Development of a Software Package for the Quantitative Analysis of Proteomic Mass Spectrometry Datasets Labelled with Nitrogen-15

Philip David Charles

Downing College

Cambridge Centre for Proteomics, Department of Biochemistry

University of Cambridge

28th September 2018

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

Development of a Software Package for the Quantitative Analysis of Proteomic Mass Spectrometry Datasets Labelled with Nitrogen-15

Philip David Charles

28th September 2018

Elemental metabolic labelling using ^{15}N stable isotopes is a technique used in peptide-centric proteomics that allows samples to be mixed before preparation and analysis (minimising technical variance) without introducing sample ambiguity to the results. Labelling with ^{15}N induces a mass shift in labelled peptides that, when analysed by mass spectrometry (MS), allows the signal associated with differently labelled samples to be differentiated.

When compared to similar labelling techniques such as Stable Isotope Labelling by Amino acids in Cell culture (SILAC), ^{15}N poses unique challenges for analysis because the level of label incorporation affects not only the relative intensity of signals in MS analysis, but also how that signal is distributed. A computational signal extraction algorithm is not easily generalised to all peptides, especially if there are differences in the level of incorporation. Analysis of ^{15}N data has been neglected by the general pace of software development in proteomic MS. Furthermore, the current ^{15}N analysis options have relatively complex installation procedures and are limited to a command-line interface.

I describe the development of a cross-platform ^{15}N quantification software package (*HeavyMetL*) which runs inside a web browser, requiring no installation procedure and providing a graphical interface for both the analysis of data and visual interrogation of results (in addition to a more typical text-format table output). The optimisation (using experimental data) of a core part of the algorithm to determine the level of ^{15}N incorporation is described in detail. Finally, the performance of *HeavyMetL* is benchmarked against published ^{15}N labelled data from *Arabidopsis* seedlings quantified by a previously published algorithm, showing that *HeavyMetL* produces quantification of equivalent or better quality.

Acknowledgements

I am hugely thankful to my supervisor, Professor Kathryn Lilley, for her encouragement, guidance and support over the course of this project. I would also like to thank the members of the Cambridge Centre for Proteomics for the friendly and welcoming atmosphere during my time in the lab.

I am particularly grateful to Dr. Sarah Martin, formerly of the Centre for Systems Biology at Edinburgh, University of Edinburgh for advice, several stimulating discussions on ^{15}N labelling and a considerable amount of her time and assistance in generation of the data presented in Chapter 3.

I have received advice and guidance throughout my PhD studies from a great many people, including Dr. Nianshu Zhang and John Roote in Cambridge, Dr. Michael Hastings and Dr. Liz Maywood from the MRC Laboratory of Molecular Biology, and Professor Charalambos Kyriacou and Karen Garner from the University of Leicester. I would like to thank all of them for their assistance.

I am also indebted to Professor Shabaz Mohammed at the University of Oxford for his very helpful guidance on manuscript structure and endless patience in explanation of some theoretical concepts of mass spectrometry, and Professor Benedikt Kessler for his support and understanding during preparation of this Thesis.

Finally, I would like to thank my parents Isabel and Andrew for some eleventh-hour proof reading, and my wife Nisha for her patience and endless support.

For Sam

Preface

The work presented here was carried out in the Department of Biochemistry at the University of Cambridge between October 2008 and December 2011, supplemented by analysis and code development through to submission of this Thesis in 2018. The raw data presented in Chapter 3 were acquired in collaboration with Dr. Sarah Martin at the University of Edinburgh. The benchmarking in Chapter 4 uses publicly available published data retrieved from a public repository. Figures and tables are entirely my own work except where indicated in the text as being derived or adapted from other sources. All web-links given were valid at time of printing. External code libraries used in the work are detailed in Appendix I.

Contents

List of Figures

Appendix II contains seven screenshots, starting on page 187.

Appendix III includes Figure AIII-I on page 194.

List of Tables

List of Abbreviations

Abbreviations are listed in alphabetical order

2DGE	Two Dimensional Gel Electrophoresis
ASB-14	3-[N,N-Dimethyl(3-myristoylaminopropyl)ammonio]propanesulfonate *(amidosulfobetaine-14)*
CEM	Chain Ejection Model
CHAPS	3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate
CID	Collision Induced Dissociation
CCA	α-Cyano-4-hydroxycinnamic acid
CRM	Charge Residue Model
CTAB	Hexadecyl-trimethyl-ammonium bromide *(cetrimonium bromide)*
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
DHB	2,5-Dihydroxybenzoic acid *(gentisic acid)*
EML	Elemental Metabolic Labelling
emPAI	Exponentially Modified Protein Abundance Index
EMT	Electron Multiplier Tube
ESI	Electrospray Ionisation
ETD	Electron Transfer Dissociation
FACS	Fluorescence-Activated Cell Sorting
FDR	False Discovery Rate

FT-ICR	Fourier-Transform Ion Cyclotron Resonance
FWHM	Full Width at Half Maximum
GUI	Graphical User Interface
HDPR	Half Decimal Place Rule
HTML	Hypertext Markup Language
ICAT	Isotope-Coded Affinity Tagging
IEF	Isoelectric Focussing
IEM	Ion Evaporation Model
iTRAQ™	Isobaric Tags for Relative and Absolute Quantification
LC	Liquid Chromatography
LC-MS-MS	Liquid Chromatography in-line with MS/MS Analysis
LFQ	Label-Free Quantification
LIT	Linear Ion Trap
LTQ	Linear Trapping Quadrupole
m/z	Mass-to-charge ratio
MALDI	Matrix Assisted Laser Desorption/Ionization
MDR	Mass Decimal Residual *(the fractional part of a mass value)*
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometer / Mass Spectrometry / Mass Spectrometric *(as appropriate)*
MS/MS	Two-stage (or 'Tandem') MS analysis *(separation of ions by m/z, selection and fragmentation of a particular precursor ion m/z, then separation and detection of the fragment ions)*

MS/MS/MS	Three-stage MS analysis *(separation of ions by m/z, selection and fragmentation of a particular precursor ion m/z, separation of fragments, selection and fragmentation of an ion m/z from the fragment ion population, then separation and detection of the resulting second generation fragment ions)*
MS^1	MS mode in which a spectrum is acquired of the ions entering the MS from the source without fragmentation *(excepting any that occurred in-source)*
MS^2	MS mode in which a spectrum of fragment ions is acquired by MS/MS
MS^3	MS mode in which a spectrum of fragment ions is acquired by MS/MS/MS
MS^n	*n*-stage MS analysis
NSAF	Normalised Spectral Abundance Factor
PAGE	Polyacrylamide gel electrophoresis
pI	Isoelectric Point
PMC	Peptide-Modification-Charge combination *(a peptidoform with particular sequence and modification state, at a particular charge)*
PRM	Parallel Reaction Monitoring
PSM	Peptide-Spectrum Match
ppm	Parts Per Million
ppt	Parts Per Thousand
QqQ	Triple Quadrupole Instrument
RP	Reversed-Phase
RT	Retention Time (specifically, retention time on an LC column)
SA	3-(4-Hydroxy-3,5-dimethoxyphenyl)prop-2-enoic acid *(sinapinic acid)*
SCX	Strong Cation eXchange

