Stable isotope labelling and computational data mining approaches in drug metabolism studies

Michael Gerard Leeming

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

June, 2017

School of Chemistry
The University of Melbourne

ORCHID: 0000-0001-8981-0701
Abstract

Many small organic molecules will be chemically modified in some way after entering the body through metabolism. Thus, metabolism plays a significant role in determining the biological properties of a compound including half-life and toxicity profile. Identifying the metabolites of a compound is an important part of drug discovery projects. This thesis describes the development and application of methodologies to detect such metabolites. A guiding principle of this work is the detection of metabolites from a complex sample without prior knowledge of their structure or formation pathways. This ‘non-targeted’ analysis approach allows unknown or unexpected metabolites to be detected, providing a complete picture of the metabolic fate of a xenobiotic.

The basis of the non-targeted approach is described in Chapter 2. Here, paracetamol (APAP) and an equal quantity of $^{13}\text{C}_6$-APAP are simultaneously administered to rats. Analysis of blood plasma extracts by liquid chromatography mass spectrometry (LC-MS) resulted in mass spectra that contained pairs of ions that eluted simultaneously with equal intensity and are unique to metabolites. To automate data analysis, software called HiTIME was written enabling the non-targeted but selective detection of metabolites that appear as twin-ions from highly complex samples.

In some cases, xenobiotics form electrophilic metabolites that can covalently react with cellular proteins. This is thought to trigger allergic and toxic side-effects. The specific nature of the protein adduct may be a determinant of the biological response. Chapter 3 describes a reactivity survey of the electrophilic APAP metabolite, N-acetyl-\(p\)-benzoquinoneimine (NAPQI), towards a panel of amino acids and peptides. In addition to the well-known reactivity toward cysteine, previously undocumented covalent adducts between chemically synthesized NAPQI and tyrosine, tryptophan and methionine were also observed, isolated and characterised.

Chapter 4 introduces a non-targeted method to identify the protein targets of reactive metabolites which uses twin-ion and HiTIME analysis to detect tryptic peptides that have been covalently modified by drug metabolites. Software called Xenophile was developed that can identify the site of modification, the mass and the chemical formula of a reactive metabolite directly from shotgun LC-MS data. In Chapter 5, Xenophile was applied to identify the protein targets of APAP and $^{13}\text{C}_6$-APAP metabolites following incubation with liver tissue extracts.
and global trypsin digest. The Xenophile software correctly identified the reactive metabolite as \( C_8H_7NO_2 \) (i.e. NAPQI) and the adduction site as Cysteine residues. Further investigation identified 7 unique proteins that were modified by APAP including those that have been previously identified as adduction targets of NAPQI and other xenobiotic reactive metabolites.

HiTIME and Xenophile are then used to assess the small molecule metabolism and protein adduction profile of the environmental contaminants benzene, bromobenzene and toluene in liver extracts. Numerous twin-ions were detected that correspond to glutathione adducts of epoxide and quinone metabolites. No modified proteins were detected following analysis of global protein digests for any sample. To rule out false negatives, targeted approaches were taken to identify protein adducts. As these did not result in the recovery of any missed peptides, we conclude that protein adducts were not formed in these experiments. This finding is rationalized based on the extent of formation of small molecule GSH adducts.
Preface

Some of the projects described herein are highly collaborative with multiple researchers contributing their time and effort. These contributions are gratefully acknowledged below:

- Ms. Heather Daykin performed animal surgeries, drug administration and blood sampling described in Chapter 2.

- Dr. Andrew P. Isaac and Dr. Bernard J. Pope wrote initial version of the HiTIME algorithm described in Chapter 2 and used throughout.

- Dr. Luke F. Gamon aided in the NAPQI synthesis optimisation and performed the purification and NMR characterisation of APAP amino acid adducts described in Chapter 3.
Publications arising from this work

Publication arising from work performed toward the completion of this degree:


- **Leeming, M.G.; Gamon, L.F.; Wille, U; Donald, W.A.; O’Hair, R.A.J.,** What are the potential sites of protein arylation by N-Acetyl-p-benzoquinone imine (NAPQI)? *Chemical Research in Toxicology, 2015*, 28, 2224.


Additional publications arising from work performed during the tenure of this degree that are not discussed in this thesis:


Declaration

This is to certify that:

1) This thesis comprises only my own original work toward completion of the Doctor of Philosophy degree except where indicated otherwise in the Preface.

2) Due Acknowledgement has been made in the text to all other materials used.

3) Thesis is less than 100,000 words in length, exclusive of tables, bibliographies and appendices.

Michael G. Leeming

June, 2017
Acknowledgements

The past few years have been a time of profound learning and discovery and I would like to take this opportunity to thank the many wonderful people who have helped me along the way. First and foremost, I would like to express my deepest gratitude to my advisors Richard O’Hair and William ‘Alex’ Donald with whom I have worked for many years. Richard and Alex paved the way for discovery with their infectious enthusiasm and outstanding scientific knowledge. They have given me great freedom to pursue my own ideas and independent goals while also reigning me in when required. Whenever I needed, they had an open door and words of wisdom. I truly would not have reached this point without their guidance.

Throughout my PhD I’ve had the good fortune to work with many scientists on highly interdisciplinary projects, some of which are the basis of the work within. In particular, I would like to than Bernie, Andrew, Heather, Linda and Luke. It has been a pleasure to work with such a talented and friendly group of people. Thanks to Spencer Williams and James Ziogas for serving as my advisory panel their valuable scientific opinions. The work described in this thesis has made heavy use of LC-MS and I would like to thank Sioe See Volaric, Nicholas Williamson and Ching-Seng Ang for their help in maintaining instrumentation.

Of course, no PhD would be complete without a fair share of frustration and failure. On the more stressful days, I was always grateful for the distraction and support of those around me. To the past and present members of the O’Hair group, and also the White, Separovic and Wille labs, thanks for the advice, the laughs, the coffees, tim tams, lunches and drinks. You’ve all helped to make this degree a fantastic experience and Bio21 a fun place to work.

A special thanks goes to my family, to whom I dedicate this thesis. To my mum, dad and sister, grandparents and extended family, thank you for your enduring love, encouragement, and understanding from the very earliest of days. You gave me the opportunity to follow my interests and sacrificed more for me than I could hope to express in words.
3 The potential protein arylation sites of NAPQI .......................................................82

Preface ........................................................................................................................................82
3.1 Summary ................................................................................................................................82
3.2 Introduction ...............................................................................................................................83
3.3 Results and discussion ..............................................................................................................85
  3.3.1 Chemical synthesis of NAPQI ..............................................................................................85
  3.3.2 Reactions of NAPQI with model amino acids .....................................................................86
  3.3.3 Bulk synthesis, isolation and characterisation of NAPQI adducts .........................................89
  3.3.4 Relative reactivity of protected amino acids toward NAPQI ................................................95
  3.3.5 NAPQI reactions with model peptides ................................................................................97
  3.3.6 APAP and GAIL-X-GAILR microsome incubations ............................................................99
3.4 Conclusion ...............................................................................................................................102

4 Xenophile: A software toolkit for the non-targeted identification of drug-protein adducts .................................................................................................................................103

Preface ........................................................................................................................................103
4.1 Summary ................................................................................................................................103
4.2 Introduction ..............................................................................................................................104
4.3 Experimental design ................................................................................................................107
  4.3.1 Synthetic data production ......................................................................................................108
  4.3.2 Initial data processing ..............................................................................................................109
4.4 Software development ............................................................................................................111
  4.4.1 The non-targeted algorithm ..................................................................................................112
  4.4.2 The targeted algorithm .........................................................................................................119
4.5 Software validation using synthetic data ................................................................................120
4.6 Xenophile features, utilisation and availability ........................................................................122
  4.6.1 Xenophile python library ......................................................................................................122
  4.6.2 Xenophile graphical user interface .......................................................................................126
  4.6.3 Xenophile licence ................................................................................................................129
4.7 Comparison to other methods .................................................................................................129
4.8 Conclusion ...............................................................................................................................131

5 Using Xenophile to identify the protein adduction targets of paracetamol in liver microsomes ..............................................................................................................................132

Preface ........................................................................................................................................132
5.1 Summary ...............................................................................................................................132
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3.6 Isolation and characterization of NAPQI amino acid adducts</td>
<td>182</td>
</tr>
<tr>
<td>8.3.7 Relative consumption of the amino acids by reaction with NAPQI</td>
<td>182</td>
</tr>
<tr>
<td>8.3.8 Preparative HPLC</td>
<td>183</td>
</tr>
<tr>
<td>8.3.9 Microsome incubations</td>
<td>183</td>
</tr>
<tr>
<td>8.3.10 Liquid chromatography mass spectrometry (LC/MS)</td>
<td>184</td>
</tr>
<tr>
<td>8.3.11 Direct infusion ESI-MS/MS</td>
<td>184</td>
</tr>
<tr>
<td>8.4 Chapter 4 methods</td>
<td>184</td>
</tr>
<tr>
<td>8.5 Chapter 5 Methods</td>
<td>187</td>
</tr>
<tr>
<td>8.5.1 Materials</td>
<td>187</td>
</tr>
<tr>
<td>8.5.2 Microsome incubations</td>
<td>187</td>
</tr>
<tr>
<td>8.5.3 Data analysis</td>
<td>188</td>
</tr>
<tr>
<td>8.5.4 Liquid chromatography-mass spectrometry</td>
<td>188</td>
</tr>
<tr>
<td>8.5.5 Production of semi-synthetic data sets</td>
<td>189</td>
</tr>
<tr>
<td>8.5.6 Formulae determination metrics</td>
<td>190</td>
</tr>
<tr>
<td>8.5.7 Post-processing of HiTIME data</td>
<td>191</td>
</tr>
<tr>
<td>8.6 Chapter 6 methods</td>
<td>191</td>
</tr>
<tr>
<td>8.6.1 Materials</td>
<td>191</td>
</tr>
<tr>
<td>8.6.2 Microsome incubations and sample preparation</td>
<td>192</td>
</tr>
<tr>
<td>8.6.3 Liquid chromatography-mass spectrometry</td>
<td>192</td>
</tr>
<tr>
<td>8.6.4 Data analysis</td>
<td>192</td>
</tr>
<tr>
<td>9 Appendix</td>
<td>194</td>
</tr>
<tr>
<td>9.1 Xenophile software installation guide</td>
<td>194</td>
</tr>
<tr>
<td>9.2 Xenophile license</td>
<td>199</td>
</tr>
<tr>
<td>9.3 Chapter 2 supplementary information</td>
<td>200</td>
</tr>
<tr>
<td>9.3.1 HiTIME shape fitting and score determination</td>
<td>200</td>
</tr>
<tr>
<td>9.3.2 Supplementary tables and figures</td>
<td>202</td>
</tr>
<tr>
<td>9.4 Chapter 3 supplementary data</td>
<td>204</td>
</tr>
<tr>
<td>9.4.1 Characterisation Data</td>
<td>204</td>
</tr>
<tr>
<td>9.4.2 $^1$H, $^{13}$C and 2D NMR Spectra for isolated compounds</td>
<td>206</td>
</tr>
<tr>
<td>9.5 Chapter 4 supplementary data</td>
<td>211</td>
</tr>
<tr>
<td>9.6 Chapter 5 supplementary data</td>
<td>218</td>
</tr>
<tr>
<td>10 References</td>
<td>224</td>
</tr>
</tbody>
</table>
List of figures

Figure 1-1. Diazepam undergoes both phase I and phase II metabolism.\textsuperscript{10} ...........................................26

Figure 1-2. Simplified overview of the relationship between metabolism, pharmacological effect, excretion and toxicity. ...........................................................................................................27

Figure 1-3. Proposed mechanism of CYP-mediated C-H bond activation. ...........................................28

Figure 1-4. Common oxidation reactions catalysed by CYP enzymes. Adapted from Guengerich.\textsuperscript{16} ...................................................................................................................30

Figure 1-5. UGT-mediated transfer of glucuronic acid from the UDPGA co-factor to a substrate RXH (X = O, N). .............................................................................................................33

Figure 1-6. Sulfotransferase-catalysed transfer of sulfonate from the PAPS co-factor to a substrate RXH (X = O, N). .............................................................................................................33

Figure 1-7. Glutathione-S-transferase catalysed conjugation of glutathione to an electrophilic α,β-unsaturated carbonyl compound. ...............................................................................35

Figure 1-8. Ester hydrolysis of enalapril forms the active ACE inhibitor enalaprilat. ...............36

Figure 1-9. Simulation of the plasma concentration of a drug during 14 consecutive doses. RHL = relative half-life. .................................................................................................................37

Figure 1-10. Optimisation of the metabolic stability of sEH inhibitors through structural changes. ............................................................................................................................38

Figure 1-11. Formation, isomerisation and protein covalent binding of acyl glucuronide conjugates through transacetylation and glycation. Stable glycation products can form from the 3- and 4-O-β-glucuronide. .................................................................................42

Figure 1-12. Metabolism pathways for sudoxicam and meloxicam. ...........................................44

Figure 1-13. Simplified schematic of the electrospray process. Adapted from Kebarle and Tang.\textsuperscript{148} ............................................................................................................................47

Figure 1-14. Schematic of a single-quadrupole mass analyser. Adapted from Paul 1990.\textsuperscript{162} 48

Figure 1-15. Schematic of a linear ion trap. Adapted from Schwartz et al.\textsuperscript{160} ..................................................................................49

Figure 1-16. Schematic of a single reflector time-of-flight mass analyser. The blue trace represents an ion with a higher initial velocity compared to the red trace. .......................51

Figure 1-17. Schematic of the Orbitrap mass analyser. Adapted from Cooks and co-workers.\textsuperscript{167} .........................................................................................................................................52

Figure 1-18. Fragmentation of drug-GSH conjugates upon collision-induced dissociation. In [1,2,\textsuperscript{13}C\textsubscript{2}, \textsuperscript{15}N-glycine]glutathione, the atoms marked with an * are enriched with \textsuperscript{15}N and \textsuperscript{13}C respectively. ...........................................55
Figure 1-19. Comparison of: A) the conventional two-group crossover design for determining absolute or relative bioavailability and B) the simultaneous dosing stable isotope approach. Adapted from Schellekens et al.\textsuperscript{194} ..................................................57

Figure 1-20. Generation of stable products from reactive metabolites through the use of trapping agents. Adapted from Ma and Subramanian.\textsuperscript{214} ..................................................59

Figure 2-1. A) Profile LCMS data mass spectrometry data in which doublet peaks assigned to APAP [M+H]\textsuperscript+ are shown. B) Representative ‘ideal’ twin-ion peaks that were used to weight the data. C) Output data after weighting centring on the light isotope. D) Visualisation of the results from HiTIME weighting as a two-dimensional heat map. RA and RT correspond to relative abundance and retention time, respectively. ...............66

Figure 2-2. Schematics showing the parameters employed for fitting the ideal twin-ion shape to raw LCMS data in the A) m/z and B) retention time dimension………………………….67

Figure 2-3. Histograms showing the distribution of scores obtained from the HiTIME algorithm for APAP-treated rats for, A) negative and B) positive ion data.…………………………..68

Figure 2-4. Positive (left column) and negative (right column) ion LCMS data from rats treated with A) and D) 1:1 APAP:[\textsuperscript{13}C\textsubscript{6}-APAP 10 mg kg\textsuperscript{-1} and, B) and E) prior to APAP administration scored by correlation with a hypothetical twin-ion model using the HiTIME algorithm. C) and F) APAP treatment data following subtraction of vehicle control. To aid in visualisation of the results, only a small segment that contains the detected twin-ion ‘hits’ is displayed. ……………………………………………………69

Figure 2-5. Negative ion LC-MS analysis of blood plasma extracts from rats administered 1:1 APAP and \textsuperscript{13}C\textsubscript{6}-APAP. A) EICs of m/z 230 and m/z 236 and B) mass spectrum corresponding to the HiTIME hit at 12.70 min and m/z 230 assigned to the elution of APAP-SO\textsubscript{3}. C) EICs and D) mass spectrum from the same data region of control data. Note the difference in y-axis intensity scales between treatment and control data. …………72

Figure 2-6. Negative ion LC-MS analysis of blood plasma extracts from rats administered 1:1 APAP and \textsuperscript{13}C\textsubscript{6}-APAP. A) EICs and B) mass spectrum corresponding to the HiTIME hit at 12.86 min and m/z 260 assigned to the elution of APAP-OMe-SO\textsubscript{3}. C) EICs and D) mass spectrum from the same data region of control data. …………………………………………………………73

Figure 2-7. Structures, nomenclature and transformation steps involved in the formation of APAP metabolites that were detected in this study. Structures in parentheses are metabolic intermediates that were not directly observed in these experiments.………………..75

Figure 2-8. Negative ion LCMS data from rats treated with 1:1 APAP:[\textsuperscript{13}C\textsubscript{6}-APAP 10 mg kg\textsuperscript{-1} weighted by correlation with hypothetical twin-ion model using doublet spacing of A) 5.0167, B) 4.0134, C) 3.0101 and D) 2.0067 m/z which correspond to the loss of 1, 2, 3 and 4 \textsuperscript{13}C labels respectively………………………………………………………77

Figure 2-9. Negative ion mass spectrum at 12.7 minutes showing twin-ions assigned to APAP-SO\textsubscript{3}. Insets show expansions of the regions around [M-H]\textsuperscript{-} and [2M-2H+Na]\textsuperscript{+}. …………………78

Figure 3-1. Extracted ion chromatograms from LC-MS analysis of NAPQI reaction solutions with protected amino acids. A) Cys, m/z 327.10092, B) Tyr, m/z 387.15506, C) Met, m/z 355.13222, and D) Trp, m/z 410.17105…………………………………………………………………………………87
Figure 3-2. Relative extent of formation of NHAc-X(APAP)-OMe as a function of pH of the buffered solution. Extent of formation is measured by EIC peak areas determined from the LC/MS data and indexed to the maximum value. The EIC m/z values used were as follows: Cys 327.10092; Met 355.13222; Trp 410.17105; Tyr 387.15506. 89

Figure 3-3. MS$^2$ CID spectra collected using the LTQ-FTICR mass spectrometer of purified products of NAPQI arylation of NHAc-X-OMe for X = A) Cys, B) Tyr, C) Met, and D) Trp. For Met and Trp, both isomers produced gave similar MS/MS spectra and only one example is provided for simplicity. The precursor ions are marked with an ‘*’ 91

Figure 3-4. Proposed mechanism of benzofuroindoline and pyrroloindoline products from reaction of NAPQI with protected tryptophan. 94

Figure 3-5. $^1$H spectra for selective NOESY experiment; A) Selected portion of the $^1$H NMR spectrum for mixture of NAc-Trp(APAP)-OMe pyrroloindoline rotamers B) 1D gradient NOE spectrum following irradiation at 3.15 ppm and C) 1D gradient NOE spectrum following irradiation at 3.08 ppm. Coloured regions indicate areas of selective irradiation. 95

Figure 3-6. LC/MS/MS spectra for NAPQI modified peptides of the form GAIL-X(APAP)-GAILR for A) X = Cys, B) X = Tyr, C) X = Trp, and D) X = Met. All peptides are in the 2+ charge state except for X = Met which is a 3+ ion. The MS$^2$ precursor ions are denoted by ‘*’ 98

Figure 3-7. A) EIC of m/z 568.3183 showing elution of the NAPQI modified GAIL-C-GAILR peptide formed in rat liver microsomes. B) CID spectrum of m/z 568.318 at 23.5 min. 100

Figure 3-8. Formation of APAP-GSH via bioactivation of APAP to NAPQI by rat liver microsomes. A) EIC traces for m/z 457.1387 showing elution of APAP-GSH, B) mass spectrum of the [APAP-GSH+H]$^+$ ion for the X = M case, C) CID spectrum of m/z 457.138 at 16.25 min for the X = M case. 101

Figure 4-1. Covalent modification of lysine side-chain amine group by flucloxacillin and subsequent MS/MS fragmentation. Adapted from Jenkins et al. 323 105

Figure 4-2. A) Raw, synthetic LC-MS data generated to simulate trypsin digestion of a mixture of 5 different proteins including one that has undergone covalent binding with TZD reactive metabolites at numerous sites. B) Results of HiTIME scoring of the synthetic LC-MS data using a mass spacing of 3.01005 Da to detect [M+2H]$^{2+}$ modified peptides. 109

Figure 4-3. Relationship between a HiTIME hit and potential native counterparts. 109

Figure 4-4. Correlation score as a function of theoretical CRM position for the modified HSA peptide ALVLIAFAQYLQCPFEDHVK 122

Figure 4-5. Example python script that uses the Xenophile utilities library to automate multiple data processing steps 125

Figure 4-6. Minimal python script to automate consecutive non-targeted CRM identification for 3 data files. 126
Figure 5-1. Venn diagrams indicating the number of proteins identified by database searching of LC-MS/MS data sets collected on the orbitrap mass spectrometer for A) paracetamol, and B) control microsomal protein digests.

Figure 5-2. Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a twin-ion spacing of 3.01005 aimed at mining doubly charged peptides modified by an APAP metabolite. A) APAP treatment, B) Vehicle control samples. C) EIC of the ‘light’ (m/z 752.33, red trace) and ‘heavy’ (m/z 755.34, blue trace) peaks indicated by the HiTIME hit at RT = 47.85 min. D) Mass spectrum at the region of EIC peak maximum. Analogous heat maps for remaining replicates of both APAP and VC treatment data are shown in Appendix Figure 9-11.

Figure 5-3. HCD MS² spectra of A) non-twin-ion peptide assigned by Mascot as (carbamidomethyl-C) EFTPCAQAFFQK²⁺ at m/z 699.3 at 37.3 minutes, and B) twin-ion peptide at m/z 745.3 at 43.9 minutes from LC-MS/MS analysis of APAP treated microsomal protein digests.

Figure 5-4. HCD MS² spectra of peptides assigned by targeted correlation of HiTIME and Mascot peptide assignments. EFTPC(X)AQAAFQK²⁺ where A) X = APAP, B) X = ¹³C₆ APAP and C) X = carbamidomethyl (C₂H₃NO, from iodoacetamide treatment of reduced proteins).

Figure 5-5. Heat maps produced by HiTIME scoring of the replicates of semi-synthetic LC-MS data sets (A-C) with a doublet spacing of 3.01005 aimed at mining doubly charged twin-ion peptides. Black boxes indicate locations of artificially introduced twin-ion signals.

Figure 5-6. Cumulative number of false positive HiTIME hits as a function of target rank for synthetic twin-in peptides.

Figure 5-7. Histogram of CRMs mass ppm errors assigned by non-targeted reactive metabolite searching from the true value of 149.04713 Da. Counts show average values from across the three replicates of semi-synthetic twin-ion data and errors are quoted as +/- one standard deviation.

Figure 5-8. Confidence in the CRM masses assigned by the non-targeted reactive metabolite detection algorithm for three replicates (A-C) of semi-synthetic twin-ion data. Confidence values are calculated from the normalized peptide correlation scores according to $C_1 - C_2 C_1 * 100$, where $C_1$ and $C_2$ are the correlation scores for the highest and second highest ranked hits respectively.

Figure 5-9. A) Number of formulae for theoretical reactive metabolites generated from 1337 molecules in the ZINC15 FDA approved drugs list. B) Histogram summarising the average number of possible formulae for molecules in 50 Da bins.

Figure 5-10. A) Isotope distribution of VFANPEDCAGFGK where the Cys residue is modified by C₂H₂O or C¹³CH₂O in equal proportions. B) Abundances of light and heavy peaks from the simulated peptide mixture in A calculated by subtraction of predicted isotope intensities. Note that heavy ions (blue peaks) have been offset from the light by m/z 0.15 to aid in visualisation. For both panels, relative peak heights are indexed to the intensity of the monoisotopic ion of the ‘light’ peptide.
Figure 6-1. Heat maps produced by HiTIME scoring Q-TOF LC-MS data from supernatants obtained following incubation of 1:1 PhX:13C6-PhX substrates with rat liver microsomes using Δm/z = 6.0201. A, C, E) incubations of X = Br, Me, H respectively without supplemental GSH. B, D, F) incubations containing 10 mM supplemental GSH for X = Br, Me, H respectively...

Figure 6-2. MS and MS/MS data for the twin-ion hit at m/z 494 and 22.2 mins in non-GSH supplemented PhBr incubations. A) Mass spectrum observed at 22.2 mins. B) EICs of m/z 494 and 500. Inset shows an expansion of the low abundance peaks and arrows indicate positions of bromine-containing twin-ions. C) MS/MS spectrum of m/z 494, and D) MS/MS spectrum of m/z 500. The precursor ions are marked with an ‘*’. ... 

Figure 6-3. Proposed pathways of formation for PhBr and PhH metabolites observed here. It should be noted that multiple isomers and formation pathways are possible for each pathway as discussed in text. For simplicity, only one possibility is drawn. ... 

Figure 6-4. Proposed mechanism of formation of QM-GSH from toluene in RLMs. 

Figure 6-5. Venn diagrams indicating the total number of proteins identified by database searching of LC-MS/MS data sets for A) PhH, B) PhMe, C) PhBr, and D) control microsomal protein digests. 

Figure 6-6. Heat maps produced by HiTIME scoring of global protein digests obtained following incubation of 1:1 PhX:13C6-PhX substrates with rat liver microsomes using Δm/z = 3.01005 to search for doubly charged peptides. A) PhH, B) PhMe, C) PhBr, and D) control. For simplicity, only 1 replicate of each treatment is shown here. 

Figure 6-7. Heat maps produced by HiTIME scoring of global protein digests obtained following incubation of 1:1 PhX:13C6-PhX substrates with rat liver microsomes using Δm/z = 2.0067 to search for triply charged peptides. A) PhH, B) PhMe, C) PhBr, and D) control. For simplicity, only 1 replicate of each treatment is shown here. 

Figure 6-8. Examples of highly scoring regions in HTIME data that were not consistent with true twin-ions upon manual verification. A) mismatched isotopic peaks from peptides of the same charge state, and B) mismatch of peptides with different charge states. 

Figure 6-9. A) An exemplary MS1 spectrum of the peptide at m/z 893.4 and 74.45 mins. B) EIC traces for the microsomal glutathione-S-transferase peptide VFANPEDCAGFGK in the 2+ charge state with either bromobenzene epoxide (m/z 763.78), bromobenzene quinone (m/z 770.77) or carbamidomethyl modifications (m/z 706.32). Note that EIC intensities for m/z 763 and 770 traces have been multiplied by a factor of 10. 

Figure 6-10. EIC peak area for the most abundant APAP and PhX metabolite GSH adducts. Figures above the bars indicate the relative proportions of these ions. 

Figure 9-1. Shell script that automates the installation of Xenophile and its dependencies for Ubuntu 14.04. 

Figure 9-2. 1H NMR spectrum for the isolated adduct NAc-Cys(APAP)-OMe. 

Figure 9-3. 13C NMR spectrum for the isolated adduct NAc-Cys(APAP)-OMe. 

Figure 9-4. 1H NMR spectrum for the isolated adduct NAc-Met(APAP)-OMe isomer A.
Figure 9-5. $^1$H NMR spectrum for the isolated adduct NAc-Met(APAP)-OMe isomer B...208

Figure 9-6. $^1$H NMR spectrum for the isolated adduct NAc-Tyr(APAP)-OMe .................208

Figure 9-7. $^1$H-$^1$H gCOSY NMR spectrum for the isolated adduct NAc-Tyr(APAP)-OMe.209

Figure 9-8. $^1$H NMR spectrum for the isolated benzofuroindoline NAc-Trp(APAP)-OMe adduct ...................................................................................................................................209

Figure 9-9. $^1$H-$^1$H gCOSY NMR spectrum for the isolated benzofuroindoline NAc- Trp(APAP)-OMe adduct ...................................................................................................................................210

Figure 9-10. $^1$H NMR spectrum for the isolated pyrroloindoline NAc-Trp(APAP)-OMe adduct ...................................................................................................................................210

Figure 9-11. Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a doublet spacing of 3.01005 aimed at mining doubly charged peptides modified by an APAP metabolite. A-C) APAP treatment, D-F) vehicle control. Data points with a weighted score less than 15 have been omitted for clarity. For convenience, the data shown here in panels A and D are reproduced from Figure 5-2A and Figure 5-2B respectively in the main text ..............................................................................................................221

Figure 9-12. Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a doublet spacing of 2.0067 aimed at mining triply charged peptides modified by an APAP metabolite. A-C) APAP treatment, D-F) vehicle control. Data points with a weighted score less than 15 have been omitted for clarity................222

Figure 9-13. Histogram of the distribution of scores from HiTIME analysis of APAP treatment and vehicle control (VC) micromal protein digest LC-MS data using twin-ion spacing settings of A) 3.01005, and B) 2.0067.................................................................223
List of tables

Table 1-1. Substrate characteristics and examples for some major CYP enzymes involved in xenobiotic metabolism. Adapted from Lin et al. and Smith et al. ............................... 29
Table 1-2. Non-CYP enzymes involved in phase I xenobiotic metabolism. ........................ 32
Table 1-3. List of enzyme classes involved in common phase II metabolic reactions. ........ 34
Table 1-4. List of some common functional groups that are known to form reactive metabolites in some instances. ................................................................. 41
Table 1-5. Mass defects of atoms that are common in organic small-molecule drugs ........ 54
Table 2-1. Non-covalent adducts of twin-ion metabolites detected by the HiTIME algorithm for negative ion LCMS data. ............................................................... 70
Table 2-2. List of APAP metabolites detected in LCMS datasets by the HiTIME algorithm. 74
Table 3-1. EIC peak areas for relative depletion of amino acids in 10 mM sodium phosphate buffer upon competitive reaction with an excess of NAPQI. ................................. 96
Table 4-1. Examples of different amino acid residues modified by various xenobiotics..... 106
Table 4-2. Program inputs for non-targeted CRM identification ........................................... 113
Table 4-3. Theoretical ‘b’ and ‘y’ series ions generated via the rolling modification function for the exemplary peptide GLCGLR. In this example, suppose that carbamidomethylation of cystene is assigned by Mascot (denoted by C\text{CAM}) and that the mass difference between HiTIME and mascot peptides is 19 Da and X\text{M} refers to placement of the hypothetical modification (M) on residue X. ..................................................................................... 115
Table 4-4. Candidate molecular formulae determined by non-targeted CRM identification of the products of in silico adduction of TZD metabolites to HSA tryptic peptides. ...... 121
Table 4-5. Functions contained within each tab within the graphical user interface of the Xenophile software. Images of the GUI are provided in the Appendix. ....................... 127
Table 5-1. Top three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 1 of 3) ................................................................. 137
Table 5-2. Summary of the peptides assigned by directed Mascot search as carrying NAPQI modification. APAPL and APAPH refer to the natural abundance APAP and 13C6 APAP respectively. ................................................................. 142
Table 6-1. Manually verified twin-ion signals detected following HiTIME scoring of PhX microsome incubation LC-MS data. ................................................................. 164
Table 6-2. PhX-derived variable modification used for targeted database searching of microsomal protein digests. ................................................................. 168
Table 8-1. MS parameters for APAP metabolite detection using the Agilent 6520 Q-TOF mass spectrometer ................................................................. 177
Table 8-2 External software libraries used in the production of the non-targeted protein-CRM adduct identification algorithm. Only libraries not included in standard Python distributions are listed here. ................................................................. 185

Table 9-1. Variables employed in the HiTIME fitting algorithm ........................................ 202
Table 9-2. Transformation parameters of MS data for APAP metabolite detection .......... 202
Table 9-3. Parameters employed in HiTIME data processing .......................................... 203
Table 9-4. Parameters employed in MetExtract data processing .................................... 203
Table 9-5. Parameters used for non-targeted reactive metabolite detection for Trazodone .. 211
Table 9-6. Atom ranges used in non-targeted CRM detection for Trazodone ................. 211
Table 9-7. Parameters used for postprocessing of HiTIME scoring results ..................... 218
Table 9-8. Parameters used for non-targeted reactive metabolite detection .................... 218
Table 9-9. Atom ranges used in non-targeted CRM detection ....................................... 219
Table 9-10. Top three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 2 of 3) ................................................................. 219
Table 9-11. Top three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 3 of 3) ................................................................. 219
Table 9-12. Parameters for targeted detection of NAPQI-modified peptides .................... 220
Table 9-13. Summary of the peptide signals created for each data set and the discovery rate of these peptides for HiTIME and NTPS analysis .................................................. 220
List of schemes

Scheme 3-1. Activation of APAP to NAPQI followed by arylation reactions with sulphydryl side chains or other reactive nucleophiles from protein side-chains. ................................. 84

Scheme 3-2. Structures of the isolated covalent adducts formed via reaction of NAPQI with NAc-X-OMe as determined via $^1$H, $^{13}$C and $^1$H-$^1$H COSY NMR experiments: A) Cys, B) Tyr, C) Met, and D) Trp. A “*” represents the new stereocentre formed, resulting in diastereomers which are separated by HPLC. ................................................................. 90

Scheme 3-3. Neighbouring group fragmentation reaction of the fixed charge NHAc-Met(APAP)-OMe adduct...................................................................................................... 93

Scheme 4-1. Bioactivation of trazodone to the electrophilic metabolite CPPQI................. 108

Scheme 4-2. Workflow diagram for A) non-targeted CRM identification, and B) twin-ion directed targeted peptide adduct identification................................................................. 112

Scheme 4-3. Fragments generated by disconnection of rotatable bonds in the TZD performed during non-targeted CRM searching................................................................. 118

Scheme 4-4. Experimental and informatics workflow for targeted identification of CRM modified peptides and proteins. ................................................................. 120
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>Angiotensin-converting-enzyme</td>
</tr>
<tr>
<td>ACN</td>
<td>Acetonitrile</td>
</tr>
<tr>
<td>ADME</td>
<td>Absorption, distribution, metabolism and excretion</td>
</tr>
<tr>
<td>AMAP</td>
<td>N-acetyl m-aminophenol (3-hydroxyacetanilide)</td>
</tr>
<tr>
<td>APAP</td>
<td>N-acetyl p-aminophenol (4-hydroxyacetanilide, paracetamol)</td>
</tr>
<tr>
<td>BSA</td>
<td>Bovine serum albumin</td>
</tr>
<tr>
<td>CID</td>
<td>Collision-induced dissociation</td>
</tr>
<tr>
<td>CNLS</td>
<td>Constant neutral loss scan</td>
</tr>
<tr>
<td>CPPQI</td>
<td>Chloro-p-piperazinequinoneimine</td>
</tr>
<tr>
<td>CRM</td>
<td>Chemically reactive metabolite</td>
</tr>
<tr>
<td>CYP</td>
<td>Cytochrome P450</td>
</tr>
<tr>
<td>DDA</td>
<td>Data-dependent acquisition</td>
</tr>
<tr>
<td>EC2</td>
<td>Elastic compute cloud</td>
</tr>
<tr>
<td>EI</td>
<td>Electron impact</td>
</tr>
<tr>
<td>EIC</td>
<td>Extracted ion chromatogram</td>
</tr>
<tr>
<td>ER</td>
<td>Endoplasmic reticulum</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray ionisation</td>
</tr>
<tr>
<td>FAD</td>
<td>Flavin adenine dinucleotide: FAD$^+$ = oxidised form, FADH$_2$ = reduced form.</td>
</tr>
<tr>
<td>FDA</td>
<td>United States food and drug administration</td>
</tr>
<tr>
<td>FHF</td>
<td>Fulminant hepatic failure</td>
</tr>
<tr>
<td>FMO</td>
<td>Flavin-containing monooxygenase</td>
</tr>
<tr>
<td>FT-ICR</td>
<td>Fourier-transform ion cyclotron resonance</td>
</tr>
<tr>
<td>GC-MS</td>
<td>Gas chromatography coupled to mass spectrometry</td>
</tr>
<tr>
<td>GIT</td>
<td>Gastro-intestinal tract</td>
</tr>
<tr>
<td>GSH</td>
<td>Glutathione</td>
</tr>
<tr>
<td>GST</td>
<td>Glutathione-S-transferase</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical user interface</td>
</tr>
<tr>
<td>HLM</td>
<td>Human liver microsomes</td>
</tr>
<tr>
<td>HPLC</td>
<td>High-performance liquid chromatography</td>
</tr>
<tr>
<td>HSA</td>
<td>Human serum albumin</td>
</tr>
<tr>
<td>IAA</td>
<td>Iodoacetamide</td>
</tr>
<tr>
<td>IADR</td>
<td>Idiosyncratic adverse drug reaction</td>
</tr>
</tbody>
</table>
ISD  Internal standard
*i.v.*  Intra venous
K_d  Dissociation constant
KIE  Kinetic isotope effect
LC-MS  Liquid chromatography coupled to mass spectrometry
MAO  Monoamine oxidase
MRM  Multiple reaction monitoring
MS/MS  Tandem mass spectrometry
*m/z*  Mass (m) to charge (z) ratio
NADP  Nicotinamide adenine dinucleotide phosphate
      NADP⁺ = oxidised form, NADPH = reduced form.
NAPQI  N-acetyl p-quinoneimine
NEM  N-ethylmaleimide
PAP  p-aminophenol
PAPS  3′-phosphoadenosine 5′-phosphosulfate
PIF  Precursor ion filtering
PTM  Post-translational modification
Q-TOF  Quadrupole-time of flight
R_f  Radio-frequency
RLM  Rat liver microsomes
RME  Residual mass error
RT  Retention time
sEH  Soluble epoxide hydrolase
SMARTS  SMILES arbitrary target specification
SMILES  Simplified molecular-input line-entry system
TCEP  Tris(2-carboxyethyl)phosphine hydrochloride
TEAB  Triethylammonium bicarbonate
TFA  Trifluoroacetate
TIC  Total ion chromatogram
TOF  Time-of-flight
TZD  Trazodone
UDPGA  Uridine diphosphate glucuronic acid
UGT  Uridine diphosphate glucuronosyltransferase
1 INTRODUCTION

1.1 Xenobiotic metabolism

Every living creature is continuously ingesting chemical compounds. This may be through food and drink, inhalation or even absorption across the skin. While the vast majority of these chemicals are harmless, they must eventually be cleared from the body to avoid the accumulation of unneeded compounds that may lead to adverse effects. Many ingested compounds will be cleared to some extent via excretion in urine, bile or faeces and some may be removed through exhaled breath. However, some of these absorbed molecules are too hydrophobic for aqueous excretion and too large to be volatile. Without some alternative means of processing these substances, accumulation and toxicity will occur.

The body has many processes that help to prevent build-up of hydrophobic exogenous compound and chief among these is ‘metabolism’. This refers to the set of biochemical transformations, usually enzyme mediated, that a given molecule undergoes. For most naturally occurring compounds, useful materials are produced in this process. For example, carbohydrates are used as a source of energy, while ingested proteins are broken down to their constituent amino acids, which are then used as the building blocks of our own proteins.

Unlike metabolism of the natural components of food, metabolism of synthetic xenobiotics generally does not produce useful nutrients or energy equivalents. ‘Xenobiotic’ is a broad term typically used to refer to synthetic compounds that are foreign to the body including medicinal drugs, industrial chemicals, air pollutants, and pesticides. In these cases, the main function of metabolism is to aid in removal of the compound from the body. Metabolism of xenobiotics
generally involves making chemical modifications to the molecule that increase polarity. This increases water solubility, which is required for filtration by the kidneys and excretion in urine.

While a wide range of enzymes and functional group transformations have been characterised, these can be broadly classified into two groups. ‘Phase I’ reactions typically result in ‘functionalisation’ of a compound via the installation or unmasking of polar moieties that enhance solubility and excretion. This most commonly occurs via oxidation but reduction or hydrolysis also can contribute in some cases. Together, these reactions often result in the formation of hydroxyl (R-OH), amino (R-NH₂) or carboxyl (R-COOH) groups and serve to both increase polarity and provide reactive sites for further metabolic reactions. Phase II processes conjugate large molecules that are water soluble to the compound which greatly increase polarity and facilitate excretion. Examples include conjugation of glucuronic acid, sulphate and taurine. In some instances, a third category can be identified. So-called ‘phase III’ processes involve trafficking of a xenobiotic or its metabolites. These are less commonly discussed than phase I and phase II reactions as they do not result in chemical changes to a xenobiotic or metabolite but rather serve to remove compounds from the cell through the action of efflux pumps and transporter proteins.

It should be noted that some researchers in the field of drug metabolism have advocated against the use of the ‘phase X’ nomenclature for describing metabolic reactions.¹ This is primarily because a xenobiotic need not be subjected to phase I transformations before phase II transformations take place, i.e. a phase I reaction is not a prerequisite for a phase II reaction. However, given that this terminology is widely understood and is still used in many modern publications,²⁻⁹ it will be used throughout this thesis.

