limited →.

These may both be due to organisational or experimental concerns.

5.2 Orthogonal Filtering

5.2.1 Spectral post-processing

There have been many efforts in the field of chemometrics, directed towards the deconvolution of pure component intensities from mixed spectral profiles. These include such procedures as CLS, multiplicative scatter correction (MSC, Geladi *et al.*, 1985), direct orthogonalisation (DO, Andersson, 1999) and self-modelling curve resolution (Lawton & Sylvestre, 1971) and its derivatives.

As outlined in § 1.5.3, CLS is limited by the need to take every pure component of the spectra in question into account. As the interfering variation in most spectroscopy of biofluids represents unknowns, the development of more advanced algorithms has included much work on models that would allow interfering variation, regardless of its characteristics, to be separated from the variation of interest.

Direct orthogonalisation is a filter intended to remove systematic interfering variation, prior to a second modelling step, such as PCR or PLS. While DO can be used as to reduce the complexity of regression models, particularly in the case of PCR, where it may in part compensate for large degrees of variation, which may obscure the variation of interest (as outlined in § 1.5.3), it still requires calibration with known data.

Multiplicative scatter correction was developed for use with near-infrared reflectance (NIRR) spectroscopy, and aims to reduce the affect of unpredictable scattering in acquired spectra by transforming each spectra onto an idealised spectrum. While this does not remove the interfering effect, the aim is to make it constant for all samples. The MSC model is simple, and assumes the relationship between the ideal spectrum and that being corrected is independent of wavelength. This has been addressed with Piecewise-MSC (PMSC, Isaksson & Kowalski, 1993), that uses a moving window to divide the spectrum into discrete wavelength regions, and models each separately. One limitation of PMSC results from the sensitivity of the model to the size of the window region, which if too small, may result in the removal of the variation of interest. Also, while MSC based methods do not require calibration on characterised data, they benefit from training with prior spectroscopic knowledge.

Self-modelling curve resolution, often also known as non-negative matrix factorisation (NMF, Lee & Seung, 1999, 2001) models a matrix in terms of a mix of components, similar

to PCs, but constrained such that they may not take negative values. In spectroscopy, these components may then be interpreted as the 'pure' spectra comprising the sample matrix. In controlled situations, deconvolution of NMR spectra by an alternating-least-squares derived multivariate curve resolution method has proved very successful at deconvolving a set of spectra into chemically-meaningful components (Winning *et al.*, 2008).

5.2.2 Application to metabolic profiling data-sets

While many of the above methods are suitable for deconvolving a set of NMR spectra, none allow the use of a limited knowledge of the sample matrix, such as the pure-spectrum of a compound of interest, to generate a relative quantification of this target across the data-set. In many metabolic profiling studies, such as the COMET project, there exists a large set of many highly overlapped spectra, in which a relative quantification of a number of well-characterised compounds may be desired.

Here I aim to improve relative quantification, that is quantifications consistent within a data-set, but not necessarily related to real-word concentrations, of compounds with known spectra by using orthogonal filtering to reduce the affect of an unknown set of systematically varying overlapping signals.

5.2.3 Orthogonal filtering for spectral quantification

To improve the estimation of metabolite concentrations in overlapped spectra, a novel method of spectral filtering based on orthogonal signal correction has been developed (see the discussion of O2-PLS in § 1.5.5). Orthogonal Filtering for Spectral Quantification (OFSQ) takes advantage of the systematic variations in overlapping resonances in a data set, modelling their contributions to the resonance of interest, and removing them before quantification takes place.

OFSQ is based on an excision and modification of the orthogonal signal correction function integrated into the O2-PLS (Trygg, 2002; Trygg & Wold, 2003) algorithm (Equation 1.8). The point of divergence concerns the generation of the correlated weight vector w. The listing below shows the section of the O2-PLS algorithm used to remove Y-orthogonal variation from X, and calculate a set of scores t, on the filtered X matrix:

$t_{temp} = Xw$	Calculate scores including orthogonal variation.
$X_{osc} = X - (t_{temp}w)$	Remove modelled variation.
$w_{osc} = \text{PCA}(X_{osc}^T t_{temp})$	Calculate the orthogonal variation.
$t_{osc} = Xw_{osc}$	Calculate orthogonal scores on full data set.
$p_{osc} = X^T t_{osc}$	Regenerate orthogonal loadings on full data set.
$X_{Fltr} = X - (t_{osc}p_{osc}^T)$	Remove orthogonal variation.
$t = X_{Fltr}w$	Regenerate scores on filtered data.

In O2-PLS, w is obtained from the first PC of the covariance matrix of the X and Y matrices, and thus models the relationship between X and Y. In the case of OFSQ, w is a synthetic vector,

generated from the pure spectrum of the compound of interest. X is the overlapped data-matrix and t correlates to levels of pure compound in X after filtering.

This technique operates under several assumptions; primarily that the pure spectrum used as w represents the compound in conditions comparable to the sample matrix; additionally we assume that the NMR acquisition is comparable. We attempt to ensure a comparable acquisition by using standardised experimental conditions and pulse-sequences, during spectral acquisition of both the standards and the samples. Despite this, variability in the sample matrix, such as ion concentrations or interactions with macromolecules →, may affect the spectrum in a manner not easily controlled for in experimental NMR.

The interaction of TSP with the proteins in serum outlined in § 3.1.2 is a good, albeit extreme example of this.

5.2.4 Implementation of OFSQ

Code to implement OFSQ was written in MATLAB , accepting a matrix of overlapped data, a pure spectrum, and the number of orthogonal components to remove. The basic algorithm used to calculate OFSQ is as follows:

1. Scale or centre X, either by MC or UV (Eriksson *et al.*, 1999; Bro & Smilde, 2003).
2. Normalise w to unit-length ($w = w^T\sqrt{w^T w}$).
3. Calculate algorithm above to remove one orthogonal component.
4. Replace X with X_{Fltr}.
5. Repeat $3-4$ for each additional orthogonal component.

The output consists of a set of scores, t, that are correlated to levels of the pure compound in X, after the removal of orthogonal variation.

In addition to the scores on the filtered data, OFSQ also outputs several diagnostic variables in an attempt to estimate model reliability: The orthogonal loadings (p_{osc}) model the orthogonal variation detected within the data. Examination of the orthogonal loadings can reveal the nature of the systematic interference within the data and may thus inform the ongoing analysis.

Cross-validated scores (as outlined in ∫ 1.7) were generated by excluding ⅐ of the data in turn, modelling the remainder, and then projecting the set-aside data into the generated model. Cross-validation can help indicate the stability of the model, by highlighting variation in the generated scores, indicating the model is over-fitted e.g. modelling features specific to the subset of data used in the cross-validation, rather than the entire data-set.

Finally, while lacking the true eigenvalue-equivalence inherent in the components of a PCA model, we can use an approximation of such to attempt to regenerate the input weights vector from the generated scores and filtered data matrix. This is achieved with the expectation that, having removed all the orthogonal variation in the data, the regenerated vector will be more similar to the original vector than that generated from overlapped data, as the situation more closely fulfils the criteria of orthogonality of components expected by PCA. The

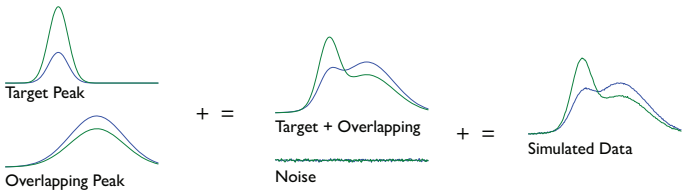

Figure 5.2: Steps in the generation of a set of simulated spectra (left to right). For each spectrum a target peak and a number of overlapping peaks are generated and summed together, low-amplitude white noise is then added to the combination to create the final simulated spectrum.

pseudo-loadings are generated by multiplying the scores t back through the filtered data X_{Fltr} (Equation 5.1):

$$p_{pseudo} = tX_{Fltr} \tag{5.1}$$

Pseudo-weights are similarly generated by multiplication of t through the Moore-Penrose pseudoinverse of the filtered data X^{+}_{Fltr} (Equation 5.2):

$$w_{pseudo} = X^{+}_{Fltr}t \tag{5.2}$$

5.2.5 Generation of simulated data

To allow testing of the OFSQ method in a well characterised situation, simulated data were generated for use as a test set. These data were generated in MATLAB in the same manner as those in § 4.2.1, as a combination of three Lorentzian peaks (Equation 3.1) of varying half-height line-widths and maximum intensities. The target peak was generated with a normally distributed set of random intensities across the set, generated by the MATLAB `randn()` function. The degree of correlation (r^2) of each overlapping peak to the target peak could be varied between 0 and 1 as required. The range of intensities of each peak could be varied by a simple scalar multiplication of the generated intensities. Finally, a low intensity white noise was added to the data to approximate instrument noise. See Figure 5.2 for an summary of this process.

This allowed the generation of data, where the true intensity of the target peak was known, along with the contributions of the overlapping interferences to each spectrum, and the overall correlation of each interfering peak to the target. For the purpose of this analysis, simulated data-sets of 100 spectra with 500 spectral data-points were generated. The size of these data were intentionally restricted to the greatest practical extent, to reduce the computational time required when repeatedly running the OFSQ algorithm, a particularly important concern when running grid searches such as those in Figure 5.8. Within these data-sets, a single target peak, and two additional peaks, one overlapping the target peak and one distinct, were generated, with both peaks having twice the half-height line-width of the target and varying over three times the intensity range. This was intended to approximate the situation in serum NMR where

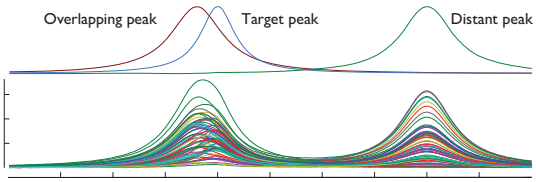

Figure 5.3: Set of example spectra, generated according the procedure laid out in § 5.2.5. The individual spectra of the pure components are shown above, with the combined spectra below.

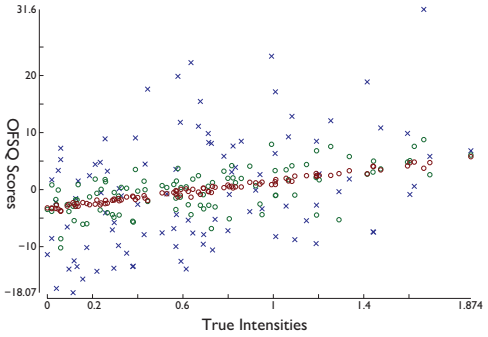

Figure 5.4: As orthogonal components are removed the correlation of the generated scores to the true intensities of the target peak improves. With no orthogonal filtering (blue ×) the scores have an r^2 to the true intensities of 0.38, after the removal of one (green o) and two (red o) orthogonal components the r^2 is 0.64 & 0.99 respectively.

the resonances of small metabolites are typically overlapped by broad obscuring signals. See Figure 5.3 for example simulated data.

5.3 Testing and Validation of OFSQ

5.3.1 Performance on simulated data

Simulated data generated by the method in § 5.2.5, with no correlation between the overlapping peaks and the target, and a simulated pure spectrum generated with identical parameters

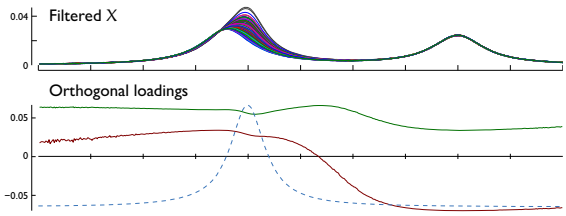

Figure 5.5: Following OFSQ, all orthogonal variation in f 5.3 has been removed from X (above). The orthogonal loadings that model this variation can be seen below, with the first component in green and the second in red. The pure weight vector w, is shown as a dashed blue line for comparison.

to the target peak, were used to test OFSQ. These demonstrated an almost perfect correlation (r^2 of 0.989) between the scores (t) and the true intensities. Figure 5.4 shows the progressive improvement in the correlation to the true intensities as orthogonal components are removed from the data. Examination of the filtered data matrix showed a complete removal of orthogonal variation (Figure 5.5). Data were unit-variance scaled prior to OFSQ, and to simplify interpretation, the obtained scores and filtered data-set were back-scaled once OFSQ was completed.

5.3.2 The effect of correlated variation on OFSQ

A limitation of any orthogonal filtering based approach arises when the interfering variation in the data-set is no longer truly orthogonal to the variation of interest. To assess this, an exhaustive grid-search was used to explore the limits of the OFSQ approach in situations where other peaks in the spectra were correlated to the peak of interest. Simulated data sets were generated as in § 5.2.5. The grid-search then varied the r^2 of each overlapping peak independently between a value of 0 and 1 in increments of 0.05←. This procedure was repeated 1,000 times with the same parameters to generate aggregate statistics.

21 × 21 correlation steps for a total of 441 simulated data-sets *per* iteration.

Figure 5.6 shows the results of this grid-search, examining the ability to correctly identify the number of orthogonal components that should be filtered from the data seen in Figure 5.6, assuming no knowledge of these data beyond the profile w. Successful selection of the number of components to be removed was defined such that the correlation of the quantifications to the true intensities could not be any worse, but may be equal to, quantification by a basic least-squares method: $t = Xw$.

In situations where the intensities of the overlapping peaks were truly orthogonal (*uncorrelated*) to that of the target peak, the OFSQ method was able to perfectly deconvolve the spectra in every case. As the correlation of each overlapping peak to the peak of interest was increased, the ability of OFSQ to improve concentration estimates declined. Figure 5.6 shows that in 50% of cases, an r^2 of above 0.45 between the overlapping peak and the target will render the OFSQ method unable to improve the correlation of the scores to the true intensities above those generated by a basic least-squares method. The correlation of the non-overlapping interfering peak to the target peak has a much weaker negative effect on the ability of OFSQ to improve quantification, reaching a 50% failure rate at an r^2 of 0.85.

The 50% failure rate is affected by the relative intensities of the overlapping and target-peaks. Where the maximum intensity of the target peak is equal to that of the overlapping peaks (as opposed to one-third in the previous example), the area of improvement contracts, to an r^2 of 0.3 to the overlapping peak, effectively preventing improvement by OFSQ. A grid-search of grid-searches, where a superset of the grid-search in ƒ 5.6, was conducted, varying the intensity of the overlapping peaks between each grid-search. In Figure 5.8 these data show that as the relative intensity of the target peak to the overlapping peak decreases, the number of situations in which OFSQ may improve the quantification of the data grows larger. This is to

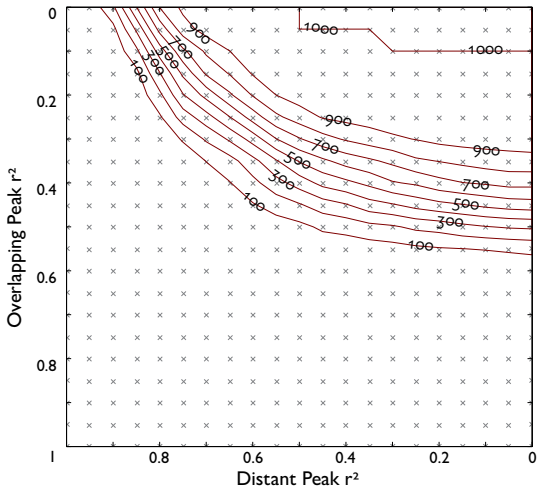

Figure 5.6: Increasing the correlation of the overlapping peak to the target hampers OFSQs ability to model the peak. Here the number of times out of 1,000 the model of a simulated data-set could be improved by OFSQ is plotted for a grid search between r^2 of 0 and 1. All three simulated peaks vary independently in the upper-right corner of the plot, while all three are perfectly correlated in the lower-left. Individual data-sets were generated at the points marked by ×, and a contour plot indicating the number of time a data-set at each point could be improved by OFSQ is overlaid.

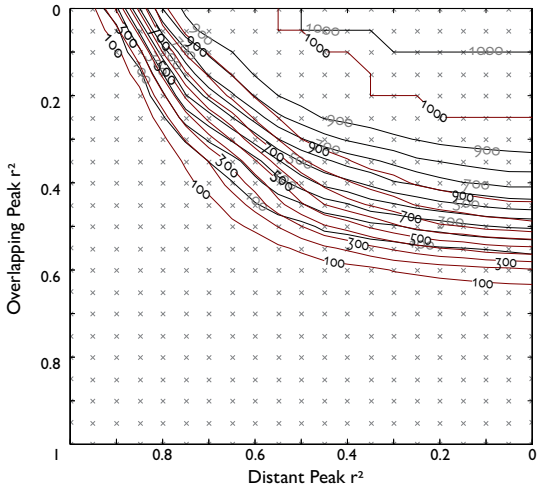

Figure 5.7: Centring the data in *f* 5.6 to the mean rather than scaling to unit-variance further expands the region of improvement. Data centred by MC are red, overlaid on the UV scaled data in black lines.

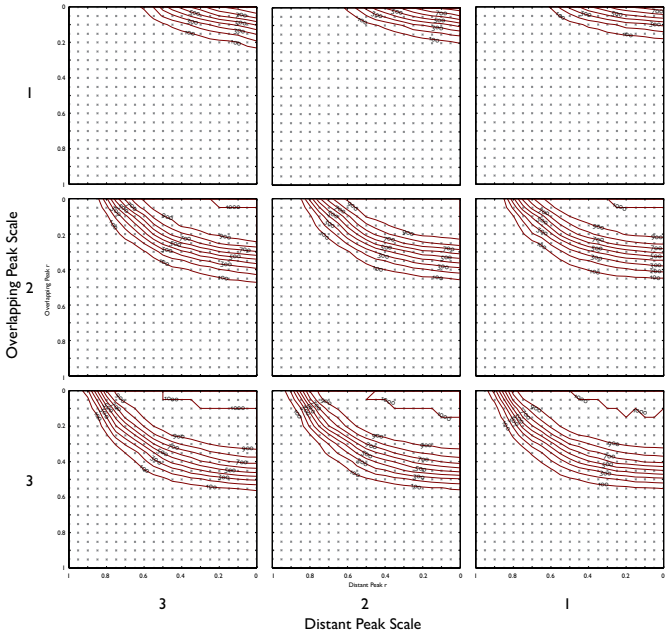

Figure 5.8: As the intensity of the overlapping peak is increased in comparison to the target peak, OFSQ is able to improve the models over a greater range of correlations. Each individual data-set above is a grid search as in ƒ 5.6 with the maximum intensities of the interfering peaks varied between 1× (equal) and 3× the intensity of the target peak.

be expected, as in such situations, the orthogonal variation comprises a larger proportion of the total variation in the data.

5.3.3 Selection of the number of orthogonal components in OFSQ models

An important question now arrises, how can we determine when OFSQ will improve our quantification in a data-set? There are two situations in which the use of OFSQ may impair the ability to quantify the target resonances; first, situations in which there is orthogonal variation in the data, but we are removing too many orthogonal components; second, in situations where one of the interfering signals is strongly correlated to the target peak. In both cases we require a metric that will allow us to gauge if we have removed too many orthogonal components.

Judgement of the number of components to remove was based on one of the three statistics described in § 5.2.4, cross-validated scores, pseudo-weights and pseudo-loadings. Each statistic was judged to have indicated the number of orthogonal components to remove if, after the removal of an additional component, the value decreased. In Figure 5.9, an additional axis is projected onto the results seen in ƒ 5.6. Here the height at each grid point indicates the number of times out of 1,000 OFSQ was able to improve prediction, and colour indicates the ability of each of the three metrics to correctly judge the number of components to remove, based on the same criteria as § 5.3.2.

Successful prediction of the number of components to remove would allow the use of OFSQ in the knowledge that the process would not be detrimental to the relative quantification of resonances.

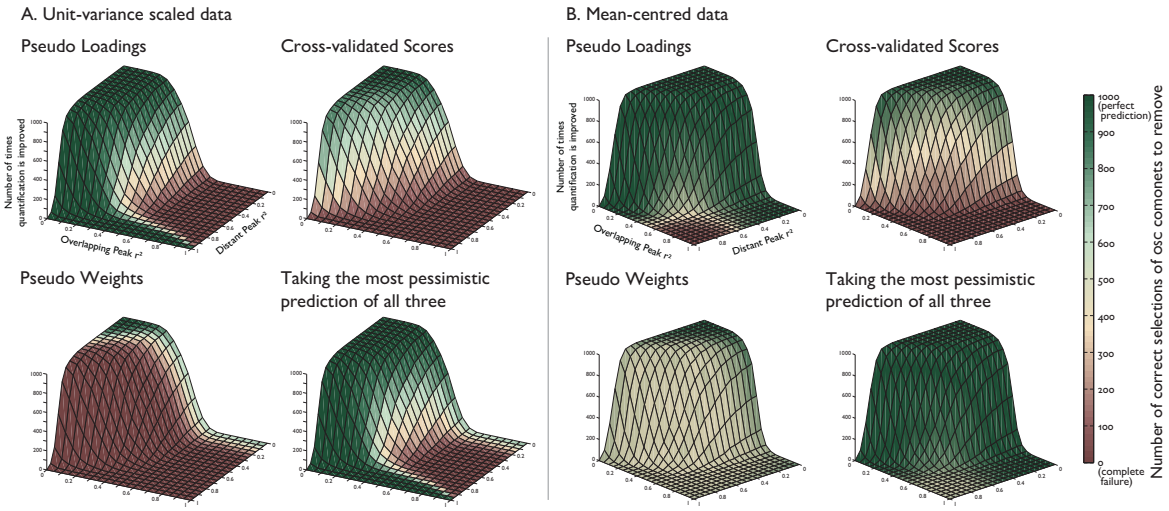

Figure 5.9: Each diagnostic variable is able to correctly indicate the correct number of orthogonal components to remove in a subset of conditions. Unfortunately, none are correct in all situations, and a combination of all still leaves us unable to correctly predict the number of components to remove. The data used are those in ƒ 5.6. In each set of axes, the height of the surface is equivalent to the number of data sets out of 1,000 at this point, where modelling may be improved by OFSQ. The colour of the surface indicates the ability of each parameter to choose a number of OFSQ components that will improve the quantification of the data set. **A**. Scaled to unit-variance. **B**. Centred by mean-centring rather than scaling to unit-variance. While the domain in which cross-validated scores are effective remains unchanged, significant differences in the performance of the pseudo-loadings and scores can be observed.

In ƒ 5.9 we can see that none of the statistics are reliable in all cases. The cross-validated scores only provide an accurate estimate of the number of orthogonal components to remove in situations where the OFSQ method is able to improve quantification, rendering it useless in other situations. The pseudo-weights are capable of distinguishing the correct number of components to remove regardless of the correlation of the overlapping peak. However they are strongly effected by the correlation of the non-overlapping peak, and lose their predictive power as this correlation rises above 0.05. This degree of correlation to a non-overlapping signal could be considered likely, $p > 0.05$ for an n of 250. This renders the pseudo-weights of marginal use at best. Finally, the pseudo-loadings show a good ability to correctly select the number of orthogonal components to be removed, being negatively effected only by overlapping-peak in situations with an r^2 of 0.5 and greater.

5.3.4 The diagnostic measures are affected by data scaling

Now if we examine the same data with mean-centring rather than unit-variance scaling (Figure 5.9B), the performance of the diagnostic measures changes markedly. While the cross-validated scores remain unchanged, both the pseudo-weights and -loadings show a marked improvement in predictive power. While the area of perfect prediction for the pseudo-weights remains the same between UV and MC scaled data, areas that were never predicted correctly after UV scaling are correctly predicted 50% of the time. The pseudo-loadings also show improved predictive ability when the data are MC scaled, only failing in the selection of components when the correlation of the overlapping and non-overlapping peaks exceeds an r^2 of 0.6.

The reason for the effect of scaling are not clear. The effect of noise will be amplified in the UV-scaled data, but identical simulations without the addition of noise exhibit the same effect in the three measures. This may be related to digital quantisation errors, as with apportionment-entropy calculations in § 4.3.

None of the three statistics used are capable of correctly selecting the number of orthogonal components to remove across all conditions tested. Even using MC data, where the validation measures proved strongest, taking the most pessimistic prediction of the three does not allow the definitive selection of the correct number of orthogonal components to remove.

5.3.5 Application of OFSQ to experimental data

As well as simulations, OFSQ was tested on a combination of data from three COMET studies. As a demonstration, glucose was chosen as the target compound, due to its position in an overlapped region of the ^{1}H spectrum, and the availability of an independent measure of glucose for validation. Glucose levels were measured *via* the hexokinase assay, and reported in mM. The final test set consisted of 90 NOESYPR1D spectra of rat serum. Target weights were generated from a NOESYPR1D spectrum of a pure glucose standard.

Table 5.1: Model performance following filtering of orthogonal components from serum spectra.

Correlation of:	**Number of Orthogonal Components Removed**					
	none	1	2	3	4	5
Model score to true-concentration	0.5320	0.5380	0.5685	0.5669	0.5482	0.2526
Pseudo-weights to weights	0.9697	0.9695	0.9685	0.9665	0.9611	0.9538
Pseudo-loadings to weights	0.3823	0.6256	0.6999	0.8300	0.8222	0.6858
Cross-validated scores to scores	N/A	0.9983	0.9701	0.9894	0.9899	0.8958

Sequential orthogonal components were removed from the data-set, for up to five components, and the resulting scores were correlated to the true values for glucose, allowing an estimation of the improvement to the modelling made by OFSQ. Table 5.1 shows the raw output of the model after the removal of each orthogonal component. Looking at the correlation of

the generated scores to the true concentration of glucose, there is a modest improvement in the model, from 0.53 to 0.57, after the removal of three orthogonal components, but after this, removing further components degrades the quality of the model.

5.3.6 Examination of the orthogonal loadings can indicate the source of interference

Examination of the orthogonal loadings gives us some clue as to the nature of the interfering variation. Examining the scores on the loadings (Figure 5.10), we see that the scores on the first three components – those that improve quantification – are strongly influenced by the study the sample was run in. For example, study two scores strongly negatively for component 1, while study three is broadly positive and study one shows no clear trend. In component 2 there is no clear trend by study, while in component 3 we see a trend of negative scores in study three, positive scores in study one and no trend in study two. The final two components show no clear trend, and have noticeably smaller scores than the previous three.

Examining the orthogonal loadings themselves, we see that generally they model broad spectral features, such as spectral-baseline and differences in line-width of resonances. Component 1 models gross changes in baseline elevation and lipid resonances. Component 2 is modelling similar effects but with most lipid resonances in anti-phase to component 1. Component 3 then models anti-phase interferences in the sugar region of the spectrum.

In combination, these observations suggest that OFSQ is improving quantification by modelling the broad differences in spectral composition between studies. These changes may result from systematic differences in sample preparation or shipping and storage (Teahan *et al.*, 2006), long term instrument drift or simply result from minor variation in sample composition resulting from the differing conditions of the experimental animals, for example Purina chow, while having a standardised level of protein, does not specify the source, and this may differ between batches of feed. While animals within a study will eat food from the same batch, and thus with same composition, this will not extend between studies or companies. Such a change in diet could then effect the broad direction of the animal's metabolism (Wang *et al.*, 2006).

5.4 Utility of Orthogonal Filtering

While demonstrating an ability to improve the relative quantification of metabolites in experimentally acquired spectra, OFSQ is hampered by the difficulty in determining the number of overlapping orthogonal components that can be safely removed *a priori*. This, coupled to the stringent requirements in terms of spectral composition, *i.e.* the operating assumptions laid out in § 5.2.1, limits the situations in which OFSQ can be safely used.

