

Improving Metabolite Annotation and Identification in Untargeted UHPLC-MS Metabolomics Studies

By

William Nash

A thesis submitted to the University of Birmingham for the degree of DOCTOR
OF PHILOSOPHY

School of Biosciences

University of Birmingham

Edgbaston

Birmingham

B15 2TT

February 2020

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This unpublished thesis/dissertation is copyright of the author and/or third parties.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

Untargeted metabolomics applying UHPLC-MS is a powerful technique capable of reporting thousands of signals from within a complex sample in one analysis and is a technique with vast potential. However, metabolite annotation is a major challenge and bottleneck in untargeted metabolomics experiments partially preventing the field from advancing further and making more substantial impact in the wider scientific community. UHPLC-MS is most typically applied and involves the collection of MS¹ and MS² data but there is much scope for improvement in the acquisition and analysis of both. This thesis characterises and addresses improvements for both data types. Firstly, the relationships between electrospray-derived MS¹ features will be investigated across 104 datasets providing new insight into the complexity of MS¹ data and new recommendations for annotation of MS¹ data. Secondly, MS² data acquisition strategies were investigated on Orbitrap based instruments, with data dependent acquisition (DDA), data independent acquisition (DIA), all ion fragmentation (AIF) and intelligent data dependent acquisition (iDDA) MS² types investigated. The volume of informative MS² information that was acquired was assessed with iDDA techniques being shown to increase biological knowledge collected from any future biological studies applying an Orbitrap analyser. Finally, a library containing chromatographic retention times and MS² data were constructed utilising the Orbitrap ID-X Tribrid Mass Spectrometer (Thermo Fisher Scientific, USA), with human biofluids subsequently analysed utilising the systems new AcquireX software capabilities for automated on-the-fly iDDA acquisition. The results and new tools developed have provided enhancements in the metabolite annotation workflow in untargeted metabolomics applying UHPLC-MS.

Acknowledgements

I would like to thank BBSRC for and Thermo Fisher Scientific for funding this work. Thanks also goes to Professor Warwick Dunn for his efforts in supervising this work and for always looking out for his students. I'd also like to thank Dr Ioanna Ntai and the team at Thermo Fisher Scientific in San Jose for hosting myself and Dr Elliott Palmer whom I also owe great thanks for his significant support in data collection for the final chapter in San Jose. Finally, I'd like to thank the rest of the metabolomics team on the 4th floor of the School of Biosciences and anyone who gave me support and a special mention to Judith Ngere for putting up with me and supporting me throughout the write up!

Contents

Abstract.....	1
Acknowledgements.....	2
List of Figures	7
List of Tables	11
List of Abbreviations	13
Publications.....	17
1.0 Introduction	18
1.1 Mass Spectrometry	19
1.2 How does a Mass Spectrometer Work?	20
1.2.1 Sample Inlets.....	21
1.2.2 Liquid Chromatography	22
1.2.3 Ion sources including electrospray ionisation.....	28
1.2.4 Mass Analysers.....	32
1.3 Metabolomics	48
1.3.1 The Untargeted Metabolomics Workflow	50
1.4 Closing Statement.....	76
2.0 Materials and Methods.....	79
2.1 Characterising the complexity of electrospray ionisation data and its impact on metabolite annotation in UPLC-MS metabolomics studies (Chapter 3.0)	80
2.1.1 Datasets applied.....	80
2.1.2 Data processing and interpretation	80
2.2 Characterisation of UHPLC-MS full scan data complexity and its influence on MS ² data collection on Q Exactive mass spectrometers (Chapter 4.0)	85
2.2.1 Sample Preparation.....	85
2.2.2 Chromatography	86
2.2.3 UHPLC-MS Data Acquisition.....	88
2.2.4 Conversion to .mzML Format.....	89
2.2.5 Isotopologue Parameter Optimisation (IPO)	89
2.2.6 XCMS data processing.....	90
2.2.7 Theoretical DIA Window Complexity Assessment	91
2.2.8 Peak Width Assessment.....	91
2.2.9 Scan Rate Estimation	91
2.3 Comparison of different MS ² acquisition strategies on the Q Exactive Plus (Chapter 5.0)	92

2.3.1 Sample Preparation.....	93
2.3.2 UHPLC-MS	93
2.3.5 Generation of Exclusion Lists	99
2.3.6 Generation of Inclusion Lists.....	100
2.3.7 Data Processing.....	100
2.3.8 Data Analysis	103
2.4 Assessment of Metabolite Annotation Using AcquireX on the Orbitrap ID-X (Chapter 6.0)	107
2.4.1 Assigning Standard Groups	108
2.4.2 Sample Preparation.....	109
2.4.3 UHPLC-MS	109
2.4.4 Acquisition of Data for Metabolite Standards	116
2.4.5 Acquisition of Data for Biological Samples	117
2.4.5 mzVault MS ² Library Construction	117
2.4.6 Data Processing and Identification	121
2.4.7 Data Analysis	124
3.0 Characterising the complexity of electrospray ionisation data and its impact on metabolite annotation in UPLC-MS metabolomics studies.....	127
3.1 Introduction	128
3.2 Results and Discussion	129
3.2.1 Determination of an Appropriate Correlation Coefficient.....	129
3.2.2 Complexity of Alternative Ion Types in Untargeted UHPLC-MS Experiments	139
3.2.3 Comparison of Mass Difference Frequencies Across Datasets.....	145
3.2.4 Are biological transformations with the same retention time for reactant and product metabolites detected in these data?	147
3.2.5 Are there homogeneous and heterogeneous dimers observed?	149
3.2.6 Which Adducts are Searched for in Other Commonly Used Annotation Software?	152
3.3 Conclusions	155
4.0 Characterisation of UHPLC-MS full scan data complexity and its influence on MS ² data collection on Q Exactive mass spectrometers	158
4.1 Introduction	159
4.2 Results and Discussion	162
4.2.1 XCMS Parameter Optimisation	162
4.2.2 Determination of appropriate MS ¹ mass resolution.....	168
4.2.3 Theoretical DIA Window Complexity Assessment	174
4.2.4 Peak Width Assessments and Scan Rate Estimations	186

4.2.5 DIA Methods Designed	189
4.3 Conclusions	190
5.0 Comparison of different MS ² acquisition strategies on the Q Exactive Plus	193
5.1 Introduction	194
5.2 Results and Discussion	196
5.2.1 Number of Features Detected	196
5.2.2 Number of Features with MS ² Data	198
5.2.3 Number of Features Annotated	200
5.2.4 Purity of Fragmentation Windows	205
5.2.5 Advantages of Repeated Injections	208
5.2.6 MS-DIAL Deconvolution Assessment	213
5.3 Conclusions	226
6.0 Assessment of Metabolite Annotation Using AcquireX on the Orbitrap ID-X	228
6.1 Introduction	229
6.2 Results and Discussion	231
6.2.1 Number of Metabolites Detected and mzVault Library Contents	231
6.2.2 Advantage of AcquireX vs Traditional DDA	233
6.2.3 How Many Repeated Injections Are Required?	244
6.3 Conclusions	260
7.0 Conclusions	263
8.0 Bibliography	270
9.0 Appendix	299
9.1 Characterising the complexity of electrospray ionisation data and its impact on metabolite annotation in UPLC-MS metabolomics studies	300
9.2 Characterisation of UHPLC-MS full scan data complexity and its influence on MS ² data collection on Q Exactive mass spectrometers	303
9.2.1 XCMS Parameter Optimisation	303
9.3 Comparison of different MS ² acquisition strategies on the Q Exactive Plus	315
9.3.1 Number of features detected (MS-DIAL)	315
9.3.2 Number of Features with MS ² Data	317
9.3.3 Number of Features Annotated	318
9.3.4 Purity of Fragmentation Windows	323
9.3.5 Value of Repeated Injections	329
9.4 Assessment of Metabolite Annotation Using AcquireX on the Orbitrap ID-X.	332

9.4.1 How Many Repetitive Injections Are Required?	332
--	-----

List of Figures

Figure 1:.....	20
Figure 2:.....	22
Figure 3:.....	25
Figure 4:.....	26
Figure 5:.....	29
Figure 6:.....	30
Figure 7:.....	36
Figure 8:.....	37
Figure 9:.....	38
Figure 10:.....	38
Figure 11:.....	42
Figure 12:.....	46
Figure 13:.....	47
Figure 14:.....	51
Figure 15:.....	61
Figure 16:.....	71
Figure 17:.....	73
Figure 18:.....	82
Figure 19:.....	94
Figure 20:.....	97
Figure 21:.....	101
Figure 22:.....	101
Figure 23:.....	102
Figure 24:.....	102
Figure 25:.....	102
Figure 26:.....	104
Figure 27:.....	104
Figure 28:.....	106
Figure 29:.....	111
Figure 30:.....	114
Figure 31:.....	115
Figure 32:.....	116
Figure 33:.....	119
Figure 34:.....	120
Figure 35:.....	120
Figure 36:.....	120
Figure 37:.....	121
Figure 38:.....	121
Figure 39:.....	122
Figure 40:.....	122
Figure 41:.....	122
Figure 42:.....	122
Figure 43:.....	123
Figure 44:.....	123
Figure 45:.....	124
Figure 46:.....	124
Figure 47:.....	131

Figure 48:.....	132
Figure 49:.....	135
Figure 50:.....	136
Figure 51:.....	137
Figure 52:.....	138
Figure 53:.....	142
Figure 54:.....	146
Figure 55:.....	150
Figure 56:.....	161
Figure 57:.....	165
Figure 58:.....	166
Figure 59:.....	167
Figure 60:.....	168
Figure 61:.....	178
Figure 62:.....	179
Figure 63:.....	180
Figure 64:.....	181
Figure 65:.....	182
Figure 66:.....	183
Figure 67:.....	184
Figure 68:.....	185
Figure 69:.....	197
Figure 70:.....	199
Figure 71:.....	201
Figure 72:.....	203
Figure 73:.....	204
Figure 74:.....	206
Figure 75:.....	208
Figure 76:.....	210
Figure 77:.....	212
Figure 78:.....	214
Figure 79:.....	216
Figure 80:.....	217
Figure 81:.....	218
Figure 82:.....	219
Figure 83:.....	220
Figure 84:.....	221
Figure 85:.....	222
Figure 86:.....	222
Figure 87:.....	224
Figure 88:.....	225
Figure 89:.....	225
Figure 90:.....	235
Figure 91:.....	237
Figure 92:.....	239
Figure 93:.....	240
Figure 94:.....	241
Figure 95:.....	243
Figure 96:.....	244

Figure 97:.....	245
Figure 98:.....	246
Figure 99:.....	247
Figure 100:.....	249
Figure 101:.....	250
Figure 102:.....	251
Figure 103:.....	253
Figure 104:.....	254
Figure 105:.....	254
Figure 106:.....	256
Figure 107:.....	257
Figure 108:.....	258
Figure 109:.....	259
Figure 110:.....	260
Figure 111:.....	260
Figure 112:.....	303
Figure 113:.....	304
Figure 114:.....	304
Figure 115:.....	305
Figure 116:.....	305
Figure 117:.....	306
Figure 118:.....	306
Figure 119:.....	307
Figure 120:.....	307
Figure 121:.....	308
Figure 122:.....	308
Figure 123:.....	309
Figure 124:.....	309
Figure 125:.....	310
Figure 126:.....	310
Figure 127:.....	311
Figure 128:.....	311
Figure 129:.....	312
Figure 130:.....	312
Figure 131:.....	313
Figure 132:.....	313
Figure 133:.....	314
Figure 134:.....	314
Figure 135:.....	315
Figure 136:.....	315
Figure 137:.....	316
Figure 138:.....	316
Figure 139:.....	317
Figure 140:.....	317
Figure 141:.....	318
Figure 142:.....	318
Figure 143:.....	319
Figure 144:.....	319
Figure 145:.....	320

Figure 146:	320
Figure 147:	321
Figure 148:	321
Figure 149:	322
Figure 150:	322
Figure 151:	323
Figure 152:	324
Figure 153:	325
Figure 154:	326
Figure 155:	327
Figure 156:	328
Figure 157:	329
Figure 158:	329
Figure 159:	330
Figure 160:	330
Figure 161:	331
Figure 162:	331
Figure 163:	332
Figure 164:	333
Figure 165:	333
Figure 166:	334

List of Tables

Table 1:.....	32
Table 2:.....	34
Table 3:.....	43
Table 4:.....	44
Table 5:.....	49
Table 6:.....	60
Table 7:.....	66
Table 8:.....	85
Table 9:.....	87
Table 10:.....	88
Table 11:.....	88
Table 12:.....	89
Table 13:.....	90
Table 14:.....	91
Table 15:.....	93
Table 16:.....	95
Table 17:.....	97
Table 18:.....	103
Table 19:.....	103
Table 20:.....	108
Table 21:.....	111
Table 22:.....	112
Table 23:.....	113
Table 24:.....	117
Table 25:.....	117
Table 26:.....	125
Table 27:.....	134
Table 28:.....	140
Table 29:.....	144
Table 30:.....	148
Table 31:.....	149
Table 32:.....	151
Table 33:.....	152
Table 34:.....	153
Table 35:.....	154
Table 36:.....	163
Table 37:.....	164
Table 38:.....	169
Table 39:.....	171
Table 40:.....	173
Table 41:.....	173
Table 42:.....	174
Table 43:.....	175
Table 44:.....	176
Table 45:.....	187
Table 46:.....	188
Table 47:.....	215

Table 48:	223
Table 49:	232
Table 50:	232
Table 51:	232
Table 52:	233
Table 53:	242
Table 54:	250
Table 55:	253
Table 56:	257
Table 57:	258
Table 58:	300
Table 59:	302

List of Abbreviations

2D-LC = Two Dimensional Liquid Chromatography

ACN = Acetonitrile

AIF = All Ion Fragmentation

ANN = Artificial Neural Networks

APCI = Atmospheric Pressure Chemical Ionisation

APPI = Atmospheric Pressure Photoionisation

AS = Aerospray

AX = AcquireX Method

BN = Bayesian Network

CC = Convergence Chromatography

CD3.0 = Compound Discoverer 3.0 (Thermo Fisher Scientific, USA)

CE = Capillary Electrophoresis

CI = Chemical Ionisation

CID = Collision Induced Dissociation

CSF = Cerebrospinal Fluid

DART = Direct Analysis in Real Time

DDA = Data Dependent Acquisition

DESI = Desorption Electrospray Ionisation

DIA = Data Independent Acquisition

DIMS = Direct Infusion Mass Spectrometry

ECD = Electron Capture Dissociation

EESI = Extractive Electrospray Ionisation

eFT = Enhanced Fourier Transform

EI = Electron Ionisation

ESI = Electrospray Ionisation

ETD = Electron Transfer Dissociation

FA = Formic Acid

FFT = Fast Fourier Transform

FT = Fourier Transform

FT-ICR = Fourier Transform Ion Cyclotron Resonance

FWHM = Full Width Half Maximum

GA-BN = Genetic algorithm-Bayesian Network

GC = Gas Chromatography

GP = Genetic Programming

HCD = Higher-energy Collisional Dissociation

HF = High Field

HILIC = Hydrophilic Interaction Chromatography

HPLC = High Performance Liquid Chromatography

ICP = Inductively Coupled Plasma

ICR = Ion Cyclotron Resonance

i.d. = Internal Diameter

iDDA = Intelligent Data Dependent Acquisition

IMS = Ion Mobility Spectroscopy

IPA = Isopropanol

IRM = Ion Routing Multipole

IT = Ion Trap

IRMPD = Infrared Multiple Photon Dissociation

KPLS = Kernel Partial Least Squares

LC = Liquid Chromatography

LC-MS = Liquid Chromatography Mass Spectrometry

LIT = Linear Ion Trap

MALDI = Matrix Assisted Laser Desorption Ionisation

MAR = Missing at Random

MBPCA = Multiblock Principal Component Analysis

MCAR = Missing Completely at Random

MeOH = Methanol

ML-PLSDA = Multilevel Partial Least Squares Discriminant Analysis

MNAR = Missing not at random and

MS = Mass Spectrometry

MSI = Metabolomics Standards Initiative

MTBE = methyl-*t*-butyl-ether

MVI = Missing Value Imputation

m/z = Mass-to-charge ratio

NMR = Nuclear Magnetic Resonance

OPLS = Orthogonal Partial Least Squares Analysis

PARAFAC = Parallel Factor Analysis

PCA = Principal Components Analysis

PC-DFA = Principal Component Discriminant Function Analysis

PI = Photo-Ionisation

PLS-DA = Partial Least Squares Discriminant Analysis

PLSR = Partial Least Squares Regression

ppm = Parts per Million

PQN = Probabilistic Quotient Normalisation

PSI = Paper Spray Ionisation

Q = Quadrupole

QA = Quality Assurance

QC = Quality Control

QqQ = Triple Quadrupole

Q-TOF = Quadrupole-Time of Flight

RF = Radio Field

RF = Random Forests

RP = Reversed Phase

RSD = Relative Standard Deviation

RT = Retention Time

SFC = Super Critical Fluid Chromatography

SVM = Support Vector Machines

TOF = Time of Flight

TS = Thermospray

UHF = Ultra High Field

UHPLC-MS = Ultra High Performance Liquid Chromatography Mass Spectrometry

UV = Ultraviolet

Publications

Nash, W.J. and Dunn, W.B. (2019) **From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data.** *TrAC - Trends in Analytical Chemistry*. doi:10.1016/j.trac.2018.11.022.

Nash, W.J., Weber, R., Dunn, W.B., **ESI Data Complexity.** *In preparation.*

Nash, W.J., Palmer, E.A., Dunn, W.B., **Optimal MS² acquisition strategies on Orbitrap Based Mass Spectrometers.** *In preparation.*

1.0 Introduction

1.1 Mass Spectrometry

A mass spectrometer (MS) is a scientific instrument which measures the mass-to-charge ratio (m/z) of ions in a vacuum through monitoring of their movement under electric and/or magnetic fields (Fenn et al., 1989). The m/z value recorded corresponds to the molecular mass of the ion detected and therefore of the non-charged biochemical. As a result, mass spectrometry is a powerful tool to help us measure and understand the world around us. It is applied across a vast range of industries, sciences and even in everyday life in airport security. The first mass spectrometer was invented by JJ Thomson in 1912 (Thomson, 1912), throughout the following 100 years new developments have gradually been made to improve the performance and capabilities of these scientific tools.

Initially, mass spectrometers were primarily used to prove the existence of isotopes. It was also quickly adopted by the chemical industry to measure the length of hydrocarbons in chemical engineering processes (Brown, 1951). Al Nier is widely credited with spreading the use of mass spectrometry to fields other than theoretical physics and by the 1940s it was being widely applied in many different industries (Griffiths, 2008). He even facilitated the nuclear age through his interest in mass spectrometry when he identified that ^{235}U was the Uranium isotope responsible for slow neutron fission (Nier, 1939). After initially being applied simply for quantitative purposes in the chemistry field, McLafferty, Biemann and Djerassi showed how the mass spectrum could be used to elucidate the structure of unknown peptides and molecules in the 1960s (McLafferty, 1962b, 1962c, 1962a; Biemann, 1962; Biemann et al., 1966; Djerassi et al., 1962; Djerassi and Fenselau, 1965; Djerassi et al., 1965). The next most important step was the application of the Fourier transform (FT) equation to mass spectrometry analyses. Ion cyclotron resonance (ICR) has been applied since 1949 as a method of measuring m/z values and could provide higher mass resolution than the previously developed time of flight (TOF) instrument. However, application of FT to the resulting data by Marshall and Comisarow in the 1970s facilitated analysis of many ions simultaneously through deconvolution of the multiple wave forms being detected (Comisarow and Marshall, 1976). Despite this, MS was still only applied in chemistry and physics, biologists were not using it due to the lack of an appropriate mechanism through which to ionise and transfer fragile biological compounds from the liquid phase into the gaseous phase without significant fragmentation and decomposition. This was until 1988 when Matrix Assisted Laser Desorption Ionisation (MALDI) was invented by Franz Hillenkamp and Michael Karas (Karas and Hillenkamp, 1988). This was shortly followed in 1989 when Electrospray Ionisation (ESI) was popularised by John Fenn who used the principles developed by Malcolm Dole in the 1960s (Dole et al., 1968) to develop his ESI source and couple it to MS analysis (Fenn et al., 1989). Both of these techniques opened the door to analysis of low and high molecular weight biological compounds through successful ionisation and transfer into the gas phase without major fragmentation. At the

start of the 21st century Makarov invented the Orbitrap (Makarov, 2000), a new type of detector based on the Kingdon Trap invented a hundred years before. Makarov's Orbitrap was capable of achieving the same mass resolution as a FT-ICR instrument but with the advantages of, being able to fit onto a benchtop, as well as not requiring liquid nitrogen or helium for operation and therefore was cheaper to run and more convenient to own.

1.2 How does a Mass Spectrometer Work?

Although there are different types of mass spectrometers available they all carry out the same basic functions (Figure 1). The sample, which may be solid, liquid or gas, is introduced to the mass spectrometer through a sample inlet and into the ion source. A mass spectrometer analysis requires gaseous phase ions which are generated in the ion source. These are then directed and filtered through a range of ion optics operating under vacuum conditions and are transferred to the mass analyser. Here ions are separated and their respective m/z values determined which are then recorded by the detector resulting in a mass spectrum as the read-out. Different options for inlet systems and ion sources will be discussed below, followed by a description of how different Orbitrap mass spectrometers function.

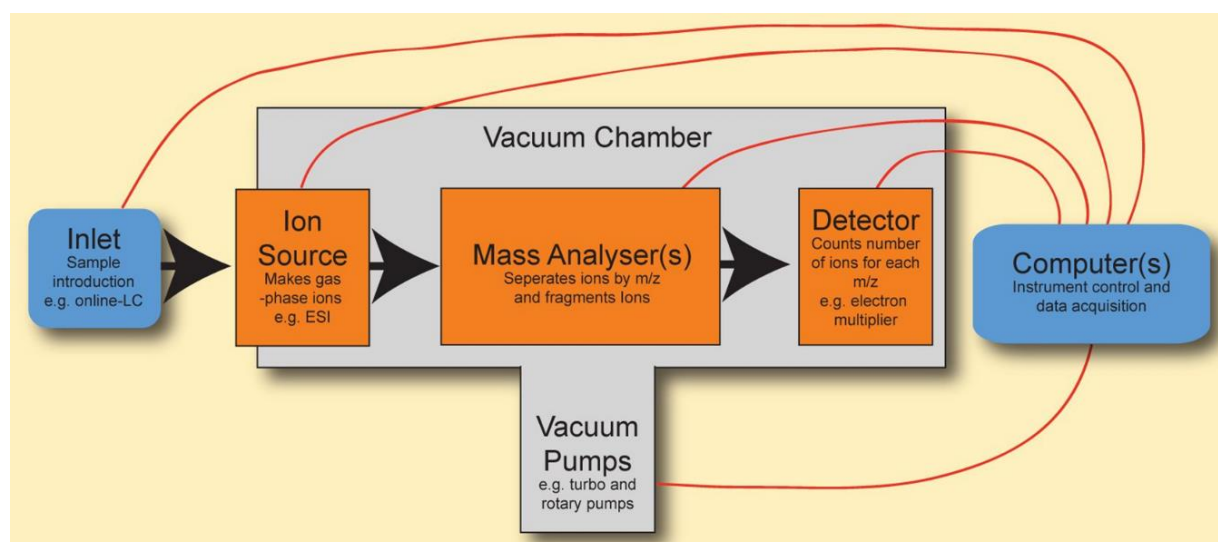


Figure 1: Schematic of the main components of a mass spectrometer. Liquid sample is injected at the inlet followed by transmission to the gas phase in the ion source before entry into the mass spectrometers vacuum system. Ion optics focus, guide and filter the ions before mass analysis is carried out in the analyser and is recorded by a detector. The detected signals are processed by a computer to provide the raw data (Murray, 2020).

1.2.1 Sample Inlets

The sample inlet can constitute a wide variety of technologies and platforms. These can include liquid chromatography (LC), gas chromatography (GC), capillary electrophoresis (CE), ion mobility spectroscopy (IMS), super critical fluid chromatography (SFC)/convergence chromatography (CC), inductively couple plasma (ICP) or be directly infused (e.g. Direct Infusion Mass Spectrometry). Each of them has their own advantages and disadvantages and may be of particular use for certain applications and not suitable for others or alternatively may be used to provide complementary data to each other. GC is well suited to analysis of volatile compounds before or after chemical derivatisation, CE is useful for highly polar and/or charged compounds, IMS is applied for separation of isomers, SFC/CC is a versatile platform that can perform chiral or achiral separations as well as polar and non-polar over a wide polarity range simultaneously (Shulaev and Isaac, 2018), ICP is applied for elemental analysis and direct infusion can allow high throughput analysis. LC is the predominant technique applied for most applications due to its versatility, strong knowledge base and well established use.

1.2.2 Liquid Chromatography

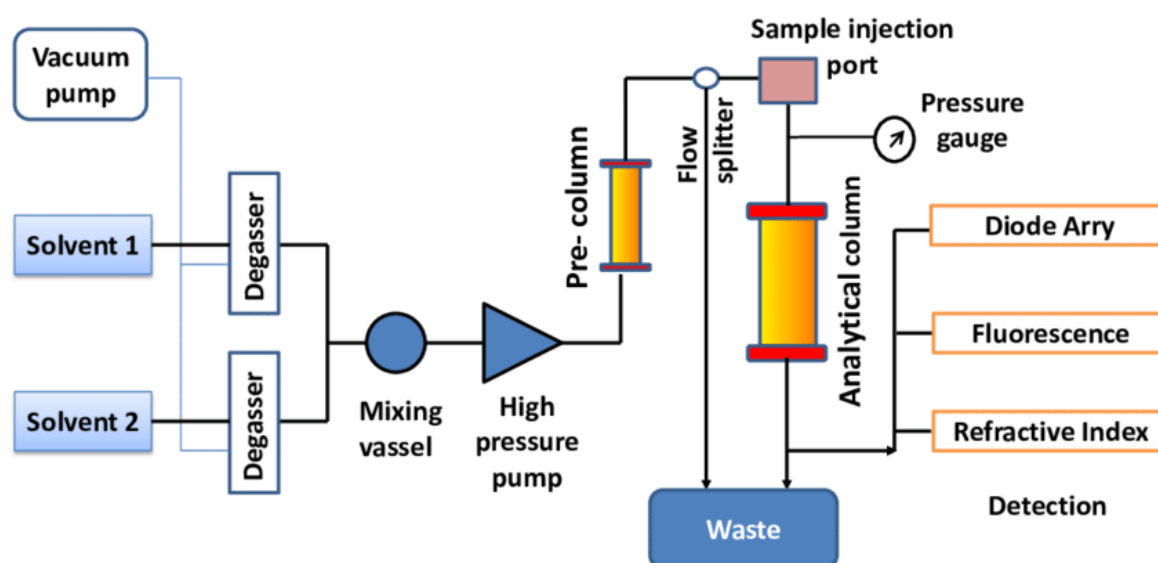


Figure 2: A liquid chromatography systems main components. Two solvent reservoirs are degassed, removal of any air bubbles helps to prevent damage to the LC pumps. The solvents are precisely mixed to the right percentage composition through a series of pumps and valves under high pressure. A pre-column may be applied to prevent dirtying and help maintain performance of the analytical column. Flow can be sent to the analytical column and detector or to waste using the flow splitter. The sample of interest is injected through the sample injection port and the resulting mixture is transferred to the analytical column which separates the components of the mixture by their physicochemical characteristics. Monitoring of the system pressure through the pressure gauge is important. Signals are recorded by a detector as components of the mixture elute from the column and go to waste unless connected to another system such as a mass spectrometer for further analysis. The detector if not a mass spectrometer can be of different types including diode array, fluorescence and refractive index amongst others (Islam, 2013).

LC systems come in many different configurations but the basic layout of most is as seen in Figure 2. Typically, there are two different solvent reservoirs utilised which sit on top of the LC system, although some LCs can use more. The solvent is mixed with the sample which is delivered in an automated fashion by the autosampler. A system of pumps (typically binary, sometimes quaternary) and valves ensure efficient and precise mixing of sample and solvents under high pressure. The resulting analyte mixture is pumped through the analytical column at high pressure for separation. The column is stored in the column compartment which may be heated to aid separation. The analyte mixture will pass through the column and components of the mixture elute at different times depending on their physicochemical characteristics. As they elute they are recorded by the detector which can constitute

a wide variety of different technologies including MS, ultraviolet (UV), diode-array, electron-capture, flame-photometric and nitrogen-phosphorous detectors (Rahman et al., 2017).

The chromatographic column is packed with a solid stationary phase. The stationary phase can vary in physicochemical characteristics to provide separation of biochemicals with different physicochemical properties; for example, HILIC columns are effective separators of water-soluble metabolites whereas C_{18} reversed-phase columns are effective separators of more hydrophobic compounds. C_{18} reversed-phase (RP) is the most commonly applied type of column. The packing material consists of silica molecules with multiple alkyl chains, 18 carbons in length, covalently bonded. These provide a hydrophobic environment for the separation of compounds. The sample is mixed with the mobile phase, which is typically provided by two solvent reservoirs. The proportions of solvents A (hydrophilic in RP, typically water) and B (hydrophobic in RP, typically methanol (MeOH) or acetonitrile (ACN)) will be modulated by a system of pumps in a controlled fashion throughout the run to achieve the desired separation. The sample is carried with the mobile phase and flows through the column. As this occurs the solvent composition will be gradually changing from hydrophilic to hydrophobic and the compounds within will be retained on the stationary phase for varying lengths of time dependent on the hydrophobicity of the solvent composition, the strength of the non-covalent bonding to the stationary phase based on the physicochemical characteristics of the compound, the stationary phase, chemical modification of silica and the current solvent composition (Figure 3 and Figure 4).

Whatever the chemistry of the column is, the fundamental concepts remain the same. The process is dependent upon the movement of the analytes within the sample between the stationary and mobile phases. For any analyte injected on to a column, a theoretical equilibrium can be established between the phases. The equilibrium state will differ based on the shifting mobile phase composition. The fact these equilibria exist allows determination of an equilibrium constant (K) for any analyte, this is also known as the partition coefficient. This can be calculated by dividing the concentration of the analyte bound to the stationary phase by the concentration of the analyte present in the mobile phase. Any analyte running through a column is present in a “band” and for good quality separation the aim is to maintain the width of these bands as much as possible but what causes a band to increase or decrease in width? The level of affinity the analyte has for the mobile and stationary phases affects the rate of transfer of the analyte between the phases with the time taken for equilibrium to be achieved, known as the resistance to mass transfer (C). Other factors that influence the movement of the analyte within the column are eddy diffusion (A), this refers to the many different possible paths that an analyte molecule can take through the stationary phase, some paths will be shorter and some longer and Longitudinal diffusion (B), this refers to how diffusion of the analyte molecules within a band of an

analyte causes lower concentrations of that analyte at the edges of the band. The last factor is the velocity of the mobile phase (u). These factors combine to provide the rate theory of chromatography and were subsequently combined to give the Van Deemter equation which calculates the height of a theoretical plate (van Deemter et al., 1956). A theoretical plate refers to the plate theory of chromatography, this is where the column is divided into a number of imagined plates, with each plate reaching its own equilibrium state (Martin and Synge, 1941). These plates are used to measure column efficiency, the more theoretical plates (the smaller the height of each plate) the better the separation provided by the column should be.

Equation 1: The Van Deemter equation for calculating the height of a theoretical plate.

$$HETP = A + \frac{B}{u} + Cu \quad (1)$$

A = Eddy diffusion

B = Longitudinal diffusion

C = Resistance to mass transfer

u = Average mobile phase velocity

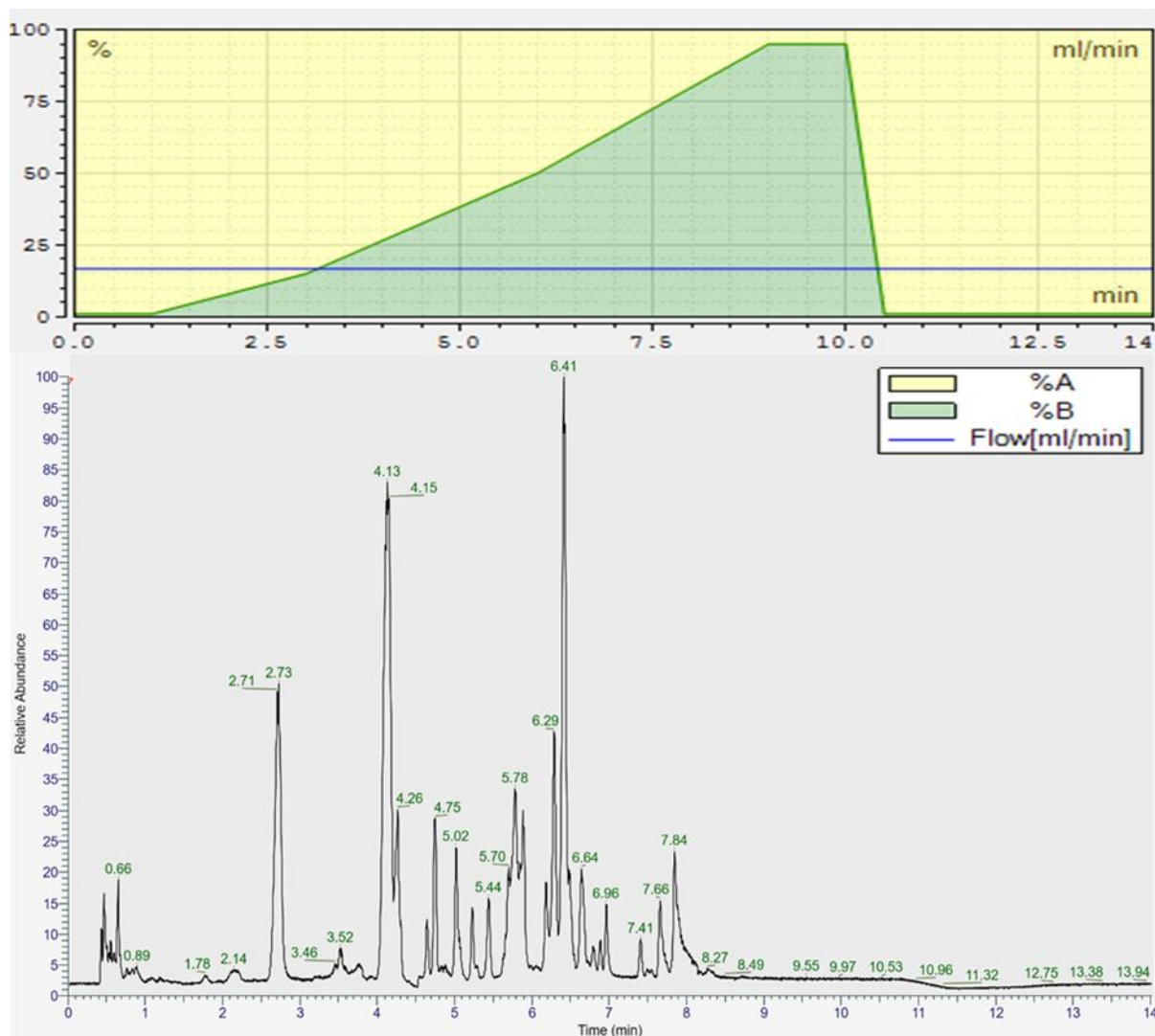


Figure 3: A chromatographic gradient with the associated chromatogram. The gradient steadily shifts throughout the 14 minute run from a high percentage of solvent A and a low percentage of solvent B to an increasingly higher percentage of solvent B and lower percentage of solvent A. Following the peak of solvent B composition it rapidly changes back to the starting composition for the equilibration phase in preparation for the next sample injection. Different components of the sample matrix elute from the column at different times depending on the solvent composition and their own physicochemical characteristics as shown in the chromatogram where the greatest density of peaks are seen between 2.5 and 9 minutes when the solvent composition is constantly changing.

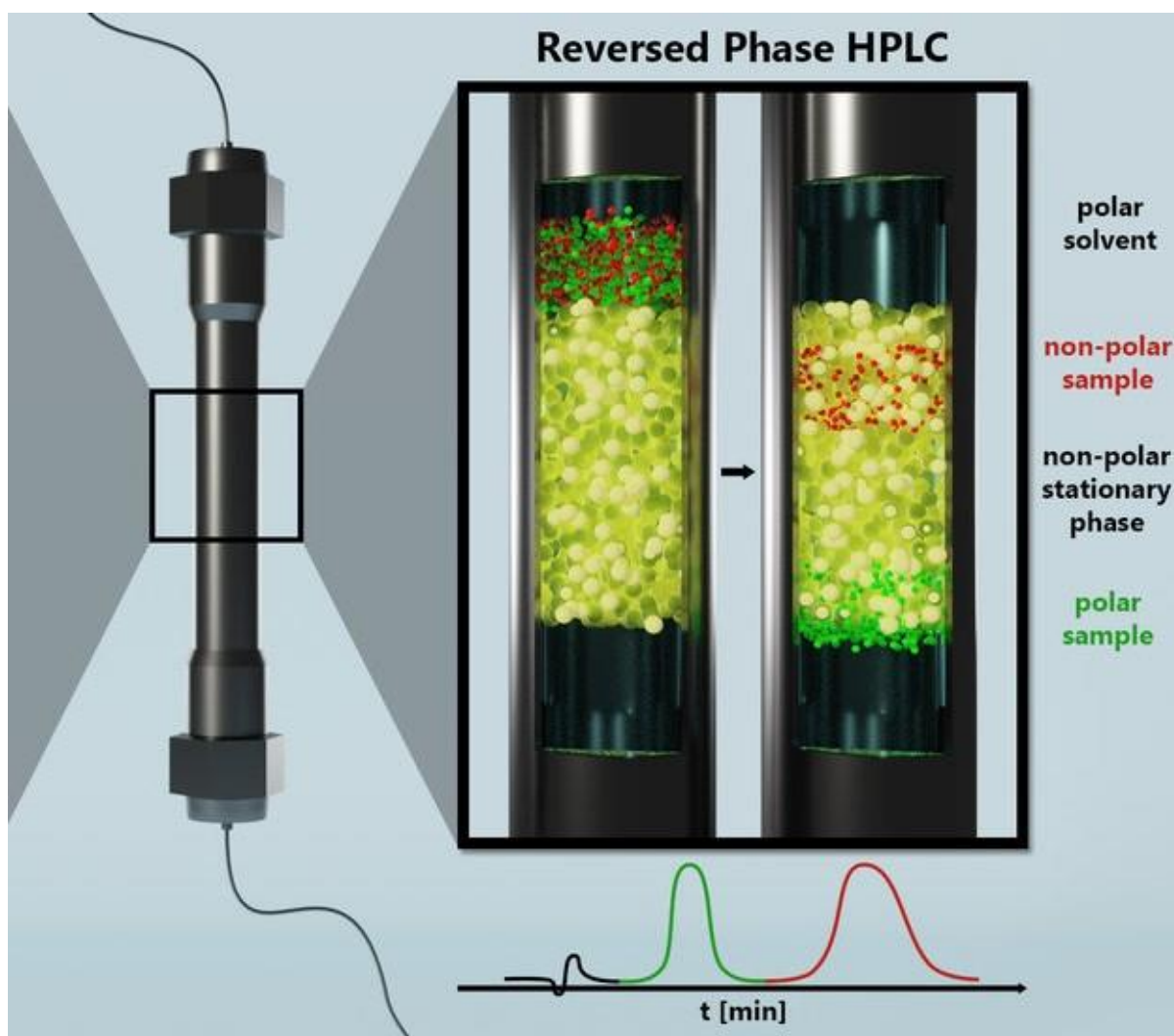


Figure 4: The separation of polar and non-polar compounds using a C_{18} reversed-phase column. The non-polar stationary phase binds the non-polar sample components and they are retained on the column. The polar sample components flow through the column. Retention is manipulated through shifting changing solvent composition throughout the run. Adapted from Max Planck Institute (no date).

RP is most commonly applied due to its suitability to separate a relatively wide range of compounds but alternative stationary phase chemistries are often used to provide complementary information to RP data for compounds not retained and resolved by RP chemistries. An example of this could be by increasing the alkyl chain length of the stationary phase to C_{30} thus increasing its hydrophobicity. Also modifying the solvent composition to be more hydrophobic, with isopropanol (IPA) for example, used in conjunction with ACN. This results in selectivity for hydrophobic and non-polar compounds such as lipids and thus the assay could be described by the umbrella term lipidomics. This term is used to describe a variety of different stationary phase chemistries and solvent compositions utilised such as

the one described to increase sensitivity for detection of lipid and other hydrophobic and non-polar species.

Hydrophilic interaction chromatography (HILIC) is another common alternative stationary phase chemistry. In a HILIC-based column, the stationary phase is typically composed of silica or another polymer with polar groups attached. The mobile phase will always contain some water, which provides an aqueous layer around the stationary phase due its hydrogen bonding interaction with the polar functional groups. The more hydrophobic portion of the mobile phase will be flowing through the cavities within the packing material creating a hydrophilic gradient from mobile phase to packing material. This a valuable technique to increase the selectivity and sensitivity for MS detection of hydrophilic and polar compounds.

Whilst modifying the stationary phase chemistry and solvent chemistry are useful ways to improve chromatographic resolution it is not the only way. The dimensions of the stationary phase can be modified including the length, internal diameter and stationary phase particle size to achieve increased chromatographic resolution. Firstly, by increasing the length of the stationary phase the user can expect to see improved chromatographic resolution due to the greater volume of stationary phase that must be navigated by the sample providing a higher number of interactions and therefore improved separation. This typically comes with the disadvantages of an increased run time and the possibility of broader peaks due to increased longitudinal diffusion though this can be combatted with a higher flow rate. Decreasing the length of the stationary phase decreases the run time of the method and the chromatographic resolution. If losing some resolution is not too important to the application the decreased run time can be valuable for high throughput applications. Secondly, the internal diameter (i.d.) of the column can be decreased. Increasingly there is a push to use columns with smaller internal diameters (for example, 1.0mm i.d. compared to the commonly used 2.1mm i.d.) which confer a number of advantages. The narrower diameter means a lower flow rate must be used to avoid over pressurising the column (for example, $<100\mu\text{L}\cdot\text{min}^{-1}$ compared to the commonly used $>300\mu\text{L}\cdot\text{min}^{-1}$, this means that solvent consumption is decreased, saving money and being a greener application but also that the sample is at a higher concentration in the mobile phase providing greater sensitivity. Thirdly, the particle size of the stationary phase can be decreased. The particle size like the column i.d. is tending to decrease at the state of the art of the field. As the particle size decreases the surface area for sample interaction with the column is increasing. This results in improved chromatographic resolution. These narrow i.d. columns (typically 2.1 mm) and small particle size ($<2.6\mu\text{m}$) columns are known as ultra-high-performance-liquid-chromatography (UHPLC) columns and started to become commercially available in 2004 by the scientific company Waters. Columns with

larger i.d. and/or stationary phase particle sizes are known as high-performance-liquid-chromatography (HPLC). The future of LC is moving towards further miniaturisation. Nano LC offers increases in sensitivity but is not widely applied in metabolomics due to its greater difficulty, lower reproducibility and the requirement for dedicated new nano-LC systems with appropriate tubing, pumps and other key infrastructure for accurate pumping and mixing of $\text{nL}\cdot\text{min}^{-1}$ volumes of solvent (Nazario et al., 2015). Micro and Pico-electrospray systems are also available but use of these is lower than with nano-LC (Marginean et al., 2014).

1.2.3 Ion sources including electrospray ionisation

In order for an analyte to be detected within any mass spectrometer it must be charged as without a charge it cannot be manipulated and transferred by the electric/magnetic fields of the MS. The majority of potential analytes present in nature are not naturally charged. Therefore, the analytes in your sample require a mechanism through which a charge can be introduced before analysis. Various methods for achieving this goal have been developed but electrospray ionisation (ESI) has become the most popular for LC-MS applications. ESI was demonstrated by Malcolm Dole in 1968 but was popularised in 1989 for MS applications by John Fenn (Fenn et al., 1989) who later won the Nobel prize for this work. Previously the hard ionisation techniques EI (Electron Ionisation) and PI (Photoionisation) were used but were too harsh, resulting in the fragmentation of the types of the fragile polar biomolecules that are measured in most biological applications. ESI is an example of a soft ionisation technique which could allow the analysis of these fragile compounds with minimal fragmentation. Soft ionisation techniques as the name suggests are less damaging to the metabolite structure. Thermospray (TS) and Aerospray (AS) were applied in the 1980s and were considered soft at the time but did not ionise the sample as effectively as ESI which has become the method of choice due to its sensitivity and suitability to a relatively wide range of compounds (Fenn et al., 1989).

This is not to say that other ionisation techniques are not applied, other soft ionisation techniques include Matrix Assisted Laser Desorption Ionisation (MALDI), Atmospheric Pressure Chemical Ionisation (APCI), Atmospheric Pressure Photoionisation (APPI), Desorption Electrospray Ionisation (DESI), Direct Analysis in Real Time (DART) Extractive Electrospray Ionisation (EESI) and Paper Spray Ionisation (PSI). MALDI was invented just before ESI but is a surface ionisation technique and so is not as suitable to LC-MS analyses as ESI but can be utilised for spatial imaging applications, and typically produces fewer multiply charged ions. APCI is also commonly applied, providing selectivity for non-polar metabolites and thus is an important tool for improved coverage of detectable metabolites (Dunn, 2008). APPI has not been applied for as long as APCI but is also a useful tool for analysis of more non-polar compounds (Marchi et al., 2009). DESI (Wiseman et al., 2008), DART (Gross, 2014),

EESI (Law et al., 2010) and PSI (Yang et al., 2012) are all ambient ionisation techniques that can all be utilised with minimum/no sample pretreatment. This means these techniques can be used for rapid analysis of a sample in the field or clinic for example (Gross, 2014).

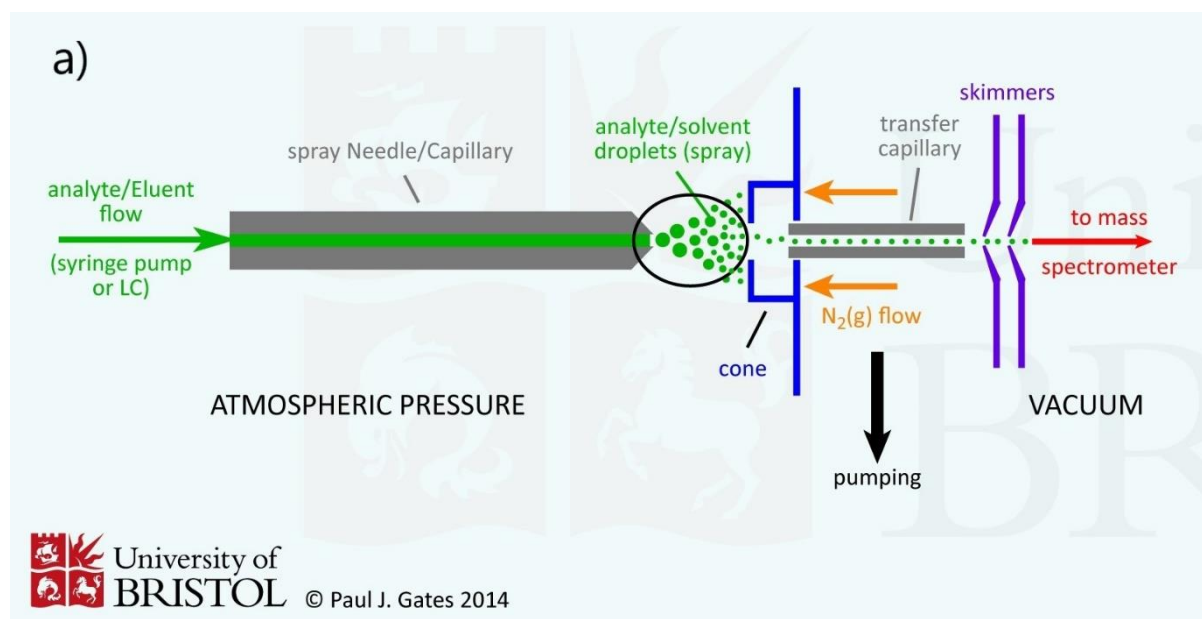


Figure 5: A schematic of a typical ESI source and how it generates gaseous ions for transfer into the mass spectrometer. High temperature and voltage is applied to the spray capillary, solvated ion droplets leave the capillary and enter the gas phase before entry into the transfer capillary and the vacuum system of the mass spectrometer. Adapted from: Gates, P.J (2014).

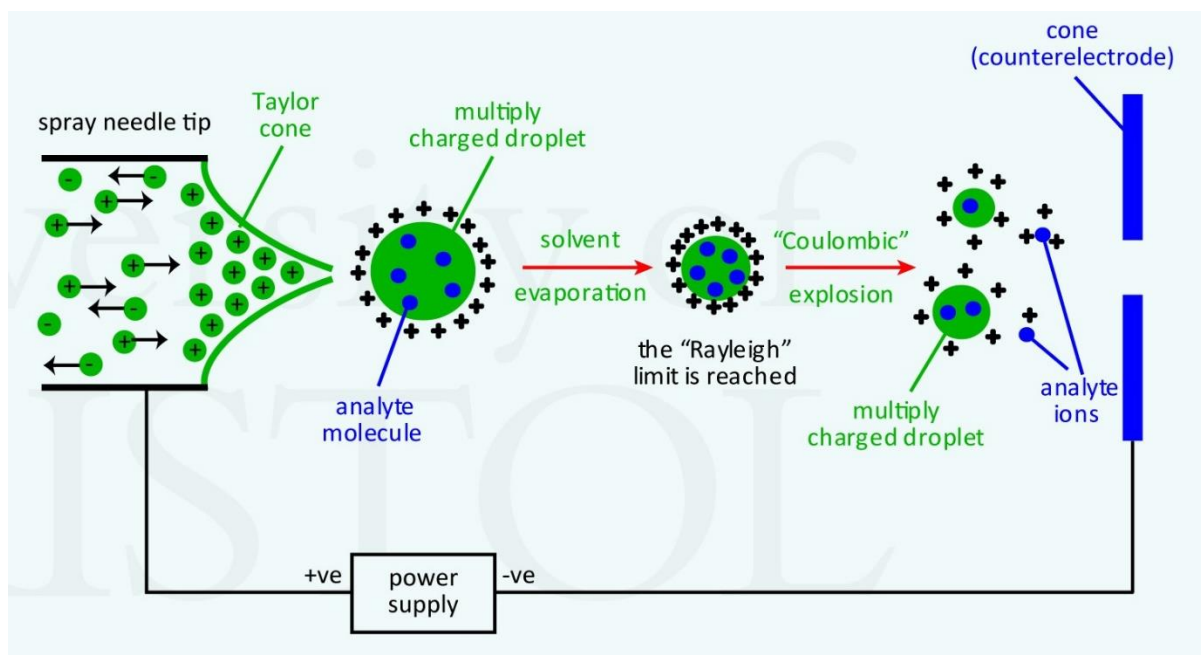


Figure 6: A schematic of how gaseous ions are formed from solvent droplets upon ejection from the Taylor cone. Droplets containing analyte ions are attracted toward the counter-electrode, elongating the Taylor cone before the liquid surface tension is exceeded and the droplet enters the gas phase. Solvent evaporation leads to increasing charge densities and exceeding of the Rayleigh limit on the droplets resulting in a Coulombic explosion. Smaller charged droplets result, the process repeats which and single analyte ions pass into the mass spectrometer. Adapted from: Gates, P.J (2014).

The mechanism through which ESI works is displayed in Figure 5 Figure 6. A typical ESI source operates as an electrochemical reactor producing a variety of ion types (e.g. $[M+H]^+$, $[M-H]^-$) for any single analyte dependent on the ion mode, solvents and modifiers used. The main components of the ESI source are the spray capillary, transfer capillary and counter electrode. The sample is received from the LC or through a manual syringe injection. As the sample passes through the spray capillary it is exposed to high temperatures (typically 250 to 400 °C) and a high voltage (typically 2-6 kV). The differential potential between the spray capillary and the counter electrode draws charged ions formed in the solution towards the end of the capillary and toward the counter electrode at the MS inlet. At the tip of the capillary the Taylor cone forms due to the surface tension of the solvent and the attraction of the charged analytes within the solvent towards the counter electrode. The tip of the cone elongates due to the attraction to the counter electrode and μm sized droplets of solvent with charged analytes within are emitted. A sheath gas is applied to the spray capillary, this helps to direct the analyte ions towards the transfer capillary and limit their spread as they exit the spray capillary. Smaller droplets are created due to the build up of charge density resulting from the multiple charges present in the droplet and the solvent evaporation. When a certain charge density is reached

Coulombic repulsion of ions within the droplet can overpower the surface tension of the solvent as the Rayleigh limit is overcome and smaller charged droplets are released through jet fission (Konermann et al., 2013). This process repeats. There are multiple mechanisms postulated for this process including the ion evaporation model (IEM), the charged residue model (CRM) and the chain ejection model (CEM). These models are applied to different molecule types with IEM applying to small molecules, CRM to larger folded protein or macro-molecular structures and CEM to larger unfolded molecules (Konermann et al., 2013). IEM will be focussed upon due to its greater relevance for small molecules which are the analytes of interest in metabolomics. The ion evaporation model describes how the energy required to extract an ion from a solvent nanodroplet is dependent upon the number of charges in the droplet as well as its overall radius, this was first proposed by Thomson and Iribarne in 1979. The theory and associated equations were refined by Labowsky et al., in 2000 who also showed this process can occur at the Taylor cone. Further molecular dynamics simulations performed since appear to fit well with the model (Higashi et al., 2015; Ichiki and Consta, 2006). The actual process of IEM is very similar to the one described for ejection of larger (μm size) charged solvent droplets. There is an activation energy barrier that must be overcome by an increase in charge density allowing the overcoming of the surface tension and release of a charged ion from the outside of the nanodroplet. Upon release from the droplet the ion moves further away but does so gradually as it remains temporarily connected by what is described as a “sticky solvent bridge” before this breaks and a smaller nanodroplet is created. This process repeats to release solvated and desolvated ions which can pass through the counter electrode into the vacuum region of the mass spectrometer. The counter electrode can thus help to keep uncharged analytes or droplets out of the MS. Once through the counter electrode the analyte ions pass into the vacuum of the mass spectrometer for analysis.

This ionisation mechanism occurs under atmospheric pressure which provides a number of advantages over older high pressure techniques such as EI or CI. Firstly, there is decreased loss and fragmentation of ions due to gas phase collisions, this means increased transfer efficiency and therefore sensitivity. Furthermore, the volume of solvent being introduced into the MS is decreased this is due to the decreased vapour pressure when operating at atmospheric pressure meaning the droplets are more easily desolvated, this should improve data quality. It is also cheaper to not use very high pressure as maintaining the pressure requires a lot of power, it also will still require a vacuum in the MS so it also means more power is required to generate the vacuum in the MS. ESI is also desirable for its suitability for operation with liquid samples which are typically used in biological studies. Another key advantage is the fact that concentration of the sample drives the intensity of features

seen in the sample not the volume of sample used such as with APCI or APPI (Hoffmann and Stroobant, 2007). This fact has driven the miniaturisation of the process.

Nano-ESI can be used in combination with nano-LC mentioned previously. Flow rates reported can be variable with 300 to 1000 nL min⁻¹ (Chetwynd and David, 2018), 200 to 800 nL min⁻¹ (Fanali, 2017) and 20 to 50 nL min⁻¹ (Banerjee and Mazumdar, 2012). Typical LC flow rates range from 100 to 500 µL min⁻¹. Injection volumes are also smaller and can be as low as 0.01 µL (Šesták et al., 2015) compared to a typical LC injection of 2 – 10 µL. This reduction in sample and solvent usage is obviously beneficial to laboratories particularly if working with a precious sample, it is also cheaper and more environmentally friendly. Nano-ESI is more sensitive, with reports of 2000 times greater sensitivity (Chetwynd et al., 2014) and limits of detection 300 times lower (García-Villalba et al., 2010) due to the smaller droplets generated at the Taylor Cone allowing faster and more efficient desolvation and thus release of analyte ions for mass analysis. This extra sensitivity can increase the comprehensiveness of biological studies. The smaller droplets also mean lower temperatures and less desolvation gas is required which should mean more analytes enter the mass spectrometer without suffering any fragmentation. Despite these advantages it is not widely applied currently due to its greater difficulty, it is more prone to blockages for example. On top of this the reproducibility is not as good and throughput is decreased (Chetwynd and David, 2018).

1.2.4 Mass Analysers

Following ionisation, the analyte ions are transferred into the mass spectrometer where mass analysis will ultimately occur and a variety of different data types can be collected (Table 1).

Table 1: The types of data that can be recorded by a mass spectrometer and they constitute and a short description of each data type. m/z = mass-to-charge ratio.

Data Type	Description
MS ¹	Records the m/z of the ions (parent ions) present in the sample.
MS ²	An m/z value or range of values that were detected in the MS ¹ data are isolated and subjected to fragmentation, the m/z values of the resulting ions are recorded.
MS ⁿ	An m/z value detected in the MS ¹ data is isolated and subjected to fragmentation, the m/z of the resulting ions is recorded, the resulting m/z values are isolated and fragmented again. This may be repeated a user defined “n” times.

Once in the mass spectrometer the ions are transmitted through a series of tubes, lenses and skimmers under vacuum in a system known as the ion optics. Electrical, RF and magnetic fields are

used to manipulate the ions paths through this system. The ion optics come in various configurations depending on the type of instrument being utilised. In general, the purpose of the ion optics is to focus ions on a specific path to maximise efficiency of transfer of ions from the source to the mass analyser. For example, the addition of the S-lens, a new improved component of the ion optics in the Orbitrap Velos instrument released in 2009 conferred a 3-5 fold increase in ion transfer efficiency in the full scan and a 10 fold increase in MS² scans (Zubarev and Makarov, 2013). The ion optics also help to remove neutral molecules which can interfere with mass analysis because of ion-molecule collisions. Therefore, the ion optics are important to ensure data quality by removal of neutral molecules and focussing of m/z values of interest before transfer to the mass analyser. Different types of mass spectrometry systems are available including quadrupole (Q), Ion Trap (IT), Time-of-Flight (TOF), Quadrupole-Time of Flight (Q-TOF), Triple Quadrupole (QqQ), Fourier Transform-Ion Cyclotron Resonance (FT-ICR) and Orbitrap (Table 2). Each system has advantages and limitations and is suitable to certain applications. The instrument types can be categorised into high and low mass resolution instruments, where high mass resolution is defined as at least 10,000 Full Width Half Maximum (FWHM) with sub 5 parts per million (ppm) accuracy. Higher mass resolutions are desirable as they provide narrower peak widths, resolve ions of similar m/z , allow detection of more unique analytes and allow greater restriction to the number of putative candidates for any m/z value recorded allowing superior confidence in annotation. The commercial release of the Orbitrap mass analyser in 2005 has made collection of high resolution data increasingly feasible to a wider range of researchers. As a result Orbitrap based systems have become one of the most popular types of mass spectrometers across a range of fields including metabolomics although other types of systems are still very popular such as Q-TOFs and QqQs (Zubarev and Makarov, 2013). The sensitivity in detection of the mass spectrometer is another important consideration. Whilst it is true that all mass spectrometers are sensitive devices some are more sensitive than others and therefore this may be an important consideration depending on the application in question. For example targeted quantitation of a trace level metabolite would benefit from application of a highly sensitive mass spectrometer type such as a triple quadrupole whereas an untargeted investigation of a sample's contents wouldn't require the same level of sensitivity and therefore a higher mass resolution option would be more useful to allow differentiation of a greater number of m/z peaks with similar monoisotopic masses.

Table 2: Different types of mass spectrometers and mass resolution, advantages and limitations.
FWHM = full width at half maximum.

Mass Analyser	Mass Resolution (FWHM at 200 m/z)	Advantages	Limitations
Ion Traps	Unit resolution	<ul style="list-style-type: none"> • Medium purchase cost • High sensitivity 	<ul style="list-style-type: none"> • Low mass resolution • Low mass accuracy
Quadrupole	Unit resolution	<ul style="list-style-type: none"> • Easier to maintain • Robust • High sensitivity 	<ul style="list-style-type: none"> • Low mass resolution • Low mass accuracy
TOF/Q-TOF	Up to 50,000	<ul style="list-style-type: none"> • Fast scan rate • Medium purchase cost 	<ul style="list-style-type: none"> • Medium-to-high mass resolution • Lower sensitivity
QqQ	Unit resolution	<ul style="list-style-type: none"> • Targeted quantitation • Medium purchase cost • Very high sensitivity 	<ul style="list-style-type: none"> • Low mass resolution • Low mass accuracy
FT-ICR	> 2,000,000, up to 10,000,000 at higher masses	<ul style="list-style-type: none"> • Very high mass resolution • Sub-ppm mass accuracy • Very high sensitivity 	<ul style="list-style-type: none"> • Slow scan rate • High purchase cost
Orbitrap (including LTQ-Orbitrap, Q-Exactive and tribrid instruments)	Up to 480,000	<ul style="list-style-type: none"> • High mass resolution • High mass accuracy • Faster scan rate than FT-ICR 	<ul style="list-style-type: none"> • Slow scan rate • Lower sensitivity

1.2.4.1 Quadrupoles and Linear Ion Traps (LITs)

A quadrupole was first applied for mass spectrometry by Paul, Reinhard and von Zahn in 1958 and they are now widely applied across a range of different configurations in different mass spectrometer types. They are primarily applied as mass filtering devices and ion guides preceding another mass analyser type however, in some cases they are used for detection such as in linear ion traps (Douglas, 2009). A quadrupole consists of four rods, the rods can be cylindrical or hyperbolic in shape and are organised as displayed in Figure 7. Manipulation of the AC and DC voltages supplied alters which ions will maintain a stable trajectory through the quadrupole and which will be discharged through collision with the rods of the quadrupole. This allows for the selection of ions of a particular m/z or of the full m/z range. At any one time the rods directly opposite each other will always have the same potential applied which can be positive or negative. The other pair of rods will have the same potential applied but with the opposite charge. Any ion on a path through the quadrupole will thus be attracted to a rod and repelled by another based on its own mass, charge, the current potential configuration of the rods and its own position within the quadrupole. The polarity of the potential on each pair of rods rapidly switches to prevent ions from crashing into a rod and being lost. Which ions are lost is based on the strength of the voltages applied and these can be changed to filter a narrow mass range of ions or a wide mass range of ions. The equations that explain the described principles of ion motion in a quadrupole are detailed by Douglas et al (2005).

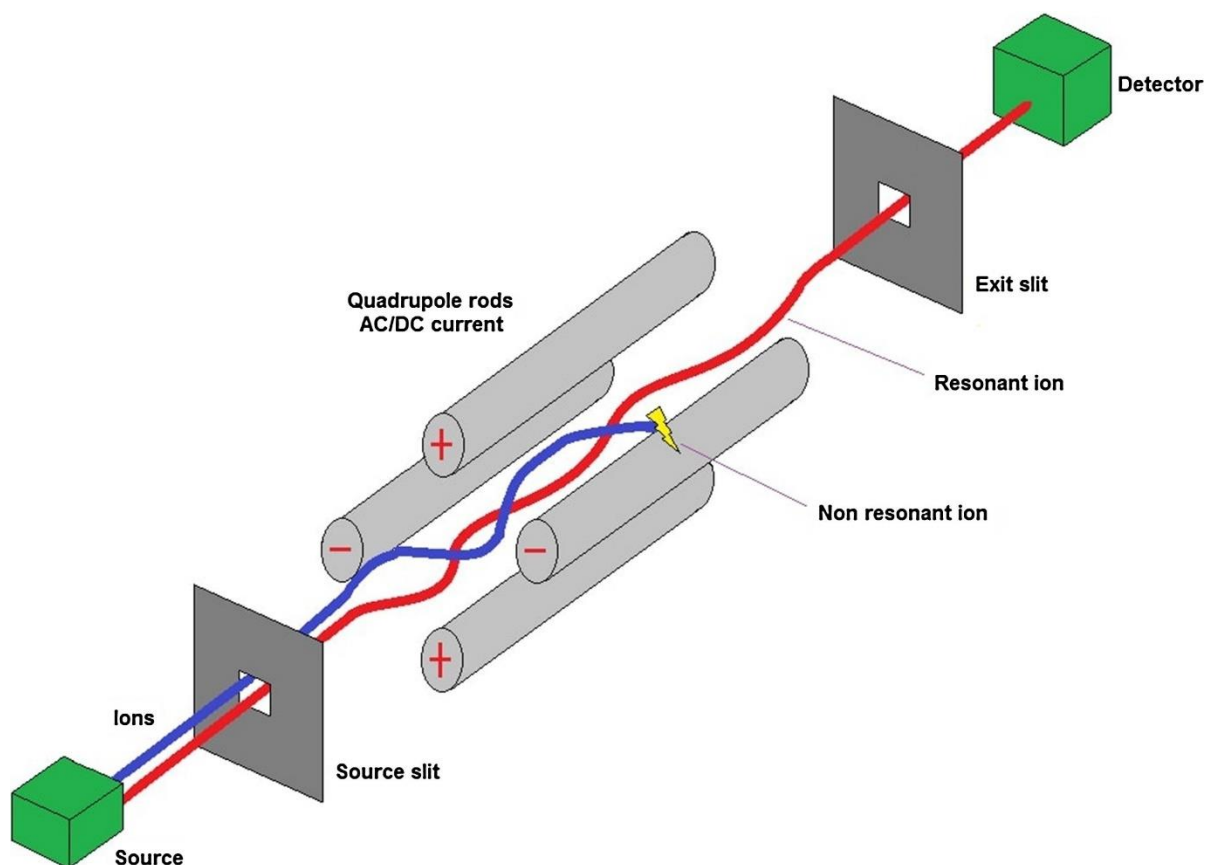


Figure 7: Structure of a cylindrical rod quadrupole showing the stable trajectory of one ion and its subsequent transmission (red line) and the unstable trajectory of another which is lost as it collides with one of the rods (blue line) (Santoiemma, 2018). The two pairs of electrodes that make the quadrupole always have alternate charge to the two adjacent rods and the potentials applied are rapidly alternated to guide ions along the quadrupole. Ion transmission is dependent on the mass of the ion and voltage strength applied.

LITs are structurally similar to quadrupoles with the distinction that they are divided in three segments and have lenses at the ends which can be used to repel ions to keep them perpetually trapped within the quadrupole providing extra functionality for ion storage and manipulation (Figure 8). An inert gas within the quadrupole such as helium is used to cool the ions by collisions with the gas molecules. Application of equal DC voltages to the end electrodes acts to trap ions axially whilst AC voltages trap the ions radially together resulting in the orbital motion of the ions in the centre of the trap (Douglas et al., 2005). These operations allow trapping of the ions with more than 50% efficiency (Hoffmann and Stroobant, 2007). Detection of ions can be performed through their ejection from one end of the trap through application of uneven voltages to the end lenses. Alternatively the LIT can be used to perform collision induced dissociation (CID) fragmentation of analyte ions to gather MS^2 data, this will be described in more detail in section 1.3.1.4.2.3 MS^2 Data and Strategies for Acquisition and Analysis

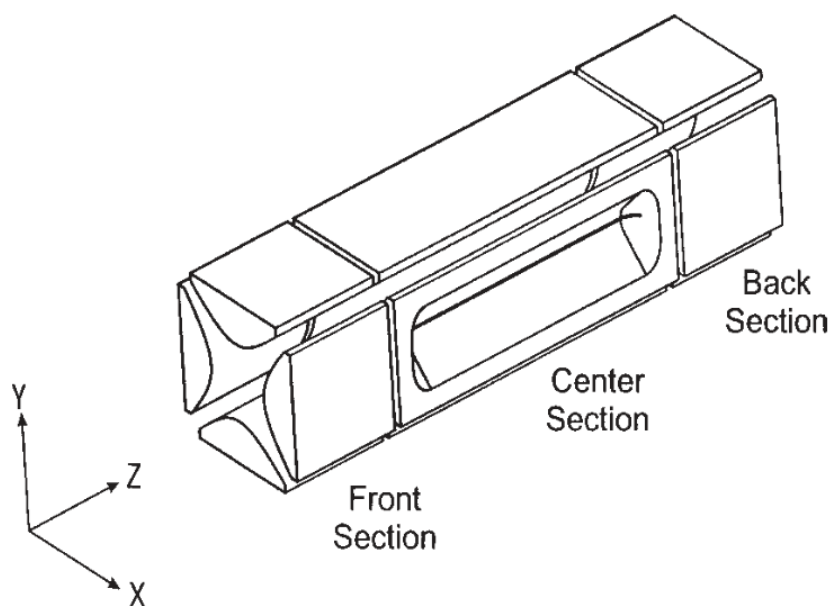


Figure 8: Structure of a LIT, the four hyperbolic shaped rods are divided into three segments. Ions can be trapped perpetually in an orbital motion in the centre section of the trap. Ions move along the y, x and z axes of the trap based on their mass, charge and the voltages applied to the rods. An inert cooling gas reduces the energy of the ions to aid trapping (Douglas et al., 2005).

1.2.4.2 Orbitrap instruments

The first Orbitrap system (Figure 9) was designed utilising three quadrupoles, the first two are utilised to focus, filter and guide analyte ions, the third is then used for storage. Manipulation of the electric fields and potentials controls the transmission and storage of the ions whilst the strength of the vacuum in each section is gradually increased (pressure is decreased) to avoid gas phase collisions resulting in loss or fragmentation of the analyte ions. Ions can then be injected into the Orbitrap mass analyser through a sudden ramping of the voltage in the storage quadrupole causing rapid ejection of the ions present in consecutive packets of m/z values. Different m/z values will eject depending on the current voltage applied during the ramping but the whole process occurs over a period of nanoseconds (Hoffmann and Stroobant, 2007).

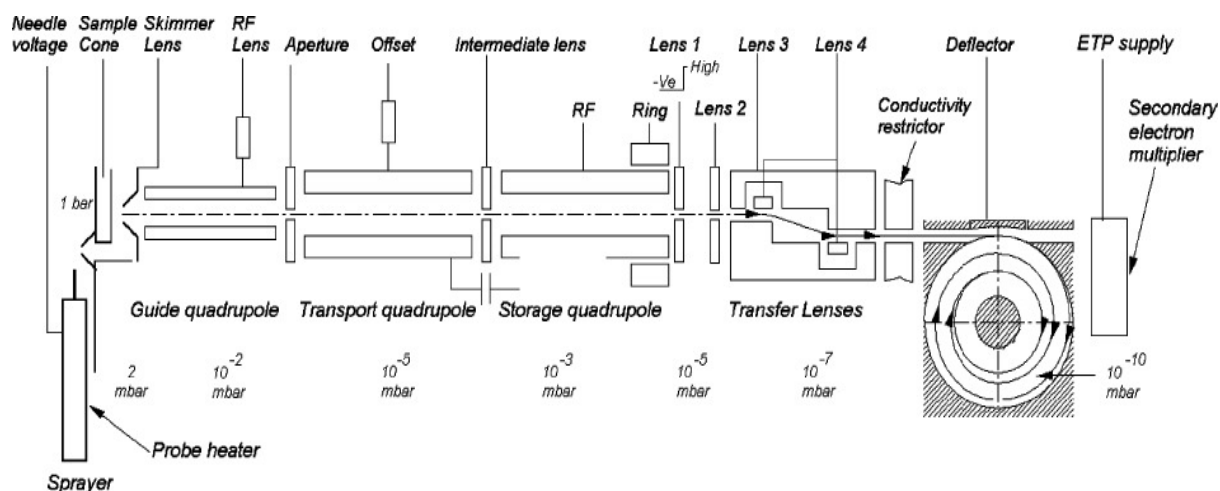


Figure 9: The first Orbitrap system as described by Hu et al, in 2005. A series of quadrupoles and lenses focus and guide ions towards the Orbitrap. The pressure is gradually reduced through the system to its lowest point in the Orbitrap itself where the ions are trapped and measured through detection of their flight paths in the Orbitrap.

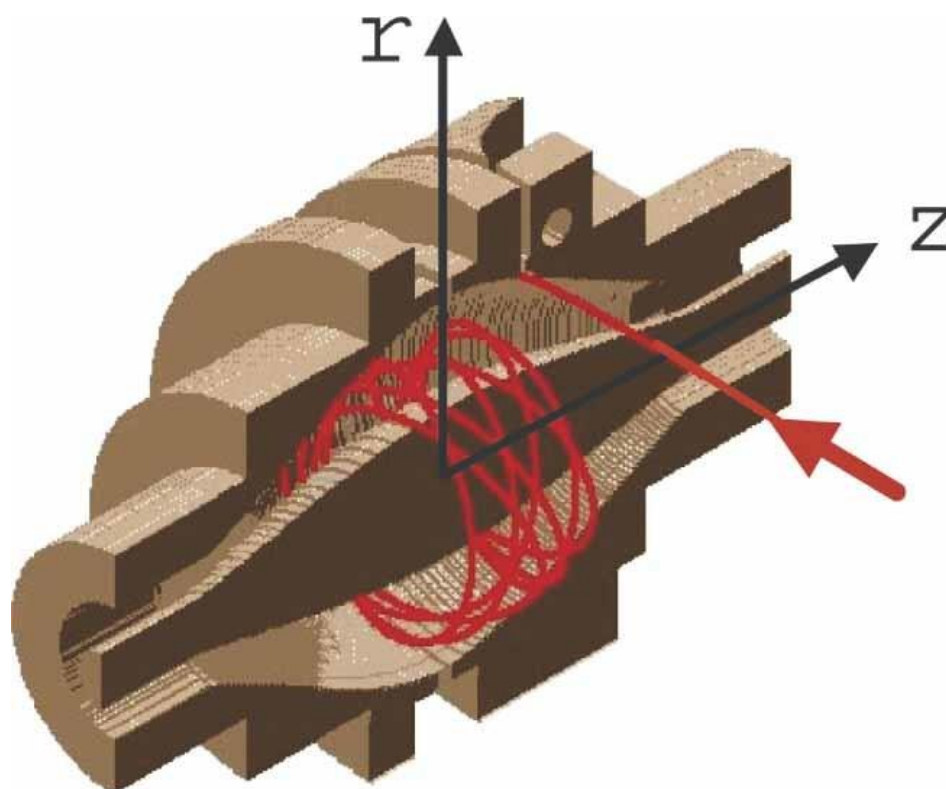


Figure 10: The cross section of an Orbitrap showing the harmonic oscillation of ions along the **z**-axis from Hu et al., (2005). The barrel shaped electrodes trap ions in the quadro-logarithmic field facilitated by the Orbitrap's unique geometry and the potential applied, ions oscillate in the radial (**r**) and axial (**z**) directions. The frequency of the axial oscillations allow calculation of the m/z through the Fourier transform algorithm.

The Orbitrap mass analyser, through which the mass analysis is carried out is based upon the Kingdon trap, invented in 1923 (Kingdon, 1923). This used a wire within a metallic cylinder with voltages applied to both to create a logarithmic potential between the inner and outer electrodes which could trap ions. It was shown that if the ions were travelling at the right speed and the right voltages were applied they would orbit the spindle. Various advancements were made upon this idea in the following years, in particular by Knight who modified the trap to allow injection of ions at the point where $z = 0$. This was shown to be useful for monitoring ions but did not produce any mass spectra (Hu et al., 2005). The Orbitrap mass analyser was proposed by Makarov in 2000 (Makarov, 2000) and the first commercially available mass spectrometer containing an Orbitrap mass analyser was released in 2005 (called the LTQ-Orbitrap) (Zubarev and Makarov, 2013). The inner and outer electrodes are modified in shape compared to the Kingdon trap (Figure 10). The inner electrode is spindle shaped whilst the outer electrodes have a split barrel shape (Hu et al., 2005). There is a narrow gap between the outer electrodes at the point where $z = 0$. The inner electrode has a high voltage applied to it (3.5 kV) whilst the outer electrode is at ground potential. The modified shape of the Orbitrap means that there is a quadro-logarithmic potential instead of a logarithmic potential. If ions are injected with the right kinetic energy (1600 eV) (Hoffmann and Stroobant, 2007) then they will oscillate around the spindle. If the kinetic energy of the ion is too low it will collide with the spindle, if it is too high it will collide

with the outer electrode. Accurate definition of the quadro-logarithmic field conferred by the design of the Orbitrap is essential for generating mass spectra. The equation is provided below.

Equation 2: The Quadro-logarithmic potential distribution in an Orbitrap

$$U(r, z) = \frac{k}{2} \left(z^2 - \frac{(r^2 - R_1^2)}{2} \right) + \frac{k}{2} \times R_m^2 \times \ln \left[\frac{r}{R_1} \right] - Ur$$

$U(r, z)$ = Quadro-logarithmic potential distribution

k = field curvature

R_1 = Radius of the spindle

R_2 = Radius of the outer electrode

R_m = The characteristic radius

r = r-axis coordinate

z = z-axis coordinate

Ur = Spindle Voltage

Once trapped the ions move around the spindle in harmonic axial oscillations along the z-axis. The signal image currents are detected by the outer electrodes. These can be converted into mass spectra through Fast Fourier Transform (FFT) providing the Quadro-logarithmic field can be related to the frequency of the harmonic oscillations by an equation. The field curvature defines the frequency of the axial oscillations this is described by the equation below.

Equation 3: The frequency of axial harmonic oscillations in the Orbitrap.

$$\omega = \sqrt{\frac{e}{m \div z} \times k}$$

ω = frequency of axial oscillation

m = mass

z = charge

k = field curvature

e = elementary charge (1.602×10^{-19} C)

The preceding two equations can thus be combined through k as seen below.

Equation 4: The frequency of axial harmonic oscillations in an Orbitrap relative to quadro-logarithmic field.

$$\omega = \sqrt{\frac{e}{m \div z} \times \frac{2 \times Ur}{R_m^2 \ln\left(\frac{R_2}{R_1}\right) - \frac{1}{2}[R_2^2 - R_1^2]}}$$

Through this equation the relationship between m/z ratio of an ion and its harmonic axial oscillation frequency is described. Subsequent use of the FFT algorithm allows conversion of the signal from the time domain to the frequency domain and into a mass spectrum (Zubarev and Makarov, 2013). This is only facilitated due to the shape of the Orbitrap and the fact that the axial oscillations are totally independent of the energy and spatial spread of the ions (Hu et al., 2005).

Two new versions of the Orbitrap mass analyser have been commercially released since the original in 2005. These are known as the High Field Orbitrap (HF-Orbitrap) and the Ultra High Field Orbitrap (UHF-Orbitrap). It is known that for a trapping style mass analyser, performance can be improved if the strength of the trapping field can be increased (Makarov et al., 2009). This can lead to improved dynamic range, scan rate, resolving power and tolerance to space charge effects (Makarov et al., 2009). With FT-ICR instruments this can only be done through increasing the size of the magnet, that drives the strong magnetic field used for trapping. For Orbitraps the strength of the electrostatic field needs increasing, this will provide an increase in the frequency of the harmonic axial oscillations which in turn will provide the benefits listed earlier in this paragraph. By looking at Equation 4 it can be determined that the only way to increase the frequency for a given mass and charge pair is to increase the spindle voltage applied (Ur), decrease the size of the trap (R_1, R_2), or by increasing the ratio of the spindle radius to the outer electrode radius ($R_1:R_2$) (Makarov et al., 2009). Increasing the voltage applied to the spindle (Ur) is a less effective method than modifying the trap dimensions as any increases in the value of Ur are subsequently square rooted. In 2011 the HF-Orbitrap was released (Zubarev and Makarov, 2013), the field strength was increased compared to a standard Orbitrap in two ways. Firstly, the radius of the spindle (R_1) was increased from 6 mm on the standard Orbitrap to 9 mm whilst the outer electrode radius (R_2) was kept the same at 15 mm. This increased the ratio of ($R_1:R_2$). The voltage applied to the spindle (Ur) was also increased from 3.5 kV to 5 kV. In combination these changes increased the frequency by a factor of 1.7 (Makarov et al., 2009). These modifications provided the improved performance expected, with an increase in possible resolving power, dynamic

range over a fixed period of time, scan rate over a fixed period of time and also improved space charge tolerance. Then in 2014 the UHF-Orbitrap was released. This Orbitrap was first contained within the Q Exactive HF model (Scheltema et al., 2014). The dimensions of both the outer and spindle electrodes were decreased compared to the standard Orbitrap. The outer electrode radius (R_2) was decreased from 15 mm to 10 mm, whilst the spindle electrode radius (R_1) was reduced from 6 mm to 5 mm (Figure 11). The general decrease in the Orbitrap dimensions alongside the increase in the ratio of the spindle electrode to the outer electrodes ($R_1:R_2$) confers increased strength of the electrostatic field and thus increased frequency of harmonic axial oscillations. These modifications confer a 1.8 fold increase in the frequency of harmonic axial oscillations than achieved with the HF-Orbitrap (Kelstrup et al., 2014). This 1.8 fold increase in frequency provides a 1.8 fold increase in resolution at a given transient length.

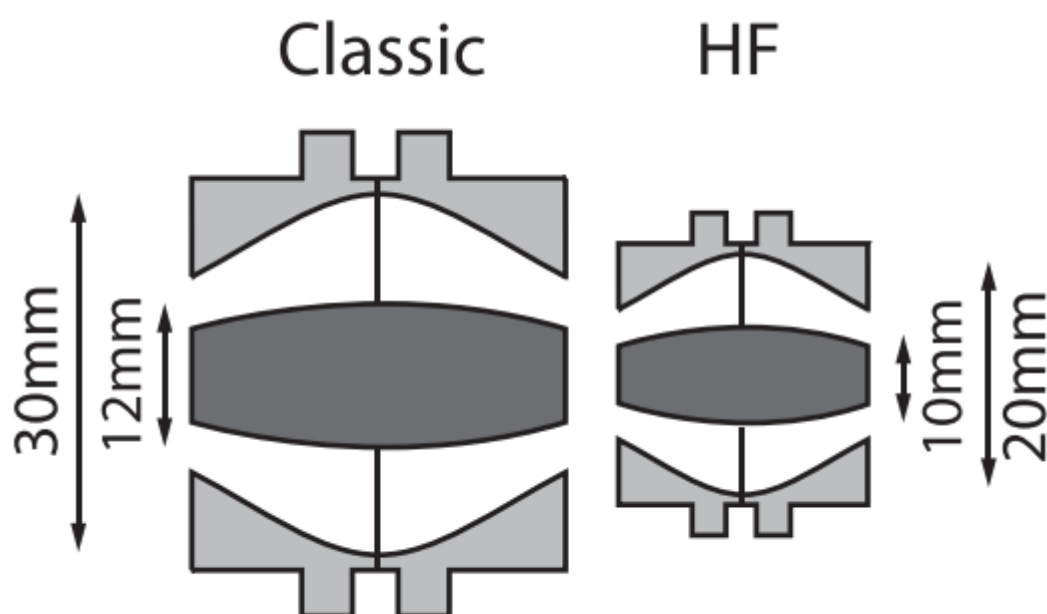


Figure 11: Comparison between the dimensions of the spindle electrode (R_1) and the outer electrodes (R_2) of the classic and UHF-Orbitrap (Scheltema et al., 2014). The reduction in size and modification of the $R_1:R_2$ allows a stronger electrostatic field which in turn allows higher mass resolution measurements to be recorded.

The frequencies of ions that are recorded are known as transients, the length of time that a transient is recorded for determines the resolution. The longer the time of the transient the higher the resolution (Table 3). The upgrades in Orbitrap technology discussed can thus be used in two different ways. By staying at the same transient length with a more powerful Orbitrap increased resolution will be achieved. Alternatively, you can decrease the transient length and still achieve the same resolution,

so a faster scan rate is achieved. The capabilities of different Orbitrap systems in terms of resolution and transient length covering the three different Orbitraps commercially available are outlined in

Table 4.

Table 3: Comparison between the resolution and transient length of the Q Exactive Plus (HF-Orbitrap detector) and the Q Exactive HF (UHF-Orbitrap detector), adapted from (Kelstrup et al., 2014). The improved UHF Orbitrap facilitates higher resolution at the same transient length compared to the HF-Orbitrap or can provide equal resolution at decreased transient length providing a decrease in the scan speed.

Transient Length (ms)	Q Exactive Plus Resolution at 200 m/z (HF Orbitrap)	Q Exactive HF Resolution at 200 m/z (UHF-Orbitrap)
32	NA	15,000
64	17,500	30,000
128	35,000	60,000
256	70,000	120,000
512	140,000	240,000

Table 4: Comparison of Orbitrap instruments and their key characteristics impacting resolution and transient length. Voltage applied to the spindle (kV), overall Orbitrap size (mm), application of enhanced Fourier transform or not, resolution achieved at 200 m/z , the transient length required to achieve that resolution, the resolution achieved per millisecond at 200 m/z . Adapted from (Kelstrup et al., 2014)

Instrument	Orbitrap central electrode voltage (kV)	Orbitrap analyser size (mm)	eFT	Specified resolution at 200 m/z	Transient Length (ms)	Resolution per millisecond at 200 m/z
LQ Orbitrap	3.5	30	No	140,000	1536	91
LQ Orbitrap XL	3.5	30	No	140,000	1536	91
Orbitrap Velos	3.5	30	No	140,000	1536	91
Orbitrap Velos Pro	3.5	30	No	140,000	1536	91
Orbitrap Elite	3.5	20	Yes	336,000	768	438
Exactive	5.0	30	No	100,000	700	143
Exactive Plus	5.0	30	Yes	140,000	512	273
Exactive EMR	5.0	30	Yes	140,000	512	273
Q Exactive	5.0	30	Yes	140,000	512	273
Q Exactive Plus	5.0	30	Yes	140,000	512	273
Q Exactive HF	5.0	20	Yes	240,000	512	469
Orbitrap Fusion	5.0	20	Yes	480,000	1024	469

Further to the hardware improvements performance was improved in both the HF-Orbitrap and UHF-Orbitrap by modification of the FT algorithm, this improved algorithm is known as enhanced-Fourier Transform (eFT) (Kelstrup et al., 2014). eFT is facilitated by the mechanism of ion injection into the Orbitrap. The ions are rapidly injected with the time it takes for an ion to travel the path length into the Orbitrap being dependent on its m/z (Lange et al., 2015). This subsequently means that the m/z and the phase of the harmonic axial oscillations are linked and this phase information can be used to improve the effectiveness of the algorithm. This enhancement confers a doubling in resolving power or scan speed at a particular resolution (Lange et al., 2015).

Various modifications have been made to different versions of the mass spectrometers in the Orbitrap series other than increasing electrostatic field strength and performance of the Orbitrap analyser itself. One of these was the introduction of the C-trap (Makarov et al., 2006). The C-trap is a modified storage quadrupole which is bent in the shape of the letter C. It is positioned adjacent to the Orbitrap mass analyser so that the ions can be injected from the centre of the C-trap through a small aperture in the same way as described before via quickly ramping the voltage. The C-trap decouples the Orbitrap from the rest of the system, providing pulsed injections of ions (Zubarev and Makarov, 2013) and storage of up to 1 million charges (Scheltema et al., 2014). This theoretically means any ion delivery/separation system can be attached before the C-trap for combination with high resolution Orbitrap analysis. It also means that 2 or more precursor ions can be stored at the same time due to the presence of the quadrupole as well (Scheltema et al., 2014) allowing more complex isolation, storage and fragmentation of ions. Another important addition has been the higher energy collision (HCD) cell. This allows HCD fragmentation data to be collected which is similar in its operation to a quadrupole collision cell observed in Q-TOF and QqQ mass spectrometers. In LTQ-Orbitrap and the newer Fusion systems, CID MS^n data can be collected generating complementary fragmentation data. MS^2 and MS^n data helps to increase the confidence in subsequent annotation accuracy. Both of these additions C-trap and HCD can be seen in the schematic of a Q Exactive Plus hybrid quadrupole mass spectrometer below (Figure 12).

In a Q Exactive system the ions are received from the source and initially are focused by the S-Lens. The injection flatapole is a low resolution variant of a typical quadrupole with flat rectangular shaped electrodes. It acts as an extra mass filter before the quadrupole, this helps to keep the segmented quadrupole clean and thus helps to maintain data quality (Scheltema et al., 2014). The bent flatapole is an elongated version of the injection flatapole with a curved shaped. Charged ions will be guided along the bent flatapole, due to interaction with the electric/RF fields applied. It filters out any neutral molecules as they will not be manipulated by the electric/RF fields and so should collide with the side

of the bent flatapole and be lost. The segmented quadrupole can be utilised to filter ions based on their mass before entry into the C-trap and subsequent injection into the Orbitrap. This quadrupole offers improved isolation performance compared to the linear ion trap (LIT) that was used for precursor selection before the release of the Velos and Elite systems (Scheltema et al., 2014). This allows for targeted and untargeted analyses to be carried out with greater precision. Ions are transferred from the segmented quadrupole to the C-trap where, as mentioned previously ions are stored and injected into the Orbitrap through a sudden ramping of the electric field. Alternatively, the ions can be shuttled into the HCD cell for fragmentation, fragments are transferred back to the C-trap and can then be injected into the Orbitrap. The Orbitrap itself offers high resolution capabilities (Up to 140,000 at 200 m/z) (

Table 4) for accurate mass analysis of the ions present.

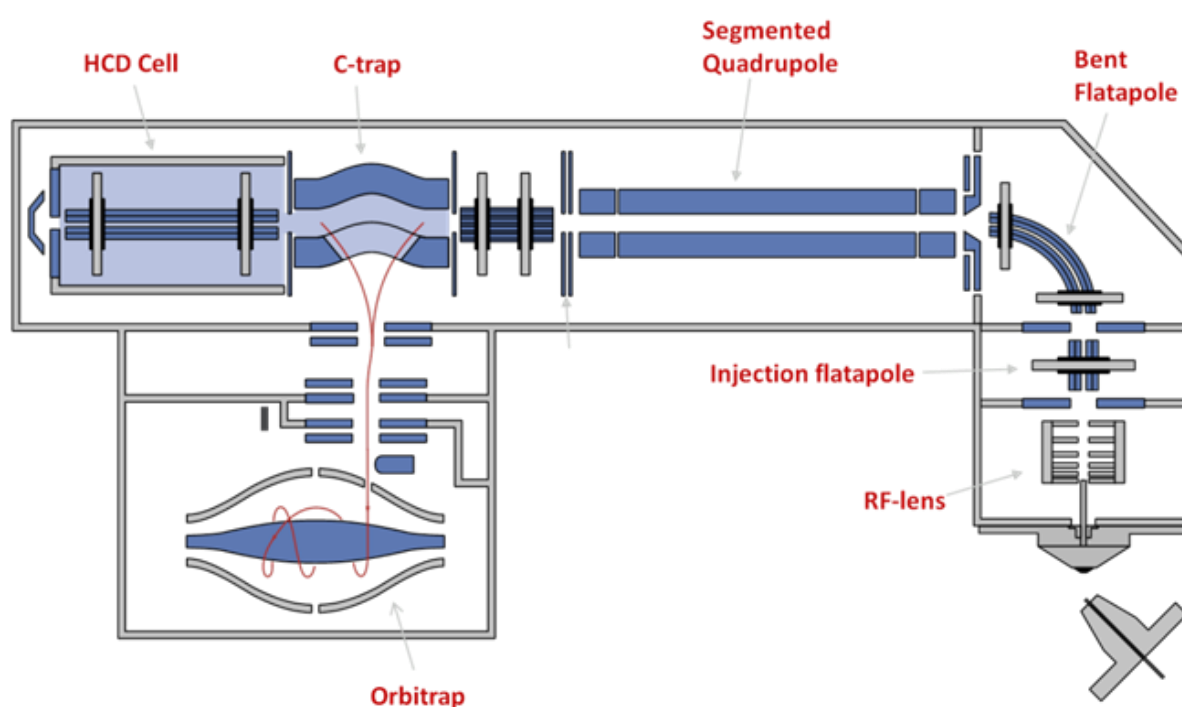


Figure 12: The schematic of a Q Exactive Plus mass spectrometer. Ions are focussed and guided by the RF-lens, the bent flatapole filters out neutral species, the segmented quadrupole filters for the specified m/z range, the C-trap then stores ions for injection into the Orbitrap. Ions can also be transferred from the C-trap to the HCD cell for HCD fragmentation before being returned to the C-trap for injection into the Orbitrap (Thermo Fisher Scientific, 2019).

In 2013, Thermo Fisher Scientific released the first in a new series of ‘tribrid’ mass spectrometers, the first of which is called the Orbitrap Fusion. The tribrid name was derived from the naming of the older hybrid instruments, so called for their dual configurations of LIT/Orbitrap or Quadrupole/Orbitrap.

The tribrid systems thus combine three types of mass analyser in a T shape configuration as seen in Figure 13, these are the quadrupole, UHF-Orbitrap and a dual pressure LIT (Senko et al., 2013). The T configuration ensures the distances and thus ion paths between each storage device and analyser is kept to a minimum helping to increase ion transfer efficiency. The addition of the LIT at the back end of the ion path allows the simultaneous collection of high resolution data in the Orbitrap and lower resolution data in the LIT. This parallelisation means more fragmentation data can be collected in a more efficient manner without compromising the quality of the MS¹ data. Higher-energy Collision Dissociation (HCD) fragmentation can be carried out in the ion-routing multipole (IRM), whilst Collision Induced Dissociation (CID) and Electron Transfer Dissociation (ETD) can be carried out in the back of the dual pressure LIT. These different fragmentation types can even be combined with the extra LIT addition allowing MSⁿ data to be collected to further aid in structural elucidation. This new configuration provides researchers with far more possibilities for fragmentation (MSⁿ), storage, transfer and analysis of their analytes than ever before in a single instrument.

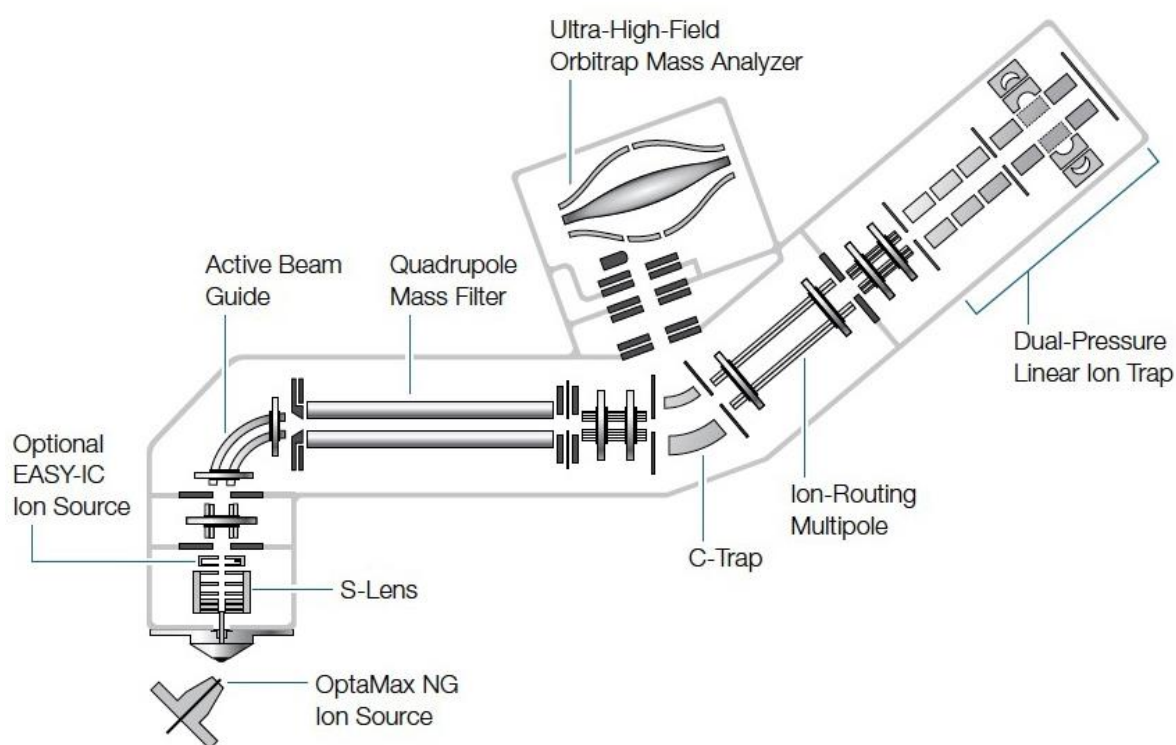


Figure 13: Schematic of an Orbitrap ID-X Tribrid mass spectrometer. The general layout and structure is the same as that seen in the Q Exactive Plus (Figure 12) but an additional dual-pressure linear ion trap in which ions can be stored and fragmented with CID and transferred back to the Orbitrap, or it can also be used to collect low resolution mass spectra in the trap itself. HCD is carried out in the ion-routing multipole, additional CID can be carried out in the dual-pressure LIT. EASY-IC has also been

added in the source to allow easy and automatic internal calibration throughout the run (Thermo Fisher Scientific, 2019).

A new tribrid mass spectrometer called the Orbitrap ID-X (Thermo Fisher Scientific, USA) was released in 2018 (Figure 13). This new system has the same basic structure as other tribrids and contains the UHF-Orbitrap like the other tribrids so is capable of very high resolution up to 480,000 (Table 4). The main difference between this system and the other tribrids is in the software which is capable of performing intelligent data acquisition in a way not seen before in any mass spectrometer. This will be discussed further in a later chapter. Continuous improvements in technology, instrumentation and software will continue to be made in the field of mass spectrometry increasing the experimental possibilities of researchers and industry.

1.3 Metabolomics

Metabolites are biochemicals found in biological systems, they act as regulators of systems and pathways, and their breakdown, construction and modification (metabolism) provide the resources needed for organisms to survive. Ultimately, they are the final output of the combined genetic, transcriptional and translational action of an organism alongside environmental and lifestyle influences. The study of the metabolome, defined as the identification and quantification of the complete collection of small compounds (typically < 1500 Da) found within a particular biological system is of vital importance if we wish to gain full understanding of that system. The metabolome was first defined in 1998 in relation to analysis of *Saccharomyces cerevisiae* and a whole organism analysis approach to studying the effects of genetic perturbations on a system (Oliver et al., 1998). Efforts to characterise metabolomes have increased rapidly since Nicholson (1999) first defined the field of study as metabonomics. Since then rapid advances in analytical and computational technology have facilitated the fast growth of the field which would have previously been impractical, it is now commonly referred to as metabolomics, the first mention of which was in 1999 as well (Dove, 1999). Upon integration with the other “-omics”, genomics, epigenomics, transcriptomics and proteomics we can theoretically characterise all components of biological systems allowing us to gain a complete global picture of that system for the first time. Global effects of perturbations on a system can be measured instead of looking at individual components or pathways in isolation, an approach which can lead to misleading results by not providing the whole picture. This is in contrast to the way in which the majority of science is done, with most science being hypothesis driven, targeted experiments (Kell and Oliver, 2004) but being able to use a global picture is a potentially powerful tool.

Metabolomics has been applied in a variety of applications and its use is set to expand greatly as the field and associated technologies are studied further and are refined and improved as new

technologies become available. In medicine it has been used as an effective tool for identification of biomarkers for a variety of serious conditions (Jansson et al., 2009; Sato et al., 2012; Osborn et al., 2013; Roede et al., 2013; Li et al., 2016c) and is also set to be integral in the burgeoning field of precision medicine (Beebe and Kennedy, 2016; Beger et al., 2016). As well as providing medical benefit it can aid metabolic engineering efforts through the study of the fine detail of the highly dynamic metabolic networks, (Johnson et al., 2016) contributing to improved biotechnological production, genetically modified microorganisms or novel crop varieties (Hill et al., 2015; Toya and Shimizu, 2013). Furthermore it can be used as a tool for discovering new beneficial natural products (Hufsky et al., 2014) such as plant secondary metabolites (Cottet et al., 2014) or as a key tool in environmental monitoring and toxicology studies (Kido Soule et al., 2015; Bundy et al., 2009). Another role for metabolomics has emerged in the field of food science and precision nutrition (Bayram and Gökırmaklı, 2018). This is just a small example of the current contexts in which metabolomics is applied.

Currently metabolomics studies are most typically carried out using UHPLC-MS (Ultra High Performance Liquid Chromatography-Mass Spectrometry) (Zhou et al., 2012), although GC-MS (Gas Chromatography-Mass Spectrometry) is widely applied (Koek et al., 2011), and NMR (Nuclear Magnetic Resonance) can also be applied (Markley et al., 2017). The early NMR based studies were defined specifically as metabonomics (Robertson, 2000) whilst metabolomics became used for other techniques. There are different advantages and disadvantages to implementing these different platforms. Some techniques are more sensitive than others whilst some are more suited to detecting a particular class of compounds (Table 5).

Table 5: Advantages and disadvantages of the three primary analytical techniques for metabolomics.

Technique	Advantages	Disadvantages
LC-MS	<ul style="list-style-type: none"> • High Sensitivity • Detects wide range of compounds 	<ul style="list-style-type: none"> • Low reproducibility between labs • Data complexity
GC-MS	<ul style="list-style-type: none"> • Good for analysis of volatiles • RT indices • Reproducibility 	<ul style="list-style-type: none"> • Many analytes require derivatisation
NMR	<ul style="list-style-type: none"> • Structural elucidation 	<ul style="list-style-type: none"> • Low sensitivity

The reason UHPLC-MS is most commonly applied is due to its high sensitivity and applicability for detection of a wide variety of compounds from water-soluble metabolites to lipids (Zhou et al., 2012).

This is particularly relevant when discussing untargeted metabolomics as its objective is to detect all metabolites in a sample. Although, this is an unrealistic goal as some metabolites do not ionise well and so cannot be detected or are present at very low levels below the limit of detection. The aim is to detect as many as is possible with the current technology. As a result of this, untargeted metabolomics is considered a hypothesis generating discipline, often researchers carry out untargeted metabolomics as a screening method. In a relatively fast manner, a vast amount of data can be gathered on the compounds in a system to subsequently provide a shortlist of dysregulated compounds for which hypotheses can be generated. These hypotheses can then be put to the test using a targeted metabolomics strategy. This is where the entire analytical chemistry set up and sample preparation is tailored to the physicochemical characteristics of the metabolite or metabolites of interest. A targeted study should be able to provide accurate quantitation of that compound and allow validation or invalidation of the hypothesis. Untargeted studies contain relative quantitative information i.e. Each detected metabolite has an intensity associated with it but due to the high number of variables and complexity of the data this is not considered as absolute quantification unlike in a targeted study. The focus of this thesis however is on untargeted metabolomics experiments and so in the next section a thorough description of a typical untargeted metabolomics workflow and the associated issues will be discussed.

1.3.1 The Untargeted Metabolomics Workflow

The untargeted metabolomics workflow (Figure 14) can be broadly characterised into the following stages:

1. Sample collection and preparation.
2. Analytical chemistry data collection.
3. Data processing.
4. Data analysis.
 - a. Univariate and multivariate statistics
 - b. Metabolite annotation/identification
 - c. Metabolic Pathway Analysis

Each of these stages will be discussed in turn with particular focus on areas relevant to the research carried out; these are MS¹ data complexity, MS² acquisition strategies and metabolite annotation.

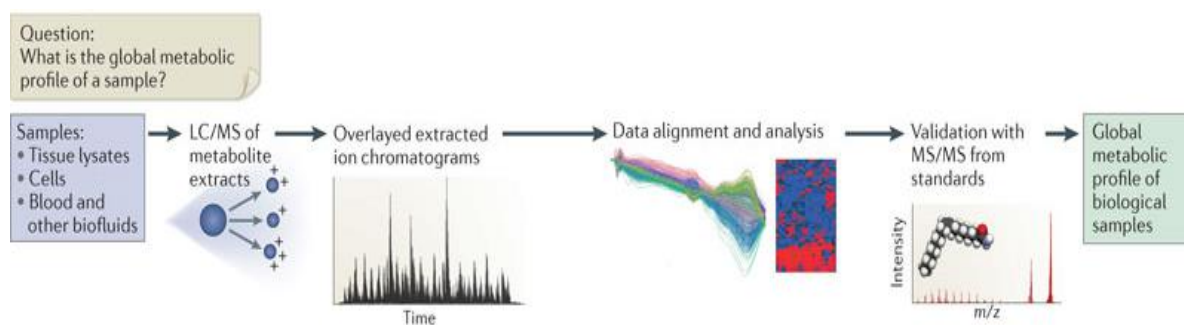


Figure 14: The untargeted metabolomics workflow. A sample is analysed using LC-MS or UHPLC-MS, raw data is processed and analysed before metabolite identification is carried out to provide a global metabolic profile of the sample (Patti et al., 2012).

