Reproducible computational tools for acquiring, analysing and managing model organism metabolomes

by
Thomas Nigel Lawson

A thesis submitted to
The School of Bioscience
The University of Birmingham
September 2018

For the degree
Doctor of Philosophy

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Mechanistic insights into the effects of pollutants on the environment and to human health can be discovered by analysing the metabolic differences between control and stressed model organisms. However, the identification or annotation of the observed metabolites can be challenging and is further complicated by the limited availability of tools that allow metabolite annotations to be reproducibly derived and managed.

The focus of this work was to create computational tools and workflows to allow for reproducible approaches for acquiring metabolomics data, deriving annotations, and managing and disseminating the results. The primary use case for these tools and workflows is the mass spectrometry (MS) component of the Deep Metabolome Annotation (DMA) of the ecotoxicological model *Daphnia magna* (the water flea).

The characteristic gas-phase fragmentation patterns of MS are commonly used within metabolomics for the purposes of metabolite annotation. A method was developed here to assess the quality of MS gas-phase fragmentation (and by extension the quality of the annotation) by determining the contribution of a selected ion's purity within a targeted isolation window. Tools were also developed for metabolite annotation including for spectral matching of fragmentation spectra to spectral libraries. Workflows were developed to aid in the acquisition of fragmentation spectra including approaches for optimising the scheduling of MS fragmentation. A tool for metadata extraction for the open source MS data file format was developed to aid in the creation of ISA-tab files for upload into public metabolomics data repositories. All the tools and workflows were integrated into the Galaxy workflow environment, where possible, and a metabolite annotation workflow was developed that incorporated both novel and existing tools. The Django MOGI application suite (Metabolomics Organisation with Galaxy and ISA) was

then designed and developed for managing metabolomic analysis and used to create the DMA Database (DMAdb). The above tools and workflow were then used to analyse and manage data generated from the liquid chromatography tandem mass spectrometry experimental assays from the DMA of *D. magna* - substantially increasing the metabolome knowledge of this ecotoxicology relevant model organism and providing a resource a valuable resource for ecotoxicological research.

The complexity in both the experimental and computational steps in obtaining a metabolite annotation necessitate careful consideration for the management of data and meta-data. By integrating the Galaxy workflow environment and the ISA framework with the tools developed for this PhD, much needed tools for elucidating model organism metabolomes in a FAIR (findable, accessible, interoperable, re-usable) and reproducible manner have been made available to the community.

You can't possibly be a scientist if you mind people thinking that you're a fool.

— Douglas Noel Adams, So Long And Thanks For All The Fish.

ACKNOWLEDGMENTS

There are many people I would like to thank that have helped throughout this journey. First off, I would like to thank my supervisor, Professor Mark Viant. I am indebted to you, not just for the opportunity, but also the enthusiasm and help you gave throughout my PhD.

Enormous thanks are also deserved for Dr Ralf Weber and Dr Martin Jones. Ralf for providing invaluable knowledge and direction in bioinformatics and always remaining positive. Martin for both his colossus mass spectrometry and *Daphnia* experiments and for his chemistry and analytical knowledge. All of which have been imperative to the work I have carried out for this PhD. Thank you, it has been a pleasure working with you both the past four years.

Thank you to my co-supervisors Professor Warwick Dunn for all his advice but specifically for his help on my research on mass spectrometry data processing. Thanks to my co-supervisor Professor Johnathon Colbourne for his advice on the genomics of *Daphnia*. I would also like to thank my CASE co-supervisor Dr Peter Li for his support throughout the PhD and for his involvement in organising a fantastic workshop and hackathon for CUDDLE. Thanks for being so welcoming at GigaScience and providing an illuminating experience of both Hong Kong and Shenzhen.

I also offer my thanks to Dr Reza Salek for his supervision during my EMBL-EBI placement, it was an amazing opportunity and I learned a great deal in a brief period of time. I extend these thanks everybody else at EMBL-EBI MetaboLights for their help and support.

I would also like to thank Martin Larralde for his work on XML metadata parsing. Also, thank you to anybody else who has contributed to the collaborative software development performed as part of this thesis.

Thank you to James and everyone else in the computational metabolomics office for providing valuable advice during the PhD. I would also like to thank everyone who worked and studied on the 4th floor (and 3rd floor) with me over the past 4 years for their advice, support and the frequent trips to staff house. Thanks also to everyone down in the basement with the *Daphnia,* my brief time culturing the critters opened my eyes to the dedication and skill involved in what you all do.

Thank you to NERC-CENTA, the University of Birmingham and my CASE funder GigaScience for their financial contributions that kindly funded this research.

Thank you to all my friends and family. Especially my Mum, Dad and brother - I am incredibly grateful for your support and encouragement throughout everything in life.

Finally, I would like to thank my wife, Laura. Thank you for keeping me sane over these years. Your love, support and patience made this PhD thesis possible.

List of Publications

Academic Journals:

1. **Lawson, T.N.**, Weber, R.J.M., Jones, M.R., Chetwynd, A.J., Rodriguez Blanco, G.A., Di Guida, R., Viant, M.R., Dunn, W.B., 2017. **msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry Based Fragmentation in Metabolomics**. Anal. Chem. acs.analchem.6b04358. https://doi.org/10.1021/acs.analchem.6b04358
2. Larralde, M., **Lawson, T.N**†., Weber, R.J.M., Moreno, P., Haug, K., Rocca-serra, P., Viant, M.R., Salek, R.M., 2017. mzML2ISA & nmrML2ISA : **Generating enriched ISA-Tab metadata files from metabolomics XML data.** Bioinformatics 4–6. https://doi.org/10.1093/bioinformatics/btx169

   † Joint first author

Invited talks:

1. **Lawson, T.N**. Database and web-framework for annotation and curation of model organism metabolomes. The Metabolic Profiling Forum: MetaboMeeting, September 2017, University of Birmingham.

DECLARATIONS

Chapter 2 of this thesis was previously published as "**Lawson, T.N.**, Weber, R.J.M., Jones, M.R., Chetwynd, A.J., Rodriguez Blanco, G.A., Di Guida, R., Viant, M.R., Dunn, W.B., 2017. **msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry Based Fragmentation in Metabolomics**. Anal. Chem. acs.analchem.6b04358. https://doi.org/10.1021/acs.analchem.6b04358"

The corresponding author "Warwick B. Dunn" agrees with the following statement of contributions:

TNL developed the metric and associated software, was lead in writing the manuscript, performed all data analysis, planned the experimental design for all isolation window experiments, and experimentally acquired data for the Thermo Fisher Q-Exactive isolation window experiments. RJMW advised on the software development. MRJ extracted metabolites and experimentally acquired data on the mass spectrometer for the *Daphnia magna* samples. AJC experimentally acquired data for the Waters Xevo G2 XS QTOF mass spectrometry isolation experiments. GRB advised and aided in the Thermo Fisher Q-Exactive isolation windows mass spectrometry experiments. RDG extracted metabolites and experimentally acquired data on the mass spectrometer for the *Ovis aries* samples. MRV and WBJ supervised and advised in all aspects of the work. All authors read and approved the final published manuscript.

Name: Prof. Warwick B Dunn

Signature: [redacted] Date: ……22/10/18……..

Chapter 4 of this thesis was previously published as "Larralde, M†., **Lawson, T.N†**., Weber, R.J.M., Moreno, P., Haug, K., Rocca-serra, P., Viant, M.R., Salek, R.M., 2017. mzML2ISA & nmrML2ISA : **Generating enriched ISA-Tab metadata files from metabolomics XML data.** Bioinformatics 4–6. https://doi.org/10.1093/bioinformatics/btx169"

† Joint first author

Both the joint first author "Martin Larralde" and corresponding author "Reza Salek" agree that the same same text from the published work "Generating enriched ISA-Tab metadata files from metabolomics XML data" can not be used in another thesis. They also both agree with the following statement of contributions:

TNL developed the mzML2ISA tool during a NERC funded placement at EMBL-EBI Cambridge in July 2015. ML further developed the mzML2ISA tool as well as developing nmrML2ISA and all associated GUIs during an internship at EMBL-EBI in July 2016. TNL was also lead in writing the manuscript and co-developed the Galaxy wrappers with RJMW. PM aided in the Galaxy wrapper development and prepared the Docker containers. KH provided advice and practical input throughout the development of the tools. PRS provided expertise regarding the ISA-Tab format. MRV, CS and RMS supervised all work carried out. The code is currently maintained at github.com/ISA-tab where maintenance is shared between TNL, ML, RJMW, KH, RMS and PRS (all contributions can be seen within the git history). All authors read and approved the final published manuscript.

Name: Mr. Martin Larralde

Signature: [redacted] Date: ……22/10/18……..

Name: Dr. Reza M. Salek

Signature: [redacted] Date: ……22/10/18……..

Chapter 3, 5 and 6 incorporated collaborative work with other researches, contributions are detailed in relevant chapters. The primary supervisor of this PhD "Mark R Viant", who supervised all work carried out for these chapters, confirms here that Thomas N. Lawson was the sole writer of the chapters 3, 5 and 6 and main contributor of the work performed.

Mark R. Viant also confirms here that Thomas N. Lawson was the sole writer and only contributor to Chapters 1 and 7.

Name: Prof. Mark R Viant

Signature: ... .. Date: ……23/10/18……..

CONTENTS

CHAPTER 1 INTRODUCTION 1

1.1 RATIONALE 1

1.2 METABOLOMICS 2

1.3 MASS SPECTROMETRY METABOLOMICS 4

1.4 METABOLITE ANNOTATION 9

1.5 REPRODUCIBLE AND FAIR METABOLOMICS 21

1.6 WORKFLOW ENVIRONMENTS 24

1.7 METABOLOMIC DATA MANAGEMENT 25

1.8 *DAPHNIA MAGNA* DEEP METABOLOME ANNOTATION (DMA) 29

1.9 AIM AND THESIS OVERVIEW 31

CHAPTER 2 MSPURITY: AUTOMATED EVALUATION OF PRECURSOR ION PURITY FOR MASS SPECTROMETRY BASED FRAGMENTATION IN METABOLOMICS 35

2.1 CHAPTER SUMMARY 35

2.2 CONTRIBUTIONS 37

2.3 PUBLICATION HISTORY 37

2.4 INTRODUCTION 38

2.5 MATERIALS AND METHODS 41

2.6 RESULTS AND DISCUSSION 53

2.7 CONCLUSION 62

CHAPTER 3 LOOK BEFORE YOU LEAP: USING PRIOR INFORMATION TO AUTOMATICALLY GUIDE MASS SPECTROMETRY FRAGMENTATION IN METABOLOMICS 63

3.1 CHAPTER SUMMARY 63

3.2 CONTRIBUTIONS 64

3.3 PUBLICATION HISTORY 64

3.4 INTRODUCTION 65

3.5 METHODS 69

3.6 RESULTS 91

3.7 DISCUSSION 99

CHAPTER 4 MZML2ISA & NMRML2ISA: GENERATING ENRICHED ISA-TAB METADATA FILES FROM METABOLOMICS XML DATA 102

4.1 CHAPTER SUMMARY 102

4.2 CONTRIBUTIONS 103

4.3 PUBLICATION HISTORY 103

4.4 INTRODUCTION 104

4.5 METHODS 105

4.6 RESULTS 107

4.7 DISCUSSION 108

4.8 ACKNOWLEDGEMENTS 109

4.9 FUNDING 109

CHAPTER 5 A SUITE OF DJANGO APPLICATIONS FOR METABOLOME ORGANIZATION WITH GALAXY AND ISA 110

5.1 CHAPTER SUMMARY 110

5.2 CONTRIBUTIONS 111

5.3 PUBLICATION HISTORY 111

5.4 FINDINGS 112

5.5 METHODS 143

5.6 AVAILABILITY OF SUPPORTING SOURCE CODE AND REQUIREMENTS 144

5.7 DISCUSSION 144

CHAPTER 6 DATA ANALYSIS AND DATA MANAGEMENT FOR THE DEEP METABOLOME ANNOTATION OF *DAPHNIA MAGNA* 146

6.1 CHAPTER SUMMARY 146

6.2 CONTRIBUTIONS 146

6.3 PUBLICATION HISTORY 147

6.4 INTRODUCTION 147

6.5 METHODS 149

6.6 RESULTS AND DISCUSSION 160

6.7 CONCLUSION 166

CHAPTER 7 CONCLUSION 168

7.1 OUTCOMES 168

7.2 FUTURE WORK 171

7.3 Final remarks 176

CHAPTER 8 APPENDIX 178

8.1 Isolation efficiency of Waters Xevo G2 XS QTOF 179

8.2 In house experimental details 180

8.3 Further information for public datasets (msPurity) 184

8.4 Violin plots for precursor purity of public datasets (mspurity) 185

8.5 Overview of assessment of precursor purity for public DDA and DIA metabolomics datasets - zero values included (msPurity) 186

8.6 Summary of positive and negative ionisation mode comparison for public DDA datasets (msPurity) 187

8.7 Comparison of centred vs most intense approach to determining precursor peak 188

8.8 Precursor purity distribution of public datasets (mspurity) 189

8.9 Original Nearline scheduling algorithms 190

8.10 adapted-LCMS-nearline Galaxy data analysis workflow 191

8.11 DIMS-nearline Galaxy data analysis workflow (for paired sample-blank analysis) 194

8.12 DIMS-nearline assessment for deep Metabolome Annotation Galaxy workflow parameters 197

8.13 XML2ISA Software Summary 211

8.14 XML2ISA Implementation (further Details) 212

8.15 XML2ISA MetaboLights Study Details 213

8.16 XML2ISA Benchmarking results 214

8.17 django-mbrowse model schema 215

8.18 LC-MS(/MS) DMA workflow settings 227

8.19 Chromatography methods for DMA 252

8.20 Chemical descriptors 254

8.21 Summary of annotations for LC-MS(/MS) DMA assays 256

8.22 PCA analysis of KEGG brite compounds 260

CHAPTER 9 REFERENCES 261

List of Figures

Figure 1-1 Gas-phase mass spectrometry fragmentation acquisition methods used for metabolomics 9

Figure 1-2 CSI:FingerID annotation approach 20

Figure 1-3 Metabolomic data management 26

Figure 1-4 Django request-response cycle 28

Figure 1-5 General overview of Deep Metabolome Annotation 30

Figure 2-1 Assessment of acquired MS^2 precursor purity for a standard DDA based experiment. 43

Figure 2-2 Assessing precursor purity of anticipated isolation windows for two XCMS features derived from an MS^2 dataset. 47

Figure 2-3 Overview of isolation efficiency experimental determination for Thermo Scientific Q Exactive Focus 49

Figure 2-4: Overview of assessment of precursor purity for public DDA and DIA metabolomics datasets 56

Figure 2-5: Overview of the precursor purity of XCMS features calculated for theoretical isolation windows of an anticipated MS^2 experiment. 61

Figure 3-1 adapted-LCMS-nearline experimental and data analysis workflow 71

Figure 3-2 Inclusion list and scheduling schematics for LC-MS-nearline 75

Figure 3-3 DIMS-nearline experimental and data analysis workflow 77

Figure 3-4 Fractionation optimisation 84

FIGURE 3-5 COMPARISON OF EFFECTIVENESS TO FRAGMENT XCMS FEATURES USING PROTYPE ADAPTED-LCMS-NEARLINE AND STANDARD DDA FRAGMENTATION 92

FIGURE 3-6 OPTIMISED FRACTIONATION COMPARED TO STANDARD 93

FIGURE 3-7 ASSESSMENT OF DIMS-NEARLINE ON A SUBSET OF THE DMA OF *D. MAGNA* - ASSAY AP_AMP1_C30_DI-MS-DI-MSN 97

FIGURE 4-1 SCHEMATIC DIAGRAM AND WORKFLOW OF MZML2ISA & NMRML2ISA SOFTWARE SUITE 107

FIGURE 5-1 DJANGO MOGI APPLICATION SUITE 116

FIGURE 5-2 OVERVIEW OF DJANGO-MISA COMPONENTS 118

FIGURE 5-3 MODEL SCHEMA FOR ISA ASSAY PROTOCOL, PROCESS AND RELATED DATA FILES (DJANGO-MISA) 122

FIGURE 5-4 MODEL SCHEMA FOR ISA STUDY-DESIGN, STUDY-SAMPLE, STUDY-FACTOR AND SAMPLE COLLECTION PROCESS AND PROTOCOL (DJANGO-MISA) 123

FIGURE 5-5 MODEL SCHEMA FOR DJANGO-GALAXY 125

FIGURE 5-6 LC-MS/MS GALAXY ANNOTATION WORKFLOW 128

FIGURE 5-7 EXAMPLE INCHIKEY STRUCTURE (MODIFIED FROM ALAN, 2009) 131

FIGURE 5-8 BROWSING AND SEARCHING METABOLOMICS DATA WITH DJANGO-METAB 134

FIGURE 5-9 MODEL SCHEMA FOR CHROMATOGRAPHIC FEATURES (DJANGO-MBROWSE) 139

FIGURE 5-10 RECOMMENDED INFRASTRUCTURE 141

FIGURE 6-1 DMA OF *D. MAGNA* EXPERIMENTAL WORKFLOW 153

FIGURE 6-2 THE DEEP METABOLOME ANNOTATION DATABASE (DMADB) OVERVIEW 159

FIGURE 6-3 UPSET PLOT SUMMARISING THE DMA OF *D. MAGNA* L2-A METABOLITE ANNOTATIONS

ACROSS 23 EXPERIMENTAL ASSAYS 163

FIGURE 6-4 PCA SCORES PLOT OF DMA EXPERIMENTAL (EXP) COMPOUND L2-A ANNOTATIONS AND KEGG COMPOUNDS 165

FIGURE 8-1: EXPERIMENTAL DETERMINATION OF ISOLATION EFFICIENCY FOR WATERS XEVO G2 XS QTOF 179

FIGURE 8-2: ASSESSMENT OF PRECURSOR PURITY FOR PUBLIC DDA AND DIA METABOLOMICS DATASETS (VIOLIN PLOTS). 185

FIGURE 8-3: OVERVIEW OF ASSESSMENT OF PRECURSOR PURITY FOR PUBLIC DDA AND DIA METABOLOMICS DATASETS (ZERO VALUES INCLUDED) 186

FIGURE 8-4: PRECURSOR PURITY (USING MOST INTENSE PEAK) DISTRIBUTION OF ALL DATA FROM PUBLIC SOURCES. 189

FIGURE 8-5 ADAPTED-LCMS-NEARLINE GALAXY DATA ANALYSIS WORKFLOW 191

FIGURE 8-6 DIMS-NEARLINE GALAXY DATA ANALYSIS WORKFLOW (FOR PAIRED SAMPLE-BLANK ANALYSIS) 194

FIGURE 8-7 METABOLOMIC DATA FILE, GENERIC FILE AND DJANGO USER MODEL SCHEMA (DJANGO-MBROWSE, DJANGO-GFILE) 215

FIGURE 8-8 ADDUCT AND ISOTOPE MODEL SCHEMA (DJANGO-MBROWSE) 216

FIGURE 8-9 FRAGMENTATION FOR CHROMATOGRAPHIC FEATURE MODEL SCHEMA (DJANGO-MBROWSE) 218

FIGURE 8-10 LIBRARY-COMPOUND-ANNOTATION MODEL SCHEMA (DJANGO-MBROWSE) 220

FIGURE 8-11 SPECTRAL MATCHING OF FRAGMENTATION SPECTRA MODEL SCHEMA (DJANGO-MBROWSE) 222

FIGURE 8-12 METFRAG AND SIRIUS CSI:FINGERID ANNOTATION MODEL SCHEMA (DJANGO-MBROWSE) 224

FIGURE 8-13 MS^1 ANNOTATION WITH PROBMETAB MODEL SCHEMA (DJANGO-MBROWSE) 226

FIGURE 8-14 PCA SCORES PLOT OF KEGG BRITE COMPOUNDS 260

List of Tables

Table 1-1 Rumsfeld's Quadrant of metabolite annotation 9

Table 2-1: Summary of public metabolomics datasets used to assess precursor purity 55

Table 3-1 Overview of LC-MS-nearline approaches 74

Table 3-2: Galaxy tools used for nearline workflows 81

Table 3-3: Assessment of adapted-LCMS-nearline on a subset of the DMA of *D. magna* - assay AP_AMP_C30_LC-MS_LCMS-MS 95

Table 5-1 Overview of new Galaxy tools 127

Table 5-2: Example metabolite annotation summary 132

Table 5-3 Example alignment of mock library and query spectra 137

Table 5-4 Weighting library and query spectra 138

Table 5-5 Overview of existing Galaxy tools 143

Table 6-1 Summary of experimental LC-MS(/MS) DMA assays 155

Table 6-2 Summary of DMA of *D. magna* unique annotations 160

Table 8-1: Further information for public datasets 184

Table 8-2 Summary of positive and negative ionisation mode comparison for public DDA datasets 187

Table 8-3: Comparison of centred vs most intense approach to determining precursor peak 188

Table 8-4 XML2ISA Software suite 211

TABLE 8-5 METABOLIGHTS STUDIES WITH XML OPEN SOURCE RAW FILE FORMATS 213

TABLE 8-6 BENCHMARKING OF SAMPLE XML DATA FILES 214

TABLE 8-7 SUMMARY OF CHEMICAL DESCRIPTORS USED FOR PCA ANALYSIS 254

TABLE 8-8 SUMMARY OF FRAGMENTATION COVERAGE FOR LC-MS(/MS) DMA ASSAYS (POSITIVE IONISATION MODE) 256

TABLE 8-9 SUMMARY OF FRAGMENTATION COVERAGE FOR LC-MS(/MS) DMA ASSAYS (NEGATIVE IONISATION MODE) 257

TABLE 8-10 SUMMARY OF ANNOTATIONS FOR LC-MS(/MS) DMA ASSAYS (POSITIVE IONISATION MODE) 258

TABLE 8-11 SUMMARY OF ANNOTATIONS FOR LC-MS(/MS) DMA ASSAYS (NEGATIVE IONISATION MODE) 259

Chapter 1
Introduction

1.1 RATIONALE

The measured metabolic profile of an organism's metabolome (complement of all small molecules) can provide mechanistic insights into how exogenous chemicals from the environment can interact with complex biochemical processes of an organism. Thus, when exposed to environmental pollutants, the measured metabolome of ecotoxicological models, such as *Daphnia magna,* can be used to assess the toxicity of chemicals on public health and the ecology of contaminated areas.

The most widely used technique for profiling a metabolome is mass spectrometry, where 1,000's of metabolite differences can be simultaneously observed. However, the identification or annotation of the observed metabolites can be challenging, limiting the effectiveness of using metabolomics as a routine procedure in human and environmental toxicology. The challenge is further complicated by the limited availability of tools that allow metabolite annotations to be reproducibly derived and managed.

One approach to improve the knowledge of the identities of metabolites for model organisms is to attempt to characterise the metabolome of a representative sample using multiple complementary physio-chemical separation and measurement techniques. Such a 'Deep

Metabolome Annotation' (DMA) could then aid in annotating and interpreting routine metabolomics experiments where time and resource constraints would make the same level of characterisation infeasible. However, these large-scale annotation projects still face similar considerable challenges due to the lack of resources to reproducibly manage and annotate untargeted mass spectrometry metabolomics analysis and results.

Galaxy (Afgan et al., 2018) is a workflow environment that has been successfully used for reproducible, accessible and transparent data analysis in metabolomics (Davidson et al., 2016; Giacomoni et al., 2015). However, many methods for metabolite annotation do not have implementation in Galaxy. Additionally, software for managing metabolomics data (Gloaguen et al., 2017; Hunter et al., 2017) does not have support for Galaxy and has limited support for the ISA framework. The ISA (The Investigation, Study, Assay) framework and ISA-tab file format is a widely used format that aids reproducibility by detailing experimental projects in a controlled and defined structure (Rocca-Serra et al., 2010).

This PhD focused on developing reproducible tools for acquiring, analysing and managing mass spectrometry derived metabolomics data and results. Where possible all tools and workflows were implemented or incorporated with either Galaxy, ISA or both. These tools and workflows were then applied to the DMA of *D. magna* for reproducible and transparent metabolite annotation.

1.2 METABOLOMICS

The metabolome was originally defined in 1998 as the full complement of small molecules (exogenous/endogenous) present in a biological sample (e.g. a cell, tissue or whole organism) at any point in time (Oliver et al., 1998). Similar to how transcriptomics provides a snapshot of expression of all genes under a certain set of conditions, metabolomics provide a snapshot of the small molecules present. There has been intensive study of metabolomics (> 3,500 papers

published with metabolomics in the title 2008-2017, PubMed), where it has been successfully implemented as an approach to identify potential biomarkers for disease (Lin et al., 2010), used to assess environmental stress and genetic permutations (Bundy et al., 2008) and used to provided novel insights into networks and systems when combined with genomics or transcriptomics (Cho et al., 2014) . However, there can be huge diversity of metabolites found in a sample depending on the organism and system, ranging from a few hundred to thousands (Dunn et al., 2012). This is among many reasons why there are still active experimental and computational challenges in the field (Cazzaniga et al., 2014; Dunn et al., 2012).

Metabolomic analytical strategies can be split into three categories: targeted, semi-targeted and untargeted (also known as profiling). The choice of strategy will depend on whether a known metabolite is required to be measured (targeted and semi-targeted) or if all metabolites including unknowns are to be identified (untargeted).

Targeted analysis allows for unambiguous identification of a small number of metabolites using chemical standards but will ignore any unknown metabolites, therefore does not provide a complete snapshot of the metabolome. For untargeted analysis a more global profile of metabolites is observed with the caveat that many metabolites will not be identified. Developments have been made to perform targeted analysis on larger sets of metabolites using kits of chemical standards e.g. with companies such as Biocrates (https://www.biocrates.com/). However the number of metabolites that are identifiable (365 lipids and 43 small molecules) is still much lower than the expected number of metabolites in a model organisms (e.g. 16,042 estimated for *Saccharomyces cerevisiae)* (Viant et al., 2017).

It is often stated that a bottleneck in untargeted metabolomics is the accurate identification (or annotation) of a metabolite in the form of a chemical composition and/or structure. A definitive identification of one metabolite is often not possible and this has given rise to a classification system determined by the Metabolomics Standards Initiative based on how accurate the

identification is (Fiehn et al., 2007; Sumner et al., 2007).

- Level 1: Confidently identified compounds
- Level 2: Putatively annotated compounds
- Level 3: Putatively annotated compound classes
- Level 4: Unknown compounds

The metabolome is incredibly sensitive and small changes in gene or protein activity can have measurable effects on metabolites (Kuile and Westerhoff, 2001). However, this high level of sensitivity comes at the cost of inherent variability observed in metabolomic samples of the same system (e.g. cell type, tissue) making it difficult to detect the unique metabolite difference that distinguish a normal and perturbed system, often the main goal of metabolomics studies. To find these underlying changes requires robust statistical analysis. Multivariate analysis is a typical approach that is made possible due to the unbiased way in which most metabolomics measure simultaneous metabolites at the same time. Principal component analysis (PCA) and partial least squares (PLS) being two popular approaches (Worley and Powers, 2013).

There is no perfect instrument for untargeted metabolomics and the identification of all metabolites in a sample is currently not possible. However, partly due to its high level of sensitivity, mass spectrometry is the most widely used platform to perform metabolomics (Dunn et al., 2012).

1.3 MASS SPECTROMETRY METABOLOMICS

1.3.1 MASS SPECTROMETRY

Simplistically, a mass spectrometer has 3 stages: Generation of gas-phase ions by an ion source, separation of the ions based on their mass-to-charge (*m/z*) ratio using a mass analyser, with the final step being detection of the ions. An additional chromatography stage is often used prior to ionisation to allow separation of complex mixtures e.g. liquid chromatography (LC-MS) or gas

chromatography (GC-MS) (Griffiths, 2007). Chromatography is not essential though and Direct Infusion Mass Spectrometry (DIMS) can be used in metabolomics (Southam et al., 2016).

Each stage of mass spectrometry can be implemented by a plethora of different instrument and method types. Chemical ionisation (CI), atmospheric pressure chemical ionisation (APCI), electron impact ionisation (EI), electrospray ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI) are all methods to achieve ionisation of the sample. Where EI is commonly used for GC-MS, most LC-MS platforms use ESI and APCI. The mass analyser stage is based on separation of *m/z* by either magnetic or electric fields, popular instruments include triple quadrupole (QQQ), Orbitrap, time-of-flight and Fourier transform ion cyclotron resonance (FT-ICR) – the latter giving the highest resolution of all mass analysers (Dunn et al., 2012; Griffiths, 2007). Detectors include electron multipliers, faraday cups and photographic plates. The sensitivity, chemical stability and speed of detection are all considerations required to determine the choice of detector (Rubert et al., 2015).

Upon ionisation of a sample molecule (M) of neutral charge forms a molecular ion ($M^{+/-}$) of either positive or negative charge. This is achieved in different ways depending on the ion source, for example EI involves bombardment of the sample molecule with high energy electrons. The molecular ion can be unstable and subsequently fragment to more stable products. This fragmentation can be useful in identifying structural information but too much fragmentation can lead to the molecular ion being unobservable, a problem often associated with EI. How extensive the fragmentation process can be dependent on the ion source, for example MALDI fragmentation is considerably less extensive than EI (Griffiths, 2007).

To obtain detailed fragmentation a collision of the ions with a background gas is typically required, this can be achieved through collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD), leading to the breaking of (normally) covalent bonds. The fragmentation pattern will be dependent on the structure of the metabolite with stronger

covalent bonds less likely to break. It should be noted that the mass spectrometer will still only be able to detect fragments that keep the ion charge (Griffiths, 2007).

The term tandem mass spectrometry (MS/MS or MS^2) is used when a single collision step is performed. There is more than one way of performing the collision step and the spectra can vary depending on whether an ion-beam or ion-trap instrumentation is used. If using an "ion trap" instrument, a fragment ion can be retained and undergo a further collision step and provide MS^3 spectra, the process can then be repeated. This method is typically called MS^n or multiple stage mass spectrometry. This is particularly useful in characterising metabolites in untargeted metabolomic experiments as it allows improved accuracy of chemical composition and can aid in structural elucidation (Scheubert et al., 2014).

The focus of this PhD project is primarily on the analysis and utilisation of DIMS, LC-MS and LC-MS/MS (liquid chromatography tandem mass spectrometry) for untargeted metabolomics.

1.3.2 FROM RAW MS DATA TO METABOLITE FEATURE

The process of transforming raw spectra to multiple interpretable metabolite “features” typically involves numerous steps and can depend on the objective of the study and the instrumentation used. The essential component of any data processing though will require “feature detection” defined here as determining the *m/z* value(s) - or *m/z* value(s) coupled to a chromatographic retention time window(s) - that represents a metabolite/compound as opposed to background noise in the spectra. The detected features can then be filtered, aligned and normalised based on the workflow being used.

As an example, a DIMS data-analysis workflow developed and optimised at the University of Birmingham involves processing the transient raw data with a method called selected-ion-monitoring (SIM) stitching, filtering for high quality peaks, removal of peaks from blanks, removal of outlier samples, data normalisation, batch correction, handling of missing values

and variance stabilisation and finally statistical analysis, for example for differential analysis. Example DIMS data and the associated workflow has previously been described in detail (Kirwan et al., 2014; Southam et al., 2016).

A popular LC-MS analysis workflow involves first a conversion of the instrument specific initial raw file format to an XML based open source file format e.g. mzML (Martens et al., 2011). Feature detection, retention time correction and clustering of similar features is performed using XCMS (Gowda et al., 2014; Smith et al., 2006; Tautenhahn et al., 2008).

Differences in *m/z* between features is a crucial step in data processing of metabolites as a single metabolite can be present in multiple ion forms in the mass spectrometer (e.g. adducts, isotopes and in-source fragmentation). Software to annotate these relationships include MI-Pack (Weber and Viant, 2010), PUTMEDID (Brown et al., 2011) and CAMERA (Kuhl et al., 2012). The deconvolution of adducts and isotopes reduces the complexity of the spectra and allows improved accuracy determination of a neutral mass of measured metabolites. The neutral mass is defined here as the sample molecule (M) before ionisation to a molecular ion. Often determining the neutral mass is the first step in annotating the measured metabolite.

1.3.3 MASS SPECTROMETRY FRAGMENTATION ACQUISITION IN METABOLOMICS

When using electrospray ionisation (ESI) mass spectrometry, gas-phase fragmentation can be used to generate product ions of an isolated precursor ion from a metabolite feature of interest. The measured characteristic product ions generated from the fragmentation of the precursor can then be used for annotating and identifying the metabolite.

There are many approaches to acquire the gas-phase fragmentation spectra, for LC-MS/MS these methods can include data-dependent acquisition (DDA), data-independent acquisition (DIA) and all ion fragmentation (AIF). Additionally, targeted approaches can be used where a known *m/z* or list of *m/z* values (and in some cases retention time) are selected for fragmentation

and other *m/z* ranges are ignored. Targeted methods will often be used when performing quantitative analysis of chemical standards. Semi-targeted methods cover a range of approaches that incorporate elements of targeted and untargeted approaches. Summary of gas-phase fragmentation mass spectrometry approaches are detailed in Figure 1-1.

All the acquisition approaches require either isolating a selected precursor ion from other unwanted contaminating ions either physically in the mass spectrometer or via some form of deconvolution of product ions to related precursor ions e.g. using correlation analysis (Geromanos et al., 2009). In proteomics it has been observed that the degree of contamination within an isolation windows (and the contribution of the selected precursor ion to the fragmentation spectra) is critical to achieving a reliable annotation when using DDA acquisition approaches (Michalski et al., 2011). For this thesis I have developed an R package (msPurity - https://doi.org/10.18129/B9.bioc.msPurity) for assessing the level of contamination with a metric termed "precursor ion purity". The level of contamination observed in metabolomics and the description of the R package are detailed in Chapter 2.

When performing gas-phase fragmentation for annotation purposes a key objective is typically to maximise the number of metabolite features that have corresponding fragmentation spectra and prioritising metabolite features which are of more importance. For example, the previously published "nearline" method (Neumann et al., 2013) can be used to automatically generate target lists of features aids to maximise the number of LC-MS features with fragmentation. For this thesis, I have adapted the nearline method with additional constraints, including "co-eluting" features within an LC-MS experiment. I have also developed tools that applied the nearline approach to other analytical techniques (fractionation and DIMS). Details of nearline developments can be found in Chapter 3.

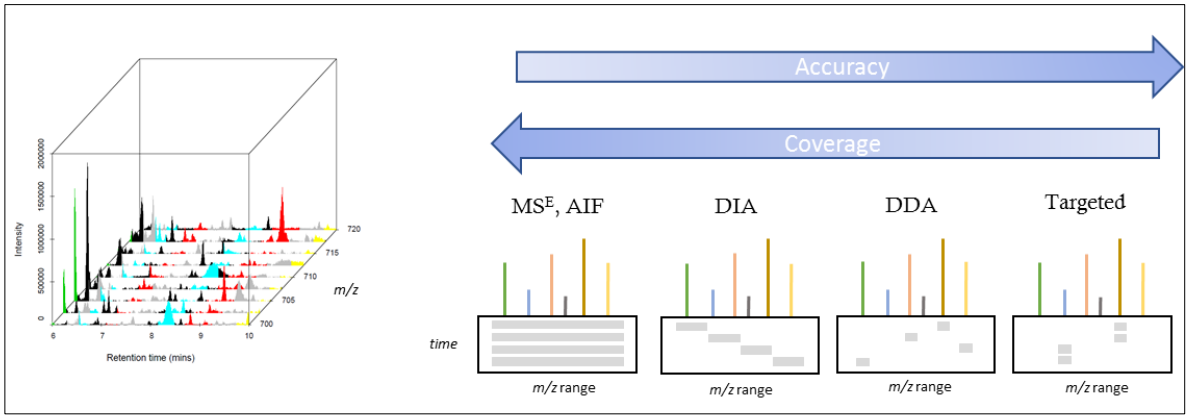


Figure 1-1 Gas-phase mass spectrometry fragmentation acquisition methods used for metabolomics

Common approaches for gas-phase fragmentation acquisition in metabolomics experiment. AIF: all ion fragmentation, DIA: Data independent acquisition. DDA: Data dependant acquisition.

1.4 METABOLITE ANNOTATION

1.4.1 KNOWN AND UNKNOWNS

We define the unknown metabolites based on the "Rumsfeld's quadrant" description of metabolite annotation (Stein, 2012), see Table 1-1. Where "known unknown" is a metabolite that is expected by an analyst but identified in a reference library, "unknown known" is a metabolite unexpected by the analyst but identified by a reference library and "unknown unknown" is where the metabolite is not identified by a library and is unexpected by the analyst.

Table 1-1 Rumsfeld's Quadrant of metabolite annotation

	Expected by Analyst	Unexpected by Analyst
Identified by library	Known Knowns: Expected and found	Unknown Knowns: Not expected but found
Not identified by library	Known Unknowns: Expected but not found	Unknown Unknowns: Not expected and not found

Approaches for metabolite annotations include searching a neutral mass or molecular formula against compound databases (see 1.4.2), searching against reference or in silico fragmentation

spectra (see 1.4.3 and 1.4.4), using fragmentation trees to infer an annotation (see 1.4.5) and/or using machine learning approaches (1.4.6).

In most cases annotation approaches require at least one known in the quadrant (depending on the definition of what constitutes a library and how one is used). "Unknown-unknowns" are either very difficult to identify a metabolite to a structure or in some cases impossible to identify. However, *de novo* annotation to a molecular formula is possible using mass spectrometry fragmentation spectra where reliance on a compound library is not a pre-requisite, see 1.4.5.

1.4.2 NEUTRAL MASS AND MOLECULAR FORMULA SEARCH

For LC-MS and DI-MS annotation the first step is often to use the calculated neutral mass of an unknown metabolite to search against compound databases such as the Human Metabolome Database (HMDB) or KEGG (Kanehisa et al., 2014; Wishart et al., 2007), where there will often be multiple annotations for one neutral mass. Alternatively, potential molecular formulae can be calculated from the neutral mass and the molecular formulas can be used to search against compound databases. In most cases the molecular formula can't be isolated to just one formula and again multiple potential annotations will typically be observed.

Prior chemical and biological knowledge can reduce the search space of potential structures. Approaches include using known ionisation behaviour and chemical information regarding the metabolite to remove any infeasible ion types (Draper et al., 2009) and using known enzymatic transformations to rationalise and reduce the metabolite identification false positive rate (Gipson et al., 2007; Rogers et al., 2009; Weber and Viant, 2010). The mapping of enzymatic transformation can also get information about the metabolic network (Rogers et al., 2009).

1.4.3 SPECTRAL MATCHING

1.4.3.1 Spectral matching for annotation

Spectral matching of experimental spectra to a reference compound (experimental or *in silico*

generated) is an essential approach for mass spectrometry identification and annotation. Indeed, to achieve level 1 identification a sufficiently unambiguous match between the metabolite feature of interest to a reference compound measured using the same mass spectrometry (and potentially chromatography) method is required.

For this thesis I developed spectral matching methods within the msPurity Bioconductor package and further implemented within Galaxy for searching experimentally collected untargeted metabolomics data against reference libraries. I also developed spectral matching methods to allow researchers to search against their own experimentally collected data (such as data collected for the DMA project), further details in Chapter 5 and Chapter 6.

1.4.3.2 Spectral matching basics

Prior to the matching of spectra, first **spectral filtering** is performed. This can involve filtering the library spectra based on various parameters including instrument type, instrument settings and the similarity between the library precursor *m/z* and the experimental precursor *m/z* value. **Spectral alignment** is then performed to attempt to ensure the correct peaks are being compared to one another in the query and library spectra.

Spectral matching is then finally performed to obtain a similarity score between the query and library spectra. Spectral matching in its simplest form can be done based on counting matching values of the query vector ($\boldsymbol{w}_Q$) and a vector from the target library ($\boldsymbol{w}_L$). A slight advancement of this, but still a relatively simplistic approach, is to calculate the dot product of the two vectors, equation (1.1).

$$F_D = \sum \boldsymbol{w}_Q \cdot \boldsymbol{w}_L \qquad (1.1)$$

The vectors can be composed of either the *m/z* or the intensity or an integration of both *m/z* and intensity in the form of a weighted intensity value, see equation (1.2) where *x* and *y* represent

weight factors.

$$\boldsymbol{w} = [\boldsymbol{peak_intensity}]^{x} \cdot [\boldsymbol{mass(m/z)}]^{y} \quad (1.2)$$

Different recommendations have been provided for optimum weight factors. The square of the *m/z* value multiplied by the square root of the intensity is proposed for ESI based data in the paper describing the mass spectra repository MassBank, i.e. x = 0.5 and y =2 (Horai et al., 2010).

Often the result reported for a dot product score is the cosine similarity. This score divides the dot-product by the magnitude of the vectors to give a bounded result between -1 and 1 see equation (1.3). As the $\boldsymbol{w}_Q$ and $\boldsymbol{w}_L$ vectors will typically always contain positive numbers, the range will in practice will be bounded between 0 and 1.

$$F_C = \frac{\sum \boldsymbol{w}_Q \cdot \boldsymbol{w}_L}{\sqrt{\sum \boldsymbol{w}_Q^2} \cdot \sqrt{\sum \boldsymbol{w}_L^2}} \quad (1.3)$$

1.4.3.3 Spectral matching confidence

For proteomics the confidence of a match is made considerably easier by having a database of known and predicted protein searches. This allows the calculation of the decoy-to-target ratio (also known as peptide False Discovery Rate) by creating of a decoy sequence database (Perkins et al., 1999). Metabolomics does not have this luxury and calculating the confidence is considered the most difficult aspect of spectral matching. The false negative risk, i.e. a query compound not found in the library, can arise from contamination of the sample from overlapping features in retention time, missing peaks due to low signal to noise and isomerization reactions in the sample not characterised in the library. The false positive risk, i.e. a match occurs in the library but for an incorrect compound, will depend on the uniqueness of the compound spectra (Stein, 2012).

It should be noted though that similar structures (and sub-structures) have similar spectra so a false positive results can also identify similar compounds and classes (Scheubert et al., 2013).

Spectral matching methods and confidence scoring are tailored for the instrumentation used and prior data analysis steps. Here I divide the types of matching based on whether the data is derived from EI or if the data is derived from gas-phase fragmentation (often used with "soft ionization techniques" such as ESI).

1.4.3.4 Spectral matching for EI spectral data

There are a variety of scoring approaches for single stage MS type data and most are predominately based around EI fragmentation using GC-MS instrumentation. An often cited 25-year-old assessment (Stein and Scott, 1994) was performed on EI based matching that tested Euclidean distance, absolute value distance, Hertz, dot-product (cosine similarity), probability based matching (PBM) (McLafferty et al., 1974) and an optimised dot-product method called "composite" developed for the paper (Stein and Scott, 1994). The optimised method performed best followed by the standard dot-product (cosine similarity) score, see equation (1.3). The widely used NIST MS search tool comes implemented with the "composite" approach often just referred to as "match factor" (see equations (1.4) and (1.5)) that incorporates relative intensity of neighbouring peaks, where F_C = equation (1.3), N= number of peaks, Q= query, L= library, $L\&Q$= peak in query and W library and is the weighted value.

$$F_R = \left(\frac{1}{N_{L\&Q} - 1}\right) \sum_{i=2}^{N_{L\&U}} \left(\frac{W_{L,i}}{W_{L,i-1}}\right)^n \left(\frac{W_{Q,i}}{W_{Q,i-1}}\right)^{-n} \quad (1.4)$$

$$F_M = \frac{1000}{N_Q + N_{L\&Q}} \left(N_Q F_C + N_{L\&Q} F_R\right) \quad (1.5)$$

The Stein and Scott paper also detailed two values used to interpret the confidence of the score, one based on an assumption that the unknown compound is in the library and the other based

on assumption that the unknown compound is not.

Minor improvements have been made since the match-factor method (Koo et al., 2011) but this modification to the dot-product method is still one of the most reliable methods for EI spectral matching (Scheubert et al., 2013).

1.4.3.5 Spectral matching for gas-phase fragmentation

Gas-phase fragmentation collected for tandem mass spectrometry (MS/MS) is typically less reproducible than EI fragmentation, therefore careful consideration needs to be made regarding instrumentation and collision energies used to create the reference library to be matched against. Similar approaches to EI scoring can be used for MS/MS data, including PBM (McLafferty et al., 1998) and dot product cosine but other approaches that have been tailored specifically for LC-MS/MS include using a machine learning approach that combine multiple similarity measures (Zhou et al., 2010) and a ranking system that can be trained on specific datasets to counteract for instrumental variation (Mylonas et al., 2009). A modification of a proteomics spectral matching method was also developed specifically for matching against *in silico* fragmentation spectra in a paper describing the automated fragmentation tool HAMMER (Zhou et al., 2014).

1.4.3.6 Matching and Clustering Fragmentation trees

Once a fragmentation tree is obtained one can then perform clustering and matching of trees to characterise unknown compounds and identify similarity. This can be achieved for tandem mass spectrometry through FT-BLAST (Fragmentation Tree Basic Local Alignment Search Tool) by searching against a spectral library (Rasche et al., 2012). The Sirius CSI:FingerID software also uses a kernel based similarity measure for both MS/MS spectra and fragmentation trees (Dührkop et al., 2015; Shawe-Taylor et al., 2004). An approach has been used for MS^n whereby trees are converted into fingerprints and compared using a Jaccard index (another measure of

similarity) (Kasper et al., 2012; Rojas-Cherto et al., 2012).

The MS^n fragmentation tree resource mzCloud uses a *non* dot-product based spectral matching algorithm (proprietary - not published) to query LC-MS, LC-MS/MS and MS^n against its experimental and i*n silico* database. The service also allows using Precursor Ion Fingerprinting (PIF) substructures of compounds can be searched (Horner et al., n.d.).

1.4.4 *IN SILICO* FRAGMENTATION

One approach for identifying the unknowns is to predict the fragmentation spectra of chemical compounds and search against this *in silico* library. This can be achieved based on a set of fragmentation rules taken from literature but can also be achieved by breaking any conceivable bonds to make up hypothetical fragments (combinatorial fragmentation).

LipidBlast is an open source library of *in silico* generated fragmentation spectra but is dedicated solely to lipids (Kind et al., 2013). Mass Frontier (http:// www.highchem.com/massfrontier/mass-frontier) is a commercial product applicable to all types of metabolites that contains the most extensive library of fragmentation rules determined from the literature. Mass frontier is however not set up for high-throughput use and does not have a command line interface (CLI). This has been overcome by the (High-throughput AutoMation of Mass frontiER) HAMMER software (Zhou et al., 2014). This software uses visual scripting to allow easier creation of *in silico* libraries, in addition it provides spectral matching and scoring of the experimental data to the generated libraries.

Our knowledge of fragmentation is incomplete though and this is a major disadvantage of software like Mass Frontier. This problem is addressed by combinatorial fragmentors using more basic rules to systematically disconnect bonds and ranking the resulting substructures. Software for this includes fragment iDentificator (FiD) (Heinonen et al., 2008) which mixed linear programming approach to enumerate all possible fragments. Unfortunately this is too

time consuming to perform on medium to large compounds (Scheubert et al., 2013). MetFrag (Ruttkies et al., 2016) addresses this problem and uses a multiple heuristic based approach and is able to enumerate possible fragments much faster.

In this thesis I have implemented workflows for annotation that incorporate both spectral matching to LipidBlast libraries and the use of MetFrag for annotation purposes, see Chapter 5 and Chapter 6.

1.4.5 FRAGMENTATION TREES

Fragmentation trees attempt to explain the fragmentation process through a network of interlinking connections where each node represents the ion (or annotation) of either a precursor or fragment and each edge represents their potential connections. Fragmentation trees can be used to determine molecular formulae of an unknown compound without the use of spectral libraries or molecular databases; approaches have been developed for gas-phase fragmentation - MS/MS (Böcker and Dührkop, 2016; Rasche et al., 2011), EI (Hufsky and Böcker, 2012) and for MS^n (Kasper et al., 2012). It should be noted that fragmentation trees are different from a spectral tree where the latter is simply explaining the connections between the MS levels of a single compound for MS/MS or MS^n spectra.

To determine a final fragmentation tree, potential empirical (or molecular) formulae are determined for each node and edge to make what can be described as a fragmentation graph. Then by using a scoring function the graph can be filtered down to a tree that best describes the fragmentation process.

Metabolite annotation (to a molecular formula) using fragmentation does not require a spectral library or molecular database because the potential molecular formulae can be identified through the isotope pattern (Böcker et al., 2009) and searching through all possible formulas (Böcker and Rasche, 2008). This can produce a large number of fragmentation trees though and

identifying the optimum tree out of all of the combinations has been determined as computationally hard (Rauf et al., 2012). The problem has been named "Maximum Colourful Subtree" where exact and heuristic approaches have been developed to solve the problem (Böcker and Dührkop, 2016; Böcker and Rasche, 2008). The problem of determining MS^n trees for the same purpose is called "Combined Colourful Subtree" and is more difficult to solve but an integer linear programming approach has been successfully applied (Scheubert et al., 2014), the paper also demonstrated that the MS^n data improves the quality of fragmentation trees.

There are now several standalone tools and services that either utilise MS^n fragmentation trees or aid in the analysis of MS^n data. mzCloud is one such resource that contains a searchable database of MS^n fragmentation trees of chemical standards derived from both experimental and in silico sources.

Isotope patterns, fragmentation trees and machine learning are used for the metabolite annotation software Sirius - CSI:FingerID (Böcker et al., 2009; Dührkop et al., 2015). The implementation of Sirius – CSI:FingerID in Galaxy is detailed in this Chapter 5. Additionally, I have developed tools to integrate the results of Sirius – CSI:FingerID with other annotation software (e.g. spectral matching with msPurity, MetFrag and Probmetab). I then applied the approaches to DMA data analysis, see Chapter 5 and Chapter 6.

1.4.6 MACHINE LEARNING

Machine learning can be broadly defined as any computer program that can use example data or past-experience to solve a problem – approaches ranging from regression analysis to neural networks. Software and packages are available specifically to perform machine learning based metabolite annotation (Dührkop et al., 2015; Laponogov et al., 2018).

The Sirius-CSI:Finger software (Dührkop et al., 2015) uses a machine learning approach for annotation and has been applied for this PhD thesis. The approach follows three stages, the

learning phase, the prediction phase and the scoring phase, see Figure 1-2.

The **learning phase** first involves processing of the spectral library and creating fragmentation trees. Several kernel similarity measures are used for all the reference compound MS/MS spectra or fragmentation trees (Kernels can be thought of as a dot product in multi-dimensional space). Weightings are then used to combine all similarities measure into a single measure using a 'multiple kernel learning' approach (Shen et al., 2014). A molecular fingerprint (a binary representation of multiple chemical properties of a compound) is then calculated for each reference compound. Support vector machines (SVM) are then trained using the calculated kernel similarity score to classify compounds that can exhibit a molecular property and those that don't.

The **prediction phase** is where molecular properties are predicted for the unknown compounds MS/MS spectra. First fragmentation trees are calculated from the unknown compounds MS/MS spectra. Kernel similarities are then calculated between the unknown fragmentation tree and MS/MS against the reference compounds. The previously trained SVMs are then used to predict molecular properties of the unknown compound and the predicted molecular fingerprint - Platt probabilities are used to replace the binary predictions of classical SVM approaches (Platt and others, 1999).

The **scoring phase** involves first searching a compound database (such as PubChem) and extracting all matching compounds based on the potential unknown molecular formula. Molecular fingerprints of the matching compounds of this query are then compared to the predicted molecular fingerprint of the unknown compound. A score is then given based on the similarity between the two fingerprints and added to a sorted list of potential candidates. The user then chooses the potential candidate out of the sorted list (e.g. a user might be expecting certain types of compounds based on the chromatography or extraction).

It should be noted that this approach only annotates a compound to the correct constitution (bond structure) of the unknown compound. The authors argue that using MS/MS data and applying search-based annotation approaches like the above are insufficient for elucidation of stereochemistry and definitely beyond the limits of full elucidation of a compound (Dührkop et al., 2015).

Machine learning in metabolomics is still limited to the reference spectral libraries used for the prediction stage. It is essential for more reference compound spectra and metabolite data in general to be made publicly available if the accuracy and reach of these approaches is to be improved.

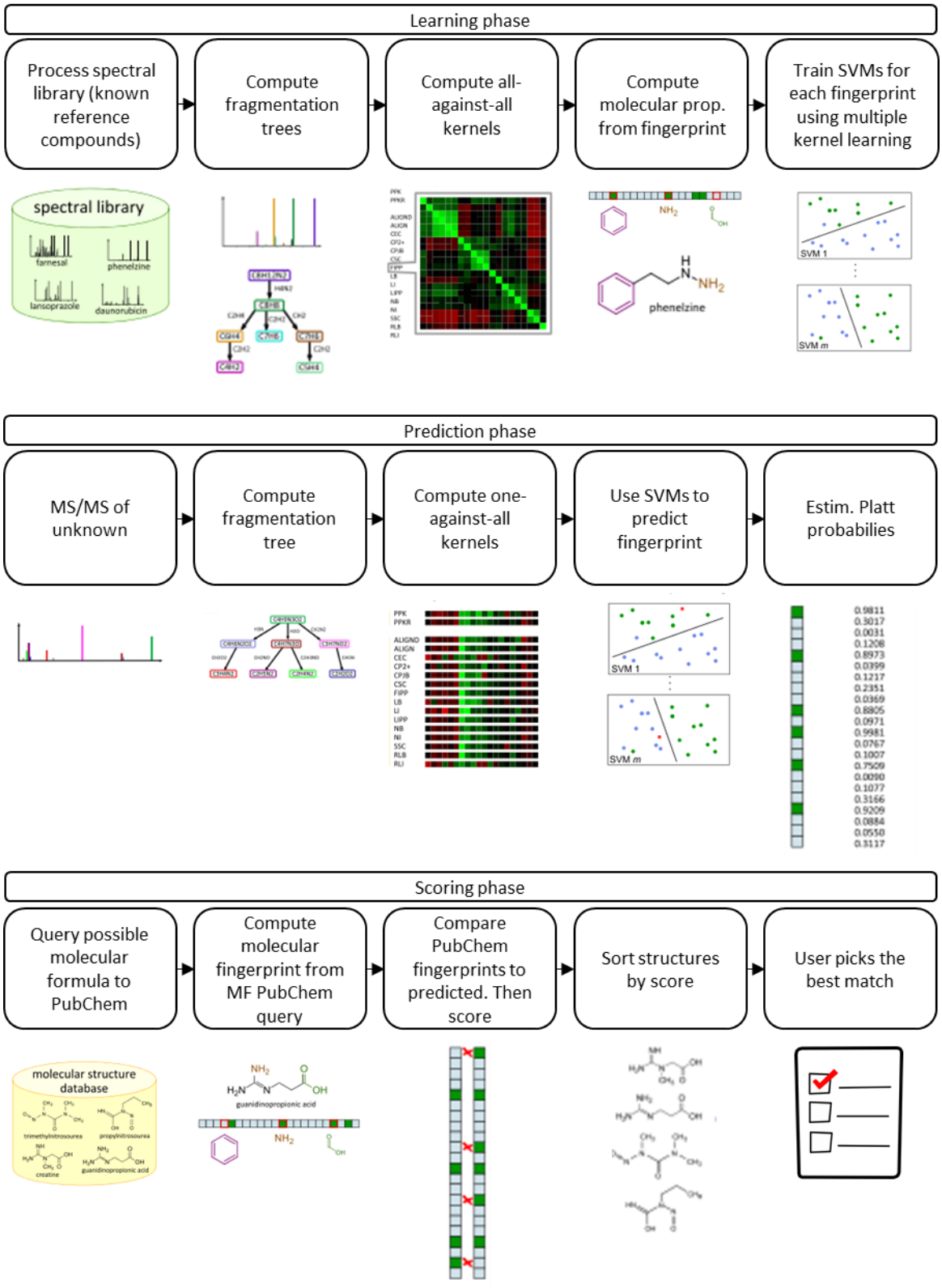


Figure 1-2 CSI:FingerID annotation approach

Modified from (Dührkop et al., 2015). SVM = support vector machine. MF = molecular formula

1.5 REPRODUCIBLE AND FAIR METABOLOMICS

Independent replication of experimental results is a cornerstone of science, despite this, many scientists struggle to replicate both their peers and their own (Baker, 2016). Additionally, as popular news providers report on the perceived reproducibility crisis (Devlin, 2018; Feilden, 2017), the impact of poor reproducibility is being considered not just within the confines of the scientific community. With a multitude of sources, many of which are difficult to overcome, the challenge being faced is not trivial. However, it is disconcerting that there are even difficulties in simply repeating the identical computational analysis of an experiment (Freedman et al., 2015; Garijo et al., 2013)

Scientific reproducibility - and particularly computational analysis - can be drastically improved by ensuring the data and metadata is shared within FAIR principles: Findable, Accessible, Interoperable, and Re-usable (Wilkinson et al., 2016) see Box 1-1. To share Metabolomics data and metadata with FAIR principles one can submit their data and metadata to the MetaboLights public data repository (Kale et al., 2016). MetaboLights uses the ISA (Investigation, Study, Assay) framework and ISA-Tab file format for describing the experimental project (Sansone et al., 2008). The investigation layer describes the objectives, factors, protocols and parameters of a project; the study layer describes the subjects and samples; and each study can have one or more assays detailing a measurement (e.g. metabolic profiling with LC-MS/MS).

Once shared in a public data repository such as MetaboLights, the data and metadata then become findable, assessible, interoperable and reusable**.** **Findable** because a Digital Object Identifier (DOI) is created that uniquely identifies the project and allowing it to be easily searched for. **Accessible** as the MetaboLights website can be visited and browsed or accessed through programmatic services. **Interoperable** as MetaboLights use both XML open source raw data objects and the ISA format metadata that are both computer readable so can

interoperate with workflows and software. Controlled vocabularies across both metabolomic and other 'omic studies should also allow the data to be integrated with other data. Finally, the data and metadata will be **reusable** - the purpose of FAIR principles. This is achieved in part from MetaboLights enforcing the minimum reporting standards for metabolomics (Fiehn et al., 2007) so that the required technical-scientific information is sufficient to (re)-use the data and metadata.

The public datasets are then available not just to ensure reproducibility but they can then be used for assessing data processing methods and new software (Lawson et al., 2017; Stanstrup et al., 2015), or for annotation purposes if the fragmentation spectra are exported to other repositories such as MassBank (Horai et al., 2010). However, the percentage of published metabolomic studies where the associated data and meta data publicly available is still smaller than one would expect considering how many metabolomics studies are being published (Rocca-Serra et al., 2016). This could be in part due to the sometimes-arduous task of creating an ISA structure of a metabolomics experiment, suggesting that new tools and integrations are still required to help researchers comply to share their data.

To aid in the submission of metabolomics datasets to public data repositories, I have co-developed the XML2ISA suite of software that utilises the metadata rich instrument open source data files to help populate ISA-tab metadata files, see Chapter 4.

To be Findable

- To make use of shared data, humans and computers first need to be able to find it. Machine-readibilty being key for automated services are to take advantages of the data, for example using Digitial Object Identifier (DOI) to uniquely identifier and search for projects. See below for official guidelines:

- F1: (meta)data are assigned a globally unique and persistent identifier
- F2: data are described with rich metadata (defined by R1 below)
- F3: metadata clearly and explicitly include the identifier of the data it describes
- F4: (meta)data are registered or indexed in a searchable resource

To be Accessible

- Afer the data has been found, a user or computer needs to be able to access it. For example "clicking" and downloading for a user on a website or more programatic access. See below for official guidelines:

- A1: (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1: the protocol is open, free, and universally implementable
  - A1.2: the protocol allows for an authentication and authorization procedure, where necessary
- A2: metadata are accessible, even when the data are no longer available

To be Interopable

- To make use of the accessed data and metadata it might need to be integrated with other data and interoperate with workflows and software. For example using data formats and standards that are suitable for the data metadata (e.g. JSON, OWL). See below for official guidlines:

- I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2: (meta)data use vocabularies that follow FAIR principles
- I3: (meta)data include qualified references to other (meta)data

To be Resuable

- The goal of FAIR is to enable data and metadata to be resuable. This means licensing should be clear and the relevant technical-scientific information should be present as defined by the domain-community (e.g. metabolomics minimium reporting standards). See below for official guidlines

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

Box 1-1 The FAIR Guiding Principles. Official guidelines from (Wilkinson et al., 2016)

1.6 WORKFLOW ENVIRONMENTS

Experimental data analysis across all scientific fields can involve complicated multi-step workflows using multiple software, data sources and parameters. This complexity can lead to considerable difficulty in reproducing the data analysis (Freedman et al., 2015; Garijo et al., 2013). To address this issue, over the past 10 years there has been a steady development and improvement in use of workflow environments to handle and process data analysis workflows in a controlled and reproducible manner (Atkinson et al., 2017; Cohen-Boulakia et al., 2017). Galaxy, KNIME and Taverna (Afgan et al., 2018; Berthold et al., 2009; Wolstencroft et al., 2013) are three popular workflow environments where a variety of tools and workflows for metabolomic analysis are available (Beisken et al., 2014b; Brown et al., 2011; Davidson et al., 2016; Giacomoni et al., 2015; Liggi et al., 2018; Verhoeven et al., 2018). The Galaxy platform being one of the most successful and popular workflow environments in bioinformatics (4,300 publications citing Galaxy usage (Atkinson et al., 2017)) and a recent survey showed a strong interest within the metabolomics community for Galaxy (Weber et al., 2017).

Key strengths for using most workflow environments include **reproducibility** - relevant and sufficient information is captured (e.g. tools, parameters, dependencies and versions) to reproduce analysis; **accessibility** - the analysis can be visually inspected and typically will not require programming or advanced skills; and **transparency** - easy to share the workflows analysis and results. These strengths all aid in sharing FAIR metabolomics data and metadata.

Many of the tools developed within this thesis were implemented in Galaxy or were developed to integrate with Galaxy.

1.7 METABOLOMIC DATA MANAGEMENT

1.7.1 DATA MANAGEMENT SOFTWARE

Data management can be broadly defined as any administrative process from acquisition to archiving in the life cycle of data. The explosion in data intensive research in all fields, especially post-genomic technologies, presents challenges in reproducible data management. For the scientific community to benefit from the plethora of data being generated, careful consideration in procedures, technology and software is required (Wilkinson et al., 2016).

Data management in the context of experimental science can be divided into four stages:

1. Planning, policies and resource allocation e.g. data storage and management plans and policies based on individual, institutional and research requirements.
2. Experimental management e.g. recording experimental design details, instrumentation information, sample tracking and experimental protocols.
3. Data analysis management e.g. recording software packages and versions, algorithms, parameters, data analysis workflows, results and data.
4. Public data management e.g. disseminated data within specialist repositories (e.g. MetaboLights, PRIDE).

Data management software can be used in all the above stages but here I focus on software that aids in the management of the experimentation and data analysis (stage 2-3). To achieve FAIR data, efficient local management of the experimentation and data analysis is essential as it allows a more reliable and automated transition to public facing ensuring relevant information has been tracked and not collected retrospectively.