It should be noted that while the improvement to glucose quantification in real data was

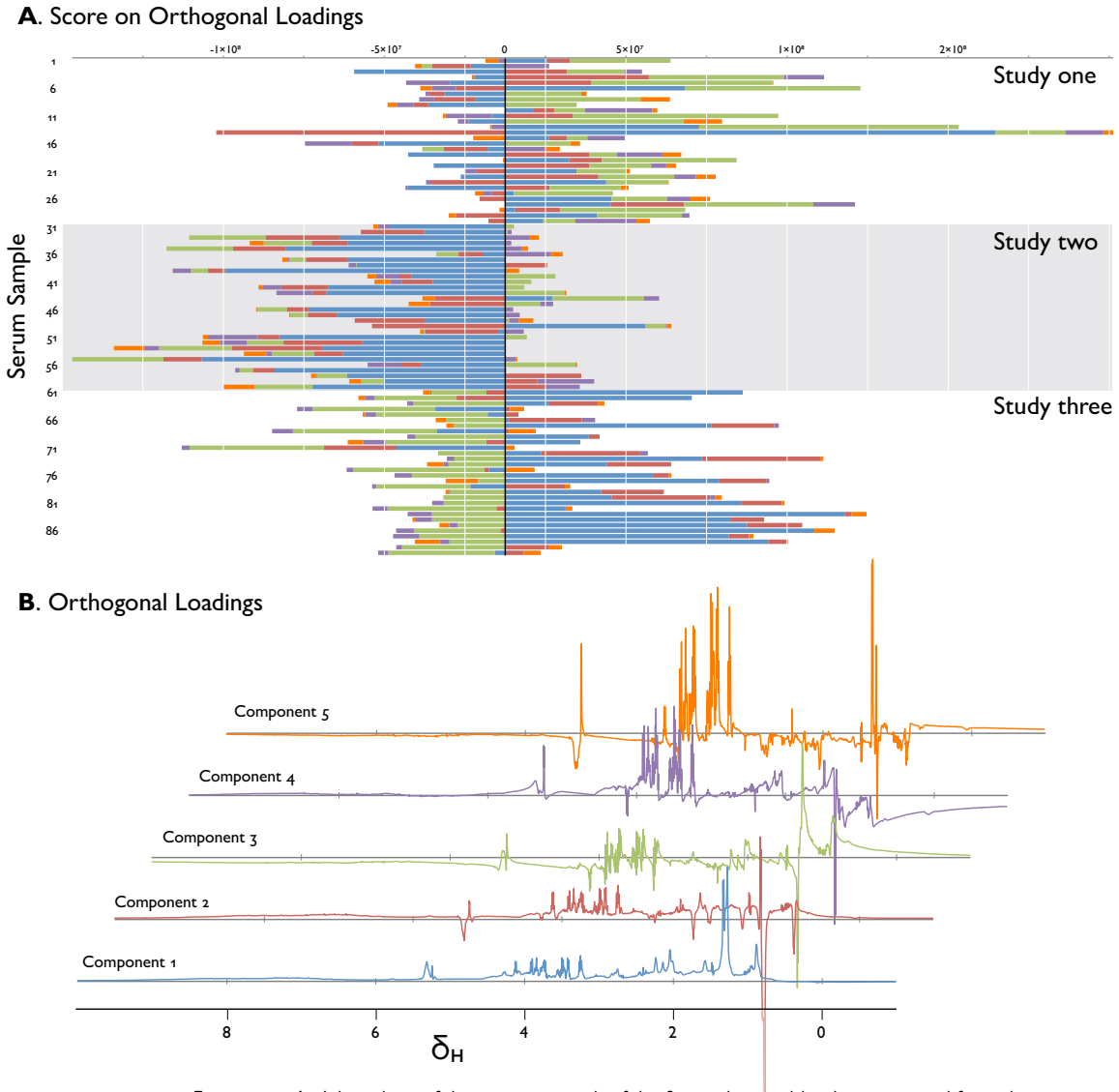

Figure 5.10: **A**. A bar-chart of the scores on each of the five orthogonal-loadings removed from the serum spectra in § 5.3.5. The orthogonal scores can be seen to be strongly related to the study of origin. **B**. The loadings for each of the orthogonal-components. All loadings are to scale, and zero is marked by the grey baseline.

modest, the results from § 5.3.1 suggest that OFSQ would perform much better on a compound that constituted a lower proportion of the total area of the sample.

Potential does remain for a cautious use of OFSQ in situations were a compound is overlapped by a well understood orthogonal interference, specifically where prioritising the quantification of specific molecules is desired. Particularly, if the number of orthogonal components can be inferred from the experimental set-up there can be more confidence in the use of OFSQ. However it cannot be recommend for general use in spectral quantification and has therefor not been pursued any further during the course of this thesis.

Chapter 6

Detecting the Effect of Liver Toxins with ^{1}H-NMR of Serum

6.1 Estimation of Liver Injury from Serum Sampling

The previous chapters have described various techniques related to the manipulation and standardisation of a large set of metabolite profiles, specifically NMR spectra and their related metadata. Along with the standardisation, each step has attempted to minimise the affect of subjective human judgement on the preparation of the data, ensuring that data processed by separate individuals is as consistent as possible. Now, to demonstrate the utility of these steps, these prepared data will be used to assess the possibility of identifying biomarkers of exposure to liver toxins in rat serum.

The affect of any toxin on a bodily organ must be assessed in terms of the degree of injury inflicted to the organ. This is typically achieved by histopathological examination of the tissue following necropsy. While highly effective, this process has limitations to the assessment of toxic insult in live subjects, where depending on the tissue, biopsy may not be an option. It also adds to the expense of toxicological studies, greatly increasing the number of experimental subjects needed to track the development of a lesion over time, as this can only be achieved with a staggered necropsy across the course of the study. The process of fixing samples for histopathology is also time-consuming and requires the attention of a trained histopathologist to make the final assessment of damage, an assessment that is also subjective.

Considering this, much research has been conducted into biomarkers that could allow a non- or minimally-invasive assessment of toxic insult, this would allow assessment of a xenobiotic's effects to be made earlier and potentially with more sensitivity. Any such biomarker may be expected to provide one of three pieces of information: to indicate exposure to a specific toxin, to indicate a response or toxic effect, or indicate susceptibility to a condition (Timbrell, 1998). Within these categories, a biomarker will be judged on two basic criteria; first, sensitivity, the ability of the diagnostic to detect the target state; and second, specificity, the degree to which a biomarker is unique to that particular state. Along with these concerns, consideration may be paid to the ease and cost of sampling, depending on the purpose of the test.

As analytical and statistical techniques have advanced, the use of biomarkers has expanded from pre-clinical toxicology to applications in risk characterisation and assessment. In their review of the COST Action B15, Gundert-Remy *et al.* (2005) outline further work to extend the current Type 0/1/2 biomarker paradigm, defined by the FDA, into a wider rage of six

classes applicable beyond toxicology. They also discuss the need for appropriate validation of biomarkers, particularly with respect to the hazards of multiple-testing. By measuring many biomarkers simultaneously NMR provides the opportunity to indicate the metabolic state of an entire system, providing additional information, when compared to more limited clinical-chemistry.

Due to the central role of the liver in detoxification, biomarkers of exposure to liver toxins are particularly valued, both for animal systems, where they may provide a useful ability to assess liver damage during a toxicological trial, without requiring a staggered series of necropsies, and in humans where non-invasive assessment of liver injury is vital in pre-clinical and clinical settings. At present, several markers of liver injury are commonly tested for in blood-serum; these include enzyme assays such as aspartate aminotransferase (AST) and alkaline phosphatase (ALP); or markers of liver function such as bilirubin levels (Zimmerman, 1999).

The present gold-standard for the non-invasive evaluation of liver injury in experimental mammals is the blood alanine aminotransferase (ALT) level (Timbrell, 2000; Amacher, 2002). ALT is an intracellular enzyme primarily expressed in the cytosol of tissues of the liver, kidney and heart. Damage to any of these tissues will result in leakage of ALT into the bloodstream. In the liver, hepatocytes express ALT at markedly higher levels than any other tissue, and so at higher levels, serum ALT is considered a liver-specific marker of damage.

One of the key limitations of ALT as an indicator of exposure to a liver toxin results from its mode of release from the cell. Because this requires damage to hepatocytes that is extensive enough to allow protein leakage into the bloodstream, or cause complete cellular necrosis, it is not a good indicator of lower levels of exposure that may result in hepatic stress rather than outright injury. Additionally, following an acute toxic-insult, ALT is rapidly cleared from the blood, with a half-life of 42 ± 11 hours in the rat (Kim, 1969), making it unsuitable for diagnosing liver-injury after the fact.

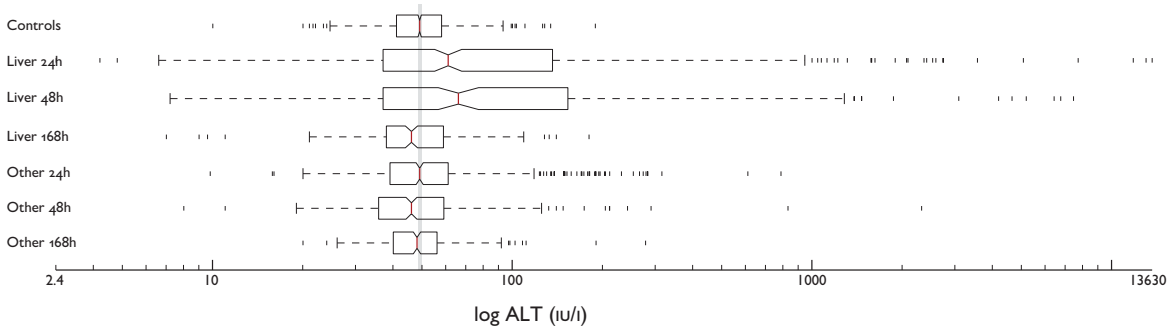

Figure 6.1: Distributions of serum ALT in control and high-dose treated rats in the COMET project. Treated animals are separated by timepoint and into those dosed with a liver toxin (liver), or any other compound (other). The elevation related to dosing of a liver toxin can be clearly seen. Whiskers extend to a maximum of 1.5 times the interquartile range, notches mark the comparison interval of the median at the 5% level.

An illustration of the distribution of serum ALT levels amongst different groups of samples in the COMET project can be seen in Figure 6.1. We see that the distribution of ALT levels in control animals and those dosed with non-liver toxins are approximately identical (all $p > 0.95$ by a Kruskal-Wallis test), the ALT levels for animals dosed with liver toxins show a markedly broader distribution, with a significantly elevated median ($p < 0.05$). ALT levels 168 hours after dosing show a similar distribution to the controls

Typical clinical chemistry measures of ALT are also susceptible to interference from other chemical-species. For example, dosing with hydrazine, a model liver toxin, may result in a decrease in observed ALT levels as the xenobiotic can act as an inhibitor of the ALT reaction used in the popular colourmetric ALT assay (Waterfield *et al.*, 1993; Lightcap & Silverman, 1996).

6.1.1 Rationale for a metabolite profiling approach

By generation of a metabolite profile based metric of the effect of a liver toxin, we have the potential to examine the link between NMR data and liver toxicity, and compare the information thus obtained to ALT.

In previous work, metabolic profiling of several biofluids has proved successful in detecting and assessing exposure to liver toxins, and the extent of injury. Urine-based studies have shown a strong association between bile acids detected in the urine and biliary toxins, allowing the separation of biliary and parenchymal injury (Beckwith-Hall *et al.*, 1998). Similar results are reported in studies based in serum and tissue data (Coen *et al.*, 2004), while further work has demonstrated detection of markers of liver regeneration (Bollard *et al.*, 2009).

From this previous work, we may hypothesise that metabolite profiling of serum has the potential to provide an assay complementary to ALT as a measure of liver toxicity, potentially detecting biochemical processes involved in liver injury that are separate to those that result in elevated ALT levels.

To generate a Metabolite Profile of Liver Effect, or MPLE metric, the entire set of COMET serum spectra will be used to generate a multivariate model of exposure to a liver toxin. This takes advantage of the breadth of toxins in COMET, averaging out the individual metabolic effects of each toxin, and thereby generating a more generalised model of exposure.

The analysis in this chapter will follow the outline in Figure 6.2. First, in step 1, the relevant spectra and metadata are extracted from the COMET database. To ensure the homogeneity of the spectra, they are then reprocessed and outliers excluded according to the rationale outlined in § 6.2.1.

In step 2 the spectra are split into two groups, one of which is held aside to provide a validation set for the models generated on the remaining data, used to train the models. Once separated into training and validation sets, each set will be independently normalised, both to MFC and apportionment-entropy →. As described in Chapter 4.

In step 3, to maximise the performance of the models, several variable restriction ap-

proaches (§ 6.2.3) are used to produce several sets of restricted spectral variables. These sets are expected to be significantly associated with the dosing of liver toxins.

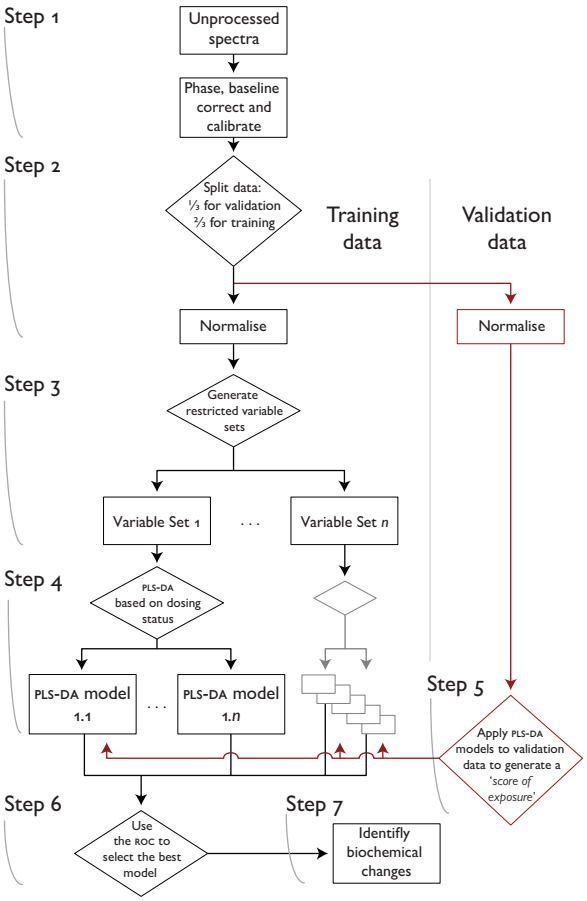

Figure 6.2: Flowchart showing the broad stages of data preparation and analysis undertaken in this chapter. Spectra in the validation set are normalised separately, and are projected into the models generated on the training set to generate MPLE scores.

In step 4 these variable sets are then used in the generation of predictive models of exposure to liver toxins. The groups of samples used to define the classes in the discriminant analysis are varied, this variation is used to generate both tightly focused models, for example, only 48 hour liver-toxin dosed samples vs 48 hour all other interventions, and more general models such as all time points liver-toxin dosed versus all time point controls.

Classes in the discriminant analysis were generated such that animals dosed with a high-dose of a liver toxin were considered one class (the positive class), while the control animals, or those dosed with other toxins were assigned to the opposing class (the negative class). To avoid confusing the models with samples that may not show a clear metabolic response, all low-dose samples were excluded from the models, regardless of the class of toxicity.

This comparative approach results in several models of exposure, each generated with a different combination of variables and samples (§ 6.2.4). In step 5, to select the model providing the strongest MPLE score, the validation data is predicted into each of the models to generate a MPLE score for each. Next in step 6 the scores generated for each sample on the models, are compared according to sensitivity and specificity using ROC analysis (§ 1.6), allowing the selection of the model that provided the MPLE score most strongly predicative of exposure to a liver toxin (§ 6.3.1).

Once the strongest model is selected, we can then generate a MPLE score for every serum sample in COMET, and compare this score to ALT and other indicators of dysfunction, such as weight change (§ 6.3.3).

Finally in step 7, by examination of the model coefficients, along with an assessment of

the performance of the MPLE score on a toxin-by-toxin basis, we can identify the source of the biochemical changes contributing to the MPLE metric (§§ 6.3.4 – 6.3.6).

6.2 Data Analysis Methods

6.2.1 Standardisation of NMR spectra

NMR spectra of serum were acquired as described in Appendix B.2. To prepare the NMR for analysis, and ensure the greatest possible consistency between data-sets acquired by many individuals over the course of several years, the acquired data were completely reprocessed from the Free Induction Decay (FID). Initially, new real and imaginary data were generated from the FID by means of the `multiefp` command in the XWin-NMR software. Following this, the entire set of Fourier-transformed spectra, both real and imaginary parts, were imported into MATLAB at the full spectral resolution of 32,768 points. Each spectrum was then automatically phased and baseline-corrected using an in-house routine, and referenced to the glucose doublet at δ_H = 5.233 with the method described in Chapter 3. These spectra were then down-sampled to 20,001 points between δ_H = -1 & 10, using linear interpolation, placing all the spectra onto a common PPM scale. Concomitant to the import of spectral data, the title data and file path of each spectrum were recorded to provide a unique identifier.

Once imported into MATLAB and processed, spectra were inspected by eye to ensure their quality. The spectra were judged on five criteria:

- Correct phasing.
- Correct baseline, *i.e.* flat with the noise centred about zero.
- Adequate water suppression, with minimal residual signal and associated baseline distortion.
- Correct frequency-calibration, judged by the position of the α-anomeric glucose doublet at δ_H = 5.233.
- Adequate line-shape, defined as the half-height line-width of the α-anomeric glucose doublet at δ_H = 5.233 falling below δ_H = 0.015.

Spectra failing any of these tests were excluded from further analysis, as were any other malformed spectra, as judged on a case-by-case basis.

At this point it was observed that spectra acquired early in the COMET project differed noticeably from those acquired later, mainly due to a lower signal-to-noise ratio. Examination of the `##$PROBHD=` parameter in the Xwin-NMR *acqu* file revealed that these spectra were those that had been acquired on a flow injection probe rather than in tubes →, as were the bulk of the samples. Despite the lower signal-to-noise, it was decided to retain theses samples, as the spectra were essentially good and the difference in intensity could be rectified by normalisation.

See Appendix B.2 for the reason for this change, and *Probehead* in Table A.1 for study-by-study details.

Following these steps, the data-set consisted of 2,539 CPMG spectra from 97 studies. Next the residual water resonance between $\delta_H = 5.13 - 4.4$ was removed. To allow for robust validation of the generated metric, at this point the spectra were split into training and validation sets. The validation spectra were then set aside during the modelling process, to provide an external test-set.

To ensure the same distribution of ALT levels in the training and validation sets, all spectra were ranked by ALT level, and every third spectra was assigned to the validation set, resulting in a training set of 1,692 spectra and a validation set of 847 spectra. The training and validation spectra were then independently normalised to median-fold-change, and apportionment-entropy, as described in § 4.1.1, to assess the performance of the method developed in Chapter 4. Finally, to reduce the computational load of dealing with such large data and lessen the effect of any shifting resonances, these spectra were reduced to a resolution of $\delta_H = 0.0090$ per point ← for the purposes of pattern recognition. This resulted in 1,100 variables between $\delta_H = 10 - -0.6113$.

Calculated to result in a whole number of spectral points per integral region. Integrals were generated starting from high PPM to low, by the addition of all data points within the range. Each region was identified by its mean PPM value.

For comparison, the above steps were repeated with the NOESYPR1D spectra, with the exception that water was only excluded between $\delta_H = 5 - 4.4$ due to improved water-suppression in the NOESYPR1D data. Due to the smaller exclusion region, variables ran between $\delta_H = 10 - -0.4957$, remaining at a resolution of $\delta_H = 0.0090$. This resulted in a set of 2,879 NOESYPR1D spectra, 1,920 for training and 959 left aside for validation.

It should be noted that the greater number of NOESYPR1D spectra compared to CPMGs is due to a greater number of CPMG spectra being rejected during the manual inspection process. This was mainly due to failed water-suppression resulting in a water-resonance that distorted the remainder of the spectrum.

6.2.2 Collation of clinical chemistry data

The *title* data of the imported spectra were then used to match each spectrum to its associated metadata in the COMET database (Chapter 2), specifically the clinical chemistry measurements and dosing information for each sample. These metadata were extracted as a comma-separated-values (CSV) spreadsheet for import into the various statistical software packages used for the numerical analysis.

The compounds dosed in each study were categorised into one of seven classes according to their expected primary locus of toxicity, physiological stressor, pancreas, liver, kidney, testicular, bladder or multiple organs, by reference to the established literature (Lindon *et al.*, 2005a). In cases where this was ambiguous, precedence was given to the intended target organ, as stated by the study organiser in the dose-justification documentation.

Animal weight change over the course of each study was expressed as percentage change between each animal's pre-dose and sacrifice weight, as taken from the database. During linear-modelling, values for ALT were expressed as the natural logarithm to linearise the extreme concentration changes.

6.2.3 Variable-removal strategies

In such a large data-set, when significance-testing variables associated with a specific outcome, we must be aware of the risk of spurious correlations to our groups of interest, which may arise due to multiple testing. To minimise the likelihood of over-fitting each model to this spurious correlation, two variable removal approaches were developed to limit the number of variables used to generate each model, and thereby reduce the likelihood of spurious correlations driving the fit.

The *p*-values indicating the significance of the association between each variable and the potential metabolic status of the liver were generated by two approaches. These *p*-values could then be used to filter the list of variables to include only those with a high-likelihood of association with the dosing of a liver toxin.

Each spectral data-point from animals dosed with a high level of a liver toxin was tested against the control spectra, with a two-tailed Student's *t*-test, to find variables with significant differences between the groups. This produced a *p*-value for each variable indicating significance with respect to its association to a perturbation from the control state.

Alternatively, to attempt to control for possible systematic interference that may have obscured a significant difference from the basic *t*-test, ALT, which as described in § 6.1 is known to be associated with liver injury, was used as a surrogate marker to identify variables that could be expected to be associated with liver damage. An initial PLS model was generated from the training set, regressing each spectrum against the natural log of its ALT level. The weights of this model were then extracted, along with a jacknife-based estimate of the standard error of each weight. A *t*-test was then used to test against the null-hypothesis that a variable that does not contribute significantly to the model will have a weight of zero.

Two methods were used to filter out false positives (type-I errors) from the variables selected as significant. First, we make use of the False Discovery Rate (FDR) a technique first described by Benjamini & Hochberg (1995) and much used subsequently (Storey, 2002). The FDR is a non-parametric measure intended to control for multiple testing that gives us a handle on the proportion of variables we can expect to have been selected due to a spurious correlation. The FDR achieves this by filtering a ranked list of hypothesis tests according their *p*-value, to generate a subset with a known (or expected) proportion of false positives. For instance, a set of *p*-values filtered with a FDR of 20% would be expected to contain approximately 20% spuriously selected variables. It is important to note that the FDR makes no predictions as to which 20% may be false positives.

The *p*-values from both the above approaches were then filtered with a FDR of 20%, to generate the final list of variables.

Second, following the significance testing, knowledge of the behaviour of biological variables and instrument noise can provide an extra rationale for excluding or including variables. In particular, many of the variables these data are divided into capture spectral regions that are empty of resonances and contain only instrument noise. Clearly if any of theses variables

are observed to show a significant association with ALT, the variable can be written-off as an entirely spurious correlation and excluded from modelling.

A good mechanism for identifying the noise level in NMR spectra is provided by the characteristic distribution of variability in biological and instrument-derived variation. Instrument noise in NMR is homoskedastic by nature, meaning the variance of the noise is constant with respect to the intensity. Meanwhile biological systems typically display heteroskadicity, as the variation in concentration of a metabolite is proportional to its concentration. This distinction provides a method of differentiating regions containing only instrument noise from those containing meaningful biological data. By plotting the standard deviation of each variable in the training-set against its mean intensity ←, as seen in Figure 6.3, we can see their is a discontinuity in the relationship, between those for which variance is constant regardless of intensity, showing a horizontal trend on the plot, and those for which variance is proportional to intensity. It is then simple to separate those containing only noise from those containing biological variation by identifying the intensity below which the variation in variables is primarily homoskedastic.

Both on a log scale for clarity.

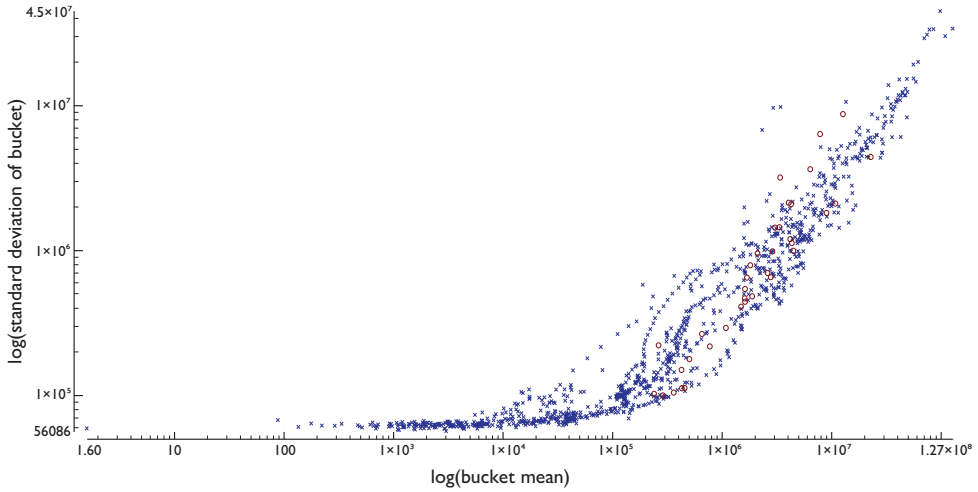

Figure 6.3: Standard deviation of variable intensities versus the mean of the variable across all the CPMG spectra in the training set. In regions dominated by noise, standard deviation is no longer proportional to intensity (instrument noise is homoskedastic, biological variation is heteroskedastic). Here the noise level is below 4×10^4. Variables identified as significantly associated with exposure to a liver toxin (set CPMG-MFC 1, see § 6.2.3 and Table D.1) are marked with red circles.

Using this approach, an intensity minimum was defined at 4×10^4, below which a datapoint was judged to represent only noise. This excluded 364 of 1,100 variables from the analysis, four of which had been identified as significant following the variable restriction conducted on the CPMG spectra by a direct t-test according to dose groups (CPMG-MFC 1 in Table D.1). None of the selected variables from the NOESYPR1D spectra were rejected by the noise criteria.

This is likely due to the globally elevated baseline observed in these spectra, which renders the majority of the spectrum above the noise threshold.

Two groups of restricted variable sets were generated from the CPMG spectra and are listed in Appendix D, one group for the MFC (Table D.1) and one for the apportionment-entropy normalised spectra (Table D.2). Variable sets for NOESYPR1D spectra were only generated from the MFC normalised data (Table D.3), as the further analysis of the CPMG data had shown the MFC normalised data to produce the stronger models.

6.2.4 Predictive PLS modelling

To generate models predictive of whether a particular exposure is causing, or likely to cause liver toxicity, by relating exposure to a liver toxin to metabolic changes observable in the serum, PLS-DA models were generated. Classes were based on whether each sample was taken from an animal dosed with a liver toxin. These PLS-DA models could then be used to generate a predicted MPLE score for each spectrum in the validation set, by projection of the validation data into the models. This score could then be directly compared to ALT as a metric of exposure. Varying permutations of groups and variable sets were used to generate several PLS-DA models, as an attempt to identify the combination providing the strongest ability to identify exposure to liver toxins.

To reduce the likelihood of generating models that focused on metabolic perturbations that were specific to one mechanism of liver injury, all liver toxins were considered together. This was intended to produce more general models of the metabolic processes, rather than modelling processes specific to one mechanism of toxicity.

The maximum number of components taken for each model was chosen by the number of components required to maximise the Q^2, on the condition that the Q^2 must always increase from one components to the next.

The various permutations of variable sets and classes produced several groups of models, listed in Table 6.1 for the MFC normalised CPMG variables, and Table 6.2 for those apportionment--entropy normalised.

When compared to the MFC normalised data, the models generated with apportionment-entropy normalisation had lower Q^2s after one component, but here the maximum Q^2 was not greater. This supports the observations in § 4.2.

Q^2 for NOESYPR1D-based models were generally lower than those derived from CPMG spectra.

ALT was included in the X block of model 7 (for the CPMG spectra, model 5 for the NOESYPR1D) to assess whether the serum spectra were conveying the same underlying information as ALT, or contained some additional information, in which case we might expect this model to perform better than the same model without ALT (model 6). Here the Q^2 is slightly higher for the model including ALT, but as seen in the next section, the models do not classify exposure any better than the equivalent without ALT.

Table 6.1: Model statistics of the PLS-DA models generated from CPMG spectra, normalised to median-fold-change, used to estimate liver damage. *N* in each class is marked in brackets, variable lists are in Table D.1.