1.3.1.1 Sample Collection and Preparation

Sample collection, transport and storage should be carefully planned and controlled to limit technical variation in the data later on. Sample preparation will be different dependent on the sample type (e.g. human biofluid vs. tissue) and the desired subsequent analysis (LC-MS vs GC-MS). A common strategy for preparation of liquid samples due its simplicity and suitability to high throughput analysis is to employ a “dilute and shoot” technique where the sample (typically a biofluid such as plasma, serum, urine or cerebrospinal fluid (CSF)) is diluted in an organic solvent such as acetonitrile or methanol, commonly a 1:3 excess of solvent is applied. This acts to precipitate proteins which could cause blockages and damage in the LC system and also be detected in the mass spectrometer causing analytical interferences for the detection of metabolites. Vortex mixing and centrifugation of the sample is performed with the supernatant being extracted and applied for analysis. Other more complex sample preparation may be required such as when a tissue is being studied instead of a biofluid. In this case homogenisation and cell lysis of the tissue must be carried out before the dilution, protein precipitation and centrifugation. Other possible steps include fractionation, pre-concentration or derivatisation (Gong et al., 2017) but these are not commonly applied in untargeted studies and are more likely to be utilised in targeted methods where the sample preparation will be tailored to enrich the metabolites of interest. The user may wish to perform a biphasic extraction, this is where a combination of immiscible solvents such as water, methanol and chloroform (Folch et al., 1957; Bligh and Dyer, 1959) are used simultaneously to create separate hydrophilic and hydrophobic phases, other solvent combinations can be applied. Samples are extracted from the two phases separately creating two samples and thus increasing analysis time but this has been shown to improve coverage compared to a monophasic extraction (Broeckling and Prenni, 2018). This is a particularly useful technique for researchers looking to enrich lipids for study of the lipidome (Matyash et al., 2008) or other lipid rich sample types such as the sphingolipidome (Han et al., 2011). Sphingolipids are

important biological lipid molecules with roles in cellular signalling, regulation and interaction with important roles in diseases such as Alzheimer's (Merrill et al., 2009)

1.3.1.2 Analytical Chemistry Data Collection

Analytical data acquisition can be performed using a range of analytical platforms. UHPLC-MS is the most frequently applied in untargeted metabolomics due to its relatively high sensitivity (compounds can be detected at pg.mL^{-1} levels (Theodoridis et al., 2012)) and its ability to detect a wide variety of metabolites simultaneously (Zhou et al., 2012). Separation of the sample type before entry into the mass spectrometer is important as it allows a reduction in ion suppression. This is a phenomenon where competition for ionisation in the source between co-eluting compounds or components of the sample matrix results in decreased signal intensity for the less competitive compound (Annesley, 2003). Reduction of ion suppression permits more metabolites to be detected. Additionally, the retention time (RT) recorded for each metabolite during the run provides an orthogonal property to increase confidence in the identification of metabolites. The more sources of orthogonal data that can be gathered the greater the chance of a confident identification. Levels of confidence in identification are very important in untargeted metabolomics and will be discussed in more detail later (1.3.1.4.2.1). One of the major challenges associated with untargeted metabolomics is metabolite annotation and this is in part due to the massive inherent chemical diversity of compounds in nature but also of synthetic drugs. This means no one analytical platform is suitable for detection of all metabolites (Dunn et al., 2013). To ensure the widest metabolite coverage a reversed phase C_{18} analytical column is normally used as it is suitable for detecting a relatively wide spectrum of semi-polar and non-polar compounds. For better detection and separation of polar compounds HILIC (hydrophilic interaction chromatography) can be used (Creek et al., 2011). For more hydrophobic compounds a reversed phase C_{30} column for example can provide increased chromatographic separation for a range of lipid classes (Houjou et al., 2005). Using these different stationary phase chemistries can ensure separation of highly polar or highly non-polar compounds that would have eluted together in the void volume towards the start or end of the analysis depending on the column chemistry utilised. Two separation chemistries can even be used in tandem in 2D-LC to provide two dimensions of separation back to back. This method is not widely applied currently due to difficulties such as solvent compatibility between separate assays (Rampler et al., 2018) but can aid identification (Sun et al., 2014). This is achieved by providing more orthogonal data as well as increasing separation and thus decreasing ion suppression in the MS improving data quality. A variety of other separation types can also be coupled to MS analysis and have their advantages for detecting certain types of compounds as described earlier (1.2.1). Following separation the sample is ionised to allow detection in the mass spectrometer and mass analysis is carried out.

Different types of data can be collected in a single UHPLC-MS analysis, these will typically always include RT and m/z from a full scan MS^1 analysis but may also include collision-induced-dissociation (CID) MS^2 , higher-energy-collision-dissociation (HCD) MS^2 and MS^n . Each of these sources of data can provide orthogonal information for increased confidence in annotation and will be discussed. The complexity of MS^1 data and its importance for metabolite annotation will be discussed in detail later (1.3.1.4.2.2) as well as MS^2 fragmentation data, acquisition strategies and their importance for metabolite annotation (1.3.1.4.2.3 MS^2 Data and Strategies for Acquisition and Analysis as they are particular focuses of this thesis.

The samples of interest in a biological study should be analysed alongside quality assurance (QA) and quality control (QC) samples. QA samples may be collected although they are not always applied and there is no consensus guidelines in the community for how they should be implemented (Broadhurst et al., 2018). Their purpose is to monitor inter-study performance and variation, for example measuring the response of the same standard sample at the beginning of each different analytical run to confirm whether the instrument performance is within specification. A full review of QA processes can be found in Dudzik *et al*, (2018). QCs are widely applied and have multiple forms and uses although primarily they are for monitoring intra-study variation and instrument performance. QC samples can be prepared in a number of different ways but are intended to be representative of the samples being analysed and so are typically prepared from a pool of the other samples in the study (Sangster et al., 2006). This might not always be a practical approach to take however, for example, if the sample is not a biofluid, or if there is extremely limited sample volume. The most appropriate QC preparation will be different from study to study (Broadhurst et al., 2018). They are run at the beginning and end of the analytical run as well as at regular intervals throughout the run. If the QCs and biological samples are plotted together using principal component analysis (PCA) (Discussed later in section 1.3.1.4.1.3) then the QCs should cluster tightly together indicating good quality data. Any drifting that might have occurred during the run such as RT drift or mass accuracy changes can be identified too. This information can then be used to perform post-acquisition correction (Dunn et al., 2011) of the data if necessary or to indicate that the data quality is too poor to use and the analysis must be repeated. The number of the QCs injected throughout the study and the interval between them are important for the efficacy of this strategy, a minimum of one QC every 10 samples is recommended (Broadhurst et al., 2018). The QC data may also be used to filter the results, for example peaks may be removed if they were not detected in at least 50% of the QC samples in the study with a relative standard deviation (RSD) $\leq 20\%$ (Godzien et al., 2015). The values selected here are arbitrary and the user should look to define appropriate values of their own. As well as QA and QC samples blanks are utilised. There are two main types of blank, an extraction/process blank and solvent/system blank. An

extraction/process blank is a blank sample that undergoes the exact same sample preparation procedure as a normal biological sample would have undergone, the solvent/system blank is run without a sample injection so only components of the solvent reservoirs and sample matrix are measured. This blank sample can be used to filter the data through what is known as blank subtraction. This is where any features detected which exceed a user determined ratio threshold of mean feature intensity within the blank samples compared to the mean feature intensity within the biological samples are removed. This threshold may vary depending upon the experimental conditions and technologies applied however a blank:sample ratio between 2 and 5 for sample filtering is common (Schiffman et al., 2019). A blank sample can also be used for the generation of an exclusion list to avoid MS² fragmentation of non-biological sample components (Broadhurst et al., 2018). Blanks are also used to check for “carry over” and dirtying of the system throughout the run, this is when signals from previous biological samples are detected in the subsequent analysis due to insufficient washing of the column between sample injections. Ultimately, QA, QC and blank samples are important to monitor data accuracy, precision and quality as well as facilitating post data acquisition correction and removal of non-biological features from further analysis. A detailed review of QA and QC processes can be found in (Broadhurst et al., 2018).

1.3.1.3 Raw Data Processing

Following collection the raw data must be processed to generate a data matrix for further univariate or multivariate analysis. This involves a number of different steps including file format conversion, peak picking, grouping of peaks and m/z and retention time alignment. The peak matrix contains information on all of the “features” within the sample. A feature is defined as a m/z (mass over charge ratio) and retention time (RT) pair. Associated with each feature will be its recorded intensity across each sample in the study. There are a host of options both commercial and open source to a researcher wishing to process their untargeted metabolomics data. A file conversion may be required depending on the chosen platform for data analysis. If using a vendor’s own software such as Compound Discoverer 3.0 (Thermo Fisher Scientific) then conversion may not be required. If using an open source software then the file should be converted to a commonly used open source file format such as .mzML beforehand. Conversion from vendor format to .mzML is often carried out using the freely available Proteowizard software (Kessner et al., 2008). After conversion the data is processed. The most popular open source software include XCMS (Smith et al., 2006), MzMine2 (Pluskal et al., 2010) and MetAlign (Lommen, 2009). Other commercial software include Compound Discoverer (Thermo Fisher Scientific, USA) and Progenesis QI (Waters Corporation, USA). These software carry out each stage from peak picking, grouping and alignment to the end result of a peak matrix. However, each piece of software is written with different algorithms, and may operate in different ways; this means the same dataset

processed with different software will produce different results (Gürdeniz et al., 2012; Coble and Fraga, 2014; Rafiei and Sleno, 2014; Myers et al., 2017). Even with a single piece of software highly different results can be seen when utilising different processing parameters. This has resulted in software such as IPO (Isotopologue Parameter Optimisation) software (Libiseller et al., 2015) which will perform optimisation of XCMS peak picking parameters when supplied with the raw data to try to ensure more accurate peak picking. Although there will be differences between the different peak picking software options available, each software broadly does the same thing. First, extracted ion chromatograms (EIC) are plotted from the raw data and peaks are extracted. Secondly the extracted peaks must be grouped together where appropriate (sometimes referred to as binning); this is done due to the fact that a single metabolite will be detected as multiple different features within the analysis. This is another of the major challenges associated with untargeted metabolomics and will be discussed in detail later (1.3.1.4.2.2). Correct grouping of these features is of great importance to help avoid spurious biological interpretation through the presence of false positives and false negative metabolite annotations at the end of the study. There are a number of dedicated grouping tools which perform further grouping on the final peak matrix and these as well as the different feature types. After grouping, alignment is carried out, this is where a feature which has been detected in multiple samples in the study has a slightly naturally shifted retention time or m/z across the run. These features can be aligned in the processing so they are not considered separate peaks in the resulting peak matrix. Another round of grouping may be performed which can now consider the newly aligned peaks and following this the peak matrix is constructed. There is then the option to perform gap filling, this is where missing peaks are imputed into some of the samples based on the peaks seen across samples and sample groups in the dataset. Once the peak matrix has been finalised then data analysis can begin. It is at this stage where blank subtraction may be performed if blank samples were collected to ensure the data analysis is focussed on features of biological origin.

1.3.1.4 Data Analysis

1.3.1.4.1 Statistics

Once the peak matrix has been generated then statistical analysis can be performed. Statistical tests can be parametric or non-parametric. A parametric test compares the means between sets of data. Non-parametric tests compare medians normally converting the data into a ranked format. Parametric tests make a number of assumptions about the data. These include that the data is normally distributed, there is homogeneity in the variances and data are independent of one another. If these are not assumed they may be accounted for through implementation of transformation and scaling, or alternatively a non-parametric test which does not make these assumptions can be applied (Delacre et al., 2019). A combination of univariate (e.g. students t -test (parametric), Mann-Whitney

test (non-parametric)), bivariate (e.g. Pearson Correlation coefficient (parametric), Spearman's rank (non-parametric)), unsupervised multivariate statistics (e.g. principal components analysis (PCA), and supervised multivariate statistics (Partial Least Squares Discriminant Analysis (PLS-DA)) are typically performed and these will be discussed in turn. To prepare the data for these analyses a number of steps may or may not be required, these include missing value imputation (MVI) (Discussed later in section $r_{xy} = \text{Pearson correlation coefficient between } x \text{ and } y$

$n = \text{The number of observations}$

$x_i = \text{Value of the } i\text{th observation of } x$

$y_i = \text{Value of the } i\text{th observation of } y$

1.3.1.4.1.2 Multivariate Statistics, normalisation, transformation and scaling. Univariate and Multivariate statistics and the necessary data pre-treatment steps will be discussed in the following sections.

1.3.1.4.1.1 Univariate Statistics

Univariate statistics such as the t -test are often applied to determine if two features are significantly dysregulated between two groups or samples. . For univariate tests normalisation of the peak matrix is recommended but no missing value imputation (MVI), transformation or scaling should be implemented (Di Guida et al., 2016). Normalisation is important to stop the highest intensity features dominating the results of any analyses performed and is particularly important in metabolomics due to the variation in feature intensities reported in any study typically spanning multiple orders of magnitude (Dunn et al., 2008). There are a wide variety of different methods available, if internal standards are utilised then they can be used to perform the normalisation (De Livera et al., 2012) or other inherent components of a sample type can be utilised such as creatinine in urine (Vollmar et al., 2019). Li et al. (2016) compared 16 different normalisation methods and found variance stabilisation normalisation (VSN) (Lin et al., 2008), probabilistic quotient normalisation (PQN) (Dieterle et al., 2006) and generalised-log (g -log) transformation (Purohit et al., 2004) to be the best performing options across a variety of different sized datasets. This is backed up by Di Guida et al., (2016) who recommend probabilistic quotient normalisation (PQN) for best performance of univariate statistical tests.

1.3.1.4.1.2 Bivariate Statistics

Bivariate statistics such as the Pearson correlation coefficient are not commonly applied in metabolomics for the statistical analysis of the final peak matrix. They are however commonly applied in the intermediary stages of the final matrix preparation through looking at the relationships between different metabolite features. This is done to help determine if two different features require grouping

together due to being derived from the same metabolite. This phenomenon will be discussed in detail later on (1.3.1.4.2.2 MS¹ and ESI data complexity and degenerate features). Related features derived from a single metabolite should have a level of analytical response that is similar across all the samples of a study. It is thus logical to perform bivariate analysis of feature pairs which may be related, this can be carried out in the annotation workflow generating a correlation coefficient score for each pair of features. A score of +1.0 indicates the strongest possible positive relationship between the two variables, a score of -1.0 indicates the strongest possible negative relationship between the two variables whilst a score of 0 indicates that there is no relationship between the two variables. There are many different methods of calculating a correlation coefficient including Pearson, Spearman's rank, Kendall rank and Point-biserial (Philip Bobko, 2001). The most widely applied however is Pearson correlation (Equation 5). Although the equation is named after Karl Pearson following his work published in 1895 (Pearson, 1895), it is based on work done previously by Francis Galton in the 1880s (Galton, 1886) which in turn was based on a formula developed by Auguste Bravais in the 1840s (Bravais, 1844). Essentially the formula assigns a value to how tightly the two continuous variables form a straight line if plotted into a scattergraph. It is a parametric test, assuming a normal distribution of data for the variables in question and it can be sensitive to outliers. If there are many outliers the results may be misleading. Removal of outliers can be beneficial, as well as comparing the results without the outliers included to the results with the outliers included. Use of an alternative test which employs a ranking system for the recorded data such as Spearman's rank or Kendall rank can overcome this problem (Philip Bobko, 2001). These are both non-parametric tests and so fewer assumptions are made about the structure of the data however by converting the true data to ranks some of the fine detail of the relationship between the variables is lost. Point-biserial correlation is similar to Pearson correlation however one of the two variables must be binary. For comparison of two continuous quantitative variables such as the intensities of two features in a mass spectrometry experiment Pearson correlation is the best option. A p-value is calculated to indicate the significance of the result and to help determine if the null hypothesis can be rejected or not. It provides an indication as to whether the correlation coefficient recorded could have occurred due to chance or whether it is likely to be deemed a statistically significant score instead. A p-value of less than or equal to 0.05 is typically considered as the threshold for a statistically significant result (Philip Bobko, 2001).

Equation 5: The Pearson correlation coefficient formula.

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson correlation coefficient between x and y

n = The number of observations

x_i = Value of the i th observation of x

y_i = Value of the i th observation of y

1.3.1.4.1.2 Multivariate Statistics

Multivariate tests are very important in untargeted metabolomics due to the high number of variables that are measured in a typical experiment. They are utilised to convert this high-dimensional data into easily interpretable 2D graphical representations of the variation within the study. Multivariate tests can determine which samples within the study are similar in composition and which are more dissimilar. The specific metabolite features that are most important for driving these differences can also be elucidated. This allows the user to identify the most important metabolite features to focus upon for further targeted analyses. Multivariate tests can be considered in two categories, supervised and unsupervised. Unsupervised tests normally comprise one of the first stages of data analysis as they offer an unbiased exploration of the data, PCA is a commonly applied example (Leal-Witt et al., 2017; Housley et al., 2018; Baptista et al., 2018). Other unsupervised methods include Hierarchical clustering (Draisma et al., 2013), K-means clustering (Torras-Claveria et al., 2014) and self-organising maps (Haddad et al., 2009). A supervised test on the other hand has prior knowledge of the sample groups and will seek to differentiate the sample groups as they were assigned, PLS-DA is a commonly applied example (García et al., 2018; Tian et al., 2018; Beale et al., 2019). Other examples include Support Vector Machines (SVM) (Mahadevan et al., 2008), Random Forests (RF) (Chen et al., 2013) and Principal Component Discriminant Function Analysis (PC-DFA) (William Allwood et al., 2010) Artificial Neural Networks (ANN) (Goodacre and Kell, 1996), Bayesian Network (BN) (Wang et al., 2013), Genetic algorithm-Bayesian Network (GA-BN) (Correa and Goodacre, 2011), Genetic Programming (GP) (Hu et al., 2018), Kernel Partial Least Squares (KPLS) (Cowcher et al., 2013), Multiblock Principal Component Analysis (MBPCA) (Xu and Goodacre, 2012), Multi Block Partial Least Squares (MBPLS) (Xu et al., 2013), Multilevel Partial Least Squares Discriminant Analysis (ML-PLSDA) (Westerhuis et al., 2010), Orthogonal Partial Least Squares Analysis (OPLS) (Westerhuis et al., 2010), Parallel Factor Analysis (PARAFAC) (Humston et al., 2011) and Partial Least Squares regression (PLSR) (Vaughan et al., 2012). When applying PLS-DA or other supervised methods it is important that the model is validated (Broadhurst and Kell, 2007; Szymańska et al., 2012; Saccenti et al., 2014). This is because supervised methods can overfit the data to the model leading to the strong likelihood of false positive results (Type I Errors). Multiple testing correction is performed to account for this, Bonferroni correction and Benjamini-Hochberg correction (Hochberg and Benjamini, 1990) are the most commonly applied methods. The Bonferroni correction is very rigorous and so is likely to lead to some false negative results (Type II Errors). Benjamini-Hochberg is a more lenient method. Any method

which looks at a very high number of variables such as metabolomics where there are hundreds or thousands of features should require multiple testing correction on the multivariate results. The very high numbers of variables can mean there are many uninformative variables too which can act to mask true positives in the data (i.e. biomarkers). Before applying any multivariate analyses whether supervised or unsupervised a number of data pre-treatment steps are required.

The first step in applying a multivariate test is to perform MVI. Missing values are a natural part of a metabolomics study and multivariate tests are not compatible with the large numbers of missing values typically found (Di Guida et al., 2016). As a result, missing values are predicted and imputed computationally into the dataset based on the data already present to allow multivariate analysis. A missing value is where a feature which is present in some samples and thus is expected in the others but is not found. This could be for any number of reasons, the metabolite may not have been present, it may have been present but at an insufficient concentration to allow detection or it may have been incorrectly omitted during the peak picking process. Missing values are considered to come in three types, missing at random (MAR), missing not at random (MNAR) and missing completely at random (MCAR) and there are at least 8 different algorithms available (Wei et al., 2018) to impute them. The best method will be dependent on the particular multivariate test that is applied but the random forest (RF) algorithm is recommended for principal component analysis (PCA) (Di Guida et al., 2016; Wei et al., 2018) whilst K Nearest Neighbour (KNN) imputation is recommended for partial least squares discriminant analysis (PLS-DA) (Di Guida et al., 2016). Following missing value imputation the data will also require normalisation, scaling and transformation too (van den Berg et al., 2006). PQN normalisation is recommended for both PCA and PLS-DA (Di Guida et al., 2016). Following normalisation scaling is performed upon the individual variables (features in the study). Each feature has a different scaling factor associated with it which is used to reduce the very large differences that will exist between the relative fold changes of high and low intensity features. An important consideration here is the amplification of any technical instrument variation when the scaling is applied to the lower intensity features that were closer to the limit of detection (van den Berg et al., 2006). Scaling methods available include autoscaling, pareto scaling, range scaling, vast scaling and level scaling. Scaling is not recommended for PCA or PLS-DA (Di Guida et al., 2016). Transformation of the data may also be applied, this can be utilised to remove any heteroscedasticity from the data, generate symmetrical data from asymmetric data and to allow identification of multiplicative biological effects (van den Berg et al., 2006). Generalised logarithm (*g*-log) transformation is recommended for both PCA and PLS-DA (Di Guida et al., 2016) although natural logarithm (*n*-log) can also be applied.

PCA and PLS-DA are by far the most commonly applied multivariate methods although other methods may be more appropriate and perhaps should be more commonly applied (Gromski et al., 2015). Ultimately whichever method is applied it is important that the user understands the assumptions that the test makes about the data and whether these are appropriate. Each test makes certain assumptions and to generate accurate and reliable results which are correctly interpreted understanding of your data characteristics and the test assumptions are essential (Saccenti et al., 2014; Gromski et al., 2015). It is always important to understand the strengths, limitations and assumptions of your statistical methods and how each step has modified the data landscape. Some false positives or false negatives are always likely to occur but by understanding the mathematical transformations and the data structure these can be minimised. Statistical analysis will hopefully have elucidated some features of interest for further investigation. The next step can then be biological interpretation providing the features of interest have been identified or annotated. But how exactly is annotation/identification carried out?

1.3.1.4.2 Metabolite annotation and identification

1.3.1.4.2.1 The Levels and Challenges of Identification

Before discussing the identification of compounds in metabolomics it is important to highlight the difference between the terms annotation and identification. Identification is reserved for an absolutely unambiguous identification applying two orthogonal properties and matching of these data to a chemical standard analysed using the same analytical chemistry method. Annotation describes a process where sufficient evidence to make the annotation unambiguous (identification) is not available. Levels of confidence in annotation and identification are very important in untargeted metabolomics. In 2007 the Metabolomics Standards Initiative (MSI) published a four tiered system for labelling annotations with the results of an untargeted study (Sumner et al., 2007) (Table 6). For a level 1 identification the compound must be compared using two or more orthogonal features to data collected from a pure authentic standard of the same compound using the same analytical chemistry method. This is difficult to achieve as pure authentic chemical standards are expensive to acquire, and for many metabolites, standards are simply not available commercially.

Table 6: The Metabolomics Standards Initiative (MSI) four levels of identification.

Level	Requirements
1	An unambiguous identification using 2 or more orthogonal properties of data collected compared to data for the standard collected in the same lab using the same analytical set up.
2	A putative annotation of a compound based on the physicochemical properties of the compound displayed in the data.
3	A putative annotation to a class of compounds based on the physicochemical properties of the compound displayed in the data.
4	Unknowns

Updates were proposed to the levels of identification by Schymanski *et al*, in 2014 to better represent the complexity of the issue with a focus only on mass spectrometry data collection (Figure 15). Level 1 remained the same but Level 2 was divided into 2a and 2b, where 2a is an unambiguous match to data from a MS² mass spectral library, whilst 2b means an unambiguous identification could be deduced from the data but no data for a reference standard is available to validate the result. A level 3 identification is an ambiguous assignment of a compound name or class. Level 4 is where an unambiguous chemical formula could be assigned but no further data is available and, a fifth level was introduced for compounds where the accurate mass has been recorded but no chemical formula has been assigned.

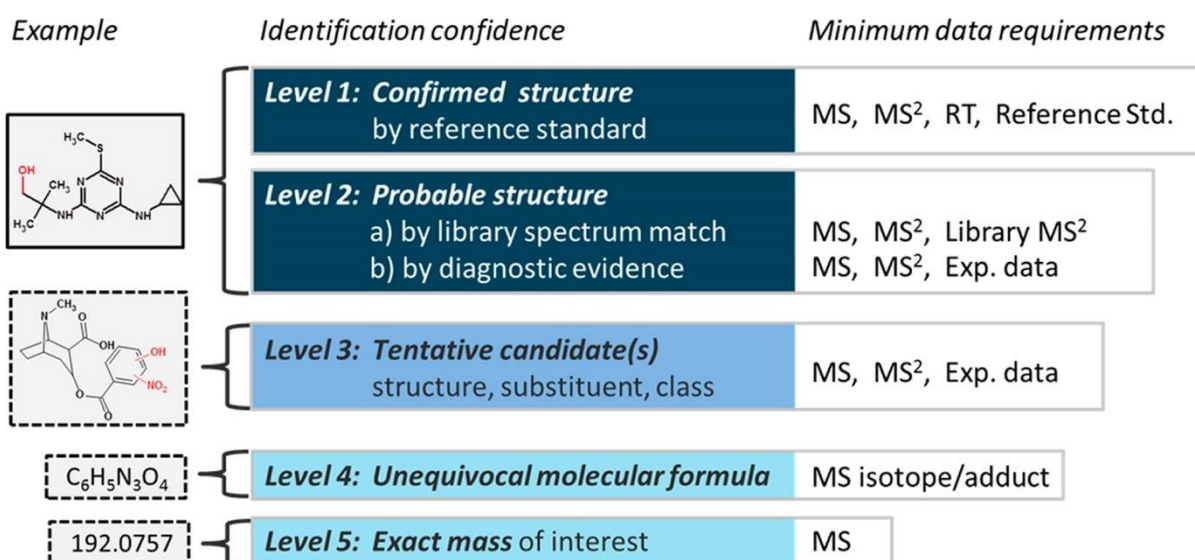


Figure 15: Schymanski's updated levels of identification (Schymanski et al., 2014). Level 1 identification requires an MS² spectral match to a pure authentic reference standard with the same analytical set up. A level 2 identification requires an MS² spectral match to an external library MS² spectrum. Level 3 requires that the MS² spectrum has revealed a unique substructure indicative of a certain class of compounds. Level 4 identification requires that the MS¹ data has provided an exclusive molecular formula assignment likely isotopic ratios and multiple adducts. Level 5 is an accurate mass with no further information.

The unambiguous identification of metabolites is one of the greatest challenges within the field of metabolomics, and the difficulties associated with this process are one of the primary reasons why metabolomics is not currently more widespread and well known. Without confident identification of metabolites, no significant biological knowledge can be gained. Therefore, for the field to gain greater traction it must overcome the challenges presented in metabolite identification.

Let us consider one of the more complex metabolomes, that of a human. The main issue behind the difficulty in annotation is the inherent chemical diversity present in nature and which humans are exposed to such as food and the air we breathe. This is then compounded further by the presence of synthetic pharmaceutical drugs and chemicals which also require consideration. Biological compounds that are endogenously produced by an organism are known as the endogenous metabolome. Synthetic drug compounds and their breakdown components although from an exogenous source are still typically considered part of the human metabolome as they can be expected to be detected but they also make up part of the exposome. The exposome was first defined by Wild in 2005 where he highlighted its importance in the context of large cohort genomics studies and how the combined effect of environmental exposures suffered by participants contribute significantly to development of

some chronic diseases. The exposome is very large and not well characterised and this needs to change to help improve metabolite annotation.

With such great chemical diversity present, there is not one analytical technique that is capable of detecting all metabolites present. This means multiple analytical platforms must be applied to gain complete biological knowledge of an organism (Dunn and Ellis, 2005). Furthermore, there is no database defining the complete contents of any metabolome. Although there are specialised databases for certain organisms which may be very useful it is unclear exactly how complete they are. Efforts have been made to try to gain complete knowledge of the expected parts through application of genome-scale metabolic network reconstructions such as in yeast (Herrgård et al., 2008) and humans (Swainston et al., 2016) to predict all endogenous metabolites. However, this method is only as reliable as our knowledge of gene function and metabolic pathways and so is not suitable to discover the complete expected parts list as there are many genes of unknown function and there may be metabolic pathways that have not been discovered yet. Furthermore some enzymes have also been shown to be promiscuous and may bind and act upon metabolites they are not normally expected to under certain conditions potentially creating previously undiscovered metabolites (Jeffries et al., 2015). Epimetabolites have also been recently discovered and add further complexity (Showalter et al., 2017) whilst other endogenous metabolites have been shown to be highly prone to non-enzymatic damage (Lerma-Ortiz et al., 2016). This complexity of metabolites, interactions and pathways contribute to the high dynamism and variability exhibited by the metabolome and help to explain why its analysis and study is such a complicated issue. With such diversity present it is not surprising that there are many isobars and isomers present and differentiation of these is the cornerstone of all the challenges discussed. This is particularly true for lipids where there is great similarity within great diversity with many lipids only varying slightly from many others present in terms of carbon chain saturation, length, double bond positions or stereochemistry for example (Rustam and Reid, 2018). This make lipidomics a particularly difficult challenge in metabolite annotation. Next the annotation of untargeted metabolomics data from the MS¹ data will be discussed and the challenges associated with this part of the annotation pipeline will be discussed in detail.

1.3.1.4.2.2 MS¹ and ESI data complexity and degenerate features

Annotation of metabolomics data from MS¹ information is confronted by a major initial hurdle, that a single metabolite will be detected in multiple different forms. These are presented by different adducts, isotopes, dimers, oligomers, multiply charged species, artefacts, in-source fragmentations and biological transformations (Brown et al., 2009). These alternative ion types can be referred to as degenerate features and add further complexity to the data and results in the possibility of false

positive annotations and subsequent incorrect biological interpretation of results. This is because the process of ionisation in the ESI source is highly complex and not well understood. A metabolite can be detected in as many as 100 different forms (Mahieu et al., 2016). This means accurate and efficient grouping of metabolites can result in major reductions in feature numbers as high as 90% (Mahieu and Patti, 2017). It is generally accepted that if operating in positive ion mode the dominant adduct type formed will be the protonated ion, whilst in negative mode it will be the deprotonated ion. Whilst there are a number of other common ion types expected including $[\text{Na}]^+$, $[\text{NH}_4]^+$ and $[\text{K}]^+$ in positive mode.

Naturally occurring isotopes will be present in the data, the most common of these will be the ^{13}C isotope which has a natural abundance of 1.109%. As a result, we expect to see ^{13}C isotopic peaks for any carbon containing compound with good signal to noise ratio. The identity of the peak can be confirmed as a ^{13}C peak by confirming the ^{12}C - ^{13}C mass difference of 1.003355 and looking at the intensity ratios between the regular ^{12}C peak and the smaller ^{13}C (normally less than 5%) peak adjacent to it (Brown et al., 2009). ^{13}C isotopes should be considered by any good piece of annotation software. However, the number of other possible isotopes considered varies between each piece of software. There are many other naturally occurring isotopes that could be present such as ^{15}N , ^{37}Cl , and ^{34}S amongst others that should be considered and accounted for during the feature grouping. The likelihood of their presence depends on the natural isotopic abundance and the number of atoms of that element present in the metabolite in question. Many software only consider the ^{13}C isotopic difference.

Dimerisation or higher order oligomerisation can occur in the source if metabolites elute at sufficiently high concentration. These can be homo or heterodimers of varying lengths (Mahieu et al., 2016). These could be detected as singly charged compounds such as $[2\text{M} + \text{H}]^+$ or as multiply charged species such as $[2\text{M} + 2\text{H}]^{2+}$. A metabolite or oligomer can become multiply charged in the ion source by picking up more than one adduct. As the m/z value recorded is a mass over charge ratio these features will be detected at roughly half the mass of the molecular ion if a doubly charged molecular ion is detected such as $[\text{M} + 2\text{H}]^{2+}$.

Artefacts are another source of confusion which are also not very well understood but commonly seen on FT based instruments such as Orbitraps (Brown et al., 2009, 2011). There are three different types of artefact, including fuzzy sites (Mitchell et al., 2018), ringing and partial ringing (Miladinović et al., 2012). They are thought to be caused by a combination of factors including radio frequency interference (RFI), saturation of the amplifier and or saturation of the digitiser (Mathur and O'Connor, 2009). This results in mistakes during the fast Fourier transform resulting in computationally

generated false peaks derived from the mixed harmonic signals of real peaks and RFIs. These are generally lower intensity peaks found around larger peaks (ringing) that do not correspond to a real chemical entity. Recently a machine learning based method shown to remove 90% of these peaks was published (Kantz et al., 2019).

Finally, the other issue giving rise to even more features derived from a single feature is presented by in-source fragmentations and biological transformations. In source fragmentation will occur for many compounds even with a soft ionisation technique such as ESI (Xu et al., 2015). Over 90% of compounds in METLIN fragment under typical ESI conditions (Domingo-Almenara et al., 2018). Tryptophan is a good example of this, in positive mode the molecular ion will be detected at 205 m/z but you should also expect to see a peak at 188 m/z which corresponds to one of the key fragments from tryptophan. These are harder to remove from the data as they require specific knowledge of what each compound is already during the data processing stage before identification has taken place. The resulting fragments will often correspond to real possible metabolites too. Some annotation software can remove them from the data if it is a common neutral loss through the usage of small neutral loss databases but for the majority of metabolites and fragments this is not the case (Domingo-Almenara et al., 2019). Also the level to which fragments form in source will be dependent on the ion source conditions utilised and the compound in question as demonstrated by Broeckling *et al*, who showed using standards that the fragments could be predicted based on compound structure. More recently (Domingo-Almenara et al., 2019) introduced a tool called MISA (METLIN-guided in-source annotation) which is capable of annotating in-source fragments through the use of low energy CID fragmentation spectra from the METLIN database. This was facilitated by Broeckling *et al* showing that ESI fragments are highly similar to those produced by low energy CID spectra. There is also the possibility of biological transformations occurring in-source but these would be difficult to differentiate from in source fragments.

Grouping of all of these alternative forms of a single metabolite is essential to ensure correct biological interpretation at the end point of a study. Lynn *et al*, predicted that 40 – 80 % of features within untargeted datasets are not molecular ions in 2015, whilst Mahieu and Patti demonstrated that up to 90% of features in their data were what they referred to as degenerate features in 2017. This number is likely to keep increasing as we continue to improve our understanding of the complex relationships involved in ion formation. This means feature lists can be dramatically reduced and thus the number of features that require identification is reduced. Insufficient grouping of these features may be a key reason why the number of features that can be successfully identified in biological studies is often disappointingly low.

Once correct grouping and reduction of degenerate features has been performed a list of predominantly protonated or deprotonated ions will remain, although for some metabolites another adduct type may have been most common. The neutral mass can be calculated and at this stage a search of the possible chemical formulae responsible for the recorded m/z will be performed. Kind and Fiehn's 7 golden rules are typically applied. These rules were developed based on the analysis of 68,327 real structures and are highly valuable as they allow a large reduction in the possible search space. There are 8 billion structures containing just the common naturally found elements C, H, N, O, P and S that are theoretically possible below 2000 Da, application of the rules reduces this to 623 million probable structures (Kind and Fiehn, 2007). They are designed to remove molecular formula assignments which represent highly chemically unlikely structures, the rules have been summarised in Table 7. Recently in this area an updated version of Van Krevelen diagrams has also be proposed to increase confidence in molecular formulae assignments (Rivas-Ubach et al., 2018). Once all features have been assigned chemical formulae they can then be searched against a metabolite database. There are a variety of different options to choose from. KEGG (Kanehisa et al., 2012), ChemSpider (Pence and Williams, 2010) and PubChem (Bolton et al., 2008) are very large databases which are freely available though they may be too large and contain many compounds that are likely to be irrelevant to any single metabolome of interest. To avoid this issue there are a number of smaller databases dedicated to particular species or sample types of interest such as the Human Metabolome Database (HMDB) (Wishart et al., 2009), Human Serum Metabolome Database (Psychogios et al., 2011), Human Urine Metabolome Database (Bouatra et al., 2013), LipidMaps (Sud et al., 2007), DrugBank (Law et al., 2014), FoodDB (Harrington et al., 2019) and Plant Metabolome Database (PMDb) (Udayakumar et al., 2012) amongst others. This step is highly unlikely to generate a single possible assignment due to the presence of isobars and isomers in nature and so the typical scenario is an extensive list of putative metabolites which match the top molecular formulae matches for each neutral mass. To be able to confidently assign one of these putative candidates as the true identification more information is required. Most typically this extra information is MS² fragmentation data.

Table 7: Summary of the seven golden rules for possible molecular formula restriction (Kind and Fiehn, 2007).

Rule 1 – Element number restriction	Restrictions on the maximum numbers of atoms of any particular element that can be considered.
-------------------------------------	--

Rule 2 – LEWIS and SENIOR rules	LEWIS – All main group elements should have completely filled s and p valence shells SENIOR – Multiple rules related to the sum of valences of the atoms, and the number of atoms making up the molecule.
Rule 3 – Isotopic pattern filter	The abundance and ratios of natural isotopic peaks in relation to the molecular ion are considered and can allow determination of the number of atoms of an element.
Rule 4 – Hydrogen/carbon ratio	H:C ratio should be less than 3 and greater than 0.125.
Rule 5 – Heteroatom ratio	Thresholds for the allowed ratio of particular heteroatoms to carbon.
Rule 6 – Element probability check	Removes molecular formulas which pass rule 5 but contain multiple high heteroatom:C ratios that are unlikely.
Rule 7 – Trimethylsilyl (TMS) check (GC-MS specific)	Ensure removal of TMS groups if derivatization has been applied and focus then on rules 4-6

1.3.1.4.2.3 MS² Data and Strategies for Acquisition and Analysis

MS² data is derived from the intentional fragmentation of components of the sample within particular m/z ranges and the subsequent mass analysis of the resulting fragments. The range of m/z values fragmented will vary depending on the method type applied as will be discussed in detail later in this section. The goal is to measure the fragments generated which hopefully provide a unique fingerprint representative of one of the candidate metabolite structures provided from the MS¹ data. If a unique fingerprint is not revealed there may be a peak which is indicative of a characteristic substructure of a certain class of compounds such as a phosphocholine head group from a glycerophosphocholine molecule (Lynn et al., 2015) allowing further restriction of putative candidates even if absolute identification was not possible. MS² fragmentation is the most commonly used method for generation of orthogonal data to provide the desired level or level 2 confidence in annotations in untargeted UHPLC-MS metabolomics experiments.

The fragmentation itself can be carried out in different ways. Higher energy collision induced dissociation (HCD) and collision induced dissociation (CID) are the most commonly applied

mechanisms (Vinaixa et al., 2016) but other methods such as Electron Transfer Dissociation (ETD), Electron Capture Dissociation (ECD), Infrared Multiple Photon Dissociation (IRMPD) and Electron Impact Dissociation (EI) can also be utilised (Aksenov et al., 2017). The fragmentation pattern that is seen upon analysis will vary depending on the technique utilised and the energy level applied in each technique (Allard et al., 2017). Multiple fragmentation techniques and energies can thus be utilised to provide orthogonal data for increased confidence in annotation (Mullard et al., 2014). In metabolomics, HCD and CID are the predominant techniques. CID is an older method of fragmentation than HCD which was first introduced on the LTQ Orbitrap XL in 2007 (Olsen et al., 2007). The mechanism of fragmentation in both techniques is the same, the isolated ions of interest are accelerated within the ion trap or quadrupole which contains an inert gas such as helium or nitrogen. The subsequent collisions with the gas molecules within result in the kinetic energy of the ions being transferred to internal energy in the ion in the form of vibrational energy in the bonds (de Graaf et al., 2011). This vibrational energy can result in fragmentation of the bond depending on the level of energy applied and the chemical bond in question. Where the two techniques differ is that CID is a resonant excitation technique, this means that the ions in the trap are rapidly excited and cooled on a millisecond timescale (Shao et al., 2014). It also suffers from low mass cut off and the one third rule. This is a phenomenon when performing CID on trapped ions that results in the loss of low m/z values that are less than a third of the precursor's m/z from the trap during the excitation process. HCD differs in that it is a beam type excitation technique, it occurs on a microsecond timescale and can generate low mass products overcoming the thirds rule (de Graaf et al., 2011). The level of energy applied in HCD can also be also higher and this results in the higher levels of the ions dissociation pathways being accessed. This means HCD will typically generate more fragments for the same ion than CID. The techniques are both valuable and can be considered complementary to each other.

Ultimately, whichever technique is applied the result will still be a mass spectrum containing peaks relating to fragments derived from the metabolite of interest. The MS^2 spectrum traditionally would then be manually inspected to elucidate the structure, possible through the presence of multiple characteristic substructures for example. Manual interpretation however, is a slow and labour intensive process requiring a high level of user expertise (Dührkop et al., 2019). More commonly, MS^2 spectral libraries containing curated reference spectra generated from pure authentic chemical standards will be searched to identify other compounds with the same or highly similar spectra. A mathematical technique such as the dot product (Allard et al., 2017), cosine similarity (Scheubert et al., 2017), reverse dot product (Tsugawa et al., 2019) or a combination of these as well as fragment presence or absence are typically used to determine the degree of similarity of two spectra resulting in a spectral similarity score. All spectra in a database can be searched automatically with matches

being ranked by their spectral similarity scores. A strong match score to a reference spectrum can be used to annotate the structure. There are a variety of different MS² spectral libraries available, some are open source (METLIN (Guijas et al., 2018), MassBank (Horai et al., 2010), MassBank of North America (MoNA) (MoNA, 2019), HMDB (Wishart et al., 2009), GNPS (Wang et al., 2016), LipidBlast (Kind et al., 2013), Golm Metabolome Database (Kopka et al., 2005), ReSpect (Sawada et al., 2012), FiehnLib (Kind et al., 2009)) and some commercially available (NIST, 2019), mzCloud (HighChem, 2019), LipidSearch (Thermo Fisher Scientific, USA), Wiley: MS for ID (Oberacher et al., 2013)). These are valuable resources with the larger databases such as METLIN and MassBank containing hundreds of thousands of spectra for tens of thousands of metabolites for thousands of different compounds, although only 5% of compounds present have data from pure authentic standards (Frainay et al., 2018). Many compounds in the largest libraries have multiple spectra derived from different fragmentation mechanisms, different fragmentation energies and different detector types (Guijas et al., 2018). Some databases are specialised and smaller such as PMDB and ReSpect and may be of interest for a certain sample type to help restrict biologically irrelevant or unlikely spectral matches. Even if one of the larger databases has been utilised researchers should still look to use multiple databases for the best coverage as Vinaixa et al, showed in their 2016 review of spectral libraries that although there is overlap between the different libraries each library contains a significant number of metabolites unique to that library. Unfortunately, many of the spectra in these databases will not be appropriate for comparison to any experimental data that is been collected particularly in the large databases. This is because there are multitude factors that will affect the spectrums characteristics alongside the fragmentation type and energy as mentioned earlier. These include but are not limited to the type of detector utilised, the modifiers used, matrix effects, artefact peaks and spectral purity. As a result, inter-laboratory variation is common and is another of the major challenges associated with untargeted metabolomics. Therefore, it is important to ensure that experimental spectra are compared to suitable reference spectra which applied the same type of fragmentation and a similar fragmentation energy. Ideally, a pure authentic standard will have been analysed previously using the same analytical method in the same lab allowing level 1 identification to be achieved but as mentioned this is not practical or feasible for many metabolites and most researchers. Despite these limitations, MS² spectral databases are still vital tools for untargeted metabolomics researchers and their careful expansion and curation is going to be important for the further advancement of the field.

Despite efforts to expand MS² libraries the reality is that there are many metabolites not present in them and much work is required by the community to continue populating them, this will be an expensive and time consuming process. In the meantime, researchers can utilise in-silico MS² tools to supplement their analyses and hopefully fill in the gaps. There are a number of different in-silico MS²

tools available to achieve this including CFM-ID (Allen et al., 2014; Djoumbou-Feunang et al., 2019), MetFrag (Ruttkies et al., 2016), HAMMER (Zhou et al., 2014), MassFrontier (Thermo Fisher Scientific, USA), MAGMa(+) (Ridder et al., 2014), MIDAS (Wang et al., 2014), FingerID (Heinonen et al., 2012), CSI:FingerID (Böcker, 2017), MS-FINDER (Tsugawa et al., 2016), ChemDistiller (Laponogov et al., 2018) SIRIUS 4 (Dührkop et al., 2019). Some operate by using a set of chemical rules to computationally determine which bonds are most likely to break at a particular fragmentation energy. This results in computer generated fragmentation spectra which can also be used for comparison of your experimental data. The performance of these software is based on the level of detail and accuracy provided in the reaction database of the software in question (Nash and Dunn, 2019). Other software work in a reverse fashion where the metabolite is predicted from the fragments present in your experimental spectra, whilst some employ machine learning methods to databases of compounds and their associated reference spectra. Other computational tools include LipidBlast (Kind et al., 2013), an entire library of in-silico spectra for lipids as well MS2LDA (van der Hooft et al., 2016) a novel approach based on text mining methods for identification of characteristic substructures in MS² data. Using in-silico spectra or other in-silico based tools can only provide a level 2 identification at best but they have improved rapidly since their introduction in part due to the CASMI challenge (Blaženović et al., 2017) and still provide a valuable alternative resource for annotating MS² data.

The importance of MS² data and how it is utilised in annotation has been discussed in this section but how the MS² data can be acquired in different ways has not been discussed thus far. Traditionally MS² data is collected in what is known as a data dependent analysis (DDA) method (Nash and Dunn, 2019). The method will be a top “*n*” method as determined by the user where *n* is the number of MS² scans between each MS¹ scan. After each MS¹ scan the top “*n*” highest intensity features from that full scan will be fragmented individually and sequentially using narrow isolation windows typically of 1 – 3 *m/z* around the *m/z* of interest (Figure 16). This technique is desirable as the MS² spectra generated should be pure or of very high purity. Purity in this sense refers to the lack of other features being fragmented simultaneously that fell within that narrow isolation window. A pure spectrum is desirable as the user can be sure that all product ions seen in the spectrum are derived from the same parent ion (Lawson et al., 2017). The resultant spectra can then be used to compare directly to an MS² spectral library as discussed earlier. While it is true that this method does generate relatively pure spectra in a simple fashion it is programmed to fragment ions of the greatest intensity. As a result, many potentially biologically relevant metabolites of lower concentration which have eluted simultaneously with multiple higher concentration metabolites will not have MS² data collected for them. With no MS² data the highest level of identification that can be achieved is level 4 (just a chemical formula) based on Schymanski’s updated levels of identification (Schymanski et al., 2014). This is far from the true and

confident level 1, or at least level 2 identification that is required to turn data into biological knowledge. Therefore, it can be said that DDA is biased toward metabolites of high concentration and the subsequent features of high intensity (Zhou et al., 2017; Renaud et al., 2017). It can then also be said that DDA is inappropriate for the goal of untargeted metabolomics which is global profiling of all metabolites within in a sample.

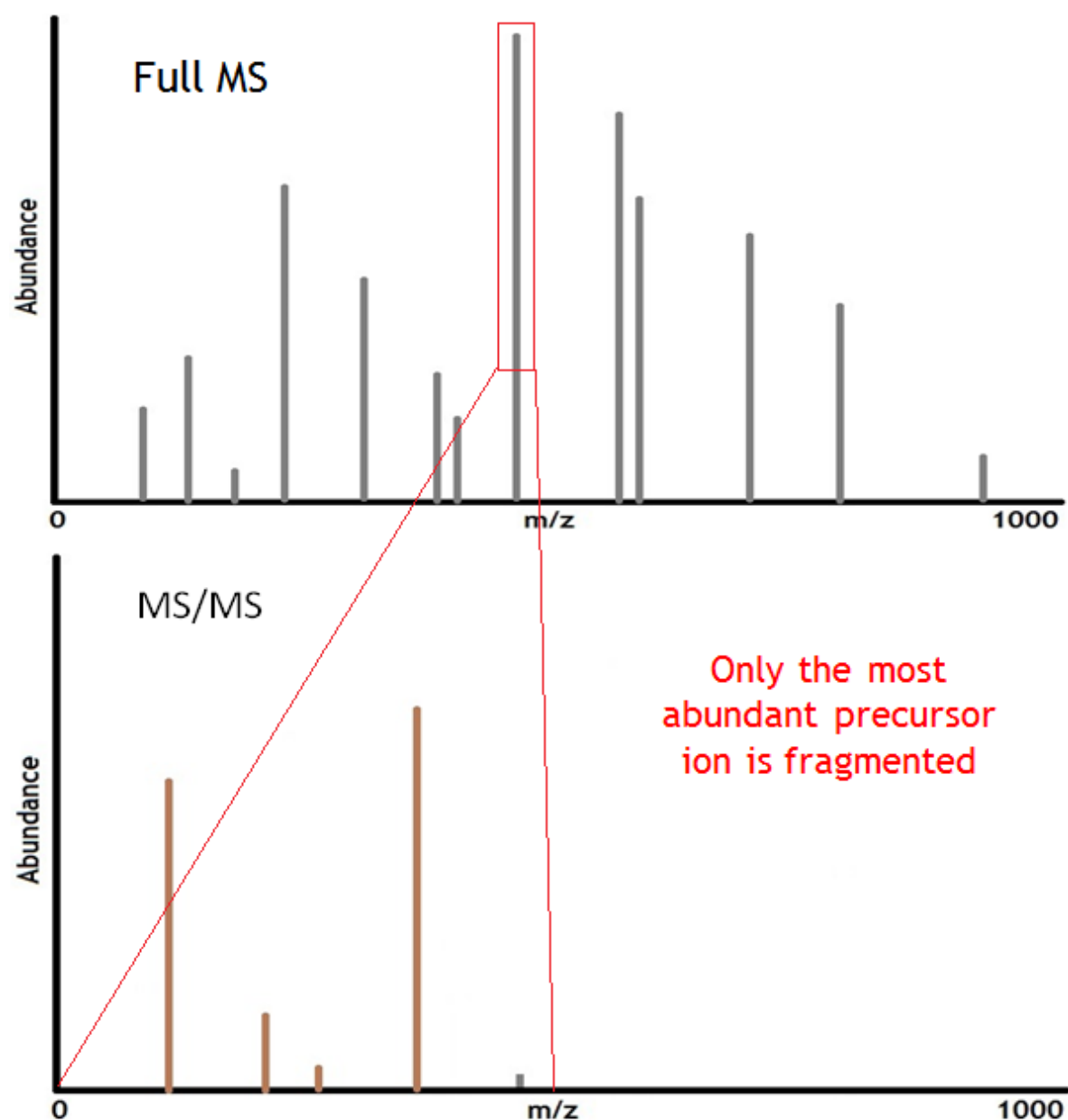


Figure 16: A representation of a data dependent MS² scan. The most intense feature from the full scan in the upper panel is fragmented generating a pure and simple MS² spectrum which can be matched to a MS² spectral library. The number of peaks fragmented between each full scan is determined by the user.

Data independent analysis (DIA), also referred to as sequential window acquisition of all theoretical fragment ion spectra (SWATH) (Ma et al., 2016) is an alternative way of collecting MS² data which can theoretically acquire MS² data for all features within the sample as is the goal of untargeted metabolomics (Nash and Dunn, 2019). It can also provide superior, reproducibility and quantitation than a traditional DDA method (Zhou et al., 2017). It is a technique that was first applied in proteomics in 2012 (Gillet et al., 2012) and become more widely applied in proteomics since (Ludwig et al., 2018) but uptake of it for metabolomics has been more limited. DIA works by using “*n*” sequential windows of a given *m/z* width. All features that fall within each window are fragmented simultaneously (Figure 17). This results in complex fragmentation spectra with product ions from many different precursor ions (Zhou et al., 2017). This creates a problem for the subsequent data analysis step, the spectra require deconvolution, the separation of fragment peaks from the different precursors into pure spectra. This can be achieved by comparing the chromatographic profiles of all precursor and product ions. A precursor’s associated product ions should have highly correlated chromatographic profiles (Bonner and Hopfgartner, 2018). This can be carried out using an open source software called MS-DIAL (Tsugawa et al., 2015). By looking at these correlations, pure, deconvoluted MS² spectra can be generated for each of the precursors that were present in the preceding MS¹ scan. To ensure that deconvolution is effective however the user should strive to limit the number of features falling into one single fragmentation window. The greater the number of features within one window the smaller the chances are of the software being able to successfully separate and differentiate the chromatographic profiles of all the precursor/product relationships correctly. Therefore, the user needs to be careful when designing a DIA experiment to ensure that the fragmentation data collected is still going to be informative. Other software for handling analysis of metabolomic data include MetDIA (Li et al., 2016b), MetaboDIA (Chen et al., 2017) and MetFamily (Treutler et al., 2016) however neither of these perform true untargeted mathematical deconvolution as is done MS-DIAL. They each rely on custom built DDA based spectral libraries to extract pure metabolite spectra from the complex DIA based on the library contents. Very recently a new software was published called DecoMetDIA (Yin et al., 2019), this is the first competitor to MS-DIAL that performs true mathematical deconvolution of DIA MS/MS spectra. This data analysis challenge has prevented significant uptake of DIA in the metabolomics community with only a handful of biological studies published.

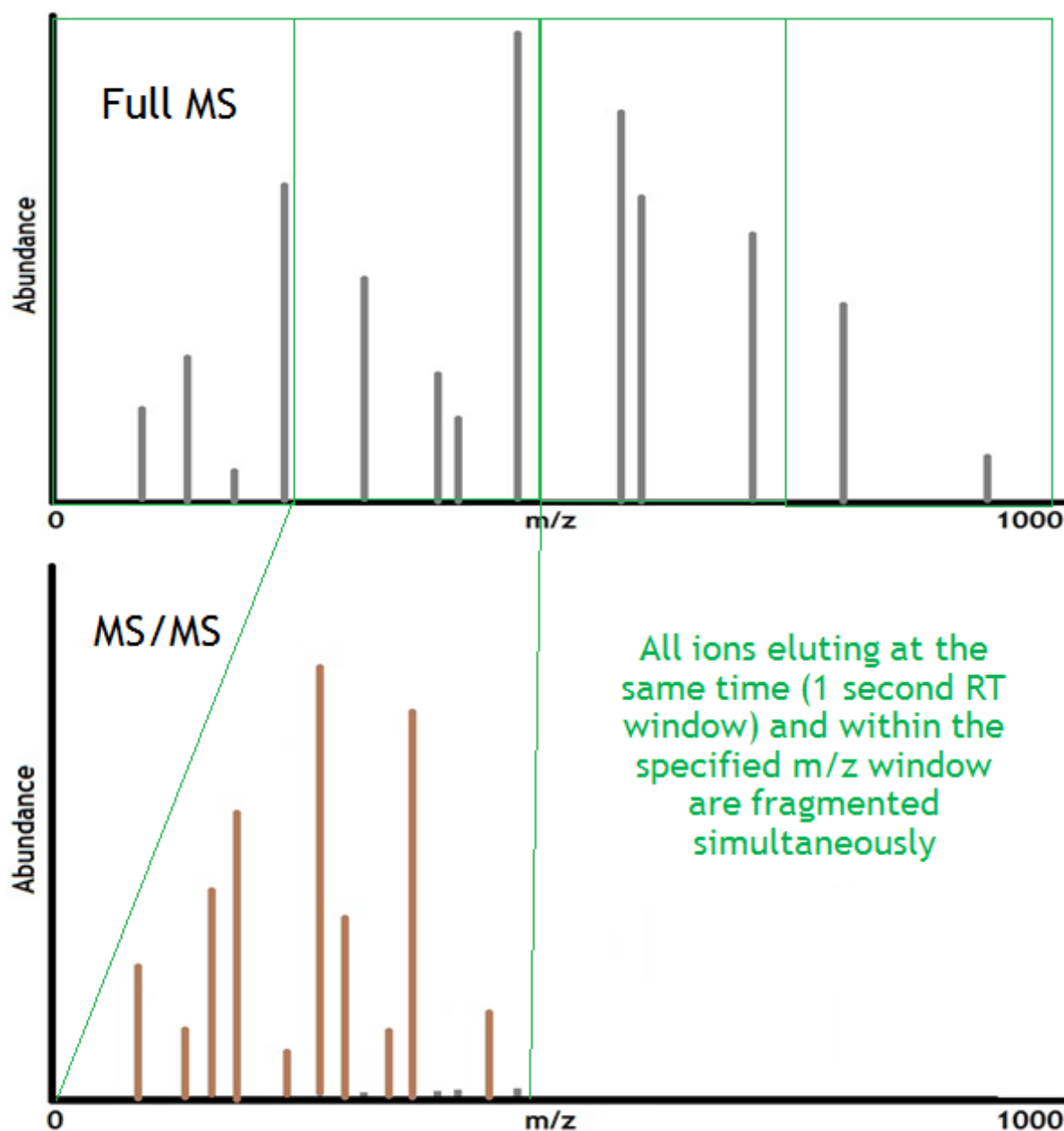


Figure 17: A representation of a DIA MS² scan. Sequential windows of a specified m/z width are fragmented between each full scan. Four sequential windows of 250 m/z from 0 to 1000 are employed in this case. All features that eluted simultaneously between 250 and 500 m/z (Scan time = 1 second) are fragmented simultaneously creating a complex MS² spectrum requiring deconvolution before matching to a MS² spectral library.

A third MS² acquisition strategy is available which also promises global coverage and is known as all-ion fragmentation (AIF) (Geiger et al., 2010) and is sometimes referred to as MS^E or MS^{ALL} (Zhou et al., 2017). It is essentially a variation of DIA but this strategy operates by collecting one MS¹ scan which is followed by just one MS² scan which covers the entire mass range of interest (Plumb et al., 2006). The

resulting spectra require deconvolution in the same manner as a DIA MS² spectrum and this can be achieved in MS-DIAL. However, the number of features falling into one fragmentation window will be far greater than a typical DIA window due to its greater width and therefore the effectiveness of the deconvolution process will be significantly decreased. This method is only likely to be effective if employed over a smaller mass range than what is typically employed in untargeted metabolomics (normally at least 500 *m/z* and up to 1500 *m/z*) or is employed upon a sample type of low complexity.

Each of the three strategies discussed has drawbacks. DDA methods are reliable and generate simple data which can be annotated immediately but suffer from lack of coverage. Whilst DIA and AIF methods provide the coverage but generate data that is difficult to analyse and requires manipulation before annotation. Another type of method that has recently been becoming more prominent is employing an intelligent-DDA (iDDA) method. This by the authors definition constitutes employing a typical DDA method but with some modifications to make overcome the coverage deficiency and/or modify the precursor selection to maximise the amount of useful biological data. These can include, segmentation of the total mass range (Mullard et al., 2014), exclusion/inclusion lists and exclusion/inclusion lists which are progressively updated with repeated injections or intelligently derived based on the data (Neumann et al., 2013). By segmenting the mass range the sensitivity of the method should be increased due to decreased ion suppression. This is achieved as there is a decrease in the number of different metabolites entering the Orbitrap cell at any one time and therefore there is less competition and interference between different ions allowing more accurate detection of lower intensity ions. It may also aid to improve data quality and accuracy through a reduction in space charge effects within the Orbitrap cell. An exclusion list is a user defined list of masses that may or may not be associated with a RT that when included in the method will cause the mass spectrometer to not fragment that mass if it is detected. The inclusion list performs the opposite operation to the exclusion list. The user submits the list and the mass spectrometer will then prioritise these masses for data dependent MS² fragmentation if detected even if the intensity is lower than other ions that are simultaneously present and would normally have been prioritised if utilising a typical DDA method. An inclusion list will not just exclusively fragment *m/z* values that are on the list however. The method can be set up so that it will operate in the normal DDA fashion when none of the inclusion list *m/z* values are being detected.

Inclusion lists and exclusion lists can be applied independently from each other or in tandem. Exclusion lists are commonly generated and applied after analysis and peak picking has been carried out on a solvent or extraction blank. This ensures that the user is not fragmenting non-biologically relevant features derived from the solvent or other contaminants such as plasticisers that are present in the

sample as a result of the sample preparation process. An inclusion list can be applied in the same way after analysing in full scan and applying peak picking on the sample of interest. Taking this a step further involves performing repeated injections of your sample with an updated exclusion list of each analysis, where the exclusion list is updated with the m/z values that were fragmented in the previous analysis. This ensures that time is not wasted fragmenting features which have already been fragmented. The issue with this methodology however is that the user should look to perform all analyses of a sample in a single run with samples being run back to back. This creates a challenge for the user to generate and update the appropriate lists and update the appropriate methods in time before the next sample with an updated method can be analysed. This is a very time pressured, labour intensive method likely requiring long hours in the lab. Recently though Thermo Fisher Scientific have released a new mass spectrometer, the Orbitrap ID³ in which the software can perform this kind of operation in a totally automated fashion. The method works by first analysing in full scan MS¹ a blank and your sample of interest. These analyses are utilised to generate the exclusion and inclusion lists respectively. The user will also determine how their DDA method will operate and over how many repetitions it will operate. The software will then perform as many DDA analyses as the user determined with the exclusion list being updated after each injection to ensure that the same features are not being fragmented repeatedly. This should ensure that through each injection new and informative MS² data is being generated and now alleviates the issue of a DDA method being biased towards ions of high intensity.

1.3.1.4.2.4 Biological Interpretation with Metabolic Pathway Analysis and -omics data Integration.

Following annotation of the MS² data the user can then begin the biological interpretation of the data. It is at this stage where the user may wish to perform metabolic pathway analysis. This is where the metabolites identified, or the metabolites of interest are mapped onto known metabolic pathways for the organism of interest. Metabolic pathway databases and/or analysis tools include KEGG (Kanehisa et al., 2012), MetaCyc/BioCyc (Caspi et al., 2016), SMPDB (Jewison et al., 2014), GeneOntology (Blake et al., 2015), WikiPathways (Slenter et al., 2018), Metaboanalyst 4.0 (Chong et al., 2018), ChemRICH (Barupal and Fiehn, 2017) and Compound Discoverer 3.0 (Thermo Fisher Scientific, USA). As well as providing context to annotations and informing biological interpretation through elucidation of parts of pathways or perhaps whole pathways that have been dysregulated. They can also help to provide extra confidence in annotation too. For example, if a putative annotation is for a metabolite in a certain pathway and there are 75% of the other metabolites in that pathway already identified then that can increase confidence. Integration of metabolomics data with information from genomics, epigenomics, transcriptomics and/or proteomics may also be included to provide a complete biological picture and allow a true systems approach for the best understanding of biological issues.

1.4 Closing Statement

Annotation of metabolomics data is still a major challenge in untargeted metabolomics though the situation has been continually improving since the field came into existence around the turn of the century. The improved resolving power, ion transfer and ion manipulation capabilities of modern mass spectrometers has facilitated the collection of more informative data more quickly. The greater mass accuracy allows a greater reduction of putative annotations, superior ion transmission allows greater sensitivity and improved ion manipulation allows MSⁿ capabilities. All of these combine to improve annotation capabilities. Other improvements in hyphenated technologies such as miniaturisation as well as totally new hyphenated technologies for prior separation of samples have allowed greater sensitivity or alternative methods of selectivity. Whilst rapid recent improvements in computational resources have helped significantly and include in-silico MS² tools which have been developed and are improving quickly. Metabolic pathway analysis tools are improving and are now commonly integrated into workflows and can increase confidence in annotation and increase the biological knowledge gained. Automated cloud-based workflows such as Galaxy-M (Davidson et al., 2016) and Workflow4Metabolomics (Giacomoni et al., 2015) also exist now and allow quicker, easier and more accurate analysis of datasets for the entire workflow in one package instead of requiring multiple pieces of software. These also make metabolomics more accessible to a wider range of researchers. This has helped to fuel the increasing size of the community which in turn means more researchers to improve the current computational resources and reference databases which are both essential for useful biological information to be acquired from each biological study performed. There is also a drive for greater cooperation and standardisation between laboratories and this could start to generate even more rapid improvements in the biological impact of studies through pooling of resources and knowledge and creation of consortiums such as mQACC (Beger et al., 2019) for promoting better QA and QC practices. Ultimately, much more work is required to improve metabolite annotation in untargeted studies, which is the focus of this thesis. It is still the case that only a small and disappointing portion of features in an experiment can be annotated, and so this thesis will focus on methods to improve metabolite annotation.

This will first be done by looking into the complexity of LC-ESI-MS full scan data. Ion formation in the ESI source is a process that is not well understood as discussed in section 1.3.1.4.2.2. A single metabolite can become a number of different adducted forms, can contain an isotope, become multiply charged, form homodimers, heterodimers or higher order oligomers, as well as potentially fragmenting in the source, or reacting with another metabolite in source. These are all detected as features and can give rise to characteristic mass differences between the features with the same retention time. These mass differences have never been characterised before across a large number

of datasets and have the potential to reveal common adducts, fragments, isotopes or combinations of these that have not previously been identified. This will be investigated in the first research chapter with the aim of providing new information that can improve annotation of MS¹ data.

Following this, the focus will shift towards MS² data and how the collection of it can be maximised to gain more useful biological information. This is first investigated with studies on a hybrid quadrupole-Orbitrap instrument, the Q Exactive Plus (Thermo Fisher Scientific, USA). This was first investigated by exploring the implementation of DIA experiments due to the promise of global MS² fragmentation of features. To ensure effective DIA experiments were planned with appropriate mass resolution, scan rate, as well as DIA window sizes, numbers, and range, the complexity of features in theoretical DIA windows was assessed along with other key data characteristics such as scan rate and peak widths. Is it possible for DIA to be an effective tool for improvement of MS² coverage on the Q Exactive Plus instrument is the focus of the second research chapter.

The work done in the second research chapter was then used to design and implement DIA experiments as well as intelligent DDA methods and AIF methods for comparison of which strategy provided the most useful biological information on a Q Exactive Plus mass spectrometer. This is the basis of the third research chapter. The fourth research chapter then capitalises on the release of the new Orbitrap ID-X and it is built in intelligent DDA capabilities. A custom built MS² library was developed to allow level 1 identifications. This library and the intelligent DDA feature of the system were implemented alongside a traditional DDA method to allow comparison of the two methods for annotation of NIST plasma and urine standard reference materials (SRM). This builds on the work of the preceding two chapters whilst factoring in state-of-the-art technology in the form of the Orbitrap ID-X. To summarise the objectives for each chapter are as follows:

Chapter 3.0 – Assess the complexity of the characteristic mass differences of related features across 104 untargeted datasets to provide new insight for more accurate future feature annotation.

Chapter 4.0 – Assess the complexity/density of features and other key data characteristics of different full scan data to allow effective implementation of DIA methods on a Q Exactive Plus (Thermo Fisher Scientific, USA).

Chapter 5.0 – Implement the DIA methods whilst also comparing them to other MS² acquisition strategies (DDA, iDDA, AIF) to determine which strategy is most suitable for the generation of the greatest volume of potentially informative MS² data on a Q Exactive Plus (Thermo Fisher Scientific, USA).

Chapter 6.0 – Use the new Orbitrap ID-X Tribrid mass spectrometer (Thermo Fisher Scientific, USA) to create a custom MS² library for annotation and subsequently compare annotation between its built in iDDA (AcquireX method) with a traditional DDA method.

Overall these investigations aim to provide new information and advice to researchers that will allow them to better understand their untargeted metabolomics data and allow more accurate and efficient feature annotation. It will also allow researchers to design better MS² experiments on hybrid quadrupole-Orbitrap instruments as well as on the Orbitrap ID-X which will generate more confident annotations than would previously have been achieved. As metabolite annotation is the biggest challenge in the field this represents important work contributing to the progress of untargeted metabolomics as a whole.

2.0 Materials and Methods

2.1 Characterising the complexity of electrospray ionisation data and its impact on metabolite annotation in UPLC-MS metabolomics studies (Chapter 3.0)

2.1.1 Datasets applied

104 UHPLC-ESI-MS datasets were collected representing a variety of different sample types, instrument types/manufacturers and assay types in both positive and negative ion mode. Some were downloaded from the publicly available data repositories MetaboLights (Kale et al., 2016) (Available at: <https://www.ebi.ac.uk/metabolights/>) and Metabolomics Workbench (Sud et al., 2015) (Available at: <https://www.metabolomicsworkbench.org/>) whilst others were acquired from The Phenome Centre Birmingham (provided by Professor Dunn). An assessment of all UHPLC-ESI-MS datasets in the MetaboLights and Metabolomics Workbench repositories was carried out. All suitable datasets that were available as of 03/11/2017 were downloaded in either .txt, .tsv, .csv or .xlsx format. Suitable datasets were defined as those that were already pre-processed into a peak intensity data matrix and also had an appropriate number of features and samples. Datasets with more than 20 samples and greater than 1000 features were deemed appropriate. These were arbitrary values that were selected however a minimum of 20 samples was sought after to try to ensure a reliable correlation coefficient could be calculated for each feature pair. 1000 features were selected to avoid the inclusion of small or previously well grouped data which would produce much smaller and less realistic numbers of correlated feature pairs for analysis. A small number of exceptions to these rules were made due to the lack of appropriate datasets available and to help provide representation of different varieties of instrument manufacturers and sample types. A list of all datasets included along with key information including sample type, ion mode, assay applied and mass spectrometer applied can be found in the electronic Appendix (2.1/2.1.1).

2.1.2 Data processing and interpretation

2.1.2.1 Correlation Data Generation

All datasets were saved as .csv files and were modified into the same structure to allow easy upload and analysis of the data utilising R version 3.2.2. All features in each dataset were divided into retention time bins of 5 seconds which overlapped by 2.5 seconds. These were arbitrary values that were selected based on the authors knowledge of typical chromatographic peak widths in LC-MS untargeted metabolomics applications. For each possible pair of features within each bin the difference between their m/z values were calculated. Pearson correlation and an associated p -value were also calculated on the intensities of the two features across all samples in the dataset. Pearson Correlation and p -value were calculated using the `rcorr` function from the `Hmisc` package. The m/z differences were calculated using the `dist` function from the `base stats` package. The resulting matrices

generated from use of rcorr and dist were converted into vertical tables using the melt function from the reshape package. The tables were combined together and stored together with the feature and bin information. The results for each dataset were stored in individual tables in an SQL database using the RSQLite package. The results tables were then filtered using R values of ≥ 0.8 , a p -value ≤ 0.05 and an $n \geq 20$. If after filtering a particular dataset the number of feature pairs left was less than 1000 then the n value filter was removed. The m/z differences that remained after filtering were rounded to 4 decimal places using the mutate function from the dplyr package. The summarize function also from dplyr was used to merge the identical rounded differences into individual rows and also calculated the frequency of each m/z difference, adding it to the table. The resulting frequency table for each dataset was saved in an SQL database. The same process was also carried out to groups of datasets as well as all of the datasets combined. The groups utilised and the datasets assigned to each group are found in the electronic Appendix (2.1/2.1.2.1).

2.1.2.2 Annotation of Correlated Mass Differences

Any mass difference with a frequency of greater than or equal to 40 from the overall dataset ($R \geq 0.8$, $p \leq 0.05$, $n \geq 20$) were utilised for annotation. There were 8,273 different mass differences to four decimal places with a frequency of 40 or more. These were grouped together manually based on the frequency distributions where generally normal distributions could be observed around the highest frequency mass differences across the whole mass range to generate 1,038 unique mass differences to be annotated. Initial annotation was carried out using commonly known adducts and isotopes in a manual fashion. Some further annotation was subsequently carried out with a list of biological transformations/fragments from the KEGG database in an automated fashion using R version 3.2.2. The final annotation step was performed by calculating theoretical m/z differences of many combinations of common adducts, isotopes, fragments and different combinations of these with each other. Some more uncommon but still plausible adducts and isotopes were considered too but with the most common adducts and isotopes, an example is shown below:

$[M + (C_{12}-C_{13} \text{ diff}) + H]^+ - [M + H]^+ = \text{an } m/z \text{ diff of } 1.0035 = \text{the } 1^{\text{st}} \text{ most prevalent } m/z \text{ diff}$

$[M + C_2H_2 + H]^+ - [M + H]^+ = \text{an } m/z \text{ diff of } 26.0156 = 14^{\text{th}} \text{ most prevalent } m/z \text{ diff}$

$[M + C_2H_2 + (C_{12}-C_{13} \text{ diff}) + H]^+ - [M + H]^+ = \text{an } m/z \text{ diff of } 27.0190 = 28^{\text{th}} \text{ most prevalent } m/z \text{ diff}$

$[M + C_2H_2 + H]^+ - [M + (C_{12}-C_{13} \text{ diff}) + H]^+ = \text{an } m/z \text{ diff of } 25.0121 = 62^{\text{nd}} \text{ most prevalent diff}$

$[M + C_2H_2 + H]^+ - [M + C_2H_2 - O + H]^+ = \text{an } m/z \text{ diff of } 10.0207 = 386^{\text{th}} \text{ most prevalent diff}$

2.1.2.3 Mass Difference Frequency Comparison

An error in Daltons was assigned manually to each mass difference based upon examination of the frequency distributions seen when performing the manual grouping of the 8273 filtered mass differences. The remaining grouped 1038 mass differences and manually assigned errors were then used to perform matching to all of the original datasets individually. This was carried out in automated fashion using R version 3.2.2. By performing this matching process the frequency of each of the 1038 commonly detected mass differences was determined for each dataset. The frequencies for all datasets were collated into tables to allow comparison of mass difference frequency between the datasets. The resulting tables were too large to easily visualise the data although the full table is included in the electronic Appendix (3.2/3.2.3) and is conditionally formatted as described in Figure 18. The tables displaying the frequency of the 1038 mass differences across the 104 datasets was used to filter the mass differences. Mass differences detected in less than 40% of all the datasets were removed leaving 264 of the original 1038 common mass differences. These 264 mass differences were used to filter the group frequency comparison tables. Data were presented by using zoomed out screenshots of the conditionally formatted Excel sheet (Figure 18).

Edit the Rule Description:

Format all cells based on their values:

Format Style: 3-Color Scale

	Minimum	Midpoint	Maximum
Type:	Lowest Value	Percentile	Number
Value:	(Lowest value)	50	100
Color:			
Preview:			

Figure 18: Conditional formatting rules for frequency cells in the frequency comparison tables.

2.1.2.4 Biotransformation Frequencies

1070 different biological transformations from the KEGG database and their associated m/z differences between reactant and product metabolites were searched for in the list of 1038 commonly detected mass differences from the 104 datasets described above. These lists were imported into R version 3.2.2 and matching between the masses in each list was performed. This was performed using ± 10 , ± 2 and ± 0.5 ppm error windows with the results for all 1070 transformations and all 104 datasets being presented in conditionally formatted, colour coded, Microsoft Excel grids. Cells were

conditionally formatted as shown in Figure 18. Excel grids are included in the electronic Appendix (3.2/3.2.4).

2.1.2.5 Dimer Identification

6 randomly selected datasets (3 positive ion mode, 3 negative ion mode) that were in the top 20% for number of significant feature pairs were considered. Raw peak matrices for each dataset were imported into R. The mean intensity was calculated for each dataset and any feature in the top 10% highest intensity features in each dataset were considered for further analysis. Only high intensity features were considered as dimerization is known to occur at high intensity and therefore by making this restriction high confidence can be had in the dimers found.

The co-eluting significant feature pairs as calculated previously (2.1.2.1) were imported from the SQL database that they were stored in using the RSQLite package. This data includes the feature ID number which relates to the feature ID from the raw peak matrices. The feature IDs for the top 10% highest intensity features were used to filter the significant feature pairs table so that only pairs where both of the features were in the top 10% highest intensity were remaining. The results table was split into separate tables based on the RT bin information already present. An assumption was made that all features were protonated or deprotonated adducts depending on the ion mode utilised. Each feature in the bin had the proton mass added or subtracted to its m/z value dependent on ion mode to give a list of potential monomer neutral masses. The m/z distance between each feature had a minimum and maximum value calculated for it presuming a ± 10 ppm error. If any of the neutral monomer masses calculated from that bin fell between any of the minimum and maximum m/z distances calculated for that bin then that feature pair could be identified as a dimer. Where a match was found the monomer neutral mass and the relevant row from the significant pairs table were joined together and added to a new data frame. Duplication of results which may have arisen from the overlapping RT bins was removed by using the distinct function from the dplyr package. The final data frame of dimers was exported as a .csv file for further analysis.

2.1.2.5.1 Homo and Hetero Dimer Identification

The .csv file was utilized for further analysis to determine how many homodimers and heterodimers were present in the data. The monomer neutral mass was subtracted from the lower mass feature from each feature pair which had been identified as featuring in a dimer. The subsequent list of values was imported into R version 3.2.2 and were searched against the list of 1038 common mass differences with a ± 0.5 Da error. If a match was found then the dimer pair was assigned as a homodimer, if no match was found the pair was assigned as a heterodimer.

2.2 Characterisation of UHPLC-MS full scan data complexity and its influence on MS² data collection on Q Exactive mass spectrometers (Chapter 4.0)

Two different reversed phase assays were utilised in this work. One utilising a reversed phase aqueous C₁₈ column which will be referred to as RP from here on. The other method was using a lipid tailored method on a reversed phase C₁₈ column and from here on will be referred to as the lipidomics or lipids method. Materials utilised in this chapter are detailed in Table 8.

Table 8: All materials and associated supplier in brackets utilised for all experiments described in chapter 4.0.

Material
Optima UHPLC Grade Acetonitrile (ACN) (Fisher Chemicals)
Optima UHPLC Grade Water (Fisher Chemicals)
Optima UHPLC Grade Isopropanol (IPA) (Fisher Chemicals)
UHPLC-MS Grade Formic Acid (FA) (Fisher Chemicals)
Ammonium Formate (Fisher Chemicals)
HILIC Accucore Amide 100mm, 2.1mm i.d., 2.6 µm particle size (Thermo Fisher Scientific)
Hypersil Gold aQ Column 100mm, 2.1mm i.d., 1.9 µm particle size (Thermo Fisher Scientific)
Hypersil Gold Column 100mm, 2.1mm i.d., 1.9 µm particle size (Thermo Fisher Scientific)
Human Plasma (BioIVT)
Human Urine (BioIVT)
Sheep Liver (Local Organic Butcher)
LC Vials (Chromatography Direct)
7 mL Glass Vials (Scientific Supplied Limited)
1.5 mL Eppendorfs (Fisher Chemicals)

2.2.1 Sample Preparation

2.2.1.1 RP Plasma/Urine Sample Preparation

Master samples of plasma and urine were thawed on ice and vortex mixed for 30 seconds. For HILIC and RP samples a 1 in 4 dilution of the plasma or urine was carried out with acetonitrile (ACN). 3 vials were used with 250 µL of sample and 750 µL of ACN added to each. All samples were vortexed for 30 seconds before being centrifuged at 14000 g and 4°C for 15 minutes. The supernatant was extracted from each vial and put into a new vial and the speed vac was used to dry the samples down. Samples

were reconstituted in 1000 µL of UHPLC grade water. As many 200 µL aliquots as possible were taken and put into LC vials and stored at -80°C for future analysis.

2.2.1.2 HILIC Plasma/Urine Sample Preparation

Master samples of plasma and urine were vortexed for 30 seconds. A 1 in 4 dilution of the sample was carried out with ACN for each replicate. 3 vials were used with 250 µL of sample and 750 µL of ACN added to each. Samples were vortexed for 30 seconds before being centrifuged at 14000 g and 4°C for 15 minutes. The supernatants were extracted from each vial and were put into 3 separate 7 mL vials and were vortexed for 30 seconds. As many 200 µL aliquots as possible were taken and put into LC vials and stored at -80°C for future analysis.

2.2.1.3 Lipidomics Plasma Preparation

Sample preparation for the lipids plasma samples was the same as for HILIC plasma except the 1 in 4 dilution at the beginning was carried out using isopropanol (IPA).

2.2.1.4 RP Tissue Preparation

Sheep liver tissue was acquired from a local butcher. Three 10 mg pieces were weighed out and placed into separate Precellys tubes. 1000 µL of ACN was added to each tube. The Precellys tubes were placed into the Precellys homogeniser and run for 10 seconds with a 10 second repeat. The supernatants were extracted into 1.5 mL Eppendorfs and were centrifuged at 14000 g and 4°C for 15 minutes. The supernatant was transferred to new 7 mL glass vials and the speed vac was used to dry the samples down. Each sample was reconstituted in 5 mL of water. As many 200 µL aliquots as possible were taken and put into LC vials and stored at -80°C for future analysis.

2.2.1.5 HILIC Tissue Preparation

HILIC tissue preparation was identical to the RP preparation until removal of the vials from the centrifuge. Supernatant was removed from each vial and added to a 7 mL vial. 4 mL of ACN was added to each vial and they were vortexed for 30 seconds. As many 200 µL aliquots as possible were taken and put into LC vials and stored at -80°C for future analysis.

2.2.1.6 Lipidomics Tissue Preparation

Lipidomics tissue preparation was identical to the HILIC tissue preparation except IPA was used instead of ACN.

2.2.2 Chromatography

All solvents used were sonicated for 20 minutes or until the solution was clear before use.

2.2.2.1 HILIC

Solvent A: 10mM Ammonium Formate in 95% ACN/Water + 0.1% FA

Solvent B: 10mM Ammonium Formate in 50% ACN/Water + 0.1% FA

UHPLC Column: Accucore 150 Amide HILIC 100 mm x 2.1mm, 2.6 μ m (Thermo Fisher Scientific)

Column temperature: 45°C

The chromatographic gradient for HILIC separation is outlined in Table 9.

Table 9: HILIC chromatography gradient

Step	RT (min)	Flow Rate (mL/min)	% Solvent A	% Solvent B	Curve
1	0.0	0.5	99	1	5
2	1.0	0.5	99	1	5
3	3.0	0.5	85	15	5
4	6.0	0.5	50	50	5
5	9.0	0.5	5	95	5
6	10.0	0.5	5	95	5
7	10.5	0.5	99	1	5
8	14.0	0.5	99	1	5

2.2.2.2 RP

Solvent A: 0.1% FA in H₂O

Solvent B: 0.1% FA in ACN

Column: Hypersil GOLD aQ 100 mm x 2.1 mm, 1.9 μ m (Thermo Fisher Scientific)

Temperature: 45°C

The chromatographic gradient for RP separations were as outlined in Table 10.

Table 10: RP chromatography gradient.

Step	RT (min)	Flow Rate (mL/min)	% Solvent A	% Solvent B	Curve
1	0.0	0.3	99	1	5
2	0.5	0.3	99	1	5
3	2.0	0.3	50	50	5
4	9.0	0.3	1	99	5
5	10.0	0.3	1	99	5
6	10.5	0.3	99	1	5
7	14.0	0.3	99	1	5

2.2.2.3 Lipidomics

Solvent A: 10mM Ammonium Formate in 60% ACN/H₂O + 0.1% FA

Solvent B: 10mM Ammonium Formate in 90% IPA/ACN + 0.1% FA

Column: Thermo Hypersil GOLD 100 mm x 2.1 mm, 1.9 µm (Thermo Fisher Scientific)

Temperature: 45°C

The chromatographic gradient for Lipids separations were as outlined in Table 11.

Table 11: Lipidomics chromatography gradient.

Step	RT (min)	Flow Rate (mL.min ⁻¹)	% Solvent A	% Solvent B	Curve
1	0.0	0.5	55	45	5
2	0.5	0.5	55	45	5
3	8.5	0.5	0	100	5
4	9.5	0.5	0	100	5
5	11.5	0.5	55	45	5
6	14.0	0.5	55	45	5

2.2.3 UHPLC-MS Data Acquisition

The parameters utilised for full scan data acquisition are shown below and were the same for each chromatography method and sample type. Analyses were performed separately at four different mass resolutions (17,500, 35,000, 70,000, 140,000 FWHM at m/z 200), in each ion mode (positive or negative). Data were collected with three biological replicates for each assay, sample and resolution

combination. All samples in each assay were analysed in a single analytical batch. Data were collected using an electrospray Thermo Q Exactive Plus mass spectrometer coupled to a Vanquish UHPLC system. Parameters utilised were as shown in Table 12.

Table 12: Mass Spectrometry parameters.

Method Duration (minutes)	14
Scan Range (<i>m/z</i>)	(100 – 1000 for HILIC and RP), (100 – 1500 for Lipidomics)
AGC Target	1×10^6
Maximum Injection Time (ms)	100
Microscans	1

2.2.4 Conversion to .mzML Format

Raw data files were converted to .mzML format using the MSconvert software available in the Proteowizard (Kessner et al., 2008) suite within R version 3.3.2 (R Core Team, 2017).

2.2.5 Isotopologue Parameter Optimisation (IPO)

IPO optimisation was carried out using the IPO package (Libiseller et al., 2015) in R version 3.2.2. The software works using a design of experiments (DOE) approach to test different XCMS parameters with their performance optimised on the number of the reliably detected peaks that are picked. Only peaks with ^{13}C isotopic peaks associated are considered to ensure optimisation is carried out upon real features. This tool is commonly applied by researchers applying XCMS for processing of their LC-MS data (Harvey et al., 2018; Roszkowska et al., 2018; Stoessel et al., 2018) The DOE approach means it is a time and computationally intensive process and therefore processing was directed through the University of Birmingham’s high powered computing (HPC) cluster known as the BlueBEAR (University of Birmingham, 2019). This was necessary due to the length of time and amount of computing power required to operate the software. Each set of triplicates which had already been converted to .mzML format as described earlier were input individually into the software. Starting parameters for the optimisation were consistent for each triplicate that was input. Default values were used for most parameters except for four parameters (Min Peak Width, Max Peak Width, PPM, Mzdiff), the remaining rows of parameters represent the default options which were utilised Table 13.

Table 13: Starting parameters utilised for the IPO optimisation software.

IPO Starting Parameter	Value
Min Peak Width	2, 8
Max Peak Width	10, 25
PPM	2, 20
Mzdiff	-0.01, 0.05
Snthresh	10
Noise	0
Prefilter	3
Value of Prefilter	100
mzCenterFun	wMean
Integrate	1
Fitgauss	FALSE
Verbose columns	FALSE

2.2.6 XCMS data processing

Peak picking, grouping and alignment were carried out using XCMS (Smith et al., 2006). XCMS was applied as it is widely applied in the field, there are expertise in the research group for its operation and a recent comparison of XCMS with MzMine2 showed similar performance of the two software (Myers et al., 2017). Data processing was carried out within R version 3.3.2. All triplicates were processed separately using the centWave algorithm for peak picking, density for grouping, and obiwrap for alignment, the fill peaks function was also implemented. Parameters shown in Table 37 were used depending on the resolution and chromatographic assay. Other parameters that are not optimised by the IPO software were set as shown in Table 14.

Table 14: XCMS parameters used.

XCMS Parameter	Value
SNR	5
Sigma	3
Prefilter	5, 1000
Minfrac	0.5
Noise	1000
Fitgauss	FALSE
Integrate	1
mzCenterFun	wMean

2.2.7 Theoretical DIA Window Complexity Assessment

Data complexity was assessed using a spreadsheet template tool, previously developed by the author. This complexity assessment tool requires a list of features (m/z and RT pairs) as input, it then automatically provides a visual representation of the complexity of theoretical DIA windows of various widths (5, 25, 50, 100 and 200 m/z) on separate sheets of the template file. Across the x axis each cell represents a transition of 1 second, starting at 30 seconds and running up to 1000 seconds. Down the y axis the cells represent one of the DIA window sizes mentioned earlier. The spreadsheet automatically calculates how many of the features would fall into each cell based on their m/z and RT and are conditionally formatted to represent the number of features present within the cell with each cell representing a theoretical DIA window. Each feature is represented once and can only be in one window. A summary table of the general window complexity for each of the DIA window sizes is provided by the tool also. All processed and RSD filtered datasets for each triplicate were input into the spreadsheet template.

2.2.8 Peak Width Assessment

The width of each peak in each triplicate of data collected at 17,500 MS^1 mass resolution was extracted in R version 3.3.2 utilising the groupval function and the saved xset objects that were saved for each triplicate when XCMS processing was performed.

2.2.9 Scan Rate Estimation

The cycle time was estimated for varying MS^1 and MS^2 mass resolution combinations on a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific). The calculations were based upon the advertised scan rate of 12 Hz at 17,500 resolution. The scan rate is proportional to the resolution therefore if the

resolution doubles the scan rate should half. Using this information the estimated time for seven full scan data points to be collected was calculated using Microsoft Excel as detailed below (Equation 6). MS¹ resolutions of 17,500, 35,000, 70,000 and 140,000 and MS² resolutions of 17,500 and 35,000 were considered. The cycle time was calculated for each of these combinations when applying 0 to 20 fragmentation events between each full scan.

Equation 6: The calculations utilised for cycle time and the time taken for 7 data points to be recorded.

$$\text{Cycle time} = MS^1 \text{ cycle time} + (n \times MS^2 \text{ cycle time})$$

Where n is equal to the number of MS² events per MS¹

$$\text{Time taken for 7 MS}^1 \text{ data points to be collected} = 7 \times \text{Cycle time}$$

2.3 Comparison of different MS² acquisition strategies on the Q Exactive Plus

(Chapter 5.0)

A pooled lithium heparin plasma sample was acquired from BioIVT (West Sussex, UK) which was pooled from 50 individuals of mixed genders. This sample was analysed using a number of different DDA, DIA and AIF methods on a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific, USA) with a Vanquish UHPLC system (Thermo Fisher Scientific, USA). HILIC and Lipidomics assays were applied in both positive and negative ion mode. Materials used are detailed in (Table 15).

Table 15: All materials and associated supplier in brackets utilised for all experiments described in chapter 5.0

Item
Optima UHPLC grade Acetonitrile (Fisher Chemicals)
Optima UHPLC grade water (Fisher Chemicals)
Optima UHPLC grade IPA (Fisher Chemicals)
Ammonium Formate (Fisher Chemicals)
Formic Acid (Fisher Chemicals)
Plasma (BioIVT)
HILIC Accucore Amide 100mm, 2.1mm i.d., 2.6 μ m particle size (Thermo Fisher Scientific)
Hypersil Gold Column 100mm, 2.1mm i.d., 1.9 μ m particle size (Thermo Fisher Scientific)
1.5 mL Eppendorfs (Fisher Chemicals)
LC Vials (Chromatography Direct)

2.3.1 Sample Preparation

The pooled plasma sample was defrosted on ice with all remaining steps done on ice where possible. Once defrosted it was vortexed for 30 seconds. Protein precipitation was performed by 1 in 4 dilution of the sample with acetonitrile (ACN) for the HILIC method and by 1 in 4 dilution of the sample with isopropanol (IPA) for the lipidomics method. Addition of a high percentage of organic solvent disrupts the folding of the proteins in the sample and causes them to aggregate together. Subsequent centrifugation of the sample causes pelleting of the protein aggregates, the supernatant is taken and the proteins have been removed from the sample as is required. 50 μ L of the plasma sample was added to a 1.5 mL Eppendorf tube followed by 150 μ L of ACN or IPA. The Eppendorf tubes were then vortexed for 30 seconds before being centrifuged at 14000 g and 4°C for 15 minutes. The supernatant was aliquoted from each Eppendorf and transferred into LC vials which were placed into the autosampler compartment of the mass spectrometer which was maintained at a temperature of 4°C.

2.3.2 UHPLC-MS

HILIC and lipidomic methods were applied, the solvents preparation and method details will be outlined below. All solvents were sonicated for 20 minutes or until clear before use. A Vanquish UHPLC system (Thermo Fisher Scientific, MA, USA) was utilised to perform all separation and an electrospray Q Exactive Plus Mass Spectrometer (Thermo Fisher Scientific, MA, USA) was utilised for mass analysis.

2.3.2.1 HILIC UHPLC-MS

HILIC samples were analysed using an Accucore Amide HILIC column 100 x 2.1 mm, 2.6 μm (Thermo Fisher Scientific, MA, USA). Mobile phase A consisted of 10mM ammonium formate in 95% ACN/H₂O and 0.1% formic acid. Mobile phase B consisted of 10 mM ammonium formate in 50% acetonitrile/water and 0.1% formic acid. The column temperature was 35 °C and the injection volume was 2 μL . Flow rate was set for 0.50 ml.min⁻¹ with the following gradient: t = 0.0, 1% B; t = 1.0, 1% B; t = 3.0, 15% B; t = 6.0, 50% B; t = 9.0, 95% B; t = 10.0, 95% B; t = 10.5, 1% B; t = 14.0, 1% B, all changes were linear with curve = 5 (Figure 19). Data were acquired in positive and negative ionisation modes separately within the mass range of 70–1000 m/z at resolution 70,000 (FWHM at m/z 200). Ion source parameters were set as follows: Sheath gas = 53 arbitrary units, Aux gas = 14 arbitrary units, sweep gas = 3 arbitrary units, Spray Voltage = 3.5 kV, Capillary temp. = 269 °C, Aux gas heater temp. = 438 °C.

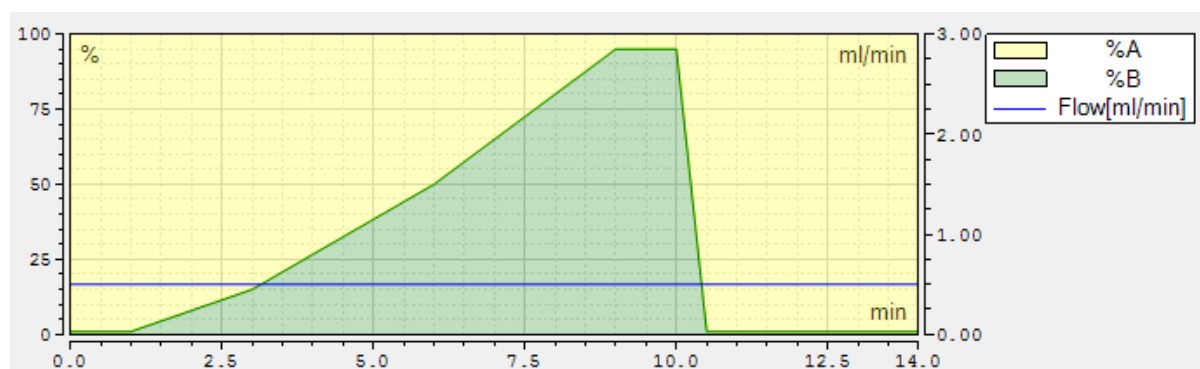


Figure 19: The HILIC chromatographic gradient applied in these studies.

A variety of HCD MS² methods were implemented. Settings which were consistent between all methods employed were, MS¹ resolution = 70,000 (FWHM at m/z 200), MS² resolution = 17,500 (FWHM at m/z 200); stepped normalised collision energies (stepped NCE) = 20, 40, 60%, AGC Target = 1e⁶, Max IT = 100 ms, Microscans = 1. All DDA based methods employed an isolation width of 3.0 m/z and operated in “discovery mode”. DIA methods employed variable window sizes over variable ranges. Method details are outlined in Table 16. Each method was employed in quadruplicate.

Table 16: Details of the MS² strategies employed in each quadruplicate for the HILIC methods. Where progressive exclusion lists were applied the sample was injected 3 times with the exclusion list updated twice.

Name	MS ² Type	Number of windows /Top "n"	Estimated Time for 7 data Points	Window Size (m/z)	Window Range /DDA Range
DIA_2_10	DIA	2	3.5	10	200 - 220
DIA_2_25	DIA	2	3.5	25	200 - 250
DIA_2_50	DIA	2	3.5	50	200 - 300
DIA_2_100	DIA	2	3.5	100	200 - 400
DIA_4_10	DIA	4	4.67	10	200 - 240
DIA_4_25	DIA	4	4.67	25	200 - 300
DIA_4_50	DIA	4	4.67	50	200 - 400
DIA_4_100	DIA	4	4.67	100	200 - 600
DIA_4_232	DIA	4	4.67	232.5	70-1000
AIF_430	AIF	1	2.92	430	70 - 500
AIF_930	AIF	1	2.92	930	70 - 1000
Tra	DDA	2	3.5	N/A	70 - 1000

Tra_exc	iDDA (Single exclusion list)	2	3.5	N/A	50 - 1000
Tra_inc	iDDA (Single inclusion list)	2	3.5	N/A	50 - 1000
Tra_p_total	iDDA (Progressive exclusion list)	2	3.5	N/A	50 - 1000
Mull_total	iDDA (Mullard/Progressive Exclusion List)	2	3.5	N/A	50 – 200 (Mull1)
		2	3.5	N/A	200 – 400 (Mull2)
		2	3.5	N/A	400 - 1000 (Mull3)

2.3.2.2 Lipidomics UHPLC-MS

Lipidomics samples were analysed using a Hypersil GOLD column (100 × 2.1 mm, 1.9 µm; Thermo Fisher Scientific, MA, USA). Mobile phase A consisted of 10mM ammonium formate in 60% ACN/H₂O and 0.1% formic acid. Mobile phase B consisted of 10 mM ammonium formate in 90% IPA/ACN and 0.1% formic acid. The column temperature was 55 °C and the injection volume was 2 µL. Flow rate was set for 0.50 ml.min⁻¹ with the following gradient: t = 0.0, 45% B; t = 0.5, 45% B; t = 8.5, 100% B; t = 9.5, 100% B; t = 11.5, 45% B; t = 14.0, 45% B, all changes were linear with curve = 5 (Figure 20) Data were acquired in positive and negative ionisation modes separately within the mass range of 200–1600 m/z at resolution 70,000 or 140,000 (FWHM at m/z 200) depending on the method applied, details found in (Table 17). Ion source parameters were set as follows: Sheath gas = 50 arbitrary units, Aux gas = 13 arbitrary units, sweep gas = 3 arbitrary units, Spray Voltage = 3.5 kV, Capillary temp. = 263 °C, Aux gas heater temp. = 425 °C.



Figure 20: Lipidomics chromatographic gradient.

A variety of HCD MS² methods were implemented. Settings which were consistent between all methods employed were: MS² mass resolution = 17,500 (FWHM at m/z 200); stepped normalised collision energies (stepped NCE) = 20, 40, 60%, AGC Target = $1e^6$, Max IT = 100 ms, Microscans = 1. All DDA based methods employed an isolation width of 3.0 m/z and operated in “discovery mode”. MS¹ mass resolution applied was variable depending on the method applied. DIA methods employed variable window sizes over variable ranges. All other method details are outlined in Table 17. Each strategy was employed in quadruplicate.

Table 17: Details of the MS² strategies employed in each quadruplicate for the lipidomics methods. Where progressive exclusion lists were applied the sample was injected 3 times with the exclusion list updated twice.

Name	MS/MS Type	Full Scan Resolution	Number of windows/Top "n"	Estimated Time for 7 data Points	Window Size (m/z)	Window Range /DDA Range (m/z)
70K_DIA_3_10	DIA	70,000	3	4.08	10	750 – 780
70K_DIA_3_25	DIA	70,000	3	4.08	25	750 – 825
70K_DIA_3_50	DIA	70,000	3	4.08	50	750 – 900

70K_DIA_3_100	DIA	70,000	3	4.08	100	750 – 1050
140K_DIA_3_10	DIA	140,000	3	6.42	10	750 – 780
140K_DIA_3_25	DIA	140,000	3	6.42	25	750 – 825
140K_DIA_3_50	DIA	140,000	3	6.42	50	750 – 900
140K_DIA_3_100	DIA	140,000	3	6.42	100	750 – 1050
70K_DIA_6_10	DIA	70,000	6	5.83	10	750 – 810
70K_DIA_6_25	DIA	70,000	6	5.83	25	750 – 900
70K_DIA_6_50	DIA	70,000	6	5.83	50	750 – 1050
70K_DIA_6_100	DIA	70,000	6	5.83	100	750 – 1350
70K_DIA_8_175	DIA	70,000	8	7	175	200 - 1600
70K_AIF_770	AIF	70,000	1	2.92	770	200 – 970
70K_AIF_1400	AIF	70,000	1	2.92	1400	200 - 1600
140K_AIF_770	AIF	140,000	1	5.25	770	200 – 970
140K_AIF_1400	AIF	140,000	1	5.25	1400	200 - 1600
NA	Full Scan	70,000	0	2.33	N/A	200 - 1600
NA	Full Scan	70,000	0	2.33	N/A	200 - 1600
NA	Full Scan	140,000	0	4.67	N/A	200 - 1600
NA	Full Scan	140,000	0	4.67	N/A	200 - 1600
70K_Tra	DDA	70,000	3	4.08	N/A	200 - 1600
140K_Tra	DDA	140,000	3	6.42	N/A	200 - 1600
70K_Tra_exc	DDA (Single exclusion list)	70,000	3	4.08	N/A	200 - 1600

70K_Tra_inc	DDA (Single inclusion list)	70,000	3	4.08	N/A	200 - 1600
70K_Tra_p_total	DDA (Progressive exclusion list)	70,000	3	4.08	N/A	200 - 1600
70K_Mull_total	DDA (Mullard/Progressive Exclusion List)	70,000	3	4.08	N/A	200-500 (Mull1)
		70,000	3	4.08	N/A	500-800 (Mull2)
		70,000	3	4.08	N/A	800-1600 (Mull3)

2.3.5 Generation of Exclusion Lists

All .RAW data files were converted to .mzML format as described in section 2.2.4 before further steps were taken to generate the lists.

2.3.5.1 Exclusion of blank related peaks from DDA triggering

The blank exclusion lists were created using the data for the blank quadruplicates from each assay. The .mzML files were processed in R version 3.2.2 using the optimised XCMS parameters for the relevant assay and resolution combination which can be found in section 2.2.6 except fill peaks was not employed. The capacity of the exclusion list is 5000 entries. After processing, the highest intensity 5000 features were added to the exclusion list, if more than 5000 were present the 5000 of highest intensity were added to the exclusion list. If less than 5000 were present, they were all added to the exclusion list. No RT data were included on the list. If being added to a progressive exclusion list method, then only the top 500 were utilised to ensure space was left for fragmented features after each pass of the progressive method.

2.3.5.2 Progressive Exclusion Lists

Each m/z value that had been fragmented during the relevant analysis was extracted from the relevant files scan header information in R version 3.2.2 using the "PrecursorMZ" value provided when using the "header" function from the package mzR (Chambers et al., 2012). The relevant .mzML files from the preceding injection and appropriate replicate for each condition were used. For example, all the m/z values which were fragmented during the first pass of Mull1, injection 1, can all be extracted from

the scan header information in an automated fashion for each MS² scan that occurred for that file. For the progressive traditional DDA methods the first exclusion was made from the regular traditional DDA method quadruplicate and so on. All m/z values were to four decimal places and any duplicated m/z values were removed from the list within Excel using the remove duplicates option from the data tools section within the data tab on the toolbar. The remaining m/z values were added onto the relevant method in the sequence. In some cases not all the fragmented m/z values could fit into the exclusion list and so in these cases the values which did not fit were left out. No RT data were considered or included.

2.3.6 Generation of Inclusion Lists

The full scan plasma quadruplicates were processed in XCMS using R version 3.2.2. using the optimised parameters for the relevant assay/resolution combination as shown in Table 37 in section 4.2.1, other parameters not detailed in this table were as detailed in Table 14. The intensity of the features across the four replicates were averaged and the 5000 m/z values of highest average intensity were added to the inclusion list. The m/z values were all rounded to four decimal places and no RT information was included.

2.3.7 Data Processing

2.3.7.1 XCMS

Peak picking, grouping and alignment was carried out for each quadruplicate using the same optimised parameters for the relevant assay/resolution combination as detailed in Table 37 in section 4.2.1 other parameters not detailed in this table were as detailed in Table 14.

2.3.7.2 MS-DIAL

Peak picking, grouping and alignment was carried out in MS-DIAL (Tsugawa et al., 2015) version 3.12 as it represented the only software package capable of processing and deconvoluting DIA data at the time the experiments were performed. Files were converted to .abf format using the ABF converter software provided with the MS-DIAL software download. Parameters used were as detailed in Figure 21, Figure 22, Figure 23, Figure 24, Figure 25 as well as Table 18 and Table 19. Each quadruplicate was processed separately. The method for selection of which file of the quadruplicate to align to was based on the XCMS processed results generated as described in 2.3.7.1. The intensity of all features was summed for each replicate, the average total intensity across the four replicates was calculated and the replicate with total intensity closest to the average was selected. For the HILIC data the MS² spectra used for identification were downloaded in .msp format from the MS-DIAL website (RIKEN, no date). For both positive and negative ion modes the .msp files containing all publicly available records were downloaded (1/10/2018). This includes spectra from MassBank (Horai et al., 2010), MassBankEU

(MassBank Consortium, no date), ReSpect (Sawada et al., 2012), GNPS (Wang et al., 2016), Fiehn lab HILIC library (Kind et al., 2009), CASMI 2016 (Schymanski et al., 2017) and RIKEN PlaSMA (Sakurai et al., 2013). For the lipidomics data, the project was set up utilising the lipidomics set up option in the software. As a result the LipidBlast (Kind et al., 2013) library was automatically used.