As an example, some of the metabolic reactions operative for diazepam, a member of the benzodiazepine class of anxiolytics, are shown in Figure 1-1. This largely hydrophobic compound undergoes a combination of the phase I reactions N-demethylation and hydroxylation resulting in the formation of oxazepam. These mildly increase polarity however, in a phase II process, oxazepam is conjugated to the carbohydrate glucuronic acid resulting in a substantial increase in water solubility. Therefore, a range of metabolites can be detected in the analysis of urine samples from patients treated with diazepam that include unchanged drug, N-desmethyl and hydroxyl diazepam, oxazepam and various glucuronic acid conjugates.¹⁰
Diazepam undergoes both phase I and phase II metabolism.\(^\text{10}\)

Once formed, xenobiotic metabolites can trigger a range of different biochemical, pharmacological and physiological events and a simplified overview of these is given in **Figure 1-2**. For instance, metabolites may be directly filtered from the bloodstream and excreted, or, alternatively, may undergo additional metabolic reactions. Metabolites are not necessarily inert and may exert their own biological effects that are different to that of the precursor compound. Compounds that show minimal biological activity but give rise to therapeutically active metabolites are known as prodrugs. Some metabolites show off-target receptor binding can result in adverse reactions. Importantly, some compounds, give rise to metabolites that are electrophilic and chemically reactive through ‘bioactivation’ reactions. These reactive metabolites can form covalent bonds to cellular macromolecules which is thought to be a key driver of toxic and allergic reactions to drugs. Each of these pathways is discussed in more detail in the following sections.
1.2 Mechanisms of xenobiotic metabolism

1.2.1 Phase I reactions
Given the structural diversity of xenobiotics encountered in everyday life, a wide variety of metabolic pathways are required to facilitate their removal from the body. Numerous enzyme families contribute to xenobiotic metabolism. Undoubtedly the most important of such enzymes are cytochrome P450s. This enzyme superfamily is ubiquitous throughout nature and catalyses oxidative reactions on a wide variety of endogenous and exogenous substrates.\(^\text{11}\) Fifty-seven individual CYP enzymes have been identified in humans that are each encoded by separate genes\(^\text{12}\) and while only about 12 of these contribute to xenobiotic metabolism, they affect almost 75% of all xenobiotic transformations.\(^\text{13}\) Understanding CYP-mediated catalysis is critical to explaining the biological effects of many compounds and, as a result, these enzymes are of significant interest to pharmaceutical and toxicology researchers.\(^\text{14-19}\)

CYP enzymes use the redox properties of an iron-porphyrin prosthetic group to catalyse the oxidation of substrates by molecular oxygen.\(^\text{20, 21}\) The generalised catalytic cycle (shown in Figure 1-3) is initiated by substrate binding followed by a one-electron reduction of the iron core by NADPH (via NADPH-P450 reductase). Molecular O\(_2\) is reduced to superoxide upon binding to the ferrous core and protonated resulting in the release of H\(_2\)O. The resulting high-valent iron-oxo complex (sometimes referred to as ‘compound 1’)\(^\text{22}\) can then abstract adventitious hydrogen atoms from the bound substrate forming ‘compound 2’.\(^\text{23, 24}\) Collapse of the enzyme substrate via radical recombination releases the substrate and leaves the iron core...
in the ferric state.\textsuperscript{15} This is known as the ‘rebound’ mechanism and it can be used to explain a wide range of transformations catalysed by CYP enzymes. Common CYP-mediated transformations include hydroxylation, epoxidation, heteroatom dealkylation and heteroatom oxidation and mechanisms for these are given in Figure 1-4.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1-3.png}
\caption{Proposed mechanism of CYP-mediated C-H bond activation.}
\end{figure}

The majority of xenobiotics have multiple sites that could be substrates for CYP-catalysed oxidation and many factors influence which specific metabolites are formed from a compound \textit{in vivo}. For example, abstraction of a hydrogen atom from the xenobiotic substrate (compound 2 in Figure 1-3) forms a short-lived radical species and thus the stability of this intermediate contributes to the reaction energetics.\textsuperscript{23} In the absence of other factors, it would therefore be expected that standard models of radical stability should be capable of predicting the site of oxidation for organic substrates. This was nicely demonstrated by Frommer \textit{et al} who analysed and quantified the oxidation products of a range of small hydrocarbons by hepatic microsomal CYP enzymes.\textsuperscript{25} A series of alkanes was selected to minimise the effects of active site substrate selectively on the reaction outcome. For all molecules studied, it was observed that tertiary
carbon atoms were most susceptible to oxidation followed by secondary and then primary positions. This reactivity order correlates well with C-H bond strengths for tertiary, secondary and primary carbon atoms of 105, 101 and 98 kcal mol\(^{-1}\) for the prototypical cases.\(^{26}\)

Rationalisation of CYP reactivity on the basis of bond strengths is complicated by the fact that binding poses of the substrate may be defined by the configuration of the enzyme active site.\(^ {27,28}\) Detailed analysis of CYP2 family by Gotoh identified 6 substrate recognition sites in the overall protein structure that appeared to dictate the substrate specificity of individual members.\(^ {29}\) In these sites, even single amino acid mutations can result in substantial variations in relative substrate preference.\(^ {30,31}\) Structure activity relationships for multiple enzymes have allowed general substrate properties to be established for major enzymes in xenobiotic metabolism (Table 1-1).\(^ {32}\) Therefore the actual oxidation pathways operative for a given molecule are determined by a combination of radical stability effects and the substrate preferences of the protein.\(^ {32}\)

**Table 1-1.** Substrate characteristics and examples for some major CYP enzymes involved in xenobiotic metabolism. Adapted from Lin *et al.*\(^ {33}\) and Smith *et al.*\(^ {32}\)

<table>
<thead>
<tr>
<th>CYP</th>
<th>Substrate Properties</th>
<th>Example substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A2</td>
<td>Planar, neutral to moderately basic</td>
<td>Caffeine</td>
</tr>
<tr>
<td>2C9</td>
<td>Weakly acidic, ionized at physiological pH, H-bond donor</td>
<td>Diclofenac</td>
</tr>
<tr>
<td>2C19</td>
<td>Weakly acidic to strongly basic, H-bond donor</td>
<td>Omeprazole</td>
</tr>
<tr>
<td>2D6</td>
<td>Basic, hydrophilic, H-bond donor</td>
<td>Amitriptyline</td>
</tr>
<tr>
<td>2E1</td>
<td>Small (&lt; 200 Da) typically lipophilic molecules</td>
<td>Ethanol</td>
</tr>
<tr>
<td>3A4</td>
<td>Lipophilic, structurally diverse, H-bond donor/acceptor</td>
<td>Terfenadine</td>
</tr>
</tbody>
</table>
Figure 1-4. Common oxidation reactions catalysed by CYP enzymes. Adapted from Guengerich.16
Other enzyme classes that operate via different mechanisms can also affect oxidation of xenobiotics. For example, the flavin-containing monooxygenases (FMOs) use the organic FAD$^+$/FADH$_2$ couple to initiate heteroatom oxidation at nitrogen, sulphur and phosphorous.$^{34}$ A total of 6 FMOs have been identified in humans and, in some cases, FMOs are the predominant oxidation system involved in metabolism, particularly for sulphur-containing drugs.$^{35,36}$ FMOs display a relatively broad substrate tolerance similar to CYPs and their role in drug metabolism is thought to be underestimated.$^{37}$ Prototypical reactions include N-oxidation of tertiary amines such as nicotine and triethylamine. Similarly, monoamine oxidases (MAOs) catalyse the oxidative deamination of primary amines forming aldehydes. Endogenous substrates include the neurotransmitter serotonin, which has made MAOs an attractive target for pharmacological intervention in mental disorders. In many cases, there is considerable overlap between the reactions catalysed by CYP enzymes, FMOs and MAOs.

While oxidative reactions are undoubtedly the most common transformation in xenobiotic metabolism, hydrolysis and even reduction reactions can play a role in some instances. For example, hydrolytic enzymes are prolific throughout nature and can catalyse the cleavage of esters, thioesters and amides and a number of different catalytic mechanisms are known.$^{38}$ Many utilise the classic serine-histidine-glutamate catalytic triad or similar variants to stabilise the tetrahedral intermediates formed during typical nucleophilic acyl substitution reactions.$^{39,40}$ Some enzymes utilise cysteine as the nucleophilic residue while others employ a coordinated Zn$^{2+}$ ion to aid in activation of H$_2$O to OH$^-$. Reduction reactions are considerably rarer but, interestingly, are sometimes catalysed by the very same enzymes that are widely known for their oxidising capacity such as the CYPs.$^{38,43}$ Nitrogen oxides, azo compounds, ketones, and aldehydes can be reduced in some cases.
**Table 1-2.** Non-CYP enzymes involved in phase I xenobiotic metabolism.⁴²

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Enzyme class</th>
<th>Subcellular location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxidation</td>
<td>Flavin-containing monooxygenase (FMO)</td>
<td>ER</td>
</tr>
<tr>
<td></td>
<td>Monoamine oxidase (MAO)</td>
<td>Mitochondria</td>
</tr>
<tr>
<td>Hydrolysis</td>
<td>Carboxylesterase</td>
<td>ER, cytosol, lysosome</td>
</tr>
<tr>
<td></td>
<td>Epoxide hydrolase</td>
<td>ER, cytosol</td>
</tr>
<tr>
<td>Reduction</td>
<td>Aldo/Keto Reductases</td>
<td>Cytosol</td>
</tr>
<tr>
<td></td>
<td>Carbonyl reductases</td>
<td>ER, cytosol, blood</td>
</tr>
</tbody>
</table>

1.2.2 Phase II reactions

Phase II reactions typically transfer relatively large, polar functional groups to a xenobiotic that substantially increase water solubility. These reactions often take place at nucleophilic sites installed or unmasked on the substrate molecule by phase I processes but it should be emphasised that many examples of direct phase II conjugation to unchanged administered drugs are known; i.e., phase I metabolism is not a necessary prerequisite for phase II transformations to take place. The products of these conjugation reactions are typically inactive and are readily excreted from the body.⁴⁴

The most common phase II reaction, glucuronidation, involves the transfer of glucuronic acid from uridine-5’-diphospho-α-D-glucuronic acid (UDPGA) to the substrate drug mediated by enzymes known as UDP-glucuronosyl transferases (UGT, Figure 1-5).⁴⁵-⁴⁷ This reaction occurs via an S₂N₂ mechanism catalysed by activation of both the substrate nucleophile and the UDPGA cofactor by proximal histidine and aspartate residues in the active site.⁴⁸, ⁴⁹ Substitution at the anomeric carbon results in an inversion of configuration from the α form in UDPGA to the β form in the glucuronide metabolite. Twenty-two different UGT isoforms have been identified and these can catalyze glucuronidation at a wide range of functional groups including alcohols, carboxylic acids, amines, thiols and, in some cases, even carbon nucleophiles.⁴⁷ These enzymes have a wide tissue distribution but are particularly concentrated in the liver.
Figure 1-5. UGT-mediated transfer of glucuronic acid from the UDPGA co-factor to a substrate RXH (X = O, N).

Figure 1-6. Sulfotransferase-catalysed transfer of sulfonate from the PAPS co-factor to a substrate RXH (X = O, N).
Sulfate conjugating enzymes, known as sulfotransferases catalyse the transfer of SO$_3$ to a substrate. These have a lower capacity than the UGTs however they are a prominent conjugation pathway. There are 13 known sulfotransferase enzymes that utilize 3’-phosphoadenosine 5’-phosphosulfate (PAPS) as the sulfate donor cofactor (Figure 1-6). The substrate specificity of sulfotransferases is more limited than for UGTs however they are more widely distributed. Sulfation at hydroxyl groups is most common however there are a number of examples of N-sulfation. Similar to glucuronidation, sulfation typically inactivates drug substrates.

Table 1-3. List of enzyme classes involved in common phase II metabolic reactions.

<table>
<thead>
<tr>
<th>Enzyme Class</th>
<th>Subcellular Location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uridine diphosphate glucuronosyltransferase</td>
<td>ER</td>
</tr>
<tr>
<td>Sulfotransferase</td>
<td>Cytosol</td>
</tr>
<tr>
<td>Glutathione S-transferase</td>
<td>ER, cytosol</td>
</tr>
<tr>
<td>N-Acetyltransferase</td>
<td>Mitochondria, ER</td>
</tr>
<tr>
<td>Amino acid conjugation enzymes (multi-enzyme system)</td>
<td>Cytosol</td>
</tr>
<tr>
<td>Methyltransferase</td>
<td>ER, cytosol, blood</td>
</tr>
</tbody>
</table>

Glutathione S-transferases (GSTs) have a prominent role in protecting the body from potentially damaging reactive metabolites. These enzymes catalyse the transfer of glutathione to the drug substrate via formation of a thioether link. This serves to detoxify xenobiotics or drug metabolites that possess an intrinsic chemical reactivity and prevent irreversible covalent modification of macromolecules. Distinct from the UGTs and sulfotransferases discussed above, the glutathione conjugation catalysed by GSTs often occurs at appreciable rates even in the absence of the enzyme. GSTs serve to increase the reaction rate still further by deprotonating the thiol group of GSH (pK$_a$ = 9) thereby activating it for the reaction. GSTs have a wide substrate specificity that is difficult to define in terms of specific functional groups but electrophilic sites on substrates are typically required. These include, but are not limited to, α,β-unsaturated carbonyl compounds, epoxides and quinones. In some cases, GSH conjugates undergo further enzymatic processing that cleaves the glutamate and glycine residues of GSH resulting in a xenobiotic-cysteine conjugate. This is sometimes referred to as the mercapturic acid pathway.
Figure 1-7. Glutathione-S-transferase catalysed conjugation of glutathione to an electrophilic α,β-unsaturated carbonyl compound.

1.3 Effects of metabolism

The implications of metabolism for the subject are far broader than simple chemical conversion, solubilisation and excretion. Indeed, metabolic reactions can mean the difference between a successful drug and an inert compound or a harmless chemical and a potent toxin. In drug discovery or xenobiotic toxicity studies, considerable effort is expended to determine the identity of the metabolites formed from a compound. Identification of metabolites can provide valuable insights that help to explain the pharmacological action of xenobiotics.

In some cases, metabolites formed from a substance will have their own biological effects that are different to those of the precursor compound. These effects may be adverse or beneficial. In the pharmaceutical arena, a compound that forms a biologically active substance upon metabolism is referred to as a ‘prodrug’ and this phenomenon is often exploited in the design of new medicines to optimise various parameters such as absorption or distribution. For example, a prominent class of antihypertensive (blood pressure-lowering) drugs are the angiotensin-converting enzyme (ACE) inhibitors. Most ACE inhibitors are short, peptide-like dicarboxylic acids such as enalaprilat (Figure 1-8). However, enalaprilat is very poorly absorbed from the gastrointestinal tract and only about 3% of an administered dose reaches systemic circulation. To increase absorption, enalaprilat is formulated as an ethyl ester derivative which shows between 53-74% absorption. Analysis of the inhibitory effects of the ester derivative against purified human ACE revealed that enalapril has essentially no inhibitory activity (IC$_{50}$, 50,000 nM). However, hepatic carboxylesterases hydrolyse the ester and release the diacid, which is ca. 3 orders of magnitude more potent (IC$_{50}$ = 34 nM, Figure 1-8). Owing to metabolism and the prodrug strategy, the overall yield of circulating active drug is approximately 40% and enalapril is one of the safest and most effective
antihypertensive agents in use today. It should also be noted that many compounds form metabolites that are thought to have toxic effects and this will be discussed in section 1.4 below.

**Figure 1-8.** Ester hydrolysis of enalapril forms the active ACE inhibitor enalaprilat.

In addition to the identity of the metabolites formed, the rate at which metabolism occurs is also a critical determinant of the biological outcome. For example, medicinal drugs generally have a well-defined plasma concentration range that is required for efficacy. If the concentration is too low, not enough active drug is present to exert the desired effect while a concentration that is too high may lead to adverse events through on-target or off-target receptors. Since metabolism aids in clearance of a xenobiotic from the body, it has a large effect on the plasma concentration and, in general, a greater rate of metabolism leads to more rapid clearance. To maintain the amount of active drug within the therapeutic range, the dosage and the frequency of administration must be matched with the rate of clearance (Figure 1-9). In this way, metabolism has a substantial role in determining how much and how often a drug needs to be administered and thereby has a direct impact on the experience of a patient with a medication. It is noteworthy that increasing the required frequency of dosing substantially reduces the percentage of patients who follow the administration directions and pharmaceutical development projects often try to ‘tune’ the metabolic profile of a drug to achieve clearance rates that are compatible with once-daily dosing.
Numerous strategies can be taken to optimise the clearance kinetics of a compound that may include alterations to the route of administration or formulation of the active compound in a slow-release capsule. However, chemical modifications to a drug structure that retain the receptor binding profile but improve kinetics are preferred in most cases. Many strategies are available to medicinal chemists that allow for optimisation of the metabolic profile. For example, Eldrup et al investigated a range of acylpyridine derivatives as potential inhibitors of soluble epoxide hydrolase (sEH).\(^6^0\) An initial lead, compound A in Figure 1-10, had a relatively short half-life \(t_{1/2}\) of 13 minutes in human liver microsomes (HLM) due to rapid hydroxylation of the pendant phenyl groups. A common approach to reduce metabolic activation of electron rich substituents is to incorporate fluorine atoms.\(^6^1\)-\(^6^3\) Fluorine has a similar Van der Waals radius to hydrogen and does not substantially change the steric environment but reduces electron density in adjacent motifs due to its high electronegativity. Incorporation of fluorine into compound A led to the identification of compound B. Measured \(t_{1/2}\) for this compound in HLM incubations was over 300 minutes indicating significant attenuation of metabolism while maintaining inhibitory activity at sEH.

**Figure 1-9.** Simulation of the plasma concentration of a drug during 14 consecutive doses. RHL = relative half-life.
Figure 1-10. Optimisation of the metabolic stability of sEH inhibitors through structural changes.

To complicate the accurate determination of the dosing schedules for a drug, individuals vary widely in their ability to metabolise xenobiotics.\textsuperscript{64, 65} This is most prominently observed for the CYP enzymes that show particularly large inter-individual variations in expression and activity.\textsuperscript{66} Many factors can influence the expression of these proteins including genetics,\textsuperscript{65, 67} gender,\textsuperscript{68} age,\textsuperscript{69} disease states\textsuperscript{70} and even psychological stress.\textsuperscript{71} Inter-individual differences in CYP expression of 5 to 10-fold are common\textsuperscript{20} and variations of up to $10^4$-fold have been identified in apparently healthy people.\textsuperscript{72, 73} These factors must be taken into account when administering a medication to a patient as the same dose may result in different plasma concentrations between individuals as a result of differences in metabolic rates. This is of particular concern for compounds where the therapeutic window (i.e., concentration range between effective and toxic doses) is small such as for many cytotoxic anti-cancer agents.\textsuperscript{70}

1.4 Idiosyncratic drug toxicity and reactive metabolites

Despite sustained research in the field of chemical toxicology and concerted effort during drug development and clinical testing, a small but significant number of approved medications have shown unacceptably high rates of adverse reactions that have no apparent cause.\textsuperscript{74} These toxic side-effects are broadly referred to as idiosyncratic adverse drug reactions (IADRs) and cannot be rationalised based on the known pharmacology of the compound.\textsuperscript{75} While IADRs are largely unpredictable and occur in a minority of patients, their effects can be severe and even life-threatening. Analyses of 548 small-molecule drugs approved between 1975 and 1999 reveal that 3\% have since been withdrawn from market as a result of adverse reactions while a further
10% were given new ‘black-box’ warning for hepatotoxicity which is the strictest possible caution applied to a medication by the United States Food and Drug Administration.\textsuperscript{76}

Clinical observations suggest that many IADRs are immune related.\textsuperscript{74, 77} A common characteristic of these reactions is a delay between the time of starting treatment with a given drug and the onset of toxicity.\textsuperscript{78} For example, troglitazone was once used to control hyperglycemia by reducing insulin resistance in patients with type II diabetes.\textsuperscript{79} However unacceptable incidence of hepatotoxicity was identified in patients after approximately 1 month of therapy and the risk of adverse reactions increased with the duration of treatment leading to the eventual withdrawal of the drug from sale.\textsuperscript{80, 81} Findings such as this support an immune-mediated mechanism as the delay between initiation of treatment and onset of toxicity would presumably reflect the time taken for proliferation of T cells and/or B cells specific for a given antigen. Immune responses can be empirically proven if the time taken for recurrence of the adverse effects is shortened upon rechallenge with the same drug\textsuperscript{82} although this test is rarely conducted in humans due to ethical concerns. Other symptoms that are commonly associated with IADR which are suggestive of an immune component include fever, skin rash and eosinophilia.\textsuperscript{75, 78, 83-87}

Unfortunately, animal models that accurately reflect human immune processes are generally not available\textsuperscript{88} and the detailed molecular events that underpin IADRs are often unclear. However, as the liver is the main site of metabolism for many drugs and is frequently involved in IADRs, mechanisms that involve biotransformation have long been suspected. In the 1940’s Miller and Miller observed that the potent hepatocarcinogen \(p\)-dimethylaminoazobenzene (better known as methyl yellow) bound covalently to cellular proteins in rats\textsuperscript{89, 90} and postulated that metabolites of the administered substance – rather than the substance itself – may be the primary toxicant. Clinical links began to emerge throughout the 1980’s with reports on the toxicity of sulfonamide antibacterial agents. For example, Shear and co-workers observed that patients experiencing sulfonamide toxicity showed markedly reduced capacity to acetylate xenobiotics than control patients.\textsuperscript{91} To explain this, it was proposed that decreased production of inert acetyl sulfonamides funnelled more of the administered dose through pathways that formed chemically reactive hydroxylamine metabolites that were known to cause cytotoxicity.\textsuperscript{92}

These observations provided a possible mechanistic link between drug administration, metabolism and immune reactions. In general, the immune system reacts predominantly to high
molecular weight compounds. Since the majority of marketed drugs have molecular weights less than 500 Da, immune effects are not a major component of the pharmacological outcome in most cases. However, formation of a reactive metabolite of a small-molecule drug that then bound covalently to an endogenous protein would produce a complex large enough for recognition by cells of the immune system. This is known as the ‘hapten hypothesis’ and the small-molecule drug is, in this case, referred to as a ‘hapten’.

In select cases, the link between covalent binding of a small molecule drug to proteins resulting in adverse immune reactions has been conclusively proven. For example, allergic reactions to penicillin-derived antibiotics are the most common of all IADRs and occur to some extent in approximately 10% of patients. In this case, human serum albumin covalently modified at specific lysine residues has been detected in the plasma of patients treated with β-lactam antibiotics and, furthermore, that this modified protein stimulated proliferation of lymphocytes from hypersensitive individuals. In light of this, predictive tests for penicillin-specific IgE antibodies have been developed to mitigate the risk of severe allergy that have ca. 97-99% accuracy. IgE-mediated immune reactions have now been proven for numerous small-molecule drugs with diverse structures including aspirin (analgesic, anti-inflammatory), trimethoprim (non-β-lactam antibiotic) and suxamethonium (muscle relaxant). Cases such as this highlight the potential for detailed analysis of the protein binding profile of a drug to deliver clinically valuable insights.

It is currently thought that formation and binding of chemically reactive metabolites (CRM) to cellular macromolecules is one of the primary events linking bioactivation to the onset of toxicity. Indeed an analysis of the covalent binding characteristics of 42 drugs revealed that the extent of protein binding is broadly correlated with the risk of IADRs. To this end, the identification of reactive metabolites, and the functional groups likely to result in their formation, has been a major driver of metabolism based research in recent decades and a vast literature has been assembled on the potential bioactivation processes operative for a wide range of organic functional groups. Common examples are listed in Table 1-4.
Table 1-4. List of some common functional groups that are known to form reactive metabolites in some instances.

<table>
<thead>
<tr>
<th>Functional Group</th>
<th>Reactive Metabolite</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>o- or p-hydroxyanilines</td>
<td>Quinone imine</td>
<td>Acetaminophen&lt;sup&gt;109&lt;/sup&gt;</td>
</tr>
<tr>
<td>o- or p- alkylphenols</td>
<td>o- or p- quinone methides</td>
<td>Eugenol&lt;sup&gt;110&lt;/sup&gt;</td>
</tr>
<tr>
<td>Nitroarenes</td>
<td>Nitrosoarenes</td>
<td>Chloramphenicol&lt;sup&gt;111&lt;/sup&gt;</td>
</tr>
<tr>
<td>Formamides</td>
<td>Isocyanates</td>
<td>N-methylformamide&lt;sup&gt;112&lt;/sup&gt;</td>
</tr>
<tr>
<td>Furans</td>
<td>α,β-Unsaturated dicarbonyls</td>
<td>Furosemide&lt;sup&gt;113&lt;/sup&gt;</td>
</tr>
<tr>
<td>Thiophene</td>
<td>Thiophene epoxide and S-oxide</td>
<td>Tienilic acid&lt;sup&gt;114&lt;/sup&gt;</td>
</tr>
<tr>
<td>Alkynes</td>
<td>Ketenes, oxirenes</td>
<td>Ethinyl estradiol&lt;sup&gt;115&lt;/sup&gt;</td>
</tr>
<tr>
<td>Aliphatic amines</td>
<td>Immonium ions</td>
<td>Haloperidol&lt;sup&gt;116&lt;/sup&gt;</td>
</tr>
<tr>
<td>Alkyl halides</td>
<td>Acyl halides</td>
<td>Halothane&lt;sup&gt;117&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

While generally rarer, some phase II metabolites are also considered to be chemically reactive. For example, Glucuronidation of carboxylic acid functional groups by UGTs forms acyl glucuronides which are considered one of the most important phase II reactive metabolites.<sup>118, 119</sup> These conjugates can undergo intramolecular acyl migration and epimerisation at the anomeric carbon forming a number of isomeric products. Covalent binding of these compounds to proteins may result from direct transacylation<sup>120</sup> or from glucuronide ring-opening and glycation (Figure 1-11).<sup>121, 122</sup> Formation of the Schiff base glycation product can occur from acetylation products at the 2, 3, or 4 positions however these reactions are reversible and production of stable adducts requires an Amadori rearrangement to the α-keto amine derivative which can only occur from the 3- and 4-O-β-glucuronide acetylation isomers. Although substantially less common, numerous polyaromatic hydrocarbons metabolites are known to be bioactivated by sulfotransferases as a result of sulphuric acid ester formation.<sup>123, 124</sup> The pronounced ability of HSO<sub>4</sub><sup>-</sup> to act as a leaving group is thought to result in generation of carbocationic species that have been demonstrated to alkylate proteins and DNA.<sup>124</sup> GSH conjugation is almost always a detoxifying process and examples of bioactivation by GSTs are very rare.<sup>125, 126</sup>
Figure 1-11. Formation, isomerisation and protein covalent binding of acyl glucuronide conjugates through transacylation and glycation. Stable glycation products can form from the 3- and 4-O-β-glucuronide.

Due to the potential for toxicity and the difficulty in predicting the outcomes of CRM formation, general practice in the drug development processes has been to minimise the formation of CRMs through chemical modifications to the drug substrate. Assays to determine the global level of protein binding by reactive metabolites are now routinely conducted on lead candidates and typically involve incubation of radiolabelled substrate with microsomal fractions or hepatocytes followed by liquid scintillation counting of washed
protein pellets. In cases where global levels of protein binding are deemed unacceptably high, chemical modifications to the drug structure are made with the aim to reduce the propensity for bioactivation.

As an example of this, consider sudoxicam and meloxicam (Figure 1-12), which differ by a single methylene group. Both drugs have anti-inflammatory properties but the clinical testing of sudoxicam was halted due to unacceptably high incidence of severe hepatotoxicity while adverse effects of meloxicam on the liver are extremely rare. Detailed studies of the metabolism of these two compounds revealed that epoxidation of the thiazole group followed by hydrolysis and ring scission produces glyoxal as well as a reactive acylthiourea which is susceptible to further oxidation and formation of electrophilic sulfenic and sulfinic acids. By contrast, a major metabolism pathway for meloxicam involved hydroxylation of the thiazole methyl group and further oxidation to the carboxylic acid that accounts for over 70% of metabolism products detected in the urine and faeces.
Figure 1-12. Metabolism pathways for sudoxicam and meloxicam.

Despite these associations and the apparent link between bioactivation and hepatotoxicity, there are many cases of CRM-protein binding that highlight the complexities inherent in this field. For example, when Roberts et al administered either acetaminophen (N-acetyl-p-aminophenol, APAP, also known as paracetamol) or the structural isomer N-acetyl-m-aminophenol (AMAP) to hamsters, similar levels of CRM-protein binding were observed but APAP treatment induced hepatotoxicity while AMAP did not. It is interesting to note that differences in the protein adduction profiles of APAP and AMAP have been identified using both radiographic and immunohistochemical analyses of protein fractions derived from various liver preparations. While it should be noted that the extent of AMAP bioactivation is species dependent, these studies indicated that protein adduction per se does not necessarily produce toxicity in all cases and any relationship between covalent binding and immune reactions may
be considerably more complicated. Due to the close structural similarity of APAP and AMAP, findings such as this have let to the ‘critical-protein hypothesis’ which asserts that the exact identity of the proteins adducted is a key determinant of the biological response and varies with even small changes in the structure of the reactive metabolite. It is interesting to note that histological studies to determine the relative nephrotoxicity of series’ of substituted quinols, catecols and o- and p- aminophenols in Wistar rats revealed that toxicity increased almost uniformly with decreasing oxidation potential despite highly similar structures.

Given that: 1) covalent adduction to proteins is thought to be one of the key events in the onset of toxicity for many drugs, 2) small numbers of specific proteins appear to be modified that can differ between drugs, and 3) the exact identity of these proteins may be an important determinant of the biological effects, numerous research groups have endeavoured to identify these specific molecular targets. To this end, efforts to catalogue the protein targets of reactive metabolites are underway. For example, Hanzlik and co-workers have set up a database of proteins reported to be modified by CRMs formed from a range of compounds.

It is thought that while different xenobiotics may modify various different proteins, there may be a common subset of these that are important determinants of adverse effects. It is hoped that identification of pivotal molecular targets may allow the toxicity of potential new drugs to be predicted in advance ultimately leading to safer medicines.

1.5 Identification of drug metabolites using mass spectrometry

The metabolism and pharmacokinetic profile of a drug is clearly one of the major determinants of the therapeutic potential of a new drug. Accordingly, elucidating the structures of metabolites is critical in pre-clinical development. As described above, advanced knowledge of the number and identity of metabolites formed may provide insights into how the drug structure could be modified to “tune” elimination kinetics or minimise the formation of reactive metabolites. However, identification of metabolites formed in cellular or animal models involves selectively detecting specific substances that may have low abundance from a matrix such as blood or urine that may contain thousands of endogenous compounds.

Prior to the 1980s, identification of metabolites was typically achieved by administration of radiolabelled (¹⁴C or ³H) drug to animals followed by measuring radioactive decay in urine and fecal samples. Where radioactivity was observed, attempts could be made to isolate pure samples of metabolites using techniques such as solvent extraction, solid-phase extraction and
thin-layer chromatography. While the introduction of cultured hepatocytes and microsomes in the early 1980s for use in metabolism studies greatly simplified the sample matrix, isolation of individual metabolites remained a challenging and slow process that frequently required many attempts.

Even when a metabolite was isolated, structural characterisation was not trivial. A key challenge was that only microgram quantities of metabolites were typically produced and isolated, which largely precluded routine characterisation using NMR. Gas chromatography coupled to mass spectrometry (GC-MS) was a mainstay of metabolite characterisation and allowed for analysis of mixtures of compounds by ionising molecules that eluted from a GC column. Ionisation was achieved by bombardment of analyte molecules with a stream of ~70 eV electrons (i.e. electron impact ionisation, EI) that typically resulted in the formation of molecular radical cations together with structurally diagnostic fragment ions. However, GC-MS was limited to the analysis of volatile compounds and chemical derivatisation of the sample prior to injection was typically required to mask highly polar functional groups. In addition, EI typically resulted in extensive ion fragmentation that could complicate interpretation of the data. While ‘softer’ ionisation techniques such as fast-atom bombardment could produce charged species without significant fragmentation, these solid-probe methods required highly pure samples and were incompatible with online chromatography. Ideally, an ionisation technique that could interface directly with HPLC and produce ions with minimal fragmentation was desired.

1.5.1 Instrumentation

A major advancement in mass spectrometry came with the introduction of electrospray ionisation (ESI), which operates at atmospheric pressure and generates ions with minimal in-source fragmentation directly from a solution containing the analyte. ESI elegantly solves the problems of sample introduction and fragmentation effectively by producing ions through gentle evaporation of charged solvent droplets containing the analyte (Figure 1-13). This breakthrough technology was rapidly interfaced with HPLC and incorporated into commercial instruments through the late 1980s and 1990s.
Many research groups recognised the potential of these instrumental developments and rapidly applied them to the analysis of drug metabolism. For example, Weidolf et al published one of the early reports of the application of liquid chromatography coupled to mass spectrometry (LC-MS) that used an ESI source to analyse the sulphate metabolites of a range of steroid hormones produced in urine samples and achieved detection limits of 10 pg of analyte on-column. Given that LC-MS is highly sensitive, does not require laborious chemical derivatization of samples and could separate compounds from complex biological matrices, it quickly became the predominant analytical method in drug metabolism studies.

Numerous types of mass spectrometers can be incorporated into LC-MS setups, and each offers different advantages and disadvantages. The key difference between these is the method by which ions are separated based on $m/z$ ratio, i.e. the type of mass analyser used. As mass spectrometry is the principal analytical technique used throughout this thesis, and several instrument types are employed, these are briefly described below. For more information and detailed descriptions of ionisation sources, mass analysers and detector types, many comprehensive books and tutorial reviews are available.

At the time LC-MS instruments were becoming commercially available, triple-quadrupole instruments were amongst the most common. These beam-type analysers consist of three quadrupole ion guides in series that each contains four parallel rods (Figure 1-14). Electrical potentials are applied to these rods that, in concert, function to stabilise or destabilise the trajectory of ions with a given $m/z$ value. Ions that have stable trajectories will be transmitted through the quadrupoles to the detector (blue trace in Figure 1-14) where an ion signal will be recorded. However, ions with unstable trajectories will collide with one of the rods and therefore not reach the detector (red trace in Figure 1-14). A spectrum is created by
systematically varying the potentials between these rods to sequentially stabilise the paths of ions of a given \( m/z \) value. In triple-quadrupole instruments, each of the three quadrupoles can be used in different ways to conduct a variety of experiments. For example, a particular ion can be mass-selected using the first quadrupole, subjected to collision-induced-dissociation in the second, while the third can be used to analyse the fragmentation products. Numerous other configurations can be used to achieve different analytical outcomes that will be discussed in the next section.

Figure 1-14. Schematic of a single-quadrupole mass analyser. Adapted from Paul 1990.\textsuperscript{162}

Ion-trap analysers are also very commonly used in LC-MS. These are ion packet-type mass analysers that use a system of rapidly alternating electrical potentials to physically confine ions with a range of \( m/z \) values in a defined special region.\textsuperscript{159, 160, 162} While different types of ion trap geometries exist, the linear ion trap (Figure 1-15) has been used for some of the experiments herein and will be described. This consists of 4 parallel hyperbolic rods that are divided into three sections. Once injected into the trap, application of DC potentials to the front (A) and back (C) electrodes produces an axial trapping field that confines a packet of ions to the central section (B) where radial trapping is achieved by applying \( R_f \) potential to the four central electrodes. Spectra are produced by scanning the \( R_f \) potential of the central section rods that causes ions of increasing \( m/z \) to be ejected from the trap through narrow slits in the axial rods where they strike an ion detector. Mass-selection of specific ions can be achieved by using this approach to expel all ions from the trap except for those of interest. Multistage mass spectrometry (MS\textsuperscript{n}) analyses of selected ions can be performed by application of an \( R_f \) ‘tickle voltage’ to the end electrodes that is calibrated to induce resonant excitation of the isolated ion. This increases the kinetic energy of the mass-selected ions causing dissociation through
multiple collisions with the helium bath (damping) gas that is maintained at a pressure of around 1mTorr in the trap. The products of fragmentation, as well as any remaining intact precursor, can be retained in the trap for further analysis. A significant advantage of ion-trap instruments is that many cycles of ion isolation, fragmentation and isolation of fragment product ions can be performed. This allows extensive data to be acquired on the fragmentation pathways of metabolites that can shed light on metabolite structure. A substantial drawback of ion trap analysers is that fragment ions formed in MS\textsuperscript{n} scans that are below approximately one-third of the precursor ion \( m/z \) are undetectable in common modes of operation. This effective low-mass cut-off arises as the potentials required for resonant excitation of the precursor cause destabilisation of the trajectories of small ions during fragmentation.

![Schematic of a linear ion trap](image)

**Figure 1-15.** Schematic of a linear ion trap. Adapted from Schwartz et al.\textsuperscript{160}

Modern commercial time-of-flight (TOF) analysers are conceptually simple and consist of three basic elements: 1) a ‘pusher’ plate, 2) a flight tube, and 3) a detector. TOF analysers differentiate between ions of varying \( m/z \) value by measuring the time taken to traverse a flight tube.\textsuperscript{158} A packet of ions is accelerated through a flight tube by application of an electric potential to the pusher plate, ions of different \( m/z \) values will travel with different velocities. Using simple Newtonian mechanics, the potential energy of a charged particle in a uniform electric field is given by Eq. 1-1. Upon acceleration, the electrical potential energy is converted to kinetic energy of the ion (Eq. 1-2).

\[
E_{el} = qV_s = zeV_s \quad 1-1
\]

\[
E_k = \frac{1}{2}mv^2 \quad 1-2
\]
Combining these equations, the velocity of the ion can be determined (Eq. 1-3). The time taken for the ion travelling at a constant velocity to traverse the fixed length of the flight tube, L, is given by \( t = L/v \). Arranging this for velocity and substituting into Eq. 1-3 gives Eq. 1-4.

\[
v = \sqrt{\frac{2zeV_s}{m}} \tag{1-3}
\]

\[
t = \sqrt{\frac{m}{z} \left( \frac{L^2}{2eV_s} \right)} \tag{1-4}
\]

Ions of lower mass or higher charge will attain higher velocities therefore reaching the end of the flight tube more rapidly. The time taken for ions to traverse the flight tube and strike the detector is measured and can then be related to \( m/z \) values. However, the resolution of this approach is hampered due to the distribution of velocities of individual particles in the ion packet at the beginning of the flight tube. To compensate for this, most modern TOF analysers are equipped with a ‘reflectron’ (also sometimes called an ‘ion mirror’, Figure 1-16). This consists of a series of electrodes placed at the end of the flight tube that repel the incoming ions. Faster moving ions (blue trace in Figure 1-16) will penetrate deeper into the field created by the ion mirror than slower moving ions (red trace in Figure 1-16) and therefore travel a slightly greater total length. A significant advantage of TOF instruments is that, in principle, the \( m/z \) range that can be surveyed is theoretically unlimited enabling analysis of large, intact entities such as viruses.\(^{163}\) Instruments with a single TOF stage cannot intrinsically perform MS/MS experiments (although post-source decay experiments can give similar data\(^{164-166}\)) however they are often combined with quadrupoles or additional TOF stages to add this capability.
Figure 1-16. Schematic of a single reflector time-of-flight mass analyser. The blue trace represents an ion with a higher initial velocity compared to the red trace.

The last types of mass analysers that will be discussed here are the Fourier transform ion cyclotron resonance (FT-ICR) and orbitrap. FTICR consists of an ion trapping cell that is located within a strong, uniform magnetic field. Ions are injected to the cell and orbit about the magnetic field owing to cyclotron motion with a frequency that is related to the $m/z$ ratio (Eq. 1-5). Here, $B_0$ is the magnetic field strength and $f_c$ is the cyclotron frequency. Clearly the cyclotron frequency of a given ion depends on both mass and charge.

$$f_c = \left(\frac{z}{m}\right) eB_0$$  \hspace{1cm} 1-5

To acquire a spectrum, a short, broadband radiofrequency ‘chirp’ pulse is applied that results in resonant excitation of ions which transition to coherent, higher radius orbits about the $z$-axis. Passage of ions near the surface of the detector plates causes induction of a small electrical current with each orbit. This ‘image current’ comprises a superposition of the currents induced by ions of different $m/z$ values that each have different cyclotron frequencies. Fourier transformation of the decaying image current then returns the $m/z$ values of the ions trapped in the cell. Fourier transform instruments have the highest resolving power of any mass spectrometers and have become a mainstay of metabolomics and proteomics owing to their ability to routinely achieve sub-ppm mass accuracies. One drawback is the relatively long time period required to acquire a spectrum (up to ~1 s), which may become problematic when rapid sampling of many compounds eluting from a column is required.
Orbitrap mass analysers are based on an electrostatic ion storage device described by Kindon in the 1920s.\textsuperscript{168} By contrast to FT-ICR, orbitrap analysers do not require R\textsubscript{f} or magnetic fields and ion trajectories are principally determined by the geometric shape of the trap components. Here, ions are electrostatically trapped between an outer barrel-shaped electrode and an inner spindle-like electrode.\textsuperscript{167, 169} The outer electrode is comprised of two axially identical halves separated by a small space that can be used for ion injection. Ion trajectories in the trap consist of both orbit about the central spindle electrode and axial oscillation in the \textit{z}-direction. Importantly, the frequency of ion oscillation in the \textit{z}-direction is directly related to \textit{m/z} ratio (Eq. 1-6) of an ion and is independent of radial position. A mass spectrum is derived by Fourier transform of the image current produced from the combined axial oscillation frequencies of ions and detected in the two halves of the outer electrode. Orbitrap analysers have a resolving power approaching that of FT-ICR but do not require superconducting magnets leading to substantial reductions in operating costs.

1.5.2 Metabolite identification

The coupling of mass spectrometry to HPLC allowed for rapid analysis of complicated biological samples that frequently contain more than a thousand different compounds and abundances that may vary by many orders of magnitude. However, when applying these technologies to xenobiotic metabolism, specific signals in LC-MS data that correspond to the metabolites of interest must somehow be differentiated from the large number of peaks corresponding to endogenous compounds.
Ideally, it is desirable to conduct LC-MS based metabolite identification studies in stages. Initially, full MS\(^1\) scans of a wide mass range should be conducted using instruments such as a Q-TOF that offers the ability to rapidly scan a very wide mass range with relatively high resolution. These initial experiments ensure that the majority of prominent components of the sample are ionised and detected. However, due to the presence of potentially very large numbers of endogenous compounds that may be significantly more abundant, signals that correspond to the drug metabolites of interest may be obscured or difficult to identify. Given that the structure of the administered molecule should be known, tables of common metabolic transformations\(^{170}\) can be used to compile a catalogue of possible metabolites which can then be used to search the data which may involve plotting extracted ion chromatograms (EICs) followed by comparison of case and control samples. To aid in this process, numerous software packages have been developed that aid the user in predicting the structure and formulae of various potential metabolites\(^{171}\).

If high-resolution mass spectrometry data is available, efforts can be made to distinguish metabolites based on their so-called ‘mass defect’.\(^{172-174}\) The mass defect of an atom, ion or molecule is defined as the difference between the particle’s exact and nominal (integer) masses. For example, the mass defect of acetylsalicylic acid (aspirin, \(\text{C}_9\text{H}_8\text{O}_4\)) that has a neutral monoisotopic mass of 180.04226 Da is +0.04226 Da or 42.26 mDa. The utility of mass defect arises from the observation that mass defects differ between atoms and generally decrease with increasing atomic number (Table 1-5). In general, small-molecule drugs have higher proportions of sulphur and halogens, greater degrees of unsaturation and lower proportions of hydrogen than natural compounds\(^{175}\) and hence a lower mass defect compared to many endogenous molecules\(^{176}\). Moreover, the majority of common biotransformations produce metabolites that are within ~50 mDa of the precursor compounds\(^{173, 177}\). For example, hydroxylation induces a change of -5 mDa, demethylation: -23 mDa, sulfation -43 mDa. Taking advantage of this, many mass spectrometry data analysis packages allow users to identify a list of signals in the data that are within defined mass defect ranges relative to the administered compounds. This can substantially reduce the number of signals from endogenous compounds that complicated the identification of drug metabolites.
Table 1-5. Mass defects of atoms that are common in organic small-molecule drugs.

<table>
<thead>
<tr>
<th>Element</th>
<th>Atomic number</th>
<th>Nominal mass (Da)</th>
<th>Mass defect (mDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>1</td>
<td>1</td>
<td>+7.825</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>7</td>
<td>14</td>
<td>+3.074</td>
</tr>
<tr>
<td>O</td>
<td>8</td>
<td>16</td>
<td>-5.085</td>
</tr>
<tr>
<td>F</td>
<td>9</td>
<td>19</td>
<td>-1.597</td>
</tr>
<tr>
<td>P</td>
<td>15</td>
<td>31</td>
<td>-26.238</td>
</tr>
<tr>
<td>S</td>
<td>16</td>
<td>32</td>
<td>-27.929</td>
</tr>
<tr>
<td>Cl</td>
<td>17</td>
<td>35</td>
<td>-31.147</td>
</tr>
</tbody>
</table>

Knowledge of the unimolecular ion fragmentation pathways that are operative for a xenobiotic (or are common to predicted metabolites) can aid in identifying biotransformation products in LC-MS data.\(^{155, 177-180}\) For example, often biotransformation of one structural moiety of a drug will leave the ion fragmentation of remote regions largely unaltered. Therefore, upon MS/MS, while the precursor ion \(m/z\) value of the metabolite may be different from the administered substance, the neutral mass lost in generation of fragment ions will be unchanged. Ions that exhibit this common mass loss can be identified by performing a ‘constant neutral loss scan’ (CNLS)\(^{181}\) on a triple-quadrupole instrument. Performing these analyses involves scanning both Q1 and Q3 while holding the difference in stable \(m/z\) trajectories between these at constant value corresponding to the expected neutral fragment loss of the chemical moiety of interest. For example, a common use of CNLS experiments is to identify compounds that have undergone conjugation with the antioxidant tripeptide glutathione (GSH). Glutathione conjugates (GSH-R) frequently exhibit neutral losses of 129 Da (Glutamic acid – H\(_2\)O) and/or 75 Da (glycine) that are common with a wide range of R groups.\(^{178, 182, 183}\) In this case, defining CNLS experiments to identify neutral losses of 129 and 75 Da are likely to detect GSH conjugates regardless of the total \(m/z\) of the metabolites. This approach is often used to detect other phase II metabolites such as glucuronic acid and sulphate conjugates. Similarly, precursor ion filtering (PIF) experiments allows all ions that form specific MS/MS fragmentation products to be readily identified regardless of the precursor mass. In this case, Q3 is set to transmit only ions of certain user-defined \(m/z\) values. This is useful for identifying the products of phase I metabolism wherein the metabolites often have appreciable structural similarity to the administered substance and give rise to common product ions.\(^{179}\)
Figure 1-18. Fragmentation of drug-GSH conjugates upon collision-induced dissociation. In [1,2-$^{13}$C$_2$, $^{15}$N-glycine]glutathione, the atoms marked with an * are enriched with $^{15}$N and $^{13}$C respectively.