#	Components	R^2Y	Q^2	Variables	Positive Class	Negative Class
1	3	0.118	0.079	CPMG-MFC 1	All liver tox HD (185)	All others (763)
2	2	0.256	0.176	CPMG-MFC 1	48h liver tox HD (89)	48h others (543)
3	2	0.256	0.176	CPMG-MFC 1	48h liver tox HD (89)	48h controls (258)
4	2	0.296	0.225	CPMG-MFC 2	All liver tox HD (185)	All others (763)
5	2	0.374	0.332	CPMG-MFC 2	48h liver tox HD (89)	48h others (543)
6	5	0.518	0.363	CPMG-MFC all	All liver tox HD (185)	All others (763)
7	2	0.522	0.368	CPMG-MFC all plus log(ALT)	All liver tox HD (158)	All others (763)
8	3	0.481	0.336	CPMG-MFC all	48h liver tox HD (89)	48h others (543)

Table 6.2: Model statistics of the PLS-DA models generated from CPMG spectra, normalised to apportionment-entropy, used to estimate liver damage. *N* in each class is marked in brackets, variable lists are in Table D.2.

#	Components	R^2Y	Q^2	Variables	Positive Class	Negative Class
1	3	0.034	0.025	CPMG-ent 1	All liver tox HD (185)	All others (763)
2	1	0.038	0.033	CPMG-ent 1	48h liver tox HD (89)	48h others (543)
3	1	0.076	0.057	CPMG-ent 1	48h liver tox HD (89)	48h controls (258)
4	4	0.274	0.206	CPMG-ent 2	All liver tox HD (185)	All others (763)
5	3	0.334	0.237	CPMG-ent 2	48h liver tox HD (89)	48h others (543)
6	3	0.285	0.187	CPMG-ent all	All liver tox HD (185)	All others (763)
7	3	0.289	0.193	CPMG-ent all plus log(ALT)	All liver tox HD (158)	All others (763)
8	3	0.393	0.294	CPMG-ent all	All liver tox HD (89)	48h others (543)

Again, these steps were repeated with the NOESYPR1D data, generating the models listed in Table 6.3.

Table 6.3: Model statistics of the PLS-DA models generated from NOESYPR1D spectra to estimate liver damage. *N* in each class is marked in brackets, variable lists are in Table D.3.

#	Components	R^2Y	Q^2	Variables	Positive Class	Negative Class
1	3	0.19	0.096	NOESYPR1D-MFC 1	All liver tox HD (236)	All others (1,033)
2	3	0.228	0.181	NOESYPR1D-MFC 1	48h liver tox HD (110)	48h others (517)
3	3	0.3	0.215	NOESYPR1D-MFC 1	48h liver tox HD (110)	48h controls (316)
4	1	0.102	0.091	NOESYPR1D-MFC 2	All liver tox HD (236)	All others (1,033)
5	3	0.153	0.102	NOESYPR1D-MFC all plus log(ALT)	All liver tox HD (236)	All others (1,033)

Once the listed models had been generated with data from the respective training set, the data from the validation sets were projected into each model, to generate a set of predicted *Y* scores, the MPLE scores, that could act as a surrogate index of exposure to liver toxins.

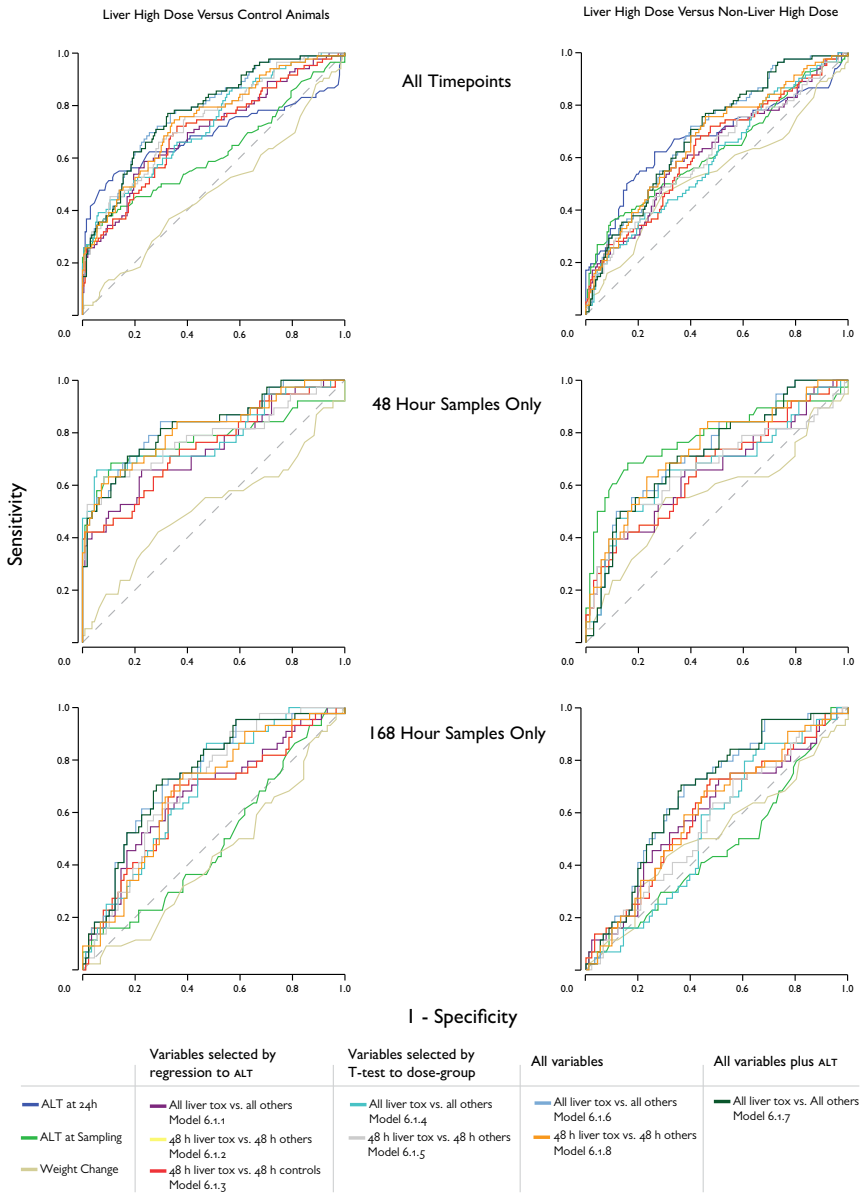

Figure 6.4: Rocs generated by projecting the validation set into the MPLE models in Table 6.1 (MFC-CPMG spectra). In each case the positive class was defined as those animals administered a high-dose of liver toxin. The majority of metrics can be seen to provide some indication of exposure to a liver toxin that is better than chance, except for percentage weight-change and ALT levels 168 hours after dosing.

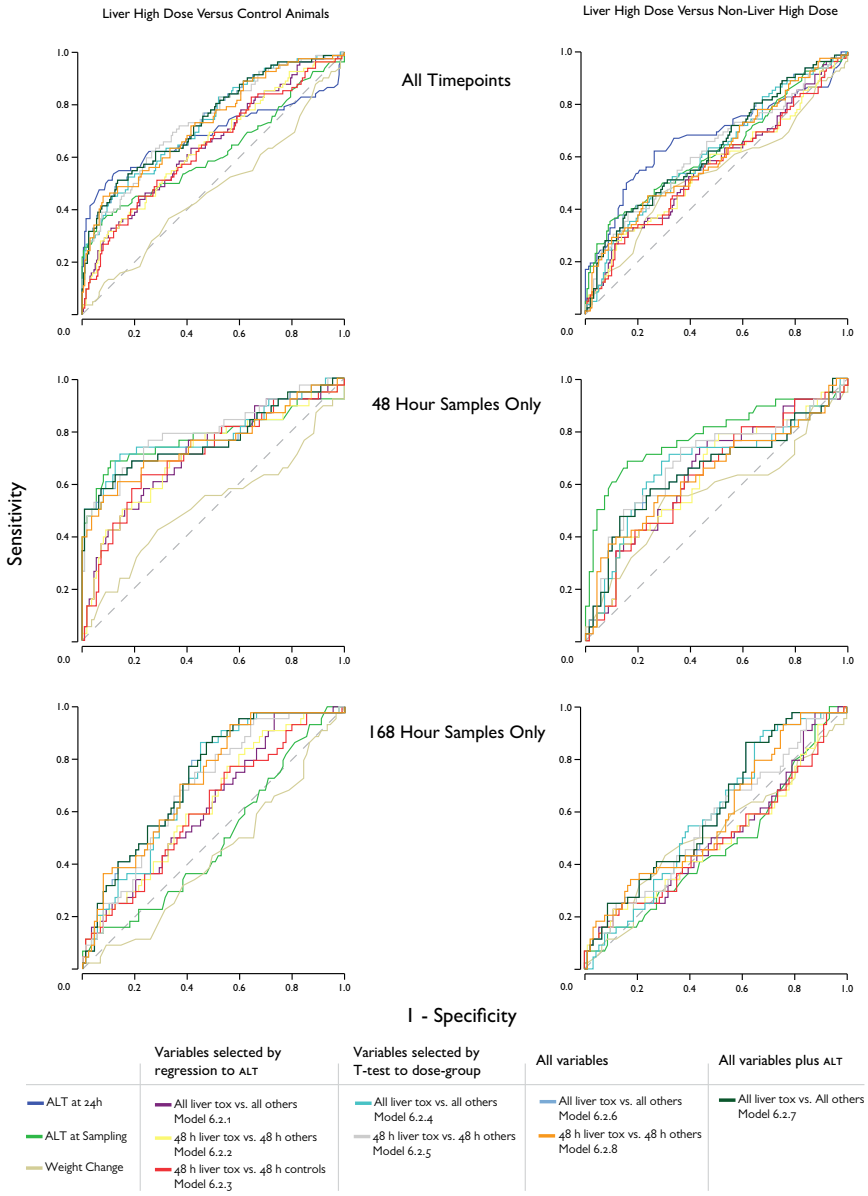

Figure 6.5: As in *f* 6.4, ROCs generated by projecting the validation set into the MPLE models in Table 6.2 (apportionment-entropy normalised CPMG spectra). In each case the positive class was defined as those animals administered a high-dose of liver toxin.

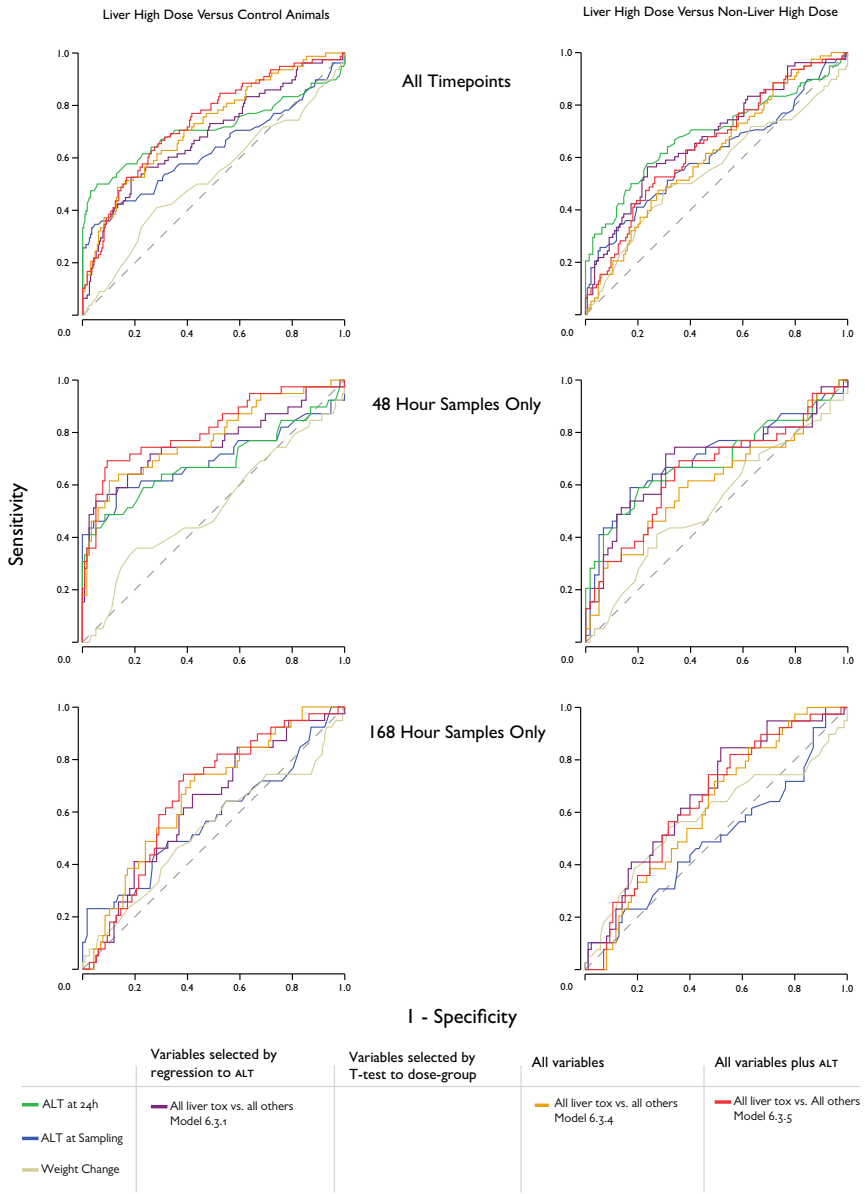

Figure 6.6: Rocs generated by projecting the validation set into the MPLE models in Table 6.3 (NOESYPR1D spectra). In each case the positive class was defined as those animals administered a high-dose of liver toxin. The majority of metrics can be seen to provide some indication of exposure to a liver toxin that is better than chance, with the exception of percentage weight-change and ALT levels 168 hours after dosing.

6.3 Analysis

6.3.1 Assessment of predictive models

Here the ROC was used to visualise the classification power of the MPLE scores, generated by each of the PLS-DA models in § 6.2.4 from spectra in the validation set. Category information was gathered from the dosing data, with those samples from animals dosed with high-doses of a liver toxin classified as positive (108 CPMG, 119 NOESYPR1D samples), while all other samples, controls (259, 318 samples) and high-doses of all other treatments (200, 200 samples), were considered negative. Again, all low-dose treatments (280, 322 samples) were excluded from the analysis.

Figures 6.4, 6.5 & 6.6 show ROC curves generated from each of the discriminant models generated in Tables 6.1 to 6.3, compared to the animals percent weight change and ALT level, both at the time of serum sampling, and 24 hours after dosing.

The ROC shows no significant predictive power for percent weight-change at any timepoint, or ALT sampled 168 hours post dose. All the other metrics showed an ability to discriminate samples from animals dosed with a liver toxin from samples from control animal and animals that were dosed with a toxin targeting any other organ. Full statistics for the ROCs can be seen in Tables D.4 & D.5 in Appendix D.

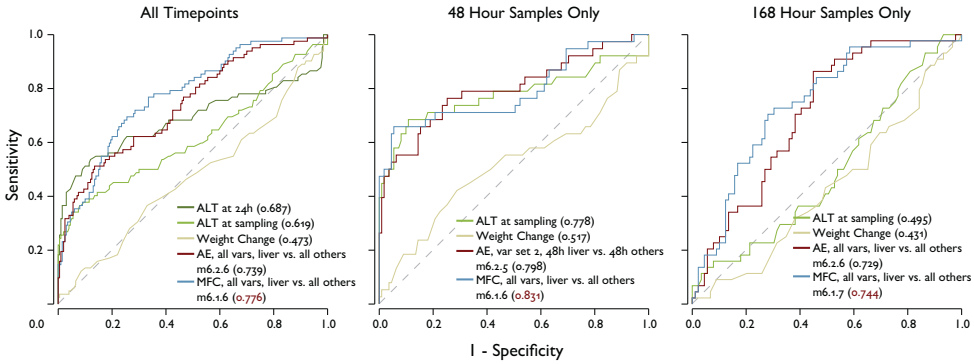

Figure 6.7: Comparison of the ROC curves of the best, CPMG-derived MPLE scores of exposure to a liver toxin versus control profiles. The area under the curve for each metric is indicated in brackets in the key. Although classification of exposure based on the MPLE based measures can be seen to improve on ALT in every case, the improvement can most clearly be seen at 168h. Lines shaded green are traditional measures, lines in red are entropy normalised, and lines in blue are MFC normalised.

To clarify the characteristics of the models, the ROCs of the best performing CPMG-based MPLE models have been isolated in Figures 6.7 and 6.8. Here best performing is defined as the model with the largest AUC.

In ∫ 6.7 the MPLE metrics are compared by their ability to distinguish liver-toxin dosed animals from controls. In every case we can see the MPLE metrics improve upon the sensitivity and specificity of ALT, including ALT measured 24 hours after dosing. This improvement

persists to the 168 hour timepoint, indicating the metabolic-profile based metrics also display a longer half-life. Of the two normalisation methods, MFC has produced the better metrics, the strongest PLS-DA model being 6.1.6, generated from MFC normalised CPMG spectra of all samples dosed with a HD of liver toxin versus all control samples and samples from animals dosed with a high-dose of a toxin affecting a different organ.

In the most general test, ignoring timepoint, this model had an AUC of 0.776 (95% confidence interval: -0.716, +0.835) compared to 0.739 (-0.674, +0.805) for the same model generated with apportionment-entropy normalised spectra, and 0.619 (-0.539, +0.886) for ALT at the time of sampling. Considering only the 48 hour timepoint, the AUC improves to 0.831 (-0.75, +0.912).

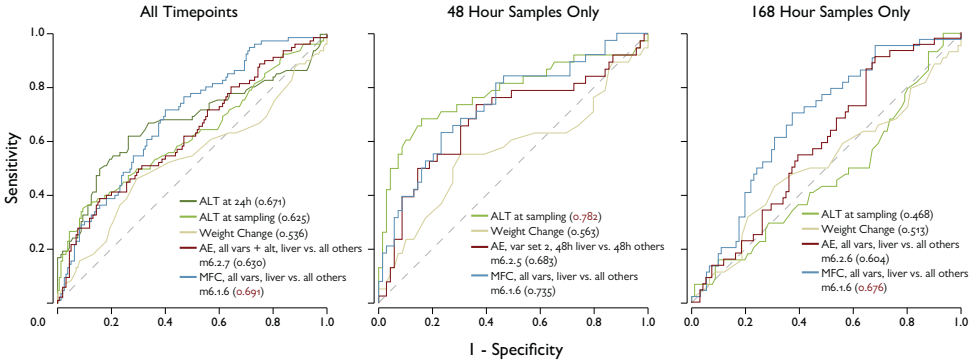

Figure 6.8: Comparison of the ROC curves of the best, CPMG-derived MPLE metrics *versus* high doses of other toxins. The AUC for each metric is indicated in brackets in the key. Classification of exposure is strongest by the MPLE based measures, and can be seen to improve on ALT in every case, with the improvement most clearly seen at 168h. The key and line shading are as *f* 6.7.

A similar situation can be seen in Figure 6.8, where the metrics were used to distinguish animals dosed with liver toxins from those dosed with a toxin affecting any other organ. The strength of discrimination is less than that versus controls, but the MPLE metrics still improve upon ALT, except at the 48 hour timepoint. Here the strongest MPLE model was once again MFC-normalised, 6.1.6, with an AUC of 0.691 (-0.623, +0.759) compared to model 6.2.7 from the apportionment-entropy normalised data 0.63 (-0.554, +0.705) and 0.625 (-0.546, +0.703) for ALT.

6.3.2 The most predictive models make use of the entire spectrum

The discriminant models that produced the best MPLE metric vary from timepoint to timepoint, but it is noteworthy that all the best models used all the variables in the data, rather than one of the sets filtered by variable-restriction. These models also showed the highest Q^2 in the MFC normalised data, indicating greater robustness.

The fact that models based on the complete variable sets showed the best performance

demonstrates the strength of chemometric techniques in teasing the underlying significant latent-variation out of data. Interestingly, despite the aims of the variable restriction, this may be due to the restriction criteria resulting in models that were over-fitted to the training data. It is also possible that the initial regression against ALT in variable set CM1 & *cetera*, drove the model toward variables that were too directly related to ALT levels, and not general to exposure.

The MPLE metrics showing the best discrimination all came from PLS-DA models generated with all spectral variables rather than one of the restricted variable sets. However, the final predictive performance is influenced by the groups of samples used to generate the initial models. Typically, the best discrimination at each timepoint was generated by models that were not created with spectra from specific timepoints, seeming to indicate that the biomarkers being modelled are consistent between the 48 and 168 hour timepoints. The only instance in which this was not the case was model 6.2.5, a model generated from 48 hour liver toxins *versus* other toxins, that outperformed other models when separating these specific groups.

The metrics from the NOESYPR1D data cannot be directly compared to the CPMG results, due to the slightly different samples-sets, caused by the differences in number and identity of spectra excluded. However examination of comparable models shows a similar, although slightly inferior, performance to those generated from the CPMG spectra. This would appear to indicate that macromolecules observed in the NOESYPR1D spectra are obscuring some of the relevant metabolic changes.

6.3.3 Performance of the strongest model on the entire COMET dataset

From this point on, reported MPLE scores were generated on Model 6.1.6, selected as the model producing the best discrimination in the greatest number of comparisons. This model consisted of MFC normalised CPMG spectra, using all variables and regressing high-dose liver toxins to control animals at all timepoints. Scores were generated on all samples, and while this did mean generating scores on samples from the training set that were used to generate the model, it allows a more detailed examination of the characteristics of individual studies, and was considered acceptable following validation of the model as described above.

In Figure 6.9 the variance of the control values for both ALT and MPLE scores have been scaled to be identical, allowing a direct comparison of the distributions of the values of MPLE and ALT in each study. A MPLE score above 0.5 can be seen to exclude the majority (approximately 99%) of control animals, and is a good cutoff between normal and 'exposed' states, for ALT, the equivalent level is approximately 90 IU/l.

Due to the sheer number of compounds this plot is rather crowded, for clarity the compounds have also been plotted in three separate Figures D.1, D.2 & D.3, in Appendix D.

The sensitivity of each metric was compared on a study-by-study basis, treating studies in which the median value of the metric in the high-dose animals was above the 99% cutoff, as a successful prediction of exposure. While both metrics give a similar number of false-positives,

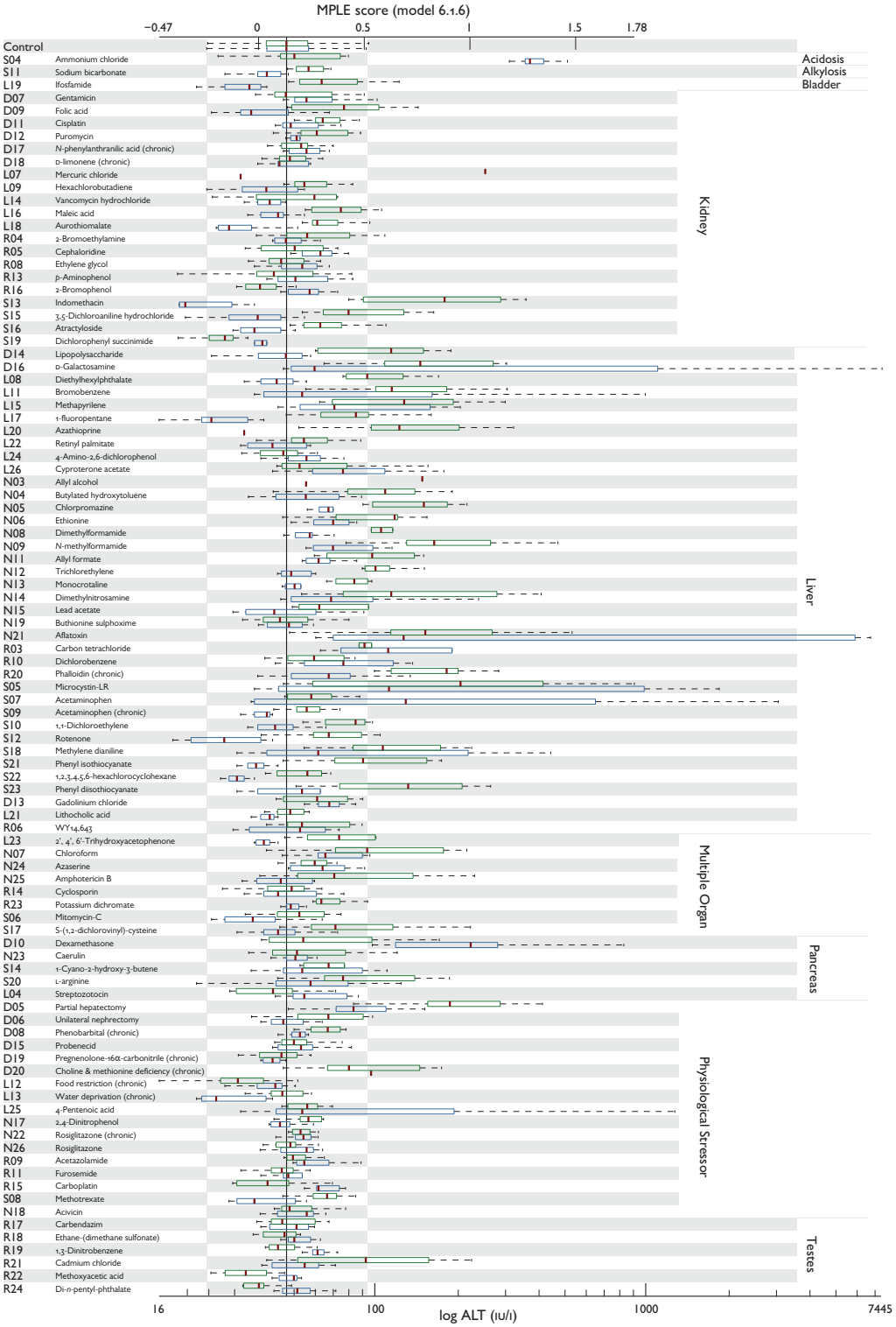

Figure 6.9: Study-by-study breakdown of the MPLE score on model 6.1.6, generated from MFC-normalised spectra of animals dosed a liver toxin, versus all other HD and control spectra at all timepoints (upper scale and box of each pair, in green) paired with the ALT level (lower scale and box of each pair, in blue) of every sample in the COMET project (training & validation). It can be seen that a MPLE score of above 0.5 is indicative of exposure to a liver toxin. Boxes and whiskers are as *f* 6.1. For comparison, both metrics have been scaled to show identical variance in the controls, their medians aligned and marked with a black line, and the 99% range about the controls on the MPLE metric indicated. See *f* D.1, D.2 & D.3, in Appendix D for expansions of this figure.

the MPLE metric detects many more of the liver-toxins compared to ALT (Figure 6.10). In total 20/38 showed an abnormal MPLE score, compared to only 1/38 for ALT.

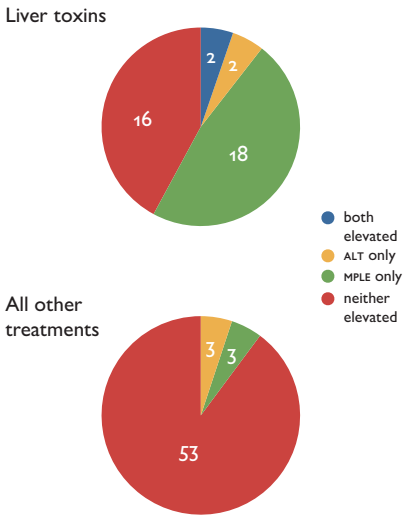

Figure 6.10: Tally of studies where the median level of either metric is elevated above the 99% cutoff, compared for high-dose liver toxins (top), and high-doses of all other interventions.

It can be seen that in some compounds targeting the liver, such as acetaminophen (S07), carbon tetrachloride (R03) and aflatoxin (N21) the ALT response is exceedingly strong, with levels elevating by an order-of-magnitude or more, while the MPLE response is more restrained.

Notably, systemic acidosis induced by ammonium chloride (S04) induces a large elevation in ALT. As acidosis may result from several common causes, often related to the dosing of non-toxic xenobiotics, this has the potential to confuse the diagnosis of liver injury by ALT.