Laboratory information management systems (LIMS) have long been used in commercial and academic laboratories to reliably trace all aspects of experimental work and data analysis and are crucial to many clinical studies where strict requirements to good clinical practice (GCP),

good laboratory practice (GLP) and good manufacturing practice (GMP) necessitate traceability and transparency to regulating bodies.

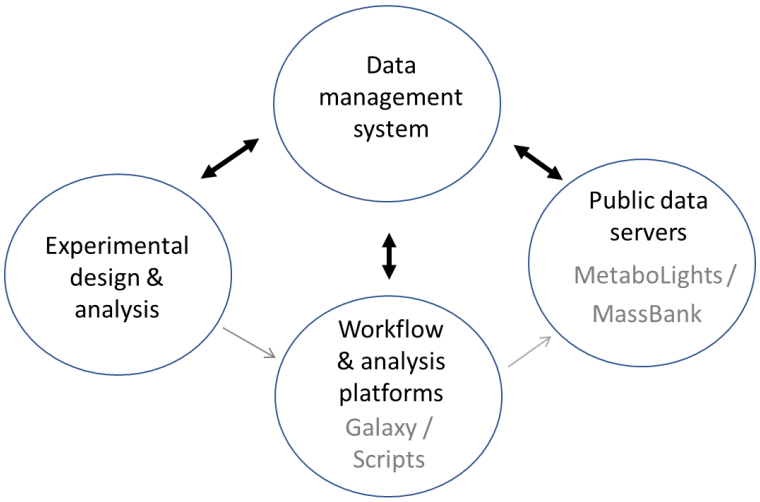


Figure 1-3 Metabolomic data management

Grey arrows indicate traditional flow of data that is typically performed with manual unrecorded processes. Black arrows show flow of data through a data management system in automated and recorded processes.

LIMS typically do not serve data to the public. However, if the LIMS tracked enough information regarding the experimental and data analysis, the exporting of data to a public resource (e.g. MetaboLights) can be automated as many of the meta-data requirements can be parsed from the LIMS database. This is possible with the MASTR-MS LIMS developed specifically for metabolomics (Hunter et al., 2017).

Full implementation of a LIMS can be a daunting task to a small research group and the benefits might seem remote. Data management systems that focus more on data analysis could be more appealing to some, such as the PiMP software (Gloaguen et al., 2017). Additionally, having a locally available resource to view, search and curate large scale data sets to interrogate data prior to publication can be extremely beneficial for researchers

In Chapter 5, a data management system was developed to manage experimental and data

analysis. The Django MOGI application suite (Metabolomics Organisation for Galaxy and ISA) uniquely integrates with both the ISA framework and Galaxy workflow environment. In Chapter 6, the Django MOGI application suite was then used to create the DMAdb (Deep Metabolome Annotation Database) to manage and analysis the data from the DMA of *D. magna.*

1.7.2 DATA MANAGEMENT WITH DJANGO

Data management software for handling experiments lends itself to web application with a central database of protocols, procedures and data. The ability to control user access via authentication and authorization is beneficial and the central storage location helps eliminate duplication of information.

A popular approach to developing a web application is the Django web-framework. Used across commercial and academic organisations (Mozilla Firefox, Bitbucket, Instagram and NASA), Django was built for fast development of secure and scalable web applications. For these reasons, Django has been successfully used in data management and data analysis in both genomics and metabolomics (Correia et al., 2015; Gloaguen et al., 2017; Hunter et al., 2017). The data management software developed in Chapter 6 was implemented in Django. Re-usable submodules (Django applications) were created for key components that can be integrated into other Django projects. See below for details of the structure of the Django framework.

A Django project consists of one or more Django applications and contains all the components for a Django website. A Django application is a submodule of a Django project that provides self-contained functionality. A re-usable Django application is a Django application that has been developed in a way that it can be imported into any Django project. Every standard Django application will follow the MTV (model, template, view) architecture. The model is a python class representing a database table. The view is either a python function or class used to receive

HTTP (Hypertext Transfer Protocol) request, perform an action - typically with a model (e.g. create, edit and delete) - and return a HTTP response. The template is the presentation layer, normally in the form of a HTML (Hyper Text Markup Language) template file and allows web pages to by dynamically loaded with different content. The request-response cycle used within Django is detailed in Figure 1-4.

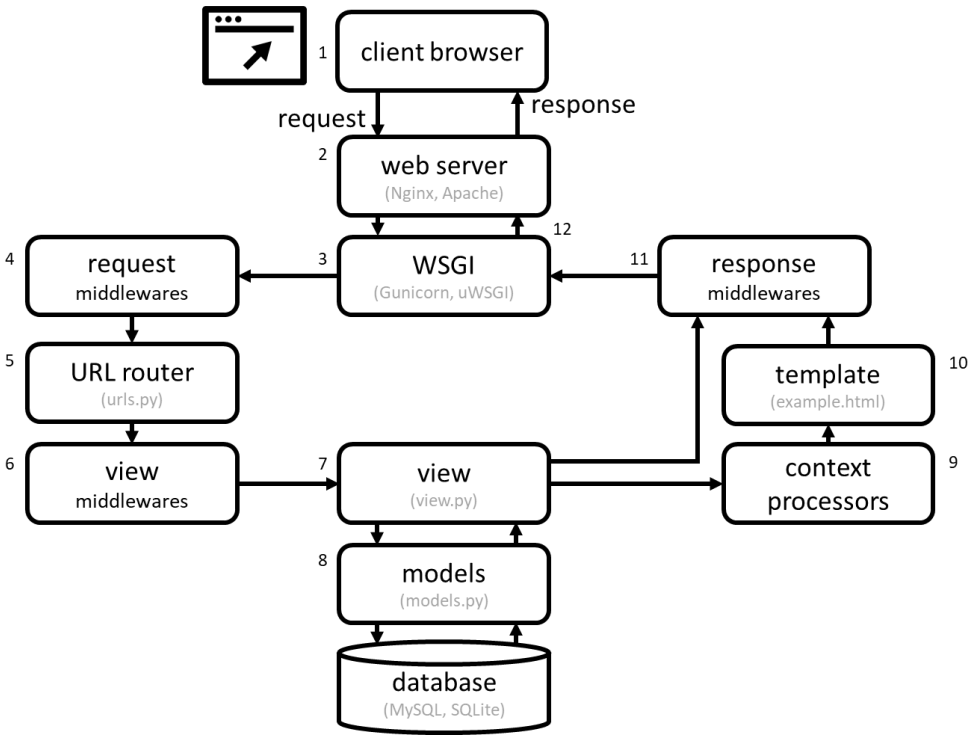


Figure 1-4 Django request-response cycle

WSGI = Web server gateway interface. URL = Uniform resource locater. 1) User requests a page on a web browser. 2) The web-server either serves a static file directly to the browser or if a python application is going to be used, the request is sent to the WSGI. 3) The WSGI provides the standard interface between web interfaces and python applications. Based on the WSGI configurations, the request will be sent to the Django request middlewares. 4) The middlewares in Django are a collection of python classes that deal with every request or response. The request middleware receives the request and sends to the URL router 5) The URL router takes the URL path from the request and checks against the patterns in urls.py file to obtain the relevant view to use. 6) The view middlewares handle the request and send the relevant information to the view. 7) The view is where the "business logic" of the application is performed (e.g. create, edit or delete) 8) The view can optionally access the database through the models depending on the operation being performed. 9) The view sends the request, including any other information based on the business logic to the context processors. 10) The context

processors allow data to be added to the template. The HTML template is rendered into a standard HTML output. 11) The view sends a HTTP response with the reference to the relevant template to the response middlewares. 12) Finally, the WSGI sends the response to the web server to serve the rendered HTML file to the user.

The view, template and models are where the majority of the development takes place when creating a new Django application.

1.8 *DAPHNIA MAGNA* DEEP METABOLOME ANNOTATION (DMA)

The lack of knowledge regarding model organism metabolomes limits the biological interpretation of metabolomics studies (Viant et al., 2017). However due to the challenges in metabolite identification and the wide physiochemical properties observed from metabolites there has been no single experiment or study that has set out to extensively characterise a metabolome. The DMA of *D. magna* is an ongoing project at the University of Birmingham and the first attempt at such an analysis, by performing these deep metabolome annotations, the intention is to use time and resources on annotation that would not be feasible for more routine metabolomics analysis. The annotation generated will then be both stored in a local database for internal analysis and disseminated to public data repositories (such as MassBank and MetaboLights). The hope being that the annotations obtained would be used as a resource for routine metabolomics somewhat analogous to how an initial in-depth sequencing of a genome is used to create a reference for further sequencing and analysis. See Figure 1-5 for a general overview of DMA.

The DMA of *D. magna* is the first of what is hoped to be many, where gradually the phyometabolomics (Edison et al., 2016; Viant et al., 2017) of metabolism could be elucidated.

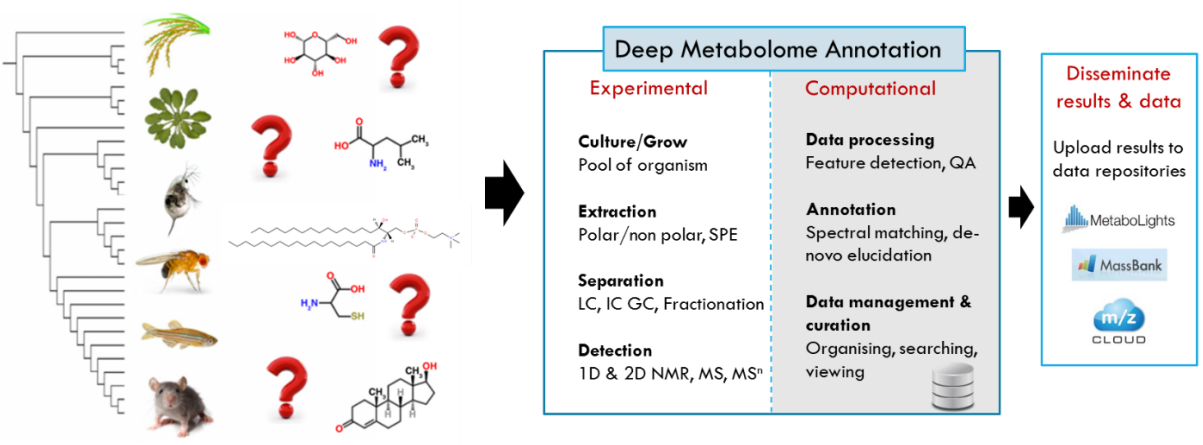


Figure 1-5 General overview of Deep Metabolome Annotation

Figure adapted from (Viant et al., 2017). SPE = Solid phase extraction. LC = Liquid chromatography. GC = Gas chromatography. 1D & 2D = 1 dimensional & 2-dimensional nuclear magnetic resonance. MS = mass spectrometry. MS^n = multiple stage mass spectrometry. MS/MS = tandem mass spectrometry. QA = quality assurance.

The genus *Daphnia* of the sub-phylum Crustacea and order Cladocera are located in a geographically widespread and diverse range of freshwater environments where they often play a crucial role as a primary consumer in the aquatic food web (Lampert, 2006). These factors, along with their relative ease of culturing and short generation time have persuaded and enabled scientists to investigate the ecology of *Daphnia* for well over a century (Weismann, 1876). The genus has also been used extensively as a model to investigate evolution (Ebert, 2005), genomic aspects of human health (NIH recognised model organism), and as an indicator genus used to set ecotoxicological regulatory standards (SOR/2002-222). (Canadian Metal Mining Effluent n.d.). The sequenced genome of *Daphnia pulex* (Colbourne et al., 2011) provided valuable information for all the above fields but to understand the system wide mechanisms of pathways and reactions for *Daphnia* it is crucial to understand its metabolome.

For the DMA of *D. magna*, ten strains of *D. magna* were cultured under multiple controlled culturing conditions to achieve a broad biological diversity, the strains were then pooled to form

a single homogenate. Apolar and polar extracts were then prepared from the homogenate and underwent two distinct solid phase extraction (SPE) methods. At least three fractions were then produced from each SPE method, each fraction then underwent separation via one or two types of liquid chromatography in both positive and negative mode whilst LC-MS/MS was collected. The elute from each chromatography column was then fractionated into 96 well plates where more extensive fragmentation analysis is to be performed.

For this thesis, computational workflows I developed and described in Chapter 3 were used to collect data for the DMA of *D. magna.* Assessment of the precursor ion purity of the spectra (using approaches described in Chapter 2) has also contributed to assessing the quality of the data.

Tools I developed and detailed in Chapter 5 were also used to manage, analyse and provide a resource so that other researchers can browse and search the DMA data and annotations. Annotation workflows for LC-MS/MS were also developed (Chapter 5) and applied to the DMA of *D. magna* dataset (Chapter 6). See Chapter 6 for a summary of all the annotations obtained so far.

1.9 AIM AND THESIS OVERVIEW

The overall aim of this PhD was to create computational tools and workflows to allow for reproducible and transparent approaches for acquiring data, deriving metabolome-wide metabolite annotations, and managing and disseminating the results. The primary use of the tools and workflows being for large-scale annotation projects such as the ongoing and extensive DMA of *D. magna*.

Available software tools and resources to perform large scale annotation analysis of metabolomics data were limited when this PhD was initiated (2014). As part of this PhD I have developed software to address these limitations and enabled large scale annotation analysis such

as the DMA of *D. magna* to be performed. See below for how I addressed key challenges presented during the PhD:

1) Are chimeric spectra problematic in untargeted metabolomics experiments and how can they be assessed?
2) Can the proportion of informative fragmentation spectra collected for DDA based mass spectrometry analysis be increased?
3) Can large scale metabolomics annotation studies be analysed and managed reproducibly?
4) Can large scale metabolomics annotation studies be applied to model organisms?

In Chapter 2, I addressed the question: *"Are chimeric spectra problematic in untargeted metabolomics experiments and how can they be assessed?"*. When collecting large volumes of mass spectrometry fragmentation data, it is often too time consuming to manually check every spectrum to ensure of its quality. One such important assessment is determining if contaminating ions have been fragmentated alongside the precursor ion of interest resulting in "Chimeric" spectra. Software provision for assessing the quality of fragmentation spectra in metabolomics is limited and prior to the work performed in this thesis there were no software or assessments for assessing chimeric spectra in metabolomics. As part of this PhD, I developed the precursor ion purity metric to assess chimeric spectra for metabolomics and applied the approach to multiple datasets to determine first assessment of chimeric spectra typically observed in metabolomics. The metric consolidates different approaches previously detailed in proteomics but also adds novel functionality including normalising the spectra based on the isolation efficiency of the mass spectrometer. The software developed (msPurity) and was used extensively for DMA of *D. magna* for both precursor ion purity assessments and other data processing and annotation (including spectral matching).

In Chapter 3, I addressed the question: *"Can the proportion of informative fragmentation spectra collected for DDA based mass spectrometry analysis be increased?"*. Often with DDA

based approaches, the fragmentation spectra only explain a small proportion of the data seen at the full scan (MS^1) level. Building on the nearline approach (Neumann et al., 2013), workflows were developed to acquire DDA fragmentation focusing on the most "informative" precursor ions, and applied to the DMA of *D. magna*. New developments included more flexible handling of co-eluting features, prioritising features based on multiple criteria (e.g. precursor ion purity and adduct type) and applying the nearline approach to DI-MS/MS and DI-MS^n data. Additionally, a genetic algorithm was considered for improving LC fractionation using the same philosophy of utilising prior information collected from LC-MS data to improve later fragmentation acquisition.

In Chapter 4, I in part addressed the question: *"Can large scale metabolomics annotation studies be analysed and managed reproducibly?"* The use of ISA-Tab structures aids in improving reproducible and FAIR metabolomics data (see section 1.5) and is key to managing metabolomics reproducibly, however the construction of these data files is non-trivial and there were no tools to help automate this process - where manual checking and retrieving technical information from thousands of data files for large studies (such as the *DMA of D. magna)* becomes increasingly time consuming and error prone. As part of this PhD, I co-developed a suite of software to semi-automate the construction of ISA-Tab file based on metabolomics data files. Although not applied directly for the DMA of *D. magna* in this Thesis - the eventual aim would be to apply these tools at the completion of the *DMA of D. magna* to aid in the submission to MetaboLights data repository.

In Chapter 5, I continue to address the question: *"Can large scale metabolomics annotation studies be analysed and managed reproducibly?"* Data analysis in metabolomics requires multiple stages typically using various software and tools, the resulting results can often be difficult to handle. To manage the data and results large scale metabolomic annotation studies a data management software suite (MOGI) was developed and applied for the DMA of *D.*

magna. The MOGI suite of applications is now one of only two open source management software that handles mass spectrometry fragmentation metabolomics data and results – the other being PIMP (Gloaguen et al., 2017) which lacked a number of utilities required, importantly, the MOGI software is unique in that it organises metabolomics data in ISA format and integrates with Galaxy – proving a valuable resource for reproducible and FAIR metabolomics data management. It should be noted as well that at the start of the PhD (2014) there was no data management software for handling metabolomics mass spectrometry fragmentation data.

Throughout this PhD, all software and workflows were developed to be integrated into Galaxy – either through a Galaxy tool wrapper or through an API (application programmatic interface) connection with Galaxy. This provides a valuable contribution to reproducible metabolomics data analysis, especially in terms of fragmentation analysis where little to no tools were available in Galaxy for metabolomics.

In Chapter 6, I addressed the question: *"Can large scale metabolomics annotation studies be applied to model organisms?"*. The tools and workflows developed in the preceding chapters were used alongside other open source and public resources for analysing and managing the LC-MS/MS component of the DMA of *D. magna* – without these newly developed tools the analysis of the DMA of *D. magna* would not have been possible.

Finally, in Chapter 7, I conclude on the progress made against these specific challenges and suggest further research.

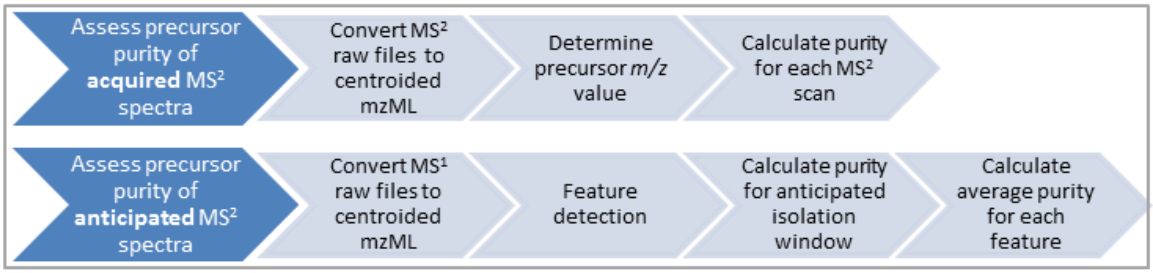
Chapter 2

msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry Based Fragmentation in Metabolomics

2.1 CHAPTER SUMMARY

Tandem mass spectrometry (MS/MS or MS^2) is a widely used approach for structural annotation and identification of metabolites in complex biological samples. The importance of assessing the contribution of the precursor ion within an isolation window for MS^2 experiments has been previously detailed in proteomics, where precursor ion purity influences the quality and accuracy of matching to mass spectral libraries. However, there has been little attention to this data-processing technique in metabolomics – where the impact of precursor ion purity is arguably more important in interpreting downstream identification and annotation of compounds. Here I present msPurity, a vendor-independent R package for liquid chromatography (LC) and direct infusion (DI) MS^2 that calculates a simple metric to describe the contribution of the selected precursor (see my contributions in section 2.2). The precursor purity metric is calculated as "intensity of a selected precursor divided by the summed intensity of the isolation window" and is interpolated at the recorded point of MS^2 acquisition using bordering full-scan spectra. Additionally, I remove low abundance peaks that are believed to have limited contribution to the resulting MS^2 spectra, isotopes are optionally removed the

isolation efficiency of the mass spectrometer can be used to normalise the intensity relative to the position of the ion within the window. The approach described here therefore allows for the most comprehensive assessment of the purity of a selected precursor ion in both metabolomics and proteomics analysis. The package was applied to Data Dependent Acquisition (DDA)-based MS^2 metabolomics datasets derived from three metabolomics data repositories. For the ten LC-MS^2 DDA datasets with > ± 1 Da isolation windows the median precursor purity score ranged from 0.67 to 0.96 (scale = 0 to +1). The R package was also used to assess precursor purity of theoretical isolation windows from LC-MS datasets of differing sample types. The theoretical isolation windows being the same width used for an anticipated DDA experiment (± 0.5 Da). The most complex sample had a median precursor purity score of 0.46 for the 64,498 XCMS determined features, in comparison to the less spectrally complex sample that had a purity score of 0.66 for 5,071 XCMS features. It has been previously reported in proteomics that a purity score of < 0.5 can produce unreliable spectra matching results, with this assumption I show that for complex samples there will be a substantial number of metabolites where traditional DDA approaches will struggle to provide reliable annotations or accurate matches to mass spectral libraries.



2.2 CONTRIBUTIONS

Thomas N. Lawson[1], Ralf J. M. Weber[1], Martin R. Jones[1], Andrew J. Chetwynd[1,2], Giovanny Rodríguez-Blanco[2], Riccardo Di Guida[1], Mark R. Viant[1,2] and Warwick B. Dunn*[1,2]

[1]School of Biosciences and [2]Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

TNL developed the metric and associated software, was lead in writing the manuscript, performed all data analysis, planned the experimental design for all isolation window experiments, and experimentally acquired data for the Thermo Fisher Q-Exactive isolation window experiments. RJMW advised on the software development. MRJ extracted metabolites and experimentally acquired data on the mass spectrometer for the *Daphnia magna* samples. AJC experimentally acquired data for the Waters Xevo G2 XS QTOF mass spectrometry isolation experiments. GRB advised and aided in the Thermo Fisher Q-Exactive isolation windows mass spectrometry experiments. RDG extracted metabolites and experimentally acquired data on the mass spectrometer for the *Ovis aries* samples. MRV and WBJ supervised and advised in all aspects of the work. All authors read and approved the final published manuscript.

2.3 PUBLICATION HISTORY

This chapter was originally published in Analytical Chemistry (Lawson et al., 2017). Section 2.5.6 was originally in the supplemental text but has been added into the main text of my thesis for clarity. Section 2.5.9 was omitted from the paper due to space but was added into the thesis for clarity and reference. Throughout the chapter grammatical corrections, updated references to sections and typos have been updated as required. Additionally, the use of the plural “we” in the original paper has been replaced with “I” when describing work that I performed.

2.4 INTRODUCTION

Mass spectrometry (MS) is routinely used to (semi)-quantify, annotate and identify small molecules (metabolites) in complex biological matrices. An MS experimental workflow can consist of infusing a sample directly into a mass spectrometer without any prior chromatographic separation (direct infusion mass spectrometry; DIMS) (Kirwan et al., 2014) but more often the sample components are spatially separated via either gas or liquid chromatography (LC) (Dunn et al., 2011). The predictable mass fragmentation patterns observed from the electron ionisation (EI) technique commonly used with gas chromatography allows for reliable matching to mass spectral libraries such as NIST (Lemmon et al., 2011) and the Golm Metabolome Database (Hummel et al., 2007). So-called soft ionisation techniques, including electrospray ionisation (ESI), used in conjunction with LC methods provide minimal fragmentation because of the ionisation process. In these cases, gas-phase fragmentation can be applied through collision-induced dissociation in hybrid quadrupole, ion trap, quadrupole time-of-flight (Q-TOF) or Orbitrap systems. The term 'tandem mass spectrometry' (MS/MS or MS^2) is used when a single collision step is used but product ions can be isolated for further collision to provide MS^n spectra where $n \geq 3$. The focus of this paper is on fragmentation resulting from tandem mass spectrometry.

A key component of any MS^2 (or higher) technology is the isolation of selected *m/z* windows for gas-phase fragmentation and the mapping back of the fragmentation (product) spectrum to the selected *m/z* window. In targeted and data-dependent acquisition (DDA) based experiments, where MS^2 is performed on a dynamic list of precursor ions, as often determined by a preceding MS full scan, an isolation window is centred on the targeted precursor peak (*m/z* value). However, the isolation window can often contain more than one distinct peak, fragmentation spectra resulting from these situations being termed "chimeric" (Houel et al., 2010) and can be problematic for interpretation of the spectra and mass spectral library searching.

The more conservative estimates of chimeric spectra in proteomic DDA-based research are calculated as over 11.2% of all MS^2 spectra acquired (Hoopmann et al., 2007), but others have calculated this to be as high as 50%, or more (Houel et al., 2010; Luethy et al., 2008). A direct comparison between different studies is difficult though as the number of chimeric spectra will vary considerably based upon sample complexity, instrumental configuration and the procedure by which spectral chimerism is calculated.

The impact of chimeric spectra on spectral matching and annotation depends on the purity of the isolation window fragmented (i.e. the ratio between the relative intensity of the precursor divided by the summed intensity of all ions within the isolation window). If the purity of the precursor ion is sufficiently low, it can often be difficult to determine the origin of resulting product ion(s). This in turn can lead to erroneous spectral matching results or no spectral matches.

Deconvolution of chimeric spectra however is possible and forms the basis of the data-analysis procedures applied to data independent acquisition (DIA) experiments. Where previous approaches in proteomics relied on comparison of fragmentation spectra to in silico-generated peptide databases (Li et al., 2009) alternative approaches have been derived from selected reaction monitoring data analysis methods (MS-SWATH, OpenSWATH (Gillet et al., 2012; Röst et al., 2014)). Although not as prevalent as DDA-based analyses, DIA-based approaches are also available in the field of metabolomics (MS-DIAL (Tsugawa et al., 2015)).

A metabolomics analytical and data analysis workflow that directly takes into consideration the purity of an isolation window has also been developed, enabling deconvolution of MS^2 spectra (Nikolskiy et al., 2013). The approach, demonstrated using an Agilent 6520 Q-TOF instrument, requires sliding isolation windows to be acquired surrounding the precursor of interest.

However, standard DDA-based approaches are still widely used and popular in metabolomics

(Beisken et al., 2014a; Chen et al., 2015; Longnecker et al., 2015) and the deconvolution techniques detailed above are not always applicable to standard DDA based acquisition techniques. In these cases, simply assessing the targeted precursor purity can be useful in interpreting the MS^2 spectra and aid in assessing the reliability of any subsequent annotation.

Some of the first software that had the functionality to count the number of chimeric spectra include the C/C++ software "hardklör" (Hoopmann et al., 2007) and the python software "ChimeraCounter" (Hoopmann et al., 2007). ChimeraCounter uses the "percent chimera intensity (PCI)" metric to count chimeric spectra when peaks that are not isotopes of the targeted precursor are within the isolation window and are above a user-defined percentage of the peak height of the precursor. Another metric has been previously described to calculate the "Precursor Ion Fragment" (PIF) (Michalski et al., 2011). This is calculated as the targeted precursor ion intensity divided by the overall intensity in the isolation window and essentially gives a score of the purity of the precursor at the closest full scan (MS^1) to the MS^2 scan. The PIF implementation is provided within the quantitative proteomics software package MaxQuant (Tyanova et al., 2015). All the software previously developed for assessing chimeric spectra has been focused on proteomics data and no such published attempts have been shown to incorporate the isolation efficiency of the instrument or consolidate the functionality of all the approaches into a single piece of software.

An automated computational method is detailed here to calculate the purity of fragmentation spectra post-acquisition from either a LC-MS^2 or $DIMS^2$ metabolomics experiment. What I call here "precursor purity" is calculated with a revised Michalski approach (Michalski et al., 2011). Advances include that the metric is interpolated at the recorded time of the MS^2 acquisition using bordering full scan spectra, the isolation efficiency of the mass spectrometer can be included within the calculation and, as per the PCI approach, isotopic peaks of the targeted precursor can be removed and peaks with intensities below a minimum percentage of the

precursor peak intensity are removed from the calculation. The software has been applied to investigate 12 DDA and one DIA metabolomics datasets for different biological samples retrieved from the data repositories MetaboLights (Haug et al., 2013), Metabolomics Workbench (Sud et al., 2015) and PRIMe Data Resource of Plant Metabolomics (DROP Met) (http://prime.psc.riken.jp/). I also detail how theoretical isolation windows can be assessed using MS^1 datasets collected independent of MS^2 acquisitions. The computational methods detailed in this paper are available in the R package "msPurity". The package has been developed to work as a standalone or to be used in conjunction with the metabolomics peak detection and processing R package: XCMS (Smith et al., 2006; Tautenhahn et al., 2008).

2.5 MATERIALS AND METHODS

2.5.1 OVERVIEW OF SOFTWARE

An R package has been developed to assess the contribution of the targeted precursor in a fragmentation isolation window using a metric called "precursor purity". The use cases of the package can be divided into two types:

1. **Assessing precursor purity of previously acquired MS^2 spectra:** A user has acquired either LC-MS2 or DIMS2 spectra and an assessment is made of the precursor purity for each MS^2 scan.
2. **Assessing precursor purity of anticipated isolation windows for MS^2 spectra:** A user has acquired either LC-MS or DIMS full scan (MS^1) data and an assessment is to be made of the precursor purity of detected features using anticipated or theoretical isolation windows. This information can then be used to guide further targeted MS^2 experiments.

2.5.2 AVAILABILITY

The msPurity package is available for Windows (both 32 and 64 bit), Mac OS, and Linux from the Bioinformatics software repository for R: Bioconductor ≥ 3.4 (2016). The code can also be downloaded directly from Github: https://github.com/computational-metabolomics/msPurity.

2.5.3 FILE PREPARATION

The input for the msPurity R package requires the raw vendor format of the mass spectrometry data to be converted into the mzML file format (Martens et al., 2011). The controlled vocabulary used for the mzML format allows the isolation window widths used by the instrument to be parsed out of the mzML file from either Thermo, Agilent, AB Sciex or Waters instruments. The conversion of instrument raw files to mzML can be achieved with the Proteowizard conversion tool MSconvert (Chambers et al., 2012) or RawConvertor (He et al., 2015).

If spectra were not collected in centroid mode it is essential that the peaks are converted from profile to centroid format before processing within msPurity.

2.5.4 ASSESSING PURITY OF ACQUIRED MS^2 SPECTRA

To assess the precursor purity of an MS^2 dataset, the m/z value of the precursor ion associated with each MS^2 spectrum must first be determined. With the assumption that the most abundant peak in the window will have the highest contribution to the fragmentation spectrum, the most abundant peak within the isolation window is used to define as the precursor *m/z* value. This is determined for the MS^1 scan closest to the MS^2 scan.

The precursor purity score of the targeted precursor ion is then calculated directly from the mzML spectra of the MS^1 scan as "target precursor intensity / summed intensity in isolation window". This simple ratio provides an easy to interpret score bounded between 0 and 1. A value of 1 should signify only the selected precursor has contributed to the resulting spectra and values closer to 0 should signify that there is little or no contribution from the selected pre-

cursor ion to the summed intensity in the isolation window. It should be noted that this approach is unable to distinguish isobaric species.

It is not possible to calculate the score at the exact time of the MS^2 acquisition as the MS^1 spectra is typically not acquired simultaneously for DDA approaches. To account for this the precursor purity is calculated using two MS^1 scans, one prior to and one post MS^2 acquisition, then interpolated at the point of MS^2 acquisition (see Figure 2-1).

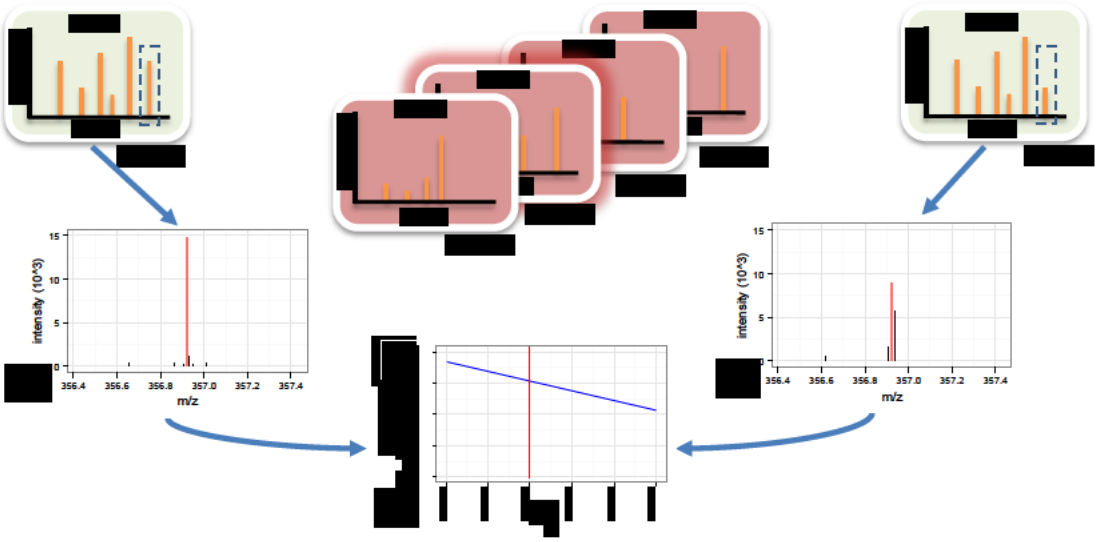


Figure 2-1 Assessment of acquired MS^2 precursor purity for a standard DDA based experiment.

Scan 1 and 6 are MS^1 scans and scans 2 to 5 are MS^2 scans for different selected precursor ions where a DDA top 4 method (fragment the top 4 most intense ions) has been applied. The example in the figure calculates the precursor purity for scan 3. The precursor purity is calculated prior to the MS^2 acquisition (1) and after MS^2 acquisition (2) and the final score is then interpolated at the time of the MS^2 acquisition using bordering full-scan spectra (3).

2.5.5 ASSESSING PURITY OF ANTICIPATED ISOLATION WINDOWS FOR MS^2 SPECTRA

The precursor purity calculation is performed using information derived solely from the MS^1 spectra, therefore the calculation can be performed independently from any MS^2 acquisition.

The steps for assessing the precursor ion purity of anticipated MS^2 spectra using an LC-MS dataset are as follows:

1. Feature detection is carried out within XCMS
2. The full width half maximum (FWHM) of the chromatographic peak is calculated for each XCMS feature
3. The median precursor purity score (intensity of target / summed intensity in anticipated isolation window) is then calculated for the FWHM of the chromatographic peak of each feature (See Figure 2-2). The median score is then calculated across any chosen samples or technical replicates.

The steps for assessing the precursor purity of anticipated MS^2 spectra using DIMS datasets consists of the following:

1. Basic processing centroid peaks performed within msPurity. This involves filtering out peaks below a user determined signal to noise level, averaging between multiple scans, filtering peaks not present in a minimum number of scans, filtering of peaks where the relative standard deviation of the intensity is below a given threshold and removal of blank peaks from the sample peaklist.
2. The median precursor purity score is determined for each feature across all scans in the dataset.

The predicted purity scores can be used to select features for MS^2 in a targeted or semi-targeted experiment e.g. using a "nearline" approach (Neumann et al., 2013). Additionally, the information can be used to assess what isolation window width would be suitable for any

subsequent MS^2 experiments.

2.5.6 REMOVAL OF ISOTOPIC PEAKS

This section was originally in the supplemental of the published msPurity paper (Lawson et al., 2017).

Peaks are removed from the precursor purity calculation by default in two use cases:

1. The targeted precursor is the ^{12}C peak of a ^{13}C peak within the isolation window
2. The targeted precursor is the ^{13}C peak of a ^{12}C peak within the isolation window.

In both cases, either the ^{12}C or the ^{13}C peak that is not the targeted precursor is omitted from the summed total of intensity used for the precursor purity calculation. Isotopic peaks are determined within an isolation window as follows:

First, the hypothetical masses of the targeted precursor isotopic peak are calculated. See Equation (2.1) for use case 1 and Equation (2.2) for use case 2 where mz_{prec} is the *m/z* value for the targeted precursor and $charge$ is an integer representing a hypothetical charge state of the ion.

$$mz_{iso13} = mz_{prec} + \frac{isotopic\ mass\ difference}{charge} \qquad (2.1)$$

$$mz_{iso12} = mz_{prec} - \frac{isotopic\ mass\ difference}{charge} \qquad (2.2)$$

The parts per million (ppm) difference between the hypothetical isotopic peak and the contaminating peaks are then calculated. See Equation (2.3) where mz_{cont} is the *m/z* value of a contaminating peak within the isolation window and mz_{iso} is either mz_{iso13} or mz_{iso12}.

$$ppm = \left| \frac{(10^6)(mz_{cont} - mz_{iso})}{mz_{iso}} \right| \qquad (2.3)$$

Any contaminating peaks less than a user defined ppm tolerance are then checked to see if the intensity ratios correspond to expected isotopic natural abundances. The calculation follows a

similar approach as the CAMERA annotation software(Kuhl et al., 2012). As the molecular formula is not known for the peak of interest a rough range is calculated based on the number of carbon atoms that could theoretically be present, see Equation (2.4) where mz_{prec} is the *m/z* value of the targeted precursor and a is rounded down to the nearest integer, $relative\ atomic\ mass$ rounded to 12 for carbon.

$$a = \frac{mz_{prec}}{relative\ atomic\ mass} \tag{2.4}$$

For use case 1, where a contaminating peak is thought to be the ^{13}C isotope, the intensity of the contaminated peak should be between the range calculated using Equation (2.5) and (2.6) where I_{prec} is the intensity of the precursor, $diff$ is the natural abundance of the isotope (1.07 for ^{13}C)).

$$I_{max} = \frac{(a \times diff)}{100} \times I_{prec} \tag{2.5}$$

$$I_{min} = \frac{(1 \times diff)}{100} \times I_{prec} \tag{2.6}$$

For use case 2, where a contaminating peak is thought to be the ^{12}C isotope, the intensity of the contaminated peak should be between the range calculated using Equation (2.7) and (2.8).

$$I_{max} = \frac{I_{prec}}{(1 \times diff)} \times 100 \tag{2.7}$$

$$I_{min} = \frac{I_{prec}}{(a \times diff)} \times 100 \tag{2.8}$$

It should be noted here that whether isotopic peaks are to be removed should depend on the downstream analysis to be performed, for consistency, no results shown in this chapter omit the isotopic peaks from the purity calculations.

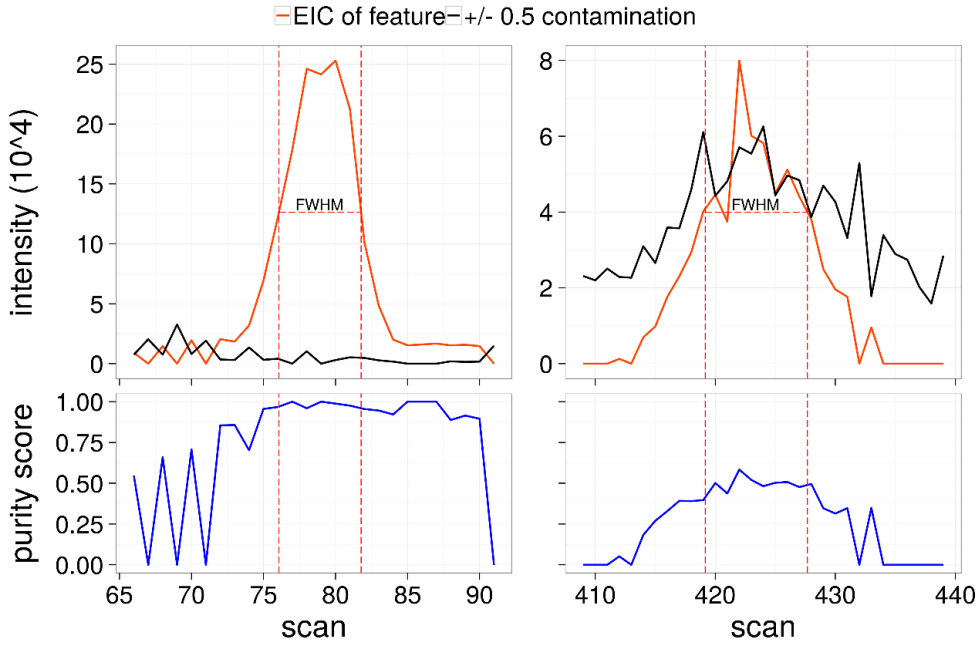


Figure 2-2 Assessing precursor purity of anticipated isolation windows for two XCMS features derived from an MS^2 dataset.

The extracted ion chromatogram (EIC) of each feature (red) and the intensity of contaminating peaks (blue) within a ± 0.5 Da window. Scan range determined by XCMS and approximate full width half maximum FWHM highlighted with red vertical lines.

2.5.7 REMOVAL OF PEAKS OF LOW RELATIVE ABUNDANCE

When using larger isolation windows there can often be numerous peaks that have a low intensity relative to the targeted precursor peak and that have negligible impact on the product ion fragmentation spectrum. When these peaks are included in the purity calculation a lower precursor purity score can be calculated, which is misleading. To accommodate for this, only peaks that have an intensity greater than 5% of the targeted precursor ion peak are included in the precursor purity calculation (the threshold of 5% can be adjusted by the user). A similar approach is used in the PCI calculation (Houel et al., 2010). The removal of these peak can be considered a noise filter – however it should be noted that a high threshold could artificially

remove real contaminating peaks that do contribute the fragmentation spectra. If the analyst suspects that these lower intense peaks are still contributing to the fragmentation spectra – a cautious approach would be lower the threshold to zero.

To demonstrate the effect of increasing the threshold a subset of five files were taken from the public dataset MTBLS144 and were processed with msPurity – using thresholds of 0, 2, 5, 10, 20, 25, 50, 75 and 100%. The following mzML files were used for the assessment: Mtab_FT_012611_11.mzML, Mtab_FT_012611_07.mzML, Mtab_FT_012611_08.mzML, Mtab_FT_012611_09.mzML and Mtab_FT_012611_10.mzML

2.5.8 ISOLATION EFFICIENCY

Isolation efficiency is defined here as the effect of the position of an ion within an isolation window on its relative intensity in corresponding fragmentation spectra. When the isolation efficiency of an instrument is known, the peak intensities within an isolation window can be normalised for the precursor purity calculation.

The isolation efficiency can be estimated experimentally by acquiring MS^2 spectra where the precursor ion of interest is experimentally measured at various locations within the isolation window. The precursor ion peak, or a known product ion peak, can then be monitored across the isolation window. The resulting intensity profile is then converted to be within a scale of 0 to 1, where 1 is the position in the isolation window with highest response and 0 is where the precursor ion is not detected. This efficiency profile can then be used to normalise the peak intensities based on their position within the isolation window.

For this study, the isolation efficiency experiment was performed using a Thermo Scientific Q Exactive Focus mass spectrometer operated with a ± 0.5 Da isolation window. Isolated ions had m/z values of 195.0875, 524.2649 and 922.0086 and were generated from constituents of a Pierce LTQ Velos ESI positive ion calibration solution (product number #88323). The mean

isolation efficiency profile was then used to create a simple linear model using B-spline polynomials (5 anchors) to predict the isolation efficiency based on isolation window position (adjusted R^2: 0.9812, F: 210, p-value < 0.001), see Figure 2-3.

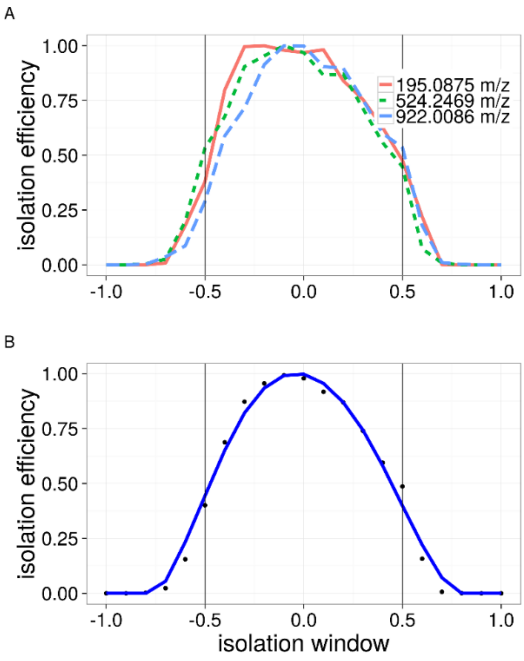


Figure 2-3 Overview of isolation efficiency experimental determination for Thermo Scientific Q Exactive Focus

Overview of isolation efficiency experimental determination for Thermo Scientific Q Exactive Focus. **A:** Isolation efficiency profile of Thermo Scientific Q Exactive Focus mass spectrometer on the following ions: 195.0875, 524.2649 and 922.0086 *m/z*. **B:** A simple linear model using B-spline polynomials to predict isolation efficiency based on isolation window position for the Thermo Scientific Q Exactive Focus mass spectrometer (negative values have been zeroed)

In addition, an assessment of isolation windows for Waters instruments was also carried out using a Xevo G2 XS QTOF mass spectrometer tuned to a mass resolution of 40,000 and run in positive ESI mode. When operated in MS^2 mode the quadrupole low mass resolution was manually tuned to 23 providing a ± 0.5 Da isolation window. Isolation windows surrounding ions at m/z 269.20, 362.95 and 906.81 from a solution of 0.5 mM sodium formate in 90:10 iso-propanol:water was assessed. The mean isolation efficiency profile observed is similar in shape

to a Gaussian curve (see Appendix 8.1).

2.5.9 FORMAL DESCRIPTION OF PRECURSOR ION PURITY CALCULATIONS

This section was absent from the published msPurity paper but has been added here in the thesis for reference.

The first step is to remove isotopic peaks that are to be ignored based on Equations (2.1) to (2.8). The following step is to adjust the *m/z* values to be relative to the center of the isolation window. See Equation (2.9), where **x** is a vector of the *m/z* values within an isolation window, w is the *m/z* value at the centre of the isolation window. A vector of normalisation values is then calculated based on the relative *m/z* values, see Equation (2.10) where $IsoNorm$ is a function to provide the isolation window normalisation. The intensity values within the isolation window can then be normalised based on their position within the isolation window, see Equation (2.11).

$$\mathbf{x}_{relative} = \mathbf{x} - w \tag{2.9}$$

$$\mathbf{n} = IsoNorm(\mathbf{x}_{relative}) \tag{2.10}$$

$$\mathbf{y}_{norm} = \mathbf{y} \cdot \mathbf{n} \tag{2.11}$$

Peaks below the relative abundance threshold from the precursor peak of interest are then removed. See Equation 2.12 where $thres$ is the relative abundance threshold and i is the position of the precursor peak of interest – default to 0.05 (e.g. 5%). The threshold used is detailed in more detail in section 2.5.7.

$$\mathbf{y}_{sml} = \mathbf{y}_{norm} > y_{norm_i} \cdot thres \tag{2.12}$$

The normalised intensities then are used to obtain the precursor ion purity for this isolation window. See Equation (2.13) where

N is the length of vector $\mathbf{y}_{sml}$.

$$p = \frac{y_{norm_i}}{\sum_{j=1}^{N} y_{sml_j}} \tag{2.13}$$

If the precursor ion purity is being calculated for MS^2 data that has been acquired, a linear interpolation can be used to obtain the precursor ion purity at the point of the MS^2 acquisition. See Equation (2.14) where p_{pre} is the precursor ion purity score prior to the MS^2 scan, p_{post} is the precursor ion purity score post the MS^2 scan, t_{pre} is the time prior to the MS scan, t_{post} the time post MS^2 scan.

$$p_{aquired} = p_{pre} + \frac{(t_{aquired} - t_{pre})(p_{post} - p_{pre})}{t_{post} - t_{pre}} \tag{2.14}$$

2.5.10 PUBLIC METABOLOMIC DATASETS

A total of 12 DDA and one DIA datasets were obtained from the public data repositories MetaboLights (Haug et al., 2013), Metabolomics Workbench (Sud et al., 2015) and PRIMe Data Resource of Plant Metabolomics (DROP Met) (http://prime.psc.riken.jp/) to be assessed for their precursor purity. See Appendix 8.2 for a summary of the studies. The limit of the 13 datasets is because, as of 31st December 2015, this was an exhaustive list of DDA and DIA datasets that include either the RAW, mzML or other XML open source data formats. Further details of the experimental setup for each study can be found in the corresponding submission entry on the relevant data repository website and within the associated papers for each study.

If not already in mzML centroided data format, the files were converted using Proteowizard MSconvert (v3.0) (Chambers et al., 2012). The purity assessments were made for all MS^2 scans for every sample using the msPurity package using the function "purityA" with standard parameter settings. Importantly though, the precursor purity was calculated for all studies using the most intense peak within the isolation window as the target precursor ion, isotopic peaks

were not removed, and no isolation efficiency normalisation was performed. The efficiency normalisation was omitted from the public data as it was not possible to retrospectively produce comparable isolation efficiency profiles for the multiple instruments used.

If possible, the isolation window widths were extracted from the mzML file. However, for datasets derived from the Agilent Q-TOF the isolation window widths are not included in the mzML file. For Agilent Q-TOF instrumentation the isolation windows are determined as either narrow (≈ 1.3 amu), medium (≈ 4 amu) or large (≈ 9 amu), the exact widths changing depending on the instrument. In the case of the Agilent datasets used here (ST000075 and ST000085), a narrow resolution of ≈ 1.3 amu was used. For these files, an approximate isolation window of ± 0.65 Da centred on the precursor ion m/z was used for the purity assessments. For Agilent Q-TOF instruments, the medium and large window widths are not currently supported within msPurity as these larger isolation window widths are not evenly centred on the recorded precursor *m/z* in the mzML file.

2.5.11 IN HOUSE METABOLOMIC DATASETS

Two LC-MS metabolomics datasets were acquired to assess XCMS-derived metabolite features for precursor purity.

Daphnia magna extract (using 2.5:1 v/v methanol:water) was separated using a reversed-phase Syncronis Phenyl LC column and mass spectral detection of the LC column eluates was performed using a Thermo Scientific Q Exactive mass spectrometer equipped with H-ESI II source acquiring in positive ionisation mode.

Additionally, *Ovis aries* (sheep) kidney tissue extract (using 70/30 methanol/water extraction solvent) was separated using Hypersil Gold C18 reversed phase column and analysed using a Thermo Scientific Q Exactive mass spectrometer. Feature detection was carried out using the XCMS package (Smith et al., 2006; Tautenhahn et al., 2008) with the following parameters:

peak picking algorithm = centWave, snthresh = 5, ppm =5, prefilter = c(3, 5000), integrate = 1, mzdiff = 0.001. Peaks were aligned by retention time using the retcor.loess algorithm and grouped using the group.density algorithm with the following settings: bw = 5, mzwid = 0.01 and minfrac = 0.5. The msPurity function "purityX" was run on the xcmsSet objects to assess the precursor purity of theoretical isolation windows of ± 0.5 Da for XCMS features using the isolation efficiency normalisation predicted for the Thermo Fisher Q Exactive mass spectrometer. See Appendix 8.2 for a more extensive description of the materials and methods for the in-house metabolomics datasets.

2.6 RESULTS AND DISCUSSION

2.6.1 ASSESSMENT OF PRECURSOR ION PURITY FOR PUBLICLY AVAILABLE DDA AND DIA DATASETS

Figure 2-4 and Table 2-1 summarise the purity of available LC-MS and DIMS based DDA datasets and a single LC-MS DIA dataset.

A direct comparison of precursor ion purity between the different datasets is difficult as the score will vary considerably based on the complexity of the sample, instrument used, isolation window width and the exact details of how the DDA was performed (e.g. if blank exclusion was applied, length of time on dynamic exclusion and number of MS^2 scans per MS^1 scan (top *n* method)). However, by using this wide range of studies a broad overview can be shown of the extent of precursor ion purity in metabolomics.

For nearly all the LC-MS DDA-based studies where the isolation window width was ≤ ± 1 Da the median pre-cursor purity score was > 0.7, the exception is study MTBLS214 where the median precursor purity was 0.67. One factor contributing to the lower score will be the larger number of consecutive MS^2 scans (up to 26) being per-formed between MS^1 scans, whereas for all other DDA studies the maximum number of consecutive MS^2 was 12. The substantial

number of MS^2 scans acquired consecutively can lead to lower intensity precursor ions selected for fragmentation that would typically be ignored with a smaller number of consecutive MS^2 scans. This is demonstrated in Figure 2-4C, where the purity of the MS^2 spectra decreases with the distance from the preceding MS^1 scan. The study MTBLS214 also highlights the need for caution when the number of consecutive MS^2 scans is high. For this study, the nearest MS^1 scan used to calculate the purity score was up to ≈ 5.2 seconds (12 scans) away from the MS^2 acquisition.

Table 2-1: Summary of public metabolomics datasets used to assess precursor purity

MS^2 method	Study	Species	Column	MS Vendor	MS Instrument	Isolation width (Da)	Precursor Purity score		
							Min	Median	Max
DDA	MTBLS26	*M. musculus*	Hypersil GOLD C18 HPLC	Thermo Scientific	LTQ-FT	± 1	0.00	0.79	1.00
DDA	MTBLS36	*S. lycopersicum*	Waters Acquity UPLC	Thermo Scientific	LTQ Orbitrap Velos	± 1	0.14	0.96	1.00
DDA	MTBLS144	*T. pseudonana*	Synergi Fusion reversed phase	Thermo Scientific	LTQ-FT	± 1.5	0.00	0.82	1.00
DDA	MTBLS154	*T. pseudonana*	Synergi Fusion reversed phase	Thermo Scientific	LTQ FT Ultra	± 1	0.00	0.84	1.00
DDA	MTBLS155	*S. elongatus*	Synergi Fusion reversed phase	Thermo Scientific	LTQ FT Ultra	± 1	0.00	0.79	1.00
DDA	MTBLS157	*R pomeroyi DSS-3*	Synergi Fusion reversed phase	Thermo Scientific	LTQ FT Ultra	± 1	0.00	0.78	1.00
DDA	MTBLS214	*H. sapiens*	Promine-nce UFLC (Shimadzu)	AB Sciex	TripleTOF 5600+	± 0.5	0.00	0.67	1.00
DDA	ST000075	*M. musculus*	Waters Acquity CSH C18 / Waters Acquity VanGuard CSH C18	Agilent	6530 QTOF	± 0.65	0.00	0.83	1.00
DDA	ST000082	*H. sapiens*	Waters Acquity CSH C18 / Waters Acquity VanGuard CSH C18	Agilent	6530 QTOF	± 0.65	0.24	0.96	1.00
DDA	ST000098	*C. elegans*	Waters Acquity UPLC	Thermo Scientific	Q Exactive	± 2	0.00	0.49	1.00
DI-MS DDA	MTBLS117	*G. max*	NA	Thermo Scientific	LTQ Orbitrap Classic	± 1	0.03	0.60	1.00
DIA	MS-DIAL: DIA	*Chlamydomonas / Chrollera*	HILIC	AB Sciex	TripleTOF 5600+	≈ ± 10	0.00	0.09	1.00
DDA	MS-DIAL: DDA	*Chlamydomonas / Chrollera*	HILIC	AB Sciex	TripleTOF 5600+	± 0.5	0.00	0.96	1.00

Footnote: See Table S1 for more information regarding each study. References: MTBLS26 (Chitraju et al., 2012), MTBLS36 (Beisken et al., 2014a), MTBLS117 (Aliferis et al., 2014), MTBLS144 (Longnecker et al., 2015), MTBLS154 (Kujawinski et al., 2017), MTBLS155 (Fiore et al., 2015), MTBLS157 (Johnson et al., 2016), MTBLS214 (Chen et al., 2015), MS-DIAL: DIA (Tsugawa et al., 2015) and MS-DIAL: DDA (Tsugawa et al., 2015).

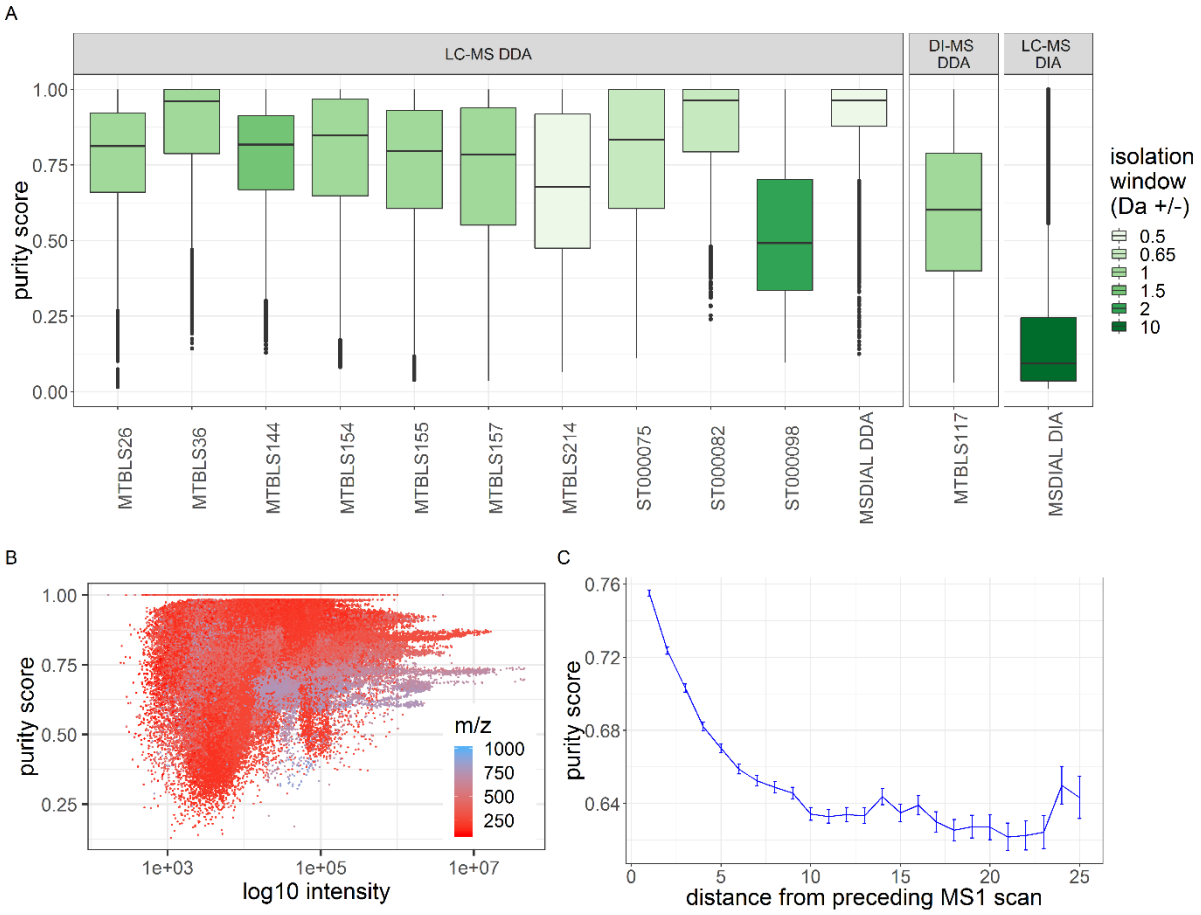


Figure 2-4: Overview of assessment of precursor purity for public DDA and DIA metabolomics datasets

A: Boxplots of calculated purity scores of all fragmentation spectra. See Appendix 8.4 for violin plots derived from the same data. **B** Summary of purity score relationship with selected precursor intensity and m/z value for study MTBLS144, *x* axis is in $\log_{10}$ scale. **C:** The mean purity relationship with the distance from the preceding MS1 scan for study MTBLS214. Error bars represent one standard error of the mean. In cases where the selected precursor value could not be found within the isolation window the purity score will be zero. These have been removed in this plot for clarity. See Appendix 8.5 for plot including zero values.

The precursor purity scores reported here are in general agreement with the field of proteomics, where the median precursor purity score (or PIF, within this field) of targeted peptides was calculated as 0.85 and 0.73 for isolation windows of ± 0.5 Da and ± 1 Da respectively (Michalski et al., 2011).

Data were acquired in both positive and negative ionisation methods for 11 of the public datasets and allowed a comparison between the precursor ion purity scores derived from positive and negative methods. In general, the positive ionisation data had higher precursor purity scores than negative ionisation data (mean difference of + 0.04) but this was not always the case. For 7 out of the 11 datasets the purity score was significantly higher for the positive ionisation mode (unpaired Wilcoxon rank sum adjusted p-value using Benjamini & Hochberg approach < 0.0001) but for 3 out of the 11 datasets the purity score was significantly smaller for the positive ionisation mode (unpaired Wilcoxon rank sum adjusted p-value using Benjamini & Hochberg approach < 0.0001) see Appendix 8.6 for further details.

It should be noted that all of the scores presented in Figure 2-4 and Table 2-1 used the most intense peak within the isolation window for the target precursor peak. When the peak that is at the centre of the isolation window is chosen for the calculation then the precursor purity scores were significantly lower for all studies (mean difference for all studies: - 0.04, paired Wilcoxon rank sum adjusted p-value using Benjamini & Hochberg approach < 0.0001). For the smaller isolation windows, the impact is marginal but for the larger isolation windows observed for DIA experiments this has a much larger impact on the purity score reported (- 0.16 difference in purity). See Appendix 8.7 for full details. The Raw Convertor (He et al., 2015) tool also uses the most intense peak in the isolation window as the precursor for the same assumption used here, the most abundant peak within the isolation window will contribute most to the fragmentation spectra.

For the only public DIA dataset available, the median purity score is < 0.1. The lower purity score for the DIA dataset is expected due to the much larger isolation window width. As the principle behind DIA is to deconvolute spectra derived from these large isolation windows, the purity score for DIA based datasets is in some regards in-consequential. Nevertheless, this dataset provides a good example to demonstrate that small precursor ion purities are observed when very large isolation window widths are employed.

For the one public DIMS DDA dataset analysed the median precursor purity score was 0.54. A lower precursor purity score was expected due to the lack of chromatography and therefore more complex MS^1 scans as well as a longer period to acquire MS^2 spectra. This leads to a higher number of low intensity precursor ions being fragmented that typically have lower precursor purity scores. It should be acknowledged here that the isobaric overlap from DIMS would also typically be higher than LC-MS if no prior fractionation has been performed.

These assessments confirm that the precursor ion purity calculation can be performed on various types of MS^2 experiment (DDA, DIA, LC-MS and DIMS), from several instrument manufactures (Thermo Scientific, AB Sciex and Agilent), and can provide an assessment of "chimeric" spectra in a dataset. If no deconvolution of the chimeric spectra is to be performed, then spectra below a certain precursor purity score can be removed to improve the accuracy of any later spectral matching and annotation.

A plot of the intensity of the selected precursor ion against the precursor purity shown in Figure 2-4B highlights a trend where the most intense precursors will have higher purity (closer to 1) but also shows that many precursors of lower intensity can still have high purity scores. This implies that simply having a higher threshold for the intensity of the precursor ion to be chosen for fragmentation is insufficient in many cases to improve the precursor ion purity of MS^2 spectra.

An acceptable level of precursor purity for DDA or targeted analysis could be at a level at where spectral matching to a library compound provides successful annotation or identification. In the field of proteomics, when the precursor purity score is at a relatively high level (≥ 0.5) the identification rate of proteins stays at nearly 60% (Michalski et al., 2011). Similar rates are difficult to calculate for metabolomics datasets due to the lack of a comprehensive database of metabolites and associated spectra. However, it is reasonable to be cautious of any MS^2 mass spectra where the target precursor ion is thought to have contributed less than half of the ions that have been fragmented. I acknowledge though that the number of other ions present in the isolation window, their relative intensity compared to the chosen precursor ion, and the complexity of the fragmentation mass spectra for each non-precursor ion will also have a significant contribution to the ability to accurately apply the MS^2 spectra for mass spectral matching.