SDS	Sodium Dodecyl Sulphate
SILAC	Stable Isotope Labelling by Amino acids in Cell culture
SINQ	Normalised Spectral INdex Quantitation
SIQE	Standardised Incorporation Quantification Error
SRM	Selected Reaction Monitoring
SS	Similarity Score
SWATH	Sequential Window Acquisition of all THeoretical spectra
TMT™	Tandem Mass Tags
ToF	Time-of-Flight

Chapter 1: Introduction

1.1 Proteomic Mass Spectrometry

1.1.1 Background

The last three decades have seen the rise of the '-omics'; large scale multivariate analysis of biological systems. From the birth of genomics in 1977 with the sequencing of bacteriophage ΦX174 (1) the field of systems biology has expanded from DNA through mRNA, proteins and metabolites to new levels of complexity. Proteomics is the study of the proteome (2); the overall state of an organism's temporal protein makeup.

Biological systems are dynamic and involve interactions between and within complexity levels (genome, transcriptome, proteome, metabolome and so on) (3). The state of the transcriptome cannot be predicted based purely on the genome, and the state of the proteome is governed not just by the current state of the transcriptome.

The biological state of the proteome is, at any point in time, encoded not just in relative protein abundance (a dynamic consequence of both protein production and degradation rates (4)), but also in their potential for activity, which depends on their current post-translational modification state ; their folded structures (5); their localisation relative to cellular spatial organisation (6) and the local availability of interaction partner molecules (7) and substrates. This last point introduces a recursive problem for metabolite substrates, since prediction of the metabolome is itself dependent on the proteome, albeit not exclusively, depending not only on (local) protein activity but also local reactant availability (8, 9).

While lagging behind the meteoric rise of genomics as a tool for scientific understanding, the more nuanced view of proteomics has demonstrated that a 'genome-centric view' of biological pathways reveals only a part of the subtle network of interactions that govern the processes of life. The complexity revealed has, as in many avenues of scientific research, resulted in increasingly specialised analyses that focus on tissues or even sub-cellular levels of organisation rather than the relative heterogeneity of whole organisms.

Small-scale proteomic analysis using classical targeted approaches such as western blotting (10), and even larger experiments based on two-hybrid models (11) have been employed for several decades, but the area of systems biology that might be thought of as modern proteomics has coalesced around two approaches that have exploited technological developments in Mass Spectrometry (MS) - the analysis of intact proteins ('top-down' proteomics), and the analysis of peptides ('bottom-up' proteomics) discussed in Section 1.1.2.2 below.

1.1.2 Mass Spectrometry as a Proteomic Analysis Tool

1.1.2.1 Analytical Constraints

As a measurement of physical phenomena, mass spectrometry is subject to three principles which apply broadly across most observational techniques.

I. More abundant entities are easier to detect, and to measure accurately, having higher signal-to-noise.

II. It is easier to analyse samples of lower complexity. Both the number of entities and the range of entity abundances contribute to complexity.

III. Difficulties stemming from points I and II require additional effort to address. For this reason, there is always a trade-off between sensitivity, robustness and time for analyses with any given instrument.

As a simple example of such principles, consider an observer counting pixels in a small image (Figure 1-I-A). There are 950 black pixels, 20 dark grey, 20 light grey, and 10 white. It is easy for the observer to conclude, just at a first glance, that there are a lot of black pixels. The observer can easily conclude that of the 1000 pixels, almost all are black; a quick estimate that all 1000 pixels are black will have relatively high accuracy. Counting the white and grey pixels is more difficult, as an over-count or under-count by one will cause a significant relative change in the count (Principle I). The 20 dark grey pixels may be easily mistaken for black (especially on a low resolution computer screen), so separating the pixels by colour before counting will make the counting of the non-black pixels much easier (Figure 1-I-B). In contrast, if the 1000 pixels were scattered at random through a large image of 10,000 pixels in another shade of dark grey, accurate evaluation of the (previously simple) black pixel abundance would itself become challenging (Principle II) (Figure 1-I-C,D). If observation time is limited to a few seconds, then the observer can conclude quickly if a specific colour of pixel is present or not, but will not necessarily be able to evaluate the actual proportion of pixel colours, except to say that there are many more black (small image) or dark grey (large image) pixels. If, however, the observer is given more time to analyse the image, it is reasonable to assume they will be able to giver a better evaluation of relative pixel proportions (Principle III). To extend the analogy, an observer with better eyesight or faster tallying method may be able to make a more accurate assessment of the image in the same time frame, or alternatively they could apply the same standard of assessment as the original observer to multiple images in the same time frame. In the context of a mass spectrometry based proteomic experiment, this might correspond to using an instrument with a more sensitive detector or a faster scan speed.