Mass accuracy

MS1 tolerance: Da

MS2 tolerance: Da

⬆ Advanced

Data collection parameters

Retention time begin: min

Retention time end: min

Mass range begin: Da

Mass range end: Da

Isotope recognition

Maximum charged number:

Multithreading

Number of threads:

RetentionTime Correction ☐

Figure 21: Parameters used for MS-DIAL processing in the data collection tab

Peak detection parameters

Minimum peak height: amplitude

Mass slice width: Da

⬆ Advanced

Smoothing method:

Smoothing level: scan

Minimum peak width: scan

Figure 22: Parameters used for MS-DIAL processing in the peak detection tab

Deconvolution parameters

Sigma window value:

MS/MS abundance cut off: amplitude

⬆ Advanced

Exclude after precursor ion: ☒

Keep the isotopic ions until: Da

Keep the isotopic ions w/o MS2Dec: ☐

Figure 23: Parameters used for MS-DIAL processing in the MS2Dec tab

Retention time tolerance: min

Accurate mass tolerance (MS1): Da

Accurate mass tolerance (MS2): Da

Identification score cut off: %

Use retention information for scoring: ☐

Figure 24: Parameters used for MS-DIAL processing in the identification tab

Retention time tolerance: min

MS1 tolerance: Da

⬆ Advanced

Retention time factor: (0-1)

MS1 factor: (0-1)

Peak count filter: %

N% detected in at least one group: %

Detected in all QCs ☐

Remove features based on blank information: ☐

Sample max / blank average: fold change

Keep 'identified' metabolite features: ☒

Keep 'annotated (wo MS2)' metabolite features: ☐

Keep removable features and assign the tag: ☒

Figure 25: Parameters used for MS-DIAL processing in the alignment tab

Table 18: The adducts searched for in positive ion mode MS-DIAL processing.

$[M + H]^+$	$[M + 2ACN + H]^+$	$[M + H + NH_4]^{2+}$
$[M + NH_4]^+$	$[M - C_6H_{10}O_4 + H]^+$	$[M + H + Na]^{2+}$
$[M + Na]^+$	$[M - C_6H_{10}O_5 + H]^+$	$[M + H + K]^{2+}$
$[M + CH_3OH + H]^+$	$[M - C_6H_8O_6 + H]^+$	$[M + ACN + 2H]^{2+}$
$[M + K]^+$	$[2M + H]^+$	$[M + 2Na]^{2+}$
$[M + Li]^+$	$[2M + NH_4]^+$	$[M + 2ACN + 2H]^{2+}$
$[M + ACN + H]^+$	$[2M + Na]^+$	$[M + 3ACN + 2H]^{2+}$
$[M + H - H_2O]^+$	$[2M + 3H_2O + 2H]^+$	$[M + 3H]^{3+}$
$[M + H - 2H_2O]^+$	$[2M + K]^+$	$[M + 2H + Na]^{3+}$
$[M + 2Na - H]^+$	$[2M + ACN + H]^+$	$[M + H + 2Na]^{3+}$
$[M + ACN + Na]^+$	$[2M + ACN + Na]^+$	$[M + 3Na]^{3+}$
$[M + 2K - H]^+$	$[M + 2H]^{2+}$	

Table 19: The adducts searched for in negative ion mode MS-DIAL processing.

$[M - H]^-$	$[M + FA - H]^-$	$[2M - H]^-$
$[M - H_2O - H]^-$	$[M + Br]^-$	$[2M + FA - H]^-$
$[M + Na - 2H]^-$	$[M + C_6H_{10}O_4 - H]^-$	$[3M - H]^-$
$[M + Cl]^-$	$[M + C_6H_{10}O_5 - H]^-$	$[M - 2H]^{2-}$
$[M + K - 2H]^-$	$[M + C_6H_8O_6 - H]^-$	$[M - 3H]^{3-}$

2.3.7.3 msPurity

The purityA function was utilised within the msPurity package (Lawson et al., 2017) in R version 3.2.2. The default parameters were utilised except for mostIntense which was set as false for DDA data and true for DIA data and nearest which was set as true for DDA and DIA data.

2.3.8 Data Analysis

2.3.8.1 Number of features detected

The number of features detected for each quadruplicate was determined without any further filtering after MS-DIAL processing was finished. Each of the four replicates was selected sequentially using the file navigation pane (Figure 26) in the MS-DIAL graphical user interface (GUI). After selection the number of unfiltered peak spots can be seen in the peak spot navigator (Figure 27). This value was

recorded for each replicate and the standard deviation was calculated. For the Mullard method total, the average for each of the three sections were summed to give the total displayed in Figure 69, section 5.2.1.

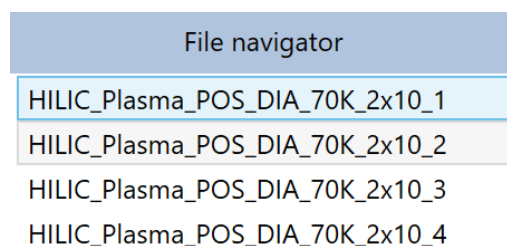


Figure 26: The file navigator in the MS-DIAL GUI.

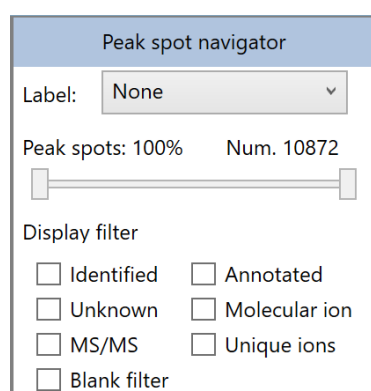


Figure 27: The peak spot navigator in the MS-DIAL GUI.

2.3.8.2 Number of Features with MS² Data

The number of features with MS² data were collected using the same method as described in 2.3.8.1 except the “MS/MS” tick box (Figure 27) was checked prior to recording the feature numbers. The total for the traditional progressive DDA method was calculated as described for the Mullard method in 2.3.8.1.

2.3.8.3 Number of Features Annotated

The number of features annotated with a score ≥ 0 were collected using the same method as described in 2.3.8.2 but the “Annotated” checkbox was checked as well as the MS/MS check box (Figure 27). The same was done for the number of features annotated with a score ≥ 70 but the Identified check box (Figure 27) was checked too. The percentage of features annotated with a score ≥ 70 (Figure 73) was calculated by taking the average just described and dividing by the average number of features with MS² data as described in 2.3.8.2.

2.3.8.4 Purity of Fragmentation Windows

The interpolated purity score for each fragmentation event in each file was extracted from the tables resulting from msPurity processing. Every fragmentation event was plotted in violin plots using the `r` package `ggplot2`.

2.3.8.5 Value of repeated Injections

Data represented in the Mullard totals and traditional DDA progressive totals which had been calculated as described in 2.3.8.2 and 2.3.8.3 were separated and displayed in their individual sections.

2.3.8.6 MS-DIAL Deconvolution Assessment

A single replicate was selected from each quadruplicate for the assessment of deconvolution. The replicate was selected after XCMS processing of the data. The intensity of all features was summed for each replicate, the average total intensity across the four replicates was calculated and the replicate with total intensity closest to the average was selected. This was also used to determine which replicate should be used for alignment in MS-DIAL processing. The appropriate replicate was selected using the file navigator (Figure 26). The features were filtered by intensity using the slider bar in the peak spot navigator (Figure 27) for low (blue/green spots) or high intensity (orange/red spots) features to be left displayed in the peak spot viewer (Figure 28). This was done first using the 2 x 10 m/z DIA window for HILIC or the 3 x 10 m/z DIA window method for lipidomics as these were the methods that would limit the features available for comparison. A high intensity feature and a low intensity feature were selected which had a suitable reference spectrum match. Suitable meant in this case multiple peaks matching between the spectra instead of a single peak. Once an appropriate feature had been selected it was searched for in the appropriate replicate from each quadruplicate, the spectra were extracted and the dot product and reverse dot product scores were recorded.

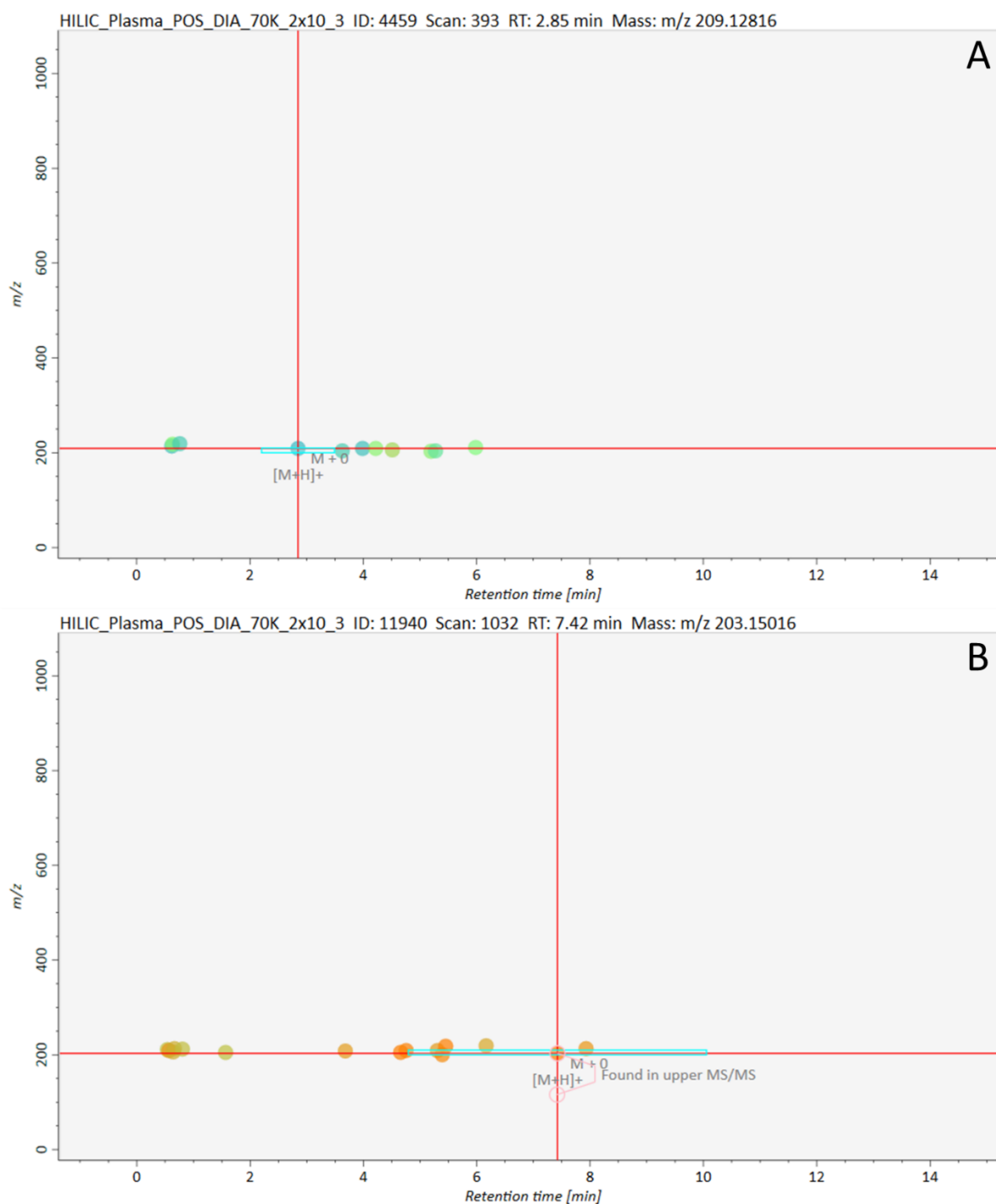


Figure 28: A) The peak spot viewer for HILIC_Plasma_POS_DIA_2_10_rep3 after filtering for low intensity identified features B) The peak spot viewer for HILIC_Plasma_POS_DIA_2_10_rep3 after - filtering for high intensity identified features. The x-axis represents RT (mins) and the y-axis represents m/z.

2.4 Assessment of Metabolite Annotation Using AcquireX on the Orbitrap ID-X

(Chapter 6.0)

All data collected in this chapter were acquired within a 2-week period at Thermo Fisher Scientific factory in San Jose, USA. Therefore, not all parts of this work could be carried out in an optimal fashion due to the inherent time constraints. All analyses were carried out using a Vanquish 2000 UHPLC system (Thermo Fisher Scientific, USA) in tandem with an electrospray Orbitrap ID-X (Thermo Fisher Scientific, USA) mass spectrometer. 875 standards were prepared in 50 groups of varying numbers of metabolites. HCD MS² data were collected in both positive and negative ion mode using an aqueous C₁₈ reversed-phase assay (from here onwards referred to as a RP method) and a HILIC assay. Materials utilised are detailed in Table 20.

Table 20: All materials and associated supplier in brackets utilised for all experiments described in chapter 6.0

Materials
Optima UHPLC Grade Acetonitrile (Fisher Chemicals)
Optima UHPLC Grade Water (Fisher Chemicals)
Optima UHPLC Grade Isopropanol (Fisher Chemicals)
Optima UHPLC Grade Methanol (Fisher Chemicals)
UHPLC-MS Grade Formic Acid (Fisher Chemicals)
UHPLC-MS Grade Acetic Acid (Fisher Chemicals)
Ammonium Formate (Fisher Chemicals)
Ammonium Acetate (Fisher Chemicals)
HILIC Accucore Amide 100mm, 2.1mm i.d., 2.6 µm particle size (Thermo Fisher Scientific) (16726-102130)
Hypersil Gold aQ Column 100mm, 2.1mm i.d., 1.9 µm particle size (Thermo Fisher Scientific) (25302-102130)
Hypersil Gold Column 100mm, 2.1mm i.d., 1.9 µm particle size (Thermo Fisher Scientific) (25002-102130)
MetaSci COMPLETE Library (MetaSci)
NIST SRM Plasma 1950 (NIST)
NIST SRM Non-Smokers Urine 3673 (NIST)
NIST SRM Smokers Urine 3672 (NIST)
LC vials (Fisher Chemicals)
7 mL Glass Vials (Fisher Chemicals)
1.5 mL Eppendorfs (Fisher Chemicals)

2.4.1 Assigning Standard Groups

The MetaSci COMPLETE human standards library contains 1027 standards and was kindly provided by Thermo Fisher Scientific. It was suitable for use as it represents a variety of metabolites normally present in human biofluids. To analyse them all individually was not feasible with the given time constraints. As a result, they were divided into groups of approximately 20 metabolites with the goal to analyse as many as possible. Groups were created manually within Microsoft Excel using the information in the reference file provided by MetaSci and the following rules. Groups were approximately sorted by hydrophobicity of the standards within with a self-imposed m/z difference

restriction of at least 4 m/z where possible between all metabolites within a single group. This was maintained in the majority of cases for all metabolites and groups. Any metabolites with a molecular mass of less than 100 were not included, all other metabolites were included and considered of equal importance. 50 groups were created in total. The details of the standards present in each group are available in the electronic Appendix (2.4/2.4.1). Groups were labelled as either hydrophobic (PHO), hydrophilic (PHI), a mixture (MIX) of the two or liquid (LIQ) for the small selection of standards provided in liquid instead of solid form.

2.4.2 Sample Preparation

2.4.2.1 Chemical standards Preparation

All standards in a single group were weighed out and placed into a 20 mL glass vial. The masses added for all solid-state standards ranged between 0.6 mg and 2.8 mg, exact amounts added can be found in the electronic Appendix file (2.4/2.4.1). Once added they were resuspended in 2 mL of either 50:50 ACN:H₂O (for groups assigned as hydrophilic or mixtures) or 50:40:10 MeOH:H₂O:IPA (for groups assigned as hydrophobic). The solution was then vortexed for 1 minute and sonicated for 10 minutes. The solution was vortexed again for 1 minute and then diluted 1:5 into either of the following 6:2:2 ACN:MeOH:H₂O for the HILIC assay or 50:50 MeOH:H₂O for the RP assay. The solution was then centrifuged at 14000 g for 20 mins at 4°C. 200 µL of the supernatant was taken and transferred into a LC vial and stored at 4°C ready for analysis. Some standards were in liquid form, for these standards 1 µL was utilised, some were also viscous, 1 µL was attempted to be added for these but the true volume added was unknown. Liquid groups were treated the same as the hydrophilic and mixture groups.

2.4.2.2 Biological Sample Preparation

Stock plasma and urine samples were vortexed for 30 seconds, 50 µL of sample was diluted with 150 µL of solvent. ACN was used for HILIC and RP and IPA was used for Lipidomics. The 1.5 mL Eppendorfs containing the sample and solvent mixtures were vortexed for 30 seconds and then centrifuged at 14,000 xg and 4°C for 20 minutes. 200 µL of supernatant was extracted and transferred to a LC vial. Samples were then dried down and reconstituted in 200 µL of a mixture of 60:20:20 ACN:MeOH:H₂O for HILIC, 50:40:10 MeOH:H₂O:IPA for RP and 60:20:20 IPA:ACN:H₂O for lipids.

2.4.3 UHPLC-MS

2.4.3.1 HILIC UHPLC-MS

The gradient for the HILIC method is outlined in Figure 29. The mobile phase was different for positive and negative ion modes otherwise all parameters were the same. ESI parameters were the same for all analyses and are displayed in Table 21. A dummy AcquireX sequence was set up to allow easy generation of a blank exclusion list that could be added to the DDA method that would be used for

analysis of the standards. Full scan analysis of a solvent blank was carried out using this sequence which subsequently automatically updated the DDA method. The sequence was cancelled after the exclusion list had been generated. The parameters for the full scan analysis are detailed in Table 21 and Table 22. The DDA method parameters are displayed in Table 23.

Column: Accucore Amide-HILIC (100mm x 2.1 mm, 2.6 μ m)

Scan Range: 70 – 1050 m/z

Injection volume: 2 μ L

Flow Rate: 0.5 mL/min

Mobile phase positive mode:

- A) 10 mM Ammonium formate 95% ACN + 0.1% Formic acid
- B) 10 mM Ammonium formate 50% ACN/Water + 0.1% Formic acid

Mobile phase negative mode:

- A) 10 mM Ammonium acetate 95% ACN + 0.1% Acetic Acid
- B) 10 mM Ammonium acetate 50% ACN/Water + 0.1% Acetic Acid

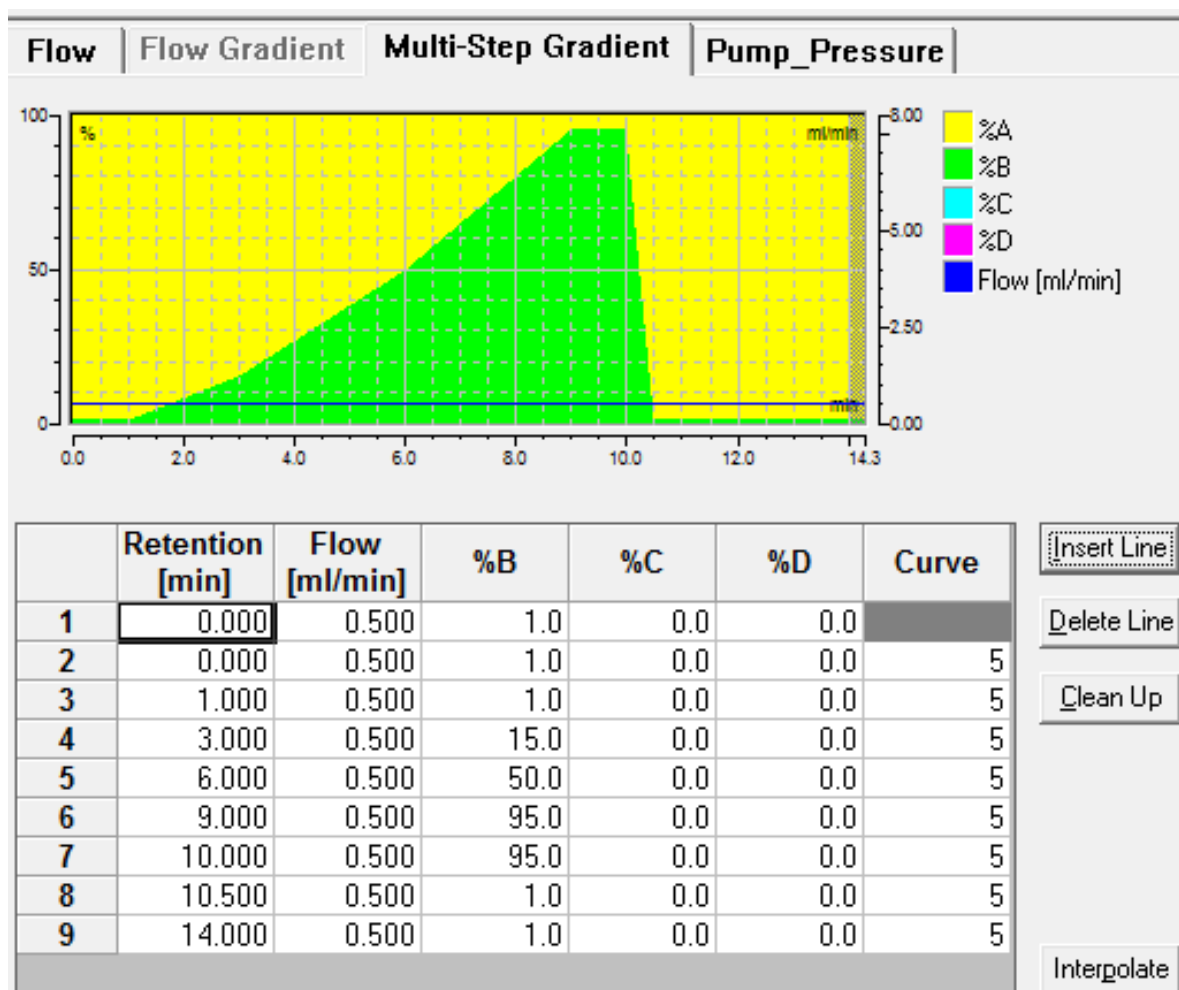


Figure 29: The gradient utilised for the HILIC method.

Full scan analyses were carried out to generate exclusion or inclusion lists during AcquireX sequences with parameters as detailed in Table 21 and Table 22.

Table 21: ESI source settings utilised for HILIC analyses.

Parameter	Value
Spray Voltage (kV)	3.2 (positive ion mode), 2.7 (negative ion mode)
Sheath Gas (Arb)	50
Aux Gas (Arb)	15
Sweep Gas (Arb)	1
Ion Transfer Tube Temperature (°C)	275
Vaporizer Temperature (°C)	320

Table 22: HILIC MS¹ parameters.

Parameter	Value
Detector	Orbitrap
MS ¹ Resolution (FWHM at 200 <i>m/z</i>)	120,000
Quadrupole Isolation	True
Scan Range (<i>m/z</i>)	70 – 1050
RF Lens (%)	30
AGC Target	1e ⁵
Maximum Injection Time (ms)	50
Microscans	1
Data Type	Profile
Source Fragmentation	Disabled
Use EASY-IC™	True

For MS² analyses the global parameters were as detailed in Table 21. MS¹ parameters were the same as detailed in Table 22 except the MS¹ resolution was set to 60k. The MS² parameters are displayed in Table 23. Dynamic exclusion was applied after an *n* of 2 within 5 seconds for a duration of 3.5 seconds. Cycle time was fixed instead of setting top “*n*”. Where exclusion/inclusion lists were applied an error of 10 ppm was set to determine exclusion/inclusion. AcquireX modifications to the method were enabled during AcquireX sequences when collecting biological data. Structure of the workflow was outlined in Figure 30.

Table 23: HILIC HCD MS² parameters.

Parameter	Value
Detector	Orbitrap
MS ² Resolution (FWHM at 200 <i>m/z</i>)	30,000
Quadrupole Isolation	True
Scan Range (<i>m/z</i>)	70 – 1050
Intensity Threshold	2.0e ⁴
AGC Target	5.0e ⁴
Inject Ions for All Available Parallelizable Time	True
Maximum Injection Time (ms)	54
Microscans	1
Data Type	Profile
Use EASY-IC™	True
Cycle Time (s)	0.7
Collision Energy Mode	Stepped
HCD Collision Energy (%)	20, 40, 60
Isolation Offset	Off
Isolation Window (<i>m/z</i>)	1.5
First Mass (<i>m/z</i>)	50

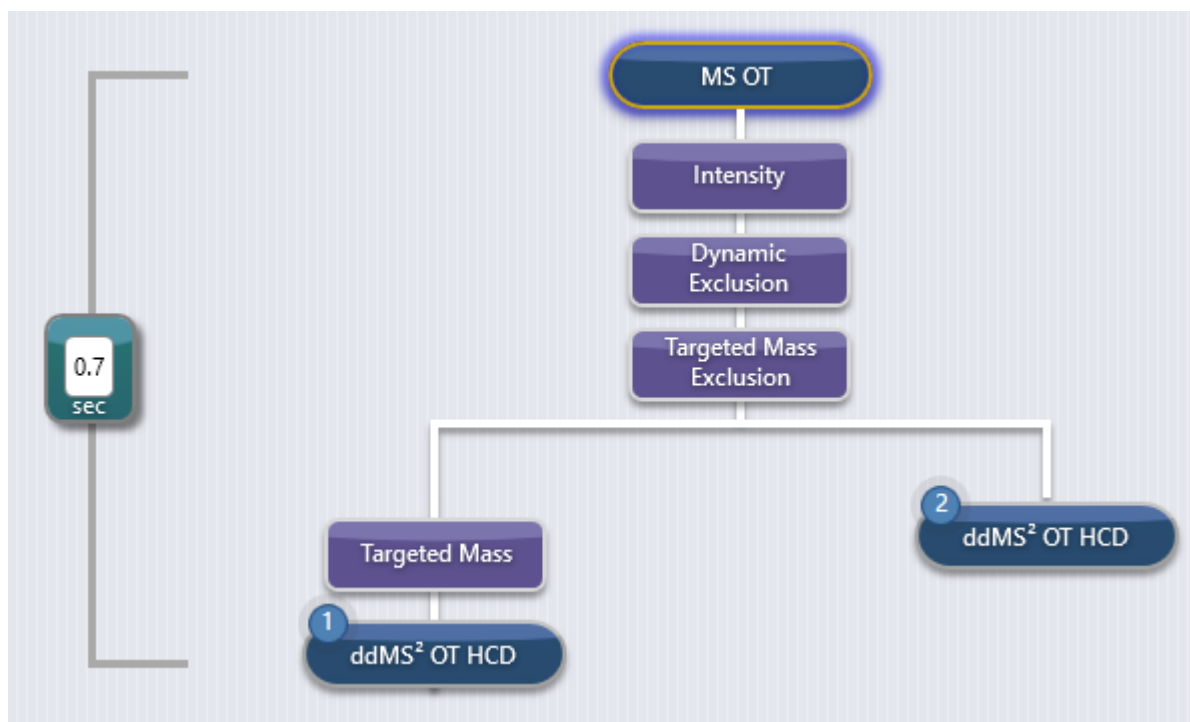


Figure 30: Structure of the HCD DDA workflow as viewed in the method Orbitrap ID-X method editing software.

2.4.3.2 RP UHPLC-MS

The gradient utilised for the RP method is displayed in Figure 31. A blank exclusion list was generated in the same manner as described in 2.4.3.1. MS parameters for the full scan analysis were as displayed in Table 21 and Table 22 except the following; sheath gas = 40, aux gas = 8, sweep gas = 0, scan range = 100-1500 m/z . All other full scan analyses used these settings. The exclusion list was automatically generated and added to the DDA method for which all parameters and workflow structure were the same as described in 2.4.3.1.

Column: Hypersil GOLD aQ (100 x 2.1 mm, 1.9 μm)

Scan Range: 100 – 1500 m/z

Injection volume: 2 μL

Flow rate: 0.3 mL/min (0.4 mL/min during equilibration)

Mobile Phase:

- A) Water + 0.1% Formic Acid
- B) Methanol + 0.1% Formic Acid

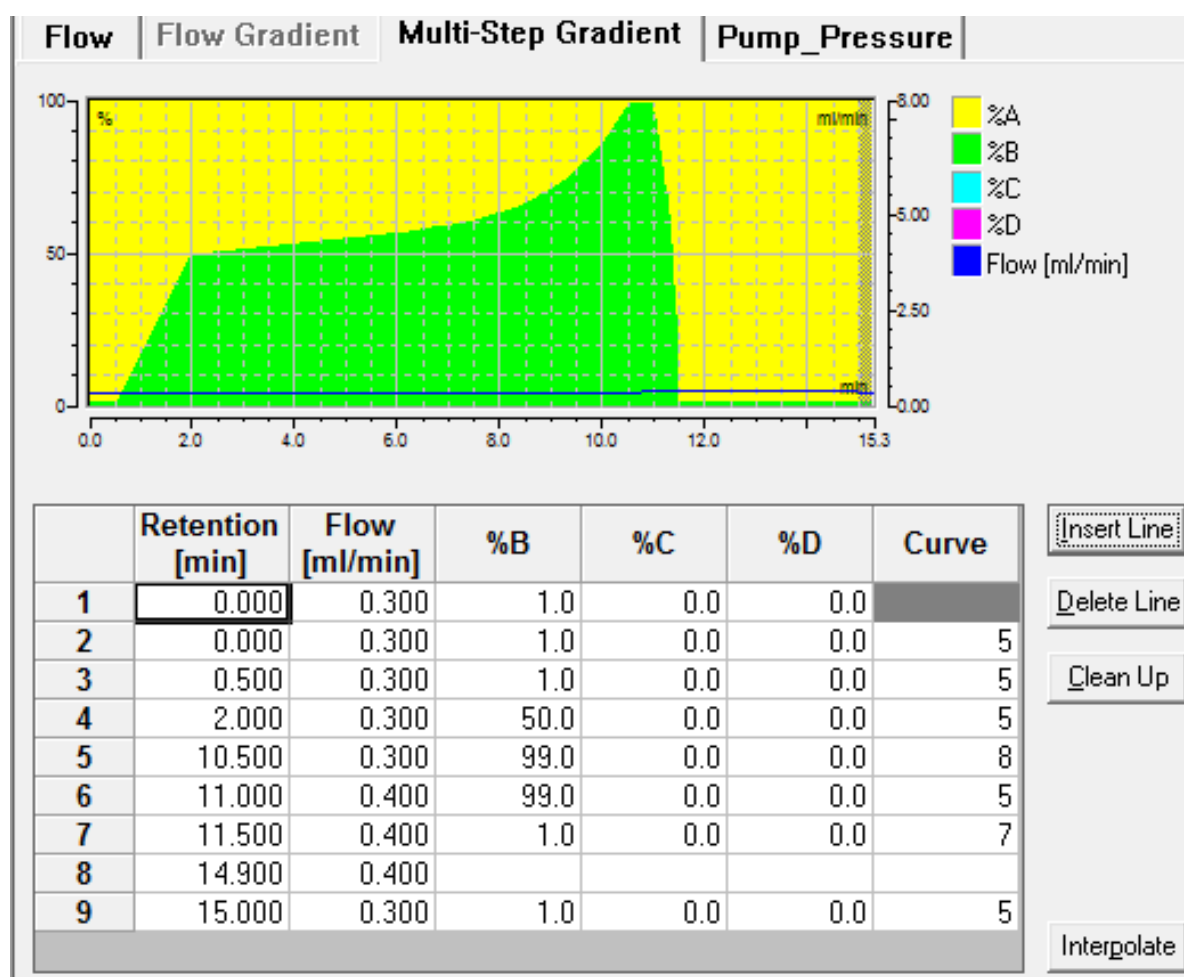


Figure 31: The gradient utilised for the RP method.

2.4.3.3 Lipidomics UHPLC-MS

The gradient utilised for the Lipidomics method is displayed in Figure 32. MS¹ parameters were the same as in Table 21 and Table 22, except the following; sheath gas = 45, sweep gas = 0, scan range = 150-2000 *m/z*. For MS² analyses global parameters and scan range were the same as described for the Lipidomics full scan all other remaining parameters and workflow structure were the same as in 2.4.3.1.

Column: Hypersil Gold column (100mm x 2.1 mm, 1.9µm)

Scan Range: 150 – 2000 *m/z*

Injection Volume: 2 µL

Flow Rate: 0.4 mL/min

Mobile Phase:

A) 60% ACN/H₂O + 10mM Ammonium Formate + 0.1% formic acid

B) 90% IPA/ACN + 10mM Ammonium Formate + 0.1% formic acid

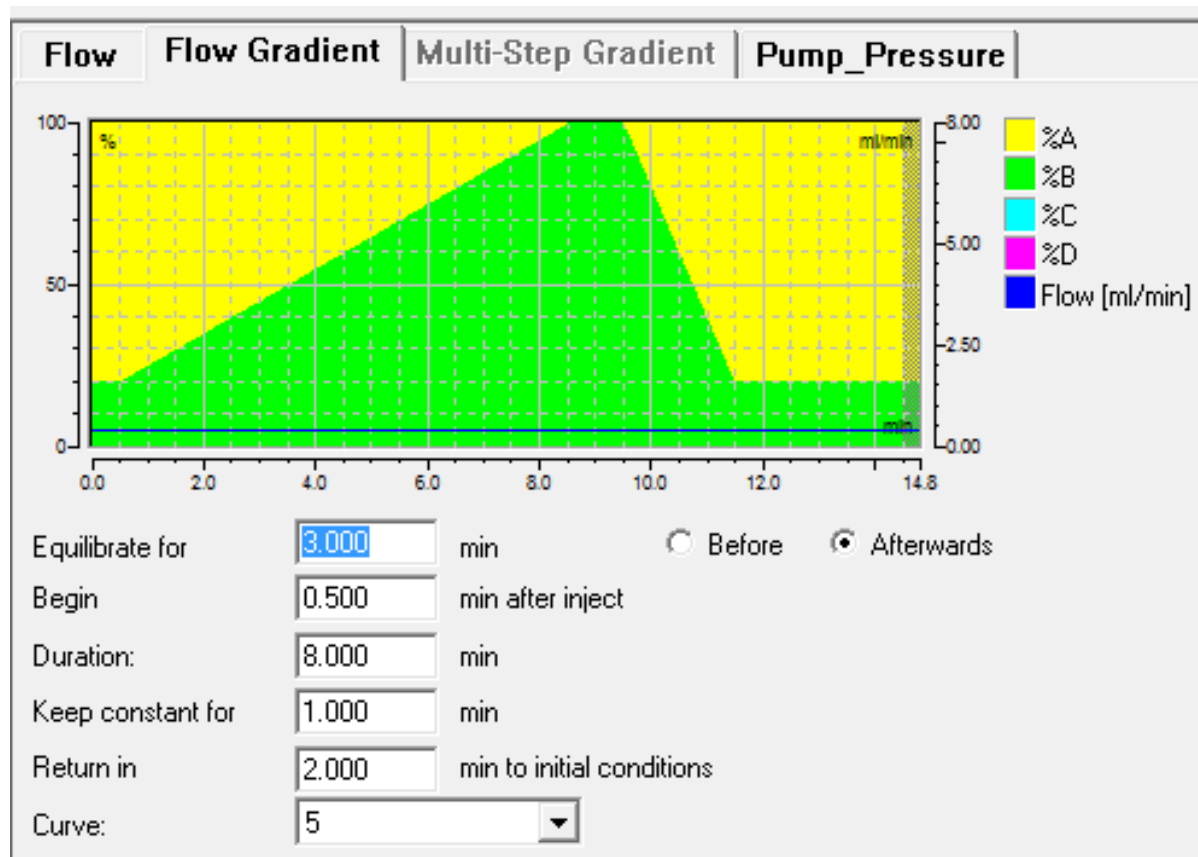


Figure 32: Gradient utilised for Lipidomics analyses.

2.4.4 Acquisition of Data for Metabolite Standards

Data were acquired for the 50 groups detailed in the electronic Appendix (2.4/2.4.1) in both positive and negative ion mode for HILIC and RP. HCD MS² data were collected using the methods described in 2.4.3.1 and 2.4.3.2. An AcquireX sequence (Table 24) was set up using the relevant full scan method to automatically generate a DDA method with a blank exclusion list. This method was then used to analyse all 50 standard groups for the assay in question using a normal sequence. RP positive and negative ion mode were carried out in a single sequence. A separate vial was used for the two injections required in RP with two blanks acquired after every 20 injections. For HILIC positive and negative ion modes were carried out in separate sequences as they utilise different mobile phases. A separate vial was used for each injection.

Table 24: AcquireX sequence used to generate the blank exclusion for analysis of standards.

Injection Number	File	Method
1	Conditioning Blank	Full Scan
2	Exclusion List Blank	Full Scan

2.4.5 Acquisition of Data for Biological Samples

Data for biological samples were acquired using the AcquireX sequence function. A separate AcquireX sequence was submitted for each sample type, assay, and ion mode combination. Sequence structure was as shown in Table 25. Three sample types were analysed, plasma, smokers urine and non-smokers urine. HILIC, RP and Lipidomics assays were applied with HCD MS² collected in both positive and negative ion modes as described in 2.4.3. A single replicate of the sequence was acquired for each combination except for the lipidomics where an *n* of 2 was applied.

Table 25: AcquireX sequence structure used for acquisition of biological data.

Injection Number	File	Method
1	Conditioning Blank	Full Scan
2	Exclusion List Blank	Full Scan
3	Inclusion List Sample	Full Scan
4	ID_01 (1st Exclusion)	HCD MS ²
5	ID_02 (2nd exclusion)	HCD MS ²
6	ID_03 (3rd exclusion)	HCD MS ²
7	ID_04 (4th exclusion)	HCD MS ²
8	ID_05 (5th exclusion)	HCD MS ²
9	ID_06 (6th exclusion)	HCD MS ²
10	Tra_DDA_01	HCD MS ²
11	Tra_DDA_02	HCD MS ²
12	Tra_DDA_03	HCD MS ²

2.4.5 mzVault MS² Library Construction

All data were processed within Compound Discoverer 3.0 (CD3.0) (Thermo Fisher Scientific, USA). The following nodes were used in the workflow, Select Spectra (Figure 33), Detect Compounds (Figure 34), Group Compounds (Figure 35), Search MassList (Figure 36), Search MzCloud (Figure 37) and Assign Compound Annotations (Figure 38). A mass list containing mol structure files for all standards analysed

was created and set as the first priority for annotation assignments. This allowed easy identification of the relevant features in each file if they had been detected. Each standards group/method/polarity combination was processed separately. The resulting compound lists were then browsed manually. When a standard had been annotated the check box adjacent to it was ticked. If MS² spectra had been acquired they were inspected manually for quality. The mzCloud match score if the standard was already present in mzCloud was checked and quality was further assessed by applying the fISH scoring algorithm. This information along with the intensity and EIC shape informed the manual decision about which spectra to export to the mzVault library. Selected spectra were exported by checking the tickbox adjacent to them and selecting export to mzVault. More than one spectrum may have been exported depending on the number, quality and variance present. A single RP library was created, separate HILIC libraries were made for positive and negative ion mode. For each group a manual record was kept in Excel of whether the spectra had been added to the library, if this was for the molecular ion or an alternative adduct form, if spectra were present but not added to the library, if just MS¹ data were recorded or if it was not detected at all. All standards detected in a group whether MS² data were acquired or not had the checkbox ticked, this allowed the export of just these compounds into an excel file. After all groups had been processed the excel files were combined to generate a csv file of all compounds detected. This was a combined file for RP but separate lists were created for HILIC in positive and negative ion mode due to the different mobile phases utilised. Where there was more than one entry in the list for a single standard the feature with the greatest intensity had the RT selected. The resulting .csv files could then be used as a mass list with RT information to provide extra confidence in the identification of features in the biological samples.

▼ 1. General Settings	
Precursor Selection	Use MS(n - 1) Precursor
Use Isotope Pattern in Precursor Reevaluation	True
Provide Profile Spectra	Automatic
Store Chromatograms	False
▼ 2. Spectrum Properties Filter	
Lower RT Limit	0
Upper RT Limit	0
First Scan	0
Last Scan	0
Ignore Specified Scans	
Lowest Charge State	0
Highest Charge State	0
Min. Precursor Mass	0 Da
Max. Precursor Mass	5000 Da
Total Intensity Threshold	0
Minimum Peak Count	1
▼ 3. Scan Event Filters	
Mass Analyzer	(Not specified)
MS Order	Any
Activation Type	(Not specified)
Min. Collision Energy	0
Max. Collision Energy	1000
Scan Type	Any
Polarity Mode	(Not specified)
▼ 4. Peak Filters	
S/N Threshold (FT-only)	1.5
▼ 5. Replacements for Unrecognized Properties	
Unrecognized Charge Replacements	1
Unrecognized Mass Analyzer Replacements	ITMS
Unrecognized MS Order Replacements	MS2
Unrecognized Activation Type Replacements	CID
Unrecognized Polarity Replacements	+
Unrecognized MS Resolution@200 Replacements	60000
Unrecognized MSn Resolution@200 Replacements	30000

Figure 33: CD3.0 Parameters in the Select Spectra node for standards processing.

1. General Settings	
Mass Tolerance [ppm]	5 ppm
Intensity Tolerance [%]	30
S/N Threshold	3
Min. Peak Intensity	50000
Ions	[2M+ACN+H] ⁺ 1; [2M+ACN+Na] ⁺ 1; [2M+FA-H] ⁻ 1; [2M+H]
Base Ions	[M+H] ⁺ 1; [M-H] ⁻ 1
Min. Element Counts	C H
Max. Element Counts	C90 H190 Br3 Ca Cl2 Co F3 K2 Li2 Mg N10 Na2 O15 P3 S5
2. Peak Detection	
Filter Peaks	False
Max. Peak Width [min]	0.5
Remove Singlets	True
Min. # Scans per Peak	5
Min. # Isotopes	1

Figure 34: CD3.0 Parameters in the Detect Compounds node for standards processing. All ions available were selected.

1. Compound Consolidation	
Mass Tolerance	5 ppm
RT Tolerance [min]	0.2
2. Fragment Data Selection	
Preferred Ions	[M+H] ⁺ 1; [M-H] ⁻ 1

Figure 35: CD3.0 Parameters in the Group Compounds node for standards processing.

1. Search Settings	
Mass Lists	\MetaSci_UB_02192019.masslist\MetaSci_UB_02192019_2.masslist
Use Retention Time	False
RT Tolerance [min]	2
Mass Tolerance	5 ppm

Figure 36: CD3.0 Parameters in the Search Mass Lists node for standards processing.

▼ 1. Search Settings	
Compound Classes	All
Match Ion Activation Type	True
Match Ion Activation Energy	Match with Tolerance
Ion Activation Energy Tolerance	40
Apply Intensity Threshold	True
Precursor Mass Tolerance	10 ppm
FT Fragment Mass Tolerance	10 ppm
IT Fragment Mass Tolerance	0.4 Da
Identity Search	HighChem HighRes
Similarity Search	Similarity Forward
Library	Reference
Post Processing	Recalibrated
Match Factor Threshold	50
Max. # Results	10

Figure 37: CD3.0 Parameters in the Search MzCloud node for standards processing.

▼ 1. General Settings	
Mass Tolerance	5 ppm
▼ 2. Data Sources	
Data Source #1	MassList Search
Data Source #2	mzCloud Search
Data Source #3	Predicted Compositions
Data Source #4	ChemSpider Search
Data Source #5	Metabolika Search

Figure 38: CD3.0 Parameters in the Assign Compound Annotations node for standards processing

2.4.6 Data Processing and Identification

2.4.6.1 AcquireX Data Processing

Each individual AcquireX sequence was processed separately in CD3.0. The second blank, that was used to generate the exclusion list, the full scan sample that was used to generate the inclusion list, and the six DDA injections were uploaded into a new project. The blank file was assigned as blank, the full scan sample was assigned as sample and the six DDA files were identified as identification only. The following nodes were used in the workflow with the parameters the same as in seen in 2.4.5, Select Spectra (Figure 33), Detect Compounds (Figure 34), Group Compounds (Figure 35), Search MzCloud (Figure 37). Extra nodes or modified nodes were, Align Retention Times node (Figure 39), Mark Background, Search MassList (Figure 41), Assign Compound Annotations (Figure 42), Search MzVault (Figure 43), Search ChemSpider (Figure 44), Predict Compositions (Figure 45), Map to Metabolika Pathways (Figure 46).

▼ 1. General Settings	
Alignment Fallback	Use Linear Model
Mass Tolerance	5 ppm
Maximum Shift [min]	1
Remove Outlier	True
Shift Reference File	True
Alignment Model	Adaptive curve

Figure 39: CD3.0 Parameters in the Align RTs node for biological data processing.

▼ 1. General Settings	
Hide Background	True
Max. Blank/Sample	0
Max. Sample/Blank	5

Figure 40: CD3.0 Parameters in the Mark Background node for biological data processing

Mass Tolerance	5 ppm
RT Tolerance [min]	0.2
Use Retention Time	True

Figure 41: CD3.0 Parameters in the Search Mass List node for biological data processing.

▼ 1. General Settings	
Mass Tolerance	5 ppm
▼ 2. Data Sources	
Data Source #1	mzVault Search
Data Source #2	mzCloud Search
Data Source #3	MassList Search
Data Source #4	ChemSpider Search
Data Source #5	Metabolika Search

Figure 42: CD3.0 Parameters in the Assign Compound Annotations node for biological data processing.

1. Search Settings	
Apply Intensity Threshold	True
Compound Classes	All
FT Fragment Mass Tolerance	10 ppm
mzVault Library	<Items> <Item SHA512="34-DF-ED-27-7E-4A-8F-0B-26-E9-93-
IT Fragment Mass Tolerance	0.4 Da
Ion Activation Energy Tolerance	20
Match Analyzer Type	True
Match Ion Activation Energy	Match with Tolerance
Match Ion Activation Type	True
Match Ionization Method	True
Match Factor Threshold	0
Max. # Results	25
Precursor Mass Tolerance	10 ppm
Remove Precursor Ion	False
RT Tolerance [min]	5
Search Algorithm	HighChem HighRes
Use Retention Time	False

Figure 43: CD3.0 Parameters in the Search mzVault node for biological data processing.

1. Search Settings	
Result Order (for Max. # of results per c	Order By Reference Count (DESC)
Database(s)	BioCyc; Human Metabolome Database; KEGG
Mass Tolerance	5 ppm
Max. # of Predicted Compositions to b	3
Max. # of results per compound	10
Search Mode	By Formula and Mass
2. Predicted Composition Annotation	
Check All Predicted Compositions	False

Figure 44: CD3.0 Parameters in the Search ChemSpider node for biological data processing.

1. Prediction Settings	
Mass Tolerance	5 ppm
Max. Element Counts	C90 H190 Br3 Ca Cl4 Co F3 K2 Li2 Mg N10 Na2 O15 P3 S5
Max. H/C	4
Max. # Candidates	10
Max. # Internal Candidates	200
Max. RDBE	40
Min. Element Counts	C H
Min. H/C	0.1
Min. RDBE	0
2. Pattern Matching	
Intensity Threshold [%]	0.1
Intensity Tolerance [%]	30
Min. Pattern Cov. [%]	90
Min. Spectral Fit [%]	30
S/N Threshold	3
Use Dynamic Recalibration	True
3. Fragments Matching	
Mass Tolerance	5 ppm
S/N Threshold	3
Use Fragments Matching	True

Figure 45: CD3.0 Parameters in the Predict Compositions node for biological data processing.

1. Search Settings	
Metabolika Pathways	<Items> <Item SHA512="E4-96-21-A9-91-58-6D-17-15-5D-12-
Search Mode	By Formula or Mass
2. By Mass Search Settings	
Mass Tolerance	5 ppm
3. By Formula Search Settings	
Max. # of Predicted Compositions to b	3
4. Display Settings	
Max. # Pathways in 'Pathways' column	10

Figure 46: CD3.0 Parameters in the Map to Metabolika Pathways node for biological data processing.

2.4.6.2 DDA Data Processing

Each triplicate of traditional DDA data that was collected was processed separately in CD3.0. All three replicates were assigned as sample. All other parameters were the same as described in 2.4.6.1.

2.4.7 Data Analysis

2.4.7.1 Counting Number of Standards Detected and mzVault Contents

The contents of the positive and negative ion mode mzVault libraries were counted using the excel reference file of the contents which was recorded as the library was constructed.

2.4.7.2 Advantage of AcquireX vs Traditional DDA

The compounds list was exported into excel, saved as a .csv file and then imported into R version 3.2.2. The length of the table was recorded to give the number of compounds detected. The number of compounds with MS² data for the preferred ion was recorded by filtering the MS² column for compounds with "ddMS2 for preferred ion". The number of compounds with MS² data for another ion was recorded by filtering the MS² column for compounds with "ddMS2 for other ion". The total number of compounds with MS² data were recorded by summing the number with MS² for the preferred and other ions. Level 1 identifications were assigned by filtering out all compounds which had an mzVault match score ≥ 70.0 and were also recorded as a "Full match" in the Mass List Annotation Source column. These same compounds were then removed from the total list of compounds so they could not be counted twice. Level 2 identifications were identified from the remaining list of compounds by filtering for all compounds with an mzVault match score of ≥ 70.0 which did not have a Mass List match, as well as any compounds which had an mzCloud match score ≥ 70.0 . The number of compounds over a certain threshold match score in mzVault and mzCloud was recorded by filtering the compounds list by the mzVault Best Match and mzCloud Best Match columns.

2.4.7.3 Progression of the Length of Inclusion/Exclusion Lists

The length of the AcquireX inclusions and exclusion lists were recorded. Each injection in the sequence has its own method saved automatically as the data is collected, each with their own lists. The lists can be exported in .csv format. This was done for each of the six injections for each of the AcquireX sequences and the length of each list was noted and compared.

2.4.7.4 msPurity

The R package msPurity was used as described for DDA data in chapter 2.3.7.3. The data were filtered by RT as shown in Table 26. The number of features with an interpolated purity of 100% was recorded by filtering the inPurity column. The precursor intensity column was utilised to display the precursor intensities with the `scale_y_log10` function utilised to transform the intensity axis.

Table 26: The minimum and maximum RTs considered for purity and intensity distribution graphs for each assay applied.

Assay	RT Start (seconds)	RT End (seconds)
HILIC	15	690
RP	30	780
Lipidomics	30	720

2.4.7.5 Identification Levels

Schymanski's five levels were utilised (Figure 15). The blank filtered compound list for each AcquireX sequence was exported into excel from Compound Discoverer 3.0. The compound list was then analysed within R version 3.2.2. The number of level 1, 2, 3, 4 and 5 identifications were recorded. Level 1 and Level 2 identifications were recorded as described in section 2.4.7.2. All level 2 identifications were then removed from the compounds list before searching for level 3 identifications. Level 3 identifications were assigned by removing any compound which did not have any name associated or substructure similarity associated with it. These were then removed from the compounds list and level 4 identifications were searched for. Level 4 identifications were assigned by searching the remaining compound list for compounds which were labelled as "Unused" in the Predicted Compositions Annotation Source column. The remaining compounds were unknown and thus assigned as level 5. The lists for each sample type were then combined to give the sample type identification lists.

3.0 Characterising the complexity of electrospray ionisation data and its impact on metabolite annotation in UPLC-MS metabolomics studies

3.1 Introduction

There are a number of different derivative ion types that can be detected for any single metabolite when it is analysed using UHPLC-ESI-MS (Ultra-High Performance Liquid Chromatography-Electrospray Ionisation-Mass Spectrometry). It has been reported that as many as 100 different signals can be detected for any single metabolite (Mahieu et al., 2016). These can include different adducts, multiply charged species, oligomers and in-source fragments (Brown et al., 2009). The grouping of all features of a single metabolite in a dataset is essential to ensure accurate annotation and subsequent biological interpretation of the dataset, this is discussed in detail in section 1.3.1.4.2.2. If these features are not grouped together the likelihood of false annotations increases which may result in inappropriate further investigations and wasted expenditure on irrelevant, incorrectly identified metabolites. For example a sodiated 1-methylhistidine molecule in positive ion mode should be detected at an m/z of 192.0743, but this could end up being identified as a protonated carbendazim molecule with a theoretical m/z of 192.0768, particularly if the mass error was shifting upwards, this would lead to incorrect biological interpretation. Successful grouping provides other benefits too including reduction of data complexity, which subsequently allows less stringent multiple testing hypothesis correction, as well as providing improved confidence in annotation if multiple features can be linked to each other (Mahieu et al., 2016). A number of different dedicated software tools are available for grouping of MS^1 data. Some of these software assign feature groups through analysis of the pairwise peak intensity correlation data, these include, PUTMEDID (Brown et al., 2011), AStream (Alonso et al., 2011), MSClust (Tikunov et al., 2012), RAMClust (Broeckling et al., 2014), MS-FLO (Defelice et al., 2017), xMSannotator (Uppal et al., 2017), findMAIN (Jaeger et al., 2017) and BINNER (Kachman et al., 2019). Others utilise the chromatographic profiles of peaks to uncover the relationships such as CAMERA (Kuhl et al., 2012), MZMine2 (Pluskal et al., 2010) and CliqueMS (Senan et al., 2019). Bayesian probabilistic sampling has also been applied to the problem (Rogers et al., 2009) this was then developed further in ProbMetab (Silva et al., 2014), MetAssign (Daly et al., 2014) and IPA (Del Carratore et al., 2019). Other approaches have included a knowledge driven tool, called CEU Mass Mediator (Gil de la Fuente et al., 2018) as well as a neural network approach (Kantz et al., 2019) although this is only for removal of artefact peaks. All of these tools are designed to uncover the relationships between MS^1 features in the data and assign a true annotation but they all rely at some point on lists of commonly expected adducts, isotopes and fragments with associated mass differences. Generally short lists of adducts, isotopes and fragments are considered ranging from 1 or 2 up to ~30 for a single ion mode and this fact alongside the requirement for improved grouping methods is perhaps an indicator as to why only 1.8% of spectra from untargeted metabolomics experiments can be annotated (da Silva et al., 2015). Research in 2009 highlighted the complexity of

electrospray full-scan data on a small scale for five sample types analysed on one UPLC-MS instrument type (Acquity UPLC chromatographic system coupled to an electrospray LTQ-Orbitrap hybrid mass spectrometer); 12 different adducts and isotopes were reported (Brown et al., 2009). Since then more work has been performed to try and elucidate the true complexity seen in the data with it becoming clear that there are a number of complex relationships in the data, involving complex combinations of different fragments and adducts and even heterodimers such as seen with the mz.unity algorithm (Mahieu et al., 2016). However, no significant characterisation of the ion types and subsequent characteristic mass differences detected across multiple instrument types, for different classes of biological samples, and for different chromatographic columns and mobile phases has been performed. Considering all annotation resources utilise lists of certain adducts, isotopes, and in-source fragments this work is important to ensure lists applied are appropriate. Whatever the method is for feature grouping, the annotation result is going to be largely dependent on the contents of these lists.

Here we present a study to characterise the complexity of electrospray ionisation full-scan (MS^1) data acquired in UPLC-MS untargeted metabolomics approaches for 104 different datasets applying different instrument and sample types and analytical methods which are publicly available through the MetaboLights and Metabolomics Workbench data repositories or were provided by Phenome Centre Birmingham.

3.2 Results and Discussion

Feature pair intensity correlation data were generated as described in 2.1.2.1. Once the correlations for each bin in each dataset had been determined an appropriate cut off for R values needed to be determined. To assess this the R values were plotted for each dataset in a histogram to visualise the distributions.

3.2.1 Determination of an Appropriate Correlation Coefficient

The R values for each feature pair in each dataset were plotted in a histogram to determine if there was a clear cut off point that should be imposed for all datasets. It was expected that a normal bell-shaped distribution around the point where $R = 0$ would have been seen with a spike towards the positive end of the correlation where the related pairs of features were being observed. Whilst this was true for some datasets there were many different distributions observed. The histogram was plotted for each dataset and were assigned to a distribution type depending on the shape of the distribution seen (Table 27). The four most predominant distributions seen are similar to the four displayed in Figure 47Figure 48. There were 23 with a normal distribution peaking between -0.1 to +0.25 with a rise at +0.9 such as in Figure 47A. There were 20 with normal distribution with a peak between -0.1 to +0.25 and no rise at the ends such as in Figure 47B. There were 17 with a peak

between -0.1 to +0.25 and a roughly normal distribution but with another sharp larger peak at +0.9 to +1.0 as in Figure 48A. These first three have a roughly normal distribution. There were 16 with a peak at +0.9 to +1.0 with gradual increases to that point such as in Figure 48B. These four distributions made up 73% of the total datasets. The peak lists used to generate the data were pre-processed in different ways. Some such as those with a normal distribution and no spike towards +1.0 have likely been de-isotoped and feature-pairs grouped already (Figure 47B). Others which have not already been grouped and de-isotoped keep the large spike at the positive end such as Figure 48A. Whilst others may have been de-isotoped/grouped to a certain extent but not as effectively as in Figure 47A, this could be due to differences in the raw data and experimental set up, different software being applied, or different data processing parameters. The PCB datasets were assessed specifically as these were all from the same research group and so should perhaps be more consistent or similar. However, they demonstrated a very similar pattern to the general pattern with, 78.8% of datasets falling into pattern 1 (11), 2 (6), 3 (5) or 4 (4) compared to 73% of all datasets.

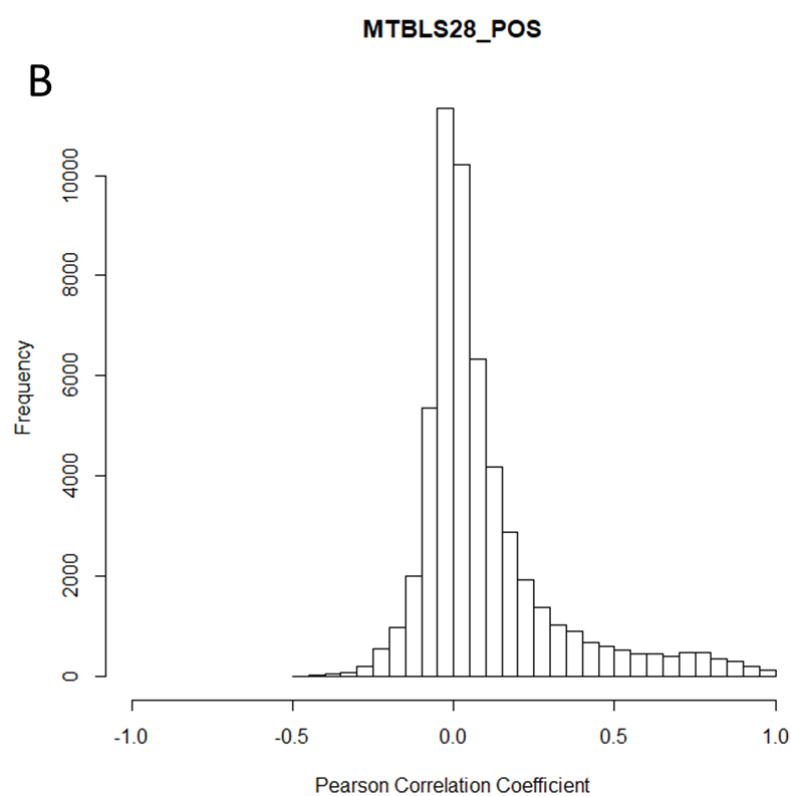
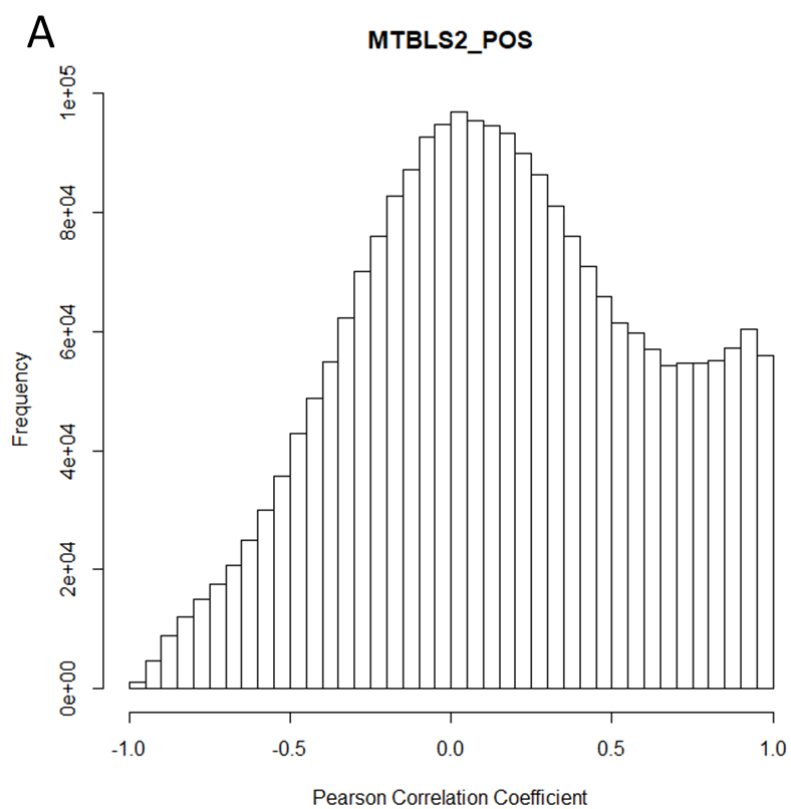


Figure 47: Histogram of Pearson correlation coefficients of intensity data for all feature pairs falling within 5 second RT windows calculated as described in 2.1.2.1 for two datasets A) MTBLS2_POS B) MTBLS28_POS

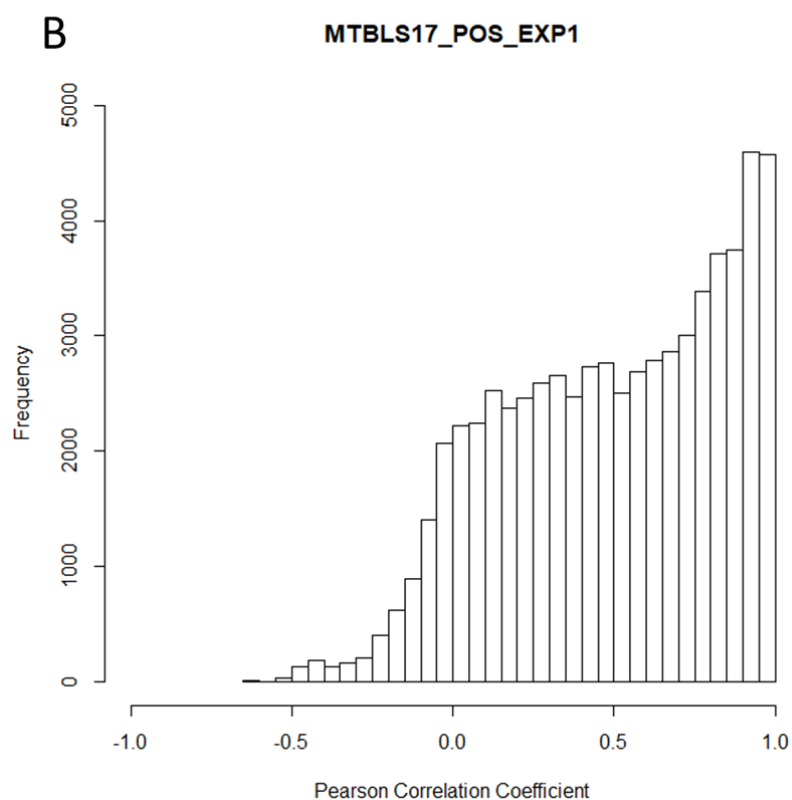
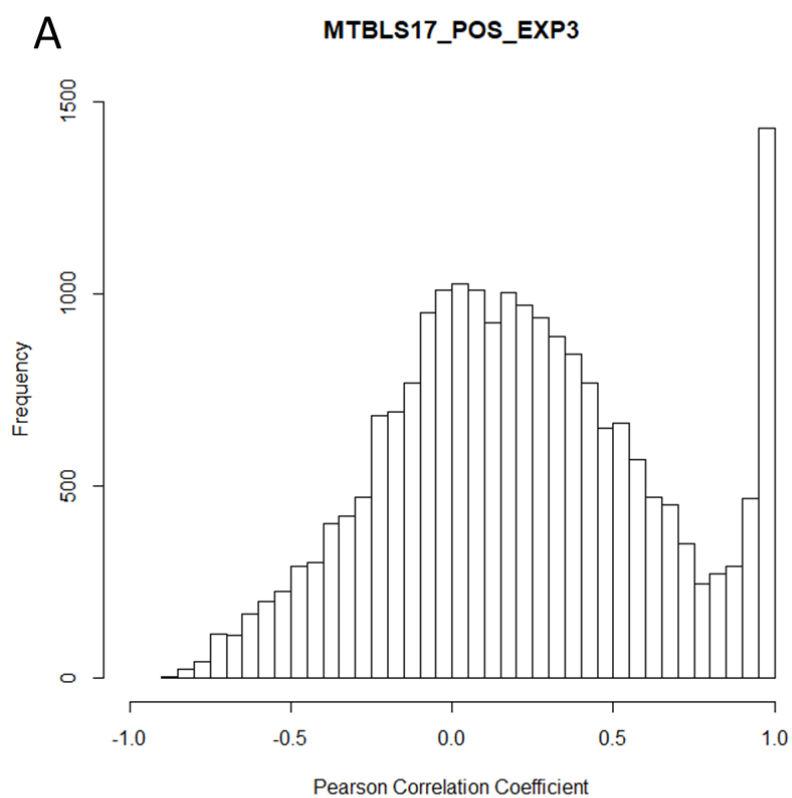


Figure 48: Histogram of Pearson correlation coefficients of intensity data for all feature pairs falling within 5 second RT windows calculated as described in 2.1.2.1 for two datasets A) MTBLS17_POS_EXP3 B) MTBLS17_POS_EXP1

The other distribution types are displayed in Figure 49, Figure 50, Figure 51, Figure 52 whilst a summary of all distribution types is found in Table 27. Overall, the variety of distributions seen was greater than expected. As discussed, this could be partially down to the data processing and grouping parameters that had been applied. However, many other factors related to the experimental set up could also be important. A key characteristic affecting the result would be the chromatographic peak widths observed in the assay. RT windows of 5 seconds were used to bin related features with each window overlapping by 2.5 seconds. The retention time window applied should ideally be the size of the chromatographic peaks recorded. This number may have been too narrow for some datasets with very wide peaks causing smaller numbers of highly correlating feature pairs to be observed. Whilst it may have been too wide for other datasets with very narrow peaks and high chromatographic resolution causing an overinflated number of correlations to be presented. The overall complexity of the dataset in terms of feature numbers reported in the matrix is also an important factor which will have been determined by the sample type and experimental parameters but also the stringency of the peak picking parameters applied. Another important factor impacting the results is the number of samples in each study, some of the studies include over 100 samples, whilst others have only 20 or slightly less, the correlations calculated in the smaller datasets are more likely to be variable and prone to error or fluctuations detrimentally affecting the results. The high variability in correlation distributions seen are indicative of the wide variety of experimental methods and data processing parameters represented by the datasets included in this study. It was determined that there was no clear single cut off value that was appropriate across all datasets which was in agreement with another study (Kachman et al., 2019). Furthermore it has been reported that unrelated features can still have high intensity correlations and related features can have very low intensity correlations (Mahieu et al., 2016), for example the reported correlation of just 0.45 between the M+H and M+Na peaks of a metabolite (Kachman et al., 2019). Considering that these are probably the two highest intensity ions expected for a metabolite in positive ion mode this is quite surprising. However, it is also important to consider that this could have been a low intensity feature closer to the noise level which would be much more likely to generate spurious results due to higher variability in peak measurement response close to the noise level of the mass spectrometer. Despite this, considering the large volume of data sampled a stringent correlation cut-off of $R \geq +0.8$ was implemented to ensure greater confidence in the results. Although this may mean missing out on a number of related features in the data it is more important to restrict false positive results. For the best results the sources of information used for grouping should be diversified to increase confidence and reliability. For example chromatographic profile correlation can also be applied in CAMERA (Kuhl et al., 2012), MZMine2 (Pluskal et al., 2010) and CliqueMS (Senan et al., 2019).

Table 27: Summary of correlation distributions plotted, their key characteristics and their frequencies across the 104 datasets utilized.

Example Dataset/Figure	Key Characteristics	Frequency
MTBLS2_POS (Figure 47A)	<ul style="list-style-type: none"> • Peaks between -0.1 and +0.25 • Normal distribution with increase at +0.8 to +1.0 	23
MTBLS28_POS (Figure 47B)	<ul style="list-style-type: none"> • Peaks between -0.1 and +0.25 • Normal distribution 	20
MTBLS17_POS_EXP3 (Figure 48A)	<ul style="list-style-type: none"> • Peaks between +0.9 and +1.0 • Normal distribution peaking between -0.1 and +0.25 	17
MTBLS17_POS_EXP1 (Figure 48B)	<ul style="list-style-type: none"> • Peaks between +0.9 and +1.0 • Gradual increase to peak from between -1.0 and -0.25 	16
PCB5_NEG (Figure 49A)	<ul style="list-style-type: none"> • Peaks between +0.9 and +1.0 • Sudden rise with slow gradual increase from between -1.0 and -0.25 	9
MTBLS372_NEG (Figure 49B)	<ul style="list-style-type: none"> • Peaks between -0.1 and +0.25 • Normal distribution with rises between +0.9 and +1.0 and between -0.9 and -1.0 	4
MTBLS291_POS (Figure 50A)	<ul style="list-style-type: none"> • Peaks +0.9 and +1.0 • Gradual increase after rapid decrease from another peak between -0.9 and -1.0 	4
MTBLS291_NEG (Figure 50B)	<ul style="list-style-type: none"> • Flat with no clear shape or peak 	4
MTBLS403_POS (Figure 51A)	<ul style="list-style-type: none"> • Peaks between +0.9 and +1.0 and between -0.9 and -1.0 • Normal distribution in between the peaks 	2
PCB9_POS (Figure 51B)	<ul style="list-style-type: none"> • Peaks between -0.1 and +0.25 • Normal distribution with second large peak between +0.6 and +0.8 	2
ST325_POS (Figure 52A)	<ul style="list-style-type: none"> • Peaks between +0.9 and +1.0 • Normal distribution with peak between -0.25 and -0.6 	2
PCB7_POS (Figure 52B)	<ul style="list-style-type: none"> • Peaks at +0.6 • Steady rise from -1.0 to the peak before rapid decline to +1.0 	1

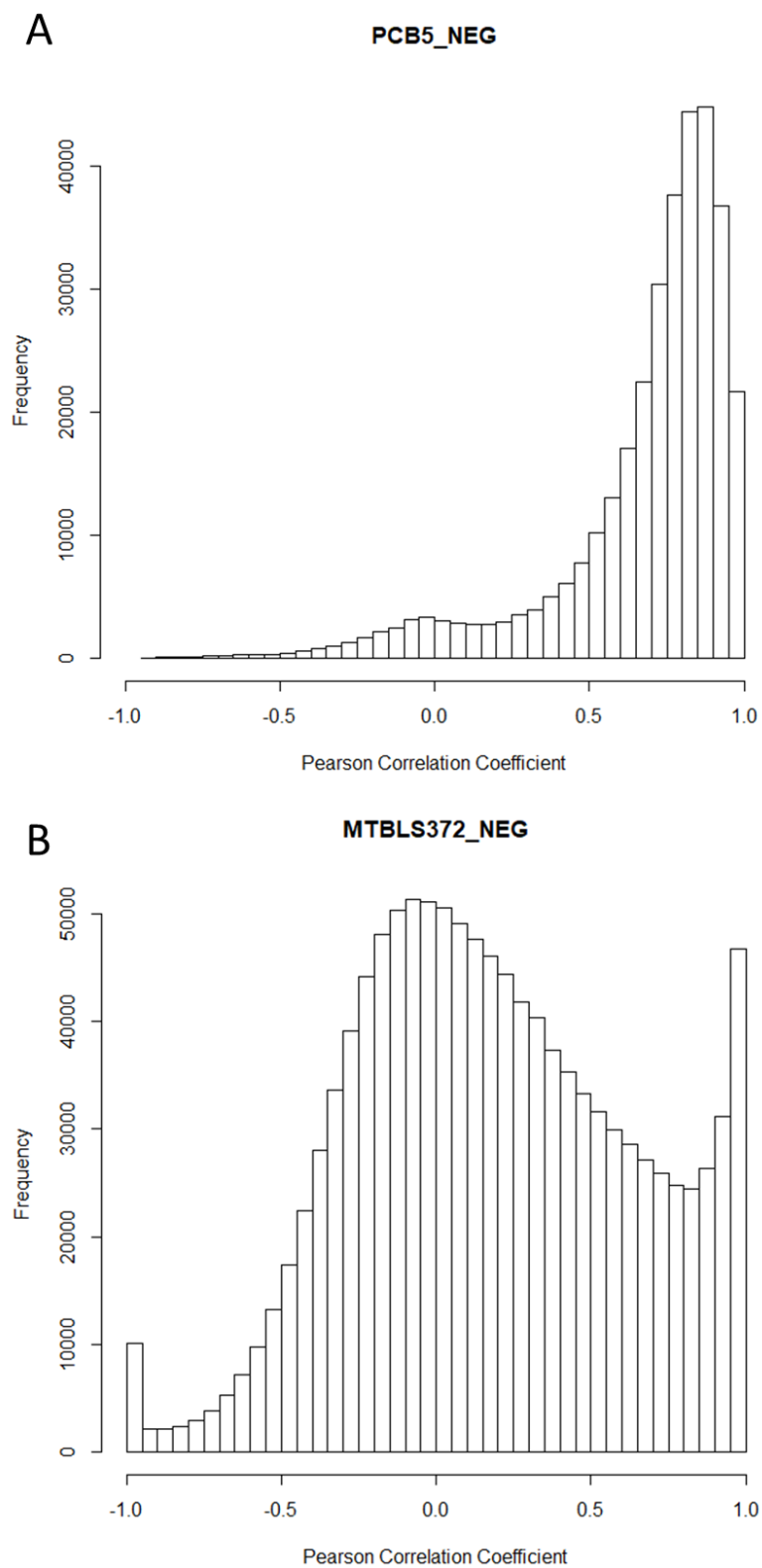


Figure 49: Histogram of Pearson correlation coefficients of intensity data for all feature pairs falling within 5 second RT windows calculated as described in 2.1.2.1 for two datasets A) PCB5_NEG B) MTBLS372_NEG

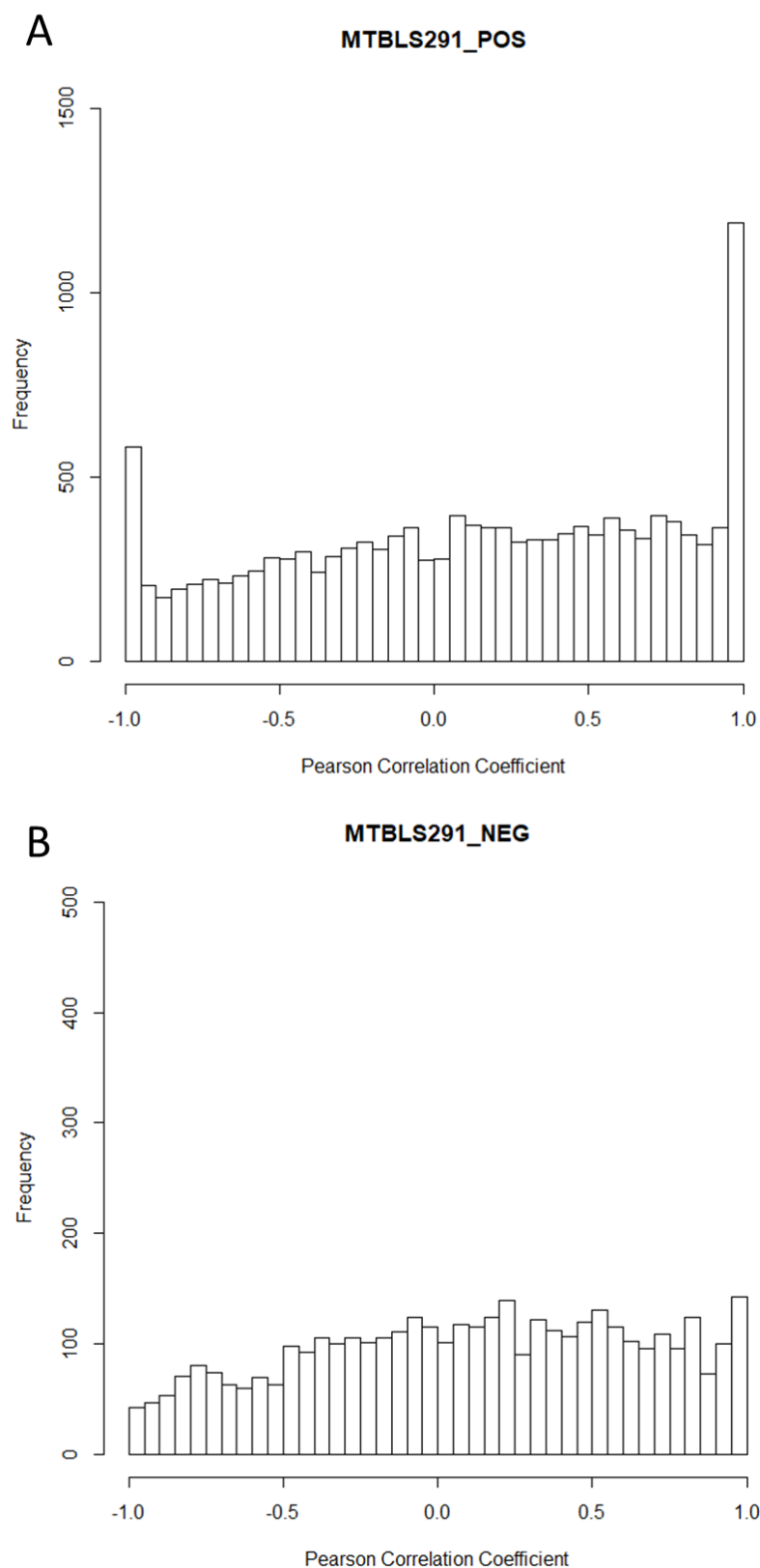


Figure 50: Histogram of Pearson correlation coefficients of intensity data for all feature pairs falling within 5 second RT windows calculated as described in 2.1.2.1 for two datasets A) MTBLS291_POS B) MTBLS291_NEG.

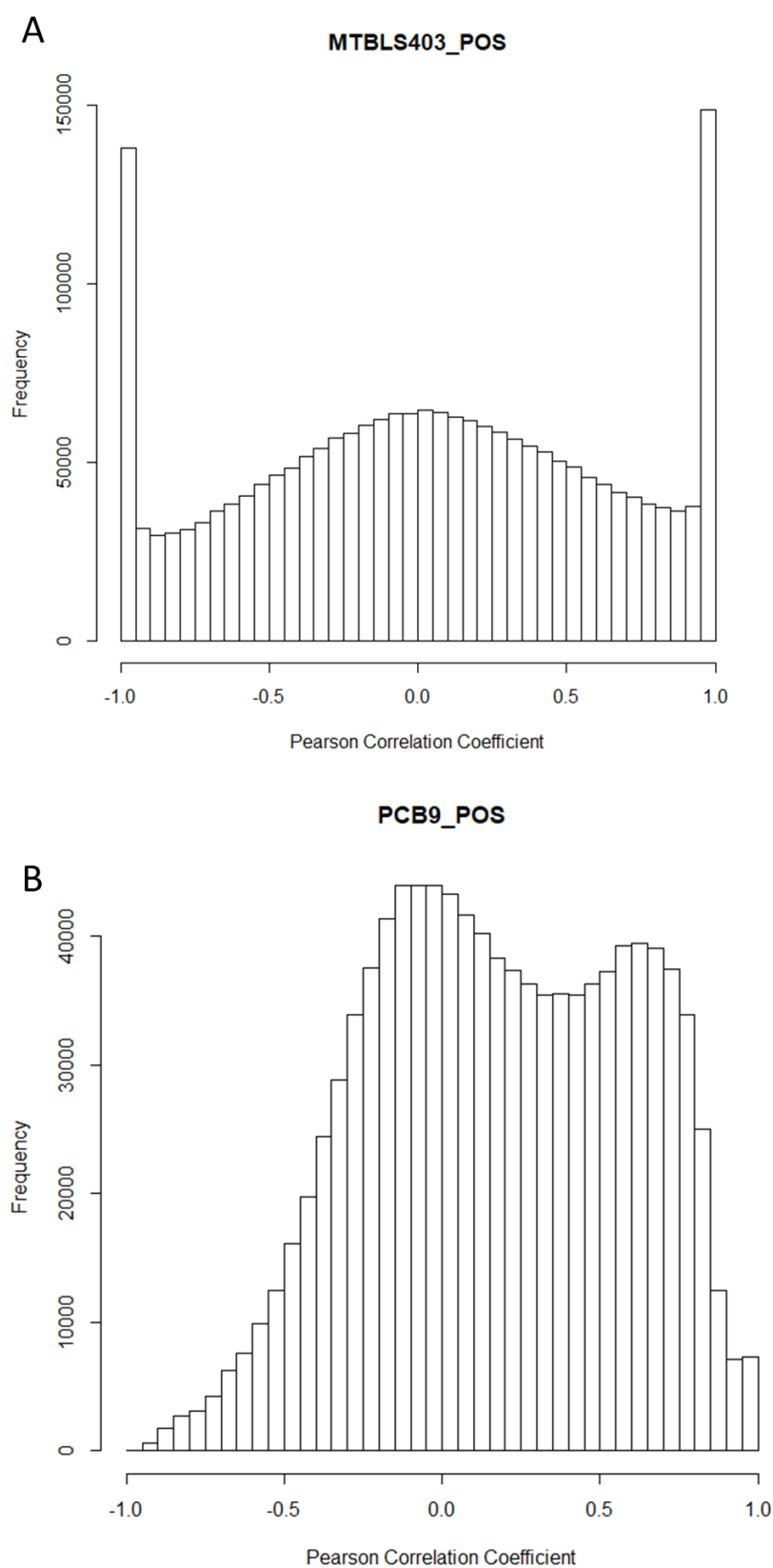


Figure 51: Histogram of Pearson correlation coefficients of intensity data for all feature pairs falling within 5 second RT windows calculated as described in 2.1.2.1 for two datasets A) MTBLS403_POS B) PCB9_POS

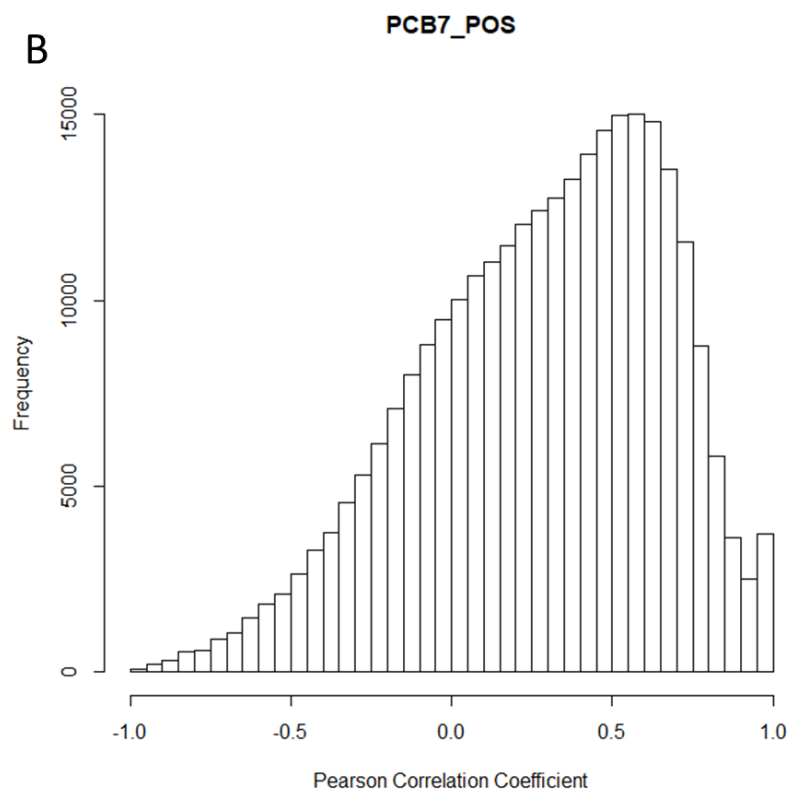
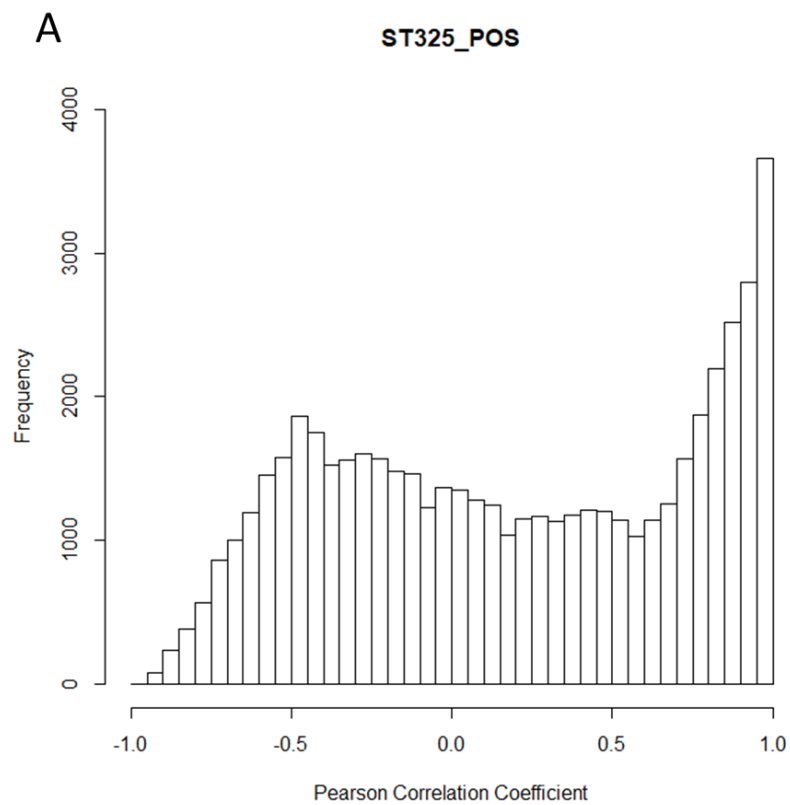


Figure 52: : Histogram of Pearson correlation coefficients of intensity data for all feature pairs falling within 5 second RT windows calculated as described in 2.1.2.1 for two datasets A) ST325_POS B) PCB7_POS.

3.2.2 Complexity of Alternative Ion Types in Untargeted UHPLC-MS Experiments

The complexity of m/z differences detected was far greater than expected. There were 8273 different m/z differences (rounded to 4 decimal places) with a frequency of 40 or more across the 104 datasets. After being manually grouped through looking at the frequency distributions these were then reduced to 1038 unique m/z differences with an associated error. Efforts were made to annotate these differences and explain them although this was not achievable for all the m/z differences, 211 of the 1038 mass differences had an annotation. A variety of different adducts, isotopes, multiply charged species, oligomers, fragments and combinations of these can explain a number of these differences. Many of the differences are represented by combinations of adducts such as $[M+H]^+$ and $[M+Na]^+$, fragments such as H_2O and multiple charges. It is easy to imagine how quickly a wide variety of different correlated mass differences can arise for any single metabolite with multiple features even if there are a small number of derivative ion types. The differences were ranked based on the maximum frequency seen for a 4 decimal place value across all datasets and the top 50 most observed are displayed in Table 28. It is not surprising to see the ^{12}C - ^{13}C isotopic difference as the most prevalent although it was not present at high frequency in all datasets as some had clearly been de-isotoped already. However, even if already deisotoped it is likely some ^{12}C - ^{13}C differences could remain. Another surprising result was the high prevalence of multiply charged species. The difference seen between the doubly charged ^{12}C isotope and the doubly charged ^{13}C isotope was the 3rd most prevalent difference whilst the equivalent triply, quadruply and quintuply charged differences also featured in the top 50. A number of isotopes other than the ^{13}C isotope also featured in the top 50 including (^{35}Cl - ^{37}Cl , 6Li - 7Li , ^{39}K - ^{41}K , ^{32}S - ^{34}S , ^{32}S - ^{33}S , ^{14}N - ^{15}N , ^{16}O - ^{18}O). It is likely that some of the differences assigned as these isotopic are correctly assigned although others may be incorrectly assigned due to the high density of mass differences present in the areas just below a m/z of 1.0000 or 2.0000 where these mass differences can be found. This makes it difficult to separate the frequency distributions from each other like in other regions of the m/z scale. Only 2 (2.0000 and 6.0170) of the top 50 were unannotated. A much wider variety of ion types, in-source fragments and biological transformations were observed than has previously been reported or which are applied in the most frequently applied open access and commercial software packages used for metabolite annotation. Overall this data highlights the incredible complexity present in full scan UHPLC-MS data of complex sample types and how more work needs to be performed to uncover the underlying patterns and relationships between derivative ions to allow more accurate annotation of complex datasets.

Table 28: The top 50 highest frequency mass differences detected where mass differences were ranked by the maximum frequency (to 4 decimal places) seen for the difference across all datasets.

Rank	ID	Type	Theoretical <i>m/z</i> Difference	Max Frequency
1	$^{12}\text{C}-^{13}\text{C}$	Isotope	1.0034	7293
2	$^{35}\text{Cl}-^{37}\text{Cl}$	Isotope	1.9970	2016
3	Doubly charged	Multiply charged	0.5017	1790
4	$[\text{M}+\text{Na}]^+ - [\text{M}+\text{H}]^+$	Adduct	21.9819	1401
5	Sodium Formate (HCOONa)	Adduct	67.9874	1385
6	H ₂ O	Fragment/Transformation	18.0106	887
7	$^6\text{Li}-^7\text{Li}$	Isotope	1.0009	864
8	Triply charged	Multiply charged	0.3344	814
9	$^{39}\text{K}-^{41}\text{K}$	Isotope	1.9981	725
10	PEG (C ₂ H ₄ O)	Fragment/Transformation	44.0262	660
11	CH ₂	Fragment/Transformation	14.0157	612
12	$(^{14}\text{N}-^{15}\text{N}) - (^{35}\text{Cl}-^{37}\text{Cl})$	Isotope/Isotope	1.0000	607
13	NaCl	Adduct	57.9586	601
14	C ₂ H ₂	Fragment/Transformation	26.0156	551
15	2H	Fragment	2.0156	547
16	NaCl - Sodium Formate (HCOONa)	Adduct/Adduct	10.0288	535
17	Ammonia (H ₃ N)	Adduct	17.0265	534
18	K - Na	Adduct/Adduct	15.9739	512
19	$(^{35}\text{Cl}-^{37}\text{Cl})/2$	Isotope/Multiply charged	0.9985	504
20	Quadruply charged	Multiply charged	0.2508	496
21	$^{33}\text{S}-^{34}\text{S}$	Isotope	0.9994	471
22	Formic Acid (CH ₂ O ₂)	Adduct	46.0055	456
23	$(^{35}\text{Cl}-^{37}\text{Cl})_2$	Isotope	3.9940	413
24	Sodium trifluoroacetate (CF ₃ CO ₂ Na)	Adduct	135.9748	407

25	Sodium Formate (HCOONa) – (^{12}C - ^{13}C)	Adduct/Isotope	66.9839	397
26	^{39}K - ^{41}K /2	Isotope/Multiply charged	0.9991	396
27	(^{12}C - ^{13}C) ₂	Isotope	2.0067	386
28	C_2H_2 + (^{12}C - ^{13}C)	Fragment/Isotope	27.019	385
29	Na - (^{12}C - ^{13}C)	Adduct/Isotope	20.9784	381
30	^{14}N - ^{15}N	Isotope	0.9970	372
31	C	Fragment	12.0000	371
32	C_2H_4	Fragment	28.0313	353
33	^{32}S - ^{34}S	Isotope	1.9956	346
33	Acetonitrile ($\text{C}_2\text{H}_3\text{N}$)	Adduct	41.0266	346
35	(^{12}C - ^{13}C) + (^{14}N - ^{15}N)	Isotope/Isotope	2.0004	343
36	[M+Na] ⁺ - [M+NH ₄] ⁺	Adduct/Adduct	4.9554	342
37	Potassium Formate (HCOOK)	Adduct	83.9614	331
38	Not Annotated	Not Annotated	2.0000	329
39	C_2	Fragment/Transformation	24.0000	323
40	(^{12}C - ^{13}C) ₂ /3	Isotope/Multiply charged	0.6690	319
41	PEG ($\text{C}_2\text{H}_4\text{O}$)/2	Fragment/Multiply charged	22.0131	318
42	CO	Fragment	27.9949	308
43	^{16}O - ^{18}O	Isotope	2.0042	295
44	(^{32}S - ^{34}S)/2	Isotope/Multiply charged	0.9978	294
45	O	Fragment/Transformation	15.9949	291
46	NaCl + (^{35}Cl - ^{37}Cl)	Adduct/Isotope	59.9556	286
47	Quintuply charged	Multiply charged	0.2007	284
48	Not Annotated	Not Annotated	6.0170	278
49	H_2O + (^{12}C - ^{13}C)	Fragment/Transformation/Isotop e	19.0141	271
50	NaCl+(^{35}Cl - ^{37}Cl) - HCOONa	Adduct/Adduct/Isotope	8.0318	269

More analysis of pure metabolite standards and the subsequent related features generated such as done by Mahieu et al (2016), would be highly beneficial to allow improved understanding of what to

expect and look for in the data. They showed that glutamate (molecular ion intensity = 1×10^9) generated 98 peaks, whilst NAD (molecular ion intensity = 3×10^8) generated 23 peaks. Analysis of more pure metabolite standards should be performed in this manner for different assays whilst also utilising dilution series to see how the relationships change at different concentrations within a simple solution. Furthermore, the work could be repeated with the standards then spiked into complex mixtures at the same concentrations to see how well conserved the different derivative feature types and relationships are.

It is hardly surprising that many of the mass differences could not be annotated when the complexity created by even a small number of derivative ions for a single metabolite is considered such as in Figure 53. A very simple, small and plausible selection of combinations of M (Tryptophan), H, Na, and ^{13}C were used and their theoretical m/z values calculated.

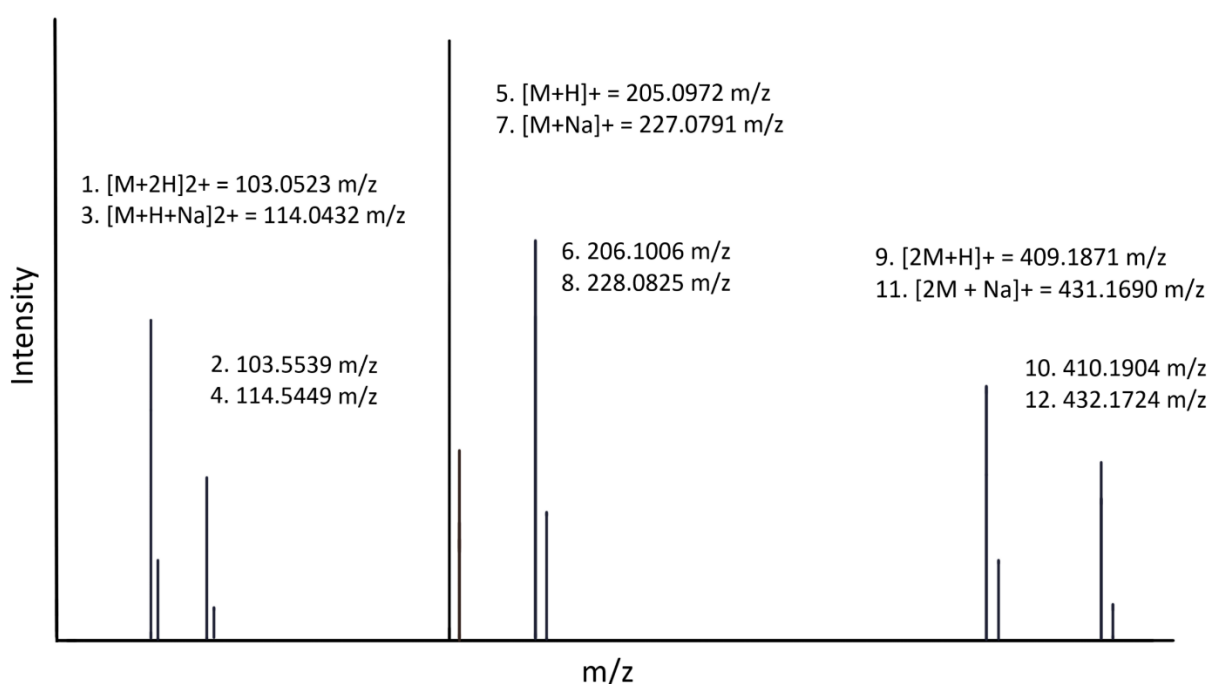


Figure 53: Theoretical spectrum of 12 possible positive ion mode peaks for tryptophan demonstrating how simple combinations of common degenerate feature types generate complexity in UHPLC-MS. Peaks are numbered from left to right. The even numbered (unlabelled) peaks all represent the ^{12}C - ^{13}C isotopic peak of the preceding odd numbered peak. Labels for all peaks from left to right, smallest to largest can be found in Table 29.

All 12 peaks should be correlated with each other as derivative ions of a single metabolite and therefore each peak would give rise to 11 correlated mass differences. The mass differences between

these theoretical peaks were calculated and are displayed in Table 29. Duplicate values and zeros were removed and replaced with x's leaving 66 correlated mass differences. Within these mass differences 18 correspond to some of the commonly detected 1038 differences discussed earlier, these cells are highlighted grey. This is because the mass difference generated between these feature pairs is not dependent on the initial m/z of the metabolite (M) in question. This is not the case for the other 48 mass differences. They are dependent on the initial m/z of M because they represent the differences between singly charged species, doubly charged species and homodimers. These differences, although generated by normal and simple adducts will not be of high frequency because of their dependence on the initial m/z of M, which will only be present once in that dataset unless isomers or isobars are present. This highlights one of the problems with this approach. This may not be the case if the M is a very commonly detected biological compound. If M is detected in most studies then the difference could have occurred more than 40 times across all datasets and so made it into the commonly detected list. This perhaps explains why so many of the list could not be annotated. Furthermore, this was a very simple selection of derivative features. When it is considered that as many as 100 different features can be observed for a single metabolite (Mahieu et al., 2016) and the number of different metabolites in a sample is expected to be at least a few hundred it is easy to see how there is so much variety in correlated mass differences across the 104 datasets. This data highlights the drastic need for a better understanding of ESI data complexity.

Table 29: m/z distance matrix for the 12 theoretical tryptophan peaks as shown in Figure 53. All possible m/z differences between the 12 features are presented. Peaks are listed in ascending order. Zeros and duplicated results have been assigned with an x. Mass differences which correspond to one of the 1038 commonly detected mass differences are highlighted in grey.

	M+2H	M+ ¹³ C+2H	M+H+Na	M+ ¹³ C+H+Na	M+H	M+ ¹³ C+H	M+Na	M+ ¹³ C+Na	2M+H	2M+ ¹³ C+H	2M+Na	2M+ ¹³ C+Na
M+2H	x	x	x	x	x	x	x	x	x	x	x	x
M+ ¹³ C+2H	0.5017	x	x	x	x	x	x	x	x	x	x	x
M+H+Na	10.9910	10.4893	x	x	x	x	x	x	x	x	x	x
M+ ¹³ C+H+Na	11.4926	10.9910	0.5017	x	x	x	x	x	x	x	x	x
M+H	102.0449	101.5433	91.0540	90.5523	x	x	x	x	x	x	x	x
M+ ¹³ C+H	103.0483	102.5466	92.0573	91.5556	1.0033	x	x	x	x	x	x	x
M+Na	124.0269	123.5252	113.0359	112.5342	21.9819	20.9786	x	x	x	x	x	x
M+ ¹³ C+Na	125.0302	124.5286	114.0393	113.5376	22.9853	21.9819	1.0034	x	x	x	x	x
2M+H	306.1348	305.6331	295.1438	294.6422	204.0899	203.0865	182.1079	181.1046	x	x	x	x
2M+ ¹³ C+H	307.1382	306.6365	296.1472	295.6455	205.0932	204.0899	183.1113	182.1079	1.0034	x	x	x
2M+Na	328.1168	327.6151	317.1258	316.6241	226.0718	225.0685	204.0899	203.0865	21.9819	20.9786	x	x
2M+ ¹³ C+Na	329.1201	328.6184	318.1291	317.6275	227.0752	226.0718	205.0932	204.0899	22.9853	21.9819	1.0034	x

3.2.3 Comparison of Mass Difference Frequencies Across Datasets

The frequency of m/z differences across the 104 datasets is displayed in Figure 54. Green cells indicate high frequency whilst red cells identify where the m/z difference was not observed, details of the conditional formatting rules are found in Figure 18. Some datasets have very few of the mass differences present whilst others show at least low levels of frequency for almost all differences with a number of high frequency differences too. The datasets used are derived from use of different samples, instruments, assays and data processing workflows. Therefore, although the amount of variation seen was not anticipated initially it is not surprising to see the large differences between the frequency levels of the mass differences found in the different datasets. The different groups (electronic Appendix, 2.1/2.1.2.1) investigated were also compared and these data can be found in the electronic Appendix (3.2/3.2.3). No clear differences could be seen between the groups (electronic Appendix, 3.2/3.2.3). The number of datasets within each group is too small and this is compounded by the fact that some of the datasets within those groups are from the same study or lab and so have high similarity, there may even be adducts found only in one lab or particular experiment (Mahieu and Patti, 2017). This could cause unusual laboratory unique differences to dominate the major mass difference frequencies in that group, skewing the data. On the other hand, it has been suggested that there might not be too much difference in complex adducts seen between different sample types (Kachman et al., 2019). The high level of variation seen in Figure 54 indicate that Mahieu and Patti are more likely to be correct than Kachman. The lack of datasets for different groups was due to the relatively small number of pre-processed untargeted metabolomics UHPLC-ESI-MS datasets with more than 1000 features and 20 samples available in the MetaboLights and Metabolomics Workbench data repositories at the time of data acquisition. This highlights the need for more researcher's data to be made available in open access online repositories. As these repositories expand the volume of data present in them, the capabilities and power of a study such as this will be greatly improved. Another issue may be the implementation of a fixed cut off value for the correlation score as well as the fixed RT windows applied despite the different chromatographic methods represented in the data.

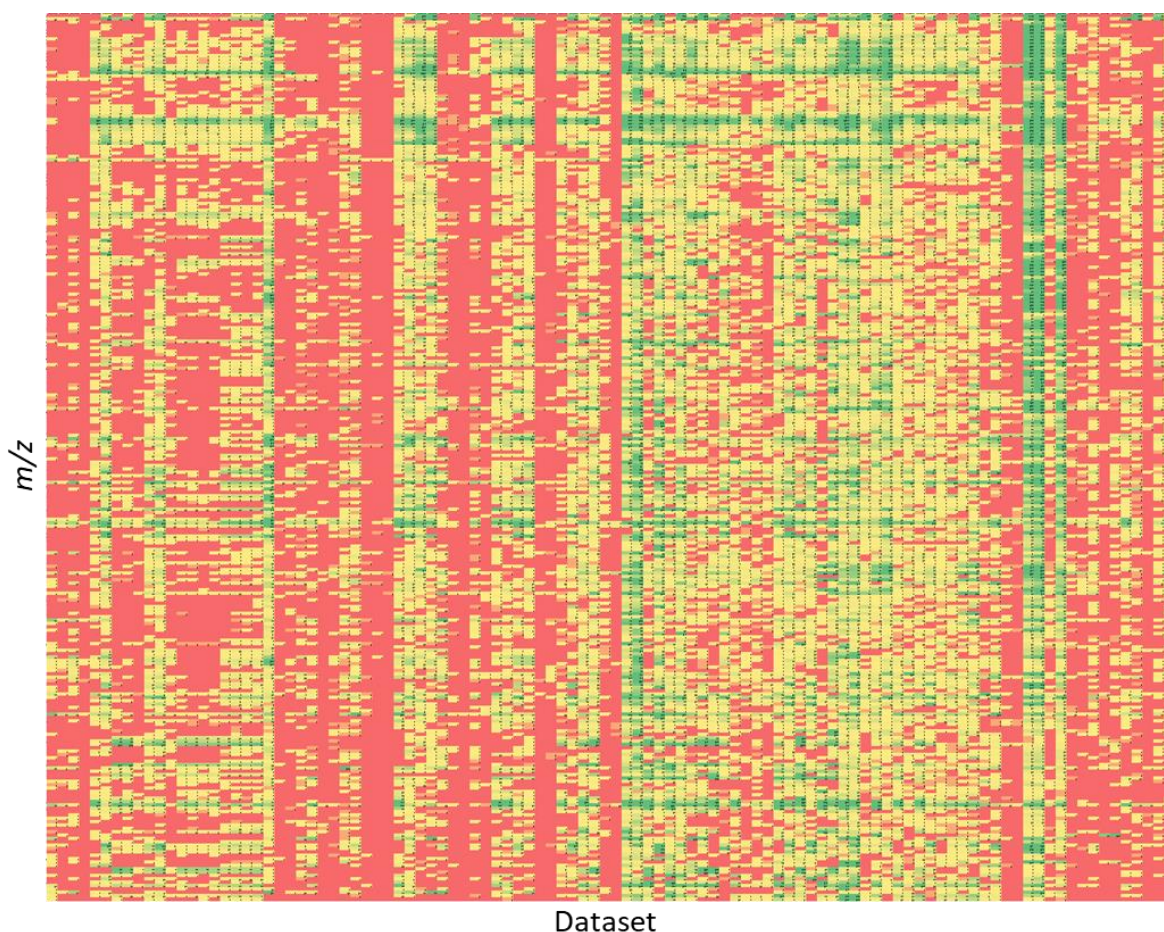


Figure 54: The comparison of the 264 common m/z differences across all 104 datasets. Each mass difference is represented by a row and each dataset by a column. Each cell contains the frequency of that mass difference in the dataset and the cells are conditionally formatted as displayed in Figure 18.