Interestingly, study D05, partial hepatectomy, shows an elevated MPLE metric, suggesting that the model is detecting metabolic effects present during liver regeneration rather than outright damage. This effect does not extend to the nephrectomy study (D06), indicating we are not modelling a common response to surgical intervention.

6.3.4 MPLE scores provides complementary information to ALT

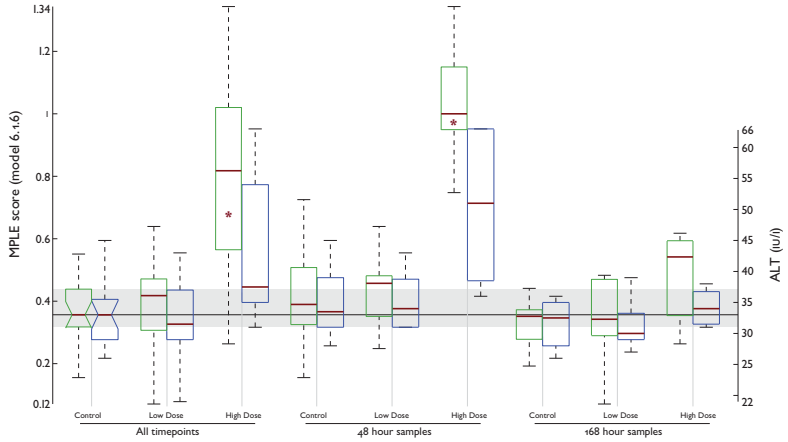

Figure 6.11: Box plot of MPLE scores on model 6.1.6 (green, left scale) and ALT values (blue, right scale) for two isothiocyanate derived compounds (Phenyl isothiocyanate and phenyl diisothiocyanate, 55 samples total). The 48 & 168 h samples have been separated, and it is possible to see that the 48h samples show a much stronger response. Comparisons marked with * are significantly different from the controls at $p < 0.05$.

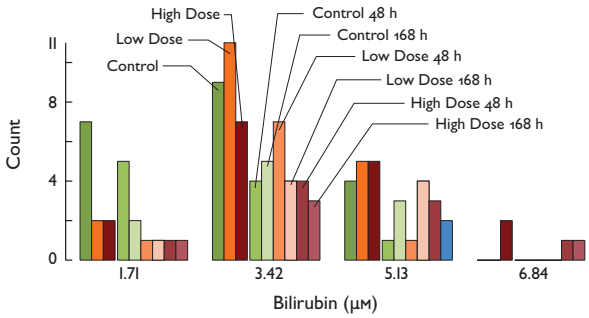

Figure 6.12: Bilirubin would be expected to be elevated in response to biliary toxicity. While this may be the case, the clinical-chemistry measures of bilirubin in studies s21 & s23 are so coarsely quantised, reporting only four unique values, that the data is hard to interpret. No group shows a significant difference from the controls ($p > 0.95$). Groups are identical to *f* 6.11, 60 samples total.

To examine these variations in greater detail, the metrics from individual studies have been selected, and expanded into the accompanying figures.

Phenyl isothiocyanate and phenyl diisothiocyanate are biliary toxins affecting the epithelial cells of the bile duct (Xu *et al.*, 2004) →. Despite the smaller range about the controls in these studies, the ALT levels never exceed the 90 IU/l upper limit of the controls seen in *f* 6.9. The MPLE-metric and ALT values for these studies are broken out in Figure 6.11, where we see a clear temporal effect, with the elevation of the MPLE-metric persisting in the 168 hour samples, while ALT levels have returned to approximately normal. Whilst non-significant ($p > 0.05$ by

While α-naphthylisothiocyanate (ANIT, N02) was also studied in COMET, the ALT values were incomplete, necessitating its exclusion from this comparison.

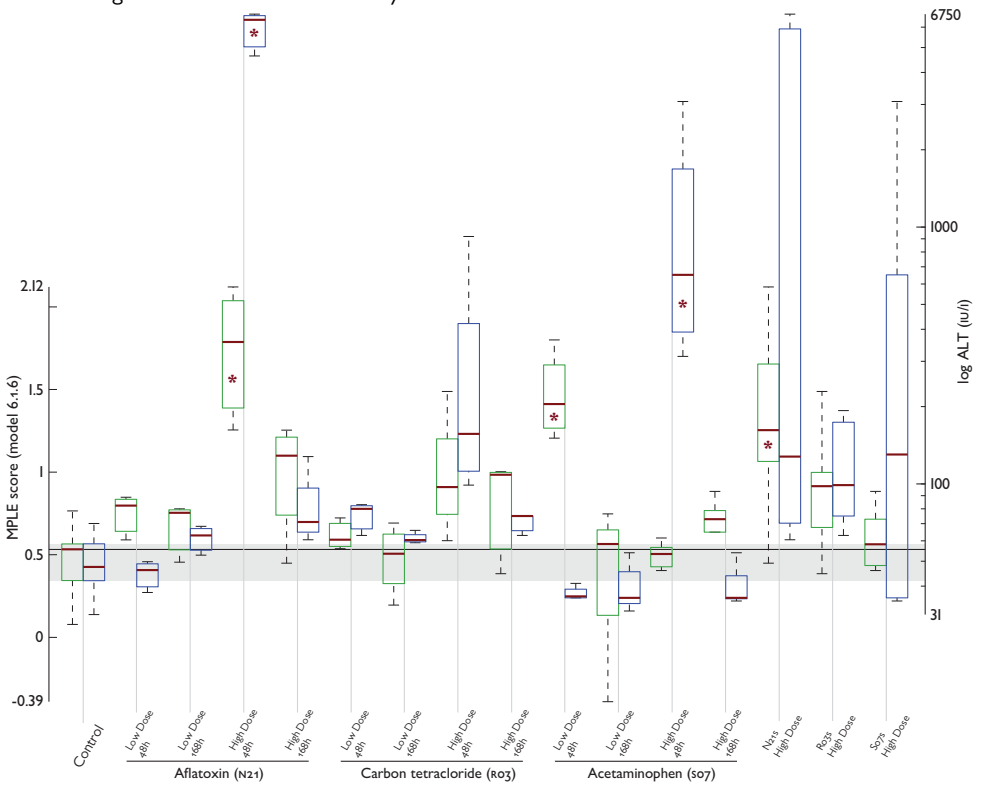

Figure 6.13: Boxplot of MPLE scores on model 6.1.6 (green, left scale) and ALT values (blue, right scale) for three studies, aflatoxin (N21), carbon tetrachloride (R03) and acetaminophen (S07), where the ALT show an extreme response. Despite not showing as extreme a response as ALT, the MPLE scores show a response to dosing. Due to the wide range of ALT values, these are plotted on a log scale. Comparisons marked by * are significantly different from the controls at $p < 0.05$.

a Kruskal-Wallis test), it also appears that the MPLE metric shows a dose-response for these compounds that is absent in ALT, with a slight elevation in the low-dose 48 hour samples.

Elevated bilirubin, a product of hæm catabolism, would be expected due to cholestasis resulting from biliary toxicity (Tjandra *et al.*, 2002). While bilirubin levels were recored in COMET (see Figure 6.12), the values for studies S21 & S23 are so quantised, taking only four unique values, that it is difficult to draw any conclusions with respect to the relationship between measured bilirubin and the MPLE score. Some high dose animals appear in line with the Tjandra *et al.* report of a doubling of control levels in animals dosed with ANIT, but not all.

The converse effect, liver toxins that result in high ALT levels but a low MPLE-metric are also observable in the aforementioned acetaminophen, carbon tetrachloride and aflatoxin studies amongst others. These specific toxins act by the formation of a reactive intermediate, which depletes glutathione. Examining the metrics on a study-by-dose-group basis in Figure 6.13, the extreme strength of the ALT response can be seen, with the median elevated by $10-100$ times in high dose animals. We also continue to see the short half-life of ALT, resulting in levels in the high-dosed group being depressed compared to that of the control animals by 168 hours. De-

spite the high specificity of ALT in these studies, MPLE scores appear more sensitive, particularly at 168 hours.

These studies also exemplify the extreme variability in ALT levels. Within each of the dose-groups that demonstrate an ALT response, the levels between individual animals may vary by an order of magnitude or more.

The inconsistency in reporting of histopathological data (§ 2.3.2) means it is risky to draw any strong conclusions with respect to the relationship between these metrics and the prevalence of observable necrosis. This variability in reporting was also a strong motivator for the use of ALT in the regression models generated in § 6.2.3, as the objectivity of this measure was superior to the judgment of the various histopathologists, whose scoring although individually reproducible, could exhibit considerable variation when considered together.

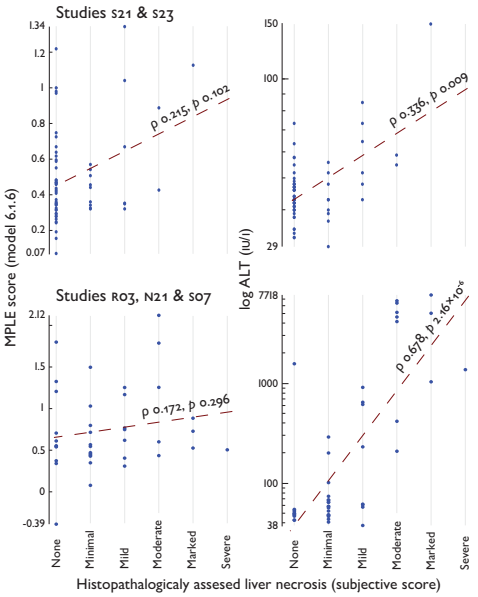

Figure 6.14: A comparison of the most severe necrotic liver-lesion, as reported by histopathology, to measured MPLE metrics (left) and ALT (right). The line of best fit is plotted in red, with the Spearmans ϱ marked. Data were taken from the studies previously presented in *f* 6.11 & 6.13.

When comparing the ALT or MPLE scores directly to the histopathologically determined extent of injury (Figure 6.14), there does not appear to be a strong relationship between the MPLE score and the degree of necrosis (max Spearmans's ϱ 0.215, p 0.102). As expected, ALT shows a much clearer relationship (up to ϱ 0.6775, p 2.16×10^{-6}). Here the severity of the injury is defined as the extent of necrotic lesions of the liver, scored from none to severe on a six-step scale (regardless of locus).

6.3.5 Effect of dietary manipulation on MPLE

Three COMET studies investigated the effect of dietary modification on metabolism in the rat, one study each on; food restriction, water restriction, and one study examining the affect of a choline & methionine deficient diet. Figure 6.15 isolates the MPLE scores and ALT levels for these studies from *f* 6.9. Restriction of food or water can been seen to decease, albeit non-significantly ($p > 0.95$ by Kruskal-Wallis test) the values of both metrics, while animals fed a choline & methionine deficient diet score show a significant elevation in the MPLE metric ($p < 0.05$), with ALT remaining low. Choline deficiency is known to cause hepatic steatosis, by reducing VLDL secretion (Li & Vance, 2008). With respect to the observed ALT levels, previous literature reported in Levin *et al.* (1993) is ambiguous, with the authors reporting elevated ALT in response to food deprivation, while previous literature reports a decrease (Oishi *et al.*, 1979).

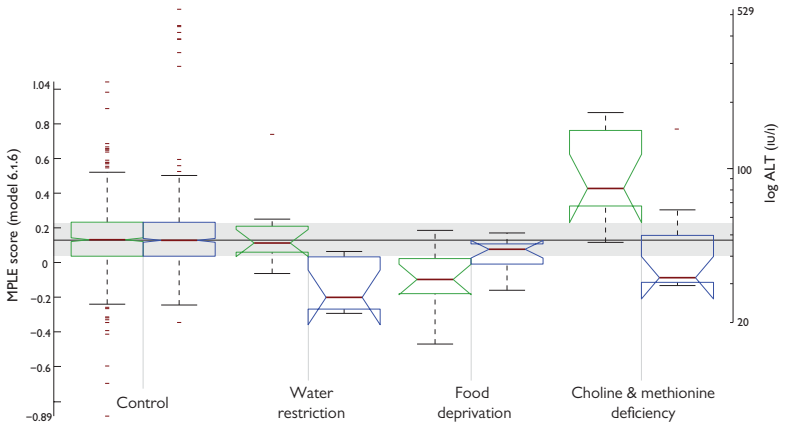

Figure 6.15: Boxplot of MPLE scores on model 6.1.6 (green, left scale) and ALT values (blue, right scale) for three studies of dietary modification. While food and water restriction do not appear to positively influence either metric, MPLE scores are sensitive to choline & methionine restriction. Due to the wide range of ALT values, these are plotted on a log scale.

6.3.6 Metabolite identification

To identify the NMR resonances, and hence metabolites, responsible for the discrimination observed in the MPLE model, the PLS coefficients were examined. To simplify the identification of resonances, Figure 6.16 shows the result of generating an O2-PLS-DA model with identical parameters ← to model 6.1.6, the discriminant model used to generate the MPLE scores, but using high resolution NMR spectra ←. As outlined in § 1.5.5, generating a new O2-PLS model for visualisation has the advantage of moving all the variation correlated with exposure to a liver toxin into the first component, allowing a simpler visualisation of the data. Model statistics for the high-resolution model were: R^2Y 0.515, Q^2 0.128. The lowering of the Q^2 with respect to model 6.1.6 is most likely due to shift variation compensated for by resolution-reduction, but present in the high-resolution data.

One correlated component, three Y-orthogonal components and no X-orthogonal components. 20,000 points per spectrum.

It can readily be seen that the majority of the significantly perturbed variables are associated with lipid resonances. Primarily, there is a decrease in the levels of serum lipids, specifically resonances attributable to the very- and low-density lipoproteins (VLDL & LDL), as evidenced by the greater weighting of the down-field sides of the broad resonances at δ_H = 0.88 & 1.25 (Ala-Korpela *et al.*, 1993; Ala-Korpela, 1995; Tukiainen *et al.*, 2008). Decreases in resonances from the backbone of fatty acids and glycerol would be consistent with a related decrease in the triacylglycerol payload of VLDL & LDL. It is notable that the resonances about δ_H = 2.8 appear to indicate the reduction in fatty-acid levels is skewed in favour of the longer chain polyunsaturated fatty acids, such as 20:4(n-6) and 22:6(n-3) rather than 18:2(n-6) ← (Willker & Leibfritz, 1998; Coen *et al.*, 2003). The lower contribution of the fatty acid resonances around δ_H = 2, which are also overlapped by 18:1(n-9) fatty acids may indicate that variance in the

20:4(n-6), Arachidonic acid or *all-cis*-5,8,11,14-eicosatetraenoic acid.
22:6(n-3), Docosahexaenoic acid or *all-cis*-4,7,10,13,16,19-docosahexaenoic acid.
18:2(n-6), Linoleic acid or *all-cis*-9,12-octadecadienoic acid.

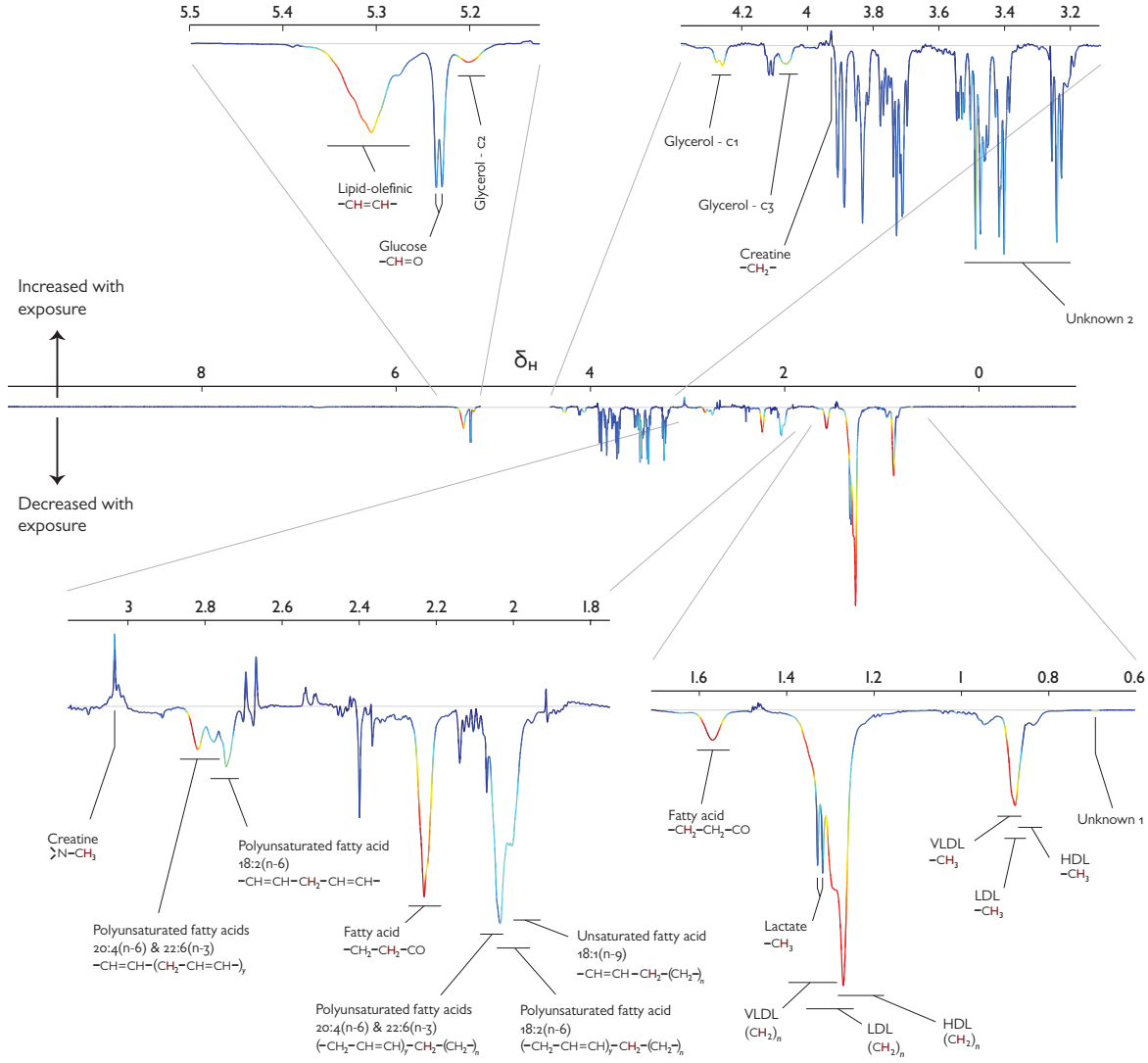

Figure 6.16: Visualisation of the O2-PLS loadings (Cloarec *et al.*, 2005b) of a high-resolution reproduction of model 6.1.6. Resonances positively associated with exposure to a liver toxin are raised above the grey baseline, while those negatively associated are below. The shading varies from blue, least significant variables, to red, the most significant variables. The majority of significant resonances are related to lipid moieties, especially those likely to be associated with lipoproteins. R^2Y 0.515, Q^2 0.128.

18:1(n-9) fatty acids is not directly related to exposure to a liver toxin and this is obscuring the changes in the poly-unsaturated fatty acids.

Considering the observed changes, the small singlet resonance marked as 'unknown 1' at $\delta_H = 0.78$ in ∫ 6.16 is most likely the C21 -CH_3 from the cholesterol component of VLDL & LDL or a derivative, such as a bile acid or ester, thereof. Further resonances from any of these species may be expected to produce the overlapped signals marked as 'unknown 2' observed in the range of $\delta_H = 3.8 - 3.2$, obscured by sugar resonances. However without additional NMR experiments this cannot be identified definitively.

The lactate doublet overlapping the lipid resonance at $\delta_H = 1.25$ is not significant to the model; nor do the glucose resonances at $\delta_H = 5.233$ and in the range of $\delta_H = 3.2 - 4$.

6.4 Discussion

By modelling the COMET serum data in terms of exposure to a liver toxin, I have been able to relate metabolic changes in the serum to an apparent impairment of liver-function. This relationship has several notable aspects:

6.4.1 MPLE scores provide information on liver function

The strength of the response of the metabolic-profile based metric in the partial-hepatectomy study suggests that the metabolic effects the model is using to predict exposure to a liver toxin are either markers of liver regeneration, or some effect caused by a restriction of liver function, perhaps representing a decreased ability to maintain some aspect of homeostatic balance.

The work of Bollard *et al.* (2009) examining the metabolic affects of partial hepatectomy has reported a decrease in serum triglycerides as a marker of liver regeneration, but only at timepoints 168 hours after dosing. Bollard *et al.* also reported increases in alanine and betaine, that do not contribute significantly to the MPLE model, although their observation of an increase in creatine is supported, albeit at a much lower significance than the lipid changes.

Despite the subjective variations in the histopathological data, the correlations seen in Figure 6.14, suggests that the MPLE-metric is less directly related to liver necrosis than ALT, reinforcing the hypothesis that we are modelling a separate process.

The metabolic changes observed in § 6.3.6, driving model 6.1.6, suggest a perturbation of hepatic lipid metabolism, with decreased levels of VLDL being secreted by the liver. Lipoproteins such as VLDL are large complexes of fatty acids, cholesterols and various binding proteins, that are the primary means of distributing lipids through the blood stream (see Figure 6.17, Pajukanta, 2008).

The observation of a preferential depletion of 20:4(n-6) and 22:6(n-3) is interesting, in that the work of Balasubramaniam *et al.* (1985) indicates that irrespective of diet, in rats, 22:6(n-3) fatty-acids are elevated in the percentage composition of VLDL in comparison to HDL, meaning a depletion of VLDL could be expected to result in the observed relative lowering of 22:6(n-

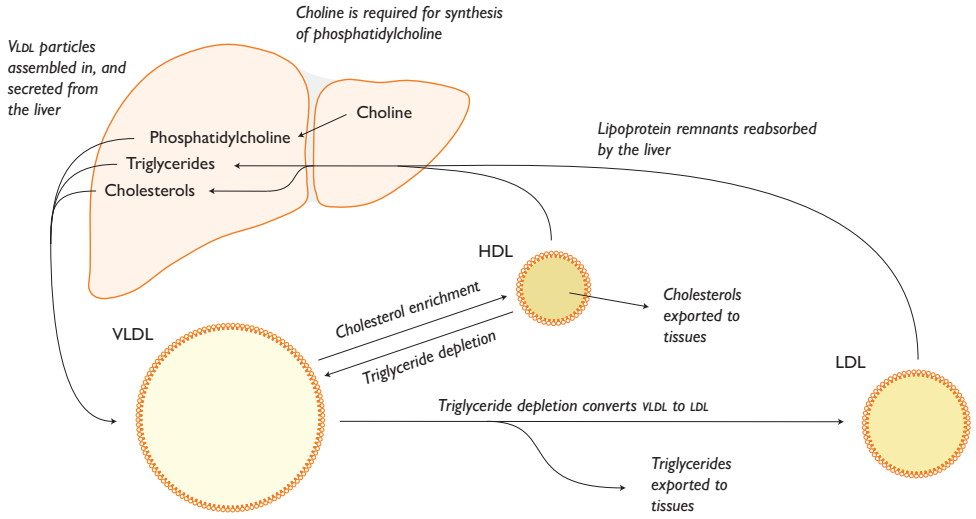

Figure 6.17: Schematic diagram of lipoprotein metabolism. VLDL secretion from the liver requires phosphatidylcholine synthesis. In the blood, cholesteryl ester transfer protein promotes the redistribution of the cholesterol component of VLDL to HDL from where it is transferred to the tissues. The transfer of triglycerides from VLDL to tissues reduces the particles density, and in combination with a modification of the protein component converts VLDL to LDL. LDL & HDL are then reabsorbed by the liver, and the components reused.

3) fatty acid levels. However, their work also indicates the same pattern for 18:2 fatty acids, and the opposite for 20:4(n-6) fatty acids. If the observed changes in resonances about $\delta_H = 2.8$ were primarily caused by the decrease in signals contributed by the constituents of VLDL, this would produce the opposite effect to that observed.

Coen *et al.* (2003) attribute decreases in levels of poly-unsaturated fatty acids to the β-oxidation activity of peroxisomes being increased in response to depleted energy levels. In the same work they also attribute the specific depletion in arachidonic acid to an inhibition of the phospholipase-A_2 pathway.

Care must be taken when considering the implications of the alterations in distribution of fatty acid species. Balasubramaniam *et al.* and further work by Schrijver *et al.* (1992) clearly show that the origin of the dietary lipids has a major effect on this distribution, and considering the source of feed in the COMET project was not controlled nor recorded, there is the possibility that the observed changes arise from dietary factors. However, it would appear unlikely that such an effect would correlate strongly with the dosing of liver toxins across 126 studies, supporting the hypothesis that the observed changes are related to toxicity.

An obvious cause of alterations to lipid metabolism is the loss of appetite that may result in sick animals eating less than their healthy counterparts. There are two pieces of evidence that suggest this is not the case, first, as seen in ∫ 6.15, animals in the food restriction (L12) study do not display an elevated MPLE metric, and second, the percentage weight change of the animals is not predicative of exposure to a liver toxin. If the observed changes did result from

a reduced appetite, it would be reasonable to expect a food restriction study to invoke the same metabolic changes, and sick animals consuming less feed than their control counterparts would be expected to show some reduction in weight-gain during the course of a study, particular those sacrificed at the 48 hour timepoint, before any recovery could take place.

It is noteworthy that the choline & methionine deficiency study (D20) also shows a high MPLE score, despite being a dietary intervention. Choline is known to be necessary for secretion of VLDL (Li & Vance, 2008), being required for the production of phosphatidylcholine, which is in turn an essential component of lipoproteins. This suggests that the observed alterations in lipoprotein metabolism caused by exposure to a liver toxin are leading to hepatic steatosis analogous to that resulting from choline deficiency.

6.4.2 MPLE scores are not directly related to lipid clinical-chemistry

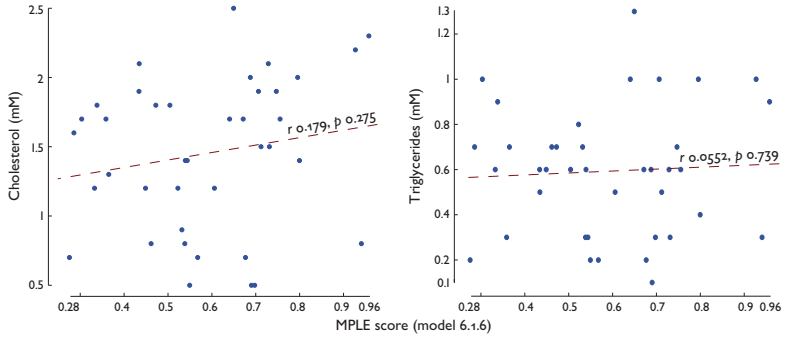

Figure 6.18: Scatterplots comparing cholesterol, and triglyceride levels in study D20, to the score on model 6.1.6. Despite modelling many lipid resonances, the metric does not appear well correlated to the clinical chemistry. A regression line has been plotted for each comparison, with the Pearson's *r* indicated.

In light of the observed changes, it is natural to question whether the observed profiles are simply indicating lipid levels, such as might be reported by a standard lipid measures. While quantification of serum lipids was not mandated in the COMET protocols, cholesterol and triglyceride levels were recored by clinical chemistry for study D20. The comparison between these measures and the score on the MPLE metric can be seen in Figure 6.18, and while this data is not taken from a study of xenobiotic dosing, the metric does not appear to correlate strongly to the traditional measures of serum lipids (cholesterol: r 0.179, p 0.275, triglycerides: r 0.0552, p 0.739).

It is unfortunate that direct measures of VLDL & LDL were not recorded, although work by Otvos *et al.* (2002) suggests that due to changes in the constituents of lipoproteins, NMR based measures of lipoprotein levels may provide a different estimate of concentrations compared to clinical chemistry. This is supported the work of Dyrby *et al.* (2005b) utilising parallel-factor analysis, that demonstrated the extraction of lipid profile information complementary to clinical chemistry from 2D diffusion-edited NMR spectroscopy of plasma.