Full-scan MS$^1$ analysis, PIF and CNLS aided by mass defect filtering and metabolite prediction software constitute a powerful set of analytical strategies that allow a broad range of drug metabolites to be detected from complex samples in many cases. However, the success of these “targeted” approaches depends on the ability of the analysis to predict the metabolic and ion fragmentation pathways operative for the analyte compound.

1.6 Stable isotope labelling in xenobiotic ADME studies

Xenobiotic compounds that are enriched with various stable isotopes such as $^2$H, $^{13}$C, $^{15}$N or $^{18}$O, have been used for many decades to aid in drug metabolism studies. Several properties of stable isotope labelled (SIL) drugs and xenobiotics make them desirable for absorption, distribution, metabolism and excretion (ADME) studies. For example, SIL compounds are essentially identical to their unlabelled counterparts allowing specific atoms to be traced through metabolic reactions. Moreover, these compounds are non-radioactive and have almost no deleterious effects on human health$^{184}$ and, as such require no specific handling precautions or administrative approvals. SIL compounds found early applications in unravelling xenobiotic metabolism pathways that appeared to cause organ toxicity.$^{185, 186}$ For example, chloroform was a major anaesthetic agent used in medical procedures from its introduction in the mid 1800s$^{187}$ to approximately the end of World War II when its use in medicine was discontinued due to numerous adverse effects and the introduction of safer alternatives.$^{188}$ Administration of
CHCl$_3$ could result in a dose-dependent cardio-, nephro- and hepatotoxicity$^{189}$ however it was unclear if chloroform itself, or a metabolite formed \textit{in vivo}, was responsible for the toxicity.

In 1978, Pohl and Krishma administered either CHCl$_3$ or CDCl$_3$ to rats$^{190}$ and found an 85% 24 h mortality rate for animals that were administered CHCl$_3$ at 4.98 mmol kg$^{-1}$. In contrast, no animal that received the same dose of CDCl$_3$ died during the study. At the time, it was known that oxidative metabolism of chloroform by P450 enzymes resulted in the highly reactive species phosgene.$^{191}$ As the first enzymatic step of this transformation is presumably abstraction of the sole hydrogen atom from chloroform, the rate of this reaction is expected to decrease upon replacement with deuterium. This is because the zero-point energy of a C-D bond is greater than that of a C-H bond in an otherwise identical environment leading to increased activation energy for the abstraction (i.e. the kinetic isotope effect). The finding that deuteration substantially lowered the toxicity of chloroform implied that hydrogen atom abstraction was a crucial mechanistic step that is entirely consistent with COCl$_2$ being the causative agent. Hydrogen/deuterium exchange has since been used to unravel pathways to toxicity of many different anaesthetics, drugs and industrial chemicals.$^{185,186}$ The observation that the KIE can make deuterium labelled drugs metabolically more inert has led to the concept of ‘leaving the label in the drug’ and deuterium labelled drugs are an active area of research.$^{192}$ Indeed, deutetrabenazine, the first drug to incorporate deuterium atoms as a means to attenuate metabolism was recently approved by the FDA for the treatment of disorders related to Huntington’s disease.$^{193}$

While atomic level investigation of metabolism and toxicity was a major early use of labelled compounds, these experiments were rapidly incorporated into a wide range of absorption and kinetics studies.$^{194-197}$ Here, the main utility of SIL compounds lies in the ability of mass spectrometry to distinguish between compounds that have the same structure but varying isotopic compositions. For example, bioavailability studies are routinely performed on experimental drugs in order to determine the fraction of an administered extravascular dose that reaches systemic circulation. The low bioavailability of drugs can be problematic in drug development and considerable effort can be expended to ensure that sufficient drug reaches the intended target for the substance to be efficacious. Bioavailability measurements are typically conducted in a 2-way crossover study where plasma concentrations measured after an \textit{i.v.} reference dose are compared to those determined following a test dose administered through a different route (typically oral) after a suitable washout period (Figure 1-19A).
An early example came in 1975 when Strong et al used an SIL approach to determine the bioavailability of the antiarrhythmic agent N-acetylprocinamide. Here, an unlabelled oral test dose was administered in the form of an immediate release capsule to each participant at the same time as an i.v. $^{13}$C-labeled N-acetylprocinamide reference dose. Blood samples extracted at various time intervals post dose were analysed and quantified via GC-MS. The labelled i.v. dose could be readily distinguished from the unlabelled oral dose by virtue of their differing isotopic composition. The peak areas of the individual isotopomers were independently used to construct concentration vs. time plots for both oral and i.v. doses which allowed simple determination of absolute oral bioavailability.

A substantial advantage of this approach is that the effects of intra-individual differences in absorption or metabolism between the test and reference doses are nullified as both can be assessed simultaneously. This reduces the number of participants required to gain meaningful data. Combined with the elimination of the second drug administration stage, many fewer participant hospital admissions are required resulting in a significant cost saving that largely offsets the added expense of isotopically enriched compounds. Indeed an economic analysis of these experiments found that reductions in cost (23%) and the number of participants required (36%) were possible by incorporation of SIL strategies into ADME studies. Since this first report, bioavailability assessments using SIL strategies have been reported for many drugs including verapamil, carbamazepine, ibuprofen, entacapone and metoprolol.
While the mass spectrometry based methods of metabolite identification such as mass defect filtering and ion fragmentation analysis are widely employed, the potential for these approaches to fail to identify true metabolites that do not conform to expected pathways has been recognised by many researchers. This has sparked interest in so-called ‘non-targeted’ approaches to metabolite detection that do not assume knowledge of the structure or fragmentation pathways for any metabolites formed.

An elegant solution to the problem of non-targeted metabolite identification was demonstrated in the early 1970’s by Knapp et al who simultaneously administered a 1:1 mixture of nortriptyline and its trideuterated counterpart to human subjects. They observed that metabolites of nortriptyline, which retained the isotopic label, eluted from the chromatographic column at the same time and with the same intensity as those derived from the unlabelled drug, but the masses of the labelled and unlabelled metabolites differed by a precisely known amount due to the isotopic label. These ‘twin-ions’ provide a signature in the LC-MS data sets unique to drug metabolites thereby allowing them to be distinguished from the multitudes of endogenous compounds without relying on predicted metabolism pathways or ion fragmentation processes. This ‘twin-ion’ method is referred to in some publications as ‘ion clustering’ and has been used to investigate the metabolism of numerous compounds including omeprazole, theophylline, atazanavir, potassium canrenoate, primaquine, fentanyl and caffeine.

It should be noted that the choice of deuterated analyte substrates as described by Knapp et al for twin-ion metabolism experiments should be carefully considered. In addition to the impacts on metabolic reactions described above for chloroform, the substantial kinetic isotope effects associated with the replacement of hydrogen with deuterium result in differences in the interaction of the labelled molecule with the stationary phases of chromatographic columns and can complicate the analysis of twin ions. With high-resolution chromatography, the retention times of deuterated and native analytes will differ and may hamper interpretation of the resulting data. To avoid these considerations, many SIL studies utilise $^{13}$C or $^{15}$N enrichment. As the mass differences between $^{13}$C and $^{12}$C or $^{15}$N and $^{14}$N are negligibly small, the enriched molecules are essentially identical to their native counterparts and any metabolic and chromatographic differences are insignificantly small.

The twin-ion studies discussed so far have all involved isotopic enrichment of the administered drug however many reports have appeared that employ SIL strategies in reactive metabolite
trapping experiments. In these studies, differentially labelled nucleophilic compounds are incorporated into an incubation mixture along with the unlabeled test compound. Reactive metabolites formed via bioactivation of the test compound will likely react with both heavy and light nucleophilic trapping agents in equal proportions and therefore produce the characteristic twin-ion signature in LC-MS or GC-MS data. Early demonstrations of this method by Mutlib et al added equimolar amounts of GSH and [1,2-\textsuperscript{13}C\textsubscript{2}, \textsuperscript{15}N-glycine]glutathione to microsomal incubations of paracetamol and other, related compounds. LC-MS analysis of the products were performed using CNLS to identify ions that gave rise to a m/z shift of 129 Da. For compounds that had reacted with the GSH and \textsuperscript{13}C\textsubscript{2}, \textsuperscript{15}N-GSH trapping agents, twin-ions eluting at the same time and with the same abundance should show a m/z difference of 3 while both exhibiting loss of 129 Da. Furthermore, the loss of glycine or \textsuperscript{13}C\textsubscript{2}, \textsuperscript{15}N glycin from the GSH moiety results in mass shifts of 75 Da or 78 Da respectively providing additional specificity for trapped metabolites (Figure 1-18). This allows confident differentiation between compounds that are true GSH-metabolite adducts from other compounds that coincidently fragment to lose the same nominal mass in MS/MS experiments.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure120.png}
\caption{Generation of stable products from reactive metabolites through the use of trapping agents. Adapted from Ma and Subramanian.\textsuperscript{215}}
\end{figure}
While glutathione is the most commonly used trapping agent for ‘soft’ electrophiles, some reactive metabolites such as aldehydes, acyl glucuronides and nitroso compounds may not form adducts with the central cysteine residue of GSH. Similar SIL techniques have been developed for analysis of these ‘hard’ electrophiles that employ native and labelled cyanide\textsuperscript{219, 220} or semicarbazide\textsuperscript{220} that allow for an enhanced range of CRMs to be trapped. To simplify the experimental process of reactive metabolite trapping, some efforts have been made to combine functional groups that are reactive toward either hard or soft electrophiles into a single molecule. For example, Huebert and co-workers synthesised a variant of GSH that replaced the glycine residue with lysine which has a side-chain primary amine group.\textsuperscript{221} This new peptide (glutamylcysteinyllysine, GSK) retained the \(\gamma\)-glutamyl group and therefore retained the characteristic loss of -129 Da produced in MS/MS analysis of many GSH conjugates. Moreover, the lysine was incorporated in either natural isotopic abundance or \(^{13}\text{C}_6, ^{15}\text{N}_2\) forms thereby providing a characteristic isotope separation of 8 Da for peaks corresponding to GSK adducts. In a twin-ion type experiment, these GSK peptides were demonstrated to form adducts with a range of compounds that formed both hard and soft electrophilic metabolites including \(p\)-cresol, and 2-methylfuran.

### 1.7 Aims of this thesis

The work presented in the following chapters describes the combination of LCMS and stable isotopic labelling with development of computer software to facilitate the non-targeted detection of xenobiotic metabolites.

Specifically, the aims of this thesis are to:

- To combine the twin-ion technique with automated computational algorithms to detect drug metabolites \textit{a priori} for the widely used analgesic paracetamol.
- Develop bioinformatic approaches and experimental workflows that allow for rapid and reliable identification of the protein targets of reactive metabolites.
- Apply these techniques to the detection of metabolites and protein adducts of a range of different model compounds.

To these ends, chemical synthesis, \textit{in vitro} and \textit{in vivo} bioassays, liquid chromatography – mass spectrometry and computer programming will be employed to (i) selectively detect the products arising from drug metabolism from endogenous compounds in LC/MS data for the compound paracetamol, (ii) determine the potential sites of protein covalent modification by
the electrophilic paracetamol metabolite NAPQI, (iii) develop an integrated workflow that employs MS$^1$ and MS$^2$ data, stable isotope labelling and automated peptide sequencing approaches to identify proteins covalently modified by reactive drug metabolites.
2 HIGH-RESOLUTION TWIN-ION METABOLITE EXTRACTION MASS SPECTROMETRY

Preface

The work presented in this chapter has been published in Analytical Chemistry:


In the following sections, the acetaminophen administration to Sprague-Dawley rats and subsequent blood sampling from was performed by Ms. Heather Daykin of The University of Melbourne and the HiTIME base code was written by Dr. Andrew Isaac and Dr. Bernard Pope.

2.1 Summary

This section describes an automated method that allows for the non-targeted detection of metabolites from complex blood samples that does not require prior knowledge of the metabolic pathways operative for a given molecule. Paracetamol (APAP) is one of the most widely used analgesic medications in the world and is known to undergo extensive metabolism. APAP and $^{13}$C$_6$ APAP were administered to rats and then blood was extracted and analysed by LC-MS. Numerous twin-ions were observed in APAP treatment groups that were absent in
control data and are thus likely to correspond to metabolites. To automate the process of finding these ions in LC-MS data, computer software called HiTIME was developed that can selectively extract data regions that fit the expected twin-ion signature. Nine doublets corresponding to twin-ions were detected by HiTIME analysis that were absent in control samples and these can be assigned to previously reported APAP metabolites. Extensive manual searching of the LC-MS data for all known metabolites of APAP did not reveal any further metabolites indicating that HiTIME detected all twin-ions present in the data. This method can be used to detect drug metabolites without knowledge of their identity.

2.2 Introduction
The metabolic fate of a drug has broad consequences for its pharmacological effects. Thus, significant effort is directed toward determining the identity, concentration and function of metabolic products throughout the drug development process. Recognising the importance of circulating metabolites, industry leaders and regulators began to establish a series of “best practice” principles\textsuperscript{222} in 2002 resulting in the release of guidelines for the safety testing of drug metabolites by the United States Food and Drug Administration\textsuperscript{223} and the International Conference on Harmonisation.\textsuperscript{224} A central tenet of the guidance is that a panel of toxicological studies be conducted on any metabolite identified in human plasma at disproportionately higher levels (>10% of drug-related material) than observed in any of the animal species used in preclinical evaluation. The FDA additionally notes many cases where “clinically relevant metabolites have not been identified”\textsuperscript{223} which has prompted pharmaceutical scientists to re-examine the various methods that can be employed to identify drug metabolites in a non-clinical setting.\textsuperscript{150, 225-227}

Liquid chromatography coupled to mass spectrometry (LCMS) has become a mainstay of drug metabolite profiling largely due to its ability to detect trace level compounds from extremely complicated biological fluids that contain thousands of endogenous compounds which are often more abundant.\textsuperscript{170, 177, 228, 229} For metabolites that are unknown and unexpected, it can be challenging to determine which of the thousands of ions detected in an LCMS experiment correspond to the downstream metabolites of precursor compounds. This “needle-in-a-haystack” problem requires the development of non-targeted methods to rapidly detect the products of compounds in complex mixtures, at trace levels, and without prior knowledge about the identity of the metabolites.
One non-targeted method to detect specific chemicals in LC-MS data is to simultaneously administer equal parts of natural abundance and isotopically enriched xenobiotic which produces a pair of co-eluting peaks that have the same abundance and are separated by an \( m/z \) interval corresponding to the mass of the isotopic label. This ‘twin-ion’ strategy was first described by Knapp et al and has been used to study the metabolism of many compounds including drugs and insecticides.\(^\text{186, 194, 206, 230-232}\) Despite the obvious practical advantages of non-targeted twin-ion metabolite detection, the use of this method has been limited\(^\text{186, 233}\) partly because (i) the sensitivity and resolution of many types of common mass spectrometers were insufficient to detect low abundance metabolite ions, and (ii) it was challenging to detect twin-ion metabolite signatures in large data sets.\(^\text{144}\)

Given the significant advances in mass spectrometer performance, there has been renewed interest in using the twin ion method for a range of biological applications including the detection of trace level metabolites of insecticides in susceptible insects \textit{in vivo}\(^\text{232}\), the determination of pharmacokinetic parameters such as bioavailability\(^\text{194}\) and the elucidation of metabolic pathways and mechanisms.\(^\text{186}\) Due to the complexity of high-resolution LCMS data sets, it is important that twin-ion metabolite detection is automated to remove the laborious and inefficient task of manually searching through such datasets.

Here, we demonstrate a non-targeted approach to detect drug metabolites that combines the twin ion method with automated data-mining software for non-targeted detection of drug metabolites. We have selected acetaminophen (APAP), a clinically relevant and widely used analgesic, to test this approach because it has a metabolic profile that is well established, has a mode of action that is not well understood, and can form reactive metabolites that can cause adverse side effects.

### 2.3 Results and discussion

Either a mixture containing equal parts of APAP and \(^{13}\text{C}_6\)-APAP or vehicle was administered to rats via \textit{i.v.} infusion at a dose of 10 mg kg\(^{-1}\), which is comparable to therapeutic APAP does used in humans. Blood samples were drawn either immediately prior to drug administration, or 30 minutes post-dose. LC-MS analysis of plasma extracts using the Q-TOF mass spectrometer (described in the experimental section 8.1.1.1) produced complex data sets. For instance, using the standard feature detection algorithms in the Agilent MassHunter software 4957 peaks were identified (average of 3 replicates) using an intensity threshold of 500 counts.
For negative ion mode, 2128 peaks are found. One strategy to identify unknown/unexpected metabolite ions is to use a mass defect filter,\textsuperscript{173} which can detect metabolites based on well-defined shifts in the differences between the accurate \textit{m/z} values and the integer \textit{m/z} values. By use of a 50 mDa filter, which is wide enough to cover the majority of common phase I and II metabolic transformations,\textsuperscript{173} an average of 1567 peaks for positive ion mode and 1460 in negative were detected. Clearly the complexity of these data presents a challenge and APAP metabolites are difficult to identify.

In these experiments, it is known that any true metabolite formed and detected by the mass spectrometer is likely to appear as a twin-ion in the LC-MS data. The assignment of any APAP metabolites would therefore be greatly enhanced if these twin-ions could be selectively detected. We therefore sought to develop computer software that can detect twin-ion signatures in raw, profile LCMS data.
Figure 2-1. A) Profile LCMS data mass spectrometry data in which doublet peaks assigned to APAP [M+H]+ are shown. B) Representative ‘ideal’ twin-ion peaks that were used to weight the data. C) Output data after weighting centring on the light isotope. D) Visualisation of the results from HiTIME weighting as a two-dimensional heat map. RA and RT correspond to relative abundance and retention time, respectively.

2.3.1 Automated twin-ion metabolite extraction algorithm.

Twin-ion pairs are defined as having: (i) identical retentions times for the heavy and light isotopologues, (ii) identical abundances for the heavy and light isotopologues, and (iii) a difference in m/z values that corresponds to the known mass difference between heavy and light isotopologues. For example, the experimental twin-ion signal assigned to the elution of unchanged APAP is shown in Figure 2-1A. Two peaks of almost the same intensity are observed that elute at 13.1 mins. The accurate mass difference between the ion at m/z 152.0718 and m/z 158.0919 is m/z 6.0201 which corresponds exactly to the theoretical mass difference between 12C6 and 13C6.

To take advantage of the full profile data produced by LC-MS instruments, a three-dimensional shape fitting approach was devised for detecting twin-ions. In this methodology, an idealised
twin ion shape is fit to each region of the LCMS data. The ideal shape consists of two three-dimensional Gaussian functions (retention time, \( m/z \) value, and ion abundance are the \( x \), \( y \) and \( z \) coordinates, respectively) that are centred at the same retention time (\( rt \)), but the \( m/z \) values differ by the accurate mass of the isotope label; i.e., \((rt,mz)\) and \((rt,mz+\Delta mz)\), where \(\Delta mz\) is the mass difference between the heavy and light isotopes (Figure 2-2). The ratio between ion abundances (\(z\)-dimension) of the two Gaussian functions are equal to the relative proportions of heavy and light drug that were administered. A summary of the parameters used is given in Appendix Table 9-1 and the algorithm is henceforth referred to as high-resolution twin-ion metabolite extraction (HITIME).

**Figure 2-2.** Schematics showing the parameters employed for fitting the ideal twin-ion shape to raw LCMS data in the A) \(m/z\) and B) retention time dimension.

The data scoring method iteratively applies the ideal shape at each data point in the dataset and assesses the goodness-of-fit. For example, the APAP/\([^{13}\text{C}_6]\)-APAP \([\text{M+H}]^+\) twin-ions (Figure 2-1A) are first compared to the ideal shape shown in Figure 2-1B. This results in a goodness-of-fit score for each mass spectral peak for a given retention time in the entire LCMS data file, which is determined by how closely the twin-ion model matches the data compared to models of either the heavy or the light isotope peaks alone. A more detailed discussion of how twin-ion scores are calculated and assigned is given in Appendix section 9.3.1 on page 200. Once the data have been processed, the scored peak profiles are centered on the retention time and \(m/z\) values of the unlabelled ion (Figure 2-1C). These results can then be visualised by the analyst as a heat map that depicts both mass spectral and chromatographic domains with confidence scores indicated by a colour gradient (Figure 2-1D).
Importantly, the HiTIME algorithm does not consider absolute peak intensity when assigning confidence scores to a data point; i.e. a very low abundance twin-ion doublet will be assigned the same score as a high abundance doublet regardless of their differing intensities (in the absence of other factors). This design feature was selected to facilitate the detection of ions observed in trace abundances. Histograms depicting the distribution of scores for a typical analysis presented in this chapter are shown in Figure 2-3. Approximately, 80% of the data in this small section of the raw files (representing tens of thousands of data points) are assigned a very low confidence scores (<2). On this basis, data points assigned a score of <2 are considered to have negligible correlation to the twin-ion model and are excluded from the plots shown here. In these experiments, a HiTIME score of >5 is strongly indicative of a twin ion hit (see below).

Figure 2-3. Histograms showing the distribution of scores obtained from the HiTIME algorithm for APAP-treated rats for, A) negative and B) positive ion data.

2.3.2 HiTIME analysis of acetaminophen metabolism in the rat
Applying this algorithm to the positive and negative ion LCMS data from the blood plasma of acetaminophen treated rats generates the heat maps shown in Figure 2-4. In these plots, confidence in the match between the twin-ion model and the raw profile data is given on a scale of 2 (low confidence) to 20 (high confidence). White space in the plots indicates a score of 2 (or less) meaning that essentially no correlation is observed between the hypothetical twin ion model and the raw data (>99.7% of all data points). This indicates that the majority of the data does not match the expected APAP metabolite twin-ion profile and has been removed by the HiTIME algorithm.
Figure 2-4. Positive (left column) and negative (right column) ion LCMS data from rats treated with A) and D) 1:1 APAP:\[^{13}C_6\]-APAP 10 mg kg\(^{-1}\) and, B) and E) prior to APAP administration scored by correlation with a hypothetical twin-ion model using the HiTIME algorithm. C) and F) APAP treatment data following subtraction of vehicle control. To aid in visualisation of the results, only a small segment that contains the detected twin-ion ‘hits’ is displayed.
Numerous ‘bright spots’ are also seen in the heat map shown in Figure 2-4A and D that are assigned scores of >10 indicating a strong correlation between the data in these regions and the model twin-ion signature. These signals likely arise from the elution of isotopically labelled drug metabolites. However, endogenous compounds unrelated to one another may by chance give an LCMS profile similar to that of the twin-ion signature. Endogenous compounds can be discounted by comparing the treatment sample against the control sample shown in Figure 2-4B and E for positive and negative ion mode data respectively. Indeed, corresponding bright spots are absent in the heat map generated from control LCMS experiments on blood drawn prior to APAP dosing. The data points that do appear in the control heat map are assigned only very low scores (<5) and are suspected to arise from endogenous compounds (or even chemical noise) that slightly overlap with the hypothetical twin-ion signature. As the HiTIME algorithm has no absolute intensity bias, these small fluctuations in the mass spectral noise will be assigned positive scores. To remove these interferences, the control data were subtracted from those produced following APAP treatment and the results are shown in Figure 2-4C and F. This process removes much of the low-level noise while leaving the high scoring data, thereby providing a sharper contrast between twin-ion signals assigned with high confidence and inevitable background noise.

Table 2-1. Non-covalent adducts of twin-ion metabolites detected by the HiTIME algorithm for negative ion LCMS data.

<table>
<thead>
<tr>
<th>Retention Time (min)</th>
<th>m/z pair</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.60</td>
<td>245.9614/251.9816</td>
<td>PAP-SO$_3$ –H + NaCl</td>
</tr>
<tr>
<td>10.60</td>
<td>255.9906/262.0098</td>
<td>PAP-SO$_3$ –H + NaHCO$_2$</td>
</tr>
<tr>
<td>10.60</td>
<td>272.9795/279.0000</td>
<td>PAP-SO$_3$ –H + NaNO$_3$</td>
</tr>
<tr>
<td>11.57</td>
<td>348.0685/354.0948</td>
<td>APAP-GLC -2H + Na</td>
</tr>
<tr>
<td>11.57</td>
<td>362.0648/368.0851</td>
<td>APAP-GLC + Cl</td>
</tr>
<tr>
<td>11.57</td>
<td>384.0472/390.0678</td>
<td>APAP-GLC –H + NaCl</td>
</tr>
<tr>
<td>11.57</td>
<td>411.0656/147.0857</td>
<td>APAP-GLC –H + NaNO$_3$</td>
</tr>
<tr>
<td>12.70</td>
<td>298.0006/304.0196</td>
<td>APAP-SO$_3$ – H + NaHCO$_2$</td>
</tr>
</tbody>
</table>

For some of the highly scoring regions in Figure 2-4C and F, similarly highly scoring data regions are seen at the same retention time but at a higher m/z value. These are not observed in the control data and, based on high-resolution mass measurements and isotope distributions, likely arise from simple non-covalent adducts between twin-ion metabolites and inorganic salts.
(e.g., [M+Na]^+ and [M+H+NaHCO_3]^+) or other endogenous products that are commonly formed in the ESI process. Assignments for some of these are provided in Table 2-1. As it is highly unlikely that unrelated endogenous compounds would produce this phenomenon simply by chance (based on the control data; Figure 2-4B and E), these provide additional confirmation that a metabolite was detected.

Analysing the subtracted heat maps, a total of 9 highly scoring regions were detected that had a unique retention time and a score greater than 5 in either positive or negative background subtracted data sets (Figure 2-4C and F). The 9 regions that were assigned a high score by the HiTIME algorithm were manually investigated by plotting extracted ion chromatograms (EICs, a plot of the ion signal at a given m/z value vs. retention time) for both the heavy and light isotopes of potential metabolites using m/z values obtained for ions that were assigned a high confidence value in Figure 2-4C and F. In this way, the presence of APAP metabolites could be readily verified by searching for regions where the EIC traces for heavy and light isotopes of suspected metabolites overlap.

For example, in Figure 2-5A the EIC traces of the m/z 230/236 ion pair, which is assigned to APAP-SO_3 and ^13C_6-APAP-SO_3, respectively, are shown. The EIC traces of each ion overlap, indicating that these ions elute with the same retention time. The accurate m/z difference between ^13C_6-APAP-SO_3 and APAP-SO_3 is 6.0202, which is very close to the ideal value of 6.0201, and the intensities of these ions are essentially the same (Figure 2-5B). In addition, if the EICs are plotted for the control data (i.e., no APAP), the twin ion signatures are not observed (Figure 2-5C and D).
Figure 2-5. Negative ion LC-MS analysis of blood plasma extracts from rats administered 1:1 APAP and $^{13}$C$_6$-APAP. A) EICs of $m/z$ 230 and $m/z$ 236 and B) mass spectrum corresponding to the HiTIME hit at 12.70 min and $m/z$ 230 assigned to the elution of APAP-SO$_3$. C) EICs and D) mass spectrum from the same data region of control data. Note the difference in y-axis intensity scales between treatment and control data.
Figure 2-6. Negative ion LC-MS analysis of blood plasma extracts from rats administered 1:1 APAP and $^{13}$C$_6$-APAP. A) EICs and B) mass spectrum corresponding to the HiTIME hit at 12.86 min and m/z 260 assigned to the elution of APAP-OMe-SO$_3$. C) EICs and D) mass spectrum from the same data region of control data.

To demonstrate for a less abundant metabolite, the twin ion signature assigned to the rarely reported APAP metabolite, 3-methoxyacetaminophen-4-sulphate (APAP-OMe–SO$_3$), is detected at m/z 260 and m/z 266 in the negative ion mode with EIC traces of the heavy and light isotopes showing the overlapping peak that is distinctive of a twin-ion ‘hit’ (Figure 2-6A). Importantly, the mass spectral signal intensity for this metabolite is ca. 1000-fold less than that of APAP-SO$_3$ discussed above. In this case the mass spectral peak for the unlabelled isotope of APAP-OMe-SO$_3$ (m/z 260.0238, marked in Figure 2-6B with a red arrow) appears immediately adjacent to an endogenous compound at m/z 260.1318. The twin ion doublets are not observed in the control data (Figure 2-6C and D).

In total, 9 acetaminophen metabolites were detected by HiTIME and confirmed by plotting the EICs of the heavy and light isotopologues. The APAP metabolites that were detected are listed...
in Table 2-2 and their nomenclature and likely formation pathways are shown in Figure 2-7. The most abundant twin-ions detected at m/z 230 and 326 are assigned to phase II sulfate and glucuronic acid conjugates respectively which are known to be the major APAP metabolites formed at therapeutic doses.\textsuperscript{234, 235} Additionally, m/z 457 is attributed to a glutathione conjugate of APAP likely arising through the electrophilic intermediate NAPQI. Further metabolism of this compound could produce the ions at m/z 271 and 313 which are assigned as cysteine and N-acetylcysteine conjugates of APAP respectively. Interestingly, the ions at m/z 246 and 260 likely correspond to APAP-OH-SO\textsubscript{3} and APAP-OMe-SO\textsubscript{3} which may be formed via either sulfate conjugation or methylation followed by sulfate conjugation respectively of the unobserved intermediate APAP-OH. Two partially resolved peaks in EIC traces of m/z 188 can be assigned to sulfate conjugates of p-aminophenol formed via deacetylation of APAP. The two apparent isomers are likely the result of regioisomeric sulfation forming either N- or O-aminophenol sulfate.

Table 2-2. List of APAP metabolites detected in LCMS datasets by the HiTIME algorithm.

<table>
<thead>
<tr>
<th>Compound</th>
<th>RT (min)</th>
<th>Ion Polarity</th>
<th>Peak Area\textsuperscript{a}</th>
<th>Measured m/z (\textsuperscript{13}C\textsubscript{6}-labelled)\textsuperscript{b}</th>
<th>Δm/z \textsuperscript{b,c}</th>
<th>HiTIME score</th>
</tr>
</thead>
<tbody>
<tr>
<td>APAP</td>
<td>13.15</td>
<td>+</td>
<td>13942546</td>
<td>152.0718</td>
<td>6.0201</td>
<td>9.4</td>
</tr>
<tr>
<td>PAP-SO\textsubscript{3}\textsuperscript{d}</td>
<td>10.60</td>
<td>-</td>
<td>1288950</td>
<td>188.0027</td>
<td>6.0199</td>
<td>11.1</td>
</tr>
<tr>
<td>APAP-SO\textsubscript{3}</td>
<td>12.70</td>
<td>-</td>
<td>94083840</td>
<td>230.0130</td>
<td>6.0202</td>
<td>17.7</td>
</tr>
<tr>
<td>APAP-OH-SO\textsubscript{3}</td>
<td>12.67</td>
<td>-</td>
<td>53731</td>
<td>246.0081</td>
<td>6.0167</td>
<td>7.7</td>
</tr>
<tr>
<td>APAP-OMe-SO\textsubscript{3}</td>
<td>12.86</td>
<td>-</td>
<td>53850</td>
<td>260.0238</td>
<td>6.0230</td>
<td>7.8</td>
</tr>
<tr>
<td>APAP-CYS</td>
<td>11.97</td>
<td>+</td>
<td>63186</td>
<td>271.0690</td>
<td>6.0206</td>
<td>9.4</td>
</tr>
<tr>
<td>APAP-NAC</td>
<td>13.08</td>
<td>+</td>
<td>74847</td>
<td>313.0890</td>
<td>6.0199</td>
<td>11.8</td>
</tr>
<tr>
<td>APAP-GLC</td>
<td>11.60</td>
<td>-</td>
<td>2776708</td>
<td>326.0886</td>
<td>6.0202</td>
<td>18.3</td>
</tr>
<tr>
<td>APAP-GSH\textsuperscript{e}</td>
<td>12.34</td>
<td>+</td>
<td>42206</td>
<td>457.1388</td>
<td>6.0242</td>
<td>3.6</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Values are the average of three replicates. \textsuperscript{b}Exemplary values from a single analysis. \textsuperscript{c}Difference in m/z values of \textsuperscript{13}C\textsubscript{6} and \textsuperscript{12}C\textsubscript{6} labelled ions. \textsuperscript{d}Combined ion abundances for both N- and O- sulfate aminophenol metabolites. \textsuperscript{e}m/z values from baseline subtracted mass spectra.
Figure 2-7. Structures, nomenclature and transformation steps involved in the formation of APAP metabolites that were detected in this study. Structures in parentheses are metabolic intermediates that were not directly observed in these experiments.

The twin ion signature that corresponds to APAP-OMe-SO\(_3\) (Figure 2-6) is representative of a low abundance metabolite ion that was detected by use of HiTIME. Of the 9 acetaminophen metabolites detected in this study, 5 were observed at \(ca. 10^3\) ion counts, which is close to the mass-spectral noise level of \(ca. 500\) ion counts. If the identity of these metabolites were not known in advance, it would be challenging to detect and assign these low abundance twin-ion signals. Overall, these data indicate that HiTIME can detect twin-ion signatures even at relatively low \(S/N\) values.

To rule out the possibility of “false-negatives” being assigned by the HiTIME algorithm (i.e., low confidence scores assigned to actual acetaminophen metabolites showing the twin-ion
signature), a list of all known APAP and 4-aminophenol metabolites was assembled from literature sources (31 metabolites in total) along with their corresponding $m/z$ values for heavy and light isotopes. EIC traces were plotted for each potential twin-ion pair not already assigned by the HiTIME algorithm. In no case was a region of clear overlap between the EIC traces of heavy and light isotopologues observed, including after subtraction of control EIC plots derived from identical treatment of blood plasma from rats that had been administered vehicle control only. These data provide confidence that HiTIME has accurately assigned all known metabolites of APAP that were present in the plasma samples, ionized, resolved and detected.

* Many of the 31 APAP metabolites reported in the literature were detected by use of hepatotoxic doses of > 200 mg kg$^{-1}$; i.e., at significantly higher doses than were employed in our experiments (10 mg kg$^{-1}$).
Figure 2-8. Negative ion LCMS data from rats treated with 1:1 APAP-[\textsuperscript{13}C\textsubscript{6}]-APAP 10 mg kg\textsuperscript{-1} weighted by correlation with hypothetical twin-ion model using doublet spacing of A) 5.0167, B) 4.0134, C) 3.0101 and D) 2.0067 m/z which correspond to the loss of 1, 2, 3 and 4 \textsuperscript{13}C labels respectively.

To account for the loss of any \textsuperscript{13}C labels through metabolic transformations involving scission of the phenyl ring (which would change the difference in m/z values between the labelled and unlabelled ions), the data were processed with doublet m/z spacings corresponding to 5, 4, 3 and 2 \textsuperscript{13}C labels. The resulting control-subtracted heat maps are shown in Figure 2-8. In general, these data are sparse and few data points attract significant HiTIME scores. Moderate scores assigned in the heat maps in Figure 2-8 largely overlap with the twin-ions already assigned in Figure 2-4 and can be readily attributed to peaks arising from isotopic impurities in the APAP and \textsuperscript{13}C\textsubscript{6} APAP samples. For example, the A+2 peak of APAP-SO\textsubscript{3} and the A+1 peak of \textsuperscript{13}C\textsubscript{6} APAP-SO\textsubscript{3} are of approximately the same intensity and elute at the same retention time but have a spacing of 3.0184 m/z rather than the 6.0201 m/z which is observed for the isotopically pure APAP-SO\textsubscript{3} and \textsuperscript{13}C\textsubscript{6} APAP-SO\textsubscript{3} as shown by the two arrows in the inset of Figure 2-9.
Thus, the region at 12.7 min and 232 \( m/z \) in Figure 2-8C, where the data were scored with a doublet spacing of 3.0101 \( m/z \), is assigned a moderate HiTIME weighted value. Similar effects are observed at the same retention time but for higher \( m/z \) values, and these arise from the complicated isotope distribution of the sodium bound dimer, \([2M-2H+Na]^+\) shown in the Figure 2-9 inset. No new highly scoring regions that could not be assigned to either chance coelution of unrelated compounds or artefacts arising from isotopic impurities in the light and heavy acetaminophen precursors were found. Overall, these data indicate that HiTIME can be used to accurately assign all metabolites of APAP that were present in the plasma samples, ionized, resolved, detected, and have been previously reported.

![Figure 2-9](image)

**Figure 2-9.** Negative ion mass spectrum at 12.7 minutes showing twin-ions assigned to APAP-SO\(_3\). Insets show expansions of the regions around \([M-H]^−\) and \([2M-2H+Na]^+\).

While the abundances of the 9 metabolites detected here vary considerably in the raw LCMS data, they are assigned similar confidence scores after weighting with the HiTIME algorithm. For example, APAP-CYS is detected with only a very low abundance (integrated peak area = 63186, Table 2-2) and is assigned a weighted intensity of 9.4 by the HiTIME algorithm. By contrast the precursor compound, APAP, was detected in high abundance (peak area = 13942546) that is \( ca. \) 200-fold greater than that for APAP-CYS and is assigned the same weighted intensity value of 9.4. These data demonstrate that the HiTIME algorithm can provide a non-targeted and unbiased assessment of twin-ion signatures present in a sample that is largely independent of absolute signal intensity over a wide range of ion abundances.
2.4 Comparison to other methods

Several computer algorithms have been developed to detect twin-ion metabolite doublets in LCMS data. The most advanced of these is MetExtract developed by Schumacher and co-workers. MetExtract initially considers each data point in each mass spectrum to be the unlabelled metabolite and the algorithm then searches for a theoretical peak at a higher m/z containing a defined number of isotopic labels. A twin-ion “hit” is assigned if this peak is found at a comparable intensity to the unlabelled peak. This differs from HiTIME as it does not initially account for the evolution of a peak profile with increasing retention time.

To benchmark our HiTIME algorithm, we compared our results to those generated by processing the same dataset with MetExtract. Using the parameters given in Appendix Table 9-4, MetExtract correctly detected 5 of the 9 twin-ion signatures known to be present in the data set. The metabolites that were not detected were APAP-GSH, APAP-CYS, APAP-OH-SO₃ and APAP-OMe-SO₃. By use of an m/z tolerance of 99 ppm (the maximum allowed value) and the ion doublet intensity ratio threshold was set to 0.5, one additional metabolite (APAP-CYS) was detected. Although the precise reason for this discrepancy is unclear, two points are worth noting. Firstly, while HiTIME is optimised to use raw, profile LC-MS data for analysis, MetExtract operates using ‘centroided’ data. Centroiding simplifies the computational complexity of the problem but has the effect of discarding most of the data describing a peak, which could conceivably be detrimental to twin-ion detection. Additionally, MetExtract is essentially a classification algorithm that decides whether a given datum is either a twin-ion or not. HiTIME, by contrast, does not classify peaks but rather weights a given data point by correlation to a twin-ion signature. All data points in an LC-MS run are given a weight that can then be interrogated by an expert user, which may be advantageous for low abundance or otherwise difficult to assign peaks.

* As MetExtract is optimized for the interpretation of LCMS data in the ‘centroided’ format, our profile data were centroided with the commonly used application MSConvert and these data were manually inspected to ensure that the signals for these metabolites were preserved throughout the data formatting process.
2.5 Availability and compatibility

HiTIME was written in the Python (v2.7) programming language and has been made available free of charge under the 3-clause BSD license. The code is hosted on GitHub which is a platform commonly used by programmers to manage software projects and distribute source code.

The source code can be accessed at: https://github.com/bjpop/HiTIME

Here, and throughout this thesis, the Python version of HTIME accessible through the link above was used to analyse LC-MS data. Python is an interpreted, high-level programming language. As a result, it is generally considered to be relatively ‘slow’ for demanding numerical calculations compared to other languages. Despite this, it is frequently used to build prototypes of software due to ease of coding. As analysis with this python software is a computationally intensive task, the processing was accelerated using distributed computing in most cases (i.e. the processing is distributed over many computers).

As not all users may have access to large computing clusters, it was desirable to increase the speed of analysis so that processing could be carried out on common desktop computer hardware. To accomplish this, the HiTIME source code has been translated into C++ which offers a >10-fold reduction in processing time making analysis of moderately sized datasets realistic on small desktop computers or even laptops.*

The source code for HiTIME C++ can be accessed at: https://github.com/bjpop/HiTIME-CPP

Both the Python and C++ HiTIME versions can accept input LC-MS data files and write analysis results in the mzML format which has become a community standard open-source data.251, 252 This enables HiTIME to be used in-line with other software that accepts mzML files such as the OpenMS library.253 OpenMS is a C++ function library containing routines for many common data processing tasks such as feature detection, baseline correction and data smoothing. The use of mzML files in HiTIME ensures that twin-ion analysis can be enriched by taking advantage of routines provided in OpenMS to pre-treat input data or conduct post-

* Development and refinement of the HiTIME C++ code is ongoing.
processing on scoring results. For example, the open-source data visualisation tool TOPPView\textsuperscript{254} can be used to visualise HiTIME results files in mzML format.

2.6 Conclusion
An automated process for the non-targeted detection of drug metabolites was demonstrated for the detection of APAP metabolites in blood plasma taken from rats that were administered a 1:1 mixture of APAP and $[^{13}\text{C}_6]$-APAP. The HiTIME algorithm successfully detected a series of doublet signals in unprocessed LCMS data that correspond to 9 metabolites of acetaminophen, including 3 which are important indicators of acetaminophen bioactivation and 4 that are rarely reported. The combination of isotopic labelling, LCMS, and the direct high performance computational processing of high resolution mass spectral data files without data reduction enabled all APAP metabolites to be rapidly and comprehensively detected in complex biological mixtures in a non-targeted and unbiased approach.
3 THE POTENTIAL PROTEIN ARYLATION SITES OF NAPQI

Preface

The work presented in this chapter has been published in *Chemical Research in Toxicology*:

- Leeming, M.G; Gamon, L.F; Wille, U; Donald, W.A; O’Hair, R.A.J. What are the potential sites of protein arylation by N-Acetyl-p-benzoquinone imine (NAPQI)? *Chemical Research in Toxicology*, 2015, 28, 2224.

In the following sections, the NAPQI synthesis optimisation, HPLC purification and NMR characterisation of the synthetically prepared APAP adducts with cysteine, methionine, tyrosine and tryptophan was performed by Dr. Luke Gamon.

3.1 Summary

This chapter investigates the reactivity of the acetaminophen metabolite N-acetyl-p-benzoquinoneimine (NAPQI) toward proteinogenic amino acids. NAPQI was synthetically prepared and reacted with a panel of protected amino acids and peptides. Synthetic NAPQI is shown to covalently react with protected cysteine (NHAc-Cys-OMe) consistent with previous literature. Interestingly, previously unidentified adducts of NAPQI to protected tyrosine, tryptophan and methionine are also observed and have been isolated and characterised by LC-MS/MS and $^1$H-NMR. Analogous products were observed when NAPQI was reacted with unprotected peptides containing these residues (H$_2$N-GAIL-X-GAILR-OH, X = C, M, Y, W)
and MS/MS data indicated that in each case the NAPQI modification was located on the central ‘X’ residue. When APAP and H₂N-GAIL-X-GAILR-OH were incubated with rat liver microsomes (a liver tissue extract that is a rich source drug-metabolising enzymes), NAPQI was produced via CYP oxidation. From these incubation mixtures, NAPQI-modified GAIL-C-GAILR was identified via LC-MS/MS analysis however no NAPQI-modified products were observed where X = M, Y and W. These products were also not identified when liver microsomes were pre-incubated with N-ethylmaleimide (an electrophilic compound used to artificially deplete reactive thiols) suggesting that, while NAPQI adducts with M, Y and W may form, their biological significance may be limited.