To demonstrate the effect of increasing the relative abundance threshold to the selected precursor (see section 2.5.7) – 5 mzML files from the public dataset MTBLS144 were processed with msPurity – using relative abundance thresholds of 0, 2, 5, 10, 20, 25, 50, 75 and 100%. The average precursor ion purity score was 0.818, 0.822, 0.822, 0.881, 0.933, and 1.00 for thresholds 0, 2, 5, 10, 15 and 20% respectively (the relative standard deviation was between 29.3 and 30.5 for each threshold). Indicating that a threshold of 5% has only has a minor impact on the purity score overall. All the above thresholds tested above 20% have a median purity score of 1 (relative standard deviation ranging from 17.0 to 28.4). As expected, an increase threshold will result in higher precursor ion scores, as the score is being artificially improved by removing real contaminating ions. This highlights the difficulty in choosing a threshold for this parameter - a similar problem is observed generally within metabolomics in choice of signal to noise ratio, where often a value is often chosen based on manual inspection of the data. Therefore, if a user has manually checked the data and believes the lower intensity ions that

have been removed are contributing to fragmentation, the threshold should be adjusted – or set to zero to ignore completely.

2.6.2 ASSESSING PRECURSOR PURITY OF ANTICIPATED MS^2 SPECTRA

The anticipated precursor purity calculated for the FWHM of the chromatographic peak for all XCMS features (following blank subtraction) determined from two in-house LC-MS datasets is summarised in Figure 5. Assessing the precursor ion purity of acquired fragmentation spectra will only produce precursor ion purity scores for the features that were able to be fragmentated in the brief time available for most DDA experiments. By calculating the anticipated (i.e. predicted) precursor ion purity score of all LC-MS XCMS features regardless of any fragmentation allows one to determine the theoretical precursor ion purity of all chromatographic features in a dataset. The resulting anticipated precursor ion purity scores will be affected by how complex the spectra of the biological sample – therefore two sample types were assessed here – one where the spectral complexity features were high (≈ 65,000 XCMS features) and one where the spectral complexity was lower (≈ 5,000 XCMS features).

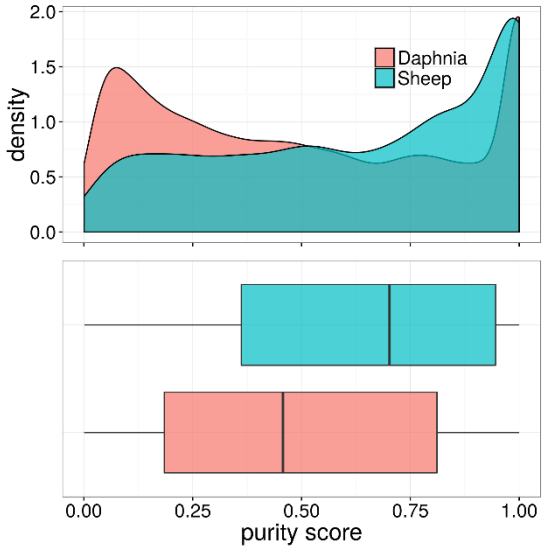


Figure 2-5: Overview of the precursor purity of XCMS features calculated for theoretical isolation windows of an anticipated MS^2 experiment.

Calculated for *D. magna* extracts (highly spectrally complex sample, i.e. ≈ 65,000 XCMS feature) and sheep kidney extracts (less spectrally complex sample, i.e. ≈ 5,000 XCMS features).

The complex *D. magna* sample consisting of 64,498 XCMS features had a median purity score of 0.46. The less spectrally complex sheep kidney (5,071 XCMS features) had a median precursor purity score of 0.66. The number of XCMS features with anticipated precursor purity scores < 0.5 was 53% for the *D. magna* sample and 39% for the sheep kidney sample. If we assume that reliable spectral matching requires a precursor purity score > 0.5, the results demonstrate there would be a considerable number of XCMS features where the precursor purity score is expected to be so low that it will be problematic for standard MS^2 spectral matching. If further MS^2 experiments were to be performed on a targeted group of these features the precursor purity score could then be used to direct the acquisition of XCMS features where reliable MS^2 matching could be performed.

Validation of anticipated score of XCMS feature. To show how stable the anticipated precursor purity scores were for the *D. magna* sample an LC-MS2 DDA experiment was conducted on

the same dataset as the original LC-MS experiment. When there was an XCMS determined feature observed between both the LC-MS and LC-MS^2 datasets it allowed a comparison to be made of both the anticipated (i.e. predicted) precursor purity score and the precursor purity score from the acquired MS^2 data. A total of 2873 features were assessed with a median of four MS^2 spectra used for each feature. The median difference between the anticipated and observed precursor purity score was 0.05.

2.7 CONCLUSION

The msPurity package has been designed to evaluate precursor purity or chimera in metabolomics MS^2 datasets by calculating an easily interpretable metric (intensity of selected precursor / summed intensity of the isolation window) similar to the previously described method used in proteomics (Michalski et al., 2011). However, the method described here differs to the Michalski approach in that I interpolate peak intensity of MS^2 acquisitions using bordering full-scan spectra, I remove low abundant peaks that are believed to have limited contribution to resulting MS^2 spectra and can optionally remove isotopic peaks and consider the isolation efficiency of the mass spectrometer.

Using the msPurity package I present here the first comprehensive characterisation of precursor ion purity or chimera in metabolomics and show that for complex samples a high proportion of metabolite features will have low (< 0.5) precursor ion purity, which could be potentially problematic for standard DDA based spectral matching analysis.

Chapter 3

Look Before You Leap: Using Prior Information to Automatically Guide Mass Spectrometry Fragmentation in Metabolomics

3.1 CHAPTER SUMMARY

Liquid chromatography mass spectrometry (LC-MS) and direct infusion mass spectrometry (DIMS) are commonly used in untargeted metabolomics to generate rich datasets derived from complex biological samples. The reliability of annotation can be improved by performing gas-phase fragmentation mass spectrometry alongside LC-MS or DIMS. The fragmentation acquisition can occur automatically online or following careful manual examination offline to determine metabolite features of interest. In addition, previously published nearline (nearly online) approaches in metabolomics have been used for semi-automated data processing of a prior LC-MS acquisition to direct the fragmentation to features of interest. Here I present a collection of "nearline" based tools and workflows that take advantage of prior LC-MS or DIMS acquisitions to optimise the effectiveness of any subsequent gas phase fragmentation (see my contributions in section 3.2). An adapted version of the previously reported nearline LC-MS workflow has been developed in Galaxy where co-eluting features can be fragmented within the same acquisition run, the same nearline principle was also been applied to DIMS acquisitions and a Galaxy workflow implemented. Additionally, a genetic algorithm was

created for optimizing fractionation of eluent from a chromatographic run based on a prior LC-MS run using the same nearline principle to optimise later fragmentation experiments. Following an assessment of prototype workflows, the LC-MS and DIMS nearline methods were applied to a large-scale metabolome annotation project (deep metabolome annotation of *D. magna,* see Chapter 6). However, the nearline workflows created here still have areas requiring further development (full automation remains difficult without vendor support and many features are still not fragmented) and it is discussed how these issues can potentially be overcome.

3.2 CONTRIBUTIONS

Thomas N. Lawson[1], Martin R. Jones[1], Ralf J. M. Weber[1,2], Mark R. Viant[1,2] and Warwick B. Dunn*[1,2]

[1]School of Biosciences and [2]Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

TNL developed the fractionation optimisation and co-developed the DIMS-nearline and the adapted-LCMS-nearline tools and workflows with RJMW and MRJ. TNL wrote this chapter, performed the assessments of the approaches and co-designed the preliminary experiment to assess the prototype adapted-LCMS-nearline approach. MRJ performed the experimental described in section 3.5.8. MRV and WBJ supervised all work performed.

3.3 PUBLICATION HISTORY

This chapter has not been previously published and is in a draft form where additional work is expected to be added before submission to a journal.

3.4 INTRODUCTION

3.4.1 GENERAL INTRODUCTION TO MASS SPECTROMETRY FRAGMENTATION METABOLOMICS

Mass spectrometry (MS) is a popular analytical technique used to study the metabolomes of complex biological matrices. Infusion of a sample onto a mass spectrometer is usually preceded by separating the sample using either gas or liquid chromatography (LC) (Dunn et al., 2011), however, direct infusion of the sample into the mass spectrometer can also be performed (direct infusion mass spectrometry; DIMS) (Kirwan et al., 2014). When electrospray ionisation (ESI) is used in conjunction with LC or direct infusion methods a gas-phase fragmentation step can be applied to generate product ions of an isolated precursor ion (or more precisely *m/z* window). The product ions can be used for annotation and identification of metabolites by either searching against mass spectral libraries (e.g. MassBank (Horai et al., 2010), METLIN (Smith et al., 2005), mzCloud) or though structural elucidation (MetFrag (Ruttkies et al., 2016), MassFrontier). The focus of this chapter is focused on gas-phase fragmentation used for DIMS or liquid chromatography mass spectrometry (LC-MS).

3.4.2 FRAGMENTATION COVERAGE AND STANDARD ACQUISITION APPROACHES

When performing LC-MS metabolomics on a complex sample the resulting features list determined can be substantial (>10,000 (Lawson et al., 2017)). “Features” defined here as a mass and retention time range (if using LC-MS) that represent a potential metabolite (or partial metabolite e.g. adduct). If there is only a limited sample volume available or there is a limit to instrument time, it is often impossible to obtain fragmentation data on all features.

Approaches used to decide which mass to fragment when performing LC-MS/MS can be determined as “online” (if performed as the instrument is running) or “offline” (if performed on a prior LC-MS run to identify a list of features). The online approach includes data-

dependent acquisition methods (DDA) and data-independent acquisition (DIA) methods. Most LC-MS instrument set-ups will accommodate a DDA approach where a survey MS scan is performed, and a precursor ion is chosen to be fragmented based on its intensity. The precursor ion can then be added to a static or dynamic exclusion list to prevent further fragmentation of the ion during a defined time. Data-independent acquisition does not select a precursor ion but instead fragments all ions within a *m/z* window, a specialised version of this called SWATH-MS (Gillet et al. 2012) has been successfully used for proteomics and more recently been applied to metabolomics (Tsugawa et al., 2015).

The disadvantage of online methods is that they do not target a feature as determined by further downstream data analysis. This can be problematic as there will a substantial number of features that will have low importance (e.g. known blank features and adducts) and should ideally be ignored for fragmentation or given low priority. The problem can be overcome by using an offline method of a prior LC-MS run and then an expert can determine which features should undergo fragmentation. This is also problematic though as if it takes more than a day to perform the offline analysis potentially leading to problems in scheduling available resources. Additionally, there might be a retention time shift of the chromatography column resulting in the feature list not corresponding to the correct chromatographic locations.

3.4.3 DIRECTED ACQUISITION OF FRAGMENTATION SPECTRA (NEARLINE)

Directed acquisition, referred to sometimes as nearline, i.e. in-between offline and online fragmentation acquisition (Neumann et al., 2013), are semi-automated acquisition approaches that can be performed at a speed fast enough so there are minimum shifts in data acquisition and an overall higher efficiency than off line. Features of interests are determined by initial feature detection and processing with XCMS (Smith et al., 2006; Tautenhahn et al., 2008) followed by isotope and adduct annotation by CAMERA (Kuhl et al., 2012). If a comparative

analysis between classes is being performed, the data is further filtered by features that are statistically different between sample groups. The resulting feature list is then scheduled into a minimum set of analytical runs to obtain quality fragmentation spectra on all features. An important consideration was made by Neumann et al. for features that "co-elute", i.e. features with overlapping retention windows, only one feature can have fragmentation acquired during the time of elution. The scheduling problem of this kind is formally called "weighted interval scheduling". The solution used by Neumann et al is based on a popular dynamic programming approach, see Appendix 8.9. The drawback to this nearline approach is that if the list of features is too large the number of analytical methods runs can exceed the limits of sample volume and instrument time.

Here we present a modified nearline approach for LC-MS that takes into consideration co-eluting features and attempts to distribute them across runs to maximise the number of features fragmented. We also rank the features based on adduct type, precursor purity, intensity and adduct cluster size to prioritize the precursors that could be the most informative to fragment.

Additionally, we also provide software to apply the same nearline principles to DIMS.

3.4.4 FRACTIONATION

Fragmentation coverage can also be improved by collecting fractions of eluent from liquid chromatography columns, fragmentation analysis on the fractions can then be performed later. This can allow considerably more time to be spent on each region of the chromatogram compared to a typical LC-MS run. When performing fractionation, LC-MS analysis can be performed prior or simultaneous to the collection providing an insight into the features that have been collected into each fraction. Some of these features in a fraction will have the same or very similar *m/z* value, this can be problematic for the subsequent MS/MS or MS^n analysis as the features will no longer by separated by retention time. This overlap in the fraction will

mean when an ion is isolated for MS/MS or MS^n analysis the ion could have arisen from multiple features. Overlapping features is defined here as features with have the same (or similar *m/z* value) eluting in the same fraction.

I detail here the feasibility of a computational method that analyses data from a prior LC-MS run to determine optimum width of each fraction to minimise overlapping features in associated fractionations.

3.4.5 DEEP METABOLOME ANNOTATION (DMA) APPLICATION

Optimising and creating methods for improving fragmentation and fractionation is of importance to large scale untargeted metabolomics annotation studies such as the DMA of *D. magna*. Some of the workflows described in this chapter were implemented for the current ongoing DMA of D. *magna*.

3.5 METHODS

3.5.1 OVERVIEW OF WORKFLOWS

This chapter covers three data analysis and fragmentation acquisition workflows. Where possible, the data analysis workflows have been implemented within the Galaxy data analysis workflow platform (Goecks et al., 2010).

1. LC-MS/MS nearline acquisition workflow (adapted-LC-MS-nearline):

   Workflow to acquire LC-MS/MS fragmentation spectra using an adapted nearline approach that schedules co-eluting features based on a preceding LC-MS experiment. Features are ranked by priority to have fragmentation acquired. The ranking is based on multiple criteria (e.g. intensity, precursor ion purity, adduct type)

2. DIMS/MS (MS^n) nearline acquisition workflow (nearline-DIMS):

   Workflow to acquire DI-MS/MS and DI-MS^n using a nearline approach based on a preceding DI-MS experiment. As with the LC-MS nearline workflow, features are ranked by fragmentation priority.

3. Fractionation Optimisation workflow:

   Workflow to optimise the number and width of fractions in a LC-MS fractionation experiment using a nearline approach. The resulting fractions can be used to perform DI-MS/MS and DI-MS^n post fraction collection.

3.5.2 ADAPTED-LC-MS-NEARLINE WORKFLOW

3.5.2.1 Workflow

The adapted-LC-MS-nearline workflow follows the same principle as the previously published nearline method (Neumann et al., 2013), a preceding LC-MS run is used to semi-automatically guide a proceeding LC-MS/MS. This approach differs to previously published work, as I provide a threshold for co-eluting features and attempt to distribute them across runs to maximise the number of features fragmented. I also rank the features based on adduct type, precursor ion purity, intensity and adduct cluster size to prioritize the precursors that could be the most informative to fragment. See Figure 3-1 for overview of experimental and data analysis workflow.

It should be noted that a prototype of the adapted-LC-MS-nearline approach was used for some preliminary assessments. See Figure 3-2 for details differences in LC-MS nearline approaches.

3.5.2.2 Software implementation

The scheduling approach has been implemented within an R package that relies on XCMS for feature picking (Smith et al., 2006; Tautenhahn et al., 2008), msPurity (Lawson et al., 2017) for precursor ion purity assessments and peak filtering, and the *Intervals* R package for interval scheduling. The nearline scheduling code can be found in (https://github.com/viant-metabolomics/xcmsWrapper - access available on request). The CAMERA (Kuhl et al., 2012) package was used to assign adducts and isotope annotations to features. A python package deconrank (https://github.com/computational-metabolomics/deconrank) was developed to rank and assign the features for fragmentation.

The workflow used was implemented into Galaxy, see Appendix section 8.10.

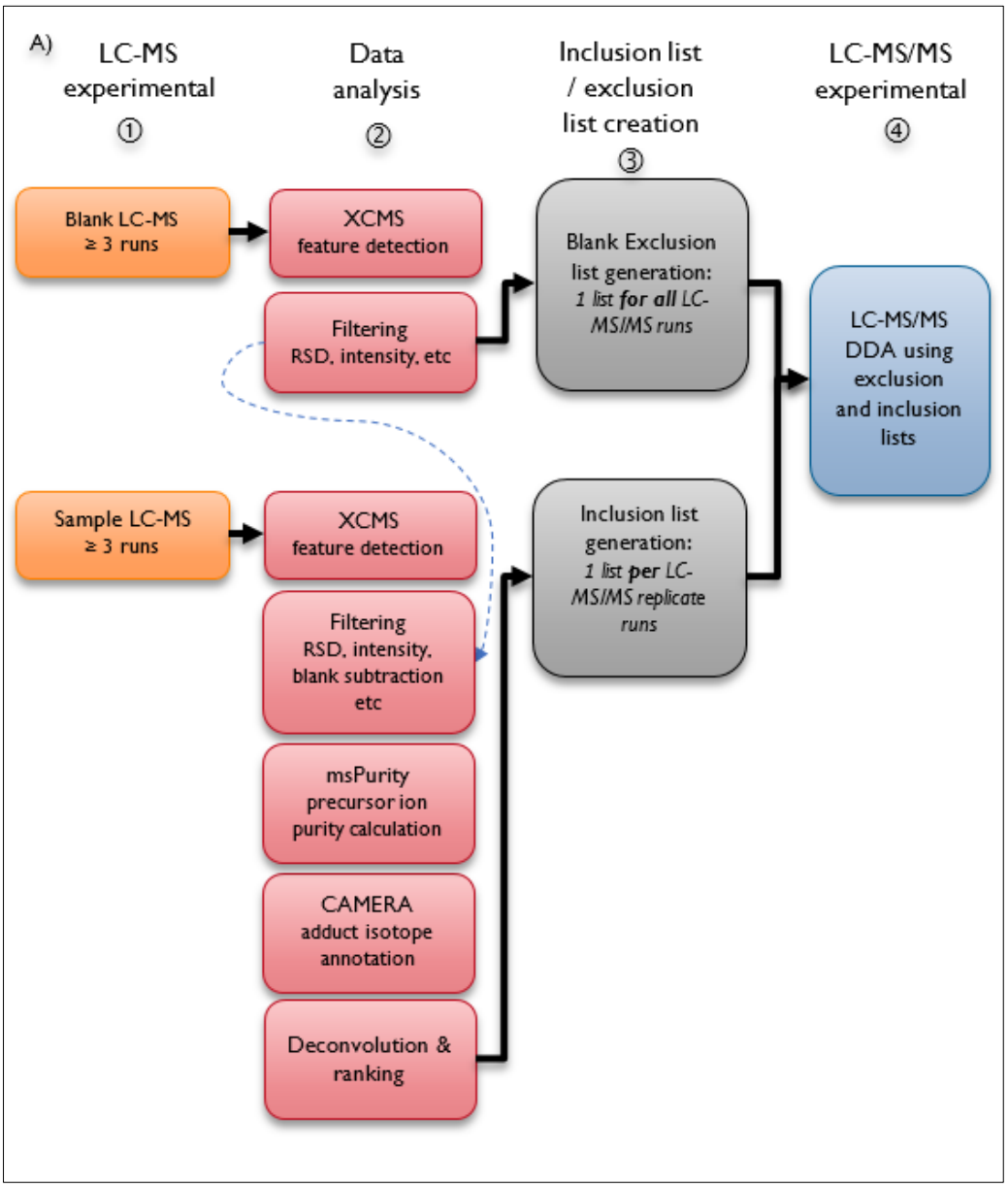


Figure 3-1 adapted-LCMS-nearline experimental and data analysis workflow

Workflow for both the experimental and computational aspects of the LC-MS-adapted-nearline approach. Step 1: Blank LC-MS and sample LC-MS runs are performed. Step 2: Data processing performed, e.g. peak feature detection, filtering, deconvolution and ranking of features. Step 3: Create inclusion lists and exclusion lists. The number of inclusion lists generated will depend on the number of LC-MS/MS runs to perform. The blank exclusion lists will be used for all subsequent LC-MS/MS runs. Step 4: Perform the LC-MS/MS runs.

3.5.2.3 Scheduling

The interval scheduling problems considered within the chapter can be categorised as

1. Weighted interval scheduling - no overlapping intervals
    - Original LC-MS-nearline method (Neumann et al., 2013)
2. Weighted interval scheduling - all overlapping intervals are allowed
    - Prototype adapted-LC-MS-nearline method
3. Weighted interval scheduling - *n* overlapping intervals are allowed
    - adapted-LC-MS-nearline method used for DMA of *D. magna*

The weighted interval scheduling problem used in the original nearline paper (Neumann et al., 2013) can be written as follows: Given a set of intervals where each interval has a start time, finish time and weight, find the subset of intervals where there are no overlapping intervals and the total weight is maximised. Intervals are the features, start and stop times represent the retention window and the weight represents the importance of the feature e.g. (level of intensity or *p-value* if performing differential analysis).

The method used to solve this problem by Neumann et al followed a well-known dynamic programming approach (Kleinberg and Tardos, 2005). The dynamic approach is then repeated until all the features have been scheduled into inclusion lists (see Appendix section 8.9 for algorithms).

When co-eluting features (overlapping intervals) are considered the problem can be described as follows: Given a set of intervals, a target of y schedules, a threshold of overlapping intervals, and each interval has a start time finish time and weight. Find y schedules (i.e. inclusion lists) where there is a maximum of x overlapping intervals (x and y being determined by the user). The first approach used for the prototype adapted-LC-MS-nearline method allowed all overlapping intervals and simply divided the features across the user defined number of

inclusion lists. This was approach was then improved to provide a limit of the number of overlapping features – where the features are first deconvoluted and ranked (see section 3.5.4) by order of preference to be fragmented. The features are then iteratively looped through and added to inclusion lists (checking before adding to the schedule that the minimum number of overlapping intervals are not exceeded). Although this approach will not determine the absolute optimum schedule (as all combinations are not assessed) it does provide acceptable schedules where speed is more important rather than achieving the optimum result.

All scheduling approaches can be used in the R package (https://github.com/viant-metabolomics/xcmsWrapper).

Another important consideration is how the mass spectrometer is able choose between the co-eluting features during the fragmentation acquisition. For experimental analysis collected for this chapter, a dynamic exclusion approach has been applied. The mass spectrometer places the *m/z* value of fragmented precursor from the inclusion list onto a dynamic exclusion list for a short length of time (length of time defined by user). In addition, the mass spectrometer can also be configured so that if a precursor on the inclusion list cannot be found then the mass spectrometer there can be an option to fragment any *m/z* value that can be isolated and has not been excluded. A static exclusion list is also used to exclude features and *m/z* values across the entire run.

Table 3-1 summaries the DDA approaches used between the nearline methods considered. Figure 3-2 demonstrates how inclusion lists and exclusion lists are distributed amongst LC-MS/MS runs and how individual features are scheduled for each run when overlapping features are permitted – this approach has been used for the DMA of *D. magna*.

Table 3-1 Overview of LC-MS-nearline approaches

	A) Original nearline method	B) Adapted-LC-MS nearline (prototype)	C) Adapted-LC-MS-nearline (applied to DMA of *D. magna*)
Interval Handling	❑ No overlapping intervals	❑ All overlapping intervals are allowed	❑ *n* overlapping intervals are allowed
Decon-volution	❑ No deconvolution	❑ Deconvolution	❑ Deconvolution
Score & rank features	❑ Features not ranked & score but schedules are optimised based on intensity or p-value (i.e. the weight)	❑ Intensity	❑ Combined score: ❑ Adduct type ❑ Anticipated precursor ion purity* ❑ Intensity ❑ Feature cluster size
Scheduling algorithm	❑ (Dynamic programming) ❑ Features are weighted by either intensity or p-value ❑ Sort features by finish times ❑ Calculate maximum weight (e.g. total intensity) achieved from intervals – using memoisation ❑ Traceback to the optimal solution ❑ Repeat until all intervals have been scheduled or until defined number of inclusion lists created	❑ Filter features > x intensity ❑ Divide all remaining features across defined number of inclusion lists	❑ Order Features by rank ❑ Iterate through features ○ Attempt to add each feature to inclusion list. ○ Check if there is overlapping intervals < than *n* ○ If TRUE, then add feature to first inclusion list ○ If FALSE, then attempt to add to next inclusion list (using the same criteria)
DDA setup	❑ Only fragment chosen precursors	❑ Only fragment chosen precursors ❑ Repeat with standard DDA with prior inclusion lists on exclusion list	❑ Fragment precursor on inclusion list – if no precursor available fragment anything not on the exclusion list

Footnote: Overview of the three different approaches used for LC-MS nearline discussed in this chapter. The original nearline method is implemented in the R package MetShot (https://github.com/sneumann/MetShot), the adapted methods were implemented (https://github.com/viant-metabolomics/xcmsWrapper).

*=Anticipated precursor ion purity was not used in the DMA of *D. magna* for LC-MS-nearline.

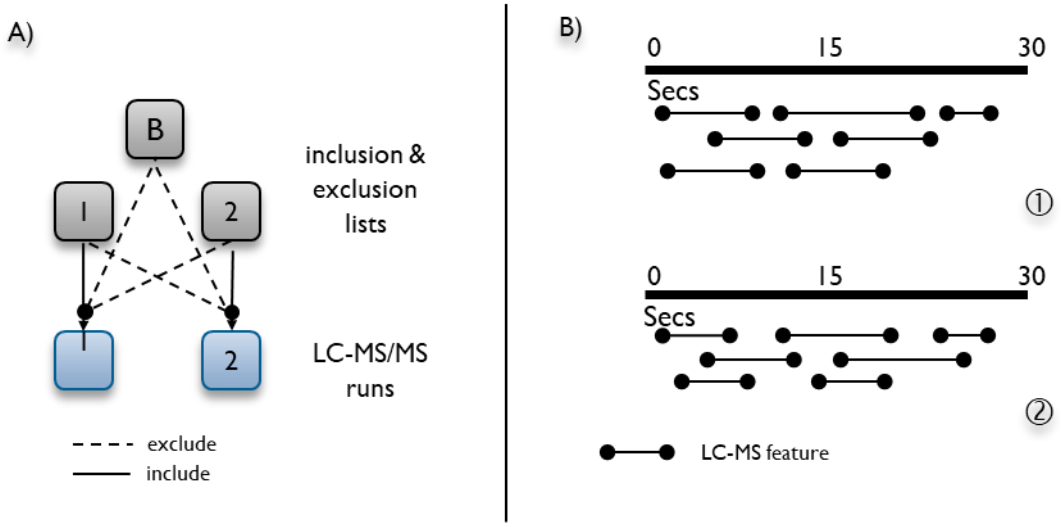


Figure 3-2 Inclusion list and scheduling schematics for LC-MS-nearline

A) Example of how the nearline uses both inclusion lists and exclusion lists. In the example the grey boxes with digits 1 & 2 are the inclusion lists have generated and single blank exclusion list has been generated from a prior LC-MS run of. The dotted lines represent the peaks that are to be excluded from fragmentation acquisition, the solid line represents the peaks that are to be included in the fragmentation acquisition. The figure demonstrates that the inclusion lists for one run will be added to the exclusion list of all other runs. **B)** Where there are co-eluting features an attempt is made to distribute them across 2 inclusion lists. This theoretical representation shows 14 LC-MS features distributed across 2 runs consisting of 30 seconds. Note that each LC-MS feature is only included in one schedule in this example.

3.5.3 DIMS-NEARLINE ACQUISITION

3.5.3.1 Workflow

The same principle used for the adapted-LC-MS-nearline workflow has been applied to DIMS. A preceding DIMS run is used to guide a DIMS/MS or DIMS^n run. Precursors are scheduled onto inclusion lists based on a pre-determined set time and rank the features based on adduct type, precursor ion purity, intensity and adduct cluster size to prioritize the precursors that could be the most informative to fragment. See Figure 3-3 for overview of experimental and data analysis workflow.

3.5.3.2 Software implementation

The peak processing or "feature picking" is performed with the python package DIMSpy (https://github.com/computational-metabolomics/dimspy - unpublished *R. Weber, 2018*), a modified version of the R package CAMERA is used to assign adducts and isotope annotations (https://github.com/computational-metabolomics/cameraDIMS - access available on request). The python package deconrank (https://github.com/computational-metabolomics/deconrank) was used to rank and choose features for fragmentation, the same package was also used to schedule the chosen features into inclusion lists.

Visual scripting was then used to automatically create mass spectrometry instrument methods that can be used directly to perform the analysis (https://github.com/viant-metabolomics/xcaliburauto - access available on request).

The workflow used was fully implemented into Galaxy, see Appendix section 8.11.

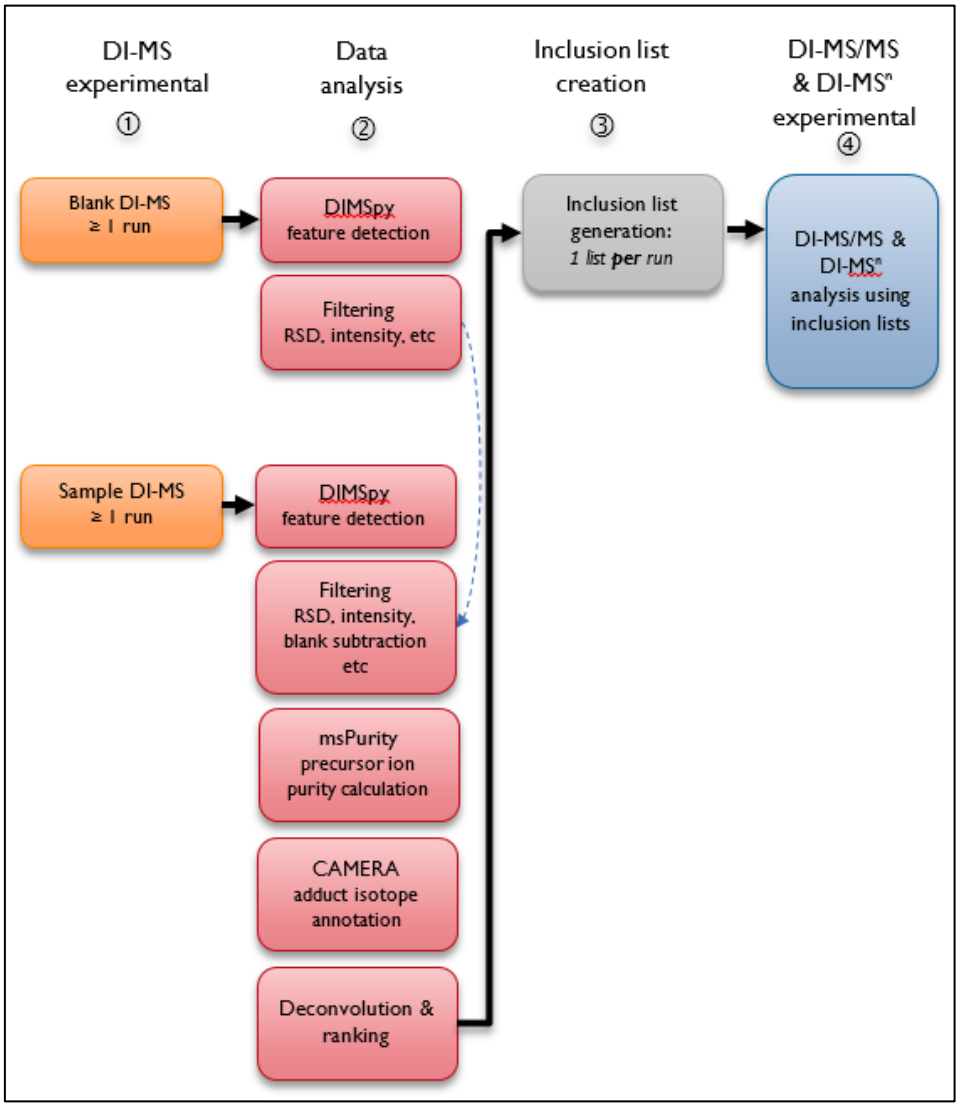


Figure 3-3 DIMS-nearline experimental and data analysis workflow

Workflow for both the experimental and computational aspects of the DI-MS-adapted-nearline approach. Step 1: Blank DI-MS and sample DI-MS runs are performed. Step 2: Data processing performed, e.g. peak feature detection, filtering, precursor ion purity calculation, adduct and isotope annotation and then deconvolution and ranking of features. Step 3: Create inclusion list based on ranked deconvoluted peaklist. The number of peaks on an inclusion list will depends on the length of time available for the proceeding fragmentation run. The instrument method file to create with the inclusion lists can be created manually or through an automated visual scripting method (https://github.com/viant-metabolomics/xcaliburauto - access available on request). Step 4: Perform the DI-MS/MS & DI-MSⁿ runs

3.5.3.3 Scheduling

The scheduling problem is simplified for DIMS compared to LC-MS. The first step is to determine the time it takes for one peak value to undergo fragmentation then based on a pre-determined limit to spend on each run the maximum number of precursors are added onto an inclusion list. If the time does not allow to fragment all precursors, then the those with a higher ranked score will be prioritized to be fragmented over those with a lower score.

3.5.4 DECONVOLUTION AND RANKING PRECURSORS

3.5.4.1 Summary

Within LC-MS and DIMS untargeted metabolomics experiment the number of features observed can be substantial - making it impractical to obtain fragmentation data on all features, a problem shared within proteomics mass spectrometry community (Michalski et al., 2011). However, even if reliable fragmentation spectra were obtained for all features within an untargeted mass spectrometry experiment there would be considerable redundancy (as one metabolite in mass spectra is represented by multiple features due to isotopes, adduct formation and in-source fragmentation). For these reasons, we have developed a method that uses the output of the isotope, adduct and annotation software CAMERA to deconvolute the features into known feature clusters (features representing the same metabolite) then rank and score these features based on multiple criteria to both prioritise features that we anticipate will fragment more successfully and prioritise features that explain more of the spectra than others.

3.5.4.2 Deconvolution and ranking procedure

The first step is to use the output from CAMERA where each feature in a dataset has either no annotation, an adduct annotation, an isotope annotation, or both an adduct and an isotope annotation. Feature clusters are then generated where features that share the same neutral mass (as determined by adduct) and have isotope relationships. In addition to these criteria, for an LC-MS experiment the features need to elute within the same time range as determined by CAMERA's peak cluster groups.

All adduct types used for annotation are previously ranked in order of how well that adduct formation can fragment (e.g. $M+H^+$ is ranked higher than $M+Na^+$). A feature can then be chosen from within each feature cluster that would provide the 'best' fragmentation data to describe that feature cluster. If there is a tie between features, then the most intense feature is chosen.

The chosen features are called “first-tier features” and the remaining peaks within the cluster are called “second-tier features”.

All features (both first and second tiers) are then scored based on weighted criteria using the following metrics (weightings can be determined by the user):

- Intensity
- Adduct score
- Size of feature cluster
- Precursor ion purity

The second-tier features can either be omitted from any fragmentation, alternatively these features can be included but with a penalised overall score (i.e. given a lower priority for fragmentation).

Hard threshold criteria are also applied to omit certain features for fragmentation:

- Precursor ion purity < *n*
- High confidence isotopes e.g. [M+1]+ and [M+2]+
- *%n* of second tier peaks removed from each feature cluster

For DI-MS experiments additional peaks will be excluded based on a given time limit of the analytical run being performed.

3.5.5 GALAXY TOOLS USED

Various tools were used for the Galaxy workflows used for the nearline approaches see Table 3-1. See Appendix section 8.10 and 8.11 for Galaxy workflow details.

Table 3-2: Galaxy tools used for nearline workflows

Galaxy tool name	New tool	Underlying package	Description
xcms.xset	N	XCMS	Data processing tool for feature detection of mass spectrometry datasets
xcms.xcmSet Merger	N	XCMS	Data processing tool to merge multiple xcms.xcmsSet data objects into a single object
xcms.group	N	XCMS	Data processing tool for grouping mass spectrometry peaks between different files
xcms.retcor	N	XCMS	Data processing tool for correcting retention time shifts in LC-MS data
xcms.fillpeaks	N	CAMERA	Data processing tool for "filling" missing peaks by re-assessing the mzML data file
CAMERA.annotate	N	CAMERA	Annotation tool for adducts and isotopes of LC-MS data
dimspy_process_scans	N	DIMSpy	Process Scans (and SIM-Stitch) - Read, filter and average MS scans
dimspy_merge_peaklists	N	DIMSpy	Merge peaklists produced by the tools 'Process scans (and SIM-Stitch)' or 'Replicate filter'
dimspy_align_samples	N	DIMSpy	Align Samples - Align peaks across Peaklists
dimspy_blank_filter	N	DIMSpy	Blank Filter - Remove 'blank' peaks from the biological mass spectra
collapse_collections	N	NA	Collapse collection into a single dataset in order of the collection
cameraDIMS	Y	cameraDIMS	Modification of the R package CAMERA to work on DIMS data
deconrank	Y	deconrank	Deconvolute adducts and isotopes then score and rank for fragmentation

Footnote: Table continued next page

Table 3-1: Galaxy tools used for nearline workflows (continued)

Galaxy tool name	New tool	Underlying package	Description
anticipated_purity_lcms	Y[x1]	msPurity	Calculate the anticipated precursor ion purity from a LC-MS XCMS dataset.
anticipated_purity_dims	Y [x1]	msPurity	Calculate the anticipated precursor ion purity from a DIMS dataset.
flag_remove_peaks	Y [x1]	msPurity	Tool to flag XCMS grouped peaks from the xcmsSet object based on various thresholds (e.g. RSD of intensity and retention time). The peaks can then be removed from the xcmsSet object and regrouped.
track_rt	Y [x1]	XCMS	Track the raw retention time in the peak slot of XCMSset objects
lcms_interval_scheduling	Y	xcmsWrapper	Generate target lists from LC-MS features for a fragmentation experiment
topn	Y	xcmsWrapper	Get the top (n) peaks in selected retention time windows of an LC-MS experiment
h_concatenate	Y	NA	Horizontal concatenation of user defined columns from multiple tables
dma_filelist_generation	Y	NA	Generate filelist for DMA DIMS nearline

Footnote: x1 These tools are newly developed but have also been used in Chapter 5 where they detailed as new tools.

3.5.6 FRACTIONATION OPTIMISATION

3.5.6.1 Summary

Figure 3-4 explains with a simplified example of how overlapping features can be minimized if the fraction width is optimised. This simplified example only collects 4 fractions over a period of 20 seconds and must distinguish between 10 features. For a fractionation of a typical LC-MS run there will be approximately 90 fractions over a period of approximately 18000 seconds and there could be over 10000 features making the problem considerably more difficult.

One approach to determine the optimum widths of the fractions is to bin the time of the chromatographic column into the smallest size the instrument can achieve. For illustrative purposes, say that this is 1 second, therefore a 10 second run would have 10 bins. We could then go through every combination of the bins to make up 5 fractions and check which combination has the lowest number of overlapping features. In this 10 second example, there would by 252 different combinations as calculated by Equation 3.1 (composition of n into k parts - COMP_ENUM - (Nijenhuis and Wilf, 2014)). For 252 combinations, this would take a minimal time to calculate the optimum combination. However, if we required 94 change points (bins) over 1500 seconds there would be over 10^{151} combinations and would be too time consuming to check every combination.

$$number\ of\ compositions = \frac{(n+k-1)!}{n! \times (k-1)!} \qquad (3.1)$$

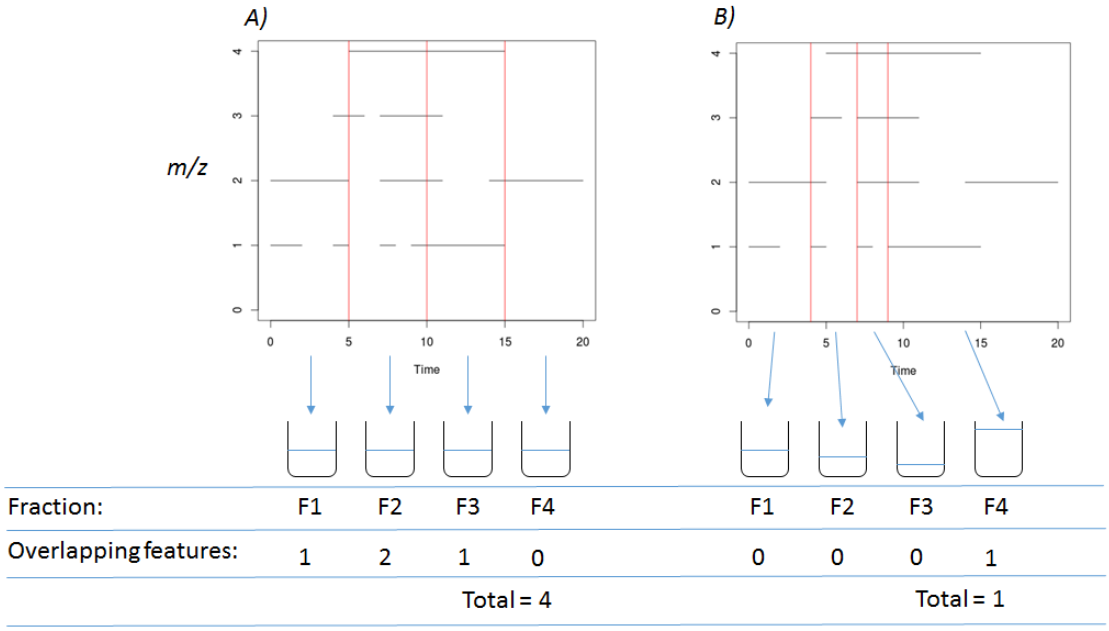


Figure 3-4 Fractionation optimisation

A) Divide the eluent from the column into 4 equally spaced fractions. Change points are at 5, 10 and 15 seconds. This results in a total of 4 overlapping features: 1 overlapping feature in fraction 1 (from *m/z* values of 1), 2 overlapping features in fraction 2 (from *m/z* values of 1 and 3) and 1 overlapping feature in fraction 3 (from *m/z* values of 2). B) This method has variable sizes for each fractions and results in only 1 overlapping feature (from fraction 4, *m/z* values of 2). The change points are at 4, 7 and 9 seconds

3.5.6.2 Change point genetic algorithm

A genetic algorithm to heuristically determine the optimum widths for fraction elution time has been developed. Genetic algorithms are modelled on evolution and start with a (typically random) population of potential solutions (or individuals). New individuals in the population are then generated from cycles of mating (or crossovers) within the population followed by removal of poor performing individuals. Random mutations in individuals (and keeping a few poor performing individuals) each cycle prevents the process getting stuck at local minima. Eventually after enough generations the population should have individuals who perform optimally (Bäck, 1996).

Genetic algorithms used for change point analysis require slightly altered crossover events

compared to standard genetic algorithms (Davis et al., 2006). This is because each change point is specific to a certain position in the time series and so the crossover between individuals needs to be ordered. The approach used is based on a previously published algorithm used for estimating optimum change points in a climatic series (Li and Lund, 2012).

In the implementation developed here, the individuals are represented as a chromatographic run split up into bins (or breakpoints). The breakpoints represent the chromosomes. The choice of breakpoints that can be used is determined from an initial analysis of feature list – where the chromatographic run is split into small bins where there are no overlapping features can be found, we call this pool of break points *bpool*. An initial population of individuals (by default 1000) is then created by randomly selecting *n* breakpoints from *bpool*. Where *n* is the number of fractions required for the fractionation experiment.

With each generation mutation and cross-over operations occurs and then a selection is made for individuals that move on to the next generation – individuals are ranked based on the number overlapping features observed (the objective function – is to lower the number of overlapping features in in the population).

The process of a mutation involves changing a randomly selected breakpoint to a neighbouring breakpoint defined in *bpool*. The distance from the original breakpoint is also randomly determined by a user defined ranges (default is 1 to 2). Following the mutation, crossover occurs by simulating reproduction. The top ranked individuals are chosen for mating and are split equally between males and females – the mating being randomly determined between the individuals. As per (Li and Lund, 2012) the breakpoints are then combined and ordered, each chromosome then has a 50% chance of being kept for the child.

It should be noted that the algorithm is optimising the overlap of the of peak apex of the overlapping features as attempts to optimise for the full width of the chromatographic feature

gave disappointing results. Reducing the overlap of the peak apex in each fraction will still reduce the subsequent complexity of the fractions though.

3.5.6.3 Software implementation

The fractionation optimisation software has been implemented in python (https://github.com/computational-metabolomics/fraction-optimiser – available on request)

3.5.7 DEEP METABOLOME ANNOTATION IMPLEMENTATIONS

The adapted-LC-MS-nearline workflow was implemented in an R package (https://github.com/viant-metabolomics/xcmsWrapper -available on request) and was run locally on a Desktop computer for the DMA of *D. magna.* Subsequently the workflow has been implemented as a Galaxy workflow for future DMA studies.

The DIMS-nearline workflow was used for the DMA of *D. magna* in three iterations the first as an R package (https://github.com/Viant-Metabolomics/msPurityBham - available on request), the second as a desktop GUI (https://github.com/Viant-Metabolomics/ftarget available on request) and finally as the Galaxy workflow described here.

The fractionation optimisation was not used for the DMA of *D. magna* due to the technicality of updating the chromatography fractionation software and the difficulty in assessing experimentally the effectiveness of the procedure, see the discussion for more details.

3.5.8 EXPERIMENTAL ASSESSMENTS

3.5.8.1 Prototype adapted-LC-MS-nearline comparisons to standard DDA methods

Preliminary analysis was performed on a prototype of the adapted-LC-MS-nearline method to compare the approach to standard DDA based fragmentation. However, a full comparison has not yet been carried out on the most recent algorithms and Galaxy implementation of the adapted-LC-MS-nearline workflow reported here.

Liquid chromatography was performed with A-phase consisting of 0.1% v/v formic acid in 95:5% v/v water/methanol. B-phase consisted of 0.1% v/v formic acid in 95:5% v/v methanol/water. The column was a Syncronis aQ (1.7 um, 2.1 x 100 mm), with a flow rate: 0.4 mL/min, column temperature: 40 °C., and injection volume: 10 µL.

Bovine serum sample was supplied from Sigma, re-suspend in 100% of A-Phase solution, vortexed for 60 seconds and centrifuged at 15000 rpm for 10 minutes.

Feature detection was carried out using the XCMS package (Smith et al., 2006; Tautenhahn et al., 2008) with the following parameters: peak picking algorithm = cent-Wave, snthresh = 5, ppm =5, prefilter = c(3, 5000), integrate = 1, mzdiff = 0.001. Peaks were aligned by retention time using the retcor-loess algorithm and grouped using the group-density algorithm with the following settings: bw = 5, mzwid = 0.01 and minfrac = 0.5. Adduct and isotope annotation as performed using the CAMERA package using default parameters with a 2 ppm error tolerance.

The acquisition was performed as follows: Three solvent blank LC-MS injections were followed by six bovine serum LC-MS injections. The LC-MS data from the resulting files was then processed using XCMS. A simplified scheduling approach was used (see column B of Table 3-1) to select features for the inclusion list. Deconvolution of the features and ranking was then performed as per section 3.5.4 – however only intensity was used to rank. All features

with > 30,000 intensity were evenly distributed between 2 sample lists. To account for shifts in retention times the "width" of the feature was increased by a factor of 4. XCMS features determine from the blank files were added to all exclusion lists.

Following the creation of the inclusion and exclusions lists, data was collected using (for 2 to 3 injections) using the

- Standard DDA approach (only a blank exclusion list)
- DDA with inclusion list 1
- DDA with inclusion list 2
- DDA with inclusion lists from 2 and 3 added to exclusion lists

3.5.8.2 Fractionation optimisation assessments

The fractionation-optimisation was performed on the features detected from three files generated from an LC-MS run (Thermo Fisher Scientific Q Exactive model in positive ionisation mode) on a *D. magna* sample – the same dataset was used for msPurity assessments in Chapter 2 - see section 2.5.11 and 8.2.1 for further chromatography and mass spectrometry details.

The genetic algorithm optimised fraction window sized approach was then compared to more conventional equally spaced fraction windows. The assessment was done for theoretical fractions of 2 to 119.

XCMS was used on the three sample LC-MS files as per section with the following parameters: peak picking algorithm = cent-Wave, snthresh = 5, ppm =5, prefilter = c(3, 5000), integrate = 1, mzdiff = 0.001. Peaks were aligned by retention time using the retcor-loess algorithm and grouped using the group-density algorithm with the following settings: bw = 5, mzwid = 0.01 and minfrac = 0.1. The exported grouped XCMS peaklist was used to optimise the fractionation.

The following parameters were used for each genetic algorithm optimisation: A starting

population of 1000 individuals underwent 300 iterations (generations), 1% of the population undergo mutation and the top 5 individuals are retained from the population for crossover operations. The was repeated 3 times to provide an average score to compare to the equally spaced fractionation.

3.5.8.3 Deep Metabolome Annotation assessments

A non-Galaxy implementation of the adapted-LCMS-nearline workflow and the Galaxy implementation of the DIMS-nearline workflow described in this chapter has been used for the DMA of *D. magna*. A small subset of this dataset is assessed here to see how effective the workflow is at fragmenting targeted features. This Chapter uses the subset of the DMA workflow stage, assay codes: AP_AMP_C30_LC-MS_LCMS-MS and AP_AMP1_C30_DI-MS-DI-MSn, where non-polar extraction of metabolites were performed followed by aminopropyl solid phase extraction, followed by reverse phase liquid chromatography using an Accucore C30 column. During the chromatographic run fractions of eluent were collected. An equivalent blank sample with no *D. magna* extract was collected in parallel to the *D. magna sample*. The fractions of eluents collected as the chromatography were dried down with liquid nitrogen and stored in a freezer for later analysis.

The adapted-LCMS-nearline workflow was performed using Thermo Fisher Scientific Q Exactive mass spectrometer in positive ionisation mode as described in Figure 3-1. The experiment consisted of five acquisition runs of sample LC-MS, five acquisition runs of blank LC-MS followed by two runs of LC-MS/MS in duplicate (four in total) of DDA acquisition using the generated target lists from the nearline processing.

Feature detection was carried out using the XCMS package with the following parameters: peak picking algorithm = cent-Wave, snthresh = 5, ppm =5, prefilter = c(3, 100), integrate = 1, mzdiff = 0.001. Peaks were aligned by retention time using the retcor-obi-warp algorithm and grouped

using the group-density algorithm with the following settings: bw = 5, mzwid = 0.01 and minfrac = 1. CAMERA was run with a 2 ppm error tolerance. No precursor ion purity assessments were made, and the ranked list of targets was based solely on adduct type and intensity. The scheduling was setup to allow 10 overlapping peaks.

The fractions were then analysed as per the workflow described in Figure 3-3 and Figure 8-6, where *D. magna* and an equivalent blank sample were analysed using a modified SIMStitch acquisition approach where only one technical replicate was acquired for each fraction for the blank and *D. magna* sample (compared to the more routine SIMstitch approach that has at least three technical replicates (Kirwan et al., 2014)), peaks were then filtered and deconvoluted and the resulting target list was used for fragmentation acquisition. The visual scripting script "XcaliburAuto" was then run to create an instrument method file to perform fragmentation of both MS^2 and MS^3 acquisitions with replication. The DIMS and $DIMS^n$ platform comprised a Triversa NanoMate coupled to an LTQ-Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). A maximum of 107 features can be added to each fragmentation acquisition (this equates to a maximum of 30 minutes per fraction).

The following parameters were used for each Galaxy tool can be found in appendix 8.12.

Full details of the experimental DMA procedure performed by M. R. Jones to be released in subsequent DMA *D. magna* publication.

All 72 fractions of the dataset were assessed to see the percentage of targeted features that were fragmented.

3.6 RESULTS

3.6.1 PROTOTYPE ADAPTED-LCMS-NEARLINE ASSESSMENTS

The results of assessing the prototype adapted-LCMS-nearline approach are summarised in Figure 3-5.

The results show the number of XCMS features determined with and without fragmentation for the following approaches:

- A: Standard DDA (with only blank exclusion list)
- B: DDA with inclusion list 1
- C: DDA with inclusion list 2
- D: DDA with inclusion lists from 2 and 3 added to blank exclusion lists
- E: Combined results of B, C and D. i.e. the adapted-LC-MS-nearline method

The standard DDA approach fragmented 1194 features (17% of the total number of XCMS features observed). The prototype adapted nearline approach fragmented 2089 features (27% of the total number of XCMS features observed). The adapted-LCMS-nearline approach therefore had a 10% increase from standard DDA approach and was able to fragment 895 more XCMS features. Following the assessments shown here adapted-LCM-nearline approach was modified further (see column C of Table 3-1 for modifications) and used for the DMA of *D. magna*.

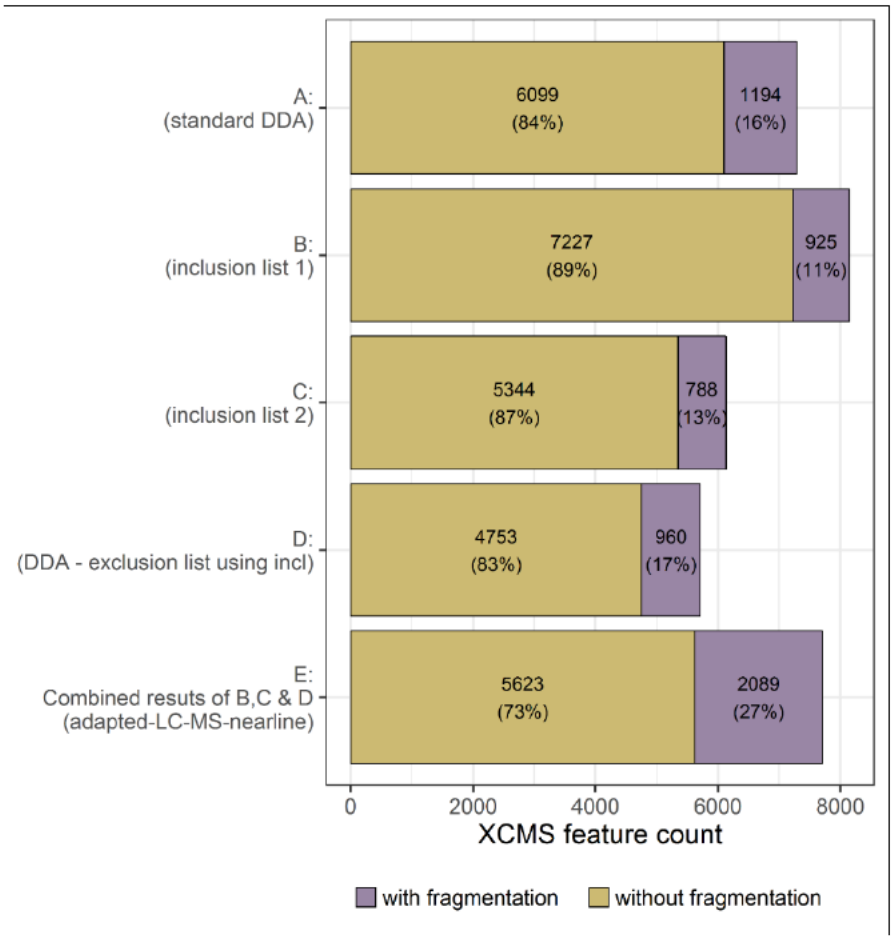


Figure 3-5 Comparison of effectiveness to fragment XCMS features using protype adapted-LCMS-nearline and standard DDA fragmentation

Comparing standard DDA based acquisition (A) against the prototype adapted-LC-MS-nearline approach. The prototype adapted LC-MS nearline approach consists of running DDA on two inclusion lists followed (result B and C), followed by DDA without any inclusion lists but with the inclusion lists from B and C and to blank exclusion list (result D). The combined result from B,C & D is considered the overall result from the prototype adapted-LCMS-nearline method. The smoke colour represents all LC-MS features with associated fragmentation spectra and the green colour represents all LC-MS features without any associated fragmentation spectra.

3.6.2 FRACTIONATION OPTIMISATION ASSESSMENTS

The standard approach (using equally sized fraction widths) was compared to the optimised approach with different limits on the maximum fraction number (Figure 3-6). The optimised method performs significantly better, i.e. fewer overlapping features than the standard approach for all the fractions (paired Wilcoxon rank sum p-value < 0.005). The mean of 341 additional overlapping features using the equal spacing approach compared the optimised approach (maximum reduction of 1182 overlapping features when 21 partitions – fractions - used). When only 96 fractions are optimised (i.e. to simulate a maximum number of fractions if 96 well plate is used for fraction collection) the number of overlapping features drops from 56 to 1.

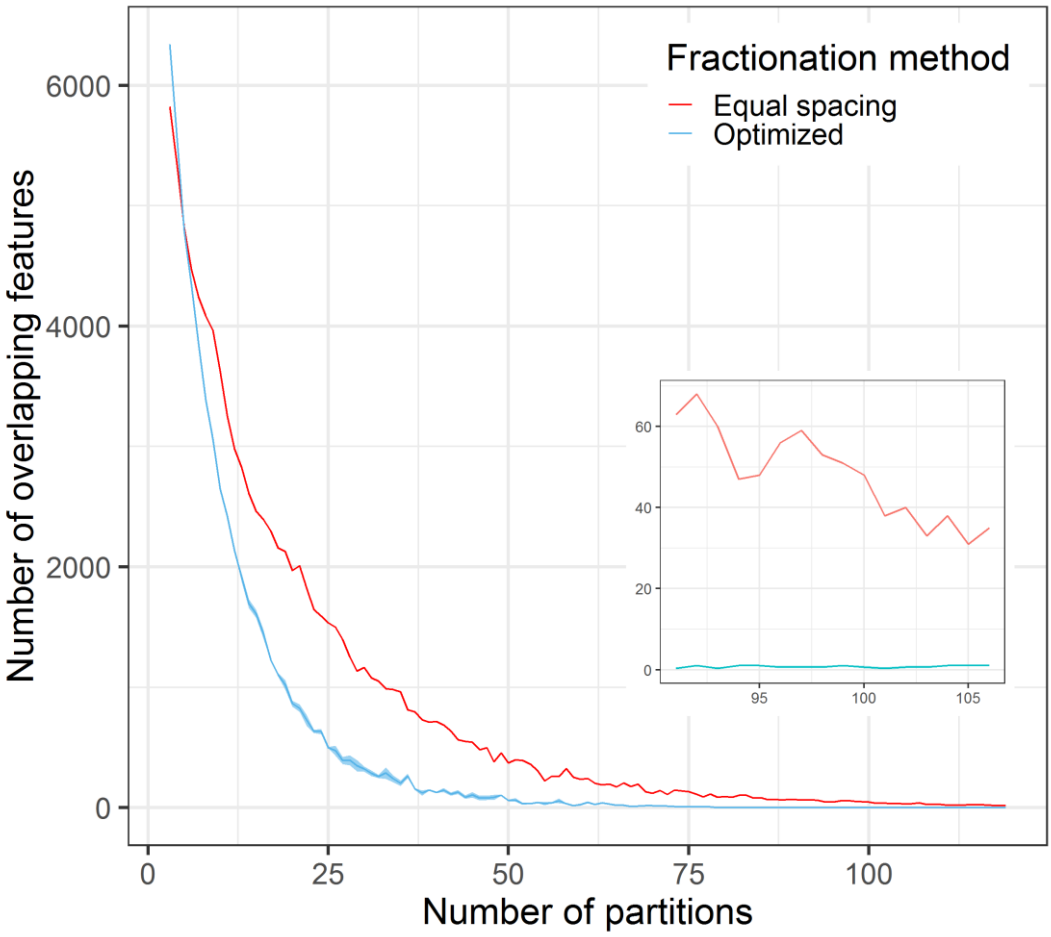


Figure 3-6 Optimised fractionation compared to standard equal spacing fractionation.

Maximum number of partitions (fractions) plotted against the total number of overlapping features. Comparing optimised approach of fractionation using genetic algorithm described here (blue line) to equally spaced fractionation (red line). The region between 96 wells to 106 wells has been magnified – as 96 well plates are commonly used for fractionation experiments – limiting the number of fractions to 96.

3.6.3 DEEP METABOLOME ANNOTATION ASSESSMENTS

3.6.3.1 Adapted-LCMS-nearline assessments for DMA

A subset of the DMA of *D. magna* has been used here to assess the effectives of the adapted-LC-MS-nearline workflow when applied to the DMA of *D. magna* (assay AP_AMP_C30_LC-MS_LCMS-MS). Table 3-3 shows the results of LC-MS(/MS) assay where 31 (%) of the features detected at the LC-MS level via XCMS had corresponding fragmentation spectra and 54 (%) of features on the inclusion lists were fragmentated.

The full annotation results of all LC-MS(/MS) component of the DMA of *D. magna* is detailed in Chapter 6.

Table 3-3: Assessment of adapted-LCMS-nearline on a subset of the DMA of *D. magna* - assay AP_AMP_C30_LC-MS_LCMS-MS

	Feature count
Total number of LC-MS features	4876
Total number of LC-MS features with fragmentation spectra	1504 (31%)
Number of LC-MS features on inclusion lists	775
Number of LC-MS features on inclusion lists with fragmentation spectra	421 (54%)

Footnote: Table describes the assessment of the adapted-LCMS-nearline workflow on a subset of the DMA of *D. magna*. The LC-MS(/MS) assay (AP_AMP_C30_LC-MS_LCMS-MS) was used to assess how many features that were detected in the full scan (MS^1) obtained equivalent fragmentation spectra. Also details how many features that were chosen to be targeted and added to the inclusion lists were fragmented.

3.6.3.2 DIMS-Nearline assessments for DMA

A subset of the DMA of *D. magna* has been used here to assess the effectives of the DIMS-nearline workflow when applied to the DMA of *D. magna* (assay AP_AMP1_C30_DI-MS-DI-MSn).

When combining all DIMS from each fraction a total of 54438 features were detected (a median of 638 features per fraction) as it was only possible to have 107 features per inclusion list – for this reason it was not possible to obtain fragmentation spectra on each fragment and only 7603 features were fragmented (7.3% of all features). It should be noted that there is redundancy between fractions as a single chromatographic feature can be eluted into multiple fractions – the effect of this however has not been accounted for as determining the origin of each feature can difficult due to changes in adduct formation from LC-MS to DIMS. Instead each fraction has been treated independently.

Overall 53% of each target list was fragmented. The maximum percentage of features fragmented per inclusion list was 82.2% and the minimum number of features fragmented per target list was 19.6%. See Figure 3-7 for breakdown for each fraction.