The clear variation in m/z differences seen across the different datasets utilised indicate how researchers should be looking to characterise the most common adducts and mass differences seen when utilising their own assays. This would be appropriate for laboratories which have developed standard methods for repeated use but would not be useful in cases where the method has been applied in a “one-off” scenario. Any method which shall undergo repeated uses in a lab should be investigated in an attempt to elucidate the keys adducts for consideration, reveal commonly searched for adducts which should not be considered and identify any uncommon or unusual adducts which appear with high regularity. This would allow the subsequent tailoring of the annotation process to the experimental set up in question and thus reduce the number of false positive and false negative annotations. This should be able to facilitate more accurate and reliable results from the MS¹ annotation phase and ultimately the study as a whole, improving the amount as well as the accuracy of biological information acquired.

3.2.4 Are biological transformations with the same retention time for reactant and product metabolites detected in these data?

Current metabolite annotation software can apply grouping of different features of the same metabolite based on retention time similarity and a positive peak intensity correlation coefficient. Two different metabolites which are metabolically linked (for example, one is the precursor and one is the product of a single metabolic reaction) will also fulfill these criteria. Here we investigated the datasets described above to determine the frequency of possible biological transformations. 1,070 different biological transformations from the KEGG database and their associated m/z differences between reactant and product metabolites were searched for in the list of significantly correlated mass differences from the 104 datasets described above. This was performed using ± 10 , ± 2 and ± 0.5 ppm error windows with the results for all 1,070 transformations and all 104 datasets being presented in colour coded Excel grids. These are too large to be visualised all at once and so are included in the electronic Appendix (3.2/3.2.4) in full with summary results displayed in Table 30. When a ± 10 ppm error was applied to the matching process 97.1 % of the transformations could be found in at least one of the datasets, compared to 92.9% with ± 2 ppm error, and 76.6% when ± 0.5 ppm error was applied. The fact that more are detected with the slightly wider ppm error alludes to the natural mass error in the datasets resulting in mass differences slightly off the true mass difference of the transformation/fragment, this is particularly likely to be true for the rare transformation products closer to the noise level of the mass spectrometer. On the other hand, the missing identified transformations at lower ppm errors could simply be due to false positive transformation identifications instead of any mass error. It is impossible to tell from this data whether these are true biological transformations or whether they are the result of in-source fragmentation. It is estimated that > 90% of metabolites in METLIN readily dissociate in-source whilst 8% are expected to generate 15 or more fragments (Domingo-Almenara et al., 2018). This indicates that its likely most of these differences will be fragments however it is reasonable to assume that a proportion of them at least could be true biological transformations. There are some common in-source fragments such as CO_2 and H_2O but most are specific to the metabolite in question (Domingo-Almenara et al., 2019). This is seen in the data presented where there are a high percentage of fragments detected overall but the majority of these possible transformations are detected at low frequency and are rare. Only 11 were detected at “high” frequency (average of ≥ 10 across all datasets) when ± 10 ppm error was applied, compared to 0 when applying lower ppm errors. This demonstrates that whilst a wide variety of biotransformations/fragmentations are present in the data and require consideration during the grouping stage of the annotation process there are only a relatively small number that could be

described as common and would make sense for routine consideration during annotation. For the majority of fragmentations/transformations which are rare a specialised method for identification may be required. The recently published MISA (Domingo-Almenara et al., 2019) uses low energy METLIN fragmentation spectra to accurately target fragment identifications and should be considered for integration into workflows for more reliable identification of in-source fragments. Particularly for the correct identification of rare fragments which as our data shows are illogical to search for normally.

Table 30: Summary of the 1070 KEGG biological transformations and the number of them detected at least once across all datasets, and the number detected above certain average thresholds across datasets when matching to the significant correlations ($R \geq 0.8$, $p \leq 0.05$, $n \geq 20$).

	± 10 ppm error	± 2 ppm error	± 0.5 ppm error
Number of Transformations Detected (% of total searched for)	1041 (97.1%)	995 (92.9%)	820 (76.6%)
Detected ≥ 10 times on average	11	0	0
Detected ≥ 5 times on average	21	5	3
Detected ≥ 1 times on average	222	42	23

Which transformations were the most commonly detected and are thus most important to consider in searching? The 21 transformations detected 5 times or more on average across all datasets when a ± 10 ppm error was applied is displayed in Table 31. These are all commonly detected and would be appropriate to select as a small list of possible transformations/fragments for use in an annotation workflow. If frequencies of biological transformations/in-source fragments are known they should also be weighted in the annotation process to favour more commonly detected differences when multiple annotations are possible such as in CliqueMS (Senan et al., 2019). Fragments which are successfully identified can be used to provide further confidence in annotations through their detection particularly if showing low ppm errors and suitable intensity relative to the parent. Parent ions and in source fragments in the MS^1 data can be differentiated through their changes in intensity relative to each other if in-source CID is applied. This was demonstrated as a useful tool for identifying in-source fragments for 75 out of 80 metabolites tested (Wang et al., 2019a). Supplementary strategies such as this and MISA (Domingo-Almenara et al., 2019) which can provide extra confidence in annotations should be implemented into workflows where possible.

Table 31: The 21 transformations detected 5 times or more on average across all datasets when a ± 10 ppm error was applied.

<i>m/z</i> Difference	Biological Transformation/ In-source Fragment	Sum across all datasets	Average in each dataset
44.02622	C ₂ H ₄ O	4039	39
18.01057	H ₂ O	2685	26
46.00548	CH ₂ O ₂	2462	24
26.01565	C ₂ H ₂	2396	23
88.05243	C ₄ H ₈ O ₂	2340	23
17.02655	H ₃ N	1535	15
162.0528	C ₆ H ₁₀ O ₅	1291	12
28.0313	C ₂ H ₄	1254	12
97.9769	H ₃ O ₄ P	1159	11
43.98983	CO ₂	1067	10
14.01565	CH ₂	1056	10
27.99492	CO	1025	10
60.02113	C ₂ H ₄ O ₂	1015	10
30.01057	CH ₂ O	880	8
42.01057	C ₂ H ₂ O	844	8
24.0000	C ₂	843	8
15.99492	O	735	7
35.9765	-CHOP	619	6
12.0000	C	608	6
180.0634	C ₆ H ₁₂ O ₆	561	5
32.02622	CH ₄ O	521	5

3.2.5 Are there homogeneous and heterogeneous dimers observed?

Throughout the data exploration, so far, we have seen that there are wide variety of adducts, isotopes, multiply charged species and possible in-source fragments/biological transformations present. Another expected source of complexity is the observation of dimers. Most annotation software utilise searches for simple homodimers but the only one which has reported and considered heterodimers is mz.unity (Mahieu et al., 2016). Dimers and higher order oligomers whether homogeneous or heterogeneous are likely to form when one or more metabolites are present on their own or together in the ESI source at high concentration (Lynn et al., 2015). 3 positive ion mode and 3 negative ion mode datasets were selected at random from all datasets in the top 20 for highest number of significantly correlated feature pairs. Then from these datasets the top 10% highest intensity features were selected. Where these features fell into a single RT bin then the possibility of dimerization was considered. The frequency of homodimers and heterodimers was estimated through investigation of the data. For each RT window the monomer neutral masses were calculated by adding or subtracting the mass of a proton depending on the ion mode being investigated (assuming all features are

protonated or deprotonated). This list of monomer neutral masses was then searched against the m/z differences between the correlated feature pairs in the window and matches were identified as dimers. If the monomer neutral mass is subtracted from the lower mass of the two features in the pair then the value left over can indicate whether the dimer is a homodimer or a heterodimer. If the value left over is equal to the mass of a proton then it is a homodimer, for example. Other results that would imply a homodimer would include if other commonly detected m/z differences are equal to the value left over such as the sodiated ion m/z difference. As a result, the list of values resulting from the subtraction of the monomer mass from the lower mass feature of each correlated pair in each window were matched to the list of 1038 commonly detected mass differences with ± 0.5 Da error applied. If a match was found to any of the 1038 differences then it was assigned as a homodimer if no match was found then it was assigned as a heterodimer. The resulting data is displayed in Table 32.

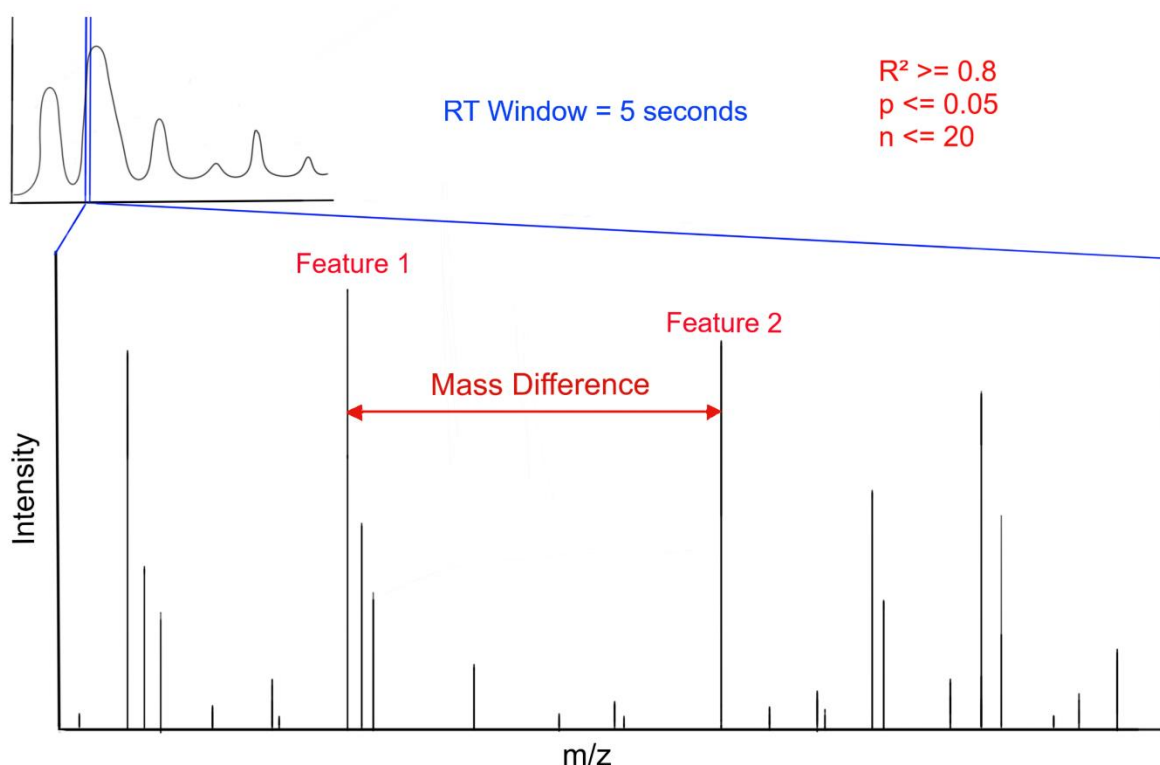


Figure 55: The mass spectrum of the correlated features in a 5 second RT window. If the mass difference between the features is equal to the mass of a monomer from the same window then dimerization was reported.

Table 32: Number of features, number of features pairs, number of dimer pairs, number of significant dimer pairs ($R \geq 0.8$, $p \geq 0.001$, $n \geq 20$), number of homo-dimers estimated and number of hetero-dimers estimated for the 6 datasets investigated.

Dataset ID	Features	Feature Pairs	Significant Pairs (% of total pairs)	Total Dimer Pairs	Homo-Dimers Estimate	Hetero-Dimers Estimate
PCB2_POS	6,177	565,571	86,605 (15.3%)	233	233	0
MTBLS372_NEG	10,401	1,199,938	71,325 (5.94%)	79	74	5
ST236_POS	13,191	1,424,706	50,668 (3.56%)	43	21	22
PCB3_NEG	11,209	1,377,949	203,687 (14.8%)	40	29	11
PCB12_POS	4,894	246,660	41,490 (16.8%)	4	4	0
PCB5_NEG	3,797	341,398	144,625 (42.36%)	0	0	0

Dimers were detected in 5 of the 6 datasets searched but not at very high frequencies, they may be more prevalent than shown here but the parameters utilised for their identification were very stringent and so this will have limited the total number identified. However, the use of these stringent parameters also means a high level of confidence has been provided in the identification of these dimers. This shows that they must be considered during the annotation process to ensure full, complete and correct annotation. The data presented does not absolutely and exclusively identify these as heterodimers but there is a high likelihood that they are heterodimers due to the stringency of the methods for dimer identification and homo/heterodimer differentiation. It is possible some dimers assigned as heterodimers may be homodimers with an unusual adduct which was not in the list of 1038 common mass differences and as a result has been incorrectly assigned as a heterodimer. Alternatively, they could represent unusual in-source fragments but considering that heterodimers have been demonstrated before it is likely some of the heterodimer assignments are correct. This analysis has demonstrated that homodimers and heterodimers are present in the data and should be

considered during the annotation process where possible. Whilst this data confirms their presence it does not truly estimate their frequency. Further work should be performed involving less stringent parameters for dimer identification on a wider number of datasets to further investigate and understand the process of dimerization in the data. Once again this highlights the need for more openly accessible data. Work should also be undertaken to investigate the prevalence of higher order oligomers where three or more molecules are present.

3.2.6 Which Adducts are Searched for in Other Commonly Used Annotation Software?

The number of adducts and mass differences searched for in the most frequently used software is insufficient to achieve full and correct annotation of datasets. Four software which perform grouping and annotation of m/z -RT data were assessed for adduct options available for annotation (Table 33; Human Metabolite Database (HMDB), METLIN, MS-DIAL and Compound Discoverer 3.0). HMDB and MS-DIAL considered the joint highest number of different adducts of the four sources investigated with 56 each when both ion modes were combined. These provide appropriate variety but are clearly insufficient when compared to number of commonly seen m/z differences across all the datasets investigated.

Table 33: The number of different adducts available for searching in some common annotation resources.

Ion Mode	HMDB	METLIN	MS-DIAL	COMPOUND DISCOVERER
Positive	39	17	38	22
Negative	17	10	18	11
Total	56	27	56	33

The individual adducts and whether they were available for each annotation source can be found in the Appendix (

9.1 Characterising the complexity of electrospray ionisation data and its impact on metabolite annotation in UPLC-MS metabolomics studies). There is no consensus between the different software for which adducts are required and therefore very different results could be achieved depending on the annotation source utilised. There were just 11 adducts considered by all four annotation sources in positive ion mode, and just 6 in negative ion mode. These 16 ubiquitous adducts are displayed in Table 34.

Table 34: The adducts considered across all four of the annotation sources assessed.

Adduct	Ion Mode	Type	Charge
M+H	Positive	Adduct	1
M+H-H ₂ O	Positive	Adduct/(Fragment/Transformation)	1
M+NH ₄	Positive	Adduct	1
M+Na	Positive	Adduct	1
M+CH ₃ OH+H	Positive	Adduct/(Fragment/Transformation)	1
M+K	Positive	Adduct	1
M+ACN+H	Positive	Adduct/(Fragment/Transformation)	1
M+ACN+Na	Positive	Adduct/(Fragment/Transformation)	1
M+2H	Positive	Adduct/Multiply Charged	2
M+H+Na	Positive	Adduct/Adduct/Multiply Charged	2
M+3H	Positive	Adduct/Multiply Charged	3
M-H	Negative	Adduct	1
M-H ₂ O-H	Negative	Adduct/(Fragment/Transformation)	1
M+Cl	Negative	Adduct	1
M+K-2H	Negative	Adduct/(Fragment/Transformation)	1
M+FA-H	Negative	Adduct/(Fragment/Transformation)	1
M-2H	Negative	Adduct/Multiply Charged	2

The theoretical m/z difference between each adduct listed in the software assessed and the protonated or deprotonated adduct was calculated. These masses were then searched against the list of 1038 common mass differences generated. For multiply charged species the theoretical mass difference between the doubly or triply charged protonated ion and the multiply charged ion in question was calculated. The mass differences for dimers could not be calculated as they are dependent on the mass of the metabolite. In positive ion mode there were 20 different singly charged adduct and fragment combinations for comparison to the acquired data. 13 of these were within 0.001

Daltons of a mass in the list of 1038 common mass differences, 3 were within 0.01 Daltons and 4 were not close to any common mass difference. In negative ion mode there were 15 singly charged adduct and fragment combinations and only 7 of these were within 0.001 Daltons of something on the common mass difference list. There were 4 more within 0.01 Daltons and 4 with no close match. For the doubly charged ions there were 8 different candidates for matching in positive ion mode, 6 of these matched within 0.001 Daltons whilst 2 were not within 0.01 of any difference in the list. There were no doubly charged negative ions. Whilst the triply charged differences calculated did not produce any matches to the common mass difference list. This data demonstrates not only that the number of adducts included is insufficient but that some of the commonly searched for adducts are not in fact commonly detected and their use may lead to false positive annotations while not providing an increase in true positive annotations (Table 35).

Table 35: The adducts which were not within 0.01 Daltons of a difference on the 1038 common mass differences list. The mass difference is the difference from the adduct displayed to the protonated/deprotonated ion for singly charged ions. For doubly and triply charged adducts it is the difference to the doubly or triply protonated ion.

Adduct	Ion Mode	Charge	Mass Difference
M+ACN+Na	Positive	1	63.0085
M+2K-H	Positive	1	75.9118
M+2ACN+H	Positive	1	82.0531
M-C ₆ H ₈ O ₆ +H	Positive	1	176.0321
M+3ACN+2H	Positive	2	61.5398
M+IsoProp+Na+H	Positive	2	41.0197
M+2H+Na	Positive	3	7.327315
M+H+2Na	Positive	3	14.65463
M+H+2K	Positive	3	25.3039
M+K-2H	Negative	1	39.971535
M+Br	Negative	1	79.9262
M+CH ₃ COO-H	Negative	1	59.0133
M-C ₆ H ₈ O ₆ -H	Negative	1	176.0321

We have shown above that a larger number of unique ion types and in-source fragments/biological transformations are present in ESI metabolomics datasets than has previously been reported or applied in metabolite annotation software. The subsequent logical conclusion is that more adducts

and other possible derivative feature types need to be considered during the annotation process. This perhaps illustrates why such difficulty is had in annotating untargeted metabolomics datasets. The number of different features in the dataset that can be identified is low (when a small adduct and molecular formula list is applied) as an insufficient number of adducts, isotopes and other derivative ion types are being considered in the feature grouping process. By increasing the number of adducts it is likely we would increase the number of true positive identifications, but the number of false positive identifications will increase far more, this effect has been previously reported (Mahieu et al., 2016). The number of possible annotations will rapidly increase as the number of different derivative feature types are considered and so the chances of a false annotation being assigned as the top annotation for a feature increases. Therefore, a balance needs to be struck between having enough adducts and isotopes considered that correct annotation is achieved but not so many that the number of spurious annotations becomes too high or the computational strain becomes too great. To counteract this there needs to be a mechanism of ranking the possible annotations assigned such as in IPA (Del Carratore et al., 2019) and CliqueMS (Senan et al., 2019).

This does not factor in the possibility that MS² data could be utilised to filter the list of putative annotations. If good quality MS² data were collected for a feature then a long list of candidates is not a problem. For all features without MS² data however it creates extra uncertainty and confusion with many possible further candidates for investigation. This lack of confidence means further mechanisms are required for improving confidence. This should involve lab/assay tailored adduct lists as the adducts appropriate to include are likely to be different depending on the sample, solvents, modifiers and analytical instrument used for analysis as well as a whole host of other parameters. If labs characterise the common mass differences seen in their systems when applying routine assays this would allow more extensive, appropriate and specific adduct, isotope and transformation/fragment lists to be considered so that more accurate and complete annotation can be achieved. Other approaches could be included to increase confidence in MS¹ annotations such as credentialing technology, large and accurate reduction in the MS¹ data complexity has been achieved with this (Mahieu et al., 2016; Wang et al., 2019a) however this can only be performed for samples generated from bacterial cultures. Some software also integrate the positive and negative ion mode datasets to provide further confidence (Wang et al., 2019a).

3.3 Conclusions

A large number of untargeted datasets were assessed for the commonly detected mass differences between related features, this had never been done before. It was shown that the complexity of UHPLC-ESI-MS full scan data is astounding, poorly understood and not appropriately considered and

accounted for in most routinely applied annotation software that uses MS¹ data. In most untargeted datasets only a small number of metabolite features can be annotated. The data presented here has clearly demonstrated the magnitude of the problem with such variety present in the 1038 commonly detected m/z differences and this should encourage the community to focus more attention to ensuring ESI data can be fully understood.

It is not surprising that such a small number of features can be annotated when the complexity of the m/z difference data is considered and compared to the simplicity of the adduct lists applied by commonly used annotation software. On the other hand it was surprising to see that some of the adducts searched for in commonly used annotation software were not found in the data presented. Approaches which consider all derivative ion types (adducts, isotopes, in-source fragments/biological transformations, multiply charged species, homodimers, heterodimers) as well as all possible combinations of these need to be applied. As there will be multiple possible groups for different features a method ranking or prioritising the most chemically probable result needs to be included. However, by considering this number of possibilities the computational strain is likely to be very high potentially causing problems for users. Furthermore, the number of false positives will likely increase if the method of annotation ranking is not highly effective. To ensure this the ranking needs to be able to consider all features simultaneously in a holistic fashion to find the best explanation for the features detected. This will require the development of new software.

The adducts present will be dependent on the specific experimental characteristics such as the assay, sample type and instrument that were applied, this was clearly seen in the mass difference frequency comparisons (3.2.3). However, the number of different possible factors influencing them are very high and the process of their formation in the ESI source is not well understood. Further work investigating a greater number of datasets should be performed to truly determine how different experimental characteristics effect the mass differences seen for different assays, sample types and instruments the volume of datasets utilised in this case was too small. Furthermore the intensity correlation cut offs and RT cut offs should be tailored for each dataset as these were likely suboptimal for many datasets and may have detrimentally impacted results

Other more targeted work involving the analysis of pure standards in solution and pure standards spiked into complex matrices should also be performed to see how many derivative feature types are detected for common metabolites representative of different metabolite classes in a clean and simple solution. Isotopically labelled standards could be utilised to provide extra confidence in identification of relevant features from noisy or contaminant peaks. The work could then be repeated in complex

matrices to provide an understanding of how the derivative feature types change with concentration and overall sample complexity.

Considering the complexity of the issue and the number of possible factors influencing the results researchers should look to characterise the commonly detected and correlated mass differences in their own data, try to understand them as best as possible and report this as key information related to the study with appropriate meta data. As time passes and the knowledge base increases, the appropriateness of the adducts selected for consideration should increase and thus so should the accuracy of annotation and thus the success of untargeted metabolomics as a whole. A reference database could be created for this purpose. Lists of adducts/correlated mass differences most commonly detected in that study could be uploaded with appropriate meta data (the solvents/modifiers/sample utilised). Gradually it would become clear which adducts, isotopes, fragments and other derivative ion types are most important for a particular solvent, column or other key experimental characteristics. For example, if water and methanol are applied as UHPLC solvents then acetonitrile adducts would not be expected and so can be excluded.

A key issue preventing progress and creating extra confusion is the differences in experimental conditions and sources of solvents and chemicals applied by different labs. UHPLC-ESI-MS is such a sensitive technology and a wide variety of experimental conditions and instrumental factors will impact the results. This variety alongside the inherent data complexity is inhibiting the progress of untargeted metabolomics. If standardised assays existed for certain applications and were commonly applied to a rigorous standard then the mass differences could be much more easily characterised and understood for these assays allowing more accurate annotation. This is unlikely to occur but in general highlights how there needs to be greater cooperation in the field to overcome the many challenges presented.

4.0 Characterisation of UHPLC-MS full
scan data complexity and its influence
on MS² data collection on Q Exactive
mass spectrometers

4.1 Introduction

The goal of untargeted metabolomics is to profile as many metabolites within a sample in an unbiased and reproducible manner. In UHPLC-MS applications applying an untargeted metabolomics approach, three types of data are normally collected; (1) chromatographic retention time data, (2) full scan accurate m/z data and (3) MS^n data where n is typically 2 but can be greater when ion trap instruments are applied. A fourth data type can be ion mobility drift/migration time on specific instruments. For increased confidence of metabolite annotations or identifications, MS^2 data can provide information on the metabolite's structure and allow discrimination of isobaric metabolites. To annotate or identify metabolites, we need to collect MS^2 data for all metabolites within the dataset as without this information a feature cannot be confidently identified. Traditionally, MS^2 data is collected applying a Data Dependent Analysis (DDA) method (Nash and Dunn, 2019). However, this strategy is inherently biased towards ions of high intensity and as a result is not suitable for the collection of MS^2 data for all ions or metabolites (Mullard et al., 2014). Data Independent Analysis (DIA) is an alternative method for collection of MS^2 data which theoretically can provide unbiased fragmentation of all metabolites within the sample followed by data deconvolution to reconstruct the MS^2 mass spectrum (Tsugawa et al., 2015). This method employs DIA fragmentation windows (typically a window width of roughly 25 m/z is applied) (Bonner and Hopfgartner, 2018) where multiple features will be fragmented simultaneously generating complex MS^2 spectra composed of the spectra of multiple ions/metabolites. The precursor/product ion relationships can theoretically be elucidated from the chromatographic profiles of the respective precursor and product ions. This means the complex MS^2 spectra generated in DIA can be deconvoluted to provide each parent ion with a "pure" MS^2 mass spectrum. This is achieved through the mathematical comparison of their chromatographic profiles, product ions should have the same chromatographic profile as their precursors and a different chromatographic profile if they are unrelated. It is logical to assume that the deconvolution process can be much more effective as long as the number of features falling within any one window is kept low although the quality of the chromatographic separation is important too to reduce co-elution of features. Therefore, it is important when designing a DIA method to provide a good understanding of the complexity and density of features within the sample to try ensure low complexity in each DIA window, which hopefully will provide good data quality.

Whilst the complexity of the data is a key aspect in the design of a DIA method there are other very important factors which require understanding, compromise and balance (Figure 56). As discussed, the sample complexity and density of features will influence the complexity of data collected for any DIA window. As the user should be looking to reduce the complexity of data for all DIA windows we can say that sample complexity influences the DIA window characteristics to apply. The window

characteristics are defined by the m/z width, number of DIA windows and therefore cycle time and m/z range of the windows. The narrower the window, theoretically, the lower the complexity and therefore the higher the quality of the MS^2 data. Decreasing the DIA window width over the same total mass range will increase the number of windows in a cycle. This detrimentally effects the MS^1 scan rate because more time is required to carry out the extra MS^2 scans in between each MS^1 scan. A decreased MS^1 scan rate negatively impacts the quality of the MS^1 data and a minimum number of data points should be acquired across each peak although the number is not necessarily agreed upon across the community. The author suggests a minimum of 7 data points across a chromatographic peak as a minimum for accurate peak picking and quantification. The minimum number required is disputed but 6 to 10 have both been recently reported (Bian et al., 2020; Melnikov et al., 2020). Ideally 11 or more data points should be collected, the more that can be collected the more accurate the peak shape and area will be. If the mass spectrometers scan rate is too slow this will not be achieved for many chromatographic peaks in the data or worse, peaks may be narrow enough that they are not detected in a single MS^1 scan. Therefore, maintenance of an appropriate MS^1 scan rate is essential. To have an understanding of the minimum scan rate that is required the peak widths detected in the data must be understood. Finally, the user must consider the mass resolution to apply for the MS^1 and MS^2 data collection. In an Orbitrap-based system the resolution is determined by the length of the transient recorded and thus the resolution increases as scan time increases (Najdekr et al., 2016). On an Orbitrap-based system the scan rate in particular is a limiting factor when compared to TOF instruments on which DIA was originally developed and is most typically applied in the literature (Zhu et al., 2014; Wang et al., 2018; Hopfgartner et al., 2012; Bonner and Hopfgartner, 2018; Yan et al., 2019; Wang et al., 2019b).

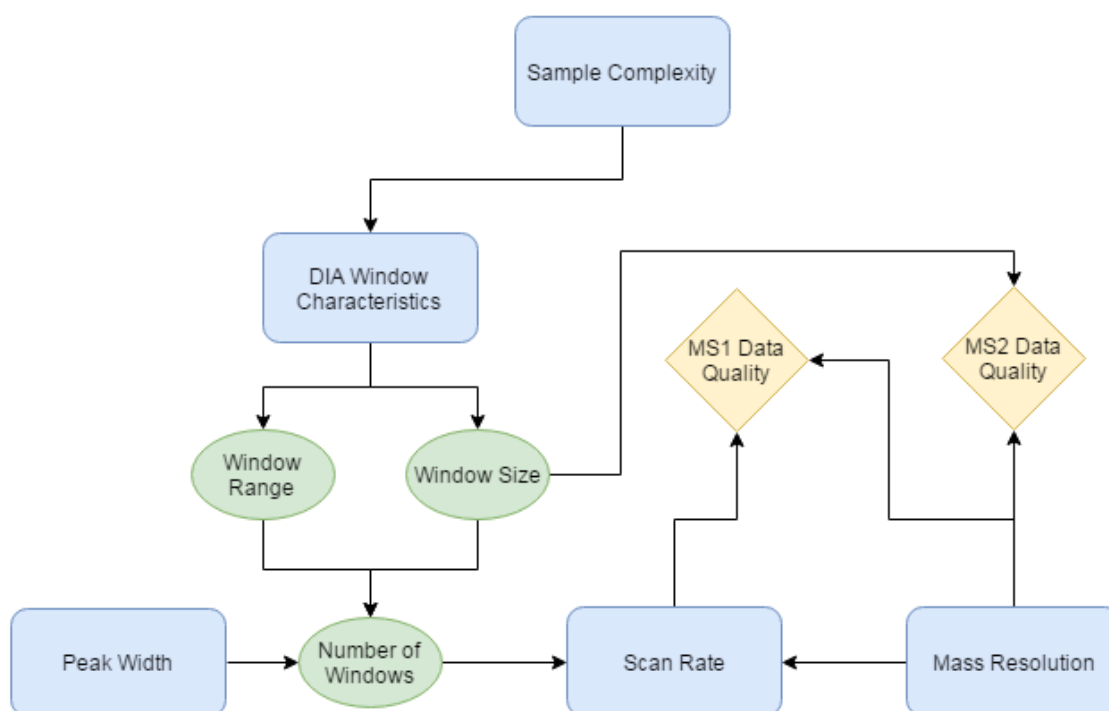


Figure 56: The main factors that influence the design of a DIA experiment and the impact on data MS¹ and MS² data quality.

In this chapter the research performed to characterise the key factors discussed above and displayed in Figure 56 for a variety of different assay, sample type, ion mode, resolution and DIA window characteristic combinations utilising a Q Exactive Plus (Thermo Fisher Scientific, USA) mass spectrometer will be presented. Full scan MS¹ data were first collected at varying resolutions (17,500, 35,000, 70,000, 140,000) for three different sample types (plasma, urine and sheep liver tissue), using three different assays (RP, HILIC, Lipidomics) in both ion modes (positive and negative). These data were collected at varying resolutions to determine which resolution was required as a minimum in the full scan to ensure adequate feature detection and that information is not being missed. To get an accurate representation of the number of true features within the data then it is important that data processing parameters utilised are appropriate. Data processing parameters are important as they can turn good raw data into bad processed data. XCMS (Smith et al., 2006) was used to perform peak picking, grouping and alignment. IPO (Libiseller et al., 2015) was used to optimise the peak picking parameters for each assay/sample/ion mode/resolution combination. Optimised parameters were utilised for processing of the data. Once the subsequent data processing had been performed the minimum resolution required to capture the true complexity of the data can be determined, defining the MS¹ resolution required in future experiments. Furthermore, the resultant feature lists were input into a data complexity spreadsheet tool previously developed by the author. This allows the

visualisation of the complexity of DIA windows of varying size (5, 25, 50, 100, 200 m/z) across the retention time (RT) axis. The peak widths seen in the data were also extracted and analysed and the estimated scan speeds at varying resolutions and number of MS² events were estimated. Ultimately this work allowed the planning of appropriate experiments to test the merits of DIA, as well as other, MS² data acquisition strategies on the Q Exactive Plus. Overall this will help to improve metabolite annotation in untargeted studies on instruments of this type or others which are closely related.

4.2 Results and Discussion

Full scan data were collected for three different sample types (plasma, urine and sheep liver tissue), using three different assays (HILIC, RP, lipidomics), in both ion modes separately and at four different mass resolutions (17,500, 35,000, 70,000, 140,000) to determine the appropriate full scan resolution for detection of most features, to allow the assessment of theoretical DIA windows, and to allow appropriate DIA method development. The first step after data had been collected was to implement optimisation of XCMS parameters to ensure reliable peak picking was being carried out. The lipidomics positive ion mode tissue data were not used due to failures during data collection and no lipidomics data were collected for urine.

4.2.1 XCMS Parameter Optimisation

8 parameters applied in XCMS were optimised using the IPO R package (Libiseller et al., 2015) for each different combination of sample type/assay/ion mode/mass resolution that had been analysed. The 8 parameters in question were the minimum peak width, maximum peak width, ppm, mzdiff, gapInit, gapExtend, bw and mzwid (Table 36). Clear differences and trends could be seen when applying different assay/mass resolution combinations on the minimum and maximum peak width parameters as well as the ppm parameter. Specific examples of this will be demonstrated later in this section. There were not clear differences in optimised parameters between ion modes and sample types. As a result the median for each assay and mass resolution combination (Table 37) was taken to provide a summary of the data collected. The median was used instead of the mean to negate the influence of the outliers in the data, of which some were strong which would have had a very strong effect due to the small sample sizes. The median values were then used to process the collected data for further analysis with the assay and mass resolution applied determining the value of the parameters as shown in Table 37. The data clearly show that no single set of XCMS parameters can be applied for all combinations of sample type UHPLC assay and mass resolution. All other supplementary data for XCMS parameter optimisation can be found in the Appendix (9.2.1).

Table 36: Definition of the XCMS parameters that were optimised (Albóniga et al., 2020) and a brief summary of their importance if known.

Parameter (Process)	Definition	Importance
Minimum Peak Width (Peak Picking)	Estimation of the width in seconds of the narrow peaks in the data.	Too small and many peaks will be split, too wide and many peaks will be missed.
Maximum Peak Width (Peak Picking)	Estimation of the width in seconds of the widest peaks in the data.	Too small and many peaks will be missed, too large has less impact.
ppm (Peak Picking)	The maximum ppm error allowed between m/z values in consecutive scans and be considered as the same peak.	Appropriate ppm for mass resolution utilised, too low and peaks are lost, too high has less impact.
mzdiff (Peak Picking)	The minimum m/z difference for peaks with overlapping RTs to be considered as different peaks.	Too small will increase feature numbers, too high will decrease feature numbers.
gapInit (RT Alignment)	Penalty for gap opening.	Unsure of impact.
gapExtend (RT Alignment)	Penalty for gap extending.	Unsure of impact.
bw (Grouping)	The RT window that determines if adjacent peaks are included in the same group across samples.	Too high increases feature numbers, too low reduces feature numbers.
mzwid (Grouping)	The width in m/z considered for grouping peaks across samples.	Too high increases feature numbers, too low reduces feature numbers.

Table 37: Median result of XCMS parameter optimisation for each assay/resolution combination after all triplicates were processed, number of triplicates used is detailed in the “n” column. A complete set of data would have included 6 triplicates for HILIC and RP assay/resolution combination and 4 triplicates for each lipidomics combination.

UPLC Method	Resolution	n	Min peak width (s)	Max peak width (s)	ppm	Mzdiff	gapInit	gapExtend	bw	mzwid
HILIC	140,000	6	7.1	32.5	10.0	0.0010	0.46	2.4	0.25	0.0027
HILIC	70,000	6	3.7	36.6	12.3	0.0054	0.42	2.4	0.25	0.0067
HILIC	35,000	6	3	78.3	14.6	-0.0219	0.62	2.4	0.25	0.0107
HILIC	17,500	2	2.5	18.3	23.2	0.0056	0.30	2.4	0.88	0.1385
Lipids	140,000	3	8.0	32.5	7.40	0.0038	0.40	2.9	0.25	0.0073
Lipids	70,000	3	5.9	25.0	13.6	0.0012	0.42	2.4	0.88	0.0107
Lipids	35,000	3	5.5	23.5	17.3	-0.0230	0.38	2.4	0.25	0.0107
Lipids	17,500	2	4.4	21.3	31.7	0.0206	0.32	2.6	1.65	0.0204
RP	140,000	6	5.4	25.0	9.10	-0.0010	0.44	2.4	0.52	0.0140
RP	70,000	6	3.0	32.5	11.1	0.0022	0.42	2.5	0.52	0.0093
RP	35,000	6	3.0	24.6	19.1	0.0070	0.40	2.4	0.39	0.0130
RP	17,500	3	3.0	25.0	34.4	0.0056	0.44	2.5	0.52	0.0140

The mass resolution applied during data acquisition was a key factor in determining the optimal parameters for the minimum peak width and the ppm parameters. This can be seen in Figure 57 where the median result for minimum peak width for a mass resolution of 140,000 was double that seen for a mass resolution of 70,000. There is a clear trend that as the mass resolution increases so does the optimised minimum peak width parameter although this is less pronounced at lower mass resolutions. A key driver in this effect is likely to be the slower scan rate at higher mass resolutions. This slower scan rate can be clearly seen and demonstrated by looking at the number of MS¹ data points across the peak of a commonly detected metabolite such as phenylalanine, in the HILIC/Plasma/Positive data (Figure 58). This is interesting as the assay/sample type stay the same as the resolution changes, meaning that the raw data should still be the same in terms of peak widths. This is clearly seen in Figure 58 where the same peak exhibits the same width in seconds across the different resolutions

but the number of data points has changed dramatically. This clearly creates differences in the way XCMS picks these peaks and as a result the minimum peak width parameter increases as the mass resolution increases. It appears that the greater the spread of MS¹ data points the wider the software will generally consider peaks to be.

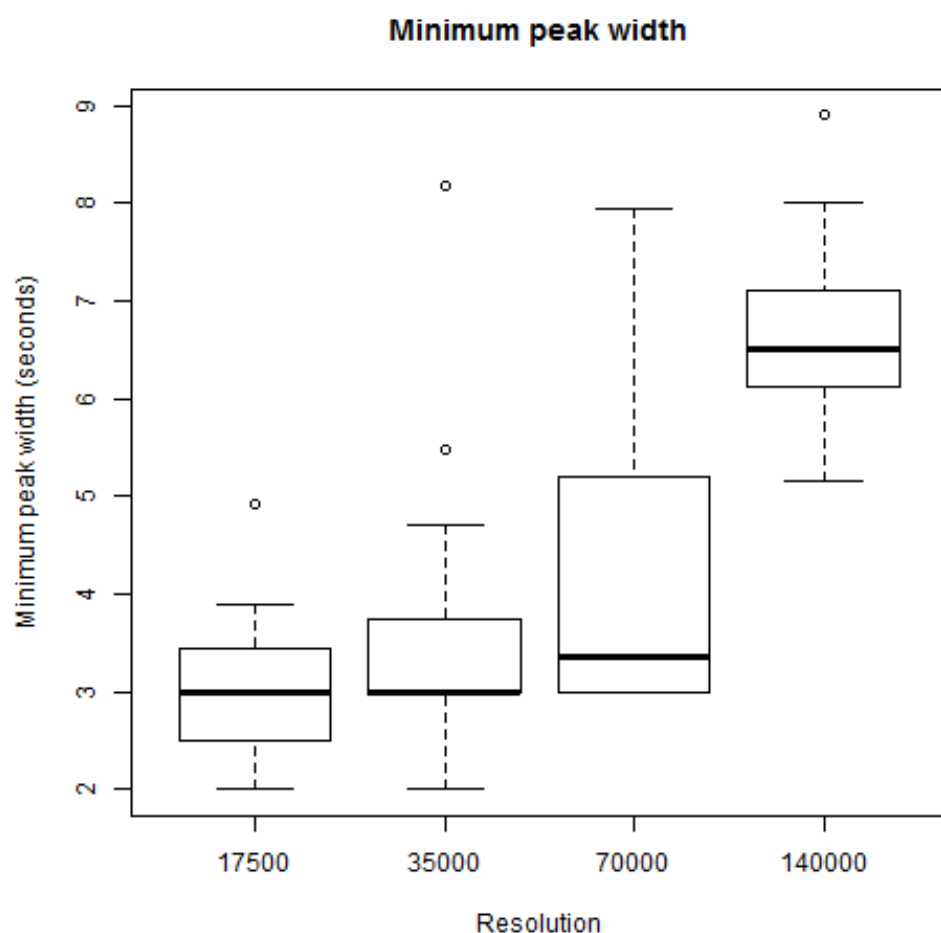


Figure 57: Boxplots representing the IPO optimised minimum peak width XCMS parameters across triplicates of varying resolution for all sample types (17,500 $n=7$, all other resolutions $n = 15$)

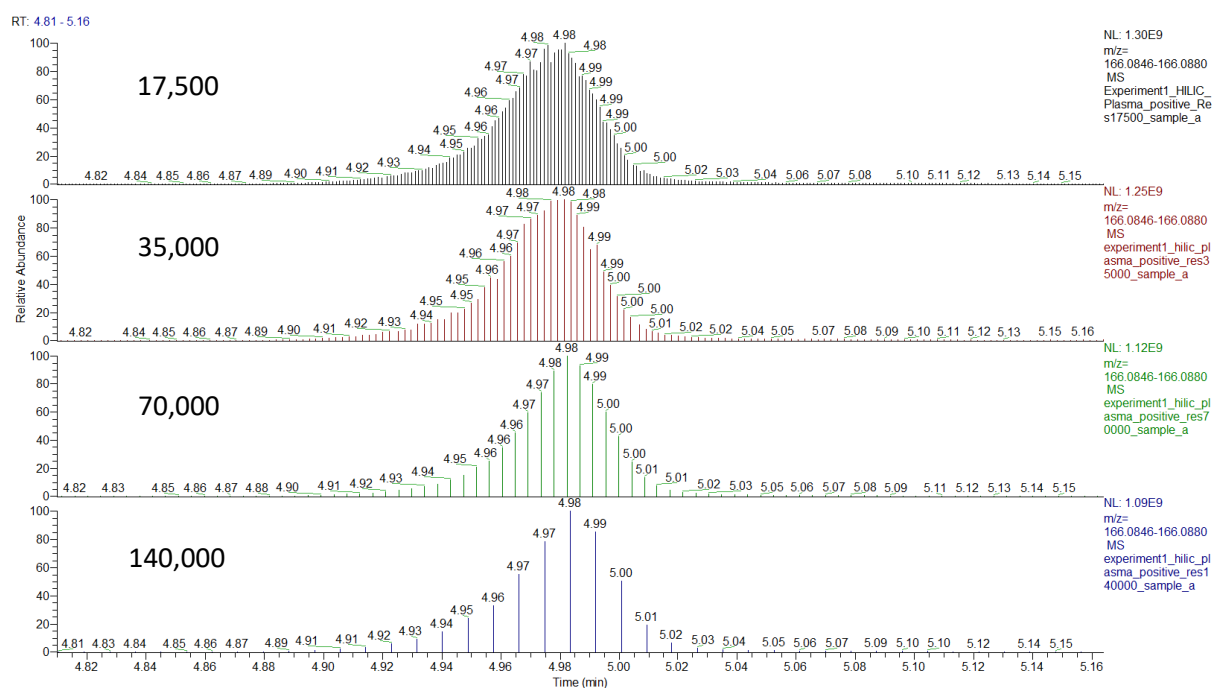


Figure 58: EIC's of phenylalanine's molecular ion from the HILIC/Plasma/positive combination at each resolution.

The analytical method utilised is also an important factor as evidenced by Figure 59. This is not surprising as the different assays employ different solvent systems and columns which result in different, physicochemical selectivity and levels of pressure. Figure 59 indicates the lipidomics data have wider chromatographic peaks than those seen with the HILIC and RP methods as the optimal minimum peak width is clearly greater. The HILIC peak widths are slightly larger than those seen with the RP method. There is some overlap between error bars which is partially due to the fact that the sample sizes were relatively small ($2 \leq n \leq 9$). Sample sizes were limited for 17,500 and 35,000 mass resolution triplicates by the computational processing time required to generate the results for these lower resolution files. The IPO software requires a high amount of computational power and the low resolution files contain more data points for the optimisation to be carried out upon and so required significantly longer processing time than the higher resolution files. Even with the use of the BlueBEAR high performance computing cluster this was in the order of days and not all triplicates were completed as a result. Despite the decreased volume of data used for the lower resolution files it does not appear to have had a negative effect on the data with the low resolution data showing a similar pattern between assays as seen at the higher resolutions with complete datasets.

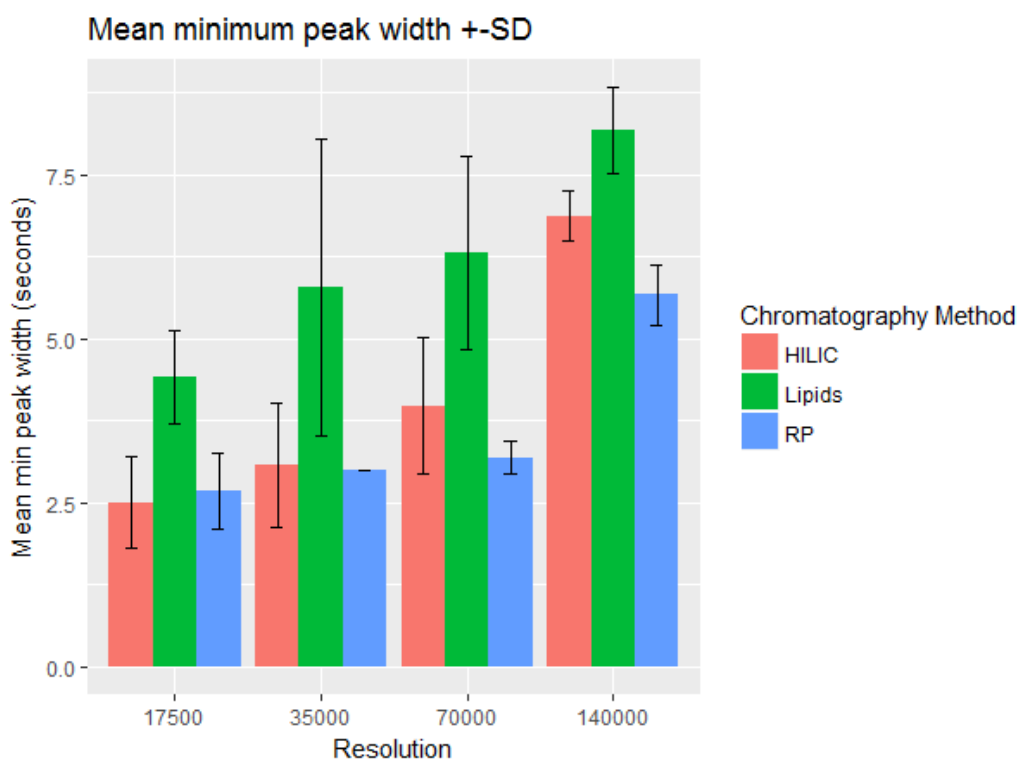


Figure 59: Mean minimum peak width \pm SD for all sample types analysed at different resolutions with different analytical methods. For values of n refer to Table 37.

Figure 60 demonstrates again the impact that the mass resolution has on determining the optimised XCMS parameters. There is a clear trend of a decreasing ppm parameter as the mass resolution increases. This is expected as ppm is a measure of mass accuracy and greater mass accuracy can be achieved by operating at a higher mass resolution (Peterson et al., 2012). The range of the plots also increases as the resolution decreases. This greater level of variation is another indicator of lower resolution as the values being recorded are less accurate due to the shorter length of time spent scanning in the Orbitrap to generate the data lead to less accurate mass measurements (Najdekr et al., 2016).

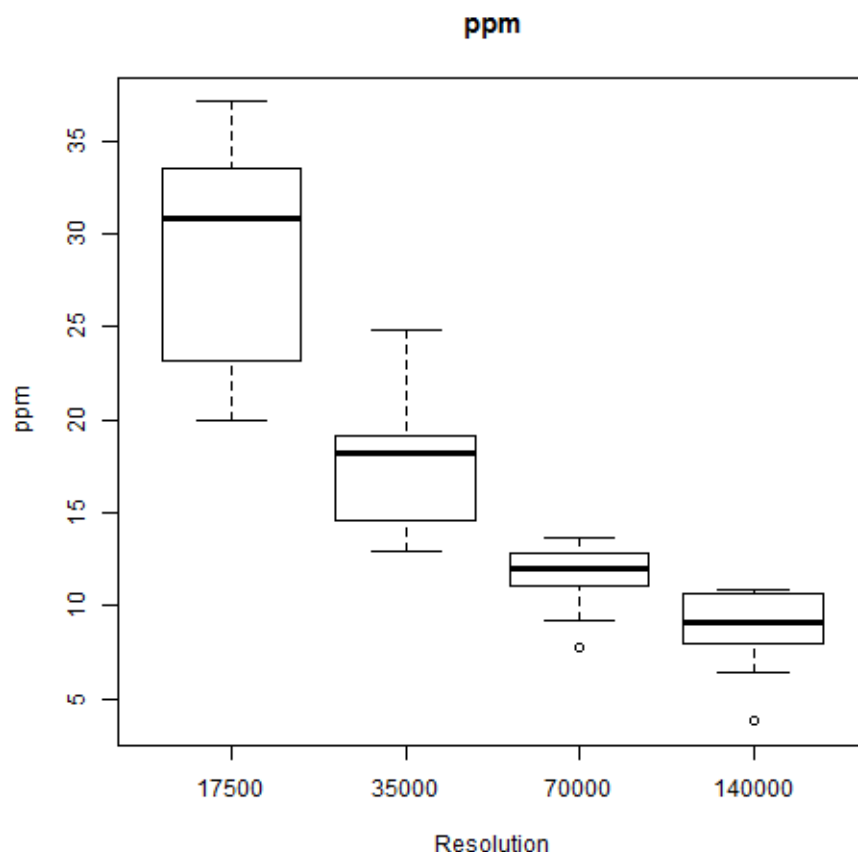


Figure 60: Distribution of the optimal ppm parameter results from the optimisations across the four resolutions used for all sample types (17,500 $n=7$, all other resolutions $n = 15$).

Although there were clear trends seen with the minimum peak width and ppm parameters no clear trends were seen with the other optimised parameters (9.2.1). It is important to note that not all 60 possible triplicates were successfully processed. 8 out of 15 of the 17,500 resolution triplicates went over the allowed duration of processing time and had to be aborted These were HILIC positive plasma and urine, HILIC negative tissue and urine, RP positive tissue and urine, RP negative tissue and lipidomics negative plasma.

Due to the fact that for the minimum peak width parameter there were clear differences between the different chromatography assay and mass resolution combinations the median result for all triplicates processed of each combination were taken as the parameters to be utilised for the final XCMS processing (Table 37).

4.2.2 Determination of appropriate MS¹ mass resolution

After optimisation of XCMS parameters the triplicates were processed using the parameters outlined in the methods section and Table 37 as determined by the chromatography assay, sample type, mass resolution and ion mode. The metabolite features in each list were filtered by their relative standard

deviation (RSD); features were removed if the $RSD \geq 30\%$. The total number of features identified for each triplicate and total number of features following RSD filtering are shown in Table 38. For some triplicates one of the three replicates was removed for the RSD calculation due to lack of similarity to the remaining two triplicates causing an unrepresentative number of features of $RSD \leq 30\%$. Triplicates with a sample removed are indicated by a * next to the RSD value.

Table 38: The total number of features detected and the number of features of $RSD \leq 30\%$ for each triplicate after processing with the IPO optimised XCMS parameters, RSD values with a * adjacent to it indicates that one sample was removed from the RSD calculation.

UHPLC Method	Sample Type	Resolution	Ion Mode	Total Features	Features with $RSD \leq 30\%$ (Percentage of total features)
HILIC	Plasma	140,000	Positive	11,202	6,438 (57.5%)
HILIC	Plasma	70,000	Positive	11,853	7,869 (66.4%)*
HILIC	Plasma	35,000	Positive	12,300	5,001 (40.7%)*
HILIC	Plasma	17,500	Positive	6,650	3,189 (48.0%)*
HILIC	Plasma	140,000	Negative	13,597	10,913 (80.3%)*
HILIC	Plasma	70,000	Negative	12,324	7,160 (58.1%)*
HILIC	Plasma	35,000	Negative	11,589	6,558 (56.6%)*
HILIC	Plasma	17,500	Negative	7,045	4,607 (65.4%)*
HILIC	Tissue	140,000	Positive	10,541	5,073 (48.1%)
HILIC	Tissue	70,000	Positive	13,922	8,427 (60.5%)*
HILIC	Tissue	35,000	Positive	9,711	2,720 (28.0%)
HILIC	Tissue	17,500	Positive	6,359	2,110 (33.2%)
HILIC	Tissue	140,000	Negative	12,282	9,110 (74.2%)*
HILIC	Tissue	70,000	Negative	17,341	4,859 (28.0%)
HILIC	Tissue	35,000	Negative	10,902	3,898 (35.8%)
HILIC	Tissue	17,500	Negative	5,381	2,965 (55.1%)*
HILIC	Urine	140,000	Positive	15,224	11,079 (72.8%)*
HILIC	Urine	70,000	Positive	18,662	12,764 (68.4%)
HILIC	Urine	35,000	Positive	14,878	7,213 (48.5%)
HILIC	Urine	17,500	Positive	9,680	3,746 (38.7%)
HILIC	Urine	140,000	Negative	17,515	12,202 (69.7%)
HILIC	Urine	70,000	Negative	17,352	12,249 (70.6%)

HILIC	Urine	35,000	Negative	13,839	6,524 (47.1%)
HILIC	Urine	17,500	Negative	9,814	6,462 (65.8%)
RP	Plasma	140,000	Positive	17,930	12,406 (69.2%)
RP	Plasma	70,000	Positive	18,583	13,375 (72.0%)
RP	Plasma	35,000	Positive	16,040	10,086 (62.9%)
RP	Plasma	17,500	Positive	9,598	5,735 (59.8%)
RP	Plasma	140,000	Negative	16,031	11,727 (73.2%)
RP	Plasma	70,000	Negative	17,592	10,566 (60.1%)
RP	Plasma	35,000	Negative	14,436	9,620 (66.6%)
RP	Plasma	17,500	Negative	8,856	6,156 (69.5%)
RP	Tissue	140,000	Positive	12,331	8,035 (65.2%)
RP	Tissue	70,000	Positive	13,765	9,155 (66.5%)
RP	Tissue	35,000	Positive	8,706	5,393 (61.9%)
RP	Tissue	17,500	Positive	6,418	3,368 (52.5%)
RP	Tissue	140,000	Negative	11,264	8,283 (73.5%)
RP	Tissue	70,000	Negative	15,121	9,509 (62.9%)
RP	Tissue	35,000	Negative	10,803	8,688 (80.4%)
RP	Tissue	17,500	Negative	6,384	5,090 (79.7%)*
RP	Urine	140,000	Positive	20,113	14,714 (73.2%)
RP	Urine	70,000	Positive	25,699	19,890 (77.4%)
RP	Urine	35,000	Positive	18,637	13,841 (74.3%)
RP	Urine	17,500	Positive	13,949	9,304 (66.7%)
RP	Urine	140,000	Negative	17,371	11,854 (68.2%)
RP	Urine	70,000	Negative	22,840	18,818 (82.4%)*
RP	Urine	35,000	Negative	15,757	8,512 (54.0%)
RP	Urine	17,500	Negative	11,252	7,953 (70.7%)
Lipidomics	Plasma	140,000	Positive	5,451	4,731 (86.8%)
Lipidomics	Plasma	70,000	Positive	5,151	4,283 (83.1%)
Lipidomics	Plasma	35,000	Positive	4,263	3,145 (73.8%)
Lipidomics	Plasma	17,500	Positive	3,017	2,085 (69.1%)
Lipidomics	Plasma	140,000	Negative	5,023	4,442 (88.4%)
Lipidomics	Plasma	70,000	Negative	4,636	3,798 (81.9%)
Lipidomics	Plasma	35,000	Negative	3,875	2,794 (72.1%)

Lipidomics	Plasma	17,500	Negative	2,617	1,684 (64.3%)
Lipidomics	Tissue	140,000	Negative	3,954	3,364 (85.1%)
Lipidomics	Tissue	70,000	Negative	2,504	2,140 (85.5%)
Lipidomics	Tissue	35,000	Negative	2,517	1,585 (63.0%)
Lipidomics	Tissue	17,500	Negative	1,222	685 (56.1%)

The values in Table 38 were grouped by chromatographic assay, averaged and are displayed in Table 39. This data demonstrates that the RP method detected the most metabolite features with an average number of 14,562. The HILIC method also produces a high number of features, 12,082 on average whilst the lipidomics method produces significantly fewer features with an average of only 3,686. The lipidomics method however produces more consistent data, this is demonstrated by the high mean percentage of features demonstrating a RSD ≤ 30 of 75.8%. The RP data is the second most consistent with an average of 68.5% whilst the HILIC data were the least consistent with an average of 55.9%. This is as expected as 11 out of the 13 triplicates which had a replicate removed were from the HILIC data. This data indicates that the lipidomics UHPLC type would be far more conducive to a DIA method due to the decreased complexity in terms of numbers of features in comparison to the HILIC and RP methods. The lower complexity means that a smaller number of features are falling into any one DIA window and the more likely it is that the subsequent MS² data collected will be informative. On the other hand, many lipids have highly similar chemical structures which would produce the same or similar product ions (O'Connor et al., 2017) which may mean they are less suitable for a DIA method than RP or HILIC despite their greater sample complexity.

Table 39: Mean total features and mean percentage of total features with RSD $\leq 30\%$ grouped by chromatography method.

UHPLC Type	Mean Total Features \pm SD	Mean Percentage of features of RSD $\leq 30\% \pm$ SD
HILIC	12,082 \pm 3664	55.9 \pm 15.2
RP	14,562 \pm 4922	68.5 \pm 7.8
Lipidomics	3,686 \pm 1308	75.8 \pm 10.9

The data in Table 38 was again grouped, this time by mass resolution, the mean values were calculated and they are displayed in Table 40. The mean number of features detected increases as the mass resolution increases up to a mass resolution of 70,000. The mean number of features detected at a

mass resolution of 70,000 is over double the number seen for a mass resolution of 17,500. There is then a decrease of nearly 2,000 features to the mean number seen at 140,000 mass resolution. The standard deviations are very large due to the fact the different UHPLC assays produce vastly different numbers of features. Although the mean shows that 70,000 mass resolution produces more features than 140,000 this is not the case in all examples as can be seen from the data in Table 38. In 5 of the 15 sets of comparable triplicates the 140,000 mass resolution triplicate produced more total features than the 70,000 triplicate. Interestingly, this was the case for all the lipids data but none of the RP data. This is logical as the lipids being detected will generally consist of a higher volume of features of highly similar m/z due to the nature of lipid classes and their inherent diversity but also high structural similarity (Rustam and Reid, 2018) compared to the RP or HILIC data where generally there should be fewer features of close m/z and therefore a higher resolution is required to resolve all metabolite features in the lipidomics method. This is consistent with literature where it has been shown that 100,000 mass resolution improves resolution of all lipid classes compared to 70,000 mass resolution, with the same being true for 280,000 compared to 100,000 (Bielow et al., 2017). The percentage of consistent features increases as the resolution increases. The mean percentage of features of RSD \leq 30% seen is lowest at 35,000 resolution, the mean is 1.9% greater at 17,500 resolution but the 35,000 resolution has a larger standard deviation. The 140,000 resolution triplicates have an average of 72.4%. 70,000 normally produces the highest number of features however using 140,000 may increase the proportion of consistent features. As discussed earlier this is not surprising as higher mass resolution analyses should result in higher mass accuracy measurements, however although the data consistency is greater the overall number of features is lower and this is indicative of the decreased scan rate at higher mass resolutions resulting in the missing of some features of narrower chromatographic peak widths. These data show 70,000 is the resolution required to capture the greatest complexity of the data for HILIC and RP but not for lipidomics, this is consistent with previous work carried out assessment of Orbitrap mass resolution (Najdekr et al., 2016). This demonstrates the value of implementing an Orbitrap based instrument as the number of features as well as the reproducibility of features is greater at $\geq 70,000$ mass resolution and this above the level of resolution possible on most TOF based instruments.

Table 40: Mean total features and mean percentage of total features with RSD <= 30% grouped by resolution.

Resolution	Mean Total Features \pm SD	Mean Percentage of features of RSD <= 30% \pm SD
140,000	12,655 \pm 4948	72.4 \pm 10.7
70,000	14,490 \pm 6527	68.3 \pm 14.4
35,000	11,217 \pm 4764	57.7 \pm 15.3
17,500	7,216 \pm 3415	59.6 \pm 12.6

Data from Table 38 was grouped by sample type and the mean values were calculated and are displayed in Table 41. This data shows no significant difference between plasma and tissue in terms of number of features detected. Urine has a much larger average of 16,411 features detected although the standard deviations with the other samples still overlap. No lipidomics data were collected for urine whilst it was for plasma and tissue and the lipidomics data produces significantly less features than HILIC or RP. This is likely to be the main contributor to the large differences seen however urine is generally regarded to be a highly complex sample type (Bouatra et al., 2013). The tissue data exhibits a lower percentage of features of RSD <= 30% and had a larger standard deviation than seen with the plasma and urine data. This could be an indicator that tissue sample preparation was less uniform than the plasma and urine samples.

Table 41: Mean total features and mean percentage of total features with RSD <= 30% grouped by sample type.

Sample Type	Mean Total Features \pm SD	Mean Percentage of features of RSD <= 30% \pm SD
Plasma	9,986 \pm 5,153	67.7 \pm 11.6
Urine	16,411 \pm 4,372	65.5 \pm 12.1
Tissue	9,071 \pm 4,556	59.8 \pm 18.0

Data from Table 38 was grouped by ion mode and the mean values were calculated and are displayed in Table 42. Positive ion mode on average generates more features than negative ion mode and so negative ion mode maybe more conducive to a DIA method. Negative ion mode also shows slightly greater consistency in its detection of peaks with on average 5.4 % more features of RSD <= 30%.

Table 42: Mean total features and mean percentage of total features with RSD <= 30% grouped by ion mode.

Ion Mode	Mean Total Features \pmSD	Mean Percentage of features of RSD <= 30% \pmSD
Positive	12,165 \pm 5,502	61.6 \pm 14.7
Negative	10,720 \pm 5,684	67.0 \pm 13.8

Considering the presented data it was determined that a mass resolution of 70,000 should be utilised when applying a HILIC or RP method. Whereas, when applying the lipidomics method using 140,000 would be more appropriate. This was determined by observing which resolution generated the greatest number of reproducible features after IPO peak picking parameter optimisation and subsequent XCMS processing using the parameters resulting from the optimisation. For RP and HILIC most of the sample type/assay/ion mode combinations saw decreasing feature numbers when the mass resolution was changed from 70,000 down to 35,000 as well as a decrease in reproducibility, whilst increasing the resolution to 140,000 resulted in less features, although higher reproducibility was recorded. For lipidomics the 140,000 mass resolution was more effective and decreasing to 70,000 resulted in fewer features detected as well as lower reproducibility. Although the largest number of features is not necessarily the best method for determining the quality of the peak picking it should ensure that biologically important information is not missed which is part of the goal of this work.

4.2.3 Theoretical DIA Window Complexity Assessment

After all the data had been processed applying XCMS and filtering based on RSD, the feature lists were input into the data complexity spreadsheet tool (2.2.7). This generates 5 separate visual representations of the complexity of the DIA windows within the sample presuming RT windows of 1 second and m/z window sizes of 5, 25, 50, 100 and 200. A table that quantifies and summarises these visual representations is also produced. This provides an idea of which size DIA window may be suitable to provide informative data for each particular UHPLC assay, sample type, mass resolution and ion mode combination. The RSD filtered lists were utilised so as to provide the most accurate representation of the true complexity of the real metabolites by eliminating any peaks that are not real such as noise peaks.

The RP/urine/70,000/positive ion mode data are displayed as it had the most features with a RSD <= 30% (Figure 61, Figure 62, Figure 63, Figure 64). The lipids/tissue/70,000/negative ion mode is also

displayed as in contrast it was one of the least complex set of results (Figure 65, Figure 66, Figure 67, Figure 68). The remaining complexity assessments can be found in the electronic Appendix (4.2/4.2.3). The visual representations demonstrate the importance of knowing the complexity of your sample. The RP/urine/70,000/positive triplicate is the most complex of all the triplicates processed. If utilising a 5 m/z window as in Figure 61 you can see that the sample is complex however there are a few orange or red cells visible indicating that the vast majority of the cells are of relatively low complexity. This indicates that a 5 m/z window may produce informative DIA MS² data of low complexity. However, once the window size is increased to 25 m/z (Figure 62) then there are many orange and red cells visible indicating that the number of features falling into these DIA windows is very high and they would be very likely to produce uninformative data. The high complexity of the windows is obviously exacerbated as the window size increases from 50 to 100 m/z (Figure 63, Figure 64) and these window sizes would not be appropriate for DIA for this triplicate. This is further backed up by Table 43, with a 25 m/z window size 0.86% of windows will be orange or red (defined as high complexity, 11 or more features in the window) this equates to 309 windows of high complexity compared to 0 windows of high complexity at 5 m/z .

Table 43: The percentage of total windows containing different number of features provided by the complexity spreadsheet template for RP, urine, 70,000, positive ion mode.

		Number of Features					
		0 (White)	1-2 (Blue)	3-6 (Green)	7-10 (Yellow)	11-20 (Orange)	>=21 (Red)
Window Size (m/z)	5	93.39	5.50	1.06	0.05	0.00	0.00
	25	84.18	8.94	4.54	1.48	0.77	0.09
	50	79.44	10.03	5.18	2.36	2.30	0.69
	100	73.76	11.92	5.71	2.53	3.25	2.83
	200	64.74	14.07	7.93	3.63	3.87	5.77

In comparison the low complexity triplicate shows very few high complexity windows even with relatively large m/z window sizes of 50 and 100. It is possible that these window sizes would produce informative MS² data with only 0.13% and 0.54% of windows being of high complexity at these window sizes (Table 44). It is best to use as large a window size as possible that will still produce informative data as this will improve the full scan acquisition rate and potentially facilitate the use of a higher mass resolution, both of which will improve overall data quality. This again shows the importance of this

complexity assessment as it will ensure that DIA experimental design for analysis of each UHPLC, sample, mass resolution and ion mode combination is appropriate.

Table 44: The percentage of total windows containing different number of features provided by the complexity spreadsheet template lipidomics, tissue, 70,000, negative ion mode.

		Number of Features					
		0 (White)	1-2 (Blue)	3-6 (Green)	7-10 (Yellow)	11-20 (Orange)	>=21 (Red)
Window Size (m/z)	5	98.83	1.06	0.11	0.00	0.00	0.00
	25	96.06	2.93	0.89	0.10	0.01	0.00
	50	93.95	4.10	1.48	0.34	0.12	0.01
	100	91.31	5.38	1.97	0.80	0.48	0.06
	200	86.36	7.46	3.34	1.34	1.24	0.27

It is important to consider that these complexity assessments are conservative ones due to the fact the RT window is set at 1 second. It is unlikely a cycle time this large would be employed during a real analysis and therefore the true window complexity is likely to be less than shown in the data below. On the other hand, the feature lists utilised were already filtered by RSD and so it could be said the true complexity of data is not represented because background chemical noise will be present in the raw data which is not being considered. However, the majority of the features filtered out are likely to be noisy and or low intensity. Noisy features are not biologically interesting while more importantly low intensity features should have a relatively small impact on a complex DIA window. The features left in the assessment should predominantly be metabolites of interest or other derivative features of them.

If the complexity of the windows looks to be too high there could be further ways to try and alleviate the issue. Modification and improvement of the LC method to provide greater chromatographic separation and resolution could provide major advantages for DIA based methods through reduction of metabolite co-elution. This could be achieved through column modification in a number of ways including a reduction in particle size or an increase in length for example. A more complex but potentially more useful modification would be to employ 2D-LC, this could provide 2-fold benefits, firstly an extra chromatographic trace to aid in mathematical deconvolution of MS² spectra, secondly by further separating the sample and reducing the complexity of the sample being detected at any one moment. This has been done in proteomics (Rodríguez-Suárez et al., 2014) but would be a

challenge to employ in metabolomics. A new deconvolution algorithm would be required as well as having to deal with the challenges typically associated with 2D-LC. Another method that has been used is in proteomics is to employ overlapping windows (Ludwig et al., 2018), this could be used to aid the deconvolution process but would result in a decrease in the mass range being covered by the windows.

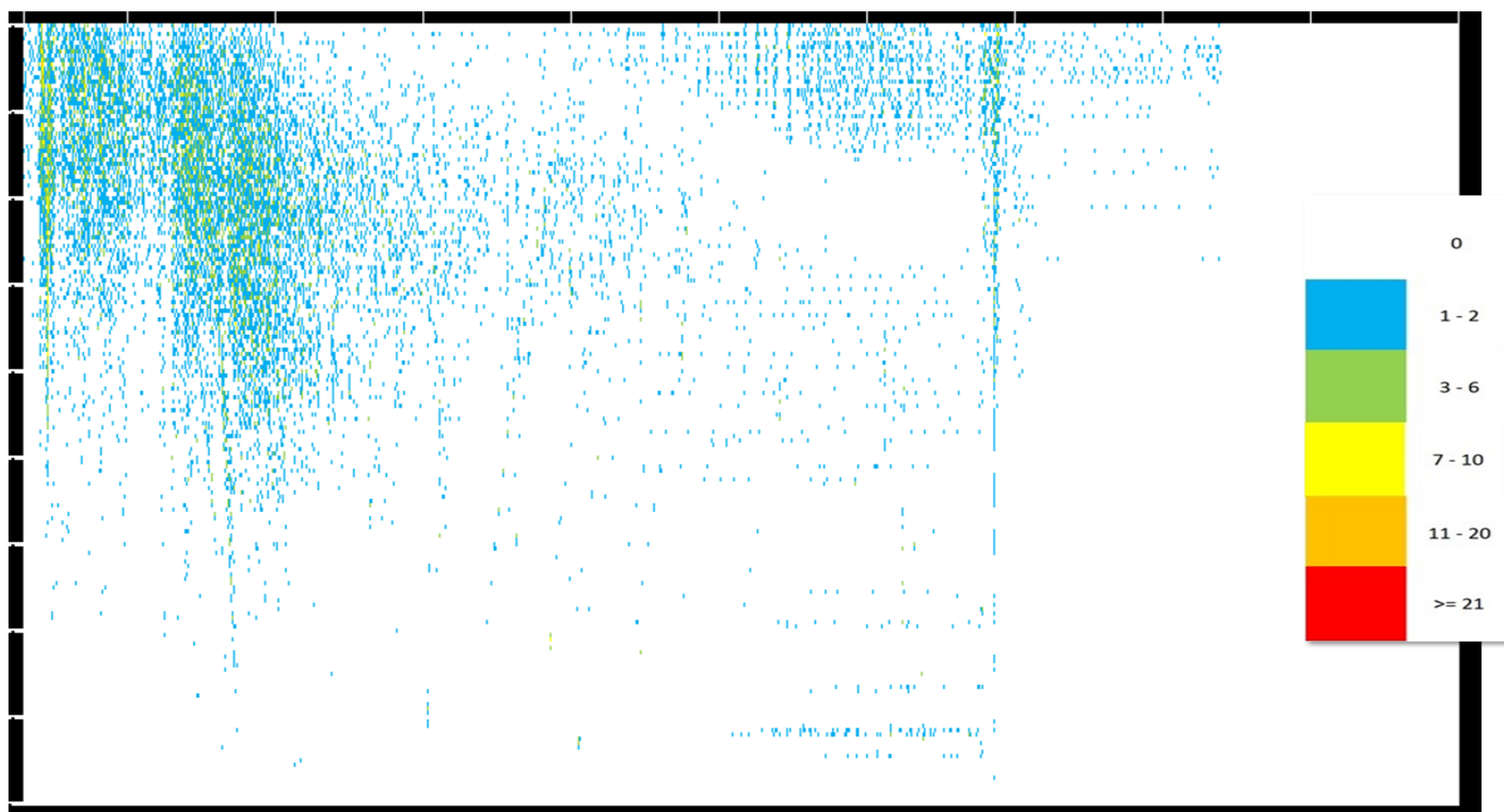


Figure 61: Visual representation of the complexity of theoretical DIA windows for urine/RP/70,000 mass resolution/positive ion mode RT windows = 1 second, DIA windows = 5 seconds. x axis represents RT with each column representing a 1 second RT window from 30 – 1000 from the left to the right. The first column of cells on the left indicated by the first white line along the x axis indicates an RT ≥ 30 and < 31 . The next white line indicates an RT ≥ 100 and < 101 . Each white line thereafter indicates another 100 seconds up to RT ≥ 999 and < 1000 . The y axis represents m/z with each row representing a 5 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 105 . Each white line after represents a 100 m/z increase. Cells are conditionally formatted as described by the figure legend, numbers represent how many features are in each window.

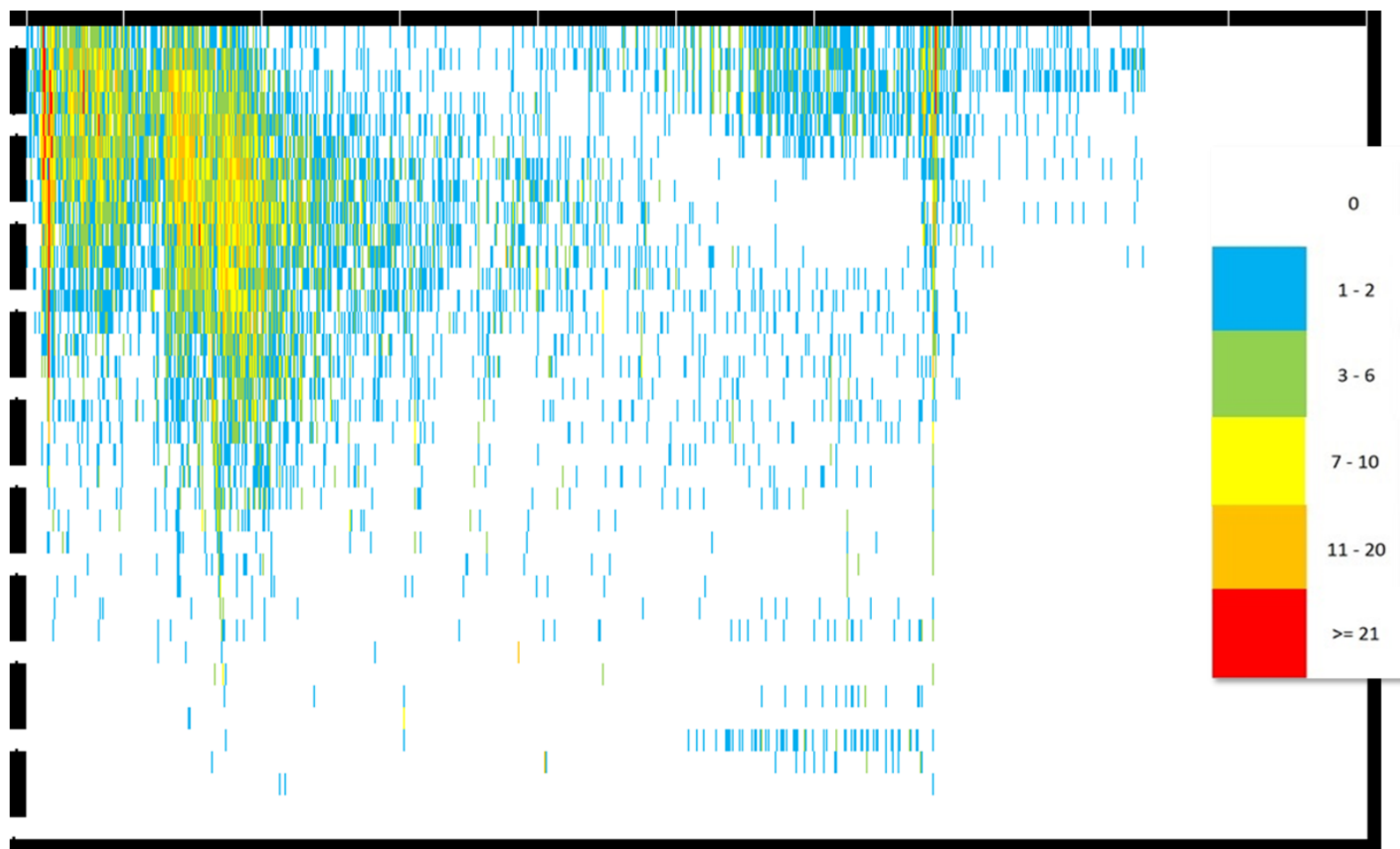


Figure 62: Visual representation of the complexity of theoretical DIA windows for urine/RP/70,000 mass resolution/positive ion mode, RT windows = 1 second, DIA windows = 25 m/z . x axis is as described in Figure 61. y axis represents m/z with each row representing a 25 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 125 . Each white line after represents a 100 m/z increase. Conditionally formatted as described in Figure 61.

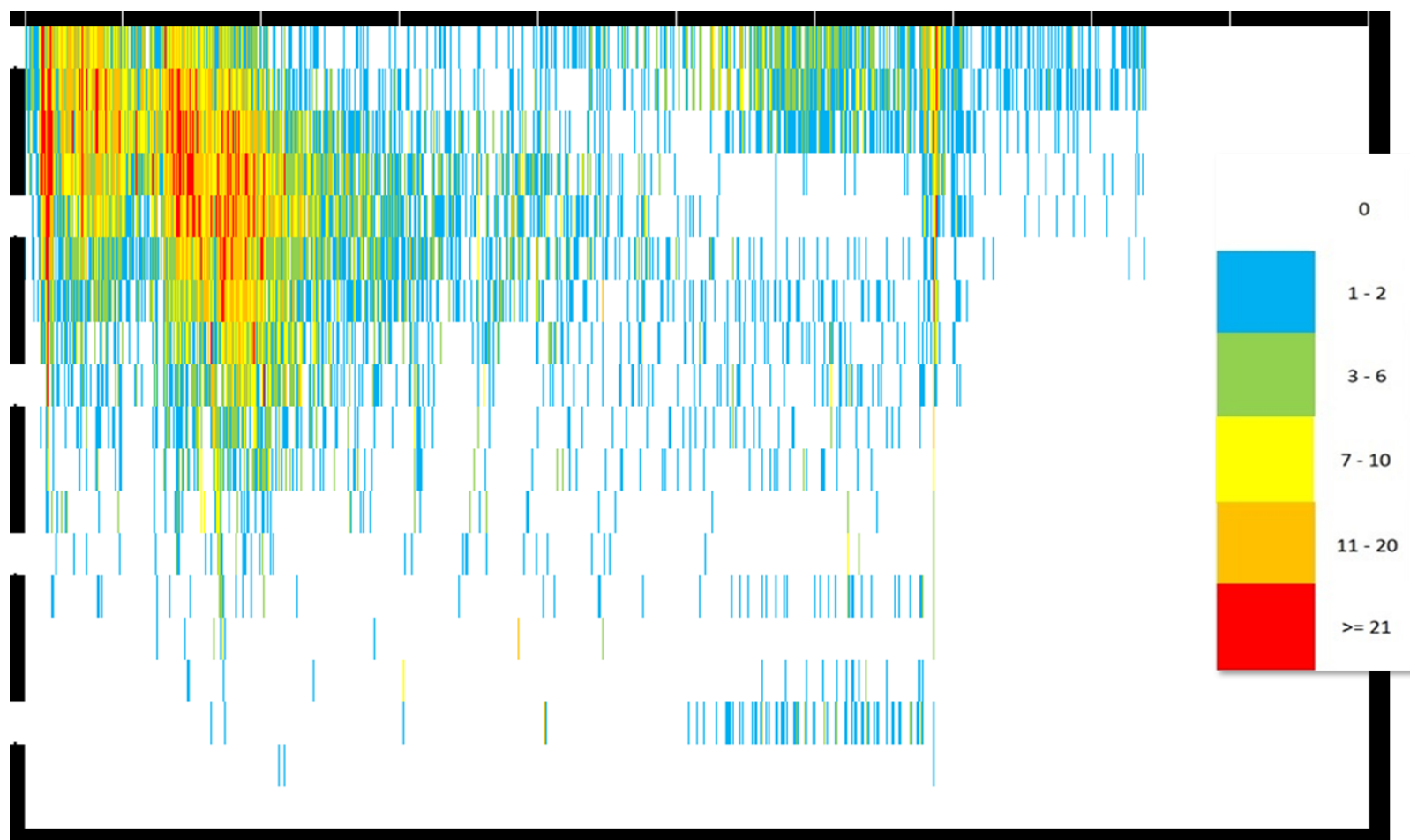


Figure 63: Visual representation of the complexity of theoretical DIA windows for urine/RP/70,000 mass resolution/positive ion mode, RT windows = 1 second, DIA windows = 50 m/z . x axis is as described in Figure 61. y axis represents m/z with each row representing a 50 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 150 . Each white line after represents a 200 m/z increase. Conditionally formatted as described in Figure 61.

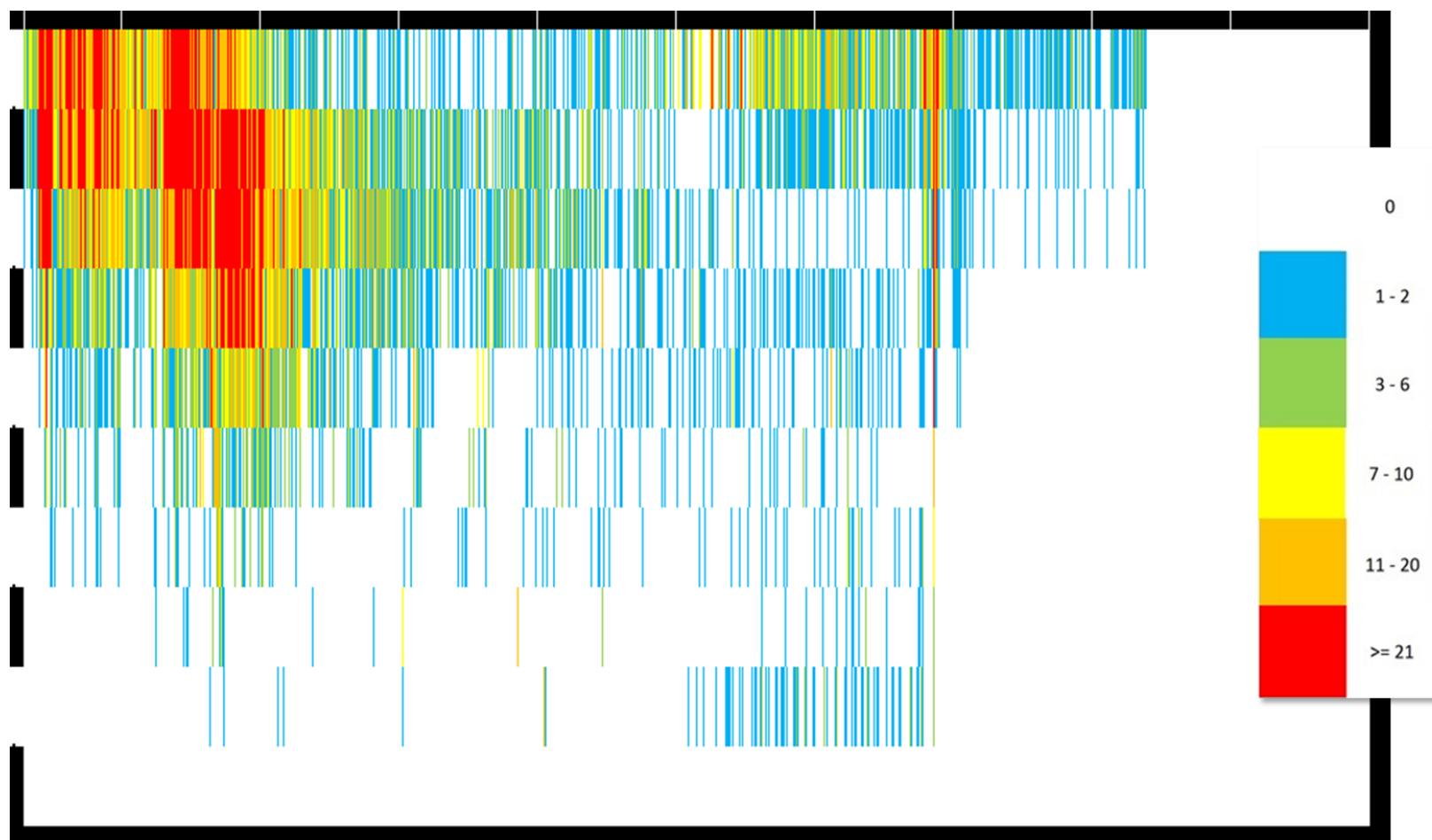


Figure 64: Visual representation of the complexity of theoretical DIA windows for urine/RP/70,000 mass resolution/positive ion mode, RT windows = 1 second, DIA windows = 100 m/z . x axis is as described in Figure 61. y axis represents m/z with each row representing a 100 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 200 . Each white line after represents a 200 m/z increase. Conditionally formatted as described in Figure 61.

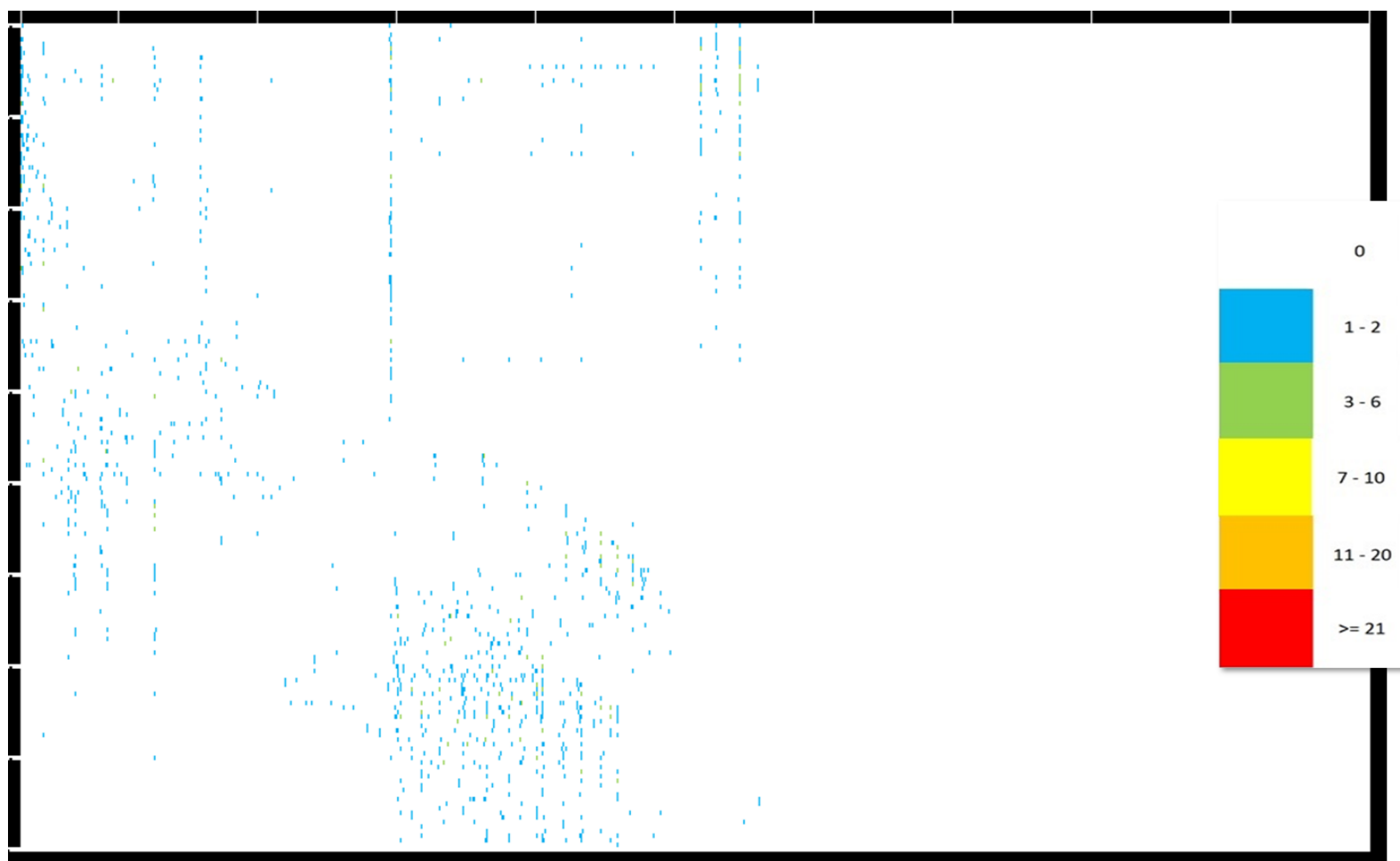


Figure 65: Visual representation of the complexity of theoretical DIA windows for tissue/lipidomics/70,000 mass resolution/negative ion mode, RT windows = 1 second, DIA windows = 5 m/z . x axis is as described in Figure 61. y axis represents m/z with each row representing a 5 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 105 . Each white line after represents a 100 m/z increase. Conditionally formatted as described in Figure 61.

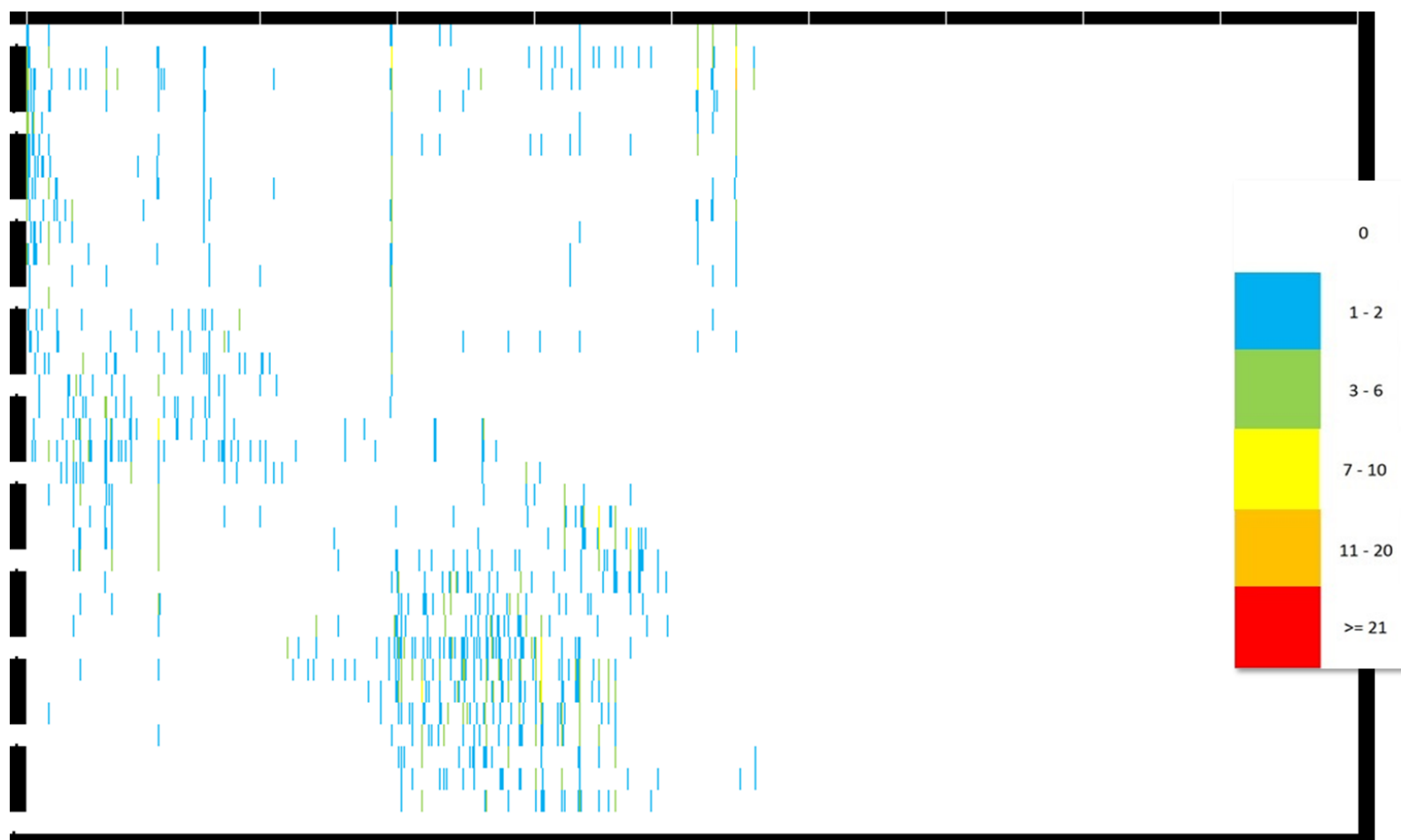


Figure 66: Visual representation of the complexity of theoretical DIA windows for tissue/lipidomics/70,000 mass resolution/negative ion mode, RT windows = 1 second, DIA windows = 25 m/z . x axis is as described in Figure 61. y axis represents m/z with each row representing a 25 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 125 . Each white line after represents a 100 m/z increase. Conditionally formatted as described in Figure 61.

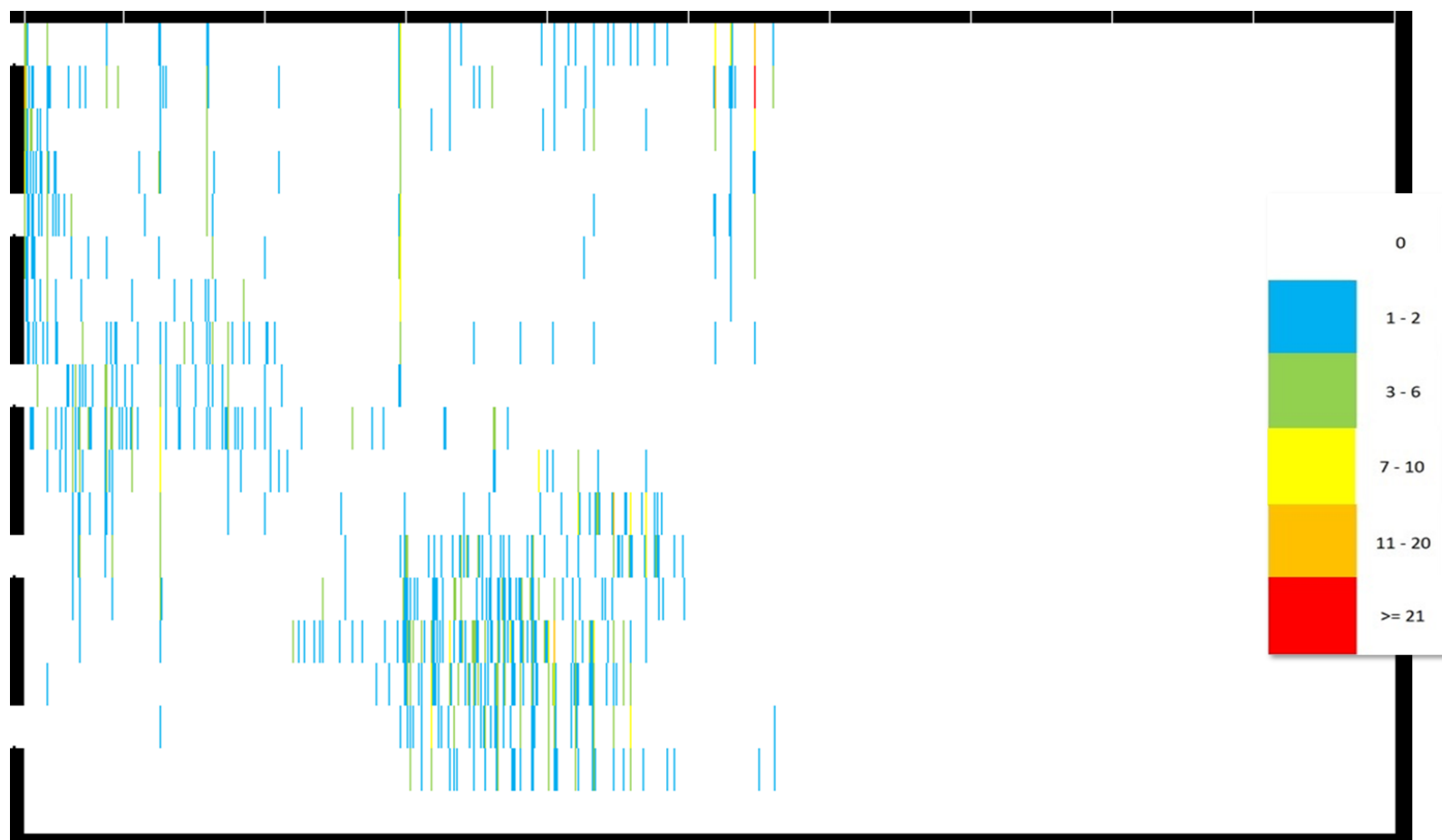


Figure 67: Visual representation of the complexity of theoretical DIA windows for tissue/lipidomics/70,000 mass resolution/negative ion mode with, RT windows = 1 second, DIA windows = 50 m/z . x axis is as described in Figure 61. y axis represents m/z with each row representing a 50 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 150 . Each white line after represents a 200 m/z increase. Conditionally formatted as described in Figure 61.

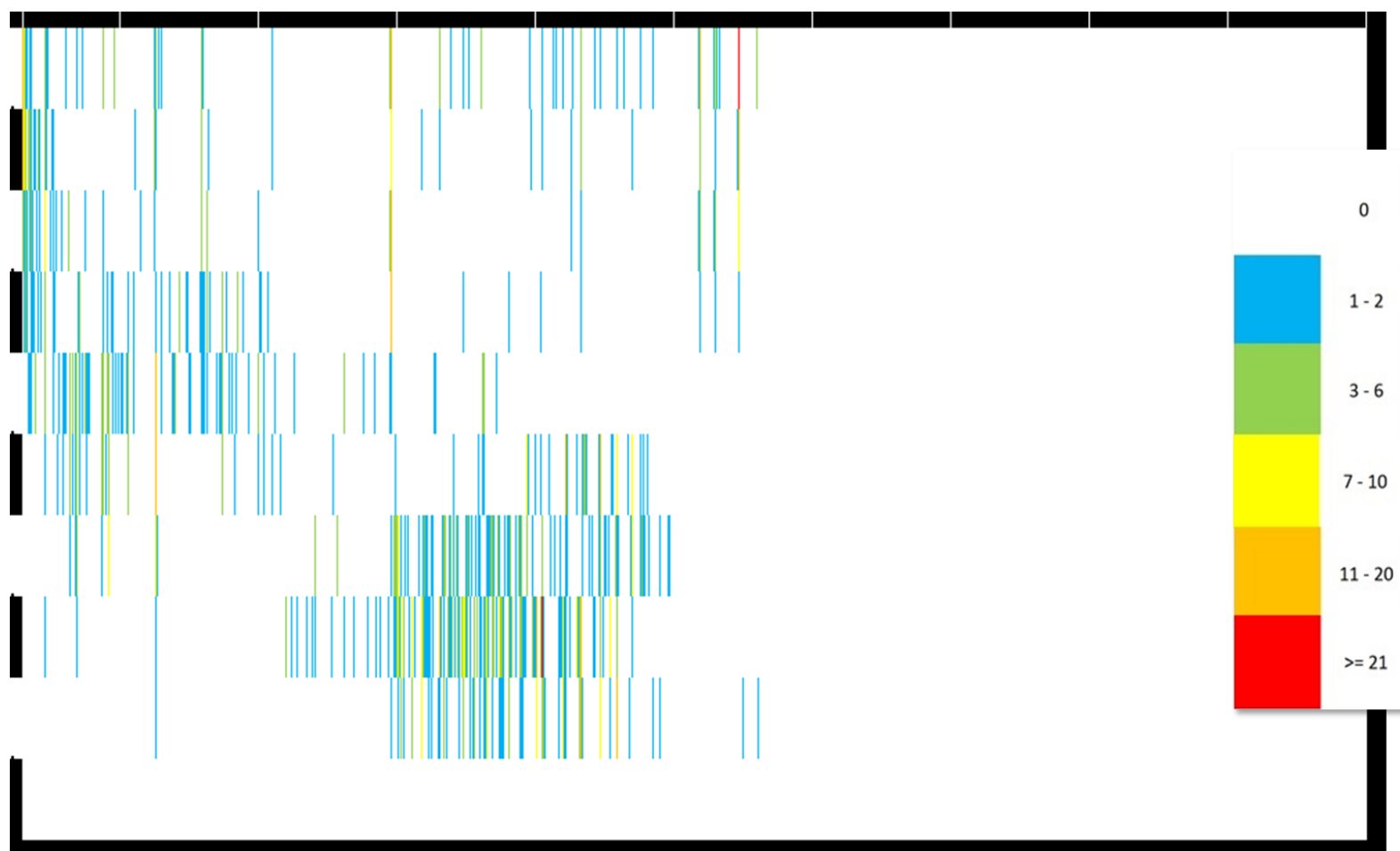


Figure 68: Visual representation of the complexity of theoretical DIA windows for tissue/lipidomics/70,000 mass resolution/negative ion mode, RT windows = 1 second, DIA windows = 100 m/z . x axis is as described in Figure 61. y axis represents m/z with each row representing a 100 m/z window from 100-1000 from top to bottom. The first white line from the top indicates an $m/z \geq 100$ and < 200 . Each white line after represents a 200 m/z increase. Conditionally formatted as described in Figure 61.

4.2.4 Peak Width Assessments and Scan Rate Estimations

Following investigation of the full scan mass resolution and data complexity, the next step to be considered in the development of a DIA method is to understand the width of the chromatographic peaks present in the full scan data for each different sample type and resolution combination. This is important as when determining the MS² method, the number of fragmentation windows or events between each MS¹ full scan must be selected by the operator. If this number is too large the resulting cycle time is too slow and the result is that features that should have been detected in the MS¹ data are not detected because they have too low a number of data points across the peak. Whether these features are missed or not is dependent on two factors, the cycle time of the method and the chromatographic width of the peaks. For a peak to be accurately detected and quantified a minimum of 7 MS¹ full scan data points (and ideally 11 or more data points) are required across the peak. Therefore, if the peak has a chromatographic width of 7 seconds then the cycle time of the method should be at most 1 second. To ensure the MS² methods to be tested later were designed in a way which ensures this is still the case the peak widths for each method were characterised and the scan rate of the Orbitrap QE Plus estimated at different full scan and MS² mass resolutions.

To best assess the peak widths seen in the data the 17,500 mass resolution full scan files were used. As they are the lowest resolution of all the triplicates, this means the cycle time is the fastest and the number of data points across each peak is greatest. This ensures that the peak widths should be more accurate and reliable than the peak widths reported if looking at a higher full scan resolution with longer intervals between each MS¹ data point.

The results for the peak width data (Table 45) show that the lipidomics method that was utilised has the widest peaks with a median peak width of 8.42 seconds compared to 6.64 for the HILIC method and 5.20 for the RP method. This data indicates that a lipidomics method would be most conducive to a DIA method on an Orbitrap system as the wider the peaks the more MS² scan events can be implemented between each full scan without negatively impacting the quality of the MS¹ data.

Table 45: The distribution of peak widths seen across all sample types for the three different UHPLC-MS assays.

Chromatography Method	Lower Limit	Lower Quartile	Median	Upper Quartile	Upper Limit
HILIC	0.23	3.96	6.64	9.49	100.55
Lipids	0.39	5.33	8.42	11.58	63.88
RP	0.23	3.13	5.20	7.49	90.38

The scan rate estimation of an Orbitrap mass analyser at different MS^1 and MS^2 mass resolution combinations with varying numbers of MS^2 events was carried out to determine how many scan events could be carried out and at what mass resolution for each assay now the peak widths had been summarised. The full scan data collected earlier had indicated that at least a 70,000 mass resolution was required for detection of the highest number of reproducible features with a significant drop off seen at 35,000 or below. Therefore, estimations with full scan mass resolution lower than 70,000 have been removed. Secondly, the data in Table 45 indicated that for the method with the highest peak widths the lower quartile was 5.33 seconds. As the goal for planning the next experiments was to achieve 7 data points across the majority of peaks anything with an estimated time for 7 data points over 5.33 seconds was not considered. Table 46 shows the MS^1 and MS^2 resolution combinations and number of MS^2 events that could be utilised for an optimal MS^2 method when considering the restrictions derived from the data.

Table 46: The estimated time taken for 7 full scan data points to be collected at different MS¹ and MS² resolution combinations on a Q Exactive Plus mass spectrometer.