3.2 Introduction

Acetaminophen (APAP) is often regarded as one of the safest medications in use today due to the extremely low risk of allergic reactions or other side effects when used as directed. Paradoxically though, APAP is a prominent hepatotoxin and doses of ca. 200 mg/kg or more will likely result in acute liver damage that can lead to fulminant hepatic failure (FHF) if not treated rapidly.²⁵⁵ Due to this, APAP has been consistently identified over the past 20 years as the single most common cause of FHF requiring liver transplantation in countries across the western world including Australia,²⁵⁶ the United States of America²⁵⁷ and the United Kingdom.²⁵⁸ 

Despite significant research, the molecular mechanisms of liver damage following APAP overdose remain largely unclear²⁵⁹, ²⁶⁰ although it is widely accepted that metabolism is involved. The results presented in Chapter 2 are consistent with many previous studies²³⁴, ²³⁵ indicating that most of a therapeutic APAP dose is converted to phase II glucuronic acid and sulfate metabolites that are highly water soluble and rapidly excreted. Additionally, a small but significant portion of the dose (ca. 5 %) undergoes oxidative metabolism catalysed by CYP enzymes producing the highly reactive and putatively toxic species N-acetyl-p-benzoquinone imine (NAPQI).¹⁰⁹, ²⁶¹ In general, any NAPQI produced is rapidly detoxified via conjugation to the cysteine-containing antioxidant tripeptide glutathione²⁶² which is highly abundant (ca. 10 mM) in the liver²⁶³ where the majority of APAP metabolism takes place. However, following an APAP overdose, or in unusually sensitive individuals, cellular reserves of GSH may become depleted and any additional NAPQI formed begins to react with adventitious nucleophiles.²⁶⁴-²⁶⁷ Numerous proteins are known to be covalently modified by NAPQI following administration of a high dose of APAP to mice.¹³⁵, ¹³⁶, ²⁶⁸ Multiple studies have
shown that the degree of protein binding is correlated with the severity of hepatotoxicity, although a direct mechanistic link has not been conclusively proven. Interestingly, relatively little effort has been directed toward understanding the precise molecular details of protein covalent modification by NAPQI or even into characterising the structural properties of the modified proteins.

**Scheme 3-1.** Activation of APAP to NAPQI followed by arylation reactions with sulfhydryl side chains or other reactive nucleophiles from protein side-chains.

In the 1980s, Baillie and co-workers established that the sulfhydryl group of cysteine residues is the major site of NAPQI modification of proteins. Subsequent investigations revealed that cysteine adducts to NAPQI form almost exclusively through initial nucleophilic attack at the 3 position (Scheme 3-1) although other sites of modification are thought to play a minor role. However the possibility that amino acids other than cysteine may react with NAPQI has received comparatively little attention. Interestingly, Streeter et al noted that a small portion of protein bound radioactivity could not be accounted for by cysteine arylation following incubation of 14C-APAP with liver tissue extracts. This suggested that NAPQI may also covalently react with residues other than cysteine. Scattered reports have appeared in the literature suggesting reactivity toward methionine, tyrosine and lysine but none of these products have been isolated and characterised and are very rarely discussed in the context of NAPQI arylation.
Here, we: (i) examine the reaction of NAPQI with \( N \)-acetyl methyl ester derivatives of selected amino acids \( \text{NHAc-X-OMe} \) (where \( X = \text{Cys, Tyr, Trp, His, Lys, Arg, Met, Gln, Glu, Thr and Ala} \)) by isolating covalent adducts using preparative high performance liquid chromatography (HPLC) followed by identification with MS and NMR methods; (ii) use HPLC coupled with HRMS and MS/MS techniques to examine the related adducts formed in the reaction of NAPQI with the peptides GAIL-X-GAILR (where \( X = \text{Cys, Met, Tyr or Trp} \)); and (iii) identify the adducts formed upon APAP incubation with the peptides GAIL-X-GAILR in rat liver microsomes.

3.3 Results and discussion

3.3.1 Chemical synthesis of NAPQI

To study the fundamental chemical reactivity of NAPQI in a tractable system, it was desirable to synthesise and isolate a pure sample of the reactive metabolite. Numerous syntheses of this metabolite have been published. For example, early work by Calder et al produced NAPQI through oxidation with \( \text{Pb(OAc)}_4 \) however the desired product could not be obtained in pure form.\textsuperscript{276} Similarly, dehydration of N-hydroxyparacetamol produces NAPQI but further reactions between the product and unreacted precursor give a mixture of compounds that are difficult to purify.\textsuperscript{277-279} Multiple groups have reported direct electrochemical oxidation of APAP yielding NAPQI.\textsuperscript{261, 274, 280} Our attempts to replicate this electrochemical synthesis were successful and solutions of NAPQI could be reproducibly obtained. However, the success of this procedure, and the magnitude of the product yield, were highly concentration dependent and only very small quantities could be obtained. This was not suitable for production of NAPQI on the synthetic scale required here.

The first practical synthesis of NAPQI was published in 1980 by Blair et al who oxidised APAP in \( \text{Pb(OAc)}_4 \) in benzene.\textsuperscript{281} However, the most widely adopted synthetic procedure is the direct oxidation of APAP by \( \text{Ag}_2\text{O} \) in chloroform, as reported by Dahlin and Nelson in 1982.\textsuperscript{282} Initial attempts to reproduce this work provided mixed results wherein small quantities of NAPQI were observed via thin layer chromatography of the reaction mixture however rapid polymerisation during purification led to complete loss of the product. The polymerisation of NAPQI is a known phenomenon\textsuperscript{283, 284} and oligomers of up to 12 APAP monomers were detected via analysis using the LTQ-FTICR mass spectrometer described in section 8.1.1.3 (data not shown). Through repeated attempts at this synthesis, aided by Dr. Luke F. Gamon, it
was determined that optimisation of three key points greatly aided in the production and isolation of NAPQI. Firstly, the Ag₂O used in this reaction must be freshly prepared on the day that the APAP oxidation is carried out. The commercial sources of Ag₂O used in preliminary reactions become denatured via formation of Ag₂CO₃ through reaction with atmospheric CO₂ that appears to inhibit reaction. Secondly, concentration of the reaction mixture must be avoided. In initial attempts, the reaction mixture was concentrated in vacuo prior to silica gel chromatography and also during attempts at vacuum sublimation. This resulted in rapid product polymerisation. Lastly, sample handling must be minimised to the greatest extent possible. These optimised conditions afforded practical quantities of a NAPQI solution in acetonitrile and were used throughout this chapter.

3.3.2 Reactions of NAPQI with model amino acids

To rapidly identify amino acids that may undergo reaction with NAPQI, and determine which residues should be pursued for synthetic scale investigations, a preliminary reactivity screen was conducted. This utilised protected NHAc-X-OMe residues of Cys, Tyr, Trp, His, Lys, Arg, Met, Gln, Glu, Thr and Ala which were selected to represent the broad range of functionalised side chains present in proteinogenic amino acids. These small-scale reactions were conducted by adding 75 μL of the NAPQI solution in acetonitrile to a solution of each protected amino acid in sodium phosphate buffer at a range of different pH values (6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0). The reaction mixtures were then agitated at room temperature for 3 h and subsequently analysed via LC-MS using the Q-TOF mass spectrometer.

Numerous intense peaks were observed in the TIC traces of each sample. Multiple intense peaks were common to all samples. For example, an intense signal at 13.1 minutes was assigned to detection of the [M+H]⁺ of APAP (m/z 152) indicating reduction of NAPQI. Numerous compounds eluting in the region of 13-16 minutes could be assigned to oligomerisation products of NAPQI. Signals at m/z 301, 450 and 599 correspond to covalent dimer, trimer and tetramer products. Interestingly, multiple peaks were observed at unique retention times for each product and this number increased with the size of the oligomer. This suggests that multiple different structural isomers could be formed with each NAPQI monomer added. Further characterisation of these oligomers was not conducted.
Figure 3-1. Extracted ion chromatograms from LC-MS analysis of NAPQI reaction solutions with protected amino acids. A) Cys, m/z 327.10092, B) Tyr, m/z 387.15506, C) Met, m/z 355.13222, and D) Trp, m/z 410.17105.

To specifically identify signals corresponding to NAPQI adducts of the protected amino acid substrates, EIC traces were plotted using predicted m/z values* for the pH 6.0 reaction mixtures (Figure 3-1). The EIC trace of m/z 327.10, which would correspond to the adduction of NAPQI to cysteine, showed a single peak at 13.6 minutes, indicative of a single regioisomer. This is consistent with the formation of a NAPQI cysteine adduct (Figure 3-1A). A single intense peak was also observed in the EIC trace of m/z 387.16 for tyrosine suggesting the formation of a NAPQI tyrosine adduct (Figure 3-1B). Interestingly, multiple heavily overlapping peaks were

* Adducts of benzoquinone (BQ, a hydrolysis product of NAPQI) were found in the ESI-MS spectra in some cases, however these were not reproducibly observed upon repetition of these experiments. One possible reason is that the synthetic NAPQI precursor partially hydrolyses to BQ prior to addition to solutions of the protected amino acids. As a result, production of BQ adducts from NAPQI hydrolysis in the reaction mixture could not be differentiated from simple BQ contamination of the NAPQI reagent. Thus adducts of BQ are not discussed in this chapter.
observed in EIC traces for NAPQI arylation products of both methionine and tryptophan (data not shown). These samples were re-analysed using a 50 minute HPLC solvent gradient to separate the overlapping compounds. For methionine, two partially resolved peaks were observed at 13.99 and 14.25 mins in EIC traces of m/z 355.13 (Figure 3-1C) suggesting that reaction of NAPQI with protected methionine produces two distinct isomers in approximately equal proportions. Similarly, two major products were observed at m/z 410.17 following NAPQI treatment of protected tryptophan (Figure 3-1C). The first eluted as a sharp peak at 27.7 while the second gave a broad signal centred at 28.2 min. Importantly, for each of the suspected product ions described, similar peaks were not observed in EIC traces of each m/z value for NAPQI-treated alanine samples. As alanine is highly unlikely to react with NAPQI due to its aliphatic side chain, this can serve as a control against which changes induced by the presence of more functionalised side-chains can be measured. Similar inspection of the EIC traces for the other amino acids studied did not reveal any signals that could be attributed to the covalent adduction of NAPQI at any of the pH values tested. These results are consistent with a partial survey of NAPQI-amino acid reactivity by Madsen et al that found Cys and Tyr covalently modified, but that Lys and Ser were unreactive.274

For the four residues that were found to be reactive toward NAPQI, the influence of solution pH on the extent of reaction was analysed by determining the relative abundance of the product ions formed. Figure 3-2 shows the relative area under the EIC peak for each product over the pH range 6.0 - 9.0 indexed to the maximal value observed. For protected cysteine, the product ion abundance was decreased by approximately 90% by increasing the solution pH from 6.0 to 9.0. For methionine and tryptophan, adduct formation was essentially abolished above pH 8.0. By contrast, NAPQI adduction to protected tyrosine was relatively insensitive to pH. The finding that NAPQI adduct formation was most efficient under acidic conditions appears counterintuitive as it would be expected that increasing pH enhances deprotonation of the amino acid substrate (particularly cysteine) therefore increasing nucleophilicity. However, detailed studies by Novak et al demonstrated that protonation of NAPQI increases reactivity substantially as measured by the rate of hydrolysis to benzoquinone.286, 287 Therefore, it appears that the enhanced reactivity observed here may be attributed to increased NAPQI electrophilicity because of protonation at acidic pH. For tyrosine, one possible explanation for the sustained reactivity across increasingly basic conditions is that deprotonation of the phenol moiety enhances the nucleophilicity of tyrosine thereby counteracting the reduced
electrophilicity of NAPQI. Indeed, such as mechanism would be analogous to the well-known Kolbe-Schmitt reaction whereby ortho substitution of phenols is promoted by the addition of base to deprotonate the hydroxyl group.²⁸⁸

![Graph](image)

**Figure 3-2.** Relative extent of formation of NHAc-X(APAP)-OMe as a function of pH of the buffered solution. Extent of formation is measured by EIC peak areas determined from the LC/MS data and indexed to the maximum value. The EIC m/z values used were as follows: Cys 327.10092; Met 355.13222; Trp 410.17105; Tyr 387.15506.

### 3.3.3 Bulk synthesis, isolation and characterisation of NAPQI adducts

Given that multiple adducts between NAPQI and protected amino acids were observed, and that multiple distinct isomeric products were suspected in some of these cases, it was desirable to increase the scale of these reactions to obtain a sufficient quantity of material for isolation and NMR characterisation. For these experiments, the NAPQI preparation described above was conducted on a 1 g scale and the NAPQI product was obtained as a solution in 10 mL of acetonitrile. This was immediately reacted with 0.2 molar equivalents (compared to the quantity of APAP used) of each protected amino acid dissolved in either H₂O (50 mL) for cysteine and methionine or 1:1 H₂O:MeOH for tyrosine and tryptophan. The reactions were stirred overnight at ambient temperature and concentrated *in vacuo* to ca. 10 mL. Precipitated solids were separated by centrifugation and the desired product in the supernatants was separated by preparative HPLC. The purified products were characterised by NMR and high-resolution mass spectrometry.
Scheme 3-2. Structures of the isolated covalent adducts formed via reaction of NAPQI with NAc-X-OMe as determined via $^1$H, $^{13}$C and $^1$H-$^1$H COSY NMR experiments: A) Cys, B) Tyr, C) Met, and D) Trp. A “*” represents the new stereocentre formed, resulting in diastereomers which are separated by HPLC.

Inspection of the splitting patterns and coupling constants in the $^1$H-NMR spectrum of the NAPQI-cysteine adduct supported substitution at the 3 position of NAPQI. No evidence of disruption of the peptide backbone or the Cβ environment was observed, indicating adduction to the reactive thiol group. High-resolution mass spectrometry using the LTQ-FTICR mass spectrometer gave a [M+H]$^+$ ion at $m/z$ 327.10097 which is within 0.2 ppm of the expected value of 327.10092. Collision induced dissociation of this ion resulted in numerous losses of small neutral molecules such as MeOH and ketene (Figure 3-3A). No evidence was obtained that suggested the formation of any other isomer in these experiments. Taken together, these
results indicate that adduction of NAPQI to NHAc-Cys-OMe affords exclusively 3-cystenyl APAP under these conditions Scheme 3-2A.

![Figure 3-3](image)

**Figure 3-3.** MS² CID spectra collected using the LTQ-FTICR mass spectrometer of purified products of NAPQI arylation of NHAc-X-OMe for X = A) Cys, B) Tyr, C) Met, and D) Trp. For Met and Trp, both isomers produced gave similar MS/MS spectra and only one example is provided for simplicity. The precursor ions are marked with an ‘**’.

The 3-cystenyl APAP has been identified as the major product of the reaction between NAPQI and thiol nucleophiles in vivo by other researchers. This apparent selectivity is interesting given that multiple sites on NAPQI may be reactive toward nucleophiles. While much of the observed product likely results from direct attack of cysteine at the 3 position it is interesting to note that Chen et al have identified products of attack at alternative positions. For example, an N-acyl-thiohemiaminal was isolated that suggests ipso attack of cysteine on NAPQI and, additionally, the existence of a C4-thiohemiketal intermediate was postulated that may undergo subsequent rearrangement to give the stable 3-substituted APAP via a 1,2-shift.
Similar reaction of NAPQI solution in acetonitrile with protected tyrosine resulted in the isolation of one major product by preparative HPLC. $^1$H-NMR analysis indicated adduction of NAPQI to NHAc-Tyr-OMe at the 3-position forming an ortho coupled phenol (Scheme 3-2B). HRMS showed [M+H]$^+$ at $m/z$ 387.15514 which is within ca. 0.2 ppm of the expected value of 387.15506. This adduct is structurally similar to the ortho-ortho linked dityrosine residues observed endogenously under conditions of oxidative stress. While Madsen et al have tentatively assigned a NAPQI-tyrosine adduct in their work on electrochemical reactive metabolite screening, to our knowledge this is the first isolation and NMR characterisation of such an adduct.

NHAc-Met-OMe was treated with NAPQI solution in acetonitrile and stirred overnight. UV-HPLC chromatography of the supernatant obtained following centrifugation of the concentrated reaction mixture showed two closely eluting peaks of approximately the same intensity. This is consistent with the results of the small-scale reactivity screen for methionine, which showed two NAPQI-methionine adducts in EIC traces. These different products were carefully isolated and each subjected to NMR analysis and MS analysis. Both compounds produced broadly similar $^1$H-NMR spectra. Small differences were observed for the aliphatic protons. For example, the $\text{C}_\alpha$ proton is observed at 4.56 ppm for one isomer and 4.62 ppm in the other. The methionine –$\text{SCH}_3$ protons were observed at significantly higher chemical shifts in both products (3.33 ppm) compared to native methionine (2.02 ppm). Interestingly, upon collisional activation of the NAPQI-methionine M$^+$ ions ($m/z$ 355) using the LTQ-FTICR mass spectrometer, both isomers produced a single, common fragment ion at $m/z$ 158 corresponding to loss of a 197 Da neutral species (Figure 3-3C). This can be assigned to a neighbouring group reaction of the methionine $N$-acetyl group on the activated $\text{C}_\beta$ resulting in loss of S-methyl APAP and formation of a 6-membered lactam derivative (Scheme 3-3). These data suggest that reaction of NAPQI has occurred at the thioether group of methionine forming a sulfonium ion. Adduction to this site would result in the formation of diastereoisomers because of lone electron pairs at the sulphur atom and the optical purity of the $\text{C}_\alpha$ stereocenter in the protected methionine starting material. These isomers are separable due to a moderate (~25 kcal mol$^{-1}$) inversion barrier that hinders rapid interconversion of the sulphur stereocentre.
Scheme 3-3. Neighbouring group fragmentation reaction of the fixed charge NHAc-Met(APAP)-OMe adduct.

Reaction of NHAc-Trp-OMe with NAPQI solution in acetonitrile followed by work-up and HPLC purification resulted in the isolation of two distinct products, which is consistent with the results observed in Section 3.3.2. HRMS analysis of both products revealed [M+H]\(^{+}\) ions at \(m/z\) 410.17112 and 410.16972 which are within 4 ppm of the expected 410.17105. Analysis of \(^{1}\)H-NMR data revealed that NAPQI adduction had indeed occurred in both cases, neither isolated compound was consistent with the simple substitution reactions observed for cysteine, tyrosine and methionine discussed above.

To aid in interpreting these spectra and determining the structure of the tryptophan products isolated, literature searches were carried out to identify potential reactivity modes of quinones and related products toward indoles. This revealed that these reactions are known to occur via three pathways: i) electrophilic aromatic substitution, 2) [3+2] annulation forming a benzofuroindoline,\(^{33-35}\) and, in the specific case of tryptophan, 3) nucleophilic attack of the indole 3 position and subsequent cyclization forming a pyrroloindoline.\(^{292-294}\) Utilising this information, the first eluting peak, was assigned to a benzofuroindoline derivative (Figure 3-3D and Figure 3-4) based on comparison of \(^{1}\)H-NMR and \(^{1}\)H-\(^{1}\)H COSY spectra to similar compounds reported by other researchers. The second eluting peak produced a substantially more complex \(^{1}\)H-NMR spectrum which could be assigned to formation of a pyrroloindoline. However, the data suggested the presence of two isomeric species. No greater purification of this compound could be obtained despite repeated attempts to separate distinct isomers of this compound via HPLC.
Figure 3-4. Proposed mechanism of benzofuroindoline and pyrroloindoline products from reaction of NAPQI with protected tryptophan.

Two possible explanations for the presence of isomeric pyrroloindoline structures were devised. Firstly, distinct diastereoisomers may be formed in the reaction of NAPQI with protected tryptophan. For example, the initial addition of NAPQI to the tryptophan C3 position could occur from above or below the indole ring. Assuming that subsequent cyclisation gives rise preferentially to the cis-fused indole/pyrrole ring system, this would result in two diastereoisomers. Alternatively, the two suspected isomers observed in the $^1$H-NMR spectrum of the pyrroloindoline product may be due to restricted rotation about a single bond (most likely the indole C3 to NAPQI C3 bond) arising from steric hindrance. If the barrier to rotation is sufficiently high, two distinct populations may be observed in $^1$H-NMR data. Indeed, rotamers such as these have been observed previously for related compounds.295

To differentiate between these, the selective NOESY technique described by Ley and co-workers296 was employed. The methyl ester signal from the pyrroloindoline (CO$_2$CH$_3$) at 3.15 ppm was selectively irradiated and chemical exchange between the two rotamers was subsequently observed (Figure 3-5). As chemical exchange between diastereoisomers should not be observed, this confirms that the two compounds that co-elute at 28.2 min are a mixture of slowly interconverting rotamers of the pyrroloindoline acetamide (Figure 3-3D).
3.3.4 Relative reactivity of protected amino acids toward NAPQI

It was desirable to determine the relative reactivity of protected amino acids toward NAPQI so that the most likely sites of protein adduction may be identified. However, given that NAPQI undergoes many hydrolysis, redox and polymerisation reactions,\textsuperscript{282, 284, 286} and that multiple product isomers are formed in some cases, detailed kinetic experiments are challenging and were deemed to be beyond the scope of this work. Therefore, an experiment was devised that provided qualitative yet useful data on the relative reactivity of NAPQI toward amino acid residues. Solutions were prepared that contained 1 mM of protected cysteine, methionine, tyrosine and tryptophan at either pH 6.0 or 7.5 in sodium phosphate buffer. As an internal standard, 1 mM of protected alanine, which was not found to react with NAPQI in Section 3.3.2, was also included in both solutions. Each solution was then treated with either an excess of NAPQI in acetonitrile or, as a control, acetonitrile alone. The products of the reactions were then analysed by LC-MS using the Q-TOF mass spectrometer.

\textbf{Figure 3-5}. $^1$H spectra for selective NOESY experiment; A) Selected portion of the $^1$H NMR spectrum for mixture of NAc-Trp(APAP)-OMe pyrroloindoline rotamers B) 1D gradient NOE spectrum following irradiation at 3.15 ppm and C) 1D gradient NOE spectrum following irradiation at 3.08 ppm. Coloured regions indicate areas of selective irradiation.
By plotting EIC traces for the precursor protected amino acids, and integrating the area under the relevant peaks for both NAPQI treatment and vehicle control samples, an estimate of the consumption of each residue could be obtained (Table 3-1). Essentially no difference in the abundance of the alanine internal standard was observed between treatment and control samples at either pH. The cysteine precursor was completely consumed at both pH values tested whereas methionine, tyrosine and tryptophan were depleted by 4%, 16% and 67% at pH 6.0 respectively and 5%, 49% and 17% respectively at pH 7.5. This data suggests reactivity orders of Cys > Trp > Tyr > Met at pH 6 and Cys > Tyr > Trp > Met at pH 7.5.

One caveat to these data is that all reactions of a given residue are represented in the consumption figures, i.e. not exclusively the formation of stable NAPQI adducts. This is expected to be most relevant to cysteine where it is thought NAPQI may induce oxidation forming cysteine disulphide (cystine). However, given that this process is proposed to occur via initial NAPQI adduction, it seems reasonable to include this reaction in the quantification.

**Table 3-1.** EIC peak areas for relative depletion of amino acids in 10 mM sodium phosphate buffer upon competitive reaction with an excess of NAPQI.

<table>
<thead>
<tr>
<th></th>
<th>pH 6</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (x10^8)</td>
<td>NAPQI (x10^8)</td>
<td>Change (%)</td>
<td>Yield (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5.5</td>
<td>5.7</td>
<td>103</td>
<td>-2.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.3</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>5.7</td>
<td>5.5</td>
<td>96.2</td>
<td>3.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>4.3</td>
<td>3.6</td>
<td>83.5</td>
<td>16.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>4.0</td>
<td>1.3</td>
<td>32.7</td>
<td>67.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pH 7.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control (x10^8)</td>
<td>NAPQI (x10^8)</td>
<td>Change (%)</td>
<td>Yield (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5.9</td>
<td>5.9</td>
<td>99.7</td>
<td>0.273</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>5.9</td>
<td>5.5</td>
<td>94.8</td>
<td>5.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>4.8</td>
<td>2.4</td>
<td>51.5</td>
<td>48.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>3.8</td>
<td>3.1</td>
<td>83.4</td>
<td>16.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A) Percentage change is defined as Area_{NAPQI}/Area_{Control} x 100
B) Yield is defined as 100 – (Change (%))
3.3.5 NAPQI reactions with model peptides

Given that products of NAPQI adduction to amino acid residues other than cysteine were observed, we investigated whether similar reactions could occur when these residues were incorporated into larger peptides. This was particularly desirable for the case of tryptophan where considerable rearrangement of the peptide backbone is required to form the pyrroloindoline adducts. For these experiments, a series of synthetic peptides of the form H₂N-GAIL-X-GAILR-OH were obtained from commercial sources for X = C, M, Y and K. Notably, the N- and C- termini of these peptides were not protected to more closely mimic in vivo proteins and peptides. Synthetic NAPQI was reacted with each of these peptides in phosphate buffer (pH 6.0 and 7.5) and reaction products were profiled by LC-MS/MS using the Q-TOF mass spectrometer.

Following examination of EIC traces for the calculated \( m/z \) values for each product, the expected products were observed in all cases for reactions at pH 6.0. For cysteine, tyrosine and tryptophan, the \([M+2H]^{2+}\) (\( m/z \) 568.3, 598.3 and 609.9 respectively) were most abundant while for methionine the \([M^{+}+2H]^{3+}\) at \( m/z \) 388.6 was the base peak. To identify the site of reaction of NAPQI with these substrates, the most abundant charge state of each product observed was subjected to MS/MS analysis and the resulting spectra are given in Figure 3-1. For cysteine, tyrosine and tryptophan, an ion at \( m/z \) 529.3 and assigned as the \( y_{5} \) GAILR fragment was observed in each spectrum. By contrast, the \( m/z \) differences between the \( y_{5} \) and \( y_{6} \) ions, that correspond to cleavage of the ‘X’ residue, varied in each case. These were \( m/z \) 252, 312 and 335 for cysteine, tyrosine and tryptophan respectively which are consistent with the mass of the modified residues. Isolation and collisional activation of GAIL-M-GAILR in the +3 charge state resulted in a spectrum that was substantially different. Here, few sequence ions were observed and low \( m/z \) fragments dominate the spectrum. The most abundant of these, \( m/z \) 198.05, 156.04 and 108.04, can be assigned to selective charge-directed fragmentation of a NAPQI-modified methionine side chain similar to that observed for protected methionine (cf. Scheme 3-3). On this basis, \( m/z \) 198.05 is assigned to protonated \( S \)-methyl APAP formed via attack of an adjacent amide group at C\( \beta \). The product ion at \( m/z \) 156.04 results from subsequent deacetylation of the protonated \( S \)-methyl APAP forming \( S \)-methyl-\( p \)-aminophenol. Taken together, this data confirms that the site of NAPQI in these peptides was the central ‘X’ residue and suggests that the reactivity observed in the protected amino acids may be transferrable to larger peptides.
Figure 3-6. LC/MS/MS spectra for NAPQI modified peptides of the form GAIL-X(APAP)-GAILR for A) X = Cys, B) X = Tyr, C) X = Trp, and D) X = Met. All peptides are in the 2+ charge state except for X = Met which is a 3+ ion. The MS² precursor ions are denoted by “*”.

Interestingly, two heavily overlapping yet partially resolved peaks were observed in EIC traces of the NAPQI modified GAIL-X-GAILR ions for methionine and tryptophan suggesting that the stereo- and structural isomers observed in the small molecule studies above were also formed in these peptides. These products could not be isolated and further characterised due to
the low quantities of material used however the relative abundance of these apparent isomers broadly reflected those observed above. Particularly for the tryptophan-containing peptide, it cannot be definitively concluded that the local structure of the NAPQI-modified residue exactly parallels that observed in Section 3.3.3, but it should be noted that rearrangement of the peptide backbone to form a pyrroloindoline does not necessarily preclude the formation of standard \( b \) and \( y \)-type sequence ions.

When the same reactions were conducted at pH 7.5, the yields of NAPQI-modified peptides were reduced. By comparison of EIC peak area, the production of modified peptides was 15 and 12% of that observed at pH 6.0 and none of the expected products were observed for methionine and tryptophan. These results are consistent with the relative reactivity data in section 3.3.4.

3.3.6 APAP and GAIL-X-GAILR microsome incubations
The propensity of NAPQI to arylate these residues was subsequently investigated in a biologically relevant system. To do this, a trapping study\(^{221, 298, 299}\) was carried out wherein APAP (the stable precursor to NAPQI) and the GAIL-X-GAILR peptide substrates studied were incubated with rat liver microsomes. Microsomes are fragments of the endoplasmic reticulum generated by homogenisation of liver tissue. They are a rich source of enzymes involved in xenobiotic metabolism (including CYPs) and are often used in the initial assessment of drug metabolism.\(^{300}\) Action of these microsomal enzymes on APAP is known to form NAPQI,\(^{301}\) which can then undergo further reactions with compounds present in the incubation. After a 90 minute incubation, proteins were removed by precipitation and centrifugation then the resulting supernatants were analysed via LC-MS/MS using the Q-TOF mass spectrometer.

The data were analysed by plotting EIC traces for the most abundant charge state of each product peptide observed in Section 3.3.5. For GAIL-C-GAILR the NAPQI modified peptide was observed at 23.45 min and \( m/z \) 568.31, corresponding to [M+2H]\(^2+\) (Figure 3-7A). The fragmentation pattern of this ion observed upon isolation and collisional activation (Figure 3-7B) was almost identical to that obtained from CID of the synthetic GAIL-C(APAP)-GAILR peptide (cf. Figure 3-6A) indicating microsomal formation of NAPQI from APAP and subsequent adduction to the central cysteine residue.
However, for methionine, tyrosine and tryptophan, analogous products to those produced with synthetic NAPQI were not observed. Upon investigating the potential fate of NAPQI formed in these incubation, an ion at 16.25 mins and m/z 457.14 (Figure 3-8A and B) was observed which can be assigned to an adduct of NAPQI with glutathione. Collisional activation of this ion produced peaks that can be assigned to loss of glycine (-75 Da) and pyroglutamic acid (-129 Da) which are characteristic of glutathione conjugates. The APAP-GSH adduct was also observed for GAIL-C-GAILR albeit at a reduced intensity (58% of the average integrated peak area for X = Met, Tyr and Trp). This indicates that adduction of NAPQI to residual glutathione present in microsome preparations has occurred in preference to reaction with the target substrates.

Figure 3-7. A) EIC of m/z 568.3183 showing elution of the NAPQI modified GAIL-C-GAILR peptide formed in rat liver microsomes. B) CID spectrum of m/z 568.318 at 23.5 min.

To examine whether NAPQI adduction to tyrosine, tryptophan and methionine probe peptides can occur under conditions of reduced glutathione concentration, GSH was artificially depleted by pre-treating the microsomes with the electrophilic compound N-ethylmaleimide (NEM)
prior to incubation. Upon analysis of the GAIL-C-GAILR LC-MS/MS data, the expected NAPQI adduct was observed and no APAP-GSH products were identified. These data indicate the concentration of free GSH was significantly reduced and that the enzyme systems required for APAP oxidation remained active following NEM pre-treatment. Despite this, no further adducts were identified following analysis of LCMS data from NEM-pretreated GALI-X-GAILR microsome incubation.

Figure 3-8. Formation of APAP-GSH via bioactivation of APAP to NAPQI by rat liver microsomes. A) EIC traces for $m/z$ 457.1387 showing elution of APAP-GSH, B) mass spectrum of the [APAP-GSH+H]$^+$ ion for the X = M case, C) CID spectrum of $m/z$ 457.138 at 16.25 min for the X = M case.

The observations of Streeter et al that a small portion of NAPQI bound to proteins that lacked cysteine residues suggests arylation of alternative sites however, to date, these have not been well characterised. The findings here that NAPQI also forms covalent adducts with tyrosine, tryptophan and methionine may therefore aid in explaining non-cysteine protein binding. However, the absence of arylated tyrosine, tryptophan and methionine peptides in these
microsome incubation data is interesting as the present studies with synthetic NAPQI conclusively showed these residues to be potential binding sites. Assuming that the non-cysteine arylation observed by Streeter et al was indeed due to arylation of tyrosine, tryptophan and methionine, the apparent discrepancy between these results may be due to differences in the liver microsomal preparations used. While the materials used here were prepared from rats that had undergone no special treatment, Streeter et al administered sodium phenobarbital for 3 days prior to sacrifice. This compound is well known to increase the expression of many drug metabolising enzymes including CYPs which would have the effect of substantially increasing the production of NAPQI in APAP incubations. This may result in greater opportunity for reaction with non-cysteine nucleophiles.

Taken together, these data indicate that, while NAPQI can form adducts to methionine, tryptophan, and tyrosine – in addition to the well-known cysteine adduct – the formation of these products is suppressed in the presence of an excess of extracellular thiols. However, this finding does not entirely preclude the possibility that NAPQI may form adducts to protein residues other than cysteine and further investigations are required.

3.4 Conclusion
The reactivity of the electrophilic acetaminophen metabolite N-acetyl-p-benzoquinoneimine toward model amino acids and peptides has been investigated using chemical synthesis, microsomal incubations, LC-MS/MS and NMR. Synthetically prepared NAPQI was found to react with NHAc-Cys-OMe consistent with previous literature. In addition, products of covalent adduction were also identified following reaction of acetyl Tyr, Trp and Met methyl esters revealing new modes of reactivity of this metabolite toward proteins, which remained operative when these residues were incorporated into GAIL-X-GAILR decapeptides. Upon CYP-mediated microsomal bioactivation of APAP to NAPQI, the NAPQI-modified cysteine-containing peptide GAIL-C-GAILR was identified following LC-MS/MS analysis however similar adducts to Tyr, Trp and Met peptides were not observed under these conditions. Despite this, the identification of multiple new reactivity channels for NAPQI is interesting given that it is one of the most intensively studied reactive metabolites known. This highlights the need for the development of rapid, non-targeted experimental methods that can identify the site of modification of a protein without prior knowledge.
4 **Xenophile: A Software Toolkit for the Non-Targeted Identification of Drug-Protein Adducts**

**Preface**

Part of the work presented in this chapter has been published in *Analytical Chemistry*:


**4.1 Summary**

An approach to the non-targeted identification of the protein adduction targets of chemically reactive metabolites is described. This method combines the twin-ion method of stable isotopic labelling with LC-MS/MS and can determine the identity of the reactive metabolite, the amino acid site(s) modified and the identity of the modified protein without the need for expensive radiotracers, laborious protein separation or prior knowledge of the reactive species or the sites modified. To analyse the vast amounts of data produced during these experiments and extract meaningful protein adduction data, a software package called Xenophile has been produced that can automate aspects of data processing and facilitate review of the results. Xenophile has been made freely available. This section describes the development of the non-targeted CRM analysis methodology, the associated software and validates the process on synthetic test data.
4.2 Introduction

The findings in Chapter 3 that the reactive paracetamol metabolite, NAPQI, can covalently modify methionine, tyrosine and tryptophan in addition to cysteine highlight the fact that new modes of reactivity can be identified for even extensively studied compounds.\textsuperscript{303} Given that adduction of reactive drug metabolites to proteins is thought to be important in initiating allergic reactions to medications,\textsuperscript{75, 103} and that small structural changes to a substrate appear to induce large changes to the protein adduction profile,\textsuperscript{135, 136, 138} experimental methodologies capable of determining which specific proteins are modified as well as the identity of the reactive metabolite and the type of amino acid modified would be beneficial. If these methods were sufficiently rapid and automated, a catalogue of proteins modified for a range of different drugs could be compiled which may lead to identification of specific modifications or adduction profiles that may increase the probability of adverse immune reactions.

Numerous experimental methodologies have been developed to identify the protein targets of reactive drug metabolites and are the subject of recent reviews.\textsuperscript{140, 141} Early approaches to the identification of drug-modified proteins involved raising antibodies against the modified amino-acid drug motif. For example, in the 1980s Satoh et al studied the hepatotoxicity induced by halothane – an inhaled anaesthetic that was commonly used at the time – and produced antisera against lysine residues modified by trifluoroacetyl chloride which was known to be formed from halothane metabolites.\textsuperscript{304, 305} Immunofluorescence studies of hepatocytes using the anti-TFA sera indicated the presence of TFA modification at the cell surface which were thought to correspond to TFA-modified membrane proteins.\textsuperscript{306}

Subsequent advances in protein separation using 2D SDS-PAGE\textsuperscript{307} and significant improvement to mass spectrometer technology with the introduction of ESI\textsuperscript{146, 308} and MALDI\textsuperscript{309} enabled routine identification of specific proteins. In terms of reactive metabolite-protein adduct detection, two-dimensional Western blots using drug-conjugate antisera allowed precise visualisations of gel regions that contained immunoreactive proteins, which could then be excised and identified using mass spectrometry. Using this methodology, protein adducts of several compounds were identified in the 1990s and 2000s including menthofuran,\textsuperscript{310} 4-hydroxynonenal,\textsuperscript{311, 312} butylated hydroxytoluene\textsuperscript{313} and tienilic acid.\textsuperscript{314} At the same time, various groups used radiolabelled drugs in combination with 2D SDS-PAGE and autoradiography to localise protein adducts on the gel, which could then be excised for MS
analysis. Numerous compounds have been studied in this way, including paracetamol,\textsuperscript{268} naphthalene,\textsuperscript{315, 316} tienilic acid\textsuperscript{317} and thioacetamide.\textsuperscript{318}

Substantial improvements through the 2000s in chromatographic separation of peptides and mass spectrometer technology enabled routine analysis of complicated shotgun digests\textsuperscript{319, 320} and gave rise to new affinity capture strategies for drug-protein adduct detection. These involve making chemical changes to the drug substrate that incorporate a functional group that undergo highly specific and selective chemical reactions such as the 1,3-dipolar cycloaddition between azides and alkynes (\textit{i.e.} the so-called ‘click’ reaction)\textsuperscript{321} or the high-affinity non-covalent binding of biotin to avidin.\textsuperscript{322, 323} This allows selective capture of the modified proteins and easy removal of the majority of proteins which are not modified.

\textbf{Figure 4-1.} Covalent modification of lysine side-chain amine group by flucloxacillin and subsequent MS/MS fragmentation. Adapted from Jenkins \textit{et al.}\textsuperscript{324}

Numerous researchers have exploited known gas-phase ion fragmentation behaviour to develop targeted screening approaches using MRM-type mass spectrometry experiments. For example, \(\beta\)-lactam antibiotics are known to covalently modify proteins by reacting with side chain primary amine of lysine residues\textsuperscript{324} (Figure 4-1) which can lead to immune reactions against haptenised peptides.\textsuperscript{325} Peptides containing this modification have been identified using MRM experiments by monitoring for fragments formed at \(m/z\) 160, which corresponds to the cleavage of the thiazolidine group.\textsuperscript{98, 324} Methods and software have been produced that allow users to search peptide MS/MS spectra according to parameters customised for a given drug substrate including fragment ions, neutral losses or ion pairs.\textsuperscript{326}

There are, however, significant drawbacks to each of these methodologies that have likely contributed to the relative rarity of studies aiming to identify the specific targets of drug-protein adducts. For example, Matthews \textit{et al} prepared antisera against APAP-modified proteins by
immunisation of rabbits with APAP bound to keyhole limpet hemocyanin (KLH)\(^*\) with different linker groups. When these were used to detect the protein adducts produced in mouse liver fractions, differences in the antibody binding were observed between the antisera for the same sample. This indicated that the local structural environment of the complex used to produce the antisera can have an impact on the target recognition by the antibodies. Use of radiolabelled drug will overcome the structural specificity of antisera although 2D SDS-PAGE and subsequent excision of radioactive gel spots and digestion of individual samples remains a labour-intensive process. Exploitation of known fragmentation behaviour in MRM experiments allows for selective detection but requires prior and specific knowledge of the reactive xenobiotic species and its gas-phase unimolecular chemistry, which runs counter to the non-targeted philosophy of this work.

**Table 4-1.** Examples of different amino acid residues modified by various xenobiotics.

<table>
<thead>
<tr>
<th>Modified Residue</th>
<th>Example</th>
<th>Reactive group or Metabolite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cysteine</td>
<td>acetaminophen(^{270})</td>
<td>quinone imine</td>
</tr>
<tr>
<td></td>
<td>diclofenac(^{329})</td>
<td>quinone imine</td>
</tr>
<tr>
<td>Lysine</td>
<td>penicillin(^{98})</td>
<td>(\beta)-lactam</td>
</tr>
<tr>
<td></td>
<td>isoniazid(^{330})</td>
<td>hydrazine</td>
</tr>
<tr>
<td>Histidine</td>
<td>nevirapine(^{331})</td>
<td>quinone methide</td>
</tr>
<tr>
<td></td>
<td>carbamazepine(^{332})</td>
<td>arene oxide, epoxide</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>dinitrochlorobenzene(^{333})</td>
<td>aryl halide</td>
</tr>
</tbody>
</table>

Detailed studies on a range of xenobiotics have shown that numerous amino acid residues can be susceptible to covalent modification (Table 4-1). Given that the reactive metabolites formed from a xenobiotic are not necessarily known, and that there are multiple potentially reactive sites on proteins, ideal methods for identifying drug-protein adducts should be insensitive to the identity of the metabolite, the amino acid modified, the structural environment surrounding the modification and the ion fragmentation pathways of the conjugate.

\(^*\) KLH is a large protein derived from marine mollusc *Megathura crenulata* that is more commonly known as the ‘giant keyhole limpet’. It is frequently used in biochemistry as a carrier protein in the production of antibodies against low molecular weight compounds. For this purpose, small molecular weight molecules are synthetically crosslinked to KLH and the complex is then injected into an experimental animal.
This chapter introduces a methodology that allows for the selective identification of reactive drug metabolites, the identity of the amino acid modified and the protein target. In keeping with the philosophy of non-targeted identification, we aim to accomplish this without making assumptions as to the identity of the reactive metabolite that may be formed from an input drug and without assuming any knowledge of the amino acid(s) that may be modified. Thus, the core aims of this chapter are:

1) Design an experiment that incorporates the stable-isotope labelling methodology to identify CRM-modified peptides displaying the twin-ion signature.

2) Develop a set of data analysis procedures that allow the protein adducts of reactive drug metabolites to be identified \textit{a priori}.

4.3 Experimental design

In Chapter 3, chemically synthesised and purified NAPQI was used to investigate amino acid arylation. While this provided valuable insights, it should be noted that this required advanced knowledge of the reactive metabolite formed \textit{in vivo} and, moreover, chemical synthesis and isolation of a reactive species will generally not be possible for most xenobiotics due to the presence of sensitive functional groups or the instability of the product. Therefore, any studies to identify reactive metabolites and their protein adduction targets must be performed in biologically representative systems such as microsomes, cells or animals where enzymatic bioactivation of the administered compound can take place.

In Chapter 2, administration of APAP and $^{13}$C$_6$-APAP in equal proportions to rats allowed the small molecule metabolites to be detected using HiTIME by virtue of the twin-ion signature produced in LC-MS data. A similar approach could be taken to detect modified proteins whereby equal parts of heavy and light drug are incubated in biological systems containing enzymes involved in xenobiotic metabolism. Any proteins modified by reactive species would have the same ratio of labelled and unlabelled as the precursor chemicals added to the incubation media. Upon global protein digestion, only those specific peptides that have been modified by the xenobiotic would appear as twin-ions in LC-MS data allowing easy detection using HiTIME. The sequence and protein origin of the modified peptides could then be determined by MS/MS. In principle, this allows for detection of peptides modified without the need for SDS-PAGE, antisera, radiotracers, specific knowledge of the modification site fragmentation behaviour or even the identity of the reactive metabolite.
4.3.1 Synthetic data production

To illustrate the concepts discussed throughout this chapter, and perform validation studies on the methods developed, a synthetic LC-MS/MS dataset was produced that simulates the results of trypsin digestion of a mixture of 5 different proteins found in human blood plasma. These proteins were: Myeloperoxidase (MPO), interleukin 6 (IL-6), α1-antitrypsin (A1AT), mitogen-activated protein kinase 12 (MAPK-12) and human serum albumin (HSA). To provide numerous artificial twin-ions, a twin-ion pair was introduced for each tryptic peptide in HSA that contained exactly one cysteine residue. This simulates adduction of a drug metabolite to HSA. The remaining 4 proteins are present in their native form only.