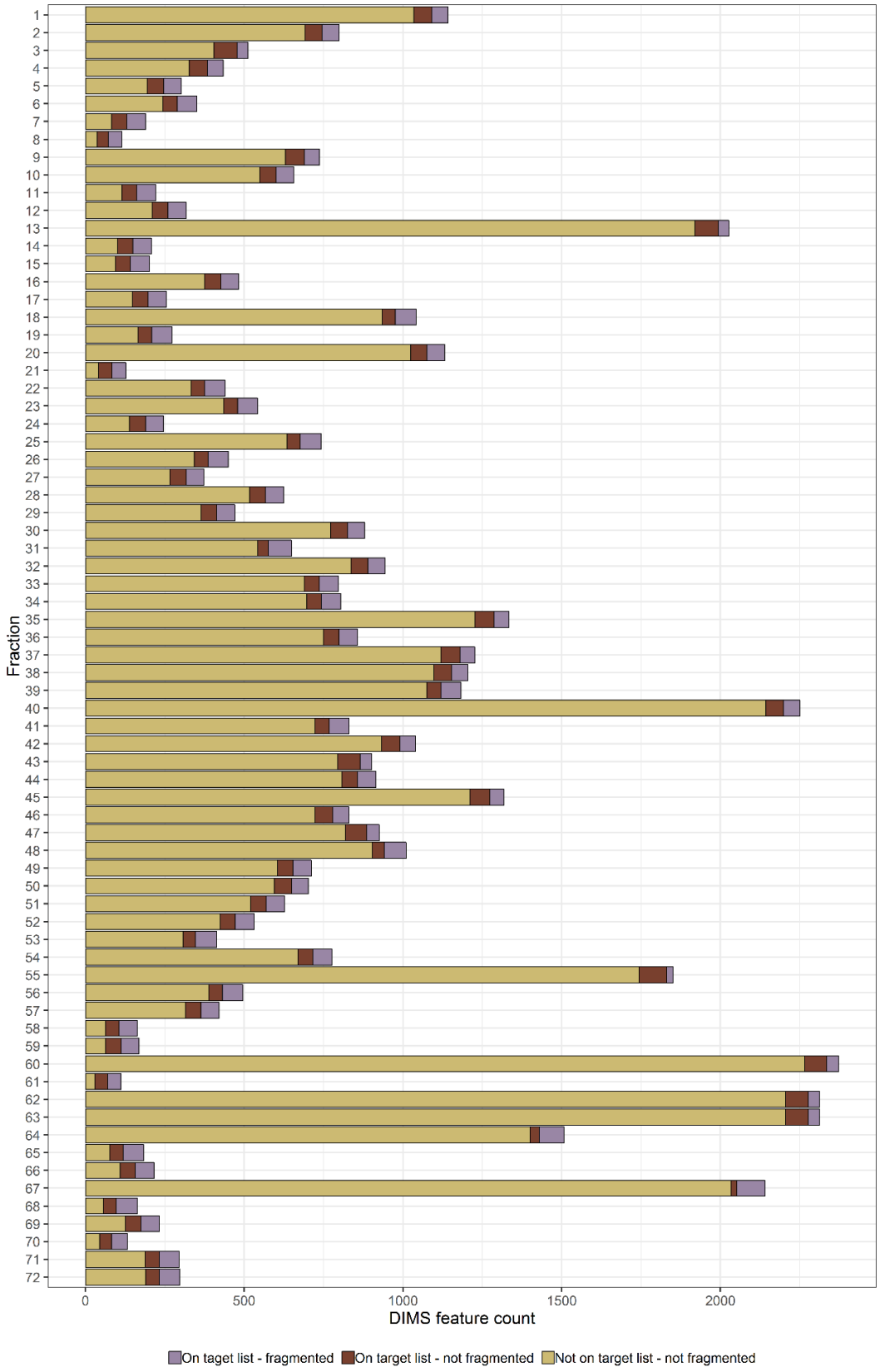


Figure 3-7 Assessment of DIMS-nearline on a subset of the DMA of *D. magna* - assay AP_AMP1_C30_DI-MS-DI-MSn

Figure details the assessment of the DIMS-nearline workflow on a subset of the DMA of *D. magna*. The DIMS(/MS) MS^n assay (AP_AMP1_C30_DI-MS-DI-MSn) was used to assess how many features that were detected in the full scan (MS^1) obtained equivalent fragmentation spectra. The smoke colour represents all features that were added to inclusion (target) list that were fragmented, the nutmeg colour represents all features that were added to the inclusion (target) list but could not be fragmented and the green colour represents all features that were not added to the inclusion (target) list and were not fragmented. The total sum of these categorisations is the total DIMS features observed for each fraction.

3.7 DISCUSSION

The preliminary assessments of the prototype adapted-LCMS-nearline approach showed moderate improvement over a standard LC-MS DDA procedure. However, more experimentation would need to be performed to assess how well the updated workflow implemented in Galaxy performs compared to standard DDA approaches. For a fair comparison to be performed, the original nearline method would also need to be assessed.

Following the preliminary assessments, a version of the adapted-LCMS-nearline method (without the Galaxy implementation) was used for the DMA of *D. magna* and performed reasonably in terms of annotations (see Chapter 6). However, the number of targets and features fragmented overall is still relatively low (e.g. 31% of the total number of LC-MS features have associated fragmentation spectra in the subset described in this chapter).

The proposed Galaxy workflow described here for the adapted-LCMS-nearline procedure could be used for future untargeted metabolomic annotation experiments of a similar nature. However, the effectiveness of the adapted-LCMS-nearline approach is limited by how many co-eluting features can be fragmented within the time it takes for the peaks to elute. For example, if there are 3 features covering a retention window of 20 seconds the mass spectrometer can easily cycle between these features but if there are over 400 features for the 20 second period this won't be possible even if it took a tenth of a second to perform a scanning event. As such, careful consideration needs to be made regarding how many coeluting features are observed, the time it takes for each scan event and the delay time observed for the instrument software to fragment precursors of the same mass.

The DIMS-nearline assessments disappointingly only fragment 53% of the features on the inclusion list for the subset of the DMA dataset assessed in this chapter. This is likely to have arisen from the target list creation using data derived from a SIM-stitch type acquisition. The

SIMstitch method has been shown to have increased sensitivity compared to standard DIMS approaches resulting in many more peaks being seen in the resulting peak list after data processing. However, the data acquisition for the fragmentation requires a prior full scan collected in the standard (typical range 50-1000 *m/z*) where ion suppression is more likely to occur causing reduced sensitivity. As the precursor from the target list must be located by the vendor software to initiate the fragmentation acquisition, in many cases the peak from the target list is either not present in this preceding full scan and as such does not get fragmented. This limitation could be overcome by ensuring the preceding scan used to trigger the fragmentation event being collected within the same SIM window that was used for the target list creation. However, it was not possible to create this dynamic method with the automated visual scripting method generation at the time of development.

The DIMS-nearline approach was used in various forms for the DMA annotation of *D. magna* with the Galaxy implementation being used for the final stage of data collection (currently being taken place). As such the annotation results were not possible to assess at the time writing. However, with adjustments to the current instrument method it is hopeful that increased number of targeted features could achieve successful fragmentation acquisitions.

Another limitation is that we used a maximum of only 107 features for each inclusion list and fraction (due to restricting the acquisition of each fraction to 30 minutes). This limit could be relaxed for other studies however as tens of thousands of fractions were to be analysed for DMA of *D. magna* it was not feasible to increase this number.

The nearline fractionation optimisation method shows a significant improvement other the non-optimised approach. It would be interesting to see how the predictions in how many features are observed in the LC-MS relate to the features in the fractions. However, it was not possible to update the available chromatography software (Dionex Chromatography MS Link 2.14 build 3818 - based on Chromeleon 6.80 SR13 Build 3818) used for the fractionation to allow different

sized windows and so it was not possible to assess beyond this hypothetical comparison.

The nearline methods in this chapter shows various attempts to improve either fragmentation or fractionation by using prior information that might be available from a preceding MS acquisition. However, there are various difficulties with the methods and areas that need improvement. One challenge is that nearline analysis is quite time consuming for the analyst as ideally a user would need to be able to run the nearline data analysis as soon as the prior MS acquisition has finished. This could be remedied by automating the workflows to be performed as soon as the required data has been acquired. A fully automated approach within the instrument vendor software would also be beneficial as to rely on either a local or Galaxy setup of the software might not be appropriate for many users. Indeed, a similar functionality has recently (2018) been launched with the Orbitrap ID-X Tribrid mass spectrometer that allows intelligent MS^n data acquisition strategy. However, this approach is only available for the Orbitrap ID-X Tribrid and using the instrument vendor proprietary software will not allow the same customisability and transparency that can be achieved using open source software and workflow systems like Galaxy.

Another consideration is the precursor ion purity of the available features to fragment. Samples in untargeted metabolomics are complex and have a high proportion of features with precursor ion purity scores < 0.5 (e.g. 53% reported for a complex sample of *D. magna*, (Lawson et al., 2017)), so even if there was an intelligent and fast way to obtain fragmentation on all the features in a dataset a high proportion would still need to be further separated from any contaminating ions to achieve reliable fragmentation for the chosen precursor. The same problem is observed in proteomics where DIA approaches have been suggested as a potential solution to help remedy the problem (Michalski et al., 2011).

Chapter 4

mzML2ISA & nmrML2ISA: Generating enriched ISA-Tab metadata files from metabolomics XML data

4.1 CHAPTER SUMMARY

Summary: Submission to the MetaboLights repository for metabolomics data currently places the burden of reporting instrument and acquisition parameters in ISA-Tab format on users, who must do it manually, a process that is time consuming and prone to user input error. Since most of these parameters are embedded in instrument raw data files, an opportunity exists to capture this metadata more accurately. Here I report a set of Python packages that can automatically generate ISA-Tab metadata file stubs from raw XML metabolomics data files. The parsing packages are separated into mzML2ISA (encompassing mzML and imzML formats) and nmrML2ISA (nmrML format only). Overall, the use of mzML2ISA & nmrML2ISA reduces the time needed to capture metadata substantially (capturing 90% of metadata on assay and sample levels), is much less prone to user input errors, improves compliance with minimum information reporting guidelines and facilitates more finely grained data exploration and querying of datasets.

Availability and Implementation: mzML2ISA & nmrML2ISA are available under version 3

of the GNU General Public Licence at https://github.com/ISA-tools. Documentation is available from http://2isa.readthedocs.io/en/latest/

Supplementary information: Available online and within the appendix pages of the thesis.

4.2 CONTRIBUTIONS

Martin Larralde[1†], Thomas N. Lawson[2†], Ralf J. M. Weber[2,3], Pablo Moreno[4], Kenneth Haug[4], Philippe Rocca-Serra[5], Mark R. Viant[2,3], Christoph Steinbeck[4], and Reza M. Salek[4]

[1]École Normale Supérieure de Cachan, 61 Avenue du Président Wilson, 94230 Cachan, France; [2]School of Biosciences and [3]Phenome Centre Birmingham, University of Birmingham, Birmingham, B15 2TT, UK; [4]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, UK; [5]University of Oxford e-Research Centre, 7 Keble Road, OX1 3QG, Oxford, UK

[†] Joint First Author for published article.

TNL developed the mzML2ISA tool during a NERC funded placement at EMBL-EBI Cambridge in July 2014. ML further developed the mzML2ISA tool as well as nmrML2ISA and all associated GUIs during an internship at EMBL-EBI in July 2015. TNL was also lead in writing the manuscript and co-developed the Galaxy wrappers with RJMW. PM aided in the Galaxy wrapper development and prepared the Docker containers. KH provided advice and practical input throughout the development of the tools. PRS provided expertise regarding the ISA-Tab format. MRV, CS and RMS supervised all work carried out. The code is currently maintained at github.com/ISA-tab where maintenance is shared between TNL, ML, RJMW, KH and PRS (all contributions can be seen within the git history). All authors read and approved the final published manuscript.

4.3 PUBLICATION HISTORY

This chapter has been previously published in Bioinformatics (Larralde et al., 2017), TNL being joint first author. Throughout the chapter grammatical corrections, updated references to sections and typos have been updated as required. Additionally, the use of the plural pronoun

"we" in the original paper has been replaced with "I" when describing work that I performed.

4.4 INTRODUCTION

MetaboLights, a database of experimental and derived metabolomics data (Kale et al., 2016), currently requires studies to be submitted in the ISA-Tab file format (Sansone et al., 2008), a hierarchical structure consisting of three components (Investigation, Study and Assay). ISA-Tab allows experimentalists to describe and record metadata in a simple format that aids reproducibility and shareability of experimental data and results. In addition to being a requirement for MetaboLights, the format is recommended for data centric journals (e.g. GigaScience and Scientific Data) and can facilitate data analysis (González-Beltrán et al., 2014; Lekschas and Gehlenborg, 2018). Data acquisition within a metabolomics experiment involves a wide range of instrument and data pre-processing parameters. Monitoring and recording those parameters is essential to maximize the reproducibility of experimental data. Currently MetaboLights utilises the ISA software suite (Rocca-Serra et al., 2010), particularly ISAcreator, to capture this information in ISA-Tab syntactic elements. Manual entry of > 40 potential parameters and associated ontology references for MetaboLights submission can be a laborious, time-consuming and error-prone process. However, most of this instrument metadata exists already within the instrument vendor data files and corresponding open source data formats. The use of freely available open source data formats for both mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy is advantageous for software development as it allows a single route to access instrument data and allows metadata to be standardised between vendors. The open source MS file format mzML (Martens et al., 2011) was created in 2008 as part of the Human Proteome Organization (HUPO) Proteomics Standards Initiative working group for mass spectrometry (PSI-MS). The mzML format amalgamated many of the benefits of the older MS formats (mzData (Orchard et al., 2004) and mzXML (Pedrioli et al.,

2004)) while adding new features. Importantly, the mzML format has a MS controlled vocabulary and is easily extendable (Martens et al., 2011). The imzML format is a relatively recent extension designed specifically for imaging MS (Schramm et al., 2012). The nmrML file format is a newly developed file format for NMR data, while not an extension of mzML, it is built around the same principles (http://nmrml.org/). Using these python packages, both the extraction of the instrument metadata and the subsequent population of an ISA-Tab structure has been automated. In so doing, I aim to both remove the submission bottleneck and help ensure the experimental data is reusable, sharable and standards compliant (Salek et al., 2015).

4.5 METHODS

4.5.1 ISA-TAB

The ISA-Tab structure consists of three file types. The Investigation file describes the overall objectives of a project as well as defining factors, protocols and parameters. The Study files describe the subjects studied and their characteristics and sampling methods. For every Study file there are one or more associated Assay files describing a measurement (e.g. metabolic profiling) and the technique (e.g. NMR). The instrument metadata extracted from either mzML, imzML or nmrML files are predominately used to automatically populate fields within the Assay file(s).

4.5.2 IMPLEMENTATION AND AVAILABILITY

The mzML2ISA & nmrML2ISA software can be used as an application programming interface (API), via the command line interface (CLI) or via a graphical user interface (GUI). To allow integration into Galaxy, a popular workflow platform for metabolomics and other 'omic data analysis (Afgan et al., 2016), the python packages have been wrapped into a Galaxy compatible tool format. Additionally, mzML2ISA & nmrML2ISA are available as tools within the PhenoMeNal Galaxy public instance and as Docker Containers as detailed in Appendix 8.14.

See Appendix 8.13 for full list of nmrML2ISA & mzML2ISA software and Appendix 8.14 for further implementation details. Instruction on how to use the tools and examples for common use cases can be found in the documentation http://2isa.readthedocs.io/en/latest/.

4.5.3 WORKFLOW

Due to the nature of the analytical technologies and structural differences between MS and NMR file formats, the software described here is divided into mzML2ISA (encompassing mzML and imzML formats) and nmrML2ISA (nmrML format only). However, the same workflow applies for both technologies (see Figure 4-1). From a user supplied collection of XML based data files and some experimental information (minimally a study identifier), the metadata are extracted from the XML data files. The Python Pronto package (https://pypi.python.org/pypi/pronto) is used to extract parameters referring to their accession number within either the HUPO PSI-MS ontology (Mayer et al., 2013), imagingMS ontology (Schramm et al., 2012) or nmrCV (http://nmrml.org/). Some instrument metadata is not available within the XML file format (e.g. chromatography parameters) such meta-data can be added by the user manually. Additionally, study design or wet lab experimental metadata (e.g. sample preparation details) can also be added by the user manually. The GUIs developed provide the easiest route for users who wish to provide additional metadata at this stage. The metadata is then converted into the ISA-Tab file format where many fields are automatically populated. Depending on the level of additional metadata already provided by the user, the ISA-Tab files may require review and expansion to meet annotation requirements (Fiehn et al., 2007; Sansone et al., 2007). These remaining fields can be manually added through the ISAcreator tool. The metadata generated from the XML file parsing is stored as a Python dictionary that can be rendered in JSON (JavaScript Object Notation), making it accessible to other software tools and analysis independent of ISA-Tab generation. See Appendix 8.15 for full details of all MetaboLights studies used for assessing software in this paper.

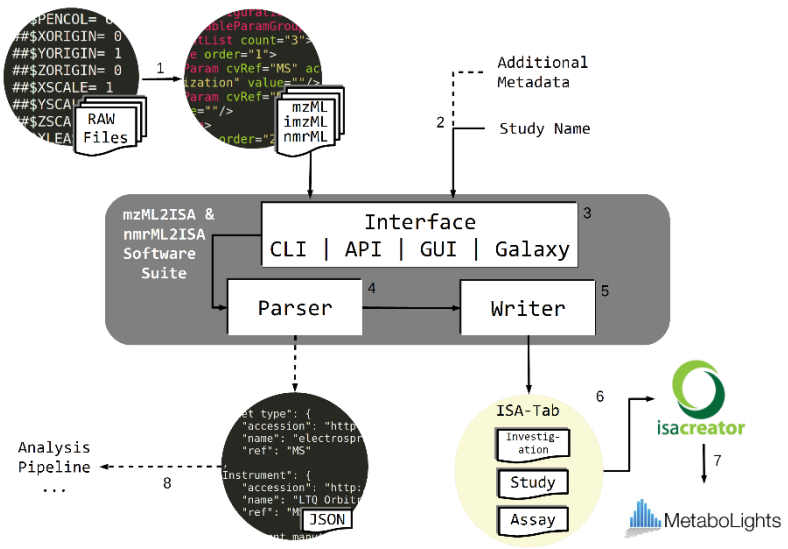


Figure 4-1 Schematic diagram and workflow of mzML2ISA & nmrML2ISA software suite

1) Experimental vendor raw files are converted into an open source XML equivalent. 2) A user provides experimental metadata (minimally a study identifier). 3) The additional metadata and open source XML raw files are submitted to the mzML2ISA & nmrML2ISA software through either a CLI, API, GUI or Galaxy interface. The time to complete this step is dependent on the extent of metadata provided. 4) Metadata is extracted from XML files. 5) ISA-Tab structure is generated with many fields automatically populated. Steps 4 and 5 take approximately 45 seconds for 50 XML files. 6) The remaining fields are then populated manually using the standard ISAcreator software. 7) The completed ISA-Tab structure can then be submitted to MetaboLights. 8) Additionally, the parsing components of mzML2ISA & nmrML2ISA can be used as standalone Python packages to extract metadata as either a python dictionary or JSON for integration in other analysis pipelines

4.6 RESULTS

4.6.1 EXTRACTED EXPERIMENTAL METADATA

Up to 23, 42 and 46 instrument metadata terms for mzML, imzML and nmrML respectively are automatically extracted and parsed into the assay component of the ISA-Tab structure. Extracted terms for the XML files include generic instrument descriptors (e.g. the instrument name, manufacturer and software), data transformation descriptors (e.g. file conversion details) and platform specific descriptors (e.g. mass analyser, detector and *m/z* range for MS derived

XML files). Where possible, extracted terms from each file format should be found within a relevant controlled vocabulary. Full details of the extracted terms can be found at http://2isa.readthedocs.io/en/latest/ (see the extracted terms sections).

4.6.2 METABOLIGHTS CASE STUDIES

The mzML2ISA & nmrML2ISA packages have parsed and created ISA-Tab structures for all MetaboLights studies that have associated mzML, imzML, nmrML files. The resulting ISA-Tab structures have then been successfully validated using the ISAvalidator (https://github.com/ISAtools/ISAvalidator-ISAconverter-BIImanager) tool. See Appendix 8.15 for details of the 21 studies tested.

Using sample sets of 50 XML files (either mzML, imzML or nmrML) derived from MetaboLights, XML to ISA-Tab conversion is completed in less than 45 seconds. The exact time will vary depending on file type and size. See Appendix 8.16 for details.

4.7 DISCUSSION

The mzML, imzML and nmrML data file formats used in metabolomics provide a parameter rich, technical layer of metadata that has been exploited here to both improve the reliability of MetaboLights submissions and increase the ease and speed of submission. Additionally, the generated ISA-Tab structures can be used with subsequent downstream analysis with software such as Risa (González-Beltrán et al., 2014) and the Refinery Platform (Lekschas and Gehlenborg, 2016).

However, to reach core information for metabolomics reporting compliance (CIMR) (BioSharing: Bsg-s000175), derived from the Metabolomics Standards Initiative (Fiehn et al., 2007; Sansone et al., 2007), a range of distinct descriptors still have to be reported into the ISA-Tab structure. Typically, these are experimental design key descriptors (predictor variables, replicate information, study group information), subject to sample relationships, as well as

biological sample characteristics, and sample separation procedures if relevant (e.g. chromatography). The development of the ISA-API (https://github.com/ISA-tools/isa-api) will further facilitate the automation of these missing descriptors, especially regarding experimental design, and aid in reducing the barrier and time required to submit metabolomics data to repositories. Indeed, instrument vendor raw files may harbour many more parameters that are often absent in the open source equivalent files.

Metadata extraction and integration into an open source file format can however be an involved process to implement for each instrument vendor API. Additionally, licensing issues can limit full exploitation of the vendor raw files. Further work that brings together the metabolomics and proteomics community, along with the instrument vendors, could help identify and integrate useful but missing parameters into the appropriate open source format.

4.8 ACKNOWLEDGEMENTS

Drs David Johnson, Alejandra Gonzalez-Beltran and Peter Li for contributions provided during the China UK data dissemination in metabolomics (CUDDEL) workshop (2016).

4.9 FUNDING

This work was supported financially through a NERC CASE PhD studentship with GigaScience, grant number NE/L002493/1; BBSRC, grant numbers BB/L005077/1, BB/M019985/1 and BB/M027635/1; MRC UK MEDical BIOinformatics partnership, grant number MR/L01632X/1; and the PhenoMeNal European Commission's Horizon2020 programme, grant number 654241.

Chapter 5

A Suite of Django Applications for Metabolome Organization with Galaxy and ISA

5.1 CHAPTER SUMMARY

Background: Reproducible and transparent metabolomics analysis has been made more achievable by using the Galaxy web-framework. However, managing multiple metabolomic assays and study data with FAIR (Findable, Accessible, Interoperable, and Re-usable) standards can still be challenging in Galaxy. Several projects have been developed using the Django web-framework to organise metabolomics studies and analysis, yet these Django projects either lack integration with Galaxy and/or have limited integration with the experimental design schema of ISA (Investigation, Study and Assay).

Additionally, a data management software that integrates with Galaxy and ISA design schema were considered key to managing large scale annotation studies such as the Deep Metabolome Annotation (DMA) of *D. magna* in a reproducible manner.

Findings: This chapter presents a suite of 4 reusable Django applications that have been developed for both developers and general users for managing metabolomic data using ISA experimental design metadata and Galaxy. The django-galaxy application is for Galaxy and Django interfacing. The django-misa application allows metabolomics projects to be stored

with an ISA framework using Django models. The results and metadata can then be exported as an ISA-JSON structure with the intention for submission to the MetaboLights metabolomics data repository. The django-mbrowse application provides general utility for browsing, visualising and searching metabolomics spectra and annotations, the initial focus being on mass spectrometry data. The final application - django-mogi - imports and connects the functionality of all the applications. Additionally, I present over 10 new Galaxy tools that have been developed for data processing and metabolite annotation.

Conclusion: By providing the tools for both users and developers to manage metabolomics data, ISA metadata and associated Galaxy processing information I have enabled metabolome studies to be performed in a more controlled and reproducible manner. As the Django applications are reusable and can be integrated into both existing and new Django projects, I aim to make metabolomic web-development in Django FAIR. The suite of Django applications and associated Galaxy annotation workflow has subsequently been applied to managing and analysing the DMA of *D. magna* (see Chapter 6).

5.2 CONTRIBUTIONS

Thomas N. Lawson[1], Martin R. Jones[1], Ralf J. M. Weber[2], Jordi Capellades[3], Peter Li[4], Mark R. Viant[1,2]

[1]School of Biosciences and [2]Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [3]Universitat Rovira i Virgili (SP) and [4]Gigascience, Hong Kong

TNL developed all Django projects and applications. TNL, JC, MRJ and RJMW developed tools and workflows within Galaxy, JC was lead in developing the MetFrag and Sirius CSI:FingerID Galaxy tools. PL and MRJ supervised all work performed.

5.3 PUBLICATION HISTORY

This chapter has not been previously published and is in a draft form. The format follows the

GigaScience guidelines on structure – where a modified version of the chapter is intended to be submitted.

5.4 FINDINGS

5.4.1 BACKGROUND

Reproducible data management of metabolomic experiments and data analysis can be challenging due to the complexities of both analytical and computational procedures. Workflow environments such as Galaxy (Afgan et al., 2018) have been previously used in metabolomics to perform reproducible and traceable data analysis (Davidson et al., 2016; Giacomoni et al., 2015). However, Galaxy does not have advanced functionality to organise and manage multiple studies. Software have been developed to provide this extra functionality and flexibility (Correia et al., 2015) but these tools do not provide the functionality for managing and searching metabolite annotations and data. Additionally, the current Galaxy data management software has limited functionality to handle experimentation within the ISA-Tab format.

For handling large scale metabolome annotations projects, such as the Deep Metabolome Annotation (DMA), the need to manage the data analysis and explore the results in a controlled and reproducible manner is key in utilising the information to its full potential. Currently there are limited options available for managing metabolomics data and no software available to handle metabolomics data that integrate with Galaxy and ISA-Tab. This chapter details the developments made that can be applied to DMA of *D. magna* but also can be utilised for any mass spectrometry-based metabolomics study.

The ISA-Tab format is a popular data structure that plays a critical role in sharing experimental metabolomics data and meta data using FAIR principles - Findable, Accessible, Interoperable, and Re-usable (Wilkinson et al., 2016). The format is required for submission to the metabolomics data repository MetaboLights (Kale et al., 2016) and is either required or strongly

encouraged for many journals (Nature Scientific Data, GigaScience, Metabolomics, Metabolites, PLoS journals and f1000Research (Rocca-Serra et al., 2016). The ISAcreator GUI (graphical user interface) and other specialist tools (Larralde et al., 2017) can be used to create ISA-Tab structures but incorporating ISA export options to existing LIMS (Laboratory Information Management Systems) or data management software provides a potentially easier route to the ISA creation process by utilising the meta-information already stored within the management framework. For example, the MASTr-MS (Hunter et al., 2017) is a metabolomics LIMS that includes the functionality to export a study directly to a ISA structure. Full implementation of a LIMS can be a daunting task to a small research group though and the benefits might seem remote. Data management systems that focus more on data analysis could be more appealing to some, such as the PiMP software (Gloaguen et al., 2017). Additionally, having a locally available resource to view, search and curate large scale data sets to interrogate data prior to publication can be extremely beneficial for researchers.

The management software MASTrMS and PiMP were developed within Django – a web framework created in 2005 by Adrian Holovaty and Simon Willison designed to enable the fast development of secure and scalable web applications and has become popular choice for a variety companies and organisation (e.g. Mozilla Firefox, Bitbucket, Instagram and NASA. See DjangoSites for a list of existing Django websites). However, the current Django metabolomics data management projects do not have Galaxy integration (MAStr-MS, PiMP) and/or do not have ISA interfacing (PiMP).

It should be noted that Django is one of many frameworks and resource that can be used to aid and develop web applications and software by providing some level of in-built classes and functions to deal with commonly required functionality (e.g. dealing with HTTP responses and requests – see section 1.7.2). Alternatives to Django include Ruby on Rails (developed in Ruby), Drupal (developed in PHP) and Flask (developed in Python). Python based web-

development frameworks and resources have the benefit that they can be more easily integrated with the wealth of data analysis resource available as Python packages (Jones et al., 2001; Kiefer et al., 2013; Landrum and others, 2013; O'Callaghan et al., 2012). Django and Flask are two of the most popular python based web-development frameworks and both are good choices to develop a web based data management system. However, although Flask has the benefit of being lightweight and flexible, we consider Django to be a more suitable approach for data management software as it provides a more structured approach to developing web applications, facilitating a community of sharing Django applications and functionality across different projects

The metabolomics community has clear benefits from increasing the number of public reproducible data and meta-data shared within FAIR principles. This can include using public datasets for developing and validating data processing methods and new software (Lawson et al., 2017; Stanstrup et al., 2015) or using publicly available mass spectrometry fragmentation spectra for metabolite annotation purposes, either directly through spectral matching (Horai et al., 2010) or as part of machine learning predictive approaches like Sirius CSI-FingerID (Böcker et al., 2009; Böcker and Dührkop, 2016). However, the percentage of published metabolomic studies where the associated data and meta data publicly available is still limited (Rocca-Serra et al., 2016) suggesting new tools and integrations are still required to help researchers comply to share their data. By locally managing metabolomics data with ISA-tab and Galaxy data can be more easily shared with the metabolomics community.

Here we present a suite of reusable applications for Metabolomics Organisation using Galaxy and ISA (MOGI) using the Django web framework. In addition, we have developed and co-developed over 10 new Galaxy tools for data analysis and metabolite annotation and have provided utilities to export relevant information into Django. Finally, by making the Django applications reusable we hope it will allow easier integration of the applications into both

existing and new metabolomics web development projects.

5.4.2 THE DJANGO-MOGI SUITE

A collection of 4 reusable Django applications (django-galaxy, django-misa, django-mbrowse django-mogi) have been created that provide general functionality useful for developers and general users for web-based metabolomics software. Documentation for all applications is hosted via ReadTheDocs: http://mogi.readthedocs.io/en/latest/. An example Django project integrating the whole suite of applications is provided https://github.com/computational-metabolomics/django-mogisite. The container platform Docker can be used for installation purposes http://mogi.readthedocs.io/en/latest/user-docs/installation.html.

Each of the Django application presented provide unique functionality intended to be integrated into existing Django projects or to be used together to create a new Django project.

The Django application 'django-mogi' imports all the other Django applications to provide an interface connecting the functionality of the different applications. See the following sections detail how the django-galaxy, django-misa and django-metab work independently and how they can be integrated.

A small application (django-gfiles) was also developed for additional file handling functionality not provided in base Django (e.g. saving files as symbolic links). This application is imported and used for all the MOGI suite of Django applications.

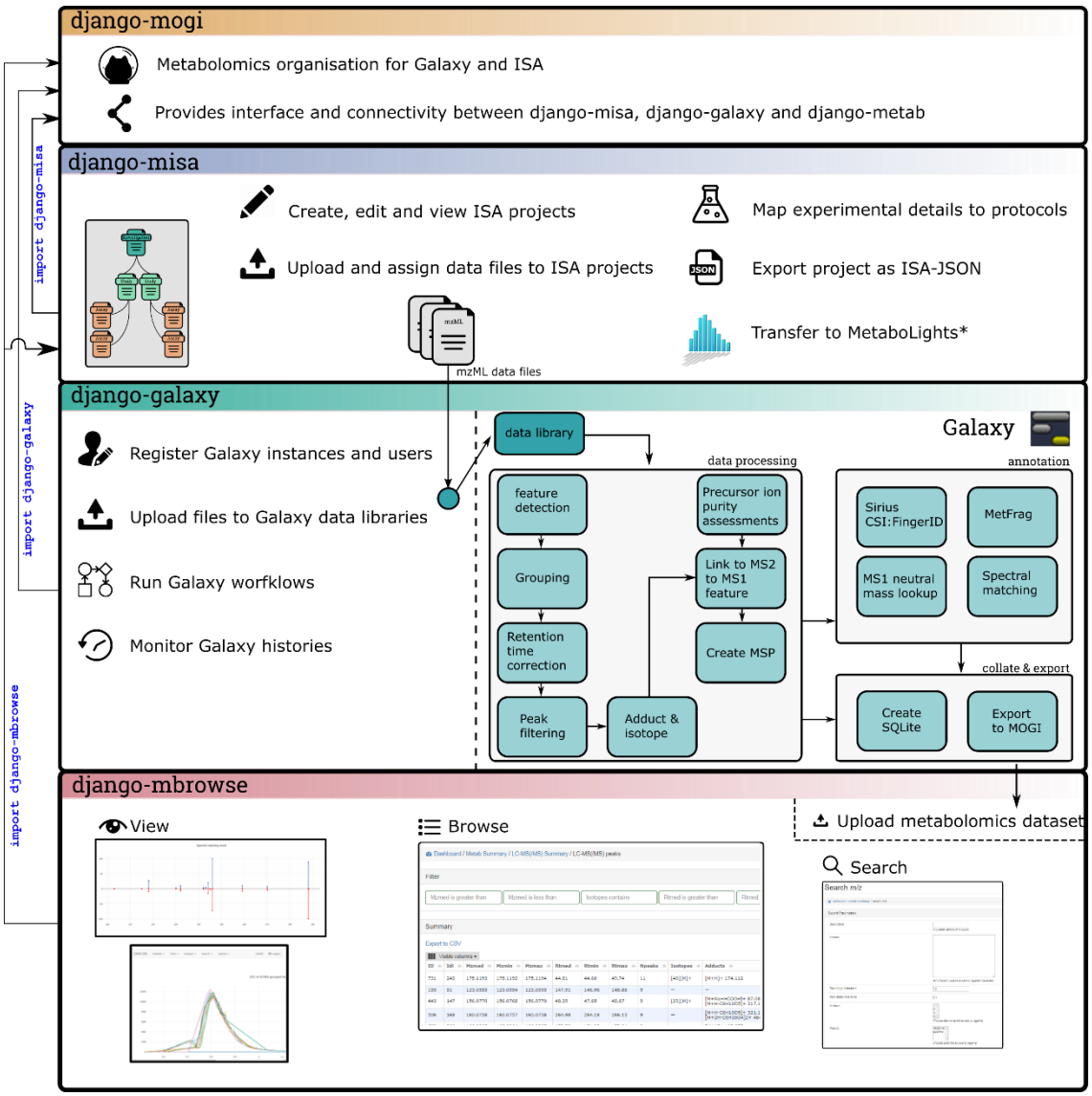


Figure 5-1 Django MOGI application suite

Overview of applications functionality. The django-mogi application provides interfacing between the other applications. The django-misa applications provides functionality for users to create, edit and view ISA projects; upload data files and assign to assays; map experimental details to defined protocols (e.g. extraction, chromatography, measurements); export meta-data to ISA-JSON and eventually there are plans to export the data and metadata to MetaboLights. The django-galaxy application allows users to upload data (e.g. from an ISA project) to a data library within Galaxy and manage histories and workflows through the Django interface. The django-mbrowse allows users to view, browse and search metabolomics data. The input datasets can be uploaded directly or from galaxy when an API call has been made through to Django.

5.4.3 DJANGO-MISA

5.4.3.1 General

The ISA framework consists of three primary layers. The Investigation layer details the global aims of a project and defines the protocols used for the project along with any associated factors and parameters. The Study layer details the characteristics and sampling methods used for the subjects involved in the project. The Assay layer details the measurement and techniques used. An Investigation can have multiple associated Studies and a Study can have multiple associated Assays. The django-misa application models the three layers and associated protocols and parameters in the Django MySQL database, allowing an ISA project to be created, edited and viewed within standard Django utilities. See Figure 5-2 for an overview of the django-misa components.

5.4.3.2 Creating ISA backbone

Basic functionality is available to create the ISA (Investigation, Study and Assay backbone). Once this has been created sample details and assay details can be uploaded and added.

5.4.3.3 Ontologies

Ontologies are essential for recording an experiment in a reproducible and searchable manner and the ISA Framework enables all meta information to be referenced to an ontological term and source. Within django-misa a Django model (and corresponding database table) is used to store all ontological terms and sources within a project. This table is then referenced whenever an ontological term is used. New ontological terms can be added manually or can be searched with django-misa using the EBI ontology Lookup service.

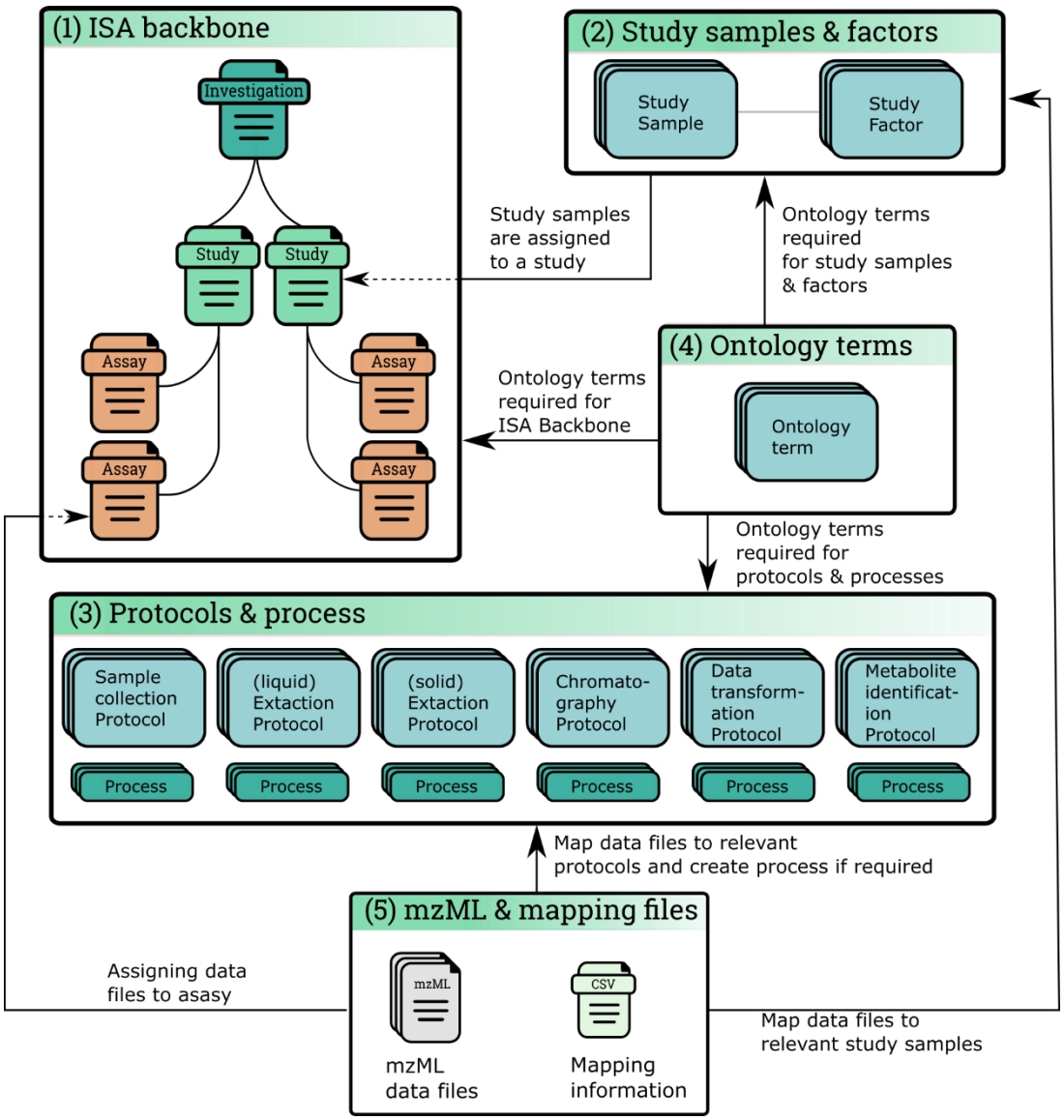


Figure 5-2 Overview of django-misa components

(1) ISA backbone of investigation, study and assay is created. (2) Study samples and relevant study factor details are recorded for the ISA project (this can be done manually or in a batch process). (3) Protocols for keys steps in the metabolomics experiments are created. When a protocol is recorded as being used, a relevant process object is created that details relevant 'run time' information. (4) Ontological terms are used for the ISA backbone, study factors and protocols and processes. (5) Metabolomics data files can be uploaded to a relevant assay, where details of the protocols and processes and samples can be mapped via a mapping file.

5.4.3.4 Sample information

Multiple samples are typically involved in an experimental study each with different study factors. This information can be added to django-misa through the web interface, where multiple study factors and associated ontological terms can be added along with organism information of the sample. As this process can be time consuming to perform for each sample, the sample details can be uploaded in a batch format.

5.4.3.5 Protocols

Protocols in the ISA framework can be used either in the Study layer (e.g. sample material preparation) and the assay layer (e.g. Extraction, and Chromatography). Models for metabolomics focused protocols have been created that represent the key components of a metabolomics experiment. These protocols are in part based on the MetaboLights descriptions for metabolomics study and consist of sample collection, (liquid phase)-extraction, (solid phase)-extraction, chromatography, measurement (i.e. mass spectrometry type), data transformation and metabolite identification.

Each protocol has an associated "process" model for recording the use of a protocol. This allows additional information to be recorded that can vary between uses of the same protocol.

5.4.3.6 Uploading and mapping experimental data files

Experimental data files (currently supporting mass spectrometry RAW and mzML files), can be uploaded to a relevant assay. The details of what protocol and sample have been used for each file can be mapped via an additional mapping file consisting of a comma separated list of "code_fields" corresponding to the relevant protocols and samples used for each file.

5.4.3.7 ISA-Tab and ISA-JSON import and export

Once an ISA project has been created and data and results have been uploaded to that project, an ISA-JSON object representing that ISA project can be exported. The general goal would be

for this ISA-JSON object be exported to MetaboLights. Further work still needs to be done on exporting and importing ISA-JSON (and ISA-Tab) files.

5.4.3.8 Database models

Figure 5-3 and Figure 5-4 show the model schema used for django-misa.

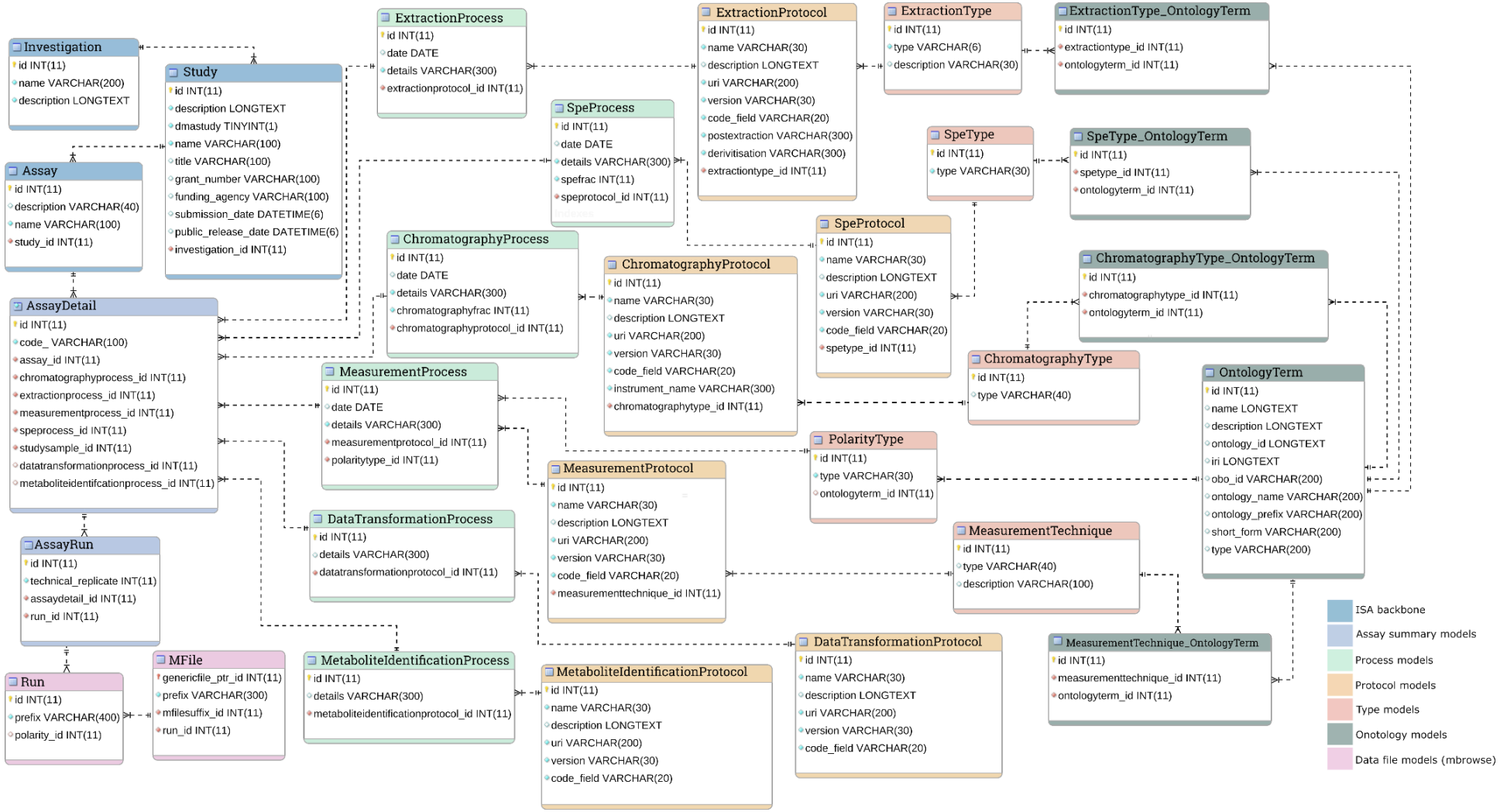

Investigation
id INT(11)
name VARCHAR(200)
description LONGTEXT
Study
id INT(11)
description LONGTEXT
dmastudy TINYINT(1)
name VARCHAR(100)
title VARCHAR(100)
grant_number VARCHAR(100)
funding_agency VARCHAR(100)
submission_date DATETIME(6)
public_release_date DATETIME(6)
investigation_id INT(11)
Assay
id INT(11)
description VARCHAR(40)
name VARCHAR(100)
study_id INT(11)
AssayDetail
id INT(11)
code_ VARCHAR(100)
assay_id INT(11)
chromatographyprocess_id INT(11)
extractionprocess_id INT(11)
measurementprocess_id INT(11)
speprocess_id INT(11)
studysample_id INT(11)
datatransformationprocess_id INT(11)
metaboliteidentifcationprocess_id INT(11)
AssayRun
id INT(11)
technical_replicate INT(11)
assaydetail_id INT(11)
run_id INT(11)
Run
id INT(11)
prefix VARCHAR(400)
polarity_id INT(11)
MFile
genericfile_ptr_id INT(11)
prefix VARCHAR(300)
mfilesuffix_id INT(11)
run_id INT(11)
ExtractionProcess
id INT(11)
date DATE
details VARCHAR(300)
extractionprotocol_id INT(11)
ExtractionProtocol
id INT(11)
name VARCHAR(30)
description LONGTEXT
uri VARCHAR(200)
version VARCHAR(30)
code_field VARCHAR(20)
postextraction VARCHAR(300)
derivitisation VARCHAR(300)
extractiontype_id INT(11)
ExtractionType
id INT(11)
type VARCHAR(6)
description VARCHAR(30)
ExtractionType_OntologyTerm
id INT(11)
extractiontype_id INT(11)
ontologyterm_id INT(11)
SpeProcess
id INT(11)
date DATE
details VARCHAR(300)
spefrac INT(11)
speprotocol_id INT(11)
SpeType
id INT(11)
type VARCHAR(30)
SpeType_OntologyTerm
id INT(11)
spetype_id INT(11)
ontologyterm_id INT(11)
SpeProtocol
id INT(11)
name VARCHAR(30)
description LONGTEXT
uri VARCHAR(200)
version VARCHAR(30)
code_field VARCHAR(20)
spetype_id INT(11)
ChromatographyProcess
id INT(11)
date DATE
details VARCHAR(300)
chromatographyfrac INT(11)
chromatographyprotocol_id INT(11)
ChromatographyProtocol
id INT(11)
name VARCHAR(30)
description LONGTEXT
uri VARCHAR(200)
version VARCHAR(30)
code_field VARCHAR(20)
instrument_name VARCHAR(300)
chromatographytype_id INT(11)
ChromatographyType_OntologyTerm
id INT(11)
chromatographytype_id INT(11)
ontologyterm_id INT(11)
ChromatographyType
id INT(11)
type VARCHAR(40)
MeasurementProcess
id INT(11)
date DATE
details VARCHAR(300)
measurementprotocol_id INT(11)
polaritytype_id INT(11)
MeasurementProtocol
id INT(11)
name VARCHAR(30)
description LONGTEXT
uri VARCHAR(200)
version VARCHAR(30)
code_field VARCHAR(20)
measurementtechnique_id INT(11)
PolarityType
id INT(11)
type VARCHAR(30)
ontologyterm_id INT(11)
MeasurementTechnique
id INT(11)
type VARCHAR(40)
description VARCHAR(100)
OntologyTerm
id INT(11)
name LONGTEXT
description LONGTEXT
ontology_id LONGTEXT
iri LONGTEXT
obo_id VARCHAR(200)
ontology_name VARCHAR(200)
ontology_prefix VARCHAR(200)
short_form VARCHAR(200)
type VARCHAR(200)
DataTransformationProcess
id INT(11)
details VARCHAR(300)
datatransformationprotocol_id INT(11)
DataTransformationProtocol
id INT(11)
name VARCHAR(30)
description LONGTEXT
uri VARCHAR(200)
version VARCHAR(30)
code_field VARCHAR(20)
MeasurementTechnique_OntologyTerm
id INT(11)
measurementtechnique_id INT(11)
ontologyterm_id INT(11)
MetaboliteIdentificationProcess
id INT(11)
details VARCHAR(300)
metaboliteidentificationprotocol_id INT(11)
MetaboliteIdentificationProtocol
id INT(11)
name VARCHAR(30)
description LONGTEXT
uri VARCHAR(200)
version VARCHAR(30)
code_field VARCHAR(20)
ISA backbone
Assay summary models
Process models
Protocol models
Type models
Onotology models
Data file models (mbrowse)

Figure 5-3 Model schema for ISA assay protocol, process and related data files (django-misa)

Models for ISA assay protocol, associated process and the relationship between the assay and metabolomics data file (mzML-MFile). “Investigation”, “Study” and “Assay” models represent the backbone of the ISA structure. “AssayDetail” and “AssayRun” models summarise the assay experimental information, where “AssayDetail” links all the experimental processes that have been performed and “AssayRun” links to the acquisition run model “Run” which in turn links to “MFile” model that represents the metabolomics data file (e.g. the mzML file). The “ExtractionProcess”, “SpeProcess”, “ChromatographyProcess”, “MeasurementProcess”, “DataTransformationProcess” and MetaboliteIdentificationProcess” models represent the process of using the corresponding protocols represented by “ExtractionProtocol”, “SpeProtocol”, ChromatographyProtocol”, “MeasurementProtocol”, “DataTransformationProtocol” and “MetaboliteIdentificationProtocol”. Protocols can be grouped together by shared types or techniques, represented by the models “ExtractionType”, “SpeType”, “ChromatographyType” , “PolarityType” and “MeasurementTechnique”. These types and techniques are stored using ontological terms represented by the model “OntologyTerm” and associated linking models that allow multiple ontology terms for each type/technique.

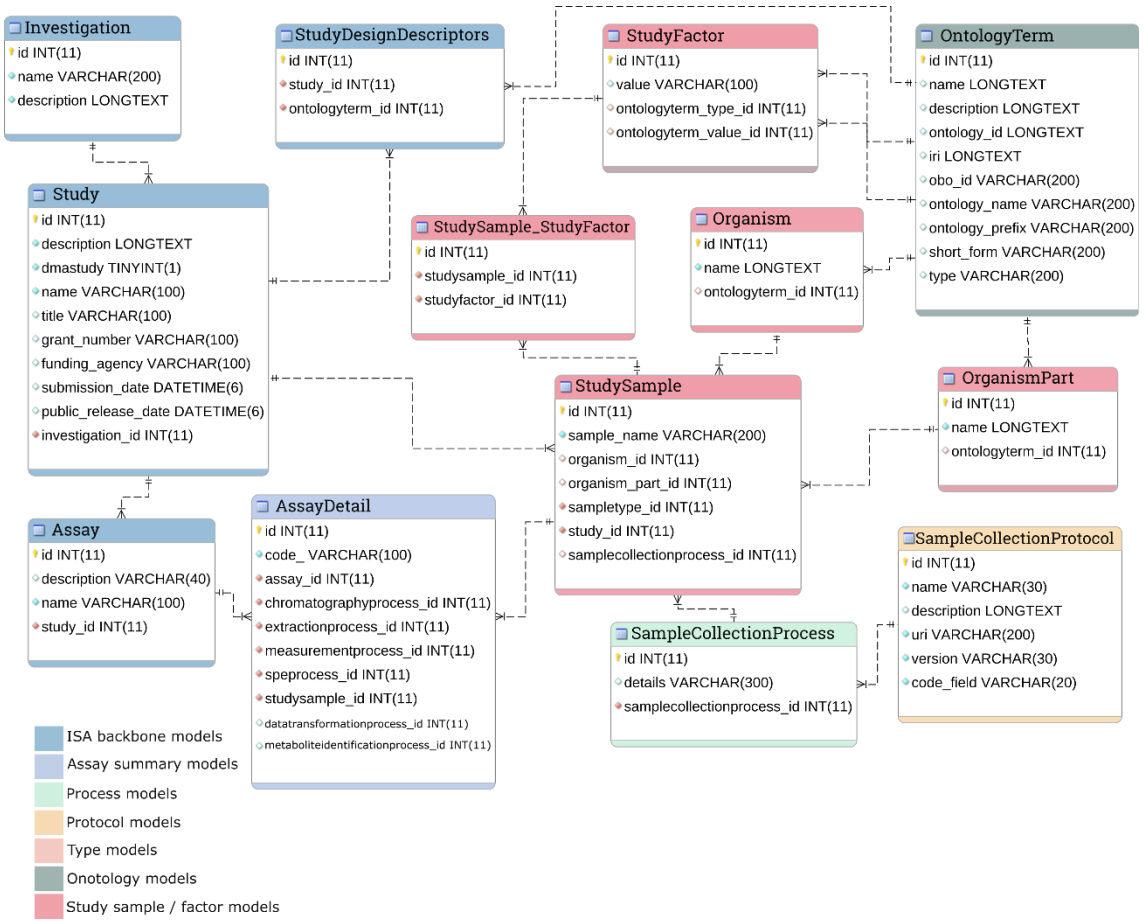


Figure 5-4 Model schema for ISA study-design, study-sample, study-factor and sample collection process and protocol (django-misa)

Models for ISA study-design, study-sample, study-factor, sample collection process and related protocol. "Investigation", "Study" and "Assay" models represent the backbone of the ISA structure. The "StudyDesignDescriptor" model is used to record any ontological term that describes the study, using the "OntologyTerm" model. The "AssayDetail" model summarise the assay experimental information and links all the experimental processes that have been performed. The "StudySample" model represents any sample that is used with the study, multiple study-factors (represented by "StudyFactor" model) can be associated with many "StudySamples". Each study sample in most cases will have a corresponding Organism and Organism part (e.g. anatomy) represented by the "Organism" and "OrganismPart" model. Each sample will have been derived from a sample collection protocol, represented by the "SampleCollectionProtocol", where the process of using the "SampleCollectionProtocol" is represented by the "SampleCollectionProcess". The "StudyFactor", "Organism" and "OrganismPart" all use ontological terms using the "OntologyTerm" model.

5.4.4 DJANGO-GALAXY (AND GALAXY DEVELOPMENTS)

5.4.4.1 General

The django-galaxy application allows users to register multiple Galaxy instances and users to a Django project. Once a Galaxy instance and User has been registered with the Django project, users can load data files into Galaxy histories and data libraries, run workflows and manage histories directly from Django.

The integration of Django and Galaxy allows Django applications to be developed where Galaxy can be used as a "backend" processor. See Figure 5-10 for recommended infrastructure.

5.4.4.2 Database model

Figure 5-5 summarises the model schema for django-galaxy

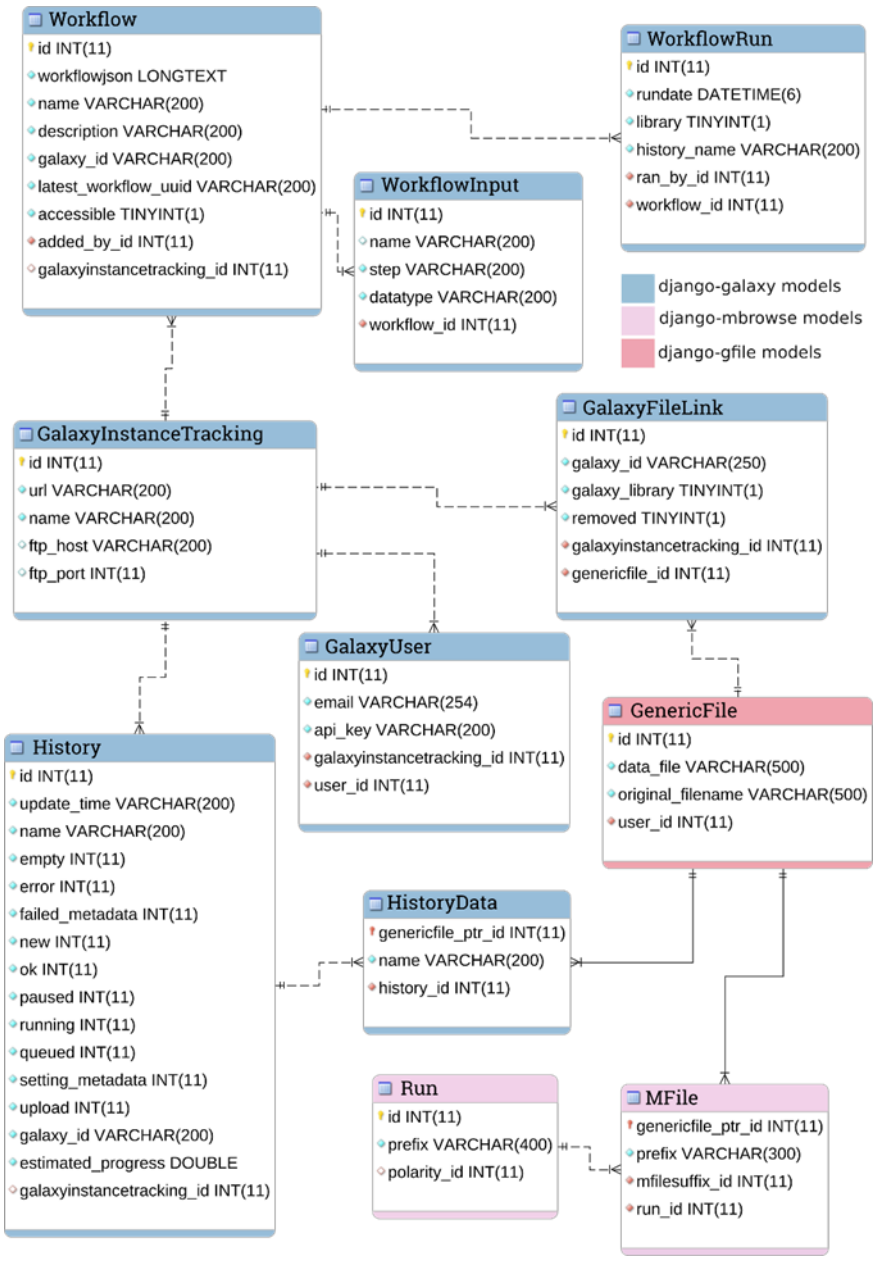


Figure 5-5 Model schema for django-galaxy

The "GalaxyInstanceTracking" model represents the Galaxy instances. The "GalaxyUser" represents an associated user for a Galaxy instance (this is linked to the Django auth user – model not shown). The "Workflow" model represents a Galaxy workflow. The "WorkflowInput" model represents the data inputs required for a workflow and their corresponding data types. The "WorkflowRun" model records the details when a workflow has been run from django-galaxy. The "History" model represents the Galaxy histories and details the current status e.g. number of running, queued, passed, failed processes. The "HistoryData" model represents a data object that can be saved transferred from the Galaxy instance to the file storage used for Galaxy. The "GenericFile" model is from the django-gfiles application and a generic file handling model. When used with django-galaxy, the "GenericFile" can either linked to a

"HistoryData" object or a "GalaxyFileLink". The "GalaxyFileLink" model is used to link files to Galaxy data libraries. When used with Django-mbrowse this allows metabolomic data files, represented with the models "MFile" and "Run", to be uploaded to Galaxy data libraries.

5.4.4.3 New Galaxy tools and workflows

Over 10 new Galaxy tools have been developed and integrated for data processing and annotation, including the following annotation software MetFrag (Ruttkies et al., 2016), Sirius CSI:FingerID (Böcker et al., 2009; Dührkop et al., 2015) and msPurity (Lawson et al., 2017).

Developed in Java, MetFrag is a combinatorial *in silico* fragmentator for MS^2 spectra, that generates *in silico* spectra from potential matches of hypothetical neutral masses or molecular formula of the fragmentation precursor ion. The Sirius software is another Java based annotation software with both a GUI and CLI, that uses isotope patterns, fragmentation trees and machine learning with CSI:FingerID, where spectral reference libraries are used to determine fingerprints of potential candidate structures. As with MetFrag, fragmentation spectra consisting of an array of *m/z* values and array of intensity values along with hypothetical neutral mass or molecular formula of the precursor ion is the required as starting point. The msPurity R package was developed to evaluate the quality of fragmentation spectra by assessing the contribution of the precursor ion within the fragmentation isolation window. Further developments to msPurity since the initial release have been made to allow fragmentation spectra to be searched against library spectra (e.g. MassBank, LipidBlast (Horai et al., 2010; Kind et al., 2013)).

The core functionality of the msPurity package along with the MetFrag and Sirius CSI:FingerID been implemented within Galaxy. The tools have been integrated into a LC-MS/MS annotation workflow, see Figure 5-6.