However, parallel work by Petersen *et al.* (2005) shows a strong correlation between clinical-chemistry derived lipoprotein levels and those obtained from NMR with a well calibrated PLS model, although this work was conducted on diffusion-edited spectra that suppressed resonances from small mobile molecules, rather than the CPMG spectra examined here. Kristensen *et al.* (2010) raise the point that most lipoprotein test kits are targeted at human samples, while the lipoprotein profile of rats may differ markedly from this, skewing the results. They also report a generally good agreement between lipoprotein levels from NMR and clinical-chemistry. A more in-depth investigation of the relationship between the information gained from the metabolite profiles versus that from traditional measures of serum lipids would be a prime candidate for any follow-up experiments arising from this work.

6.4.3 Conclusion

We can see that the MPLE metric is at least comparable to ALT levels as an indication of exposure to liver toxins, particularly in cases where high sensitivity is desired. As expected, ALT performs poorly as time between sampling and exposure lengthens, the short half-life resulting from ALTs status as a marker of cellular damage and the swift rate of proteolysis putting it at a distinct disadvantage to the metabolic-profiling based approaches. As seen in ∫ 6.7, the MPLE-metrics demonstrate improved sensitivity and specificity across almost the entire classification range, when all timepoints are considered together. In situations were the precise moment of toxic-insult is unknown, these characteristics would give the MPLE metric an advantage over ALT levels in the diagnosis of exposure to a liver toxin.

Besides their differing temporal profiles, the two metrics provide contrasting and complimentary sets of information. While ALT is a good marker of necrosis, the information provided by the MPLE score gives a valuable insight into the status of hepatic lipid metabolism, that could be of use in future studies. The two metrics together provide a greater insight into the extent of exposure and response to a liver toxin than either alone.

Finally, the ability to successfully generate and validate a model generated out of these data further supports the legitimacy of the data preparation steps taken in the previous chapters. The undirected approach also substantiates the use of such methods for large, top-down explorations of such data.

Chapter 7

Discussion

7.1 Concluding Remarks

7.1.1 The importance of data standardisation

Analysis of large volumes of data, regardless of the form, is a complex and time consuming task. Biological data provides a particular challenge, where, despite the use of standardised protocols, any project too large to be run as a single batch will inevitably be subject to variance between batches, and these effects must be accounted for during further analysis.

Providing adequate records have been retained with respect to sample handling and origin, such batch variance is often easily diagnosed. However, if such data has not been retained, batch effects may be attributed some biological relevance, or obscure a factor of interest in the data. In situations where the record of sample handling is adequate, any unknown variance can be correlated to the treatment of the sample, and this effect accounted for. One good example of this is the signal-to-noise variation seen in § 6.2.1. Because this variation could be attributed to the NMR probe used to acquired the data, the effect could be disregarded, in the knowledge that the differences would be mitigated by normalisation.

These effects may impose one of two forms of error into the data. There may be a systematic effect, that varies according to some other parameter, such as time of sample acquisition, the instrument used for acquisition & *cetera*. Alternatively, the error introduced may be essentially random, simply decreasing the precision of these data. The effects causing this variation may arise at any point in the analytical procedure, and include the processes outlined below.

In Robosky *et al.* (2005) two groups of otherwise identical Sprague-Dawley rats were seen to display different metabolic profiles prior to dosing. The groups were traced to two different breeding rooms in the animal house the rats were sourced from, and potentially attributed to a shift in the populations of gut microflora, but no further. The divergent phenotype mainly demonstrated a modulated metabolism of aromatic molecules, particularly in the breakdown of products of chlorogenic acid, and these changes in metabolism have implications for the metabolism of xenobiotics.

Errors may be introduced during sample preparation by several factors; differences in reagent composition between batches, human error while following the experimental protocol, or even be attributed to differences in lab technique between scientists (Teahan *et al.*, 2006). While NMR is a highly reproducible technique (Keun *et al.*, 2002b) there is still a degree

of analytical drift between acquisitions (as outlined in Chapter 3 and § 6.2.1) and instruments, and while out of the scope of this thesis, this problem is compounded in most MS-based analysis, where changes in sample composition can radically alter both chromatography and ion-suppression and thus strongly perturb the acquired data in a manner that often requires complex post-processing before these data can be effectively analysed (Smith *et al.*, 2006; van der Kloet *et al.*, 2009). Further variance may be introduced to both of these kinds of data, as a result of processing raw data with differing parameters.

Finally, conducting statistical analysis in software packages that implement common algorithms as black-boxes may result in significant differences in the reported output, if the analyst is not aware of differing assumptions with respect to data-processing that may be coded into the software (Lindon *et al.*, 2005b).

Following the introduction and implementation of the database of COMET metadata in Chapter 2, this thesis has concentrated on reducing the affect of both types of interference on NMR based metabolite profiles. I have attempted to eliminate or mitigate as many of these sources of confounding variation as possible, in particular those that may result from the affect of subjective human judgment on a process, that with a suitable algorithm, might be accomplished computationally. While events occurring prior to data-acquisition in COMET are immutable, by completely reprocessing the spectra from raw data, including phasing, baseline correcting, calibrating and normalising the spectra with automated methods, the effect of subjectively estimated parameters on the processing of the data can be minimised.

In general, there is a degree of trade off between automated systems that may fail in some edge-cases, and human processing, which can often conduct adequate phasing or calibration in situations were automated systems fail, at the expense of a general decrease in reproducibility. In situations were the number of samples is not a limiting factor, I would consider the greater reproducibility of the automated systems is more important to the conduct of good analysis, than the increase in sample numbers obtained by manual processing of the borderline samples.

In Chapter 3, the drift along the chemical shift axis caused by fluctuations in sample temperature is detected and corrected for by means of an automated search for the α-anomeric glucose doublet. The automation of this process also has the advantage of removing the subjectivity in the calibration process, as exemplified by the results in ∫ 3.4. This technique has proved robust to spectra with relatively low signal-to-noise ratios and is already being applied to further studies, due to both its labour-saving nature and improved reproducibility.

The gross intensity changes that may result from differences in instrument sensitivity between acquisitions, or sample concentration are addressed in Chapter 4, with an assessment of the current normalisation methodology and the development of a novel approach to the problem. Apportionment-entropy normalisation is intended to restrict the variance in the data into the fewest variables. Compared to the current methodology, apportionment-entropy proved superior, both in simulations and when applied to limited sets of experimental data.

It is however, interesting to note that while apportionment-entropy did prove superior in small scale tests, when applied to the COMET dataset as a whole, median-fold-change normalisa-

tion proved to generate the strongest models. There are several potential explanations for this, but possibly most relevant, is that apportionment-entropy was not tested in situations where the signal-to-noise varied as markedly as it did across the entire COMET, and in light of the problems outlined in § 4.4, this might explain the superior performance of MFC normalisation.

The development of OFSQ to remove batch effects in the quantification of metabolites, generated a technique with a good ability to detect and remove orthogonal variation from data. Unfortunately, no metric for reliably judging the amount of variation that should be removed from the data could be generated, and this restricts the use of OFSQ to situations were the degree of interfering variation is known *a posteriori*. Clearly this is a problem in most experimental data, meaning OFSQ was not applied to the final analysis of the COMET serum data.

7.1.2 Mining the COMET data

With the exception of OFSQ, each of the data processing steps developed in this thesis succeeded in proving their worth in simulations or with small sets of data. However, the final arbiter of their quality must be whether it is possible to obtain biologically relevant results from profiles in a large project, such as COMET, prepared with them.

In Chapter 6, I have applied the techniques developed in the preceding chapters to an investigation into possible biomarkers of exposure to a liver toxin. By taking a broadly undirected approach, dealing only with the known dosing status of animals, rather than basing the groups on histopathology outcomes or clinical-chemistry, Chapter 6 demonstrates that is indeed possible to obtain biologically meaningful results while minimising subjective intervention. By targeting metabolic changes associated with the dosing of liver toxins I have generated a model of exposure that performs equivalently to ALT levels in differentiating these groups.

Indeed, while considered in terms of sensitivity and specificity the MPLE metric is comparable to ALT, the two appear to detect very different biological processes, and provide complimentary information as to the state of the liver. While ALT is reporting an affect powerful enough to compromise the integrity of the cell wall, essentially damage related to necrosis, the MPLE metric appears to model a perturbation in lipid metabolism of some form. Examination of the relevant model weights suggests a reduction in VLDL secretion from the liver, but the lack of correlation between the NMR profiles and clinical chemistry in ∫ 6.18 suggest that this change is more complex than that detected by basic lipid measures, leaving its precise nature unclear.

It is important to consider the possibility of non-toxic compounds having a similar affect on lipid metabolism in the liver. Many stimuli, ranging from dietary factors, to pharmacologically active xenobiotics such as the statins prescribed to lower cholesterol in humans (Ginsberg, 1998), may affect lipid metabolism in a manner that would not be expected to progress to any toxic pathology. Due to the focus on toxicity in COMET, it is not possible, to rule out these effects without further experiments.

By generating models on such a heterogeneous collection of liver toxins, it has been possible

to average-out the individual metabolic effects of each compound. This has allowed the MPLE model to focus on a more general indicator of hepatic stress, one that appears to indicate a modulation of overall liver function, rather than gross toxic insult.

It is worthwhile to note that the variable restriction strategies outlined in § 6.2.3 proved too restrictive in their filtering of the data, and resulted in models over-fitted to the training data. In any continuation of this work, it would be enlightening to investigate whether relaxation of the FDR, or an improved estimate of the variability in the population would improve matters. For instance, a bootstrap-based (Harrell Jr., 2001, Chapter 5) estimate of the p-value of the test groups, might indicate a bias in the training set.

The outcome of the variable restriction does reiterate the need for the comparison of models generated with differing parameters, and the use of independent validation data, that has been completely set aside from the modelling process. Without validation, it would be trivial to generate over-fitted models, and while assessing models by their biological plausibility will always be necessary, the comparison of model performance on validation data allows for a rapid and objective assessment of the quality of a family of models.

7.2 Wider Scope

There are several avenues that any continuation of the work in this thesis could follow.

In regards to the broader theme of removing subjectivity from the analytical process, while much manual intervention has been removed from the processing of the spectral data, there are still many areas in the statistical modelling of these data where human intervention could be mitigated or removed.

Or *outliers*, that may result from poor spectral-acquisition, or extreme biological variation.

Notably, the detection and removal of malformed spectra ←, despite being systemised as far as possible, still requires human intervention. A fully automated system of outlier detection is one of the primary stumbling blocks in the development of a fully automated 'one click' system that could rapidly explore the links between a large dataset and its metadata. While Ebbels *et al.* (2003) showed the potential of using outliers in metabolic space to exclude samples, this approach is only applicable to situations, like controls samples, were there is the expectation of some consistency between profiles. Here, the potential of apportionment-entropy in similarity matching or outlier detection merits further investigation. The brief analysis summarised in ∫ 4.8 gives a strong indication that this method could provide a simple method of summarising the similarity between spectral profiles, but the examination of this aspect of apportionment-entropy in Chapter 4 is too rudimentary to make any specific statements.

However, I feel further work should not only pursue apportionment-entropy based outlier detection, but also investigate the use of other spectral fitting techniques, such as line-shape fitting, to assess and quantify parameters such as line-width, or detect asymmetries resulting from poor phasing. An ideal mechanism of outlier detection should not only indicate a sample differs significantly from normal, but also how. For NMR this could include all the expected processing stages, such as 'is the spectrum badly phased?', 'is the baseline not flat?' essentially those

factors considered in § 6.2.1. Without a reliable estimate of the quality of these parameters, some manual checking will always be required, as the reason for an exclusion often provides useful information. For example, it might not be considered unusual to see 50 outliers in a collection of 1,000 samples, but if 49 of these were excluded due to failed shift calibration, or any other single cause, it would merit further investigation.

While model selection during the predictive modelling stage was driven by the area under the ROC curve, there are situations, as outlined in § 1.6, where if the ROC curves diverges from a simple parabola, two models may have a comparable AUC, while offering very different classification performance.

Beyond the continuing development of the data-processing methodology, there is a great deal of information still latent within the COMET serum data. There is obvious potential to generate global models of damage to other organs, in particular renal damage, as the kidney is the next most represented target organ in COMET. Other target organs in the project, such as the pancreas or testes may be too under represented to be amenable to this approach, while the class of physiological stressors is probably too diffuse in the range of biological effects it encompasses.

Any further analysis of these data, specifically that relating to the effect modelled here, would do well to concentrate on some of the compounds in COMET known to result in liver steatosis. In addition to the choline & methionine restriction study discussed in § 6.4, compounds such as ethionine, an *S*-ethyl analogue of methionine, which induces steatosis following acute dosing, have been examined using similar pattern-recognition techniques (Skordi *et al.*, 2007). While only dealing with urine and tissue samples in their study, Skordi *et al.* observed broader metabolic changes than those modelled here, and they hypothesise that some of these changes may be related to several secondary toxicities caused by ethionine. This supports the concept that by averaging out the individual side-effects across all the compounds in COMET, the liver metric developed here has focused specially on the metabolic changes related to steatosis, or a related dysfunction in lipid metabolism in the liver.

Beyond this, the scores on model 6.1.6 seen in ∫ 6.9 provide an interesting method of clustering the compounds, and it would be interesting to investigate any potential mechanistic or physiological similarities between compounds that score similarly. There are hints that this may be the case in the isothiocyanate compounds, that show a very similar distribution of scores on the model, and in the previously established conservation of disturbances in metabolic space by mechanism outlined in § 1.3.2.

If this approach were to prove fruitful, there is the potential combine it with models based on other target organs and to use the scores on several such models to estimate a compound's affects on each of the metabolic systems. Such an approach would have an advantage over CLOUDS as it would skip the time-consuming density-superstition process →, in favour of projecting the metabolite profiles through a small number of models, each defining a functional metabolic change.

CLOUDS requires a multidimensional overlap to be calculated in a global PC space for every class of toxicity to be tested.

The MPLE approach shares an interesting commonality with the concept of perturbations

in toxic pathways outlined in the National Research Council report described in § 1.2. Here the MPLE metric provides an indication of the activity of a key unit of hepatic function, combining the information derived from several resonances. Indeed the desire to expand biomarker use, particularly with respect to metabolite profiling, is currently one of the greatest challenges in toxicology.

Biomarkers are already supporting the development of pharmaceuticals, as outlined in § 1.3.2. However, Spear *et al.* (2001) have described the efficacy of many common pharmaceuticals as lying in the 25 – 80% range, and many see biomarkers arising from the various —omics fields providing a solution to this lack of efficacy (Mendrick & Schnackenberg, 2009). This improvement may be achieved on several fronts, for example; by being diagnostic of an early stage disease, where dosing could be more effective; indicating potential greater efficacy or susceptibility to toxic side-effects (the pharmacometabonomics approach) in an individual; or being prognostic of the long-term benefit of a treatment.

7.3 Conclusion

To conclude, in this thesis I have undertaken the curation, archiving, and standardisation of a large dataset of metabolite profiles. To validate the data standardisation techniques, I have then explored these data with the aim of extracting general markers of exposure to a liver toxin.

In the process of standardising the data I have developed novels methods of calibration and normalisation, in addition to exploring avenues related to orthogonal filtering of data. The analysis has been successful in detecting perturbations in hepatic liver metabolism that are strongly linked to the dosing of xenobiotics targeting the liver. These observations have potential for use as a diagnostic biomarker of hepatic lipid metabolism. I hope the work in this thesis adequately demonstrates the potential of an undirected approach to biomarker discovery in large datasets of metabolite profiles.

Appendix A

COMET Study Details

The table in this appendix contains details of the dosing regime and other pertinent information for each study in the COMET project. Compounds were administered as a single acute dose unless otherwise noted. The first study run by each company (X01) was run to a matched protocol to assess inter-site variability, however the complete COMET protocols were not finalised at this point, and so these studies differ from the bulk of the COMET studies in terms of data-collection. All Pfizer (FXX) studies were run in the mouse rather than the rat, with the exception of F01. The presented data were extracted from the COMET database (see Chapter 2) or directly from raw instrument files (probehead information). The *excluded* column indicates any studies omitted from the bulk serum analysis in Chapter 6, and the *comments* column contains my own observations where relevant.

Abbreviations:	I.P.	Intraperitoneal injection.
	P.O.	*Per os* (Oral gavage).
	I.V.	Intravenous injection.
	S.C.	Subcutaneous injection.
	Diet	Diet modification.
	Surgical	Surgical intervention.

Table A.1

Study	Compound	Dose Route	Target	Probehead	Excluded	Comments
D01	Hydrazine	P.O.	Liver	5 mm FI TXB	Yes	Incomplete clinical chemistry.
D02	Cisplatin	I.P.	Kidney	5 mm FI TXB	No	Bacterial contamination of samples, study repeated as D11.
D03	Puromycin	I.P.	Kidney	5 mm FI TXB	Yes	Bacterial contamination of samples, study repeated as D12.
D04	Surgical pilot	Surgical			Yes	Pilot study for D05 & D06, no samples sent to Imperial College.
D05	Partial hepatectomy	Surgical	Physiological stressor	5 mm FI TXB	No	
D06	Unilateral nephrectomy	Surgical	Physiological stressor	5 mm FI TXB	No	
D07	Gentamicin	I.P.	Kidney	5 mm FI TXB	No	Also run in the mouse as F11.
D08	Phenobarbital (chronic)	I.P.	Physiological stressor	5 mm BBI	No	Also run in the mouse as F23.
D09	Folic acid	I.P.	Kidney	5 mm BBI	No	
D10	Dexamethasone	I.P.	Pancreas	5 mm FI TXB	No	
D11	Cisplatin	I.P.	Kidney	5 mm FI TXB	No	Repeat of D02. Also run in the mouse as F14.
D12	Puromycin	I.P.	Kidney	5 mm FI TXB	No	Repeat of D03.
D13	Gadolinium chloride	I.V.	Liver	5 mm BBI	No	
D14	Lipopolysaccharide	I.V.	Liver	5 mm TXI	No	Also run in the mouse as F17.
D15	Probenecid	I.P.	Physiological stressor	5 mm TXI	No	Competitve inhibitor of anion transport.
D16	D-galactosamine	I.P.	Liver	5 mm BBI	No	Also run in the mouse as F07.
D17	N-phenylanthranilic acid (chronic)	Oral gavage	Kidney	5 mm BBI	No	
D18	D-limonene (chronic)	Oral gavage	Kidney	5 mm BBI	No	
D19	Pregnenolone-16α-carbonitrile (chronic)	Oral gavage	Physiological stressor	5 mm BBI	No	Induces liver P450s.
D20	Choline & methionine deficiency (chronic)	Diet	Physiological stressor	5 mm BBI	No	Low-dose group fed a choline deficient diet, high-dose fed a choline & methionine deficient diet. Also run in the mouse as F26.
F01	Hydrazine	P.O.	Liver	5 mm FI TXB	Yes	Incomplete clinical chemistry.
F02	Hydrazine	P.O.	Liver		Yes	Mouse counterpart to x01, serum data not acquired.
F03	Potassium dichromate	I.P.	Multiple organ		Yes	Mouse counterpart to R23, serum data not acquired.

Continued on next page

Study	Compound	Dose Route	Target	Probehead	Excluded	Comments
F04	Thioacetamide	I.P.	Multiple organ		Yes	Mouse counterpart to R07, serum data not acquired.
F05	Bromobenzene	P.O.	Liver		Yes	Mouse counterpart to L11, serum data not acquired.
F06	Tamoxifen	P.O.			Yes	Mouse study, serum data not acquired.
F07	D-Galactosamine	I.P.	Liver		Yes	Mouse counterpart to D16, serum data not acquired.
F08	Hexachlorobutadiene	P.O.	Kidney		Yes	Mouse counterpart to L09, serum data not acquired.
F09	Carbon tetrachloride	P.O.	Liver		Yes	Mouse counterpart to R03, serum data not acquired.
F10	Amphotericin B	I.P.	Multiple organ		Yes	Mouse counterpart to N25, serum data not acquired.
F11	Gentamicin	I.P.	Kidney		Yes	Mouse counterpart to D07, serum data not acquired.
F12	*p*-Aminophenol	I.P.	Kidney		Yes	Mouse counterpart to R13, serum data not acquired.
F13	2-Bromoethylamine	I.P.	Kidney		Yes	Mouse counterpart to R04, serum data not acquired.
F14	Cisplatin	I.P.	Kidney		Yes	Mouse counterpart to D11, serum data not acquired.
F15	Transplatin	I.P.	Non-toxic control		Yes	Mouse study, serum data not acquired.
F16	Allyl alcohol	I.P.	Liver		Yes	Mouse counterpart to N03, serum data not acquired.
F17	Lipopolysaccharide	I.P.	Liver		Yes	Mouse counterpart to D14, serum data not acquired.
F18	α-Naphthylisothiocyanate	P.O.	Liver		Yes	Mouse counterpart to N02, serum data not acquired.
F19	Acetaminophen	I.P.	Liver		Yes	Mouse counterpart to S07, serum data not acquired.
F20	Adriamycin	I.P.	Kidney		Yes	Mouse counterpart to L06, serum data not acquired.
F21	Methotrexate	I.P.	Physiological stressor		Yes	Mouse counterpart to S08, serum data not acquired.

Continued on next page

Study	Compound	Dose Route	Target	Probehead	Excluded	Comments
F22	Mercuric chloride	I.P.	Kidney		Yes	Mouse counterpart to L07, serum data not acquired.
F23	Phenobarbital (chronic)	I.P.	Liver		Yes	Mouse counterpart to D08, serum data not acquired.
F24	Food restriction (chronic)	Diet	Physiological stressor		Yes	Mouse counterpart to L12, serum data not acquired.
F25	Butylated hydroxytolulene	P.O.	Liver		Yes	Mouse counterpart to N04, serum data not acquired.
F26	Choline & methionine deficiency (chronic)	Diet	Physiological stressor		Yes	Mouse counterpart to D20, serum data not acquired.
L01	Hydrazine	P.O.	Liver	5 mm FI TXB	Yes	Incomplete clinical chemistry.
L02	Insulin	S.C.	Physiological stressor	5 mm FI TXB	No	
L03	Chlorethanamine	I.P.	Kidney	5 mm FI TXB	No	
L04	Streptozotocin	I.P.	Pancreas & Endocrine	5 mm FI TXB	No	
L05	Clofibrate	P.O.	Liver	5 mm FI TXB	No	
L06	Adriamycin	I.V.	Kidney	5 mm FI TXB	N	Also run in the mouse as F20.
L07	Mercuric chloride	I.P.	Kidney	5 mm FI TXB	No	Also run in the mouse as F22.
L08	Diethylhexylphthalate	P.O.	Liver	5 mm FI TXB	No	
L09	Hexachlorobutadiene	P.O.	Kidney	5 mm FI TXB	No	Also run in the mouse as F08.
L10	Sodium valproate	P.O.	Liver	5 mm FI TXB	No	
L11	Bromobenzene	P.O.	Liver	5 mm FI TXB	No	Also run in the mouse as F05.
L12	Food restriction (chronic)	Diet	Physiological stressor	5 mm FI TXB	No	Also run in the mouse as F24.
L13	Water deprivation (chronic)	Diet	Physiological stressor	5 mm FI TXB	No	
L14	Vancomycin hydrochloride	I.V.	Kidney	5 mm FI TXB	No	
L15	Methapyrilene	P.O.	Liver	5 mm FI TXB	No	
L16	Maleic acid	P.O.	Kidney	5 mm FI TXB	No	
L17	1-Fluoropentane	I.P.	Liver	5 mm BBI	No	
L18	Aurothiomalate	I.P.	Kidney	5 mm BBI	No	
L19	Ifosfamide	P.O.	Bladder	5 mm FI TXB	No	
L20	Azathioprine	P.O.	Liver	5 mm FI TXB	No	
L21	Lithocholic acid	P.O.	Liver	5 mm BBI	No	
L22	Retinyl palmitate	P.O.	Liver	5 mm BBI	No	
L23	2',4',6'-Trihydroxyacetophenone	I.P.	Multiple organ	5 mm TXI	No	
L24	4-Amino-2,6-dichlorophenol	I.P.	Liver	5 mm TXI	No	

Continued on next page

Study	Compound	Dose Route	Target	Probehead	Excluded	Comments
L25	4-Pentenoic acid	P.O.	Physiological stressor	5 mm TXI	No	
L26	Cyproterone acetate	I.P.	Liver	5 mm TXI	No	
N01	Hydrazine	P.O.	Liver	5 mm FI TXB	Yes	Incomplete clinical chemistry.
N02	α-Naphthylisothiocyanate	P.O.	Liver	5 mm FI TXB	Yes	Also run in the mouse as F18
N03	Allyl alcohol	P.O.	Liver	5 mm FI TXB	No	Also run in the mouse as F16.
N04	Butylated hydroxytoluene	P.O.	Liver	5 mm FI TXB	No	Also run in the mouse as F25.
N05	Chlorpromazine	I.P.	Liver	5 mm FI TXB	No	
N06	Ethionine	P.O.	Liver	5 mm FI TXB	No	
N07	Chloroform	P.O.	Multiple organ	5 mm FI TXB	No	
N08	Dimethylformamide	I.P.	Liver	5 mm FI TXB	No	
N09	N-methylformamide	I.P.	Liver	5 mm FI TXB	No	
N10	Ferrous sulphate	P.O.	Liver	5 mm FI TXB	Yes	
N11	Allyl formate	I.P.	Liver	5 mm FI TXB	No	
N12	Trichlorethylene	I.P.	Liver	5 mm FI TXB	No	
N13	Monocrotaline	P.O.	Liver	5 mm FI TXB	No	
N14	Dimethylnitrosamine	P.O.	Liver	5 mm BBI	No	
N15	Lead acetate	I.P.	Liver	5 mm TXI	No	
N16	Cadmium chloride	Oral gavage	Testes	5 mm FI TXB	No	
N17	2,4-Dinitrophenol	Oral gavage	Physiological stressor	5 mm TXI	No	
N18	Acivicin	I.P.	Physiological stressor	5 mm TXI	No	
N19	Buthionine sulphoxime	I.P.	Liver	5 mm TXI	No	
N20	1,1-Dichloroethylene & maleic acid	P.O.	Physiological stressor	5 mm TXI	No	
N21	Aflatoxin	P.O.	Liver	5 mm BBI	No	
N22	Rosiglitazone (chronic)	P.O.	Physiological stressor	5 mm BBI	No	Chronic dosing counterpart to N26.
N23	Caerulin	I.P.	Pancreas	5 mm TXI	No	
N24	Azaserine	I.P.	Multiple organ	5 mm BBI	No	
N25	Amphotericin B	I.P.	Multiple organ	5 mm FI TXB	No	Also run in the mouse as F10.
N26	Rosiglitazone	P.O.	Physiological stressor	5 mm FI TXB	No	Liver effect. Acute dosing counterpart to N22.
R01	Hydrazine	P.O.	Liver	5 mm FI TXB	Yes	Incomplete clinical chemistry.
R02	Hydrazine	P.O.	Liver	5 mm FI TXB	Yes	Repeat of R01. Incomplete clinical chemistry.
R03	Carbon tetrachloride	P.O.	Liver	5 mm FI TXB	No	Also run in the mouse as F09.
R04	2-Bromoethylamine	I.P.	Kidney	5 mm BBI	No	Also run in the mouse as F13.