MS ¹ Mass Resolution	MS ² Mass Resolution	Number of Windows	Time taken for 7 data points (secs)
70,000	17,500	1	2.92
70,000	17,500	2	3.50
70,000	35,000	1	3.50
70,000	17,500	3	4.08
70,000	17,500	4	4.67
70,000	35,000	2	4.67
70,000	17,500	5	5.25
140,000	17,500	1	5.25

The results show that the scan rate of the Q Exactive Plus is a limiting factor in designing an effective DIA method for the instrument. The largest number of windows compatible with capturing the majority of peaks with 7 MS¹ data points (a time for 7 data points below the LQ for that assay) for the RP assay is just 1 at 17,500 mass resolution. With the HILIC method 2 MS² events can be afforded at 17,500 or 1 at 35,000. Whilst with the lipidomics method up to 5 MS² events at 17,500 could be performed or 2 MS² events at 35,000. If the aim is to get fragmentation information for all features within the dataset and in the best case scenario the greatest number of windows that could be utilised is 5. Then assuming the mass range investigated is at least 500 *m/z* in size the windows utilised already have to be at least 100 *m/z* wide. As the complexity data displayed earlier demonstrates, this is unlikely to generate informative MS² data unless the sample type is of low complexity (Table 44). The prospects for DIA on the Q Exactive Plus as a result do not look good when applying current short chromatographic analysis times with narrow peak widths. The scan rate of Orbitrap mass analysers is limited due to the inherent method of data recording, higher mass resolutions require a longer ion path to be traversed which requires a longer time. This can and has been improved on recent Orbitrap mass analysers, by decreasing the size of the Orbitrap mass analyser, slightly modifying its geometry and increases the voltages applied. Despite these improvements they are not to a level which would suddenly make DIA analyses significantly more suitable. A TOF instrument is more suitable for DIA work as they are capable of significantly higher scan rates. The majority of the relatively little previous work that has been carried out for DIA based metabolomics has been carried out on TOF based systems (Tsugawa et al., 2017; Chen et al., 2017; Bonner and Hopfgartner, 2018).

Utilising this information optimal MS² methods for improved annotation of metabolites in biological studies were planned. DIA methods were still considered despite the shortcomings described. To alleviate the issue of large windows being required to cover the entire mass range experiments were designed with varying window widths the majority of which only cover a portion of the mass range.

4.2.5 DIA Methods Designed

The data collected and presented in this chapter was then used to develop 2 sets of methods, 1 set for HILIC plasma and one for lipidomics plasma in both ion modes. Only two assay and sample type combinations were selected to reduce the number of analyses required in the next part of the study and increase the number of different MS² acquisition strategies that could be tested. Considering the challenge presented and the limitations in developing a DIA method on the Q Exactive Plus the highest complexity assay and sample type combinations were not selected. HILIC/plasma was selected due to its medium level of complexity, whilst lipidomics/plasma was selected to provide a lower complexity combination for assessment. Within each set a number of different DIA methods for implementation on the Q Exactive Plus were developed. For each method the data from sections 4.2.2, 4.2.3, and 4.2.4 was considered. MS¹ resolutions were determined by the data presented in 4.2.2, the MS² resolution was set at the lowest possible (17,500) due to the scan rate being the limiting factor. Considering these pieces of information the number of MS² scans that could be implemented per cycle whilst still maintaining 7 data points across the majority of peaks was calculated and implemented. The quality of data from different sized DIA windows was to be assessed and so DIA methods had varying window sizes across varying overall mass ranges. These would then be compared to a variety of different DDA based method to compare the merits of these different MS² acquisition strategies.

4.2.5.1 HILIC

The lower quartile of peak widths seen with the HILIC assay was 3.96 seconds (Table 45 Table 46). It had also been shown that 70,000 mass resolution should be applied for the MS¹ data as a minimum for HILIC analyses (Table 38). This meant that the highest number of MS² events (17,500 mass resolution) that could be carried out whilst still achieving 7 data points across the majority of peaks for the assay was 2. Therefore, the DIA methods employed used 2 DIA windows (3.5 seconds for 7 data points), whilst some with 4 windows (4.67 seconds for 7 data points) were also utilised to assess the impact of sacrificing MS¹ scan rate for more MS² information. Using these different number of windows will allow comparison between the number of features identified as well as comparison of MS¹ data quality and number of features detected initially. Using this small number of windows means for most DIA methods the entire MS¹ mass range being measured was not covered by the windows. This meant an *m/z* region to focus on required selection. The complexity assessments (electronic

Appendix 4.2/4.2.3) carried out showed that the region above an m/z of 200 should be a highly complex m/z region and therefore was selected to help provide a rigorous test. Windows of 10, 25, 50, 100 and 232.5 m/z were implemented. The full details of MS² type, window size, number and range are found in Table 16.

4.2.5.2 Lipidomics

The same strategy as discussed for HILIC was applied for lipidomics. The lower quartile of peak widths seen with the lipidomics assay was 5.33 seconds. It had also been shown that 140,000 mass resolution would detect the greatest number of reliable features for the lipidomics assay (Table 38). However due to inherent scan rate limitations some methods implemented an MS¹ mass resolution of 70,000 instead. As with the HILIC method MS² mass resolution was restricted to the lowest level (17,500) due to inherent limiting factor of the instrument scan rate. However, the wider peak widths of the lipidomics method facilitate the use of more windows whilst still achieving an appropriate MS¹ scan rate. DIA methods with 3 windows were employed at varying sizes and with both 140,000 (4.08 seconds for 7 data points) and 70,000 (6.42 seconds for 7 data points) MS¹ mass resolution (Table 45 Table 46). DIA methods with 6 windows were also employed however only with a full scan resolution of 70,000 (5.83 seconds for 7 data points) due to scan rate limitations. An 8 window method covering the entire mass range was also implemented, full details of all methods developed can be found in Table 17. The complexity assessment (electronic Appendix 4.2/4.2.3) was used to select a region to focus the DIA windows upon. 750 m/z and above was the region chosen due to its high complexity. DIA windows of 10, 25, 50, 100 and 175 m/z were tested.

4.3 Conclusions

Firstly, in this chapter XCMS optimisation was carried out using the R package IPO, this was an important step to take to ensure peak picking was reliable and accurate. This was performed on the data collected from the different sample types (plasma, urine, tissue), assays (HILIC, RP, lipidomics), ion modes (positive, negative) and mass resolutions (17,500, 35,000, 70,000, 140,000) tested. The impact of the method changes to the resulting optimised parameters was investigated and it was determined that assay and mass resolution were key drivers of some of the optimal processing parameters (ppm, min peak width, max peak width). This is logical as their modification should result in differences to the data characteristics (especially chromatographic peak width) in a broader manner than changing the ion mode or sample type would. As a result, the median results for each assay and mass resolution combination were applied for data processing. The number of reproducibly detected features was then assessed to determine which mass resolution was required to maximise the amount of biological knowledge that could be gained from a study carried out on a Q Exactive Plus. It was

determined that 70,000 was required for HILIC and RP whilst 140,000 was more appropriate for lipidomics analyses. Following determination of the required MS¹ resolution it was then possible to move onto investigation of appropriate DIA methods for future comparison to other MS² acquisition strategies (DDA, iDDA, AIF) on the Q Exactive Plus which will be the focus of the next chapter. The complexity of theoretical DIA windows of varying size (5, 25, 50, 100, 200 *m/z*) was assessed for all datasets collected to provide a guideline as to the overall complexity of the data, what percentage of windows will be of high complexity when different window sizes are employed? It allows elucidation of regions of high complexity/density all of which informed the subsequent DIA method designs. The data showed that with narrower windows of 5 or 25 *m/z* that the DIA windows will be predominantly of low complexity and successful deconvolution would be likely for the majority of windows. Larger window sizes would not be recommended unless the sample type was of very low complexity.

The peak width characteristics as well as scan rate estimations of varying MS¹ mass resolutions, MS² mass resolutions and numbers of MS² events per cycle combinations were then investigated to ensure a minimum of 7 MS¹ data points across the majority of chromatographic peaks in the data. This showed when developing a DIA method on the Q Exactive Plus that the slow scan rate due to the nature of the Orbitrap mass analyser is a major limiting factor. DIA experiments designed needed to fulfil the following criteria:

1. A scan rate fast enough to achieve 7 MS¹ data points across the majority of chromatographic peaks in the method.
2. A DIA window size small enough that MS² spectra are not too complex and thus deconvolution can be successful.
3. A DIA window range that covers the whole mass range typically analysed.

As a result, the DIA methods planned will each be suboptimal in different ways and many will not be capable of providing MS² data for all features in a global unbiased manner as is the goal of untargeted metabolomics. For example, some lipidomics methods were included without their optimal MS¹ resolution of 140,000 due to the scan rate being a limiting factor, and almost all DIA methods covered a small portion of the mass range. Any method covering the whole mass range will have to apply a window size which has been indicated the theoretical DIA window complexity assessments to be far too large for successful deconvolution. An interesting way to overcome this could be to develop an intelligent DIA (iDIA) method. This could comprise a surveying full scan analysis. Peak picking could be performed on-the-fly as is done with the new Orbitrap ID-X software. The density and complexity across the *m/z*-RT axes could also be assessed on the fly. The subsequent analysis will then be carried out with DIA windows varying in size throughout the analysis depending on the mass range and RT.

Resolution could even be increased or decreased at certain time points to increase data quality or the scan speed where appropriate but this is not something which is currently feasible.

Despite the discussed limitations in terms of scan rate the complexity assessments indicate DIA could be effective in terms of the data quality provided and therefore DIA methods were developed to test alongside other possible optimal MS² acquisition strategies and this work will be presented in the chapter 5.

5.0 Comparison of different MS²

acquisition strategies on the Q

Exactive Plus

5.1 Introduction

Multiple types of data are collected when performing untargeted metabolomics studies applying UHPLC-MS. Chromatographic retention time is recorded and is determined by the physicochemical characteristics of the metabolites measured. The measurement of the mass-to-charge ratio (m/z) of metabolites provides data to calculate their molecular formula (Nash and Dunn, 2019). However, whilst these data provide some information to reduce the number of possible metabolites matched more structural information is required for a confident identification and this is typically achieved through collection of MS² data (Nash and Dunn, 2019). Without MS² data or another type of further structural information (for example, 1D or 2D NMR spectroscopic data), a confident identification cannot possibly be assigned and without confident identifications biological information and conclusions cannot be drawn from the data. This is not to say MS² acquisition will automatically result in an identification, an identification will be dependent on the quality of the spectrum acquired and whether there is a good quality reference spectrum available. A strong match to an external library can provide a level 2 identification, or ideally a strong match to a pure standard collected in the same lab with the same analytical method for a level 1 identification (Schymanski et al., 2014). Therefore, the volume and quality of MS² data need to be maximised whilst still maintaining MS¹ data quality too for maximum biological knowledge to be gained from the study. As a result, the strategy employed for acquisition of MS² data is vitally important.

The method of acquisition traditionally applied is data dependent acquisition (DDA) (Nash and Dunn, 2019). This method selects the top “n” highest intensity features from each MS¹ scan and fragments them individually, recording the data in separate “pure” MS² spectra with a narrow isolation window. This method is biased towards ions of high intensity, the majority of features within the dataset have no MS² information acquired for them and so DDA is not appropriate for the goal of profiling all metabolites within a sample (Mullard et al., 2014) which is the ultimate goal of untargeted metabolomics. The advantage of DDA is that the spectra collected should be pure and can be directly compared to an MS² spectral library. Data independent acquisition (DIA) methods offer the opportunity to overcome the flaw in DDA and acquire fragmentation data for all metabolites (and all features of all metabolites) regardless of intensity (Wang et al., 2019b). DIA involves the collection of an MS¹ scan followed by “n” fragmentation windows of a user defined m/z width potentially providing complete coverage of MS² data for all features if the windows cover the entire mass range of interest. The limitation of this method is that DIA MS² result in spectra made up of multiple precursor ions and thus before comparison to an MS² library the spectra must be deconvoluted beforehand. There are a number of different factors that require balancing if a DIA method for MS² data acquisition is going to be effective as outlined in the previous chapter; these include the sample complexity, the DIA window

characteristics (m/z width, number of windows, m/z range of windows), the scan rate and the mass resolution employed. The data presented so far have clearly demonstrated the limitations in applying Orbitrap mass analysers to DIA methods due to their relatively slow scan rate. Despite this DIA methods were developed for two different chromatographic assays (HILIC, Lipidomics) and run in both positive and negative ion mode to determine if DIA can provide any advantages over other MS² acquisition strategies. The methods applied had varying size fragmentation windows (10, 25, 50, 100 m/z), covering varying portions of the total mass range. This would allow assessment of different sized windows and the subsequent quality of the MS² data as well as the quality of MS² deconvolution when these different window sizes are applied. Larger window sizes will provide the advantage that MS² data will be collected for more features, however the deconvolution process may be less effective. Narrow window sizes will mean fewer features will have MS² data collected but the deconvolution process should be more effective. As the DIA window sizes already mentioned did not cover the entire mass range required, other methods were also applied covering the total mass range with wider windows than those specified above in both DIA and All Ion Fragmentation (AIF) styles. AIF provides an advantage in that the MS¹ scan rate is faster than with the DIA and DDA methods and will also provide fragmentation data for all features. However, with fragmentation windows hundreds of m/z values wide the deconvolution process is certainly in doubt. Considering the limitations of DIA for the chromatographic characteristics of the methods and Orbitrap analysers other MS² acquisition methods were tested too. With the goal to assess which style of MS² acquisition is most appropriate for generation of the greatest volume of biologically useful data. Therefore, DDA methods were implemented too as well as different forms of intelligent-DDA (iDDA). These included “traditional” DDA methods where a simple top n strategy is employed. The traditional method was then adapted in different ways to form the different iDDA methods.

These adaptations included the employment of inclusion lists, exclusion lists, progressive exclusion lists and segmentation of the total mass range covered. These can each be used to try to increase the amount of informative MS² data collected. Exclusion/inclusion lists can be utilised to improve MS² coverage of metabolite features. Exclusion lists are employed to ensure components of the sample matrix such as solvent peaks do not have MS² data collected for them as this is a waste of the limited MS² acquisition time available which should be spent on acquisition of biological sample components, the lists are generated after full scan analysis of a process blank. m/z values in the list are ignored and these can also be progressively updated over repeated injections of the same sample to include m/z values already fragmented (Neumann et al., 2013), this should ensure increased coverage by avoiding repetitive acquisition of high intensity sample components. Inclusion lists can also be applied, to try to ensure biological components of the sample have MS² data acquired, the list is generated after full

scan analysis of the sample. m/z values in the list will be fragmented if detected regardless of intensity if present in an MS^1 scan. Segmentation of the total mass range, referred to as the ‘Mullard method’ (Mullard et al., 2014) from herein can help to improve coverage of MS^2 spectra for metabolite features. Segmentation of the m/z range considered for MS^2 acquisition means lower intensity features are more likely to be within the “top n ” in any MS^1 scan collected and thus more low intensity metabolite features will have MS^2 collected. These methods can help to increase the MS^2 coverage of biologically important features using the reliable DDA method for which coverage is the main limiting factor.

In this chapter data will be presented comparing which MS^2 method provides the most useful biological information and the merits of these different method types for MS^2 data acquisition (traditional “top n ” DDA, DDA w/ inclusion list, DDA w/ exclusion list, ‘Mullard method’ w/ inclusion/exclusion lists, DIA of varying window m/z widths and total m/z window range and AIF). The pros and cons of each method will be assessed with recommendations of how they might be applied for different applications on a Q Exactive Plus MS. This would then be directly applicable to some other Orbitrap instruments with very similar scan rates too as well as providing a useful guideline for other more dissimilar Orbitrap instruments.

5.2 Results and Discussion

Throughout this results section, the HILIC/positive ion mode dataset will be focused upon. The trends seen in this dataset are generally repeated in the other three datasets. The remaining data for HILIC/negative ion mode, Lipidomics/positive ion mode and Lipidomics/negative ion mode can be found in the Appendix (9.3). Each method shall be referred to by a code such as DIA_4_10 for full information on the method details refer to Table 16. DIA_4_10, means a DIA method with 4 windows of 10 m/z width.

5.2.1 Number of Features Detected

The number of features (m/z -RT pairs) detected across each method was first assessed as an indicator of MS^1 data quality. As the different methods employed had different MS^1 scan rates this was important to assess to determine if any of the methods showed an insufficient scan rate which would be clear from a reduction in the number of features detected. The number of features detected is relatively consistent despite the different MS^2 strategies employed (Figure 69). The relatively small differences that can be seen between the different method types can be explained through looking at the number of MS^2 events. The AIF methods (AIF_430, AIF_930) have one MS^2 scan per MS^1 scan with an average of 13,097 features across the two methods employed. This is compared to the DDA methods (Tra, Tra_exc, Tra_inc, Tra_p_exc_1, Tra_p_exc_2) and 2 window DIA methods (DIA_2_10,

DIA_2_25, DIA_2_50, DIA_2_100) which each have 2 MS² scans per MS¹ scan, these methods average 12,004 features per file. Whilst the 4 window DIA methods (DIA_4_10, DIA_4_25, DIA_4_50, DIA_4_100, DIA_4_232) average 10,820 features per file. This is logical as the peak width data from the previous chapter had shown the scan rate that would be achieved using four windows with the MS¹ (70,000) and MS² (17,500) mass resolutions utilised would result in less than 75% of the features having 7 MS¹ data points across the peak. This proportion increases as the number of MS² scans is decreased and thus the increase in feature numbers described occurs when applying 2 DIA windows compared to 4.

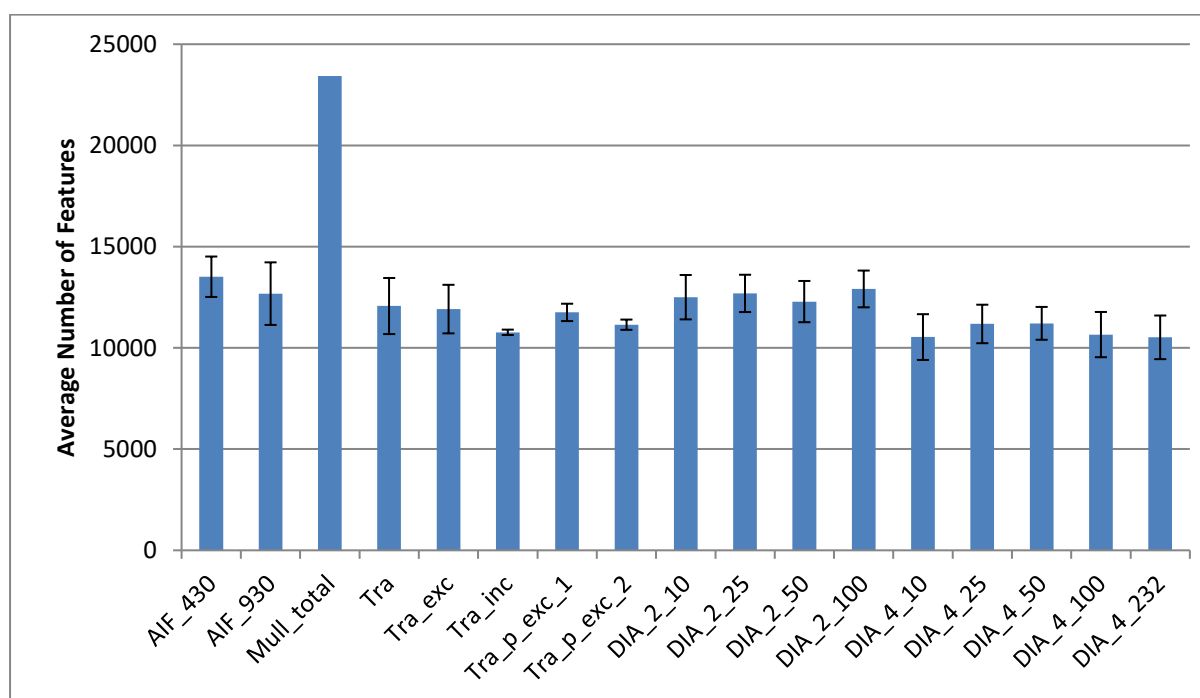


Figure 69: The average number of features detected for each quadruplicate in the HILIC positive ion mode data after MS-DIAL processing \pm SD. Mull_total does not have a SD due to the way it was calculated.

The Mullard method (Mull_total) allowed the detection of the greatest number of features. The average number of features detected in each of the three full scan segments were summed to give the total displayed. Just under 10,000 more features were detected with the Mullard method than the next best method. This was as expected as the mass range had been segmented into multiple and smaller ranges, this has previously been shown to increase the number of detectable features. This was demonstrated in the development of the SIM-stitch method on an FT-ICR instrument (Southam et al., 2007; Kang et al., 2019). It has also been demonstrated on an Orbitrap system with increases in feature numbers of 42 – 102 % reported (Ranninger et al., 2016). This increase can be explained when thinking about the accumulation of charges in the C-trap prior to injection into the Orbitrap. The C-

trap has a limited charge capacity, this is normally set by the user when they determine the AGC target. By setting a smaller m/z range for full scan analysis ions that fell outside of the region of interest that are present in the sample and would normally be transported to the C-trap will be filtered out beforehand by the quadrupole. As a result, the relative concentrations of ions that were within the region of interest are now increased due to the reduction in competition for space in the C-trap. This allows an increase in the intensity of features within the region of interest compared to what would be achieved when employing a wider mass range and thus metabolite features which were previously below the limit of detection can now be detected. These lower intensity ions being detected should also benefit from increased mass accuracy due to decreased space-charge effects in the Orbitrap (Kang et al., 2019). Although there are the benefits mentioned to segmenting the mass range it does also require more time as a separate UHPLC-MS analysis is required for each segment. Therefore, if employing segmentation it is important to consider the practical implications of running these extra analyses. Is the sample limited? How much time is available? How much money is available? Is the extra information important for the desired goal and worth the extra time and money required to achieve it? Another downside to segmentation of the mass range is the loss of relationships and correlations between different adducted, fragmented and multiply charged versions of the same metabolite which are no longer grouped together and thus give rise to an artificial increase in the number of features detected. However, both of these issues could be overcome if this method type was applied on just the pooled QC samples of a biological study. This handful of QC analyses can be used for the studies feature identification whilst the remaining biological samples are analysed with full scan MS^1 data only. This would mean the extra time required for analysis is only a matter of an hour or two and feature relationships for all other biological samples would be preserved. Data for the other assays can be found in the Appendix (9.3.1).

5.2.2 Number of Features with MS^2 Data

The number of features with associated MS^2 data were investigated to see how the different methods affected the coverage of MS^2 data for all features in the dataset (Figure 70).

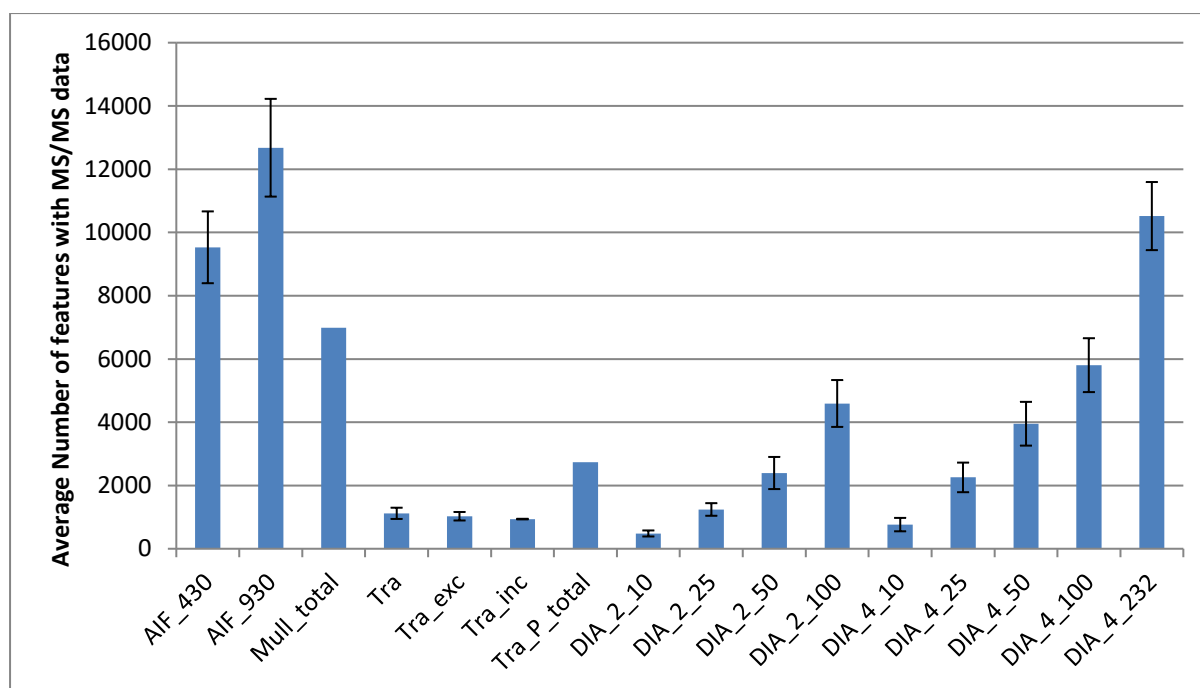


Figure 70: The average number of features that have MS² data associated with them for each MS² strategy after processing in MS-DIAL in the HILIC/positive ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

The method which provided the greatest number of features with associated MS² data is the AIF_930 method (average of 12,679 features with MS² data) (Figure 70). This method covers the entire mass range (70 – 1000 m/z) and collects fragmentation data for all features within this window therefore this is as expected as the majority of other methods tested cover only a portion of the mass range or are data dependent methods. The second greatest is provided by the DIA_4_232 method (average of 10,518). This method also covers the entire mass range however suffers from a decreased scan rate in comparison to the AIF_930 method due to increased number of MS² events between each MS¹ scan (4 compared to 1). This allowed AIF_930 to detect more features due to the faster MS¹ scan rate allowing more data points to be collected across each chromatographic peak and therefore it can associate more features with MS² data. The AIF_430 method (average of 9,531) also provided a high number of features with MS² data associated but was lower compared to the previously mentioned two methods due to only performing fragmentation on half the total mass range (70 – 500 m/z). Following this there is a large decrease to the next best method, this is the Mullard method (average of 6,986). The number of features with MS² data for the AIF_930, DIA_4_232 and AIF_430 methods were much higher but the quality of that MS² data is in doubt due to the large size of the fragmentation windows, the number of features that are likely to be falling into these windows and the requirement for successful deconvolution of the many signals detected. This is not the case for the data dependent

based methods which includes the Mullard method. This method has required nine separate injections (not including any replication), these extra injections have conferred great advantages over any of the single injection DDA methods (Tra, Tra_exc, Tra_inc) with the highest number of features with MS² data for any of those three methods being the traditional DDA method (Tra) with an average of 1,120. Therefore, the Mullard method is providing a potentially massive increase in identifications and thus biological information if these MS² mass spectra can be confidently annotated. However, it is important to consider the cost of this increase in information in terms of sample, time and money required to achieve it. This also applies to the progressive traditional DDA method (Tra_P_total). Three injections were required for this method (not including any replication) and this generated an average of 2,736 features with MS² data associated. The traditional DDA methods are the 3rd, 4th and 5th lowest. Only the DIA methods with a window width of 10 *m/z* have less features with MS² data and these only cover a total MS¹ *m/z* range of 20 and 40 *m/z* respectively. As the window size of the DIA methods increases so does the number of features with MS² data, whilst more DIA windows will also have the same effect providing the MS¹ scan rate has been maintained as was the case with all the methods in this study. Greater coverage of features is a good thing but is only useful if the spectra can be successfully deconvoluted and are of good enough quality to provide a match upon comparison to a reference library. Data for the other assays can be found in the Appendix (9.3.2).

5.2.3 Number of Features Annotated

Features were annotated in MS-DIAL. The number of features with an MS² spectral match of any quality is displayed in Figure 71. The number of features with a high quality spectral match (≥ 70) is displayed in Figure 72.

The number of features with MS² data which could be matched to a reference spectrum was greatest for the full mass range AIF method (AIF_930) with AIF_430 and DIA_4_232 close behind. This is logical considering the number of features which had MS² data associated with them (Figure 70) however it still does not elucidate whether the quality of these MS² data has been maintained and is still useful. The distribution seen in Figure 71 is very similar to the distribution seen in Figure 70 but the values are about 6-fold lower. Therefore, for each method roughly 1 in 6 spectra could be matched to a reference spectrum. This provides an idea as to the quality of the spectra collected.

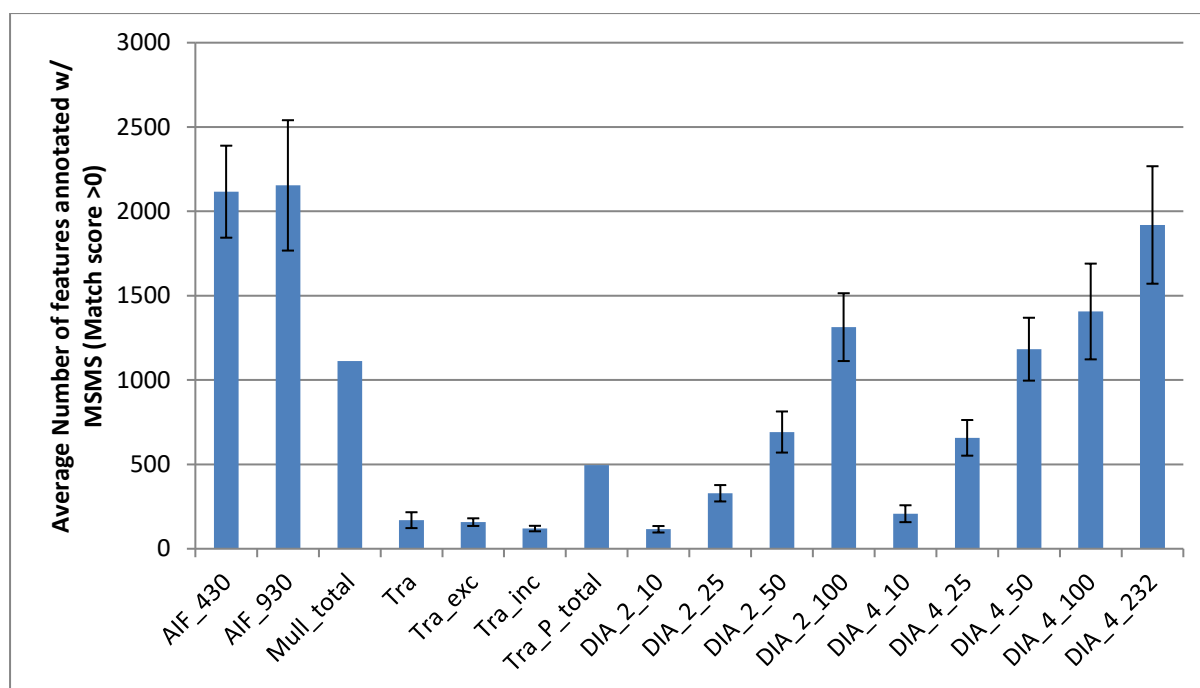


Figure 71: Average number of features with an MS² spectral match of any quality for each quadruplicate after processing in MS-DIAL for the HILIC/positive ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

A better indication of spectral quality is the number of high quality (match score ≥ 70) (Figure 72) matches. The distribution and heights of the bars has shifted considerably from the data in (Figure 70 and Figure 71). The DIA and AIF based methods show large decreases in comparison to the DDA methods. This is as expected as DIA MS² spectra require deconvolution and the effectiveness of this process is likely to be hindered at higher window sizes, this effect can be more clearly seen in Figure 73 where the percentage of total number of features which had MS² data associated which had a good match score (≥ 70) is displayed. The decrease in good matches seen with the AIF methods is not as pronounced as the decrease seen with the DIA_4_232 method. This is illogical as they had very similar numbers of features with MS² data associated and the smaller window size of 232.5 m/z of the DIA_4_232 method should have ensured data were of higher quality than achieved with the larger window sizes in the AIF methods of 430 and 930 m/z . This is reinforced by Figure 73 where the percentage of spectra with a good spectral match is just 0.78% for DIA_4_232 compared to 1.87% for AIF_930 and 2.64% for AIF_430. The two AIF methods have followed the expected pattern where the larger window size results in a lower percentage of spectra with a good match. A theory to explain the unexpected difference between the DIA_4_232 method and AIF methods is that the two AIF methods are always fragmenting the lower half of the mass range (70- 500 m/z). This is where the majority of metabolites are detected in the HILIC assay, whereas the DIA_4_232 method spends half of its MS²

scans fragmenting above this region. These spectra acquired in that region are less likely to contain fragments of good intensity that are derived from real metabolites that could be found in a reference database. At these large window sizes the deconvolution is always likely to be ineffective but the AIF method windows should at least always contain a high intensity feature whereas this might not be the case for many of the DIA_4_232 windows and thus the lower percentage is seen. The Mullard method (Mull_total) was the best performing DDA based method. It produced the greatest number of features with MS² data and a good match score with an average of 245 good matches. This is 91% more than what was achieved with the traditional DDA method (Tra) which again outperformed the traditional methods with an inclusion or exclusion list (Tra_exc, Tra_inc). The traditional progressive method (Tra_P_total) outperformed the traditional DDA method by 34% with an average of 173 good matches. These are certainly useful improvements to provide increased biological knowledge from a study but did require extra injections as previously mentioned and this needs to be considered when designing experiments. The traditional DDA method outperformed all the DIA methods except the two AIF methods, although considering the DIA methods (except DIA_4_232) only cover a portion of the mass range the performance is appropriate, particularly for the 10 *m/z* window methods (DIA_2_10 and DIA_4_10). So, it would appear that AIF methods are the best and most efficient for identification of most metabolite features in the most efficient manner. However, there are still major question marks over the quality of these MS² data and the deconvolution process. Using a DDA based strategy at least provides confidence in the composition of the MS² spectra with a direct link between precursor and product ions. Considering confidence in annotations is important in untargeted metabolomics, keeping this link is very valuable.

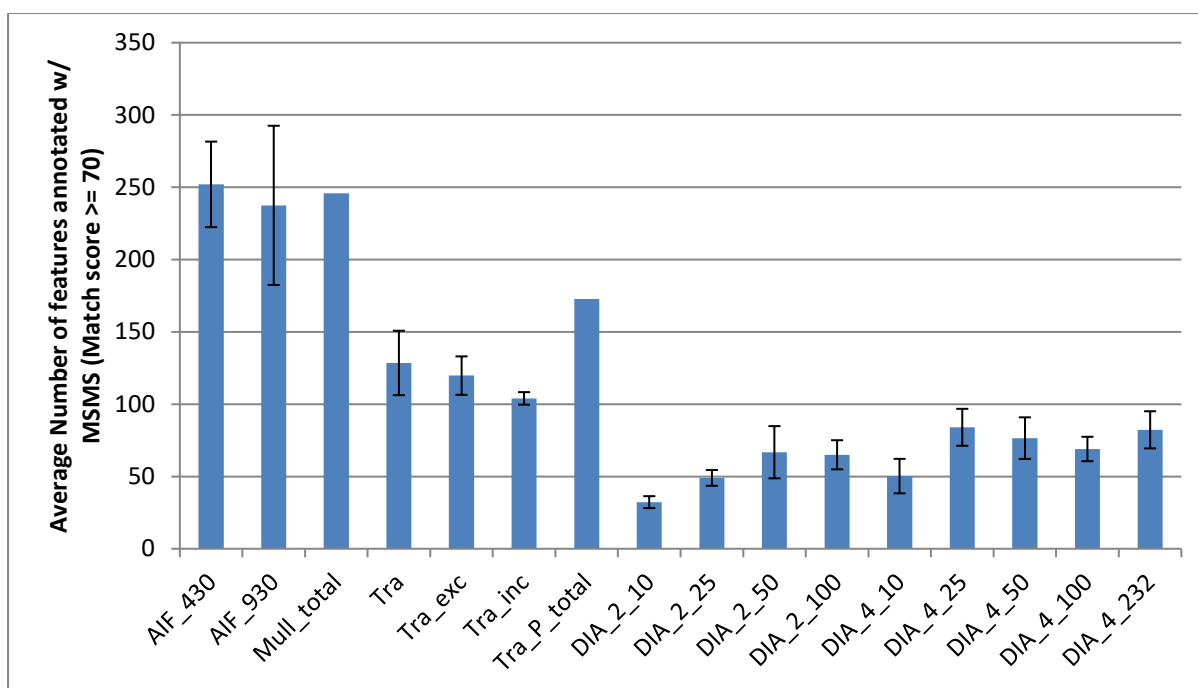


Figure 72: Average number of features with an MS² spectral match score ≥ 70 for each quadruplicate after processing in MS-DIAL for the HILIC/positive ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

Time spent collecting spectra that cannot be annotated is time that has been wasted. Making untargeted metabolomics more effective and reliable is part of the goal of this work. By employing a traditional DDA method the least time is wasted collecting MS² spectra which are subsequently unannotated and therefore could be said to be uninformative. The three single injection traditional DDA methods (Tra, Tra_exc, Tra_inc) averaged between 11.0 and 11.8 % of MS² mass spectra which could be annotated. There is a decrease from the single injection DDA methods to the progressive DDA methods (Mull_total, Tra_P_total). A theory to explain this is the decrease in intensity of features fragmented through each pass of the method as more of the higher intensity features are being added to the exclusion list. This explains why the traditional progressive DDA method (Tra_P_total) has a higher percentage than the progressive Mullard method (Mull_total) as these methods used 3 and 9 injections respectively. The DIA methods show the expected trend of a decreasing percentage of good matches as the window size increases. Interestingly using a narrow window of 10 m/z the percentage of spectra that could be annotated was greater than that seen with both progressive DDA method types. This indicates that DIA could be appropriate if the window size is narrow enough. However, the scan rate is insufficient on an Orbitrap instrument for a 10 m/z window method to be utilised that would still cover a typical untargeted metabolomics experiment mass range. Using a 25 m/z window also resulted in a similar percentage to the progressive DDA methods and so could be considered as a

reasonable m/z window width. This is backed up by literature in which DIA based metabolomics methods have employed window widths ranging from 18 to 65 m/z (Ma et al., 2016; Li et al., 2016a; Zhou et al., 2017; Tsugawa et al., 2015; Chen et al., 2017; Yan et al., 2018, 2019). Using larger m/z window sizes results in a very low percentage of spectra with a good match and this is making the assumption that the deconvolution of those spectra was good and the subsequent match can be trusted. However, if the sample type is of low complexity large window sizes for example of 100 m/z could still be appropriate (Zhou et al., 2017).

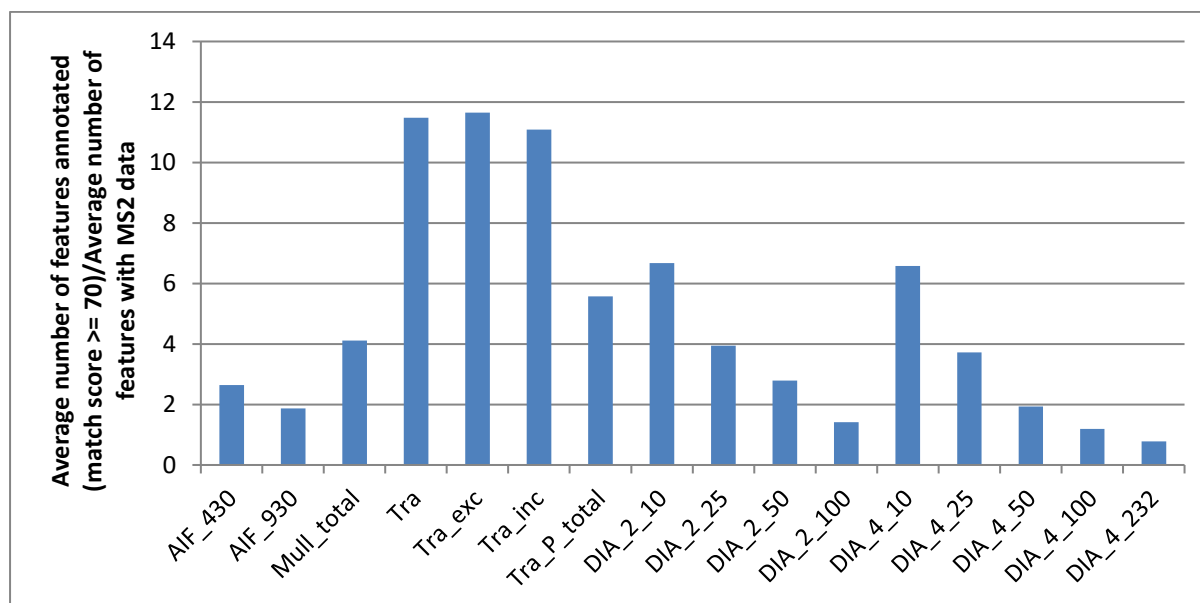


Figure 73: The percentage of average number of features annotated (match score ≥ 70)/average number of features with MS² data for each MS² strategy for the HILIC/positive ion mode dataset.

It is important to also consider that the libraries used for spectral matching were large and will have contained a number of spectra that may not be relevant to the sample type studied for example those from GNPS, a large natural products database and data analysis and sharing platform (Wang et al., 2016). As a result, the number of annotations achieved may be inflated by matching of plant metabolites or non-biological compounds. Individual annotations were not manually checked due to the volume of data analysed. Of the DIA workflows previously published, MS-DIAL is the only one which relies on external spectral databases. Both MetaboDIA (Chen et al., 2017) and MetDIA (Li et al., 2016b) rely on the previous construction of a custom made DDA based MS² spectral database. This, while a reliable approach to take for more confidence in annotations requires a lot of time and money to allow construction of the library. It also means the method is not truly untargeted as spectral deconvolution is not carried out, instead pseudo spectra are generated based on the spectra already

in the library and therefore only metabolites in the library can have pseudo spectra created for them. Data for the other assays can be found in the electronic Appendix (9.3.3).

5.2.4 Purity of Fragmentation Windows

The quality of spectra had been assessed through number of annotations, number of good quality annotations and percentage of MS² spectra providing a good quality annotation. The purity of spectra collected was assessed next. Purity in this context refers to the mathematical estimation of MS² spectral purity. It looks to provide the answer to the question, what percentage of the product ions seen in the MS² scan are products of the desired precursor? The estimation is performed by looking at the isolation window applied as well as the intensity of any m/z values found within that window in the preceding and following MS¹ mass spectra. Interpolation of m/z intensities between the two MS¹ scans allows an estimation of the intensities of each m/z value within the isolation window at the time that the MS² spectrum was actually collected. This subsequently allows an estimation of the purity of the spectrum, for example an isolation window with just the m/z value being fragmented and no other peaks would give a purity score of 100%. The interpolated purity of each fragmentation event was estimated using the R package msPurity (Lawson et al., 2017) (Figure 74). The purity of the MS² data is important to increase the quality of spectral matching (reducing false positives and increasing true positives) when applying a data dependent method. The highest median purity is seen when applying the traditional DDA method (Tra) (mean inPurity = 0.60). The traditional method with an inclusion list (tra_inc) performed similarly well, with a similar distribution seen and a similar mean (0.58). The violin plots for both of these methods are relatively wide at the point where the purity is 100%. The traditional method with an exclusion list (Tra_exc) shows a significantly lower distribution of purities. This is perhaps due to the greater number of lower intensity features that are being fragmented as a result of no exclusion list being applied. The exclusion list method has not spent time fragmenting any of the high intensity features derived from the sample matrix and solvent and therefore the average purity of fragmentation windows has decreased. Major decreases in purity are also seen when performing repeated injections, particularly from the first to the second injection. This demonstrates that the exclusion lists are working as desired and “deeper” annotation is being performed by collecting data for more features whose intensity is nearer to the noise level. However, this also highlights the need to take measures to increase the purity of fragmentation windows for lower intensity features. The fragmentation window size applied in this study was 3 m/z (1.5 m/z either side of the central point). A window width of this size is typically employed so that the ¹³C isotope is included in the fragmentation as this can aid annotation. In a complex sample however, this width is likely to lead to chimeric spectra particularly in high complexity regions of the RT: m/z axes. Furthermore, at this low intensity, the ¹³C isotope peak is present at an even lower intensity than the

low intensity parent being fragmented and therefore the successful transmission and detection of the fragments of the ^{13}C isotopic peak is unlikely. It may be beneficial to be able to employ different isolation widths depending on the intensity of the precursor ion being fragmented to achieve more identifications, however this is not currently possible with the current mass spectrometers available. Decreasing the isolation window width can have negative consequences, if the window is too narrow the ion transmission efficiency will be decreased and the fragmentation spectrum may end up being noisier and less informative as a result.

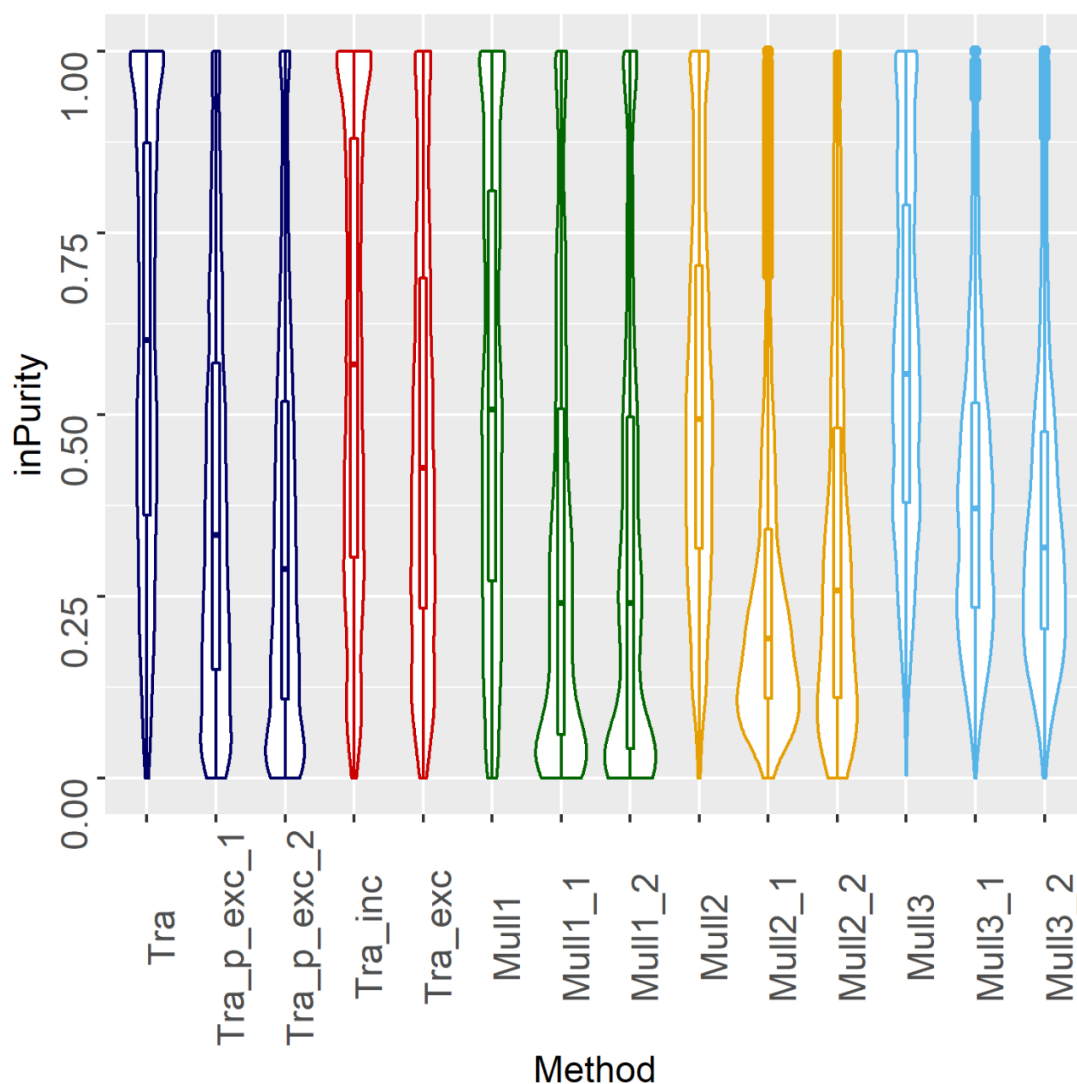


Figure 74: The distribution of interpolated purity (inPurity) scores for all DDA based methods for the HILIC_POS data.

Whilst purity cannot be truly assessed for DIA data as there is no precursor selection, the purity can still be calculated based on whichever feature within the window was present at the highest intensity. The purity distributions for the DIA data are displayed in Figure 75. As expected, the purity of the data

increases as the window width decreases. When applying the same window size either 2 or 4 times, using 2 windows consistently confers greater average purity. This is likely due to the way in which the purity is calculated whereby the preceding and subsequent MS¹ scans either side of the relevant MS² scan are used to estimate the purity. The greater interval between MS¹ scans should lead to a decrease in the average purity estimates. Another reason could be that the extra two windows added are tending to be of lower complexity than the first two windows which were specifically selected to cover a high complexity region. An unexpected result was that the AIF methods showed higher purity than a number of the DIA methods. This can be explained by considering that there must always be a high intensity feature being fragmented by the AIF method which will likely dominate the fragmentation spectrum and ensure the purity score is not too low despite the very large size of the isolation window. This is much less likely to be the case for the DIA methods. An important consideration is that the most intense feature was used to determine the purity and so these purity estimations are a best case scenario for all features within the window and emphasises the amount of work required during deconvolution. Data for the other assays can be found in the Appendix (9.3.4).

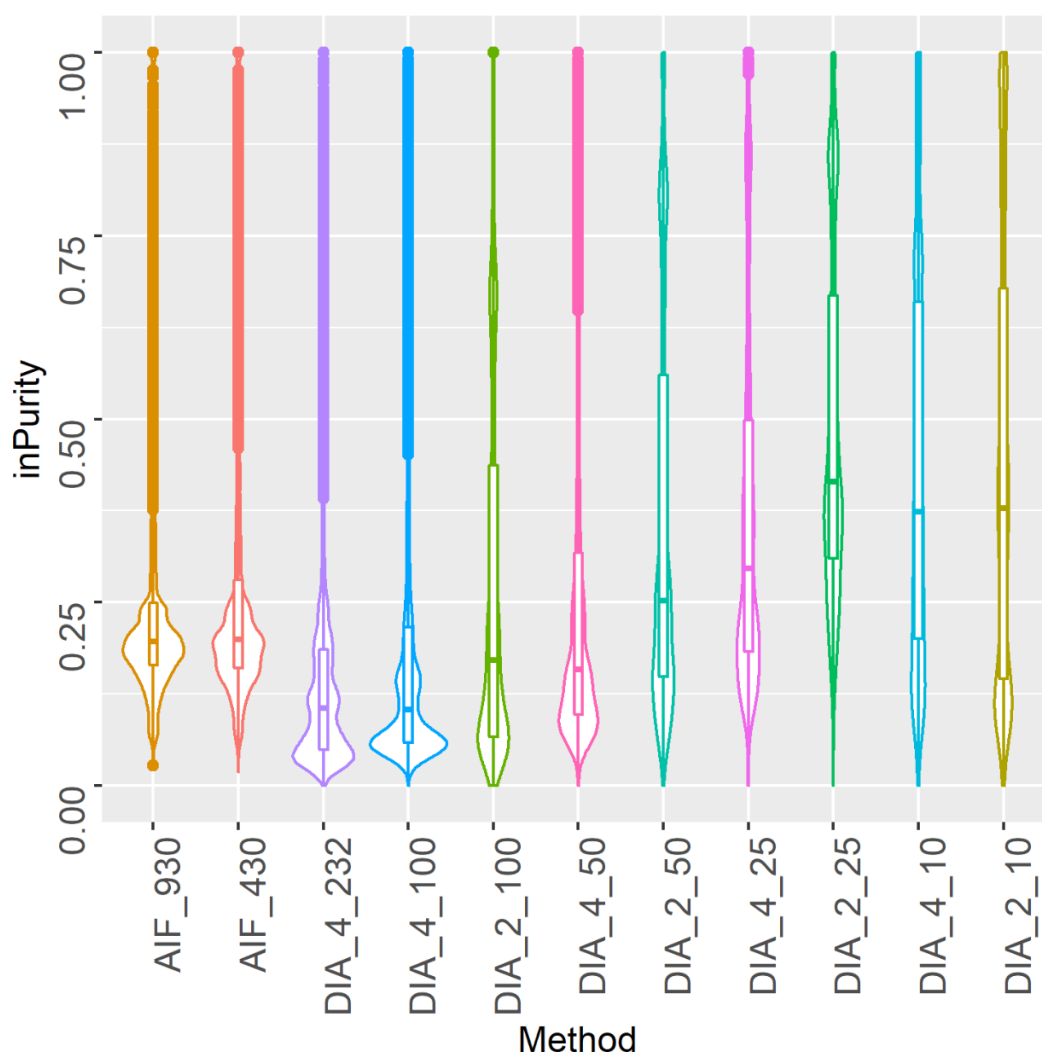


Figure 75: The distribution of interpolated purity (inPurity) scores for all DIA based methods for the HILIC_POS data.

5.2.5 Advantages of Repeated Injections

Following the assessment of volume and quality of the MS² data collected through match scores and purity scores the value of repeated injections for the progressive DDA methods was assessed. If performing repetitive injections with updated exclusion lists an important aspect of the method to consider is when to stop repeated injections particularly if time, money or sample is limiting. How valuable are the repeated injections for metabolite annotation? The number of features with MS² data gradually decreases through each pass of the progressive traditional DDA method (Figure 76A). This is logical because as more *m/z* values are added to the exclusion list the number of features available in the data that meet the required intensity threshold to trigger a fragmentation event decreases. The decrease seen in the number of good spectral matches through each injection however is much more dramatic (Figure 76B). Average features with a good spectral match plummet from 129 in the first

pass to just 32 in the second and 12 in the third. This rapid decrease in good matches (Figure 76B) compared to the gradual decrease in the features with MS² data (Figure 76A) indicates that the quality of the MS² spectra being collected is decreasing. As the intensity of the features fragmented decreases so does the quality of the mass spectra, as it can be expected that the intensities of MS² mass spectra are at least an order of magnitude lower than those of MS¹ spectra (Neumann et al., 2013). This is backed up by the spectral purity data (Figure 74) which shows that the majority of MS² spectra collected in the second and third passes were below 25% purity and thus unlikely to generate a high match score.

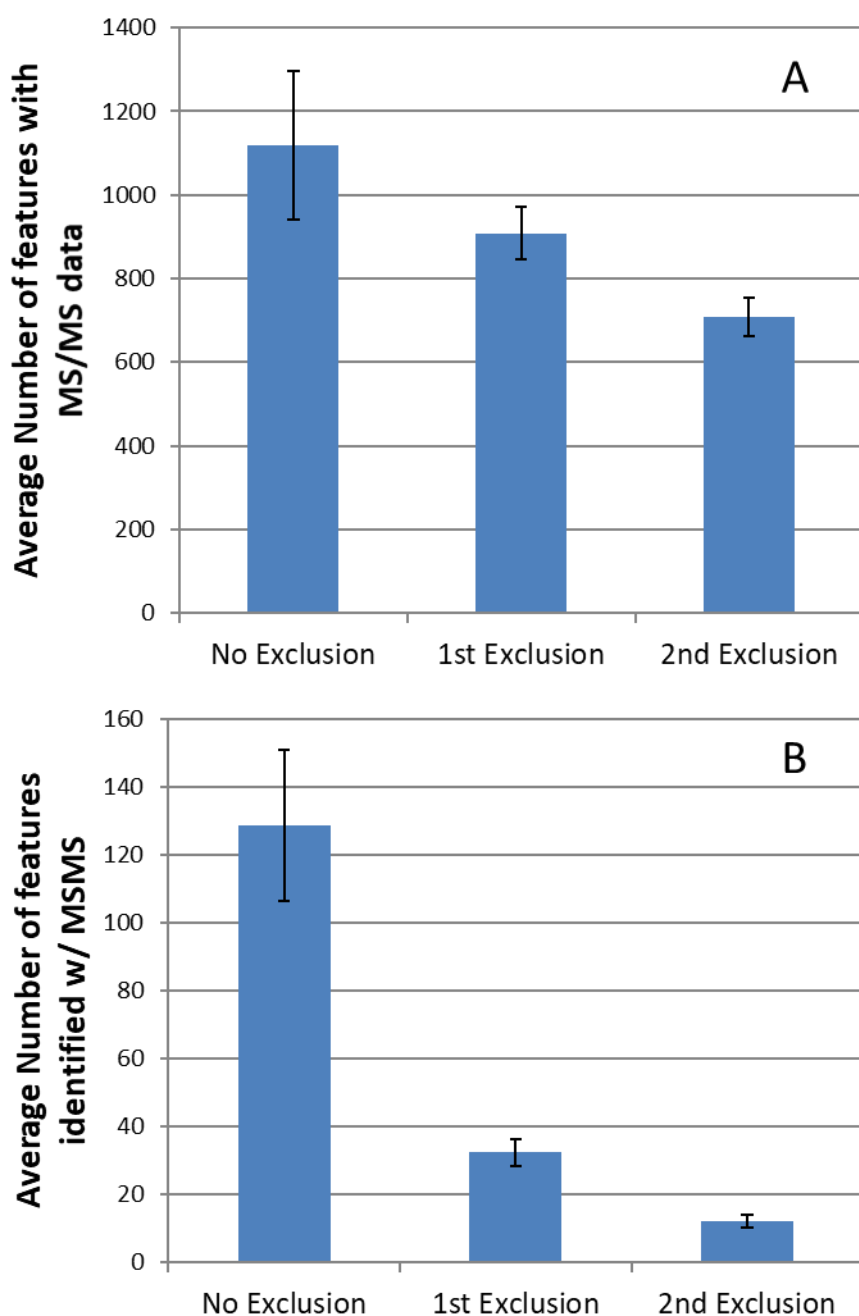


Figure 76: A) The number of features with MS² data associated with them through each pass of the traditional progressive DDA method for the HILIC positive ion mode dataset. B) The number of features with spectral match score ≥ 70 (MS-DIAL) for the same dataset.

The number of features with MS² data associated with them decreases at different rates depending on the width of the Mullard segment applied and the assay applied. In this case the segments were Mull1 (70-200 m/z), Mull2 (200-400 m/z), Mull3 (400-1000 m/z). These segments were selected based on the complexity of the data seen in the complexity assessments in the electronic Appendix (4.2/4.2.3) to try to provide more coverage between each segment. The data indicates that this

segmentation was too heavily weighted in the favour of the lower mass regions. This is demonstrated by the average number of features detected in each segment which were 2,498, 5,667 and 13,496 respectively for segments Mull1, Mull2 and Mull3. This explains why we have the most rapid decrease proportionally in Mull1, a more gradual decrease in Mull2 and no decrease at all in Mull3 (Figure 77A). The number of features with a good spectral match decreases rapidly through each pass of segments Mull1 and Mull2 with average numbers falling from 102 to 35 to 7 in Mull1 and 76 to 14 to 3 in Mull2 (Figure 77B). A decrease is seen with Mull3 from the first pass to the second from 6 to 2 features with a good spectral match, it then remains at 2 in the third pass. The number of features with good spectral matches in Mull3 is extremely low, this might be as expected considering it is a HILIC method and the majority of features being detected will be present in the lower regions of the mass range. The decreases in number of good spectral matches through each pass is consistent with what was seen with the traditional progressive DDA data (Figure 76B). It is also consistent with the decrease in purity seen through each pass of each segment (Figure 74).

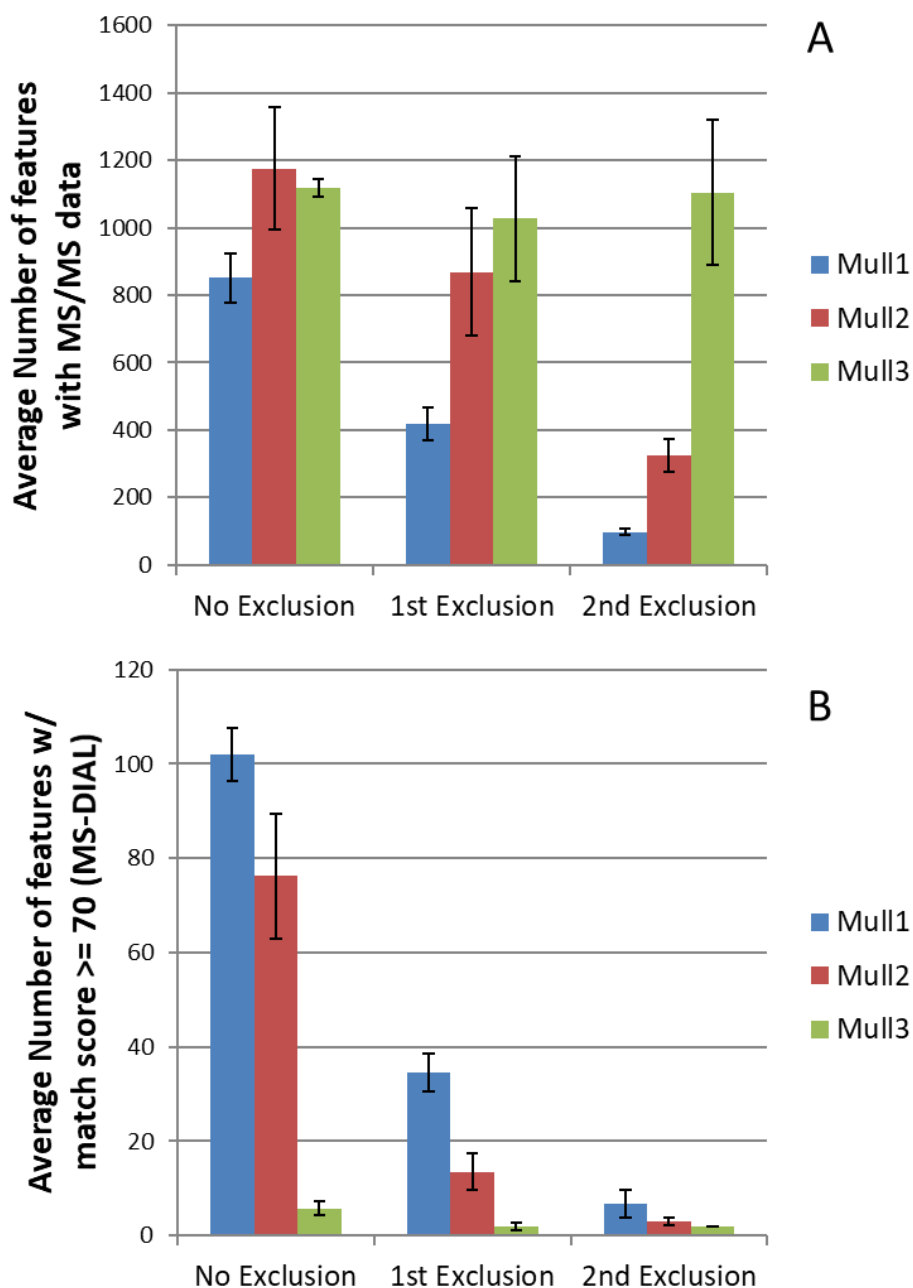


Figure 77: A) The number of features with MS² data associated with them through each pass of the three progressive Mullard segments for the HILIC positive ion mode dataset. B) The number of features with spectral match score >= 70 (MS-DIAL) for the same dataset.

In general, the amount of high quality MS² spectral information provided by the third pass of each method was very low, particularly in the Mullard segments. If performing Mullard segmentation then there is little value in performing the third injection. Without segmentation it is more beneficial but still of relatively low impact. This further highlights the need for ways to improve the quality of fragmentation spectra acquired for low intensity ions. Further segmentation could be implemented to try to increase sensitivity, or modification of the chromatography to achieve the same goal. Apex

triggering can also be performed on some types of mass spectrometer and should be implemented to ensure quality of MS² data for low intensity features (Neumann et al., 2013). This is where fragmentation is delayed until the peak reaches its point of greatest intensity and is understandably of greater importance for lower intensity features.

Whilst there is reduced value in the repeated injections, there are still good spectral matches being made. A high number of injections could be required for complete coverage at high density points on the RT axis where there are many co-eluting features whilst other regions may only require one or two (Neumann et al., 2013). This highlights the importance of using multiple assays for separation as metabolites which co-elute in one assay may separate well in another. Alternatively, 2D-LC could be applied to provide an extra dimension of separation and thus reduce co-elution and thus the number of injections required for full coverage. Performance of the 2nd and 3rd injections could also have been improved through introduction of RT values into the exclusion lists. It is likely that some isobars and isomers of features fragmented in the 1st injection were ignored in subsequent injections due to the RT of the features to be excluded not being specified and this is important to consider. Furthermore, the capacity of the exclusion list in the software is only 5000 and this capacity was not enough to include all fragmented *m/z* values for some methods after the 2nd pass which will have resulted in some repetitive data being collected. This highlights the importance of software improvement as well as continuing improvement of mass spectrometers themselves.

A final consideration is that the goal of gaining MS² information on all features is not necessary. When the number of different features of a single metabolite can be as high as 100 (Mahieu et al., 2016). Collecting MS² data 100 times for a single metabolite is not efficient and is pointless when it is considered that the vast majority of reference spectra in MS² databases are acquired for the protonated molecular ion. Efforts should be focused towards ensuring only data for the protonated ion is collected however this requires some kind of survey and prior processing of the data such as with the nearline algorithm (Neumann et al., 2013). This is something that has not been routinely feasible until recently with the release of the Orbitrap ID-X (Thermo Fisher Scientific, USA) which can perform online processing of a survey full scan analysis, followed by repetitive injections of the same sample with intelligent feature selection for fragmentation of molecular ions based on the processing results. Data for the other assays can be found in the Appendix (9.3.5).

5.2.6 MS-DIAL Deconvolution Assessment

The number of identifications for the DIA data collected had been discussed earlier but the actual deconvolution of the original complex DIA spectra had not been investigated. The effect of different window sizes on the annotation of high and low intensity features was investigated for all window

sizes and was compared to the DDA annotations of the same features. Examples are included with the full analysis included in the electronic Appendix (5.2/5.2.6).

5.2.6.1 High Intensity Features

The quality of the deconvolution process at different window sizes was assessed by comparing the dot product and reverse dot product scores that are produced during MS-DIAL processing. The dot product score considers all m/z peaks in the experimentally recorded spectrum when calculating the match score. The reverse dot product score will ignore all peaks in the experimentally recorded spectrum that are not found in the reference spectrum. As a result of this the reverse dot product can still produce a high spectral similarity score even if there are many contaminant or noisy peaks in the experimental spectrum. This makes it useful for assessing DIA spectra and the quality of deconvolution. A well deconvoluted spectrum should have a high dot product and reverse dot product score. Whilst a poorly deconvoluted spectrum should produce a poor dot product score and a high reverse dot product score. The effect of the changing window sizes on the detection of a high intensity feature annotated as kynurenine (Figure 78) in the HILIC positive ion mode data set was assessed at varying DIA/AIF window sizes in comparison to a traditional DDA method (Table 47).


RT [min]	Mz [Da]	Type	Area	Intensity	Gaussian Sim	Chromatogram
4.74	209.0919	[M+H] ⁺	1.305377E+08	4.878444E+07	0.927346	

Figure 78: Details of the feature annotated as kynurenine in the HILIC positive ion mode dataset.

Table 47: The dot product and reverse dot product scores for a feature annotated as kynurenine in a traditional DDA method and the DIA/AIF based methods.

File	Dot Product	Reverse Dot Product	Status
HILIC_POS_DDA_Tra_rep2	786	955	Annotated as kynurenine
HILIC_POS_DIA_2_10_rep3	730	918	Annotated as kynurenine
HILIC_POS_DIA_4_10_rep3	698	910	Annotated as kynurenine
HILIC_POS_DIA_2_25_rep3	716	917	Annotated as kynurenine
HILIC_POS_DIA_4_25_rep3	703	916	Annotated as kynurenine
HILIC_POS_DIA_2_50_rep3	651	873	Annotated as kynurenine
HILIC_POS_DIA_4_50_rep2	665	899	Annotated as kynurenine
HILIC_POS_DIA_2_100_rep3	671	886	Annotated as kynurenine
HILIC_POS_DIA_4_100_rep3	658	894	Annotated as kynurenine
HILIC_POS_DIA_4_232_rep1	204	689	Annotated as chalcone
HILIC_POS_AIF_430_rep1	281	747	Annotated as chalcone
HILIC_POS_AIF_930_rep3	282	739	Annotated as chalcone

The dot product score of the DDA method as would be expected was greatest at 786, the experimentally recorded spectrum is shown in comparison to the reference in Figure 79. There are some noisy or contaminant peaks present in the spectrum, however, the dot product score of 786 is still a high score, whilst the reverse dot product is close to a perfect match score of 1000. This was classified as identified by the software, this means that a spectral match score of at least 70 out of 100 was achieved. However, it is not clear how this score is calculated from the dot product and reverse dot product scores and the score out of 100 does not appear to be displayed anywhere in the version of the software that was utilised. This is despite the fact that an identification cut off point must be selected out of 100, not 1000 when going through the processing parameters. It does seem though that it must be more heavily weighted towards the quality of the reverse dot product score.

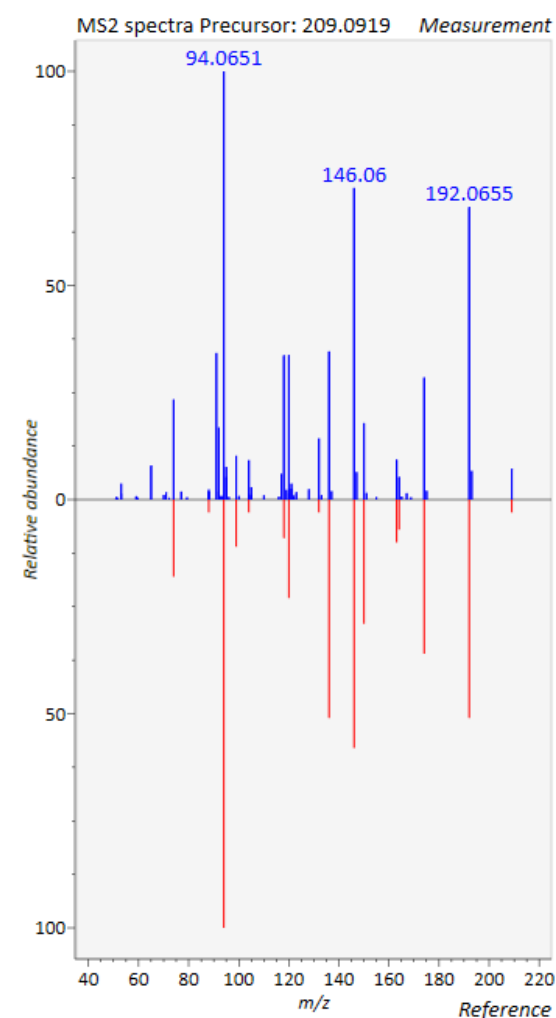


Figure 79: The traditional DDA spectrum for Kynurenine is displayed in blue, the reference spectrum used to assign the identification is displayed in red below.

The narrower window DIA methods (10 and 25 m/z) produce good quality matches only slightly below the quality of match achieved using the DDA method. The pre- and post-deconvolution spectra for these methods are displayed in Figure 80 and Figure 81 respectively. The raw spectra are not significantly more complex than the DDA spectrum (Figure 79). A small number of low intensity contaminant peaks have been removed during the deconvolution process however it appears that at this window size the feature annotated as kynurenine has constituted the majority of the ions fragmented in these windows. Some contaminant peaks still remain and this can be seen by the slightly decreased dot product scores for these methods in comparison to the DDA method. This demonstrates the capability of DIA to perform accurate identification at narrow window sizes for high intensity features.

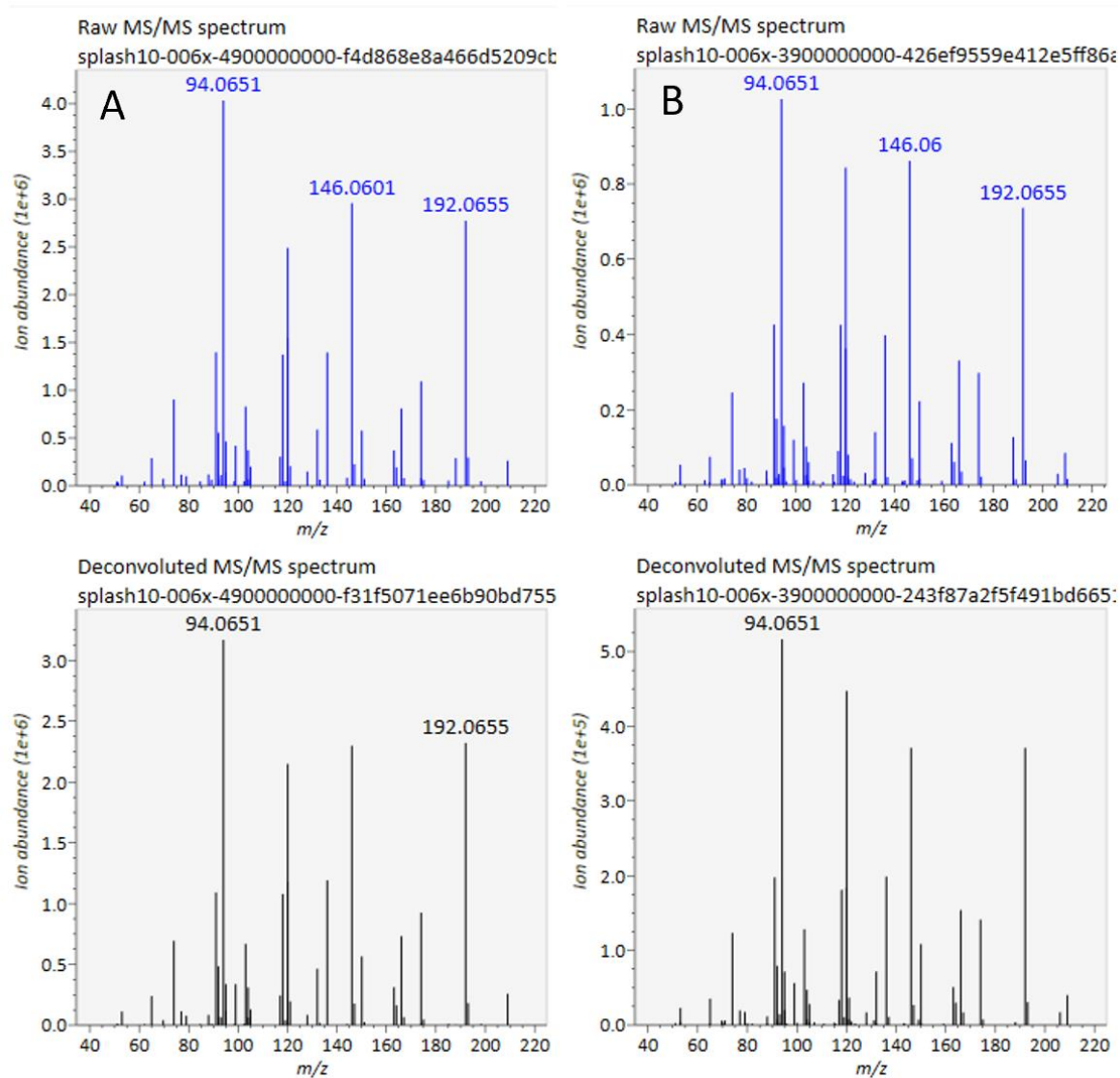


Figure 80: The raw and deconvoluted MS² spectra for the feature annotated as kynurenine in the HILIC positive ion mode dataset for 10 m/z window files A) HILIC_POS_DIA_2_10_rep3, dot product = 730, reverse dot product = 918 B) HILIC_POS_DIA_4_10_rep3, dot product = 698, reverse dot product = 910.

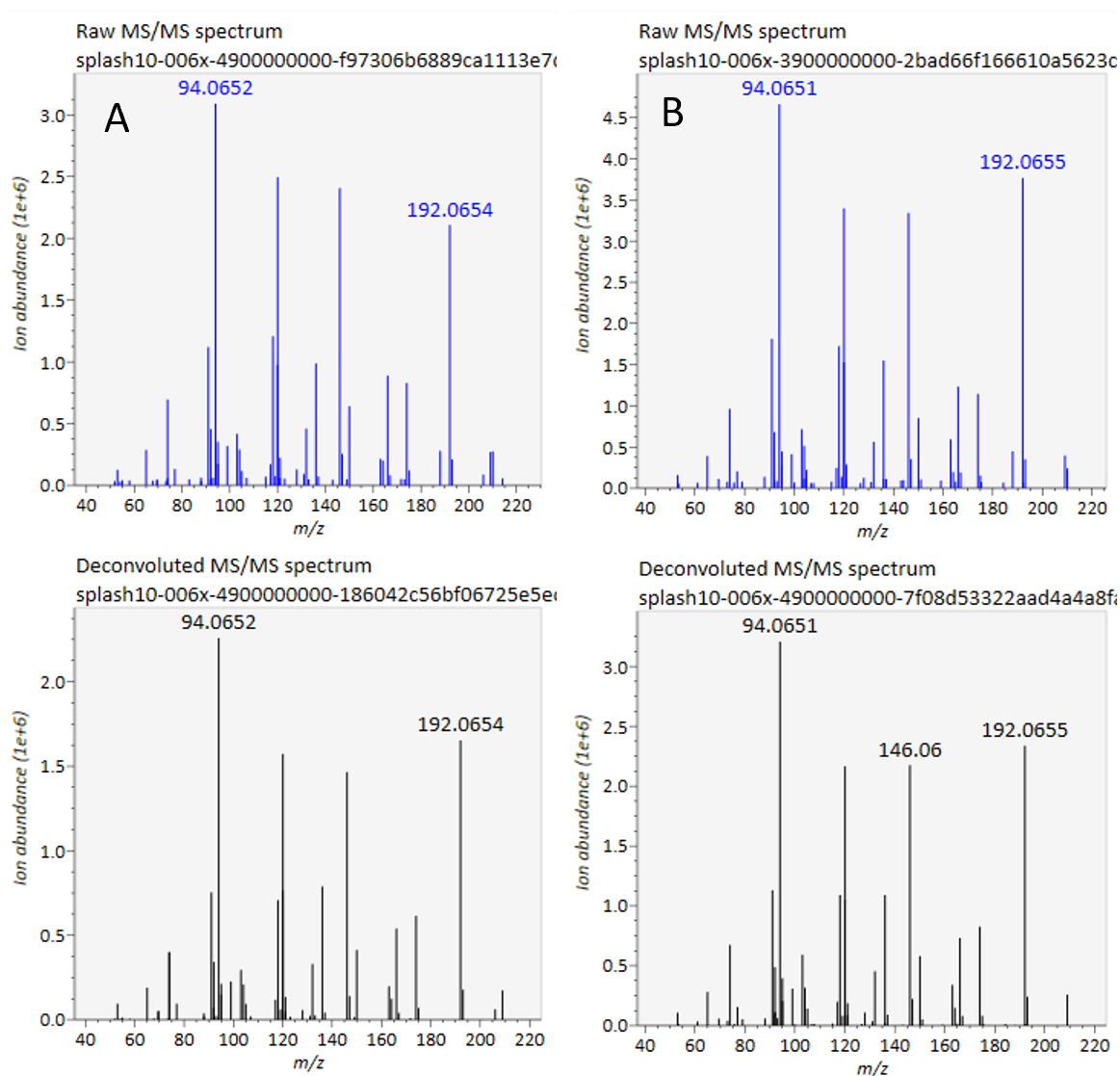


Figure 81: The raw and deconvoluted MS² spectra for the feature annotated as kynurenine in the HILIC positive ion mode dataset for 25 *m/z* window files A) HILIC_POS_DIA_2_25_rep3, dot product = 716, reverse dot product = 917, B) HILIC_POS_DIA_4_25_rep3, dot product = 703, reverse dot product = 916.

As the window size increases further to 50 (Figure 82) and 100 *m/z* (Figure 83) it is clear the complexity of the raw spectra has increased significantly particularly in the low mass region to the spectra seen with lower window sizes and the DDA method. The deconvolution appears to have worked quite well when looking at the spectra and this is reflected in the dot product scores which have decreased but only by about 50 in comparison to those seen with the 10 and 25 *m/z* window methods. This shows that even with a relatively wide window of 50 or 100 *m/z* that high intensity features can be annotated correctly, assuming that the DDA annotation is correct. The ratios of the 4 dominant peaks in the spectra are more likely to be shifted at higher window sizes as seen in Figure 83A where the

peak at 120.0808 m/z has become dominant which is not the case in any of preceding spectra displayed. These disruptions to the ratios of the reference spectra peak intensities cause the slight decreases in reverse dot product scores that are occurring as the window size increases.

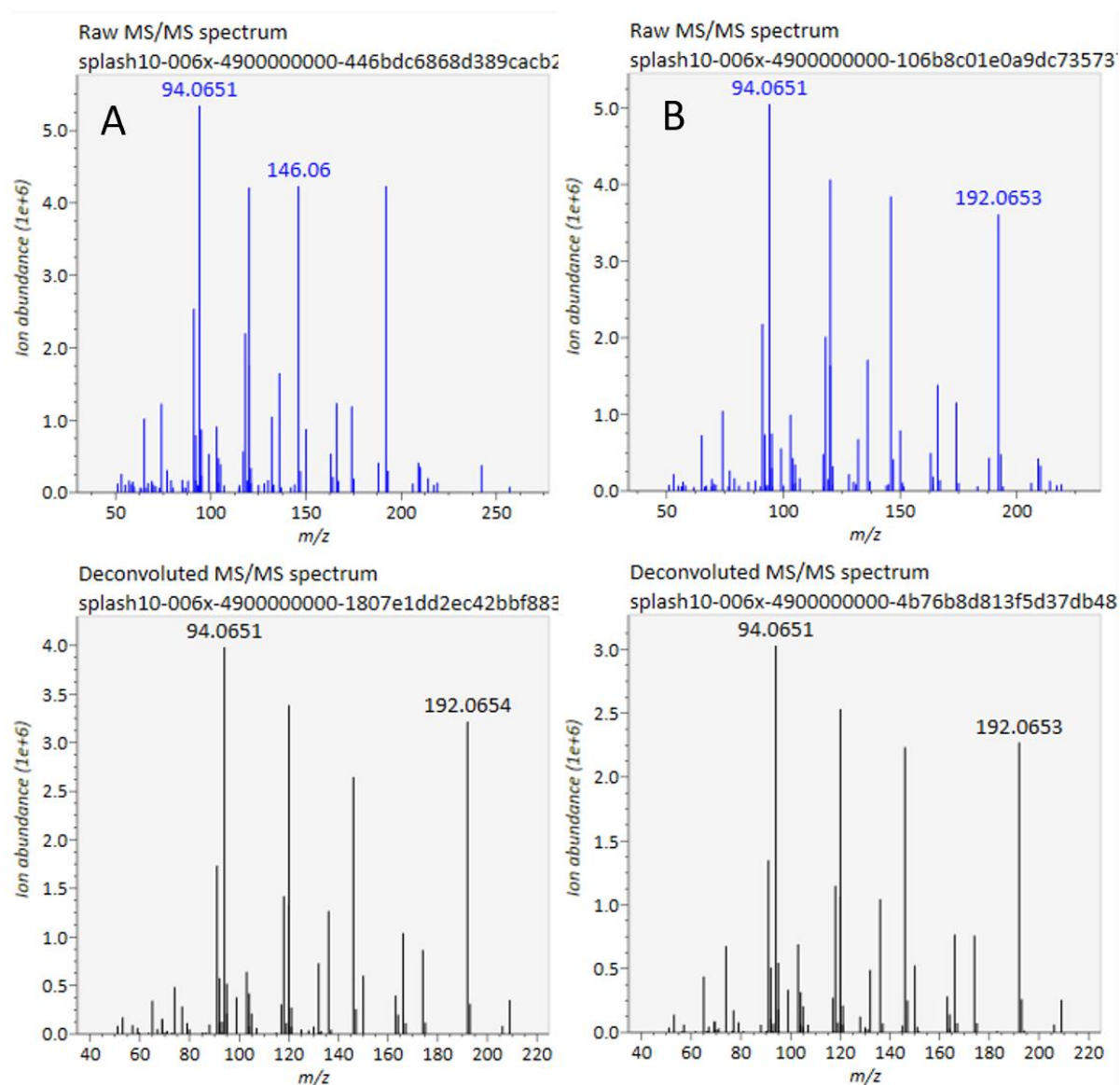


Figure 82: The raw and deconvoluted MS² spectra for the feature annotated as kynurenine in the HILIC positive ion mode dataset for 50 m/z window files A) HILIC_POS_DIA_2_50_rep3, dot product = 651 , reverse dot product = 873, B) HILIC_POS_DIA_4_50_rep2, dot product = 665, reverse dot product = 899.

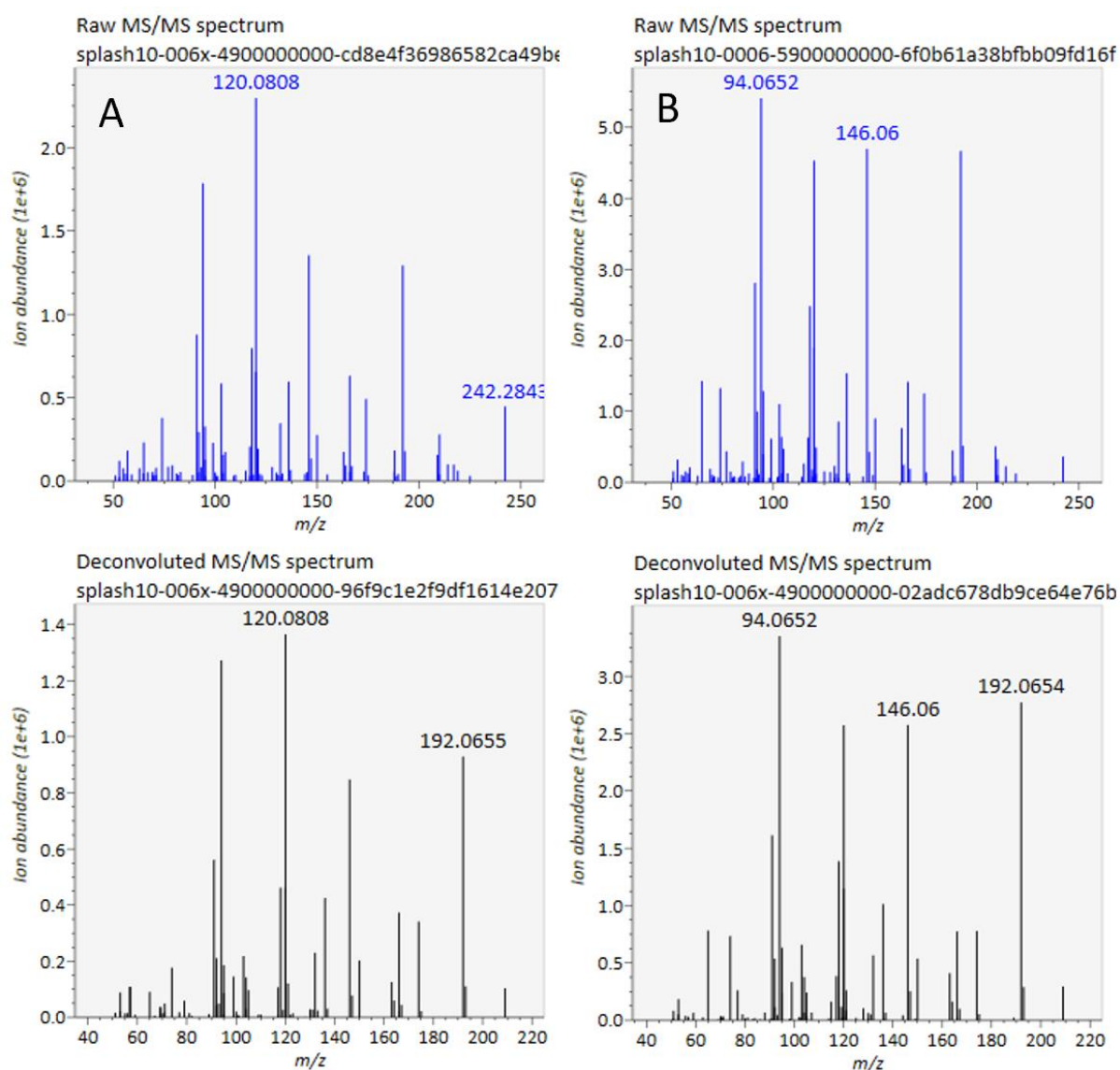


Figure 83: The raw and deconvoluted MS² spectra for the feature annotated as kynurenine in the HILIC positive ion mode dataset for 100 m/z window files A) HILIC_POS_DIA_2_100_rep3 , dot product = 671, reverse dot product = 886, B) HILIC_POS_DIA_4_100_rep3, dot product = 658, reverse dot product = 894.

When utilising a very large window size such as those in Figure 84 the same feature was identified as chalcone instead of kynurenine. The raw spectra are very different to all those collected at lower window sizes with the peak at 120.0808 becoming very high intensity and dominant within the spectra. This shows how when utilising these large window sizes the annotations cannot be reported as confidently.

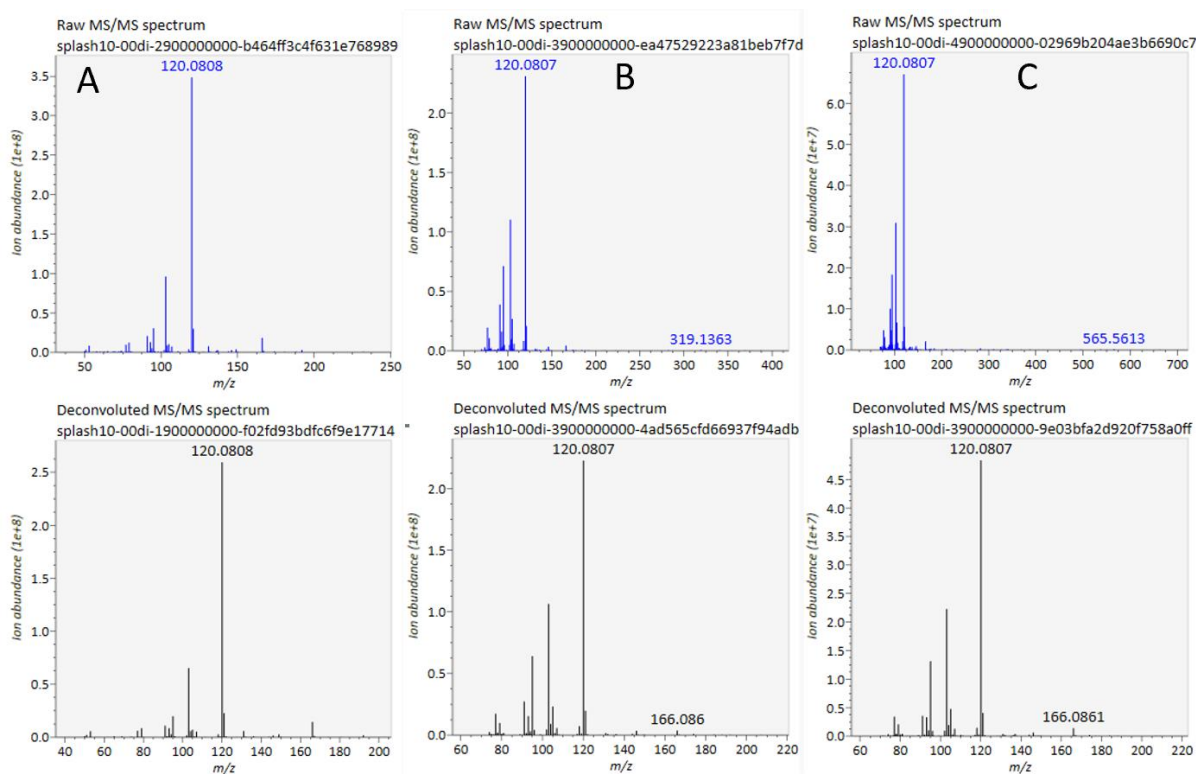


Figure 84: The raw and deconvoluted MS² spectra for the feature annotated as kynurenine in the HILIC positive ion mode dataset for the 232.5 m/z window file, and AIF methods A) HILIC_POS_DIA_4_232_rep1, dot product = 204, reverse dot product = 689, B) HILIC_POS_AIF_430_rep1, dot product = 281, reverse dot product = 747, C) HILIC_POS_AIF_930_rep3, dot product = 282, reverse dot product = 739.

5.2.6.2 Low Intensity Features

The main advantage to using a DIA method over a DDA method is to collect MS² data on all features detected within the dataset. Identification of low intensity features needs to be achieved for a DIA method to be providing any advantage over a DDA method. Therefore, whilst it was important to assess if DIA methods correctly identified high intensity features the true test is to determine if low intensity features for which fragmentation data would not normally be acquired if applying a DDA based method can be identified using a DIA method. The features which had been “identified” (score of ≥ 70 after MS-DIAL processing) were filtered by intensity. The DIA method which should be best suited to identifying a low intensity feature is the method which applies a narrow mass window. Therefore the 10 m/z window methods were investigated first. Very few low intensity features were “identified” (match score ≥ 70). One low intensity feature identified was cyclo(proline-leucine), the extracted ion chromatogram (EIC) and feature information are displayed in Figure 85 and Figure 86. This feature was detected but had no fragmentation information collected for it across any of the DDA based methods.

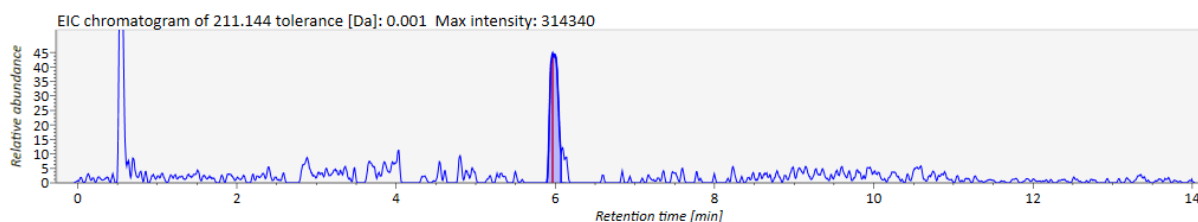


Figure 85: EIC for m/z 211.1440 in HILIC_POS_DIA_2_10_rep3.

RT [min]	Mz [Da]	Type	Metabolite Name	Area	Intensity	Gaussian Sim	Chromatogram
5.97	211.1440	[M+H] ⁺	Cyclo(proline-leucine)	1020421	140273.8	0.6782201	

Figure 86: Feature information from HILIC_POS_DIA_2_10_rep3 for a low intensity identified feature.

The DIA_2_10 method was investigated. The raw MS² spectrum has a number of high intensity features which have been removed in the deconvoluted spectrum (Figure 87A). Deconvolution has been relatively effective in this case by removing those high intensity features however there are a high number of noisy looking peaks in the spectrum still and so the deconvolution process could be still be improved. The deconvoluted spectrum is shown in comparison to the cyclo(proline-leucine) reference spectrum in Figure 87B. There are six peaks in the reference spectrum all of which are present in the deconvoluted spectrum except the peak just below an m/z of 155. The only one of the 5 matching peaks with a suitable looking ratio is the peak at m/z 211 and at least 3 of the matching peaks look like they could be noise. This spectral match resulted in an “identification” with a dot product score of 408 and a reverse dot product score of 780. A relatively high dot product score is recorded due to the presence of 5 out of the 6 peaks found in the experimental spectrum that are in the reference spectrum. The dot product score is relatively low due to the high number of noisy contaminant peaks in the spectrum. Although assigned as an identification by MS-DIAL, upon manual inspection it is clear that this feature could not be confidently annotated as cyclo(proline-leucine).

Table 48: The dot product and reverse dot product scores for a feature annotated as Kynurenine in a traditional DDA method and the DIA/AIF based methods.

File	RT (mins)	Dot Product	Reverse Dot Product	Status
HILIC_POS_DDA_Tra_rep2	NA	NA	NA	No MS ²
HILIC_POS_DIA_2_10_rep3	5.97	408	780	Annotated as cyclo(proline-leucine)
HILIC_POS_DIA_4_10_rep3	0.55	143	926	Annotated as cyclo(proline-leucine)

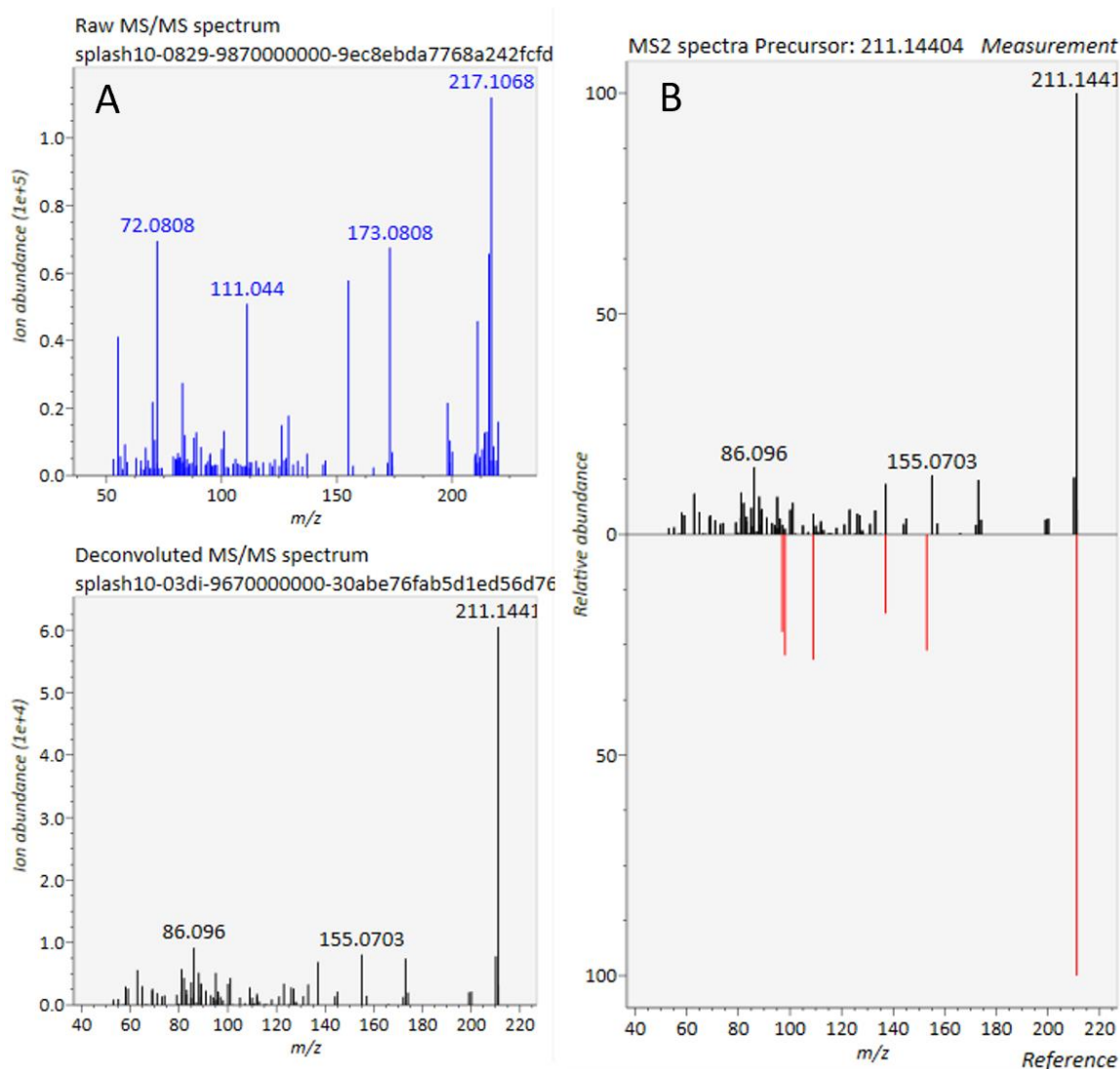


Figure 87: A) The raw vs deconvoluted spectra for the feature annotated as cyclo(proline-leucine) in the file HILIC_POS_DIA_2_10_rep3. B) The deconvoluted spectrum versus the reference spectrum for cyclo(proline-leucine)

The DIA_4_10 method was investigated. The same m/z was also identified as cyclo(proline-leucine) in the 4 x 10 m/z window DIA method however at a different retention time (Figure 88), a feature with the same m/z and RT as seen in the DIA_2_10 method was not present in this method. The level of noise present in the raw spectrum however was much higher, and the deconvolution failed to remove the majority of this noise (Figure 89A). This demonstrates the difficulty for the deconvolution in a high-density area of the m/z -RT axes in comparison to the lower density area where the deconvolution occurred and how the deconvolution process will not be effective even with a narrow window size in a high-density region. This further demonstrates DIAs unsuitability to the task at hand.

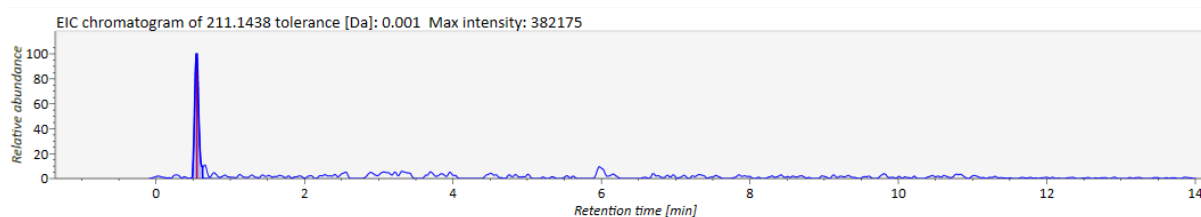


Figure 88: EIC for m/z 211.1440 in HILIC_POS_DIA_4_10_rep3.

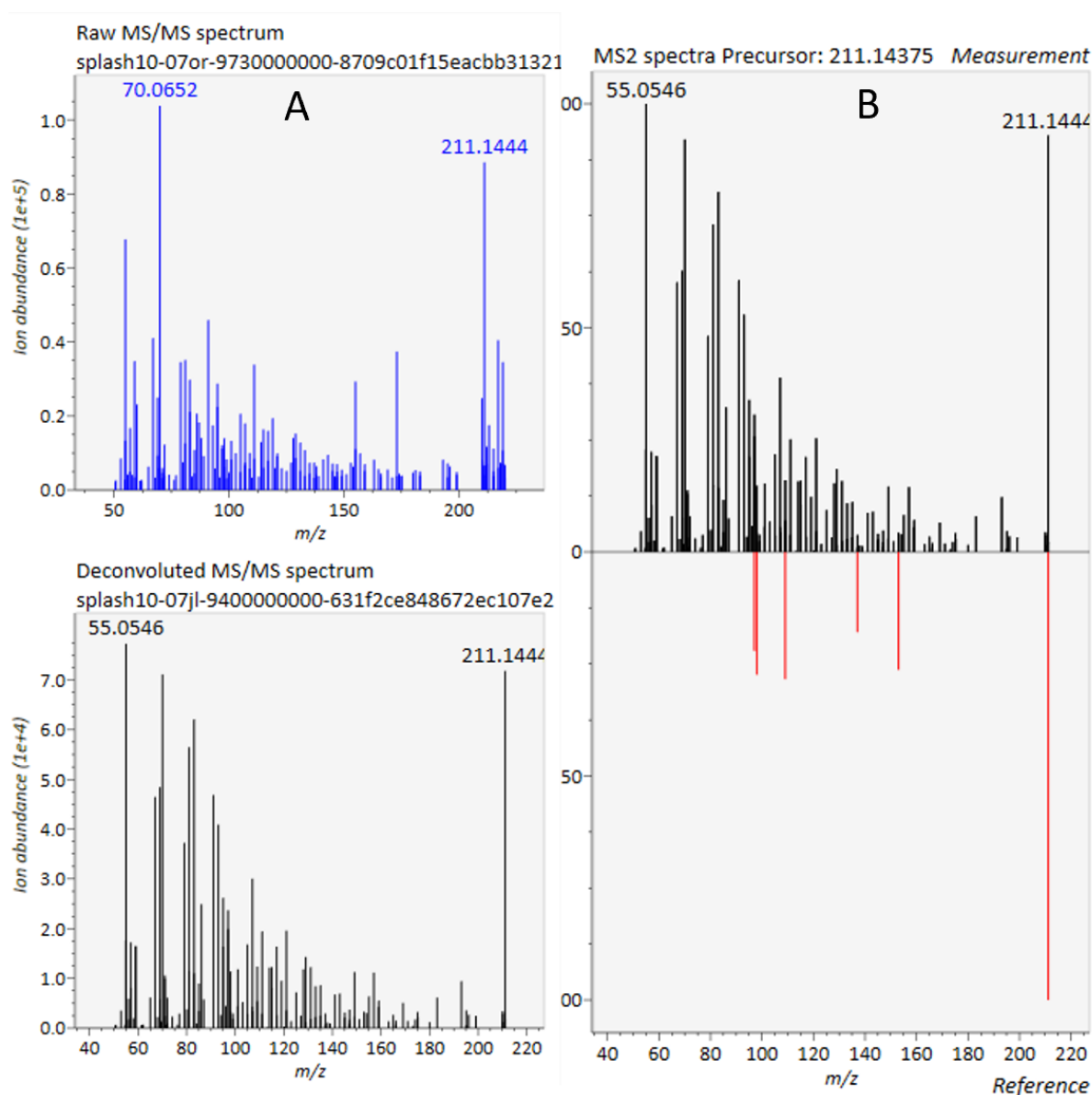


Figure 89: A) The raw vs deconvoluted spectra for the feature annotated as cyclo(proline-leucine) in the file HILIC_POS_DIA_4_10_rep3. B) The deconvoluted spectrum versus the reference spectrum for cyclo(proline-leucine)

To make DIA feasible the quality of deconvolution needs to be improved. This requires improved algorithms, improved chromatography to allow easier differentiation of different precursor/product

chromatographic profiles or both simultaneously. However, whether there is any value in pursuing a DIA method is highly questionable. While it may be suitable for analyses on TOF based instruments due to their superior scan rate, the scan rate of Orbitrap instruments is too low to allow small DIA mass windows to be used while collecting data across a wide MS^1 range. This is reflected by the fact that of the few DIA metabolomics papers published only one (Zhou et al., 2017) is Orbitrap-based whilst all others have utilised a TOF based system. Furthermore, deconvolution even at narrow window sizes does not appear to be reliable and this manipulation of the raw data introduces an extra element of uncertainty into the untargeted metabolomics process which already suffers from a lack of confidence. Therefore, it is not recommended to pursue this method particularly on an Orbitrap-based instrument. A new deconvolution algorithm, DecoMetDIA (Yin et al., 2019) has very recently been published, this is the first alternative algorithm to MS^2 Dec from MS-DIAL (Tsugawa et al., 2015) that has been published and claims to offer improved deconvolution performance but there has not been time to perform a comparison on the data presented here. Deconvolution performance could also be improved by employing some form of 2D-LC, for example LC-IM-MS (Zheng et al., 2017). This would provide orthogonal information which could be used to aid in deconvolution, however, would also require a new algorithm which is capable of using the secondary information. It may be more beneficial to just employ 2D-LC with a progressive DDA method, the extra separation would dramatically decrease co-elution and thus decrease the number of repeated injections required to acquire MS^2 data for all metabolites.