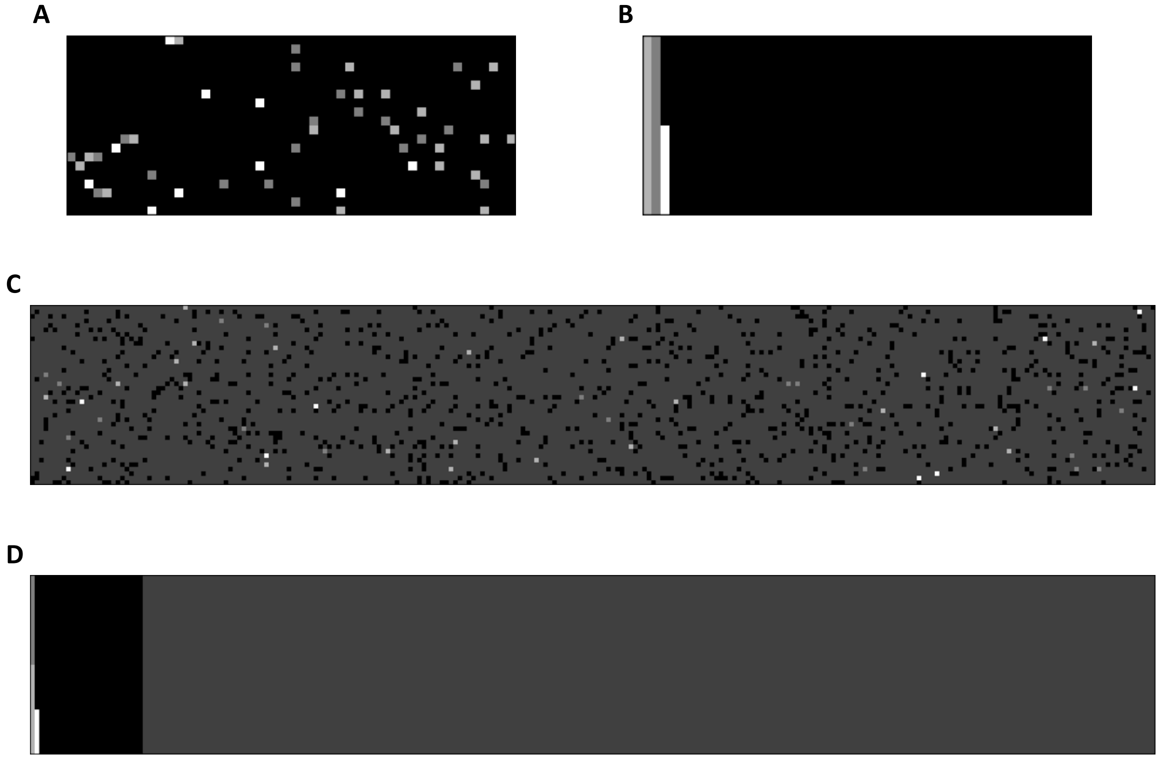


Figure 1-I. *Principles of Observation.* ***A*** *&* ***B:*** *These images each contain 1000 pixels in the ratio 950:20:20:10 for black:medium grey:light grey:white respectively. Estimation of the number of black pixels is relatively straightforward as they are the majority of the data, even an estimate that all 1000 pixels are black will be relatively accurate. Assessment of the other colours cannot be performed as easily with a similar level of relative error, but the task can be rendered easier if the pixels are sorted by colour first (****B****).* ***C*** *&* ***D:*** *These images contain the same 1000 pixels as A and B, and an additional 9,000 dark grey pixels. Now, even counting the black pixels becomes challenging (****C****), and while the image may again be sorted to separate the colours (****D****), an at-a-glance estimate of anything but the number of the dominant dark grey pixels is more difficult.*

Samples analysed in proteomics are typically a protein mixture obtained from a cell or tissue preparation (12), although analyses of exogenous proteins and peptides in fluids (e.g. urine (13), blood (14, 15)) or extracellular matrices (e.g. plaques (16), biofilms (17)) are also widespread. A simple preparative protocol might, for example, involve lysis of tissue sample cells in a buffered detergent, followed by precipitation of the protein content and re-suspension of the pellet in a low pH buffer. Biochemical techniques for the general extraction of the protein mixture are not a focus of this thesis, but the same considerations apply as to the elements discussed below – minimisation of sample loss to keep the absolute and relative amount of protein high (Principle I), minimisation of sample complexity, such that the proteome or proteome components of interest are not swamped by other biomolecules (Principle II), and minimisation of technical variance so that accuracy of measurement is as high as possible within a feasible number of replicates (Principle III).

Even with optimised preparation of the protein mixture, proteomic samples are generally relatively complex, so ‘modern’ proteomic mass spectrometry routinely involves some strategy to reduce sample complexity before mass spectral analysis, to maximise the resolution and depth of obtained spectra. Descriptions of exemplary experimental designs can be found as early as 1993 when Henzel and colleagues described a strategy of separation of proteins by two dimensional electrophoresis, tryptic digestion and MS analysis of excised spots, and identification of proteins by comparison against *in silico* digestion of a database of possible candidates (18).

1.1.2.2 Methodologies

There are two broad methodological approaches to the analysis of proteins by MS, which distinguish between the MS analysis of intact proteins, or alternatively the MS of constituent peptides (predominantly generated by protease digestion). Respectively, these are commonly referred to as ‘top-down’ (19) and ‘bottom-up’ (20). Initially, this distinction in analytical method referred to the point of separation, so the methodology of Henzel *et al.* would be considered to be top-down as the separation step (by 2D electrophoresis) is performed on intact proteins. The definition has shifted over time to refer to ‘*the composition of the sample at the point of introduction to the MS’* (21). Since in Henzel *et al.* the samples are digested before MS analysis, the analyte is a mixture of peptides, so their methodology would now be considered bottom-up.

Preserving protein compositional integrity in top-down proteomics facilitates discrimination between isoforms and modification states and is often preferred when such distinctions are key to understanding results. However, the heterogeneous proteomic composition of many biological samples presents many challenges for intact analysis. The diverse array of physiochemical characteristics can frustrate attempts to find an acceptable set of analytical conditions, or to achieve sufficient separation such that the acquired mass spectra can be de-convoluted and interpreted.

Conversion of the protein mixture to a population of derived peptides effectively imposes soft limits on the length of the polypeptide chains present in the sample, which narrows the distribution of the physiochemical characteristics in the population. In effect, population physiochemical complexity is reduced at the expense of population sequence complexity, as each protein is converted into a mixture of peptides. Peptides with a particular amino acid sequence may be generated by digestion of several different protein species, resulting in a loss of information that can frustrate the inference of protein-level results based on the peptide data. This so-called Protein Inference Problem (22) is discussed further in Section 1.1.2.10.2. Nevertheless, the reduction in molecular weight and physiochemical complexity broadens the range of applicable sample separation techniques (in particular, the use of liquid chromatography) and allows MS to be optimised for a 'general case' peptide sample scenario.

Both top-down and bottom-up strategies have been employed in various incarnations for studies of the proteome; however, the bottom-up approach is the sole methodology applicable to all the experiments described in this thesis. Further discussion of proteomic mass spectrometry and the associated workflow will be confined to this area.

1.1.2.3 Sample Preparation

As discussed in the pixel analogy, better results are obtained when analyte complexity is as low as possible. Techniques to separate intact proteins (i.e. prior to digestion), while playing an important role in top-down proteomics, have been somewhat overshadowed in bottom-up analyses due to the ascendancy of in-line liquid chromatography peptide separations. Nevertheless, such techniques continue to play an important role in bottom-up strategies, as they allow discrimination between protein isoforms that would otherwise be impossible to distinguish post-digestion due to very high sequence overlap resulting in many, if not all detectable peptide products being shared. Separation by polyacrylamide gel chromatography is generally employed on intact proteins, since they have a much greater spread of masses than a peptide mixture. By the same token, the relatively narrow distribution of peptide lengths which makes them so amenable to liquid chromatographic separation (see below) places high requirements on gel chromatography in terms of mass resolution and reproducibility.

Protein (and peptide) separation techniques may also be employed on relatively simple samples for ‘clean up’ purposes in order to make them more amenable to MS analysis. Many contaminants do not have a significant interaction with common separation mechanics, remaining in the loading buffer or eluting before the first protein/peptide fraction is collected. This section will discuss key bottom-up workflow steps in preparing a sample for MS analysis starting from a point of a relatively pure protein mixture, as the preceding steps relating to cell lysis/extracellular protein collection and depletion of non-protein biological components are specific to experimental design and biological context.