Continued on next page

Study	Compound	Dose Route	Target	Probehead	Excluded	Comments
R05	Cephaloridine	I.P.	Kidney	5 mm FI TXB	No	
R06	WY14,643 ([4-chloro-6-(2,3-xylidino)-2-pyrimidinylthio]acetic acid)	Oral gavage	Liver	5 mm FI TXB	No	Acute dosing counterpart to R25.
R07	Thioacetamide	I.P.	Multiple organ	5 mm BBI	Yes	Also run in the mouse as F04.
R08	Ethylene glycol	Oral gavage	Kidney	5 mm BBI	No	
R09	Acetazolamide	Oral gavage	Physiological stressor	5 mm BBI	No	
R10	Dichlorobenzene	Oral gavage	Liver	5 mm BBI	No	
R11	Furosemide	Oral gavage	Physiological stressor	5 mm BBI	No	
R12	Hydrazine	P.O.	Liver	5 mm TXI	Yes	Hans Wistar rats (strain: HanBrl:WIST(SPF)).
R13	p-Aminophenol	I.P.	Kidney	5 mm BBI	No	Also run in the mouse as F12.
R14	Cyclosporin	Oral gavage	Multiple organ	5 mm BBI	No	
R15	Carboplatin	I.V.	Physiological stressor	5 mm TXI	No	
R16	2-Bromophenol	I.P.	Kidney	5 mm FI TXB	No	
R17	Carbendazim	Oral gavage	Testes	5 mm BBI	No	
R18	Ethane-(dimethane sulfonate)	I.P.	Testes	5 mm TXI	No	
R19	1,3-Dinitrobenzene	Oral gavage	Testes	5 mm BBI	No	
R20	Phalloidin (chronic)	I.P.	Liver	5 mm TXI	No	
R21	Cadmium chloride	I.P.	Testes	5 mm BBI	No	
R22	Methoxyacetic Acid	P.O.	Testes	5 mm BBI	No	
R23	Potassium dichromate	I.P.	Multiple organ	5 mm BBI	No	Also run in the mouse as F03.
R24	Di-n-pentyl-phthalate	P.O.	Testes	5 mm BBI	No	
R25	WY14,643 ([4-chloro-6-(2,3-xylidino)-2-pyrimidinylthio]acetic acid) (chronic)	P.O.	Other	5 mm TXI	No	Chronic dosing counterpart to R06.
S01	Hydrazine	P.O.	Liver	5 mm FI TXB	Yes	Incomplete clinical chemistry.
S02	N,N'-dimethyl-4,4'-bipyridinium dichloride	I.P.	Other	5 mm FI TXB	Yes	
S03	Ketoconazole	P.O.	Liver	5 mm FI TXB	Yes	
S04	Ammonium chloride	P.O.	Acidosis	5 mm FI TXB	No	All animals including controls show elevated ALT.
S05	Microcystin-LR	I.P.	Liver	5 mm FI TXB	No	
S06	Mitomycin-C	I.P.	Multiple organ	5 mm TXI	No	

Continued on next page

Study	Compound	Dose Route	Target	Probehead	Excluded	Comments
S07	Acetaminophen	P.O.	Liver	5 mm BBI	No	Acute dosing counterpart to S09. Also run in the mouse as F19.
S08	Methotrexate	P.O.	Physiological stressor	5 mm TXI	No	Also run in the mouse as F21.
S09	Acetaminophen (chronic)	P.O.	Liver	5 mm TXI	No	Chronic dosing counterpart to S07.
S10	1,1-Dichloroethylene	P.O.	Liver	5 mm TXI	No	
S11	Sodium bicarbonate	P.O.	Alkalosis	5 mm TXI	No	
S12	Rotenone	P.O.	Liver	5 mm BBI	No	
S13	Indomethacin	P.O.	Kidney	5 mm FI TXB	No	
S14	1-Cyano-2-hydroxy-3-butene	I.P.	Pancreas	5 mm TXI	No	
S15	3,5-Dichloroaniline hydrochloride	I.P.	Kidney	5 mm TXI	No	
S16	Atractyloside	I.P.	Kidney	5 mm TXI	No	
S17	S-(1,2-dichlorovinyl)-cysteine	I.P.	Multiple organ	5 mm BBI	No	
S18	Methylene dianiline	P.O.	Liver	5 mm TXI	No	
S19	Dichlorophenyl succinimide	P.O.	Kidney	5 mm BBI	No	
S20	L-arginine	I.P.	Pancreas	5 mm TXI	No	
S21	Phenyl isothiocyanate	I.P.	Liver	5 mm TXI	No	
S22	1,2,3,4,5,6-Hexachlorocyclohexane	I.P.	Liver	5 mm TXI	No	
S23	Phenyl diisothiocyanate	I.P.	Liver	5 mm TXI	No	

Appendix B

Experimental Methods

B.1 COMET metadata collection

Metadata collected during each COMET study included a full set of clinical chemistry data and histopathology reports →. Metadata were collected in-house at each of the participating companies, or by their designated experimental house, using recognised current gold-standard methodology, and submitted to Imperial College in formatted Excel templates.

See Chapter 2 for a full list of metadata acquired.

B.2 Acquisition of COMET spectra

Urine samples were prepared for NMR into covered 96-well plates using a Bruker SampleTrack (Bruker Biospin, Rheinstetten, Germany) system running a Gilson 215 Liquid Handler. 200 µL of rat urine → was mixed with 200 µL of buffer solution (0.04 M solution of NaH_2PO_4 in H_2O mixed 4:1 in D_2O with TSP is added at 1 mM and sodium azide at 3 mM). The first and last sample of each plate were blanks, with H_2O substituted in place of urine, this allowed a check for carry-over between samples, and flushed the probe at the end of the run. Prior to NMR, each well plate was centrifuged at 4,000 rpm for 5 minutes to remove insoluble material.

For the mouse samples, 400 µL of urine plus 200 µL of H_2O due to lower sample volumes.

NMR spectra of urine samples were acquired on a 600.13 MHz Bruker BioSpin Avance 600 Ultrashield spectrometer using a 5 mm TXI flow-injection probe. Sample temperature was set at 300 K. Urine spectra were acquired with the standard Bruker NOESYPR1D → presaturated pulse-sequence of the form $RD-90°-t_1-90°-t_m-90°-AQ$, with an acquisition time of 1.36 s and a t_1 of 3 µs. Suppression of the water signal was achieved using radiofrequency presaturation at the water resonance (δ_H = 4.701) during RD (2 s) and t_m (100 ms). During the acquisition period the FID was recorded in the time domain into 32k data-points with a spectral width of 12,019.2 Hz. For each experiment the sum of 64 transients were recorded following 4 dummy scans. Each FID was multiplied by an exponential line-broadening function of 1 Hz prior to Fourier-transformation.

Referred to as NOESYPR1D spectra throughout.

Serum was prepared by the addition of 200 µL of plasma to 400 µL of saline solution (0.9% NaCl in H_2O mixed 9:1 with D_2O). Spectra were initially acquired from 96-well plates with the same flow-injection probe as the urine samples, but technical difficulties caused by coagulation of proteins in the flow-system required the majority of samples to be acquired in 5 mm tubes with a 5 mm TXI or 5 mm BBI probe. NOESYPR1D spectra were acquired identically to the

urine spectra, with the exception that 128 transients were collected per sample. Additionally serum spectra were acquired with a Carr-Purcell-Meiboom-Gill pulse-sequence ← as follows: $RD-90°-(\tau/2-180°-\tau/2)_n-AQ$, the RD was 2 s, with an acquisition time of 1.36 s, total echo time was 64 ms ($n = 80$, $\tau = 400$ µs). During the acquisition period the FID was recorded in the time domain into 32k data-points with a spectral width of 12,019.2 Hz. For each experiment the sum of 128 transients were recorded following 16 dummy scans. Each FID was multiplied by an exponential line-broadening function of 1 Hz prior to Fourier-transformation. This CPMG pulse-sequence had a comparatively short echo time, suppressing the very largest macromolecules, while retaining resonances from smaller molecules such as fatty acids.

Referred to as CPMG spectra throughout.

A small subset of the serum samples were also acquired with a diffusion edited pulse-sequence (Wu *et al.*, 1995)←, that resulted in a suppression of resonances associated with smaller molecules, focusing on macromolecules, or as J-resolved 2D spectra, where these spectra were considered appropriate.

Referred to as LEDPBGS1SPR spectra.

B.3 Human blood-serum spectrum acquisition

As part of a separate study, blood serum was acquired from healthy adults into vacutainers, allowed to clot for 30 mins, before centrifugation and removal of 250 µL aliquots of the supernatant for storage at -40 °C. Serum was prepared for NMR by addition of 200 µL serum to 400 µL saline (0.9% NaCl in 10% D_2O: 90% water). Samples were then centrifuged at 12,000 g for 5 min, 550 µL of the resulting supernatant were pipetted into standard 5 mm NMR tubes.

1D ^{1}H-NMR spectra were acquired with a 600.13 MHz NMR system (Bruker BioSpin, Rheinstetten, Germany), at 300 K. One-dimensional ^{1}H spectra were acquired using the pulse sequence $RD-90°-t_1-90°-t_m-90°-AQ$, with an acquisition time of 1.36 s and a t_1 of 3 µs. Suppression of the water signal was achieved using radiofrequency presaturation at the water resonance ($\delta_H = 4.701$) during RD (1 s) and t_m (100 ms). During the acquisition period the free induction decay was recorded in the time domain into 64k data-points with a spectral width of $\delta_H = 20.0173$. For each experiment the sum of 128 transients were recorded following 8 dummy scans. Each FID was multiplied by an exponential linebroadening function of 1 Hz prior to Fourier-transformation. Spectra were phased and baseline corrected with the Xwin-NMR (Bruker Biospin 2004) software. Data were imported into MATLAB release 2007b between $\delta_H = 13$ & -3 at full resolution

B.4 Human urine spectrum acquisition

One hundred and eighty adult participants were recruited to the study described in Thomas *et al.* (2009), and 178 samples were analysed by NMR. Briefly, each individual was asked to provide one spot morning urine sample; each sample was frozen on the day of collection. Ethical approval for the study was granted by Bristol South and Central Ethics Committee, and all participants provided their written informed consent before data collection. Each volunteer

completed a lifestyle questionnaire that recovered information on a number of parameters including smoking status (current, past, never), age and gender.

Samples were thawed on ice and prepared in a single batch. For each sample, an aliquot (200 µL) was mixed with buffer (400 µL, 0.2 M sodium phosphate, pH 7.4) containing D_2O (8:1 v/v) and TSP, 0.3 mM and centrifuged at 16,000 g for 10 minutes to remove insoluble material.

High-resolution ^{1}H-NMR spectroscopy was conducted using a 600.13 MHz Bruker Avance 600 spectrometer fitted with a flow-injection mode probehead (5 mm FI TXB ^{1}H-^{13}C/^{15}N-^{2}H Z-GRD H8432/K0201 Z8432/0201, Bruker) at a field strength of 14.1 T. The probe temperature was set to 300 K. Samples were delivered to the probehead in 96-well plates using an automated flow injection system. Following introduction to the probe, samples were left to equilibrate (3 min) prior to gradient shimming using the ^{1}H channel to ensure good magnetic field homogeneity.

One-dimensional NMR spectra were acquired using a standard pulse sequence using excitation sculpting with gradients to suppress the water resonance (Hwang & Shaka, 1995). Each spectrum was acquired into 64k data-points over a spectral width of δ_H = 20 as the sum of 64 transients, recorded following 8 dummy scans. The acquisition time was 2.73 s, giving a native FID resolution of 0.183 Hz. The relaxation delay was set at 2 s. The total acquisition time was 6 min per sample. An apodisation function equivalent to a line-broadening of 0.3 Hz was applied to each FID prior to Fourier-transformation.

Appendix C

Database Implementation

C.1 Database implementation

The COMET database was implemented in MySQL v4.0.24 (Sun Co., formerly MySQL), running on Xserve G5 (Apple Inc.) hardware running Mac OS X Server 10.3.9. SQL code was generated with the CocoaMySQL v0.7b4 application.

C.2 MySQL data types

Attributes in MySQL must be assigned a specific data-type that governs the type of information capable of being stored within them. Numeric information may be stored in: an **`int(n)`**, for storing an integers of up to n digits; or a **`float(n, m)`**, for floating point numbers with n digits before the decimal-point and m after.

Text data may be stored; as **`char(n)`** capable that stores exactly n ASCII characters; a **`varchar(n)`**, capable of storing up to n characters; or as a **`text`** attribute, that may store an undefined amount of text. The **`enum(x, y, z)`** data-type is a special case that will only store one of the values entered when it is defined, in this case one of 'x','y' or 'z'.

Careful consideration must be paid to the data-types chosen for each attribute. While it may seem more flexible and simpler to define every attribute as the largest variant of its type, every text attribute as **`text`**, and every numerical value as the widest possible **`float()`**, doing so will result in a database that is massively wasteful of storage space and slow to perform most of the common database operations such as searching and table matching.

The 'allow null' property of each attribute controls whether SQL will allow this attribute to be blank (or *null*) for a record, and the default value governs what value will be given to an attribute when no value is specified on creation. Typically the default value is set to either a common value, or a custom value that is not expected to be found in the data-set which can then be used to diagnose any errors in data import. Attributes set to *increment* are numbered consecutively as they are created, each giving each record a unique ID based on the order of its creation.

C.3 Core table definitions

The following tables outline the attributes and data-types defined for each of the tables in the COMET database, along with relevant comments, refer back to § 2.3.2 for in-depth discussion of the table design. Attributers highlighted in **bold** are keys to other tables, *i.e.* they define the relations in the database. These attributes are named with the format *tablename*_ref, where *tablename* is the table the attribute is a key to. Relevant units for physical measures are recorded in the *Coments*.

Table C.1: Study table definition: 146 records

Attribute	Type	Allow Null	Default	Comments
study_id	char(3)	No	'err'	Key to table, indexed.
compound	text	No		Compound dosed / kind of intervention.
dose_route	text	No		Method of dosing / experimental implementation.
vehical	text	No		Dosing vehicle (misspelled in the database implementation).
volume_administered	float(4,2)	Yes	Null	Vehicle volume administered (**ml/kg**)
no_animals	int(4)	No	30	No of experimental animals involved.
company_ref	int(11)	No	1	Consortium member conducting the study.
internal_study_no	varchar(15)	Yes	Null	Companies internal reference for the study
study_director_ref	int(11)	No	1	Company representative with responsibility for the *in vivo* aspect of the study.
study_contact_ref	int(11)	No	1	Person to contact with questions regarding the *in vivo* component of the study.
suplier_id	int(11)	Yes	0	Supplier of experimental animals
species	varchar(5)	No	'Rat'	Model species
strain	text	No		Model strain
comments	text	Yes	Null	Any comments from the study director regarding the progress of the study.
pr_comments	text	Yes	Null	Analysts comments following PR.
complete	enum('Y','N')	No	N	Indicates whether the study is complete.

Table C.2: Animal table definition: 4,169 records

Attribute	Type	Allow Null	Default	Comments
animal_id	int(11)	No	*increment*	Table Key. Unique ID number for each animal.
study_ref	char(3)	No	'err'	Key to study table. Reference to to the study the animal took part in.
study_animal_no	int(3)	No	1	Animal number within the study.
predose_condition	text	Yes	Null	Any comments on the predose condition from the study director.
age	varchar(10)	No	'err'	Age (**weeks**) at the start of the study.
sex	enum('M', 'F')	No	M	Sex of the animal.
adaption_time	float(4,2)	Yes	Null	Time (**days**) between delivery and the start of the study.
sacrafice_group	char(1)	No	1	Time of sacrifice
dose_group	int(1)	No	1	Controls, or level of dose
dose_level	float(4,2)	No	9999.00	Dose level (**mg/kg**).
predose_weight	float(4,2)	Yes	Null	(**g**)
sacrafice_weight	float(4,2)	Yes	Null	(**g**)
brain_weight	float(4,2)	Yes	Null	(**g**)
comments	text	Yes	Null	Any observations specific to this animal.

Table C.3: Histopathology table definition: 10,037 records

Attribute	Type	Allow Null	Default	Comments
histopathology_id	int(11)	No	*increment*	
animal_ref	int(11)	No	0	
tissue	varchar(20)	No	'Error'	
weight	float(4,2)	Yes	*Null*	(g)
morphological_diagnosis	text	Yes	*Null*	
primary_pathological_process	text	Yes	*Null*	
secondary_pathological_process	text	Yes	*Null*	
localisation	text	Yes	*Null*	
severity	text	Yes	*Null*	
distribution	text	Yes	*Null*	
other_modifiers	text	Yes	*Null*	
comments	text	Yes	*Null*	

Table C.4: PR project table definition: 170 records

Attribute	Type	Allow Null	Default	Comments
project_id	int(11)	No	*increment*	
iteration	int(11)	No	1	Number of previous projects
study_ref	varchar(4)	No	'n/a'	
regions_excluded	text	Yes	*Null*	Areas of spectra excluded from PR
regions_merged	text	Yes	*Null*	Areas of spectra merged during spectral reduction
reduced_data	text	Yes	*Null*	Reduced resolution data
unmasked_data	text	Yes	*Null*	
coments	text	Yes	*Null*	

Table C.5: PR model table definition: 774 records

Attribute	Type	Allow Null	Default	Comments
model_id	int(11)	No	*increment*	
project_ref	int(11)	No	1	
model_name	text	Yes	*Null*	
important	enum('Y', 'N')	No	N	Flags a notable model
pr_type	varchar(10)	No	'PCA'	Type of model
scaling	varchar(20)	No	'Mean Centre'	Type of data-scaling
sample_exclusions	text	Yes	*Null*	Samples excluded from this model
exclusion_coments	text	Yes	*Null*	Reasons for exclusion
no_components	int(3)	No	999	
r2x_cumalative	float(4,2)	Yes	*Null*	
r2y_cumalative	float(4,2)	Yes	*Null*	
q2_cumalative	float(4,2)	Yes	*Null*	
comments	text	Yes	*Null*	

Table C.6: Urine sample table definition, with default clinical chemistry measures: 30,930 records

Attribute	Type	Allow Null	Default	Comments
sample_id	int(11)	No	*increment*	

Continued on next page

Attribute	Type	Allow Null	Default	Comments
title	varchar(13)	No	'default'	
animal_ref	int(11)	No	1	
timepoint	int(11)	No	999	
present	enum('N', 'Y')	No	N	
quantity_collected	float(4,2)	Yes	Null	(**ml**)
osmolarity	float(4,2)	Yes	Null	(**mOsm/l**)
ph	float(4,2)	Yes	Null	
protein	float(4,2)	Yes	Null	(**g/l**)
glucose	float(4,2)	Yes	Null	(**mM**)
bar_code	varchar(7)	Yes	Null	
reduced_spectra	text	Yes	Null	
comments	text	Yes	Null	

Table C.7: Serum sample table definition, with default clinical chemistry measures: 7,128 records

Attribute	Type	Allow Null	Default	Comments
sample_id	int(11)	No	*increment*	
title	varchar(13)	No	'default'	
animal_ref	int(11)	No	1	
timepoint	int(11)	No	999	
present	enum('N', 'Y')	No	N	
creatinine	float(4,2)	Yes	Null	(**µM**)
urea	float(4,2)	Yes	Null	Serum urea nitrogen (**µM**)
alt	float(4,2)	Yes	Null	alanine-aminotransferase (**IU/I**)
ast	float(4,2)	Yes	Null	Aspartate-aminotransferase (**IU/I**)
alp	float(4,2)	Yes	Null	Alkaline-phosphatase (**IU/I**)
ggt	float(4,2)	Yes	Null	Gamma-glutamyltransferase (**IU/I**)
glucose	float(4,2)	Yes	Null	(**mM**)
sodium	float(4,2)	Yes	Null	(**mM**)
potassium	float(4,2)	Yes	Null	(**mM**)
calcium	float(4,2)	Yes	Null	(**mM**)
phosphorus	float(4,2)	Yes	Null	(**mM**)
albumin	float(4,2)	Yes	Null	(**g/l**)
total_protein	float(4,2)	Yes	Null	(**g/l**)
bilirubin	float(4,2)	Yes	Null	(**µM**)
bar_code	varchar(7)	Yes	Null	
path_to_spectra	text	Yes	Null	
comments	text	Yes	Null	

Table C.8: Tissue sample table definition: 3,219 records

Attribute	Type	Allow Null	Default	Comments
sample_id	int(11)	No	*increment*	
title	varchar(13)	No	'default'	
animal_ref	int(11)	No	1	
timepoint	int(11)	No	999	
type	text	No		
present	enum('N', 'Y')	No	N	
bar_code	varchar(7)	Yes	Null	
path_to_spectra	text	Yes	Null	

Continued on next page

Attribute	Type	Allow Null	Default	Comments
comments	text	Yes	Null	

Table C.9: Outliers table, a record of all the NMR spectra excluded from pattern recognition and the reason for this: 1,078 records

Attribute	Type	Allow Null	Default	Comments
outlier_id	int(11)	No	*increment*	
title	varchar(13)	No	'X01r01h+999'	Title text of excluded sample
class	enum('NMR', 'missing', 'other')	No	missing	Class of exclusion
comments	text	Yes	Null	Reason for exclusion

C.4 Supplementary table definitions

In addition to the core database table modelling the COMET data-set, several support tables generated. These tables are intended to integrate miscellaneous information into the database, such as: company names and addresses; contact details for participants; and affiliations.

Table C.10: Company table definition: 10 records

Attribute	Type	Allow Null	Default	Comments
company_id	int(11)	No	*increment*	
name	text	No		
study_code	char(1)	Yes	Null	One letter code that identifies COMET participants studies.
primary_contact_ref	int(11)	Yes	Null	Key to the Person table.

Table C.11: Person table definition: 23 records

Attribute	Type	Allow Null	Default	Comments
person_id	int(11)	No	*increment*	
name	text	Yes	Null	
company_ref	int(11)	No	1	Key to company table. Company this individual is associated with.
phone_no	text	Yes	Null	Contact phone number, with formatting (To allow for international numbers).
email	text	No		email address
address	text	Yes	Null	Postal address.

C.5 Queries

This section lists SQL queries used to extract data from the COMET database.

To select serum clinical chemistry together with dosing information:

```
SELECT * FROM animal JOIN serum_sample WHERE animal.animal_id =
    serum_sample.animal_ref
```

To extract histopathology details for select studies (here R03):

```
SELECT animal.study_ref, animal.study_animal_no, animal.sacrafice_group,
    animal.dose_group, serum_sample.alt, histopathology.* FROM animal JOIN
    histopathology ON animal.animal_id = histopathology.animal_ref JOIN
    serum_sample ON animal.animal_id = serum_sample.animal_ref WHERE
    (animal.study_ref = 'N21' OR animal.study_ref = 'S07' OR animal.study_ref =
    'R03') AND histopathology.tissue = 'liver'
```

Extraction of animals displaying liver necrosis from database:

```
SELECT animal.study_ref, animal.study_animal_no FROM histopathology INNER JOIN
    animal ON animal.animal_id = histopathology.animal_ref JOIN study ON
    animal.study_ref = study.study_id WHERE histopathology.tissue = 'Liver' AND
    histopathology.morphological_diagnosis LIKE '%Necro%' AND study.species =
    'Rat'
```

Appendix D

Data Tables

This appendix lists data-tables referred to from the body of the text.

Table D.1: Sets of selected CPMG variables, following median-fold-change normalisation. Variables common between both restricted variable sets are in bold.

#	Variables	Variable Names	Comments
CPMG-MFC 1	35	**7.827**-7.817, 7.412-**7.404**, 7.313, 7.250-7.232, **7.161**, **5.217**, 3.363, 3.273 – 3.264, **3.200** – **3.173** – **3.155**, **3.002**, **2.615** – 2.607, 2.535, **2.408** – 2.391, 2.373, 2.076 - 2.067, **1.823** – **1.769** – 1.770, 1.446, 1.014	20% FDR after PLS regression to ALT.
CPMG-MFC 2	249	8.736, 8.518, 8.458 – 8.448, **7.827** – 7.818, 7.728 – 7.719, 7.430 – **7.404**, 7.377 – 7.350, 7.332 – 7.305, 7.206, 7.18 – **7.161**, 6.96 – 6.945, 6.882 – 6.873, 6.351 – 6.342, 6.045, 5.352 – **5.271**, 5.235 – 5.190, 4.280 – 4.253, 4.208 – 4.181, 4.082 – 4.046, 3.947 – 3.920, 3.902, 3.884, 3.848, 3.830 – 3.821, 3.731 – 3.713, 3.677 – 3.614, 3.578, 3.533 – 3.524, 3.506 – 3.452, 3.425 – 3.398, 3.335 – 3.326, 3.245 – 3.236, 3.218, **3.200**, **3.173**, **3.155**, 3.137, 3.092 – 2.975, 2.849 – 2.723, 2.696, 2.669, 2.642 – **2.615**, 2.498 – 2.471, 2.435 – **2.408**, 2.300, 2.255 – 2.165, 2.156 – **1.823** – **1.769**, 1.652 – 1.535, 1.472 – 1.454, 1.391 – 1.328, 1.310 – 1.247, 1.220 – 1.211, 1.103 – 1.085, 1.013, 0.950 – 0.941, 0.905 – 0.869, 0.698 – 0.689, 0.662, 0.509, 0.437, 0.419, -0.120, -0.589	20% FDR by direct t-test between controls and high-dose liver toxins. Note two variables selected upfield of $\delta_H = 0$.
CPMG-MFC all	1,100	All	No filtering applied.

Table D.2: Sets of selected CPMG variables, following normalisation to apportionment-entropy. Variables common between both restricted variable sets are in bold.

#	Variables	Variable Names	Comments
CPMG-ent 1	21	7.827, 5.217, 3.362, 3.191, 3.173, 3.164, 3.155, **2.615**, 2.606, 2.408, 2.399, 2.390, 2.381, 2.075, 1.823, **1.814**, 1.805, 1.796, 1.787, 1.670, 1.661, 1.652, 1.643, 1.445	20% FDR after PLS regression to ALT.

Continued on next page

#	Variables	Variable Names	Comments
CPMG-ent 2	140	8.736, 7.818, 7.728, 7.422 – 7.404, 7.368 – 7.359, 7.323 – 7.314, 7.188 – 7.161, 6.882 – 6.873, 5.352 – 5.271, 5.199, 4.289 – 4.244, 4.199 – 4.181, 4.082 – 4.046, 3.947 – 3.929, 3.902, 3.884, 3.848, 3.830, 3.722, 3.704, 3.641, 3.479, 3.461, 3.398, 3.056 – 3.002, 2.849 – 2.777, 2.750 – 2.732, 2.696, 2.669, 2.633 – **2.615**, 2.498, 2.489, 2.435 – 2.417, 2.255 – 2.201, 2.030 – 1.985, 1.868 – **1.814**, 1.733 – 1.697, 1.607 – 1.535, 1.472 – 1.454, 1.382 – 1.337, 1.310 – 1.247, 0.896 – 0.869, -0.121, -0.589	20% FDR by direct t-test between controls and high-dose liver toxins.
CPMG-ent all	1,100	All	No filtering applied.

Table D.3: Sets of selected NOESYPR1D variables. All spectra are MFC normalised

#	Variables	Variable Names	Comments
NOESYPR1D-MFC 1	338	9.636, 9.348, 9.330, 9.321, 9.285 – 9.267, 9.240 – 9.222, 9.195, 9.177, 9.069, 9.033, 8.943 – 8.844, 8.826 – 8.808, 8.781 – 8.763, 8.745, 8.331 – 8.277, 8.250 – 8.223, 8.205, 8.187 – 8.151, 8.106, 8.052, 7.989 – 7.962, 7.926, 7.917, 7.899, 7.773, 7.737, 7.629 – 7.548, 7.467 – 7.440, 6.801 – 6.720, 6.684 – 6.621, 6.468 – 6.414, 5.901, 5.874, 5.856 – 5.811, 5.775 – 5.721, 5.703, 5.676, 5.532, 5.496, 5.442, 5.226, 4.130 – 4.094, 3.914, 3.896, 3.860, 3.842, 3.779, 3.761, 3.707, 3.671 3.662, 3.590 – 3.572, 3.536, 3.518, 3.500, 3.482, 3.446, 3.428, 3.365 – 3.338, 3.275 – 3.257, 3.239, 3.221, 3.203 – 3.185, 3.122 – 3.095, 3.077, 3.059 – 3.005, 2.897, 2.870 2.861, 2.717 – 2.699, 2.618, 2.546, 2.519, 2.501 2.492, 2.429 – 2.303, 2.186 – 2.168, 2.096 – 2.060, 1.970 1.961, 1.853 – 1.610, 1.538 – 1.493, 1.457 – 1.385, 1.331 1.322, 1.205 – 1.052, 1.034 – 0.926, 0.863 – 0.836, 0.719 – 0.674, 0.611 – 0.566	20% FDR after PLS regression to ALT.
NOESYPR1D-MFC 2	189	9.978, 9.915, 9.609, 9.519, 9.150, 7.422 – 7.404, 7.341, 7.179 – 7.161, 7.125, 6.999 – 6.972, 6.882 – 6.864, 6.369 – 6.333, 6.126 – 6.108, 6.072 – 6.045, 6.009, 5.631, 5.388, 5.352 – 5.262, 5.244 – 5.190, 4.283 – 4.247, 4.202 4.193, 4.085 – 4.031, 4.004, 3.941 – 3.923, 3.905, 3.887, 3.851, 3.833 – 3.815, 3.734 3.725, 3.707, 3.680 – 3.626, 3.491 – 3.383, 3.239, 3.194, 3.050 – 3.005, 2.843 – 2.807, 2.780, 2.744 – 2.726, 2.438 – 2.411, 2.249 – 2.213, 2.159 – 2.123, 2.096 – 2.078, 2.024 – 1.988, 1.862 – 1.835, 1.736 – 1.691, 1.601 – 1.547, 1.475 – 1.457, 1.358 – 1.340, 1.313 – 1.259, 1.034 1.025, 1.007 – 0.971, 0.890 – 0.872, 0.674, 0.422 – 0.377, 0.188, -0.118, -0.298	20% FDR by direct t-test between controls and high-dose liver toxins. Note two variables selected upfield of $\delta_H = 0$.
NOESYPR1D-MFC all	1,100	All	No filtering applied.