The extra information that was gathered on low intensity ions using the DIA methods that would not have been acquired with a DDA method is low quality and untrustworthy and although DIA methods can identify high intensity features, this can be easily achieved with less manipulation of the raw data using a traditional DDA method. Efforts should be focused towards collecting reliable DDA data and enhancing quality and coverage of fragmentation for low intensity ions.

5.3 Conclusions

One important goal of untargeted metabolomics is to structurally identify all detected metabolites within a dataset. To meet this goal MS^2 fragmentation data (as well as RT data) is required for as many metabolites as possible but current methodologies have proved insufficient for this task. This work compared different MS^2 strategies on an Orbitrap-based instrument to maximise the amount of *informative* MS^2 data that could be acquired.

DIA and AIF methods promise global coverage and it is true that by using them MS^2 data can be gathered on all features compared to if DDA based methods are used. However, the quality of the MS^2 data is insufficient to allow for confident annotation of low intensity ions even when applying narrow

window sizes. Furthermore, when utilising a narrow window the m/z range covered on an Orbitrap mass analyser is too small to come close to satisfying the m/z ranges typically applied in an untargeted metabolomics experiment. When applying a window wide enough to cover a typically used mass range the annotation performance is poor even for high intensity ions. Therefore, there are no advantages to applying DIA or AIF on an Orbitrap system. This may not be the case on TOF based instruments as they have significantly faster scan rates. DIA could with a window size of 100 m/z or smaller be useful for profiling high intensity metabolites within a sample. Whilst narrow window methods could offer advantages if the deconvolution process could be improved.

With the current Orbitrap technology, DDA methods should be employed with progressive exclusion/inclusion lists to increase the number of metabolites with informative MS^2 data collected and therefore potentially increase the number of metabolite identifications that can be achieved for a sample type. Whilst segmentation of the mass range should be applied to increase signal intensities for lower intensity features. The narrower the segments are that are applied the more good quality fragmentation information should be acquired for low concentration metabolites within the sample. The more segments that are applied however the more time consuming the method becomes so how far this is carried depends on if the goal is routine profiling or deep annotation. Efforts should also be focused towards providing orthogonal data and increased separation with 2D-LC techniques. Also, intelligent acquisition should be implemented so that data is being collected on protonated ions such as can be done with Orbitrap ID-X, this will reduce time wasted on fragmenting features which are derived from other unusual adducts, multiply charged species or oligomers. If routine profiling is being carried out traditional DDA analyses should still be performed to ensure reliability of the resulting data.

6.0 Assessment of Metabolite

Annotation Using AcquireX on the

Orbitrap ID-X

6.1 Introduction

The vital importance of MS² data acquisition for metabolite identification in untargeted metabolomics and the deficiencies of data dependent analysis (DDA) and data independent analysis (DIA) approaches for its collection were highlighted throughout the previous chapter. The data collected demonstrated that on Orbitrap based instruments the implementation of an intelligent data dependent acquisition (iDDA) method for MS² data collection is the best method for acquiring the greatest volume of informative MS² data. This can allow a greater number of identifications and thus more biological knowledge to be gained from any biological study. iDDA methods can employ progressively updating inclusion and/or exclusion lists and repeated injections of the same sample to provide increased MS² coverage for lower intensity ions and biologically important sample components whilst ensuring the same features are not repeatedly fragmented (Neumann et al., 2013). To perform this work manually in a single analytical run is a challenge and not feasible for routine profiling experiments. Recently however, the release of the Orbitrap ID-X mass spectrometer and associated AcquireX software by Thermo Fisher Scientific has changed this. The AcquireX software is capable of performing a user defined number of repeated injections, with a user defined DDA method that automatically updates the exclusion and inclusion lists after each injection. Using this strategy, the user can set up their sequence as they would have done for a traditional DDA method and leave the system to independently and automatically perform iDDA. This opens the door to researchers to routinely perform deeper annotation or identification of their samples.

In an AcquireX sequence a blank sample is injected first and analysed in full scan, this is automatically processed and an exclusion list is generated ensuring that any components of the sample matrix, or contaminants such as plasticisers which are not of biological interest are not fragmented. Following this the sample of interest is injected and analysed in full scan. Automatic, peak picking and grouping is carried out and an inclusion list is generated. Next, a user defined number of sample injections is carried out utilising the exclusion/inclusion lists generated with DDA data collected for each injection. After each sample analysis, the features fragmented are automatically added onto the exclusion list and a further injection is performed. The sequence is finished after the user defined number of sample injections has been carried out. The software not only progressively works through all the features of interest detected but also prioritises fragmentation of the protonated ion as feature grouping is carried out when generating the inclusion list. This ensures time is not wasted collecting MS² data for unusual adducts or multiply charged species which may give rise to false positive results or no identification at all when performing spectral matching against mass spectral libraries which are constructed predominantly with MS² data for protonated and deprotonated ions only (Domingo-Almenara et al., 2018). The method can also be set to “pick others”, this means that features which

are not on the inclusion or exclusion list can still be fragmented if features on the inclusion list are not detected. This new innovative system provides a satisfactory method of MS² acquisition without any major weaknesses as discussed and demonstrated with the other method types (manual iDDA, DDA, DIA, AIF) in the previous chapter.

While this provides a good solution for MS² data collection it does not aid the actual annotation/identification process. A key issue in the field is the development of mass spectral libraries containing MS² data. Mass spectral libraries often contain spectra derived from a number of different chromatographic assays, analyser types, fragmentation mechanisms, fragmentation energies and a host of other MS settings which can impact on the metabolites fragmentation and hence the MS² mass spectrum making many spectra automatically incomparable (Jaeger et al., 2017). Furthermore, the coverage of metabolites present in biology is low in mass spectral libraries and limits identification of metabolites where no chemical standard is available to collect MS² data (Frainay et al., 2018). Many compounds present will also be irrelevant to the sample of interest. Using these libraries also limits the user to at best a level 2 identification (Schymanski et al., 2014). Libraries must be carefully curated and users need to be aware of what type of metabolites are present and not present, is the library being utilised too broad in its chemical space coverage or is it too narrow? These are just some of the reasons why MS² spectral libraries are insufficient in their current state although they do provide highly valuable resources. To move untargeted metabolomics into being a totally robust science, true level 1 confidence in identification must be provided more regularly. This requires experimental spectra and chromatographic retention times to be matched against spectra and retention times acquired using a pure authentic standard under the same experimental conditions, on the same liquid chromatograph and mass spectrometer, in the same laboratory (Schymanski et al., 2014).

The acquisition of pure authentic standards is expensive, many metabolites do not have standards commercially available and building a large custom library from standards is time consuming (Dunn et al., 2013). However, for well-established metabolomics laboratories with standard assays that expect to be performing a high number of routine analyses it makes sense for custom libraries containing MS² mass spectra and retention times to be constructed. In this chapter it will be described how a custom library was built using a standardised aqueous C₁₈ reversed-phase (from here-on referred to as RP) method and using a standardised hydrophilic interaction chromatography (HILIC) assay; both assays were developed and validated in Phenome Centre Birmingham. The MetaSci COMPLETE human metabolite standard library was used as the source of 875 metabolites. This work was performed on an Orbitrap ID-X system in San Jose, USA with the assistance of Dr Ioanna Ntai who is a Thermo Fisher Scientific employee and another PhD student, Mr Elliott Palmer. The same system and assays were

subsequently used for the analysis of three samples (NIST 1950 Plasma SRM (standard reference material), NIST smokers urine SRM and NIST non-smokers urine SRM) to perform deep annotation of these common sample types to demonstrate the advantage of the AcquireX method over a traditional DDA approach.

6.2 Results and Discussion

6.2.1 Number of Metabolites Detected and mzVault Library Contents

A number of different categories were assigned for each metabolite standard as the library was constructed:

- The total number of metabolites detected.
- The number of metabolites with high quality MS² mass spectra and retention times added to the mzVault library.
- The number of metabolites with MS² spectra deemed unsuitable for addition to the library.
- The number of metabolites detected but with no MS² spectra acquired.
- The number of metabolites incorrectly designated due to charge.

These data are displayed in Table 49 for positive ion mode and Table 50 for negative ion mode. The number of metabolites with suitable quality MS² mass spectra for addition to the library was at roughly half of the total number (875) analysed for each assay and ion mode combination. This resulted in 546 different metabolites with MS² mass spectra in positive ion mode (Table 51) and 564 different metabolites in the negative ion mode (Table 52) in the MS² library. This was roughly as expected, as the metabolites represented a range of different physicochemical characteristics and were prepared and analysed in groups, not individually. The physicochemical characteristics of a metabolite may have resulted in it not being retained on the column and thus co-eluted with many other metabolites and was subsequently not detected due to ion suppression. Or due to the nature of the groups in sample preparation metabolites may have reacted with other metabolites in the mixture and thus were no longer present for detection. Another possibility is that the metabolite was detected but errors were made during the data processing which resulted in it being misidentified or missed. The details of which metabolites were detected can be found in electronic Appendix (2.4/2.4.1).

Table 49: Summary of the number of metabolite standards detected in positive ion mode.

Type of detection	RP	HILIC
MS ² mass spectra added to library	493	414
MS ² mass spectra collected but not added to library	67	7
Metabolite was detected in full-scan mode but no MS ² data were acquired	118	92
Metabolite was detected, MS ² data were acquired but the correct molecular structure was not assigned due to it being a charged metabolite	0	2
Total Detected	678	515

Table 50: Summary of the number of metabolite standards detected in negative ion mode.

Type of detection	RP	HILIC
MS ² mass spectra added to library	440	500
MS ² mass spectra collected but not added to library	46	50
Metabolite was detected in full-scan mode but no MS ² data were acquired	84	76
Metabolite was detected, MS ² data were acquired but the correct molecular structure was not assigned due to it being a charged metabolite	0	2
Total Detected	570	628

Table 51: The number of different metabolite standards with HCD MS² spectra added to the mzVault Library in positive ion mode.

Assay	Number of Different Metabolite Standards Added to Library
RP	493
HILIC	414
Combined	546

Table 52: The number of different metabolite standards with HCD MS² spectra added to the mzVault Library in negative ion mode.

Assay	Number of Different Metabolite Standards Added to Library
RP	440
HILIC	500
Combined	564

6.2.2 Advantage of AcquireX vs Traditional DDA

Different assay/sample type/ion mode combinations produced datasets of varying complexity. A high and low complexity example have been selected and their characteristics will be discussed in turn in the following section. Data for the remaining datasets will be presented in the electronic Appendix (6.2/6.2.2).

6.2.2.1 High Complexity Biological Sample

The dataset collected for RP smokers urine applying a RP positive ion mode assay was selected to demonstrate the difference in performance between the iDDA AcquireX method and a traditional DDA method. This dataset had the greatest number of compounds detected after processing in CD3.0 (Figure 90A). Therefore, it should be suitable to demonstrate the advantages of the AcquireX method for analysis of high complexity biological samples. The number of features will not be discussed as in CD3.0 the data is displayed as compounds which are defined as groups of features which have been assigned to a single molecular ion group and thus are different grouped adducts, fragments, oligomers or other derivative features of a single metabolite. 5,471 compounds were detected with the AcquireX iDDA method and 5,288 were detected with the DDA method. A similar number of features being detected indicates that the peak picking that has occurred in the two methods has performed similarly despite the AcquireX method peak picking being carried out on a single full scan file compared to the DDA method being carried out on all three samples where DDA data were acquired. The fact that the AcquireX method detected slightly more compounds could indicate that it is less stringent as there can be no filtering based on the percentage of files a feature has been detected in and features cannot be removed through alignment of multiple replicates as would normally be done. However, it may also be a result of the decreased full scan rate of the DDA method as they have a decreased duty cycle and number of data points detected for each chromatographic peak. In reality it is probably a combination of both factors.

The main advantage of performing iDDA such as demonstrated with the AcquireX method is increasing the number of features/compounds which ultimately have MS² data acquired. This effect is clearly demonstrated in Figure 90B where the number of compounds which have MS² data is nearly 4 times higher when applying the AcquireX method compared to DDA (4,331 and 1,142 respectively). This is a major and highly valuable increase. However, does this result in more identifications is the most important question. Also, the extra time, money and sample volume used to generate this data must also be considered. Whether this was efficient will be revealed by the number of identified metabolites but will also depend on the task that the user was hoping to carry out. Is this routine profiling or deep annotation?

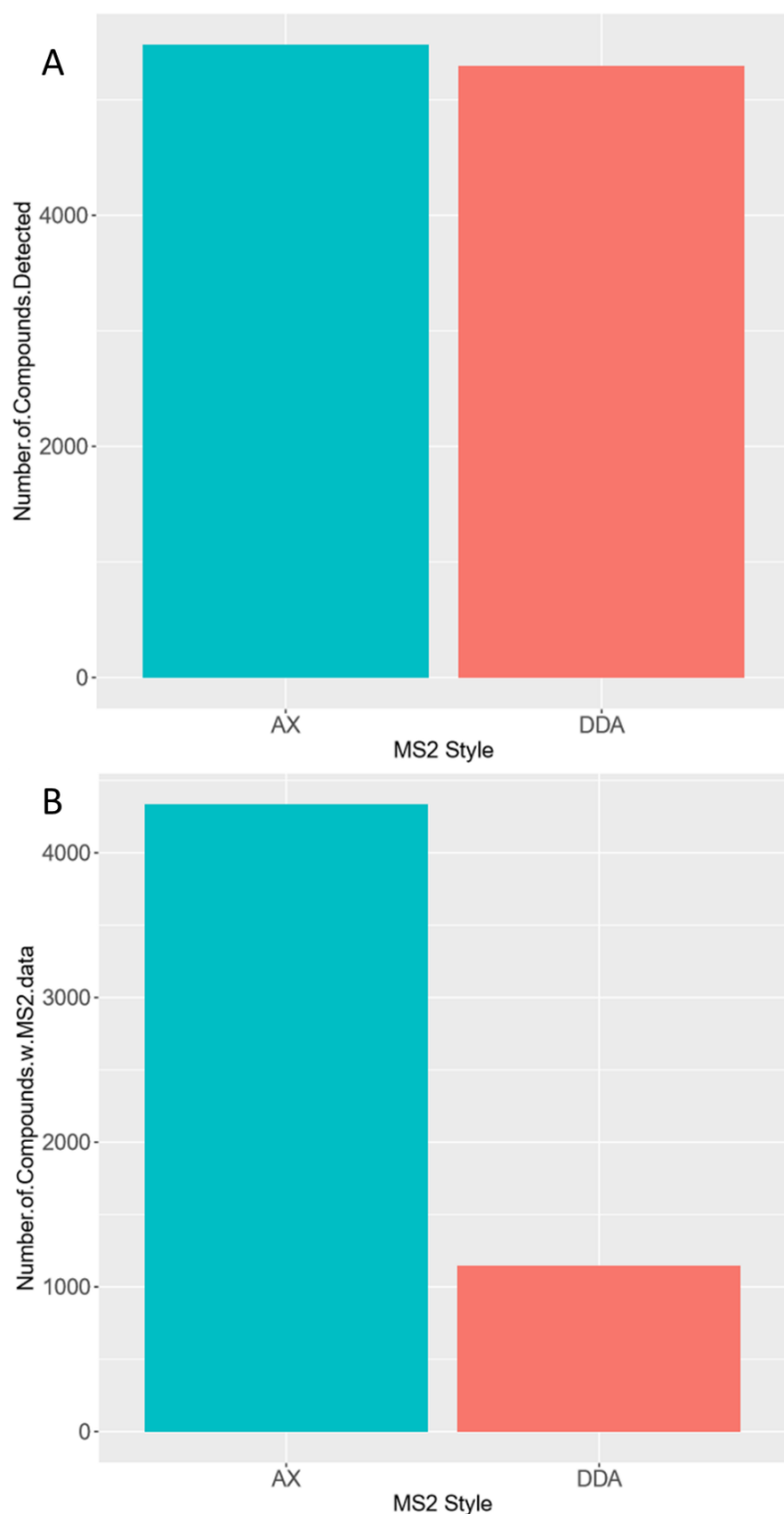


Figure 90: Comparison of AcquireX (AX) and DDA methods for RP/positive ion mode/smokers urine after processing in CD3.0 A) Number of compounds detected. B) Number of compounds with MS² data.

An important consideration and another of the advantages provided by an iDDA method such as the AcquireX method is the preferential acquisition of MS² data for the protonated ion. This is desirable

as the majority of spectra available in MS² databases are for the protonated ion for positive ion mode (or deprotonated ion in negative ion mode) and so collecting MS² spectra for other adduct types will increase the likelihood of no match or a false positive match. This advantage is clearly demonstrated in Figure 91 where part A displays the number of compounds with MS² data for the “preferred ion” which can be set as one or more adducts of the users choice in CD3.0. In this case the preferred ion was set as the protonated ion in positive ion mode. The AcquireX method collected MS² spectra for 4,187 compounds over the 6 progressive injections whilst the DDA method only achieved this for 852 compounds. AcquireX achieved nearly a 5-fold increase over traditional DDA. Furthermore, part B of Figure 91 shows the number of compounds with MS² spectra acquired for “another ion”. This means MS² spectra acquired for any adduct that was not the protonated ion. The DDA method has slightly over twice as many spectra for other adduct types than the AcquireX method with 290 and 144 respectively. This equates as a percentage to 25.4% of all MS² spectra acquired with DDA being for non-protonated ions and therefore 25.4% of the 4 times lower amount of MS² spectra acquired are of decreased likelihood of generating a true positive match when MS² spectral matching is performed. This is compared to just 3.3% for the AcquireX method. This clearly demonstrates the intelligence of the selection being performed during the AcquireX sequence.

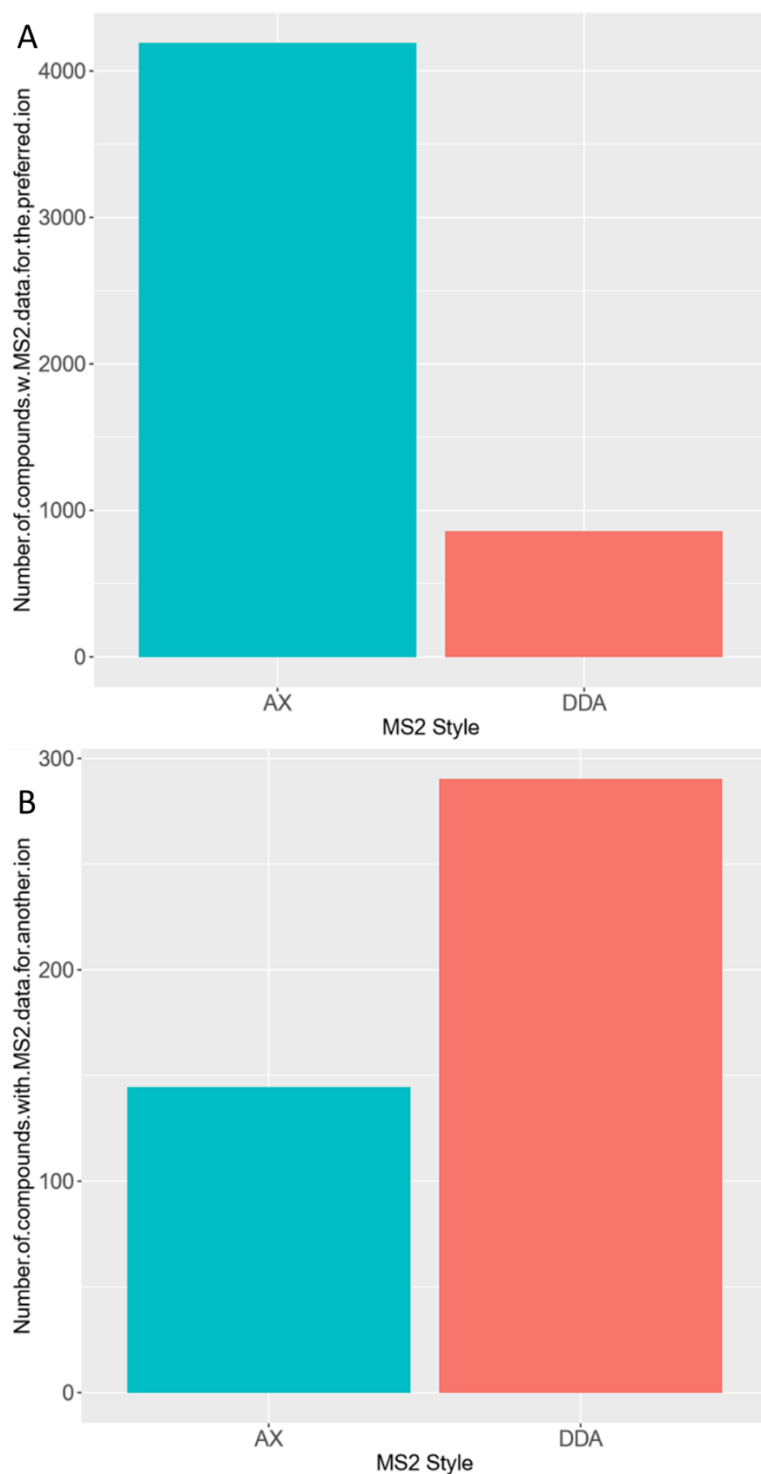


Figure 91: Comparison of AcquireX method (AX) and DDA method for RP/positive ion mode/smokers urine after processing in CD3.0 A) Number of compounds with MS² data collected for the preferred ion (M + H)⁺ B) Number of compounds with MS² data collected for another ion.

Collecting more MS² data in general and particularly for protonated ions is a good advantage, but how did those extra MS² mass spectra translate into spectral matches and identifications? The number of level 1 and level 2 identifications had been determined as described in section 2.4.7 and are displayed

in Figure 92A and B. The AcquireX method only slightly outperformed the DDA method with 55 and 46 respectively equating to a modest 19.6% increase. However, this is perhaps not surprising when the overall size of the library is considered with only 544 different metabolites present in the positive ion mode custom library. On the other hand, the library is constructed from the MetaSci COMPLETE Human metabolite standards kit and the sample is human urine and so perhaps only getting matches to roughly 10% of metabolite standards in the library is disappointing. Increasing the size of the library by comparing the number of level 2 identifications shows the significant advantages of the AcquireX method over the traditional DDA method. Level 2 identifications were determined using mzCloud. The AcquireX method achieved 198 level 2 identifications compared to 83 with the DDA method. A greater than 2-fold increase is not surprising and is perhaps disappointing considering 4-fold advantage in the number of MS² mass spectra collected and 5-fold increase in the number of MS² mass spectra for the protonated ion. This is likely indicative of the decreasing quality of MS² mass spectra that are acquired as the number of repeated injections increases and the intensity of the features being fragmented gradually decreases and thus so does the purity and the likelihood of a high quality spectral match. However, the doubling of level 2 identifications is still very beneficial and demonstrates the superior performance of the AcquireX method. Are these advantages consistent for lower complexity datasets though?

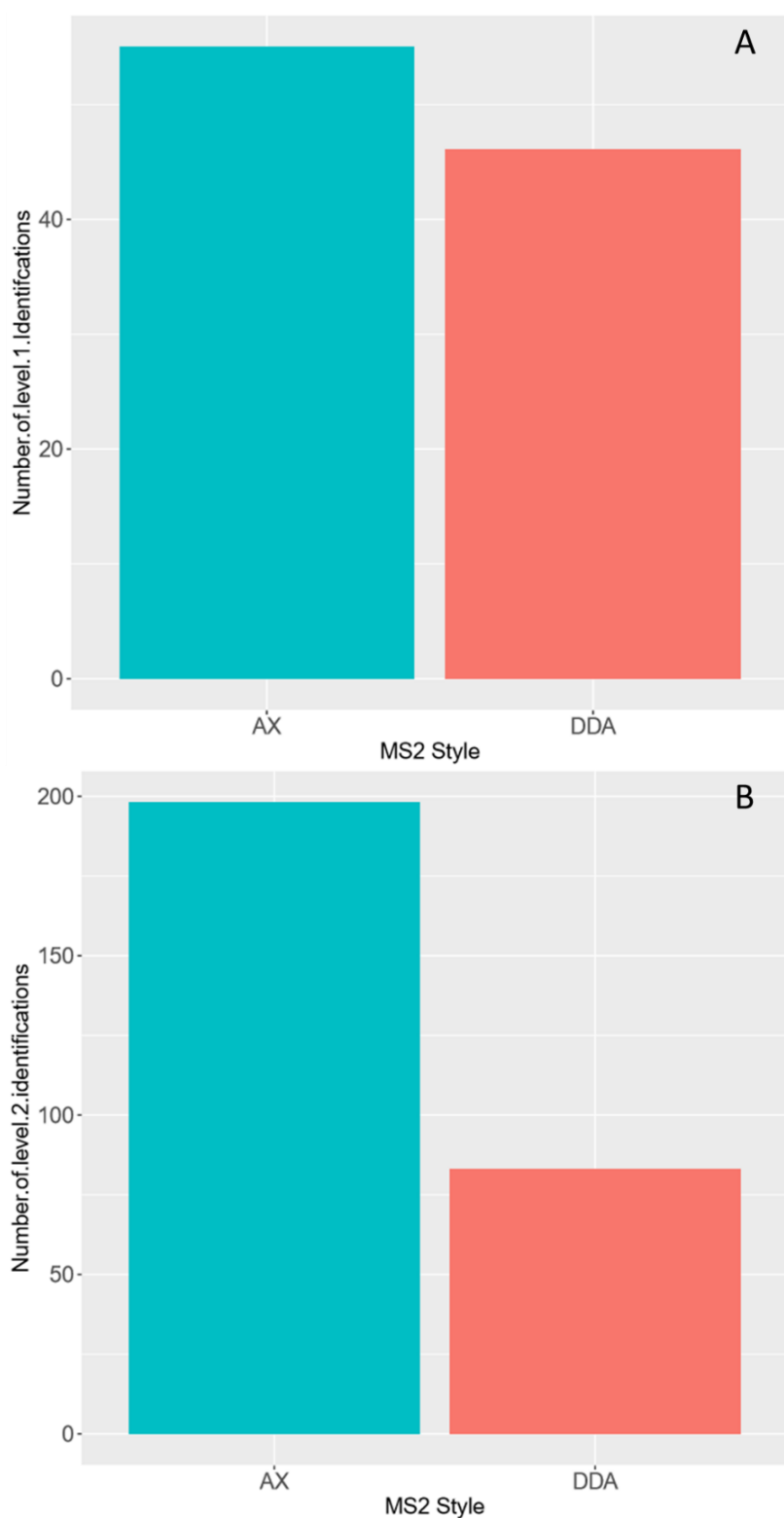


Figure 92: Comparison of AcquireX (AX) and DDA methods for RP/positive ion mode/smokers urine after processing in CD3.0 A) The number of compounds with a level 1 identification B) Number of compounds with a level 2 identification.

The number of compounds with mzVault best match scores \geq (0, 10, 20, 30, 40, 50, 60, 70, 80 and 90) is displayed in Figure 93. The AcquireX method has more compounds than DDA at all thresholds except ≥ 90 . This shows that although the number of compounds with MS² mass spectra is much higher with AcquireX, the majority of these extra compounds do not produce high quality matches. Although as discussed earlier the number of compounds in the mzVault library is relatively small compared to the number of metabolites present in typical MS² spectral libraries being in the thousands. mzCloud is an example of one of these and the same data but for mzCloud annotations is displayed in Figure 94.

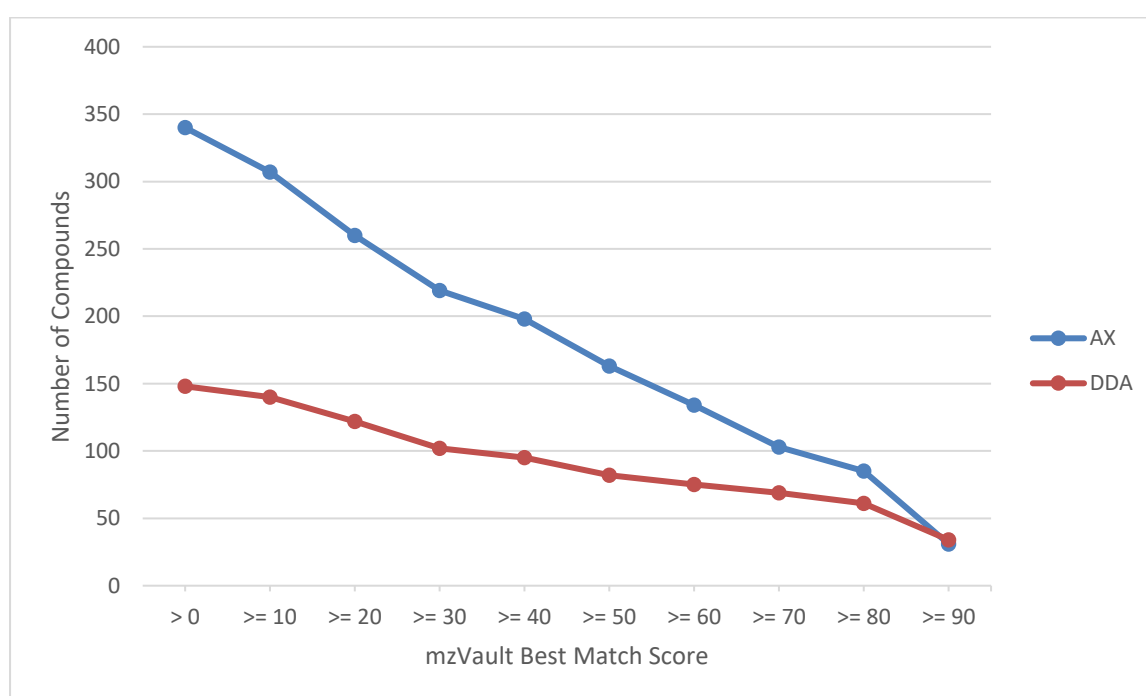


Figure 93: The number of compounds with an mzVault best match score \geq (0, 10, 20, 30, 40, 50, 60, 70, 80 and 90) for AcquireX and traditional DDA methods in the RP/smokers urine/positive ion mode dataset after processing in CD3.0.

The number of spectra with lower quality matches \geq (30, 40, 50, 60) is a lot higher when applying the AcquireX method compared to DDA. Whilst the number of very high quality matches (≥ 90) is still nearly equal. High quality matches \geq (70, 80) are superior when using the AcquireX method. This is perhaps indicative of the difficulty of acquiring high quality spectra for lower intensity ions. Other factors however may also be lack of coverage in MS² library. This highlights the importance of the further development and improvement of in-silico fragmentation tools/libraries. This is the same trend as seen with mzVault matches (Figure 93) which shows that if the data complexity is high enough the number of these spectra with good quality matches (likely due to higher intensity) is high enough

to justify the extra injections and thus time, money and sample spent when applying the AcquireX method although a large portion of time is still spent collecting spectra of poor quality or that do not have appropriate reference spectra available.

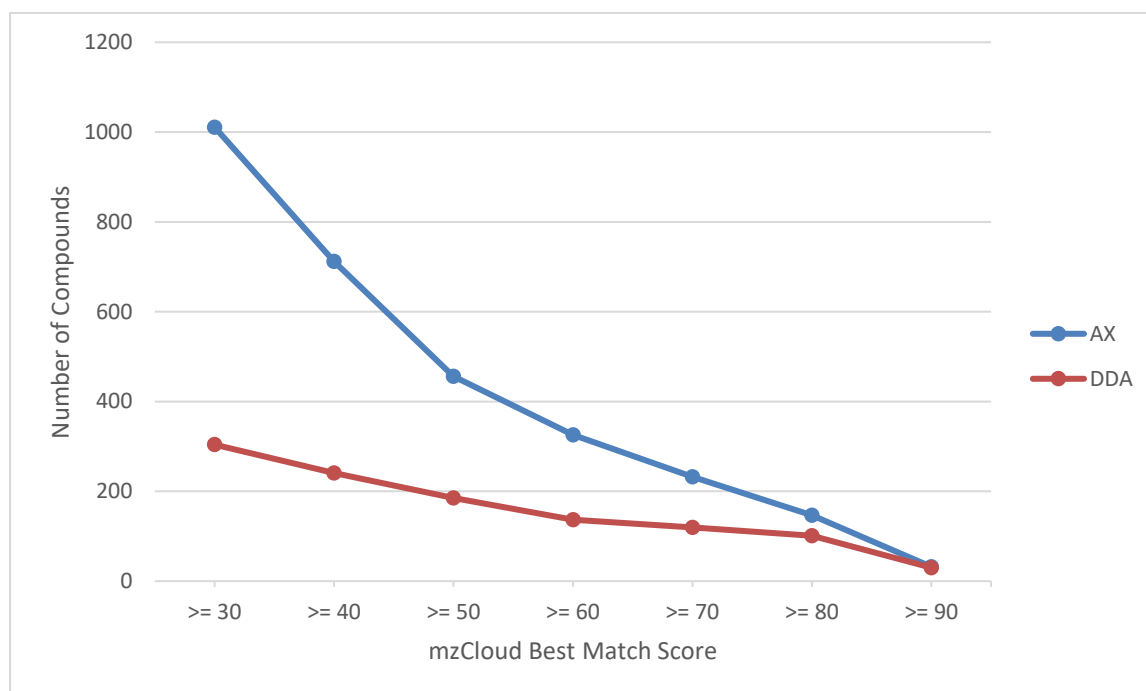


Figure 94: The number of compounds with an mzCloud best match score \geq (30, 40, 50, 60, 70, 80 and 90) for AcquireX and traditional DDA methods in the RP/smokers urine/positive ion mode dataset after processing in CD3.0.

6.2.2.2 Lower Complexity Biological Sample

The smokers urine sample analysed applying the HILIC positive ion mode assay was selected as an example of a lower complexity dataset. Very similar trends were seen across all of the HILIC methods applied. The same categories discussed for the high complexity data (6.2.2.1) are all displayed in Table 53 alongside the fold change when using the AcquireX method compared to DDA.

Table 53: A comparison of AcquireX and traditional DDA from the HILIC/smokers urine/positive ion mode dataset based on CD3.0 data processing. The fold change provided by utilising AcquireX in comparison to DDA is displayed.

MS ² Style	AX	DDA	AX/DDA Fold Change
Number of Compounds Detected	2059	1870	1.10
Number of Compounds w/MS ² data	1656	948	1.75
Number of compounds w/MS ² data for the preferred ion	1614	873	1.85
Number of compounds with MS ² data for another ion	42	75	0.56
Number of level 1 Identifications	53	50	1.06
Number of level 2 Identifications	140	129	1.09

The number of compounds detected was roughly 1/3rd of the total seen with the high complexity dataset in section 6.2.2.1 with once again a relatively small difference in the number of compounds detected between the AcquireX method and DDA. Similar trends to those seen with the high complexity dataset are observed for the number of compounds with MS² data, the number of compounds with MS² data for the preferred ion and the number of compounds with MS² data for another ion. AcquireX nearly doubled the number of compounds with MS² data and the number of compounds with MS² data for the preferred ion. In the high complexity data these increases were nearly 4 and 5 fold respectively. So, it is clear that the advantage of using iDDA is not as great with a lower complexity dataset however the advantage is still there. When the number of identifications are considered however we see almost no advantage to using AcquireX. There were 3 more level 1 and 11 more level 2 identifications when applying AcquireX. Whilst this is still useful it is not significantly higher than the DDA method. Considering the number of injections used to acquire the AcquireX sequence in this case it was unlikely that this extra data were worth the time taken. The higher number of MS² spectra being acquired in this case when applying the AcquireX method for the most part have not been annotated. This can be for two reasons, either the spectrum is too noisy and of poor spectral quality, or there is no suitable mass spectrum in the MS² spectral library. This in itself could be for a variety of reasons but in general highlights the insufficiency of MS² spectral databases. Poor quality of the extra MS² mass spectra acquired is evidenced by the data displayed in Figure 95 and Figure 96. The number of compounds with high (≥ 70 , 80) or very high quality (≥ 90) spectral

matches to mzVault and mzCloud are not significantly higher when applying the AcquireX method. The number of compounds with low quality spectral matches ($\geq 30, 40, 50, 60$) are greater but not by very large amounts when compared to the difference in number of MS² spectra acquired. Further highlighting the assumed decreasing purity and thus quality of the extra spectra acquired with the AcquireX method.

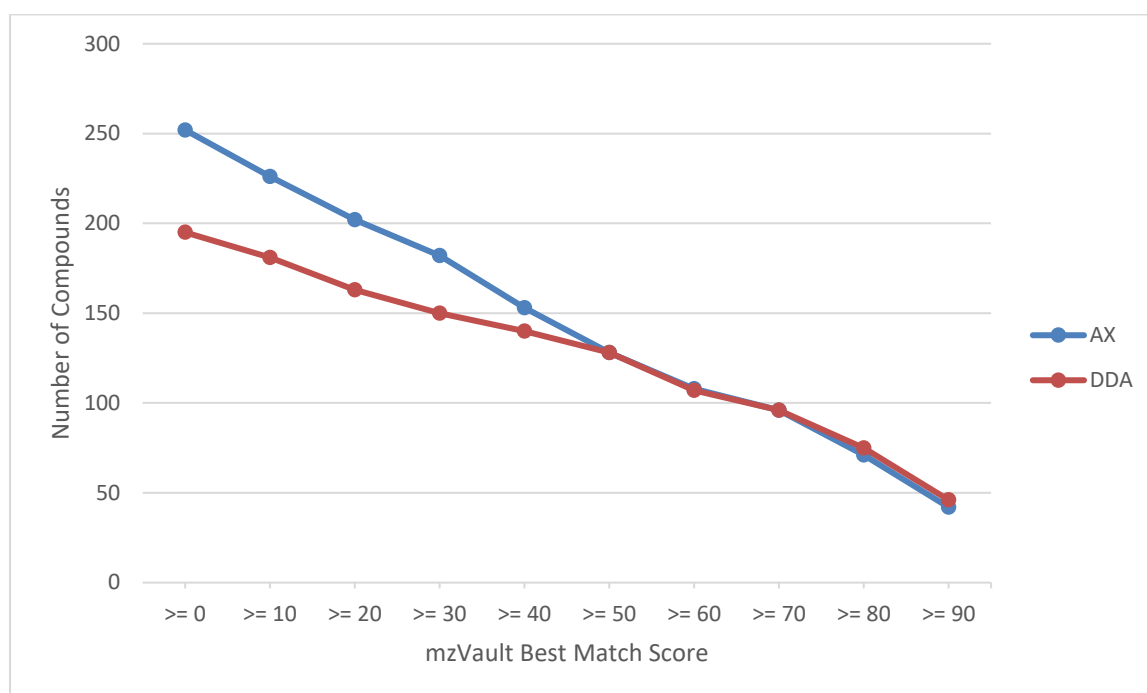


Figure 95: The number of compounds with an mzVault best match score \geq (0, 10, 20, 30, 40, 50, 60, 70, 80 and 90) for AcquireX and traditional DDA methods in the HILIC/smokers urine/positive ion mode dataset after processing in CD3.0.

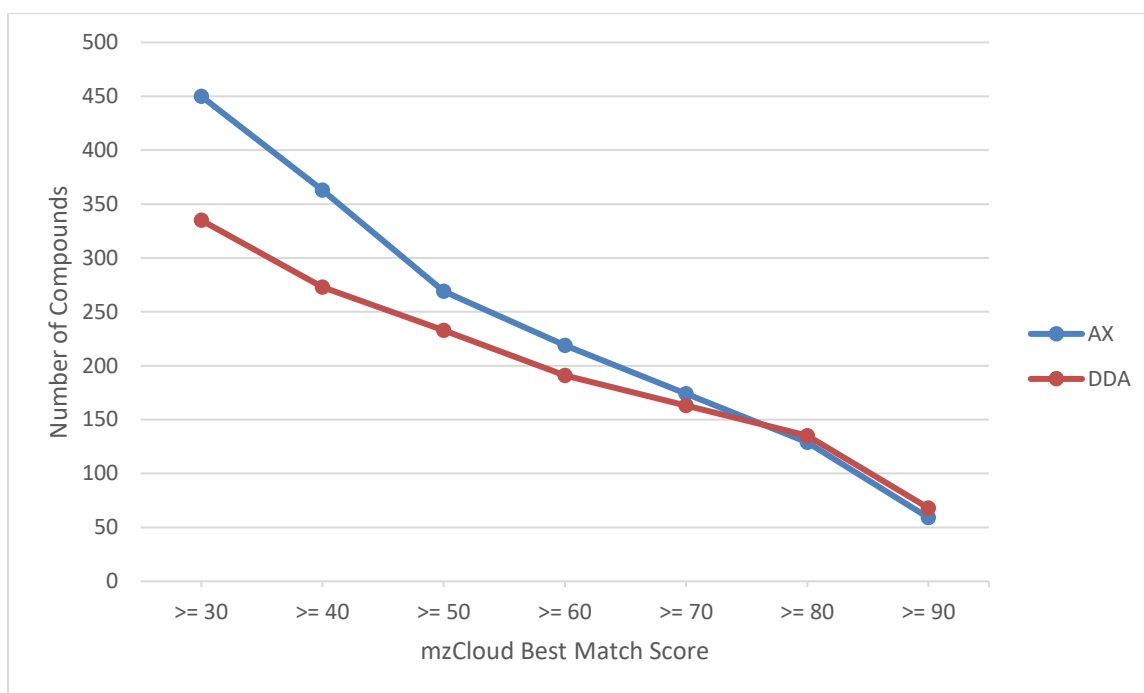


Figure 96: The number of compounds with an mzCloud best match score \geq (30, 40, 50, 60, 70, 80 and 90) for AcquireX and traditional DDA methods in the HILIC/smokers urine/positive ion mode dataset after processing in CD3.0.

Overall the data in this section demonstrates that the AcquireX method is valuable if you expect your method to generate data with a high number of compounds detected. However, if the feature/compound number expected is low, the advantages gained are likely to be small and the extra time, money and sample utilised may not be appropriate. Ultimately this is up to the users discretion, as for example, the extra 10 level 1 identifications provided by AcquireX could be deemed highly valuable or not so valuable depending on the desired research goal.

6.2.3 How Many Repeated Injections Are Required?

The last section demonstrated the advantage of the AcquireX method. However, an important question to be answered is how many repeated injections are required? This was assessed by investigating the number of features on the automatically updated inclusion/exclusion lists after each repeated injection in the AX sequence, as well as the purity of spectra collected and the intensity of precursors fragmented.

6.2.3.1 Length of Progressive Exclusion/Inclusion Lists

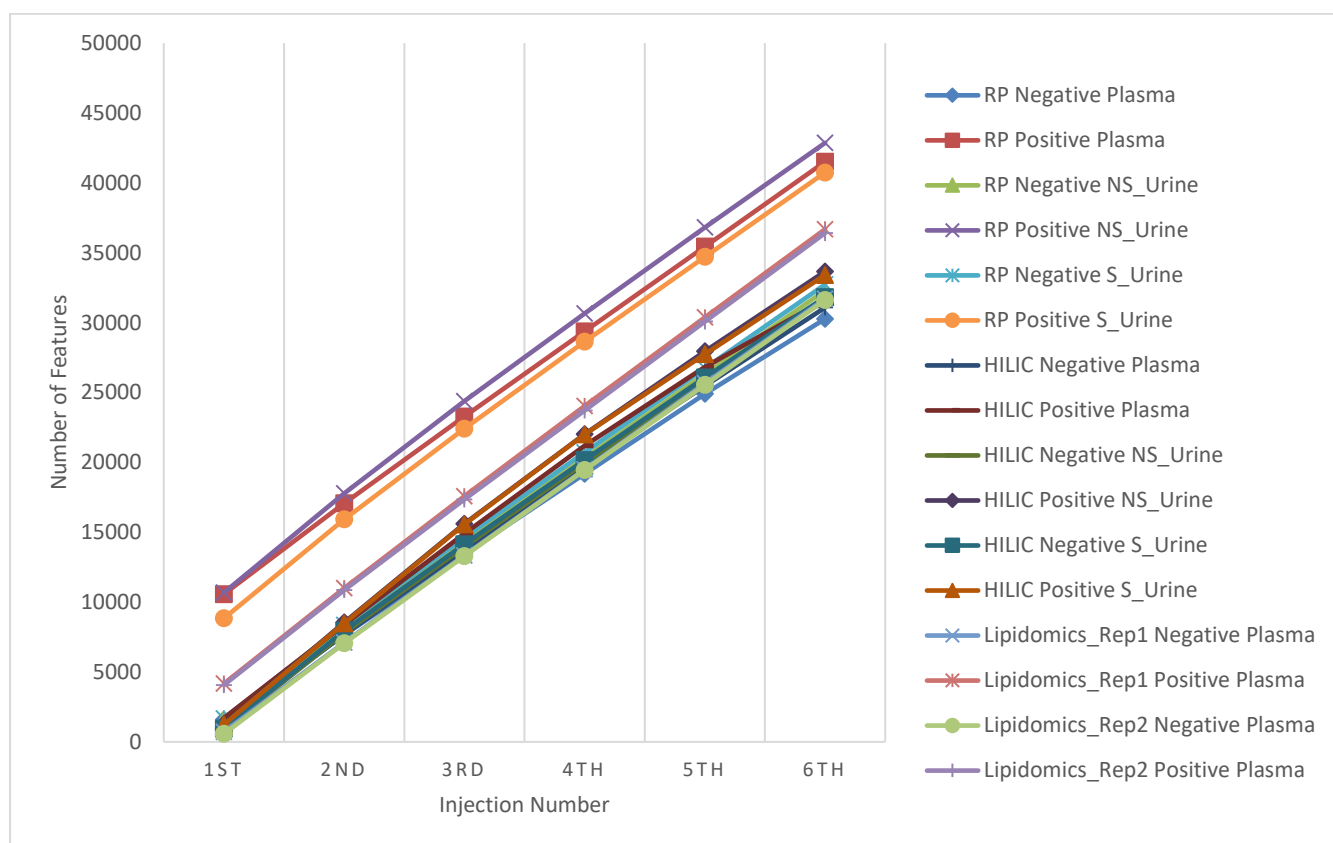


Figure 97: The number of features saved on the automatically updated exclusion list for each injection of the AcquireX sequence for all assay/sample type/ion mode combinations.

No matter what assay, sample type and ion mode combination was applied, the trend was the same (Figure 97) regardless of the overall dataset complexity as the instrument was set to “pick others”. Whilst not fragmenting features contained on the inclusion list it continued to fragment other features at a nearly constant rate, these must have been lower intensity noisy features or features which should have been on the inclusion list and were not detected in the full scan at the beginning of the AcquireX sequence. There are some fluctuations but overall the increase in list length was roughly consistent from injection to injection.

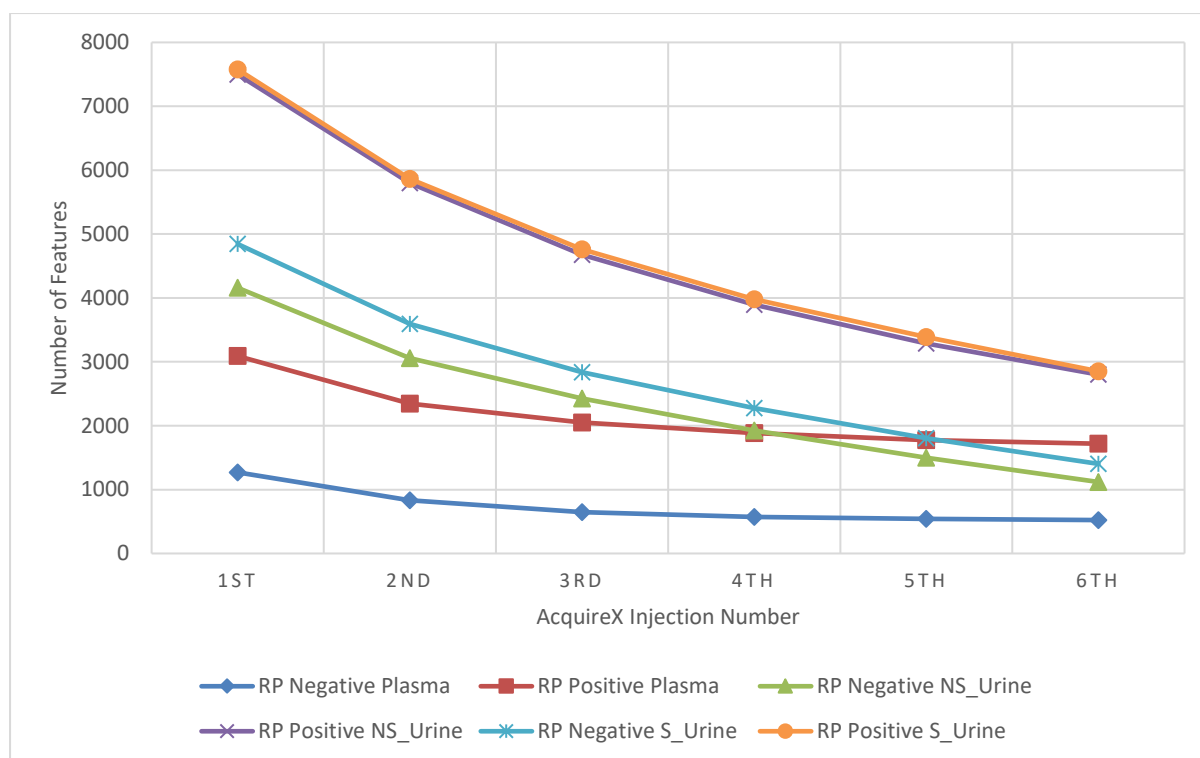


Figure 98: The number of features saved on the automatically updated inclusion list for each injection of the AcquireX sequence for all RP/sample type/ion mode combinations.

Figure 98 shows the decrease in the number of features on the inclusion list from injection to injection for all the RP datasets. This clearly shows the impact of the sample complexity on the number of injections that may be valuable. RP smokers urine was the most complex dataset in any assay and sample type combination across both ion modes. A high number of features (~7,500) were present on the initial inclusion list and this ensures that the rate of progress through the list during the method although steadily decreasing never plateaus or becomes zero. A significant number of features (538) from the inclusion list were fragmented during the fifth injection and were thus added to the 6th exclusion list when applying RP smokers urine in positive ion mode. In contrast, RP/plasma/negative ion mode data were far less complex. Only 1,269 features were included on the initial inclusion list and as a result the inclusion list length plateaus after the third injection. Interestingly, none of the methods inclusion lists appear like they would ever reach zero. This could be because some features on the list are of an intensity below the threshold required to trigger fragmentation but it also highlights an imperfection in the design of the AcquireX sequence in the fact that the inclusion and exclusion lists are generated through processing one full scan sample. Without replication the likelihood of noisy features being included on the inclusion list is more likely and so perhaps these remaining features on the list represent errors from the data processing step. Alternatively, some of the features still remaining on the list could be present in areas of high co-elution and as a result

require further injections to be selected for MS² DDA fragmentation. If fragmented however, due to being in an area of high co-elution they would likely produce spectra of low purity. In order to get good fragmentation for these features superior separation is required, whether this be through improvement of the current chromatographic method or through addition of an additional separation method thus turning it into a 2D-LC method.

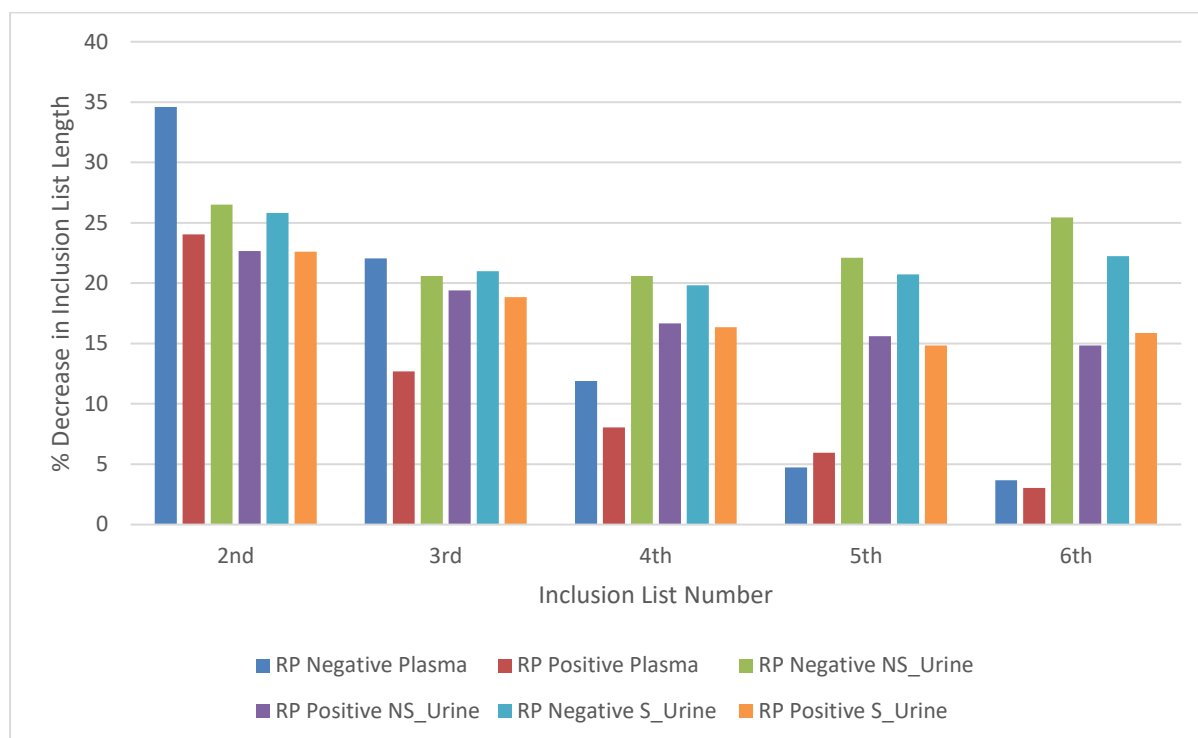


Figure 99: The percentage decrease in length of the inclusion list compared to the length of the previous inclusion list on the method for each updated AcquireX injection for all RP/sample type/ion mode combinations.

Figure 99 displays the same data as in Figure 98 but shows the percentage decrease in the length of the inclusion list on each injection in comparison to the length of the previous inclusion list. For the urine methods at least 15% of the features are fragmented in every injection except the 4th injection of the positive/smokers urine method and the 5th injection of the positive non-smokers urine method. For plasma less than 10% are fragmented from the 3rd or 4th injection. The urine rate of decrease stays fairly consistent. This highlights the importance of the sample type again in the number of valuable injections.

The assay/sample type/ion mode determine the sample complexity and this in turn determines the number of injections which are valuable in an AcquireX sequence. The data for the remaining datasets are not discussed and can be found in the electronic Appendix (6.2/6.2.3.1).

6.2.3.2 MS² Spectral Purity and Precursor Intensities

The purity of MS² mass spectra acquired through each injection of the AcquireX sequence was investigated to determine how the quality of spectra collected throughout the sequence changes for the different assay/sample type/ion mode combinations. This will help to inform the number of repeated injections which are suitable. Comparison between iDDA and DDA data were also performed. Any data not included in this section can be found in the electronic Appendix (6.2/6.2.3.2).

Purity scores are dependent on a number of different factors. These are the method applied (AcquireX or DDA) and the complexity/density of the data which is defined by the assay, sample type and ion mode combination that was applied. The interpolated purity tended to decrease through each injection of the AcquireX sequence as can be seen in Figure 100A. It can also be seen in Figure 100B how much higher the purity is when applying a DDA method which was true across all datasets. This is not necessarily a good thing, as no blank exclusion list has been applied to these methods so all the highest intensity features are being fragmented. Many of these features could be components of the solvent matrix or represent multiple features of a single metabolite and so time spent collecting these will have been wasted. From the first AcquireX injection the purity distribution is much lower than in any of the DDA injections. Therefore, most of these features have likely been included on the exclusion from the blank processing. Or alternatively they might be more evenly spread throughout the AcquireX sequence due to co-elution. This was a relatively high complexity dataset, with 3,457 compounds detected. However, the separation appears to be of good quality with a number of different peaks clearly visible and separable across the chromatogram (Figure 101) helping to ensure the median purity never dropped below 50%. A good number of high intensity features are still being fragmented during the later injections with each injection having many spectra of 100% purity with the first injection having only 89 more than the last (Table 54). This could be due to the fact that the chromatographic separation appears to be good and therefore there is reduced co-elution of compounds meaning that despite the decreasing intensity of the features being fragmented (Figure 102) the purity scores are not very badly affected.

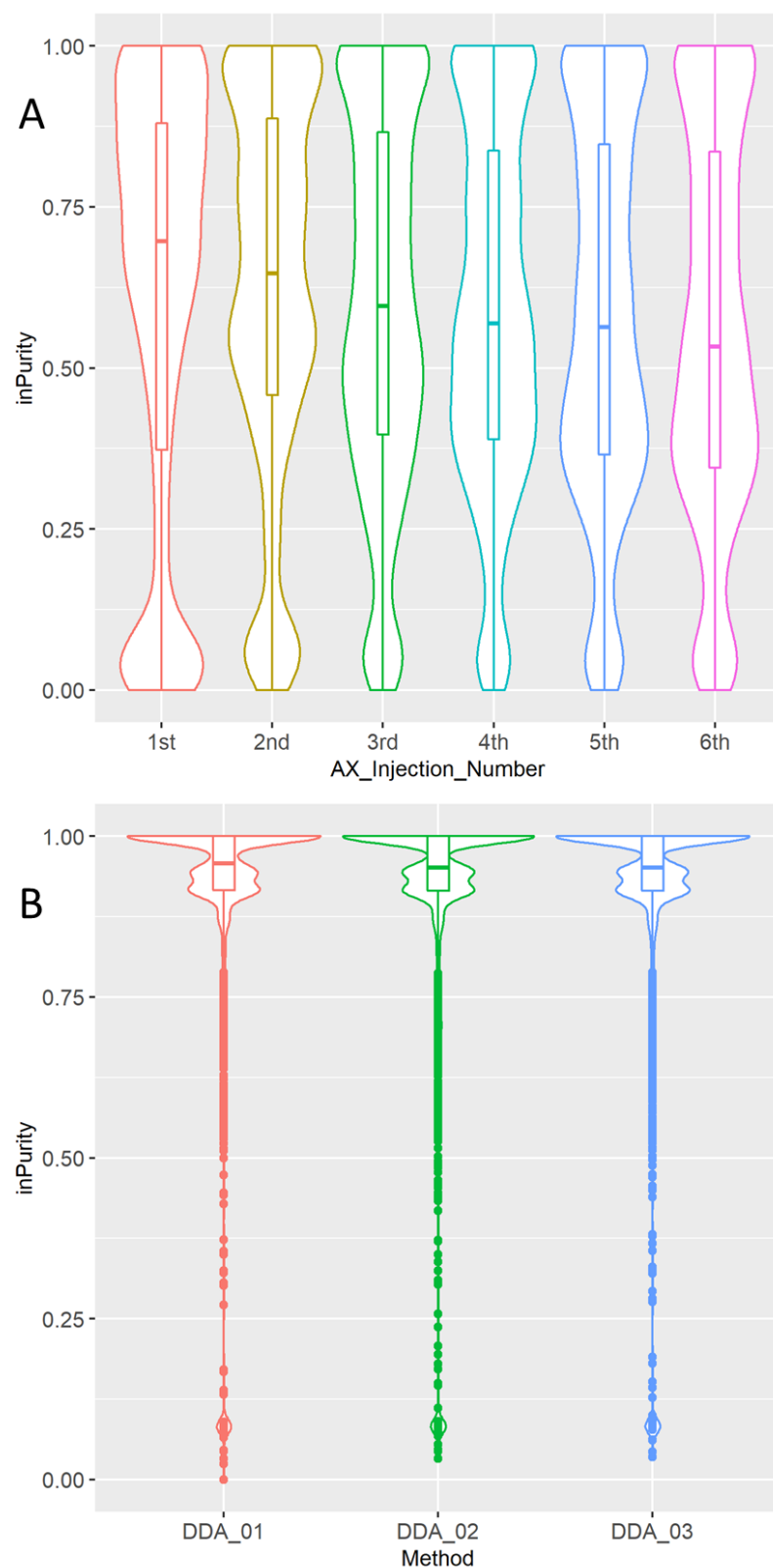


Figure 100: The interpolated purity scores for all MS² spectra acquired between 30 and 780 seconds in the RP/Non-smokers Urine/Negative ion mode dataset A) Scores for each injection of the AcquireX sequence. B) Scores for the 3 traditional DDA injections.

Table 54: The number of MS² spectra with a purity score of 100 % for each injection during the AcquireX sequence in the RP/Non-Smokers Urine/Negative ion mode dataset.

Dataset	1st	2nd	3rd	4th	5th	6th
RP/Non-Smokers-Urine/Negative	1214	1101	992	1040	1008	1125

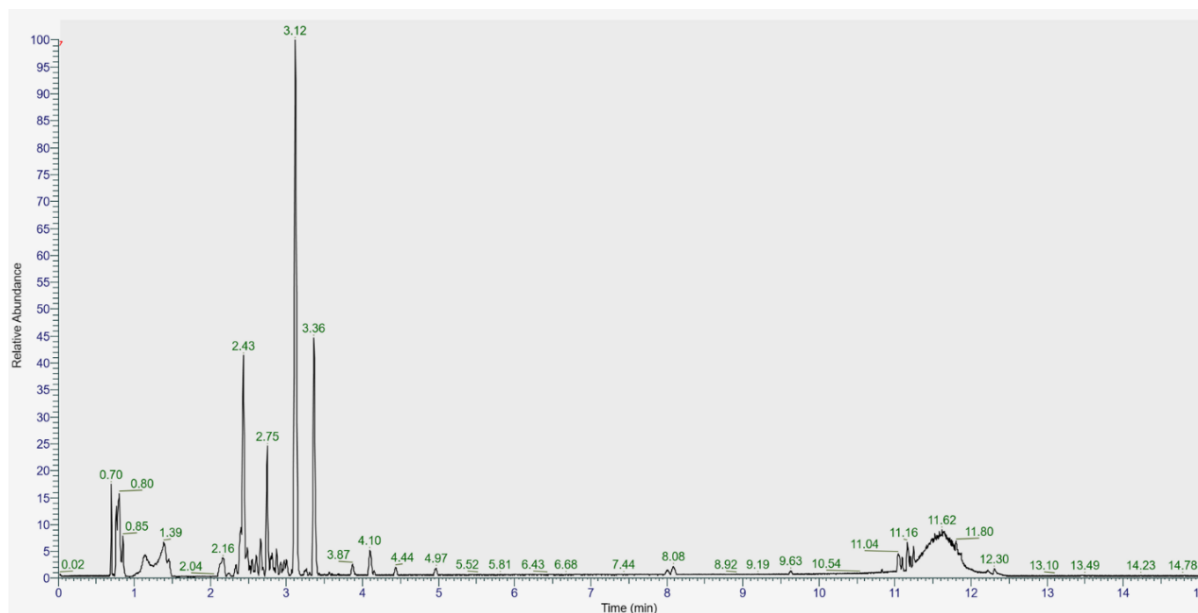


Figure 101: Base Peak Chromatogram (BPC) for the RP/Non-Smokers Urine/Negative ion mode full scan acquisition from the AcquireX sequence.

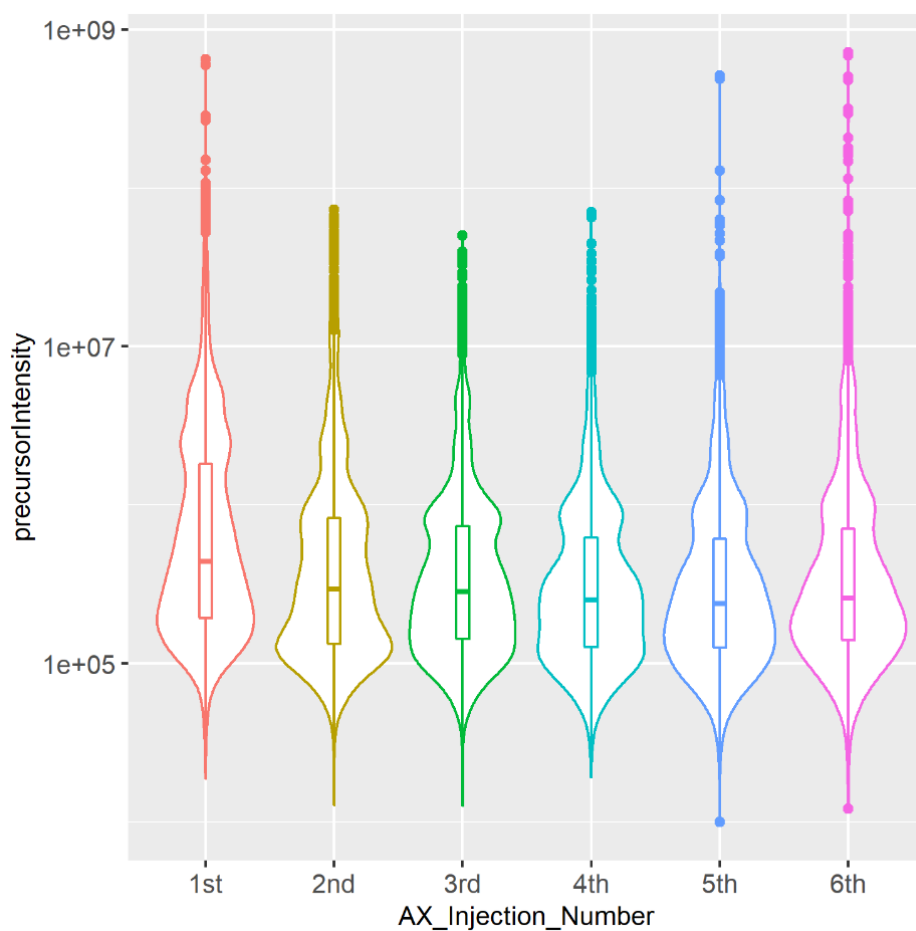


Figure 102: The distribution of precursor intensities acquired between 30 and 780 seconds with log10 transformed y axis through each injection of the AcquireX sequence for the RP/Non-Smokers Urine/Negative ion mode dataset

If a different sample type and ion mode is applied and the complexity of the dataset is lower and the quality of separation is lower, a different story can be seen. 1,757 compounds were detected in the AcquireX full scan acquisition for the RP/Plasma/Positive ion mode dataset. This decreased complexity of features alongside the decreased quality of separation (Figure 104) resulted in a distribution of lower purity scores across every injection (Figure 103). The median purity of the first injection was roughly equal to the median purity of the 6th injection of the RP/Non-Smokers Urine/Negative ion mode dataset (Figure 100A). The median purity then rapidly drops to around 32% by the third injection. Interestingly the purity distribution then increases again over the next two injections. This is likely due a number of low intensity features being fragmented in the second and third injections as the number of higher intensity features on the inclusion list decreases. During the 4th and 5th injections when the purity increases again the instrument is probably “picking others” where higher intensity features that were not on the inclusion or exclusion list are now being fragmented more regularly. The purity then falls in the 6th injection to roughly the level seen in the 3rd. Despite the lower average purity of spectra

being acquired there are still a high number of spectra being acquired of high purity. The number of spectra acquired in each injection with purity of 100% (Table 55) are not much lower than the numbers seen for the RP/Non-Smokers Urine/Negative ion mode dataset (Table 54) and also stay fairly constant throughout the sequence. It seems then that even with the lower average quality there are enough high quality spectra still being acquired to make six injections potentially valuable as it appeared to be for the RP/Non-Smokers Urine/Negative ion mode dataset. This is assuming however that a high purity spectrum will produce good quality data and this is not necessarily true as if the precursor is of too low intensity the fragmentation spectrum may be too noisy to produce a good spectral match. This could occur with a high purity score as the calculation just looks at the intensity of all features that fell within the fragmentation window in the preceding and following MS¹ scans without factoring in MS² information. And therefore, this data could be misleading. However, the intensity distribution does not change significantly throughout the AX sequence (Figure 105). The number of low intensity features being fragmented does increase slightly in the later injections but overall the distribution is very similar and the number of features of decent intensity being fragmented remains fairly high. Many of these features however are likely to be co-eluting as can be seen by the large cluster of high intensity peaks in the BPC (Figure 104) towards the end of the analysis and this has led to decreased average purity.

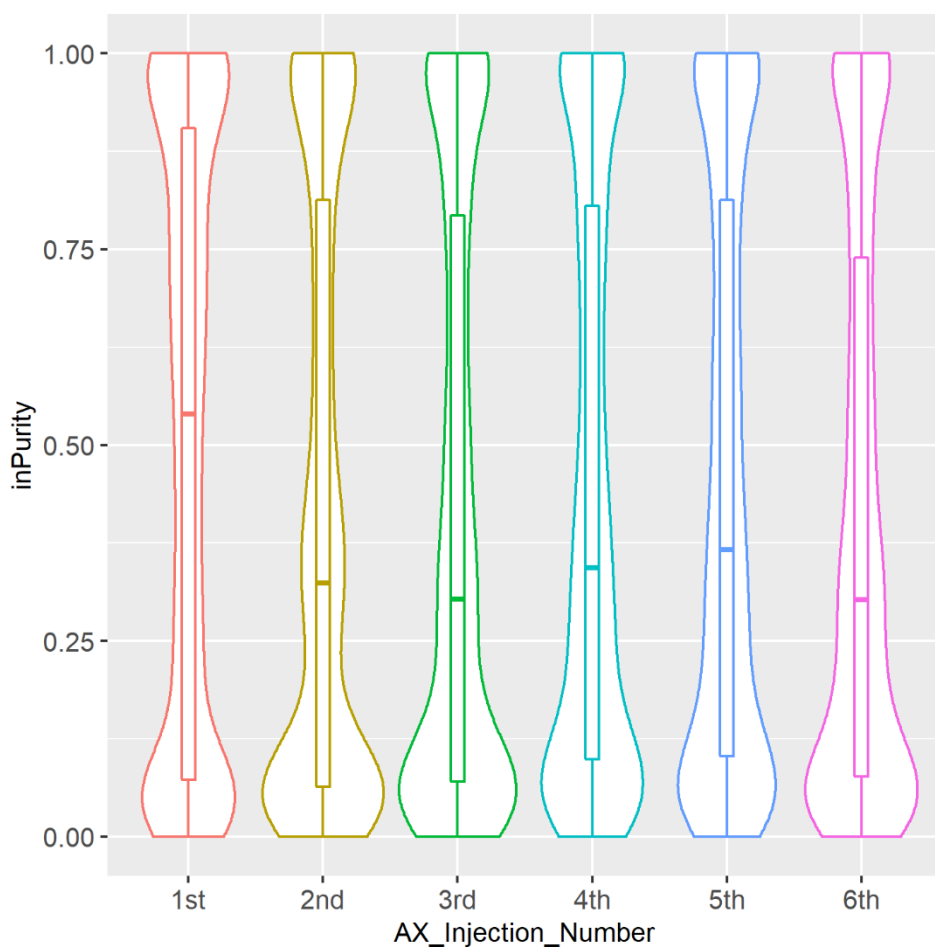


Figure 103: The distribution of interpolated purity scores acquired between 30 and 780 seconds through each injection of the AcquireX sequence for the RP/Plasma/Positive ion mode dataset.

Table 55: The number of MS² spectra with a purity score of 100 % for each injection during the AcquireX sequence in the RP/Plasma/Positive ion mode dataset.

Dataset	1st	2nd	3rd	4th	5th	6th
RP/Plasma/Positive ion mode	1076	1032	962	1029	1044	959

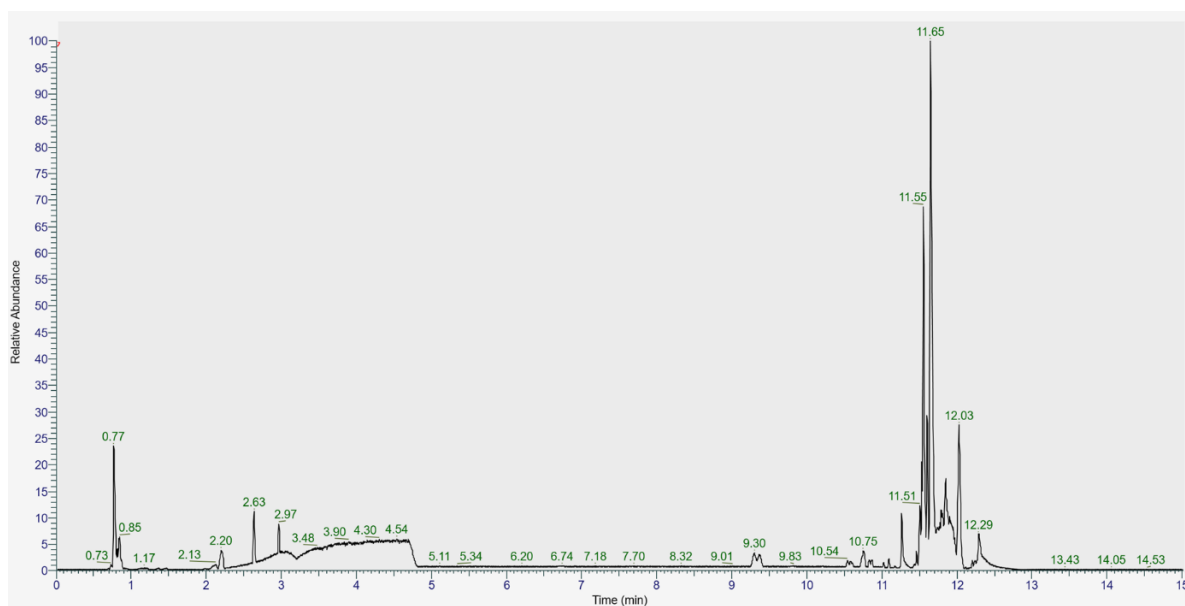


Figure 104: Base peak chromatogram (BPC) for the RP/Plasma/Positive ion mode full scan acquisition from the AX sequence.

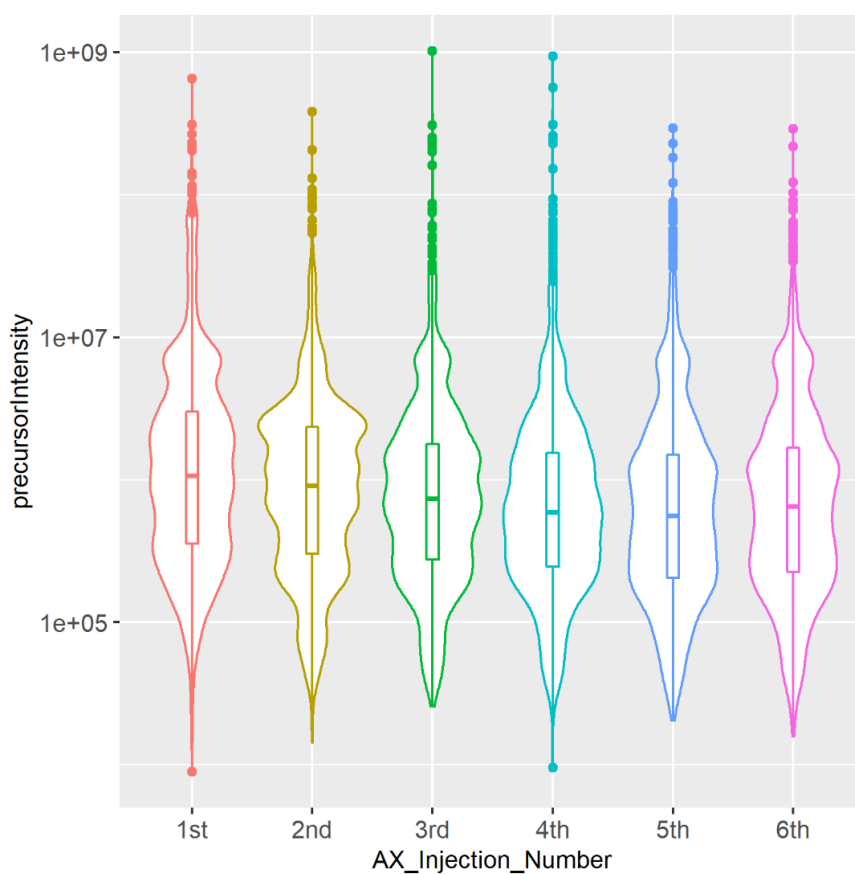


Figure 105: The distribution of precursor intensities acquired between 30 and 780 seconds with log10 transformed y axis through each injection of the AX sequence for the RP/Plasma/Positive ion mode dataset.

Through these two examples we have seen low and medium purity levels as determined by the assay, sample type and ion mode combination. A third example of the HILIC/Non-Smokers Urine/Positive ion mode is presented of higher purity (Figure 106) and clearly confirms the effects discussed. The median purity does not decrease below 75% even during the 6th injection. This was a dataset of medium complexity, with 2,089 compounds being detected, with this being only slightly more than seen in the low purity RP/Plasma/positive ion mode dataset. A key difference between the datasets is the quality of chromatographic separation which appears to be very good (Figure 107). The high quality of separation means that despite similar overall data complexity the purity score distributions are extremely different. The purity as expected decreases through each injection but is still high at the 6th injection. The number of 100% purity spectra are very high across every injection (Table 56) further demonstrating the quality of separation seen, the importance of this and the value of 6 injections for this assay. The median intensity has actually increased through each injection of the sequence (Figure 108) with a general increase in the range of intensities being fragmented too. This will be due to lower intensity features being on the inclusion list for preferential fragmentation selection in the initial injections of the sequence.

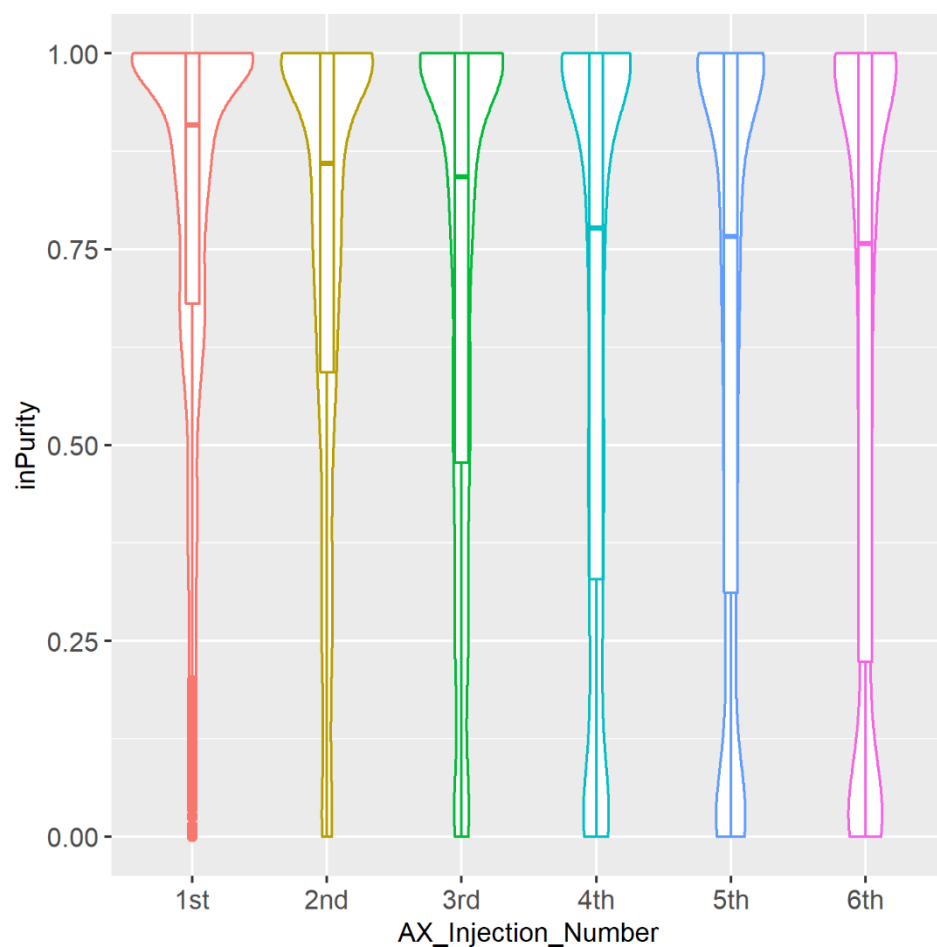


Figure 106: The distribution of interpolated purity scores acquired between 15 and 690 seconds through each injection of the AcquireX sequence for the HILIC/Non-Smokers Urine/Positive ion mode dataset.

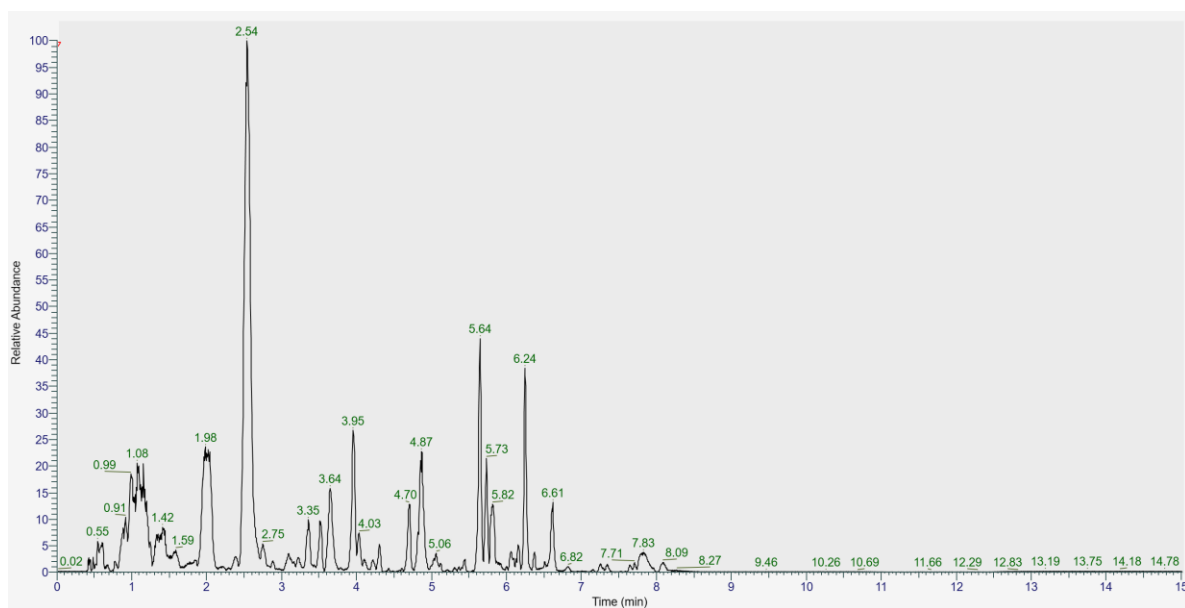


Figure 107: Base peak chromatogram (BPC) for the HILIC/Non-Smokers Urine/Positive ion mode full scan acquisition from the AX sequence.

Table 56: The number of MS² spectra with a purity score of 100 % for each injection during the AcquireX sequence in the HILIC/Non-Smokers/Positive ion mode dataset.

Dataset	1 st	2 nd	3 rd	4 th	5 th	6 th
HILIC/Non-Smokers Urine/Positive Ion Mode	2469	2197	2260	2185	2385	1923

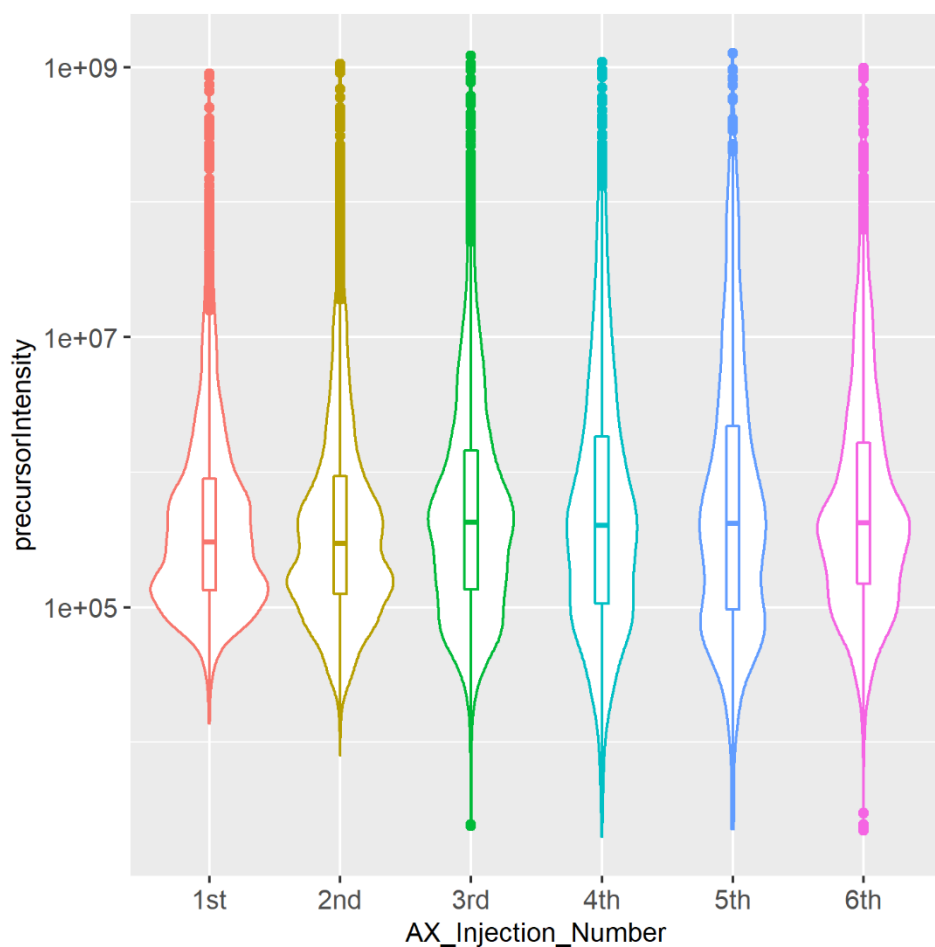


Figure 108: The distribution of precursor intensities acquired between 15 and 690 seconds with log10 transformed y axis through each injection of the AcquireX sequence for the HILIC/Non-Smokers Urine/Positive ion mode dataset.

Table 57: Summary of the three datasets presented and their key characteristics including the number of compounds detected in the full scan mode of the AcquireX sequence, the quality of separation seen in the BPC, the lowest median purity of any injection and the intensity trend through the AcquireX sequence.

Dataset	Number of Compounds	Separation Quality	Lowest Median Purity (%)	Intensity Through AX
RP/Non-Smokers Urine/Negative	3457	Good	55	Decreases
RP/Plasma/Positive	1757	Poor	32	Decreases
HILIC/Non-Smokers Urine/Positive	2089	Excellent	75	Increases

With the high number of high purity spectra being acquired it is perhaps disappointing that the numbers of compounds being identified is not higher. This could be due to poor spectral quality despite apparent purity for lower intensity ions, perhaps due to poor ion transmission. However, another important possible reason is the need for better and larger MS² spectral libraries. To help determine if the observations described above are caused by poor spectral quality as a result of low precursor purity all the mzCloud best match scores were plotted against the log transformed precursor maximum peak area values for each of the three datasets discussed (Figure 109, Figure 110, Figure 111). Linear regression was carried out and the R and *p* values generated demonstrate that there is a significant positive correlation between spectral match score and precursor intensity. The data clearly shows that although good spectral match scores are more likely at higher peak intensities they can still be observed for lower intensity features. Although an important factor to consider in this case is what was the intensity of features at the point of ion isolation? Was fragmentation triggered close to the peak apex or not? Whether this is true would dramatically effect the results as the maximum peak area is plotted in all cases. This data indicates that even lower intensity features may still be appropriate for collecting fragmentation data for but the likelihood of a confident identification is decreased.

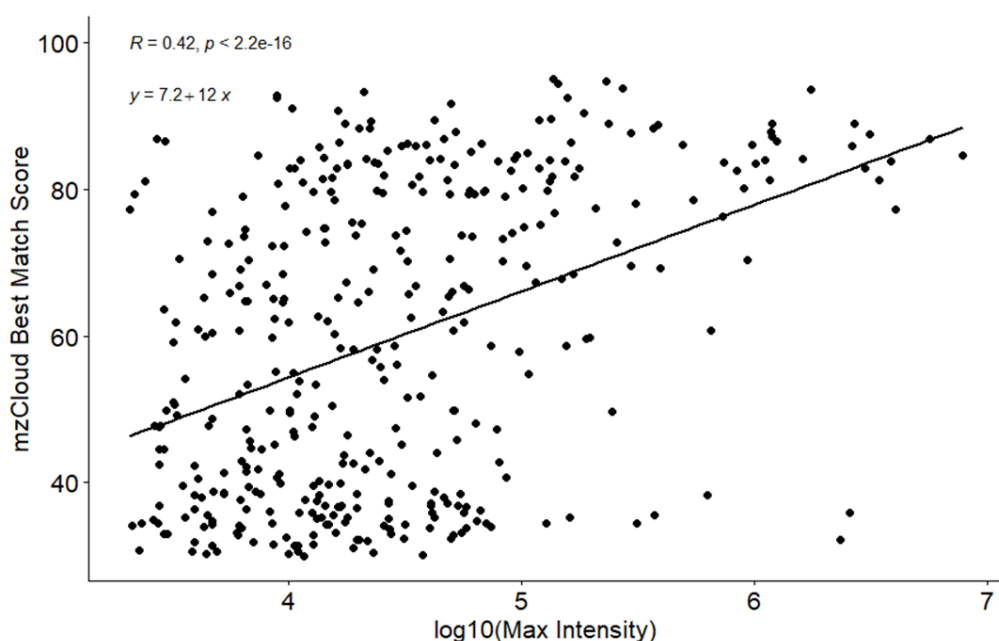


Figure 109: All mzCloud best match scores vs \log_{10} (Max Precursor Peak Area) for RP/Non-Smokers Urine/Negative Ion Mode. $R = 0.42$.

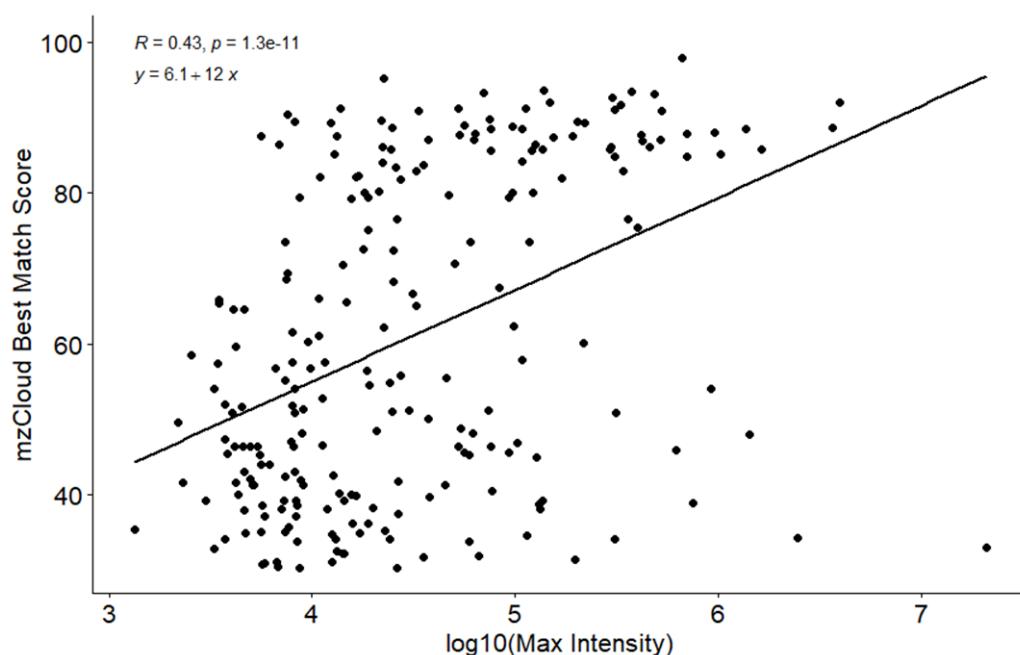


Figure 110: All mzCloud best match scores vs \log_{10} (Max Precursor Peak Area) for RP/Plasma/Positive Ion Mode. $R = 0.43$.

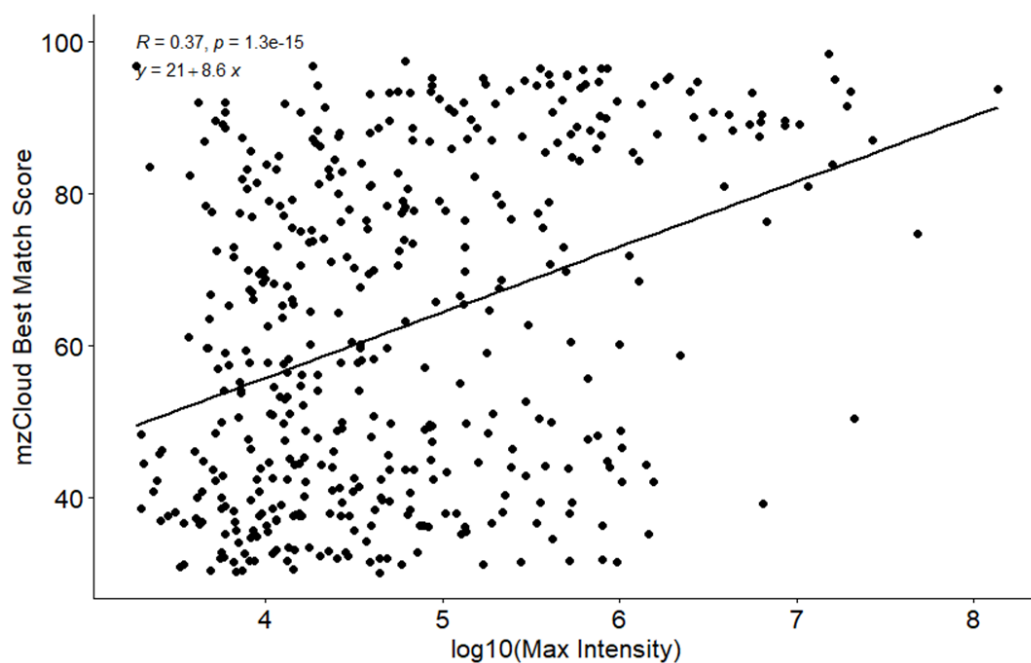


Figure 111: All mzCloud best match scores vs \log_{10} (Max Precursor Peak Area) for HILIC/Non-Smokers Urine/Positive Ion Mode. $R = 0.37$.

6.3 Conclusions

Throughout this chapter a comparison of a new iDDA method called AcquireX, inbuilt into the new Orbitrap ID-X (Thermo Fisher Scientific) was compared to a traditional DDA approach for annotation of metabolites in untargeted metabolomics alongside the development of a custom MS^2 library. A

number of metrics were utilised for assessment of its effectiveness including the number of level 1 and level 2 identifications but other metrics such as the volume of compounds with MS² data associated with them was recorded too. The advantage of the AcquireX method was dependent on the complexity of the data generated which is determined by the chromatographic assay, sample type and ion mode utilised. If the sample type, assay, ion mode combination resulted in a high number of features then the AcquireX method was very beneficial over a traditional DDA method, more level 1 identifications were recorded, over 2 times the number of level 2 identifications were achieved, over 5 times the number of compounds with MS² information were provided. Alongside a decrease in the number of compounds with MS² information for a non-protonated ion, clearly demonstrating the intelligence of the precursor selection. On the other hand, if the sample dataset was of lower complexity then although an advantage was still provided it was of much less significance. AX provided marginally more level 1 and level 2 identifications, however there was still a significant increase in the number of compounds with MS² spectra associated with them and so more potentially useful data were collected at least. So, the complexity of the dataset impacts the benefits of AcquireX and thus if known beforehand should be used to inform the number of repeated injections that might be required as this was also seen to be variable depending on sample complexity as well as quality of chromatographic separation.

The fact that so many more MS² spectra were acquired with AX compared to DDA but the increase in the numbers of compounds that were actually identified is much smaller is indicative of some of the other major conclusions from this work. One is that the precursor intensity and purity of the subsequent fragmentation spectra are very important, these should look to be maximised wherever possible and practical as there is a positive correlation between intensity and spectral match quality. No Mullard segmentation was carried out to the total mass range in these studies but based on the data in the previous chapter and the flaws seen in the data in this chapter a series of Mullard segmented AcquireX methods would be the authors recommendation for the greatest number of annotations. The improved sensitivity resulting from segmentation of the mass range will mean greater intensities for features across the mass range and this could be vital when performing MS² fragmentation. Purity of fragmentation events can also be increased by improving chromatographic methods, introducing nano-LC for example should improve sensitivity and feature intensities, or more importantly introducing a second dimension of separation may be the key to rapidly increasing the number of identifications that are achievable by increasing the purity of the spectra acquired. Secondly the improvement and expansion of spectral databases is key, there is still large scope for improvement of these databases and the number of spectra within, perhaps many spectra collected went unannotated despite good purity and quality due to the lack of a reference spectrum. In-silico

fragmentation tools/libraries will also be important to cover the gaps that remain in the MS² libraries. Overall however the AcquireX software facilitates easy, on the fly intelligent DDA acquisitions which would otherwise only be possible with significant and rushed user intervention with rapid processing and method modification to ensure the correct result. AcquireX changes this and alongside its MSⁿ capabilities as a tribrid instrument for further fragmentation and structural information promises to push untargeted metabolomics forward.

Whilst it has been recommended to perform a segmented AcquireX sequence for best annotation results the time taken for these analyses to be performed must also be considered, it is unlikely that it would be practical to perform an AcquireX sequence on all biological samples in a study for example as the overall analysis time would then likely be too long. Application of the method to a pooled QC sample would allow deep annotation to be performed on the QC whilst other samples in the study undergo routine full scan analyses. The annotations from the QC can then be applied to the full scan features from the individual samples analysed in the study.

7.0 Conclusions

Untargeted metabolomics is still a relatively young field and it has thus far been hindered by the many difficulties presented by the annotation of metabolites and the resulting “dark matter” of metabolomics data found at the identification bottleneck. There are a host of issues which make metabolite annotation particularly challenging including but not limited to:

1. The vast and inherent physicochemical diversity of natural compounds observed in biological systems as well as synthetic drug compounds and chemicals which may be present in samples – No one analytical method is appropriate for detection of all compounds.
2. The presence of isobaric and isomeric species in biology – Assigning a confident annotation requires detailed structural information which is difficult to obtain for all compounds detected.
3. The complexity of ESI data - A single analyte may be detected as multiple different species including different adducts, isotopes, in-source fragments, oligomers, multiply charged species, biotransformations and radicals which require grouping to avoid false positive annotations. MS artefacts also add complexity and require removal.
4. Incomplete and insufficient MS² spectral libraries – MS² data acquired may not match to anything in the libraries utilised as many compounds do not have reference spectra present yet. Furthermore, data or metadata quality in the library may not always be satisfactory.
5. Authentic standards are not commercially available for all compounds and they are expensive to acquire – Comparison of experimental data to standards analysed on the same analytical setup is the only way to achieve a level 1 (highest confidence) annotation.
6. Difficulty in the comparison of MS data between labs – Comparison may not be appropriate due to differences in assays, instrumentation, and system parameters. There can also be a lack of reproducibility between different labs using the same analytical setup.
7. There is no complete parts list for metabolomes – It is unclear what a complete metabolome for any sample type looks like.

These factors combine to inhibit metabolite annotation but annotation must be improved in terms of the number of metabolites that can be confidently and reliably annotated within a typical study for the field to gain more traction and make more impact. As a result, the work presented throughout this thesis had the goal of providing new knowledge and insights that may improve metabolite annotation capabilities in untargeted metabolomics experiments.

This was first done by looking at problem number 3 from the above list, the unconsidered complexity in ESI data and subsequent feature relationships at the MS¹ level. It was shown through the unprecedented analysis of 104 publicly available datasets that there is great variety in the highly

correlated m/z differences detected. These differences are derived from a variety of adducts, isotopes, in-source fragments, biotransformations, homo and heterodimers and multiply charged species alongside many possible combinations of these. The primary conclusion from this chapter is that the true variety and diversity of derivative feature types formed in UHPLC-ESI-MS data is far beyond the level previously considered. Grouping of these features is ultimately essential for correct annotation and biological interpretation of the resulting data. The second major conclusion was provided in the investigation of which particular adducts are considered in commonly applied metabolite annotation resources. It was demonstrated that most annotation software use different adducts and ion types, do not consider sufficient variety of adducts, and that some commonly considered adducts that perhaps should not be commonly searched for as they are infrequently detected.

Overall, this chapter provides an important contribution to the field by acting as a warning and reminder to researchers that many more features are derived from the same metabolite than may have historically been expected and more attention is required to this neglected area of metabolomics. Errors made at the MS^1 level can easily lead to false positive or false negative annotations and most commonly used annotation software are generally too simplistic in the variety and types of derivative features considered in their approach to correctly identify all related features. Whilst this represents an important contribution there are a number of limitations to the work carried out in this chapter. These include the arbitrary implementation of cut off values for Pearson correlation coefficient and RT window size, a more ideal approach would have involved tailoring of these to each individual datasets characteristics. Secondly, some of the datasets may have had their feature matrix grouped or de-isotoped already and this was not clear in the repositories, elimination of these datasets from the overall data before processing would improve the results. Thirdly, a small number of datasets that were included in the study had a lower number of samples than was ideally wanted and so the correlation calculations may be less reliable for these datasets. Ultimately though these datasets were deemed necessary to include to provide sufficient representation in the different groups and assays being investigated to allow a global rather than a focussed investigation and provide conclusions applicable to the metabolomics community in full. It would have been very beneficial to have access to more datasets to allow effective statistical comparison of different groups, however there were a limited number of suitable datasets available in the publicly available data repositories. This was not only a problem for certain groups such as Shimadzu mass spectrometers but also for the entire study in general. Many researchers are reluctant to upload their raw data to repositories. This may be because they do not want to share their data but it may also be due to time consuming and cumbersome data submission processes. Many other researchers have chosen to upload data in a way so that it is incomplete and becomes inappropriate for the kind of large scale data analysis applied in

this chapter. Within all science, but in metabolomics in particular, there is a great need for less fierce competition between researchers and a greater culture of data sharing and collaboration. Metabolights have recently updated their data submission process to make it more straightforward and steps such as this which reduce the time required for busy researchers to upload their data should improve the volume of data that is publicly available. If repeating this work tailored cut offs for correlation coefficient and RT windows would be employed, whilst the volume of datasets utilised for the analysis would be increased as far as possible. With a much larger volume of datasets present statistical analyses could be carried out to determine the differences between groups in a robust manner which was not possible with the volume of data currently present. Further work could also include analysing standards (or isotopically labelled standards for greater confidence in annotations) in solution as well as in complex mixtures at varying concentrations to investigate the formation of these derivative feature types in a more targeted manner.