1.1.2.3.1 Protein Separation

Protein separation by mass using Polyacrylamide Gel Electrophoresis (PAGE) is a venerable (23) and adaptable technique employed widely in proteomics (24). In the most widely used incarnation, protein samples are prepared for PAGE separation by boiling in Sodium Dodecyl Sulphate (SDS). The combination of the anionic surfactant SDS and the disruption due to heat results in general denaturation of the protein mixture; this popular pairing of techniques is referred to as SDS-PAGE. Denatured proteins are loaded onto a polyacrylamide gel and a constant electric field is applied across the gel, causing proteins to migrate through the gel towards the anode. In their denatured state, lower mass proteins travel more rapidly through the gel matrix, thus the population becomes spread over the length of the gel according to mass. In cases where the use of SDS results in poorly resolved bands (for example, glycoprotein samples), alternative denaturing reagents such as the cationic surfactant hexadecyl-trimethyl-ammonium bromide (cetrimonium bromide, CTAB) may be used instead. After separation, proteins may be visualised generally on the gel by Coomassie or Silver Stain, or specifically by techniques that use antibody detection (such as western blotting). Sections of a gel lane may also be excised, and the proteins then extracted for further investigation. For analysis by MS, the proteins may be digested in-gel to peptides, which are able to diffuse out of the gel matrix.

In principle, the ability to visualise the mass distribution of proteins prior to MS analysis yields information regarding the relative abundance of different protein isoform/modification states which is frequently intractable to ‘bottom-up’ analysis. However, a single dimension of separation by mass alone provides insufficient resolution for analysis of all components in a protein mixture, even for relatively low complexity examples. For this reason, the pre-eminent approach to proteomics in the late 1990s/early

2000s involved two dimensions of separation wherein proteins were first separated by isoelectric point (pI) on an immobilized pH gradient gel (termed isoelectric focussing, IEF) and then subsequently by mass on an SDS-PAGE gel, termed two-dimensional gel analysis (2DGE). This approach has retained some popularity, particularly for isoform and modification state visualisation, but for general proteomic analysis has been found to have a number of disadvantageous qualities (25). In particular, the technique yields very poor coverage of membrane proteins, which frequently resist solubilisation in IEF-compatible zwitterionic detergents such as 3-[N,N-dimethyl(3-myristoylaminopropyl)ammonio]propanesulfonate (amidosulfobetaine-14*,* ASB-14) or 3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) due to hydrophobic transmembrane domains. Furthermore, the resolution in both 2DGE dimensions is insufficient to separate proteoforms with similar pI and mass, thus a single 2DGE spot may contain more than one protein, frustrating quantitative comparisons. The reproducibility of separation is further degraded by even small inconsistencies in gel casting, an effect which increases with gel size (larger gels are otherwise desirable for better resolution). There are also experimental practicality limits on the size of the gel apparatus and the feasibility of performing a large cohort study. These factors have driven 2DGE to be largely supplanted by liquid chromatography peptide separation (Section 1.1.2.4) as the predominant proteomic separation approach.

1.1.2.3.2 Digestion

Peptides must contain enough information to be distinctive whilst being short enough to limit their physiochemical characteristics to an acceptable range. In general, modern bottom-up proteomic MS analysis techniques are optimised for peptides containing between 6 and 20 residues (26). The most widely used method of generating peptides from parent proteins is by digestion by proteolytic enzymes. The enzyme trypsin is frequently employed as it has a well characterised, highly specific cleavage pattern (C-terminal to lysine or arginine residues not followed by a proline) that tends to generate a large number of peptides favourable to MS analysis. The distribution of arginine and lysine in the proteome sequence of most organisms is typically such that peptides of desirable length are generated from most proteins, whilst ensuring every peptide has at least one proton accepting group at each peptide terminus (favouring a minimum charge state of 2+ after positive ionisation). Trypsin is generally utilised in a commercially available modified form whereby reductive methylation of lysine residues reduces the autolytic properties of the protease.

While trypsin is overwhelmingly popular due to agreeable cleavage properties, efficient kinetics (close to 100% digestion may be achieved within an hour or less (27)), and amenability to a broad range of buffer conditions, other proteases may also be used in conjunction with or instead of trypsin, particularly when the distribution of lysine and arginine in the proteins of interest lead to a suboptimal distribution of peptide lengths (and thus achievable coverage). Historically, this would typically be an avenue explored if initial tests with trypsin indicated poor coverage of proteins of interest. However, very recent advances in the ability to predict the amenability to MS identification of particular peptide sequences may ultimately lead to more frequent selection of trypsin alternatives (28). The extent to which a proteome is covered by observable peptides resulting from tryptic digestion varies substantially from species to species. In humans, the average tryptic peptide length is 14 amino acids (29), but in yeast, the average length is only 8.4, with 56% of tryptic peptides generated having a length of 6 amino acids or less (30), yielding a high proportion of non-observable peptides. Common alternative proteases include LysC, LysN, AspN, GluC and chymotrypsin. The properties of popular proteases are listed in Table 1-I. Proteases with varying degrees of non-specificity (i.e. a propensity to cleave at random or loosely defined residue motifs) including elastase and pepsin may also be used; these present an additional challenge for data interpretation as the partially random cleavage means the range of possibly generated peptides is much larger and the resulting peptide mixture less comparable between experiments. Increasing interest has also been paid to the analysis of peptides generated by *in vivo* protein cleavage, particularly in the context of immunology, for example, the endogenous peptides displayed on major histocompatibility complexes (class I and II) (31, 32).

Protease	Type	Specificity
ArgC	Cysteine Protease	C-terminal to R (high efficiency) C-terminal to K (lower efficiency)
AspN	Metalloprotease	N-terminal to D (high efficiency) N-terminal to E, in presence of detergent (low efficiency)
Chymotrypsin	Serine Protease	C-terminal to FLMWY (varying efficency, many missed cleavages)
GluC	Serine Protease	C-terminal to E C-terminal to D (at pH 8)
LysargiNase	Metalloprotease	N-terminal to KR
LysC	Serine Protease	C-terminal to K
LysN	Metalloprotease	N-terminal to K
Pepsin	Aspartic Protease	C-terminal to FWY pH-dependent broader specificity
Trypsin	Serine Protease	C-terminal to K when not followed by P (lower efficiency) C-terminal to R when not followed by P higher efficiency) In addition, negatively charged residues e.g. DE & phospho-ST adjacent or proximal to cleavage site reduce efficiency.
WaLP and MaLP	Serine Protease	C-terminal to aliphatic residues

Table 1-I. *Site-specific Proteases Popular in Proteomics. Adapted from Giansanti, P.* et al.*, 2016 (33).*

1.1.2.4 Peptide Separation

While proteolytic digestion of a typical proteome is effective for reducing the range of physiochemical properties by limiting maximum polypeptide length, this still results in a highly complex sample. For example, tryptic digestion of the human proteome (approximately 20000 genes) results in millions of peptides with abundances ranging across seven orders of magnitude (34), and thousands of these peptides share very similar *m/z* ratios. Consequently, it is widespread practice to include one or more stages of peptide separation before MS analysis in order to reduce complexity (as per Principle 2 in the pixel analogy) and increase the resolving power of the analysis. Improved resolving power decreases the number of peptide species competing for ionisation at the same time and thus reduces the possibility of failing to observe low abundance (or poorly competitive) ion species whose signal would otherwise be suppressed below detection limit (35).