Table 5-1 Overview of new Galaxy tools

Galaxy tool name	Lang-uage	Under-lying tool	Description
assess_purity_msms	R	msPurity	Assess the precursor ion purity for fragmentation spectra acquired from either a LC-MS/MS or DI-MS/MS (MSn) experiment
anticipated_purity_lcms	R	msPurity	Calculate the anticipated precursor ion purity from a LC-MS XCMS dataset
anticipated_purity_dims	R	msPurity	Calculate the anticipated precursor ion purity from a DIMS dataset.
frag4feature	R	msPurity	Assign fragmentation spectra to XCMS features using msPurity
create_msp	R	msPurity	Create MSP files after frag4feature function
flag_remove_peaks	R	msPurity	Tool to flag XCMS grouped peaks from the xcmsSet object based on various thresholds (e.g. RSD of intensity and retention time). The peaks can then be removed from the xcmsSet object and regrouped.
create_sqlite	R	msPurity	Create SQLite database of an XCMS-CAMERA dataset
spectral_matching	R	msPurity	Perform spectral matching to spectral libraries using dot product cosine on a MS/MS dataset
track_rt	R	xcms	Track the raw retention time in the peak slot of XCMSset objects
submit_to_mogi	Bash	NA	Submit Galaxy datasets to MOGI django database
metfrag	Java / Python	MetFrag	Run MetFrag CLI in batch mode
SIRIUS-CSI:FingerID	Java / Python	Sirius CLI	Run SIRIUS-CSI:FingerID CLI in batch mode

Footnote: See section 5.5.1 for other Galaxy tools used for this chapter

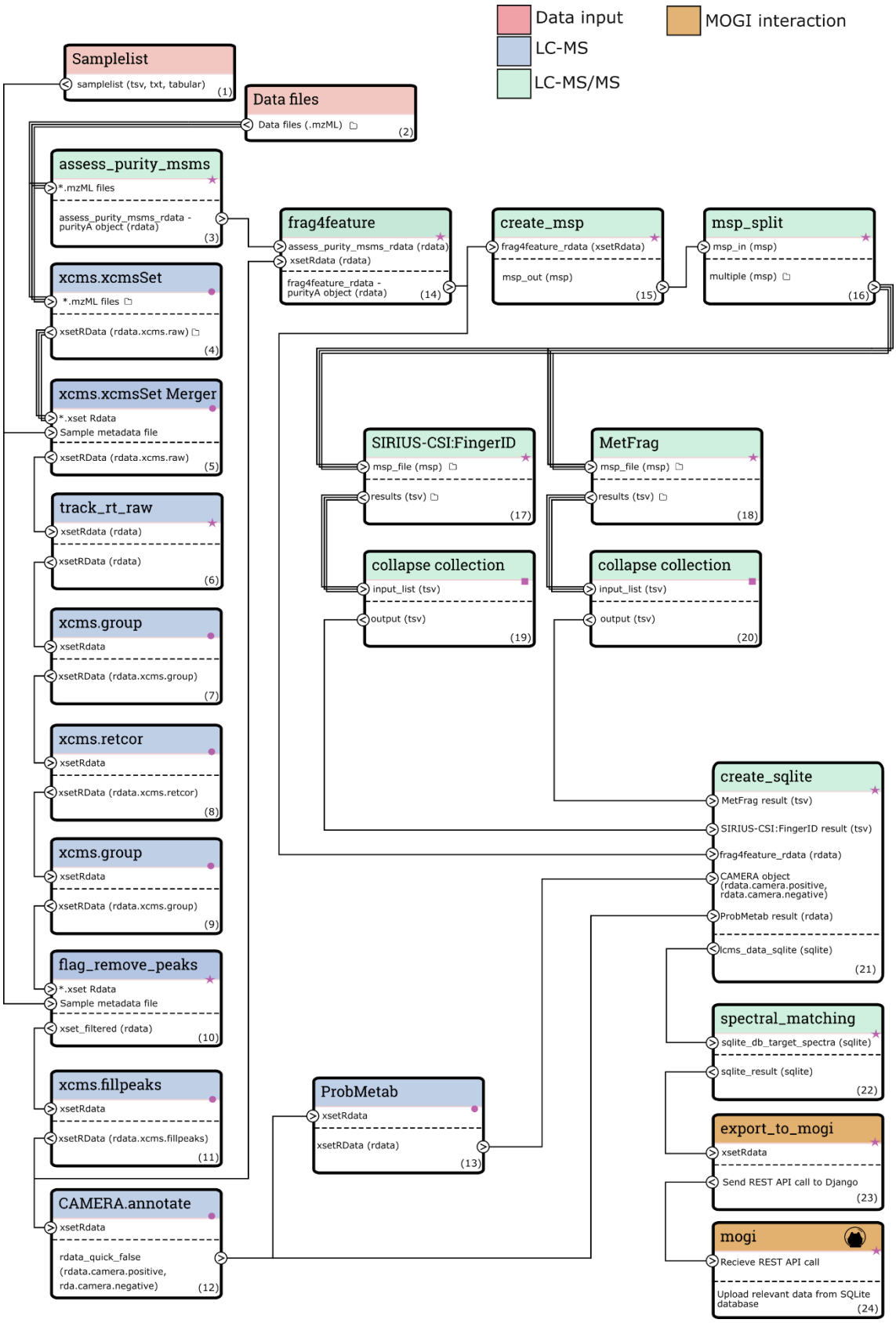


Figure 5-6 LC-MS/MS Galaxy annotation workflow

In depth schema for Galaxy workflow for LC-MS/MS annotation – for basic overview see figure 5-1.

(1) Sample list identifying the class for each mzML file and identifying which class to be used for the blank. (2) Data files (mzML) used in the study, these can be either loaded from the Galaxy history or Galaxy data library. (3) Using the tool "assess_purity_msms", the precursor ion purity of the fragmentation spectra is assessed. (4) Using the tool "xcms.xcmsSet", perform XCMS feature detection on each file. This is done via a data collection object in Galaxy where each file is run independently. (5) Using the tool "xcms.xcmsSet Merger", the individual features that have been detected for each file are merged along with a samplelist to determine any sample classes for each file. (6) Using the tool "track_rt", the raw retention time is tracked – this information can also be retrieved within the xcmsSet object after retention time correction, but the information is saved here for each peak for both convenience and troubleshooting purposes. (7) Using the tool "xcms.group", individual peaks are grouped across files. (8) Using the tool "xcms.retcor", retention time correction is performed. (9) Using the tool "xcms.group", regrouping is performed, a requirement following retention time correction. (10) Using the tool "flag_remove_peaks", features that do not pass relevant criteria for various thresholds are flagged (e.g. fold change between sample and blank, relative standard deviation of retention time or intensity), the features can then be removed from the xcmsSet object, regrouped and checked again. This process is repeated until only peaks within acceptable thresholds are present within the xcmsSet object. (11) Using the tool "xcms.fillpeaks", missing peaks are 'filled' by checking the region of interest in the mzML files. (12) Using the tool "CAMERA.annotate", adducts and isotopes are annotated. (13) Using the tool "ProbMetab" XCMS grouped features are annotated with metabolites (this does not use any fragmentation spectra). (14) Using the tool "frag4feature", fragmentation spectra are assigned to XCMS grouped features. (15) Using the tool "create_msp", an MSP file is created for all the fragmentation spectra associated with the XCMS grouped features. (16) To optimise the parallelization of processes on Galaxy it is often useful to run datasets as "collections". For this reason, the "msp_split" tool is used to split the single MSP file into multiple smaller MSP files stored in a data collection (e.g. one MSP file containing 10,000 spectra can be split into 100 MSP containing 100 spectra each). (17) Using the tool "SIRIUS-CSI:FingerID", the SIRIUS CLI (command line interface) is run for all spectra within each MSP file. (18) Using the tool "MetFrag", the MetFrag CLI tool is run for all spectra within each MSP file. (19, 20) Using the tool "Collapse collection" the tab separated results files generated from either MetfFrag or SIRIUS-CSI:FingerID are collapsed into a single tab separated file. (21) Using the tool "create_sqlite", the results from frag4feature, CAMERA.annotate, SIRIUS-CSI:FingerID, MetFrag and ProbMetab are combined into a single SQLite database. (22) Using the tool "spectral_matching", spectral matching of all collected fragmentation spectra is performed against library spectra. The results can either update the original SQLite database produced at the previous step or a new SQLite database can be created. (23) Using the tool "submit_to_mogi", a REST API signal is sent to the django-mogi project. (24) The user confirms the upload to the django-mogi project.

The colour coding represents process in the workflow. Triple lines connecting tools represent the use of

a Galaxy data collection. The circle symbol indicates if the tool originated from Workflow4metabolomics, the square indicates if the tool originated from standard Galaxy tools and the star symbol indicates if the tool was developed as part of this chapter.

5.4.5 DJANGO-MBROWSE

5.4.5.1 General

The django-mbrowse application provides functionality to view, search and browse metabolomics data. Currently the focus is for mass spectrometry spectra and metabolite annotation results.

5.4.5.2 Data input

The application currently accepts data from LC-MS and LC-MS/MS datasets and that can be generated programmatically through the msPurity R package (Bioconductor) or through the associated tool within Galaxy (create_sqlite). The generated input file is a collated output from the results of XCMS (Smith et al., 2006; Tautenhahn et al., 2008), CAMERA (Kuhl et al., 2012), msPurity, ProbMetab (Silva et al., 2014), MetFrag and Sirius CSI:FingerID in an SQLite database. The SQLite database can then be easily read with both R and Python as opposed to programming language specific data objects. The structure also has the benefit of being able to store the large results generated and their relationships instead of handling multiple text files. Additionally, the schema used is flexible enough for any other LC-MS peak processing and annotation software to be incorporated into the output if required.

The SQLite input file can be uploaded manually or if django-mogi is installed can be uploaded via a REST (representational state transfer) API (application programming interface) from Galaxy to Django. See Figure 5-6 for Galaxy workflow.

5.4.5.3 Summarising metabolite annotation results

The metabolite annotation results are summarised from the following pieces of annotation software: MetFrag, Sirius CSI:FingerID, msPurity spectral-matching and ProbMetab. ProbMetab is an R package for annotation, that searches a hypothetical neutral mass determined from MS^1 LC-MS data to a compound database (e.g. KEGG (Kanehisa et al., 2014)) and uses a Bayesian model where known biochemical reactions have been used to determine a prior distribution, providing a probability score for each metabolite annotation. See section 5.4.4, for details of MetFrag, Sirius CSI:FingerID and msPurity.

All the annotations across the multiple approaches are aligned via matching IUPAC International Chemical Identifier hash key (InChIKey) (Heller and McNaught, 2009). InChiKeys are 27-character string to (nearly) uniquely identify a chemical substance whilst still being readily interpretable by computers and easily searched, see Figure 5-7 for an example.

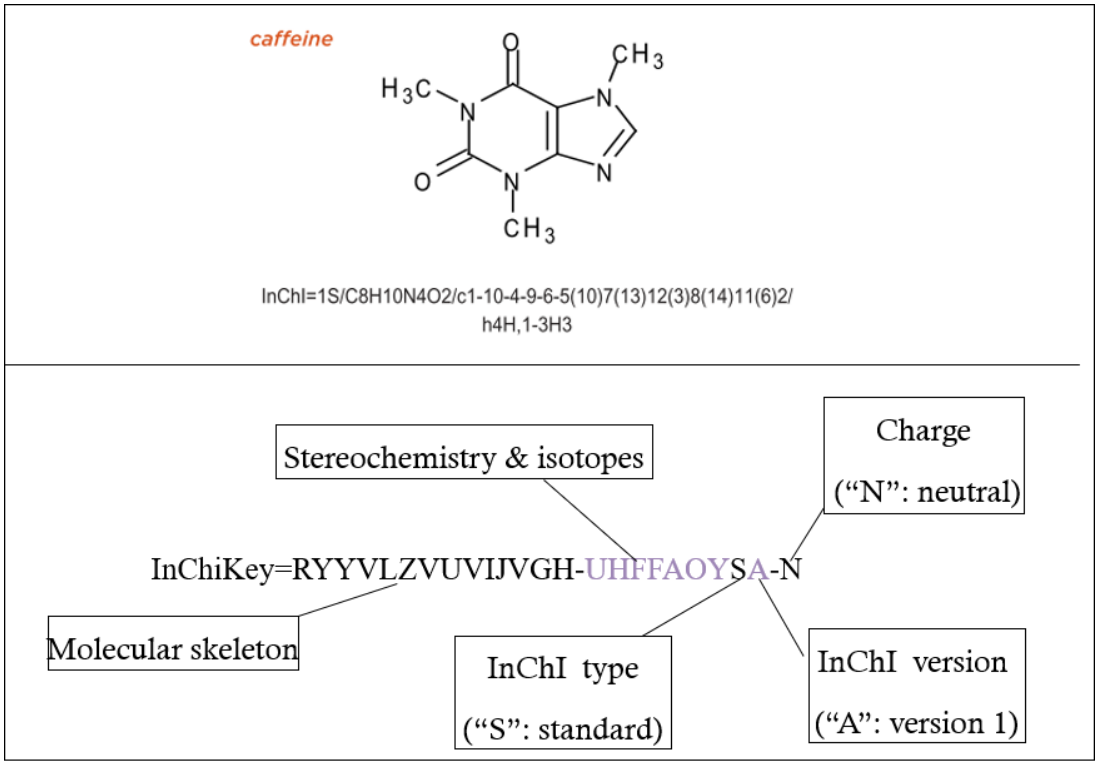


Figure 5-7 Example InChIKey structure (modified from Alan, 2009)

A) Representation of caffeine with InChI and chemical diagram. B) InChIKey representation of caffeine.

The initial InChIKey compound database used for django-metab was extended from (Madhavan et al., 2014) to include all the InChIKeys in PubChem. This allows compounds annotated across different tools to be integrated if an InChIKey can be determined.

When an annotation has been aligned via an InChiKey, an overall score is given by scaling the annotation score of each approach between 0 and 1. Each annotation approach can then be weighted, where more weight can be given to a more reliable approach. The annotation score where the sum of the scores is closest to 1 is then considered the "best" ranked annotation. The approach can help a user determine what is the most realistic metabolite annotation. See below for an example.

Table 5-2: Example metabolite annotation summary

Compound (InChIKey)	**Spectral matching**	**MetFrag**	**Sirius CSI:FingerID**	**ProbMetab**	**Overall weighted score**
	weight: 0.30	weight: 0.25	weight 0.25	weight: 0.20	
2-(Propylamino)ethanol BCLSJHWBDUYDTR-UHFFFAOYSA-N	0.8	1	0.4	0.2	(0.8 * 0.3) + (1 * 0.25) + (0.4 * 0.25) + (0.2 * 0.2) = 0.630
N-propylaminoethanol YLLVPHVZPCGVQN-UHFFFAOYSA-N	0.3	1	0.4	0.6	(0.3 * 0.3) + (1 * 0.25) + (0.4 * 0.25) + (0.6 * 0.2) = 0.560
N-isopropyl aminoethanol LZAMLBFWTLFDKM-UHFFFAOYSA-N	0.5	1	0.1	0.2	(0.5 * 0.3) + (1 * 0.25) + (0.1 * 0.25) + (0.2 * 0.2) = 0.465

5.4.5.4 Browsing and searching spectra and annotations

Once the input data has been uploaded and extracted into the Django project database there are various option to browse, view and search the spectra and metabolite annotations.

Browsing the LC-MS(/MS) datasets directly allows a user to view the extracted ion chromatogram, fragmentation spectra, peak information (*m/z*, retention time) and metabolite annotation results for each peak of a dataset. Alternatively, all the metabolite annotations can be viewed in a single table. See Figure 5-8 (part A) for demonstration of user experience.

Data can be searched either by individual *m/z and* neutral mass values, or via fragmentation spectra using spectral matching.

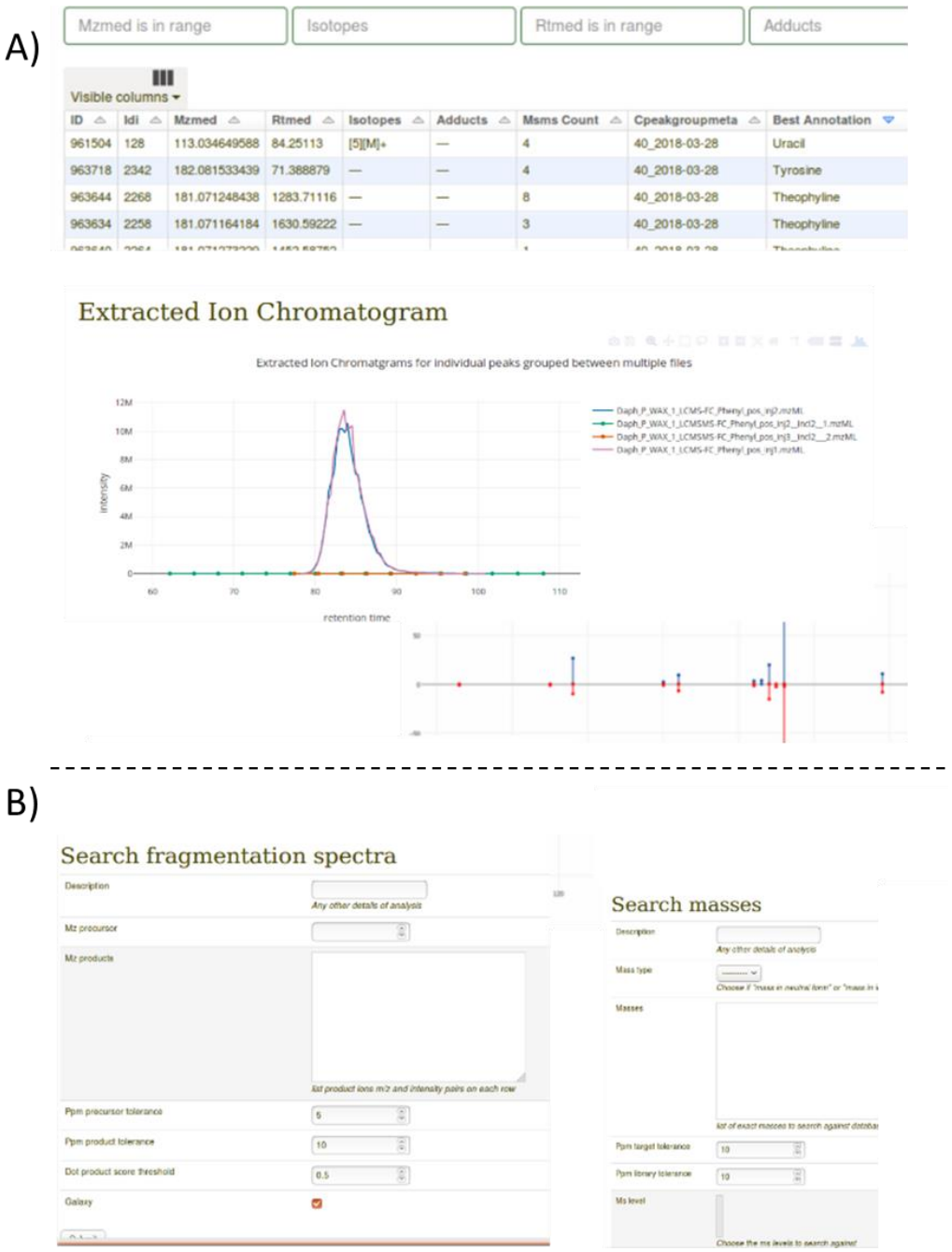


Figure 5-8 Browsing and searching metabolomics data with django-metab

A) User interface showing how a data table of spectra associated with an assay can be viewed. The chromatographic peak and fragmentation spectra (and any corresponding matching spectra) can also be viewed. **B)** The fragmentation spectra, m/z and neutral mass can also be searched using the user interface.

5.4.5.5 *m/z* and neutral mass search

The *m/z* search will match a query *m/z* value to either the chromatographic peak (e.g. the XCMS picked peak) or the *m/z* values in the fragmentation spectra. The neutral mass search will match one or more query neutral mass values to the pre-calculated neutral mass values determined by CAMERA for the chromatographic peaks. Both the *m/z* and neutral mass search allow a parts-per-million (ppm) error tolerance to be set for the 'library' spectra (e.g. data that is stored within the database) and the query values, as the error tolerance is potentially different for both. It should be noted that this approach is slightly different to searching against a reference compound library, where the mass of a reference compound would have an error tolerance of zero. See Equation (5.1) to (5.5) where qmz is the query *m/z*, lmz_i is the *m/z* of an individual library peak, $qtol$ is the query ppm error tolerance and $ltol$ is the library ppm error tolerance.

$$qt_{low} = qmz - (qmz \times 10^{-6}) \times qtol \quad (5.1)$$

$$qt_{high} = qmz + (qmz \times 10^{-6}) \times qtol \quad (5.2)$$

$$lt_{i_{low}} = lmz_i - (lmz_i \times 10^{-6}) \times ltol \quad (5.3)$$

$$lt_{i_{high}} = lmz_i + (lmz_i \times 10^{-6}) \times ltol \quad (5.4)$$

$$H_{peaks} = \{ Library\ peaks\ where\ (qt_{high} \geq lt_{i_{low}})\ \&\ (lt_{i_{high}} \geq qt_{low}) \} \quad (5.5)$$

5.4.5.6 Fragmentation search

Spectral matching of the fragmentation spectra can be searched via the user interface. The spectral matching process can be broken down into 3 components: spectral filtering, spectral alignment and the spectral matching.

The **spectral filtering** stage determines what spectra from the library to be searched against. The *m/z* of the precursor ion of the query to the library spectra is compared with the *m/z* of the precursor ion of the library spectra, tolerances calculated as per Equation (5.1) to (5.4) where qmz is the query precursor *m/z*, lmz_i is the precursor *m/z* of an individual library spectra, $qtol$ is the query ppm error tolerance and $ltol$ is the library ppm error tolerance. A collection of filtered library spectra can be determined using (5.6).

$$H_{spectra} = \{\, Library\ spectra\ where\ (qt_{high} \geq lt_{i_{low}})\ \&\ (lt_{i_{high}} \geq qt_{low}) \qquad (5.6)$$

The precursor ion purity score can also be used as a filter. Alternatively, the fragmentation of all spectra in the library can be searched regardless of the precursor ion information.

The **spectral alignment** of the fragmentation (i.e. MS^2) query peaks to library fragmentation peaks is performed using a similar approach to the modified pMatch algorithm described in (Zhou et al., 2014). The aim of an alignment is to obtain two sets of corresponding peaks from the library where they have been aligned together minimising the differences between the spectra, see Table 5-3 for an example alignment.

Table 5-3 Example alignment of mock library and query spectra

Peak id	mz_Q	$intensity_Q$	mz_L	$intensity_L$
1	500.232	100	500.233	100
2	248.431	50	248.431	47
3	Missing	missing	140.42	10
4	120.312	6	120.313	8
5	102.130	10	missing	missing
6	100.210	5	100.210	15

Footnote: Mock peaks for both library and query peaks showing an ideal alignment based on an initial assessment of similarity between spectra. Where mz_Q is the query *m/z* values $intensity_Q$ is the query intensity values, mz_L is the library spectra *m/z* values and $intensity_L$ is the intensity values of library spectra.

An alignment is made for the query spectra against every library spectra within $H_{spectra}$. For each alignment (i.e. one query spectra vs one library spectra) the query peaks values and the library peak values are sorted in descending order of intensity. Starting with the most intense query peak, all library peaks within a defined ppm tolerance between the library and query peak are filtered. The library peak that has the closest difference in relative abundance is then selected as the best match for the query peak, the process is first repeated for all the query peaks. Once a match has been made the peak cannot be used for any other match. If a match cannot be found for either a library or query peak a 'missing' flag should be recorded at the peak position. See Table 5-2.

The **spectral matching** is then performed on a weighted vector of the library query spectra. See (5.7) where x and y represent weight factors as per the mass spectra MassBank (Horai et al. 2010) approach the x = 0.5 and y = 2.

$$\boldsymbol{w} = [\boldsymbol{peak_intensity}]^{\boldsymbol{x}} \cdot [\boldsymbol{mass(m/z)}]^{\boldsymbol{y}} \qquad (5.7)$$

The weighted values are then used in a dot-product cosine similarity calculation, see (5.8) where $\mathbf{w}_Q$ is the weighted values of the query and $\mathbf{w}_L$ are the weighted values of the library. See Table 5-4 for example calculation.

$$dpc = \frac{\sum \boldsymbol{w}_Q \cdot \boldsymbol{w}_L}{\sqrt{\sum \boldsymbol{w}_Q^2} \cdot \sqrt{\sum \boldsymbol{w}_L^2}} \quad (5.8)$$

Table 5-4 Weighting library and query spectra

Peak id	mz_Q^2	$intensity_Q$^0.5	mz_L^2	$intensity_L$ ^0.5	w_Q = mz_Q^2 * $intensity_Q$^0.5	w_L = mz_L^2 * $intensity_L$^0.5
1	250232.1	10.00	250233.1	10.00	2502320.54	2502330.54
2	61718.0	7.07	61718.0	6.86	436411.89	423117.03
3	Missing	Missing	19717.8	3.16	0	62353.08
4	14475.0	2.45	14475.2	2.83	35456.31	40942.10
5	10430.5	3.16	Missing	Missing	32984.25	0
6	10042.0	2.23	100042.0	3.87	22454.69	38892.67

Footnotes: This table uses the aligned spectra from Table 5-2. The dot-product cosine using Equation (5.8) is 0.9996

5.4.5.7 Database models

Figure 5-9 shows a general relationship between a Chromatographic grouped feature, the input dataset, metabolomics data files and individual chromatographic features. Full model description for django-mbrowse including details of fragmentation spectra, metabolite annotation, library spectra and isotope-adduct relationships detailed in the appendix section 8.17.

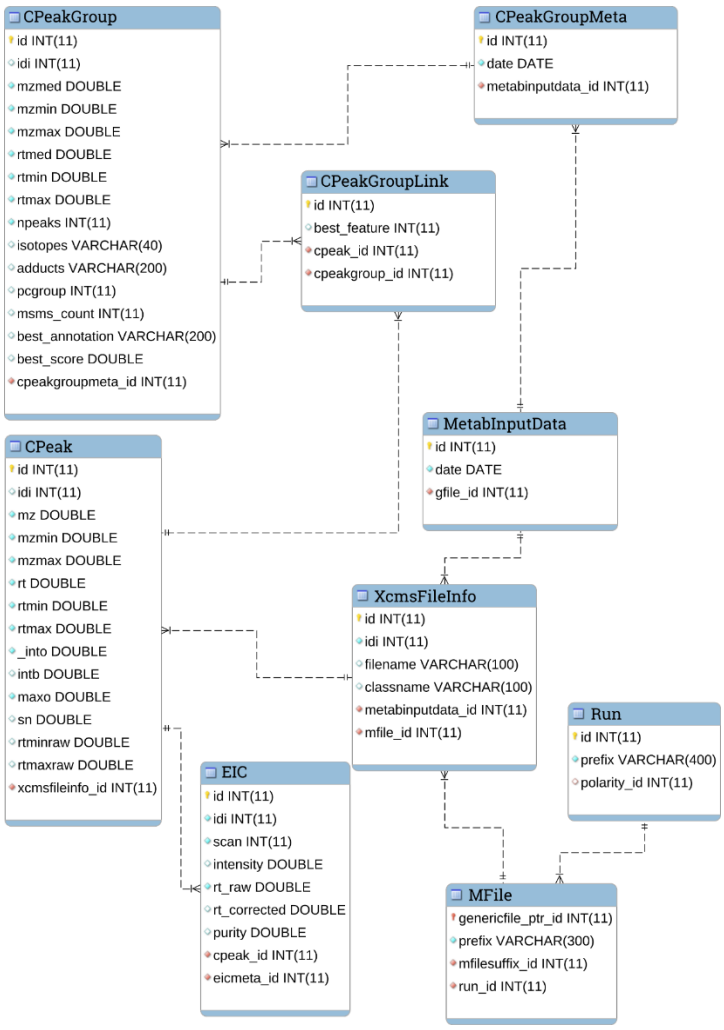


Figure 5-9 Model schema for chromatographic features (Django-mbrowse)

Models for chromatographic features showing the relationship between the LC-MS input file, the metabolomics data files, the chromatographic grouped features and the individual chromatographic features. "MetabInputDataset" represents an uploaded LC-MS dataset, "CPeakGroupMeta" represents the analysis performed to create the grouped chromatographic features. "CPeakGroup" represents the grouped feature. Each "CPeakGroup" has associated individual peak features represented by the "CPeak" model, each "CPeak" model can have an extracted ion chromatogram represented by the "EIC" model. Each "CPeak" model can be traced back to the metabolomics data file used via the "XcmsFileInfo" model to the "MFile" (representing the metabolomics file, e.g. mzML) and the associated "Run" model

5.4.6 USER GUIDE

The documentation (http://mogi.readthedocs.io/en/latest/) has both an API documentation and a user guide for an end user (not a developer) to explain how to install and use a Django project created using the MOGI suite of Django applications. The user guide details how use the MOGI suite for managing ISA projects, interacting with Galaxy and browsing, viewing and searching metabolomics data. The example data used for the user guide was taken from MetaboLights study MTBLS144.

5.4.7 RECOMMENDED INFRASTRUCTURE

The infrastructure to run the suite of applications can be divided into 4 sections. (1) The file system used to store the raw files (and any other large files). (2) The SQL database to store all other data and meta data. (3) The Python-Django server for providing the web-interface and interactions with the SQL database and Galaxy. (4) The Galaxy server ideally set up so that it is connected to a high-performance computer. See Figure 5-10 for overview.

The Django backend database storage system should be either MySQL or MariaDB. Large files are not typically loaded into SQL database structures due to performance and space limitations. For this reason, the metabolomic raw files are referenced within the database and can either by uploaded to designated file system and a mapping to file-path is saved or a file-path to accessible file system can be used.

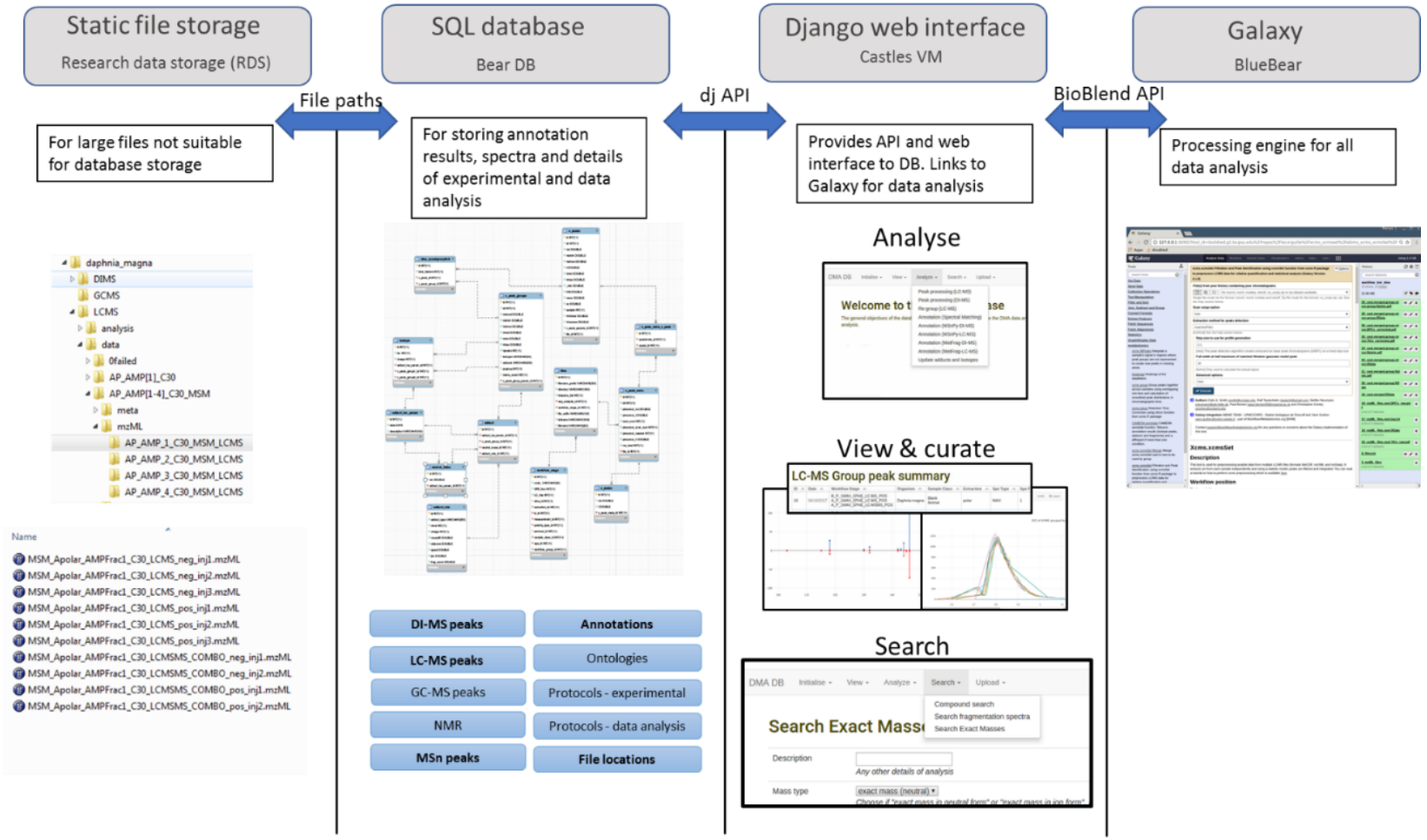


Figure 5-10 Recommended infrastructure

The figure details the recommended infrastructure to use with the MOGI suite of Django applications. The figure also highlights the implementation used at the University of Birmingham. Raw data and large files are stored on a static file storage (e.g. Research data storage at University of Birmingham). An SQL database stores all other information e.g experimental information (e.g. ISA), results of the analysis, peak information, details of Galaxy users and instances. For the University of Birmingham this consists of the MariaDB SQL database (BearDB). The database being the backend of the Django interface. The interface for the user is provided via HTML and Javascript views on the data created using Django and connects to the database via the standard Django API. The Django instance is running on a CaStLeS virtual machine at the university of Birmingham. Finally the Django interface is connected via the BioBlend API to a Galaxy instance that can be sent files for analysis and upload the results back to the Django instance. The University of Birmingham Galaxy instance is running on the BlueBear high performance computer.

5.5 METHODS

5.5.1 EXISTING GALAXY TOOLS USED

The chapter made use of Galaxy tools that have been previously developed either through Workflow4Metabolomics or were available freely from the Galaxy toolshed.

Table 5-5 Overview of existing Galaxy tools

Galaxy tool name	Underlying tool name	Language	Description
xcms.xset	XCMS	R	Data processing tool for feature detection of mass spectrometry datasets
xcms.xcmSet Merger	XCMS	R	Data processing tool to merge multiple xcms xcmsSet data objects into a single object
xcms.group	XCMS	R	Data processing tool for grouping mass spectrometry peaks between different files
xcms retcor	XCMS	R	Data processing tool for correcting retention time shifts in LC-MS data
xcms fillpeaks	XCMS	R	Data processing tool for "filling" missing peaks that by re-assessing the mzML data file
CAMERA.annotate	CAMERA	R	Annotation tool for adducts and isotopes of LC-MS data
ProbMetab	ProbMetab	R	Annotation tool that searches a hypothetical neutral mass determined from MS^1 LC-MS data to a compound database and uses a Bayesian model where known biochemical reactions to determine a prior distribution is used to calculate a probability for each potential metabolite annotation
Collapse Collection	awk	NA	Data wrangling tool for combining data collections into a single dataset

5.6 AVAILABILITY OF SUPPORTING SOURCE CODE AND REQUIREMENTS

- Project name: Django MOGI application suite
- Project home page: http://mogi.readthedocs.io/en/latest/
- Operating system(s): Linux
- Programming language: Python 3.6 and Python 2.7
- Other requirements: Celery, Redis, RabbitMQ and MySQL, Galaxy.
- License: GNU GPL 3.0

5.7 DISCUSSION

Current software to organise metabolomics data and experimental information (MAStr-MS and PIMP) have no Galaxy interfacing and have limited ISA interactions. The Django MOGI application suite provides a collection of applications to organise metabolomics data and experiments using ISA and Galaxy with the aim to further encourage metabolomics experimentation to become FAIRer. Additionally, as the Galaxy and ISA functionality has been developed as reusable Django applications, the functionality is not limited to a single web project can be potentially integrated into existing systems and projects.

Essential further work is required to improve the Docker for the example Django project and Galaxy instance. Additionally, all new Galaxy tools should be migrated from the Galaxy test-toolshed to the stable Galaxy toolshed (the Galaxy toolshed host Galaxy tools in a way that makes installing the tools easy and traceable). The django-misa application also requires further updates to support ISA more extensive export and import functionality. Documentation of use cases with example data should also be included to further demonstrate the MOGI applications.

Further work to provide additional functionality could include updating the django-mbrowse and django-misa applications to support DIMS (direct infusion mass spectrometry) and NMR

(nuclear magnetic resonance) spectroscopy metabolomics data. Also, to develop more distinct public and private separation of ISA projects for user groups (e.g. to hide certain projects that should not be shared) and to provide functionality to submit a finished ISA project to the MetaboLights data repository. The performance of the annotation should also be assessed by running the Galaxy annotation workflow on dataset that contains a mix of chemical standards – allowing a false positive/negative rate to be calculated for the workflow. This could also be used to determine appropriate weights for each annotation tool score. Future work should also be done to provide a user testing period where a working website is trialled by a subset of users that complete surveys on usability and feedback on where improvements are required. Future long-term maintenance for web-based tools and application can be challenging due to the fast-changing nature of web-development. However, as all code is available publicly and made as modular packages – it will hopefully encourage others to contribute as an open source resource.

The MOGI suite has subsequently been used to create the DMAdb (a web-server and database) for handling the DMA of *D. magna* where the Galaxy annotation workflow described in this chapter was also used for the metabolite annotation data processing, (see Chapter 6).

Chapter 6

Data analysis and data management for the deep metabolome annotation of *Daphnia magna*

6.1 CHAPTER SUMMARY

The Deep Metabolome Annotation (DMA) of *Daphnia magna* is an on-going project to characterise the metabolites in a pooled sample of *D. magna* using multiple analytical separation and measurement technologies. In this chapter, I detail how the (Metabolomics Organisation with Galaxy and ISA) MOGI suite of Django applications detailed in Chapter 5 have been used to create a searchable web interface (DMAdb) to manage and analyse the DMA of *D. magna.* I then summarise the compounds that have been annotated to date for the DMA of *D. magna,* where from 23 liquid chromatography tandem mass spectrometry experimental assays a total of 5,469 high quality level 2 metabolite annotations have been obtained.

6.2 CONTRIBUTIONS

Thomas N. Lawson[1], Martin R. Jones[1], Ralf J. M. Weber[1], Warwick B. Dunn[1,2], Mark R. Viant[1,2]

[1]School of Biosciences and [2]Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

TNL developed the DMAdb and performed the data analysis presented in this chapter. Nearline workflows were developed between TNL, MRJ and RJMW (see Chapter 3). The LC-MS(/MS)

annotation workflow used was developed between TNL and MRJ (see Chapter 5 for full details). MRJ performed experimental analysis, including animal husbandry, and developed the overall experimental design of the DMA workflow. MRV and WBD supervised all work carried out.

6.3 PUBLICATION HISTORY

The content of this chapter has not been published. The information is intended to be included into a future publication that encompasses all the work carried out for the DMA of *D. magna*.

6.4 INTRODUCTION

The lack of knowledge on the count and composition of commonly studied metabolomes limits the biological interpretation of an untargeted metabolomics analysis. However, to date, there has been no extensive experimental characterisation of a metabolome due to the analytical complexities of measuring the wide physicochemical properties and concentration ranges exhibited by an organism's metabolome (Viant et al., 2017). This chapter details the data analysis and data management used for the ongoing "Deep Metabolome Annotation" of the model organism *Daphnia magna*, where extensive separations, analytical measurements and computational techniques have been used to annotate as many metabolites as possible in a biological sample of interest and thus increasing our knowledge of the metabolome of *D. magna*.

The DMA of *D. magna* is the first of what is hoped to be many organisms to be characterised by a DMA (or DMA like) experimental workflow. The benefits of increasing the metabolome knowledge of *D. magna* is of particular importance due to its role as an ecotoxicological model used in regulatory standards (SOR/2002-222 - Canadian Metal Mining Effluent) and its use as a model used for evolutionary biology (Ebert, 2005) and human health (NIH recognised model

organism).

Identification of metabolites is the often cited “bottle neck” (Dunn et al., 2012; Wishart, 2011) in metabolomics and the challenge of metabolite annotation is reflected in the lack of knowledge of model organism metabolomes. Different analytical and computational methods provide varying levels of reliability of a metabolite identification. The Metabolomics Standard Initiative (MSI) clarified this into 4 identification levels:

- Level 1 identification requirements differ depending on if the metabolite is novel or non-novel. For non-novel metabolites this requires a chemical standard matching the metabolite of interest on the same instrument (ideally in the same laboratory) that the observed metabolite has been measured on. A level 1 identification of a novel metabolite requires considerably more work, requiring isolation and purification of the metabolite and the use of multiple measurement technologies and computational techniques to elucidate the compound.
- Level 2 identification is a putative annotation and does not require a chemical standard but would rely on the similarities to physicochemical properties of a compound and spectral similarity to compounds in external spectral libraries.
- Level 3 identification is an annotation of the class of a compound by physiochemical properties or spectral similarity.
- Level 4 identification is an unknown compound.

The DMA approach described here focuses on achieving MSI level 2 identifications (e.g. putative annotations) by using mass spectrometry gas-phase fragmentation to annotate both novel and non-novel metabolites. I define here non-novel metabolites as compounds that are either present in metabolite databases (e.g. KEGG, HMDB) and/or have spectral libraries available (e.g. from MassBank, mzCloud). I define novel metabolites as compounds that are

not in a metabolite database and are not available in spectral libraries.

The reliance of using chemical standards has intentionally been avoided for two reasons: 1) The available chemical standards do not cover the entirety of any metabolome. 2) Using chemical standards requires prior knowledge about what is expected in the metabolome.

Although not detailed here, gas chromatography mass spectrometry (GC-MS) and 1D, 2D nuclear magnetic resonance (NMR) spectroscopy has also been carried out on a subset of the DMA of *D. magna* providing the potential for high level 2 (or low level 1) identification of metabolites. The work described in this chapter is focusing on the liquid chromatography tandem mass spectrometry (LC-MS/MS) component of the DMA of *D. magna.*

Multiple aspects of this thesis have been used as part of the DMA of *D. magna*, this includes: Computational workflows and experimental workflows described in Chapter 3 to acquire the mass spectrometry (MS) fragmentation data. Precursor ion purity calculation and data processing software detailed in Chapter 2 have been used for mass spectrometry data analysis. Tools and workflows detailed in Chapter 5 to manage and analyse metabolomics experiments have been used to create the deep metabolome annotation database (DMAdb).

6.5 METHODS

6.5.1 EXPERIMENTAL

6.5.1.1 DMA Experimental workflow

Figure 6-1 summarises the experimental workflow of the DMA of *D. magna.*

6.5.1.2 Daphniid culturing

Daphnia magna from ten genetically distinct strains were cultured for at least 14 days to adulthood in a modified high-hardness combination medium. The green algae *Chlorella vulgaris* was added at regular occurring intervals. An additional two days of culturing was then

performed where no additional algae were added. For these final two days of culture, half of the *Daphnia* from each strain was cultured under normal conditions (20 ºC, day 1 – 16:8 light:dark, day 2 -8:6) and half under stress conditions (10 ºC, day1 - 16:8 light:dark, day2 - 8:16 light:dark).

6.5.1.3 Pooling and homogenisation

Following the 2 days of culturing without algae, the *Daphnia* were filtered from their growth environment washed in fresh culturing medium, blotted dry, transferred to microfuge tubes, submerged in liquid nitrogen to quench metabolism and stored frozen at -80 °C.

Samples were submerged in liquid nitrogen and homogenised using a pestle and mortar over dry ice prior to metabolite extraction. The homogenised *Daphnia* was split into equal parts for the polar and apolar arms of the workflow.

6.5.1.4 Metabolite extraction

Apolar metabolite extraction was performed as per the Bligh & Dyer extraction procedure (Bligh and Dyer, 1959). Polar metabolites were extracted using 2.5:1 v/v methanol:water (4 °C) extraction solvent comprising at a ratio of 15 parts solvent per 1 part biological wet mass. Blank samples were prepared following the same steps but without any biological material. *Daphnia* samples were mixed and centrifuged (10 minutes, 3000 rpm, 4 °C) and supernatant divided into microfuge tubes where each tube approximates to 20 *Daphnia*.

6.5.1.5 Solid phase extraction

Polar samples homogenate and undergo two distinct solid phase extraction (SPE) method where each SPE extraction procedure produced four fractions. See below for polar SPE details:

- WCX - Weak cation-exchange SPE cartridge (HyperSep WCX, 3 mL, 500 mg).
- WAX - Weak anion-exchange SPE cartridge (HyperSep WAX, 3 mL, 500 mg).

Apolar samples homogenate also undergo two SPE methods (C18 and AMP). The C18 method produced 3 fractions and the AMP method produced four fractions. See below for apolar SPE details:

- C18 - alkyl-bonded SPE cartridge (HyperSep C18, 3 mL, 500 mg)
- AMP - Weak anion-exchange SPE cartridge (HyperSep WAX, 3 mL, 500 mg). It should be noted that this is the same cartridge as the WAX SPE. The AMP id was given to distinguish the two methods.

6.5.1.6 Chromatography

Liquid chromatography was then performed on the SPE fractions. For the polar SPE fractions resulting from either the WAX or WCX SPE, hydrophilic interaction chromatography (HILIC) was performed using a Syncronis Phenyl (PHE) column and reverse phase chromatography (RPC) was performed using an Accucore Amide column (AMD). For the apolar SPE fractions, RPC was performed using an Accucore C30 column (C30). Full details of chromatography methods can be found in Appendix 8.19.

6.5.1.7 Mass spectrometry measurements

Mass spectral detection of the liquid chromatography eluent was performed using an Orbitrap Q Exactive mass spectrometer equipped with H-ESI II source, operating under the following parameters: 70000 resolution (FWHM at 200 m/z); 3e6 AGC target; 100-1000 m/z scan range. The mass spectrometer was performed in both positive and negative ionisation mode. The nearline procedure described in Chapter 3 (adapted-LC-MS-nearline) was used to generate target lists to guide the LC-MS/MS fragmentation acquisition.

The eluant from each chromatography column was then fractionated into 96 well plates where more extensive fragmentation analysis is performed using direct-infusion mass spectrometry with gas-phase fragmentation, including multi-stage fragmentation (DI-MS/MS and DI-MS^n).

The nearline procedure described in Chapter 3 (nearline-DIMS) was used to generate target lists to guide the DI-MS/MS and DI-MS^n fragmentation acquisition.

It should be noted that the DI-MS, DI-MS/MS and DI-MS^n data is in the process of being acquired for the DMA of *D. magna* but at the time of writing the full dataset had not been collected. As the data analysis workflow has not been formally developed for the DI-MS/MS and DI-MS^n and the data has not been fully acquired, the data analysis is not reported in this chapter.

6.5.1.8 Experimental Assays

The data used for this chapter is derived from 24 separate experimental assays (and 4 experimental repeats). See

Assay name	LPE	SPE	SPE frac	Chromato-graphy	Chrom type
AP_AMP_1_C30	Apolar	Aminopropyl	1	Accucore C_{30}	RPC
AP_AMP_2_C30	Apolar	Aminopropyl	2	Accucore C_{30}	RPC
AP_AMP_3_C30	Apolar	Aminopropyl	3	Accucore C_{30}	RPC
AP_AMP_4_C30	Apolar	Aminopropyl	4	Accucore C_{30}	RPC
AP_C18_1_C30	Apolar	Syncronis C_{18}	1	Accucore C_{30}	RPC
AP_C18_2_C30	Apolar	Syncronis C_{18}	2	Accucore C_{30}	RPC
AP_C18_3_C30	Apolar	Syncronis C_{18}	3	Accucore C_{30}	RPC
P_WCX_1_AMD	Polar	Weak Cation-eXchange	1	Accucore Amide	HILIC
P_WCX_2_AMD	Polar	Weak Cation-eXchange	2	Accucore Amide	HILIC
P_WCX_3_AMD	Polar	Weak Cation-eXchange	3	Accucore Amide	HILIC
P_WCX_4_AMD	Polar	Weak Cation-eXchange	4	Accucore Amide	HILIC
P_WAX_1_AMD	Polar	Weak Anion-eXchange	1	Accucore Amide	HILIC
P_WAX_2_AMD	Polar	Weak Anion-eXchange	2	Accucore Amide	HILIC
P_WAX_3_AMD	Polar	Weak Anion-eXchange	3	Accucore Amide	HILIC
P_WAX_4_AMD	Polar	Weak Anion-eXchange	4	Accucore Amide	HILIC
P_WCX_1_PHE	Polar	Weak Cation-eXchange	1	Syncronis phenyl	RPC
P_WCX_2_PHE	Polar	Weak Cation-eXchange	2	Syncronis phenyl	RPC
P_WCX_3_PHE	Polar	Weak Cation-eXchange	3	Syncronis phenyl	RPC
P_WCX_4_PHE	Polar	Weak Cation-eXchange	4	Syncronis phenyl	RPC

P_WAX_1_PHE	Polar	Weak Anion-eXchange	1	Syncronis phenyl	RPC
P_WAX_2_PHE	Polar	Weak Anion-eXchange	2	Syncronis phenyl	RPC
P_WAX_3_PHE	Polar	Weak Anion-eXchange	3	Syncronis phenyl	RPC
P_WAX_4_PHE	Polar	Weak Anion-eXchange	4	Syncronis phenyl	RPC

Table 6-1.

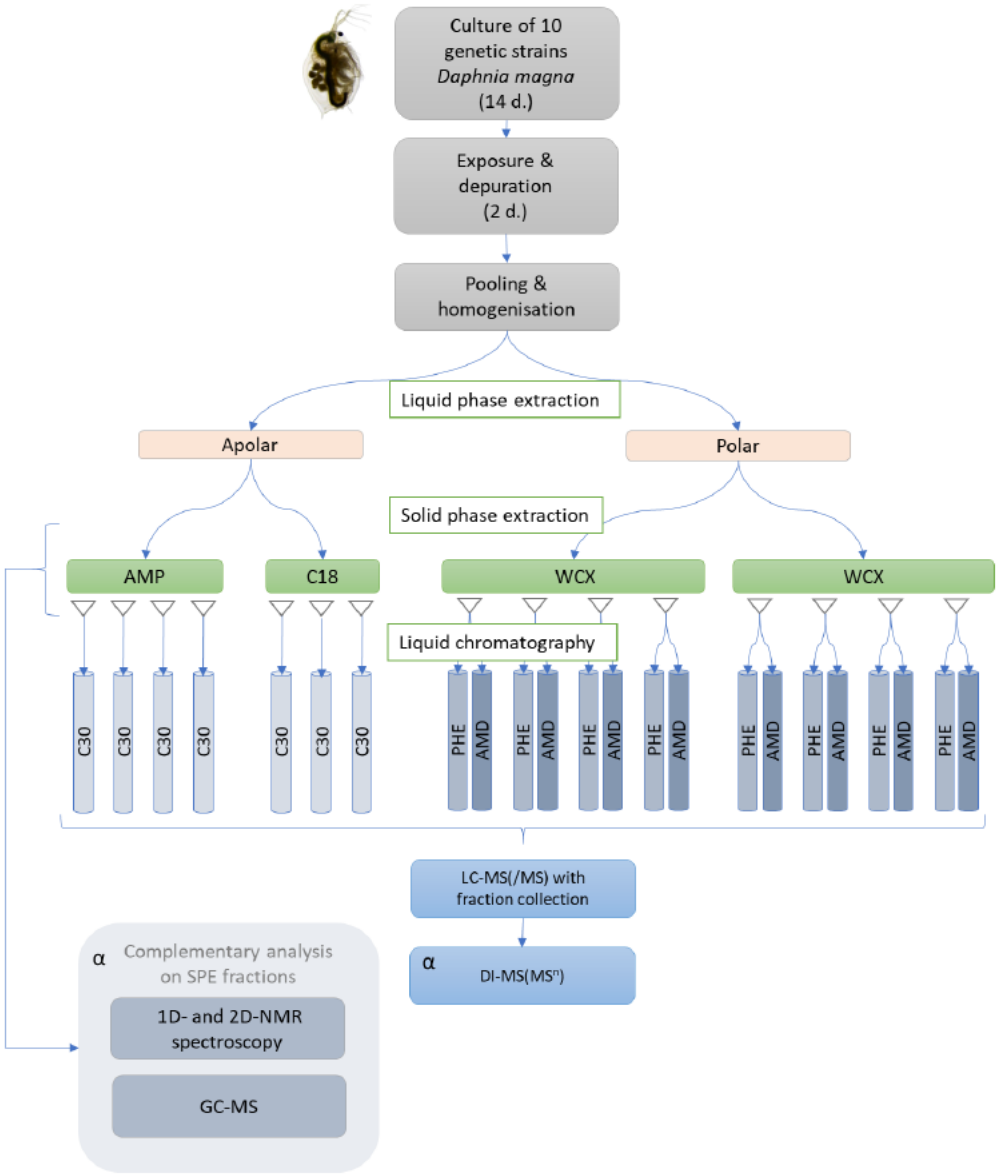


Figure 6-1 DMA of *D. magna* experimental workflow

1) Strains of Daphnia are cultured under controlled conditions for >14 days with periodic additions of algae 2) For 2 days the *Daphnia* are cultured with the absence of green algae under both stressed and normal conditions. 3) The Daphnia are then pooled and homogenised. 3) Apolar and polar metabolites

are then extracted from the homogenate. 4) Solid phase extraction (SPE) is performed on the apolar and polar extracts resulting in a least 3 fractions from each SPE cartridge 5) Liquid chromatography is then performed on each SPE fraction. 6) As the eluent from the chromatography is measured via liquid chromatography tandem mass spectrometry (LC-MS/MS) and the eluent is also fractionated into 96 well plates for later analysis. 7) Each fraction in the plate then undergoes extensive direct infusion mass spectrometry with tandem mass spectrometry and multi stage mass spectrometry (DI-MS/MS and DI-MS^n). α = details of direct infusion mass spectrometry fragmentation (DI-MS(MS^n)); 1D, 2D nuclear magnetic spectroscopy; and gas spectrometry not included in this thesis.

Table 6-1 Summary of experimental LC-MS(/MS) DMA assays

Assay name	LPE	SPE	SPE frac	Chromato-graphy	Chrom type
AP_AMP_1_C30	Apolar	Aminopropyl	1	Accucore C_{30}	RPC
AP_AMP_2_C30	Apolar	Aminopropyl	2	Accucore C_{30}	RPC
AP_AMP_3_C30	Apolar	Aminopropyl	3	Accucore C_{30}	RPC
AP_AMP_4_C30	Apolar	Aminopropyl	4	Accucore C_{30}	RPC
AP_C18_1_C30	Apolar	Syncronis C_{18}	1	Accucore C_{30}	RPC
AP_C18_2_C30	Apolar	Syncronis C_{18}	2	Accucore C_{30}	RPC
AP_C18_3_C30	Apolar	Syncronis C_{18}	3	Accucore C_{30}	RPC
P_WCX_1_AMD	Polar	Weak Cation-eXchange	1	Accucore Amide	HILIC
P_WCX_2_AMD	Polar	Weak Cation-eXchange	2	Accucore Amide	HILIC
P_WCX_3_AMD	Polar	Weak Cation-eXchange	3	Accucore Amide	HILIC
P_WCX_4_AMD	Polar	Weak Cation-eXchange	4	Accucore Amide	HILIC
P_WAX_1_AMD	Polar	Weak Anion-eXchange	1	Accucore Amide	HILIC
P_WAX_2_AMD	Polar	Weak Anion-eXchange	2	Accucore Amide	HILIC
P_WAX_3_AMD	Polar	Weak Anion-eXchange	3	Accucore Amide	HILIC
P_WAX_4_AMD	Polar	Weak Anion-eXchange	4	Accucore Amide	HILIC
P_WCX_1_PHE	Polar	Weak Cation-eXchange	1	Syncronis phenyl	RPC
P_WCX_2_PHE	Polar	Weak Cation-eXchange	2	Syncronis phenyl	RPC
P_WCX_3_PHE	Polar	Weak Cation-eXchange	3	Syncronis phenyl	RPC
P_WCX_4_PHE	Polar	Weak Cation-eXchange	4	Syncronis phenyl	RPC
P_WAX_1_PHE	Polar	Weak Anion-eXchange	1	Syncronis phenyl	RPC
P_WAX_2_PHE	Polar	Weak Anion-eXchange	2	Syncronis phenyl	RPC
P_WAX_3_PHE	Polar	Weak Anion-eXchange	3	Syncronis phenyl	RPC
P_WAX_4_PHE	Polar	Weak Anion-eXchange	4	Syncronis phenyl	RPC

Footnote: LPE = liquid phase extraction, SPE = solid phase extraction, RPC = Reverse phase chromatography, HILIC = Hydrophilic interaction chromatography.

6.5.2 DMA COMPUTATIONAL ANNOTATION WORKFLOW

The data analysis workflow described in Chapter 5 Figure 5-6 was performed in the Galaxy workflow environment. The settings used for the Galaxy workflows are detailed in the Appendix 8.18.

The workflow has four methods of metabolite annotation: 1) Spectral matching using msPurity R package to search against MassBank experimental and LipidBlast *in silico* mass spectrometry spectra. 2) Searching against *in silico* spectra calculated using the combinatorial fragmentor software MetFrag. 3) Using the Sirius CSI:FingerID software that incorporates isotope patterns, fragmentation trees and machine learning to match to molecular fingerprints to predicted fingerprints of the experimental spectra. 4) Using Probmetab to search the hypothetical neutral mass determined from MS^1 LC-MS against the KEGG compound database.

Compounds are aligned across approaches based on InChIKeys (Heller and McNaught, 2009) and a weighted score was calculated as per Chapter 5 section 5.4.5.3, the following weightings were used: spectral-matching: 0.4, metfrag: 0.2, Sirius-CSI:fingerID: 0.2 and Probmetab: 0.1. Although somewhat arbitrary, the weightings used attempt to prioritise all methods that use fragmentation spectra compared to just MS^1 LC-MS (i.e. Probmetab). Additionally, spectral matching has been prioritised over MetFrag and Sirius CSI:FingerID, with the assumption that real experimental spectra (or *in silico* lipid spectra) is more reliable that either the predictions from CSI:FingerID or the *in silico* spectra generated from MetFrag. Where multiple annotations were available for a single LC-MS *m/z* and retention-time based feature, the annotation ranked with the highest weighted score was chosen as the annotation

Three categories of level 2 putative annotations are defined here (L2-A, L2-B and L2-C):

L2-A) Requires the compound to be annotated in at least two out of the four annotation approaches.

L2-B) Requires the compound to be annotated in at least one of the MS/MS based annotation approaches (MetFrag, Sirius CSI:FingerID or spectral-matching).

L2-C) Requires the compound to be annotated in any of the annotation approaches.

Additional criteria included: spectral-matching and MetFrag non-weighted scores above 0.6 were considered valid and only the top 6 Sirius-CSI:FingerID annotations were used (this was due to time constraints of data processing as the Sirius-CSI:FingerID analysis resulted in > 10^6 annotations per experimental assay).

It should be noted that the annotations have not been filtered to remove compounds that are not thought to be endogenous to biological systems. When so little is known about the metabolome of a model organism it is difficult to make this assessment. However, as the Probmetab annotation approach only searches through KEGG compounds (that are predominantly endogenous), the approach described above will have some favourable bias to compounds already known to be involved in a biological process of some form.

6.5.3 DJANGO PROJECT AND APPLICATION FOR DEEP METABOLOME ANNOTATION

The deep metabolome annotation database (DMAdb), implemented in Python-Django using the Django MOGI application suite detailed in Chapter 6, is a web portal and database that has been developed for management and analysis of DMA projects. The DMAdb allows for a user to create and organise DMA experiments within the ISA framework, allows Galaxy to be used as a data processing backend for extensive annotation analysis and allows the results of the annotations be viewed and searched either directly from the database or through a graphical user interface (GUI). The DMAdb is accessible locally within the University of Birmingham (https://dmadb.researchcloud.bham.ac.uk) where it currently stores the data for the DMA of *D. magna*. The infrastructure used is the same as detailed in Chapter 5 section 5.4.7, the relevant code to run the Django project is available on request (https://github.com/computational-

metabolomics/django-mogisite-dmadb).

6.5.4 DATA ANALYSIS AND VISUALISATION OF DMA ANNOTATIONS

To visualise and assess the overlap of annotations (internal category L2-A, see section 6.5.2) between the experimental assays, the UpSet plots were created using the UpSetR package (Conway et al., 2017).

To visualise and assess similarities between the annotated (L2-A) compounds, chemical descriptors calculated from the Chemoinformatics library RDKit (Landrum and others, 2013) in python were used for principal component analysis (PCA). Using PCA on chemical descriptors is an stablished method to view similarities between groups of compounds (Medina-Franco et al., 2008). For the method performed for this chapter, 202 chemical descriptors were used to compare the 5,389 L2-A annotated compounds along with 5,441 from KEGG (the total number of compounds from KEGG (Kanehisa et al., 2014) with have an associated pathway). See Appendix 8.20 for list of chemical descriptors used. Each chemical descriptor was centred and scaled as per (Wenderski et al., 2015) prior to performing PCA. Additionally, to assess if the approach was capable of distinguishing different compound classes, PCA analysis on the chemical descriptors was performed on all compounds in KEGG with a BRITE biological definition, see Appendix 8.22. The KEGG BRITE (Kanehisa, 2005) definitions are manually curated ontological terms describing the functional hierarchies of a biological object – for example: D-Glucose belongs to the “Aldoses” term, which belongs to the “Monosaccharides” term which belongs to the “Carbohydrates” term which belongs to the “Compound with biological role” term.