If individual research labs have standard assays which are implemented routinely for a variety of different biological analyses then the mass differences and adducts most typically associated with that particular assay should be recorded and the annotation pipeline be modified to account for this. The author also proposes an online mass difference database, where researchers can input the most commonly seen differences and adducts in their study with appropriate meta data to build up a knowledge base that can subsequently reveal the adducts which are most appropriate for consideration for different sample types, assays or other key experimental characteristics. This will be built on analysis of large volumes of publicly available data and will become more powerful the more data that is uploaded to it. More community cooperation such as this is required for faster advancement of the field as well as greater standardisation between labs where possible. Standardisation of analytical methods whether that is just the chromatographic step or the whole UHPLC-MS pipeline across multiple labs would increase comparability of data between labs for both MS^1 , MS^2 and MS^n data. This increase in comparability would allow more effective collaborations but most importantly a more robust understanding of the intricacies of the resulting data for example the different degenerate feature types that would be expected to be found.

Once the MS^1 data has been processed and annotated the MS^2 data can be utilised. MS^2 data is often essential for confident identification of isobaric and isomeric compounds but acquiring good quality MS^2 information for all features in a study is a challenge. This is problem number 2 from the list above and across the second, third and fourth research chapters methods for solving this issue were investigated. MS^2 is the most commonly used structural information for confident identification of metabolites. Traditionally a DDA approach is used for collection of MS^2 data however this approach is

biased towards ions of high intensity and its use results in many low intensity features being unidentifiable due to lack of structural information acquired. Therefore, mechanisms for improving the coverage of MS² information was an area of interest when looking to improve metabolite annotation. The possibility of applying DIA, a method which fragments all features in an unbiased fashion was thus investigated on Orbitrap based instruments. It was determined that a key factor was likely to be complexity of the DIA windows, and that the number of features falling within any single window should be restricted to maintain quality. This was assessed for theoretical DIA windows using a data complexity visualisation tool after optimisation of data processing parameters. The scan rate of the Q Exactive Plus mass spectrometer applied was also investigated in relation to the chromatographic peak widths of the assay applied. These analyses together indicated DIA may be applicable, particularly with lower complexity datasets. Unfortunately however the scan rate of Orbitrap instruments is inherently limited by the Orbitrap itself and thus to design DIA experiments with reasonable window sizes that still maintain a suitable MS¹ scan rate with the chromatographic assays in question the window range must be reduced. This results in DIA not providing the global coverage sought after. Whilst traditional DDA methods have their limitations they can also be modified to increase the level of coverage of MS² data that they provide. DDA methods can be modified through the use of inclusion and/or exclusion lists as well as through segmentation of the mass range, these DDA types were said to be intelligent. Considering the data gathered and analysed iDDA methods were planned for comparison to traditional DDA as well as DIA and AIF methods on an Orbitrap Q Exactive Plus (Thermo Fisher Scientific, USA). The major limitation for this section of the work was the application of Orbitrap mass analysers due to their slower scan rate. If a TOF based instrument was applied their significantly greater scan rate would have facilitated a greater number of MS² scans (and thus smaller DIA window sizes) to be carried out between each MS¹ scan whilst still collecting the MS¹ data with sufficient regularity to allow accurate peak identification and quantification. Therefore, if performing this work again a TOF based instrument would have been applied. Another method to facilitate use of more MS² scans but whilst still applying Orbitrap analysers could be to employ a longer chromatographic method with broader chromatographic peak widths or to employ another Orbitrap instrument with a faster scan rate than that of the Q Exactive Plus. Both of these options will be considered for future work. The main conclusion from this chapter was that DIA was unlikely to be appropriate or superior to DDA methods on the Orbitrap Q Exactive Plus; however, if the experiment is carefully planned perhaps DIA could demonstrate some advantages such as the capacity to annotate lower intensity features. The following research chapter applied the methods developed in this chapter to determine which of them would provide the most coverage of MS² data and also the greatest volume of good quality informative MS² data.

Comparison of these methods showed that intelligent DDA acquisition comprising of Mullard segmentation of the total mass range with updated exclusion lists to ensure non-repetitive feature fragmentation, superior coverage of good quality MS² data and thus the greatest number of good annotations, was the best strategy. DIA methods are inappropriate on Orbitrap based systems due to the scan rate although may still be beneficial on TOF based instruments. AIF methods provide annotations and good coverage but the data is not trustworthy. Simple traditional DDA methods are reliable but as discussed suffer from lack of coverage. The issue with applying intelligent DDA on the Orbitrap Q Exactive Plus is that it required significant time in planning and rushing through processing of updated exclusion lists during the analytical run. Therefore, it is difficult to implement manually. This has since changed with the release of the Orbitrap ID-X (Thermo Fisher Scientific, USA) and its AcquireX method which is capable of applying intelligent DDA methods on the fly with no user intervention. This chapter provided valuable recommendations to the many Orbitrap users in the metabolomics community for how to get the most valuable structural information out of their MS² data collection. In the subsequent and final research chapter some of the findings of this chapter were applied in the comparison of the AcquireX method and a traditional DDA method on an Orbitrap ID-X.

The Orbitrap ID-X was implemented to build an in-house MS² library using the MetaSci COMPLETE human standards kit. This is the only way to generate level 1 identifications and labs performing a high volume of biological analyses using standard assays should develop their own libraries to facilitate truly confident identifications. The AcquireX method was then used in comparison to traditional DDA for annotation of human plasma and urine to demonstrate the advantage an intelligent DDA method can provide over traditional analyses. The subsequent conclusion was that the advantage is dependent on the complexity of the sample type, if the complexity is high then the advantage is large. If the sample complexity is low then the advantage is not as significant. Precursor intensity and purity is also very important, the greater they are the greater the chance there is of a good spectral match and future work should look at ways to maximise these. This can be easily done initially through Mullard segmentation of the mass range in combination with an AcquireX method and implementation of apex triggering. This was not applied in these experiments due to time constraints which were the main limitation in this work but would certainly be a focus for a future work. Investigation of different precursor window sizes and AGC target values could also be carried out. Modification to chromatographic methods should also be investigated in future work, implementation of nano and 2D-LC can improve sensitivity and separation to improve intensity and purity of precursors and thus improve MS² quality. Novel software ideas should be pursued too such as whether or not it is possible for a mass spectrometer to be designed so that it can perform on the fly calculations to vary the size

of the precursor isolation window in accordance with the feature intensity and surrounding feature density? Other possible improvements to the AcquireX method could be introduction of an automatic complexity assessment, the RT-*m/z* axes of the data collected can be automatically assessed for peak density with automatically generated Mullard segmented methods tailored to the distribution of the features across the RT-*m/z* axes. This would allow the most efficient use of time and generation of informative MS² data. This chapter clearly demonstrated to the community the advantage of the AcquireX method, progressively removed from inclusion lists and progressively added to exclusion lists in general over a traditional DDA method if applied to a complex sample type.

Whilst AcquireX and the suggested improvements can keep increasing the volume of informative data gathered in biological studies there still needs to be continual expansion of current MS² libraries with high quality data provided from the analysis of pure standards. The greater the variety of standards analysed the more biological knowledge can be gained from studies using those reference libraries. Within those libraries there are still significant gaps in biologically important metabolites missing whilst many other spectra present in databases are not derived from standards. Improvement of annotation software such as in-silico fragmentation tools and integration of these into easy to navigate pipelines such as Galaxy will be important.

Metabolomics is heavily dependent on technology and computational data processing. The field has advanced rapidly throughout the last two decades and will continue to do so as technologies improve. Advances in separation technologies such as nano and 2D-LC as well as emerging techniques such as Ion Mobility (IM) will be important to increase the amount of orthogonal data available for identification and improve separation of metabolites in complex mixtures for superior annotation. Improvements in mass spectrometers will be important too to further improve the scan rate and or resolution capabilities, superior data could be gathered. Improvement of computational tools and processing pipelines will also be important.

Overall the work presented in this thesis has provided important information and a number of recommendations that can be used to improve annotation in untargeted metabolomics but much more work is required in the field to improve the annotation success rate. If all features could be correctly identified in untargeted metabolomics experiments it would be an incredibly powerful tool but this is going to require continued research into improving both the analytical and computational processes involved.

8.0 Bibliography

- Aksenov, A.A., da Silva, R., Knight, R., et al. (2017) Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry*, 1 (7): 0054. doi:10.1038/s41570-017-0054.
- Albóniga, O.E., González, O., Alonso, R.M., et al. (2020) Optimization of XCMS parameters for LC–MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics*, 16 (1): 14. doi:10.1007/s11306-020-1636-9.
- Allard, P.M., Genta-Jouve, G. and Wolfender, J.L. (2017) Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. *Current Opinion in Chemical Biology*, 36: 40–49. doi:10.1016/j.cbpa.2016.12.022.
- Allen, F., Pon, A., Wilson, M., et al. (2014) CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Research*, 42 (W1): 94–99. doi:10.1093/nar/gku436.
- Alonso, A., Julià, A., Beltran, A., et al. (2011) AStream: An R package for annotating LC/MS metabolomic data. *Bioinformatics*, 27 (9): 1339–1340. doi:10.1093/bioinformatics/btr138.
- Annesley, T.M. (2003) Ion suppression in mass spectrometry. *Clinical Chemistry*. 49 (7) pp. 1041–1044. doi:10.1373/49.7.1041.
- Banerjee, S. and Mazumdar, S. (2012) Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *International Journal of Analytical Chemistry*, 2012: 40. doi:10.1155/2012/282574.
- Baptista, R., Fazakerley, D.M., Beckmann, M., et al. (2018) Untargeted metabolomics reveals a new mode of action of pretomanid (PA-824). *Scientific Reports*, 8 (1). doi:10.1038/s41598-018-23110-1.
- Barupal, D.K. and Fiehn, O. (2017) Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Scientific Reports*, 7 (1): 14567. doi:10.1038/s41598-017-15231-w.
- Bayram, M. and Gökırmaklı, Ç. (2018) Horizon Scanning: How Will Metabolomics Applications Transform Food Science, Bioengineering, and Medical Innovation in the Current Era of Foodomics? *OMICS: A Journal of Integrative Biology*, 22 (3): 177–183. doi:10.1089/omi.2017.0203.
- Beale, D.J., Oh, D.Y., Karpe, A. V, et al. (2019) Untargeted metabolomics analysis of the upper respiratory tract of ferrets following influenza A virus infection and oseltamivir treatment. *Metabolomics*, 15 (3): 33. doi:10.1007/s11306-019-1499-0.

- Beebe, K. and Kennedy, A.D. (2016) Sharpening Precision Medicine by a Thorough Interrogation of Metabolic Individuality. *Computational and Structural Biotechnology Journal*, 14: 97–105. doi:10.1016/j.csbj.2016.01.001.
- Beger, R.D., Dunn, W., Schmidt, M.A., et al. (2016) Metabolomics enables precision medicine: “A White Paper, Community Perspective.” *Metabolomics*, 12 (10): 149. doi:10.1007/s11306-016-1094-6.
- Beger, R.D., Dunn, W.B., Bandukwala, A., et al. (2019) Towards quality assurance and quality control in untargeted metabolomics studies. *Metabolomics*, 15 (1): 4. doi:10.1007/s11306-018-1460-7.
- van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., et al. (2006) Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7. doi:10.1186/1471-2164-7-142.
- Bian, Y., Zheng, R., Bayer, F.P., et al. (2020) Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. *Nature Communications*, 11 (1). doi:10.1038/s41467-019-13973-x.
- Bielow, C., Mastrobuoni, G., Orioli, M., et al. (2017) On Mass Ambiguities in High-Resolution Shotgun Lipidomics. *Analytical Chemistry*, 89 (5): 2986–2994. doi:10.1021/acs.analchem.6b04456.
- Biemann, K. (1962) The Application of Mass Spectrometry in Organic Chemistry: Determination of the Structure of Natural Products. *Angewandte Chemie International Edition in English*, 1 (2): 98–111. doi:10.1002/anie.196200981.
- Biemann, K., Cone, C., Webster, B.R., et al. (1966) Determination of the Amino Acid Sequence in Oligopeptides by Computer Interpretation of Their High-Resolution Mass Spectra. *Journal of the American Chemical Society*, 88 (23): 5598–5606. doi:10.1021/ja00975a045.
- Blake, J.A., Christie, K.R., Dolan, M.E., et al. (2015) Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43 (D1): D1049–D1056. doi:10.1093/nar/gku1179.
- Blaženović, I., Kind, T., Torbašinović, H., et al. (2017) Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: Database boosting is needed to achieve 93% accuracy. *Journal of Cheminformatics*, 9 (1): 1–12. doi:10.1186/s13321-017-0219-x.
- Bligh, E.G. and Dyer, W.J. (1959) A rapid method of total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology*, 37 (8): 911–917. doi:10.1139/o59-099.

Böcker, S. (2017) Searching molecular structure databases using tandem MS data: are we there yet? *Current Opinion in Chemical Biology*, 36: 1–6. doi:10.1016/j.cbpa.2016.12.010.

Bolton, E.E., Wang, Y., Thiessen, P.A., et al. (2008) *PubChem: Integrated Platform of Small Molecules and Biological Activities*. Elsevier B.V. doi:10.1016/S1574-1400(08)00012-1.

Bonner, R. and Hopfgartner, G. (2018) SWATH data independent acquisition mass spectrometry for metabolomics. *TrAC - Trends in Analytical Chemistry*, October. doi:10.1016/j.trac.2018.10.014.

Bouatra, S., Aziat, F., Mandal, R., et al. (2013) The Human Urine Metabolome. *PLoS ONE*, 8 (9). doi:10.1371/journal.pone.0073076.

Bravais, A. (1844) *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Paris: Mem. Acad. Roy. Sci. Inst. France. Available at: [https://books.google.co.uk/books?hl=en&lr=&id=y3s_AAAcAAJ&oi=fnd&pg=PA19&dq=Analyse+mathématique+sur+les+probabilités+des+erreurs+de+situation+d%27un+point&ots=Plcth5rs6M&sig=BnGbtYH5_eqdQESJ-rfGYA6Qags#v=onepage&q=Analyse mathématique sur les probabilités](https://books.google.co.uk/books?hl=en&lr=&id=y3s_AAAcAAJ&oi=fnd&pg=PA19&dq=Analyse+mathématique+sur+les+probabilités+des+erreurs+de+situation+d%27un+point&ots=Plcth5rs6M&sig=BnGbtYH5_eqdQESJ-rfGYA6Qags#v=onepage&q=Analyse+mathématique+sur+les+probabilités) (Accessed: 30 October 2020).

Broadhurst, D., Goodacre, R., Reinke, S.N., et al. (2018) Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, 14 (6): 72. doi:10.1007/s11306-018-1367-3.

Broadhurst, D.I. and Kell, D.B. (2007) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2 (4): 171–196. doi:10.1007/s11306-006-0037-z.

Broeckling, C.D., Afsar, F.A., Neumann, S., et al. (2014) RAMClust: A novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Analytical Chemistry*, 86 (14): 6812–6817. doi:10.1021/ac501530d.

Broeckling, C.D., Ganna, A., Layer, M., et al. (2016) Enabling Efficient and Confident Annotation of LC-MS Metabolomics Data through MS1 Spectrum and Time Prediction. *Analytical Chemistry*, 88 (18): 9226–9234. doi:10.1021/acs.analchem.6b02479.

Broeckling, C.D. and Prenni, J.E. (2018) Stacked Injections of Biphasic Extractions for Improved Metabolomic Coverage and Sample Throughput. *Analytical Chemistry*, 90 (2): 1147–1153. doi:10.1021/acs.analchem.7b03654.

Brown, M., Dunn, W.B., Dobson, P., et al. (2009) Mass spectrometry tools and metabolite-specific

databases for molecular identification in metabolomics. *The Analyst*, 134 (7): 1322.
doi:10.1039/b901179j.

Brown, M., Wedge, D.C., Goodacre, R., et al. (2011) Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics*, 27 (8): 1108–1112. doi:10.1093/bioinformatics/btr079.

Brown, R.A. (1951) Compound Types in Gasoline by Mass Spectrometer Analysis. *Analytical Chemistry*, 23 (3): 430–437. Available at: <https://pubs.acs.org/sharingguidelines> (Accessed: 9 November 2020).

Bundy, J.G., Davey, M.P. and Viant, M.R. (2009) Environmental metabolomics: A critical review and future perspectives. *Metabolomics*, 5 (1): 3–21. doi:10.1007/s11306-008-0152-0.

Del Carratore, F., Schmidt, K., Vinaixa, M., et al. (2019) Integrated Probabilistic Annotation: A Bayesian-Based Annotation Method for Metabolomic Profiles Integrating Biochemical Connections, Isotope Patterns, and Adduct Relationships. *Analytical Chemistry*, 17: 29.
doi:10.1021/acs.analchem.9b02354.

Caspi, R., Billington, R., Ferrer, L., et al. (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44 (D1): D471–D480. doi:10.1093/nar/gkv1164.

Chambers, M.C., MacLean, B., Burke, R., et al. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*. 30 (10) pp. 918–920. doi:10.1038/nbt.2377.

Chen, G., Walmsley, S., Cheung, G.C.M., et al. (2017) Customized consensus spectral library building for untargeted quantitative metabolomics analysis using data independent acquisition mass spectrometry and MetaboDIA workflow. *Analytical Chemistry*, p. acs.analchem.6b05006.
doi:10.1021/acs.analchem.6b05006.

Chen, T., Cao, Y., Zhang, Y., et al. (2013) Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-based Complementary and Alternative Medicine*, 2013. doi:10.1155/2013/298183.

Chetwynd, A.J. and David, A. (2018) A review of nanoscale LC-ESI for metabolomics and its potential to enhance the metabolome coverage. *Talanta*. 182 pp. 380–390. doi:10.1016/j.talanta.2018.01.084.

Chetwynd, A.J., David, A., Hill, E.M., et al. (2014) Evaluation of analytical performance and reliability of direct nanoLC-nanoESI-high resolution mass spectrometry for profiling the (xeno)metabolome.

Journal of Mass Spectrometry, 49 (10): 1063–1069. doi:10.1002/jms.3426.

Chong, J., Soufan, O., Li, C., et al. (2018) MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, 46 (W1): W486–W494. doi:10.1093/nar/gky310.

Coble, J.B. and Fraga, C.G. (2014) Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. *Journal of Chromatography A*, 1358: 155–164. doi:10.1016/j.chroma.2014.06.100.

Comisarow, M.B. and Marshall, A.G. (1976) Theory of Fourier transform ion cyclotron resonance mass spectroscopy. I. Fundamental equations and low-pressure line shape. *The Journal of Chemical Physics*, 64 (1): 110–119. doi:10.1063/1.431959.

Correa, E. and Goodacre, R. (2011) A genetic algorithm-Bayesian network approach for the analysis of metabolomics and spectroscopic data: Application to the rapid identification of *Bacillus* spores and classification of *Bacillus* species. *BMC Bioinformatics*, 12. doi:10.1186/1471-2105-12-33.

Cottet, K., Genta-Jouve, G., Fromentin, Y., et al. (2014) Comparative LC–MS-based metabolite profiling of the ancient tropical rainforest tree *Symphonia globulifera*. *Phytochemistry*, 108: 102–108. doi:10.1016/j.phytochem.2014.09.009.

Cowcher, D.P., Xu, Y. and Goodacre, R. (2013) Portable, quantitative detection of *Bacillus* bacterial spores using surface-enhanced Raman scattering. *Analytical Chemistry*, 85 (6): 3297–3302. doi:10.1021/ac303657k.

Creek, D.J., Jankevics, A., Breitling, R., et al. (2011) Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Analytical Chemistry*, 83 (22): 8703–8710. Available at: <http://pubs.acs.org/doi/abs/10.1021/ac2021823>.

Daly, R., Rogers, S., Wandy, J., et al. (2014) MetAssign: Probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics*, 30 (19): 2764–2771. doi:10.1093/bioinformatics/btu370.

Davidson, R.L., Weber, R.M.J., Liu, H., et al. (2016) Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *GigaScience*, 5: 1–9. doi:10.1186/s13742-016-0115-8.

van Deemter, J.J., Zuiderweg, F.J. and Klinkenberg, A. (1956) Longitudinal diffusion and resistance to

mass transfer as causes of nonideality in chromatography. *Chemical Engineering Science*, 5 (6): 271–289. doi:10.1016/0009-2509(56)80003-1.

Defelice, B.C., Mehta, S.S., Samra, S., et al. (2017) Mass Spectral Feature List Optimizer (MS-FLO): a tool to minimize false positive peak reports in untargeted LC-MS data processing. *Analytical Chemistry*, p. acs.analchem.6b04372. doi:10.1021/acs.analchem.6b04372.

Delacre, M., Leys, C., Mora, Y.L., et al. (2019) Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA. *International Review of Social Psychology*, 32 (1): 13. doi:10.5334/irsp.198.

Dieterle, F., Ross, A., Schlotterbeck, G., et al. (2006) Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1 H NMR Metabonomics. *Analytical Chemistry*, 78 (13): 4281–4290. doi:10.1021/ac051632c.

Djerassi, C. and Fenselau, C. (1965) Mass Spectrometry in Structural and Stereochemical Problems. LXXXV. 1 The Nature of the Cyclic Transition State in Hydrogen Rearrangements of Aliphatic Amines 2,3. *Journal of the American Chemical Society*, 87 (24): 5752–5756. doi:10.1021/ja00952a040.

Djerassi, C., Karliner, J. and Aplin, R.T. (1965) Mass spectrometry in structural and stereochemical problems LXXVIII steroidal Δ^4 -3,6-diketones. *Steroids*, 6 (1): 1–8. doi:10.1016/0039-128X(65)90029-2.

Djerassi, C., Wilson, J.M., Budzikiewicz, H., et al. (1962) Mass Spectrometry in Structural and Stereochemical Problems. XIV. 1 Steroids with One or Two Aromatic Rings 2. *Journal of the American Chemical Society*, 84 (23): 4544–4552. doi:10.1021/ja00882a034.

Djombou-Feunang, Y., Pon, A., Karu, N., et al. (2019) Cfm-id 3.0: Significantly improved esi-ms/ms prediction and compound identification. *Metabolites*, 9 (4): 72. doi:10.3390/metabo9040072.

Dole, M., Mack, L.L., Hines, R.L., et al. (1968) Molecular Beams of Macroions. *The Journal of Chemical Physics*, 49 (5): 2240–2249. doi:10.1063/1.1670391.

Domingo-Almenara, X., Montenegro-Burke, J.R., Benton, H.P., et al. (2018) Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Analytical Chemistry*, 90 (1): 480–489. doi:10.1021/acs.analchem.7b03929.

Domingo-Almenara, X., Montenegro-Burke, J.R., Guijas, C., et al. (2019) Autonomous METLIN-Guided In-source Fragment Annotation for Untargeted Metabolomics. *Analytical Chemistry*, 91 (5): 3246–3253. doi:10.1021/acs.analchem.8b03126.

- Douglas, D.J. (2009) Linear quadrupoles in mass spectrometry. *Mass Spectrometry Reviews*, 28 (6): 937–960. doi:10.1002/mas.20249.
- Douglas, D.J., Frank, A.J. and Mao, D. (2005) Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews*, 24 (1): 1–29. doi:10.1002/mas.20004.
- Dove, A. (1999) Proteomics: translating genomics into products? *Nature Biotechnology*, 17 (3): 233–236. doi:10.1038/6972.
- Draisma, H.H., Reijmers, T.H., Meulman, J.J., et al. (2013) Hierarchical clustering analysis of blood plasma lipidomics profiles from mono-and dizygotic twin families. *European Journal of Human Genetics*, 21 (1): 95–101. doi:10.1038/ejhg.2012.110.
- Dudzik, D., Barbas-Bernardos, C., García, A., et al. (2018) Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. *Journal of Pharmaceutical and Biomedical Analysis*, 147: 149–173. doi:10.1016/j.jpba.2017.07.044.
- Dührkop, K., Fleischauer, M., Ludwig, M., et al. (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16 (4): 299–302. doi:10.1038/s41592-019-0344-8.
- Dunn, W.B. (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, 5 (1): 011001. doi:10.1088/1478-3975/5/1/011001.
- Dunn, W.B., Broadhurst, D., Begley, P., et al. (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 6 (7): 1060–1083. doi:10.1038/nprot.2011.335.
- Dunn, W.B., Broadhurst, D., Brown, M., et al. (2008) Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system. *Journal of Chromatography B*, 871 (2): 288–298. doi:10.1016/j.jchromb.2008.03.021.
- Dunn, W.B. and Ellis, D.I. (2005) Metabolomics: Current analytical platforms and methodologies. *TrAC - Trends in Analytical Chemistry*, 24 (4): 285–294. doi:10.1016/j.trac.2004.11.021.
- Dunn, W.B., Erban, A., Weber, R.J.M., et al. (2013) Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9 (SUPPL.1): 44–66. doi:10.1007/s11306-012-0434-4.

- Fanali, S. (2017) An overview to nano-scale analytical techniques: Nano-liquid chromatography and capillary electrochromatography. *Electrophoresis*, 38 (15) pp. 1822–1829. doi:10.1002/elps.201600573.
- Fenn, J.B., Mann, M., Meng, C.K., et al. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246 (4926) pp. 64–71. doi:10.1126/science.2675315.
- Folch, J., Lees, M. and Sloane Stanley, G.H. (1957) A simple method for the isolation and purification of total lipides from animal tissues. *The Journal of biological chemistry*, 226 (1): 497–509. doi:10.3989/scimar.2005.69n187.
- Frainay, C., Schymanski, E., Neumann, S., et al. (2018) Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas. *Metabolites*, 8 (3): 51. doi:10.3390/metabo8030051.
- Galton, F. (1886) Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15: 246. doi:10.2307/2841583.
- García-Villalba, R., Carrasco-Pancorbo, A., Zurek, G., et al. (2010) Nano and rapid resolution liquid chromatography-electrospray ionization-time of flight mass spectrometry to identify and quantify phenolic compounds in olive oil. *Journal of Separation Science*, 33 (14): 2069–2078. doi:10.1002/jssc.201000184.
- García, C.J., Gil, M.I. and Tomas-Barberan, F.A. (2018) LC–MS untargeted metabolomics reveals early biomarkers to predict browning of fresh-cut lettuce. *Postharvest Biology and Technology*, 146 (May): 9–17. doi:10.1016/j.postharvbio.2018.07.011.
- Gates, P.J (2014) *Electrospray Ionisation (ESI)*. Available at: <http://www.chm.bris.ac.uk/ms/esi-ionisation.xhtml> (Accessed: July 2019)
- Geiger, T., Cox, J. and Mann, M. (2010) Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation* □ S. *Molecular & Cellular Proteomics*, 9: 2252–2261. doi:10.1074/mcp.M110.001537.
- Giacomoni, F., Le Corguille, G., Monsoor, M., et al. (2015) Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics*, 31 (9): 1493–1495. doi:10.1093/bioinformatics/btu813.
- Gil de la Fuente, A., Godzien, J., Fernández López, M., et al. (2018) Knowledge-based metabolite annotation tool: CEU Mass Mediator. *Journal of Pharmaceutical and Biomedical Analysis*, 154: 138–

149. doi:10.1016/j.jpba.2018.02.046.

Gillet, L.C., Navarro, P., Tate, S., et al. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Molecular and Cellular Proteomics*, 11 (6): O111.016717. doi:10.1074/mcp.O111.016717.

Godzien, J., Alonso-Herranz, V., Barbas, C., et al. (2015) Controlling the quality of metabolomics data: new strategies to get the best out of the QC sample. *Metabolomics*, 11 (3): 518–528. doi:10.1007/s11306-014-0712-4.

Gong, Z.G., Hu, J., Wu, X., et al. (2017) The Recent Developments in Sample Preparation for Mass Spectrometry-Based Metabolomics. *Critical Reviews in Analytical Chemistry*. 47 (4) pp. 325–331. doi:10.1080/10408347.2017.1289836.

Goodacre, R. and Kell, D.B. (1996) Rapid identification of streptococcus and enterococcus species using diffuse reflectance-absorbance fourier transform infrared spectroscopy and artificial neural networks". *FEMS Microbiology Letters*, 364 (10): 1–4. doi:10.1093/femsle/fnx018.

de Graaf, E.L., Maarten Altelaar, A.F., van Breukelen, B., et al. (2011) Improving SRM Assay Development: A Global Comparison between Triple Quadrupole, Ion Trap, and Higher Energy CID Peptide Fragmentation Spectra. *J. Proteome Res*, 10: 19. doi:10.1021/pr200156b.

Griffiths, J. (2008) A brief history of mass spectrometry. *Analytical Chemistry*. 80 (15) pp. 5678–5683. doi:10.1021/ac8013065.

Gromski, P.S., Muhamadali, H., Ellis, D.I., et al. (2015) A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, 879: 10–23. doi:10.1016/j.aca.2015.02.012.

Gross, J.H. (2014) Direct analysis in real time-a critical review on DART-MS. *Analytical and Bioanalytical Chemistry*. 406 (1) pp. 63–80. doi:10.1007/s00216-013-7316-0.

Di Guida, R., Engel, J., Allwood, J.W., et al. (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*, 12 (5): 93. doi:10.1007/s11306-016-1030-9.

Guijas, C., Montenegro-Burke, J.R., Domingo-Almenara, X., et al. (2018) METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Analytical Chemistry*, 90 (5): 3156–3164. doi:10.1021/acs.analchem.7b04424.

- Gürdeniz, G., Kristensen, M., Skov, T., et al. (2012) The effect of LC-MS data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats. *Metabolites*, 2 (1): 77–99. doi:10.3390/metabo2010077.
- Haddad, I., Hiller, K., Frimmersdorf, E., et al. (2009) An emergent self-organizing map based analysis pipeline for comparative metabolome studies. *In Silico Biology*, 9 (4): 163–178. doi:10.3233/ISB-2009-0396.
- Han, X., Rozen, S., Boyle, S.H., et al. (2011) Metabolomics in early Alzheimer’s disease: Identification of altered plasma sphingolipidome using shotgun lipidomics. *PLoS ONE*, 6 (7). doi:10.1371/journal.pone.0021643.
- Harrington, R.A., Adhikari, V., Rayner, M., et al. (2019) Nutrient composition databases in the age of big data: FoodDB, a comprehensive, real-time database infrastructure. *BMJ Open*, 9 (6). doi:10.1136/bmjopen-2018-026652.
- Harvey, C.J.B., Tang, M., Schlecht, U., et al. (2018) HEx: A heterologous expression platform for the discovery of fungal natural products. *Science Advances*, 4 (4). doi:10.1126/sciadv.aar5459.
- Heinonen, M., Shen, H., Zamboni, N., et al. (2012) Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28 (18): 2333–2341. doi:10.1093/bioinformatics/bts437.
- Herrgård, M.J., Swainston, N., Dobson, P., et al. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol*, 26 (10): 1155–1160. doi:10.1038/nbt1492.
- Higashi, H., Tokumi, T., Hogan, C.J., et al. (2015) Simultaneous ion and neutral evaporation in aqueous nanodrops: Experiment, theory, and molecular dynamics simulations. *Physical Chemistry Chemical Physics*, 17 (24): 15746–15755. doi:10.1039/c5cp01730k.
- HighChem (2019) Available at: www.mzcloud.org (Accessed: July 2019)
- Hill, C.B., Czauderna, T., Klapperstück, M., et al. (2015) Metabolomics, Standards, and Metabolic Modeling for Synthetic Biology in Plants. *Frontiers in Bioengineering and Biotechnology*, 3 (October): 167. doi:10.3389/fbioe.2015.00167.
- Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9 (7): 811–818. doi:10.1002/sim.4780090710.

Hoffmann, E. de. and Stroobant, V. (2007) *Mass spectrometry : principles and applications*. J. Wiley. Available at: <https://www.wiley.com/en-us/Mass+Spectrometry%3A+Principles+and+Applications%2C+3rd+Edition-p-9780470033104> (Downloaded: 13 August 2019).

van der Hooft, J.J.J., Wandy, J., Barrett, M.P., et al. (2016) Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, p. 201608041. doi:10.1073/pnas.1608041113.

Hopfgartner, G., Tonoli, D. and Varesio, E. (2012) High-resolution mass spectrometry for integrated qualitative and quantitative analysis of pharmaceuticals in biological matrices. *Analytical and Bioanalytical Chemistry*, 402 (8): 2587–2596. doi:10.1007/s00216-011-5641-8.

Horai, H., Arita, M., Kanaya, S., et al. (2010) MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45 (7): 703–714. doi:10.1002/jms.1777.

Houjou, T., Yamatani, K., Imagawa, M., et al. (2005) A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, 19 (5): 654–666. doi:10.1002/rcm.1836.

Housley, L., Magana, A.A., Hsu, A., et al. (2018) Untargeted Metabolomic Screen Reveals Changes in Human Plasma Metabolite Profiles Following Consumption of Fresh Broccoli Sprouts. *Molecular Nutrition and Food Research*, 62 (19): 1–6. doi:10.1002/mnfr.201700665.

Hu, Q., Noll, R.J., Li, H., et al. (2005) The Orbitrap: A new mass spectrometer. *Journal of Mass Spectrometry*, 40 (4): 430–443. doi:10.1002/jms.856.

Hu, T., Oksanen, K., Zhang, W., et al. (2018) “Analyzing Feature Importance for Metabolomics Using Genetic Programming.” In Castelli, M., Sekanina, L., Zhang, M., et al. (eds.). *Genetic Programming*. Cham, 2018. Springer International Publishing. pp. 68–83.

Hufsky, F., Scheubert, K. and Böcker, S. (2014) New kids on the block: novel informatics methods for natural product discovery. *Natural product reports*, 31 (6): 807–817. doi:10.1039/c3np70101h.

Humston, E.M., Dombek, K.M., Tu, B.P., et al. (2011) Toward a global analysis of metabolites in regulatory mutants of yeast. *Analytical and Bioanalytical Chemistry*, 401 (8): 2387–2402. doi:10.1007/s00216-011-4800-2.

Ichiki, K. and Consta, S. (2006) Disintegration mechanisms of charged aqueous nanodroplets studied

by simulations and analytical models. *Journal of Physical Chemistry B*, 110 (39): 19168–19175. doi:10.1021/jp062222a.

Islam, T. (2013) *Generation and use of ceramic uorapatite binding peptides for the Novel strategies for the purification of biomolecules by affinity chromatography*. Tanta University. Available at: https://www.researchgate.net/figure/2-Schematic-diagram-of-a-high-performance-liquid-chromatography_fig2_334152072.

Jaeger, C., Méret, M., Schmitt, C.A., et al. (2017) Compound annotation in liquid chromatography/high-resolution mass spectrometry based metabolomics: robust adduct ion determination as a prerequisite to structure prediction in electrospray ionization mass spectra. *Rapid Communications in Mass Spectrometry*, 31 (15): 1261–1266. doi:10.1002/rcm.7905.

Jansson, J., Willing, B., Lucio, M., et al. (2009) Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS ONE*, 4 (7). doi:10.1371/journal.pone.0006386.

Jeffryes, J.G., Colastani, R.L., Elbadawi-Sidhu, M., et al. (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *Journal of cheminformatics*, 7 (1): 44. doi:10.1186/s13321-015-0087-1.

Jewison, T., Su, Y., Disfany, F.M., et al. (2014) SMPDB 2.0: Big improvements to the small molecule pathway database. *Nucleic Acids Research*, 42 (D1). doi:10.1093/nar/gkt1067.

Johnson, C.H., Ivanisevic, J. and Siuzdak, G. (2016) Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17 (7): 451–459. doi:10.1038/nrm.2016.25.

Kachman, M., Habra, H., Duren, W., et al. (2019) Deep annotation of untargeted LC-MS metabolomics data with Binner Kelso, J. (ed.). *Bioinformatics*. doi:10.1093/bioinformatics/btz798.

Kale, N.S., Haug, K., Conesa, P., et al. (2016) MetaboLights: An open-access database repository for metabolomics data. *Current Protocols in Bioinformatics*, 2016 (1): 14.13.1-14.13.18. doi:10.1002/0471250953.bi1413s53.

Kanehisa, M., Goto, S., Sato, Y., et al. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40 (D1). doi:10.1093/nar/gkr988.

Kang, W.-Y., Thompson, P.T., El-Amouri, S.S., et al. (2019) Improved segmented-scan spectral stitching for stable isotope resolved metabolomics (SIRM) by ultra-high-resolution Fourier transform mass spectrometry. *Analytica Chimica Acta*, 1080: 104–115. doi:10.1016/j.aca.2019.06.019.

- Kantz, E.D., Tiwari, S., Watrous, J.D., et al. (2019) Deep Neural Networks for Classification of LC-MS Spectral Peaks. *Analytical Chemistry*, 91 (19): 12407–12413. doi:10.1021/acs.analchem.9b02983.
- Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60 (20): 2299–2301. doi:10.1021/ac00171a028.
- Kell, D.B. and Oliver, S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, 26 (1): 99–105. doi:10.1002/bies.10385.
- Kelstrup, C.D., Jersie-Christensen, R.R., Batth, T.S., et al. (2014) Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field orbitrap mass spectrometer. *Journal of Proteome Research*, 13 (12): 6187–6195. doi:10.1021/pr500985w.
- Kessner, D., Chambers, M., Burke, R., et al. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24 (21): 2534–2536. doi:10.1093/bioinformatics/btn323.
- Kido Soule, M.C., Longnecker, K., Johnson, W.M., et al. (2015) Environmental metabolomics: Analytical strategies. *Marine Chemistry*, 177: 374–387. doi:10.1016/j.marchem.2015.06.029.
- Kind, T. and Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8: 105. doi:10.1186/1471-2105-8-105.
- Kind, T., Liu, K.-H., Lee, D.Y., et al. (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nature Methods*, 10 (8): 755–758. doi:10.1038/nmeth.2551.
- Kind, T., Wohlgemuth, G., Lee, D.Y., et al. (2009) FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry*, 81 (24): 10038–10048. doi:10.1021/ac9019522.
- Kingdon, K.H. (1923) A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Physical Review*, 21 (4): 408–418. doi:10.1103/PhysRev.21.408.
- Koek, M.M., Jellema, R.H., van der Greef, J., et al. (2011) Quantitative metabolomics based on gas chromatography mass spectrometry: Status and perspectives. *Metabolomics*. 7 (3) pp. 307–328. doi:10.1007/s11306-010-0254-3.
- Konermann, L., Ahadi, E., Rodriguez, A.D., et al. (2013) Unraveling the mechanism of electrospray

ionization. *Analytical Chemistry*, 85 (1): 2–9. doi:10.1021/ac302789c.

Kopka, J., Schauer, N., Krueger, S., et al. (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, 21 (8): 1635–1638. doi:10.1093/bioinformatics/bti236.

Kuhl, C., Tautenhahn, R., Böttcher, C., et al. (2012) CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84 (1): 283–289. doi:10.1021/ac202450g.

Labowsky, M., Fenn, J.B. and Fernandez de la Mora, J. (2000) A continuum model for ion evaporation from a drop: effect of curvature and charge on ion solvation energy. *Analytica Chimica Acta*, 406 (1): 105–118. doi:10.1016/S0003-2670(99)00595-4.

Lange, O., Damoc, E., Wieghaus, A., et al. (2015) Reprint of “enhanced fourier transform for orbitrap mass spectrometry.” *International Journal of Mass Spectrometry*, 377 (1): 338–344. doi:10.1016/j.ijms.2014.07.040.

Laponogov, I., Sadawi, N., Galea, D., et al. (2018) ChemDistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics*, (June): 1–7. doi:10.1093/bioinformatics/bty080.

Law, V., Knox, C., Djoumbou, Y., et al. (2014) DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research*, 42 (D1): 1091–1097. doi:10.1093/nar/gkt1068.

Law, W.S., Wang, R., Hu, B., et al. (2010) On the mechanism of extractive electrospray ionization. *Analytical Chemistry*, 82 (11): 4494–4500. doi:10.1021/ac100390t.

Lawson, T.N., Weber, R.J.M., Jones, M.R., et al. (2017) MsPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Analytical Chemistry*, 89 (4): 2432–2439. doi:10.1021/acs.analchem.6b04358.

Leal-Witt, M.J., Ramon-Krauel, M., Samino, S., et al. (2017) Untargeted metabolomics identifies a plasma sphingolipid-related signature associated with lifestyle intervention in prepubertal children with obesity. *Nature Publishing Group*, 42: 72–78. doi:10.1038/ijo.2017.201.

Lerma-Ortiz, C., Jeffryes, J.G., Cooper, A.J.L., et al. (2016) Nothing of chemistry disappears in biology': The Top 30 damage-prone endogenous metabolites. *Biochemical Society Transactions*, 44 (3): 961–971. doi:10.1042/BST20160073.

Li, B., Tang, J., Yang, Q., et al. (2016a) Performance Evaluation and Online Realization of Data-driven Normalization Methods Used in LC/MS based Untargeted Metabolomics Analysis. *Scientific Reports*,

6 (1): 38881. doi:10.1038/srep38881.

Li, H., Cai, Y., Guo, Y., et al. (2016b) MetDIA: Targeted Metabolite Extraction of Multiplexed MS/MS Spectra Generated by Data-Independent Acquisition. *Analytical Chemistry*, p. acs.analchem.6b02122. doi:10.1021/acs.analchem.6b02122.

Li, S., Todor, A. and Luo, R. (2016c) Blood transcriptomics and metabolomics for personalized medicine. *Computational and Structural Biotechnology Journal*, 14: 1–7. doi:10.1016/j.csbj.2015.10.005.

Libiseller, G., Dvorzak, M., Kleb, U., et al. (2015) IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics*, 16 (1): 118. doi:10.1186/s12859-015-0562-8.

Lin, S.M., Du, P., Huber, W., et al. (2008) Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Research*, 36 (2): 11. doi:10.1093/nar/gkm1075.

De Livera, A.M., Dias, D.A., De Souza, D., et al. (2012) Normalizing and integrating metabolomics data. *Analytical Chemistry*, 84 (24): 10768–10776. doi:10.1021/ac302748b.

Lommen, A. (2009) MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry*, 81 (8): 3079–3086. doi:10.1021/ac900036d.

Ludwig, C., Gillet, L., Rosenberger, G., et al. (2018) Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 14 (8): 8126. doi:10.15252/msb.20178126.

Lynn, K.S., Cheng, M.L., Chen, Y.R., et al. (2015) Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. *Analytical Chemistry*, 87 (4): 2143–2151. doi:10.1021/ac503325c.

Ma, Y., Tanaka, N., Vaniya, A., et al. (2016) Ultrafast Polyphenol Metabolomics of Red Wines Using MicroLC-MS/MS. *Journal of Agricultural and Food Chemistry*, 64 (2): 505–512. doi:10.1021/acs.jafc.5b04890.

Mahadevan, S., Shah, S.L., Marrie, T.J., et al. (2008) Analysis of metabolomic data using support vector machines. *Analytical Chemistry*, 80 (19): 7562–7570. doi:10.1021/ac800954c.

Mahieu, N.G. and Patti, G.J. (2017) Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Analytical Chemistry*, 89 (19): 10397–

10406. doi:10.1021/acs.analchem.7b02380.

Mahieu, N.G., Spalding, J.L., Gelman, S.J., et al. (2016) Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. *Analytical Chemistry*, 88 (18): 9037–9046. doi:10.1021/acs.analchem.6b01702.

Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, 72 (6): 1156–1162. doi:10.1021/ac991131p.

Makarov, A., Denisov, E., Kholomeev, A., et al. (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical Chemistry*, 78 (7): 2113–2120. doi:10.1021/ac0518811.

Makarov, A., Denisov, E. and Lange, O. (2009) Performance evaluation of a high-field orbitrap mass analyzer. *Journal of the American Society for Mass Spectrometry*, 20 (8): 1391–1396. doi:10.1016/j.jasms.2009.01.005.

Marchi, I., Rudaz, S. and Veuthey, J. (2009) Atmospheric pressure photoionization for coupling liquid-chromatography to mass spectrometry: A review. *Talanta*, 78 (1): 1–18. doi:10.1016/j.talanta.2008.11.031.

Marginean, I., Tang, K., Smith, R.D., et al. (2014) Picoelectrospray ionization mass spectrometry using narrow-bore chemically etched emitters. *Journal of the American Society for Mass Spectrometry*, 25 (1): 30–36. doi:10.1007/s13361-013-0749-z.

Markley, J.L., Brüschweiler, R., Edison, A.S., et al. (2017) The future of NMR-based metabolomics. *Current Opinion in Biotechnology*. 43 pp. 34–40. doi:10.1016/j.copbio.2016.08.001.

Martin, A.J.P. and Synge, R.L.M. (1941) A new form of chromatogram employing two liquid phases. *Trends in Biochemical Sciences*, 2 (11): N245. doi:10.1016/0968-0004(77)90204-3.

Mathur, R. and O'Connor, P.B. (2009) Artifacts in Fourier transform mass spectrometry. *Rapid Communications in Mass Spectrometry*, 23 (4): 523–529. doi:10.1002/rcm.3904.

Matyash, V., Liebisch, G., Kurzchalia, T. V., et al. (2008) Lipid extraction by methyl-terf-butyl ether for high-throughput lipidomics. *Journal of Lipid Research*, 49 (5): 1137–1146. doi:10.1194/jlr.D700041-JLR200.

Max Planck Institute (no date) Available at: <http://www.mpip-mainz.mpg.de/5230139/original-1518448538.jpg?t=eyJ3aWR0aCI6ODQwLCJvYmpfaWQiOiUyMzAxMzI9--db56ecba9ee8f84a000be5b09407dea376331112> (Accessed July 2019)

McLafferty, F.W. (1962a) Mass Spectrometric Analysis. Aliphatic Halogenated Compounds. *Analytical Chemistry*, 34 (1): 2–15. doi:10.1021/ac60181a003.

McLafferty, F.W. (1962b) Mass Spectrometric Analysis. Aliphatic Nitriles. *Analytical Chemistry*, 34 (1): 26–30. doi:10.1021/ac60181a005.

McLafferty, F.W. (1962c) Mass Spectrometric Analysis. Aromatic Halogenated Compounds. *Analytical Chemistry*, 34 (1): 16–25. doi:10.1021/ac60181a004.

Melnikov, A.D., Tsentalovich, Y.P. and Yanshole, V. V (2020) Deep Learning for the Precise Peak Detection in High-Resolution LC-MS Data. *Analytical Chemistry*, 92 (1): 588–592. doi:10.1021/acs.analchem.9b04811.

Merrill, A.H., Stokes, T.H., Momin, A., et al. (2009) Sphingolipidomics: A valuable tool for understanding the roles of sphingolipids in biology and disease. *Journal of Lipid Research*. 50 (SUPPL.). doi:10.1194/jlr.R800073-JLR200.

Miladinović, S.M., Kozhinov, A.N., Tsybin, O.Y., et al. (2012) Sidebands in Fourier transform ion cyclotron resonance mass spectra. *International Journal of Mass Spectrometry*, 325–327: 10–18. doi:10.1016/j.ijms.2012.08.009.

Mitchell, J.M., Flight, R.M., Wang, Q.J., et al. (2018) New methods to identify high peak density artifacts in Fourier transform mass spectra and to mitigate their effects on high-throughput metabolomic data analysis. *Metabolomics*, 14 (10): 125. doi:10.1007/s11306-018-1426-9.

MoNA (2019) *Welcome to MoNA!* Available at: <https://mona.fiehnlab.ucdavis.edu> (Accessed: July 2019)

Mullard, G., Allwood, J.W., Weber, R., et al. (2014) A new strategy for MS/MS data acquisition applying multiple data dependent experiments on Orbitrap mass spectrometers in non-targeted metabolomic applications. *Metabolomics*, 11 (5): 1068–1080. doi:10.1007/s11306-014-0763-6.

Murray, K. (2020) *Mass Spectrometer Overview*. Available at: <https://www.pinterest.com.au/pin/388224430356469936/> (Accessed: 8 November 2020).

Myers, O.D., Sumner, S.J., Li, S., et al. (2017) Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal. Chem*, 89: 8695. doi:10.1021/acs.analchem.7b01069.

- Najdekr, L., Friedecký, D., Tautenhahn, R., et al. (2016) Influence of mass resolving power in orbital ion-trap mass spectrometry-based metabolomics. *Analytical Chemistry*, p. acs.analchem.6b02319. doi:10.1021/acs.analchem.6b02319.
- Nash, W.J. and Dunn, W.B. (2019) From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC - Trends in Analytical Chemistry*. doi:10.1016/j.trac.2018.11.022.
- Nazario, C.E.D., Silva, M.R., Franco, M.S., et al. (2015) Evolution in miniaturized column liquid chromatography instrumentation and applications: An overview. *Journal of Chromatography A*. 1421 pp. 18–37. doi:10.1016/j.chroma.2015.08.051.
- Neumann, S., Thum, A. and Böttcher, C. (2013) Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. *Metabolomics*, 9 (S1): 84–91. doi:10.1007/s11306-012-0401-0.
- Nicholson, J.K., Lindon, J.C. and Holmes, E. (1999) “Metabonomics”: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29 (11): 1181–1189. doi:10.1080/004982599238047.
- Nier, A.O. (1939) The isotopic constitution of uranium and the half-lives of the uranium isotopes. I. *Physical Review*, 55 (2): 150–153. doi:10.1103/PhysRev.55.150.
- NIST (2019) *NIST Standard Reference Database 1A v17*. Available at: <https://www.nist.gov/srd/nist-standard-reference-database-1a-v17> (Accessed: July 2019)
- O’Connor, A., Brasher, C.J., Slatter, D.A., et al. (2017) LipidFinder: A computational workflow for discovery of lipids identifies eicosanoid-phosphoinositides in platelets. *JCI Insight*, 2 (7): 1–18. doi:10.1172/jci.insight.91634.
- Oberacher, H., Whitley, G., Berger, B., et al. (2013) Testing an alternative search algorithm for compound identification with the “Wiley Registry of Tandem Mass Spectral Data, MSforID.” *Journal of Mass Spectrometry*, 48 (4): 497–504. doi:10.1002/jms.3185.
- Oliver, S.G., Winson, M.K., Kell, D.B., et al. (1998) Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16 (9): 373–378. doi:10.1016/S0167-7799(98)01214-1.
- Olsen, J. V., Macek, B., Lange, O., et al. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4 (9): 709–712. doi:10.1038/nmeth1060.

Osborn, M.P., Park, Y., Parks, M.B., et al. (2013) Metabolome-Wide Association Study of Neovascular Age-Related Macular Degeneration. *PLoS ONE*, 8 (8). doi:10.1371/journal.pone.0072737.

Paul, W., Reinhard, H.P. and von Zahn, U. (1958) Das elektrische Massenfiter als Massenspektrometer und Isotopentrenner. *Zeitschrift für Physik*, 152 (2): 143–182. doi:10.1007/BF01327353.

Pearson, K. (1895) VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58 (347–352): 240–242. doi:10.1098/rspl.1895.0041.

Pence, H.E. and Williams, A. (2010) Chemspider: An online chemical information resource. *Journal of Chemical Education*. 87 (11) pp. 1123–1124. doi:10.1021/ed100697w.

Peterson, A.C., Russell, J.D., Bailey, D.J., et al. (2012) Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Molecular & Cellular Proteomics*, 11 (11): 1475–1488. doi:10.1074/mcp.o112.020131.

Philip Bobko (2001) *Correlation and Regression: Applications for Industrial Organizational Psychology and Management*. 2nd ed. Sage Publications. Available at: [https://books.google.co.uk/books?id=CvDL7YdZos0C&pg=PR4&lpg=PR4&dq=Bobko,+P.+\(2001\).+Correlation+and+regression:+Applications+for+industrial+organizational+psychology+and+management+\(2nd+ed.\).+Thousand+Oaks,+CA:+Sage+Publications&source=bl&ots=ZOYTHMOavm&](https://books.google.co.uk/books?id=CvDL7YdZos0C&pg=PR4&lpg=PR4&dq=Bobko,+P.+(2001).+Correlation+and+regression:+Applications+for+industrial+organizational+psychology+and+management+(2nd+ed.).+Thousand+Oaks,+CA:+Sage+Publications&source=bl&ots=ZOYTHMOavm&) (Downloaded: 2 November 2020).

Plumb, R.S., Johnson, K.A., Rainville, P., et al. (2006) UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry*, 20 (13): 1989–1994. doi:10.1002/rcm.2550.

Pluskal, T., Castillo, S., Villar-Briones, A., et al. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, 11: 395. doi:10.1186/1471-2105-11-395.

Psychogios, N., Hau, D.D., Peng, J., et al. (2011) The Human Serum Metabolome. *PLoS ONE*, 6 (2): 16957. doi:10.1371/journal.

Purohit, P. V., Rocke, D.M., Viant, M.R., et al. (2004) Discrimination models using variance-stabilizing transformation of metabolomic NMR data. *OMICS A Journal of Integrative Biology*, 8 (2): 118–130. doi:10.1089/1536231041388348.

Rafiei, A. and Sleno, L. (2014) Comparison of peak-picking workflows for untargeted liquid

chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Communications in Mass Spectrometry*, 29 (1): 119–127. doi:10.1002/rcm.7094.

Rahman, M.M., Abd El-Aty, A.M., Kim, S.W., et al. (2017) Quick, easy, cheap, effective, rugged, and safe sample preparation approach for pesticide residue analysis using traditional detectors in chromatography: A review. *Journal of Separation Science*. 40 (1) pp. 203–212. doi:10.1002/jssc.201600889.

Rampler, E., Schoeny, H., Mitic, B.M., et al. (2018) Simultaneous non-polar and polar lipid analysis by on-line combination of HILIC, RP and high resolution MS. *The Analyst*, 143 (5): 1250–1258. doi:10.1039/C7AN01984J.

Ranninger, C., Schmidt, L.E., Rurik, M., et al. (2016) Improving global feature detectabilities through scan range splitting for untargeted metabolomics by high-performance liquid chromatography-Orbitrap mass spectrometry. *Analytica Chimica Acta*, 930: 13–22. doi:10.1016/j.aca.2016.05.017.

Renaud, J.B., Sabourin, L., Topp, E., et al. (2017) Spectral Counting Approach to Measure Selectivity of High-Resolution LC-MS Methods for Environmental Analysis. *Analytical Chemistry*, 89 (5): 2747–2754. doi:10.1021/acs.analchem.6b03475.

Ridder, L., van der Hooft, J.J.J. and Verhoeven, S. (2014) Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrometry*, 3 (Special_Issue_2): S0033–S0033. doi:10.5702/massspectrometry.s0033.

RIKEN (no date) Available at: http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/ (Accessed: 1st October 2018)

Rivas-Ubach, A., Liu, Y., Bianchi, T.S., et al. (2018) Moving beyond the van Krevelen Diagram: A New Stoichiometric Approach for Compound Classification in Organisms. *Analytical Chemistry*, 90 (10): 6152–6160. doi:10.1021/acs.analchem.8b00529.

Robertson, D.G. (2000) Metabonomics: Evaluation of Nuclear Magnetic Resonance (NMR) and Pattern Recognition Technology for Rapid in Vivo Screening of Liver and Kidney Toxicants. *Toxicological Sciences*, 57 (2): 326–337. doi:10.1093/toxsci/57.2.326.

Rodríguez-Suárez, E., Gonzalez, E., Hughes, C., et al. (2014) Quantitative proteomic analysis of hepatocyte-secreted extracellular vesicles reveals candidate markers for liver toxicity. *Journal of Proteomics*, 103: 227–240. doi:10.1016/j.jprot.2014.04.008.

Roede, J.R., Uppal, K., Park, Y., et al. (2013) Serum Metabolomics of Slow vs. Rapid Motor

Progression Parkinson's Disease: A Pilot Study. *PLoS ONE*, 8 (10): 1–11.

doi:10.1371/journal.pone.0077629.

Rogers, S., Scheltema, R.A., Girolami, M., et al. (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25 (4): 512–518.

doi:10.1093/bioinformatics/btn642.

Roszkowska, A., Yu, M., Bessonneau, V., et al. (2018) Tissue storage affects lipidome profiling in comparison to in vivo microsampling approach. *Scientific Reports*, 8 (1). doi:10.1038/s41598-018-25428-2.

Rustam, Y.H. and Reid, G.E. (2018) Analytical Challenges and Recent Advances in Mass Spectrometry Based Lipidomics. *Analytical Chemistry*. 90 (1) pp. 374–397. doi:10.1021/acs.analchem.7b04836.

Ruttkies, C., Schymanski, E.L., Wolf, S., et al. (2016) MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, 8 (1): 1–16. doi:10.1186/s13321-016-0115-9.

Saccenti, E., Hoefsloot, H.C.J., Smilde, A.K., et al. (2014) Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10 (3): 361–374. doi:10.1007/s11306-013-0598-6.

Sakurai, T., Yamada, Y., Sawada, Y., et al. (2013) PRIME Update: Innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant and Cell Physiology*, 54 (2). doi:10.1093/pcp/pcs184.

Sangster, T., Major, H., Plumb, R., et al. (2006) A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. *Analyst*. 131 (10) pp. 1075–1078. doi:10.1039/b604498k.

Santoemma, G. (2018) Recent methodologies for studying the soil organic matter. *Applied Soil Ecology*. 123 pp. 546–550. doi:10.1016/j.apsoil.2017.09.011.

Sato, Y., Suzuki, I., Nakamura, T., et al. (2012) Identification of a new plasma biomarker of Alzheimer's disease using metabolomics technology. *The Journal of Lipid Research*, 53 (3): 567–576. doi:10.1194/jlr.M022376.

Sawada, Y., Nakabayashi, R., Yamada, Y., et al. (2012) RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. *Phytochemistry*, 82: 38–45. doi:10.1016/j.phytochem.2012.07.007.

- Scheltema, R.A., Hauschild, J.P., Lange, O., et al. (2014) The Q exactive HF, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field orbitrap analyzer. *Molecular and Cellular Proteomics*, 13 (12): 3698–3708. doi:10.1074/mcp.M114.043489.
- Scheubert, K., Hufsky, F., Petras, D., et al. (2017) Significance estimation for large scale metabolomics annotations by spectral matching. *Nature Communications*, 8 (1). doi:10.1038/s41467-017-01318-5.
- Schiffman, C., Petrick, L., Perttula, K., et al. (2019) Filtering procedures for untargeted lc-ms metabolomics data. *BMC Bioinformatics*, 20 (1): 1–10. doi:10.1186/s12859-019-2871-9.
- Schymanski, E.L., Jeon, J., Gulde, R., et al. (2014) Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environmental Science & Technology*, 48 (4): 2097–2098. doi:10.1021/es5002105.
- Schymanski, E.L., Ruttkies, C., Krauss, M., et al. (2017) Critical Assessment of Small Molecule Identification 2016: automated methods. *Journal of Cheminformatics*, 9: 22. doi:10.1186/s13321-017-0207-1.
- Senan, O., Aguilar-Mogas, A., Navarro, M., et al. (2019) CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics*, 35 (March): 4089–4097. doi:10.1093/bioinformatics/btz207.
- Senko, M.W., Remes, P.M., Canterbury, J.D., et al. (2013) Novel parallelized quadrupole/linear ion trap/orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. *Analytical Chemistry*, 85 (24): 11710–11714. doi:10.1021/ac403115c.
- Šesták, J., Moravcová, D. and Kahle, V. (2015) Instrument platforms for nano liquid chromatography. *Journal of Chromatography A*. 1421 pp. 2–17. doi:10.1016/j.chroma.2015.07.090.
- Shao, C., Zhang, Y. and Sun, W. (2014) Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ Orbitrap Velos mass spectrometer. *Journal of Proteomics*, 109: 26–37. doi:10.1016/j.jprot.2014.06.012.
- Showalter, M.R., Cajka, T. and Fiehn, O. (2017) Epimetabolites: discovering metabolism beyond building and burning. *Current Opinion in Chemical Biology*, 36: 70–76. doi:10.1016/j.cbpa.2017.01.012.
- Shulaev, V. and Isaac, G. (2018) Supercritical fluid chromatography coupled to mass spectrometry –

A metabolomics perspective. *Journal of Chromatography B*, 1092: 499–505.

doi:10.1016/j.jchromb.2018.06.021.

da Silva, R.R., Dorrestein, P.C. and Quinn, R.A. (2015) Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, 112 (41): 201516878.

doi:10.1073/pnas.1516878112.

Silva, R.R., Jourdan, F., Salvanha, D.M., et al. (2014) ProbMetab: An R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics*, 30 (9): 1336–1337.

doi:10.1093/bioinformatics/btu019.

Slenter, D.N., Kutmon, M., Hanspers, K., et al. (2018) WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46 (D1): D661–D667. doi:10.1093/nar/gkx1064.

Smith, C.A., Want, E.J., O'Maille, G., et al. (2006) XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 78 (3): 779–787. doi:10.1021/ac051437y.

Southam, A.D., Payne, T.G., Cooper, H.J., et al. (2007) Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. *Analytical Chemistry*, 79 (12): 4595–4602. doi:10.1021/ac062446p.

Stoessel, D., Schulte, C., Teixeira dos Santos, M.C., et al. (2018) Promising metabolite profiles in the plasma and CSF of early clinical Parkinson's disease. *Frontiers in Aging Neuroscience*, 10 (MAR): 1–14. doi:10.3389/fnagi.2018.00051.

Sud, M., Fahy, E., Cotter, D., et al. (2007) LMSD: LIPID MAPS structure database. *Nucleic Acids Research*, 35 (SUPPL. 1): 527–532. doi:10.1093/nar/gkl838.

Sud, M., Fahy, E., Cotter, D., et al. (2015) Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44 (8). doi:10.1093/nar/gkv1042.

Sumner, L.W., Amberg, A., Barrett, D., et al. (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3 (3): 211–221. doi:10.1007/s11306-007-0082-2.

Sun, C., Zhao, Y.Y. and Curtis, J.M. (2014) Elucidation of phosphatidylcholine isomers using two dimensional liquid chromatography coupled in-line with ozonolysis mass spectrometry. *Journal of*

Chromatography A, 1351: 37–45. doi:10.1016/j.chroma.2014.04.069.

Swainston, N., Smallbone, K., Hefzi, H., et al. (2016) Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12 (7): 109. doi:10.1007/s11306-016-1051-4.

Szymańska, E., van Dorsten, F.A., Troost, J., et al. (2012) A lipidomic analysis approach to evaluate the response to cholesterol-lowering food intake. *Metabolomics*, 8 (5): 894–906. doi:10.1007/s11306-011-0384-2.

Theodoridis, G.A., Gika, H.G., Want, E.J., et al. (2012) Liquid chromatography–mass spectrometry based global metabolite profiling: A review. *Analytica Chimica Acta*, 711: 7–16. doi:10.1016/j.aca.2011.09.042.

Thermo Fisher Scientific (2019) *Q Exactive Plus Hybrid Quadrupole-Orbitrap Mass Spectrometer*. Available at: <https://planetOrbitrap.com/q-exactive-plus#tab:schematic> (Accessed: July 2019)

Thermo Fisher Scientific (2019) *Orbitrap ID-X Tribrid Mass Spectrometer System*. Available at: <https://planetOrbitrap.com/Orbitrap-id-x> (Accessed July 2019)

Thomson, B.A. and Iribarne, J. V. (1979) Field induced ion evaporation from liquid surfaces at atmospheric pressure. *The Journal of Chemical Physics*, 71 (11): 4451–4463. doi:10.1063/1.438198.

Thomson, J.J. (1912) XIX. Further experiments on positive rays. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 24 (140): 209–253. doi:10.1080/14786440808637325.

Tian, M., Xu, X., Liu, F., et al. (2018) Untargeted metabolomics reveals predominant alterations in primary metabolites of broccoli sprouts in response to pre-harvest selenium treatment. *Food Research International*, 111 (April): 205–211. doi:10.1016/j.foodres.2018.04.020.

Tikunov, Y.M., Laptinok, S., Hall, R.D., et al. (2012) MSClust: A tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics*, 8 (4): 714–718. doi:10.1007/s11306-011-0368-2.

Torras-Claveria, L., Berkov, S., Codina, C., et al. (2014) Metabolomic analysis of bioactive Amaryllidaceae alkaloids of ornamental varieties of *Narcissus* by GC-MS combined with k-means cluster analysis. *Industrial Crops and Products*, 56: 211–222. doi:10.1016/j.indcrop.2014.03.008.

Toya, Y. and Shimizu, H. (2013) Flux analysis and metabolomics for systematic metabolic engineering of microorganisms. *Biotechnology Advances*, 31 (6): 818–826. doi:10.1016/j.biotechadv.2013.05.002.

- Treutler, H., Tsugawa, H., Porzel, A., et al. (2016) Discovering regulated metabolite families in untargeted metabolomics studies. *Analytical Chemistry*, 88 (16): 8082–8090. doi:10.1021/acs.analchem.6b01569.
- Tsugawa, H., Cajka, T., Kind, T., et al. (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods*, 12 (6): 523–526. doi:10.1038/nmeth.3393.
- Tsugawa, H., Ikeda, K., Tanaka, W., et al. (2017) Comprehensive identification of sphingolipid species by in silico retention time and tandem mass spectral library. *Journal of Cheminformatics*, 9 (1): 19. doi:10.1186/s13321-017-0205-3.
- Tsugawa, H., Kind, T., Nakabayashi, R., et al. (2016) Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Analytical Chemistry*, 88 (16): 7946–7958. doi:10.1021/acs.analchem.6b00770.
- Tsugawa, H., Nakabayashi, R., Mori, T., et al. (2019) A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nature Methods*, 16 (4): 295–298. doi:10.1038/s41592-019-0358-2.
- Udayakumar, M., Chandar, D.P., Arun, N., et al. (2012) PMDB: Plant metabolome database-A metabolomic approach. *Medicinal Chemistry Research*, 21 (1): 47–52. doi:10.1007/s00044-010-9506-z.
- University of Birmingham (2019) *BlueBEAR (Linux HPC)*. Available at: <https://intranet.birmingham.ac.uk/it/teams/infrastructure/research/bear/bluebear/index.aspx> (Accessed: November 2019).
- Uppal, K., Walker, D.I. and Jones, D.P. (2017) xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Analytical Chemistry*, 89 (2): 1063–1067. doi:10.1021/acs.analchem.6b01214.
- Vaughan, A.A., Dunn, W.B., Allwood, J.W., et al. (2012) Liquid chromatography-mass spectrometry calibration transfer and metabolomics data fusion. *Analytical Chemistry*, 84 (22): 9848–9857. doi:10.1021/ac302227c.
- Vinaixa, M., Schymanski, E.L., Neumann, S., et al. (2016) Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC - Trends in Analytical Chemistry*, 78: 23–35. doi:10.1016/j.trac.2015.09.005.
- Vollmar, A.K.R., Rattray, N.J.W., Cai, Y., et al. (2019) Normalizing untargeted periconceptional urinary

metabolomics data: A comparison of approaches. *Metabolites*, 9 (10). doi:10.3390/metabo9100198.

Wang, H., Correa, E., Dunn, W.B., et al. (2013) Metabolomic analyses show that electron donor and acceptor ratios control anaerobic electron transfer pathways in *Shewanella oneidensis*. *Metabolomics*, 9 (3): 642–656. doi:10.1007/s11306-012-0488-3.

Wang, L., Su, B., Zeng, Z., et al. (2018) Ion-Pair Selection Method for Pseudotargeted Metabolomics Based on SWATH MS Acquisition and Its Application in Differential Metabolite Discovery of Type 2 Diabetes. *Analytical Chemistry*, 90 (19): 11401–11408. doi:10.1021/acs.analchem.8b02377.

Wang, L., Xing, X., Chen, L., et al. (2019a) Peak Annotation and Verification Engine for Untargeted LC-MS Metabolomics. *Analytical Chemistry*, 91 (3): 1838–1846. doi:10.1021/acs.analchem.8b03132.

Wang, M., Carver, J.J., Phelan, V. V, et al. (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, 34 (8): 828–837. doi:10.1038/nbt.3597.

Wang, R., Yin, Y. and Zhu, Z.J. (2019b) Advancing untargeted metabolomics using data-independent acquisition mass spectrometry technology. *Analytical and Bioanalytical Chemistry*. 411 (19) pp. 4349–4357. doi:10.1007/s00216-019-01709-1.

Wang, Y., Kora, G., Bowen, B.P., et al. (2014) MIDAS: A database-searching algorithm for metabolite identification in metabolomics. *Analytical Chemistry*, 86 (19): 9496–9503. doi:10.1021/ac5014783.

Wei, R., Wang, J., Su, M., et al. (2018) Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, 8 (1): 663. doi:10.1038/s41598-017-19120-0.

Westerhuis, J.A., van Velzen, E.J.J., Hoefsloot, H.C.J., et al. (2010) Multivariate paired data analysis: Multilevel PLSDA versus OPLSDA. *Metabolomics*, 6 (1): 119–128. doi:10.1007/s11306-009-0185-z.

Wild, C.P. (2005) Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiology Biomarkers & Prevention*, 14 (8): 1847–1850. doi:10.1158/1055-9965.EPI-05-0456.

William Allwood, J., Clarke, A., Goodacre, R., et al. (2010) Dual metabolomics: A novel approach to understanding plant-pathogen interactions. *Phytochemistry*, 71 (5–6): 590–597. doi:10.1016/j.phytochem.2010.01.006.

Wiseman, J.M., Ifa, D.R., Zhu, Y., et al. (2008) Desorption electrospray ionization mass spectrometry: Imaging drugs and metabolites in tissues. *Proceedings of the National Academy of Sciences*, 105 (47):

18120–18125. doi:10.1073/pnas.0801066105.

Wishart, D.S., Knox, C., Guo, A.C., et al. (2009) HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, 37 (SUPPL. 1): 603–610. doi:10.1093/nar/gkn810.

Xu, Y., Correa, E. and Goodacre, R. (2013) Integrating multiple analytical platforms and chemometrics for comprehensive metabolic profiling: application to meat spoilage detection. *Analytical and Bioanalytical Chemistry*, 405 (15): 5063–5074. doi:10.1007/s00216-013-6884-3.

Xu, Y. and Goodacre, R. (2012) Multiblock principal component analysis: An efficient tool for analyzing metabolomics data which contain two influential factors. *Metabolomics*, 8: 37–51. doi:10.1007/s11306-011-0361-9.

Xu, Y.F., Lu, W. and Rabinowitz, J.D. (2015) Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. *Analytical Chemistry*, 87 (4): 2273–2281. doi:10.1021/ac504118y.

Yan, L., Zhou, J., Wang, D., et al. (2018) Unbiased lipidomic profiling reveals metabolomic changes during the onset and antipsychotics treatment of schizophrenia disease. *Metabolomics*, 14 (6): 1–13. doi:10.1007/s11306-018-1375-3.

Yan, Z., Li, T., Wei, B., et al. (2019) High-resolution MS/MS metabolomics by data-independent acquisition reveals urinary metabolic alteration in experimental colitis. *Metabolomics*, 15 (5): 70. doi:10.1007/s11306-019-1534-1.

Yang, Q., Wang, H., Maas, J.D., et al. (2012) Paper spray ionization devices for direct, biomedical analysis using mass spectrometry. *International Journal of Mass Spectrometry*, 312: 201–207. doi:10.1016/j.ijms.2011.05.013.

Yin, Y., Wang, R., Cai, Y., et al. (2019) *DecoMetDIA: Deconvolution of Multiplexed MS/MS Spectra for Metabolite Identification in SWATH-MS based Untargeted Metabolomics*. doi:10.1021/acs.analchem.9b02655.

Zheng, X., Wojcik, R., Zhang, X., et al. (2017) Coupling Front-End Separations, Ion Mobility Spectrometry, and Mass Spectrometry For Enhanced Multidimensional Biological and Environmental Analyses. *Annual Review of Analytical Chemistry*, 10 (1): 71–92. doi:10.1146/annurev-anchem-061516-045212.

Zhou, B., Xiao, J.F., Tuli, L., et al. (2012) LC-MS-based metabolomics. *Mol. BioSyst.*, 8 (2): 470–481. doi:10.1039/C1MB05350G.

Zhou, J., Li, Y., Chen, X., et al. (2017) Development of data-independent acquisition workflows for metabolomic analysis on a quadrupole-orbitrap platform. *Talanta*, 164 (November 2016): 128–136. doi:10.1016/j.talanta.2016.11.048.

Zhou, J., Weber, R.J.M., Allwood, J.W., et al. (2014) HAMMER: Automated operation of mass frontier to construct in silico mass spectral fragmentation libraries. *Bioinformatics*, 30 (4): 581–583. doi:10.1093/bioinformatics/btt711.

Zhu, X., Chen, Y. and Subramanian, R. (2014) Comparison of information-dependent acquisition, SWATH, and MS All techniques in metabolite identification study employing ultrahigh-performance liquid chromatography-quadrupole time-of-flight mass spectrometry. *Analytical Chemistry*, 86 (2): 1202–1209. doi:10.1021/ac403385y.

Zubarev, R.A. and Makarov, A. (2013) Orbitrap Mass Spectrometry. *Anal Chem*, 85 (11): 5288–5296. doi:10.1021/ac4001223.

9.0 Appendix

9.1 Characterising the complexity of electrospray ionisation data and its impact on metabolite annotation in UPLC-MS metabolomics studies

Table 58: The different adducts searched for in positive ion mode for different metabolite annotation resources. A y indicates that software considers that adduct and an x means that the adduct is not considered by the software.

Adduct	Type	Charge	HMDB	METLIN	MS-DIAL	CD3.0
M+H	Adduct	1	y	y	y	y
M+H-2H ₂ O	Adduct/(Fragment/Transformation)	1	y	y	y	x
M+H-H ₂ O	Adduct/(Fragment/Transformation)	1	y	y	y	y
M+NH ₄ -H ₂ O	Adduct/(Fragment/Transformation)	1	y	x	x	x
M+Li	Adduct	1	y	y	y	x
M+NH ₄	Adduct	1	y	y	y	y
M+Na	Adduct	1	y	y	y	y
M+CH ₃ OH+H	Adduct/(Fragment/Transformation)	1	y	y	y	y
M+K	Adduct	1	y	y	y	y
M+ACN+H	Adduct/(Fragment/Transformation)	1	y	y	y	y
M+2Na-H	Adduct/(Fragment/Transformation)	1	y	y	y	x
M+IsoProp+H	Adduct/(Fragment/Transformation)	1	y	x	y	x
M+ACN+Na	Adduct/(Fragment/Transformation)	1	y	y	y	y
M+2K-H	Adduct/(Fragment/Transformation)	1	y	x	y	x
M+DMSO+H	Adduct/(Fragment/Transformation)	1	y	x	y	y
M+2ACN+H	Adduct/(Fragment/Transformation)	1	y	x	y	x
M+IsoProp+Na+H	Adduct/Adduct/Multiply Charged/(Fragment/Transformation)	2	y	x	y	x
M+H+HCOONa	Adduct/(Fragment/Transformation)	1	y	x	x	x
2M+H	Dimer/Adduct	1	y	x	y	y
2M+NH ₄	Dimer/Adduct	1	y	x	y	y
2M+Na	Dimer/Adduct	1	y	x	y	y
2M+2H+3H ₂ O	Dimer/Adduct/Multiply Charged/(Fragment/Transformation)	2	y	x	y	x

2M+K	Dimer/Adduct	1	y	x	y	y
2M+ACN+H	Dimer/Adduct/(Fragment/Transformation)	1	y	x	y	y
2M+ACN+Na	Dimer/Adduct/(Fragment/Transformation)	1	y	x	y	y
2M+H-H ₂ O	Dimer/Adduct/(Fragment/Transformation)	1	y	x	x	x
M+2H	Adduct/Multiply Charged	2	y	y	y	y
M+H+NH ₄	Adduct/Adduct/Multiply Charged	2	y	x	y	y
M+H+Na	Adduct/Adduct/Multiply Charged	2	y	y	y	y
M+H+K	Adduct/Adduct/Multiply Charged	2	y	x	y	y
M+ACN+2H	Adduct/Multiply Charged/(Fragment/Transformation)	2	y	x	y	y
M+2Na	Adduct/Multiply Charged	2	y	y	y	x
M+2ACN+2H	Adduct/Multiply Charged/(Fragment/Transformation)	2	y	x	y	x
M+3ACN+2H	Adduct/Multiply Charged/(Fragment/Transformation)	2	y	x	y	x
M+3H	Adduct/Multiply Charged	3	y	y	y	y
M+2H+Na	Adduct/Adduct/Multiply Charged	3	y	y	y	x
M+H+2Na	Adduct/Adduct/Multiply Charged	3	y	y	y	x
M+3Na	Adduct/Multiply Charged	3	y	x	y	x
M+H+2K	Adduct/Adduct/Multiply Charged	3	y	x	x	x
M-C ₆ H ₁₀ O ₄ +H	Adduct/(Fragment/Transformation)	1	x	x	y	x
M-C ₆ H ₁₀ O ₅ +H	Adduct/(Fragment/Transformation)	1	x	x	y	x
M-C ₆ H ₈ O ₆ +H	Adduct/(Fragment/Transformation)	1	x	x	y	x
M+H-NH ₃	Adduct/(Fragment/Transformation)	1	x	x	x	y

Table 59: The different adducts searched for in positive ion mode for different metabolite annotation resources. A y indicates that software considers that adduct and an x means that the adduct is not considered by the software.

Adduct	Type	Charge	HMDB	METLIN	MS-DIAL	CD3.0
M-H	Adduct	1	y	y	y	y
M-H ₂ O-H	Adduct/(Fragment/Transformation)	1	y	y	y	y
M+F	Adduct	1	y	y	x	x
M+Na-2H	Adduct/(Fragment/Transformation)	1	y	y	y	x
M+Cl	Adduct	1	y	y	y	y
M+K-2H	Adduct/(Fragment/Transformation)	1	y	y	y	y
M+FA-H	Adduct/(Fragment/Transformation)	1	y	y	y	y
M+Hac-H	Adduct/(Fragment/Transformation)	1	y	x	y	y
M+Br	Adduct	1	y	x	y	x
M+TFA-H	Adduct/(Fragment/Transformation)	1	y	x	y	y
M-H+HCOONa	Adduct/(Fragment/Transformation)	1	y	x	x	x
2M-H	Dimer/Adduct	1	y	x	y	y
2M+FA-H	Dimer/Adduct/(Fragment/Transformation)	1	y	x	y	y
2M+Hac-H	Dimer/Adduct/(Fragment/Transformation)	1	y	x	y	y
3M-H	Dimer/Adduct	1	y	x	y	x
M-2H	Adduct/Multiply Charged	2	y	y	y	y
M-3H	Adduct/Multiply Charged	3	y	y	y	x
M+CH ₃ COO-H	Adduct/(Fragment/Transformation)	1	x	y	x	x
M-C ₆ H ₁₀ O ₄ -H	Adduct/(Fragment/Transformation)	1	x	x	y	x
M-C ₆ H ₁₀ O ₅ -H	Adduct/(Fragment/Transformation)	1	x	x	y	x
M-C ₆ H ₈ O ₆ -H	Adduct/(Fragment/Transformation)	1	x	x	y	x

9.2 Characterisation of UHPLC-MS full scan data complexity and its influence on MS² data collection on Q Exactive mass spectrometers

9.2.1 XCMS Parameter Optimisation

9.2.1.1 Parameters by UHPLC Assay

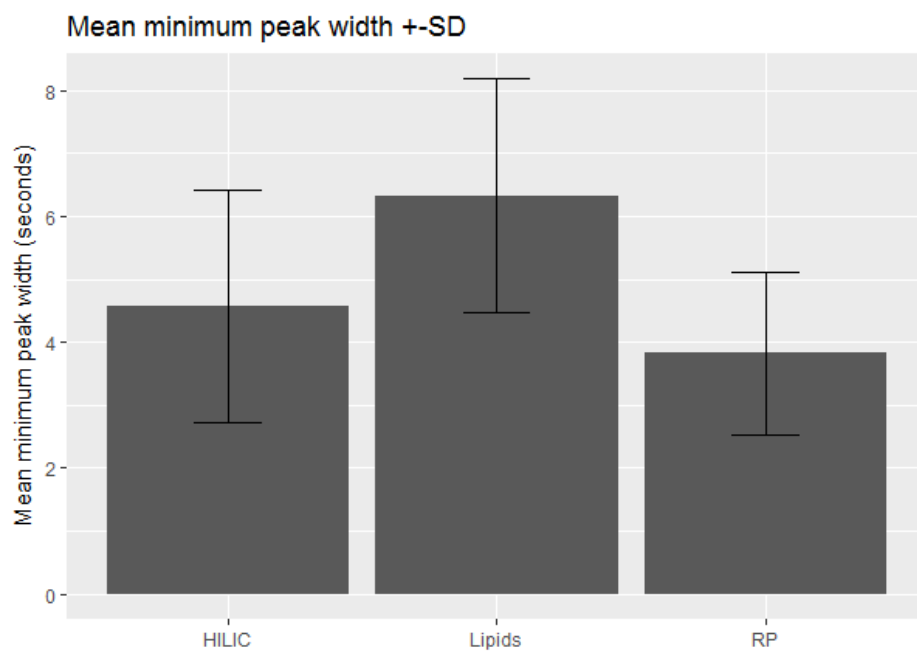


Figure 112: Mean minimum peak width \pm SD for all triplicates analysed with each chromatographic assay.

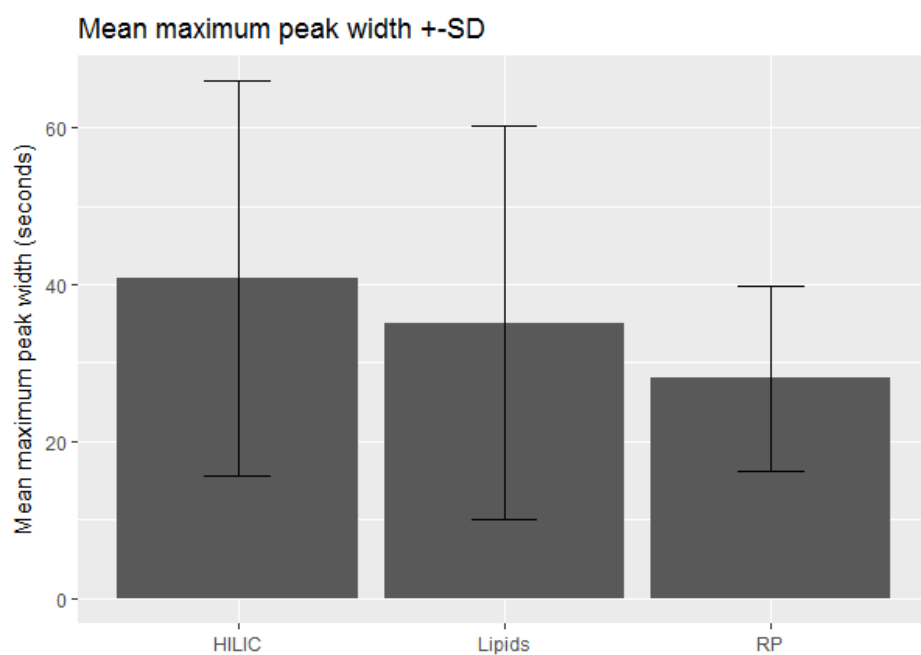


Figure 113: Mean maximum peak width \pm SD for all triplicates analysed with each chromatographic assay.

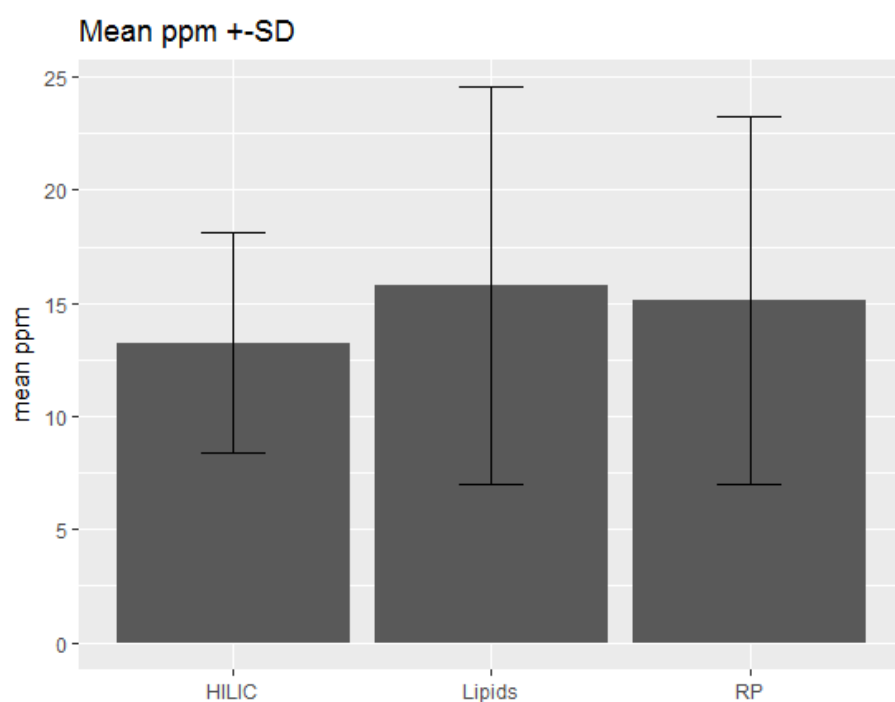


Figure 114: Mean ppm \pm SD for all triplicates analysed with each chromatographic assay.

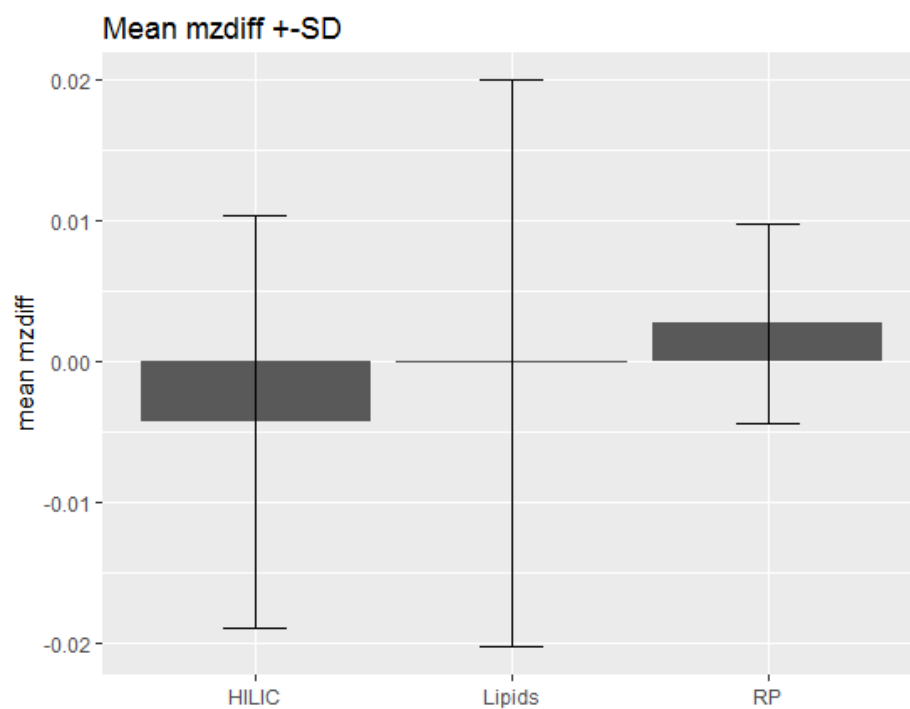


Figure 115: Mean mzdifff \pm SD for all triplicates analysed with each chromatographic assay.

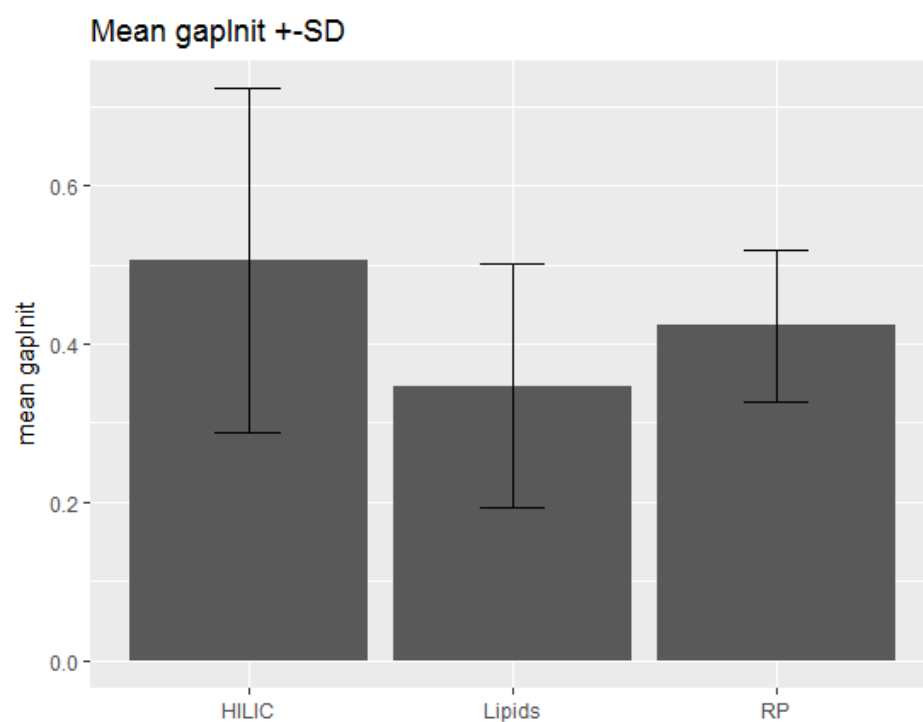


Figure 116: Mean gaplnit \pm SD for all triplicates analysed with each chromatographic assay.

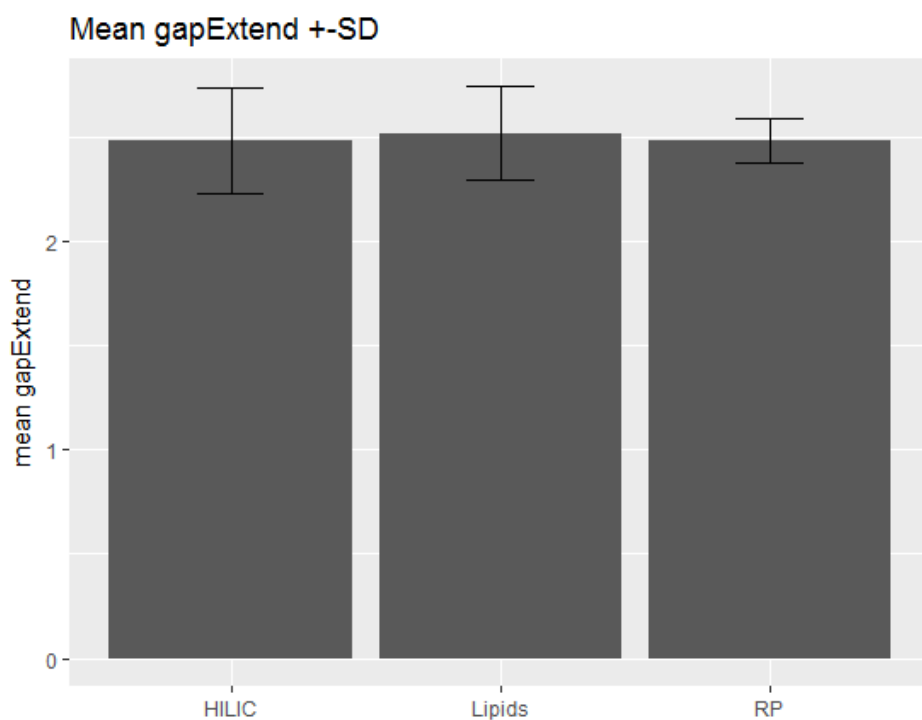


Figure 117: Mean gapExtend \pm SD for all triplicates analysed with each chromatographic assay.

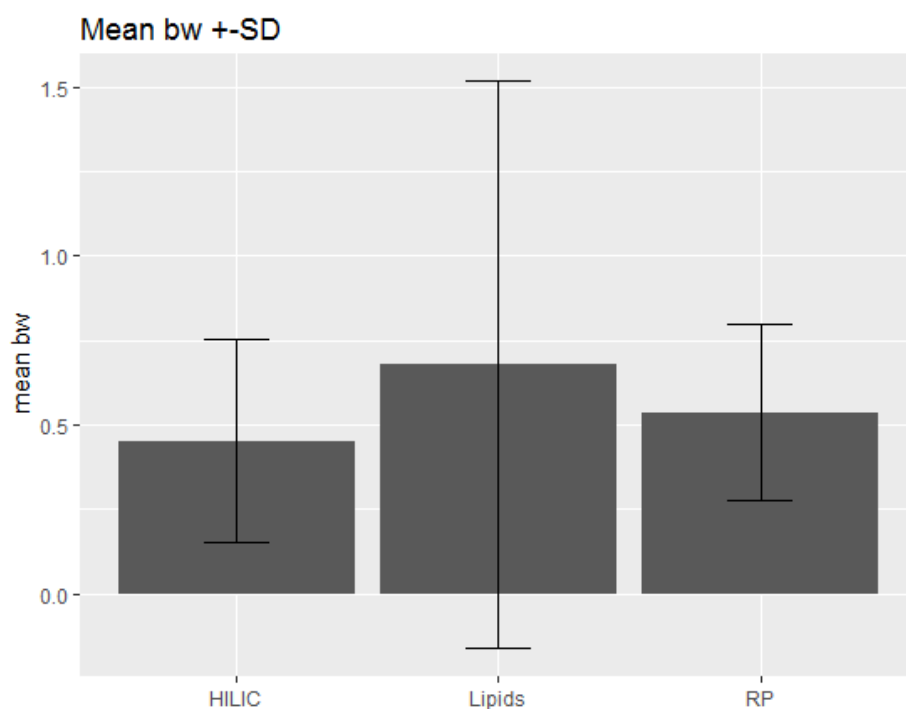


Figure 118: Mean bw \pm SD for all triplicates analysed with each chromatographic assay.

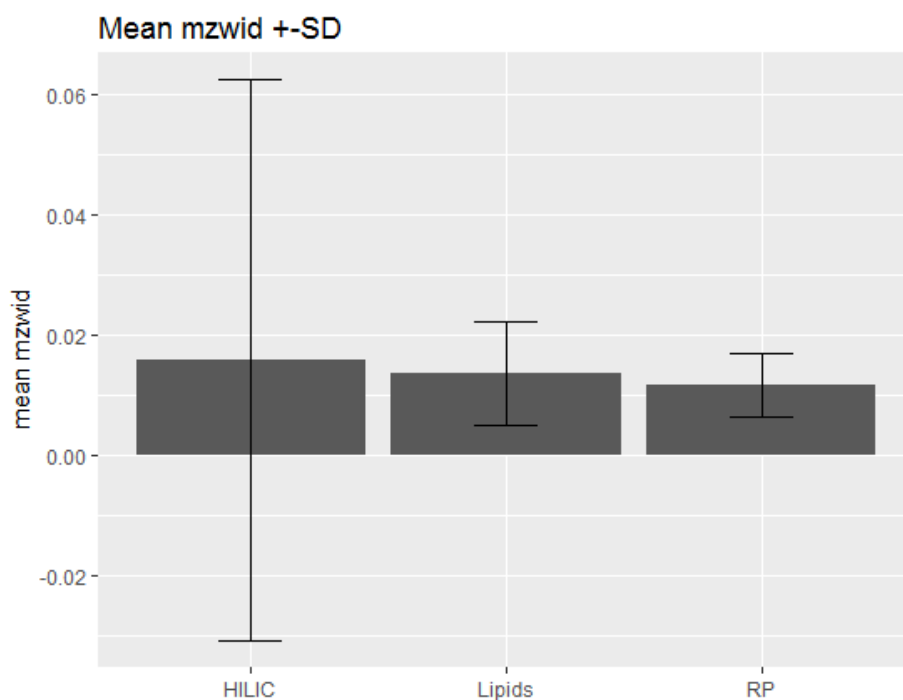


Figure 119: Mean mzwid \pm SD for all triplicates analysed with each chromatographic assay.

9.2.1.2 Parameters by Mass Resolution

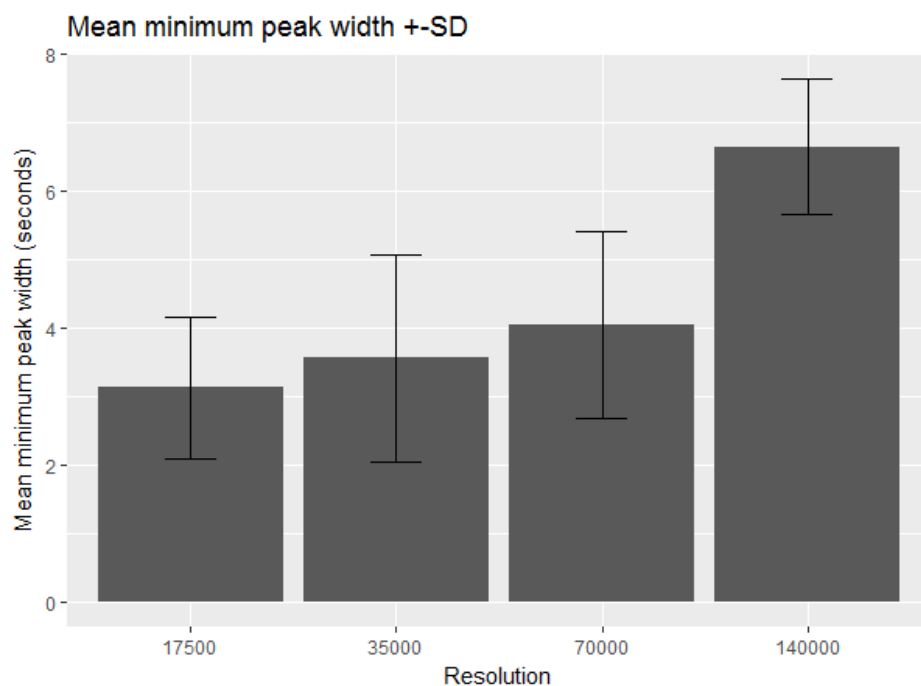


Figure 120: Mean minimum peak width \pm SD for all triplicates analysed at different mass resolutions.

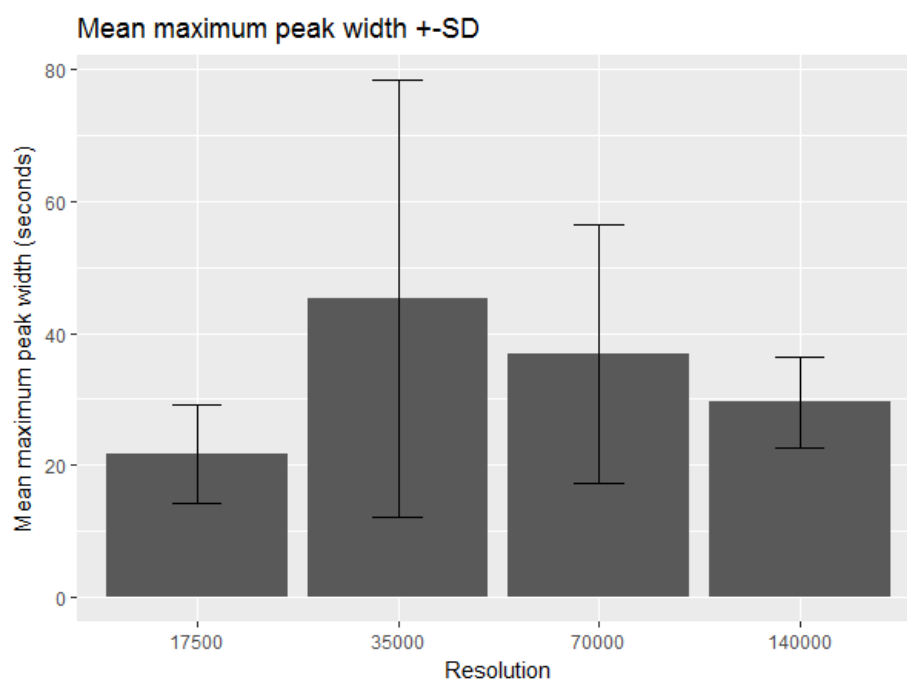


Figure 121: Mean maximum peak width \pm SD for all triplicates analysed at different mass resolutions.

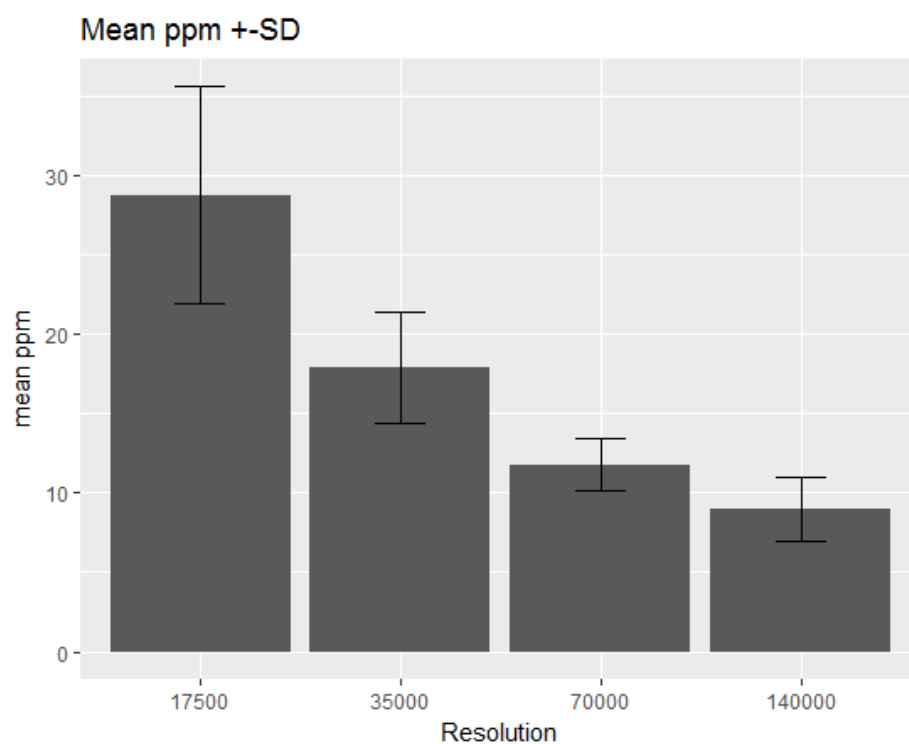


Figure 122: Mean ppm \pm SD for all triplicates analysed at different mass resolutions.

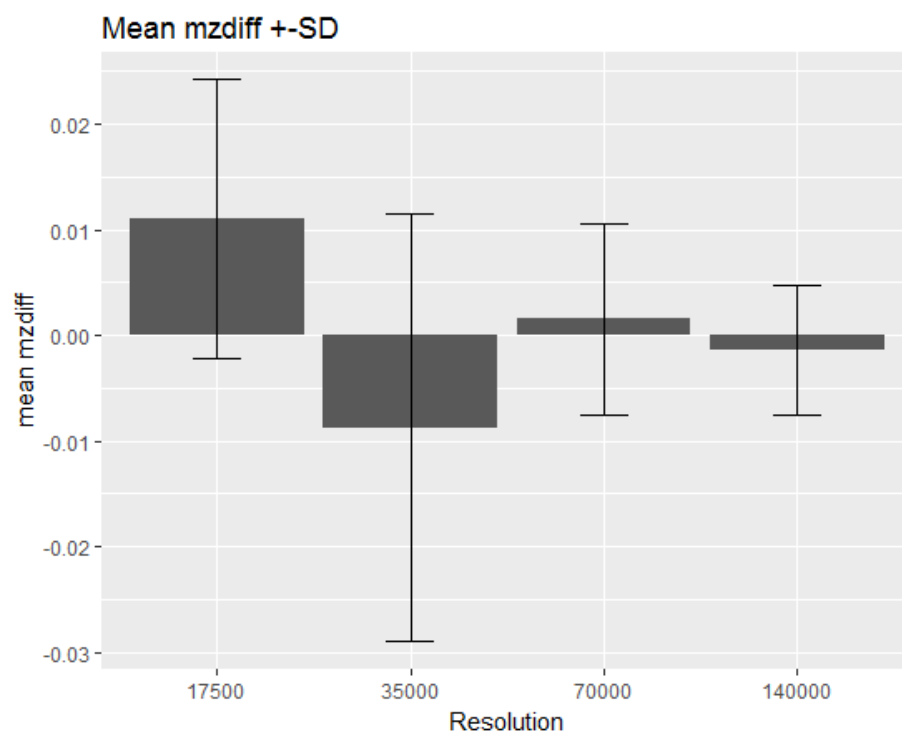


Figure 123: Mean mzdiff \pm SD for all triplicates analysed at different mass resolutions.