The most widely used separation technique in bottom-up proteomics is Liquid Chromatography (LC). Peptides are soluble in a wide range of solvents and LC separation may be combined in-line with Electrospray Ionisation (see below) to feed eluting peptides directly into the mass spectrometer.

LC separates a sample between a column-immobilised matrix (the stationary phase), and a solvent passed through the column (the mobile phase). Analytes exhibit a range of affinities for the stationary phase relative to the mobile phase and thus can be eluted over time during which the composition of the mobile phase can be changed (the 'gradient'). In proteomics, the column is usually prepared by 'packing' with the stationary phase in the form of silica beads that have the desired surface chemistry. LC is a powerful tool for proteomic separation as it can be highly optimised for the separation of a particular mixture. As well as stationary phase chemistry, separation resolution is also affected by flow rate, column diameter and length, stationary phase bead size and the design of the solvent gradient. Two LC techniques are widely used in proteomics: Reversed-Phase and Strong Cation Exchange, both of which are straightforwardly compatible with Electrospray Ionisation, either directly in-line (Reversed-Phase) or with minimal further processing (Strong Cation Exchange).

1.1.2.4.1 Reversed-Phase Chromatography

Reversed-Phase *(*RP) chromatography separates analytes by their hydrophobicity. Peptides generally contain a high proportion of hydrophobic or uncharged amino acids and thus are particularly suited to this form of separation. Analytes are partitioned between a hydrophobic stationary phase and a polar hydrophilic mobile phase. The mixture is loaded onto the stationary phase under low-organic solvent conditions and analytes selectively return to the mobile phase as the organic component of the mobile phase is increased (either as a continuous gradient or in a series of steps). The order of elution relates to the strength of the hydrophobic interactions of each peptide with the stationary phase. The stationary phase typically consists of a C_{18} resin, i.e. silica beads derivatised to present alkane chains of 18 carbons. The mobile phase is usually based on a mixture of water and acetonitrile (as the organic component). One drawback of silica-based matrices is the interaction of residual silanol groups with the positive charges of peptides. This effect can be minimized by lowering the pH below 4, as silanol groups then become protonated. The column surface can also be treated to 'end-cap' the polar surface silanol groups with a non-polar trimethylsilyl group.

The use of a low pH aqueous/acetonitrile mobile phase is ideal for electrospray ionisation (see below), so low pH RP-LC is a very popular technique for in-line separation immediately prior to introduction into the mass spectrometer. It has become increasingly popular to precede this with an initial high-pH RP-LC separation step for greater resolution (see Section 1.1.2.4.3)

Advances in LC technology have enabled the development of systems which operate at high pressures, allowing smaller bead sizes in the column (increasing the relative surface area for hydrophobic interactions) and lower flow rates (improving MS ionisation efficiency and reducing the amount of sample required for sensitive analysis). High-Pressure/High-Performance and, more recently, Ultra-High-Performance LC systems have been rapidly adopted in proteomics.

1.1.2.4.2 Strong Cation Exchange Chromatography

Strong Cation eXchange (SCX) chromatography separates molecules by the number of positively charged residues they contain. Just as free amino acids are zwitterionic, so are peptides, thus they have predominantly net positive charge when the pH is lower than the pK_a values of both the N-terminal amino group conjugate acid (such that it is protonated) and the C-terminal carboxylic acid group (such that it is neutral). This threshold is determined by the lower of the two pK_a values which is that of the carboxylic acid, approximately 3.1 (36). At this threshold, arginine, lysine and histidine will also carry a second positive charge due to protonation of their guanidino (pK_a 12.5), lysyl (pK_a 10), and imadazole (pK_a 6) side chains respectively (36). Under typical SCX buffer conditions (pH 2.5 to 3), tryptic peptides (containing one arginine or lysine) will (on average) carry a net charge of +2, further increased by one for each histidine residue or missed cleavage (additional arginine or lysine). The peptides are loaded onto a stationary phase consisting of exposed negatively charged groups, which interact with the positively charged peptides. Peptides may then be eluted by increasing the concentration of salt ions in the mobile phase such that the peptides are outcompeted for electrostatic interaction with the stationary phase.

Relative to RP methodology, SCX does not offer as fine a control over the elution process. The quantised nature of charge means a large number of peptides tend to elute closely together, particularly groups of doubly and triply charged peptides that contain zero or one histidine or missed cleavage. An additional problem is that samples must be desalted after separation, which can be a relatively high variance step. This prevents salt ions from competing for charge as the sample is introduced into the mass spectrometer, causing ion suppression.

1.1.2.4.3 Multidimensional Separation

Resolution in chromatographic separation may be improved by increasing the number of fractions into which the sample is divided. Many peptides have properties that are sufficiently similar that they co-elute under particular LC conditions, imposing a practical limit to peptide separation. One solution to this problem is to further separate each sample fraction using an orthogonal fractionation method, producing a series of fractions-of-fractions with reduced complexity. Separation efficiency is measured in peak capacity, defined as the number of peaks which may be separated without overlap, assuming peaks to extend four standard deviations from the apex (37). The maximum achievable separation efficiency will be the product of the peak capacity of each separation step. This can enable a much higher resolution of peptides to be achieved if separation steps are sufficiently orthogonal, either by separating differently according to the same physiochemical properties, or by separating on different physiochemical properties. Assuming that the final stage of separation is in-line low pH RP chromatography, then options for prior dimensions include alternative aqueous/organic partitioning-based methods (such as high pH RP or Hydrophilic Interaction Chromatography(38)), or charge based methods such as SCX.

Two-dimensional chromatographic separation of peptides has for many years been restricted to SCX followed by low pH RP. SCX-low pH RP is not ideal as the orthogonality of SCX to RP (and thus the peak capacity) can be limited when there is a relatively small range of peptide charges (which is the case when employing tryptic digestion). Recently, the use of high pH RP instead of SCX as a first dimension of separation has been gaining traction as a strategy. At high pH, the set of amino acids contributing to RP chromatographic behaviour is different to the set contributing at low pH. At low pH, positively charged residues are surrounded by counterions from the buffer which strongly affect peptide retention. Due to the different protonation and deprotonation pK_a values for the various side chains, a large change in pH substantially alters the charge distribution within the peptide, which in turn affects ion pairing and thus retention (39), resulting in different separation profiles. Studies of separation at pH 10 have shown high pH RP to be orthogonal in terms of separation to the more usual low pH approach (39, 40), with high-pH RP–low pH RP approaches demonstrating a greater peak capacity than SCX-low pH RP. Two dimensional RP is now frequently employed in modern studies where high protein coverage is desirable and the additional MS time required to analyse the separate fractions can be justified (41).

1.1.2.5 Mass Spectrometer Architecture

Mass spectrometry evaluates the ionic composition of a gas phase analyte, reporting the mass-to-charge (*m/z*) ratios of constituents, and their relative intensity. Only charged species can be observed. The primary functions of MS analysis are therefore (i) conversion of the analyte to gas phase, with the molecules of interest within the mixture carrying charge, (ii) manipulation of the population of ionised species to maximise sensitivity of the detection method, and (iii) detection of each ion species. The components of a mass spectrometer fulfilling these functions are described respectively as (i) Ion Sources, (ii) Mass Analysers and (iii) Detectors. Mass analysis and detection may be combined within the same device and are discussed together here. Technological developments in instruments and the associated prior (sample preparation) and posterior (data analysis) workflows have largely been driven by the principles of observation described at the start of this chapter.