Table D.4: Sets of ROC statistics for CPMG models discriminating liver toxins from controls. Model numbers refer to lines in Tables 6.1 & 6.2.

Model	Area under the curve			Std error			Asymptotic 95% Confidence Interval					
							all		48h		168h	
	All	48h	168h	All	48h	168h	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
24 h ALT	0.687	0.686	0.687	0.042	0.063	0.057	0.605	0.77	0.563	0.81	0.574	0.799
ALT	0.619	0.778	0.495	0.041	0.055	0.053	0.539	0.7	0.669	0.886	0.39	0.599
Weight change	0.473	0.517	0.431	0.039	0.06	0.052	0.395	0.55	0.398	0.635	0.328	0.534
6.1.1	0.698	0.742	0.667	0.036	0.051	0.05	0.628	0.768	0.643	0.842	0.568	0.766
6.1.2	0.697	0.748	0.654	0.036	0.05	0.051	0.626	0.767	0.65	0.845	0.554	0.755
6.1.3	0.697	0.748	0.654	0.036	0.05	0.051	0.626	0.767	0.65	0.845	0.554	0.755
6.1.4	0.724	0.791	0.69	0.034	0.05	0.046	0.658	0.79	0.692	0.889	0.599	0.78
6.1.5	0.738	0.784	0.707	0.033	0.051	0.045	0.673	0.803	0.684	0.883	0.619	0.795
6.1.6	0.776	0.831	0.742	0.03	0.041	0.044	0.716	0.835	0.75	0.912	0.655	0.829
6.1.7	0.775	0.828	0.744	0.03	0.041	0.044	0.716	0.835	0.747	0.909	0.657	0.83
6.1.8	0.74	0.817	0.681	0.033	0.044	0.048	0.675	0.805	0.731	0.904	0.587	0.775
6.2.1	0.647	0.714	0.624	0.037	0.051	0.05	0.575	0.719	0.614	0.814	0.526	0.722
6.2.2	0.644	0.704	0.63	0.037	0.053	0.05	0.571	0.716	0.6	0.808	0.532	0.728
6.2.3	0.636	0.708	0.61	0.037	0.052	0.051	0.562	0.709	0.605	0.81	0.51	0.711
6.2.4	0.731	0.786	0.706	0.033	0.051	0.045	0.665	0.796	0.685	0.886	0.617	0.795
6.2.5	0.735	0.798	0.69	0.033	0.048	0.047	0.671	0.8	0.705	0.892	0.598	0.781
6.2.6	0.739	0.768	0.73	0.033	0.053	0.044	0.674	0.805	0.665	0.871	0.643	0.817
6.2.7	0.739	0.768	0.729	0.033	0.053	0.044	0.674	0.804	0.665	0.871	0.642	0.816
6.2.8	0.726	0.756	0.715	0.03	0.053	0.046	0.66	0.793	0.652	0.86	0.626	0.804

Table D.5: Sets of ROC statistics for CPMG models discriminating liver from other toxins. Model numbers refer to lines in Tables 6.1 & 6.2.

Model	Area under the curve			Std error			Asymptotic 95% Confidence Interval					
							all		48h		168h	
	All	48h	168h	All	48h	168h	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
24 h ALT	0.671	0.67	0.671	0.04	0.063	0.053	0.592	0.75	0.547	0.793	0.567	0.775
ALT	0.625	0.782	0.468	0.04	0.052	0.054	0.546	0.703	0.679	0.884	0.363	0.574
Weight change	0.536	0.563	0.513	0.042	0.063	0.056	0.454	0.618	0.44	0.686	0.405	0.622
6.1.1	0.622	0.654	0.596	0.039	0.058	0.053	0.546	0.699	0.541	0.767	0.492	0.701
6.1.2	0.628	0.661	0.598	0.038	0.057	0.052	0.552	0.703	0.55	0.772	0.496	0.701
6.1.3	0.628	0.661	0.598	0.038	0.057	0.052	0.552	0.703	0.55	0.772	0.496	0.701
6.1.4	0.598	0.671	0.543	0.039	0.058	0.051	0.522	0.674	0.558	0.784	0.443	0.642
6.1.5	0.61	0.665	0.565	0.039	0.06	0.052	0.533	0.687	0.548	0.782	0.463	0.666
6.1.6	0.691	0.727	0.676	0.035	0.051	0.047	0.623	0.759	0.627	0.827	0.583	0.769
6.1.7	0.69	0.721	0.676	0.035	0.051	0.047	0.622	0.759	0.62	0.822	0.583	0.769
6.1.8	0.664	0.735	0.599	0.037	0.052	0.051	0.591	0.736	0.633	0.836	0.499	0.7
6.2.1	0.567	0.641	0.504	0.04	0.057	0.055	0.489	0.646	0.529	0.752	0.396	0.611
6.2.2	0.558	0.628	0.495	0.041	0.058	0.056	0.478	0.638	0.515	0.741	0.385	0.604
6.2.3	0.556	0.631	0.49	0.041	0.057	0.056	0.476	0.635	0.519	0.743	0.38	0.599
6.2.4	0.615	0.666	0.582	0.039	0.059	0.05	0.539	0.691	0.549	0.782	0.484	0.681
6.2.5	0.62	0.683	0.56	0.039	0.058	0.053	0.543	0.697	0.57	0.797	0.456	0.664
6.2.6	0.629	0.652	0.604	0.038	0.06	0.051	0.554	0.705	0.535	0.769	0.505	0.703
6.2.7	0.63	0.652	0.604	0.038	0.06	0.051	0.554	0.705	0.535	0.769	0.505	0.703
6.2.8	0.606	0.638	0.579	0.039	0.06	0.053	0.529	0.683	0.521	0.755	0.476	0.682

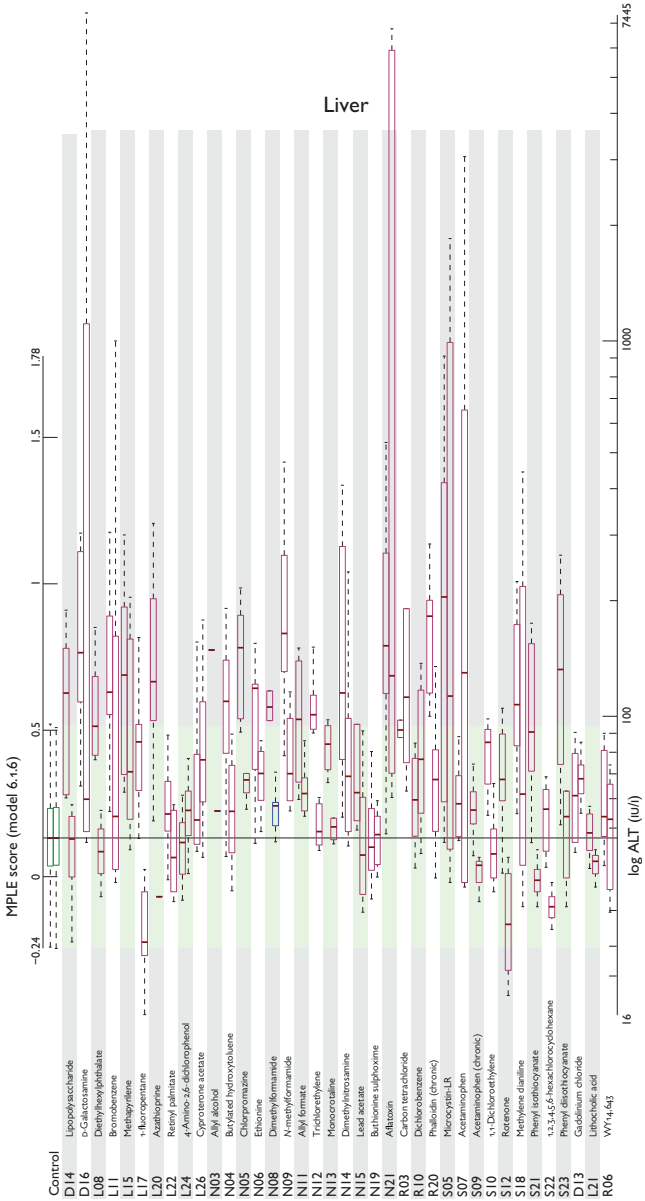

Figure D.1: Breakdown *f* 6.9, showing only the MPLE score for liver toxins. For comparison, both metrics have been scaled to show identical variance in the controls, their medians aligned and marked with a black line, and the 99% range about the controls on the MPLE metric shaded green.

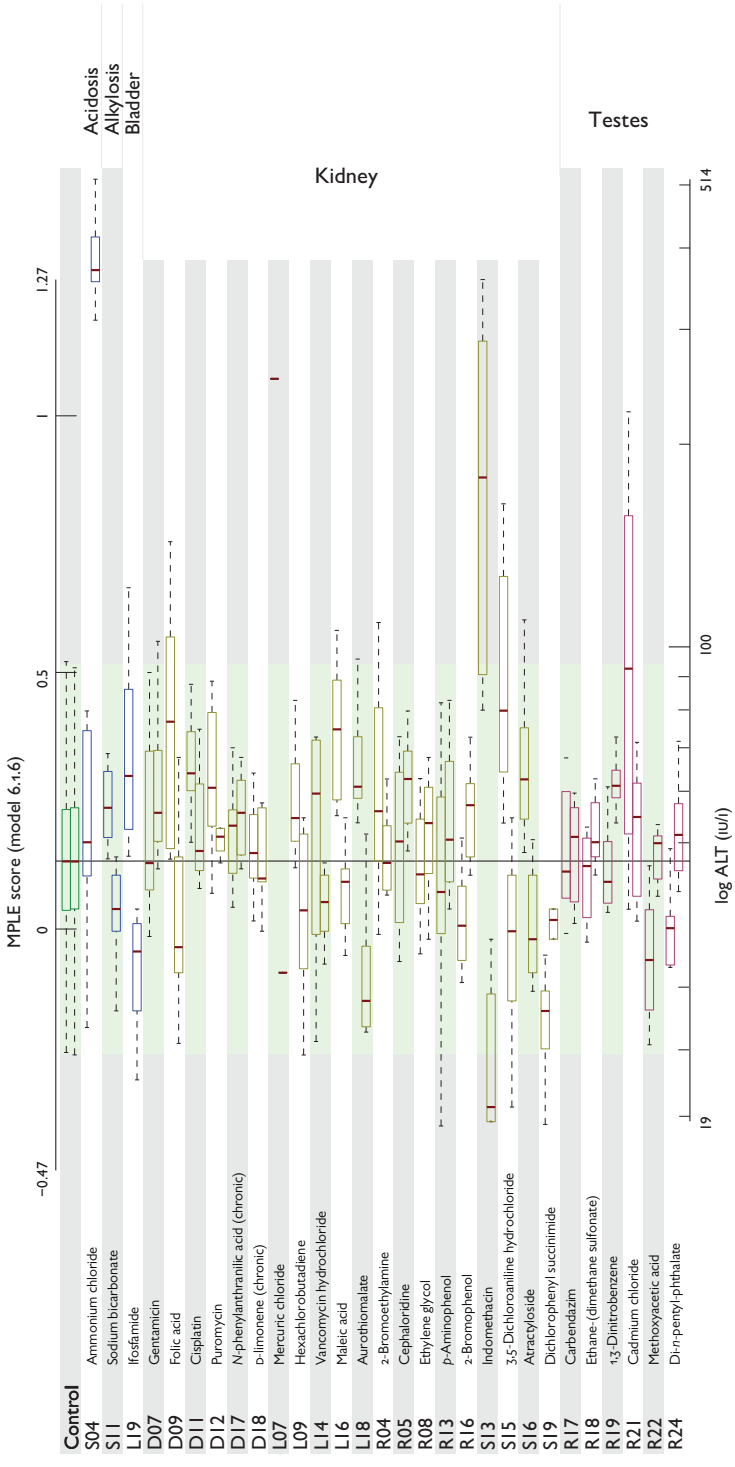

Figure D.2: Breakdown *f* 6.9, showing only the MPLE score for kidney & testes toxins. For comparison, both metrics have been scaled to show identical variance in the controls, their medians aligned and marked with a black line, and the 99% range about the controls on the MPLE metric shaded green.

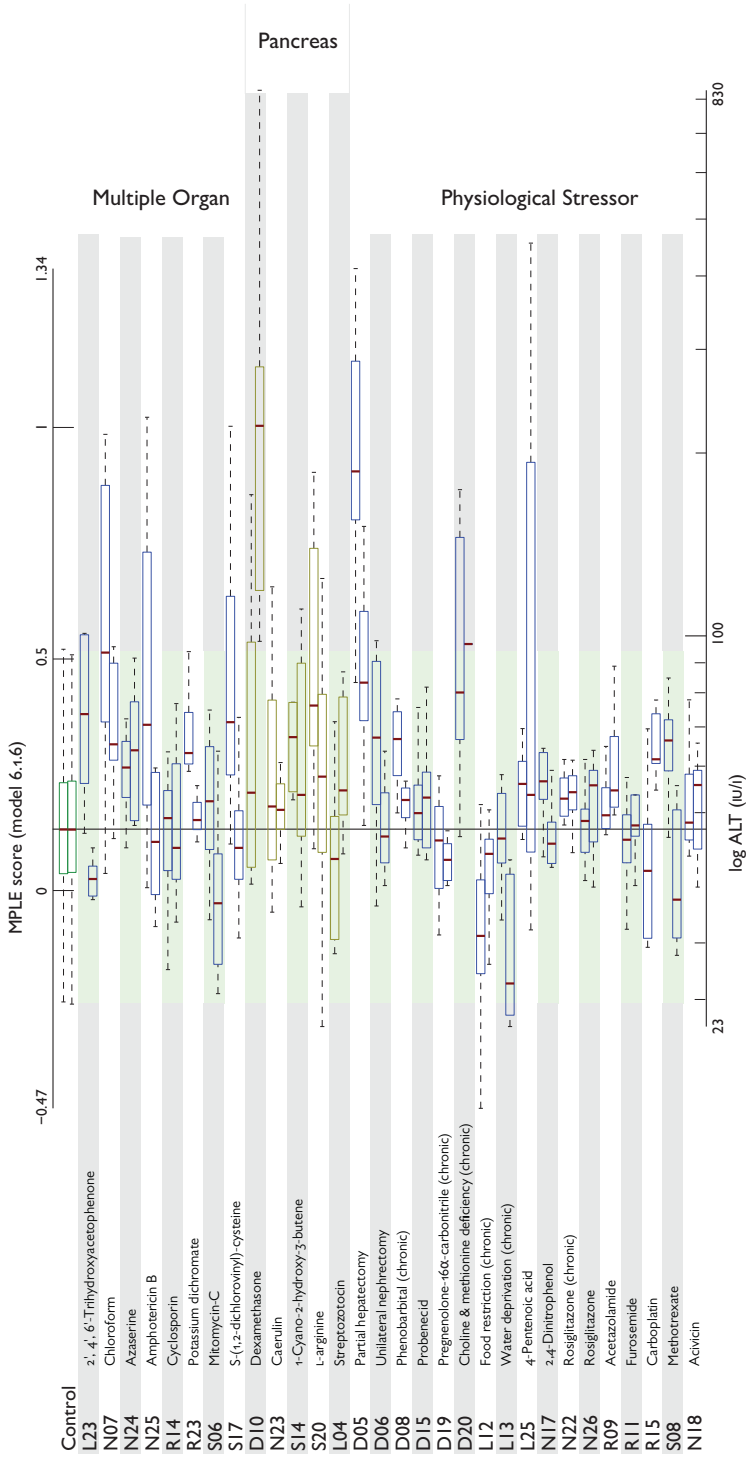

Figure D.3: Breakdown *f* 6.9, showing only the MPLE score for other stressors. For comparison, both metrics have been scaled to show identical variance in the controls, their medians aligned and marked with a black line, and the 99% range about the controls on the MPLE metric shaded green.

Bibliography

Aardema, M. J. & MacGregor, J. T. Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "—omics" technologies. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis*, **499**(1):13 – 25, 2002.

Ala-Korpela, M. ^{1}H NMR spectroscopy of human blood plasma. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **27**:475 – 554, 1995.

Ala-Korpela, M., Hiltunen, Y., Jokisaari, J., Eskelinen, S., Kivinitty, K., Savolainen, M. J. & Kesäniemi, Y. A. A comparative study of ^{1}H NMR lineshape fitting analyses and biochemical lipid analyses of the lipoprotein fractions VLDL, LDL and HDL, and total human blood plasma. *NMR in Biomedicine*, **6**:225 – 33, 1993.

Ala-Korpela, M., Korhonen, A., Keisala, J., Hörkkö, S., Korpi, P., Inpan, L. P., Jokisaari, J., Savolainen, M. J. & Kesäniemi, Y. A. ^{1}H NMR-based absolute quantitation of human lipoproteins and their lipid contents directly from plasma. *Journal of Lipid Research*, **35**:2292 – 304, 1994.

Alum, M. F., Shaw, P. A., Sweatman, B. C., Ubhi, B. K., Haselden, J. N. & Connor, S. C. 4,4-dimethyl-4-silapentane-1-ammonium trifluoroacetate (DSA), a promising universal internal standard for NMR-based metabolic profiling studies of biofluids, including blood plasma and serum. *Metabolomics*, **4**(2):122 – 7, 2008.

Amacher, D. E. A toxicologist's guide to biomarkers of hepatic response. *Human and Experimental Toxicology*, **21**(5):253 – 62, 2002. URL http://het.sagepub.com/cgi/content/abstract/21/5/253.

Andersson, C. A. Direct orthogonalization. *Chemometrics and Intelligent Laboratory Systems*, **47**(1):51 – 63, 1999.

Anthony, M. L., Sweatman, B. C., Beddell, C. R., Lindon, J. C. & Nicholson, J. K. Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine. *Molecular Pharmacology*, **46**(1):199 – 211, 1994.

Antti, H., Ebbels, T. M. D., Keun, H. C., Bollard, M. E., Beckonert, O., Lindon, J. C., Nicholson, J. K. & Holmes, E. C. Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects. *Chemometrics and Intelligent Laboratory Systems*, **73**(1):139 – 49, 2004.

Balasubramaniam, S., Simons, L. A., Chang, S. & Hickie, J. B. Reduction in plasma cholesterol and increase in biliary cholesterol by a diet rich in n-3 fatty acids in the rat. *Journal of Lipid Research*, **26**:684 – 9, 1985.

Beckwith-Hall, B. M., Nicholson, J. K., Nicholls, A. W., Foxall, P. J. D., Lindon, J. C., Connor, S. C., Abdi, M., Connelly, J. & Holmes, E. C. Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins. *Chemical Research in Toxicology*, **11**(4):260–72, 1998.

Beckwith-Hall, B. M., Thompson, N. A., Nicholson, J. K., Lindon, J. C. & Holmes, E. A metabonomic investigation of hepatotoxicity using diffusion-edited ^{1}H NMR spectroscopy of blood serum. *The Analyst*, **128**(7):814–8, 2003.

Benjamini, Y. & Hochberg, Y. Controling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society Series B*, **57**(1):289–300, 1995.

Bock, J. L. Analysis of serum by high-field proton nuclear magnetic resonance. *Clinical Chemistry*, **28**(9):1873–7, 1982.

Bollard, M. E., Contel, N. R., Ebbels, T. M. D., Smith, L., Beckonert, O., Cantor, G. H., Lehman-McKeeman, L., Holmes, E. C., Lindon, J. C., Nicholson, J. K. & Keun, H. C. NMR-based metabolic profiling identifies biomarkers of liver regeneration following partial hepatectomy in the rat. *Journal of Proteome Research*, 2009. URL http://dx.doi.org/10.1021/pr900200v.

Bollard, M. E., Keun, H. C., Beckonert, O., Ebbels, T. M. D., Antti, H., Nicholls, A. W., Shockcor, J. P., Cantor, G. H., Stevens, G., Lindon, J. C., Holmes, E. C. & Nicholson, J. K. Comparative metabonomics of differential hydrazine toxicity in the rat and mouse. *Toxicology and Applied Pharmacology*, **204**(2):135–51, 2005.

Boole, G. The calculus of logic. *Cambridge and Dublin Mathematical Journal*, **3**:183–98, 1848. URL http://www.maths.tcd.ie/pub/HistMath/People/Boole/CalcLogic/CalcLogic.html.

Boxenbaum, H. Interspecies pharmacokinetic scaling and the evolutionary-comparative paradigm. *Drug Metabolism Reviews*, **15**(5):1071 – 121, 1984.

Brindle, K. M., Brown, F. F., Campbell, I. D., Grathwohl, C. & Kuchel, P. W. Application of spin-echo nuclear magnetic resonance to whole-cell systems. Membrane transport. *Biochemical Journal*, **180**(1):37–44, 1979.

Bro, R. & Smilde, A. K. Centering and scaling in component analysis. *Journal of Chemometrics*, **17**(1):16–23, 2003. doi:10.1002/cem.773.

Brown, T. R. & Stoyanova, R. NMR spectral quantitation by principal-component analysis. ii. determination of frequency and phase shifts. *Journal of Magnetic Resonance, Series B*, **112**:32–43, 1996.

Butler, W. L. & Hopkins, D. W. Higher derivative analysis of complex absorption spectra. *Photochemistry and Phototbiology*, **12**(6):439–525, 1970.

Carr, H. Y. & Purcell, E. M. Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Physical Review*, **94**(3):630–8, 1954. doi:10.1103/PhysRev.94.630.

Chayes, F. & Trochimczyk, J. An effect of closure on the structure of principal components. *Mathematical Geology*, **10**(4):323–33, 1978.

Claridge, T. D. W. *High-Resolution NMR Techniques in Organic Chemistry*. Pergamon, 1999.

Clayton, T. A., Baker, D., Lindon, J. C., Everett, J. R. & Nicholson, J. K. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proceedings of the National Academy of Sciences*, **106**(34):14,728–33, 2009.

Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., Holmes, E. C. & Nicholson, J. K. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Analytical Chemistry*, **77**(5):1282–9, 2005a.

Cloarec, O., Dumas, M.-E., Trygg, J., Craig, A., Barton, R. H., Lindon, J. C., Nicholson, J. K. & Holmes, E. C. Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^{1}H NMR spectroscopic metabonomic studies. *Analytical Chemistry*, **77**(2):517–26, 2005b.

Coen, M., Lenz, E. M., Nicholson, J. K., Wilson, I. D., Pognan, F. & Lindon, J. C. An integrated metabonomic investigation of acetaminophen toxicity in the mouse using NMR spectroscopy. *Chemical Research in Toxicology*, **16**(3):295–303, 2003.

Coen, M., Ruepp, S. U., Lindon, J. C., Nicholson, J. K., Pognan, F., Lenz, E. M. & Wilson, I. D. Integrated application of transcriptomics and metabonomics yields new insight into the toxicity due to paracetamol in the mouse. *Journal of Pharmaceutical and Biomedical Analysis*, **35**(1):93–105, 2004.

Council, N. R. *Toxicity Testing in the 21st Century: A Vision and a Strategy*. The National Academies Press, 500 Fifth Street, NW Washington, DC 20001, 2007. ISBN 978–0–309–10992–5.

Couto Alves, A., Rantalainen, M., Holmes, E. C., Nicholson, J. K. & Ebbels, T. M. D. Analytic Properties of Statistical Total Correlation Spectroscopy Based Information Recovery in ^{1}H NMR Metabolic Data Sets. *Analytical Chemistry*, **81**(6):2075–84, 2009.

Crockford, D. J., Holmes, E. C., Lindon, J. C., Plumb, R. S., Zirah, S., Bruce, S. J., Rainville, P., Stumpf, C. L. & Nicholson, J. K. Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabonomic toxicology studies. *Analytical Chemistry*, **78**(2):363–71, 2006. ISSN 0003-2700.

Crockford, D. J., Keun, H. C., Smith, L. M., Holmes, E. C. & Nicholson, J. K. A curve-fitting method for direct quantitation of compounds in complex biological mixtures using ^{1}H NMR: application in metabonomic toxicology studies. *Analytical Chemistry*, **77**(14):4556–62, 2005.

Date, C. J. *A guide to the SQL standard*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1986. ISBN 0–201–05777–8.

De Beer, R., Van Den Boogaart, A., Cady, E., Graveron-Demilly, D., Knijn, A., Langenberger, K., Lindon, J. C., Ohlhoff, A., Serra, H. & Wylezinska-Arridge, M. Absolute metabolite quantification by in vivo NMR spectroscopy: V. multicentre quantitative data analysis trial on the overlapping background problem. *Magnetic Resonance Imaging*, **16**(9):1127–37, 1998.

Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in ^{1}H NMR metabonomics. *Analytical Chemistry*, **78**(13):4281–90, 2006.

Doull, J., Klaassen, C. D. & Amdur, M. O., editors. Casarett and Doull's *Toxicology: The Basic Science of Poisons*. Macmillan Publishing, New York, second edition, 1980. ISBN 0–02–330040–X.

Dumas, M.-E., Wilder, S. P., Bihoreau, M.-T., Barton, R. H., Fearnside, J. F., Argoud, K., D'Amato, L., Wallis, R. H., Blancher, C., Keun, H. C., Baunsgaard, D., Scott, J., Sidelmann, U. G., Nicholson, J. K. & Gauguier, D. Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. *Nature Genetics*, **39**(5):666–72, 2007.

Dyrby, M., Baunsgaard, D., Bro, R. & Engelsen, S. B. Multiway chemometric analysis of the metabolic response to toxins monitored by NMR. *Chemometrics and Intelligent Laboratory Systems*, **76**(1):79–89, 2005a.

Dyrby, M., Petersen, M., Whittaker, A. K., Lambert, L., Nørgaard, L., Bro, R. & Engelsen, S. B. Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Analytica Chimica Acta*, **531**(2):209–16, 2005b.

Ebbels, T. M. D., Keun, H. C., Beckonert, O., Antti, H., Bollard, M. E., Holmes, E. C., Lindon, J. C. & Nicholson, J. K. Toxicity classification from metabonomic data using a density superposition approach: 'CLOUDS'. *Analytica Chimica Acta*, **490**:109–22, 2003.

Ebbels, T. M. D., Keun, H. C., Beckonert, O., Bollard, M. E., Lindon, J. C., Holmes, E. C. & Nicholson, J. K. Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: The consortium on metabonomic toxicology screening approach. *Journal of Proteome Research*, **6**:4407–22, 2007.

Efron, B. & Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, **37**(1):36–48, 1983.

El-Deredy, W. Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy. *NMR in Biomedicine*, **10**:99–124, 1997.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. & Wold, S. *Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS)*, chapter Scaling, pages 213–25. Umetrics, 1999.

Esbensen, K. & Geladi, P. The start and early history of chemometrics: Selected interviews. part 2. *Journal of Chemometrics*, **4**(6):389–412, 1990.

Evilia, R. F. Quantitative NMR spectroscopy. *Analytical Letters*, **34**(13):2227–36, 2001.

Farrant, R. D., Lindon, J. C. & Nicholson, J. K. Internal temperature calibration for ^{1}H NMR spectroscopy studies of blood plasma and other biofluids. *NMR in Biomedicine*, **7**(5):243–7, 1994.