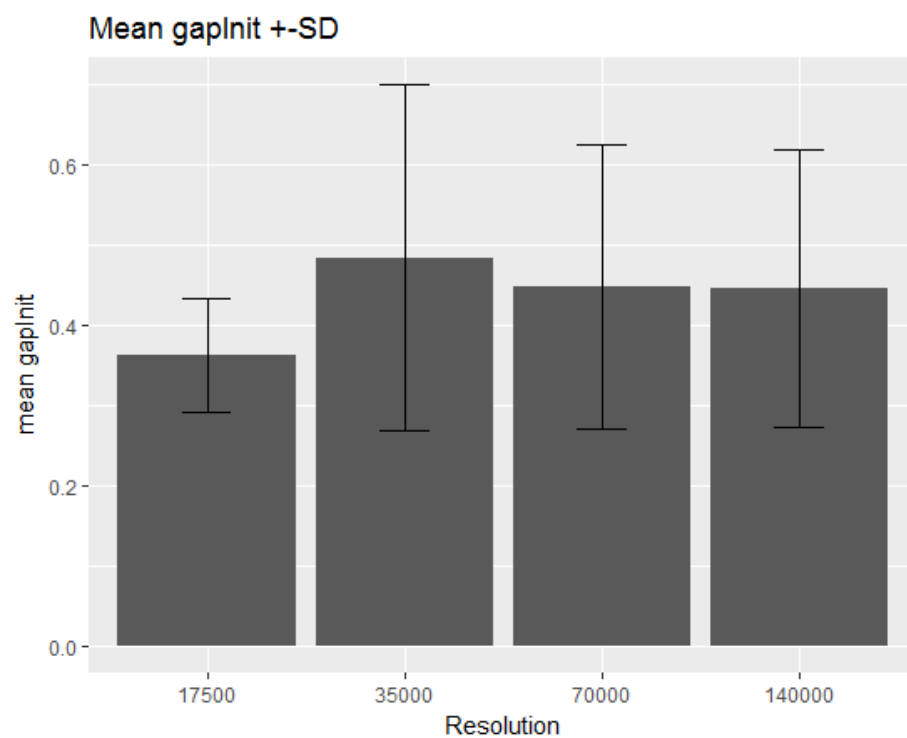


Figure 124: Mean gaplnit \pm SD for all triplicates analysed at different mass resolutions.

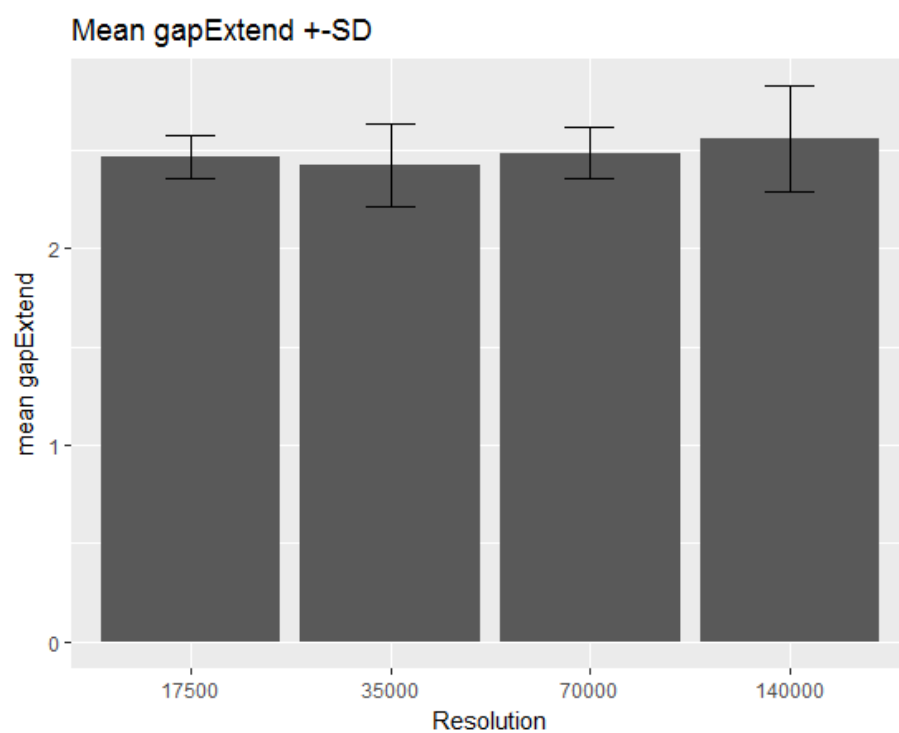


Figure 125: Mean gapExtend \pm SD for all triplicates analysed at different mass resolutions.

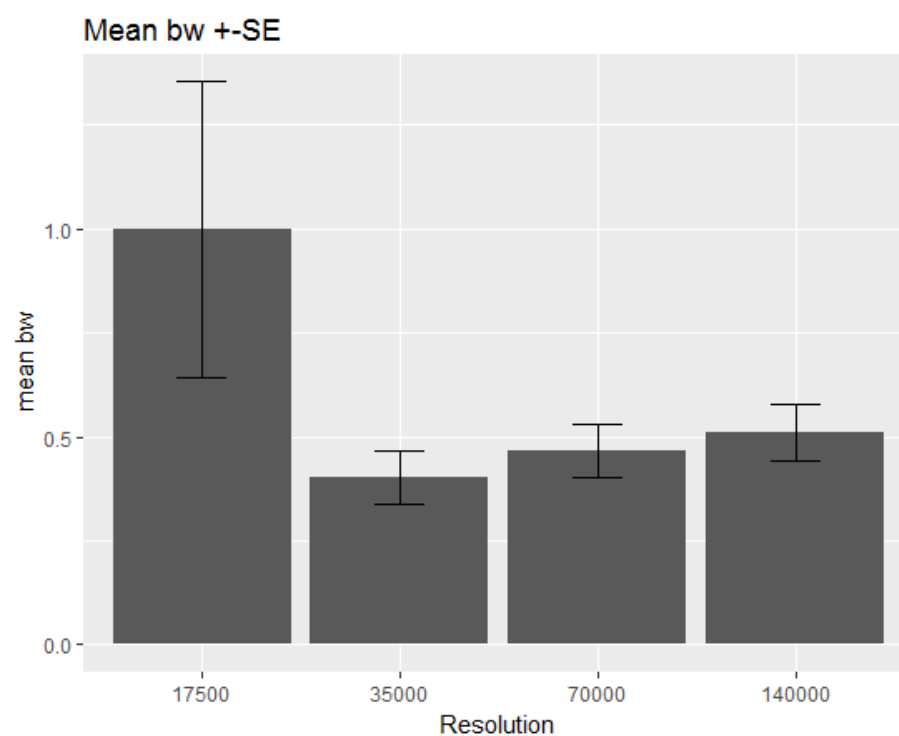


Figure 126: Mean bw \pm SD for all triplicates analysed at different mass resolutions.

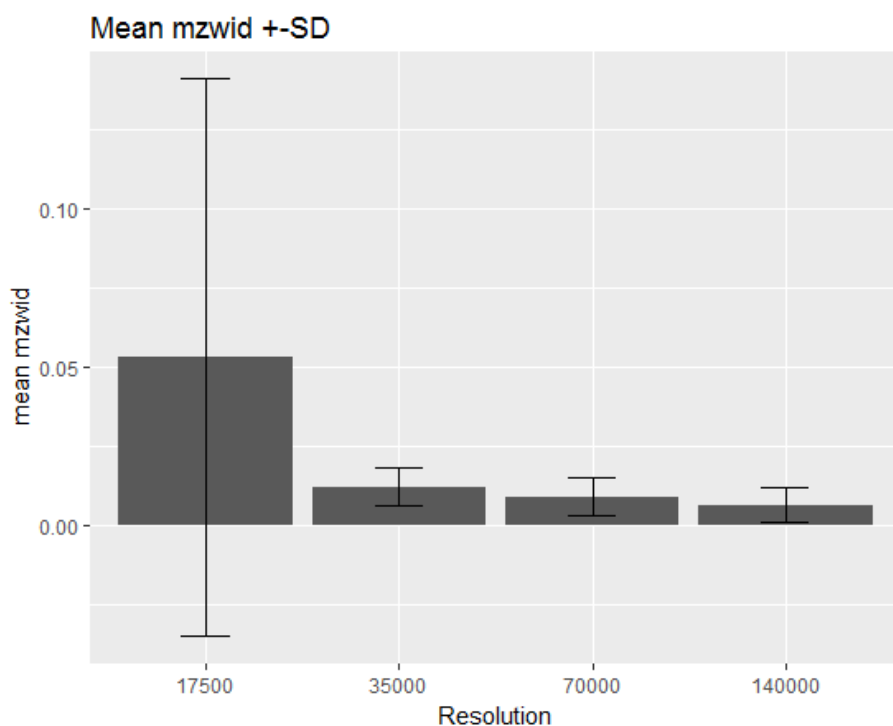


Figure 127: Mean mzwid \pm SD for all triplicates analysed at different mass resolutions.

9.2.1.3 Parameters by Mass Resolution/UHPLC Assay

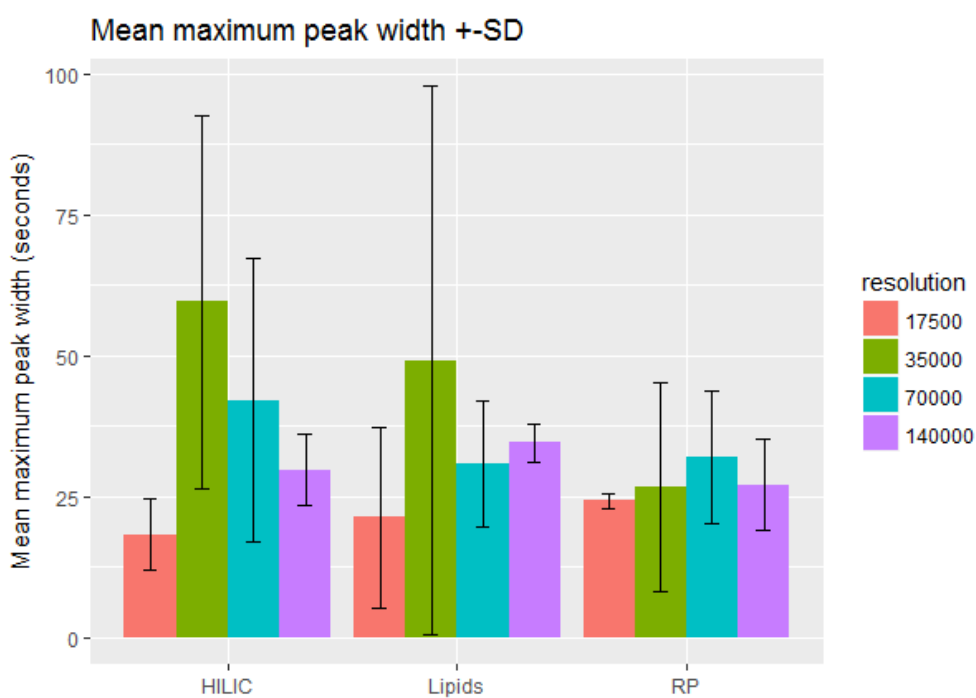


Figure 128: Mean maximum peak width \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

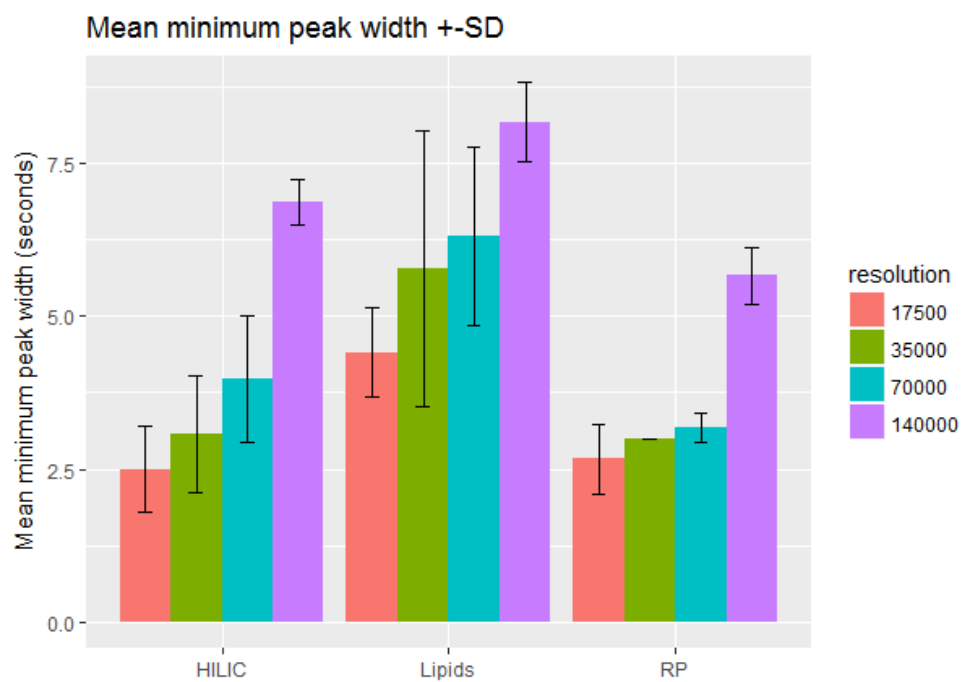


Figure 129: Mean minimum peak width \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

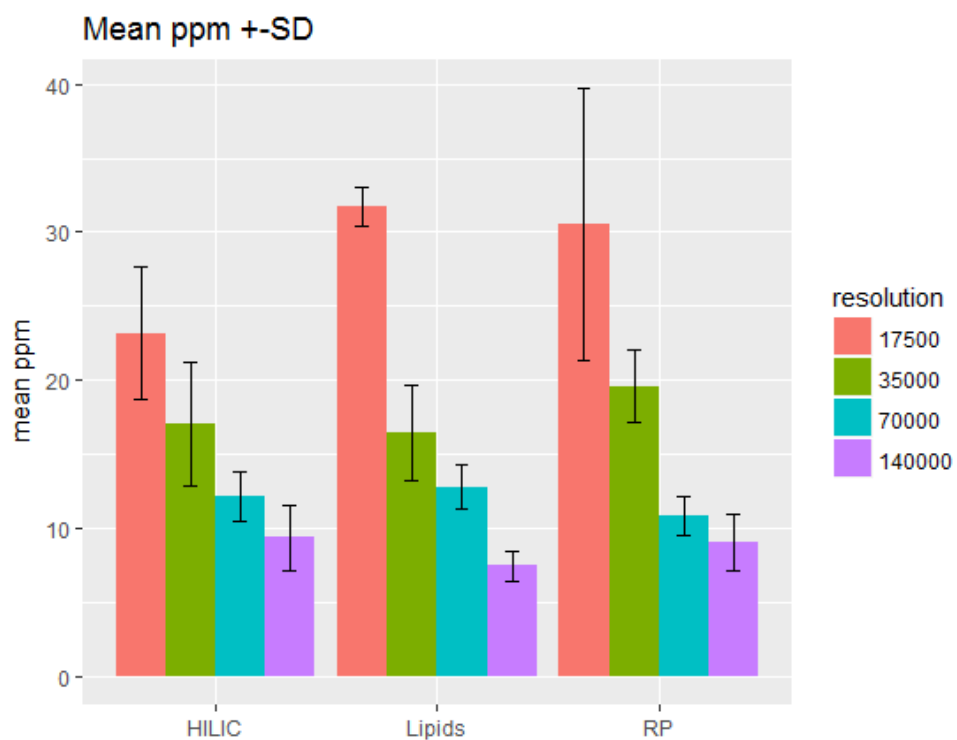


Figure 130: Mean ppm \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

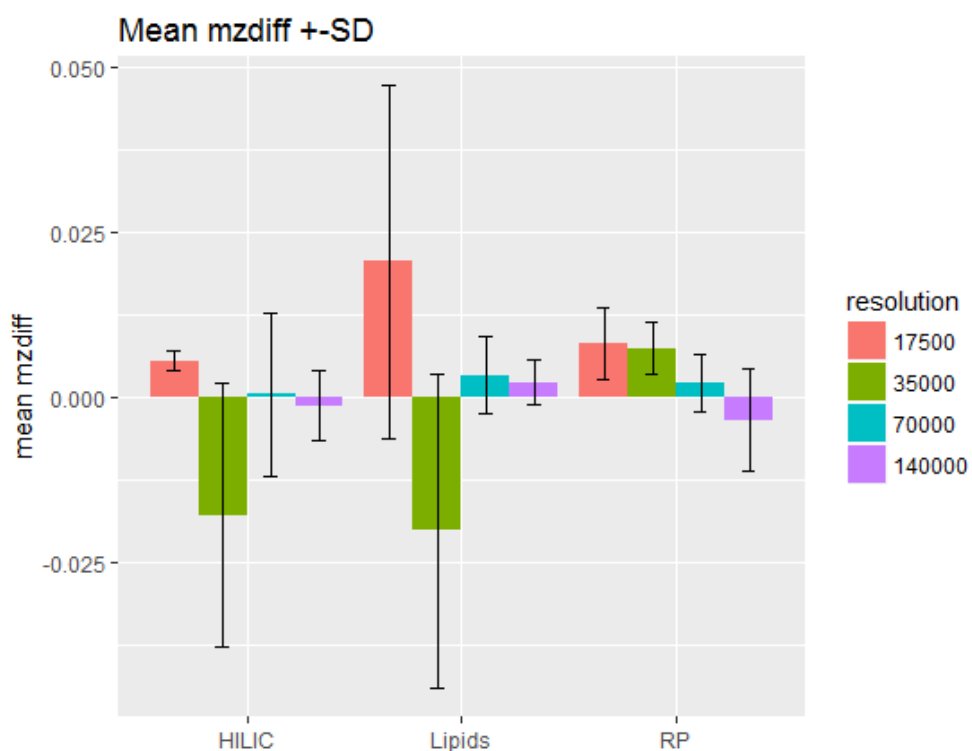


Figure 131: Mean mzdiff \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

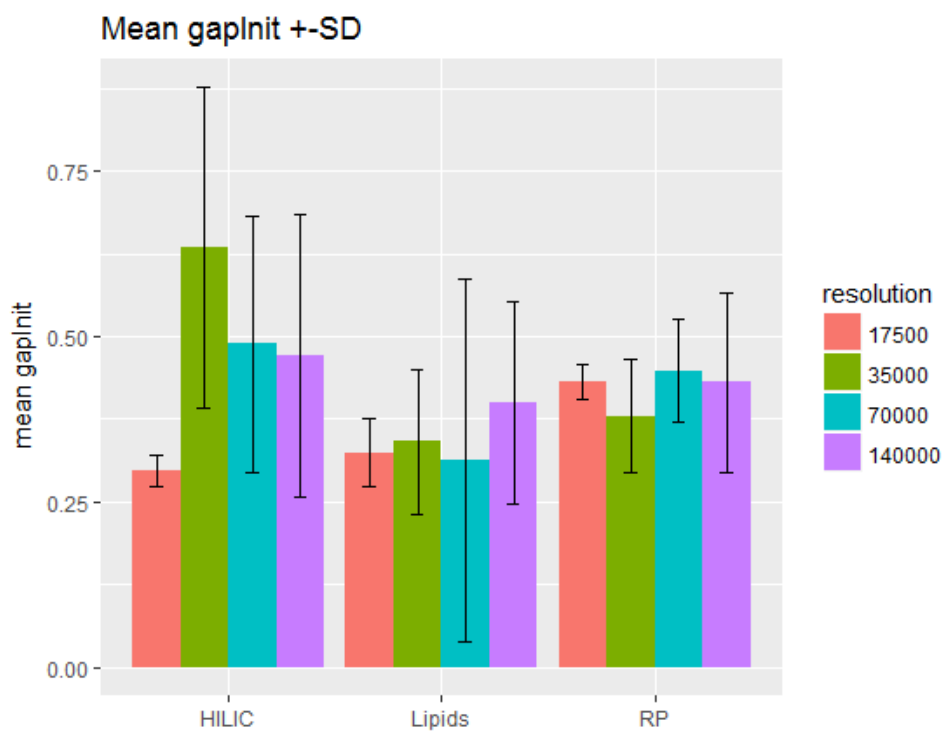


Figure 132: Mean gaplnit \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

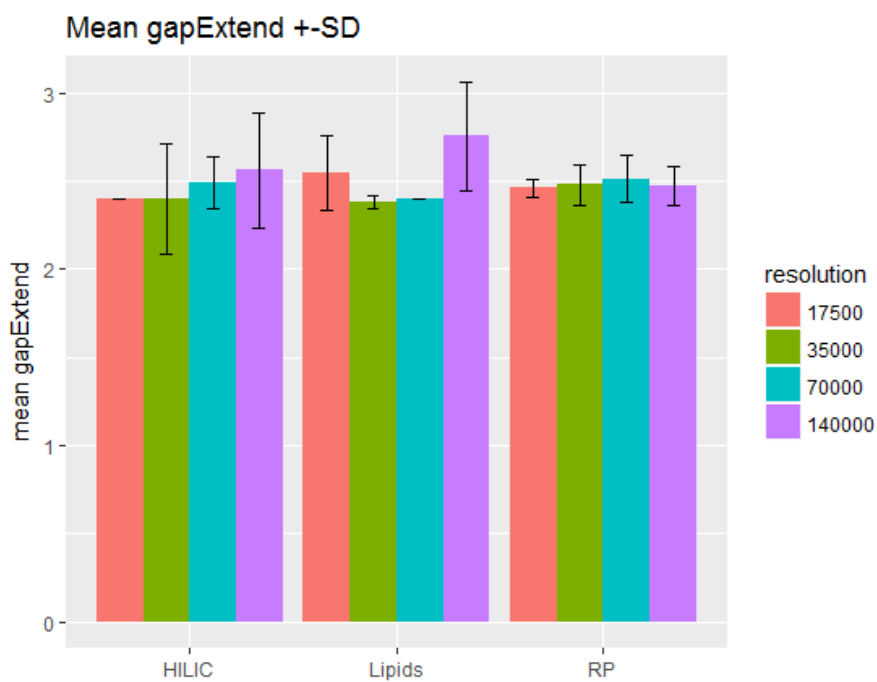


Figure 133: Mean gapExtend \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

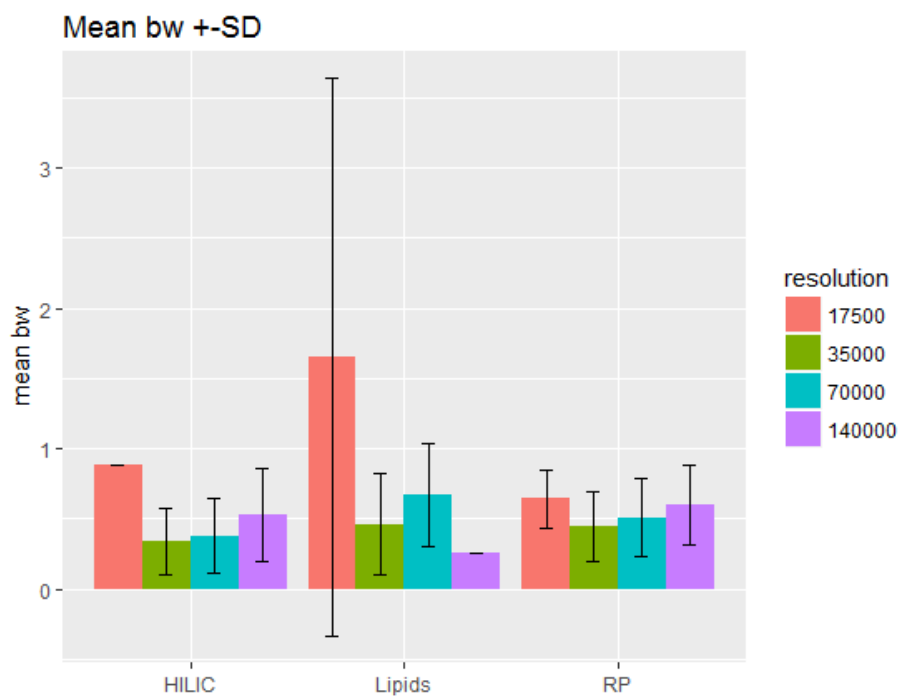


Figure 134: Mean bw \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

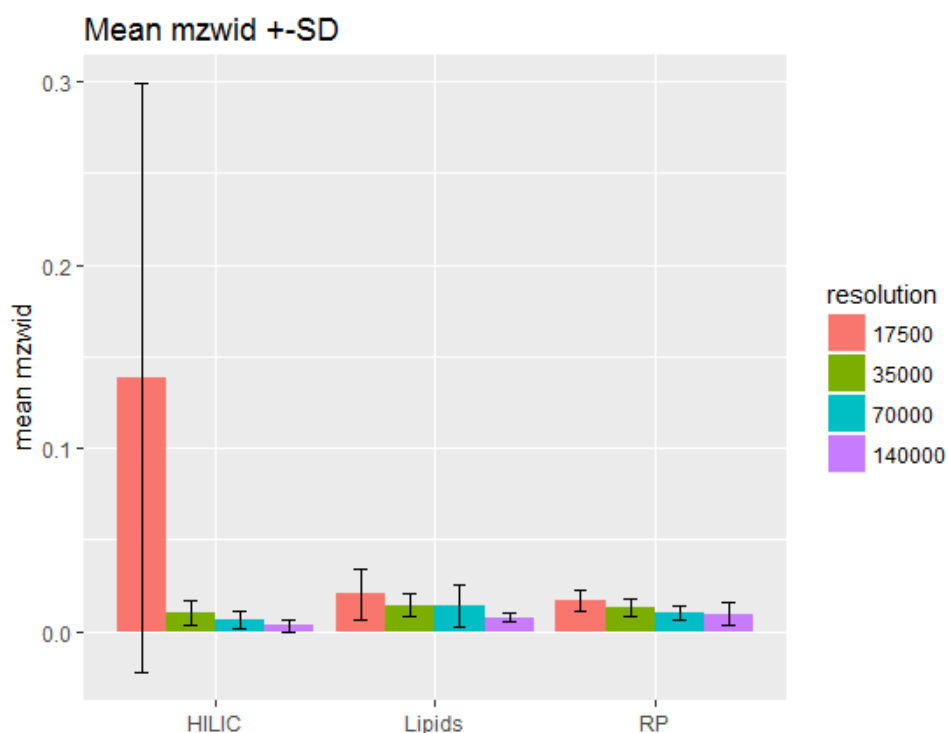


Figure 135: Mean mzwid \pm SD for all triplicates analysed at different mass resolutions with different chromatographic assays.

9.3 Comparison of different MS² acquisition strategies on the Q Exactive Plus

9.3.1 Number of features detected (MS-DIAL)

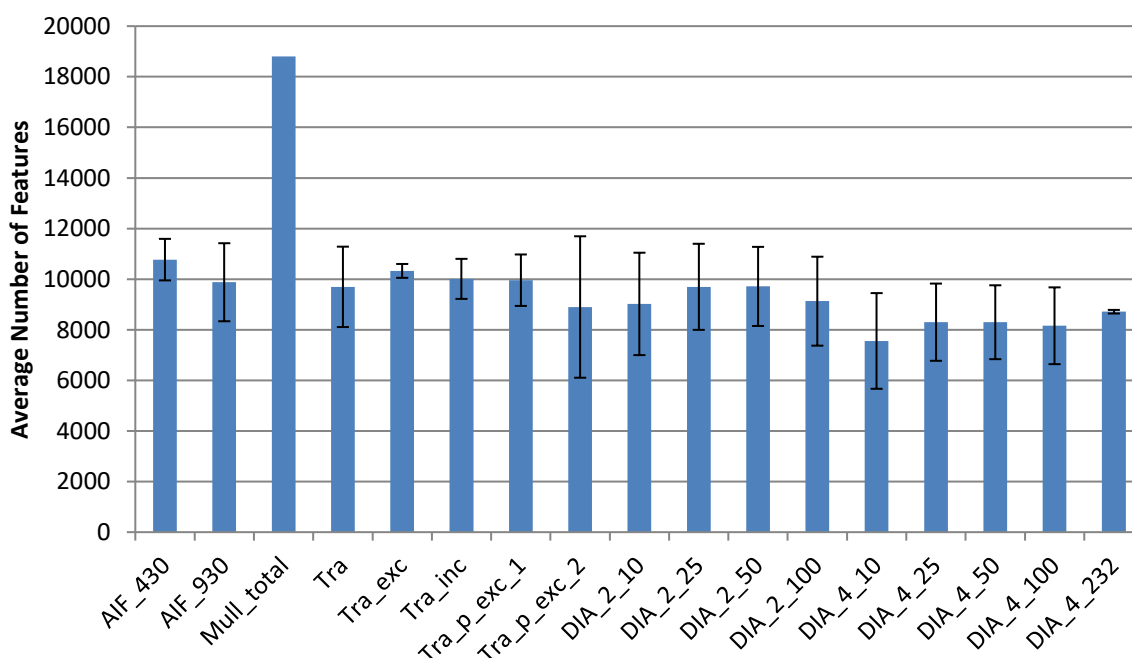


Figure 136: The average number of features detected for each quadruplicate in the HILIC negative ion mode data after MS-DIAL processing \pm SD.

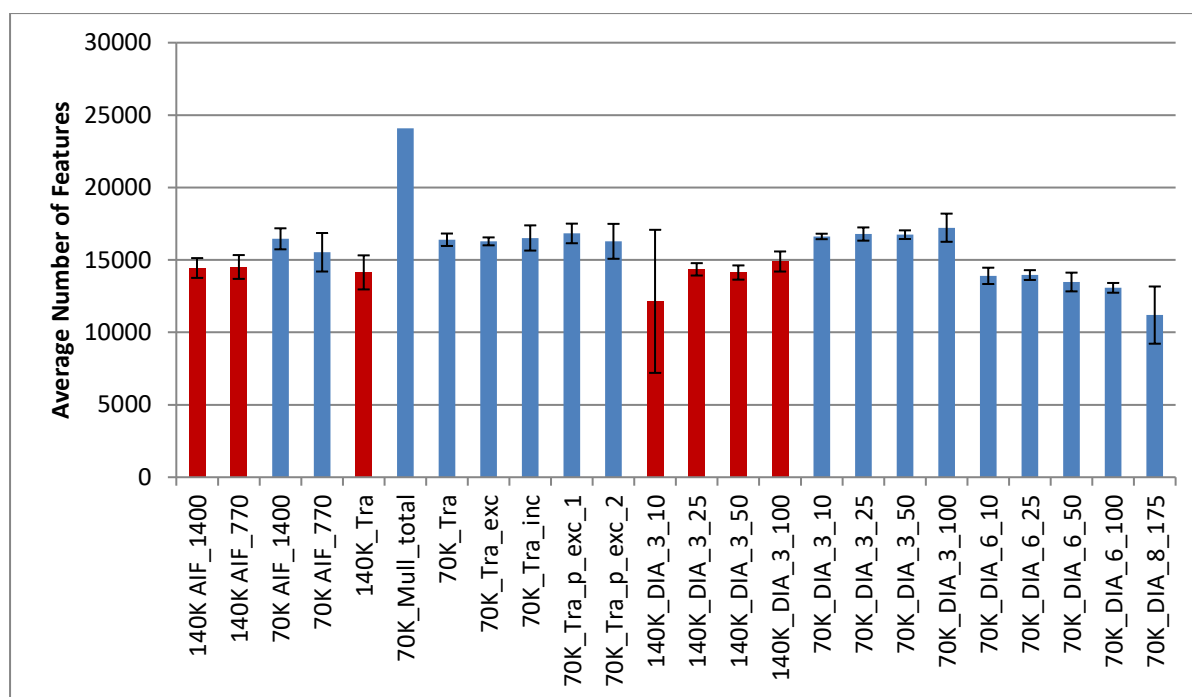


Figure 137: The average number of features detected for each quadruplicate in the Lipidomics positive ion mode data after MS-DIAL processing \pm SD.

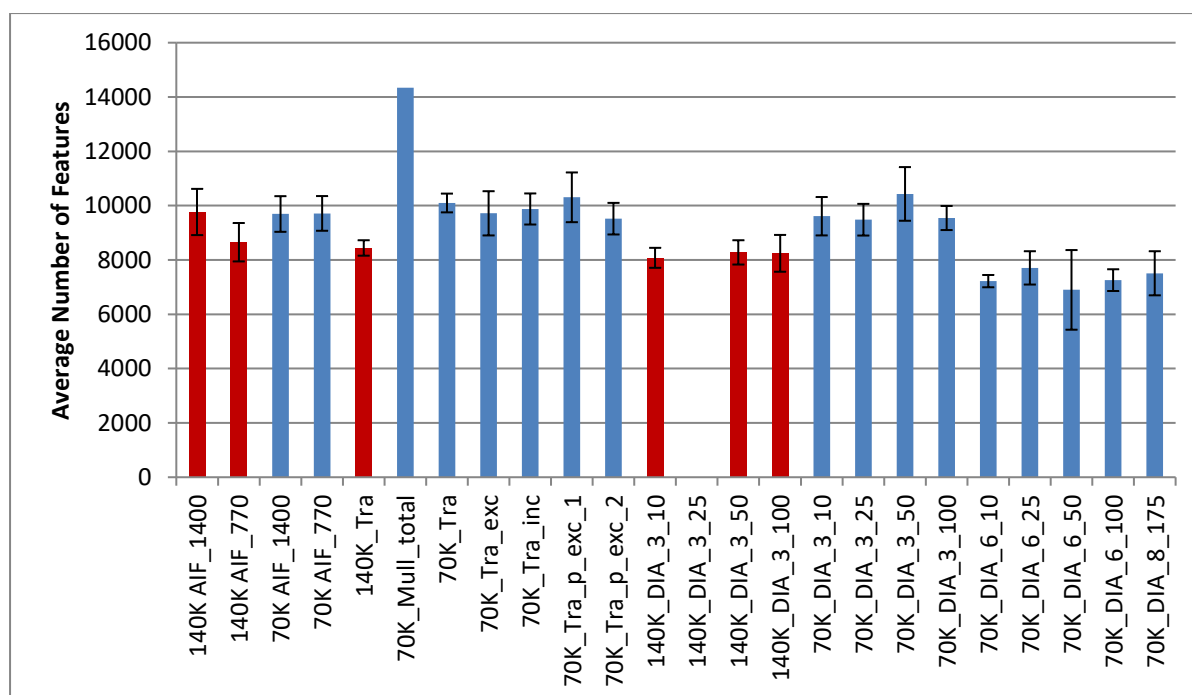


Figure 138: The average number of features detected for each quadruplicate in the Lipidomics negative ion mode data after MS-DIAL processing \pm SD.

9.3.2 Number of Features with MS² Data

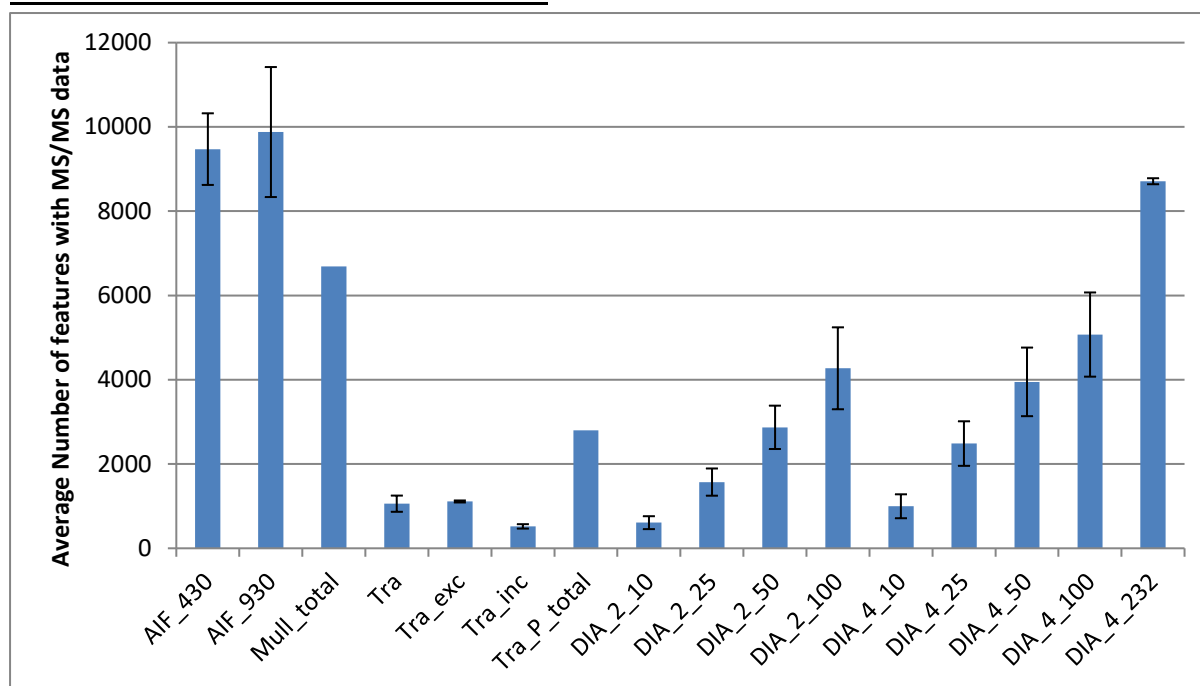


Figure 139: The average number of features that have MS² data associated with them for each quadruplicate after processing in MS-DIAL in the HILIC/negative ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

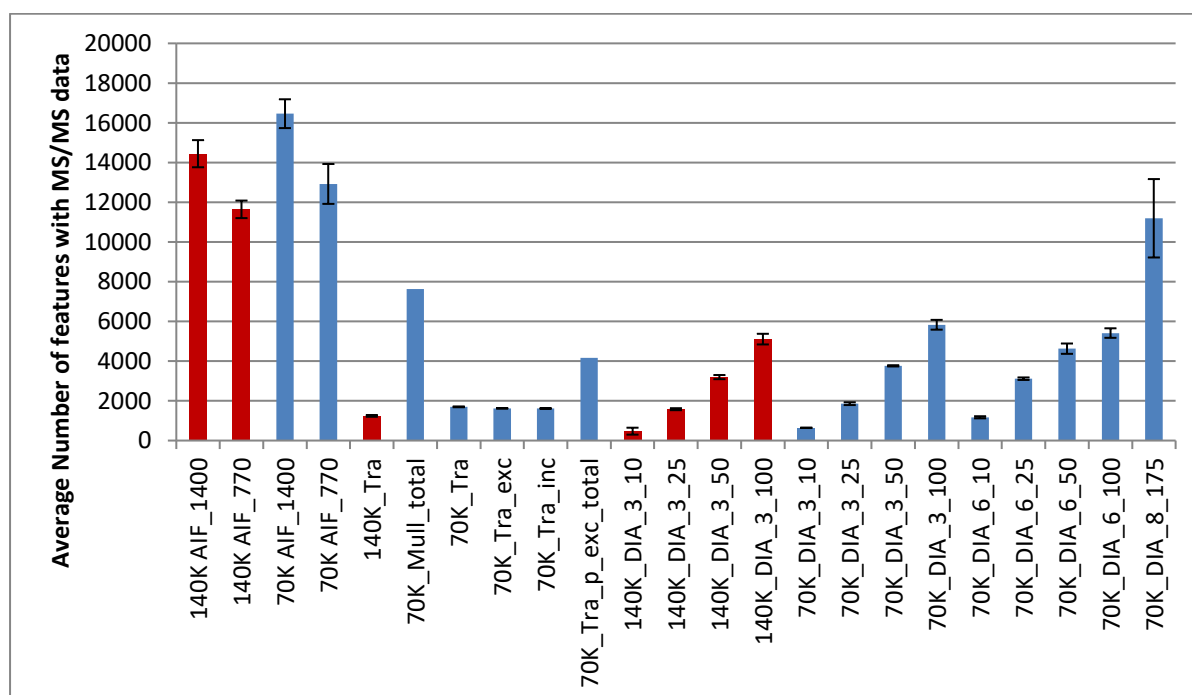


Figure 140: The average number of features that have MS² data associated with them for each quadruplicate after processing in MS-DIAL in the Lipidomics/positive ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

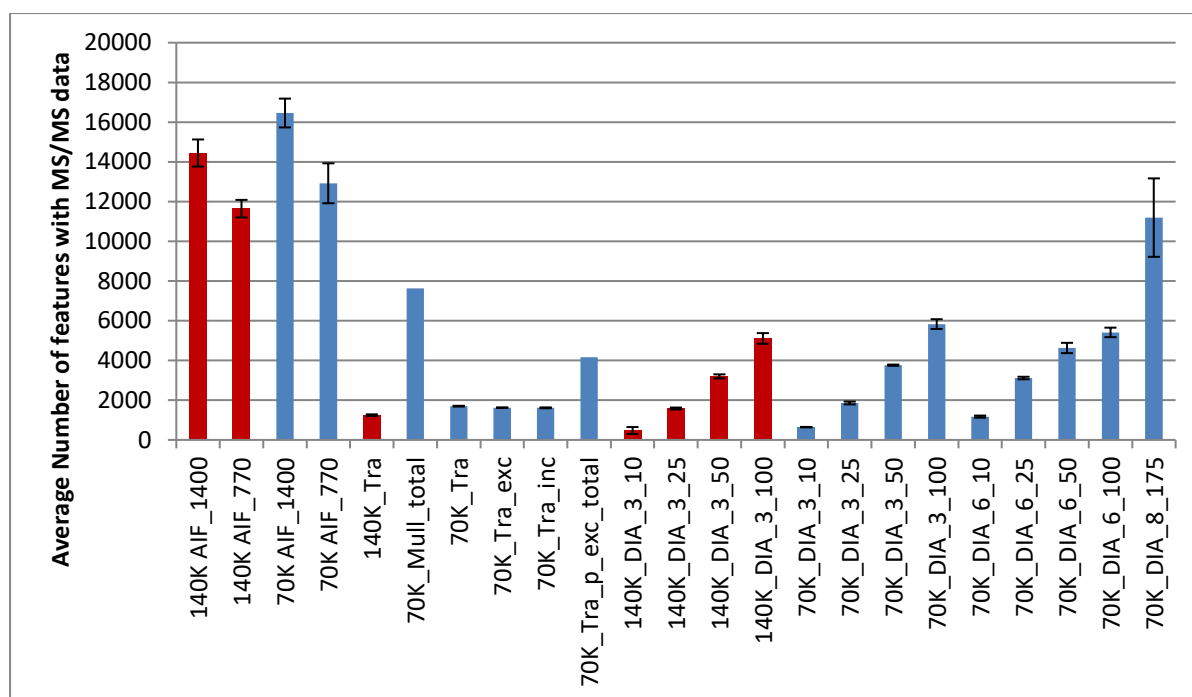


Figure 141: The average number of features that have MS² data associated with them for each quadruplicate after processing in MS-DIAL in the Lipidomics/negative ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

9.3.3 Number of Features Annotated

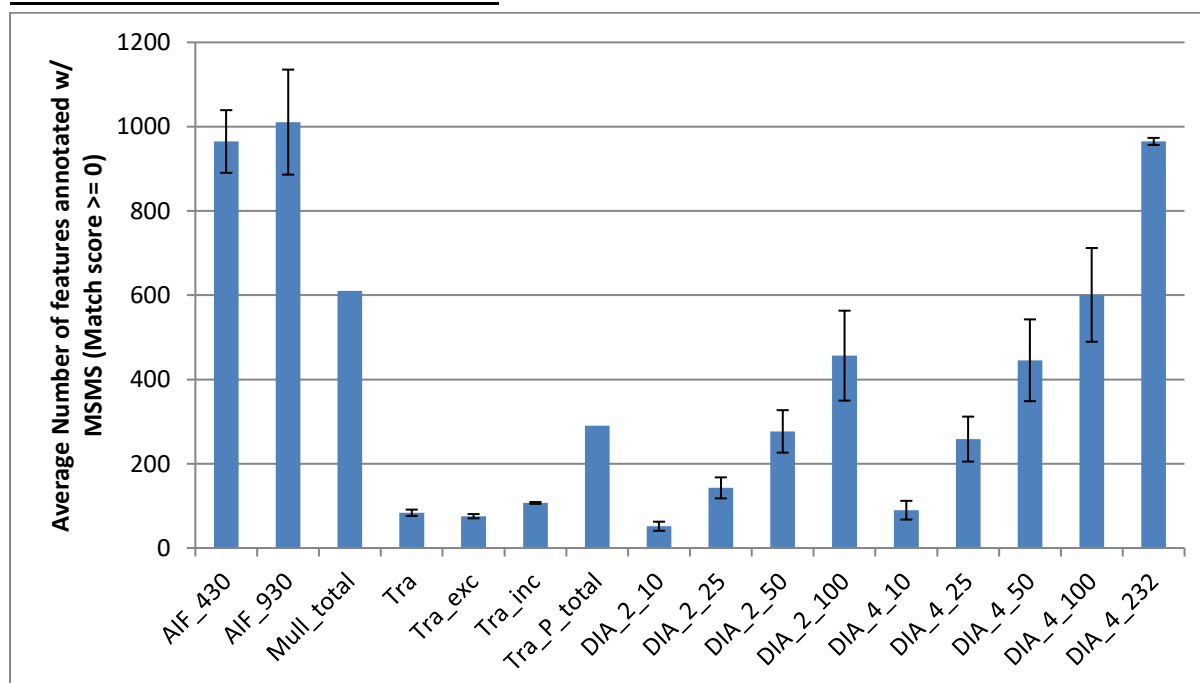


Figure 142: Average number of features with an MS² spectral match of any quality for each quadruplicate after processing in MS-DIAL for the HILIC/negative ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

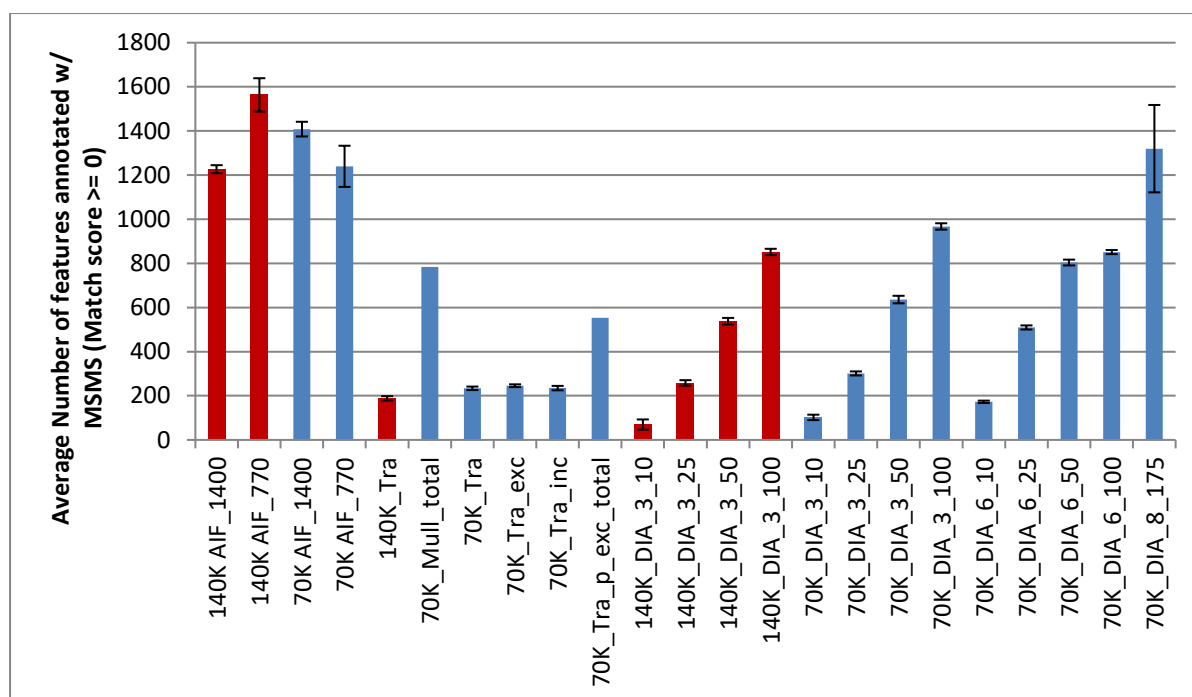


Figure 143: Average number of features with an MS² spectral match of any quality for each quadruplicate after processing in MS-DIAL for the Lipidomics/positive ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

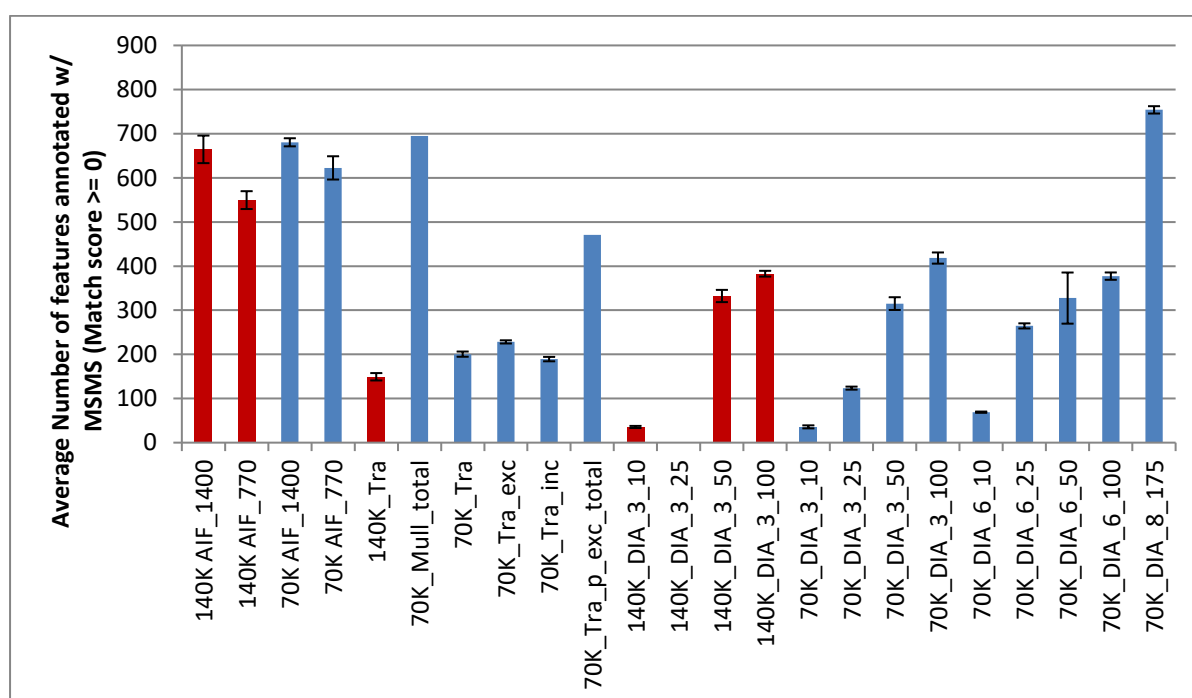


Figure 144: Average number of features with an MS² spectral match of any quality for each quadruplicate after processing in MS-DIAL for the Lipidomics/negative ion mode data \pm SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

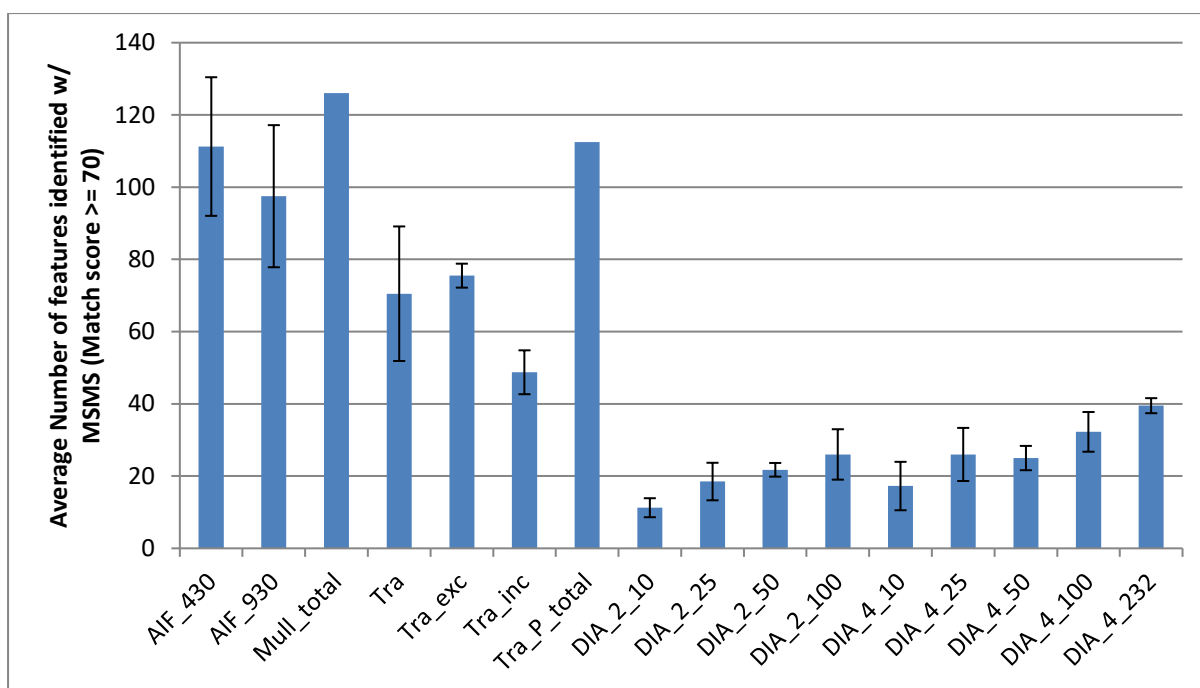


Figure 145: Average number of features with an MS² spectral match score ≥ 70 for each quadruplicate after processing in MS-DIAL for the HILIC/negative ion mode data ±SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

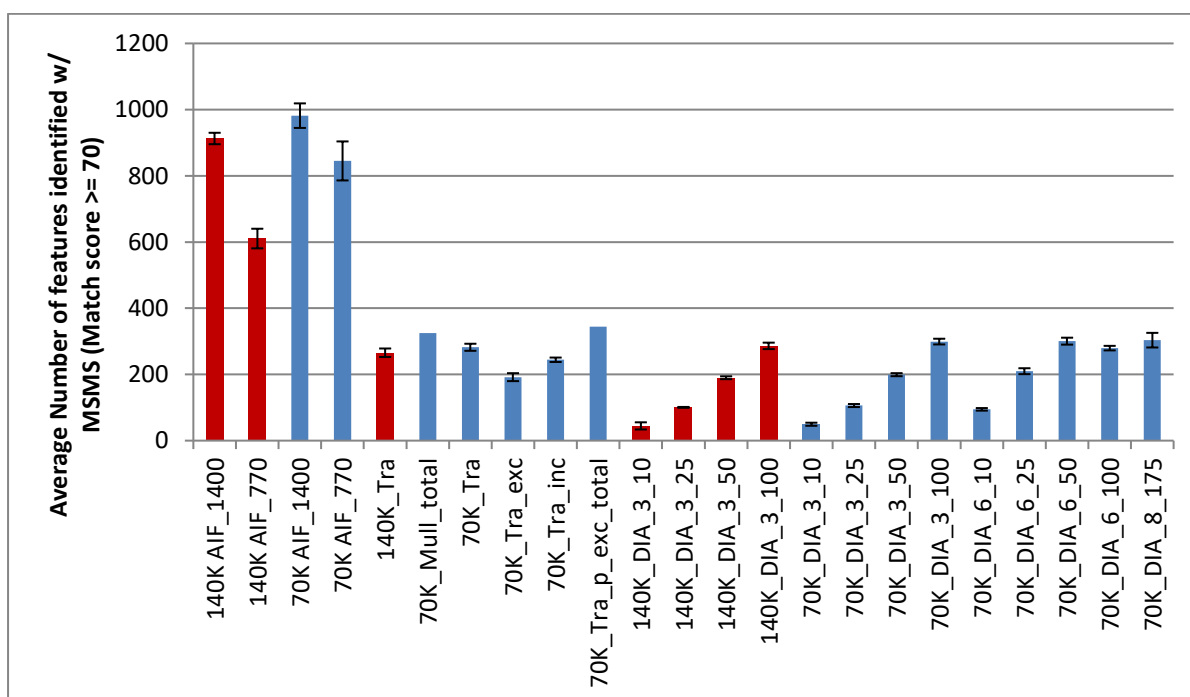


Figure 146: Average number of features with an MS² spectral match score ≥ 70 for each quadruplicate after processing in MS-DIAL for the Lipidomics/positive ion mode data ±SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

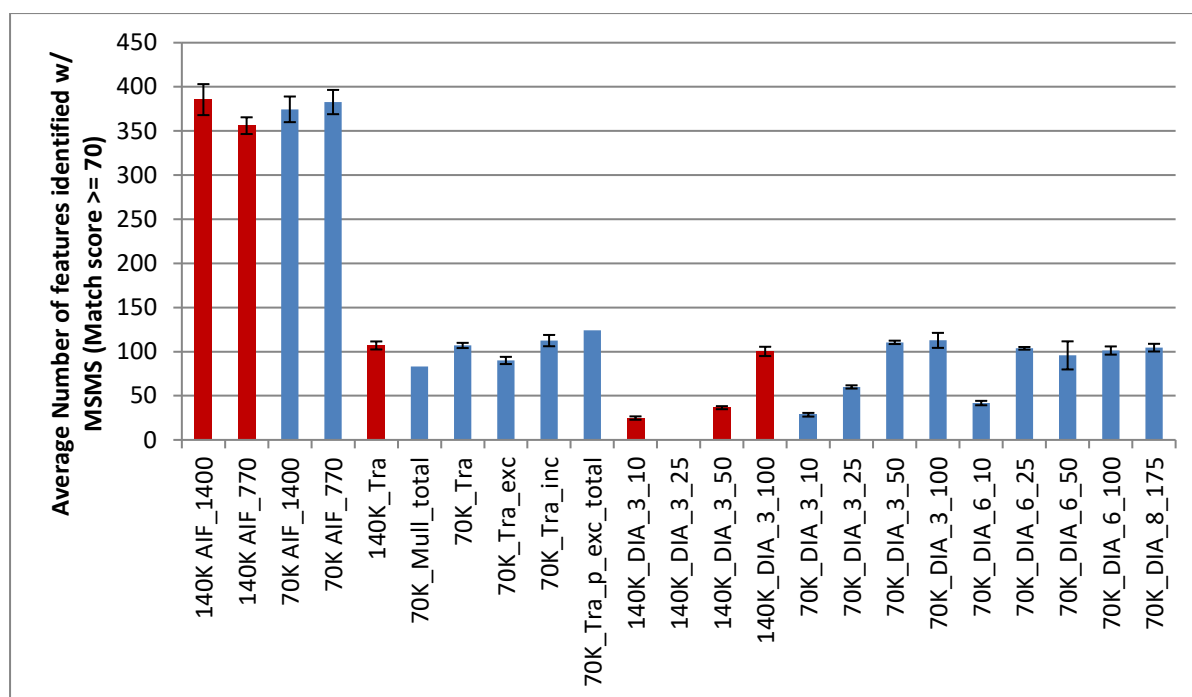


Figure 147: Average number of features with an MS² spectral match score ≥ 70 for each quadruplicate after processing in MS-DIAL for the Lipidomics/negative ion mode data ±SD. Mull_total and Tra_p_total do not have a SD due to the way they were calculated.

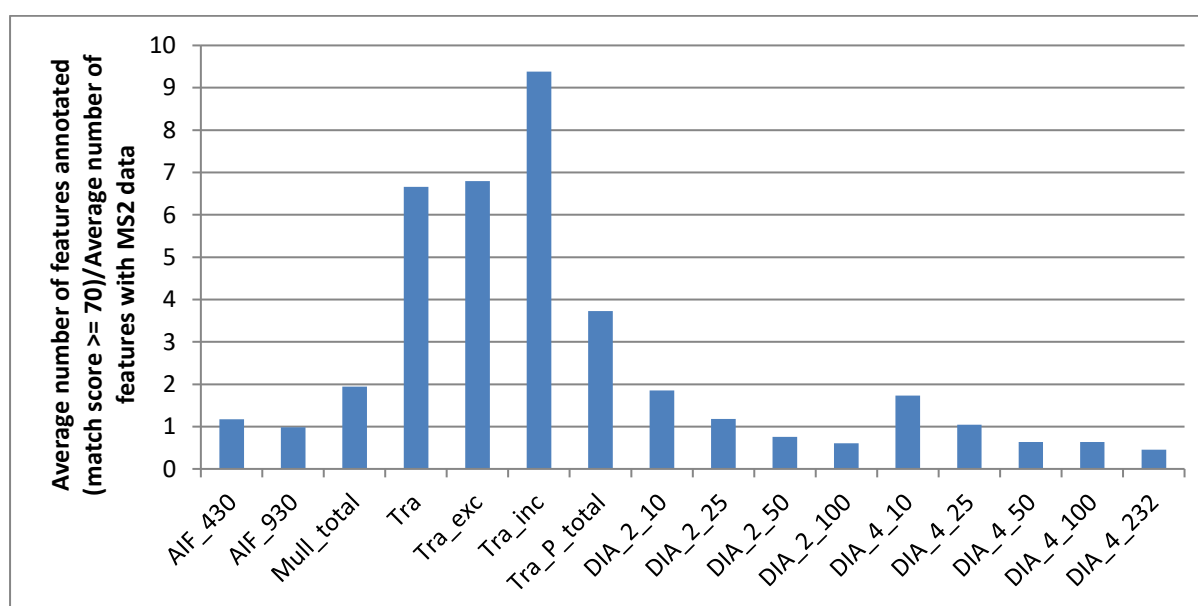


Figure 148: The percentage of average number of features annotated (match score ≥ 70)/average number of features with MS² data for each MS strategy for the HILIC/negative ion mode dataset

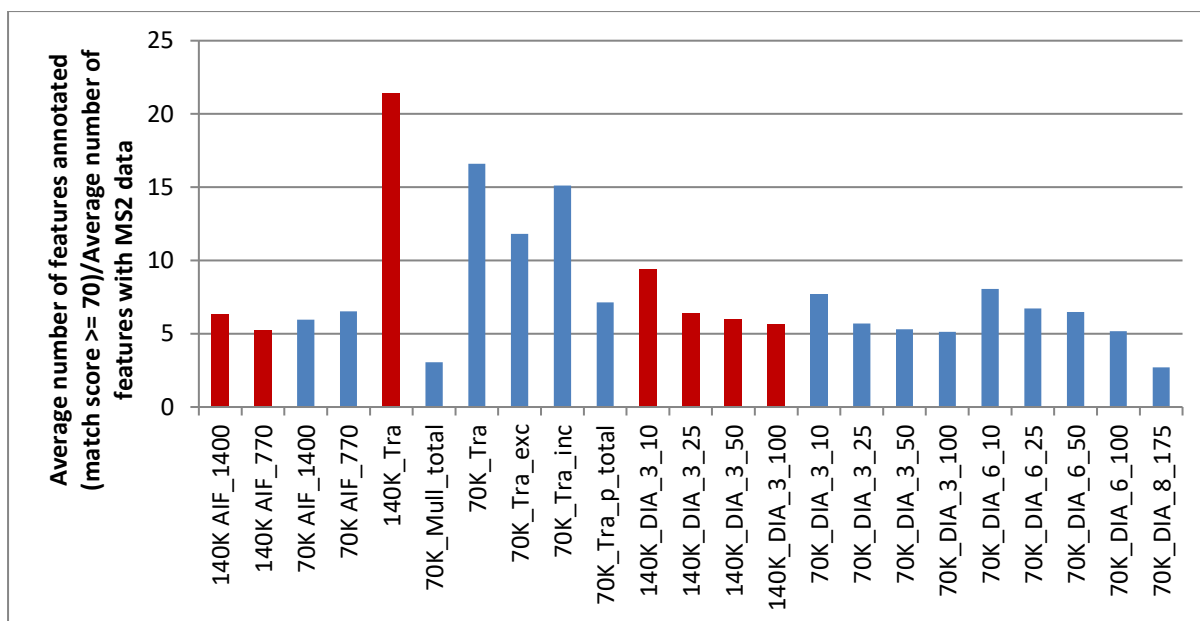


Figure 149: The percentage of average number of features annotated (match score ≥ 70)/average number of features with MS² data for each MS² strategy for the Lipidomics/positive ion mode dataset

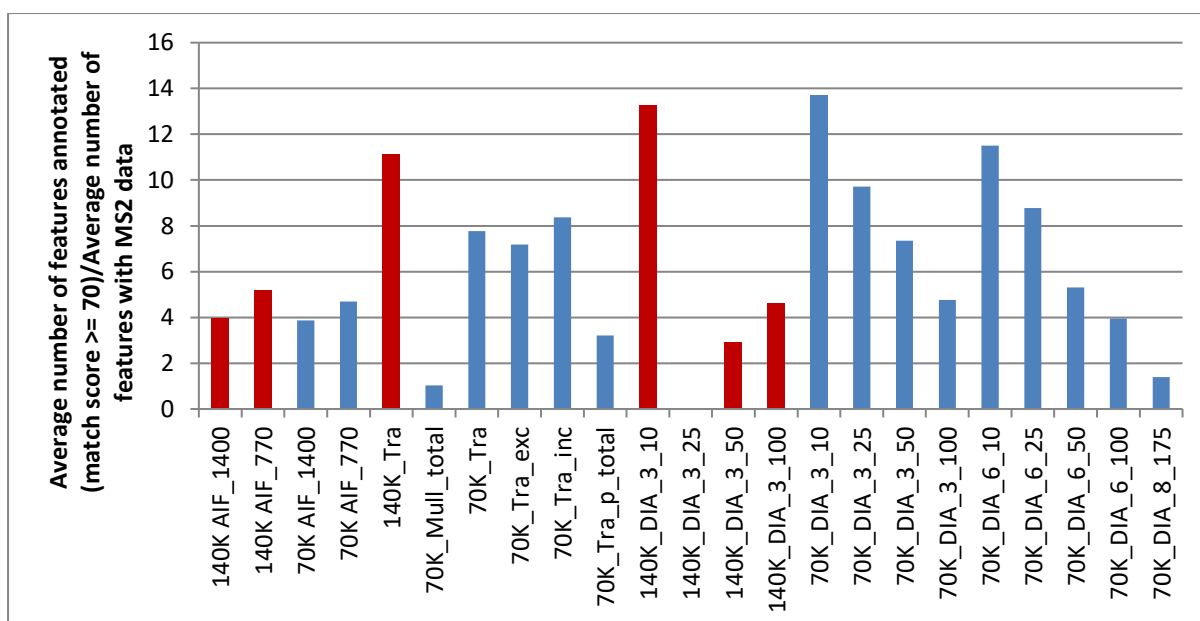


Figure 150: The percentage of average number of features annotated (match score ≥ 70)/average number of features with MS² data for each MS² strategy for the Lipidomics/negative ion mode dataset

9.3.4 Purity of Fragmentation Windows

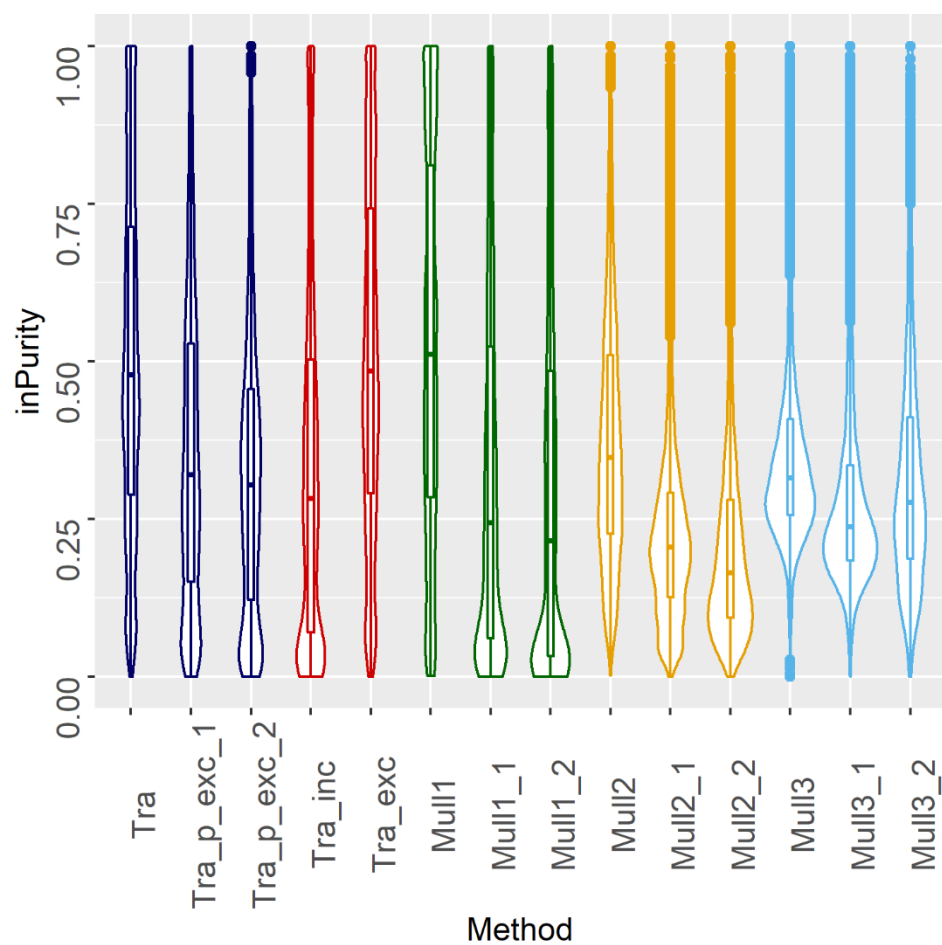


Figure 151: The distribution of interpolated purity (inPurity) scores for all DDA based methods for the HILIC_NEG data.

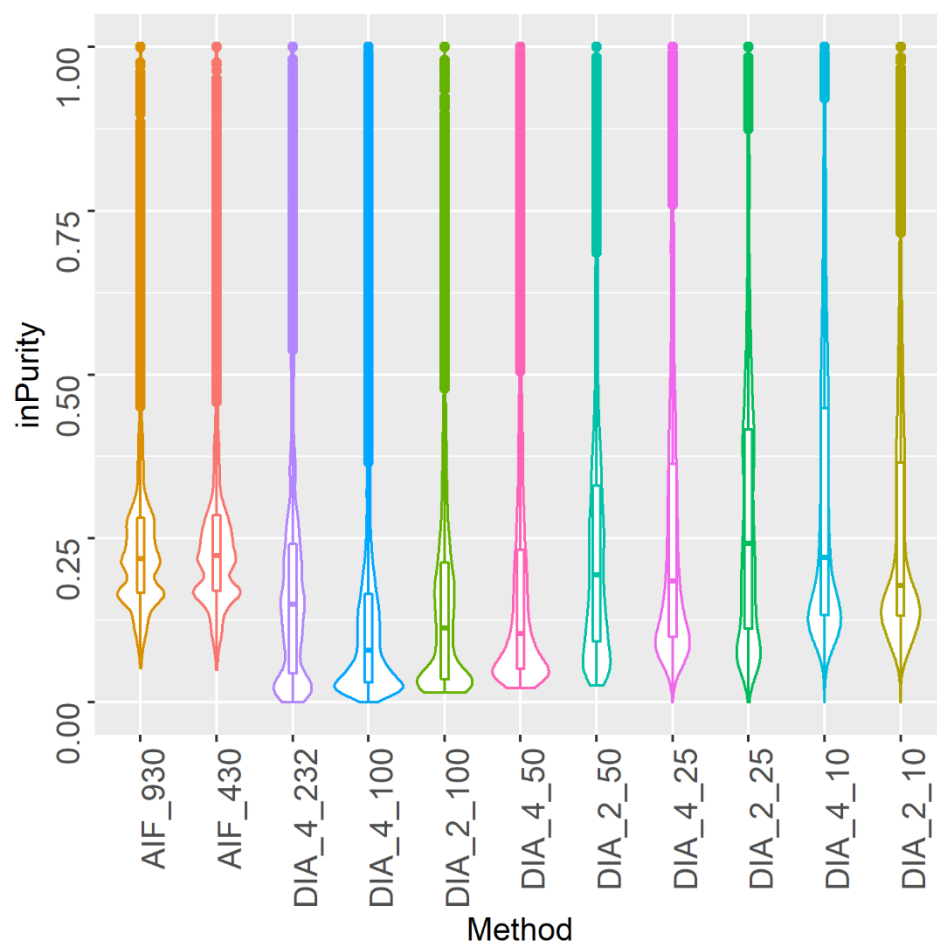


Figure 152: The distribution of interpolated purity (inPurity) scores for all DIA based methods for the HILIC_NEG data.

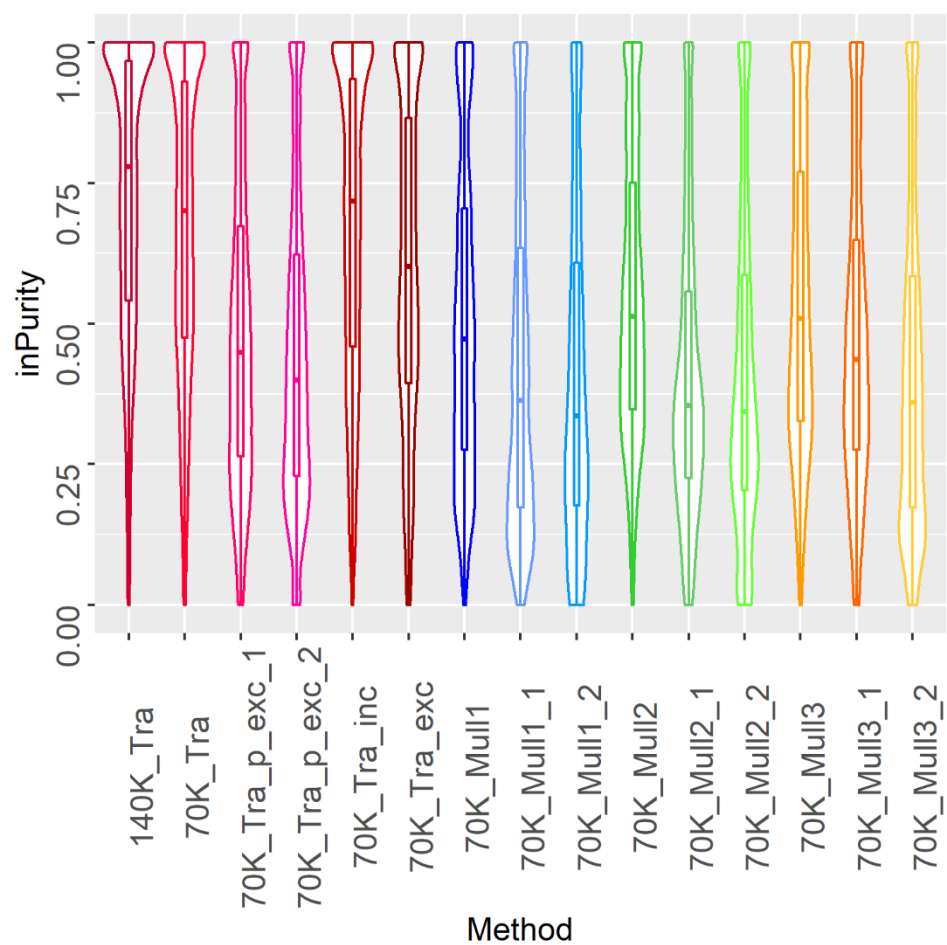


Figure 153: The distribution of interpolated purity (inPurity) scores for all DDA based methods for the Lipids_POS data.

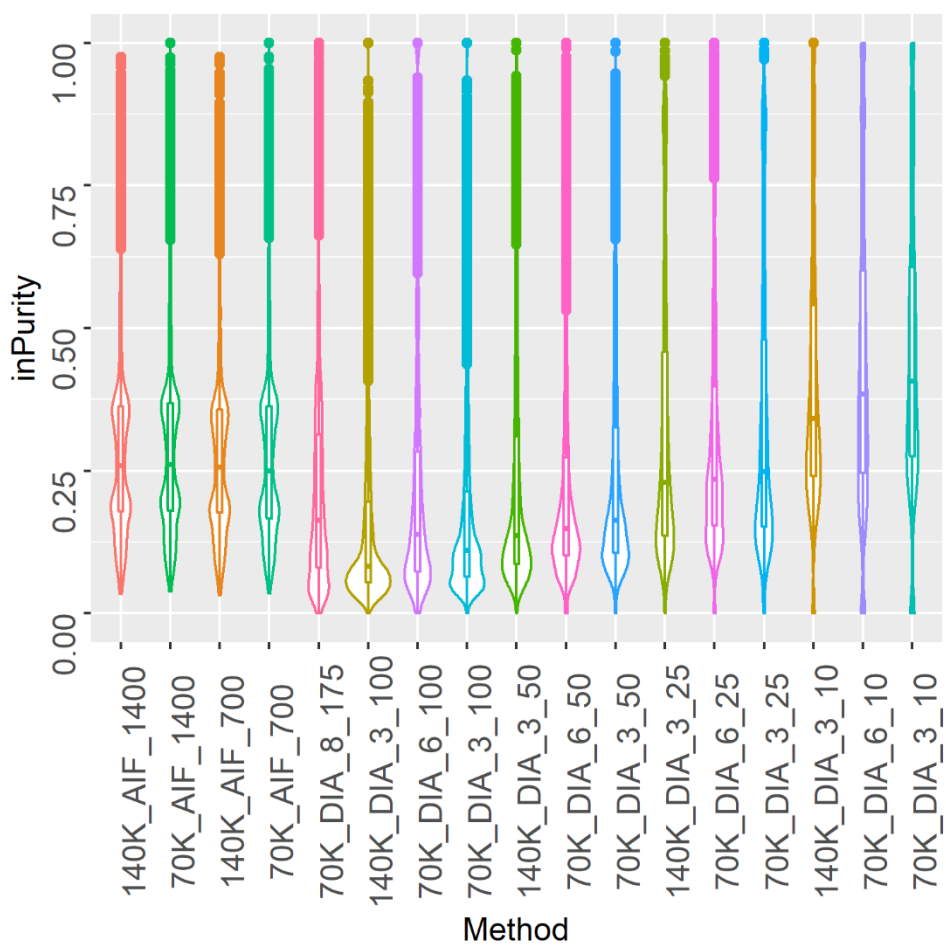


Figure 154: The distribution of interpolated purity (inPurity) scores for all DIA based methods for the Lipids_POS data.

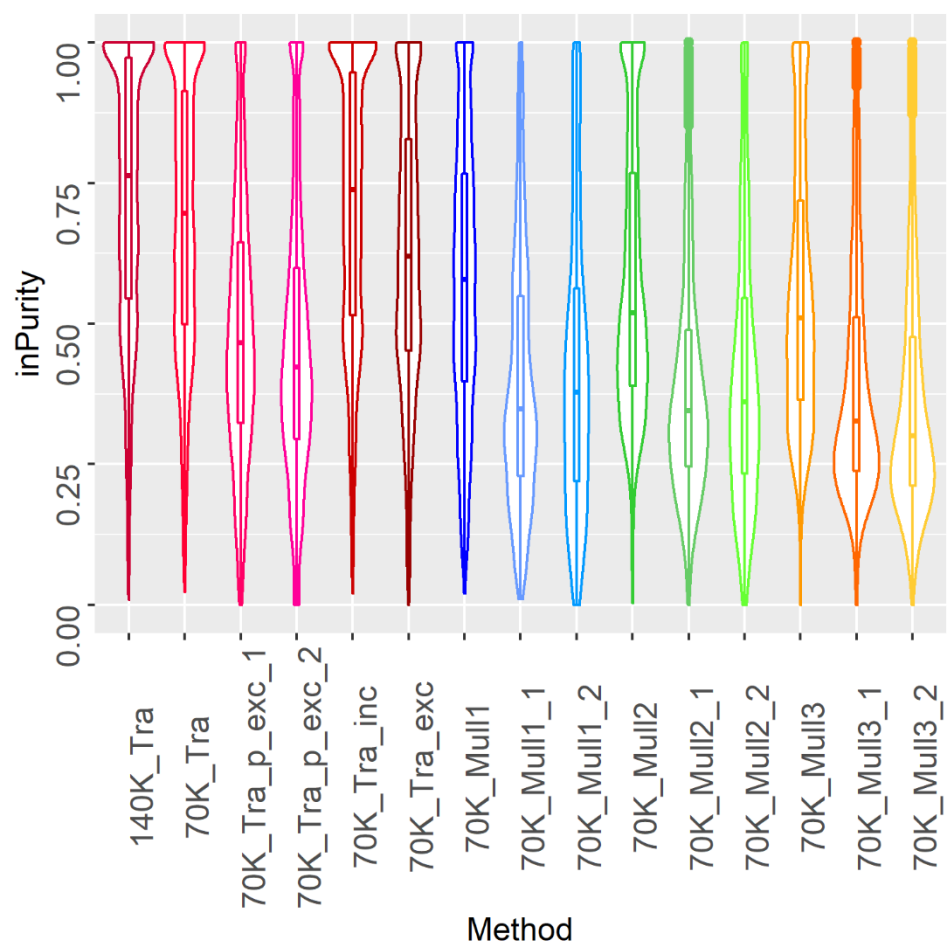


Figure 155: The distribution of interpolated purity (inPurity) scores for all DDA based methods for the Lipids_NEG data.

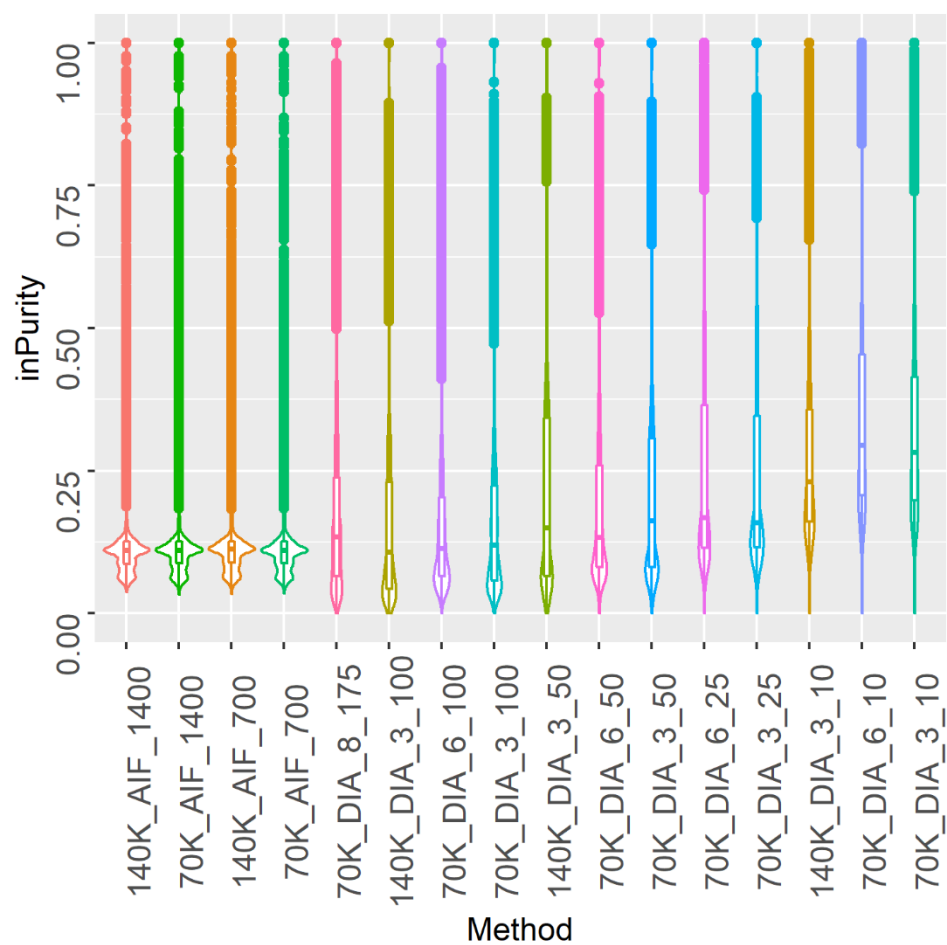


Figure 156: The distribution of interpolated purity (inPurity) scores for all DIA based methods for the Lipids_NEG data

9.3.5 Value of Repeated Injections

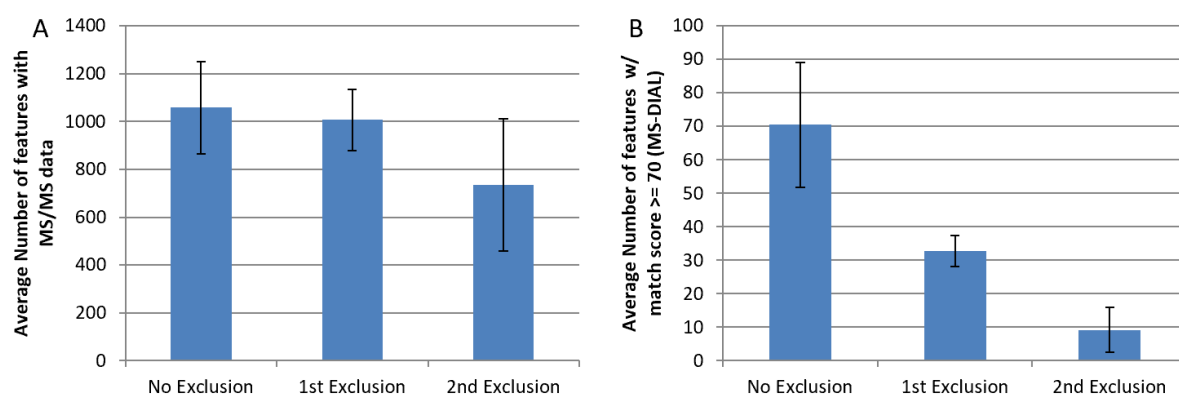


Figure 157: A) The number of features with MS² data associated with them through each pass of the traditional progressive DDA method for the HILIC negative ion mode dataset. B) The number of features with spectral match score ≥ 70 (MS-DIAL) for the same dataset.

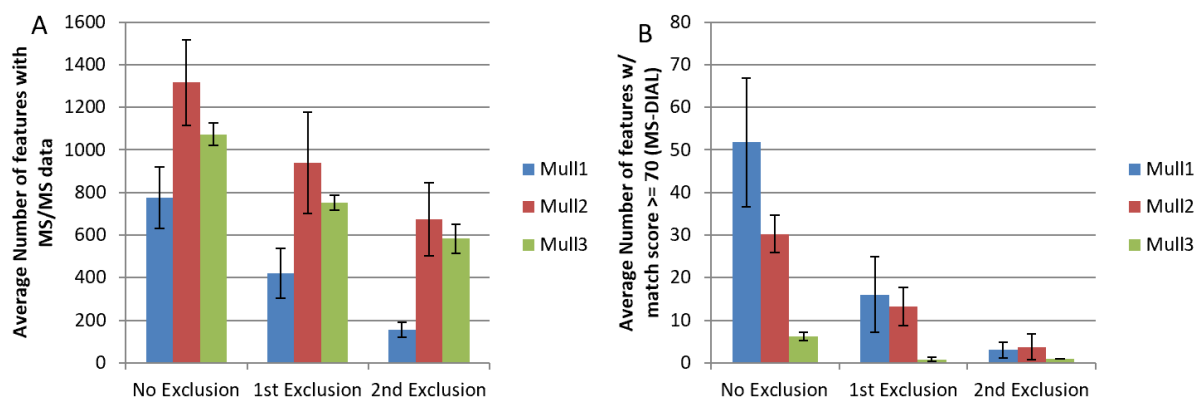


Figure 158: A) The number of features with MS² data associated with them through each pass of the three progressive Mullard segments for the HILIC negative ion mode dataset. B) The number of features with spectral match score ≥ 70 (MS-DIAL) for the same dataset.

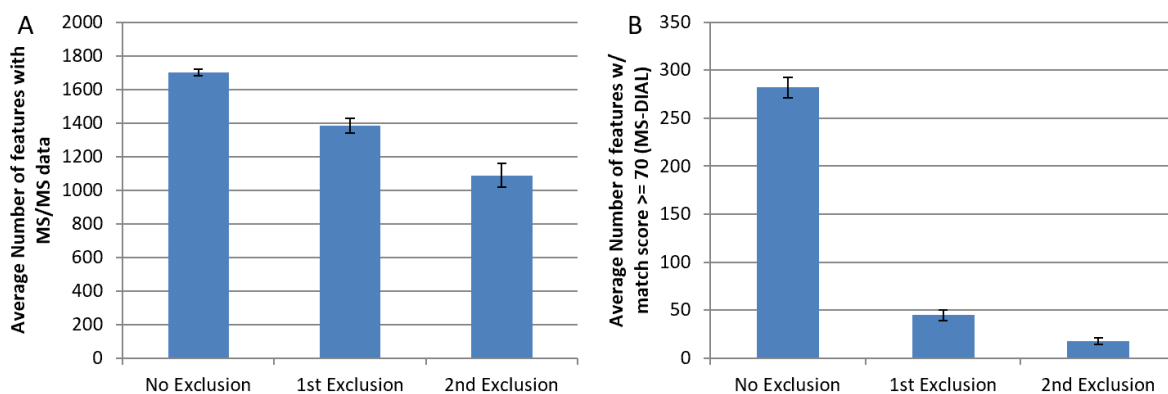


Figure 159: A) The number of features with MS² data associated with them through each pass of the traditional progressive DDA method for the Lipids positive ion mode dataset. B) The number of features with spectral match score ≥ 70 (MS-DIAL) for the same dataset.

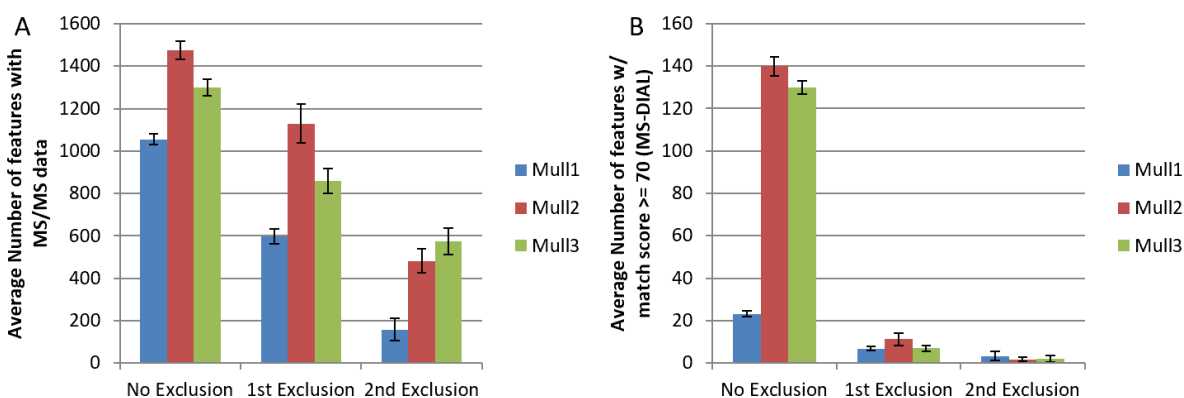


Figure 160: A) The number of features with MS² data associated with them through each pass of the three progressive Mullard segments for the Lipids positive ion mode dataset. B) The number of features with spectral match score ≥ 70 (MS-DIAL) for the same dataset.

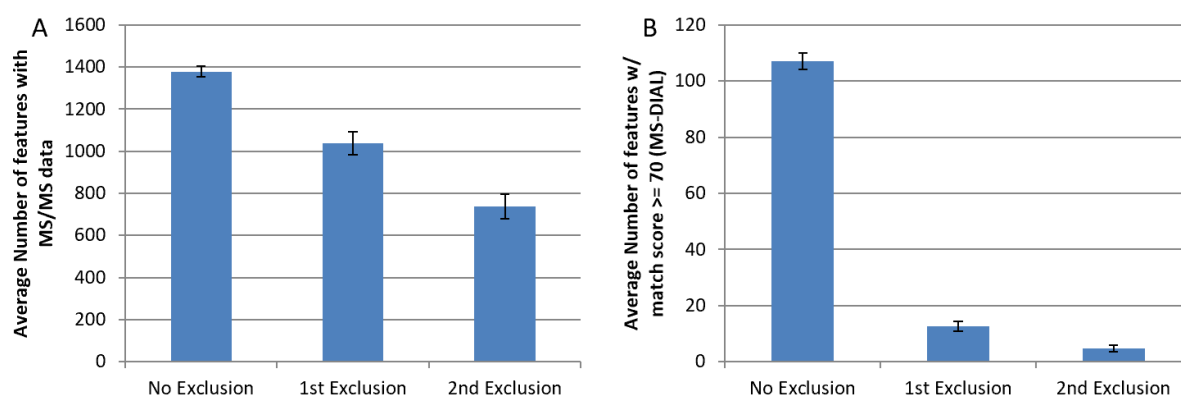


Figure 161: A) The number of features with MS² data associated with them through each pass of the traditional progressive DDA method for the Lipids negative ion mode dataset. B) The number of features with spectral match score ≥ 70 (MS-DIAL) for the same dataset.

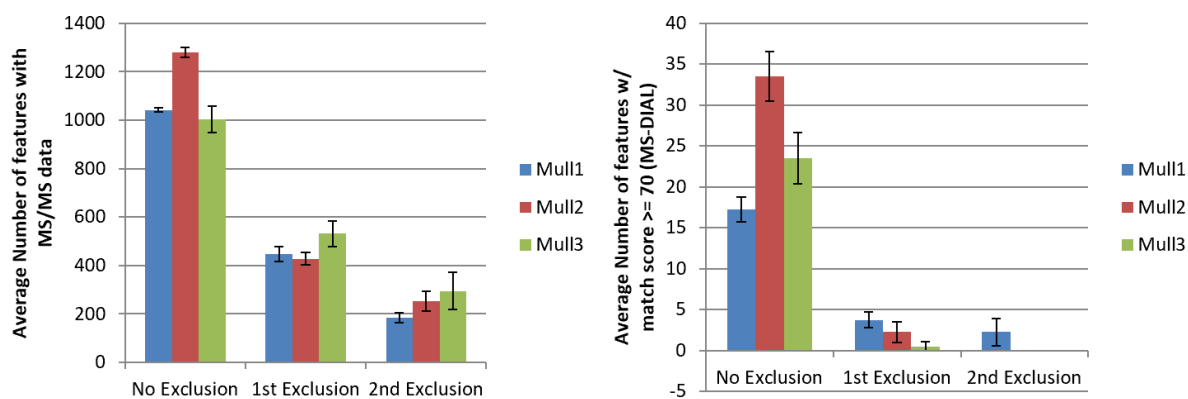


Figure 162: A) The number of features with MS² data associated with them through each pass of the three progressive Mullard segments for the Lipids positive ion mode dataset. B) The number of features with spectral match score ≥ 70 (MS-DIAL) for the same dataset.

9.4 Assessment of Metabolite Annotation Using AcquireX on the Orbitrap ID-X.

9.4.1 How Many Repetitive Injections Are Required?

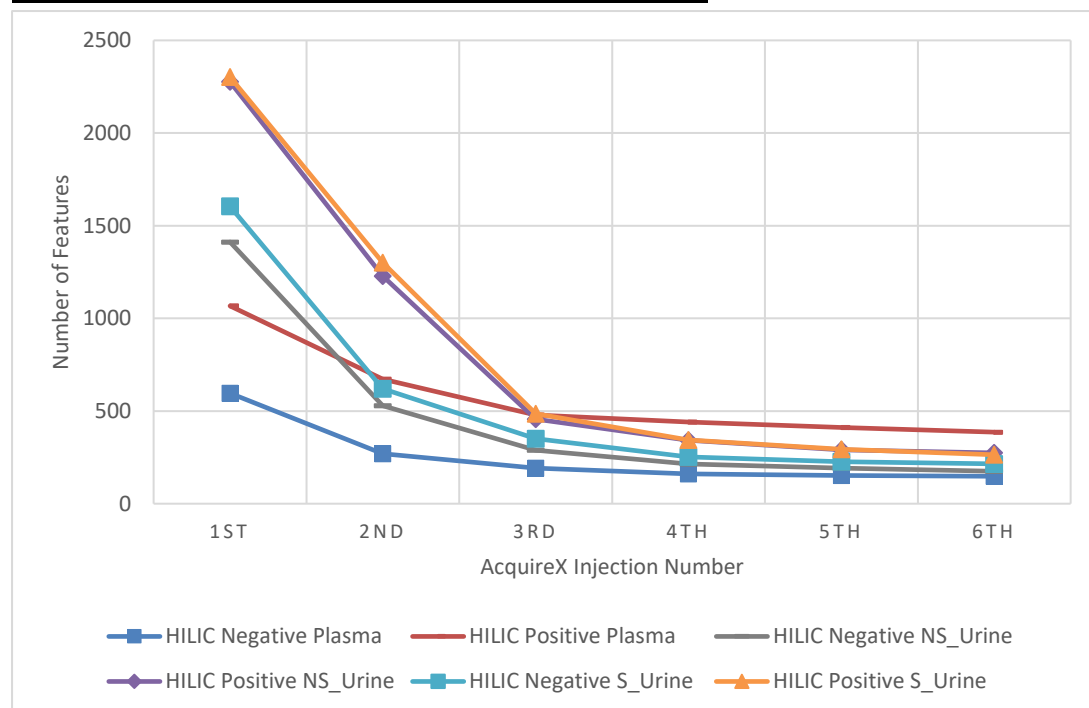


Figure 163: The number of features saved on the automatically updated inclusion list for each injection of the AcquireX sequence for all HILIC/sample type/ion mode combinations.

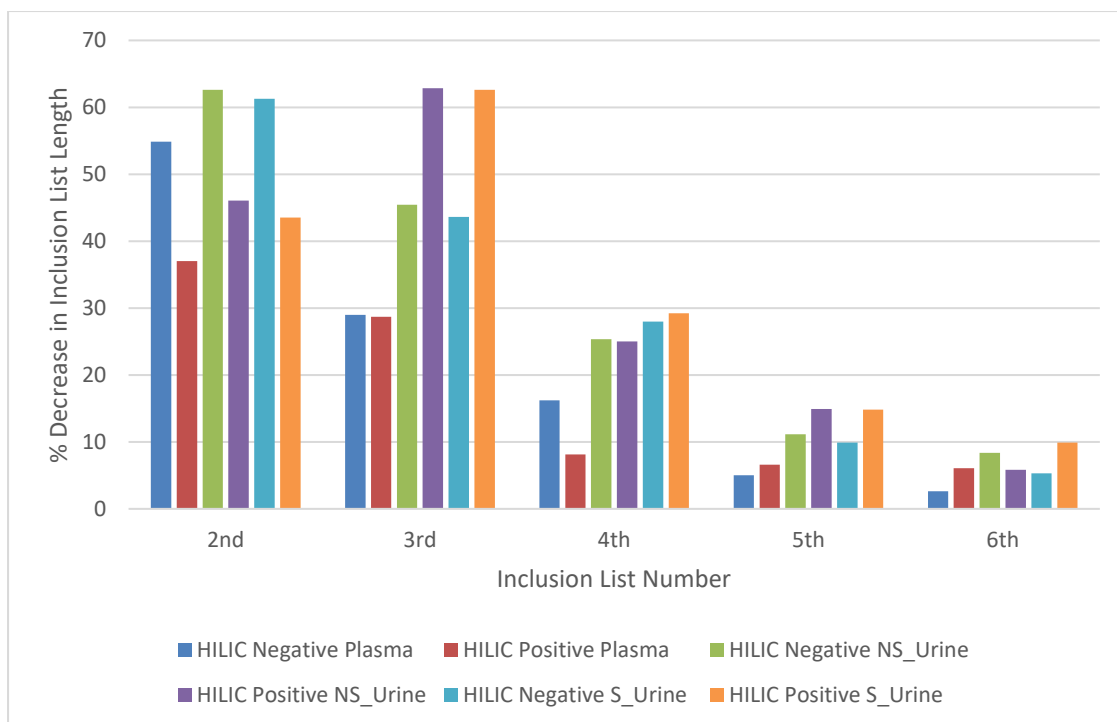


Figure 164: The percentage decrease in length of the inclusion list compared to the length of the previous inclusion list on the method for each updated AcquireX injection for all HILIC/sample type/ion mode combinations.

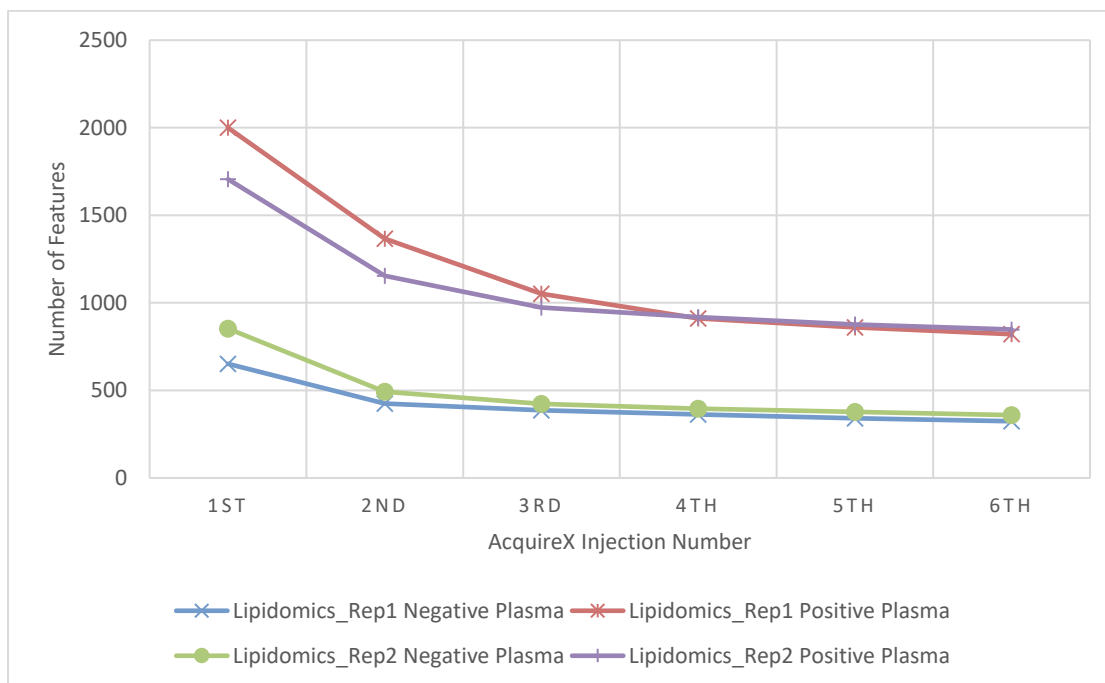


Figure 165: The number of features saved on the automatically updated inclusion list for each injection of the AcquireX sequence for all Lipidomics/sample type/ion mode combinations.

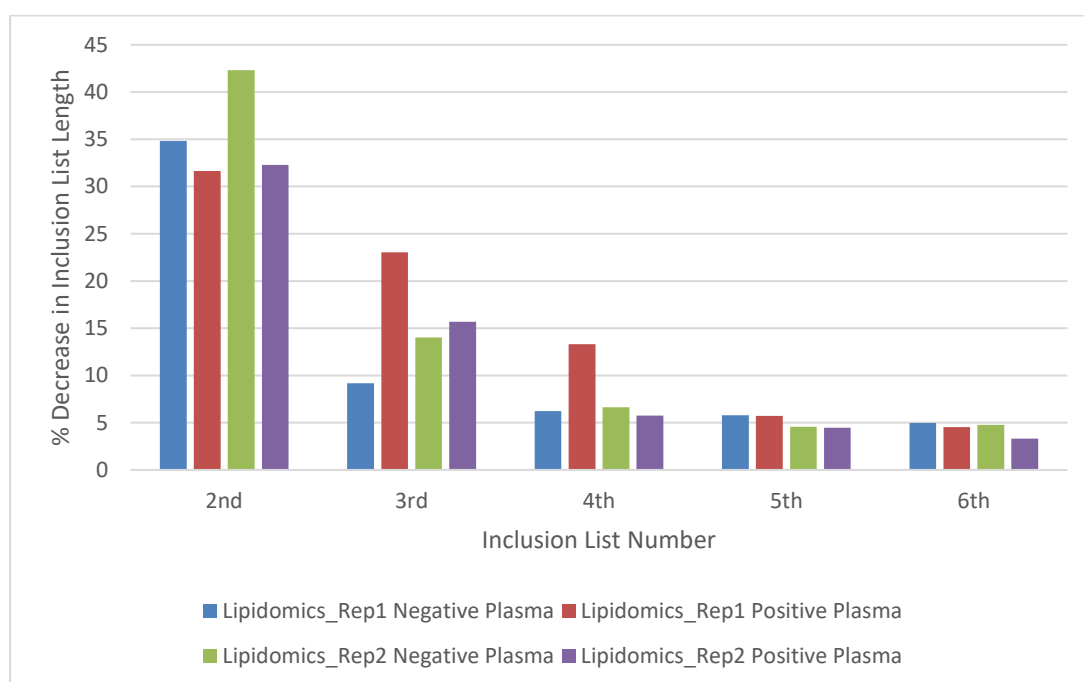


Figure 166: The percentage decrease in length of the inclusion list compared to the length of the previous inclusion list on the method for each updated AcquireX injection for all Lipidomics/sample type/ion mode combinations.