Scheme 4-1. Bioactivation of trazodone to the electrophilic metabolite CPPQI.

The non-tricyclic antidepressant trazodone was selected for this example. Trazodone functions by inhibiting serotonin reuptake in the central nervous system and is indicated for the treatment of major depression and also has anxiolytic and sedative properties. In rare cases, severe hepatotoxicity has been reported following administration of trazodone and multiple CRMs have been identified following detailed metabolic studies including chloro-p-piperazinequinoneimine (CPPQI) produced as a result of multiple CYP-mediated oxidations shown in Scheme 4-1. To produce the training data, the CPPQI metabolite in both natural abundance and 13C6 forms was computationally applied to the cysteine residues of HSA. This created twin-ion peptides that differed in mass by 6.0201 Da.

This procedure resulted in the generation of 159 native peptides (total for all 5 proteins) including 17 CRM-modified twin-ions. The raw LC-MS data (i.e. before any processing) is depicted as a heat map in Figure 4-2A. The output of this data synthesis was an mzML file.
containing LC-MS\(^1\) data for all simulated peptides and an MGF file containing MS\(^2\) fragment ion lists. Specific details about the data production can be found in section 8.4.1.2 on page 185 and the python script written to carry out these steps can be obtained from https://github.com/mgleeming/phdScripts.

**Figure 4-2.** A) Raw, synthetic LC-MS data generated to simulate trypsin digestion of a mixture of 5 different proteins including one that has undergone covalent binding with TZD reactive metabolites at numerous sites. B) Results of HiTIME scoring of the synthetic LC-MS data using a mass spacing of 3.01005 Da to detect [M+2H]\(^{2+}\) modified peptides.

4.3.2 Initial data processing

From this point forward, it will be assumed that we do not know the number or identity of individual proteins in the synthetic dataset, the nature of the reactive metabolite formed from TZD, the amino acid residue bound by this metabolite or the protein modified.

Many peptides are observed in the synthetic LC-MS data that are not twin-ions. To identify these peptides and the proteins they originated from, LC-MS/MS data was searched against the Uniprot database\(^3\(^4\)\)\(^0\) using Mascot.\(^3\(^4\)\(^1\)\) This resulted in the identification of 5 unique proteins: MPO, IL-6, A\(_1\) AT, MAPK-12 and HSA. It should be noted that, in these searches, adduction of any reactive metabolites would change both the total mass of the peptide and induce an offset in the ladder sequence ions formed in MS\(^2\) experiments resulting in a mismatch between the experimental spectra and the theoretically predicted fragments generated during database searching. Adducted peptides will therefore remain unassigned and, accordingly, this would not account for the possibility of reactive metabolite covalent binding.
Several methods could be used to recover the ‘missed’ peptides that have undergone metabolite adduction. For example, Mascot searches could be directed to consider variable modifications at specific residues for predicted metabolites. However, this would require knowledge of both the identity of the reactive metabolite and the amino acid residue(s) likely to be modified, which runs counter to the goal of non-targeted identification. ‘Blind’ PTM search algorithms that aim to identify modifications by various approaches based on sequence tags may be adequate in some instances, but since protein-CRM adducts are expected to be rare events relative to the thousands of proteins that can be detected in shotgun experiments, the abundances of modified peptides are likely to be low (often leading to ‘noisy’ MS/MS spectra) and CRM modification for a compound of interest are unlikely to be recorded in PTM databases such as UniMod. Any search algorithm would therefore be required to handle arbitrary mass offsets and account for modification on any residues which would result in increased search times, false-negatives and false-positives. However reactive metabolites can be considered a distinct subset of all PTMs present in a given sample and as HiTIME searching allows these peptides to be selectively detected global PTM searching becomes unnecessary.

However, in these in silico experiments, it is known that any peptide that is modified by a reactive metabolite derived from TZD should be observed as a twin-ion. To identify twin-ions in this LC-MS data, HiTIME was performed using an m/z spacing of 3.01005 (corresponding to the 6.0201 Da mass difference between $^{12}$C and $^{13}$C$_6$ trazodone in the +2 charge state) and the resultant heat map is shown in Figure 4-2B. Upon manual inspection of these results, 17 data regions that attract high HiTIME scores (cf. 9.3.1) are identified indicating the presence of CRM-modified peptides.

Close inspection and analysis of the MS$^2$ spectra associated with these twin-ion peptides may allow for sequence assignment. However, given that numerous twin-ions are observed, and that the site and mass of the modification are unknown, this task is likely to become prohibitively laborious. To partially automate the task of assigning the MS$^2$ spectra of twin-ion peptides, and simultaneously deriving the formula of the reactive metabolite modification, a custom suite of software was developed.
4.4 Software development

The process of identifying the protein targets of reactive metabolites directly from shotgun LC-MS/MS data is daunting but can be broken up into three distinct computational tasks:

1) Identification of the reactive species,
2) Identification of the type of amino acid(s) modified by this species, and
3) Identification of the specific proteins modified.

If the identity and reactivity profile (tasks 1 and 2) of the reactive metabolite can be established, then identification of the proteins adducted (task 3) can be achieved via standard protein database searching.

The following sections describe the development and implementation of a non-targeted workflow that rapidly processes the data generated by LC-MS and LC-MS/MS analysis of twin-ion experiments using microsomes, cells or in vivo (Scheme 4-4). The output provides the user with a list of the modified peptides identified, the proteins that these originate from and candidate formulae for the reactive drug metabolite.

The software consists of two core elements in addition to many smaller tools. The core elements are:

1) A non-targeted reactive metabolite identification algorithm that allows for the molecular formula of a CRM to be determined as well as the type of amino acid modified.
2) A targeted peptide and protein analysis algorithm that correlates peptides that are assigned as having reactive metabolite modifications by Mascot with twin-ion data produced by HiTIME.

The various tools described below have been grouped together and are collectively referred to hereafter as ‘Xenophile’. 
4.4.1 The non-targeted algorithm

The non-targeted reactive metabolite identification algorithm (Scheme 4-2A) aims to determine the identity of a reactive metabolite that is covalently bound to a peptide, and the type of amino acid residues modified, directly from LC-MS/MS data. Since reactive metabolite modifications of proteins are generally not quantitative and are formed in relatively low abundances, it is reasonable to assume that some quantity of its unmodified ‘native’ counterpart will remain. As the native and CRM-modified peptides share the same base structure and differ only in the configuration of various PTMs, MS/MS spectra of these peptides are likely to display sequence ions that are either common to both modified and native peptides or offset from one-another by a constant amount that is determined by the mass of the adducted metabolite. Therefore, given a list of twin-ion hits produced by HiTIME searching of LC-MS data, MS/MS spectra of these twin-ions can be compared to other non-twin-ion peptides assigned by Mascot attempting to identify correlations.

4.4.1.1 Initial data processing

Given raw LC-MS and LC-MS/MS data, HiTIME searches are performed to detect candidate twin-ion peptides that are likely to correspond to CRM modifications. MS$^2$ spectra are processed using the peptide search engine Mascot. Importantly, when specifying fixed and
variable residue modifications during Mascot search setup, it is not necessary to input any modification relating to CRM adduction. Only modifications based upon the underlying biology and sample preparation (which are generally known) must be entered. For example, these may include methionine oxide (arising from oxidation of the methionine side chain thioether group) and carbamidomethyl cysteine (arising from alkylation of reduced cysteine residues with iodoacetamide). The output of Mascot analysis is a file containing peptide search results for each MS² query spectrum submitted.

It is important to note that Mascot searching in this way will not directly identify reactive metabolite-adducted peptides as the CRM modifications will cause mismatches between the calculated and experimental peptide mass and fragment ion m/z values. However, given that some amount of the native peptide is expected to be present in the sample, these should be assigned by database searching. It is assumed that for every HiTIME hit representing a CRM-modified peptide, the corresponding non-CRM-modified peptide will be assigned by database searching. The non-targeted CRM identification algorithm therefore attempts to find the native peptide assigned by Mascot that most closely matches the MS/MS spectrum of each twin-ion peptide.

Table 4-2. Program inputs for non-targeted CRM identification

<table>
<thead>
<tr>
<th>Input</th>
<th>Type</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mascot .dat file</td>
<td>Required</td>
<td>Results for search using typical variable modifications</td>
</tr>
<tr>
<td>HiTIME hit file</td>
<td>Required</td>
<td>Results for searches using appropriate Δm/z values</td>
</tr>
<tr>
<td>Xenobiotic structure</td>
<td>Required</td>
<td>Used to aid in determination of m/z bands and final reactive metabolite formulae.</td>
</tr>
<tr>
<td>Allowed atom ranges</td>
<td>Required</td>
<td>Allowed number of each atom in molecular formulae of candidate reactive metabolite. Used to determine the m/z band searched for potential peptide matches</td>
</tr>
<tr>
<td>Reactive residues</td>
<td>Optional</td>
<td>Restrict residues that are considered possible modification sites to those specified.</td>
</tr>
</tbody>
</table>

4.4.1.2 Reactive metabolite assignment

For each HiTIME hit, a subset of Mascot-assigned peptides is extracted that fall within m/z ranges defined by the user (Table 4-2). A schematic of this process is shown in Figure 4-3. MS² spectra are then extracted for both the HiTIME hit and each identified peptide in the Mascot subset ready for correlation and scoring. Using the assigned sequence of each peptide in the Mascot-identified peptide subset, a theoretical series of fragment ions is generated. The type
of fragment ions generated can be selected by the user to match the ion fragmentation method used to acquire experimental data. For example, slow ion activation methods such as CID and HCD typically produce ‘b’, ‘y’ and ‘a’ type ions while electron based methods such as ETD and ECD predominantly form ‘c’ and ‘z’ type ions.\(^{347}\)

**Figure 4-3.** Relationship between a HiTIME hit and potential native counterparts.

The theoretical fragment ion spectrum is then compared to the experimental MS\(^2\) spectrum associated with the HiTIME hit. A correlation score that quantifies the similarity of the two spectra is calculated by summing the normalised intensities of ions in the theoretical spectrum that overlap with ions in the experimental spectrum to within a user-defined \(m/z\) tolerance and this sum is then multiplied by the integer number of ions that are matched (Eq. 4-1). This metric provides higher scores for matches of highly abundant ions in the MS/MS spectra while also incorporating a multiplier to elevate the scores for spectra that match numerous fragments.

\[
Score = n \sum i_m
\]

Correlation of the MS\(^2\) spectra in this way is complicated by the fact that fragment ions in the HiTIME hit spectrum that contain the CRM modification will be offset from those of the native spectrum by a mass related to CRM. The impact of this is minimal if the modification occurs near N-terminus of the peptide, however this cannot be assumed. To address this problem, a ‘rolling modification’ model was developed that, for each pair of HiTIME and native peptides, generates multiple theoretical ion sets that correspond to the fragmentation behaviour expected for the CRM at different sites.
<table>
<thead>
<tr>
<th>Native modifications only</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

**Modification at site 1**
Modification mass = 19 Da

| 1 | G^M | 77.0 | LCCAMGLR | 618.3 |
| 2 | G^ML | 190.1 | C^CAMGLR | 505.3 |
| 3 | G^MGLC^CAM | 350.1 | GLR | 345.2 |
| 4 | G^MGLC^CAMG | 407.2 | LR | 288.2 |
| 5 | G^MGLC^CAMGL | 520.2 | R | 175.1 |

**Modification at site 2**
Modification mass = 19 Da

| 1 | G | 58.0 | L^MC^CAMGLR | 637.3 |
| 2 | GL^M | 190.1 | C^CAMGLR | 505.3 |
| 3 | GL^MC^CAM | 350.1 | GLR | 345.2 |
| 4 | GL^MCC^CAMG | 407.2 | LR | 288.2 |
| 5 | GL^MCC^CAMGL | 520.2 | R | 175.1 |

**Modification at site 3**
Modification mass = 57 + 19 = 76 Da

| 1 | G | 58.0 | LCMGLR | 637.3 |
| 2 | GL | 171.1 | C^MGGLR | 524.2 |
| 3 | GLCM | 350.1 | GLR | 345.2 |
| 4 | GLCMG | 407.1 | LR | 288.2 |
| 5 | GLCMGL | 520.2 | R | 175.1 |

**Table 4-3.** Theoretical ‘b’ and ‘y’ series ions generated via the rolling modification function for the exemplary peptide GLCGLR. In this example, suppose that carbamidomethylation of cysteine is assigned by Mascot (denoted by C^CAM) and that the mass difference between HiTIME and mascot peptides is 19 Da and X^M refers to placement of the hypothetical modification (M) on residue X.

It should be noted that adduction of a CRM at a given residue is likely to preclude modification at that residue that may otherwise form. For example, CRM adduction at a cysteine residue will likely block alkylation of the thiol group by iodoacetamide treatment during sample preparation. This possibility must be considered when generating theoretical fragment ion spectra using the rolling modification approach. For a given position, if no modification is assigned to a given site the CRM mass (M_{CRM}) is determined by Eq. 4-2 which is simply the
mass difference between the twin-ion peptide \((M_T)\) and the native peptide \((M_N)\). However, if a modification is present at a given site, the mass of the CRM is given by Eq. 4-3, which incorporates the mass of the native modification \((M_{\text{mod}})\) into the CRM.

\[
M_{\text{CRM}} = z(M_T - M_N) \\
M_{\text{CRM}} = z(M_T - M_N) + M_{\text{mod}}
\]

As an example, suppose that the peptide GLCGLR is covalently modified by a hypothetical reactive metabolite benzyne derived from benzene resulting in the net addition of \(\text{C}_6\text{H}_4\) (76 Da) to the cysteine residue and that the ‘native’ equivalent is assigned by database searching as \(\text{GLC}^{\text{CAM}}\text{GLR}\) where \(\text{C}^{\text{CAM}}\) indicates carbamidomethyl cysteine. The \(m/z\) difference between these two ions (in a 1+ charge state) is 19 Da. The first three theoretical ion sets generated in this case are shown in Table 4-3. For theoretical ion sets 1 and 2, placement of a hypothetical modification on the first and second residues does not obstruct a native modification and modification mass is taken to be 19 Da. However, for the third set, modification by a reactive metabolite is likely to preclude carbamidomethylation and the modification mass is calculated as the sum of 19 and 57 Da. Each set is compared to the MS/MS spectrum of the twin-ion peptide and the correlation is quantified according to Eq. 4-1. The theoretical ion set giving the highest correlation score is taken as the lead match for this ion pair. In this process, the user is given the option of restricting reactive metabolite modifications to only specific residues. The user may, for example, define glycine as ‘unreactive’ toward CRM modification and theoretical ion sets with the modification placed at glycine residues will not be calculated thereby narrowing the search space.

This process is repeated for each potential native peptide and the peptide and theoretical ion set that is the overall highest scoring of any comparison made is taken as the lead match candidate. The modification site and calculated CRM mass that gave rise to this ion set is considered as a potential CRM.
4.4.1.3 Chemical formula assignment

Chemical formulae are then computed that may correspond to the assigned CRM mass. Although the number of formulae that are determined for a given $\Delta m/z$ may be large, many of these will likely fall beyond the limits of chemically reasonable compositions. To restrict the search space only compositions that fall within user-defined elemental compositions are calculated. Further refinement can be achieved by noting that the chemical structure and composition of the assigned reactive metabolite should have elements in common with the administered xenobiotic \textit{i.e.} there should be some degree of similarity between the formula of the input drug and the formula of the reactive metabolite.

To take advantage of this relationship, the user is given the options of specifying the structure of the test compound as a SMILES string.\textsuperscript{348} It is known that various reactions in xenobiotic metabolism such as heteroatom dealkylation may cleave the input molecule into separate fragments that may each have their own metabolic and pharmacokinetic profile. To approximate this process, the user-defined input molecule is split into fragments at rotatable bonds that are identified using the SMARTS string:

$$[!$(\text{NH}@C(=O))&!D1&!$(***))]&!@ [!$(\text{NH}@C(=O))&!D1&!$(***))]$$ 4-4

As an example, the fragments generated by disconnection of rotatable bonds of TZD are given in Scheme 4-3. It is assumed that reactive metabolites may be derived from either the precursor molecule or any of the fragments. To quantify the difference between a given candidate molecular formula and the input molecule and its fragments a metric was developed that is hereafter referred to as ‘Residual Mass Error’ (RME). This is a numerical value calculated according to Eq. 4-5 where $|n_f^i - n_d^i|$ is the difference between the number of atoms of element \textit{i} in the fragment formula \textit{n}_f and the determined formula \textit{n}_d and \textit{m}^i represents the atomic mass of element \textit{i}. The RME value represents the total mass (in Da) of the administered molecule fragment that is not explained by the lead candidate formula.

$$RME = \sum |n_f^i - n_d^i| \cdot m^i$$ 4-5

For each molecular formula that may correspond to an assigned reactive metabolite, a residual mass error is calculated between the formula and each molecular fragment generated by rotatable bond disconnection. The fragment that gives the lowest RME value is that which is most similar to the candidate formula. Thus, for an assigned CRM, the CRM mass, assigned
formulae, ppm error, RME and the structure of the most similar input molecule fragment are then presented to the user for review.

Scheme 4-3. Fragments generated by disconnection of rotatable bonds in the TZD performed during non-targeted CRM searching.

The RME value, combined with the candidate CRM formulae and mass, can be used to aid in determination of the metabolic processes that occurred to a molecule to generate the reactive
metabolite that is observed covalently bound to proteins. RME values should generally correspond to the mass offsets introduced to a molecule by the various metabolic processes that are known to generate reactive metabolites. For example, hydroxylation and oxidation of electron rich phenyl groups generates quinones, quinone imines, or quinone methides depending on the substituents. RME values for all these cases should equate to 14 Da, i.e. +16 Da introduced by addition of oxygen in the hydroxylation step followed by -2 Da corresponding to the loss of 2 hydrogen atoms in quinone formation.

This section of the overall CRM protein-identification workflow is executed first as an initial data exploration step. The aim is to identify 1) candidate formulae for the CRM and 2) the sites at which this CRM reacts. The output of the non-targeted identification algorithm is therefore:

1) A list of peptides that may have been modified via CRM adduction
2) For each peptide, a list of possible molecular formulae for the CRM
3) Candidate structures for the CRM.
4) The amino acid residue sites at which the CRM reacts

4.4.2 The targeted algorithm

The non-targeted CRM identification algorithm above aims to identify the formulae of any reactive metabolites and the type of amino acid residue modified however the identity of the peptides and proteins modified remains to be established. This information on the nature and reactivity of the CRM can then be used to direct further protein database searches to explicitly consider the possibility of these modifications. Results of these database searches should now include peptides that have been modified by CRMs.

However, the potential for false-assignments is likely to increase as additional variable modifications are incorporated into peptide database searches as a result of the combinatorial expansion of the number of PTM configurations that must be searched.\(^{349}\) In these experiments, it is known that true modified peptides should be assigned as such by database searching and also display the expected twin-ion shape. This allows CRM-peptide assignments to be confirmed using two independent data sources. To aid in the correlation of HiTIME data with CRM-peptide assignments from database searching, the targeted CRM correlation algorithm was developed.
The targeted adduct identification algorithm is a conceptually simple process that checks if a peptide assigned by database searching as a CRM appears as a twin-ion in MS\textsuperscript{1} data. To do this, the \(m/z\) and retention time values of the peptide from database searches are compared to the locations of twin-ions identified by HiTIME searching. Where overlaps are observed, the same peptide has been assigned as a twin-ion by HiTIME and as a CRM adduct by database searching providing a high level of confidence that this is a product of covalent protein modification by a reactive metabolite. This process is schematically depicted in Scheme 4-4.

Scheme 4-4. Experimental and informatics workflow for targeted identification of CRM modified peptides and proteins.

4.5 Software validation using synthetic data

To ensure that the functions and algorithms described above have been implemented correctly and verify that these processes are able to identify reactive metabolites, the non-targeted CRM identification algorithm was applied to the synthetic LC-MS/MS data generated via \textit{in silico} digest of TZD-treated proteins. The parameters used for these test searches are given in Appendix Table 9-5 and Table 9-6.

After running the non-targeted CRM identification algorithm, the highest scoring peptide matches (determined by the similarity of MS/MS spectra, \textit{cf.} Eq. 4-1) for all 17 of the validated HiTIME hits indicated an assigned reactive metabolite mass of 211.0638 Da and, in each case, the modification was assigned at a cysteine residue. The possible formulae that are within the
user-specified ppm error limits and fall within the allowed stoichiometries are given in Table 4-4. The first listed hit, C\textsubscript{10}H\textsubscript{12}N\textsubscript{2}OCl, is assigned the lowest RME score of 16 indicating a close structural similarity to its nearest fragment, structure 2b in Scheme 4-3. The remaining candidate formulae are all assigned both higher RME values and ppm errors and can be eliminated by simple comparison to their candidate fragments.

\textbf{Table 4-4.} Candidate molecular formulae determined by non-targeted CRM identification of the products of \textit{in silico} adduction of TZD metabolites to HSA tryptic peptides.

<table>
<thead>
<tr>
<th>Hit</th>
<th>Formula</th>
<th>ppm error</th>
<th>RME</th>
<th>Nearest Fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C\textsubscript{10}H\textsubscript{12}N\textsubscript{2}OCl</td>
<td>0.003</td>
<td>16</td>
<td>2b</td>
</tr>
<tr>
<td>2</td>
<td>C\textsubscript{7}H\textsubscript{15}N\textsubscript{3}Cl\textsubscript{2}</td>
<td>2.3</td>
<td>88</td>
<td>2b</td>
</tr>
<tr>
<td>3</td>
<td>C\textsubscript{13}H\textsubscript{9}N\textsubscript{2}OCl\textsubscript{3}</td>
<td>-2.3</td>
<td>81</td>
<td>1a</td>
</tr>
<tr>
<td>4</td>
<td>C\textsubscript{8}H\textsubscript{10}N\textsubscript{5}Cl</td>
<td>6.4</td>
<td>68</td>
<td>2b</td>
</tr>
<tr>
<td>5</td>
<td>C\textsubscript{9}H\textsubscript{17}OCl\textsubscript{2}</td>
<td>-8.7</td>
<td>96</td>
<td>2b</td>
</tr>
<tr>
<td>6</td>
<td>C\textsubscript{11}H\textsubscript{7}N\textsubscript{4}O</td>
<td>8.7</td>
<td>41</td>
<td>2a</td>
</tr>
</tbody>
</table>

To examine the modification site assignment, the correlation score for each theoretical peptide set generated for the HSA peptide ALVLIAFAQYLQQCPFEDHVK was plotted as a function of nominal CRM position (Figure 4-4). Evidently, the maximal score is obtained when the CRM offset is placed at the 14\textsuperscript{th} residue which corresponds to a cysteine residue.
The accepted formula, \( C_{10}H_{12}N_2OCl \), corresponds exactly to the molecular formula of the CPPQI metabolite of TZD which has been identified as a possible hepatotoxic compound. The 16 Da RME can be reconciled with the nearest fragment (structure 2b in Scheme 4-3) by addition of an oxygen atom that accounts for the CYP2D6-mediated oxidation of chlorophenylpiperazine to form the immediate precursor to quinoneimine generation Scheme 4-1. Taken together, these results provide confidence that xenophile can identify CRMs from synthetic twin-ion LC-MS/MS data without prior knowledge of the metabolite identity.

4.6 Xenophile features, utilisation and availability

The non-targeted CRM identification algorithm and the targeted correlation algorithm, in addition to numerous other tools, have been combined into a single software suite and the source code can be accessed via the internet at https://github.com/mgleeming/Xenophile and an installation guide for the open-source Linux-based operating system Ubuntu (v14.04) is included in section 9.1 on page 194. The Xenophile bundle provides two ways for users to access the core analysis and review functions: 1) a Python programming library, and 2) a graphical user interface (GUI).

4.6.1 Xenophile python library

To facilitate incorporation of Xenophile methods into custom software, a python library is provided that give users easy access to various functions that enable data analysis and review.
The following code examples are executed directly from an interactive Python interpreter (i.e. don’t enter the ‘>>>’ at the beginning of each line).

In some cases, it is desirable to interactively inspect the results of HiTIME scoring. To import the Xenophile functions, load a HiTIME results file (named ‘exampleHT.dat’ in this case) and print some summary statistics:

```
>>> import xenophile.xenophile as xe
>>> htData = xe.loadResults('exampleHT.dat')
>>> xe.summarise(htData)

Number of points: 162785
Score Min, Max, Average: 1.62e-05, 20.20, 1.82
m/z Min, Max: 300.19015, 996.83270
rt Min, Max: 10.00, 39.82

10th percentile: 0.17
25th percentile: 0.47
50th percentile: 1.17
75th percentile: 2.43
95th percentile: 5.79
99th percentile: 9.76
```

These can be viewed as a heat map and interactively explored:

```
>>> import xenophile.xenophile as xe
>>> htData = xe.loadResults('exampleHT.dat')
>>> xe.showHeatMap(htData)
```

This produces window shown below. Clicking the button allows the user to pan and zoom allowing for close inspections of interesting data points. The image can be saved by clicking the button or, alternatively, by using the ‘saveFig’ option when calling ‘showHeatMap’.
The majority of these data points in this figure are very low scoring and are unlikely to correspond to twin-ions. It is desirable to remove these points for further processing and subsequent non-targeted protein adduct identification. Using the Xenophile library:

```python
>>> import xenophile.xenophile as xe
>>> htData = xe.loadResults('exampleHT.dat')
>>> htSubset = xe.removeLow(htData, 5)
>>> xe.showHeatMap(htSubset)
```

This produces the following figure containing only points with HiTIME scores greater than 5:

While interactive exploration of individual files is helpful to gain an understanding of a data set, in many cases multiple files will share similar characteristics and it is desirable to write scripts that automate the processing. This is particularly useful when many very similar samples need to be processed such as in the case of multiple replicates of different treatment groups.

If one wished to develop a python script to automate processing of multiple files, this could be accomplished using the xenophile utilities library.
import xenophile.xenophile as xe

mzMLfiles = ['treatment1.mzML', 'treatment2.mzML', 'treatment3.mzML']

for inFile in mzMLfiles:
    # run HT search
    resultsFile = xe.runHTSearch(inFile, mzDelta = 6.0201)

    # load HT results file
    data = xe.loadResults(resultsFile)

    # find 50th percentile
    pc50 = xe.getPercentile(data, 50)

    # remove data below this point
    dataPc50 = xe.removeLow(data, pc50)

    # find highest scoring peaks
    peaks = xe.findPeaks(dataPc50)

    # get top 10 scoring data points
    top10 = xe.getTopX(data, 10)

    print(top10)

Figure 4-5. Example python script that uses the Xenophile utilities library to automate multiple data processing steps.

A minimal script that can be used to automate the non-targeted reactive metabolite search given HiTIME and Mascot files is provided in Figure 4-6.
import xenophile.xenophile as xe

def main():
    # Input/output files
    htFiles = ['egHT1.dat', 'egHT2dat', 'egHT3.dat']
mascotFiles = ['egMascot1.dat', 'egMascot2.dat', 'egMascot3.dat']
outFiles = ['egOut1.dat', 'egOut2.dat', 'egOut3.dat',]

    # Input drug SMILES string
    APAP = 'CC(NC1=CC=O)C=C1)=O'

    # Allowed atom ranges
    atomDict = {
        'C': range(6,16),  'N': range(0,5),  
        'O': range(0,5),  'H': range(3,31)
    }

    for i in range(len(htFiles)):
        htFile = htFiles[i]
mascotFile = mascotFiles[i]
outFile = outFiles[i]

        xe.nonTargetedMetID(
            htInput = [[htFile, 2]],
mascotInput = [mascotFile],
mzBand = (76,360),
atomDict = atomDict,
drugSMILES = APAP,
outFile = outFile,
toMins = False,
HT_ms2_mz_tol = 0.5,
HT_ms2_rt_tol = 30,
ppmTol = 100)

    return

if __name__ == '__main__':  
    main()  

Figure 4-6. Minimal python script to automate consecutive non-targeted CRM identification for 3 data files.

4.6.2 Xenophile graphical user interface
While the Python scripting library enables the most customised and high-performance utilisation of Xenophile, it is recognised that users may be unfamiliar with command line environments. To lower the barrier to entry for the Xenophile software, a GUI has been developed to allow execution of data analysis procedures and review of results in a software environment that is intuitive to users that do not wish to incorporate these functions into custom scripts. The GUI is implemented using PyQt4, which are a set of python binding for the Qt4 C++ library. The GUI is divided into 4 main sections:
1) HiTIME searching and results post-processing
2) Non-targeted reactive metabolite identification
3) Targeted CRM-protein identification
4) LC-MS data file viewer

Each of these 4 sections contains a number of subsections that are accessed via various tabs, which are summarised in Table 4-5. Various screenshots of the GUI are provided in the Appendix.

**Table 4-5.** Functions contained within each tab within the graphical user interface of the Xenophile software. Images of the GUI are provided in the Appendix.

<table>
<thead>
<tr>
<th>Xenophile GUI tab</th>
<th>Key functions</th>
<th>Image(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HiTIME methods</strong></td>
<td>• Setup and run HiTIME searches</td>
<td>9-1,</td>
</tr>
<tr>
<td></td>
<td>• Subtract control sample HiTIME data from treatment data</td>
<td>9-2</td>
</tr>
<tr>
<td></td>
<td>• Perform results post-processing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Identify local maxima, plot EICs and extract MS spectra</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Interactively review results of HiTIME post-processing and save a list of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>accepted hits.</td>
<td></td>
</tr>
<tr>
<td><strong>Non-targeted CRM identification</strong></td>
<td>• Setup non-targeted CRM search</td>
<td>9-3,</td>
</tr>
<tr>
<td></td>
<td>• Review results</td>
<td>9-4</td>
</tr>
<tr>
<td><strong>Targeted CRM identification</strong></td>
<td>• Setup targeted protein adduct search</td>
<td>9-5,</td>
</tr>
<tr>
<td></td>
<td>• Review results</td>
<td>9-6</td>
</tr>
<tr>
<td><strong>mzML viewer</strong></td>
<td>• Interactively view LC-MS/MS files and perform common data analysis functions</td>
<td>4-1</td>
</tr>
<tr>
<td></td>
<td>such as plotting EICs.</td>
<td></td>
</tr>
</tbody>
</table>

An mzML file viewer is contained within the GUI (Image 4-1, not accessible through the Xenophile command line interface). This allows users to interactively interrogate raw mass spectrometry data files in the mzML format which is the common file type used in this software as well as almost all open source software of this nature. This provides users with the ability to view TICs and scroll through mass spectra at any point. Users can save images of spectra or chromatograms in a number of different file formats or export Cartesian coordinates that describe these traces for later reconstructions with specialist graphing software. To facilitate
analysis of twin-ion features, which is expected to be a common task in this research, functions that allow for plotting of EIC traces are also included.

For best results when using this file browser, it is highly recommended that indexed mzML files employed. These include specialised headers within the data file that allow specific spectra to be randomly accessed without the need to iterate through the entire file allowing for smooth and responsive scrolling through the data.\textsuperscript{251, 252} Headers can be reconstructed if not present but this can be a time-consuming process and it is generally recommended that headers be included with all files.

![Image 4-1. Screen capture of the mzML data file browser included in the Xenophile graphical user interface.](image)

Image 4-1. Screen capture of the mzML data file browser included in the Xenophile graphical user interface.
4.6.3 Xenophile licence
The development of the software presented here has benefited significantly from the utilisation of numerous open source function libraries including pymzml, pyteomics and RdKit. As such, the Xenophile source code has been made freely available under the 3-clause BSD license. The 3-clause BSD license allows any person, company or organisation to freely use, modify and re-distribute the Xenophile software almost unrestricted. This is advantageous as other developers are free to incorporate Xenophile into different software or make improvements to the existing code for the benefit of the scientific community.

4.7 Comparison to other methods
There have been reports from other researchers that employ stable isotopes in a twin-ion-like fashion to identify the protein adducts of reactive metabolites and it is useful to compare and contrast these methods to that presented here.

For example, Hanzlik and co-workers have identified covalent protein adducts produced in vitro following thioacetamide S-Oxide (TASO) incubation with rat hepatocytes and in vivo with thiobenzamide (TB) administration to rats. In these studies, parallel cell or animal treatments were performed with either 1) radiolabelled (14C) drug, or 2) a 1:1 mixture of unlabelled:stable isotope labelled drug (TB:2H5TB or TASO:13C22H3TASO). Following treatments, protein extracts were separated via 2-dimensional SDS-PAGE. Phosphorimaging of gels containing proteins from 14C drug treatment revealed the locations of proteins that carried covalent modification. This data was then used as a guide to excise the equivalent spot on the gel containing proteins from the combined heavy/light drug treatment groups which were then subjected to protease digestion and LC-MS/MS analysis. Identification of LC-MS peaks corresponding to peptides separated by a nominal mass difference of 5 Da then indicated specific sites of drug modification. In both cases, many proteins were identified by this method. For example, 63 proteins were found to be covalently modified by reactive metabolite derived from TB and 88 from TASO.

* The only restrictions imposed by this license are that the names and affiliations of the original authors cannot be used to endorse derived products and that the authors cannot be held liable for any damages arising from use of this software. Note that Xenophile includes the JSME and MascotParser libraries for convenience. If redistributing this code, the licensing restrictions of these packages must also be considered. The licenses for JSME and MascotParser are included in the ‘license.txt’ file in the Xenophile GitHub repository.
However the absolute requirement for radiolabelled drug, gel-electrophoresis and in-gel protein digestion pose significant resource requirements and, in addition, the necessity to conduct replicate parallel treatments with radiolabelled and 1:1 unlabelled:stable isotope labelled drug essentially doubles the number of experiments that need to be conducted. While this methodology is clearly capable of identifying drug-modified proteins, the significant labour and resource requirements pose a substantial barrier to widespread use.

An interesting method published recently by Humphreys and co-workers sought to identify the proteins modified by acetaminophen (APAP) formed in mouse liver microsomes. These experiments involved conducting separate microsomal incubations with either APAP or $^{13}\text{C}_2\text{^{15}}\text{N}$ APAP followed by shotgun LC-MS/MS analysis of digested protein extracts. To selectively detect peptides modified by APAP, base-peak chromatograms (BPC) for APAP and $^{13}\text{C}_2\text{^{15}}\text{N}$ APAP were subtracted from one another leaving only peptides that differ between samples. As the only artificial difference between the two samples was the use of either labelled or unlabelled APAP or enriched APAP, peptides that are not modified during this treatment should be identically detected in both samples and then eliminated by subtraction. Modified peptides, however, would be retained due to the $m/z$ offset between heavy and light APAP and the resulting data sets should contain peaks corresponding to drug-peptide adducts. This analysis identified three proteins that were covalently modified following incubation of APAP or $^{13}\text{C}_2\text{^{15}}\text{N}$ APAP in mouse liver microsomes.

This method is advantageous in that it avoids the use of radiolabelled drug and does not require separation of proteins via SDS-PAGE. However, the ability of this background subtraction method to identify modified peptides is highly dependent on the uniformity of the two data sets and how well the matrix of one sample is replicated by the other. In this case, inconsistencies between the samples would be detected as peaks in the subtracted data sets. Liver microsome incubations that were used in this proof-of-principle study are generally reproducible providing high-quality data that enables this analysis. In cases where there are differences between heavy and light samples BPCs may not align and many peaks may be observed upon background subtraction resulting in data sets that are difficult to interpret. This may become problematic if subtle variations in LC-MS analysis occur (e.g. retention time offsets) or upon expansion of this method to in vivo experiments that frequently have large variation between biological replicates resulting from inter-individual differences in protein expression and post-translational modification.
By contrast, the methodology presented here that combines the use of twin-ion experiments with the comprehensive Xenophile data analysis software suite does not require prior separation of proteins by SDS-PAGE or radiolabelled tracers. Additionally, as all twin-ion information necessary to assign reactive metabolites and modified proteins is present within a single sample, subtle differences between data sets resulting from biological or technical variability are not likely to have a significant impact on the results. This also eliminates the duplication of biological replicates necessary in both cases above wherein two separate sets of biological replicates with drug treatment were required to accumulate the data necessary for assigning proteins.

4.8 Conclusion
In summary, a combined experimental and computational workflow has been described that allows for the identification unknown reactive metabolites and the protein adduction targets of these compounds at residue resolution without prior knowledge of the metabolic pathways or residue reactivity profiles. This method takes advantage of relationships between twin-ion and native peptides within a given sample to assign then identity of reactive metabolites, which can then be used to parameterise standard database searches. To aid in the analysis of these data, a software toolkit, called Xenophile, has been developed that facilitates the processing and reviewing of both LC-MS/MS data obtained from twin-ion experiments and peptide database searching algorithms. Using synthetic data, Xenophile correctly identified both the structure of the reactive metabolite of trazodone as chloroquinoneimine and the site of adduction to HSA digest peptides as a Cys residue. This methodology simplifies the experimental approach for identification of protein adducts by eliminating the need for radiotracers, antisera, 2D SDS-PAGE and specific knowledge of the structure, reactivity and fragmentation pathways of the reactive metabolite(s).
Preface

The work presented in this chapter has been published in *Analytical Chemistry*:


5.1 Summary

This chapter describes the use of the twin-ion methodology described in the preceding sections to identify the protein targets of paracetamol adduction in rat liver microsomes. APAP and $^{13}$C$_6$-APAP were incubated with rat liver microsomes followed by global trypsin digestion and LC-MS/MS analysis. Numerous HiTIME hits were detected in MS$^1$ data for peptides in both the +2 and +3 charge states. After Mascot analysis, processing of these combined datasets using the non-targeted CRM identification algorithm contained within the Xenophile software resulted in the confident assignment of a reactive metabolite with the formula C$_8$H$_7$NO$_2$ bound to cysteine residues which corresponds to the well-studied electrophilic paracetamol metabolite N-acetyl-p-quinoneimine (NAPQI). Using this data to parameterise subsequent peptide database searches, 8 peptides were found to be assigned as having NAPQI modifications that also corresponded to LC-MS data regions displaying twin ions. These 8 peptides represent 7
unique proteins including many that have been identified in previous studies as targets for adduction by NAPQI or reactive metabolites of other xenobiotics.

5.2 Introduction

Given the complexity of any potential link between protein adduction and immune reactions to medications, numerous researchers have suggested that surveying the protein adduction profile of a range of drug molecules may allow patterns to be discerned that link adduction of specific targets to pharmacological outcomes. Indeed, construction of such a catalogue is underway. Hanzlik and co-workers curate a website that contains proteins known to be adducted by a range of compounds. However common methods to determine the protein adduction targets of reactive metabolites including incubation of radiolabeled drug followed by 2D SDS-PAGE and LC-MS/MS remain laborious and, as such, studies to identify specific targets are relatively rare. Introduction of an experimental method to identify modified proteins that is simple, accurate and rapid would significantly lower the barriers to the routine identification of drug modification targets enabling large-scale collection of this data that would allow critical targets to be identified.

The combined experimental methodology and data analysis tools presented in Chapter 5 overcome numerous hurdles encountered using alternative methods. This twin-ion and Xenophile workflow does not require radiolabelled compounds, antisera, SDS-PAGE or specific knowledge of metabolism or ion fragmentation pathways and can determine the identity and adduction site of reactive metabolites a priori. The Xenophile non-targeted CRM analysis algorithm correctly identified the reactive metabolite CPPQI from simulated trazadone-adducted HSA indicating successful development and implementation of the functionality.

However, even simple in vitro bioactivation assays utilising liver microsomes or S9 fractions can contain over one thousand different proteins. Global protein digests performed on extracts of these systems will likely produce thousands of unique peptides leading to highly complex LC-MS data. In this case, thousands of peptides are likely to be detected and, inevitably, some of these will be of low abundance, have overlapping isotope distributions and chromatographic profiles and imperfect MS/MS fragmentation patterns. The complexity of LC-MS data from shotgun proteomics experiments may create challenges in the implementation of the Xenophile algorithms. The key to the success of the non-targeted CRM
identification algorithm is the ability to accurately match the MS/MS spectrum of a twin-ion peptide to that of its native counterpart. With increasing complexity of the sample, more combinations of peptides will need to be searched and it is unclear if this method will be successful when the search space becomes large. One possibility is that matches by random chance for unrelated peptides produce high correlation scores resulting in many false assignments that would render this method unreliable. Moreover, co-elution of peptides of similar mass may result in overlapping LC-MS signals. If this occurs for twin-ion peptides, distortions to the peak shape will arise that may hinder identification of CRM-modified peptides by HiTIME.

Given these possibilities, we sought to test the Xenophile against complicated experimental data. The next sections describe the assessment of the covalent protein adduction targets of APAP and $^{13}$C$_6$ APAP in rat liver microsomes that are routinely used in xenobiotic metabolism assays.

5.3 Results and discussion

5.3.1 Microsomal bioactivation of paracetamol

APAP and $^{13}$C$_6$ APAP were incubated with liver microsomes that are a rich source of phase 1 metabolic enzymes and are known to bioactive APAP to the reactive metabolite NAPQI.$^{322,353}$ Following a 3 h incubation, the proteins were separated by precipitation, digested with trypsin, and analysed by LC-MS/MS using the Orbitrap mass spectrometer (described in section 8.1.1.2). Mascot analysis of the data against the Uniprot database allowing variable oxidation of methionine and carbamidomethylation of cysteine resulted in the identification of 720 and 660 proteins that were common to all three replicates in APAP and VC treatment groups respectively (Figure 5-1). However, adduction of any metabolites formed from APAP would change both the total mass of the peptide and induce an offset in the ladder sequence ions formed in MS$^2$ experiments resulting in a mismatch between the experimental spectra and the theoretically predicted fragments generated by database searching. Adducted peptides will remain unassigned and, accordingly, this data does not account for the possibility of reactive metabolites covalently reacting with microsomal proteins.
Figure 5-1. Venn diagrams indicating the number of proteins identified by database searching of LC-MS/MS data sets collected on the orbitrap mass spectrometer for A) paracetamol, and B) control microsomal protein digests.