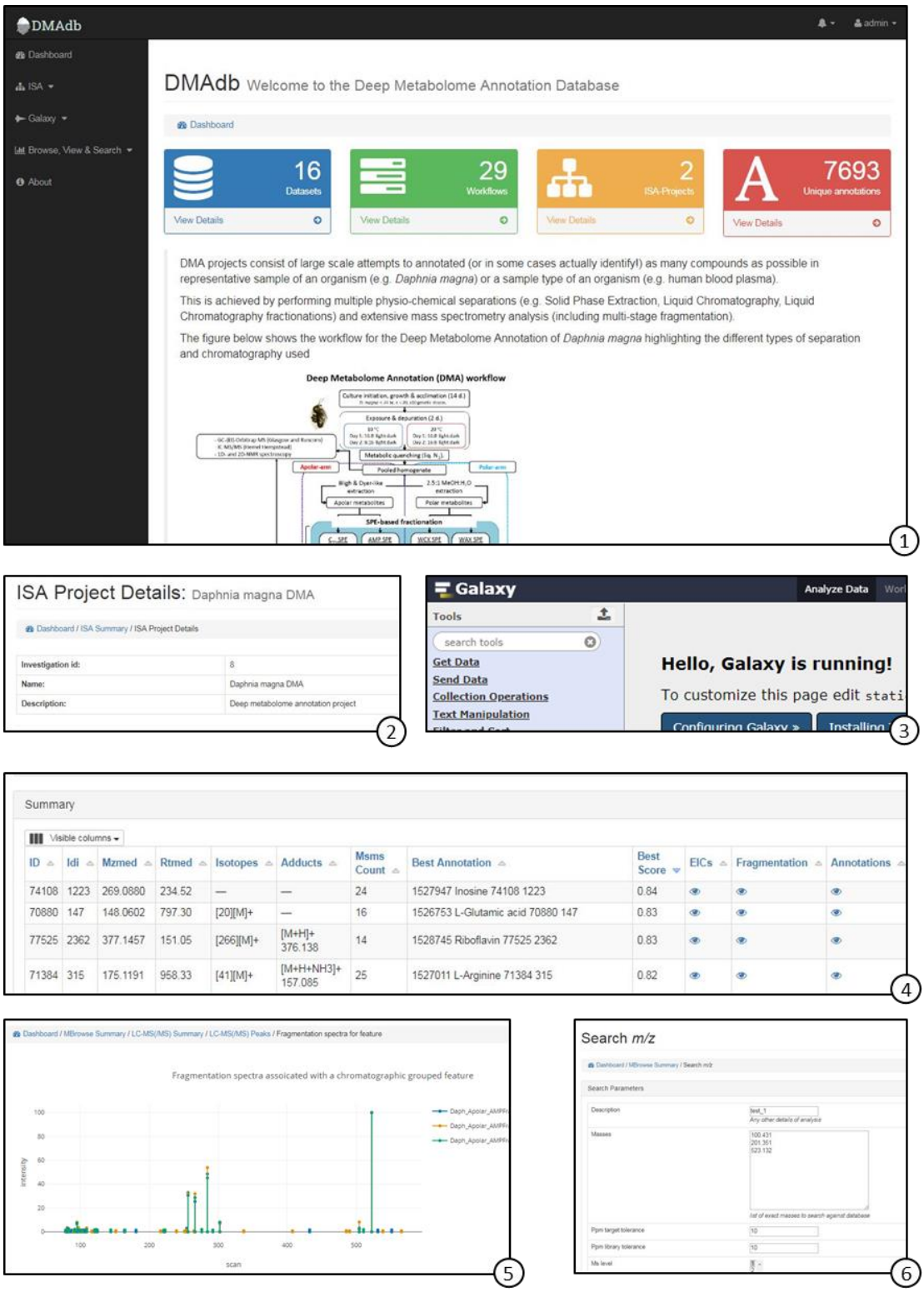


Figure 6-2 The deep metabolome annotation database (DMAdb) overview

1) Dashboard for the DMAdb. Developed using the Django MOGI application suite. 2) Multiple DMA projects can be handled within the ISA framework where relevant ontologies, protocols and processes can be recorded. 3) Data processing and annotation analysis can be performed using Galaxy. The

resulting output can then be uploaded into the DMAdb. Once uploaded into the DMAdb, mass spectrometry datasets can be browsed (4), visualised (5) and searched (6).

6.6 RESULTS AND DISCUSSION

6.6.1 OVERALL LC-MS(/MS) ANNOTATION

We define a unique annotation here based on a unique InChiKey. If multiple LC-MS features obtain the same annotation, we only count this as one unique annotation. It should also be noted that if multiple annotations are available for single LC-MS feature the annotation with the highest weighted score is only used as the annotation.

The overall unique annotations from the 23 experimental assays are detailed in Table 6-2, where a total of 5,469 (2,292 polar, 3,248 apolar, 71 overlapping) unique L2-A annotations and up to 16,896 unique annotations are observed if using less stringent criteria.

Table 6-2 Summary of DMA of *D. magna* unique annotations

Annotation level	# unique annotations (polar)	# unique annotations (apolar)	# unique annotations (total)
L2-A	2292	3248	5469
L2-B (or higher)	7666	6972	14522
L2-C (or higher)	9912	7171	16896

Footnote: Unique annotations based on a unique InChiKey for the compound. If multiple annotations are available for single LC-MS feature the annotation with the highest weighted score is used as the annotation. Note that the same compound can be observed in both polar and apolar analysis.

As the fragmentation spectra of lipids compounds can be more readily predicted (Kind et al., 2013) than that of polar compounds it was anticipated that more annotations would be observed for the polar arm of the DMA workflow. This trend was generally observed up until the L2-C level annotations where more annotations were obtained for the polar arm.

A summary of the number of annotations per experimental assay is detailed in Appendix Table 8-8 to Table 8-11. When summarised across all assays, 22.4% of all measured features have associated fragmentation spectra, where for each assay a median of 18.8% (positive) and 23.0%

(negative) features had associated fragmentation spectra. For all measured features, 6.9% had L2-A annotations, 18.8% had L2-B or higher annotations, and 22.3% had L2-C or higher annotations.

There are far fewer negative L2-A annotations than positive (5,404 positive, 83 negative), due to Sirius CSI:FingerID being optimised for positive ionisation mode, and as a result no successful annotations were obtained from Sirius CSI:FingerID for any negative ionisation data.

As less than a quarter of the LC-MS features have associated fragmentation spectra the observed metabolic space from each assay is not covered. However, as many of the LC-MS features will be adducts, isotopes or in-source fragmentation, the coverage of both features with fragmentation spectra and annotations could be improved if the annotation was based on LC-MS feature clusters rather than individual features. Even with these improvements, there will still be a substantial number of features where the precursor ion purity will be too low to achieve reliable annotation, see Chapter 2. Future analysis that incorporates data independent acquisition (DIA) could help improve coverage (Tsugawa et al., 2015).

It should be noted that this analysis does not include the annotations from the DI-MS/MS and DI-MS^n. This will improve both the fragmentation spectra collected and the annotations observed. It should also be noted that although there is some favourable bias in the annotation approach to try and rank compounds that have known biological functionality higher than those that do not, known synthetic and plant compounds have not been removed from the annotation procedure. However, as shown in Figure 6-4 the compounds annotated to the L2-A inhabit a similar chemical space to those of known KEGG compounds.

As mentioned, when multiple annotations are observed for a single LC-MS feature the annotation with the highest weighted score is used as the annotation. However, all annotation

results are stored in the DMAdb so a user can investigate other annotations that have been observed for any of the LC-MS features.

To improve this further work to validate the annotation approach is required to assess type I and type II errors. This would involve running the same annotation workflow but for a biological sample spiked with a wide range of chemical standards.

6.6.2 OVERLAP AND REDUNDANCY BETWEEN ASSAYS

The overlap of L2-A annotations between assays is described using an UpSet plot (Figure 6-3). The plot highlights that each assay contributes at least 18 unique L2-A annotations, and in the cases for the AP_AMP_1_C30 assay, up to 485 unique L2-A annotations are observed from a single assay. Overlap is also shown between assays, where assays with the same experimental techniques typically exhibiting more overlap of annotations.

The unique annotations observed for each experimental assay in part justify the multiple physiochemical separations performed by liquid phase extraction (LPE), SPE and LC of the assays.

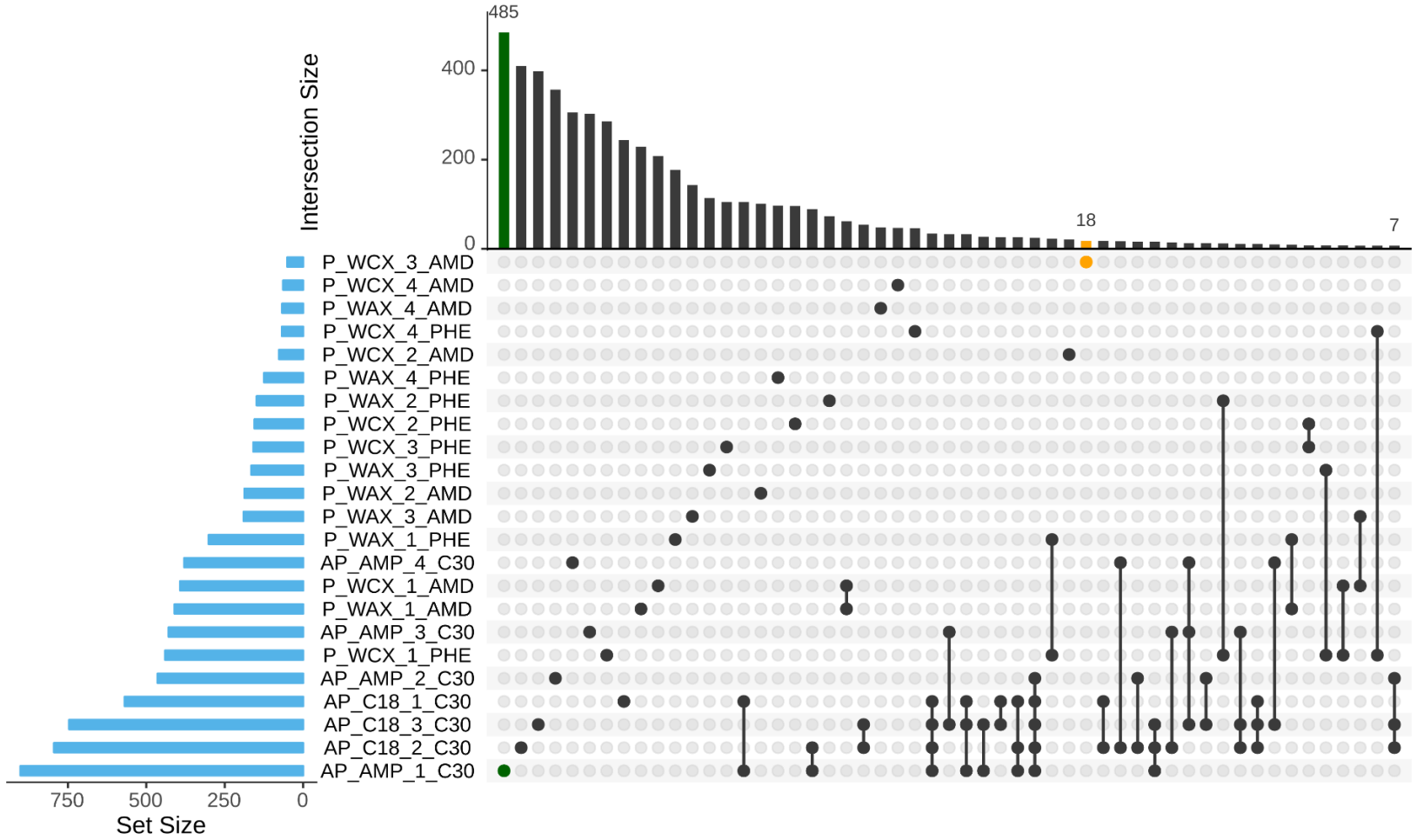


Figure 6-3 UpSet plot summarising the DMA of *D. magna* L2-A metabolite annotations across 23 experimental assays

All annotations were aligned based on InChIKeys. The dark green bar highlights the assay with the highest number of annotations unique to that assay. Orange highlights the assay with the lowest number of annotations unique to that assay. Only showing intersection sizes of 7 and above.

6.6.3 COMPOUND SIMILARITY

To assess the effectiveness of using the chosen chemical descriptors to distinguish different compound types, all KEGG compounds that contained a BRITE biological definition were processed, see Appendix 8.22. Figure 6-4 shows that the PCA of chemical descriptors can separate biological compounds classes.

The chemical descriptors for all L2-A annotations were used to perform PCA. Figure 6-4 displays the scores plot where PC1 and PC2 cover 20.61% and 7.89% variance respectively. Most of the compounds from the DMA experiment are clustered with the KEGG compounds, however there is a subset of compounds that appear to be sufficiently different to the KEGG compounds based primarily on separation on the PC1 axis.

This preliminary analysis of the compounds from DMA of *D. magna* is encouraging, with the co-clustering with KEGG indicating that compounds with biological function are being annotated. Further work using molecular fingerprints of compounds to show clustering and similarity could be an interesting alternative to using chemical descriptors (Owen et al., 2011).

As the compound similarity analysis is only being performed on annotated compounds, the analysis overlooks features that have fragmentations spectra but no annotation. Approaches that cluster fragmentation spectra (such as hierarchal clustering (Frank et al., 2009)) could be used for future analysis to show similarity between features. This might also provide a better understanding of the overlap and redundancy between the assays.

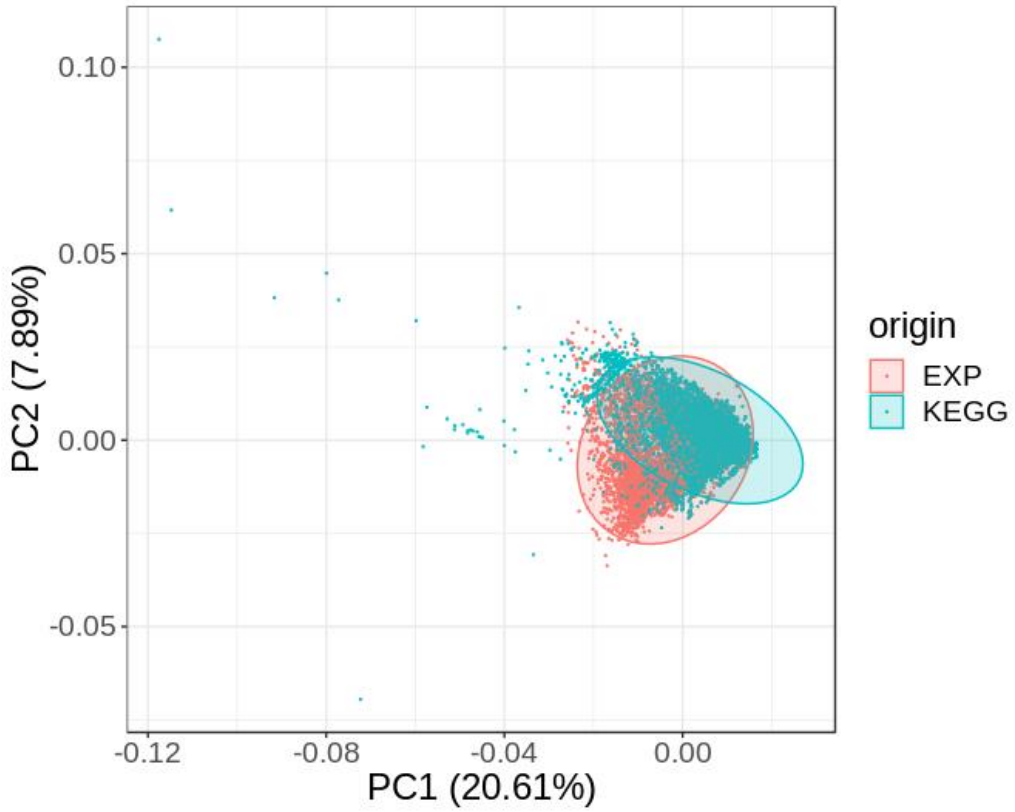


Figure 6-4 PCA scores plot of DMA experimental (EXP) compound L2-A annotations and KEGG compounds

EXP=DMA of *D. magna* experimental annotations (L2-A). KEGG = All KEGG compounds with an associated pathway.

6.6.4 DJANGO PROJECT AND APPLICATION FOR DEEP METABOLOME ANNOTATION

The DMAdb is accessible locally at the University of Birmingham (https://dmadb.researchcloud.bham.ac.uk) and provides a resource for users to search and browse the spectra and annotations from the DMA of *D. magna.* The database is 16 Gb in size and currently contains 1,139 files; 2,078,109 individual LC-MS features; 116,856 grouped LC-MS features; 22,470 neutral masses derived from grouped features; 2,650,815 individual fragmentation spectra with 78,507,671 individual m/z intensity peak pairs. The database also includes experimental details stored within database tables modelled on the ISA framework, including >23 assays, 5 chromatography protocols, 2 LPE protocols, 4 SPE, protocols and 4 measurement protocols.

Essential further work to provide more automated export to MetaboLights and MassBank would help to ensure the data and data analysis for the DMA of *D. magna* is more reproducible and FAIR-er. Additionally, as the diversity of instrumentation used for DMA goes beyond LC-MS/MS, further developments are required to handle data from DI-MS, DI-MS/MS, DI-MS^n and NMR.

6.7 CONCLUSION

The LC-MS/MS component of the DMA of *D. magna* has been processed and managed using DMAdb, Galaxy and ISA. The DMAdb contains the experimental, spectra and compound details of 16,896 metabolite annotations of which 5,469 have been annotated using at least two independent computational approaches.

The results detailed in this chapter provide only the LC-MS(/MS) component of the DMA of *D. magna* and future work is required to develop the computational tools and workflows to analyse the remainder of the DMA of *D. magna* dataset. However, the DMAdb and the

annotations workflows developed will be essential for all future development.

Chapter 7
Conclusion

7.1 OUTCOMES

7.1.1 RECAP ON AIMS

Overall aim: Create computational tools and workflows to allow for reproducible and transparent approaches for acquiring data, deriving metabolome-wide metabolite annotations, and managing and disseminating the results. The primary use of the tools and workflows being for large-scale annotation projects such as the ongoing and extensive DMA of D. magna.

Key questions:

1) Are chimeric spectra problematic in untargeted metabolomics experiments and how can they be assessed?
    - See Chapter 2
2) Can the proportion of informative fragmentation spectra collected for DDA based mass spectrometry analysis be increased?
    - See Chapter 3
3) Can large scale metabolomics annotation studies be analysed and managed reproducibly?
    - See Chapter 4, 5 and 6
4) Can large scale metabolomics annotation studies be applied to model organisms?

- See Chapter 6

7.1.2 OVERVIEW

In Chapter 2, a data processing R package (msPurity) was developed to assess the quality of the MS fragmentation spectra for LC-MS/MS and DI-MS/MS. The software calculates a metric to describe the contribution of the selected precursor ion to assess either the acquired fragmentation spectra or the anticipated fragmentation spectra that has yet to be acquired. The R package was further developed to be implemented into Galaxy and to provide functionality for spectral matching against library spectra to be used later in the thesis as a complement to other Galaxy tools used for metabolite annotation. The R package is currently hosted at the Bioconductor bioinformatic repository, where it has had 797 and 623 unique downloads in 2017 and 2018 (Jan-Aug), the associated paper (Lawson et al., 2017) has also been cited 6 times since its release (Aksenov et al., 2017; Fenaille et al., 2017; Kind et al., 2017; Misra, 2018; Petras et al., 2017; Wang et al., 2018). The chapter addressed the question "*Are chimeric spectra problematic in untargeted metabolomics experiments and how can they be assessed?*" and shows that chimeric spectra are present within metabolomics datasets and should be assessed using the novel metric developed as part of this PhD. This important assessment was necessary to ensure the quality of the spectra from the DMA of *D. magna* where the R package msPurity was used extensively in the data processing and annotation.

In Chapter 3, nearline (nearly-online) data analysis workflows were developed to optimise the scheduling of MS acquisition of gas-phase fragmentation spectra for both liquid chromatography tandem mass spectrometry (LC-MS/MS), direct infusion tandem mass spectrometry (DI-MS/MS) and multistage fragmentation (DI-MS^n) for untargeted metabolomics experiments. An assessment was also made to see if it would be feasible to optimise the LC-MS fractionation (collected eluent from an LC-MS run for later offline

extensive fragmentation analysis) by using a genetic algorithm. Where appropriate the nearline workflows were implemented into Galaxy. These workflows were then used to acquire the fragmentation spectra for the DMA of *D. magna.* The chapter addressed the question "*Can the proportion of informative fragmentation spectra collected for DDA based mass spectrometry analysis be increased?*" and shows that while it is possible and new approaches have been used here for the DMA of *D. magna* - more rigorous assessments are required to accurately determine if the approaches described in the chapter outperform other approaches.

In Chapter 4, to aid in creating ISA structures from metabolomics experiments a suite of convertors (mzML2ISA and nmrML2ISA) to create ISA-tab structures from mzML, imzML and nmrML was developed. These tools were implemented as command line interfaces, python packages, desktop graphical user interfaces and within Galaxy. The associated paper ((Larralde et al., 2017), joint first authorship with Lawson), has been cited 4 times since publication (Gustafsson et al., 2018; Misra, 2018; Schober et al., 2018; Spicer et al., 2017). The chapter addressed in part the question *"Can large scale metabolomics annotation studies be analysed and managed reproducibly?"* and details that ISA-Tab is key for managing metabolomics studies reproducibly but the recording of technical information could be time consuming and error prone – especially for larger studies with many files. The development of parsing tools provided a much-needed automation to help in managing metabolomics studies (such as the DMA of *D. magna*).

In Chapter 5, a suite of Django reusable web applications was developed for both developers and general users for managing metabolomic annotation analysis and their associated Galaxy workflows within an Investigation Study Assay (ISA) framework. This Django MOGI suite (Metabolome Organisation with Galaxy and ISA) was used to create a database and interface to manage the data and annotations for the DMA of *D. magna* (DMAdb). This chapter addressed the question *"Can large scale metabolomics annotation studies be analysed and*

managed reproducibly?" and show that appropriate novel tools and workflows have been developed to perform such experiments – the application of which is shown in Chapter 6.

In Chapter 6, the liquid chromatography tandem mass spectrometry (LC-MS/MS) component of the DMA of *D. magna* was processed using a combination of the tools developed for this thesis and publicly available tools where a total of 5,469 (2,292 polar, 3,248 apolar) metabolites were annotated to MSI (metabolomics standard initiative) level 2. If less stringent selection criteria are used, >14,000 MSI level 2 metabolite annotation are observed. This increase in knowledge of metabolites for the *D. magna* will help aid researchers in biological interpretation of both metabolomics and for the wider understanding of the biology of the model organism *D. magna*. This chapter addressed the question *"Can large scale metabolomics annotation studies be applied to model organisms?"* and shows that using a combination of publicly available tools and those developed in the preceding chapters that the DMA of *D. magna* can be achieved.

As many of the tools and workflows developed as part of this thesis were integrated within the Galaxy workflow environment, I hope that the analysis used to obtain the metabolite annotation can be transparent, reproducible and accessible. Additionally, through XML2ISA and MOGI software and application, work has been developed to aid in recording an experiment within the ISA framework. In doing so, the hope is that metabolome annotation experiments - particularly data analysis - can become easier to be shared in public data repositories under FAIR guidelines.

7.2 FUTURE WORK

7.2.1 PRECURSOR ION PURITY

In proteomics it was determined that a precursor ion purity (or precursor ion fragment) of < 0.5 is insufficient for a reliable protein annotation/identification (Michalski et al., 2011). It would be interesting to perform an experiment to determine the adequate purity cut-off in metabolomics for a reliable annotation. This could involve spiking metabolite standards into a

complex biological sample. When measured on a mass spectrometer there would hopefully be range of precursor ions of the metabolite standards that have varying levels of purity (this isolation window could also be changed on the instrument to adjust the purity of a single precursor). The precursor of the metabolite standard could then be fragmented, and an annotation derived by spectral-matching could be attempted to be determined. The point at which the purity is too low to achieve an annotation would then be used as the recommended cut-off.

Data independent acquisition approaches can use correlation analysis to deconvolute chimeric spectra (Geromanos et al., 2009). However, the same approaches are not possible for data derived from data dependent acquisition (DDA) approaches as only a small proportion of the MS^1 spectra has associated fragmentation spectra. It would be interesting to explore approaches to deconvolute DDA chimeric spectra, a starting point could be to use fragmentation spectra that has been successfully determined as being derived from multiple precursor ions and using this data to be build a model for predicting if newly acquired fragmentation spectra is likely to have arisen from more than one precursor.

7.2.2 NEARLINE

As detailed in the Chapter 3, the nearline approaches performed adequately to acquire fragmentation spectra for both LC-MS and DI-MS, but further work is required to assess the effectiveness of the approaches and to improve the percentages of fragmentation spectra obtained for each MS^1 metabolite feature. Additionally, the move for this type of acquisition approach to be incorporated into vendor instrumentation and associated software is welcomed (e.g. Thermo Fisher Scientific's Orbitrap ID-X Tribrid Mass Spectrometer released in 2018 has intelligent MS^n data acquisition). As the time and skills required both analytically and computationally to perform the nearline procedure described in Chapter 3, limits its use considerably.

7.2.3 MOGI

As detailed in the Chapter 5 section 5.7, essential work is required to ensure the Docker instance of the example MOGI project for interacting with a local Galaxy instance. Additionally, all Galaxy tools developed should be migrated to the stable Galaxy toolshed and the django-misa application needs further developments to allow support for ISA-tab import and export. It would also be beneficial to integrate the XML2ISA suite of metadata parsers to aid in the creation of the ISA-tab files.

Further work that allows data from other instrumentation types including DI-MS, GC-MS and NMR would be beneficial especially as these techniques were used in the DMA of *D. magna*.

7.2.4 DMA

The DMA of *D. magna* is ongoing and future work in both the experimental and data analysis approaches are required to process the remaining data before the full extent of the annotation can be assessed. Any further developments for the MOGI suite of Django applications will also need to be incorporated into the DMAdb. This includes additional functionality to help submit the DMAdb to the MetaboLights data repository. This would drastically improve *Daphnia*'s representation in MetaboLights where at present there is one *Daphnia* specific study that comprises only one associated experimental assay and 69 experimentally annotated metabolites.

As detailed in Chapter 6 section 6.6, only preliminary analysis was performed on the assessment of the compound classes and similarity observed in the data collected so far for the DMA of *D. magna.* Further work clustering molecular fingerprints (Owen et al., 2011) and clustering fragmentation spectra (Frank et al., 2009) would be essential to investigate more extensively.

The hope is that DMA type analysis will be performed on other model organisms (and human samples) which will allow trends and networks to be developed across multiple species and

sample types. In particular it would be interesting to see how organism metabolic pathways are related in evolutionary context, i.e. phylometabolomics (Edison et al., 2016; Viant et al., 2017).

In terms of new instrumentation that might be used, as mentioned in section 3.7 and 7.2.2, the Orbitrap ID-X Tribrid mass spectrometer is well equipped to perform the types of data acquisition described in DMA in a more streamlined and automated manner than was possible even half a year ago.

7.2.5 BEYOND THE METABOLOME

Improving the knowledge of the metabolome of any organism benefits not just the metabolomics community but any researcher interested in the biology of an organism. For biological research to fully benefit from newly discovered *D. magna* metabolites, especially in the context of ecotoxicological testing - it is essential that data can be integrated with other data sources.

7.2.5.1 Multi-omics

Consider the bottom of the "pyramid of life" (Oltvai and Barabási, 2002) of any biological system to be the genome followed by the transcriptome, proteome, metabolome and phenotype. The connections between levels becomes more difficult to infer the further up the pyramid we go. At the bottom there is a near-binary relationship between DNA and RNA molecules. When we move up one step, RNA molecules then encode for 20 amino acids that can generate much more complex structures (proteins), but despite this complexity there is still an unambiguous relationship between the primary sequence of a protein and the RNA sequence. This transfer of information breaks down when we try to understand how a protein, or collection of proteins, are involved in the generation of a metabolite. The problem is then compounded by the diversity of the physiochemical properties of metabolites (Fernie and Stitt, 2012).

Despite the difficulties of integrating metabolites and other layers in the "pyramid of life",

multi-omic approaches incorporating metabolomics are often used successfully to infer biological insights from a system. One example of a multi-omic study on *D. magna* modelled molecular changes associated with the toxicity of cadmium using MS based metabolomics to identify differentially perturbed metabolite classes and microarray analysis to identify differentially expressed genes (Poynton et al., 2011) The challenge for metabolomics approaches in these studies is that untargeted approaches that have the capacity to screen multiple samples in a high-throughput manner in mass spectrometry will typically not have the luxury of extensive fragmentation analysis. This means often only a select few *m/z* features of interest will be chosen for further investigation. The information generated from the DMA of *D. magna* can help in these cases by providing the opportunity to determine whether the *m/z* (or calculated neutral mass) has already been measured and/or annotated before. This can help with preliminary annotation purposes and help the researcher decide if further investigation is required.

7.2.5.2 Genomic scale metabolic network reconstructions and biological networks

Information derived from genome, transcriptome and metabolome are often used to generate "genomic scale metabolic network reconstructions". The two largest resources of genomic scale metabolic pathways are MetaCyc (Caspi et al., 2007) and KEGG (Kanehisa et al., 2014), see (Altman et al., 2013) for a comparison.

Prior knowledge of the metabolites or validation is required to develop a more robust reconstruction. This is problematic for an organism like *Daphnia* with limited metabolite knowledge available (ChEBI database currently only holds 120 entries for *Daphnia* metabolites) this will be overcome to some extent from the annotation provided from the DMA of *D. magna.* The *D. magna* genomic scale reconstruction is currently being developed at the University of Birmingham, it will be interesting to see if the reconstruction can be improved

when using the additional metabolite annotations from the DMA of *D. magna.*

There have also been a few studies that use a single mass spectrometry dataset to reconstruct biological networks, something that should be implemented on the data generated from the DMA dataset. Approaches include using mass difference between metabolites as evidence for biochemical reactions (Breitling et al., 2006), a similar result can also be achieved at level of molecular formulae (Rogers et al., 2009). Spectral alignments have also been used group supposed similar metabolites in a network (Watrous et al., 2012).

7.2.5.3 Linking to phenotype

Behavioural and imaging assays can be considered to provide a final 'downstream' phenotype associated with a perturbed system. Initial work at the *Daphnia* facility at the University of Birmingham has been performed to assess a custom imaging system to gather automated behaviour data, similar to (Matthias Grote, 2014), with the goal of identifying changes in behaviour. Modelling the observed phenotype to the underpinning molecular mechanisms by integrating imaging data with other 'omic data could be extremely useful in ecotoxicological testing. Future work to use phenotypic information as an additional annotation layer of the DMAdb could be interesting. This phenotypic annotation for unknown metabolites would be useful in targeting biologically important metabolites where a compound annotation is not possible to achieve.

7.3 FINAL REMARKS

Model organism metabolome knowledge can be improved by performing large scale metabolite annotation studies that incorporate multiple techniques and technologies that go beyond the resources of routine untargeted metabolomics analysis. However, reproducible tools and workflows to acquire, analyse and manage such large datasets are limited. In this thesis I have developed and applied various reproducible focused software and workflows to a large-scale

metabolome annotation study and have successfully aided in substantially increasing the metabolome knowledge of the model organism *D. magna*. Future ecotoxicological research can hopefully use this information to better inform the mechanisms by which chemical pollutants effect the biology of *Daphnia* and the environments it inhabits.

Chapter 8
Appendix

8.1 ISOLATION EFFICIENCY OF WATERS XEVO G2 XS QTOF

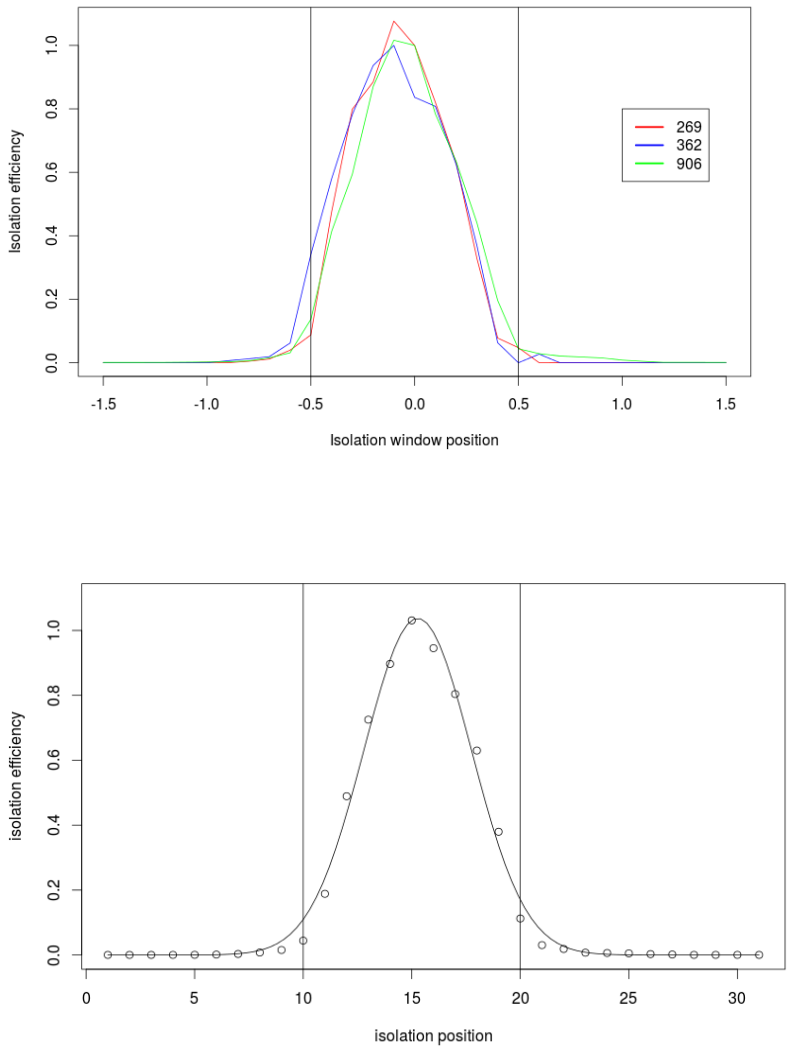


Figure 8-1: Experimental determination of isolation efficiency for Waters Xevo G2 XS QTOF

Top : Isolation efficiency profile of Xevo G2 XS QTOF Mass Spectrometer on the following ions: 269.20, 362.95 and 906.81 m/z. **Bottom:** A Gaussian curve (represented as Y = K*exp(-1/2*(X-μ)^2/σ^2)) was fitted on the mean isolation efficiency profile, using a non-linear least squares minimization method (R package "nls") (Residual standard error 0.03017, P<0.0001 for μ, K and σ)

8.2 IN HOUSE EXPERIMENTAL DETAILS

8.2.1 DAPHNIA MAGNA DATASET

8.2.1.1 Culturing and metabolic quenching

Daphnia magna were cultured to adulthood (>14 days) in a modified high-hardness combination medium, to which the green algae *Chlorella vulgaris* was periodically added. Prior to metabolic quenching, daphniids were filtered from their growth environment on to a fabric gauze mesh, upon which they were washed in fresh culturing medium and, thereafter, blotted dry. Finally, daphniids were transferred to 1.5 mL microfuge tubes and submerged in liquid nitrogen in order to quench metabolism. Samples were thereafter stored frozen at -80 °C until utilised in metabolite extractions.

8.2.1.2 Metabolite extraction

Metabolically quenched samples were submerged in liquid nitrogen before homogenisation over dry ice by grinding with a ceramic mortar and pestle. The resulting homogenate was transferred to a 28 mL glass vial on dry-ice, to which was added ice-cold (4 °C) extraction solvent comprising 2.5:1 v/v methanol:water (corrected for estimated biological water content) at a ratio of 15 parts solvent per 1 part biological wet mass. Blank samples were similarly prepared, with extraction solvent being added to a separate 28 mL glass vial, in addition to a small correction for biological water content. Both blank and *Daphnia* samples were vortex mixed for 30 seconds, followed by centrifugation over 10 minutes at 3000 rpm, 4 °C. Blank and *Daphnia s*upernatants were transferred to separate pre-cooled glass vials using glass Pasteur pipettes and then divided in to microfuge tubes such that each tube contained 20 *Daphnia* equivalents. Samples were dried by vacuum centrifugation at 35 °C.

8.2.1.3 Solid phase extraction

Samples were fractionated using a weak cation-exchange solid phase extraction sorbent

(HyperSep WCX, 3 mL, 500 mg). In this work, we include only the first of these fractions in our liquid chromatography fraction purity assessment calculations.

8.2.1.4 LC-MS and LC-MS/MS

Separations were performed using a Dionex Ultimate 3000 liquid chromatograph equipped with fraction collector, in-line degasser and column oven. A reversed-phase Syncronis Phenyl LC column (2.1 x 100 mm, Thermo Fisher Scientific) fitted with in-line filter cartridge (0.2 µm pore size, 2.1 mm ID). A gradient elution protocol was applied, in which mobile phase A comprised 90:5:5% v/v/v water:methanol:100 mM ammonium acetate buffer (pH 5.80, aq.), while mobile phase B comprised 95:5% v/v methanol:100 mM ammonium acetate buffer (p 5.80, aq.). The elution gradient was as follows: 0/3/21/24/24.1/25/30 min at 0/0/100/100/0/0/0% B.

Mass spectral detection of LC column eluates was performed using an Orbitrap Q Exactive mass spectrometer equipped with H-ESI II source, operating under the following parameters: 70000 resolution (FWHM at 200 m/z); 3e6 AGC target; 100-1000 m/z scan range. Performed in positive ionisation mode only.

On the same *Daphnia magna* sample, three consecutive LC-MS runs were performed followed by four consecutive LC-MS/MS runs. The first LC-MS/MS run was started within 4 hours of the last LC-MS run to have finished.

The three *Daphnia* LC-MS, three blank LC-MS and four LC-MS/MS thermo .RAW files were converted into the mzML file format using the vendor centroided peak picking with MSConvert (v3.0). Feature detection was carried out using the XCMS R package with the following parameters: peak picking algorithm = centWave, snthresh = 5, ppm = 5, prefilter = c(3, 5000), integrate = 1, mzdiff = 0.001. Peaks were aligned by retention time using the retcor.loess algorithm and grouped using the group.density algorithm with the following settings: bw = 5,

mzwid = 0.01 and minfrac = 0.5. Blank features were then subtracted from the sample features. Additionally, only sample features were considered when the peak was present in 85% of the LC-MS Daphnia replicates (i.e. 3 out of 3).

Where an XCMS feature was found across both the LC-MS and LC-MS/MS runs the anticipated (predicted) and assessed precursor scores were compared and validated.

8.2.2 SHEEP TISSUE DATASET

8.2.2.1 Metabolite extraction

Sheep liver was extracted applying a monophasic extraction. Accurately weighed 50-60 mg of tissue was homogenised in a monophasic extraction solvent of 2.5:1:1 of methanol: chloroform: water with a volume normalised to the amount of tissue added, to give an overall concentration of 9.2µl extraction solvent per mg of tissue. The homogenised sample was centrifuged at 13,500 g for 5 minutes at 4°C. 200µL of the supernatants obtained after homogenisation were pooled together and centrifuged followed by drying under a nitrogen gas stream at 40°C and subsequently stored at -80°C until further analysis.

8.2.2.2 LC-MS

Ultra High Performance Liquid Chromatography-Mass Spectrometry (UHPLC-MS) analysis was performed applying an Ultimate3000 RSLC UHPLC system coupled to an electrospray Q Exactive mass spectrometer operating in positive/negative ion switching mode. Sample were reconstituted in 100µl of 50/50 methanol/water and centrifuged. 5µl of each sample was injected on to a Hypersil Gold C18 reversed phase column at a flow rate of 400µl min-1. A gradient elution was applied as previously described(Gehmlich et al., 2015). Three injections were performed applying a top 5 DDA method with three different precursor m/z ranges – 100-400, 400-700 and 700-1000 at a mass resolution of 35,000 (FWHM at m/z 200). A further three injections were performed to acquire full-scan data in the m/z range of 100-1000 at a mass

resolution of 35,000 (FWHM at m/z 200).

A total of 2 blank LC-MS samples and 12 sheep tissue LC-MS runs were performed. The thermo .RAW files were converted into the mzML file format using the vendor centroided peak picking with MSConvert (v3.0). Feature detection was carried out using the XCMS R package with the following parameters: peak picking algorithm = centWave, snthresh = 5, ppm = 5, prefilter = c(3, 5000), integrate = 1, mzdiff = 0.001. Peaks were aligned by retention time using the retcor.loess algorithm and grouped using the group.density algorithm with the following settings: bw = 5, mzwid = 0.01 and minfrac = 0.5. Blank features were then subtracted from the sample features. Additionally, only sample features were considered when the peak was present in 85% of the tissue sample replicates (i.e. 10 out of 12).

8.3 FURTHER INFORMATION FOR PUBLIC DATASETS (MSPURITY)

Table 8-1: Further information for public datasets

ID	Data Repository	Submission Date	Paper DOI	Repository entry link	Number of files	Number of MS/MS scans
MTBLS26	MetaboLights	11/14/2012	10.1194/jlr.M028902	www.ebi.ac.uk/metabolights/MTBLS26	71	140124
MTBLS36	MetaboLights	02/07/2014	10.1038/sdata.2014.29	www.ebi.ac.uk/metabolights/MTBLS36	48	338457
MTBLS117	MetaboLights	10/29/2014	Not available	www.ebi.ac.uk/metabolights/MTBLS117	4	321
MTBLS144	MetaboLights	11/25/2014	10.1016/j.marchem.2014.11.003	www.ebi.ac.uk/metabolights/MTBLS144	54	249013
MTBLS154	MetaboLights	1/27/2015	https://doi.org/10.1002/lol2.10043	www.ebi.ac.uk/metabolights/MTBLS154	68	185499
MTBLS155	MetaboLights	1/15/2015	10.1111/1462-2920.12899	www.ebi.ac.uk/metabolights/MTBLS155	120	269223
MTBLS157	MetaboLights	1/29/2015	10.1038/ismej.2016.6	www.ebi.ac.uk/metabolights/MTBLS157	21	472382
MTBLS214	MetaboLights	7/31/2015	10.1021/acs.jproteome.5b00443	www.ebi.ac.uk/metabolights/MTBLS214	442	131113
ST000075	Metabolomics Workbench	7/23/2014	Not available	http://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST000075&StudyType=MS&ResultType=1	2	18727
ST000082	Metabolomics Workbench	7/25/2014	Not available	http://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST000082&StudyType=MS&ResultType=1	3	4732
ST000098	Metabolomics Workbench	9/15/2014	Not available	http://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST000098&StudyType=MS&ResultType=1	2	18960
MS-DIAL: DIA	PRIME DROP Met	03/07/2015	10.1038/nmeth.3393	http://prime.psc.riken.jp/?action=drop_index	6	26444
MS-DIAL: DDA	PRIME DROP Met	03/07/2015	10.1038/nmeth.3393	http://prime.psc.riken.jp/?action=drop_index	23	331302

8.4 VIOLIN PLOTS FOR PRECURSOR PURITY OF PUBLIC DATASETS (MSPURITY)

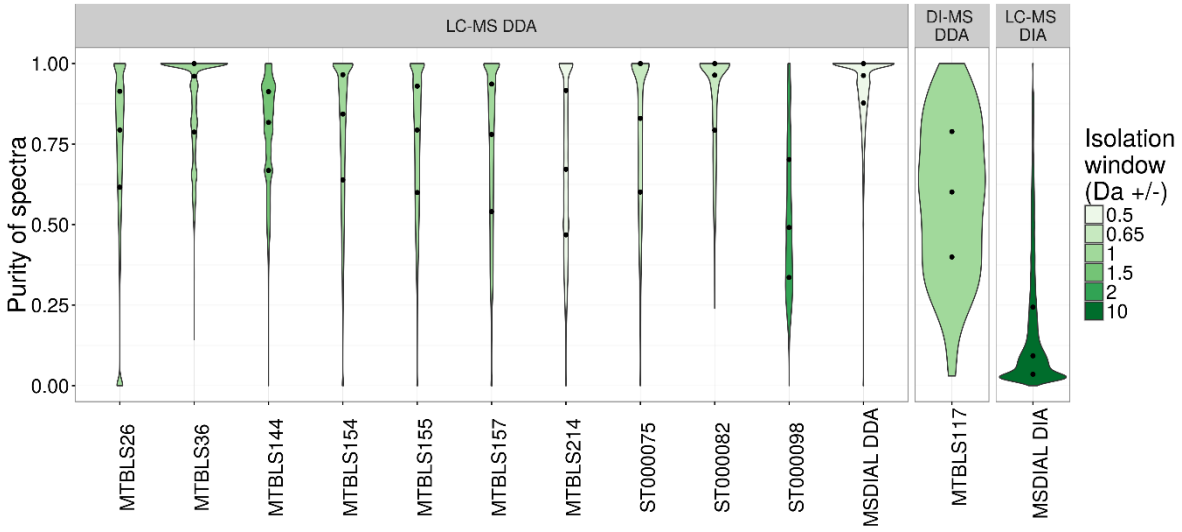


Figure 8-2: Assessment of precursor purity for public DDA and DIA metabolomics datasets (violin plots).

Violin plots of calculated precursor purity scores of all fragmentation spectra. Black dots represent upper quartile, median and lower quartile.

8.5 OVERVIEW OF ASSESSMENT OF PRECURSOR PURITY FOR PUBLIC DDA AND DIA METABOLOMICS DATASETS - ZERO VALUES INCLUDED (MSPURITY)

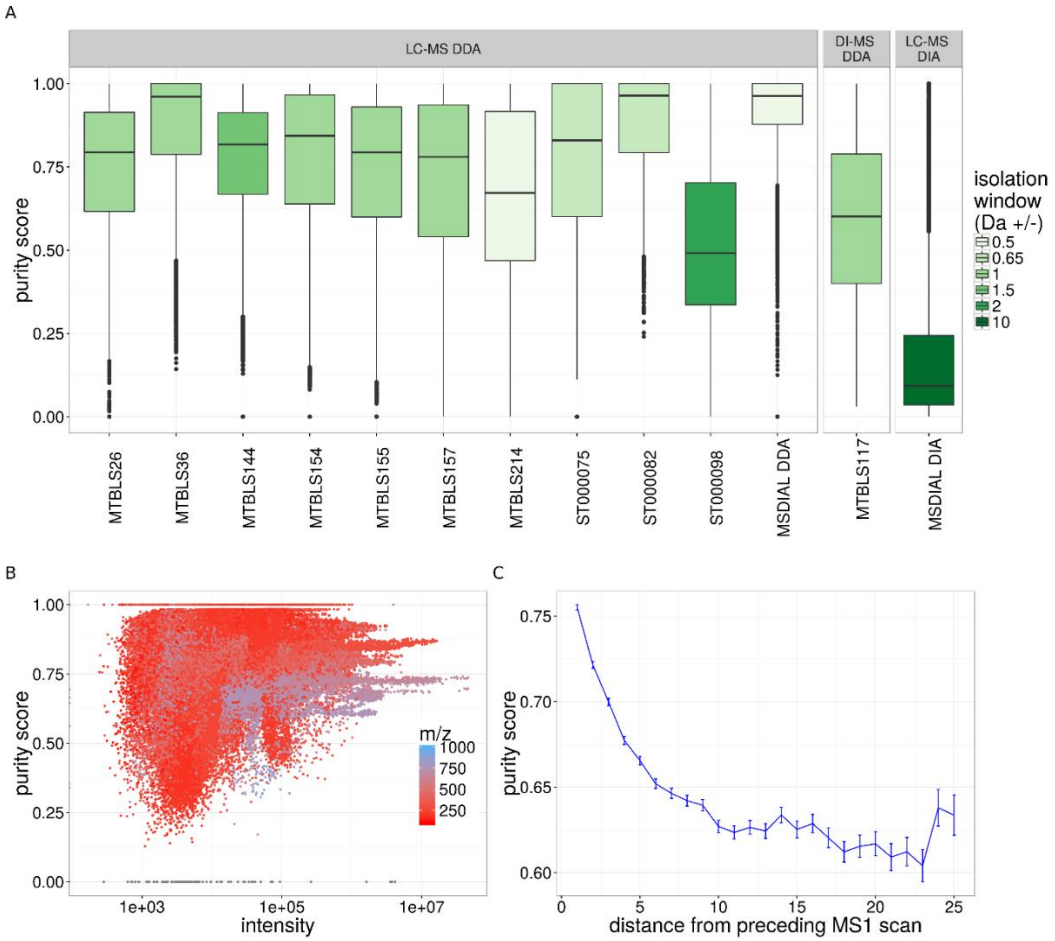


Figure 8-3: Overview of assessment of precursor purity for public DDA and DIA metabolomics datasets (zero values included)

A: Boxplots of calculated purity scores of all fragmentation spectra. See Figure S2 for violin plots derived from the same data. **B** Summary of purity score relationship with selected precursor intensity and m/z value for study MTBLS144, *x* axis is in $\log_{10}$ scale. **C:** The mean purity relationship with the distance from the preceding MS1 scan for study MTBLS214. Error bars represent one standard error of the mean.

8.6 SUMMARY OF POSITIVE AND NEGATIVE IONISATION MODE COMPARISON FOR PUBLIC DDA DATASETS (MSPURITY)

Table 8-2 Summary of positive and negative ionisation mode comparison for public DDA datasets

Study	Mean purity score for MS/MS scans		Total scan number		p-value *	q-value †	Difference in purity score (pos-neg)
	pos	neg	pos	neg			
MTBLS26	0.75	0.70	69407	70717	0.86484	0.88389	+0.05
MTBLS36	0.91	0.84	200911	137546	<0.0001	<0.0001	+0.07
MTBLS117	0.58	0.59	111	210	0.88389	0.88389	+0.01
MTBLS144	0.84	0.73	122257	126756	<0.0001	<0.0001	+0.11
MTBLS154	0.76	0.8	101276	84223	<0.0001	<0.0001	-0.03
MTBLS155	0.75	0.74	149023	120200	<0.0001	<0.0001	+0.01
MTBLS157	0.76	0.68	260246	212136	<0.0001	<0.0001	+0.08
MTBLS214	0.65	0.79	109284	21829	<0.0001	<0.0001	-0.14
ST000075	0.94	0.72	4384	14343	<0.0001	<0.0001	+0.22
ST000082	0.95	0.82	2212	2520	<0.0001	<0.0001	+0.13
ST000098	0.49	0.54	5370	13590	<0.0001	<0.0001	-0.05
Mean	0.76	0.72	93134.64	73097.27	0.16	0.16	+0.04

Footnotes:

- *Unpaired Wilcoxon rank sum test with continuity correction
- †q-value (adjusted p) calculated using the Benjamini & Hochberg approach
- pos = acquisition performed in positive ionisation mode
- neg = acquisition performed in negative ionisation mode
- Non-parametric used based on non-normally distributed assumption of data (see S6)
- Comparison of centred vs most intense approach to determining precursor peak (msPurity)

8.7 COMPARISON OF CENTRED VS MOST INTENSE APPROACH TO DETERMINING PRECURSOR PEAK

Table 8-3: Comparison of centred vs most intense approach to determining precursor peak

Study id	Precursor purity score (centre peak)				Precursor purity score (most intense peak)				p-value *	q-value †	Difference in median purity score (centre peak - most intense)
	min	mean	median	max	min	mean	Median	max			
MTBLS26	0.000	0.693	0.788	1.000	0.000	0.724	0.794	1.000	<0.0001	<0.0001	-0.031
MTBLS36	0.000	0.841	0.946	1.000	0.143	0.880	0.961	1.000	<0.0001	<0.0001	-0.040
MTBLS117	0.003	0.563	0.597	1.000	0.030	0.587	0.601	1.000	<0.0001	<0.0001	-0.024
MTBLS144	0.000	0.733	0.800	1.000	0.000	0.781	0.817	1.000	<0.0001	<0.0001	-0.048
MTBLS154	0.000	0.730	0.829	1.000	0.000	0.778	0.843	1.000	<0.0001	<0.0001	-0.048
MTBLS155	0.000	0.708	0.779	1.000	0.000	0.747	0.794	1.000	<0.0001	<0.0001	-0.039
MTBLS157	0.000	0.689	0.765	1.000	0.000	0.721	0.780	1.000	<0.0001	<0.0001	-0.032
MTBLS214	0.000	0.630	0.652	1.000	0.000	0.672	0.672	1.000	<0.0001	<0.0001	-0.042
ST000075	0.000	0.768	0.830	1.000	0.000	0.771	0.830	1.000	<0.0001	<0.0001	-0.002
ST000082	0.052	0.880	0.964	1.000	0.240	0.881	0.964	1.000	<0.0001	<0.0001	-0.001
ST000098	0.000	0.416	0.376	1.000	0.000	0.527	0.491	1.000	<0.0001	<0.0001	-0.111
MSDIAL DDA	0.000	0.916	0.963	1.000	0.000	0.918	0.963	1.000	<0.0001	<0.0001	-0.002
MSDIAL DIA	0.000	0.018	0.008	1.000	0.000	0.180	0.093	1.000	<0.0001	<0.0001	-0.162

Footnotes:

- *Paired Wilcoxon rank sum test with continuity correction
- †q-value (adjusted p) calculated using the Benjamini & Hochberg approach
- Most intense peak = The precursor purity calculation was performed for the most intense peak within an isolation window
- Centre peak = The precursor purity calculation was performed for the central peak within an isolation window
- Non-parametric used based on non-normally distributed assumption of data (see S6)

8.8 PRECURSOR PURITY DISTRIBUTION OF PUBLIC DATASETS (MSPURITY)

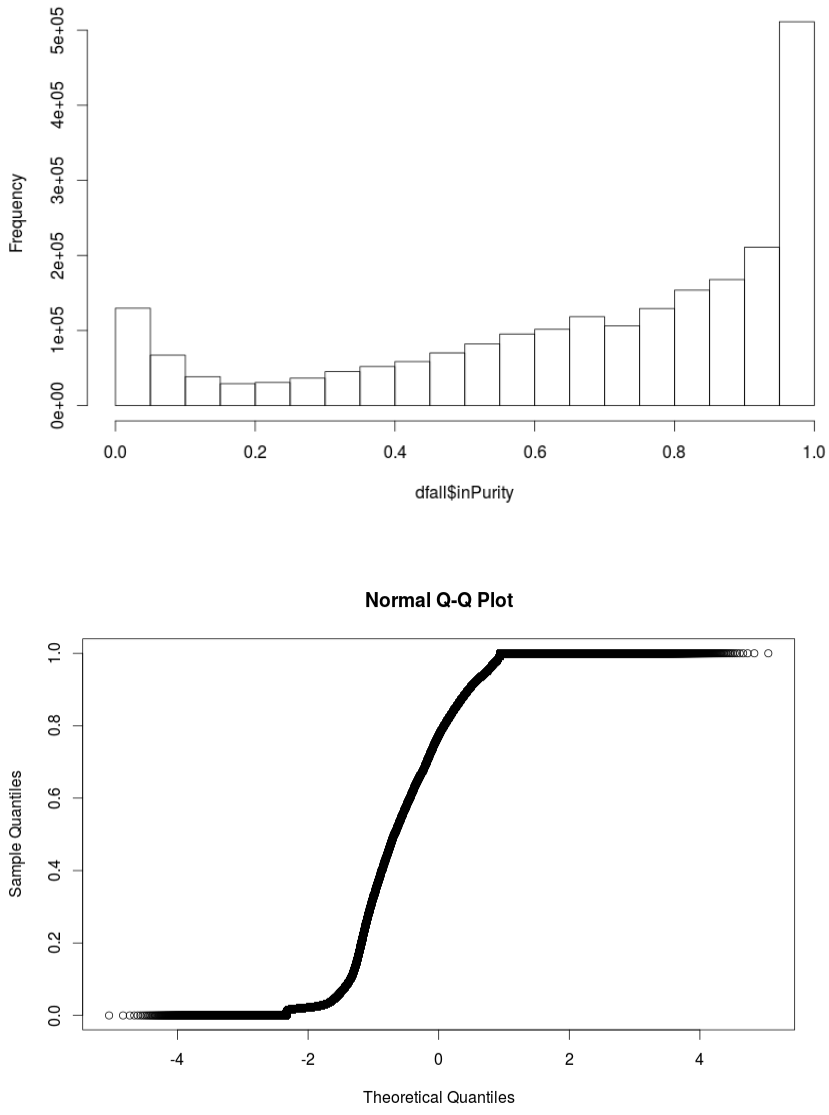


Figure 8-4: Precursor purity (using most intense peak) distribution of all data from public sources.

Top: Histogram. ***Bottom:*** Q-Q plot

8.9 ORIGINAL NEARLINE SCHEDULING ALGORITHMS

Algorithm 1 Weighted Interval Scheduling (dynamic programming)

Objective: Given a set of n intervals I where each n of I has a start time s_i, finish time f_i and weight w_i.
Find the subset of I where there are no overlapping intervals and the total weight is maximised.

procedure P(j)
Moving from start time to finish time. Get the closest interval to j that is not-overlapping with j.

function ITERATIVE-COMPUTE-OPT(n, w, p)
OPT[0]= 0
j = 1
for j to n **do**
OPT[j] = max(w_j+ OPT[p[j]], OPT[$j-1$])
return OPT

function FIND-SOLUTION(OPT, p)
j = 0
if $j = 0$ **then**
output nothing
else if (w_j + OPT[p[j]] > OPT[$j-1$]) **then**
output j
FIND-SOLUTION(p[j])
else
FIND-SOLUTION($j-1$)

function SCHEDULE(I, s, f, w)
Sort intervals by finish times $f_1 \leq f_2 \leq \ldots f_n$
n = length(I)
p = P(1)...P(n) ▷ compute array of previous intervals
OPT = ITERATIVE-COMPUTE-OPT(n, w, p) ▷ compute opt array
S = FIND-SOLUTION(OPT, p) ▷ compute optimum subset of I
return S

Algorithm 2 Nearline scheduling approach

function NEARLINE(I, s, f, w)
Repeat SCHEDULING procedure until all intervals have been scheduled
while length(I)> 0 **do**
Get optimum subset
S = SCHEDULE(I, s, f, w)
Remove S from I, s, f, w

8.10 ADAPTED-LCMS-NEARLINE GALAXY DATA ANALYSIS WORKFLOW

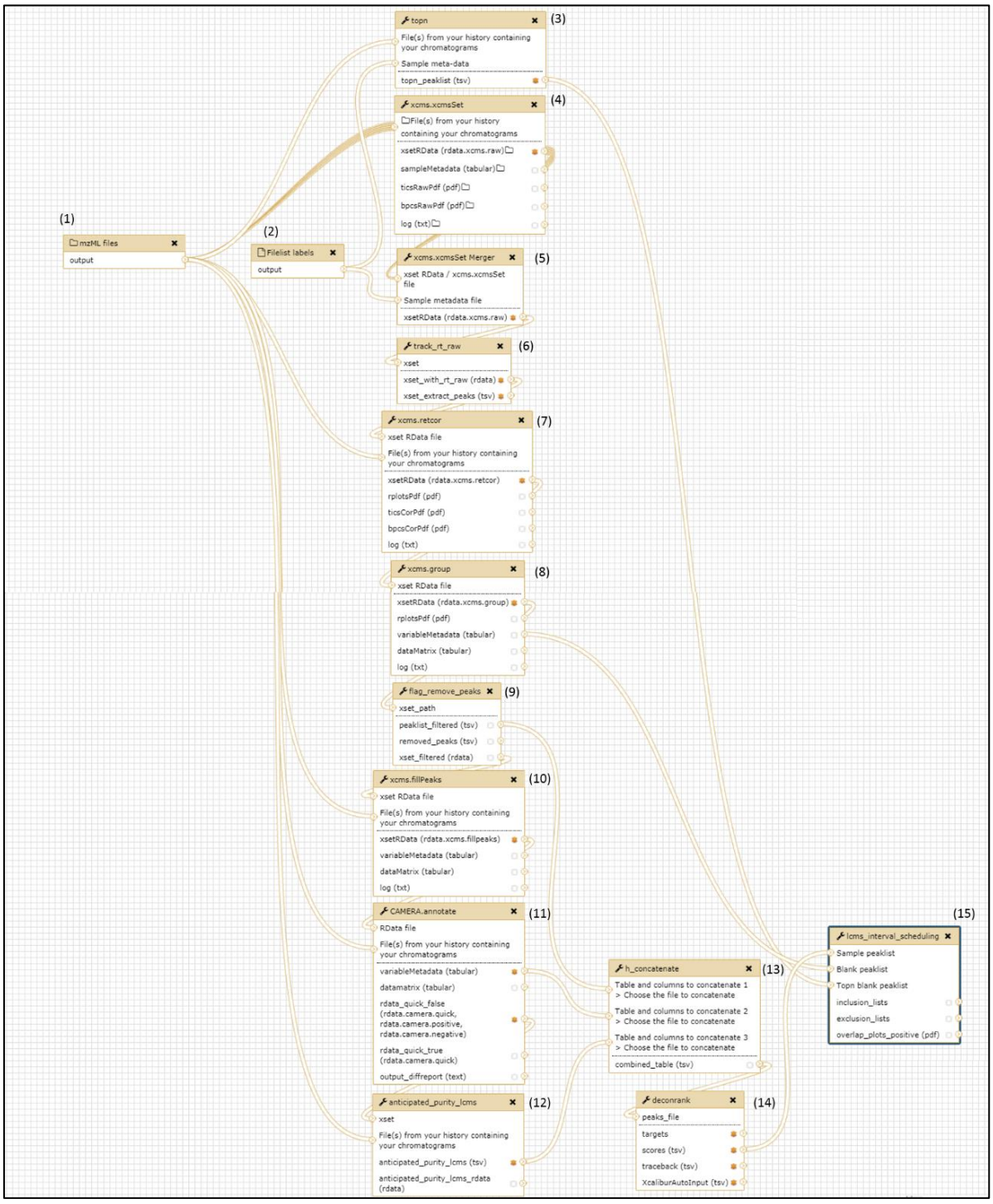


Figure 8-5 adapted-LCMS-nearline Galaxy data analysis workflow

1) Dataset collection: mzML data files of the LC-MS acquisition - minimally 3 runs for the blank and the sample

2) Data: File-list labels mapping if the files to either a blank or a sample to be used to generate the target

list.

3) Topn tool: This tool takes as input the mzML data files (from step 1) and the file-list mapping (from step 2) and determines the top *n* peaks for the blank within defined time windows of the chromatographic run. This step is required as not all blank peaks are picked during XCMS feature detection. The output is a list of the constant blank peaks that are observed throughout the run (used for step 15).

4) xcms.xcmsSet tool: This tool takes as input a dataset collection of mzML files (from step 2) and perform XCMS feature detection to obtain peaks for all files. The output is a dataset collection of xcmsSet objects of each mzML file in rdata file (used for step 5).

5) xcms.xcmsSet Merger tool: This tool takes as input a sample metadata file (file-list labels, from step 2) and a dataset collection of xcmsSet objects (from step 4) and merges them into a single xcmsSet object where the file-list labels are used to map what file is from either blank or sample origin. One of the outputs from the tool is a xcmsSet object in a rdata file (used for step 6).

6) track_rt_tool tool: This tool takes as input an xcmsSet object in a rdata file (from step 5) and tracks the raw retention time. This information can also be retrieved within the xcmsSet object after retention time correction, but the information is saved here for each peak for both convenience and troubleshooting purposes. The output is a xcmsSet object in a rdata file (used for step 7).

7) xcms.retcor tool: This tool takes as input an xcmsSet object in a rdata file (from step 6) and the mzML dataset collection (from step 1) and performs retention time (Note: for this workflow the obi-warp algorithm is used and so retention time correction can be performed without a preceding grouping procedure). The outputs contain an xcmsSet object in a rdata file (used for step 8) and the variableMetadata peak matrix (used for step 15).

8) xcms.group tool: This tool takes as input an xcmsSet object in a rdata file (from step 7) and groups peaks with similar *m/z* and retention time. One of the outputs of this tool is an xcmsSet object in a rdata file (used for step 10).

9) flag_remove_peaks tool: This tool takes as input an xcmsSet object in a rdata file (from step 8) and flags XCMS grouped peaks from the xcmsSet object based on various thresholds (e.g. RSD of intensity and retention time). The peaks can then be removed from the xcmsSet object and regrouped. The output contains an xcmsSet object in a rdata file (used for step 10) and a tsv file of the filtered peaks (used for step 13).

10) xcms.fillpeaks tool: This tool takes as input a xcmsSet object in a rdata file (from step 9) and the mzML dataset collection (from step 1) and "fills" missing peaks by checking the region of interest in the mzML files. This is a requirement for using the CAMERA.annotate tool. One of the outputs is an xcmsSet object in a rdata file (used for step 11).

11) CAMERA.annotate tool: This tool takes in as input a xcmsSet object in a rdata file (from step 10)

and the mzML data collection (from step 1) and calculates the adducts and isotopes annotation. The outputs contain a CAMERA.annotate object in a rdata file (used for step 12) and a "variableMetadata" tab separated file with adduct and isotope details (used for step 13).