1.1.2.5.1 Ion Sources

To be analysed by the mass spectrometer, the analyte must be both charged and in the gas phase so that it can be easily manipulated by electromagnetic fields inside the instrument. The optimum method of ionisation depends on the starting phase of the analyte and the stability of the target molecules. An early technique developed for ionisation was brute force electron bombardment of samples already in the gas phase (Electron Ionisation - EI). EI is an example of a ‘hard’ ionisation technique, in that it imparts a large amount of energy to the ionised molecules, frequently resulting in the fragmentation of the molecule to a series of smaller ions. In relatively low complexity samples (such as small molecules) this may be desirable, as fragmentation can yield additional information about the molecular structure. However, for complex samples such as peptide mixtures, the resulting mixture of fragments cannot be readily deciphered, or even analysed in a robust fashion. For the analysis of peptides, it is desirable that the analytes survive ionisation intact. In proteomics, there are two popular so-called ‘soft ionisation’ techniques that can achieve this result: Electrospray Ionisation (ESI) and Matrix Assisted Laser Desorption (MALDI).

1.1.2.5.1.1 Electrospray Ionisation

In Electrospray Ionisation (ESI) a strong electric field (typically 2-3 kV in modern nanolitre flow range LC) is applied under atmospheric pressure to a liquid passing through a capillary tube with low flow rate. Charge accumulates at the liquid surface located at the end of the capillary (the 'emitter'). The liquid at the tip of the emitter forms a Taylor cone (42) and a jet of liquid is ejected from the centre. This jet rapidly disintegrates into small droplets, which are dispersed radially by Coulomb repulsion, and the droplet plume is sampled into the first vacuum stage of the mass spectrometer, where the droplets evaporate leaving charged analyte ions (Figure 1-II). Evaporation may be assisted by use of an inert nebulising gas such as nitrogen.

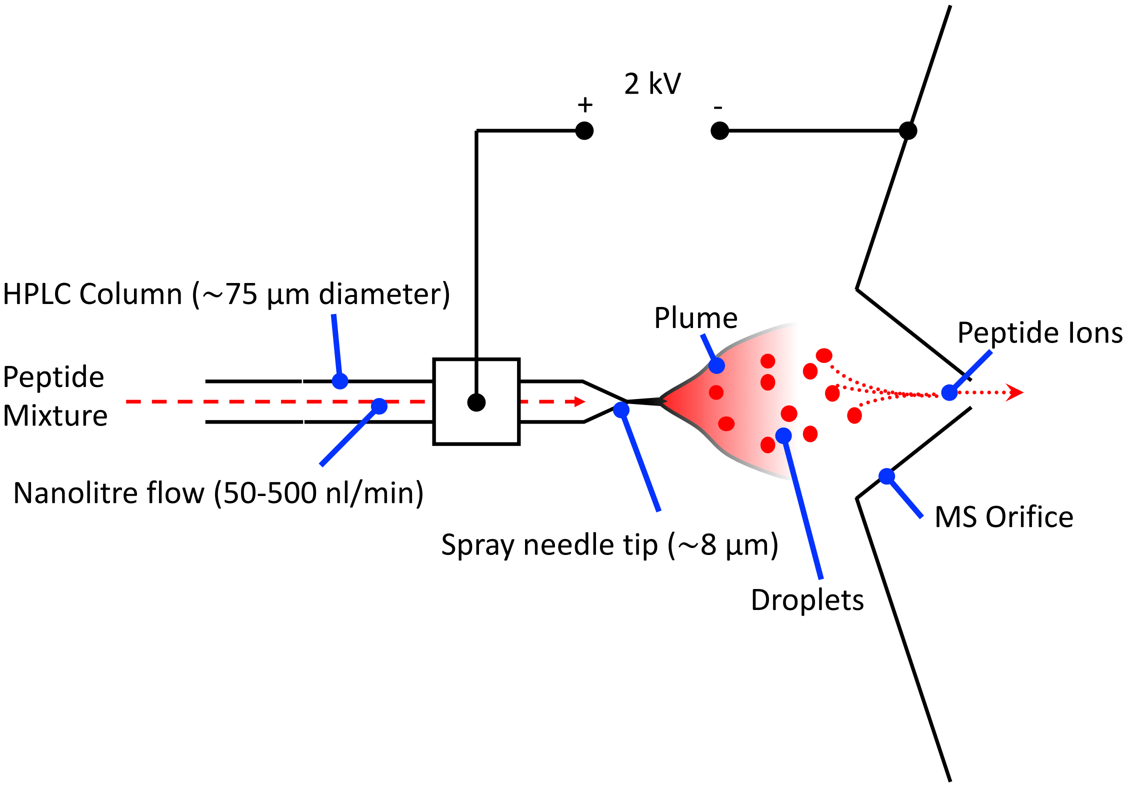


Figure 1-II. *Electrospray Ionisation of Peptides from Nano-Flow HPLC Column. Adapted from Steen, H., and Mann, M., 2004 (43).*

ESI was originally described by Fenn *et al.* (44) for the analysis of intact proteins, as the charge states achieved allow high mass analytes (such as proteins) to be brought within an *m/z* range amenable to MS analysis. ESI has remained a popular technique in proteomics as the analyte mixture is charged in a liquid state. Because of this, ESI can be performed in-line with LC, allowing protein and peptide mixtures to be separated (to reduce mixture complexity) and then charged and introduced into the MS in a single unified workflow. ESI in-line with an LC system is the most popular and widely used method of sample introduction in proteomic mass spectrometry. Ionisation efficiency is improved by reducing the initial size of the droplets formed, yielding greater sensitivity for a smaller amount of sample. This also reduces competition and suppression effects where the total available charge is insufficient to ionise all eluting species and due to preferential ionisation of the most kinetically amenable species, may result in low abundance less amenable species being masked. The drive towards smaller droplet sizes and lower flow rate lead to the development of micro- and nano-electrospray (45) techniques. Modern nano-ESI is usually operated at flow rate of 50 to 500 nL/min and the use of a nebulising gas to maximise droplet evaporation is no longer necessary.

The exact mechanism by which solvated ions are transferred to gas phase is not completely understood, and there are two competing theories that explain it. Under the vacuum conditions, solvent evaporates from the droplets until they reach their Rayleigh limit (46). At the point of instability, the droplet deforms and undergoes Coulomb fission, emitting charged jets of liquid. According to the Ion Evaporation Model (IEM), the surface area of the droplet eventually becomes small enough to support field desorption of ions from the droplet surface (47). According to the Charge Residue Model (CRM), fission of the droplets continues until they contain an average of only one analyte molecule, at which point evaporation of the remaining solvent molecules leaves the analyte molecule in the gas phase with the remaining charge carried by the droplet (48).