The MS data were thus processed using HiTIME software capable of identifying twin-ion signatures which are defined as a pair of co-eluting peaks that have the same abundance and are separated by a \( m/z \) spacing corresponding to the mass of the isotopic label.\(^{334}\) The heat maps produced after scoring with a mass delta of 3.01005 and 2.0067 Da (corresponding to the 6.0201 Da difference in the +2 and +3 charge states) for all replicates are shown in Appendix Figure 9-11 and Figure 9-12 respectively. HiTIME scoring reveals multiple ‘bright spots’ in the APAP incubations (Figure 5-2A) that lack comparable points control data (Figure 5-2B) indicating that the data in those regions closely match the expected twin-ion signature. Investigation of highly scoring points reveals that these are true twin-ion hits. For example, the HiTIME hit at \( m/z \) 752 and 47.8 minutes suggests a twin ion of \( m/z \) 752 and \( m/z \) 755. The EIC traces of these ions overlap significantly in the region of 47.8 minutes (Figure 5-2C) and two peptide signals, both of the 2+ charge state, are evident in the mass spectrum at this point (Figure 5-2D). These data indicate that this is a true twin-ion corresponding to a peptide modified by a reactive metabolite derived from APAP in the 2+ charge state. For vehicle control samples, similarly high-scoring data regions as those corresponding to true twin-ion peptides in APAP treatment data were not observed (Appendix Figure 9-11 and Figure 9-12) and no convincing case of a twin-ion peptide could be found upon inspection of EIC traces and mass spectra following post-processing (Appendix Table 9-7).
5.3.2 Non-targeted searching of APAP-microsome data

To detect the APAP-derived reactive metabolite, the APAP treatment HiTIME and Mascot data was analysed using the non-targeted CRM detection algorithm of Xenophile using the parameters provided in Appendix Table 9-8 and Table 9-9. In this process, MS² spectra associated with peptides observed as twin-ions (unassigned by peptide database searching) are correlated with their non-CRM-modified counterparts and the mass difference between these highest ranked pairs is taken as a potential CRM. The three highest scoring hits for each of the three APAP treatment replicates are provided in Table 5-1 and Appendix Table 9-10 and Table

**Figure 5-2.** Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a twin-ion spacing of 3.01005 aimed at mining doubly charged peptides modified by an APAP metabolite. A) APAP treatment, B) Vehicle control samples. C) EIC of the ‘light’ (m/z 752.33, red trace) and ‘heavy’ (m/z 755.34, blue trace) peaks indicated by the HiTIME hit at RT = 47.85 min. D) Mass spectrum at the region of EIC peak maximum. Analogous heat maps for remaining replicates of both APAP and VC treatment data are shown in Appendix Figure 9-11.
9-11 respectively. Taken together, eight of these nine results detect a CRM with a mass of 149.05 Da that, in each case, arises from covalent modification of a cysteine residue. Two possible molecular formulae are identified that fit within the allowed atom stoichiometries and ppm error tolerances. These are C$_8$H$_7$N$_2$O$_2$, identified as a possibility in all 8 cases, and C$_6$H$_5$N$_4$O which is identified in 5 out of the 8 cases. The average ppm error for these possible formulae are similar (8.0 and 6.1 respectively) however the residual mass errors calculated versus the input APAP structure differ considerably (2 and 86 respectively). This indicates that the former candidate, C$_8$H$_7$N$_2$O$_2$, is very similar in composition to the APAP input molecule (APAP: C$_8$H$_9$NO$_2$). Given that (i) C$_8$H$_7$N$_2$O is consistently identified in multiple HiTIME hits and across all replicates, (ii) this formula is closely related to the APAP input molecule, and (iii) no chemically reasonable structure can be devised for the alternative, C$_6$H$_5$N$_4$O is taken as the stoichiometry of the reactive metabolite formed from APAP in these incubations.

Table 5-1. Top three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 1 of 3)

<table>
<thead>
<tr>
<th>Hit</th>
<th>CRM mass (Da)</th>
<th>Sequence</th>
<th>Modification Site</th>
<th>Formula</th>
<th>ppm</th>
<th>Residual mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>149.0494</td>
<td>VFANPEDCAGFGK</td>
<td>C (8)</td>
<td>C$_8$H$_7$N$_2$O$_2$</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>149.0481</td>
<td>EFTPCAQAAFQK</td>
<td>C (5)</td>
<td>C$_8$H$_7$N$_2$O$_2$</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C$_6$H$_5$N$_4$O</td>
<td>12</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>149.0464</td>
<td>TIQLNVCNSEEVEK</td>
<td>C (7)</td>
<td>C$_6$H$_5$N$_4$O</td>
<td>0.4</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C$_8$H$_7$N$_2$O$_2$</td>
<td>-9</td>
<td>2</td>
</tr>
</tbody>
</table>

To interrogate the data underpinning this result, the MS/MS spectra for the CRM-modified and native peptides were extracted and compared. For example, the native peptide EFTPC$_{CAM}$AQAAFQK (m/z 699.3) and its twin-ion counterpart (m/z 745.3) that were used to assign a CRM mass of 149.0481 Da are shown in Figure 5-3. A prominent y-ion series is observed and the m/z values of these sequence ions are common to both spectra for y$_2$ through to y$_7$. However, these diverge beyond the y$_7$ ion at m/z 763 (assigned as [AQAAFQK+H]$^+$). For the native peptide spectrum (Figure 5-3A), the y$_8$ and y$_9$ ions, that now include a cysteine residue, are located at m/z 923 and 1020 respectively. While no prominent fragments are observed at these positions in the MS/MS spectrum of the twin-ion peptide (Figure 5-3B), two
additional fragments appear offset by +92 Da at m/z 1015 and 1112 indicating that the cysteine residue is differentially modified in the two peptides. Because the unmodified peptide is alkylated at the Cys residue (+57 Da relative to unalkylated peptide) and the twin ion peptide has a CRM adducted to the Cys residue (+92 Da relative to the alkylated cysteine), the mass of the CRM (149 Da) is calculated from the sum of 57 and 92 Da. The assigned formula, C₈H₇N₂O, is the molecular formula of the well-studied electrophilic NAPQI metabolite of APAP that is known to bind covalently to the side-chains of reduced cysteine residues in proteins.²⁶⁰ It should be emphasised that this metabolite has been identified directly from shotgun proteomics data while making essentially no assumptions as to the identity of the metabolite, the protein identity, or the amino acid target of the modification.

**Figure 5-3.** HCD MS² spectra of A) non-twin-ion peptide assigned by Mascot as (carbamidomethyl-C) EFTPCAQAAFQK²⁺ at m/z 699.3 at 37.3 minutes, and B) twin-ion peptide at m/z 745.3 at 43.9 minutes from LC-MS/MS analysis of APAP treated microsomal protein digests.
5.3.3 Targeted search of APAP-microsome data

Having determined the stoichiometry of the reactive metabolite formed from APAP, and the amino acid that was modified by this metabolite, this data can be used to direct subsequent protein database searches. Mascot searches of the LC-MS/MS data were replicated with the addition of C$_8$H$_7$NO$_2$ and $^{13}$C$_6$C$_2$H$_7$NO$_2$ (corresponding to the molecular formulae of NAPQI and $^{13}$C$_6$ NAPQI) as variable modifications in order to directly identify peptides carrying the NAPQI modification.

In all cases, a large number of peptides were identified by Mascot searching as APAP adducts. For the APAP treatment samples, an average of 55 peptides were assigned as having either APAP or $^{13}$C$_6$ APAP and, interestingly, an average of 41 were assigned for vehicle control samples that were not exposed to APAP. It therefore seems highly unlikely that these can correspond to legitimate APAP-peptide adducts. For VC samples, no APAP assigned peptides displayed the twin-ion signature upon manual inspection their associated MS$^1$ spectra and this was also true of many hits in the drug treatment samples.

Attempting to understand the origin of these apparent false assignments, the peptides from control samples that had been assigned APAP modifications were analysed in detail. For these peptides, 100% of the APAP assignments (41 peptides on average per sample) were in the form of $^{13}$C$_6$ APAP, i.e. modifications were consistently assigned as the heavy rather than light forms. This suggests that these false assignments are likely to be a systematic event arising from certain combinations of PTMs rather than random and unpredictable matches.

It was suspected that some combinations of carbamidomethylation and methionine oxidation could produce a mass offset that would be close to that induced by $^{13}$C$_6$-APAP. A large number of peptides were assigned modifications of 1 $^{13}$C$_6$ APAP and 1 methionine oxidation which have a combined mass of 171.0627 Da. Interestingly, the combined mass of 3 carbamidomethyl groups is 171.0644 Da, which is similar to the mass of $^{13}$C$_6$APAP and oxoMet modifications (within ca. 1 ppm for a 1500 Da peptide in the 2+ charge state). It appears that many of these peptides have been carbamidomethylated at a site other than cysteine explaining the false assignment in the first instance. Indeed, reactivity of iodoacetamide to multiple non-cysteine sites has been identified previously.$^{354}$ These data indicate that targeted database searching alone may not always reliably identify the products of CRM-peptide adduction, and may in fact give rise to false positives.
In these experiments, it is known that peptides modified by APAP CRMs will approximate a twin ion shape. Thus, true modified peptides should be assigned as such by database search algorithms and also display the expected twin-ion shape allowing APAP-modified peptides to be identified despite the presence of multiple false-assignment by Mascot. To restrict the data to only those peptides that meet these criteria, the targeted CRM correlation algorithm included in the Xenophile software bundle was employed.

Running this algorithm with HiTIME and Mascot data from microsomal protein digests using the parameters supplied in Appendix Table 9-13, 8 peptides were found to be assigned as CRM adducts by Mascot that also appear as twin-ions (Figure 5-4, Table 5-2) indicating that these are true CRM-modified peptides. For 5 of these 8 cases, both the ‘light’ and the ‘heavy’ peptides were separately mass-selected during data-dependent MS² acquisition and subjected to HCD. For example, the MS² spectra of the APAP and ¹³C₆APAP adducts of the peptide EFTPCAQAQAFQK²⁺ are shown in Figure 5-4A and Figure 5-4B respectively. The series of ‘y’ ions at m/z 1015, 1112 and 1213 in Figure 5-4A is offset from the analogous ions in Figure 5-4B at 1021, 1118, 1219 by 6 Da which is equal to the nominal mass difference between APAP and ¹³C₆APAP. Taken together, these data provide a high level of confidence that the peptides assigned as APAP adducts by database searching and as twin-ions in MS¹ data are indeed true products of covalent bond formation between drug reactive metabolites and proteins.
Figure 5-4. HCD MS$^2$ spectra of peptides assigned by targeted correlation of HiTIME and Mascot peptide assignments. EFTPC(X)AQAFQK$^{2+}$ where A) X = APAP, B) X = $^{13}$C$_6$ APAP and C) X = carbamidomethyl (C$_2$H$_3$NO, from iodoacetamide treatment of reduced proteins).
Table 5-2. Summary of the peptides assigned by directed Mascot search as carrying NAPQI modification. APAPL and APAPH refer to the natural abundance APAP and $^{13}$C$_6$ APAP respectively.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Modification</th>
<th>Site</th>
<th>Accension</th>
<th>Protein</th>
<th>Unique Peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCDLGGLWR</td>
<td>APAP L/H</td>
<td>C2</td>
<td>FMO1_RAT</td>
<td>Hepatic flavin-containing monooxygenase 1</td>
<td>16</td>
</tr>
<tr>
<td>EFTPCAQAFAQK</td>
<td>APAP L/H</td>
<td>C5</td>
<td>HBB1_RAT</td>
<td>Hemoglobin beta-1 chain</td>
<td>13</td>
</tr>
<tr>
<td>VFANPEDCAGFGK</td>
<td>APAP L</td>
<td>C8</td>
<td>MGST1_RAT</td>
<td>Microsomal glutathione S-transferase-I</td>
<td>8</td>
</tr>
<tr>
<td>VFANPEDCAGFGKGENAK</td>
<td>APAP L/H</td>
<td>C8</td>
<td>MGST1_RAT</td>
<td>Microsomal glutathione S-transferase-I</td>
<td>8</td>
</tr>
<tr>
<td>HIGDGCHLTR</td>
<td>APAP L/H</td>
<td>C6</td>
<td>MET7B_RAT</td>
<td>Methyltransferase-like protein 7B</td>
<td>8</td>
</tr>
<tr>
<td>TIQLNVCNSEEVEK</td>
<td>APAP L/H</td>
<td>C7</td>
<td>BDH_RAT</td>
<td>3-Hydroxybutyrate dehydrogenase</td>
<td>12</td>
</tr>
<tr>
<td>KHHLCGETEEER</td>
<td>APAP H</td>
<td>C5</td>
<td>GSTM1_RAT</td>
<td>Glutathione S-transferase Mu 1</td>
<td>18</td>
</tr>
<tr>
<td>QVADEGDALVAGGVSQPSYLSCK</td>
<td>APAP L</td>
<td>C23</td>
<td>BHMT1_RAT</td>
<td>Betaine-homocysteine S-methyltransferase 1</td>
<td>11</td>
</tr>
</tbody>
</table>
These 8 modified peptides comprise partial sequences of 7 unique proteins that are listed in Table 5-2. Several of these proteins have been identified by other researchers as targets of reactive electrophiles, which provides additional confidence in these assignments. For example, glutathione-S-transferase was first identified as a target for reactive APAP metabolites by radiometric counting in 1981\textsuperscript{355} and its constituent peptides VFANPEDCAGFGK and VFANPEDCAGFGKGENAK (and homologous sequences from different species) have since been identified numerous times to be modified at the cysteine residue by NAPQI\textsuperscript{301, 322, 351} and electrophillic metabolites of other drug compounds.\textsuperscript{356} Furthermore, the triazine herbicide atrazine has been found to form covalent adducts with the haemoglobin \(\beta\) chain peptide EFTPCAQAADFQK in Sprague Dawley rats\textsuperscript{357} and the FMO1 peptide SCDLGGLWR was identified as a target of reactive metabolites formed from model furan-containing compounds in rat liver microsome incubations.\textsuperscript{356} The global biological and toxicological effects of reactive metabolite adduction to these specific proteins are largely unclear. Previous studies have demonstrated dose dependent inhibition of the glutathione conjugation activity of GST by APAP which has been linked to NAPQI formation and adduction to Cys47\textsuperscript{358} which may have an impact on GSH/GSSG balance. Other proteins identified have roles in drug metabolism (FMO1), methionine synthesis (BHMT) and fatty acid catabolism and energy balance (BDH).

The number of proteins identified is similar to and slightly greater than the number of proteins identified in closely related studies that used more targeted methodologies to identify APAP-protein adducts from rat liver microsome incubations.\textsuperscript{301} In addition, many reactive metabolites, such as those derived from diclofenac and halothane are known to form specific adducts with selected targets despite the presence of large numbers of other proteins.\textsuperscript{143} These previous studies are consistent with our results, which indicate that NAPQI forms relatively specific adducts with a limited number of microsomal proteins under these conditions.

Interestingly, the non-targeted PTM identification algorithm implemented here did not identify any adducts of APAP to tyrosine, tryptophan and methione residues despite the finding in Chapter 3 that NAPQI was reactive toward these sites. This is consistent with the incubations of GAIL-X-GAILR (X = C, Y, W, M) wherein NAPQI adducts of these peptides were observed for X = C but not for any other target residue even under conditions of artificial thiol depletion. The reason for this was not investigated in detail but may be related to the reaction rate of NAPQI toward these targets and the size of the microsomal thiol pool.
5.3.4 False assignments and limitations

Importantly, it is difficult to determine the rates of false-negative twin-ion and CRM assignments directly from the experimental data presented above, i.e. have any true twin-ions and modified peptides been missed by this analysis? The non-targeted detection of reactive metabolites by Xenophile relies on two main events: 1) ionisation, fragmentation and detection of peptide ions by the mass spectrometer, and 2) successful identification of twin-ions and correlation of modified and native peptides. The former point is highly dependent on the instrumental setup and operating parameters employed for LC-MS analysis while the latter is a function of the Xenophile software.

To investigate the ability of the Xenophile suite to identify twin-ion signals and correctly assign CRM masses and elemental compositions in a controlled manner, semi-synthetic datasets were produced by superimposing artificially generated twin-ion signals onto experimental shotgun LC-MS data from the vehicle control treatments described above which contain no modified peptides. Briefly, for each doubly-charged cysteine-containing peptide assigned by Mascot searching of VC data sets an LC-MS signal was generated and superimposed on the experimental data at both the locations of the light and heavy peaks expected following NAPQI modification. This results in data sets that contain experimental signals for native peptides and corresponding synthetic twin-ions. MS² spectra were produced based on experimental data and written to MGF files. Detailed information on the production of these data sets is given in Section 8.5.5.

Semi-synthetic data were generated for each of the three control replicates, resulting in the incorporation of 560, 610 and 620 artificial peptide signals respectively. These data were scored with HiTIME and the resulting heat maps are shown in Figure 5-5. Comparing the target twin-ion locations with maxima in the HITIME data it is found that approx. 95% of the synthetic signals were identified (Appendix Table 9-13). Manual analysis of the remaining 5% revealed that many of these signals heavily overlapped with surrounding data resulting in large distortions to the peak shape. False-positive rates were ca. 10% for the top ca. 90% of semi-synthetic twin-ions as judged purely by HiTIME score (Figure 5-6) although it should be stressed that the utilities for rapid user review of hits provided in the Xenophile suite should improve these figures substantially.

Running the non-targeted CRM detection algorithm resulted in ca. 88% of the synthetic twin-ion singles assigned a mass consistent with NAPQI modification to within 20 ppm which
increases to ca. 98% upon relaxing the error threshold to 100 ppm (Figure 5-7, Appendix Table 9-13). The confidence in these assignments was investigated by analysing the difference in correlation score between the highest and second highest ranked peptide correlations which provides a metric of how easily the top ranked hit was chosen over next closest candidate (Figure 5-8). Confidence in CRM assignments was ca. 90% or better for the highest scoring hits across three replicates and decreased to ca. 60% for lower scoring hits indicating that the correct peptide correlations are well separated from their next closest counterparts.
**Figure 5-5.** Heat maps produced by HiTIME scoring of the replicates of semi-synthetic LC-MS data sets (A-C) with a doublet spacing of 3.01005 aimed at mining doubly charged twin-ion peptides. Black boxes indicate locations of artificially introduced twin-ion signals.
**Figure 5-6.** Cumulative number of false positive HiTIME hits as a function of target rank for synthetic twin-in peptides.

**Figure 5-7.** Histogram of CRMs mass ppm errors assigned by non-targeted reactive metabolite searching from the true value of 149.04713 Da. Counts show average values from across the three replicates of semi-synthetic twin-ion data and errors are quoted as +/- one standard deviation.
Figure 5-8. Confidence in the CRM masses assigned by the non-targeted reactive metabolite detection algorithm for three replicates (A-C) of semi-synthetic twin-ion data. Confidence values are calculated from the normalized peptide correlation scores according to $\frac{C_1 - C_2}{C_1} \times 100$, where $C_1$ and $C_2$ are the correlation scores for the highest and second highest ranked hits respectively.
Figure 5-9. A) Number of formulae for theoretical reactive metabolites generated from 1337 molecules in the ZINC15 FDA approved drugs list. B) Histogram summarising the average number of possible formulae for molecules in 50 Da bins.

The selection of the lead chemical formula corresponding to a candidate CRM may be hampered by large numbers of possible formulae with similar masses and this may become particularly pronounced for large modifications. The search space may be refined using known characteristics of the input molecule and metabolic processes that lead to reactive metabolite formation. To profile the change in the number of formulae for a range of molecules, candidate formulae were generated for compounds in the ZINC15 database of FDA approved small molecule drugs\(^{359}\) that had been modified by randomly selected metabolic reactions that are known to lead to electrophilic metabolites. The number of hits within 100 ppm of the target mass rises rapidly with molecular weight reaching \textit{ca.} 100 for 500 Da candidates however this decreases to \textit{ca.} 20 hits when tolerances are tightened to 20 ppm (Figure 5-9). Additional refinements can be made to further reduce the number of possibilities. For example, the RME value represents the portion of the molecular mass that is not explained by input structure and is expected to be closely related to the mass changes induced by bioactivation processes. Therefore, these values should only lie in certain regions that correspond to common metabolic transformations. For example, if the RME values were restricted to < 20 and between 160 and 200 (representing hydroxylation and glucuronidation) the number of hits is reduced to \textit{ca.} 5 for a 500 Da compound which is feasible for manual user review.

It should be noted that reproducing the complexity and nuances of experimental data with synthetic spectra is difficult and therefore these results likely represent a best-case scenario. However, fundamentally, the detection limit for CRM-modified peptides using the software
approach presented here should be similar to that for any standard peptides using typical workflows. That is, detection is limited by the ability of the mass spectrometer to ionize, fragment and detect a peptide. As an approximation of the sensitivity of this experiment, *ca.* 4 µg of protein was injected on to the column resulting in the detection of 1021 proteins in each sample (on average) with a mean molecular weight of 52,794 Da. This equates to an average protein loading of *ca.* 75 fmol. It should be noted that CRM-modified peptides identified here were likely of substantially lower abundances as CRM adduction generally results in low yields. For example, the APAP-modified peptide VFANPEDCAGFGK (Figure 5-2C and Figure 5-2D) is detected at *ca.* 20-fold lower abundance than its native counterpart.

One limitation of the software presented here is that reactive metabolites with a heavy/light offset of *ca.* < 4 Da will produce twin-ion signals with overlapping isotope distributions which may distort the target shape. This would hinder detection of twin-ion signals using HiTIME, which form the basis of the non-targeted detection algorithm. Computational strategies are currently being explored that may account for these cases. One such technique is described on page 151 below and may be included in future releases of the Xenophile software.

5.4 Future work

During the development of the Xenophile software and the associated methodologies presented here, numerous concepts were explored and tested in preliminary studies that may offer the potential for enhanced functionality. Many simple improvements could be made. For example, different functions and approaches for matching sequences to MS² spectra of twin-ion peptides could be investigated and the Xenophile source code could be refined for greater speed and efficiency. Two more substantial points that offer the potential for significant improvement are discussed in detail below.

5.4.1 Improving metabolite prediction techniques:

In the non-targeted detection module, CRMs are considered to be derived from discrete fragments of the administered drug molecule that are generated by simple disconnection of rotatable bonds. This is important as the RME of candidate CRM formulae – which aids in assignment of metabolites – is calculated based on the formulae of these fragments. Such a simple fragmentation model may be a reasonable approximation of the formation of many CRMs but a more comprehensive and accurate list of structure candidates could be produced by employing more advanced metabolite prediction software. For example, MetaPrint2D-
React developed by Adams et al.\textsuperscript{360} analyses circular atom environment descriptors to determine likely phase I and II metabolites of an input structure. Unfortunately, this tool is not currently available as a command line application and would be difficult to incorporate into the software developed here. Alternative software packages exist such as SyGMA,\textsuperscript{361} TIMES\textsuperscript{362} and MetaSite\textsuperscript{363} but these are either commercial, lack some necessary functionality or are difficult to integrate into custom software.\textsuperscript{364}

5.4.2 Isotope washout effects for small modifications

In the case of APAP-protein adducts presented above, the mass difference between ‘light’ and ‘heavy’ compounds was 6.0201 Da which causes the separation between the twin-ions to be greater than the natural isotope distribution of most tryptic peptides facilitating detection by HiTIME. However, in some cases, the nature of the reactive metabolite or the labelling strategy employed may limit the attainable mass separation between the twin-ions. In these cases, the mass separation between the ions would be lower than the width of the isotope distribution of the light peak resulting in overlapping signals. This would interfere with HiTIME scoring as the ratio of the intensities between heavy and light peaks will differ from the expected values.

One method that could be used to recover twin-ion data in these cases is to employ the ‘averagine’ model\textsuperscript{365-367} to subtract expected isotope intensities from measured values and thus obtain estimates of deconvoluted intensities. As an example of a simple implementation of this, consider the peptide VFANPEDCAGFGK that is modified by an acetyl group with one carbon enriched as $^{13}$C. This gives rise to monoisotopic masses of the light and heavy isotopes are 1396.62018 and 1397.62353 Da respectively for [M+H]$^+$ ions. However, upon simulating the isotope distributions of this mixture, heavy overlap of multiple $^{13}$C isotope peaks is observed (Figure 5-10A).

Using the isotope peak ratio model developed by Valkenborg and co-workers,\textsuperscript{365} the ratio of heights between a peak $x$ and $x + 1$ is given by:

$$R(x + 1, m) = \frac{H(x + 1, m)}{H(x, m)}, \text{where } x = 0, 1, ..., n$$  \hspace{1cm} 5-1

with $x = 0$ referring to the monoisotopic peak. The intensity ratio of a given isotopic peak to the monoisotopic peak can then be determined by:
\[ R(x + 1, m) = \sum_{i=0}^{i=4} \beta_i \left( \frac{m}{1000} \right)^i \]

where the parameters for the values \( \beta_0 \) through \( \beta_4 \) are provided in the model description. The height of peak \( x + 1 \) is then given by:

\[ H(x + 1) = H(x, m) \sum_{i=0}^{i=4} \beta_i \left( \frac{m}{1000} \right)^i \]

The observed intensity of the monoisotopic ‘light’ ion could be used to predict the intensities of the light ion isotopic peaks which could then be subtracted from the observed intensities to give expected values for the light and heavy peaks individually. Applying this strategy to the peptide VFANPEDCAGFGK, the calculated isotope distributions and intensities are shown in Figure 5-10B. The corresponding heavy and light isotopes have approximately the same intensity (to within ca. 3%) which would be acceptable for HiTIME analysis.

**Figure 5-10.** A) Isotope distribution of VFANPEDCAGFGK where the Cys residue is modified by C\(_2\)H\(_2\)O or C\(^{13}\)CH\(_2\)O in equal proportions. B) Abundances of light and heavy peaks from the simulated peptide mixture in A calculated by subtraction of predicted isotope intensities. Note that heavy ions (blue peaks) have been offset from the light by \( m/z \) 0.15 to aid in visualisation. For both panels, relative peak heights are indexed to the intensity of the monoisotopic ion of the ‘light’ peptide.

### 5.5 Conclusion

The application of the Xenophile non-targeted protein CRM adduct identification methodology has been demonstrated using the bioactivation of acetaminophen to the reactive metabolite
NAPQI in hepatic microsomes. Non-targeted searching identified a reactive metabolite adducted to cysteine residues with a formula of $C_8H_7NO_2$ that is consistent with the molecular formula and known residue-reactivity of NAPQI. Upon subsequent Mascot searches, 8 unique NAPQI-modified peptides were identified that arise from 7 distinct proteins that are known targets of NAPQI adduction. Thus, this approach should prove useful for identifying the protein targets of chemically reactivate metabolites to within a single amino acid residue without prior knowledge of the metabolite(s) or protein target(s). It is anticipated that this powerful method should be generally applicable to study the mechanisms of toxicity induced by many different drugs and other xenobiotics.
6 METABOLISM OF AROMATIC ENVIRONMENTAL POLLUTANTS

Summary

The bioactivation of the industrial chemicals and environmental aromatic, PhX, pollutants benzene (X = H), toluene (X = Me) and bromobenzene (X = Br) have been examined using the twin-ion technique in combination with the HiTIME and Xenophile analysis tools developed throughout the preceding chapters. LC-MS and HiTIME analysis of small molecule fractions from rat liver microsome of PhX and ^13^C_6 PhX substrates resulted in the identification of at least one unique glutathione conjugate for each compound suggesting bioactivation to chemically reactive metabolites. Protein fractions from the same incubations were digested and analysed by LC-MS. HiTIME analysis and post processing through the Xenophile suite did not result in the identification of any twin-ion peptides for any sample. To verify that this is not the result of false-negatives in the HiTIME analysis, targeted database searches were conducted against the reactive metabolites identified from small molecule GSH adducts. No peptides identified by this analysis also presented as twin-ions in MS\(^1\) data. This suggests that no peptides were modified by PhX reactive metabolites, ionised and detected in LC-MS experiments. These results are rationalised on the basis of the extent of bioactivation of the PhX substrates studied here.
6.1 Introduction

With widespread incorporation of purified compounds and synthetic chemicals into industrial processes and consumer products throughout the 20th century, many parts of our natural environment have become contaminated with chemical waste and byproducts. These may include effluent from landfills or chemical runoff from agricultural practices. In some cases, these contaminants are long-lived chemical species that persist in the environment for years or even decades. Persistent pollutants, as these compounds are known, can enter into the food chain and become widely distributed from their point of origin.

Chemical pollution can have harmful health effects to humans. One well-known case of environmental contamination leading to human toxicity is the so-called Minamata syndrome. This was observed in Minamata City, Japan where, in the mid 1950’s, an unusually large number of patients began presenting to local hospitals with unexplained neurological symptoms. Subsequent house-to-house searches revealed many more people with similar symptoms and sparked concerns of an environmental cause. An investigation revealed that a nearby chemical plant was releasing large quantities of the extremely toxic compound, methyl mercury, into the Minamata Bay. Ingestion of this compound by marine life in the bay, that were a major food source for the local population, led to its transfer to humans where concentration in the central nervous system is known to result in neurological disorders.

While this is clearly an extreme example, many different chemicals released into the environment are likely to have health implications for wildlife and humans. Importantly, the pharmaceutical compounds, that have been the focus of this thesis thus far, undergo extensive safety and toxicity studies, whereas non-pharmaceutical compounds, that are often more widely distributed and are used in greater abundances, often do not. The concepts that are used to rationalise the toxicity and adverse effects of pharmaceutical compounds may also be applied to environmental contaminants. That is, processes such as bioactivation and protein covalent binding, which are through to be critical for the adverse responses to pharmaceuticals may also operate for pollutants. For example, the organic hydrocarbon benzene is present in petrol (gasoline), engine exhaust and tobacco smoke amongst others and is widely distributed throughout the environment. Benzene is a known carcinogen and relatively potent inducer of leukemia. Interestingly, it is postulated that benzene bioactivation and subsequent protein and DNA covalent binding is one of the primary events in the molecular pathogenesis of benzene-induced toxicities.
The various methodologies developed throughout this thesis could apply equally well to studying the metabolism of environmental contaminants as they do to the medicinal compounds investigated thus far. Accordingly, this section applies the techniques developed in the preceding sections to examining the small molecule metabolism and protein covalent binding of three different compounds released into the environment through industrial activity: benzene, toluene and bromobenzene. These compounds were selected as they each display toxic effects that are thought to be linked to metabolic bioactivation.

6.2 Results and discussion

Equal proportions of benzene (PhH), toluene (PhMe) or bromobenzene (PhBr) and their $^{13}\text{C}_6$ labelled isotopologues were incubated with rat liver microsomes in triplicate and the reactions were quenched after 3 h. For each PhX (X = H, Me, Br) or control treatments, a separate incubation was performed with 5 mM supplemental GSH added to the incubation mixture to specifically identify glutathione conjugates. To assess small molecule metabolite formation as well as protein adduction, the incubation mixtures were centrifuged and both the supernatants and protein pellets were independently analysed.

6.2.1 Microsomal PhX bioactivation and small molecule metabolism

The supernatants produced by centrifugation of microsome samples were analysed by LC-MS using the Q-TOF mass spectrometer. To detect twin-ions, LC-MS data were processed with HiTIME and the resultant heat maps are provided in Figure 6-1. Numerous hits were evident for non-GSH supplemented PhBr data (Figure 6-1A) that were confirmed upon manual inspection (Table 6-1). Many of these twin-ions corresponded to metabolites that retained the bromine atom which were readily identifiable due to the distinctive isotopic quartet created by the presence of both artificial $^{13}\text{C}_6$ label and natural isotope distribution of bromine. For example, the mass spectrum of the HiTIME hit at $m/z$ 494 and 22.2 mins in non-GSH supplemented PhBr incubations is shown in Figure 6-2A and a doublet of doublets is clearly visible. Interestingly, numerous twin-ions were observed at the same $m/z$ ratio but different retention times. For example, EIC traces for $m/z$ 494 and 500 (Figure 6-2B) showed one major peak at 22.2 mins and a series of 4 additional resolved or partially resolved peaks eluting between 19 and 22 mins suggesting the presence of multiple isomers. While some of these ions are of low abundance, inspection of the mass spectra associated with each of these EIC maxima clearly show the isotope quartet expected for a bromine-containing $^{12}\text{C}_6/^{13}\text{C}_6$ twin-ion
providing confidence that these signals do originate for PhBr despite their low intensity. HiTIME hits were not observed for the elution of the compounds at 19.0, 20.5 and 21.4 mins in Figure 6-2B which likely arises due to the extremely low abundance of these peaks. Two distinct maxima are also observed at 19.70 and 19.95 mins in the HiTIME data and EIC traces for the twin-ion at m/z 480/486. An isotope quartet is observed in the mass spectrum of this compound indicating retention of the bromine atom. Upon PhBr incubation with rat liver microsomes containing 10 mM supplemental GSH, the prominent ion at m/z 494 is not observed (Figure 6-1E) while the abundance of m/z 480/486 is markedly increased. Furthermore, two isomeric bromine-containing twin-ions are observed upon GSH supplementation at m/z 462, 19.95 mins and 19.70 mins that were not observed in non-supplemented samples. These data indicate that PhBr metabolism in rat liver microsomes gives rise to a complicated mixture of isomeric products and that the relative extent of formation of these depends on the concentration of glutathione.

To aid in elucidating the structure of these metabolites, MS/MS spectra of the twin-ions were recorded where possible. The core features of these spectra were similar in many cases and an example spectrum of the twin-ion at m/z 494 and 22.2 min is shown in Figure 6-2C. The precursor ion at m/z 494 loses 75 and 129 Da neutral species upon CID resulting in m/z 419 and 365 respectively. Peaks at m/z 76 and 130 are assigned to protonated forms of these fragments. The same fragmentation behaviour is observed upon MS/MS of m/z 500 (i.e. the $^{13}$C$_6$ isotopologue of this twin-ion (Figure 6-2C) wherein peaks at m/z 425 and 371 correspond to loss of 75 and 129 Da fragments respectively. Given that these neutral losses of 75 and 129 Da are known to be characteristic fragments of GSH conjugates$^{178, 180, 182, 183}$ and that GSH is present in even unsupplemented microsomes,$^{303}$ these data suggest glutathione conjugation to metabolites of PhX compounds.

On this basis, manually verified HiTIME hits that appear at substantially higher m/z than the PhX starting material could be assigned to GSH conjugates in all cases. For PhBr, one of the major metabolites (m/z 480/486) corresponds to mono-oxygenated bromobenzene followed by GSH conjugation. This suggests that mono-oxygenation installed a reactive functional group and, on this basis, the 172 Da intermediate is assigned to bromobenzene epoxide (BBE). Similarly, the twin-ion at m/z 494/500 could be assigned to GSH conjugation with a 186 Da metabolite consistent with bromobenzoquinone (BBQ-GSH). The twin-ion at m/z 462/468 retains an isotope quartet indicating retention of the bromine atom and appears to correspond
to formal addition of bromobenzene to GSH (BB-GSH). The lowest mass twin-ion observed at $m/z$ 416/422, does not appear as a quartet indicating the absence of bromine and can be assigned to a benzoquinone-GSH conjugate (BQ-GSH).

The metabolism of benzene and toluene appear to be considerably simpler than that of PhBr. For PhMe, a single low abundance twin-ion could be assigned following manual inspection of HiTIME data at $m/z$ 414 and 19.1 mins (Figure 6-1B). Unfortunately, attempts to acquire an MS/MS spectrum of this ion were unsuccessful however, given that the retention time and $m/z$ value are similar to those of glutathione conjugates formed from PhBr, it seems likely that this compound is also formed via GSH adduction. Subtracting the mass of GSH from the $m/z$ value of the observed ion indicates a metabolite of mass 106 Da which is 14 Da greater than the toluene precursor (92 Da) suggesting hydroxylation followed by oxidation with loss of H$_2$. On this basis, the metabolite is assigned to the product toluene hydroxylation and subsequent oxidation forming a quinone methide (QM). Similarly, a single metabolite was observed in PhH incubations at $m/z$ 416/422 and 15.8 mins. Upon CID, peaks with formation of 75 and 129 Da neutral fragments are observed (data not shown) suggesting a GSH conjugate of a 108 Da compound which can be assigned to benzoquinone (BQ).
Figure 6-1. Heat maps produced by HiTIME scoring Q-TOF LC-MS data from supernatants obtained following incubation of 1:1 PhX:13C6-PhX substrates with rat liver microsomes using Δm/z = 6.0201. A, C, E) incubations of X = Br, Me, H respectively without supplemental GSH. B, D, F) incubations containing 10 mM supplemental GSH for X = Br, Me, H respectively.
Figure 6-2. MS and MS/MS data for the twin-ion hit at \( m/z \) 494 and 22.2 mins in non-GSH supplemented PhBr incubations. A) Mass spectrum observed at 22.2 mins. B) EICs of \( m/z \) 494 and 500. Inset shows an expansion of the low abundance peaks and arrows indicate positions of bromine-containing twin-ions. C) MS/MS spectrum of \( m/z \) 494, and D) MS/MS spectrum of \( m/z \) 500. The precursor ions are marked with an ‘\*\’.

The formation of the PhX metabolites observed here for all 3 compounds can be rationalised using a single scheme wherein initial oxidation of the PhX substrate produces an arene oxide intermediate. While arene oxides are generally not directly observed due to their inherent instability, these compounds are thought to be the initially formed CYP oxidation product for
a range of aromatic hydrocarbons.\textsuperscript{382-384} Indeed, the epoxide derivative of naphthalene has been isolated and characterised from liver microsome incubation.\textsuperscript{385}

Arene oxide derivatives of PhX substrates could undergo a range of different reactions.\textsuperscript{386} The first is direct trapping of the arene oxide by GSH which can occur via both chemical and enzyme-mediated reactions.\textsuperscript{381, 384, 386} For PhBr, this would give rise to the metabolite at \textit{m/z} 480/486 and accounts for the observation that GSH supplementation of these incubations substantially increased the abundance of \textit{m/z} 480 formed (Figure 6-3). Another possibility is that the initially formed arene oxide undergoes rearrangement to a phenol through an intramolecular hydrogen atom transfer in a process known as the ‘NIH shift’.\textsuperscript{381, 387*} This rearrangement is operative for a wide range of arene oxides and the rates and distributions of products are found to depend on ring substituent patterns and solution pH.\textsuperscript{384} For PhMe, NIH shift of an arene oxide would produce a cresol which, upon further oxidation, would afford a quinone methide giving rise to the observed metabolite at \textit{m/z} 414 upon GSH trapping. Indeed, detailed studies of toluene metabolism in the rat have revealed that hydroxylation proceeds with deuterium migration confirming the presence of an arene oxide intermediate.\textsuperscript{389, 390} While multiple epoxides may be formed, it has been found that toluene 3,4-epoxide undergoes an NIH shift almost exclusively producing \textit{p}-cresol\textsuperscript{391} which is the major regioisomer observed in rats.\textsuperscript{392} On this basis, the sole metabolite observed for toluene is assigned to an adduct of \textit{p}-quinonemethide to GSH (QM-GSH, Figure 6-4).

\textsuperscript{*} The term “NIH shift” used in literature to describe this rearrangement stands for the United States National Institute of Health where the reaction was discovered.\textsuperscript{388}
Figure 6-3. Proposed pathways of formation for PhBr and PhH metabolites observed here. It should be noted that multiple isomers and formation pathways are possible for each pathway as discussed in text. For simplicity, only one possibility is drawn.

Figure 6-4. Proposed mechanism of formation of QM-GSH from toluene in RLMs.

Interestingly, no evidence of GSH-trapped arene oxides was observed for either PhH or PhMe incubations. While the precise reason for this is unclear, it may be due to the relative instability of the arene oxide intermediate. Indeed, numerous metabolites have been identified for other halobenzene substrates suggesting that halobenzene oxides are considerably more stable than analogous benzene or alkylbenzene compounds.\textsuperscript{386, 393} Therefore, rapid rearrangement of benzene and toluene arene oxides to phenol derivatives may prevent the formation of GSH-trapped epoxides.

Numerous quinones are observed as conjugates with glutathione for both benzene and bromobenzene that could be formed through various pathways in a complex network of
chemical and enzymatic reactions. For example, an initially formed arene oxide may undergo hydrolysis forming a 1,2-dihydrodiol mediated by epoxide hydrolase.\textsuperscript{394} Subsequent action of dihydrodiol dehydrogenase may restore aromaticity forming a catechol\textsuperscript{386,395} and further CYP-mediated oxidation may produce an \textit{ortho} quinone.\textsuperscript{110} An alternative pathway has also been demonstrated whereby an NIH shift affords a phenol and subsequent net hydroxylation at the \textit{ortho} or \textit{para} positions produce either 1,2-hydroquinone (i.e. catechol) or 1,4-hydroquinone\textsuperscript{396} which may then undergo oxidation affording an \textit{ortho} or \textit{para} quinone respectively.\textsuperscript{106}

The range of isomeric products observed for metabolites of bromobenzene at $m/z$ 494, 480 and 462 likely arises from two key points in this metabolic scheme. Firstly, CYP-mediated epoxidation of PhBr may conceivably produce both a 2,3-oxide and a 3,4-oxide. While arene oxides such as these are often highly reactive and difficult to directly observe, \textit{in vivo} and \textit{in vitro} studies have identified apparent chemical derivatives of both isomers by comparison to synthetic standards.\textsuperscript{397,398} The relative extent of formation of these isomers is difficult to quantify. Numerous studies have demonstrated that different chemical inducers of CYP expression lead to selective increases (ca. 5-30 fold) in the formation of either the 2,3- or 3,4-epoxide suggesting that these isomers are predominantly formed by distinct CYP isoforms.\textsuperscript{399} Moreover, regioisomeric addition of GSH to either epoxide isomer may give rise to numerous dihydrobromophenol structural and stereoisomers.\textsuperscript{397,400,401} Given that a range of isomers could be produced at a number of stages, absolute assignment of structure is not possible without the synthesis of authentic standards which was deemed to be beyond the scope of this work.
Table 6-1. Manually verified twin-ion signals detected following HiTIME scoring of PhX microsome incubation LC-MS data.

<table>
<thead>
<tr>
<th>Retention Time (min)</th>
<th>m/z (monoisotopic)</th>
<th>Transition</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.8</td>
<td>416/424</td>
<td>PhH+2O-2H+GSH</td>
<td>BQ-GSH</td>
</tr>
<tr>
<td>PhMe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19.1</td>
<td>414/420</td>
<td>PhMe+O-2H+GSH</td>
<td>QM-GSH</td>
</tr>
<tr>
<td>PhBr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22.20</td>
<td>494/500</td>
<td>PhBr+2O-2H+GSH</td>
<td>BBQ-GSH</td>
</tr>
<tr>
<td>21.90</td>
<td>494/500</td>
<td>PhBr+2O-2H+GSH</td>
<td>BBQ-GSH</td>
</tr>
<tr>
<td>21.40</td>
<td>494/500</td>
<td>PhBr+2O-2H+GSH</td>
<td>BBQ-GSH</td>
</tr>
<tr>
<td>20.40</td>
<td>494/500</td>
<td>PhBr+2O-2H+GSH</td>
<td>BBQ-GSH</td>
</tr>
<tr>
<td>19.00</td>
<td>494/500</td>
<td>PhBr+2O-2H+GSH</td>
<td>BBQ-GSH</td>
</tr>
<tr>
<td>20.05</td>
<td>480/486</td>
<td>PhBr+O+GSH</td>
<td>BBE-GSH</td>
</tr>
<tr>
<td>19.95</td>
<td>480/486</td>
<td>PhBr+O+GSH</td>
<td>BBE-GSH</td>
</tr>
<tr>
<td>19.95</td>
<td>462/468</td>
<td>PhBr-2H+GSH</td>
<td>BB-GSH</td>
</tr>
<tr>
<td>19.80</td>
<td>462/468</td>
<td>PhBr-2H+GSH</td>
<td>BB-GSH</td>
</tr>
<tr>
<td>15.8</td>
<td>416/422</td>
<td>PhBr-Br+2O-H+GSH</td>
<td>BQ-GSH</td>
</tr>
</tbody>
</table>

6.2.2 PhX adducts to microsomal proteins

Microsomal protein extracts from PhX and control treatments were reduced, alkylated and digested with trypsin. Following solid-phase clean-up and analysis using the Orbitrap mass spectrometer, Mascot searches were carried out on LC-MS/MS data and summaries of the protein assignments for each treatment group are given in Figure 6-5. Approximately 1300 unique proteins were assigned in all 3 replicates of each treatment group. This number is higher than that obtained for similar Mascot analysis of the APAP microsome incubation discussed in Chapter 5 (approximately 1000) which is largely attributed to the use of a longer solvent gradient used for the experiments here.
Figure 6-5. Venn diagrams indicating the total number of proteins identified by database searching of LC-MS/MS data sets for A) PhH, B) PhMe, C) PhBr, and D) control microsomal protein digests.

To identify peptides that appear as twin-ions and thus indicate PhX metabolite adducts, LC-MS data were scored with HiTIME. The heat maps obtained using a $\Delta m/z$ value of 3.01005 and 2.0065 (corresponding to the 6.0201 Da mass difference between heavy and light isotopologues in the +2 and +3 charge states respectively) are shown in Figure 6-6 and Figure 6-7 respectively. These data were subjected to isotope filtering and post processing through the Xenophile software. The results of these analyses (EIC traces, mass spectra and isotope-weighted HiTIME scores) were manually reviewed using the included visualisation tools.
Figure 6-6. Heat maps produced by HiTIME scoring of global protein digests obtained following incubation of 1:1 PhX:13C₆-PhX substrates with rat liver microsomes using Δm/z = 3.01005 to search for doubly charged peptides. A) PhH, B) PhMe, C) PhBr, and D) control. For simplicity, only 1 replicate of each treatment is shown here.