12) anticipated_purity_lcms tool: This tool takes as input a xcmsSet object or a CAMERA.annotatate object (from step 11) and the mzML dataset collection (from step 1). The precursor ion purity for an anticipated fragmentation acquisition is calculated. Default settings is to expect a ± 0.5 Da isolation window. One of the outputs is a tsv file containing the precursor ion purity of each grouped XCMS peak (used for step 13).

13) h_concatenate tool: This tool takes as input the CAMERA.annotate output table "variableMetadata" that contains columns for the adduct and isotopes (from step 11), the flag_remove_peaks table "peaklist_filtered" that contains peak intensity and relative standard deviation metrics for retention time and intensity (from step 9), and the "anticipated_purity_lcms" table that contains the precursor ion purity for each grouped feature (from step 12). The tool joins the multiple datasets tables by appending the selected columns together. The output is a tsv file with the selected columns (used for step 14).

14) deconrank tool: This tool takes as input the combined tsv table generated from the h_concatenate tool (from step 13). The peaks in the tsv file are deconvoluted based on the CAMERA adduct and isotope information. Peaks are then ranked and filtered based on intensity, precursor ion purity, adduct and isotope form. One of the outputs is tsv file containing all the peaks and details if each peak should be included in the target list or not (used for step 15)

15) lcms_interval_scheduling tool: This tool takes as input the deconrank scored list of features for the target list (from step 14), the blank grouped peaks determined from XCMS (this can be extracted from the grouped peaks prior to blank removal, from step 8) and top intense peaks determined by binning across set retention time ranges for the blank (from step 3). This final step in the workflow is to schedule the ranked and filtered peaklist into a defined number of experimental fragmentation runs with a defined limit on the number of overlapping peaks. Additionally, the grouped blank peaks determined either from XCMS (from step 8) and the blank peaks calculated by using the top *n* most intense blank peaks with the topn tool (from step 3) are required as input to be used for the exclusion lists.

8.11 DIMS-NEARLINE GALAXY DATA ANALYSIS WORKFLOW (FOR PAIRED SAMPLE-BLANK ANALYSIS)

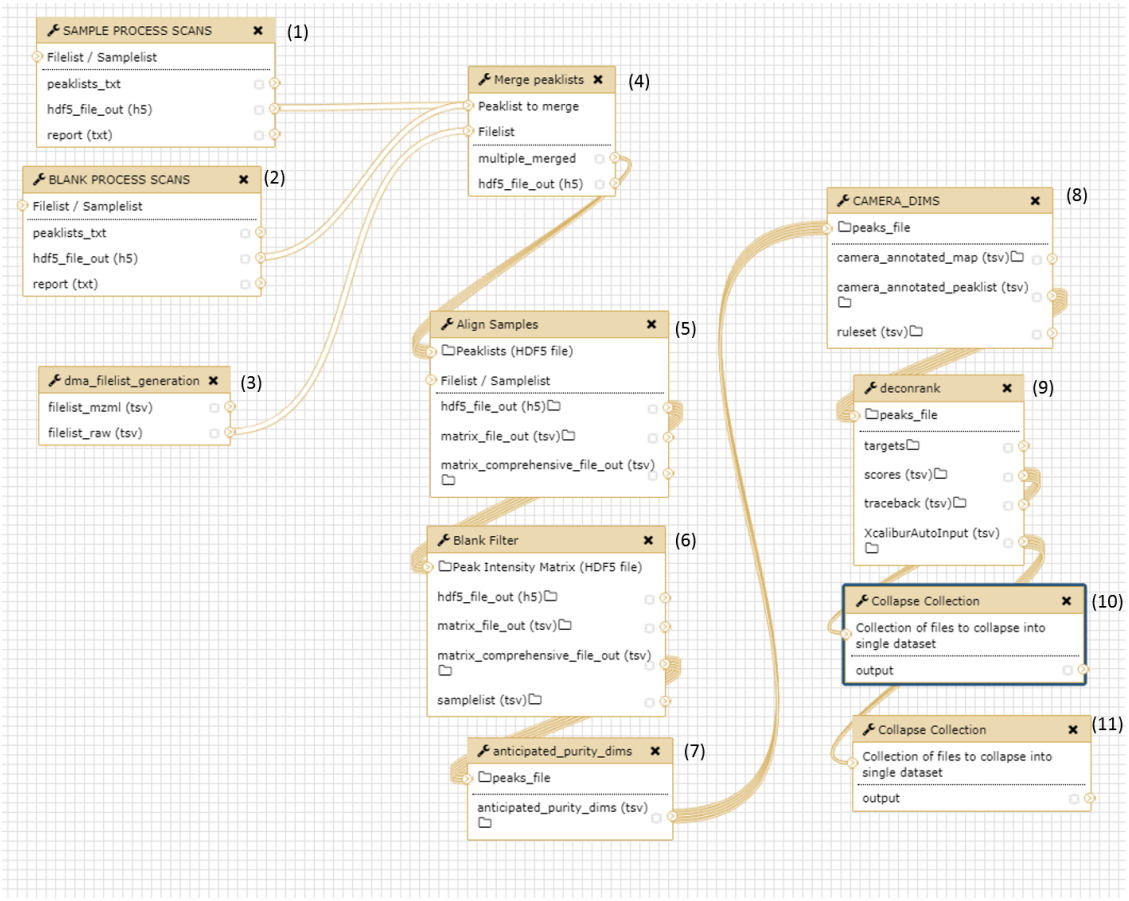


Figure 8-6 DIMS-nearline Galaxy data analysis workflow (for paired sample-blank analysis)

This workflow is intended at processing multiple paired analysis DIMS datasets where a blank has been collected for each sample e.g. sample1.mzML, blank1.mzML, sample2.mzml, blank2.mzML. The blank can be processed with different parameters to the sample (e.g. the signal-noise-threshold can be lower).

1) SAMPLE PROCESS SCANS (dimspy_process_scans tool): Process the scans of a DIMS acquisition files. The input can be either a single file, a data collection of files or a reference to a directory path that contains the relevant data files. In the case for this workflow a reference to a directory containing the sample data files in RAW mass spectrometry vendor data format (Thermo Fisher Scientific) should be used. When multiple scans of the same mass range are present in the data file, the peaks are clustered within the file using one-dimensional hierarchal clustering to obtain an averaged peaklist for each data file. The tool can be used for "full scan" data where the same *m/z* range is used for multiple scans or can

be used for "SIMstitch" where smaller windows of *m/z* ranges have been collected and resulting windows are "stitched" together. The output used for the next step is a HDF5 file that contains a list of DIMSpy "process_scan" python data objects for each file.

2) BLANK PROCESS SCANS (dimspy_process_scans tool): The same as step 1 but for the blank data files.

3) dma_filelist_generation tool: This tool takes in as input the reference to the blank and sample data files directories. The output used for the next step is a sample list that has been automatically created mapping raw data files to blank or sample class.

4) Merge peaklists (dimspy_merge_peaklist tool): This tool takes as input the blank and sample HDF5 files (from step 1 and 2) as well as the sample list (from step 3). The paired datasets are merged together, this allows the paired sample-to-blank files to be processed together for the remaining steps in the workflow. The tool outputs a dataset collection of the paired sample-to-blank files are then created.

5) Align samples (dimspy_align_samples tool): This tool takes as input a dataset collection of the paired peaklists in HDF5 format (from step 4). The peaks between the paired files are then aligned using one-dimensional hierarchal clustering. The output is data collection of HDF5 files each containing a dimspy "peak matrix" object of the aligned peaks of each sample-to-blank pair.

6) Blank filter (dimspy_blank_filter tool): This tool takes as input a dataset collection of the HDF5 dimspy "peak matrix" objects (from step 5) and removes the blank peaks from the peak matrix object if the fold of the sample/blank intensity is greater than *n* (set by user). The outputs contain a dataset collection of HDF5 files each containing a dimspy "peak matrix" object of the aligned peaks from each sample-to-blank pair with the blank peaks removed as well as a data collection containing tab separated (tsv) file of each peak matrix.

7) anticipated_purity_dims tool: This tool takes as input a data collection of peak matrices in tsv format (from step 6). A reference to the directory containing the sample files should also be set in the parameters. For each peak within the peak matrixes, the anticipated precursor ion purity is calculated for an anticipated fragmentation acquisition (by default is expected that a ± 0.5 Da isolation window will be used).

8) CAMERA_DIMS tool: This tool takes as input a dataset collection of tsv peak matrixes (from step 7). For each matrices a modified version of CAMERA is run to calculate the adducts and isotopes for each matrix. The output is a dataset collection of each original input matrix with columns for adduct and isotope appended.

9) deconrank tool: This tool takes as input a dataset-collection of tsv files (from step 8) with the following columns: peakID, mz, intensity, purity, adducts and isotope. For each peak matrix, the peaks are deconvoluted based on the CAMERA adduct and isotope information and then ranked and filtered

based on intensity, precursor ion purity, adduct and isotope form. The outputs contain: A dataset collection of each peak matrix in tsv format where details of the ranking of each peak and a dataset collection of the target lists in a format suitable to be used the Visual scripting "XcaliburAuto" method.

10, 11) Collapse Collection (collapse_collection tool): This tool simply converts the data collection into a single file for easier readability.

8.12 DIMS-NEARLINE ASSESSMENT FOR DEEP METABOLOME ANNOTATION GALAXY WORKFLOW PARAMETERS

8.12.1 STEP 1: PROCESS SCANS (AND SIM-STITCH) - SAMPLE

1. Choose the source for the dataset
    a. Library directory name
2. Library directory containing *.mzml or *.raw files
    a. DMA_FinalDataset/NonPolar_C30LC_DIMS_MSn/2018_SAMPLES_FOR_PROCESSING
3. Filelist / Samplelist
    a. (NA)
4. Function to calculate the noise from each scan
    a. As shown in Xcalibur Qual Browser (Available for *.RAW files only)
5. Signal-to-noise ratio threshold
    a. 10.0
6. Filter specific windows or scan events?
    a. True
7. Include / Exclude scan event(s)
    a. Exclude
8. Descriptions
    a. Description 1
        i. Start m/z for scan event
            1. 190.0
        ii. End m/z for scan event
            1. 1200.0
        iii. Scan type

1. Full scan

9. Show options for multiple scans:
    a. Minimum number of scans required for each m/z window or event
        i. 1
    b. Ppm error tolerance
        i. 5.0
    c. Minimum fraction (i.e. percentage) of scans a peak has to be present in.
        i. 0.7
    d. Relative standard deviation threshold
        i. 0.0
10. Show advanced options:
    a. Skip SIM-Stitching?
        i. False
    b. Relative intensity threshold used to remove ringing artifacts
        i. 0.0

8.12.2 STEP 2: PROCESS SCANS (AND SIM-STITCH) - BLANK

1. Choose the source for the dataset
    a. Library directory name
2. Library directory containing *.mzml or *.raw files
    a. DMA_FinalDataset/NonPolar_C30LC_DIMS_MSn/2018_BLANKS_FOR_PROCESSING
3. Filelist / Samplelist
    a. select at runtime (NA)
4. Function to calculate the noise from each scan
    a. As shown in Xcalibur Qual Browser (Available for *.RAW files only)
5. Signal-to-noise ratio threshold
    a. 10.0
6. Filter specific windows or scan events?
    a. True
7. Include / Exclude scan event(s)
    a. Exclude
8. Descriptions
    a. Description 1
        i. Start m/z for scan event
            1. 190.0
        ii. End m/z for scan event
            1. 1200.0
        iii. Scan type
            1. Full scan
9. Show options for multiple scans:

a. Minimum number of scans required for each m/z window or event
    i. 1
b. Ppm error tolerance
    i. 5.0
c. Minimum fraction (i.e. percentage) of scans a peak has to be present in.
    i. 0.2
d. Relative standard deviation threshold
    i. 0.0

10. Show advanced options:
    a. Skip SIM-Stitching?
        i. False
    b. Relative intensity threshold used to remove ringing artifacts
        i. 0.0

8.12.3 STEP 3: DMA_FILELIST_GENERATION

1. Library directories containing *.mzml or *.raw files
    a. Library directories containing *.mzml or *.raw files 1
        i. library
            1. DMA_FinalDataset/NonPolar_C30LC_DIMS_MSn/2018_SAMPLES_FOR_PROCESSING
2. Library directories containing *.mzml or *.raw files 2
    a. library
        i. DMA_FinalDataset/NonPolar_C30LC_DIMS_MSn/2018_BLANKS_FOR_PROCESSING
3. Check mzML files?
    a. False
4. Check raw files?
    a. True
5. Create filelist for mzML files?
    a. False
6. Create filelist for raw files?
    a. True

8.12.4 STEP 4: MERGE PEAKLISTS

1. Peaklist to merge
    a. Output dataset 'hdf5_file_out' from step 1,Output dataset 'hdf5_file_out' from step 2
2. Create multiple merged peaklists?
    a. True
3. Filelist
    a. Output dataset 'filelist_raw' from step 3

8.12.5 STEP 5: ALIGN SAMPLES

1. Peaklists (HDF5 file)
    a. Output dataset 'multiple_merged' from step 4
2. Filelist / Samplelist
    a. NA
3. ppm
    a. 5.0
4. Show options for additional output (*.tsv files):
    a. Standard output?
        i. False
    b. Comprehensive output?
        i. False
    c. Should the rows or columns represent the samples?
        i. Rows
    d. The Peak Matrix should contain ... values
        i. Intensity

8.12.6 STEP 6: BLANK FILTER

1. Peak Intensity Matrix (HDF5 file)
    a. Output dataset 'hdf5_file_out' from step 5
2. Label for the blank samples
    a. blank
3. Minimum fraction (percentage)
    a. 1.0
4. Function
    a. mean
5. min_fold_change
    a. 10.0
6. Remove blank samples (rows)
    a. True
7. Show options for addtional output (*.tsv files):
    a. Standard output?
        i. False
    b. Comprehensive output?
        i. True
    c. Should the rows or columns represent the samples?
        i. Columns
    d. The Peak Matrix should contain ... values
        i. Intensity

8.12.7 STEP 7: ANTICIPATED_PURITY_DIMS

1. peaks_file

    a. Output dataset 'matrix_comprehensive_file_out' from step 6

2. Choose the source for the dataset
    a. Library directory name (to be used with dimspy workflows) or path to an individual .mzML file
3. Library directory containing *.mzml files or path to an individual *.mzML
    a. DMA_FinalDataset/NonPolar_C30LC_DIMS_MSn/2018_mzML_FOR_PROCESSING
4. minoffset
    a. 0.5
5. maxoffset
    a. 0.5
6. ilim
    a. 0.05
7. Normalisation for isolation efficiency
    a. No normalisation
8. Handling of isotopic peaks
    a. Exclude C12/C13 isotopes in precursor ion purity calculation
9. ppm
    a. 5.0
10. Remove rows where mz value is NA or NaN?
    a. False
11. SIM-Stitch experiment?
    a. True
12. dimspy peak matrix text file usage?
    a. dimspy used to prepare mz file

i. File number

8.12.8 STEP 8: CAMERA_DIMS

1. peaks_file
    a. Output dataset 'anticipated_purity_dims' from step 7
2. ppm (adduct)
    a. 3.5
3. ppm (iso)
    a. 3.5
4. mzabs (adduct)
    a. 0.001
5. mzabs (iso)
    a. 0.001
6. maxmol
    a. 3
7. maxiso
    a. 4
8. maxcharge
    a. 3
9. polarity
    a. positive
10. Intensity filter
    a. 5000.0
11. Rules for adducts
    a. Extended adducts
12. Export the adduct ruleset used?
    a. False

8.12.9 STEP 9: DECONRANK

1. peaks_file
    a. Output dataset 'camera_annotated_peaklist' from step 8
2. technology
    a. DI-MS
3. polarity
    a. positive
4. Scores for adducts
    a. Default
5. precursor ion purity threshold
    a. 0.05
6. Second tier percentage
    a. 0.5
7. isotope removal
    a. [M+1]+ [M+2]+ [M+1]2+ [M+2]2+ [M+1]- [M+2]- [M+1]2- [M+2]2-
8. Use extended output for the scores tsv file?
    a. False
9. Weightings for scores:
    a. adduct
        i. 0.3
    b. intensity
        i. 0.3
    c. precursor-ion-purity
        i. 0.2
    d. clustern

i. 0.2

10. Scheduling features for DIMS fragmentation experimental run:

a. max_time

i. 1800.0

b. min_time

i. 120.0

c. max_cid_time

i. 300.0

d. peak_time_hcd

i. 14.0

e. peak_time_cid

i. 12.0

f. percentage_cid

i. 0.333

g. delay_time

i. 24.0

h. Use default method template file names?

i. Default

8.12.10 STEP 10: COLLAPSE COLLECTION

1. Collection of files to collapse into single dataset
    - Output dataset 'XcaliburAutoInput' from step 9
2. Keep one header line
    - True
3. Append File name
    - False

8.12.11 STEP 11: COLLAPSE COLLECTION

1. Collection of files to collapse into single dataset
    a. Output dataset 'scores' from step 9
2. Keep one header line
    a. True
3. Append File name
    a. True
4. Where to add dataset name
    a. Same line and each line in dataset

8.13 XML2ISA SOFTWARE SUMMARY

Table 8-4 XML2ISA Software suite

Interface	Technology	Software Name	Repository	XML Format
CLI and API	MS	mzMl2ISA	https://github.com/ISA-tools/mzML2ISA	mzML, imzML
	NMR	nmrML2ISA	https://github.com/ISA-tools/nmrML2ISA	nmrML
GUI	MS	mzML2ISA-qt	https://github.com/ISA-tools/mzML2ISA-qt	mzML
		imzML2ISA-qt	https://github.com/ISA-tools/imzML2ISA-qt	imzML
	NMR	nmrML2ISA-qt	https://github.com/ISA-tools/nmrML2ISA-qt	nmrML
Galaxy Tool	MS	mzML2ISA-galaxy	https://github.com/ISA-tools/mzML2ISA-galaxy	mzML, imzML
	NMR	nmrML2ISA-galaxy	https://github.com/ISA-tools/nmrML2ISA-galaxy	nmrML

Footnote:

CLI: Command Line Interface, GUI: Graphical User Interface, API: Application Program Interface

8.14 XML2ISA IMPLEMENTATION (FURTHER DETAILS)

Both mzML2ISA and nmrML2isa packages are fully compatible with Python 2.7 and 3.5 (CPython implementation). Continuous Integration relies on Travis CI (https://travis-ci.org/) and AppVeyor (https://appveyor.com/) for automated building and testing of packages on multiple operating systems, using MetaboLights data repository as model data (available open and free via FTP). Development is being carried out on the experimental branch of the Github repository of each project while distribution is assured via Github releases and PyPI (https://pypi.python.org/pypi). Code reference for citation is available via Zenodo DOIs (https://zenodo.org/). Documentation is generated using Sphinx (http://www.sphinx-doc.org/) and hosted on readthedocs.io (https://readthedocs.org/). The mzML2ISA, nmrML2ISA and imzML2ISA GUIs are built using PyQt5 (https://pypi.python.org/pypi/PyQt5), and only work with Python 3.

8.15 XML2ISA METABOLIGHTS STUDY DETAILS

Table 8-5 MetaboLights Studies with XML Open Source RAW File Formats

Study ID	Title	XML
MTBLS1	A metabolomic study of urinary changes in type 2 diabetes in human compared to the control group	nmrML
MTBLS32	Lipid mediators of inflammation in BALF 6-19 days after infection with influenza.	mzML
MTBLS36	Metabolic differences in ripening of Solanum lycopersicum 'Ailsa Craig' and three monogenic mutants	mzML
MTBLS38	Metabolite Standards for the development and validation of MassCascade	mzML
MTBLS67	Metabolomic Analysis of Fission Yeast at the Onset of Nitrogen Starvation	mzML
MTBLS87	Unexpected similarities between the Schizosaccharomyces and human blood metabolomes, and novel human metabolites (Blood fraction)	mzML
MTBLS88	Unexpected similarities between the Schizosaccharomyces and human blood metabolomes, and novel human metabolites (Blood plasma and RBC fractions)	mzML
MTBLS125	Distribution of RESV and its metabolite peaks in mouse tissues after oral and skin administration	mzML
MTBLS126	Absorption efficiency of RESV through mouse skin using 3 bases in different tissues	mzML
MTBLS127	Resveratrol metabolism in HepG2 (human hepatocytes), HaCaT (human keratinocytes), and C2C12 (mouse myoblasts)	mzML
MTBLS137	MetaDB a Data Processing Workflow in Untargeted MS-Based Metabolomics Experiments	mzML
MTBLS140	Metabolome analysis via an HPLC-ESI-MS-based experimental and computational pipeline for chronic nephron toxicity profiling	mzML
MTBLS228	Untargeted extraction of metabolites 13C labelling profiles from time course labelling switch experiment	mzML
MTBLS229	Untargeted extraction of metabolites 13C labelling profiles from time course labelling switch experiment	mzML
MTBLS263	Individual variability in human blood metabolites identifies age-related differences - determination of coefficients of variation for each metabolite (3 injections of same sample, 3 independent sample preparations).	mzML
MTBLS265	Individual variability in human blood metabolites identifies age-related differences (30 persons, whole blood data).	mzML
MTBLS266	Individual variability in human blood metabolites identifies age-related differences (30 persons, plasma data).	mzML
MTBLS273	Metabolic phenotyping of ex-vivo breast samples by DESI mass spectrometry imaging	imzML
MTBLS289	Analysis of colorectal adenocarcinoma tissue samples by DESI mass spectrometry imaging	imzML
MTBLS315	Towards improving point-of-care diagnosis of non-malaria febrile illness: a metabolomics approach	mzML
MTBLS341	Piriformospora indica stimulates root metabolism of Arabidopsisthaliana	mzML

8.16 XML2ISA BENCHMARKING RESULTS

Assessment ran on Ubuntu 16.04.1 LTS 64bit Virtual Machine (Virtual Box 5.0.3) with a maximum of 4 CPUs and 4 GB of memory. The times shown are a mean calculated from 5 repeat measurements.

Table 8-6 Benchmarking of Sample XML Data Files

MetaboLights ID	Mean File size (MB)	Python version	CPUs	Mean Time (seconds)	StdDev
MTBLS1 (nmrML)	0.51	2.7	1	1.848	1.22
			4	1.202	0.27
		3.5	1	0.984	4.22
			4	2.028	0.06
MTBLS127 (mzML)	78.8	2.7	1	75.02	0.29
			4	44.47	0.72
		3.5	1	75.05	0.72
			4	43.42	1.04
MTBLS137 (mzML)	51.9	2.7	1	77.01	1.98
			4	46.19	4.97
		3.5	1	62.84	4.80
			4	38.12	1.58
MTBLS315 (mzML)	25.6	2.7	1	32.07	0.29
			4	23.61	0.72
		3.5	1	31.20	0.72
			4	29.01	1.04
MTBLS273 (imzML)	7.8	2.7	1	20.27	1.70
			4	22.87	2.06
		3.5	1	20.65	1.63
			4	19.88	1.28

8.17 DJANGO-MBROWSE MODEL SCHEMA

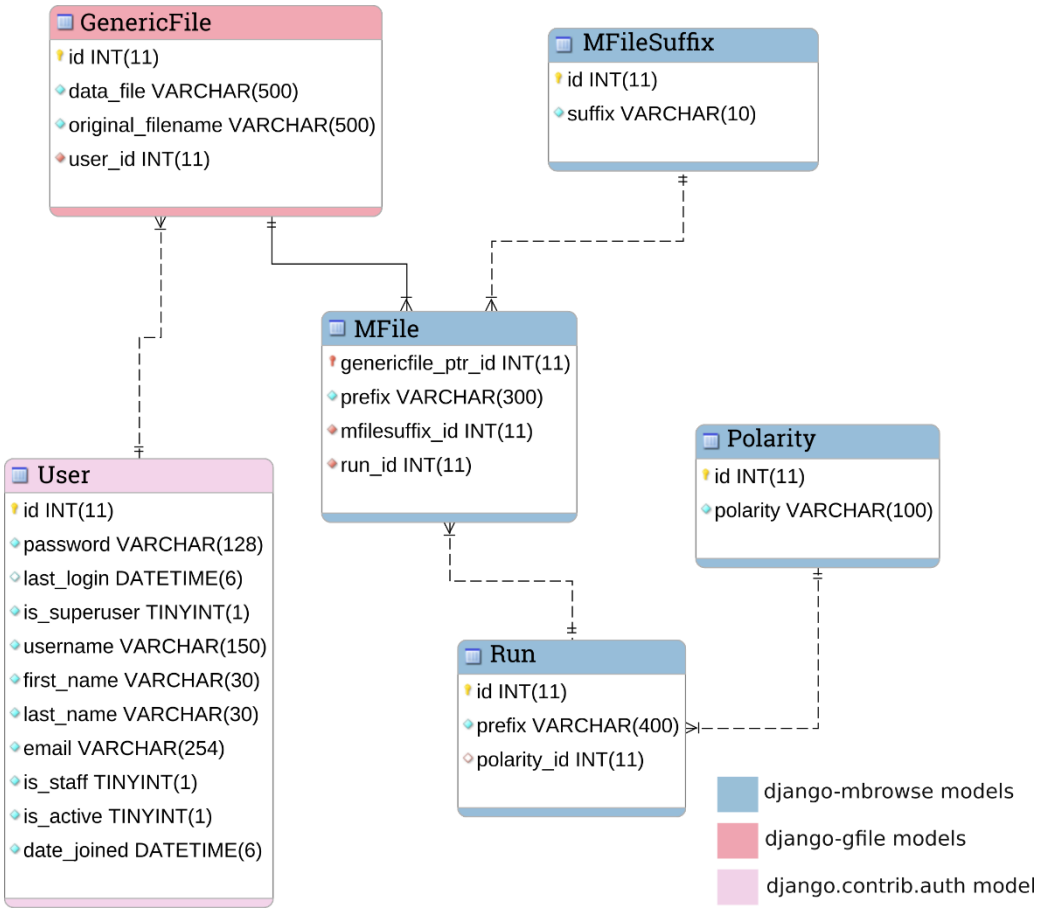


Figure 8-7 Metabolomic data file, generic file and Django user model schema (django-mbrowse, django-gfile)

The "MFile" model represents a metabolomics file, the "MFileSuffix" represents the suffix of the file (e.g. mzML), the "Run" model represents the acquisition on the instrument (as single acquisition run can in some cases generate multiple files), the "Polarity" model represents the polarity setting of the instrument that was used. The "MFile" model inherits the "GenericFile" model from the Django-gfiles application. The "GenericFile" model allows the management of different files to be handled and save in a consistent location where is file is linked to an authorised Django user via the "User" model (from django.contrib.auth).

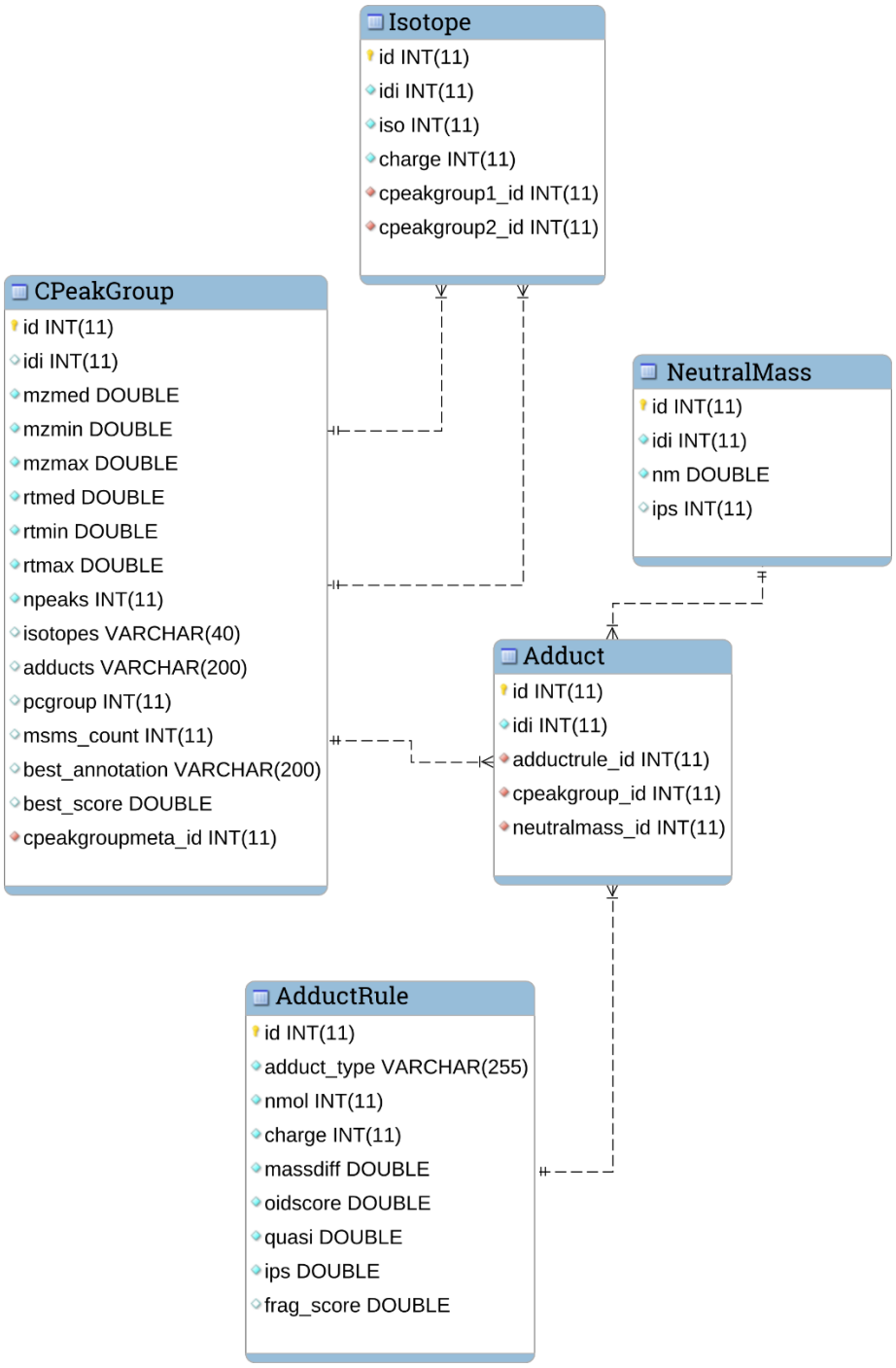


Figure 8-8 Adduct and isotope model schema (django-mbrowse)

Models for adduct and isotope relationships. The "CPeakGroup" can be associated with multiple adducts via the "adduct" model. Each adduct is linked to a neutral mass (represented as the NeutralMass model) and adduct rule (represented by the "AdductRule" model. The isotope relationships are also tracked via

the “Isotope” model.



Figure 8-9 Fragmentation for chromatographic feature model schema (django-mbrowse)

Models for fragmentation and associated chromatographic features. The "SPeak" model represents the scan peaks where many scan peaks are derived from one scan represented by the "SPeakMeta" model that also includes all other information regarding that scan (precursor ion purity, retention time, etc). For space considerations, only the scan peaks of fragmentation spectra are stored in the database. The "CPeak" model representing and individual chromatographic feature has a many-to-many relationship

with the “SPeakMeta” model (using the SPeakMetaCPeakFragLink model to link the models). Each “SPeakMeta” is associated with a “Run” model that represents the instrument acquisition.

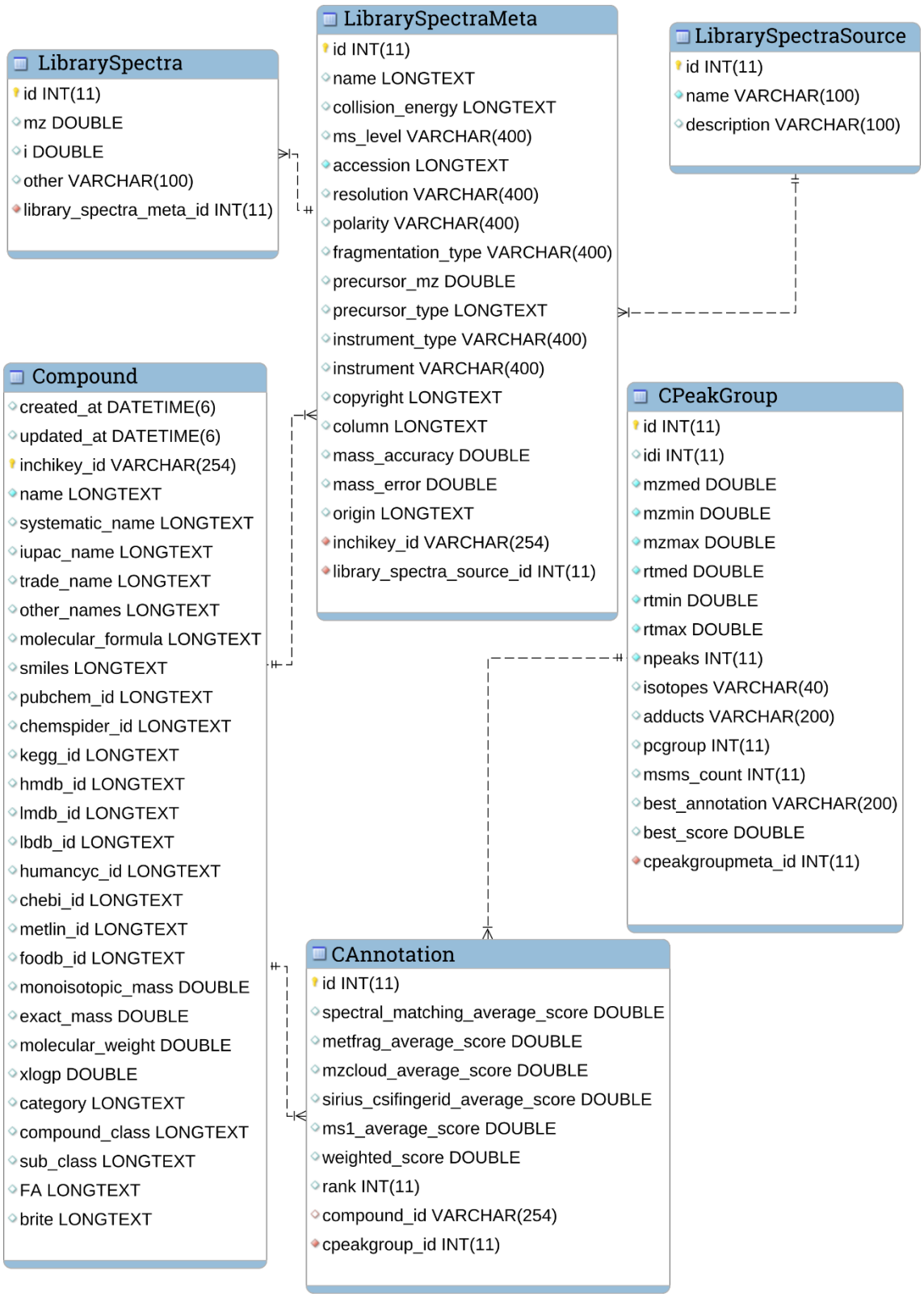


Figure 8-10 Library-Compound-Annotation model schema (django-mbrowse)

Models representing spectral libraries, compounds, annotations and how they relate to a

chromatographic feature. The “Compound” model represents a unique InChiKey compound and the schema is an extension of the MetPlus SQLite database schema (Madhavan et al., 2014). The “LibrarySpectra” model represents the library spectra where many spectra can be associated with one “LibrarySpectraMeta” detailing information regarding the spectra’s acquisition. The “LibrarySpectraMeta” model is an associated with a “LibrarySpectraSource” detailing the spectra’s general origin (e.g. MoNA MassBank). The “CAnnotation” model represents a summary of all the annotation approaches that are currently being used within the django-mbrowse application and allows multiple “Compound” objects to be associated with multiple “CPeakGroup” objects.

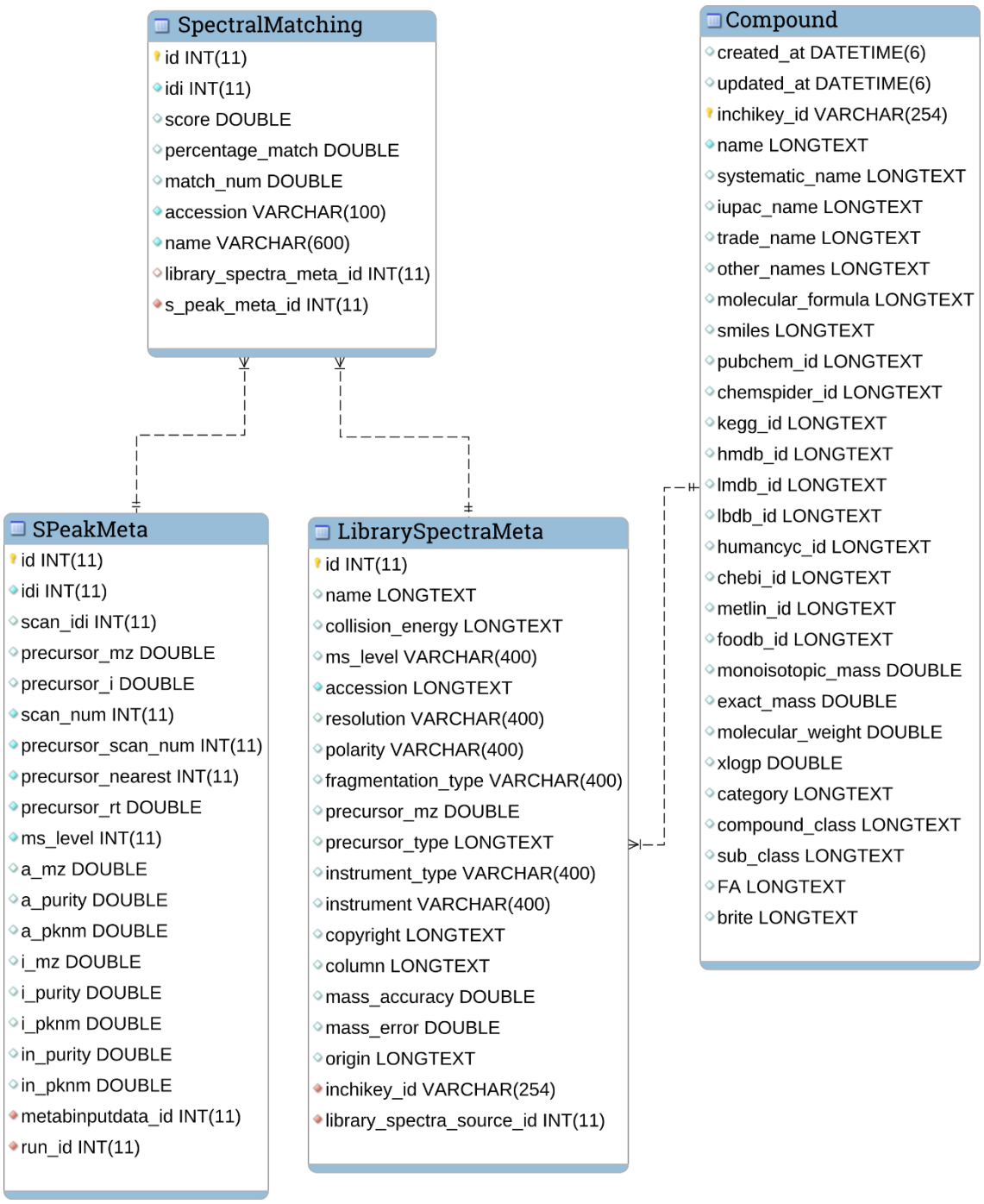


Figure 8-11 Spectral matching of fragmentation spectra model schema (django-mbrowse)

Models representing per fragmentation scan spectral matching annotation. “SPeakMeta” represent the scan information for a user generated fragmentation spectrum, see Figure 8-9 for more details. “Compound” represents a unique InChiKey compound and “LibrarySpectraMeta” represents the acquisition information regarding a “LibrarySpectra” object, see Figure 8-10 for more details. The “SpectralMatching” model represents the spectral matching information (e.g. spectral matching score), where many “SPeakMeta” objects can be associated to many “LibrarySpectraMeta” objects via the

"SpectralMatching" model.

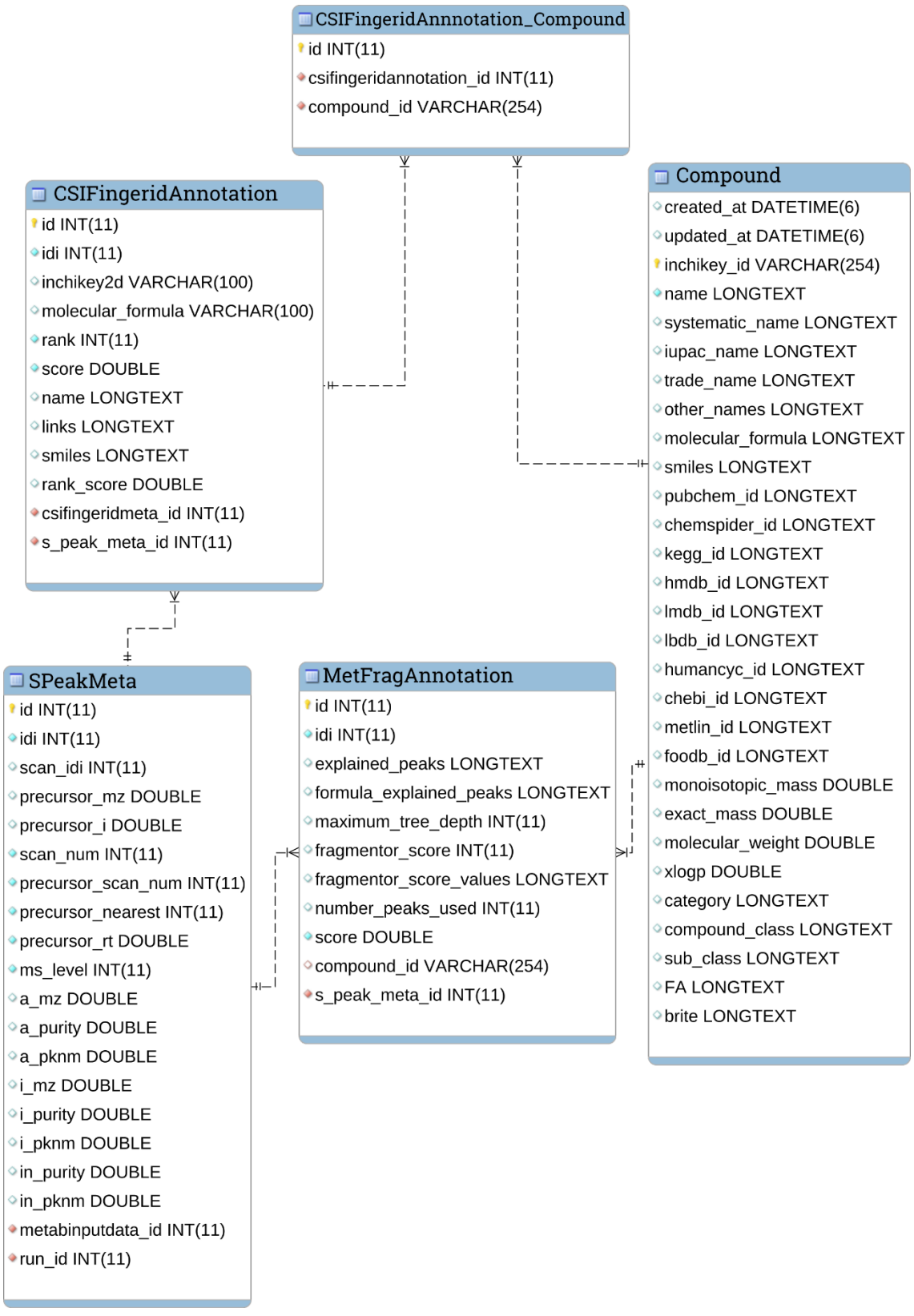


Figure 8-12 MetFrag and Sirius CSI:FingerID annotation model schema (django-mbrowse)

Models representing the per fragmentation scan annotations for MetFrag and Sirius CSI:FingerID. “SPeakMeta” represent the scan information for a user generated fragmentation spectrum, see Figure 8-9 for more details. “Compound” represents a unique InChiKey compound, see Figure 8-10 for more details. The “MetFragAnnotation” model represents the annotation information from MetFrag, where many “SPeakMeta” objects can be associated to many “Compound” objects depending on what spectral matching has been performed. The “CSIFingeridAnnotation” model represents the annotation information from Sirius CSI:FingerID, where each annotation from Sirius CSI:FingerID is represented as only the first two components of an InChiKey. This means that one many CSIFingeridAnnotation can correspond to multiple Compounds via the “CSIFingeridAnnotation_Compound” linking model.

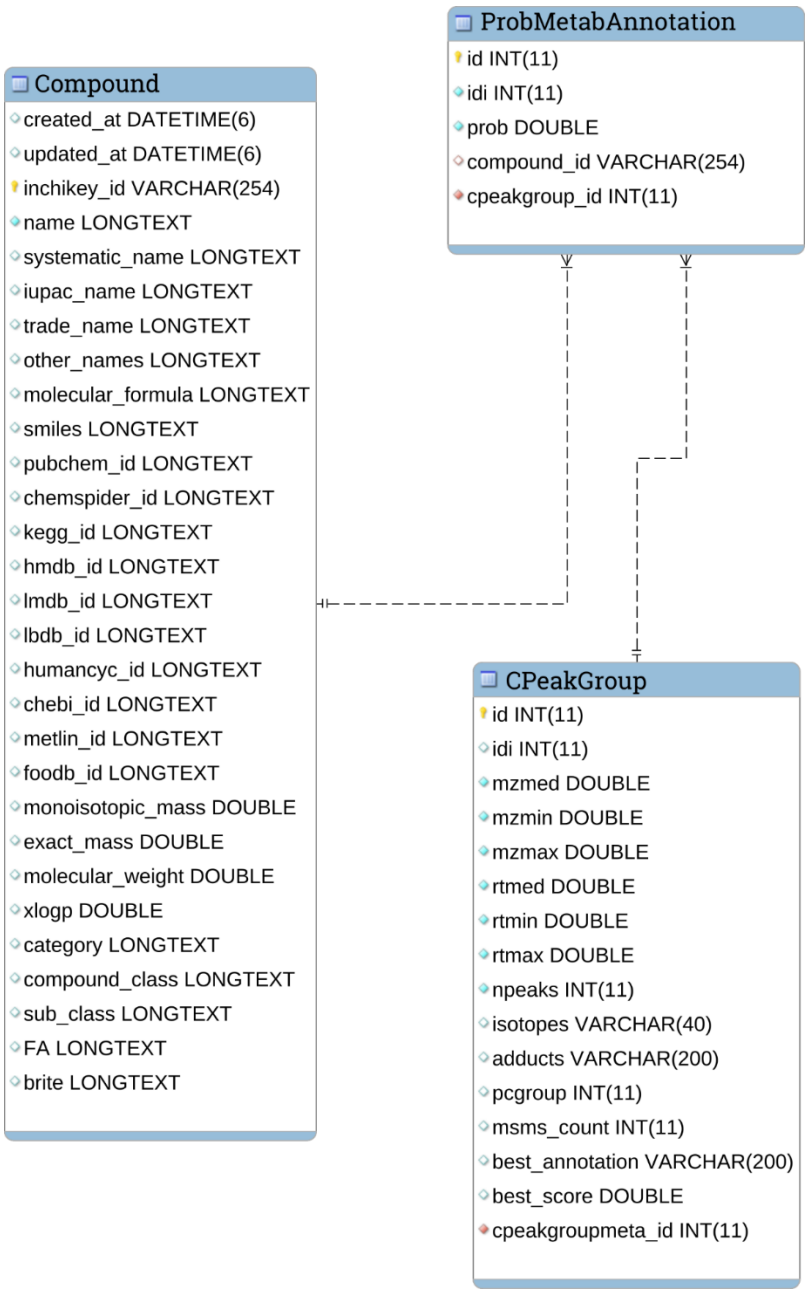


Figure 8-13 MS[1] annotation with ProbMetab model schema (django-mbrowse)

Models representing the MS[1]annotation using the ProbMetab software. The "CPeakGroup" model represents the grouped chromatographic feature, see **Error! Reference source not found.** and Figure 8-8 for more information. The "ProbMetabAnnotation" represents the annotation information from ProbMetab where many "CPeakGroup" objects can be annotated with many "Compound" objects through the "ProbMetabAnnotation" model.

8.18 LC-MS(/MS) DMA WORKFLOW SETTINGS

8.18.1 STEP 1: INPUT DATASET COLLECTION

1. Data files (mzML)
    a. Select at runtime

8.18.2 STEP 2: INPUT DATASET

1. Samplelist (tsv, tabular, txt)
    a. Select at runtime

8.18.3 STEP 3: ASSESS_PURITY_MSMS

1. Choose the source for the dataset
    a. .mzML files
2. *.mzML files
    a. Output dataset 'output' from step 1
3. Use most intense peak within isolation window for precursor?
    a. True
4. Use nearest full scan to determine precursor?
    a. True
5. Offsets
    a. Uses offsets determined in the mzML file
6. Ilim
    a. 0.05
7. Normalisation for isolation efficiency
    a. Calculated from Q-Exactive for +/- 0.5 Da windows
8. Handling of isotopic peaks
    a. Exclude C12/C13 isotopes in precursor ion purity calculation

8.18.4 STEP 4: XCMS.XCMSSET

1. File(s) from your history containing your chromatograms
    a. Output dataset 'output' from step 1
2. Scan range option
    a. Hide
3. Extraction method for peaks detection
    a. centWave
4. Max tolerated ppm m/z deviation in consecutive scans in ppm
    a. 5
5. Min,Max peak width in seconds
    a. 20,50
6. Advanced options
    a. Show
7. Signal/Noise threshold
    a. 100
8. Minimum difference in m/z for peaks with overlapping retention times
    a. 0.001
9. peak limits method
    a. peak limits based on smoothed 2nd derivative (less precise)
10. Prefilter step for the first phase
    a. 3,100
11. Noise filter
    a. 0

8.18.5 STEP 5: XCMS.XCMSSET MERGER

1. xset RData / xcms.xcmsSet file
    a. Output dataset 'xsetRData' from step 4
2. Sample metadata file
    a. Output dataset 'output' from step 2

8.18.6 STEP 6: TRACK_RT_RAW

1. xset
    a. Output dataset 'xsetRData' from step 5
2. Save peaks to a tsv file
    a. False

8.18.7 STEP 7: XCMS.RETCOR

1. xset RData file
    a. Output dataset 'xset_with_rt_raw' from step 6
2. Method to use for retention time correction
    a. obiwarp
3. Step size (in m/z)
    a. 1.0
4. Resubmit your raw dataset or your zip file:
    a. Resubmit your dataset or your zip file
        i. yes
    b. File(s) from your history containing your chromatograms
        i. Output dataset 'output' from step 1

8.18.8 STEP 8: XCMS.GROUP

1. xset RData file
    a. Output dataset 'xsetRData' from step 7
2. Method to use for grouping
    a. density
3. Bandwidth
    a. 5.0
4. Minimum fraction of samples necessary
    a. 0.5
5. Width of overlapping m/z slices
    a. 0.025
6. Advanced options
    a. show
7. Maximum number of groups to identify in a single m/z slice
    a. 50
8. Get a Peak List
    a. False

8.18.9 STEP 9: FLAG_REMOVE_PEAKS

1. xset_path
    a. Output dataset 'xsetRData' from step 8
2. Change biological sample flag parameters?
    a. Update biological sample flag parameters
3. rsd_i_sample
    a. NA
4. minfrac_sample
    a. 0.7
5. rsd_rt_sample
    a. NA
6. ithres_sample
    a. NA
7. Change blank flag parameters?
    a. Use default blank flag parameters
8. Remove peaks from xcmsSet object?
    a. Remove peaks and re-group
9. minfrac_xcms
    a. 0.5
10. mzwid
    a. 0.025
11. bw
    a. 5.0
12. Advanced parameters
    a. Use default advanced parameters

13. Samplelist
    a. Use samplelist
14. samplelist
    a. Output dataset 'output' from step 2

8.18.10 STEP 10: XCMS.FILLPEAKS

1. xset RData file
    a. Output dataset 'xset_filtered' from step 9
2. Filling method
    a. chrom
3. Get a Peak List
    a. False
4. Resubmit your raw dataset or your zip file:
    a. Resubmit your dataset or your zip file
        i. yes
    b. File(s) from your history containing your chromatograms
        i. Output dataset 'output' from step 1

8.18.11 STEP 11: FRAG4FEATURE

1. xcmsSet object
    a. Output dataset 'xsetRData' from step 10
2. purityA object
    a. Output dataset 'assess_purity_msms_rdata' from step 3
3. For matching fragmentation to a feature, use most intense peak within isolation window for precursor?
        i. True
4. Was retention time correction used?
    a. False (Note: retention time correction was not working at the time of writing. This will have some minimal impact on the annotations of certain peaks)
5. ppm
    a. 10.0
6. plim
    a. 0.0
7. Create SQLite database of spectra?
    a. False
8. Add a different grouped peaklist to database?
    a. Use default grouped peaklist
9. Resubmit your dataset
    a. no need

8.18.12 STEP 12: CAMERA.ANNOTATE

1. RData file
    a. Output dataset 'xsetRData' from step 10
2. Group co-eluted peaks based on RT [groupFWHM]:
    a. Multiplier of the standard deviation
        i. 6
    b. Percentage of FWHM width
        i. 0.6
3. Annotation general options:
    a. General ppm error
        i. 5
    b. General absolute error in m/z
        i. 0.015
4. Annotate Isotopes [findIsotopes]:
    a. Max. ion charge
        i. 3
    b. Max. number of expected isotopes
        i. 4
    c. The percentage number of samples, which must satisfy the C12/C13 rule for isotope annotation
        i. 0.5
5. Mode
    a. All functions
6. Verifying grouping co-eluted peaks [groupCorr]:
    a. groupCorr: correlation threshold (0..1)

b. 0.75

c. groupCorr: Method selection for grouping peaks after correlation analysis into pseudospectra

    i. hcs

d. groupCorr: significant correlation threshold

    i. 0.05

e. groupCorr: Use correlation inside samples for peak grouping

    i. True

f. groupCorr: Use isotopic relationship for peak grouping

    i. False

g. groupCorr: Use correlation across samples for peak grouping

    i. False

7. Annotate Adducts [findAdducts]:

    a. Which polarity mode was used for measuring of the ms sample

        i. (Note: this was set as either “positive” or “negative” depending on the polarity of the data)

8. How much peaks will be calculated in every thread using the parallel mode

    a. 100

9. Use a personal ruleset file

    a. FALSE

10. If no ruleset is provided, calculate ruleset with max. number n of [nM+x] clusterions

    a. 3

11. Statistics and results export: [diffreport]:

    a. Number of condition

    b. One condition

12. Export options:
    a. Convert retention time (seconds) into minutes
        i. False
    b. Number of decimal places for mass values reported in ions' identifiers.
        i. 4
    c. Number of decimal places for retention time values reported in ions' identifiers.
        i. 0
13. General used intensity value
    a. into
14. Resubmit your raw dataset or your zip file:
    a. Resubmit your dataset or your zip file
        i. yes
    b. File(s) from your history containing your chromatograms
        i. Output dataset 'output' from step 1

8.18.13 STEP 13: CREATE MSP FILES FROM MSPURITY

1. msPurity frag4feature dataset
    a. Output dataset 'frag4feature_rdata' from step 11
2. Spectra use mode (if precursor was fragmented in >1 scans)
    a. All
3. Allowed mass deviation between fragment MZs for spectra averaging (in ppm)
    a. Not available.

8.18.14 STEP 14: PROBMETAB TOOL

1. Choose your acquisition mode
    a. One acquisition charge mode
2. Annotate RData
    a. Output dataset 'rdata_quick_false' from step 12
3. Resubmit your raw dataset or your zip file:
    a. Resubmit your dataset or your zip file
        i. yes
    b. File(s) from your history containing your chromatograms
        i. Output dataset 'output' from step 1
4. Extraction of CAMERA annotation [get.annot]:
    a. Retrieves peaks with no evidence of adduct or isotope
        i. True
    b. Exclude samples
        i. hide
5. Database matching [create.reactionM]:
    a. Search on KEGG database or multiple organisms
        i. KEGG
    b. Parts per million mass tolerance allowed in the mass search
        i. 8
6. Probability calculations matrix export [export.class.table]:
    a. Calculation of the probability to attribute a mass to a compound
        i. Count
    b. Logical, check if you want to generate a HTML ProbMetab report
        i. True

7. Calculate the correlations and partial correlations and cross reference then with reactions [reac2cor]:
    a. Correlation option
        i. cor
    b. Probability that the correlation is considered significant
        i. 0.8
    c. Probability that the partial correlation is considered significant.
        i. 0.8
    d. Correlation intensity threshold
        i. 0.75

8.18.15 STEP 15: MSP_SPLIT

1. Input MSP file
    a. Output dataset 'msp_out' from step 13
2. Number of new files
    a. 100

8.18.16 STEP 16: MSP_SPLIT

1. Input MSP file
    a. Output dataset 'msp_out' from step 13
2. Number of new files
    a. 250

8.18.17 STEP 17: SIRIUS-CSI:FINGERID

1. MSP file (Output from Create MSP tool)
    a. Output dataset 'multiple' from step 15
2. Select SIRIUS-CSI:FingerID Database
    a. PubChem
3. Mass deviation of the fragment peaks in ppm
    a. 10
4. The maximum number of candidates in the output
    a. 5
5. Ion Mode
    a. (Note: this was set as either “positive” or “negative” depending on the polarity of the data)
6. Analysis used
    a. Orbitrap

8.18.18 STEP 18: METFRAG

1. MSP file (Output from Create MSP tool)
    a. Output dataset 'multiple' from step 16
2. Local Database (CSV)
    a. select at runtime
3. Available MetFrag Databases
    a. PubChem
4. Relative Mass Deviation
    a. 5
5. Fragment Peak Match Relative Mass Deviation
    a. 5
6. Fragment Peak Match Absolute Mass Deviation
    a. 0.001
7. Ion Mode
    a. (Note: this was set as either “positive” or “negative” depending on the polarity of the data)

8.18.19 STEP 19: COLLAPSE COLLECTION

1. Collection of files to collapse into single dataset
    a. Output dataset 'results' from step 17
2. Keep one header line
    a. True
3. Append File name
    a. False

8.18.20 STEP 20: COLLAPSE COLLECTION

1. Collection of files to collapse into single dataset
    a. Output dataset 'results' from step 18
2. Keep one header line
    a. True
3. Append File name
    a. False

8.18.21 STEP 21: CREATE_SQLITE_DB

1. xcmsSet or CAMERA object
    a. Output dataset 'rdata_quick_false' from step 12
2. Use CAMERA object or XCMS object from RData?
    a. CAMERA (xa)
3. purityA object (frag4feature output)
    a. Output dataset 'frag4feature_rdata' from step 11
4. Include EIC data?
    a. False
5. Use user defined Raw RT columns?
    a. True
6. Add a different grouped peaklist to database?
    a. Use default grouped peaklist
7. Resubmit your dataset
    a. no need
8. Include MetFrag results?
    a. yes
9. MetFrag result
    a. Output dataset 'output' from step 20
10. Include Sirius CSI-FingerID results?
    a. Yes (Note: all negative analysis for Sirius CSI:FignerID resulted in no annotations, as such for negative analysis this was assigned “No”)
11. Sirius CSI-FingerID result
    a. Output dataset 'output' from step 19 (Note: for negative analysis this was assigned blank – see above)

12. Include ProbMetab results?

    a. yes

13. ProbMetab result

    a. Output dataset 'variableMetadata' from step 14

8.18.22 STEP 22: SPECTRAL_MATCHING

1. SQLite DB of target spectra
    a. Output dataset 'lcms_data_sqlite' from step 21
2. Use custom library SQLite Database?
    a. no
3. polarity
    a. (Note: this was set as either “positive” or “negative” depending on the polarity of the data)
4. ppm_tol_prod
    a. 10.0
5. ppm_tol_prec
    a. 5.0
6. score_thres
    a. 0.1
7. instrument_types
    b. CE-ESI-TOF ESI-ITFT ESI-ITTOF ESI-QTOF LC-ESI-IT LC-ESI-ITFT LC-ESI-ITTOF LC-ESI-QFT LC-ESI-QIT LC-ESI-QQ LC-ESI-QTOF LC-ESI-TOF (Note: for apolar lipids this was left blank so that in silico LipidBlast fragmentation spectra could be used)
8. library_sources
    a. LipidBlast MassBank
9. Create a new database for the results?
    a. False
10. advanced:
    a. Only use the top n spectral matching hits?

    i. False

b. ra_thres_l

    i. 0.0

c. ra_thres_t

    i. 2.0

d. Specific scan ids

    i. NA

8.18.23 STEP 23: SUBMIT_TO_MOGI

1. Data to submit to MOGI
    a. Output dataset 'lcms_data_sqlite' from step 21
2. MOGI URL
    a. https://dmadb.researchcloud.bham.ac.uk
3. Name of Galaxy instance in MOGI
    a. brum_galaxy_dev
4. Name of the investigation in submit to in MOGI
    a. daphnia-magna-dma
5. Galaxy URL
    a. https://metabolomics-dev.galaxy.bham.ac.uk

8.19 CHROMATOGRAPHY METHODS FOR DMA

Details provided by M. R. Jones.

8.19.1 SYNCRONIS PHENYL (PHE)

A reversed-phase gradient elution procedure was applied using a Syncronis Phenyl LC column (2.1 x 100 mm [ID x length], Thermo Fisher Scientific) fitted with an in-line filter cartridge (0.2 μm pore size, 2.1 mm ID). Mobile phase A comprised 90:5:5% v/v/v water:methanol:100 mM ammonium acetate buffer (pH 5.80, aq.), while mobile phase B comprised 95:5% v/v methanol:100 mM ammonium acetate buffer (p 5.80, aq.). The elution gradient was as follows: 0/3/21/24/24.1/25/30 min at 0/0/100/100/0/0/0% B. Throughout the analysis, the following were held constant: mobile phase flow rate at 0.4 mL/min; column oven temperature at 40 °C; autosampler module temperature at 10 °C. A total of 10 μL of sample was injected. Injection was performed using a 20 μL fused-silica loop in combination with a user-defined injection program, in which 10 μL of sample was plugged between aliquots of mobile phase A. Note The injection needle was rinsed at the end of the gradient elution procedure using a solution comprising 80:20% v/v methanol:water

8.19.2 ACCUCORE C30 (C30)

A reversed-phase gradient elution procedure was applied using an Accucore C30 LC column (2.1 x 100 mm [ID x length], 2.6 μm particle diameter) fitted with an in-line filter cartridge (0.2 μm pore size, 2.1 mm ID). Mobile phase A comprised 50:50% v/v ACN:[100 mM ammonium formate + 1% v/v formic acid (aq. solution)], while mobile phase B comprised 10:88:2% v/v/v ACN:IPA:[100 mM AmFm + 1% FA (aq. solution)]. The elution gradient was as follows: 0/6/14/23/26/26.1/30 mins at 22/60/85/100/100/22/22% B. Throughout the analysis, the following were held constant: mobile phase flow rate at 0.4 mL/min; column oven temperature at 35 °C; autosampler module temperature at 10 °C. A total of 10 μL of sample was injected.