When investigations of the mechanism have examined large macromolecules like proteins, the results have generally supported CRM (49) (50), whilst investigations examining small inorganic and organic ions have indicated support for the IEM (51). Although the pre-eminence of a single model has been asserted for both theories (52, 53), there is now some consensus that the process is a combination of multiple mechanisms, with low molecular weight analytes following the IEM, while the CRM dominates for larger species (54, 55). A third proposition, the Chain Ejection Model (CEM), has been suggested to apply when changes in species conformation result in the exposure of nonpolar, hydrophobic moieties previously buried within a folded structure, for example in protein unfolding. While the folded protein would follow the CRM, unfolding forces the now non-polar version to migrate to the surface of the droplet, where it is ejected stepwise in a manner more similar to IEM (55).

1.1.2.5.1.2 Matrix Assisted Laser Desorption/Ionization

In Matrix Assisted Laser Desorption/Ionization (MALDI), a matrix containing the analyte mixture is bombarded with laser pulses. The energy is absorbed by the matrix causing the top layer to be ablated as a microplasma plume of matrix and analyte molecules which is sampled into the mass spectrometer. The laser pulses thus result in both vaporization and ionization of the sample. While laser desorption from a variety of surfaces is possible, the mass spectral data that can be obtained depend on the specific physiochemical proprieties of the analyte, particularly photoabsorption and volatility. The key advance for widespread adoption of this technique was to present the analyte inside a matrix which had excellent photoabsoption and proton-donation properties to increase the efficiency of energy absorption from the laser and encourage ionisation. This approach was pioneered primarily by Karas and Hillenkamp in early 1985 (56, 57).

The analyte is prepared in a mixture of water and organic solvent to encourage both hydrophobic and hydrophilic molecules to dissolve, then mixed with a suitable matrix molecule solution. For biomolecules such as peptides, this is usually 3,5-dimethoxy-4-hydroxycinnamic (sinapinic) acid (SA), α-cyano-4-hydroxycinnamic acid (CCA) or 2,5-dihydroxybenzoic (gentisic) acid (DHB). The choice of matrix determines the amount of internal energy transferred to the analyte during desorption and ionisation. SA is a ‘softer’ matrix than CCA and DHB as less energy is transferred resulting in reduced ion fragmentation during laser ablation (termed post-source decay) and is generally the matrix of choice for intact protein analysis. CCA and DHB are more commonly used for analysis of peptides where fragmentation due to post-source decay may be more easily deconvoluted to gain further sequence information. DHB is particularly used for the preparation of glycopeptides (58, 59). The analyte-matrix molecule mixture is precipitated onto the target surface so that the analytes are presented within the crystallised matrix molecules. The target is then bombarded with nanosecond laser pulses, typically in the ultraviolet frequency range. The precise mechanisms of the microplasma plume generation and analyte ionisation are not fully understood. It is known that irradiation by the laser imparts localised excitation of the matrix molecules, resulting in rapid heating and sublimation of the matrix crystals, ablation of a portion of the crystal surface and expansion of the matrix into the gas phase. This plume contains intact analyte along with protonated, deprotonated and neutral matrix molecules. Within the plume, protons are transferred to the analyte molecules, resulting in a quasi-molecular charged analyte which is sampled into the mass spectrometer (Figure 1-III).

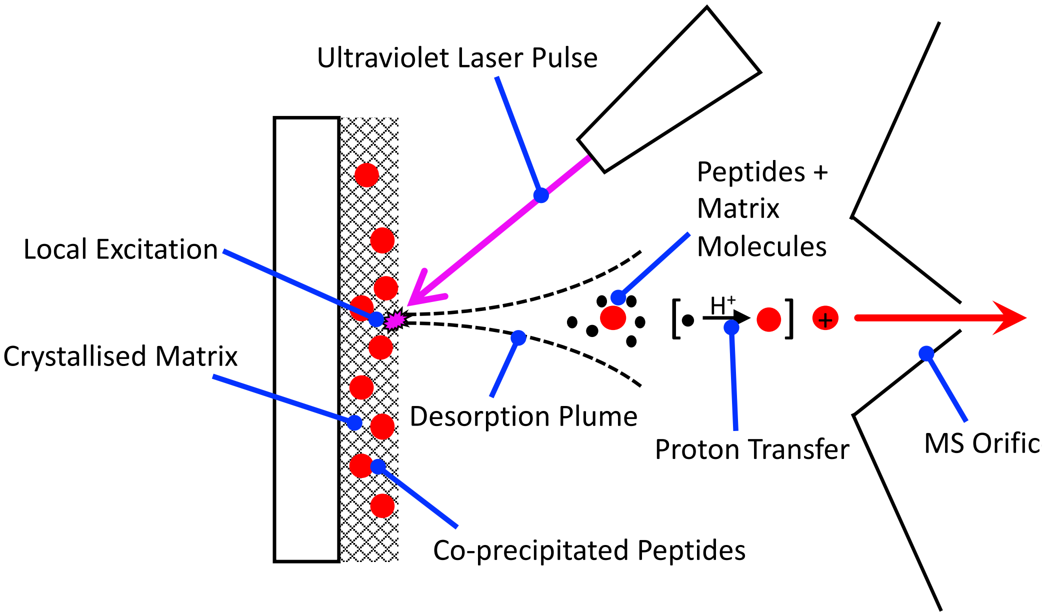
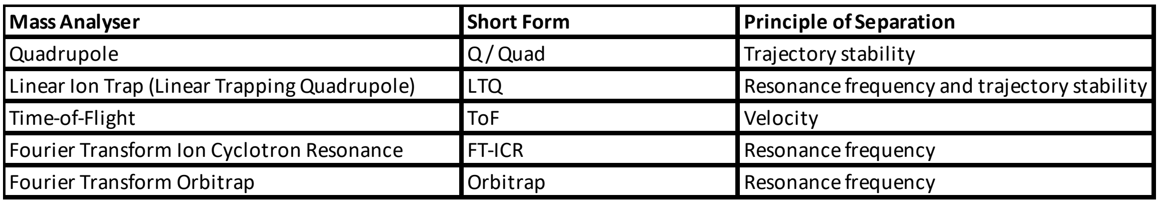


Figure 1-III. *MALDI Ionisation of Peptides from a Photoabsorbant Matrix. Adapted from de Hoffmann, E., and Stroobant, V., 2007 (60).*

MALDI is more resilient to higher concentrations of detergents and salts than electrospray with the added benefit that, after precipitation onto the matrix, the sample is preserved in the crystal structure of the matrix and can be re-probed for future analyses. The sampling of ions into the instrument is more efficient than ESI. However, MALDI is less suited for some experiments. The nature of the matrix deposition means that the analyte is not homogenously deposited throughout the matrix, so the total amount of analyte sampled may vary between probings (laser discharges). The presence of matrix ions in the sampled plume can also mask low molecular weight signals due to noise and ion suppression, and the fact that ESI generates multiply charged ions increases the total range of biomolecules that can be sampled.