Fawcett, T. ROC graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories, HP Laboratories, MS 1143, 1501 Page Mil l Road, Palo Alto, CA 94304, 2004. URL `http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf`.

Fossel, E. T., Carr, J. M. & McDonagh, J. Detection of malignant tumors. Water-suppressed proton nuclear magnetic resonance spectroscopy of plasma. *New England Journal of Medicine*, **315**(22):1369–76, 1986.

Gartland, K. P. R., Bonner, F. W. & Nicholson, J. K. Investigations into the biochemical effects of region-specific nephrotoxins. *Molecular pharmacology*, **35**(2):242 – 50, 1989.

Gartland, K. P. R., Sanins, S. M., Sweatman, B. C., Beddell, C. R., Lindon, J. C. & Nicholson, J. K. Pattern recognition analysis of high resolution ^{1}H NMR spectra of urine. a nonlinear mapping approach to the classification of toxicological data. *NMR in Biomedicine*, **3**(4):166 – 72, 1990.

Gavaghan, C. L., Holmes, E. C., Lenz, E. M., Wilson, I. D. & Nicholson, J. K. An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk: ApfCD mouse. *FEBS letters*, **484**(3):169 – 74, 2000.

Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. *Journal of Chemometrics*, **2**(4):231 – 46, 1988.

Geladi, P. Chemometrics in spectroscopy. part 1. classical chemometrics. *Spectrochimica Acta Part B*, **58**:767 – 82, 2003.

Geladi, P. & Esbensen, K. The start and early history of chemometrics: Selected interviews. part 1. *Journal of Chemometrics*, **4**(5):337 – 54, 1990.

Geladi, P., MacDougall, D. & Martens, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, **39**(3):491 – 500, 1985.

Geladi, P., Sethson, B., Nyström, J., Lillhonga, T., Lestander, T. & Burger, J. Chemometrics in spectroscopy. part 2. examples. *Spectrochimica Acta Part B*, **59**:1347 – 57, 2004.

Gibb, S. Toxicity Testing in the 21st Century: A Vision and a Strategy. *Reproductive Toxicology*, **25**(1):136 – 8, 2008.

Giese, A. T. & French, C. S. The analysis of overlapping spectral absorption bands by derivative spectrophotometry. *Applied Spectroscopy*, **9**(2):78 – 96, 1955.

Ginsberg, H. N. Effects of statins on triglyceride metabolism. *The American journal of cardiology*, **81**(4):32B – 35B, 1998.

Gundert-Remy, U., Dahl, S. G., Boobis, A., Kremers, P., Kopp-Schneider, A., Oberemm, A., Renwick, A. & Pelkonen, O. Molecular approaches to the identification of biomarkers of exposure and effect — report of an expert meeting organized by COST Action B15. *Toxicology Letters*, **156**(2):227 – 40, 2005.

Haaland, D. M. & Thomas, E. V. Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, **60**:1193 – 202, 1988.

Harrell Jr., F. E. *Regression Modeling Strategies*. Springer - Verlag, 175 Fifth Avenue, New York, NY 10010, USA, 2001. ISBN 0 – 387 – 95232 – 2.

Harwood, L. M. & Claridge, T. D. W. *Introduction to Organic Spectroscopy*. Number 43 in Oxford Chemistry Primers. Oxford University Press, Oxford, 1997. ISBN 0 – 19 – 855755 – 8.

Harwood, P. D. Therapeutic dosage in small and large mammals. *Science*, **139**(3555):684 – 85, 1963.

Hausser, K. H. & Kalbitzer, H. R. *NMR in Medicine and Biology: Structure Determination, Tomography, In Vivo Spectroscopy*. Springer - Verlag, Berlin, english language edition, 1991. ISBN 3 – 540 – 53195 – 5.

Hawkins, D. M. On the investigation of alternative regressions by principal component analysis. *Journal of The Royal Statistical Society. Series C (Applied Statistics)*, **22**(3):275 – 86, 1973. ISSN 00359254. URL http://www.jstor.org/stable/2346776.

Hayes, A. W., editor. *Principles and Methods of Toxicology*. Raven Press, New York, second edition, 1989. ISBN 0 – 88167 – 439 – 7.

Higham, N. J. *Accuracy and stability of numerical algorithms*. Society for Industrial Mathematics, second edition, 2002. ISBN 978 – 089871521 – 7.

Holmes, E., Nicholls, A. W., Lindon, J. C., Connor, S. C., Connelly, J. C., Haselden, J. N., Damment, S. J. P., Spraul, M., Neidig, P. & Nicholson, J. K. Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chemical Research in Toxicology*, **13**(6):471 – 8, 2000.

Holmes, E. C., Bonner, F. W. & Nicholson, J. K. Comparative studies on the nephrotoxicity of 2-bromoethanamine hydrobromide in the fischer 344 rat and the multimammate desert mouse (*Mastomys natalensis*). *Archives of Toxicology*, **70**(2):89 – 95, 1995.

Holmes, E. C., Bonner, F. W., Sweatman, B. C., Lindon, J. C., Beddell, C. R., Rahr, E. & Nicholson, J. K. Nuclear magnetic resonance spectroscopy and pattern recognition analysis of the biochemical processes associated with the progression of and recovery from nephrotoxic lesions in the rat induced by mercury(II) chloride and 2-bromoethanamine. *Molecular Pharmacology*, **42**:922 – 30, 1992.

Hore, P. J. *Nuclear Magnetic Resonance*. Number 32 in Oxford Chemistry Primers. Oxford University Press, Oxford, 1995. ISBN 0 – 19 – 855682 – 9.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**(6):417 – 41, 1933.

Hwang, T.-L. & Shaka, A. J. Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *Journal of Magnetic Resonance, Series A*, **112**(2):275 – 9, 1995.

Höskuldsson, A. PLS regression methods. *Journal of Chemometrics*, **2**(3):211 – 28, 1988.

Isaksson, T. & Kowalski, B. Piece-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products. *Applied Spectroscopy*, **47**(6):702 – 9, 1993.

Johansson, E., Wold, S. & Sjöedin, K. Minimizing effects of closure on analytical data. *Analytical Chemistry*, **56**(9):1685 – 8, 1984.

Jolyon West, L., Pierce, C. M. & Thomas, W. D. Lysergic acid diethylamide: Its effects on a male asiatic elephant. *Science*, **138**(3545):1100 – 3, 1962.

Kawachi, T., Maruyama, T. & Singh, V. P. Rainfall entropy for delineation of water resources zones in Japan. *Journal of Hydrology*, **246**:36–44, 2001.

Keun, H. C. & Athersuch, T. J. Application of metabonomics in drug development. *Pharmacogenomics*, **8**(7):731–41, 2007. doi:10.2217/14622416.8.7.731.

Keun, H. C., Beckonert, O., Griffin, J. L., Richter, C., Moskau, D., Lindon, J. C. & Nicholson, J. K. Cryogenic probe ^{13}C NMR spectroscopy of urine for metabonomic studies. *Analytical Chemistry*, **74**(17):4588–93, 2002a.

Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M. E., Beckonert, O., Holmes, E. C., Lindon, J. C. & Nicholson, J. K. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Analytica Chimica Acta*, **490**(1-2):265–76, 2003.

Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M. E., Beckonert, O., Schlotterbeck, G., Senn, H., Niederhauser, U., Holmes, E. C., Lindon, J. C. & Nicholson, J. K. Analytical reproducibility in ^{1}H NMR-based metabonomic urinalysis. *Chemical Research in Toxicology*, **15**(11):1380–6, 2002b.

Keun, H. C., Ebbels, T. M. D., Bollard, M. E., Beckonert, O., Antti, H., Holmes, E. C., Lindon, J. C. & Nicholson, J. K. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chemical Research in Toxicology*, **17**(5):579–87, 2004.

Keun, H. C., Sidhu, J., Pchejetski, D., Lewis, J. S., Marconell, H., Patterson, M., Bloom, S. R., Amber, V., Coombes, R. C. & Stebbing, J. Serum molecular signatures of weight change during early breast cancer chemotherapy. *Clinical Cancer Research*, **15**(21):6716–23, 2009.

Kim, J.-H. Spurious correlation between ratios with a common divisor. *Statistics & Probability Letters*, **44**:383–6, 1999.

Kim, Y. S. The half-life of alanine aminotransferase and of total soluble protein in livers of normal and glucocorticoid-treated rats. *Molecular Pharmacology*, **5**(2):105–108, 1969.

Kitano, H. Systems biology: a brief overview. *Science*, **295**(5560):1662–4, 2002.

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–43. 1995.

Kowalski, B. R. & Bender, C. F. Pattern recognition. Powerful approach to interpreting chemical data. *Journal of the American Chemical Society*, **94**(16):5632–5639, 1972.

Krewski, D., Andersen, M. E., Mantus, E. & Zeise, L. Toxicity testing in the 21st century: Implications for human health risk assessment. *Risk Analysis*, **29**(4):474–9, 2009.

Kristensen, M., Savorani, F., Ravn-Haren, G., Poulsen, M., Markowski, J., Larsen, F. H., Dragsted, L. O. & Engelsen, S. B. NMR and interval PLS as reliable methods for determination of cholesterol in rodent lipoprotein fractions. *Metabolomics*, **6**(1):129–36, 2010. doi: 10.1007/s11306-009-0181-3.

Krzanowski, W. J. Cross-validation in priciple component analysis. *Biometrics*, **43**:575–84, 1987.

Lawton, W. H. & Sylvestre, E. A. Self modeling curve resolution. *Technometrics*, **13**(3):617 – 33, 1971.

Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755):788 – 91, 1999. doi:10.1038/44565.

Lee, D. D. & Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pages 556 – 62. MIT Press, 2001.

Levin, S., Semler, D. & Ruben, Z. Effects of two weeks of feed restriction on some common toxicologic parameters in sprague-dawley rats. *Toxicologic Pathology*, **21**(1):1 – 14, 1993.

Li, Z. & Vance, D. E. Phosphatidylcholine and choline homeostasis. *Journal of Lipid Research*, **49**(6):1187 – 94, 2008.

Lightcap, E. S. & Silverman, R. B. Slow-Binding Inhibition of [gamma]-Aminobutyric Acid Aminotransferase by Hydrazine Analogues. *Journal of Medicinal Chemistry*, **39**(3):686 – 94, 1996.

Lindgren, F., Hansen, B., Karcher, W., Sjöström, M. & Eriksson, L. Model validation by permutation tests: applications to variable selection. *Journal of Chemometrics*, **10**:521 – 32, 1996.

Lindon, J. C., Holmes, E. C. & Nicholson, J. K. Metabonomics in pharmaceutical R&D. *The FEBS Journal*, **274**(5):1140 – 51, 2007. ISSN 1742-464X (Print). doi:10.1111/j.1742-4658.2007.05673.x.

Lindon, J. C., Keun, H. C., Ebbels, T. M. D., Pearce, J. T., Holmes, E. C. & Nicholson, J. K. The consortium for metabonomic toxicology (COMET): aims, activities and achievements. *Pharmacogenomics*, **6**(7):691 – 9, 2005a. ISSN 1462-2416 (Print). doi:10.2217/14622416.6.7.691.

Lindon, J. C., Nicholson, J. K., Holmes, E. C., Antti, H., Bollard, M. E., Keun, H. C., Beckonert, O., Ebbels, T. M. D., Reily, M. D., Robertson, D. G., 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. Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicology and Applied Pharmacology*, **187**(3):137 – 46, 2003.

Lindon, J. C., Nicholson, J. K., Holmes, E. C., Keun, H. C., Craig, A., Pearce, J. T. M., Bruce, S. J., Hardy, N., Sansone, S.-A., Antti, H., Jonsson, P., Daykin, C., Navarange, M., Beger, R. D., Verheij, E. R., Amberg, A., Baunsgaard, D., Cantor, G. H., Lehman-McKeeman, L., Earll, M., Wold, S., Johansson, E., Haselden, J. N., Kramer, K., Thomas, C., Lindberg, J., Schuppe-Koistinen, I., Wilson, I. D., Reily, M. D., Robertson, D. G., Senn, H., Krotzky, A., Kochhar, S., Powell, J., van der Ouderaa, F., Plumb, R. S., Schaefer, H. & Spraul, M. Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotechnology*, **23**(7):833 – 8, 2005b.

Lucas, L. H., Larive, C. K., Wilkinson, P. S. & Huhn, S. Progress toward automated metabolic profiling of human serum: Comparison of CPMG and gradient-filtered NMR analytical methods. *Journal of Pharmaceutical and Biomedical Analysis*, **39**:156 – 63, 2005. doi:10.1016/j.jpba.2004.09.060.

Mainardi, L. T., Origgi, D., Lucia, P., Scotti, G. & Cerutti, S. A wavelet packets decomposition algorithm for quantification of in vivo ^{1}H-MRS parameters. *Medical Engineering & Physics*, **24**:201 – 8, 2002.

Mandel, J. Use of the singular value decomposition in regression analysis. *American Statistician*, **36**(1):15 – 24, 1982.

Maniara, G., Rajamoorthi, K., Rajan, S. & Stockton, G. W. Method performance and validation for quantitative analysis by ^{1}H and ^{31}P NMR spectroscopy. applications to analytical standards and agricultural chemicals. *Analytical Chemistry*, **70**:4921 – 8, 1998.

Maruyama, T., Kawachi, T. & Singh, V. P. Entropy-based assessment and clustering of potential water resources availability. *Journal of Hydrology*, **309**:104 – 13, 2005.

Meiboom, S. & Gill, D. Modified spin-echo method for measuring nuclear relaxation times. *Review of Scientific Instruments*, **29**(8):688 – 91, 1958. doi:10.1063/1.1716296. URL http://link.aip.org/link/?RSI/29/688/1.

Mendrick, D. L. & Schnackenberg, L. Genomic and metabolomic advance in the identification of disease and adverse event biomarkers. *Future Medicine*, **3**(5):605 – 15, 2009.

Mishraa, A. K., Özgera, M. & Singh, V. P. An entropy-based investigation into the variability of precipitation. *Journal of Hydrology*, **370**:139 – 54, 2009.

Nicholson, J. K., Buckingham, M. J. & Sadler, P. J. High resolution ^{1}H NMR studies of vertebrate blood and plasma. *Biochemical Journal*, **211**(3):605 – 15, 1983.

Nicholson, J. K., Connelly, J., Lindon, J. C. & Holmes, E. C. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, **1**(2):15 – 61, 2002.

Nicholson, J. K., Foxall, P. J. D., Spraul, M., Farrant, R. D. & Lindon, J. C. 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. *Analytical Chemistry*, **67**(5):793 – 811, 1995.

Nicholson, J. K. & Lindon, J. C. Systems biology: Metabonomics. *Nature*, **455**(7216):1054 – 6, 2008. URL http://dx.doi.org/10.1038/4551054a.

Nicholson, J. K., Lindon, J. C. & Holmes, E. C. `Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**(11):1181 – 9, 1999.

Nicholson, J. K., O'Flynn, M. P., Sadler, P. J., Macleod, A. F., Juul, S. M. & Sönksen, P. H. Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *Biochemical Journal*, **217**(2):365 –75, 1984.

Nicholson, J. K., Timbrell, J. A. & Sadler, P. J. Proton NMR spectra of urine as indicators of renal damage. mercury-induced nephrotoxicity in rats. *Molecular Pharmacology*, **27**(6):644 – 51, 1985.

Oishi, S., Oishi, H. & Hiraga, K. The effect of food restriction for 4 weeks on common toxicity parameters in male rats. *Toxicology and Applied Pharmacology*, **47**(1):15 – 22, 1979.

Otvos, J. D., Jeyarajah, E. J. & Cromwell, W. C. Measurement issues related to lipoprotein heterogeneity. *The American Journal of Cardiology*, **90**(8):22 – 9, 2002.

Pajukanta, A. J. L. . P. A treasure trove for lipoprotein biology. *Nature Genetics*, **40**(2):129 – 30, 2008.

Pearce, J. T. M., Athersuch, T. J., Ebbels, T. M. D., Lindon, J. C., Nicholson, J. K. & Keun, H. C. Robust algorithms for automated chemical shift calibration of 1D ^{1}H NMR spectra of blood serum. *Analytical Chemistry*, **80**(18):7158 – 62, 2008. ISSN 0003-2700. doi:10.1021/ac801114g. URL http://dx.doi.org/10.1021/ac8011494.

Pearson, K. Mathematical contributions to the theory of evolution. — on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, **60**:489 – 98, 1886 – 87.

Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(6):559 – 72, 1901.

Petersen, M., Dyrby, M., Toubro, S., Engelsen, S. B., Nørgaard, L., Pedersen, H. T. & Dyerberg, J. Quantification of lipoprotein subclasses by proton nuclear magnetic resonance-based partial least-squares regression models. *Clinical Chemistry*, **51**(8):1457 – 61, 2005.

Rantalainen, M., Cloarec, O., Beckonert, O., Wilson, I. D., Jackson, D., Tonge, R., Rowlinson, R., Rayner, S., Nickson, J., Wilkinson, R. W., Mills, J. D., Trygg, J., Nicholson, J. K. & Holmes, E. Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. *Journal of Proteome Research*, **5**(10):2642 – 55, 2006.

Rietjens, M. Reduction of error propagation due to normalization: Effect of error propagation and closure on spurious correlations. *Analytica Chimica Acta*, **316**(2):205 – 15, 1995.

Robosky, L. C., Wells, D. F., Egnash, L. A., Manning, M. L., Reily, M. D. & Robertson, D. G. Metabonomic identification of two distinct phenotypes in sprague-dawley (crl:CD(SD)) rats. *Toxicological Sciences*, **87**(1):277 – 84, 2005. URL http://toxsci.oxfordjournals.org/cgi/content/abstract/87/1/277.

Rose, W. C. The metabolism of creatine and creatinine. *Annual Review of Biochemistry*, **2**:187 – 206, 1933.

Sammon Jr., J. W. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, **18**(5):401 – 9, 1969.

Sarpal, A. S., Kapur, G. S., Mukherjee, S. & Jain, S. K. Estimation of oxygenates in gasoline by ^{13}C NMR spectroscopy. *Energy & Fuels*, **11**(3):662 – 7, 1997.

Schrijver, R. D., Vermeulen, D. & Daems, V. Dose-response relationships between dietary (n-3) fatty acids and plasma and tissue lipids, steroid excretion and urinary malondialdehyde in rats. *The Journal of Nutrition*, **122**(10):1979 – 87, 1992.

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, **27**:379 – 423, & 623 – 56, 1948.

Shaw, D. *Fourier transform N.M.R. spectroscopy*. Elsevier Scientific Pub. Co., 1976. ISBN 978 – 0444414663. URL http://books.google.com/books?id=EEiBAAAAIAAJ.

Skordi, E., Yap, I. K. S., Claus, S. P., Martin, F.-P. J., Cloarec, O., Lindberg, J., Schuppe-Koistinen, I., Holmes, E. & Nicholson, J. K. Analysis of time-related metabolic fluctuations induced by ethionine in the rat. *Journal of Proteome Research*, **6**(12):4572 – 81, 2007. URL http://dx.doi.org/10.1021/pr070268q.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, **78**(3):779 – 87, 2006.

Song, Y.-Q. Categories of coherence pathways for the CPMG sequence. *Journal of Magnetic Resonance*, **157**:82 – 91, 2002.

Spear, B. B., Heath-Chiozzi, M. & Huff, J. Clinical application of pharmacogenetics. *Trends in Molecular Medicine*, **7**(5):201 – 4, 2001.

Spratlin, J. L., Serkova, N. J. & Eckhardt, S. G. Clinical applications of metabolomics in oncology: A review. *Clinical Cancer Research*, **15**(2):431 – 40, 2009.

Spraul, M., Neidig, P., Klauck, U., Kessler, P., Holmes, E. C., Nicholson, J. K., Sweatman, B. C., Salman, S. R., Farrant, R. D. & Rahr, E. Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *Journal of Pharmaceutical and Biomedical Analysis*, **12**(10):1215 – 25, 1994.

Storey, J. D. A direct approach to false discovery rates. *Journal of The Royal Statistical Society Series B*, **64**(3):479 – 98, 2002. doi:10.1111/1467-9868.00346.

Stoyanova, R., Kuesel, A. C. & Brown, T. R. Application of principal-component analysis for NMR spectral quantitation. *Journal of Magnetic Resonance, Series A*, **115**(2):265 – 9, 1995.

Styles, P., Soffe, N., Scott, C., Cragg, D. A., Row, F., White, D. & White, P. A high-resolution NMR probe in which the coil and preamplifier are cooled with liquid helium. *Journal of Magnetic Resonance*, **60**:397 – 404, 1984.

Suehring, S. *MySQL Bible*. Wiley Publishing, Inc, 2002.

Tauler, R., Kowalski, B. & Fleming, S. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Analytical Chemistry*, **65**(15):2040 – 7, 2002. URL http://dx.doi.org/10.1021/ac00063a019.

Teahan, O. *A Metabonomic Approach to Biomarker Discovery in Prostate Cancer*. Ph.D. thesis, Department of Biomolecular Medicine, Imperial College London, 2009.

Teahan, O., Gamble, S., Holmes, E. C., Waxman, J., Nicholson, J. K., Bevan, C. & Keun, H. C. Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Analytical Chemistry*, **78**(13):4307 – 18, 2006. URL http://dx.doi.org/10.1021/ac051972y.

Thomas, L. D. K., Hodgson, S., Nieuwenhuijsen, M. & Jarup, L. Early Kidney Damage in a Population Exposed to Cadmium and Other Heavy Metals. *Environmental Health Perspectives*, **117**(2):181 – 4, 2009.

Timbrell, J. A. Biomarkers in toxicology. *Toxicology*, **129**:1 – 12, 1998.

Timbrell, J. A. *Principles of Biochemical Toxicology*. Taylor & Francis, 11 New Fetter Lane, London EC4P 4EE, third edition, 2000. ISBN 0−748−40736−7.

Tjandra, K., Le, T. & Swain, M. G. Experimental colitis attenuates development of toxin-induced cholangitis in rats. *Digestive Diseases and Sciences*, **47**(6):1216−23, 2002.

Torgrip, R. J. O., Åberg, K. M., Alm, E., Schuppe-Koistinen, I. & Lindberg, J. A note on normalization of biofluid 1D ^{1}H-NMR data. *Metabolomics*, **4**:114−21, 2008.

Trygg, J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics*, **16**(6):283−93, 2002.

Trygg, J. & Wold, S. O2-PLS, a two-block ($X-Y$) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*, **17**(1):53−64, 2003.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**(3):279−311, 1966.

Tukiainen, T., Tynkkynen, T., Mäkinen, V.-P., Jylänki, P., Kangas, A., Hokkanen, J., Vehtari, A., Gröhn, O., Hallikainen, M., Soininen, H., Kivipelto, M., Groop, P.-H., Kaski, K., Laatikainen, R., Soininen, P., Pirttilä, T. & Ala-Korpela, M. A multi-metabolite analysis of serum by ^{1}H NMR spectroscopy: Early systemic signs of alzheimer's disease. *Biochemical and Biophysical Research Communications*, **375**:356−61, 2008.

van der Kloet, F. M., Bobeldijk, I., Verheij, E. R. & Jellema, R. H. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *Journal of Proteome Research*, **8**(11):5132−41, 2009. doi:10.1021/pr900499r.

Vanhamme, L., Sundin, T., Van Hecke, P., Van Huffel, S. & Pintelon, R. Frequency-selective quantification of biomedical magnetic resonance spectroscopy data. *Journal of Magnetic Resonance*, **143**(1):1−16, 2000.

Vanhamme, L., van den Boogaart, A. & Huffel, S. V. Improved method for accurate and efficient quantification of MRS data with use of prior knowledge. *Journal of Magnetic Resonance*, **129**:35−43, 1997.

Varmuza, K. & Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, 6000 Broken Sound Parkway, NW, Boca Raton, FL 33487, USA, 2009. ISBN 978−1420059472.

Viant, M. R. Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochemical and Biophysical Research Communications*, **310**:943−8, 2003.

Viau, C., Lafontaine, M. & Payan, J. Creatinine normalization in biological monitoring revisited: the case of 1-hydroxypyrene. *International Archives of Occupational and Environmental Health*, **77**(3):177−85, 2004.

Wang, Y., Holmes, E., Tang, H., Lindon, J. C., Sprenger, N., Turini, M. E., Bergonzelli, G., Fay, L. B., Kochhar, S. & Nicholson, J. K. Experimental metabonomic model of dietary variation and stress interactions. *Journal of Proteome Research*, **5**(7):1535−42, 2006.

Watanabe, S. Karhunen-Loeve expansion and factor analysis: theoretical remarks and applications. In *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 635−60. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1965.

Watanabe, S. Pattern recognition as a quest for minimum entropy. *Pattern Recognition*, **13**(5):381 – 7, 1981.

Waterfield, C. J., Turton, J. A., Scales, M. D. C. & Timbrell, J. A. Investigations into the effects of various hepatotoxic compounds on urinary and liver taurine levels in rats. *Archives of Toxicology*, **67**(4):244 – 54, 1993.

Waters, N. J., Holmes, E., Williams, A., Waterfield, C. J., Farrant, R. D. & Nicholson, J. K. NMR and pattern recognition studies on the time-related metabolic effects of α-naphthylisothiocyanate on liver, urine, and plasma in the rat: An integrative metabonomic approach. *Chemical Research in Toxicology*, **14**(10):1401 – 12, 2001. URL http://dx.doi.org/10.1021/tx010067f.

Weljie, A. M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. M. Targeted profiling: Quantitative analysis of ^{1}H NMR metabolomics data. *Analytical Chemistry*, **78**(13):4430 – 42, 2006. ISSN 0003-2700.

Willker, W. & Leibfritz, D. Assignment of mono- and polyunsaturated fatty acids in lipids of tissues and body fluids. *Magnetic Resonance in Chemistry*, **36**:S79 – 84, 1998.

Winning, H., Larsen, F. H., Bro, R. & Engelsen, S. B. Quantitative analysis of NMR spectra with chemometrics. *Journal of Magnetic Resonance*, **190**(1):27 – 32, 2008.

Wishart, D. S., Bigam, C. G., Yao, J., Abildgaard, F., Dyson, H. J., Oldfield, E., Markley, J. L. & Sykes, B. D. ^{1}H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. *Journal of Biomolecular NMR*, **6**(2):135 – 40, 1995.

Wold, H. Causal flows with latent variables: Parting of the ways in light of NIPALS modeling. *European Economic Review*, **5**:67 – 87, 1974.

Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**:37 – 52, 1987.

Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**:109 – 30, 2001.

Wu, D., Chen, A. & Johnson Jr., C. S. An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. *Journal of Magnetic Resonance, Series A*, **115**(2):260 – 4, 1995.

Xu, J., Lee, G., Wang, H., Vierling, J. M. & Maher, J. J. Limited role for CXC chemokines in the pathogenesis of α-naphthylisothiocyanate-induced liver injury. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, **287**:734 – 41, 2004.

Zimmerman, H. J. *Hepatotoxicity: The Adverse Effects of Drugs and Other Chemicals on the Liver*. Lippincott Williams & Wilkins, Philadelphia, second edition, 1999. ISBN 0 – 7817 – 1952 – 6.

Zolnai, Z., Macura, S. & Markley, J. L. Phasing two- and three-dimensional NMR spectra by use of the hilbert transform can save computer time and space. *Journal of Magnetic Resonance*, **89**(1):94 – 101, 1990.

Zweig, M. & Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine [published erratum appears in clin chem 1993 aug;39(8):1589]. *Clinical Chemistry*, **39**(4):561 – 77, 1993. URL http://www.clinchem.org/cgi/content/abstract/39/4/561.

Šárka Mierisová & Ala-Korpela, M. MR spectroscopy quantitaion: A review of frequency domain methods. *NMR in Biomedicine*, **14**:247 – 59, 2001.

This work is set in 10pt Joanna MT Pro with headings in Gill Sans MT Pro, monospaced text in `Consolas` and maths in Euler.

Text was set with XeLaTeX, version 0.99991

Document Statistics

Text Revision: b10
Text was complied on: Friday 29th May, 2015

Words in main text: 33,177
Words in appendices: 6,570
Words in captions: 1,291
Number of sections: 118
Words in section headings: 486
Number of illustrations: 49
Number of math inlines: 224
Number of math displayed: 22