Upon close manual inspection of this data, no twin-ion peptides could be assigned for any treatment group. In many cases, highly scoring regions in HiTIME data were found to correspond to unrelated peptide signals. For example, many apparent HiTIME hits were found to correspond to two closely eluting peptides that were of different charge states. In other cases, peptides of the same charge state were observed but HiTIME scoring had highly weighted mismatched isotopic peaks of two independent signals. Examples of these features are shown in Figure 6-8.
Figure 6-7. Heat maps produced by HiTIME scoring of global protein digests obtained following incubation of 1:1 PhX:\textsuperscript{13}C\textsubscript{6}-PhX substrates with rat liver microsomes using $\Delta m/z = 2.0067$ to search for triply charged peptides. A) PhH, B) PhMe, C) PhBr, and D) control. For simplicity, only 1 replicate of each treatment is shown here.

As no twin-ion peptides were assigned, further analysis through the Xenophile pipeline became redundant and was not conducted. The lack of twin-ion peptides detected by this analysis may result from either 1) false-negative assignments by HiITME, or 2) the absence of PhX modified peptides in microsomal protein digests. A third possibility is that the \textsuperscript{13}C\textsubscript{6} label incorporated into each PhX substrate was lost or scrambled throughout the incubation and/or analysis. Given that: 1) this process would require rearrangement and/or cleavage of aromatic rings, and 2) retention of the \textsuperscript{13}C\textsubscript{6} label was observed for all of the small molecule metabolites discussed above, loss of the \textsuperscript{13}C label appears exceedingly unlikely.
Figure 6-8. Examples of highly scoring regions in HITIME data that were not consistent with true twin-ions upon manual verification. A) mismatched isotopic peaks from peptides of the same charge state, and B) mismatch of peptides with different charge states.

Table 6-2. PhX-derived variable modification used for targeted database searching of microsomal protein digests.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Residue</th>
<th>Metabolite</th>
<th>Formula</th>
<th>m/z (monoisotopic)</th>
<th>Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhH</td>
<td>C</td>
<td>L</td>
<td>C₆H₄O₂</td>
<td>108.02113</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H</td>
<td>¹³C₆H₄O₂</td>
<td>114.04126</td>
<td>994</td>
</tr>
<tr>
<td>PhMe</td>
<td>C</td>
<td>L</td>
<td>C₇H₆O</td>
<td>106.04186</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H</td>
<td>¹³C₆CH₆O</td>
<td>112.06199</td>
<td>10</td>
</tr>
<tr>
<td>PhBr</td>
<td>C</td>
<td>L</td>
<td>C₆H₃OBr</td>
<td>171.95183</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H</td>
<td>¹³C₆H₃OBr</td>
<td>177.97196</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>L</td>
<td>C₆H₃O₂Br</td>
<td>185.93164</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H</td>
<td>¹³C₆H₃O₂Br</td>
<td>191.95177</td>
<td>0</td>
</tr>
</tbody>
</table>
To differentiate between the remaining possibilities of false-negative HiTIME assignments and absence of twin-ion signals in the data, targeted database searching was carried out to search for suspected metabolite adducts. For these searches, putative reactive metabolites that may be formed from PhX precursors were derived from supernatant LC-MS data presented in Section 6.2.1. These reactive metabolites are shown in Table 6-2. Mascot searches were conducted for each PhX treatment and replicate allowing for variable post-translational modification of cysteine residues by these metabolites. The number of peptides that were assigned as having at least one of these PTMs are given in Table 6-2. For PhBr and PhMe, 4 and 10 peptides were assigned as covalent adducts of reactive metabolites derived from these precursors however these assignments were also similarly prevalent in vehicle control samples that were not exposed to either material. Moreover, no evidence of twin peptide ions was observed upon manual inspection of the MS\(^1\) spectra associated with these hits indicating that these may likely arise from erroneous assignments. Nearly 1000 modified peptides were assigned as benzoquinone adducts in PhH treatments, which is significantly higher than any other treatment or reactive metabolite investigated. Interestingly, the overwhelming majority of these are assigned to adducts of the ‘heavy’ \(^{13}\)C\(_6\) benzoquinone metabolite while only 2 are assigned as adducts of ‘light’ benzoquinone. Again, manual inspection of the MS\(^1\) spectra revealed no evidence of twin-ion signals (Figure 6-9A). Given that these assignments do not correspond to twin-ion peptides, and that the number of assignments is unreasonably high, it seems likely that a combination of more common events may coincidently equate to the mass of \(^{13}\)C\(_6\) benzoquinone. Indeed, the combined mass of two carbamidomethyl modifications (2 x 57.02146 Da) is 114.04293 Da which is very similar to the mass of \(^{13}\)C\(_6\) BQ (114.04126 Da). As most peptides observed here are in the range of 1000-2000 Da, the difference between these two modifications becomes small. For example, this difference translates to a \textit{ca.} 1 ppm error for a 1500 Da peptide in the 2+ charge state.
Figure 6-9. A) An exemplary MS$^1$ spectrum of the peptide at m/z 893.4 and 74.45 mins. B) EIC traces for the microsomal glutathione-S-transferase peptide VFANPEDCAGFGK in the 2+ charge state with either bromobenzene epoxide (m/z 763.78), bromobenzene quinone (m/z 770.77) or carbamidomethyl modifications (m/z 706.32). Note that EIC intensities for m/z 763 and 770 traces have been multiplied by a factor of 10.

To further explore this data, EIC traces were plotted for the calculated m/z values of adducts of each metabolite in Table 6-2 adducted to each peptide found to be modified by the APAP in Chapter 5. As an example, the EIC traces for bromobenzene quinone adduction to the microsomal glutathione-S-transferase peptide VFANPEDCAGFGK, which was the most abundant APAP adduct observed in Chapter 5, is shown in Figure 6-9. No significant peaks are observed in the EIC trace of the calculated product m/z despite a clear signal for the native peptide. Upon close inspection of the minor peaks in the product trace, no mass spectrum displayed twin-ions at these locations.

Taken together, these data indicate that no detected peptides have been covalently modified by PhX reactive metabolites. The finding that numerous peptides appear to be falsely assigned as PhX adducts in Mascot searches (particularly for X = H) highlight the potential pitfalls of conducting targeted proteomics studies without the aid of isotopic signatures on systems where many PTMs are possible. In the absence of the twin-ion signature, which serves to confirm the presence of a reactive metabolite-modified peptide, the CRM adduct assignments may be difficult to validate and it is likely that further experiments would be necessary.

The lack of PhX adducts following targeted searches is consistent with the absence of twin-ions in HiTIME results. Therefore, HiTIME and the associated post-processing routines
implemented through Xenophile have correctly assigned a ‘true negative’. It should be emphasised that correctly assigning negative results (i.e. the absence of protein adducts) is just as important as correctly assigning positive results (i.e. the presence of protein adducts). This is because many xenobiotics do not undergo metabolic bioactivation and do not form covalent macromolecular adducts.

Nonetheless, it is interesting to note that other authors have previously identified specific protein adducts of these compounds, especially bromobenzene. The specific reason for the discrepancy between these previous findings and the results presented here is unclear. However, it is worth noting that the vast majority of studies that have investigated bioactivation and protein binding of PhX compounds have conducted chemical pre-treatments of the experimental animals employed. This has the effect of increasing CYP enzyme expression thereby increasing the production of putative reactive metabolites. These include phenobarbital and 3-methylcholoranthrene which have both been shown to increase PhBr bioactivation by >10 fold compared to uninduced controls.

To estimate the extent of PhBr bioactivation in these experiments relative to that of APAP (for which numerous protein adducts were identified, Chapter 5), the integrated EIC peak area for all GSH adduct metabolites observed in GSH-supplemented microsomal incubations was summed and is given in Figure 6-10. In general, the intensity of signals in mass spectra for different compounds should not be used as a direct measure of relative concentration in the sample due to varying ionisation efficiencies of different compounds. However, in this case, the metabolites of interest are structurally similar arylated glutathione derivatives and thus mass spectral signal intensity should a reasonable approximation of concentration. The EIC peak area for APAP-GSH is observed to be ca. 50-fold greater than that for PhBr which is the most extensively metabolised PhX xenobiotic studied here. This indicates that PhX bioactivation by microsomal proteins proceeded to a far lesser extent than for APAP.

*The bioactivation of APAP to NAPQI and subsequent GSH trapping has been used as a positive control in microsome incubations throughout this thesis. Accordingly, these APAP incubations were conducted simultaneously with the PhX incubations using microsomes from the same batch and the products were analysed under identical conditions.*
Numerous factors may contribute to the reduced production of PhX reactive metabolites compared to APAP. For example, bioactivation of APAP to the quinoneimine metabolite, NAPQI, can be accomplished in a single enzymatic step. By contrast, several biochemical reactions are generally required for production of the electrophilic metabolites of the PhX compounds studied here. It is likely that the overall yield of any given metabolite decreases with each additional metabolic step required for its formation and, therefore, the total reactive metabolite yield may be low. Moreover, the energetic barriers to the initial oxidation step of the PhX metabolites studied here may be substantially higher than that for APAP. Indeed, Calder et al demonstrated that the nephrotoxicity of a series of substituted hydroquinones, catechols and aminophenols increased sharply with increasing ease of oxidation.\textsuperscript{407} Similarly, Blumer and co-workers showed that the ease of oxidation for a series of benzene metabolites was strongly correlated with DNA damage\textsuperscript{408} which can be rationalised using known substituent effects on the redox potentials of aryl systems. For example, oxidation involves removal of electrons and is facilitated by the presence of activating, electron-donating substituents.\textsuperscript{138, 409} Interestingly, APAP possesses the activating hydroxyl and acetamide groups while the PhX substrates examined here (especially PhH and PhMe) are essentially unactivated. Therefore, it seems likely that the markedly reduced formation of PhX reactive metabolites compared to that of APAP can be attributed a combination of reduced initial oxidation arising from electronic effects and decreasing product yields across the numerous reaction steps required for some reactive PhX metabolites. Finally, the lack of protein adducts of reactive electrophiles derived from PhX is consistent with their low yields, which allows the preferential glutathione adduction.

**Figure 6-10.** EIC peak area for the most abundant APAP and PhX metabolite GSH adducts. Figures above the bars indicate the relative proportions of these ions.
There may be merit in studying the bioactivation of these compounds under conditions of artificially increased xenobiotic metabolising enzyme expression. For example, the well-established inter-individual variability of CYP enzymes may result in natural expression levels in some subjects that are similar to those obtained after chemical induction. However, in general, increasing chemical perturbation of the native system prior to study may push the biological processes beyond the limits of living systems, especially when artificially extracted and enriched sources of xenobiotic metabolising enzymes such as microsomes are used.

6.3 Conclusion

The metabolic activation of benzene, toluene and bromobenzene has been studied in rat liver microsomes using the combination of the twin-ion method and HiTIME analysis. For PhBr, numerous small molecule metabolites were detected via HiTIME analysis of microsomal incubation supernatants and these could be assigned to glutathione conjugates of epoxide and quinone metabolites. For both PhH and PhMe, a single metabolite was observed that was assigned to GSH conjugates of a quinone and quinone methide respectively. Despite formation of these electrophilic metabolites for each compound, no twin-ion signals were observed following HiTIME analysis of microsomal protein digests. Targeted attempts to identify peptides modified by putative PhX reactive metabolites yielded a range of hits which, when manually inspected, did not reveal signals that also displayed a twin-ion signature in MS$^1$ data. Taken together, these data indicate that PhX protein adducts were not formed and detected in these experiments and highlight the pitfalls of using targeted proteomics approaches that do not use ‘twin ions’ to identify the products of covalent attachment of reactive metabolites to protein targets.
7 CONCLUDING REMARKS

In the ongoing search for new medications that are safer and more effective, industrial and academic scientists, as well as regulatory bodies, have recognised the importance of drug metabolism in determining the pharmacological effects of a compound. For example, metabolism, and pharmacokinetics more broadly, largely dictates the dose and frequency with which a therapeutic agent must be administered to a patient. Additionally, bioactivation, another process affected by metabolism, is thought to be critical in the onset of many toxic reactions. For these reasons, xenobiotic metabolism remains an active area of research in many laboratories around the world.

Identifying the specific chemical species formed in metabolic reactions for a given xenobiotic is the key starting point to explaining these biological effects. Metabolic insights also provide the foundation for the rational optimisation of a drug structure to minimise toxicity and tune kinetic parameters leading to the development of safer, more effective and convenient medication. This thesis has described methodologies that aim to detect metabolites formed from a drug without requiring prior knowledge of the types of products that are likely to be formed. These ‘non-targeted’ methods ensure, to the greatest extent possible, that analyses are not biased by the limitations of current knowledge so that the maximum possible information can be derived from experimental data sets.

Specifically, the methods developed here combined stable isotope labelling, liquid chromatography mass spectrometry and computational data mining in various forms to detect the products of xenobiotic metabolism from complex biological samples. The metabolic fate of a molecule could take multiple different forms and, accordingly, methods have been
described to account for as many of these as possible. A software package, called HiTIME, was developed that uses designer isotopic signatures to detect signals corresponding to xenobiotic metabolites in LCMS data. The use of this software, and the associated experimental techniques, has been successfully applied for detecting small molecule metabolites of numerous analytes in a range of biological systems.

Despite the apparent importance of reactive metabolite adduction to macromolecules in the onset of toxic reactions to a xenobiotic, identification of these adducts is not yet routinely conducted, which may be related to the practical difficulty in conducting such experiments. Given that this is an important aspect of drug metabolism, the latter half of this thesis described application of these techniques to characterisation of the protein targets of reactive drug metabolites. This resulted in the development of Xenophile, a software package capable of identifying the formula, binding site and protein target of a drug directly from shotgun proteomics data. Xenophile proved successful at characterising protein adducts in both small in silico validation studies as well as complex experimental datasets produced using liver microsomes that contained more than one thousand unique proteins.

The methodologies described here require further development and refinement but have shown promise in realising the goal of non-targeted metabolite identification. It should be noted that these are not specific to a particular substrate but rather should constitute general approaches that could be employed to study the metabolism of a great many substances with widely varying structures. In this sense, it is hoped that these techniques may be adopted by the community at large so that an ever-clearer picture of drug metabolism may be developed.
8 EXPERIMENTAL AND COMPUTATIONAL METHODS

8.1 General methods and instrumentation

8.1.1 Mass spectrometry
Three different mass spectrometers were used for the experiments herein and are described below.

8.1.1.1 Quadrupole-time of flight mass spectrometer
The majority of experiments were performed on an Agilent 1100 liquid chromatography system coupled to an Agilent 6520 Quadrupole-Time of Flight (Q-TOF) mass spectrometer controlled via the Agilent MassHunter software package version B.05.00. Compounds were separated by an Agilent ZORBAX Eclipse XDB C18 reversed phase column (150 x 4.6mm, 5µm) using 100% Milli-Q water as mobile phase A and 70% acetonitrile in 0.1% formic acid (v/v) as mobile phase B (flow rate of 0.3 ml min⁻¹). The column was maintained at 25 ± 0.8°C throughout all analytical runs and samples were held in the autosampler chamber at 4°C prior to analysis. The autosampler needle was washed with 100% mobile phase B for 5 s prior to sample injection into the LC solvent stream. In general, the LC eluent was diverted to waste for the first 2 minutes immediately after injection. Components that elute from the column were
transferred into the gas phase as ions by use of electrospray ionization (ESI). ESI-MS parameters (Table 8-1) were optimized by directly infusing an aqueous APAP solution.

### Table 8-1. MS parameters for APAP metabolite detection using the Agilent 6520 Q-TOF mass spectrometer

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drying gas flow rate</td>
<td>12 L min⁻¹</td>
<td>12 L min⁻¹</td>
</tr>
<tr>
<td>Gas temperature</td>
<td>300°C</td>
<td>300°C</td>
</tr>
<tr>
<td>Nebuliser</td>
<td>50 psig</td>
<td>50 psig</td>
</tr>
<tr>
<td>Capillary voltage</td>
<td>3.5 kV</td>
<td>5.5 kV</td>
</tr>
<tr>
<td>Skimmer voltage</td>
<td>65 V</td>
<td>65 V</td>
</tr>
<tr>
<td>Octopole RFᵱᵥ-ᵥ</td>
<td>750 V</td>
<td>750 V</td>
</tr>
<tr>
<td>Fragmentor</td>
<td>75 V</td>
<td>75 V</td>
</tr>
<tr>
<td>Mass scan range</td>
<td>50-1500 m/z</td>
<td>50-1500 m/z</td>
</tr>
<tr>
<td>Acquisition rate</td>
<td>2 spectra s⁻¹</td>
<td>2 spectra s⁻¹</td>
</tr>
<tr>
<td>Scan time</td>
<td>500 ms spectrum⁻</td>
<td>500 ms spectrum⁻</td>
</tr>
</tbody>
</table>

8.1.1.2 Orbitrap fusion lumos mass spectrometer

Nano LC-MS/MS analyses of microsomal protein digests were performed on a hybrid linear ion trap mass spectrometer and orbitrap mass spectrometer (Orbitrap Fusion Lumos; ThermoFisher Scientific) equipped with a nano-HPLC (UltiMate 3000 RSLC; Dionex). The nano-LC system was equipped with a C₁₈ nano-trap column (Acclaim Pepmap, Dionex, 100 Å, 75 μm × 2 cm) and a C₁₈ analytical column (Acclaim Pepmap RSLC, Dionex, 100 Å, 75 μm × 50 cm). Peptides were separated using gradient elution with mobile phases of 0.1% (v/v) aqueous formic acid (solvent A) and 0.1% (v/v) formic acid in CH₃CN (solvent B). Analytes were introduced into the gas phase via nano-ESI using a spray voltage of 1.9 kV, capillary temperature of 275 °C and all MS¹ mass spectra were acquired in the positive ion mode over a range of 400-1500 m/z in the Orbitrap analyser operating at a resolution of 120,000. Data-dependent MS² spectra were collected by subjecting 2+ to 5+ precursor peptide ions with a measured intensity greater than 10⁴ to higher-energy collision induced dissociation (HCD) with a normalised collision energy of 35%. Dynamic exclusion was used to prevent the selection of the same precursor ion more than once for 45 s. For each run, 2 μL of sample was injected and samples were held in the autosampler at 5 °C prior to analysis.

8.1.1.3 LTQ-FTICR mass spectrometer
Selected CID and high-resolution mass spectrometry experiments were performed on a Thermo hybrid LTQ-FTICR mass spectrometer (Finnigan, LTQ-FT, Bremen, Germany). The instrument consists of an electrospray ionisation source followed by a linear ion trap coupled to a 7 Tesla FT-ICR mass analyser which enables high resolution and high accuracy mass determination. Samples were prepared in methanol unless stated otherwise at a concentration of ca. 0.5 mM and introduced into the ESI source via an automated syringe pump at an injection rate of ca. 5 μL min⁻¹ with typical needle voltages in the range of 3.5-5.0kV with capillary temperatures of 280-300 °C using N₂ as the sheath and auxiliary gases. Mass calibration was performed with the recommended calibration solution consisting of caffeine, the tetrapeptide MRFA and, Ultramark 1621.

8.1.2 Nuclear magnetic resonance
All experiments (¹H, ¹³C, ¹H-¹H COSY) were carried out at 296 K with either a Varian Unity Inova 500 MHz spectrometer [499.688 MHz (¹H), 125.646 MHz (¹₃C)] or Agilent MR 400 MHz spectrometer [400 MHz (¹H), 125 MHz (¹₃C)] in deuterated dimethylsulfoxide (DMSO-ᴅ₆) or deuterated chloroform (CDCl₃). The ¹H and ¹³C chemical shifts are expressed in ppm relative to TMS and referenced against the residual solvent signal. The 1D gradient nuclear overhauser effect (NOE) spectra were obtained at 500 MHz using a standard double-pulsed-field-gradient-spin-echo (DPFGSE) method and a 500 ms mixing time at 323 °K.

8.2 Chapter 2 methods

8.2.1 Materials.
N-Acetyl p-aminophenol (APAP; Sigma-Aldrich), [¹³C₆]-APAP (Isosciences), [²H₄]-APAP (Isosciences), and HPLC solvents (Ajax Finechem) were obtained from commercial sources. The labelled compounds were 99.5% pure with >99% isotope enrichment. High purity water (18 MΩ) from a Milli-Q filtration device was used throughout.

8.2.2 In vivo APAP metabolism studies.
The University of Melbourne Animal Ethics Committee approved experiments (approval #1212592) in accordance with the Australian code for the care and use of animals for scientific purposes (8th edition, 2013, National Health and Medical Research Council, Canberra). Rats were housed in climate-controlled conditions with 12 h light/dark cycle and had free access to
normal pellet diet and drinking water. Male adult Sprague-Dawley rats \( (n = 24, \text{327–366 g}) \) were randomly allocated to either APAP treatment or vehicle control groups. For \text{i.v.\ drug administration and blood sampling, a jugular vein catheter and a carotid artery catheter, respectively, were implanted under general anaesthesia (3.5% isoflurane in O}_2\ via a nose cone) so that the following procedures could be performed later whilst the animal was conscious. From conscious rats, following 2 h of recovery post-catheter implantation, a 500 µL baseline blood sample was collected (mixed with 2 U heparin; 2 µL of 1000 U ml\(^{-1}\)), which was subsequently centrifuged (2600 rpm for 10 min at 4°C). The plasma supernatant was collected, immediately frozen, and stored at -80°C. Rats received the treatments: vehicle (saline 2 ml kg\(^{-1}\)) or 1:1 APAP:[\text{\textsuperscript{13}C}_6\text{-APAP} 10 mg kg\(^{-1}\) \( (i.e., \) 5 mg kg\(^{-1}\) of each isotopologue, comparable to a therapeutic dose). Plasma samples were collected at 30, 60, 120 and 240 min post-treatment administration and stored at -80°C.

8.2.3 LCMS sample preparation.
The sample preparation method was adapted from that published by Jung and co-workers with minor modifications. Methanol (900 µL) and \text{\textsuperscript{2}H}_4\text{-APAP internal standard (10 µL, 1.0 µg ml}^\text{-1}) were added to rat plasma (100 µL). After vortex mixing \( (ca. \text{30 s}) \) and centrifugation (10 min, 14,000 rcf), an 800 µL aliquot of the supernatant was evaporated to complete dryness. The residue was reconstituted in water (100 µL), vortexed \( (ca. \text{30 s}) \), and transferred to autosampler vials (200 µL) for LCMS analysis.

8.2.4 LCMS analysis
All LC-MS experiments were performed using the Q-TOF mass spectrometer. For each analysis, 10 µL was injected and compounds were separated using gradient elution as follows \[ \text{[time (min), %B]: [0, 0], [0, 1], [100, 10], [100, 60], [0, 61], [0, 65].} \] Analyses were performed in both positive and negative ion mode for each sample and spectra were acquired over the range \( m/z \text{ 50-1500}. \)

8.2.5 Data analysis.
For HiTIME analysis, data files that were obtained in the Agilent ‘.d’ file format were converted to the open source mass spectrometry mzML data format by use of MSConvert (see Appendix Table 9-1 for the transformation parameters) which produced \( ca. \) 14 gigabyte data files that each contains \( ca. \) 900 million data points. Processing the data is a computationally
intensive task that takes the equivalent of 60 hours of processing time for the datasets that were used here. Consequently, an intensity filter of 500 counts is applied and, furthermore, the software implementation makes use of distributed computing to spread the processing across many computers, which increases the rate of data processing by a factor of over 60.

The HiTIME score is derived from the goodness-of-fit of Gaussian functions to the experimental data. As the resolution (width) of both mass spectral and chromatographic peaks is dependent on both the instrument used and the operating parameters of that instrument, the parameters describing the dimensions of the target Gaussian shape used in the fitting procedure need to be adjusted to match. This process involves measuring the peak width in both \( m/z \) and RT dimensions for an observed experimental signal which need not be a known twin-ion.

Extracted ion chromatograms (EICs) were produced by plotting integrated ion abundances for a specified \( m/z \) value (with a width of ± 0.03 \( m/z \)) vs. retention time. 3D data plots and 2D heat maps were generated using Gnuplot. Centroiding of profile datasets was performed using MSConvert.

8.3 Chapter 3 methods

8.3.1 Materials

Unless otherwise stated, all chemicals were purchased from commercial sources and used as received. HPLC grade acetonitrile, methanol and diethyl ether were from Merck (Kenilworth, NJ, USA), dichloromethane was from Honeywell (Morristown, NJ, USA) and 18 MΩ H₂O was obtained using a Milli-Q apparatus. \( N \)-Acetyl amino acid methyl esters were from BAChem (Bubendorf, Switzerland) except for \( N \)-acetyl cysteine methyl ester and \( N \)-acetyl glutamine methyl ester, which were synthesized as described below. Peptides GAIL-X-GAILR (where X = W, Y, C or M) were purchased from Mimotopes (Victoria, Australia) and used without further purification. Rat liver microsomes (lot number RT053C) were purchased from Life Technologies (Carlsbad, CA, USA), acetaminophen, AgNO₃, NADPH tetrasodium salt and \( N \)-ethylmaleimide were from Sigma-Aldrich (St. Louis, MO, USA).

8.3.2 \( N \)-Acetyl cysteine methyl ester

\( N \)-Acetyl cysteine (1.00 g) was dissolved in a solution of acetyl chloride (0.4 mL) in methanol (10 mL) and stirred at ambient temperature for 36 h. The solvent was then removed \textit{in vacuo} and the off-white residue reconstituted in EtOAc (ca. 30 mL) and washed with brine (3x30
The organic layer was evaporated to dryness then purified by silica gel chromatography using 95:5 CHCl₃:MeOH as the eluent. The combined fractions were dried (MgSO₄) and evaporated to dryness affording a white powder (0.30 g). HRMS (ESI): Calculated for C₆H₁₁NO₃S+H+: 178.05324, observed: 178.05320.

8.3.3 N-Acetyl glutamine methyl ester

Glutamine (4.00 g 27.4 mmol) was dissolved in hot water (30 mL) then acetic anhydride (6.4 mL) was added at once. The solution was stirred at ca. 90 °C for 15 mins then allowed to stand at room temperature overnight. The solvent was evaporated in vacuo then the off-white solid was recrystallised from hot water and washed with a small quantity of cold H₂O yielding fine pure white crystals (2.29 g, 12.2 mmol, 44%). HRMS (ESI): Calculated for [C₇N₂O₄H₁₂ + H]⁺, 189.08698 observed: 189.08667

Acetyl chloride (0.2 ml) was added to a suspension of acetyl glutamine (0.5 g, 2.66 mmol) in methanol (25 ml) and the solution stirred overnight at room temperature. The reaction mixture was neutralised with NaHCO₃ solution, the solvent was removed in vacuo and the residue was triturated with 95:5 CHCl₃:MeOH. The solids were removed by filtration then the filtrate was concentrated in vacuo to ca. 3 ml then purified via silica gel column with 95:5 CHCl₃:MeOH. The solvent was evaporated to give a white powder (60 mg, 0.3 mmol, 11%): HRMS (ESI): Calculated for [C₈N₂O₄H₁₄ + Na]⁺ = 225.08458 observed: 225.08452

8.3.4 Synthesis of N-acetyl p-benzoquinone imine (NAPQI)

NAPQI was synthesized using the method of Dahlin and Nelson²⁸² with some modifications. Freshly prepared Ag₂O (1.12 g, 4.8 mmol) was added to a suspension of APAP (0.50 g, 3.3 mmol) in dry dichloromethane (35 mL). The reaction mixture was covered to exclude light and stirred at room temperature for 1 hour. Excess Ag₂O was removed by filtration and the filtrate was immediately purified by silica gel chromatography using dry diethyl ether as the eluent. NAPQI was isolated as a single bright yellow band. The combined ethereal fractions were diluted via the addition of dry acetonitrile (10 mL) and the resultant solution was concentrated in vacuo to ca. 10 mL. This NAPQI solution in acetonitrile was used for all subsequent reactions.

The silver oxide used in the above reaction was prepared as follows. Silver nitrate (2.25 g) was dissolved in H₂O (15 ml) then a solution of potassium hydroxide (0.95 g) in H₂O (10 ml) was added with vigorous stirring. The brown precipitate was filtered and washed with H₂O (3x10
ml) followed by MeOH (3x10 ml). The product was dried under vacuum at 50°C for ca. 3 hours prior to use.

8.3.5 General procedure for the reaction of NAPQI with amino acids and peptides
A series of reaction vials containing either NHAc-X-OMe (where X = Cys, Tyr, Trp, His, Lys, Arg, Met, Gln, Glu, Ser and Val) (10 µg mL⁻¹) were prepared in sodium phosphate buffer (10 mM) and then adjusted to various different pH values (6.0, 6.5, 7.0, 7.5, 8.0, 8.5 and 9.0) with aqueous NaOH. Aliquots of the NAPQI solution in acetonitrile (75 µL) were then added to each of the reaction vials, which were shaken for 3 h and then stored at 4 °C prior to liquid chromatography mass spectrometry (LC/MS) analysis. For the studies involving model peptides, NHAc-X-OMe were replaced by unprotected GAIL-X-GAILR (where X = C, M, Y, W), and the reactions were conducted at both pH 6.0 and pH 7.5.

8.3.6 Isolation and characterization of NAPQI amino acid adducts
For residues that were found to react, the above synthesis was scaled up to produce sufficient material for NMR characterization. For this, one quantity of NAPQI in acetonitrile (10 mL) was added to a stirred suspension of NHAc-X-OMe (0.2 equiv. relative to the original quantity of APAP) in either H₂O (50 mL, for NHAc-Cys-OMe, NHAc-Met-OMe) or 1:1 H₂O:MeOH (50 mL, for NHAc-Tyr-OMe, NHAc-Trp-OMe) and stirred overnight at ambient temperature with the exclusion of light. The reaction mixture was then concentrated to ca. 10-15 mL in vacuo and any precipitate separated by centrifugation. The crude supernatant was purified via preparative HPLC. Characterization data for synthetic compounds are provided in the Appendix.

8.3.7 Relative consumption of the amino acids by reaction with NAPQI.
NAPQI solution in acetonitrile (75 µL) was added to a mixture containing NHAc-X-OMe where X = Cys, Met, Tyr, and Trp (10 µM each) in sodium phosphate buffer (500 µL, 10 mM) at either pH 6.0 or pH 7.5. Reaction solutions also contained NHAc-Ala-OMe (10 µM) as an internal standard. For control experiments, acetonitrile (75 µL) that did not contain NAPQI was added. The reaction solutions were continuously shaken for 2 h and analyzed by LC/MS. EICs were plotted for each protonated NHAc-X-OMe starting material, and the yield of NAPQI
adduct was taken as the percentage consumption of NHAc-X-OMe between the control and NAPQI-treated reactions.

8.3.8 Preparative HPLC
Purification of the NAPQI amino acid adducts was performed using preparative HPLC on an Agilent Technologies 1200 Series system consisting of a Rheodyne manual injector, Agilent G1311A Quaternary Pump and Agilent G1365B MWD UV detector. Separations were performed on a reverse phase Phenomenex Kinetex 5 µm C18 100 Å, AXIA packed LC column 150 x 21.2 mm. Elution was carried out with acetonitrile and H2O acidified with 0.1 % TFA. Runs were performed in a gradient fashion from 20-90 % acetonitrile over 70 min at a flow rate of 8 mL min\(^{-1}\). Separation of tryptophan adducts required isocratic flow at 15 % acetonitrile without TFA buffer. Fractions were collected with an Agilent G1328B fraction collection system based on a set UV absorption threshold. Each fraction was analyzed by analytical RP-HPLC (Phenomenex Kinetex 5 µm C18 100 Å 150 x 4.6 mm, 0-100 % acetonitrile over 25 min at a flow rate of 1 mL min\(^{-1}\)) and ESI-MS before removal of the solvent by freezing at -80 °C and subsequent lyophilization. Product identification was performed by NMR spectroscopy.

8.3.9 Microsome incubations
APAP (10 µL, 20 mM), test GAIL-X-GAILR peptide (10 µL, 100 mM) and microsomes (5 µL, 20 mg mL\(^{-1}\) protein) were combined in phosphate buffer (165 µL, pH 7.44), and the solution was pre-incubated at 37°C for 5 min with shaking. The reactions were initiated by the addition of NADPH (10 µL, 20 mM) bringing the total incubation volume to 200 µL. The incubation was quenched after 90 min by the addition of ice cold acetone (800 µL) followed by vortex mixing. The mixture was centrifuged at 18000 x g for 10 minutes and 900 µL of the resultant supernatant was lyophilized. The residue was resuspended in H2O (100 µL) and stored at 4 °C prior to LC/MS analysis.

Thiol depletion studies followed the above procedure, however, microsomes (5 µl, 20 mg mL\(^{-1}\) protein) and APAP (10 µL, 20 mM) in phosphate buffer (160 µL, pH 7.4) were pre-incubated with N-ethylmaleimide (NEM, 5 µL, 20 mM) at 37 °C for 10 min prior to addition of reactant peptides and NADPH.
8.3.10 Liquid chromatography mass spectrometry (LC/MS)
All LC/MS and LC-MS/MS experiments were performed using the Q-TOF mass spectrometer. For analysis of NAPQI reaction products with single amino acid residues the following solvent gradient was used [time (min), %B solvent]: [0,0], [1,0], [11,100], [15,100], [16,0], [21,0]. For separation of the peptides, microsome incubation extracts, competition reaction mixtures and selected single residue reaction products, a longer gradient was used: [0,0], [1,0], [41,100], [44,100], [45,0], [50,0]. In each case, 10 µL of sample was injected. For data-dependent collision-induced dissociation (CID) experiments (LC/MS/MS), a collision energy of 25 V was applied, and ion isolation windows were typically set to 4.0 \textit{m/z}.

8.3.11 Direct infusion ESI-MS/MS
High-resolution mass spectrometry analysis and low-energy CID spectra were collected using the LTQ-FTICR mass spectrometer. High resolution analysis was performed via transferring ions from the ion trap to the FT-ICR cell with measured \textit{m/z} values within 1 ppm of those theoretically predicted. For CID experiments, an isolation width of 1.5 \textit{m/z} was used to select a single isotope peak and ions were activated for 30 ms with a normalized collision energy (NCE) sufficient to deplete approximately 90% of the precursor ion (typically a NCE of about 20%).

8.4 Chapter 4 methods
8.4.1 Software development
8.4.1.1 General
The CRM-peptide adduct identification software was written in Python (v2.7) on a machine running Ubuntu 14.04 LTS. Numerous external libraries\textsuperscript{411-414} are employed to facilitate common functions and a list of these is provided in Table 8-2.
Table 8-2 External software libraries used in the production of the non-targeted protein-CRM adduct identification algorithm. Only libraries not included in standard Python distributions are listed here.

<table>
<thead>
<tr>
<th>Library</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>pymzml</td>
<td>Extracting LC-MS/MS data form mzML files</td>
</tr>
<tr>
<td>MascotParser</td>
<td>Parsing of Mascot .dat search output files</td>
</tr>
<tr>
<td>Rtree</td>
<td>Spatial indexing and local maxima detection for HiTIME output files</td>
</tr>
<tr>
<td>RDkit</td>
<td>Creation and manipulation of chemical structures via SMILES/SMARTS</td>
</tr>
<tr>
<td>Pyteomics</td>
<td>Determination of theoretical peptide MS$^2$ spectra</td>
</tr>
<tr>
<td>Numpy</td>
<td>General array functions</td>
</tr>
<tr>
<td>Matplotlib</td>
<td>Selected graphical plotting actions</td>
</tr>
<tr>
<td>Qt/PyQt</td>
<td>Development and implementation of the graphical user interface</td>
</tr>
<tr>
<td>JSME</td>
<td>Graphical chemical structure editor and SMILES string generator</td>
</tr>
<tr>
<td>Pyqtgraph</td>
<td>Generation of data plots displayed in the graphical user interface</td>
</tr>
<tr>
<td>MSSimulator</td>
<td>\textit{in silico} production of LC-MS data sets for algorithm evaluation</td>
</tr>
</tbody>
</table>

8.4.1.2 Validation data

In order to develop the software in a controlled and methodical manner, a simulated LC-MS/MS dataset was produced using MSSimulator\textsuperscript{415}, a program that produces synthetic LC-MS/MS data which is distributed as part of OpenMS.\textsuperscript{253} This software takes a FASTA file, performs an \textit{in silico} trypsin digestion of the supplied sequence, simulates retention times, MS and MS/MS spectra for each tryptic peptide then writes an mzML file\textsuperscript{251} that can be used for testing and validation. While MSSimulator does allow the user to include various endogenous and exogenous PTMs to the \textit{in silico} peptides, the possible modifications are pre-defined and cannot be customised to suit the purposes at hand. Modifications could be made to the MSSimulator C++ source code to include functions that allow for the application of arbitrary PTMs however this task was deemed to be beyond the scope of this work. Therefore, a custom software pipeline was developed that is capable of producing synthetic LC-MS/MS data from input FASTA files that also allows user-specified arbitrary PTMs to be applied.

The core user inputs to this pipeline are one or more FASTA files, the chemical formula of the desired PTM, and the mass difference (in Da) between natural abundance and isotope enriched forms. The FASTA file is initially parsed and segmented into tryptic peptides using functions provided in the pyteomics library.\textsuperscript{412} For simplicity in the charge state distribution of peptides
in the final data, peptides that have >1 arginine or lysine residue (i.e. missed tryptic cleavages) as well as those that have <1 (i.e. a C-terminal peptide) are discarded. Two processes are then conducted for each remaining peptide:

1. A FASTA file is then produced for each tryptic peptide and these are sequentially passed to MSSimulator to generate an mzML file. This produces one separate mzML file for each tryptic peptide that contains the features for that peptide only.

2. Mascot MGF file entries are produced for each peptide. ‘y’-type sequence ions are generated by sequentially cleaving residues from the peptide N-terminus and determining the \( m/z \) value of the remaining fragment using the pyteomics fast_mass function. For peptides that contain the user-specified modification target residue, a second MGF entry is written where the mass of the desired PTM is applied to the relevant sequence ions.

The separate mzML files (containing only features for 1 specific tryptic peptide) are then combined to produce the final output file. Where these correspond to PTM target peptide, the LC-MS feature for the native peptide is copied and translated through the \( m/z \) dimension by a distance equal to mass of the PTM divided by the charge state (+2 by default in all cases).

Synthetic data was generated for human serum albumin (HSA). To simulate the effect of protein adduction by heavy and light forms of a CRM, the unlabelled \((^{12}\text{C})\) and \(^{13}\text{C}_6\) forms of the CPPQI metabolite of the nontricyclic antidepressant trazodone were applied to cysteine residues. The datasets that will be taken as inputs to the software were produced by taking the MS\(^1\) spectra and processing these with HiTIME using a heavy/light \( m/z \) shift of 3.01005 corresponding to a peptide in the +2 charge state. The MS\(^2\) spectra were searched with Mascot\(^{416}\) against the UniProt database.\(^{340}\) The immediate result of these two procedures is: 1) a HiTIME data file containing a list of twin-ion candidates, and 2) a Mascot data file containing peptide and protein assignments for each query MS\(^2\) spectrum. These two files then serve as input for the software to be developed.

The python script and support files used to generate the above data can be obtained from [https://github.com/mgleeming/phdScripts/tree/master/dataGen](https://github.com/mgleeming/phdScripts/tree/master/dataGen). The dependencies for this script are: pyteomics, numpy, pymzml, openMS and pyopenms.
8.5 Chapter 5 Methods

8.5.1 Materials

Acetaminophen (APAP), sequencing grade modified trypsin, iodoacetamide, triethylammonium bicarbonate (TEAB) buffer, urea and reduced nicotinamide adenine dinucleotide 2’-phosphate tetrasodium salt (NADPH) were purchased from Sigma-Aldrich (St. Louis, MO, USA). Pooled rat liver microsomes (20 mg mL\(^{-1}\) protein, 0.462 nmol p450 mg\(^{-1}\) protein, from 83 uninduced, male Sprague-Dawley donors, lot number RT053C) were from Life Technologies (Carlsbad, CA, USA), and stored at -80 °C until use. Ring-labelled \(^{13}\)C\(_6\)-APAP [>99% isotope enrichment] and tris-(2-carboxyethyl)phosphine hydrochloride were from IsoSciences (King of Prussia, PA, USA) and Thermo Fisher Scientific (Waltham, MA, USA), respectively. Oasis HLB Solid phase extraction cartridges (10 mg) were from Waters (Milford, MA, USA), formic acid was from Fluka. HPLC grade acetone and acetonitrile were from Merck (Kenilworth, NJ, USA), and 18 MΩ H\(_2\)O was obtained using a Milli-Q apparatus.

8.5.2 Microsome incubations

Either aqueous APAP/\(^{13}\)C\(_6\)-APAP solution 1:1 (10 µL, 20 mM) or water (10 µL) were combined with microsomes (10 µL, 20 mg mL\(^{-1}\) protein) in potassium phosphate buffer (170 µL, 100 mM, pH 7.4) and the mixtures were pre-incubated at 37 °C for 10 mins with continuous agitation. The reactions were initiated by the addition of NADPH (10 µL, 20 mM) and incubated for 3h at 37 °C. At both \(t = 60\) and \(t = 120\) mins, an additional aliquot of NADPH (10 µL, 20 mM) was added. Reactions were quenched at \(t = 180\) min via the addition of ice-cold acetone (1 mL). Samples were vortex mixed for 30 s then stored at 4 °C overnight while precipitated proteins pelleted under gravity. The samples were centrifuged (16,000 x g, 10 mins, 4 °C), the supernatant was removed and the remaining protein pellet gently washed by overlaying acetone (1 mL). After removal of the supernatant, the pellet was then resuspended in TEAB buffer (200 µL, 50 mM, pH 8.5) containing urea (8 M) and TCEP (10 mM) and sonicated until complete dissolution. The samples were incubated at 37 °C for 45 mins and allowed to cool to ambient temperature before iodoacetamide (122 µL, 100 mM) was added. Samples were incubated in darkness for 60 mins then diluted to a final urea concentration of 1M with TEAB (25 mM). Trypsin (4 µg) was added to each sample and digestions proceeded.
at 37 °C for 24 h. The samples were acidified with formic acid to a final concentration of 1% (v/v) and purified by solid phase extraction as follows: (i) the cartridge was pre-conditioned with 1 mL of 80% acetonitrile containing 0.1% TFA followed by 2 x 1.2 mL of 0.1% aqueous TFA; (ii) the sample was loaded onto the cartridge; (iii) the cartridge was washed with 2 x 1.2 mL 0.1% TFA; (iv) The peptides were eluted from the cartridge with 1 mL 80% acetonitrile containing 0.1% TFA; and (v) The eluent was concentrated on a centrifugal evaporator to approximately half the original volume then lyophilised overnight and resuspended in 0.1% formic acid (100 µL) and stored at -20 °C until the time of analysis via nano LC-MS/MS experiments (described below). Experiments were performed in triplicate for both APAP and vehicle control (VC) treatments.

8.5.3 Data analysis
Raw MS$^1$-only LCMS data files were first converted to the open source ‘mzML’ format$^{251}$ with a lower intensity threshold of 5,000 counts using MSConvert.$^{250}$ MS$^1$ data were scored with HiTIME$^{334}$ using a heavy/light spacings of 3.01005 and 2.0067 Da (corresponding to the 6.0201 Da difference between APAP and $^{13}$C$_6$ APAP in the 2+ and +3 charge states respectively). For database searching of MS$^2$ LC-MS runs, MGF files were produced using MSConvert that contained the 100 most intense fragment ions in each HCD spectrum. Peptide and protein identifications were performed using Mascot$^{341}$ (v2.4.1) and searches of the Uniprot database$^{340}$ (release 2015_07) accounted for the possibility of 1 missed tryptic cleavage, carbamidomethyl cysteine and oxo-methionine variable modifications. The MS$^1$ and MS$^2$ m/z tolerance were set to 20 ppm and 0.2 Da respectively. Further variable modifications were used in some searches as described in-text. Significance thresholds were set to achieve a global false discovery rate of 1% using a target-decoy approach against sequence-reversed proteins.$^{417}$ Proteins with two or more sequence-unique peptides were retained and single-peptide proteins have been omitted. Venn diagrams were produced using matplotlib-venn 0.11.4.