Injection was performed using a 20 µL fused-silica loop in combination with a user-defined injection program, in which 10 µL of sample was plugged between aliquots of mobile phase A. The injection needle was rinsed at the end of the gradient elution procedure using a solution comprising 47.5:47.5:5% v/v/v acetonitrile:propan-2-ol:water.

8.19.3 ACCUCORE C30 (C30)

A HILIC-based gradient elution procedure was applied using an Accucore Amide LC column (2.1 x 100 mm [ID x length], 2.6 µm particle diameter) fitted with an in-line filter cartridge (0.2 µm pore size, 2.1 mm ID). Mobile phase A comprised 90:5:5% v/v/v acetonitrile:water:[98:2% v/v 100 mM ammonium acetate (aq.):acetic acid], while mobile phase B comprised 60:35:5% v/v/v ACN:H2O:[98:2% v/v 100 mM ammonium acetate (aq.):acetic acid]. The elution gradient was as follows: 0/1.5/21/25/26/30 min at 0/0/100/100/0/0% B. Throughout the analysis, the following were held constant: mobile phase flow rate at 0.4 mL/min; column oven temperature at 30 °C; autosampler module temperature at 10 °C. A total of 5 µL of sample was injected. Injection was performed using a 5 µL stainless steel loop in combination with a user-defined injection program, in which the 5 µL sample aliquot was forced in to the injection loop by aspirating an aliquot of mobile phase A. The injection needle was rinsed at the end of the gradient elution procedure using a solution comprising 80:20% v/v water:acetonitrile.

8.20 CHEMICAL DESCRIPTORS

Table 8-7 Summary of chemical descriptors used for PCA analysis

ID	Name	ID	Name	ID	Name	ID	Name
1	MaxEStateIndex	29	Chi3v	57	SMR_VSA5	85	VSA_EState2
2	MinEStateIndex	30	Chi4n	58	SMR_VSA6	86	VSA_EState5
3	MaxAbsEStateIndex	31	Chi4v	59	SMR_VSA7	87	VSA_EState6
4	MinAbsEStateIndex	32	HallKierAlpha	60	SMR_VSA9	88	VSA_EState8
5	qed	33	Ipc	61	SlogP_VSA1	89	VSA_EState9
6	MolWt	34	Kappa1	62	SlogP_VSA10	90	FractionCSP3
7	HeavyAtomMolWt	35	Kappa2	63	SlogP_VSA11	91	HeavyAtomCount
8	ExactMolWt	36	Kappa3	64	SlogP_VSA12	92	NHOHCount
9	NumValenceElectrons	37	LabuteASA	65	SlogP_VSA2	93	NOCount
10	NumRadicalElectrons	38	PEOE_VSA1	66	SlogP_VSA3	94	NumAliphaticCarbocycles
11	MaxPartialCharge	39	PEOE_VSA10	67	SlogP_VSA4	95	NumAliphaticHeterocycles
12	MinPartialCharge	40	PEOE_VSA11	68	SlogP_VSA5	96	NumAliphaticRings
13	MaxAbsPartialCharge	41	PEOE_VSA12	69	SlogP_VSA6	97	NumAromaticCarbocycles
14	MinAbsPartialCharge	42	PEOE_VSA13	70	SlogP_VSA7	98	NumAromaticHeterocycles
15	FpDensityMorgan1	43	PEOE_VSA14	71	SlogP_VSA8	99	NumAromaticRings
16	FpDensityMorgan2	44	PEOE_VSA2	72	TPSA	100	NumHAcceptors
17	FpDensityMorgan3	45	PEOE_VSA3	73	EState_VSA1	101	NumHDonors
18	BalabanJ	46	PEOE_VSA4	74	EState_VSA10	102	NumHeteroatoms
19	BertzCT	47	PEOE_VSA5	75	EState_VSA11	103	NumRotatableBonds
20	Chi0	48	PEOE_VSA6	76	EState_VSA2	104	NumSaturatedCarbocycles
21	Chi0n	49	PEOE_VSA7	77	EState_VSA3	105	NumSaturatedHeterocycles
22	Chi0v	50	PEOE_VSA8	78	EState_VSA4	106	NumSaturatedRings
23	Chi1	51	PEOE_VSA9	79	EState_VSA5	107	RingCount
24	Chi1n	52	SMR_VSA1	80	EState_VSA6	108	MolLogP
25	Chi1v	53	SMR_VSA10	81	EState_VSA7	109	MolMR
26	Chi2n	54	SMR_VSA2	82	EState_VSA8	110	fr_Al_COO
27	Chi2v	55	SMR_VSA3	83	EState_VSA9	111	fr_Al_OH
28	Chi3n	56	SMR_VSA4	84	VSA_EState10	112	fr_Al_OH_noTert

Footnote: Definitions available at https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors

Table 8-8 Summary of chemical descriptors used for PCA analysis (continued)

ID	Name	ID	Name	ID	Name	ID	Name
113	fr_ArN	141	fr_azide	169	fr_nitro_arom	197	NPR2
114	fr_Ar_COO	142	fr_azo	170	fr_nitro_arom_nonortho	198	RadiusOfGyration
115	fr_Ar_N	143	fr_barbitur	171	fr_nitroso	199	InertialShapeFactor
116	fr_Ar_NH	144	fr_benzene	172	fr_oxazole	200	Eccentricity
117	fr_Ar_OH	145	fr_benzodiazepine	173	fr_oxime	201	Asphericity
118	fr_COO	146	fr_bicyclic	174	fr_para_hydroxylation	202	SpherocityIndex
119	fr_COO2	147	fr_diazo	175	fr_phenol		
120	fr_C_O	148	fr_dihydropyridine	176	fr_phenol_noOrthoHbond		
121	fr_C_O_noCOO	149	fr_epoxide	177	fr_phos_acid		
122	fr_C_S	150	fr_ester	178	fr_phos_ester		
123	fr_HOCCN	151	fr_ether	179	fr_piperdine		
124	fr_Imine	152	fr_furan	180	fr_piperzine		
125	fr_NH0	153	fr_guanido	181	fr_priamide		
126	fr_NH1	154	fr_halogen	182	fr_pyridine		
127	fr_NH2	155	fr_hdrzine	183	fr_quatN		
128	fr_N_O	156	fr_hdrzone	184	fr_sulfide		
129	fr_Ndealkylation1	157	fr_imidazole	185	fr_sulfonamd		
130	fr_Ndealkylation2	158	fr_imide	186	fr_sulfone		
131	fr_Nhpyrrole	159	fr_isocyan	187	fr_term_acetylene		
132	fr_SH	160	fr_isothiocyan	188	fr_tetrazole		
133	fr_aldehyde	161	fr_ketone	189	fr_thiazole		
134	fr_alkyl_carbamate	162	fr_ketone_Topliss	190	fr_thiophene		
135	fr_alkyl_halide	163	fr_lactam	191	fr_unbrch_alkane		
136	fr_allylic_oxid	164	fr_lactone	192	fr_urea		
137	fr_amide	165	fr_methoxy	193	PMI1		
138	fr_amidine	166	fr_morpholine	194	PMI2		
139	fr_aniline	167	fr_nitrile	195	PMI3		
140	fr_aryl_methyl	168	fr_nitro	196	NPR1		

Footnote: Definitions available at https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors

8.21 SUMMARY OF ANNOTATIONS FOR LC-MS(/MS) DMA ASSAYS

Table 8-8 Summary of fragmentation coverage for LC-MS(/MS) DMA assays (positive ionisation mode)

Assay name	# features	# features with frag	% features with frag
AP_AMP_1_C30	4875	1504	30.9
AP_AMP_2_C30	2522	896	35.5
AP_AMP_3_C30	2203	812	36.9
AP_AMP_4_C30	2382	754	31.7
AP_C18_1_C30	2998	976	32.6
AP_C18_2_C30	3993	1328	33.3
AP_C18_3_C30	3365	1230	36.6
P_WAX_1_AMD	4498	1395	31.0
P_WAX_2_AMD	1531	353	23.1
P_WAX_3_AMD	1868	343	18.4
P_WAX_3_AMD§	1764	305	17.3
P_WAX_4_AMD	868	138	15.9
P_WAX_4_AMD§	895	109	12.2
P_WCX_1_AMD	4285	1011	23.6
P_WCX_2_AMD	1401	270	19.3
P_WCX_3_AMD	957	152	15.9
P_WCX_4_AMD	1091	185	17.0
P_WCX_4_AMD§	1229	207	16.8
P_WAX_1_PHE	4877	963	19.7
P_WAX_2_PHE	3725	567	15.2
P_WAX_3_PHE	5460	418	7.7
P_WAX_4_PHE	2823	327	11.6
P_WCX_1_PHE	5984	1157	19.3
P_WCX_2_PHE	2486	438	17.6
P_WCX_3_PHE	2966	360	12.1
P_WCX_4_PHE	1278	156	12.2
total	72324	16354	22.6*
median	2504.0	428.0	18.8
mean	2781.7	629.0	21.7
standard deviation	1526.54	445.31	8.85

Footnotes: : * = Total calculated across all assays (not a sum of the column). § = Experimental repeat performed for this assay

Table 8-9 Summary of fragmentation coverage for LC-MS(/MS) DMA assays (negative ionisation mode)

Assay name	# features	# features with frag	% features with frag
AP_AMP_1_C30	1267	425	33.5
AP_AMP_2_C30	314	103	32.8
AP_AMP_3_C30	857	360	42.0
AP_AMP_4_C30	665	242	36.4
AP_C18_1_C30	556	73	13.1
AP_C18_2_C30	1476	439	29.7
AP_C18_3_C30	949	333	35.1
P_WAX_1_AMD	3141	1094	34.8
P_WAX_2_AMD	890	290	32.6
P_WAX_3_AMD	1070	265	24.8
P_WAX_3_AMD§	919	192	20.9
P_WAX_4_AMD	775	170	21.9
P_WAX_4_AMD§	673	152	22.6
P_WCX_1_AMD	2968	922	31.1
P_WCX_2_AMD	689	176	25.5
P_WCX_3_AMD	484	107	22.1
P_WCX_4_AMD	560	155	27.7
P_WCX_4_AMD§	523	108	20.7
P_WAX_1_PHE	6044	866	14.3
P_WAX_2_PHE	3768	698	18.5
P_WAX_3_PHE	5465	677	12.4
P_WAX_4_PHE	2973	386	13.0
P_WCX_1_PHE	4306	1010	23.5
P_WCX_2_PHE	1462	298	20.4
P_WCX_3_PHE	1130	138	12.2
P_WCX_4_PHE	608	111	18.3
total	44532	9790	22.0*
median	934.0	277.5	23.0
mean	1712.8	376.5	24.6
standard deviation	1619.39	306.92	8.40

Footnotes: * = Total calculated across all assays (not a sum of the column). § = Experimental repeat performed for this assay

Table 8-10 Summary of annotations for LC-MS(/MS) DMA assays (positive ionisation mode)

Assay name	# features with annotation			% features with annotation			# unique annotations		
	L2-A	L2-B	L2-C	L2-A	L2-B	L2-C	L2-A	L2-B	L2-C
AP_AMP_1_C30	970	1459	1504	19.9	29.9	30.9	927	1397	1442
AP_AMP_2_C30	475	850	882	18.8	33.7	35.0	468	823	850
AP_AMP_3_C30	486	779	789	22.1	35.4	35.8	449	739	749
AP_AMP_4_C30	378	727	745	15.9	30.5	31.3	377	713	729
AP_C18_1_C30	625	947	993	20.8	31.6	33.1	595	905	946
AP_C18_2_C30	820	1305	1330	20.5	32.7	33.3	806	1204	1229
AP_C18_3_C30	791	1194	1194	23.5	35.5	35.5	766	1115	1115
P_WAX_1_AMD	460	1325	1511	10.2	29.5	33.6	431	704	886
P_WAX_2_AMD	189	349	390	12.3	22.8	25.5	185	245	285
P_WAX_3_AMD	179	339	395	9.6	18.1	21.1	170	225	280
P_WAX_3_AMD§	123	299	381	7.0	17.0	21.6	118	161	242
P_WAX_4_AMD	64	133	149	7.4	15.3	17.2	61	79	95
P_WAX_4_AMD§	32	104	122	3.6	11.6	13.6	31	56	73
P_WCX_1_AMD	450	997	1111	10.5	23.3	25.9	416	940	1053
P_WCX_2_AMD	81	256	311	5.8	18.3	22.2	77	252	304
P_WCX_3_AMD	51	148	174	5.3	15.5	18.2	50	147	173
P_WCX_4_AMD	40	171	180	3.7	15.7	16.5	38	168	177
P_WCX_4_AMD§	51	174	174	4.1	14.2	14.2	49	129	129
P_WAX_1_PHE	304	910	1146	6.2	18.7	23.5	302	436	664
P_WAX_2_PHE	152	531	710	4.1	14.3	19.1	147	286	462
P_WAX_3_PHE	168	398	627	3.1	7.3	11.5	163	237	455
P_WAX_4_PHE	118	306	396	4.2	10.8	14.0	117	208	294
P_WCX_1_PHE	448	1112	1285	7.5	18.6	21.5	439	675	844
P_WCX_2_PHE	156	405	495	6.3	16.3	19.9	155	257	345
P_WCX_3_PHE	167	336	454	5.6	11.3	15.3	164	247	363
P_WCX_4_PHE	67	152	193	5.2	11.9	15.1	66	131	172
total	7845	15706	17641	0.6	14.1	18.8	5404*	9472*	10780*
median	173.5	401.5	561.0	7.2	18.2	21.5	167.0	254.5	409.0
mean	301.7	604.1	678.5	10.1	20.8	23.2	291.0	480.0	552.2
standard deviation	268.16	432.40	453.13	6.74	8.67	7.90	258.00	391.52	392.60

Footnotes: * = Total calculated across all assays (not a sum of the column). § = Experimental repeat performed for this assay

Table 8-11 Summary of annotations for LC-MS(/MS) DMA assays (negative ionisation mode)

Assay name	# features with annotation			% features with annotation			# unique annotations		
	L2-A	L2-B	L2-C	L2-A	L2-B	L2-C	L2-A	L2-B	L2-C
AP_AMP_1_C30	0	395	395	0.0	31.2	31.2	0	392	392
AP_AMP_2_C30	2	75	89	0.6	23.9	28.3	2	75	89
AP_AMP_3_C30	4	341	343	0.5	39.8	40.0	4	338	340
AP_AMP_4_C30	3	221	230	0.5	33.2	34.6	3	219	225
AP_C18_1_C30	1	64	80	0.2	11.5	14.4	1	64	78
AP_C18_2_C30	5	399	412	0.3	27.0	27.9	5	396	409
AP_C18_3_C30	2	299	304	0.2	31.5	32.0	2	299	304
P_WAX_1_AMD	37	478	736	1.2	15.2	23.4	31	451	688
P_WAX_2_AMD	13	145	191	1.5	16.3	21.5	13	140	183
P_WAX_3_AMD	11	91	165	1.0	8.5	15.4	10	86	157
P_WAX_3_AMD§	7	73	157	0.8	7.9	17.1	7	71	150
P_WAX_4_AMD	9	102	128	1.2	13.2	16.5	8	95	120
P_WAX_4_AMD§	10	87	103	1.5	12.9	15.3	9	81	97
P_WCX_1_AMD	35	360	552	1.2	12.1	18.6	31	336	517
P_WCX_2_AMD	10	121	154	1.5	17.6	22.4	10	120	153
P_WCX_3_AMD	5	75	75	1.0	15.5	15.5	5	74	74
P_WCX_4_AMD	2	103	119	0.4	18.4	21.3	2	95	111
P_WCX_4_AMD§	1	80	94	0.2	15.3	18.0	1	79	92
P_WAX_1_PHE	8	487	910	0.1	8.1	15.1	6	475	875
P_WAX_2_PHE	5	529	677	0.1	14.0	18.0	5	524	669
P_WAX_3_PHE	11	464	710	0.2	8.5	13.0	11	460	694
P_WAX_4_PHE	27	259	362	0.9	8.7	12.2	21	251	348
P_WCX_1_PHE	34	636	846	0.8	14.8	19.6	30	617	822
P_WCX_2_PHE	15	196	270	1.0	13.4	18.5	12	192	266
P_WCX_3_PHE	4	101	168	0.4	8.9	14.9	4	100	167
P_WCX_4_PHE	3	82	116	0.5	13.5	19.1	3	77	110
total	264	6263	8386	10.8	21.7	24.4	84*	5139*	6443*
median	6.0	170.5	210.5	0.6	14.4	18.5	5.5	166.0	204.0
mean	10.2	240.9	322.5	0.7	17.0	20.9	9.1	234.9	312.7
standard deviation	10.89	176.41	258.01	0.47	8.76	7.23	9.22	172.59	248.88

Footnotes: * = Total calculated across all assays (not a sum of the column). § = Experimental repeat performed for this assay

8.22 PCA ANALYSIS OF KEGG BRITE COMPOUNDS

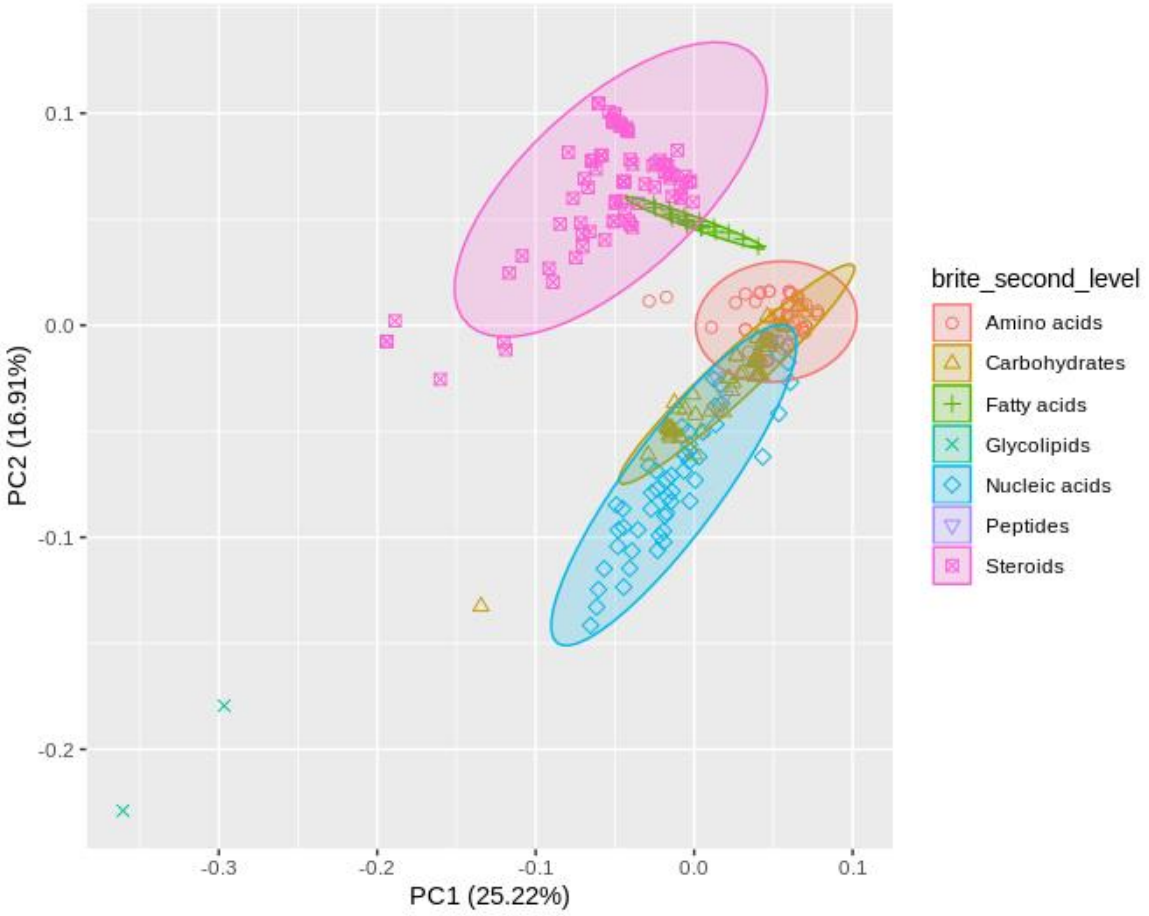


Figure 8-14 PCA scores plot of KEGG Brite compounds

PCA scores plot of KEGG Brite compounds. Comparison of 153 chemical descriptors for 324 compounds.

Chapter 9
References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D., 2018. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update." *Nucleic Acids Research* 46, W537–W544. https://doi.org/10.1093/nar/gky379

Aksenov, A.A., da Silva, R., Knight, R., Lopes, N.P., Dorrestein, P.C., 2017. "Global chemical analysis of biology by mass spectrometry." *Nature Reviews Chemistry* 1, 0054. https://doi.org/10.1038/s41570-017-0054

Aliferis, K.A., Faubert, D., Jabaji, S., 2014. "A metabolic profiling strategy for the dissection of plant defense against fungal pathogens." *PloS one* 9, e111930. https://doi.org/10.1371/journal.pone.0111930

Altman, T., Travers, M., Kothari, A., Caspi, R., Karp, P.D., 2013. "A systematic comparison of the MetaCyc and KEGG pathway databases." *BMC bioinformatics* 14, 112. https://doi.org/10.1186/1471-2105-14-112

Atkinson, M., Gesing, S., Montagnat, J., Taylor, I., 2017. "Scientific workflows: Past, present

and future." *Future Generation Computer Systems* 75, 216–227.

Bäck, T., 1996. Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford university press.

Baker, M., 2016. "1,500 scientists lift the lid on reproducibility." *Nature* 533, 452–454. https://doi.org/10.1038/533452a

Beisken, S., Earll, M., Baxter, C., Portwood, D., Ament, Z., Kende, A., Hodgman, C., Seymour, G., Smith, R., Fraser, P., Seymour, M., Salek, R.M., Steinbeck, C., 2014a. "Metabolic differences in ripening of Solanum lycopersicum 'Ailsa Craig' and three monogenic mutants." *Scientific data* 1, 140029. https://doi.org/10.1038/sdata.2014.29

Beisken, S., Earll, M., Portwood, D., Seymour, M., Steinbeck, C., 2014b. "MassCascade: Visual programming for LC-MS data processing in metabolomics." *Molecular Informatics* 33, 307–310. https://doi.org/10.1002/minf.201400016

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., 2009. "KNIME - the Konstanz information miner." *ACM SIGKDD Explorations Newsletter* 11, 26. https://doi.org/10.1145/1656274.1656280

Bligh, E.G., Dyer, W.J., 1959. "A rapid method of total lipid extraction and purification." *Canadian Journal of Biochemistry and Physiology* 37, 911–917.

Böcker, S., Dührkop, K., 2016. "Fragmentation trees reloaded." *Journal of Cheminformatics* 8, 5. https://doi.org/10.1186/s13321-016-0116-8

Böcker, S., Letzel, M.C., Lipták, Z., Pervukhin, A., 2009. "SIRIUS: Decomposing isotope patterns for metabolite identification." *Bioinformatics* 25, 218–224. https://doi.org/10.1093/bioinformatics/btn603

Böcker, S., Rasche, F., 2008. "Towards de novo identification of metabolites by analyzing

tandem mass spectra." *Bioinformatics* 24, 49–55. https://doi.org/10.1093/bioinformatics/btn270

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M.L., Barrett, M.P., 2006. "*Ab initio* prediction of metabolic networks using Fourier transform mass spectrometry data." *Metabolomics : Official journal of the Metabolomic Society* 2, 155–164. https://doi.org/10.1007/s11306-006-0029-z

Brown, M., Wedge, D.C., Goodacre, R., Kell, D.B., Baker, P.N., Kenny, L.C., Mamas, M.A., Neyses, L., Dunn, W.B., 2011. "Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets." *Bioinformatics* 27, 1108–1112. https://doi.org/10.1093/bioinformatics/btr079

Bundy, J.G., Davey, M.P., Viant, M.R., 2008. "Environmental metabolomics: a critical review and future perspectives." *Metabolomics* 5, 3–21. https://doi.org/10.1007/s11306-008-0152-0

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang, P., Karp, P.D., 2007. "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." *Nucleic Acids Research* 36, D623–D631. https://doi.org/10.1093/nar/gkm900

Cazzaniga, P., Damiani, C., Besozzi, D., Colombo, R., Nobile, M., Gaglio, D., Pescini, D., Molinari, S., Mauri, G., Alberghina, L., Vanoni, M., 2014. "Computational Strategies for a System-Level Understanding of Metabolism." *Metabolites* 4, 1034–1087. https://doi.org/10.3390/metabo4041034

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N.,

Frewen, B., Baker, T.A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., Mallick, P., 2012. "A cross-platform toolkit for mass spectrometry and proteomics." *Nature Biotechnology* 30, 918–920. https://doi.org/10.1038/nbt.2377

Chen, L., Li, J., Guo, T., Ghosh, S., Koh, S.K., Tian, D., Zhang, L., Jia, D., Beuerman, R.W., Aebersold, R., Chan, E.C.Y., Zhou, L., 2015. "Global metabonomic and proteomic analysis of human conjunctival epithelial cells (IOBA-NHC) in response to hyperosmotic stress." *Journal of Proteome Research* 14, 3982–3995. https://doi.org/10.1021/acs.jproteome.5b00443

Chitraju, C., Trotzmuller, M., Hartler, J., Wolinski, H., Thallinger, G.G., Lass, a., Zechner, R., Zimmermann, R., Kofeler, H.C., Spener, F., 2012. "Lipidomic analysis of lipid droplets from murine hepatocytes reveals distinct signatures for nutritional stress." *The Journal of Lipid Research* 53, 2141–2152. https://doi.org/10.1194/jlr.M028902

Cho, K., Evans, B.S., Wood, B.M., Kumar, R., Erb, T.J., Warlick, B.P., Gerlt, J. a., Sweedler, J. V., 2014. "Integration of untargeted metabolomics with transcriptomics reveals active metabolic pathways." *Metabolomics* 503–517. https://doi.org/10.1007/s11306-014-0713-3

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Bras, Y. Le, Lemoine, F., Mareuil, F., Ménager, H., Pradal, C., Blanchet, C., 2017. "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities." *Future Generation Computer Systems* 75, 284–298. https://doi.org/10.1016/j.future.2017.01.012

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Cáceres, C.E., Carmel, L., Casola, C., Choi, J.-H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., Fröhlich, T., Geiler-Samerotte, K. a, Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E. V, Kültz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J.R., Muller, J., Pangilinan, J., Patwardhan, R.P., Pitluck, S., Pritham, E.J., Rechtsteiner, A., Rho, M., Rogozin, I.B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y.I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J.R., Andrews, J., Crease, T.J., Tang, H., Lucas, S.M., Robertson, H.M., Bork, P., Koonin, E. V, Zdobnov, E.M., Grigoriev, I. V, Lynch, M., Boore, J.L., 2011. "The ecoresponsive genome of Daphnia pulex." *Science (New York, N.Y.)* 331, 555–61. https://doi.org/10.1126/science.1197761

Conway, J.R., Lex, A., Gehlenborg, N., 2017. "UpSetR: An R package for the visualization of intersecting sets and their properties." *Bioinformatics* 33, 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Correia, D., Doppelt-Azeroual, O., Denis, J.-B., Vandenbogaert, M., Caro, V., 2015. "MetaGenSense : A web application for analysis and visualization of high throughput sequencing metagenomic data." *F1000Research* 1–11. https://doi.org/10.12688/f1000research.6139.1

Davidson, R.L., Weber, R.J.M., Liu, H., Sharma-Oates, A., Viant, M.R., 2016. "Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data." *GigaScience* 5, 10. https://doi.org/10.1186/s13742-016-0115-8

Davis, R. a, Lee, T.C.M., Rodriguez-Yam, G. a, 2006. "Structural break estimation for

nonstationary time series models." *Journal of the American Statistical Association* 101, 223–239. https://doi.org/10.1198/016214505000000745

Devlin, H., 2018. "Attempt to replicate major social scientific findings of past decade fails." *The Guardian*.

Draper, J., Enot, D.P., Parker, D., Beckmann, M., Snowdon, S., Lin, W., Zubair, H., 2009. "Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'." *BMC bioinformatics* 10, 227. https://doi.org/10.1186/1471-2105-10-227

Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., 2015. "Searching molecular structure databases with tandem mass spectra using CSI:FingerID." *Proceedings of the National Academy of Sciences* 112, 12580–12585. https://doi.org/10.1073/pnas.1509788112

Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J.D., Halsall, A., Haselden, J.N., Nicholls, A.W., Wilson, I.D., Kell, D.B., Goodacre, R., 2011. "Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry." *Nature protocols* 6, 1060–1083. https://doi.org/10.1038/nprot.2011.335

Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., Viant, M.R., 2012. "Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics." *Metabolomics* 9, 44–66. https://doi.org/10.1007/s11306-012-0434-4

Ebert, D., 2005. Introduction to the ecology, epidemiology, and evolution of parasitism in Daphnia. National Center for Biotechnology Information (US).

Edison, A.S., Hall, R.D., Junot, C., Karp, P.D., Kurland, I.J., Mistrik, R., Reed, L.K., Saito, K., Salek, R.M., Steinbeck, C., Sumner, L.W., Viant, M.R., 2016. "The time is right to focus

on model organism metabolomes." *Metabolites* 6. https://doi.org/10.3390/metabo6010008

Feilden, T., 2017. "Most scientists can't replicate studies by their peers'." *BBC News*.

Fenaille, F., Saint-Hilaire, P.B., Rousseau, K., Junot, C., 2017. "Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: where do we stand?" *Journal of Chromatography A* 29–30. https://doi.org/10.1016/j.chroma.2017.10.043

Fernie, a. R., Stitt, M., 2012. "On the discordance of metabolomics with proteomics and transcriptomics: Coping with increasing complexity in logic, chemistry, and network interactions." *Plant Physiology* 158, 1139–1145. https://doi.org/10.1104/pp.112.193235

Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau, B., Morrison, N., Sumner, L.W., Goodacre, R., Hardy, N.W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., van Ommen, B., Lindon, J.C., Sansone, S.-A., 2007. "The metabolomics standards initiative (MSI)." *Metabolomics* 3, 175–178. https://doi.org/10.1007/s11306-007-0070-6

Fiore, C.L., Longnecker, K., Kido Soule, M.C., Kujawinski, E.B., 2015. "Release of ecologically relevant metabolites by the cyanobacterium, *Synechococcus elongatus* CCMP 1631." *Environmental Microbiology* 17, n/a--n/a. https://doi.org/10.1111/1462-2920.12899

Frank, A.M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S.P., Smith, R.D., Pevzner, P.A., 2009. "Clustering Millions of Tandem Mass Spectra." *Proteome* 7, 113–122. https://doi.org/10.1021/pr070361e.Clustering

Freedman, L.P., Cockburn, I.M., Simcoe, T.S., 2015. "The economics of reproducibility in preclinical research." *PLoS Biology* 13, 1–9. https://doi.org/10.1371/journal.pbio.1002165

Garijo, D., Kinnings, S., Xie, L.L., Xie, L.L., Zhang, Y., Bourne, P.E., Gil, Y., 2013. "Quantifying reproducibility in computational biology: The case of the tuberculosis drugome." *PLoS ONE* 8. https://doi.org/10.1371/journal.pone.0080278

Gehmlich, K., Dodd, M.S., Allwood, J.W., Kelly, M., Bellahcene, M., Lad, H. V, Stockenhuber, A., Hooper, C., Ashrafian, H., Redwood, C.S., Carrier De, L., Dunn, W.B., 2015. "Changes in the cardiac metabolome caused by perhexiline treatment in a mouse model of hypertrophic cardiomyopathy." *Mol. BioSyst. Mol. BioSyst* 564, 564–573. https://doi.org/10.1039/c4mb00594e

Geromanos, S.J., Vissers, J.P.C., Silva, J.C., Dorschel, C. a, Li, G.-Z., Gorenstein, M. V, Bateman, R.H., Langridge, J.I., 2009. "The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS." *Proteomics* 9, 1683–1695. https://doi.org/10.1002/pmic.200800562

Giacomoni, F., Le Corguille, G., Monsoor, M., Landi, M., Pericard, P., Petera, M., Duperier, C., Tremblay-Franco, M., Martin, J.-F., Jacob, D., Goulitquer, S., Thevenot, E.A., Caron, C., 2015. "Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics." *Bioinformatics* 31, 1493–1495. https://doi.org/10.1093/bioinformatics/btu813

Gillet, L.C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., Aebersold, R., 2012. "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis." *Molecular & Cellular Proteomics* 11, O111.016717-O111.016717. https://doi.org/10.1074/mcp.O111.016717

Gipson, G.T., Tatsuoka, K.S., Sokhansanj, B. a., Ball, R.J., Connor, S.C., 2007. "Assignment of MS-based metabolomic datasets via compound interaction pair mapping."

Metabolomics 4, 94–103. https://doi.org/10.1007/s11306-007-0096-9

Gloaguen, Y., Morton, F., Daly, R., Gurden, R., Rogers, S., Wandy, J., Wilson, D., Barrett, M., Burgess, K., 2017. "PiMP my metabolome: An integrated, web-based tool for LC-MS metabolomics data." *Bioinformatics* 33, 4007–4009. https://doi.org/10.1093/bioinformatics/btx499

Goecks, J., Nekrutenko, A., Taylor, J., 2010. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome biology* 11, R86. https://doi.org/10.1186/gb-2010-11-8-r86

González-Beltrán, A., Neumann, S., Maguire, E., Sansone, S.-A., Rocca-Serra, P., 2014. "The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again." *BMC Bioinformatics* 15, S11. https://doi.org/10.1186/1471-2105-15-S1-S11

Gowda, H., Ivanisevic, J., Johnson, C.H., Kurczy, M.E., Benton, H.P., Rinehart, D., Nguyen, T., Ray, J., Kuehl, J., Arevalo, B., Westenskow, P.D., Wang, J., Arkin, A.P., Deutschbauer, A.M., Patti, G.J., Siuzdak, G., 2014. "Interactive XCMS online: Simplifying advanced metabolomic data processing and subsequent statistical analyses." *Analytical Chemistry* 86, 6931–6939. https://doi.org/10.1021/ac500734c

Griffiths, W.J., 2007. Metabolomics, metabonomics and metabolite profiling. Royal society of chemistry.

Gustafsson, O.J.R., Winderbaum, L.J., Condina, M.R., Boughton, B.A., Hamilton, B.R., Undheim, E.A.B., Becker, M., Hoffmann, P., 2018. "Balancing sufficiency and impact in reporting standards for mass spectrometry imaging experiments." *GigaScience* 7. https://doi.org/10.1093/gigascience/giy102

Haug, K., Salek, R.M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone,

S.-A., Griffin, J.L., Steinbeck, C., 2013. “MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data.” *Nucleic acids research* 41, D781--6. https://doi.org/10.1093/nar/gks1004

He, L., Diedrich, J.K., Chu, Y.-Y., Yates, J.R., 2015. “Extracting accurate precursor information for tandem mass spectra by RawConverter.” *Analytical Chemistry* acs.analchem.5b02721. https://doi.org/10.1021/acs.analchem.5b02721

Heinonen, M., Rantanen, A., Mielikäinen, T., Kokkonen, J., Kiuru, J., Ketola, R.A., Rousu, J., 2008. “FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data.” *Rapid Communications in Mass Spectrometry* 22, 3043–3052. https://doi.org/10.1002/rcm.3701

Heller, S.R., McNaught, A.D., 2009. “The IUPAC International Chemical Identifier (InChI).” *Chemistry International* 31, 7–9. https://doi.org/10.1515/ci.2009.31.1.7

Hoopmann, M.R., Finney, G.L., MacCoss, M.J., 2007. “High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry.” *Analytical Chemistry* 79, 5620–5632. https://doi.org/10.1021/ac0700833

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., Nishioka, T., others, Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito,

K., Nishioka, T., 2010. “MassBank: A public repository for sharing mass spectral data for life sciences.” *Journal of Mass Spectrometry* 45, 703–714. https://doi.org/10.1002/jms.1777

Horner, J. a, Thakur, R. a, Scientific, T.F., Jose, S., n.d. “PIF : Precursor Ion Fingerprinting – Searching for a Structurally Diagnostic Fragment Using Combined Targeted and Data Dependent MSn.” *Thermo Scientific Application Note* 441, 1–6.

Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N.G., Old, W.M., 2010. “Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies.” *Journal of proteome research* 9, 4152–60. https://doi.org/10.1021/pr1003856

Hufsky, F., Böcker, S., 2012. “Comparing fragmentation trees from electron impact mass spectra with annotated fragmentation pathways,” in: OASIcs-OpenAccess Series in Informatics.

Hummel, J., Selbig, J., Walther, D., Kopka, J., 2007. “The Golm Metabolome Database: a database for GC-MS based metabolite profiling,” in: Metabolomics. Springer, pp. 75–95. https://doi.org/10.1007/4735_2007_0229

Hunter, A., Dayalan, S., De Souza, D., Power, B., Lorrimar, R., Szabo, T., Nguyen, T., O’Callaghan, S., Hack, J., Pyke, J., Nahid, A., Barrero, R., Roessner, U., Likic, V., Tull, D., Bacic, A., McConville, M., Bellgard, M., 2017. “MASTR-MS: a web-based collaborative laboratory information management system (LIMS) for metabolomics.” *Metabolomics* 13, 1–9. https://doi.org/10.1007/s11306-016-1142-2

Johnson, W.M., Kido Soule, M.C., Kujawinski, E.B., 2016. “Evidence for quorum sensing and differential metabolite production by a marine bacterium in response to DMSP.” *ISME Journal* 10, 2304–2316. https://doi.org/10.1038/ismej.2016.6

Jones, E., Oliphant, T., Peterson, P., others, 2001. "SciPy: Open source scientific tools for Python."

Kale, N.S., Haug, K., Conesa, P., Jayseelan, K., Moreno, P., Rocca-Serra, P., Nainala, V.C., Spicer, R.A., Williams, M., Li, X., Salek, R.M., Griffin, J.L., Steinbeck, C., 2016. "MetaboLights: An open-access database repository for metabolomics data." *Curr. Protoc. Bioinformatics* 53, 14.13.1-14.13.18. https://doi.org/10.1002/0471250953.bi1413s53

Kanehisa, M., 2005. "From genomics to chemical genomics: new developments in KEGG." *Nucleic Acids Research* 34, D354–D357. https://doi.org/10.1093/nar/gkj102

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2014. "Data, information, knowledge and principle: back to metabolism in KEGG." *Nucleic acids research* 42, D199-205. https://doi.org/10.1093/nar/gkt1076

Kasper, P.T., Rojas-Chertó, M., Mistrik, R., Reijmers, T., Hankemeier, T., Vreeken, R.J., 2012. "Fragmentation trees for the structural characterisation of metabolites." *Rapid communications in mass spectrometry : RCM* 26, 2275–86. https://doi.org/10.1002/rcm.6340

Kiefer, P., Schmitt, U., Vorholt, J.A., 2013. "eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows." *Bioinformatics* 29, 963–964. https://doi.org/10.1093/bioinformatics/btt080

Kind, T., Liu, K.-H., Lee, D.Y., DeFelice, B., Meissen, J.K., Fiehn, O., 2013. "LipidBlast in silico tandem mass spectrometry database for lipid identification." *Nature methods* 10, 755–758. https://doi.org/10.1038/nmeth.2551

Kind, T., Tsugawa, H., Cajka, T., Ma, Y., Lai, Z., Mehta, S.S., Wohlgemuth, G., Barupal, D.K., Showalter, M.R., Arita, M., others, Fiehn, O., 2017. "Identification of small molecules using accurate mass MS/MS search." *Mass spectrometry reviews*.

https://doi.org/10.1002/mas.21535

Kirwan, J. a, Weber, R.J.M., Broadhurst, D.I., Viant, M.R., 2014. "Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control." *Scientific Data* 1, 1–13. https://doi.org/10.1038/sdata.2014.12

Kleinberg, J., Tardos, É., 2005. Algorithm Design. Pearson Education.

Koo, I., Zhang, X., Kim, S., 2011. "Wavelet- and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry." *Analytical Chemistry* 83, 5631–5638. https://doi.org/10.1021/ac200740w

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T.R.R., Neumann, S., Bo, C., Larson, T.R.R., Neumann, S., 2012. "CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets." *Analytical Chemistry* 84, 283–289. https://doi.org/10.1021/ac202450g

Kuile, B.H., Westerhoff, H. V., 2001. "Transcriptome meets metabolome: Hierarchical and metabolic regulation of the glycolytic pathway." *FEBS Letters* 500, 169–171. https://doi.org/10.1016/S0014-5793(01)02613-8

Kujawinski, E.B., Longnecker, K., Alexander, H., Dyhrman, S.T., Fiore, C.L., Haley, S.T., Johnson, W.M., 2017. "Phosphorus availability regulates intracellular nucleotides in marine eukaryotic phytoplankton." *Limnology and Oceanography Letters* 2, 119–129. https://doi.org/10.1002/lol2.10043

Lampert, W., 2006. "Daphnia: Model herbivore, predator and prey." *Polish Journal of Ecology* 54, 607–620.

Landrum, G., others, 2013. "RDKit: cheminformatics and machine learning software." *RDKIT.ORG*.

Laponogov, I., Sadawi, N., Galea, D., Mirnezami, R., Veselkov, K.A., 2018. "ChemDistiller: An engine for metabolite annotation in mass spectrometry." *Bioinformatics* 34, 2096–2102. https://doi.org/10.1093/bioinformatics/bty080

Larralde, M., Lawson, T.N., Weber, R.J.M., Moreno, P., Haug, K., Rocca-serra, P., Viant, M.R., Salek, R.M., 2017. "mzML2ISA & nmrML2ISA : Generating enriched ISA-Tab metadata files from metabolomics XML data." *Bioinformatics* 4–6. https://doi.org/10.1093/bioinformatics/btx169

Lawson, T.N., Weber, R.J.M., Jones, M.R., Chetwynd, A.J., Rodriguez Blanco, G.A., Di Guida, R., Viant, M.R., Dunn, W.B., 2017. "msPurity: Automated evaluation of precursor ion purity for mass spectrometry based fragmentation in metabolomics." *Analytical Chemistry* acs.analchem.6b04358. https://doi.org/10.1021/acs.analchem.6b04358

Lekschas, F., Gehlenborg, N., 2018. "SATORI : A system for ontology-guided visual exploration of biomedical data repositories." *Bioinformatics* 34, 1200–1207. https://doi.org/10.1093/bioinformatics/btx739

Lemmon, E.W., McLinden, M.O., Friend, D.G., Linstrom, P.J., Mallard, W.G., 2011. "NIST chemistry WebBook, NIST standard reference database number 69." *National Institute of Standards and Technology, Gaithersburg*.

Li, G.Z., Vissers, J.P.C., Silva, J.C., Golick, D., Gorenstein, M. V., Geromanos, S.J., 2009. "Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures." *Proteomics* 9, 1696–1719. https://doi.org/10.1002/pmic.200800564

Li, S., Lund, R., 2012. "Multiple changepoint detection via genetic algorithms." *Journal of Climate* 25, 674–686. https://doi.org/10.1175/2011JCLI4055.1

Liggi, S., Hinz, C., Hall, Z., Santoru, M.L., Poddighe, S., Fjeldsted, J., Atzori, L., Griffin, J.L.,

2018. “KniMet: a pipeline for the processing of chromatography–mass spectrometry metabolomics data.” *Metabolomics* 14, 1–4. https://doi.org/10.1007/s11306-018-1349-5

Lin, L., Yu, Q., Yan, X., Hang, W., Zheng, J., Xing, J., Huang, B., 2010. “Direct infusion mass spectrometry or liquid chromatography mass spectrometry for human metabonomics? A serum metabonomic study of kidney cancer.” *The Analyst* 135, 2970. https://doi.org/10.1039/c0an00265h

Longnecker, K., Kido Soule, M.C., Kujawinski, E.B., 2015. “Dissolved organic matter produced by Thalassiosira pseudonana.” *Marine Chemistry* 168, 114–123. https://doi.org/10.1016/j.marchem.2014.11.003

Luethy, R., Kessner, D.E., Katz, J.E., MacLean, B., Grothe, R., Kani, K., Faça, V., Pitteri, S., Hanash, S., Agus, D.B., Mallick, P., 2008. “Precursor-ion mass re-estimation improves peptide identification on hybrid instruments.” *Journal of proteome research* 7, 4031–4039. https://doi.org/10.1021/pr800307m

Madhavan, S., Gauba, R., Robert, C., Gusev, Y., 2014. “Integrative analysis workflow for untargeted metabolomics in translational research.” *Metabolomics : Open Access* 4, 1–11. https://doi.org/10.4172/2153-0769.1000130

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpp, A., Neumann, S., Pizarro, A.D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., Deutsch, E.W., 2011. “mzML--a community standard for mass spectrometry data.” *Molecular & cellular proteomics : MCP* 10, R110.000133. https://doi.org/10.1074/mcp.R110.000133

Matthias Grote, J.C., 2014. “A new multi-cell exposure system for continuous tracking of Daphnia behavior for toxicity assessments.” *Journal of Environmental & Analytical Toxicology* 05. https://doi.org/10.4172/2161-0525.1000246

Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D., Jones, A.R., Binz, P.-A., Deutsch, E.W., Chambers, M., Kallhardt, M., Levander, F., Shofstahl, J., Orchard, S., Antonio Vizcaino, J., Hermjakob, H., Stephan, C., Meyer, H.E., Eisenacher, M., 2013. "The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary." *Database* 2013, bat009-bat009. https://doi.org/10.1093/database/bat009

McLafferty, F.W., Hertel, R.H., Villwock, R.D., 1974. "Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures." *Organic Mass Spectrometry* 9, 690–702.

McLafferty, F.W., Zhang, M.-Y., Stauffer, D.B., Loh, S.Y., 1998. "Comparison of algorithms and databases for matching unknown mass spectra." *Journal of the American Society for Mass Spectrometry* 9, 92–95.

Medina-Franco, J., Martinez-Mayorga, K., Giulianotti, M., Houghten, R., Pinilla, C., 2008. "Visualization of the chemical space in drug discovery." *Current Computer Aided-Drug Design* 4, 322–333. https://doi.org/10.2174/157340908786786010

Michalski, A., Cox, J., Mann, M., 2011. "More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS." *Journal of Proteome Research* 10, 1785–1793. https://doi.org/10.1021/pr101060v

Misra, B.B., 2018. "New tools and resources in metabolomics: 2016-2017." *Electrophoresis* 39, 909–923. https://doi.org/10.1002/elps.201700441

Mylonas, R., Mauron, Y., Masselot, A., Binz, P.-A., Budin, N., Fathi, M., Viette, V., Hochstrasser, D.F., Lisacek, F., 2009. "X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry." *Analytical chemistry* 81, 7604–10. https://doi.org/10.1021/ac900954d

Neumann, S., Thum, A., Böttcher, C., 2013. "Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data." *Metabolomics* 84–91. https://doi.org/10.1007/s11306-012-0401-0

Nijenhuis, A., Wilf, H.S., 2014. Combinatorial algorithms: for computers and calculators. Elsevier.

Nikolskiy, I., Mahieu, N.G., Chen, Y.J., Tautenhahn, R., Patti, G.J., 2013. "An untargeted metabolomic workflow to improve structural characterization of metabolites." *Analytical Chemistry* 85, 7713–7719. https://doi.org/10.1021/ac400751j

O'Callaghan, S., De Souza, D.P., Isaac, A., Wang, Q., Hodkinson, L., Olshansky, M., Erwin, T., Appelbe, B., Tull, D.L., Roessner, U., others, 2012. "PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools." *BMC bioinformatics* 13, 115.

Oliver, S.G., Winson, M.K., Kell, D.B., Baganz, F., 1998. "Systematic functional analysis of the yeast genome." *Trends in biotechnology* 16, 373–378.

Oltvai, Z.N., Barabási, A.-L., 2002. "Life's complexity pyramid." *Science (New York, N.Y.)* 298, 763–764. https://doi.org/10.1126/science.1078563

Orchard, S., Taylor, C.F., Hermjakob, H., Weimin-Zhu, Julian, R.K., Apweiler, R., 2004. "Advances in the development of common interchange standards for proteomic data." *Proteomics* 4, 2363–2365. https://doi.org/10.1002/pmic.200400884

Owen, J.R., Nabney, I.T., Medina-Franco, J.L., López-Vallejo, F., 2011. "Visualization of molecular fingerprints." *Journal of Chemical Information and Modeling* 51, 1552–1563. https://doi.org/10.1021/ci1004042

Pedrioli, P.G.A., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B.,

Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W., Aebersold, R., 2004. "A common open representation of mass spectrometry data and its application to proteomics research." *Nature Biotechnology* 22, 1459–1466. https://doi.org/10.1038/nbt1031

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999. "Probability-based protein identification by searching sequence databases using mass spectrometry data." *Electrophoresis* 20, 3551–3567. https://doi.org/10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2

Petras, D., Koester, I., Da Silva, R., Stephens, B.M., Haas, A.F., Nelson, C.E., Kelly, L.W., Aluwihare, L.I., Dorrestein, P.C., 2017. "High-Resolution Liquid Chromatography Tandem Mass Spectrometry Enables Large Scale Molecular Characterization of Dissolved Organic Matter." *Frontiers in Marine Science* 4, 405. https://doi.org/10.3389/fmars.2017.00405

Platt, J., others, 1999. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." *Advances in large margin classifiers* 10, 61–74.

Poynton, H.C., Taylor, N.S., Hicks, J., Colson, K., Chan, S., Clark, C., Scanlan, L., Loguinov, A. V., Vulpe, C., Viant, M.R., 2011. "Metabolomics of microliter hemolymph samples enables an improved understanding of the combined metabolic and transcriptional responses of Daphnia magna to cadmium." *Environmental Science and Technology* 45, 3710–3717. https://doi.org/10.1021/es1037222

Rasche, F., Scheubert, K., Hufsky, F., Zichner, T., Kai, M., Svatoš, A., Böcker, S., 2012. "Identifying the unknowns by aligning fragmentation trees." *Analytical Chemistry* 84, 3417–3426. https://doi.org/10.1021/ac300304u

Rasche, F., Svatoš, A., Maddula, R.K., Böttcher, C., Böcker, S., 2011. "Computing Fragmentation Trees from Tandem Mass Spectrometry Data." *Analytical Chemistry* 83, 1243–1251. https://doi.org/10.1021/ac101825k

Rauf, I., Rasche, F., Nicolas, F., Böcker, S., 2012. "Finding Maximum Colorful Subtrees in Practice," in: Journal of Computational Biology. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, pp. 213–223. https://doi.org/10.1007/978-3-642-29627-7_22

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., Sansone, S.-A., 2010. "ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level." *Bioinformatics* 26, 2354–2356. https://doi.org/10.1093/bioinformatics/btq415

Rocca-Serra, P., Salek, R.M., Arita, M., Correa, E., Dayalan, S., Gonzalez-Beltran, A., Ebbels, T., Goodacre, R., Hastings, J., Haug, K., Koulman, A., Nikolski, M., Oresic, M., Sansone, S.-A., Schober, D., Smith, J., Steinbeck, C., Viant, M.R., Neumann, S., 2016. "Data standards can boost metabolomics research, and if there is a will, there is a way." *Metabolomics* 12, 14. https://doi.org/10.1007/s11306-015-0879-3

Rogers, S., Scheltema, R. a, Girolami, M., Breitling, R., 2009. "Probabilistic assignment of formulas to mass peaks in metabolomics experiments." *Bioinformatics* 25, 512–518. https://doi.org/10.1093/bioinformatics/btn642

Rojas-Cherto, M., Peironcely, J.E., Kasper, P.T., van der Hooft, J.J.J., de Vos, R.C.H., Vreeken, R., Hankemeier, T., Reijmers, T., 2012. "Metabolite identification using automated comparison of high-resolution multistage nass spectral trees." *Analytical Chemistry* 84, 5524–5534. https://doi.org/10.1021/ac2034216

Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L., Aebersold, R., 2014. "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data." *Nature biotechnology* 32, 219–223. https://doi.org/10.1038/nbt.2841

Rubert, J., Zachariasova, M., Hajslova, J., 2015. "Advances in high-resolution mass spectrometry based on metabolomics studies for food – a review." *Food Additives & Contaminants: Part A* 32, 1685–1708. https://doi.org/10.1080/19440049.2015.1084539

Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S., 2016. "MetFrag relaunched: Incorporating strategies beyond in silico fragmentation." *Journal of Cheminformatics* 8, 1–16. https://doi.org/10.1186/s13321-016-0115-9

Salek, R.M., Neumann, S., Schober, D., Hummel, J., Billiau, K., Kopka, J., Correa, E., Reijmers, T., Rosato, A., Tenori, L., Turano, P., Marin, S., Deborde, C., Jacob, D., Rolin, D., Dartigues, B., Conesa, P., Haug, K., Rocca-Serra, P., O'Hagan, S., Hao, J., van Vliet, M., Sysi-Aho, M., Ludwig, C., Bouwman, J., Cascante, M., Ebbels, T., Griffin, J.L., Moing, A., Nikolski, M., Oresic, M., Sansone, S.-A., Viant, M.R., Goodacre, R., Günther, U.L., Hankemeier, T., Luchinat, C., Walther, D., Steinbeck, C., 2015. "COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access." *Metabolomics* 11, 1587–1597. https://doi.org/10.1007/s11306-015-0810-y

Sansone, S.-A., Fan, T., Goodacre, R., Griffin, J.L., Hardy, N.W., Kaddurah-Daouk, R., Kristal, B.S., Lindon, J., Mendes, P., Morrison, N., Nikolau, B., Robertson, D., Sumner, L.W., Taylor, C., van der Werf, M., van Ommen, B., Fiehn, O., 2007. "The Metabolomics Standards Initiative." *Nature Biotechnology* 25, 846–848. https://doi.org/10.1038/nbt0807-846b

Sansone, S.-A., Rocca-Serra, P., Brandizi, M., Brazma, A., Field, D., Fostel, J., Garrow, A.G.,

Gilbert, J., Goodsaid, F., Hardy, N., Jones, P., Lister, A., Miller, M., Morrison, N., Rayner, T., Sklyar, N., Taylor, C., Tong, W., Warner, G., Wiemann, S., 2008. "The First RSBI (ISA-TAB) Workshop: 'Can a Simple Format Work for Complex Studies?'" *OMICS: A Journal of Integrative Biology* 12, 143–149. https://doi.org/10.1089/omi.2008.0019

Scheubert, K., Hufsky, F., Böcker, S., 2014. "Multiple Mass Spectrometry Fragmentation Trees Revisited: Boosting Performance and Quality," in: Algorithms in Bioinformatics. Springer, pp. 217–231. https://doi.org/10.1007/978-3-662-44753-6_17

Scheubert, K., Hufsky, F., Böcker, S., 2013. "Computational mass spectrometry for small molecules." *Journal of Cheminformatics* 5, 12. https://doi.org/10.1186/1758-2946-5-12

Schober, D., Jacob, D., Wilson, M., Cruz, J.A., Marcu, A., Grant, J.R., Moing, A., Deborde, C., de Figueiredo, L.F., Haug, K., Rocca-Serra, P., Easton, J., Ebbels, T.M.D., Hao, J., Ludwig, C., Günther, U.L., Rosato, A., Klein, M.S., Lewis, I.A., Luchinat, C., Jones, A.R., Grauslys, A., Larralde, M., Yokochi, M., Kobayashi, N., Porzel, A., Griffin, J.L., Viant, M.R., Wishart, D.S., Steinbeck, C., Salek, R.M., Neumann, S., 2018. "nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data." *Analytical Chemistry* 90, 649–656. https://doi.org/10.1021/acs.analchem.7b02795

Schramm, T., Hester, A., Klinkert, I., Both, J.-P., Heeren, R.M.A., Brunelle, A., Laprévote, O., Desbenoit, N., Robbe, M.-F., Stoeckli, M., Spengler, B., Römpp, A., 2012. "imzML — A common data format for the flexible exchange and processing of mass spectrometry imaging data." *Journal of Proteomics* 75, 5106–5110. https://doi.org/10.1016/j.jprot.2012.07.026

Shawe-Taylor, J., Cristianini, N., others, 2004. Kernel methods for pattern analysis. Cambridge university press.

Shen, H., Dührkop, K., Böcker, S., Rousu, J., 2014. "Metabolite identification through multiple kernel learning on fragmentation trees." *Bioinformatics* 30, i157–i164. https://doi.org/10.1093/bioinformatics/btu275

Silva, R.R., Jourdan, F., Salvanha, D.M., Letisse, F., Jamin, E.L., Guidetti-Gonzalez, S., Labate, C.A., Vêncio, R.Z.N.N., 2014. "ProbMetab: An R package for Bayesian probabilistic annotation of LC-MS-based metabolomics." *Bioinformatics* 30, 1336–1337. https://doi.org/10.1093/bioinformatics/btu019

Smith, C.A., Maille, G.O., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G., 2005. "METLIN: A Metabolite Mass Spectral Database." *Therapeutic Drug Monitoring* 27, 747–751. https://doi.org/10.1097/01.ftd.0000179845.53213.39

Smith, C.A., Want, E.J., Maille, G.O., Abagyan, R., Siuzdak, G., 2006. "XCMS : Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment , Matching , and Identification." *ACS Publications* 78, 779–787. https://doi.org/10.1021/ac051437y

Southam, A.D., Weber, R.J.M., Engel, J., Jones, M.R., Viant, M.R., 2016. "A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics." *Nature protocols* 12, 310–328. https://doi.org/10.1038/nprot.2016.156

Spicer, R.A., Salek, R., Steinbeck, C., 2017. "Compliance with minimum information guidelines in public metabolomics repositories." *Scientific data* 4, 170137.

Stanstrup, J., Neumann, S., Vrhovšek, U., 2015. "PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems." *Analytical Chemistry* 87, 9421–9428. https://doi.org/10.1021/acs.analchem.5b02287

Stein, S., 2012. “Mass spectral reference libraries: An ever-expanding resource for chemical identification.” *Analytical chemistry* 84, 7274–7282. https://doi.org/10.1021/ac301205z

Stein, S.E., Scott, D.R., 1994. “Optimization and testing of mass spectral library search algorithms for compound identification.” *Journal of the American Society for Mass Spectrometry* 5, 859–866.

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K.S., Sumner, S., Subramaniam, S., 2015. “Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools.” *Nucleic Acids Research* gkv1042. https://doi.org/10.1093/nar/gkv1042

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.-M., Fiehn, O., Goodacre, R., Griffin, J.L., Hankemeier, T., Hardy, N., Harnly, J., 2007. “Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Inititative (MSI).” *Metabolomics* 3, 211–221. https://doi.org/10.1007/s11306-007-0082-2.Proposed

Tautenhahn, R., Bottcher, C., Neumann, S., 2008. “Highly sensitive feature detection for high resolution LC/MS.” *BMC Bioinformatics* 9, 1–16. https://doi.org/10.1186/1471-2105-9-504

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., Arita, M., 2015. “MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis.” *Nature Methods* 12, 523–526. https://doi.org/10.1038/nmeth.3393

Tyanova, S., Temu, T., Carlson, A., Sinitcyn, P., Mann, M., Cox, J., 2015. “Visualization of LC-MS/MS proteomics data in MaxQuant.” *Proteomics* 15, 1453–1456.

https://doi.org/10.1002/pmic.201400449

Verhoeven, A., Giera, M., Mayboroda, O.A., 2018. "KIMBLE: A versatile visual NMR metabolomics workbench in KNIME." *Analytica Chimica Acta* 1–11. https://doi.org/10.1016/j.aca.2018.07.070

Viant, M.R., Kurland, I.J., Jones, M.R., Dunn, W.B., 2017. "How close are we to complete annotation of metabolomes?" *Current Opinion in Chemical Biology* 36, 64–69. https://doi.org/10.1016/j.cbpa.2017.01.001

Wang, W., Dong, M., Song, C., Cai, X., Liu, Y., Liu, Z., Tian, S., 2018. "Structural information of asphaltenes derived from petroleum vacuum residue and its hydrotreated product obtained by FT-ICR mass spectrometry with narrow ion isolation windows." *Fuel* 227, 111–117. https://doi.org/10.1016/j.fuel.2018.04.064

Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C., 2012. "Mass spectral molecular networking of living microbial colonies." *Proceedings of the National Academy of Sciences of the United States of America* 109, E1743--52. https://doi.org/10.1073/pnas.1203689109

Weber, R.J.M., Lawson, T.N., Salek, R.M., Ebbels, T.M.D., Glen, R.C., Goodacre, R., Griffin, J.L., Haug, K., Koulman, A., Moreno, P., Ralser, M., Steinbeck, C., Dunn, W.B., Viant, M.R., 2017. "Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy." *Metabolomics* 13, 1–5. https://doi.org/10.1007/s11306-016-1147-x

Weber, R.J.M.M., Viant, M.R., 2010. "MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways." *Chemometrics and Intelligent Laboratory Systems* 104, 75–82.

https://doi.org/10.1016/j.chemolab.2010.04.010

Weismann, A., 1876. Beiträge zur Naturgeschichte der Daphnoiden. W. Engelmann.

Wenderski, T.A., Stratton, C.F., Bauer, R.A., Kopp, F., Tan, D.S., 2015. "Principal Component Analysis as a Tool for Library Design: A Case Study Investigating Natural Products, Brand-Name Drugs, Natural Product-Like Libraries, and Drug-Like Libraries Todd." *Chemical Biology* 1263, 225–242. https://doi.org/10.1007/978-1-4939-2269-7

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.. C. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. a., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wishart, D.S., 2011. "Advances in metabolite identification." *Bioanalysis* 3, 1769–1782. https://doi.org/10.4155/bio.11.155

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., Macinnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J., Querengesser, L., 2007. "HMDB: the Human

Metabolome Database." *Nucleic acids research* 35, D521--6. https://doi.org/10.1093/nar/gkl923

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., Nieva de la Hidalga, A., Balcazar Vargas, M.P., Sufi, S., Goble, C., 2013. "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud." *Nucleic acids research* 41, 557–561. https://doi.org/10.1093/nar/gkt328

Worley, B., Powers, R., 2013. "Multivariate Analysis in Metabolomics." *Current Metabolomics* 1, 92–107. https://doi.org/10.2174/2213235X11301010092

Zhou, B., Cheema, A.K., Ressom, H.W., 2010. "SVM-based spectral matching for metabolite identification." *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10* 756–759. https://doi.org/10.1109/IEMBS.2010.5626337

Zhou, J., Weber, R.J.M.M., Allwood, J.W., Mistrik, R., Zhu, Z., Ji, Z., Chen, S., Dunn, W.B., He, S., Viant, M.R., 2014. "HAMMER: Automated operation of mass frontier to construct in silico mass spectral fragmentation libraries." *Bioinformatics* 30, 581–583. https://doi.org/10.1093/bioinformatics/btt711