8.5.4 Liquid chromatography-mass spectrometry
Analyte peptides were introduced into the gas phase via nano-ESI using a spray voltage of 1.9 kV, capillary temperature of 275 °C and all MS$^1$ mass spectra were acquired in the positive ion mode over a range of 400-1500 m/z in the Orbitrap analyser operating at a resolution of 120,000. Data were collected for each sample in both MS and MS/MS modes. For MS/MS
runs, data-dependent MS$^2$ spectra were collected by subjecting 2+ to 5+ precursor peptide ions with a measured intensity greater than $10^4$ to higher-energy collision induced dissociation (HCD) with a normalised collision energy of 35%. Dynamic exclusion was used to prevent the selection of the same precursor ion more than once for 45 s. For each run, 2 μL of sample was injected and samples were held in the autosampler at 5 °C prior to analysis.

8.5.5 Production of semi-synthetic data sets

Semi-synthetic data sets were produced by computationally creating synthetic data features that mimic the elution and fragmentation of APAP/$^{13}$C$_6$APAP modified peptides and superimposing these onto experimental data sets produced from analysis of vehicle control samples. There are numerous advantages to this approach that include: 1) the number and location of ‘true’ twin-ion signals is precisely known, 2) the complexity and confounding features of experimental data are maintained and accounted for, and 3) numerous properties and parameters of the data analysis described here can be rapidly assessed.

Semi-synthetic data sets were produced for each of the three vehicle control replicates. For each replicate, two data files were created: 1) an mzML (LCMS) data file containing the synthetic twin-ion features superimposed on experimental MS$^1$ data, and 2) an MGF file containing all entries from experimental data in addition to those for synthetic peptide signals.

Briefly, the procedure was as follows:

1) Parse the experimental mascot results file and obtain a list of peptides that were confidently assigned as having exactly: (i) no missed enzymatic cleavages, and (ii) exactly one carbamidomethyl-cysteine residue. These requirements were applied simply to constrain the number of peptides.

2) For each peptide, identify the experimentally observed ions in the MS$^2$ spectrum that were assigned by mascot as corresponding to fragments that contain the cysteine residue.

3) Produce a synthetic MS$^2$ peak list by incrementing the $m/z$ value of these fragments by the desired mass of the reactive metabolite modification. The result of this is an MS$^2$ spectrum identical to the experimental entry except that mascot-assigned fragments containing cysteine have been offset by the modification mass.

4) MGF entries were created for these synthetic MS$^2$ spectra which were then combined into the MGF file derived from experimental LC-MS/MS data.
5) For each peptide identified in step 1, a synthetic LC-MS signal was produced using MSSimulator.\textsuperscript{415} These signals take into account the predicted isotope distributions from different elemental compositions and intensity values were matched to that of the experimental peptide. Retention times were set at a randomly chosen point between 0.5 and 5 minutes after elution of the experimental peptide.

6) Each synthetic signal was then added into the experimental mzML file. The synthetic data were superimposed on existing experimental data in the target region to replicate aspects of LC-MS experiments such as overlapping peaks. A new mzML file containing the final semi-synthetic data was written and used in subsequent processing steps.

A Python script utilising the OpenMS library\textsuperscript{253, 413} was written to automate the data creation process which can be obtained from the Xenophile project GitHub repository at https://github.com/mgleeming/Xenophile.

The MGF and mzML files were then processed with Mascot and HiTIME respectively using the parameters described in the main text and the results of these were then subjected to the same postprocessing routines described in-text.

8.5.6 Formulae determination metrics
Estimates of the number of candidate formulae assigned to a given CRM mass were produced for a wide variety of molecules. The ZINC15 database\textsuperscript{359} of small molecule drugs approved by the United States Food and Drug Administration (containing 1385 chemical entities) was used. Molecules containing fewer than 5 carbon or hydrogen atoms were removed resulting in a list of 1337 compounds ranging in molecular weight from \textit{ca.} 90–1000 Da. For each of these compounds, fragments were generated by disconnection of rotatable bonds as described in the main text. Allowed elemental compositions were computed by taking the lowest and highest number of a given element observed in any fragment generated. These upper limits of these ranges were then extended by addition of C\textsubscript{8}H\textsubscript{10}O\textsubscript{8} to allow for the possibility of metabolic reactions that substantially increase the mass of the CRM such as Glucuronidation. The lower limits were decreased by H\textsubscript{2} to allow for oxidations such as formation of NAPQI from APAP. The mass of all combination of stoichiometries within these ranges were then computed to give the set of allowed CRM compositions.

A theoretical CRM mass was then produced by incrementing the mass of the precursor molecule by a value randomly selected from 4 possibilities: 1) -2.01565 (-2H), 2) 13.97926
(+O, -2H), 3) 15.99491 Da (+O), and 4) 176.03209 (+6C 8H 6O). Formulae within 100 ppm of the CRM target mass were then retrieved from the target list and residual mass errors were calculated for each.

8.5.7 Post-processing of HiTIME data

Raw HiTIME data was subjected to various post-processing steps using the in order to derive a list of possible twin-ion locations. This was achieved using the “Post-processing” tab of the “HiTIME search” tools of the Xenophile software. Specifically, a baseline noise level was estimated by inspecting data heat maps and score distribution histograms then point below this level were removed. Local maxima detection of the remaining data provided sites of probably twin-ions.

It is known that peptide signals in LC-MS data typically display multiple $^{13}$C isotope peaks in addition to the monoisotopic peak and, given this, unique HiTIME local maxima should be observed corresponding to each peak in a peptide isotope distribution. This feature can be used to elevate the scores of signals where multiple HiTIME maxima are observed separated by a distance corresponding to the charge state under analysis. For the current data, this feature was enabled by selecting the ‘Peptide Isotope Scaling’ checkbox in the ‘Post-processing’ setup tab of the Xenophile software.

The final set of HiTIME local maxima points, with scores that had been scaled according to the presence of $^{13}$C isotope peaks, were then manually reviewed using the ‘Results Viewer’ tab of the Xenophile software. In general, hits that approximately satisfied the condition: $H_{SC} - H_{max}^{raw} > 0$, where $H_{SC}$ is the score of a given hit following peptide isotope scaling and $H_{max}^{raw}$ is the maximum score or the raw (unscaled) HiTIME data were manually reviewed. This ultimately selects data points that also have $^{13}$C isotope peaks and are likely to be due to modified peptide signals.

8.6 Chapter 6 methods

8.6.1 Materials

$^{13}$C$_6$-benzene, $^{13}$C$_6$-toluene and $^{13}$C$_6$ bromobenzene and benzene were purchased from Sigma-Aldrich. Toluene was from Merck and rat liver microsomes (lot number RT053L) were
purchased from Life Technologies. All other chemicals were from commercial sources listed previously.

8.6.2 Microsome incubations and sample preparation
PhH/\(^{13}\)C\(_6\)-PhH, PhMe/\(^{13}\)C\(_6\)-PhMe, PhBr/\(^{13}\)C\(_6\)-PhBr incubations with rat liver microsomes were carried out in triplicate as described in section 8.5.2 with minor changes. In these experiments, the concentration of each individual substrate isotopologue was increased to 2 mM.

Glutathione trapping studies were carried out similarly to the above but a portion of the phosphate buffer (10 µL) was replaced with GSH (10 µL, 100 mM) in the incubation mixture. These reactions were quenched after 1 h by the addition of ice-cold acetone (1 mL). Samples were vortex mixed (ca. 30 s) and centrifuged (16,000 x g, 10 mins, 4 °C). The supernatant was then removed, evaporated to dryness and resuspended in H\(_2\)O (100 µL) and stored at 4 °C until analysis.

8.6.3 Liquid chromatography-mass spectrometry
Incubation supernatants largely containing small molecules were analysed using the Q-TOF mass spectrometer in both MS-only and MS/MS modes. The LC-MS method was the same as that describe for NAPQI GAIL-X-GAILR reaction analysis (section 8.3.10). For data-dependent MS\(^2\) analysis runs, ions with an absolute intensity greater than 500 counts were selected and activated with a collision energy of 25 V and selection preferences were given to twin-ions identified by HiTIME searching. Ions were excluded from further selection for a period of 45 seconds after 5 MS\(^2\) acquisitions.

For protein digest analysis, LC-MS and LC-MS/MS were carried out using the Orbitrap mass spectrometer as described for APAP microsomal protein digest analysis above (Section 8.5.4) with minor differences to the analytical method. In this case, the solvent gradient was: [Time (mins), %B]: [0, 3], [6, 3], [95, 22], [105, 40], [110, 80], [115, 80], [117, 3], [125, 3]. All other parameters were set as previously.

8.6.4 Data analysis
For HiTIME analysis of LC-MS data, the raw Agilent ‘.d’ or Thermo ‘.raw’ data files were converted to the open source mzML\(^{251}\) format using MSConvert\(^{250}\) with a lower intensity threshold of 500 counts. HiTIME analysis was performed with a doublet mass spacing (\(\Delta m/z\))
of 6.0201 for incubation supernatants or 3.01005 and 2.0067 for protein digests. HiTIME search and post processing parameters were the same as those used in Chapter 5. MGF files were produced by selecting the 6 most abundant ions per 100 Th for each MS$^2$ spectrum. Initial protein identification was performed using Mascot (v2.4.1) implemented through Proteome Discoverer (v2.1 SP1) against the Uniprot database$^{340}$ (release 2015_07). One missed tryptic cleavage was allowed and MS$^1$ and MS$^2$ m/z tolerances were set to 10 ppm and 0.6 Da respectively. Variable modifications were applied as described in the main text. Peptide-spectral matches were validated with percolator$^{418, 419}$ and protein level assignments were verified using the FidoCT algorithm$^{420}$ using default parameters. Extracted ion chromatograms were plotted with an m/z tolerance of +/- 0.03 Da.
9 Appendix

9.1 Xenophile software installation guide

Many common analysis functions that are performed on shotgun LC-MS/MS based proteomics data require the ability to process multiple large data files, which can be a computationally intensive process. As such, it is common for laboratories that routinely conduct these analyses to use remote computing services such as high-performance compute clusters or cloud-based infrastructure. Examples of this include the Nectar research cloud (https://nectar.org.au/research-cloud/) available to researchers in Australia as well as numerous commercial equivalents such as Elastic-Compute-Cloud (EC2) operated by Amazon. These services are most commonly utilised with Linux-based operating systems.

To enable easy deployment of the Xenophile software on cloud computing infrastructure, scripts have been developed to automate the installation of Xenophile and its dependencies on Ubuntu, which is the most commonly utilised Linux distribution.

Beginning from a fresh install of Ubuntu 14.04, Xenophile and all dependencies can be downloaded and installed by executing the commands in Figure 9-1.
# install foundation stuff
sudo apt-get update
sudo apt-get upgrade
sudo apt-get install python-pip
sudo pip install --upgrade pip
sudo apt-get install python-setuptools python-dev build-essential automake
autoconf libtool git

# install scipy stack
sudo apt-get install python-tk
sudo pip install --user numpy scipy matplotlib ipython jupyter pandas sympy
nose

# install pyteomics and pymzml
sudo pip install lxml pyteomics pymzml

# install RDkit
sudo apt-get install python-rdkit librdkit1 rdkit-data

# install libspatialindex rtree
git clone https://github.com/libspatialindex/libspatialindex.git
cd libspatialindex/
./autogen.sh
./configure; make; make install
sudo apt-get install libspatialindex-dev
ldconfig
sudo pip install rtree
cd

# install qt4
sudo apt-get install libglew-dev libcheese7 libcheese-gtk23 libclutter-gst-2.0-0 libcogl15 libclutter-gtk-1.0-0 libclutter-1.0-0

# install pyqt4
sudo apt-get install python-qt4 qt4-dev-tools python-qt4-dev pyqt4-dev-tools

# install xenophile
git clone https://github.com/mgleeming/Xenophile.git
cd Xenophile
sudo python setup.py install
cd

Figure 9-1. Shell script that automates the installation of Xenophile and its dependencies for Ubuntu 14.04.

In selected cases described in this thesis, it was desirable to deploy the Xenophile software on a cloud computing environment to allow simultaneous processing of large data files. Given that processing of many large data files (for example, biological and technical replicates of different treatment groups) may be a common necessity for potential users of this software, a detailed description of how this was accomplished is provided below.
The instructions provided here are for deployment on Amazon’s EC2 cluster. This infrastructure was selected for this demonstration as it is accessible to anyone who desires to use this software and is not a private cluster limited to one research institute, state or country and has been employed by others in many biomedical cloud-computing tasks. While EC2 is commercial, pricing is generally quite accessible and in the order of cents-per-usage-hour. The total cost for all usage toward this thesis was less than AU$10.

First, create an account with Amazon Web Services (AWS).

The StarCluster framework was used to manage job submission and resource allocation as well as load balancing and transfer of data to and from the cluster. This is a third-party application installed on a local machine that largely automates the process of initialising the setup of the cloud-computing environment. It should be emphasised that usage of StarCluster is not necessary and instances can be readily configured using the Amazon web console and accessed using standard methods. To begin, install StarCluster. Note that as of January 2017, it is important to retrieve the package from the development branch in order to access new generation EC2 instance types such as “c4.large”. The commands provided here are for installation of StarCluster using a machine running OSX or Linux. Similar procedures are available for Windows machines.

```
sudo -H pip install https://github.com/jtriley/StarCluster/archive/develop.zip
```

Next, modify the StarCluster configuration file as desired. Details can be obtained from the quick start guide at http://star.mit.edu/cluster/docs/latest/quickstart.html. Important points to consider when modifying the configuration file that are not discussed in this guide are:

1) The default geographical region for EC2 resource allocation is us-east-1. It is desirable to select the closest region to the location of the user to increase performance and minimise costs. If a different region is required, consult the official list from Amazon.

For example, based in Melbourne, Australia, the ‘ap-southeast-2’ region is selected.

* Link to Amazon Web Services homepage: https://aws.amazon.com/
† Link to StarCluster installation guide: http://star.mit.edu/cluster/docs/latest/installation.html
‡ Amazon regions and endpoints list: http://docs.aws.amazon.com/general/latest/gr/rande.html#ec2_region
2) Update the default instance type if desired. The instance type determines the amount of resources (CPUs, memory etc.) available in one node of the cluster formed and also the usage price per hour.

3) Update the default machine image if desired. The machine image is the operating system that will be deployed on the instances created.

Machine images can be customised so that software and tools specific to a given project are installed and available for immediate deployment on newly created instances. To install Xenophile and its dependencies, execute:

```
starcluster start -o -s 1 -I t2.micro -m ami-807876e3 imagehost
```

This creates a single-node t2.micro instance called “imagehost” that used the Ubuntu Server 14.04 machine image “ami-807876e3”. Configuring a new machine image using a single node of a small instance type is desirable to reduce costs. Note that this machine image must also be specified in the StarCluster configuration file.

Once the instance is running, log into the cluster:

```
starcluster sshmaster imagehost
```

Make any desired modifications. For Xenophile installation, execute the commands in Figure 9-1. Once all modifications are done, create a new machine image containing all of the software installed. From the local machine (that is running starcluster), execute:

```
starcluster listclusters
```

This will produce an output similar to the below:

```
Note that machine image availability may differ between geographical regions. Consult the EC2 marketplace for similar machine images in your region.
```
In this example, the value ‘i-1s144df68741d6’ on the third to last line is the identification number of the imagehost instance. To save a new machine image of an EBS-backed instance, execute:

```
starcluster ebsimage i-1s144df68741d6 myNewMachineImage
```

This generates the output below and saves the new image. The identification number of the new image in this case is ‘ami-4k83sfh4’.

```
StarCluster - (http://star.mit.edu/cluster) (v. 0.95.6)
Software Tools for Academics and Researchers (STAR)
Please submit bug reports to starcluster@mit.edu

>>> Removing private data...
>>> Cleaning up SSH host keys
>>> Cleaning up /var/log
>>> Cleaning out /root
>>> Cleaning up /tmp
>>> Creating new EBS AMI...
>>> New EBS AMI created: ami-4k83sfh4
>>> Fetching block device mapping for ami-4k83sfh4
>>> Waiting for snapshot to complete: snap-
    snap-1: |QUIT|QUIT|QUIT|QUIT|QUIT|QUIT|QUIT|QUIT|QUIT|QUIT|100% Time: 00:02:35
>>> Waiting for ami-4k83sfh4to become available...
>>> create_image took 3.547 mins
>>> Your new AMI id is: ami-4k83sfh4
```

Updating the default machine image ID in the StarCluster configuration file will allow creation of large multi-node clusters that are each running the customised image.

To start a new cluster, execute:
starcluster start myClusterName

where ‘myClusterName’ is the desired name of the new cluster. Once initialisation is complete, data can be transferred to the cluster by executing:

starcluster put myClusterName /path/to/local/data /remote/target/path

To retrieve data on the cluster and download to the local machine:

starcluster get myClusterName /path/to/remote/data /local/target/path

If use of the Xenophile GUI is desired, enable X11 forwarding by specifying the -X flag when accessing the instance:

starcluster sshmaster -X myClusterName

9.2 Xenophile license

Xenophile is released under the 3-clause BSD license. If redistributing this code, licensing restriction of the included JSME and MascotParser libraries must be considered. The licenses for these packages are included in the ‘license.txt’ file in the Xenophile GitHub repository.
9.3 Chapter 2 supplementary information

9.3.1 HiTIME shape fitting and score determination

Each data point is described by a retention time, m/z ratio and abundance (denoted here as \(t, m, a\)). For a given spectrum, the data consists of \(d = (m, a)\). The Gaussian distribution (eq 1) where the mean is the m/z value of the point under consideration and the variance is related to the ppm error tolerance.

\[
g(m) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{m - \mu}{\sigma} \right)^2}
\]

9-1

The sum of squared errors (SSE) can then be computed according to:

\[
SSE(a, g(m)) = \sum (a - g(m))^2
\]

9-2

The goodness of fit of the Gaussian function model to this data is quantified by calculating Pearson’s correlation coefficient. These values range from 0 to 1 with higher coefficients indicating greater correlation between the data and the target shape.

\[
r = \frac{\sum (a - \bar{a})(g(m) - \bar{g})}{\sqrt{\sum (a - \bar{a})^2 \sum (g(m) - \bar{g})^2}}
\]

9-3

While the correlation coefficient describes the quality of the fit between the data and the applied Gaussian distribution it cannot be used directly for HiTIME scoring. This is because a non-twin-ion peak (i.e. an LC-MS signal derived from an endogenous compound that has no ‘heavy’ counterpart) attains a moderate score as a result of matching to one of the twin-ions despite the absence of the other.

To account for this, correlations between the data and each of three different target shapes are computed. The three target shapes are: 1) both heavy and light peaks (denoted as AB), 2) the light peak alone (A0), and 3) the heavy peak alone (B0). Using this approach, if a high correlation is observed for either the light or heavy isotope while the other is low, then this data is likely not a twin-ion. By contrast if similar correlations are observed for light and heavy this data strongly matches the twin-ion shape.

To condense these metrics into a single value representing the HiTIME score, Steiger’s Z-test (a.k.a. Meng’s Z-test) for correlated correlation coefficients\(^{422, 423}\) is applied.
\[
Z = (z_{r1} - z_{r2}) \sqrt{\frac{N - 3}{2(1 - r^2)h}}
\]

The value of \( h \) is computed according to:

\[
h = \frac{1 - f r^2}{1 - r^2}
\]

where:

\[
f = \frac{1 - r_x}{2(1 - r^2)}
\]

and,

\[
r = \frac{r_1^2 + r_2^2}{2}
\]

Additionally, \( N \) is the number of data points, \( z_{r1} \) and \( z_{r2} \) are the test correlation coefficients following Fisher z-transformation defined by:

\[
z = \frac{1}{2} \ln \left( \frac{1 + r}{1 - r} \right) = \text{arctanh}(r)
\]

Z-values are computed for the correlations between AB-A0 and AB-B0 and the minimum of these is taken as the final HiTIME score for the data region.
9.3.2 Supplementary tables and figures

Table 9-1. Variables employed in the HiTIME fitting algorithm.

<table>
<thead>
<tr>
<th>Description</th>
<th>Parameter</th>
<th>Units</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retention time</td>
<td>$rt$</td>
<td>minutes, seconds or scans</td>
<td>data variable</td>
</tr>
<tr>
<td>Mass to charge ratio</td>
<td>$mz$</td>
<td>$m/z$</td>
<td>data variable</td>
</tr>
<tr>
<td>Signal intensity</td>
<td>$intensity$</td>
<td>dimensionless</td>
<td>data variable</td>
</tr>
<tr>
<td>Difference between twin-ion masses</td>
<td>$delta_{mz}$</td>
<td>Dalton</td>
<td>known masses of compounds</td>
</tr>
<tr>
<td>Full width at half maximum height of elution peak in retention time dimension</td>
<td>$FWHM_{rt}$</td>
<td>minutes, seconds or scans</td>
<td>sampled peaks in the data</td>
</tr>
<tr>
<td>Full width at half maximum height of elution peak in mass to charge dimension</td>
<td>$FWHM_{mz}$</td>
<td>$m/z$ parts per million</td>
<td>sampled peaks in the data</td>
</tr>
<tr>
<td>Retention time boundary</td>
<td>$rt_{bound}$</td>
<td>scalar multiple of $FWHM_{rt}$</td>
<td>estimated from sampled peaks in the data</td>
</tr>
<tr>
<td>Mass to charge boundary</td>
<td>$mz_{bound}$</td>
<td>scalar multiple of $FWHM_{mz}$</td>
<td>estimated from sampled peaks in the data</td>
</tr>
<tr>
<td>Intensity ratio</td>
<td>$intensity_{ratio}$</td>
<td>dimensionless</td>
<td>known ratio of compounds</td>
</tr>
</tbody>
</table>

Table 9-2. Transformation parameters of MS data for APAP metabolite detection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary encoding precision</td>
<td>32 bit</td>
</tr>
<tr>
<td>Use zlib compression</td>
<td>False</td>
</tr>
<tr>
<td>Package in gzip</td>
<td>False</td>
</tr>
<tr>
<td>Write index</td>
<td>False</td>
</tr>
<tr>
<td>TPP compatibility</td>
<td>False</td>
</tr>
</tbody>
</table>
Table 9-3. Parameters employed in HiTIME data processing.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>intensityRatio</td>
<td>1.0</td>
</tr>
<tr>
<td>rtWidth</td>
<td>17</td>
</tr>
<tr>
<td>rtSigma</td>
<td>1.5</td>
</tr>
<tr>
<td>mzWidth</td>
<td>150</td>
</tr>
<tr>
<td>ppm</td>
<td>4</td>
</tr>
<tr>
<td>mzSigma</td>
<td>1.5</td>
</tr>
<tr>
<td>mzDelta</td>
<td>6.0201</td>
</tr>
<tr>
<td>minSample</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Table 9-4. Parameters employed in MetExtract data processing.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity threshold</td>
<td>500</td>
</tr>
<tr>
<td>Number of X atoms</td>
<td>Min = 6, Max = 6</td>
</tr>
<tr>
<td>Max mass deviation (ppm)</td>
<td>20</td>
</tr>
<tr>
<td>Isotopic pattern count</td>
<td>1</td>
</tr>
<tr>
<td>Intensity abundance error</td>
<td>Non-Labelled ion = 0.1, Labelled ion = 0.1</td>
</tr>
<tr>
<td>EIC m/z width (ppm)</td>
<td>5</td>
</tr>
<tr>
<td>Correlation threshold</td>
<td>0.75</td>
</tr>
<tr>
<td>All others</td>
<td>Default</td>
</tr>
</tbody>
</table>
9.4 Chapter 3 supplementary data

9.4.1 Characterisation Data

**NAC-Cys(APAP)-OMe**

$^1$H NMR (500 MHz, DMSO-$d_6$) $\delta$ 9.70 (s, 1H, N$_b$H), 8.40 (d, $J$ = 7.6 Hz, 1H, N$_a$H), 7.46 (d, $J$ = 2.5 Hz, 1H, 2-H), 7.27 (dd, $J$ = 8.7, 2.5 Hz, 1H, 5-H), 6.75 (d, $J$ = 8.7 Hz, 1H, 6-H), 4.35 (td, $J$ = 8.1, 5.4 Hz, 1H, CH), 3.58 (s, 3H, CO$_2$CH$_3$), 3.20 (dd, $J$ = 13.3, 5.4 Hz, 1H, CH$_2$), 3.06 (dd, $J$ = 13.3, 8.5 Hz, 1H, CH$_2$), 1.98 (s, 3H, N$_a$HCOC$_H_3$), 1.84 (s, 3H, N$_b$HCOC$_H_3$).

$^{13}$C NMR (126 MHz, DMSO-$d_6$) $\delta$ 171.10 (CO$_2$CH$_3$), 169.34 (N$_a$HCOCH$_3$), 167.61 (N$_b$COCH$_3$), 151.91 (C-5), 131.50 (C-2), 122.17 (C-1), 119.87 (C-3), 119.47 (C-6), 114.97 (C-4), 51.98 (CH), 51.69 (CO$_2$CH$_3$), 33.18 (C$_H$), 23.76 (N$_b$COCH$_3$), 22.25 (N$_a$COCH$_3$).

HRMS: Calculated for [C$_{14}$H$_{18}$N$_2$O$_5$S + H]$^+$ = 327.10092, observed = 327.10097

**NAC-Met(APAP)-OMe**

Isomer A

$^1$H NMR (400 MHz, Methanol-$d_4$) $\delta$ 8.14 (d, $J$ = 2.6 Hz, 1H, 2-H), 7.56 (dd, $J$ = 8.9, 2.5 Hz, 1H, 5-H), 7.08 (d, $J$ = 8.9 Hz, 1H, 6-H), 4.56 (dd, $J$ = 8.3, 5.5 Hz, 1H, CH), 3.93 (ddd, $J$ = 12.8, 9.7, 5.9 Hz, 1H, $\beta$-CH$_2$), 3.73 (s, 3H, CO$_2$CH$_3$), 3.72 – 3.64 (m, 1H, $\beta$-CH$_2$), 3.33 (s, 3H, S-CH$_3$), 2.13 (s, 3H, N$_a$HCOC$_H_3$), 2.11 – 2.03 (m, 1H, $\gamma$-CH$_2$), 1.98 (s, 3H, N$_a$HCOC$_H_3$).

Isomer B

$^1$H NMR (400 MHz, Methanol-$d_4$) $\delta$ 8.15 (d, $J$ = 2.6 Hz, 1H, 2-H), 7.57 (dd, $J$ = 8.9, 2.5 Hz, 1H, 5-H), 7.09 (d, $J$ = 8.9 Hz, 1H, 6-H), 4.62 (dd, $J$ = 8.9, 4.9 Hz, 1H, CH), 3.89 (ddd, $J$ = 12.8, 9.3, 5.1 Hz, 1H, $\beta$-CH$_2$), 3.71 (s, 3H, CO$_2$CH$_3$), 3.68 (ddd, $J$ = 12.7, 9.2, 6.8 Hz, 1H, $\beta$-CH$_2$), 3.33 (s, 3H, S-CH$_3$), 2.28 (ddd, $J$ = 14.2, 9.3, 6.8 Hz, 1H, $\gamma$-CH$_2$), 2.13 (s, 3H, N$_b$HCOC$_H_3$), 2.00 (s, 3H, N$_a$HCOC$_H_3$), 2.07 – 1.93 (m, 1H, $\gamma$-CH$_2$).

HRMS: Calculated for [C$_{16}$H$_{23}$N$_2$O$_5$S]$^+$ = 355.13222, observed = 355.13208
N.B.: the purified NAc-Met(APAP)-OMe diastereomers isolated via preparative HPLC epimerised prior to analysis on the LTQ-FTICR mass spectrometer as indicated by LCMS analysis on the Q-TOF mass spectrometer. The HRMS value reported above is the aggregate value assigned to both isobaric ions on the ICR instrument. For comparison, m/z values for each isomer determined by LC-MS were 355.1327 and 355.1323 for isomers at retention time 13.8 and 14.1 min respectively.

NAc-Tyr(APAP)-OMe

$^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ 9.66 (s, 1H, $N_b$H), 8.27 (d, $J = 7.7$ Hz, 1H, $N_a$H), 7.36 – 7.28 (m, 2H, Ar-H), 6.98 (dd, $J = 8.2, 2.3$ Hz, 1H, Ar-H), 6.94 (d, $J = 2.2$ Hz, 1H, Ar-H), 6.78 (dd, $J = 8.2, 3.8$ Hz, 2H, Ar-H), 4.36 (td, $J = 8.2, 5.8$ Hz, 1H, CH), 3.59 (s, 3H, CO$_2$CH$_3$), 2.89 (dd, $J = 13.8, 8.8$ Hz, 1H, CH$_2$), 1.98 (s, 3H, N$_a$HCOC$_3$H$_3$), 1.80 (s, 3H, N$_b$COC$_3$H$_3$).

HRMS: Calculated for [C$_{20}$H$_{22}$N$_2$O$_6$ + H]$^+$ = 387.15506, observed = 387.15514

NAc-Trp(APAP)-OMe

**Benzofuroindoline**

$^1$H NMR (500 MHz, DMSO-$d_6$) $\delta$ 9.76 (s, 1H, $N_b$H), 8.21 (d, $J = 8.0$ Hz, 1H, $N_a$H), 7.63 (d, $J = 2.2$ Hz, 1H, 12-H), 7.28 (d, $J = 2.8$ Hz, 1H, indole-NH$_2$), 7.18 (obsc. d, $J = 7.3$ Hz, 1H, 6-H), 7.17 (obsc. dd, $J = 8.5, 2.2$ Hz, 1H, 13-H), 6.96 (td, $J = 7.6, 1.2$ Hz, 1H, 5-H), 6.65 (d, $J = 8.5$ Hz, 1H, 10-H), 6.62 (td, $J = 7.5, 1.0$ Hz, 1H, 4-H), 6.52 (d, $J = 7.3$ Hz, 1H, 3-H), 6.36 (d, $J = 2.8$ Hz, 1H, 7-H), 4.13 (td, $J = 8.5, 4.2$ Hz, 1H, CH), 3.53 (s, 3H, CO$_2$CH$_3$), 2.58 – 2.53 (m, 1H, CH$_2$), 2.28 (dd, $J = 14.7, 8.8$ Hz, 1H, CH$_2$), 1.99 (s, 3H, N$_a$HCOCH$_3$), 1.69 (s, 3H, N$_b$HCOCH$_3$).

HRMS: Calculated for [C$_{22}$H$_{23}$N$_2$O$_5$ + H]$^+$ = 410.17105, observed = 410.17112

**NAc-Trp(APAP)-OMe**

**Pyrroloindoline***
$^1$H NMR (400 MHz, DMSO-$d_6$) δ 9.56 (s, 1H, $N_b$H), 7.43 (dd, $J = 5.8, 2.5$ Hz, 0.4H, Ar-$H$), 7.41 (dd, $J = 6.1, 2.5$ Hz, 0.6H, Ar-$H$), 7.21 – 7.17 (m, 0.6H, Ar-$H$), 7.12 (d, $J = 7.5$ Hz, 0.4H, Ar-$H$), 7.03 (t, $J = 7.6$ Hz, 0.4H, Ar-$H$), 7.01 – 6.93 (m, 1.6H), 6.78 (d, $J = 7.0$ Hz, 0.6H, Ar-$H$), 6.75 (d, $J = 7.0$ Hz, 0.4H, Ar-$H$), 6.70 – 6.60 (m, 1.4H, Ar-$H$), 6.58 (d, $J = 6.1$ Hz, 0.6H, Ar-$H$), 6.56 (d, $J = 6.7$ Hz, 0.4H, Ar-$H$), 5.82 (s, 0.4H, 13-$H$), 5.78 (s, 0.6H, 13-$H$), 4.90 (d, $J = 8.7$ Hz, 0.6H, $CHCO_2CH_3$), 4.79 (dd, $J = 9.3, 1.6$ Hz, 0.4H, $CHCO_2CH_3$), 3.21 – 3.06 (overlapping m, 4H, $CO_2CH_3/CH_2$), 2.96 (d, $J = 12.7$ Hz, 0.6H, $CH_2$), 2.60 (dd, $J = 12.8, 1.7$ Hz, 0.4H, $CH_2$), 2.16 (s, 1.4H, NCOCH$_3$), 1.93 (s, 1.6H, NCOCH$_3$), 1.89 (s, 3H, N$_b$HCOCH$_3$).

HRMS: Calculated for [C$_{22}$H$_{23}$N$_2$O$_5$ + H]$^+$ = 410.17105, observed = 410.16972

* Present as mixture of rotamers in a ratio of approximately 4:6 (see Figure 3-5)

9.4.2 $^1$H, $^{13}$C and 2D NMR Spectra for isolated compounds

Figure 9-2. $^1$H NMR spectrum for the isolated adduct NAc-Cys(APAP)-OMe
Figure 9-3. $^{13}$C NMR spectrum for the isolated adduct NAc-Cys(APAP)-OMe

Figure 9-4. $^1$H NMR spectrum for the isolated adduct NAc-Met(APAP)-OMe isomer A
Figure 9-5. $^1$H NMR spectrum for the isolated adduct NAc-Met(APAP)-OMe isomer B

Figure 9-6. $^1$H NMR spectrum for the isolated adduct NAc-Tyr(APAP)-OMe
Figure 9-7. $^1$H-$^1$H gCOSY NMR spectrum for the isolated adduct NAc-Tyr(APAP)-OMe

Figure 9-8. $^1$H NMR spectrum for the isolated benzofuroindoline NAc-Trp(APAP)-OMe adduct
**Figure 9-9.** $^1$H-$^1$H gCOSY NMR spectrum for the isolated benzofuroindoline NAc-Trp(APAP)-OMe adduct

**Figure 9-10.** $^1$H NMR spectrum for the isolated pyrroloindoline NAc-Trp(APAP)-OMe adduct
9.5 Chapter 4 supplementary data

Table 9-5. Parameters used for non-targeted reactive metabolite detection for Trazodone.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Peptide Selection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threshold</td>
<td>Identity</td>
<td></td>
</tr>
<tr>
<td>Charge state range</td>
<td>2-4</td>
<td></td>
</tr>
<tr>
<td><strong>HiTIME Selection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS² m/z offset</td>
<td>0.5</td>
<td>m/z</td>
</tr>
<tr>
<td>MS² RT offset</td>
<td>20</td>
<td>seconds</td>
</tr>
<tr>
<td><strong>MS² correlation parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Match m/z Tolerance</td>
<td>0.2</td>
<td>m/z</td>
</tr>
<tr>
<td>Match Ion Types</td>
<td>y</td>
<td>N/A</td>
</tr>
<tr>
<td>Reactive Residues</td>
<td>C, W, Y, M, K</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>CRM parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m/z band</td>
<td>40-650</td>
<td>Da</td>
</tr>
<tr>
<td>Max. RME</td>
<td>100</td>
<td>Da</td>
</tr>
<tr>
<td>Ppm tolerance</td>
<td>10</td>
<td>Ppm</td>
</tr>
</tbody>
</table>

Table 9-6. Atom ranges used in non-targeted CRM detection for Trazodone.

<table>
<thead>
<tr>
<th>Element</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>H</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>Cl</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Image 9-1. Screen capture of the Postprocessing setup dialog from the Xenophile software.
Image 9-2. Screen capture of non-targeted CRM search setup dialog from the Xenophile software.
Image 9-3. Screen capture of non-targeted CRM search setup dialog from the Xenophile software.
Image 9-4. Screen capture of non-targeted CRM search results analysis dialog from the Xenophile software.
Image 9-5. Screen capture of targeted correlation search setup dialog from the Xenophile software.
**Image 9-6**, Screen capture of targeted correlation search results analysis dialog from the Xenophile software.
### 9.6 Chapter 5 supplementary data

**Table 9-7.** Parameters used for postprocessing of HiTIME scoring results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum HiTIME Score</td>
<td>15</td>
<td>Dimensionless</td>
</tr>
<tr>
<td>m/z Width</td>
<td>0.15</td>
<td>m/z</td>
</tr>
<tr>
<td>RT Width</td>
<td>1</td>
<td>Minutes</td>
</tr>
<tr>
<td>RT Exclusion</td>
<td>0</td>
<td>Minutes</td>
</tr>
<tr>
<td>mzDelta</td>
<td>3.01005</td>
<td>Daltons</td>
</tr>
<tr>
<td>EIC Width</td>
<td>0.03</td>
<td>m/z</td>
</tr>
</tbody>
</table>

**Table 9-8.** Parameters used for non-targeted reactive metabolite detection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Peptide Selection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threshold</td>
<td>Identity</td>
<td></td>
</tr>
<tr>
<td>Charge state range</td>
<td>2-4</td>
<td></td>
</tr>
<tr>
<td><strong>HiTIME Selection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS² m/z offset</td>
<td>0.5</td>
<td>m/z</td>
</tr>
<tr>
<td>MS² RT offset</td>
<td>2</td>
<td>minutes</td>
</tr>
<tr>
<td><strong>MS² correlation parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Match m/z Tolerance</td>
<td>0.5</td>
<td>m/z</td>
</tr>
<tr>
<td>Match Ion Types</td>
<td>b,y</td>
<td>N/A</td>
</tr>
<tr>
<td>Reactive Residues</td>
<td>C, W, Y, M, K</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>CRM parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m/z band</td>
<td>76-360</td>
<td>Da</td>
</tr>
<tr>
<td>Max. RME</td>
<td>100</td>
<td>Da</td>
</tr>
<tr>
<td>Ppm tolerance</td>
<td>20</td>
<td>Ppm</td>
</tr>
</tbody>
</table>
### Table 9-9. Atom ranges used in non-targeted CRM detection

<table>
<thead>
<tr>
<th>Element</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>30</td>
</tr>
</tbody>
</table>

### Table 9-10. Top three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 2. of 3)

<table>
<thead>
<tr>
<th>Hit</th>
<th>CRM mass (Da)</th>
<th>Sequence</th>
<th>Modification Site</th>
<th>Formula</th>
<th>Ppm</th>
<th>Residual mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>149.0471</td>
<td>VFANPEDCAGFGK</td>
<td>C (8)</td>
<td>C&lt;sub&gt;8&lt;/sub&gt; H&lt;sub&gt;7&lt;/sub&gt; N O&lt;sub&gt;2&lt;/sub&gt;</td>
<td>-4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C&lt;sub&gt;6&lt;/sub&gt; H&lt;sub&gt;5&lt;/sub&gt; N O</td>
<td>5</td>
<td>86</td>
</tr>
<tr>
<td>2</td>
<td>149.0475</td>
<td>EFTPCAQAQAFK</td>
<td>C (5)</td>
<td>C&lt;sub&gt;8&lt;/sub&gt; H&lt;sub&gt;7&lt;/sub&gt; N O&lt;sub&gt;2&lt;/sub&gt;</td>
<td>-2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C&lt;sub&gt;6&lt;/sub&gt; H&lt;sub&gt;5&lt;/sub&gt; N O</td>
<td>7</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>149.0468</td>
<td>TIQLNVCNSEEVEK</td>
<td>C (7)</td>
<td>C&lt;sub&gt;6&lt;/sub&gt; H&lt;sub&gt;5&lt;/sub&gt; N O</td>
<td>3</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C&lt;sub&gt;8&lt;/sub&gt; H&lt;sub&gt;7&lt;/sub&gt; N O&lt;sub&gt;2&lt;/sub&gt;</td>
<td>-6</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 9-11. Top three highest ranked reactive metabolite assignments produced for one of the APAP replicates using the non-targeted reactive metabolite identification algorithm. (APAP replicate 3 of 3)

<table>
<thead>
<tr>
<th>Hit</th>
<th>CRM mass (Da)</th>
<th>Sequence</th>
<th>Modification Site</th>
<th>Formula</th>
<th>Ppm</th>
<th>Residual mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>149.0506</td>
<td>VFANPEDCAGFGK</td>
<td>C (8)</td>
<td>C&lt;sub&gt;8&lt;/sub&gt; H&lt;sub&gt;7&lt;/sub&gt; N O&lt;sub&gt;2&lt;/sub&gt;</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>149.0493</td>
<td>TIQLNVCNSEEVEK</td>
<td>C (7)</td>
<td>C&lt;sub&gt;8&lt;/sub&gt; H&lt;sub&gt;7&lt;/sub&gt; N O&lt;sub&gt;2&lt;/sub&gt;</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>151.0261</td>
<td>KDAQTLYDAGEK</td>
<td>G(10)</td>
<td>C&lt;sub&gt;7&lt;/sub&gt; H&lt;sub&gt;3&lt;/sub&gt; N O&lt;sub&gt;3&lt;/sub&gt;</td>
<td>-6</td>
<td>32</td>
</tr>
</tbody>
</table>
### Table 9-12. Parameters for targeted detection of NAPQI-modified peptides

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide Selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threshold</td>
<td>Identity</td>
<td></td>
</tr>
<tr>
<td>Charge state range</td>
<td>2-4</td>
<td></td>
</tr>
<tr>
<td>HiTIME Selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS2 m/z offset</td>
<td>0.5</td>
<td>m/z</td>
</tr>
<tr>
<td>MS2 RT offset</td>
<td>2</td>
<td>minutes</td>
</tr>
<tr>
<td>EIC tolerance</td>
<td>0.03</td>
<td>m/z</td>
</tr>
<tr>
<td>Neutral isotope mass difference</td>
<td>6.0201</td>
<td>Da</td>
</tr>
</tbody>
</table>

### Table 9-13. Summary of the peptide signals created for each data set and the discovery rate of these peptides for HiTIME and NTPS analysis.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Simulated Peptides</th>
<th>Post HT-Processing</th>
<th>NTPS &lt; 20 ppm</th>
<th>NTPS &lt; 100 ppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control 1</td>
<td>620</td>
<td>590 (95 %)</td>
<td>526 (89 %)</td>
<td>579 (98 %)</td>
</tr>
<tr>
<td>Control 2</td>
<td>560</td>
<td>539 (96 %)</td>
<td>467 (87 %)</td>
<td>530 (98 %)</td>
</tr>
<tr>
<td>Control 3</td>
<td>610</td>
<td>582 (95 %)</td>
<td>523 (90 %)</td>
<td>574 (99 %)</td>
</tr>
</tbody>
</table>
Figure 9-11. Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a doublet spacing of 3.01005 aimed at mining doubly charged peptides modified by an APAP metabolite. A-C) APAP treatment, D-F) vehicle control. Data points with a weighted score less than 15 have been omitted for clarity. For convenience, the data shown here in panels A and D are reproduced from Figure 5-2A and Figure 5-2B respectively in the main text.
Figure 9-12. Heat maps produced by HiTIME scoring of microsomal protein digest LC-MS data sets with a doublet spacing of 2.0067 aimed at mining triply charged peptides modified by an APAP metabolite. A-C) APAP treatment, D-F) vehicle control. Data points with a weighted score less than 15 have been omitted for clarity.
Figure 9-13. Histogram of the distribution of scores from HiTIME analysis of APAP treatment and vehicle control (VC) micromal protein digest LC-MS data using twin-ion spacing settings of A) 3.01005, and B) 2.0067.
10 REFERENCES


(94) Langman, R. E.; Cohn, M., Semin. Immunol. 2000, 12, 159-162.


(121) Smith, P. C.; Benet, L. Z.; McDonagh, A. F., Drug Metab. Disposition 1990, 18, 639-644.


(414) RDKit: Open-source cheminformatics http://www.rdkit.org