1.1.2.5.2 Mass Analysers

Once gas-phase ions have been produced, it is necessary to separate them according to their masses so that the mass values (and relative abundance) can be determined. Mass analysers measure the *m/z* of ions. Several types of mass analysers have been developed based on a number of principles, although in general they involve the manipulation and separation of ions by a combination of electromagnetic fields (generated either by electric induction or by actual magnets). The primary differences between the various types of mass analyser are the methods by which these fields are used to achieve separation, each with advantages and limitations. These methods typically fall into two broad categories. ‘Scanning’ analysers separate ions of different *m/z* values successively over time by limiting transmission of ions to a restricted window which ‘scans’ over time through the *m/z* range of detection. ‘Simultaneous’ analysers allow co-transmission of ions of any *m/z*, which are then resolved according to differential behaviour (e.g. flight time or angular frequency) within the analyser. Analysers may also be grouped on the basis of other properties, for example analysis of a continuous ion beam versus a discrete ion packet, or by the typical kinetic energy of the ions during the analysis (Table 1-II).



Mass Analyser	Short Form	Principle of Separation
Quadrupole	Q / Quad	Trajectory stability
Linear Ion Trap (Linear Trapping Quadrupole)	LTQ	Resonance frequency and trajectory stability
Time-of-Flight	ToF	Velocity
Fourier Transform Ion Cyclotron Resonance	FT-ICR	Resonance frequency
Fourier Transform Orbitrap	Orbitrap	Resonance frequency

Table 1-II. *Summary of Mass Analysers Typically used in Proteomics. Adapted from de Hoffmann, E., and Stroobant, V., 2007 (60).*

Simultaneous analysers that trap the ions being detected can manipulate the trapped ions to induce a current differential which carries information about the *m/z* values of the trapped ion population in the waveform. Detection of ions transiting or ejected from scanning analysers, and in simultaneous analysers where the ions are not trapped (e.g. Time-of-Flight devices), is generally performed by Electron Multiplier Tubes/Plates (EMTs). Ions colliding with the plate induce an electric current, proportional to the number of ions striking the plate. The signal is amplified by secondary emission, in that ions striking the detector surface release further electrons towards detector surface deeper into the tube. This causes a cascade effect which propagates through the tube, resulting in a detectable electric current. By ‘scanning’ a population of ions separated by *m/z* over time into an EMT, the signal observed over time yields the intensity across the corresponding *m/z* range. It is possible to saturate EMTs with abundant ions, as they have a small recovery period after transduction of a signal, the result of which is under-sampling of the ion current and thus underreporting of the true ion intensity. Multiple small EMTs around 10 microns in diameter may be combined in an array to form multichannel plates (61) offering greater sensitivity and a higher saturation threshold.

Mass analysers may be compared on the basis of five primary characteristics: mass range, scan speed, sensitivity, mass accuracy and resolution. Mass range is the range of *m/z* values over which the MS can record ion intensity as a spectrum, and scan speed is the time taken to produce a spectrum over a given mass range (assumed to be the full range of the instrument unless otherwise stated). Sensitivity is generally given in terms of the transmission ratio between the number of ions entering the mass analyser and the number reaching the detector; a higher transmission ratio results in more ions reaching the detector which facilitates detection of lower abundance species. Mass accuracy indicates the difference that is observed between the theoretical *m/z* values of ion species and the *m/z* values measured in the mass analyser. It is usually expressed in parts per million (ppm). Finally, resolution (in mass spectrometry terms) is the ability of a mass analyser to yield distinct signals for two ions with a small *m/z* difference. The technical definition of resolution, according to the International Union of Pure and Applied Chemistry (62), is the ratio of the mass of a mass spectral peak to the resolving power (or ‘peak separation’) at that mass. The resolving power may be defined in terms of either the minimum distinguishable mass difference between two peaks, or the width of a peak at that mass. In the former case (‘valley definition’), the minimum distinguishable mass difference is the mass difference at which the signal intensity between two peaks of equal height is no more than 10% of the maximum height of either peak. In the latter case (‘peak width definition'), the peak width is taken as the width at a specified fraction of the maximum intensity, either 50%, 5% or 0.5%. In the case of the width at 50%, this value is typically referred to as the ‘Full-Width at Half Maximum’ (FWHM) (Figure 1-IV).

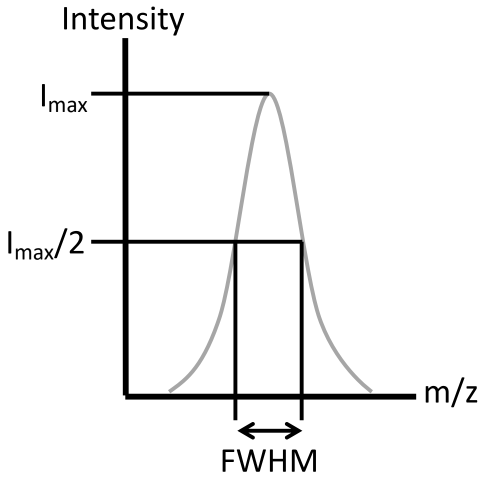


Figure 1-IV. *The Full-Width at Half Maximum Property of a Spectral Peak.*

1.1.2.5.2.1 Quadrupoles

A quadrupole mass analyser consists of four electrode rods arranged in an equidistant diamond configuration around a central channel (Figure 1-V). Each rod is electrically connected to its counterpart on the opposite side of the channel, creating two pairs at 90° to each other. The rods may be cylindrical or have a hyperbolic cross-section (compare Figure 1-V-A and Figure 1-VI-B); the latter design yields a more optimal distribution of electric field but is harder to fabricate. Ions are streamed through the device *via* the central channel. Ions whose trajectory intersects the edge of the channel will discharge on the rods or the surrounding surfaces and will not pass through the quadrupole.

An alternating current (AC) in the radiofrequency range is applied across each electrode pair, with the potentials of each pair exactly out-of-phase with the other (in this case, inverted). A constant direct current (DC) is also applied between the two electrode pairs. Because both AC and DC voltages are applied orthogonal to the channel axis, the ion velocities along the channel axis are unaffected. Ions are attracted to the electrode pair with opposite charge and repelled from the electrode pair with the same charge. The oscillation of the AC component deflects the ions alternately in the two dimensions orthogonal to the direction of travel, such that they describe a helical path through the quadrupole, with a radius which stabilises over the course of the path towards a constant value *r* (Figure 1-V-B). In a typical quadrupole analyser, the AC component may vary from 500 to 2000 V while the DC component may vary from 0 to 3000 V.

Ions are affected by electric fields in proportion to their *m/z* ratio. This property is exploited to differentially affect their path though the quadrupole, by varying the AC frequency and holding AC and DC voltages constant, or by varying AC and DC voltages (while preserving the relative ratio) fixed for a constant AC frequency. The trajectory of low *m/z* ions is affected substantially by the AC component of the field, while as *m/z* increases (relative to the AC and DC voltages), the destabilising effect of the DC component dominates.

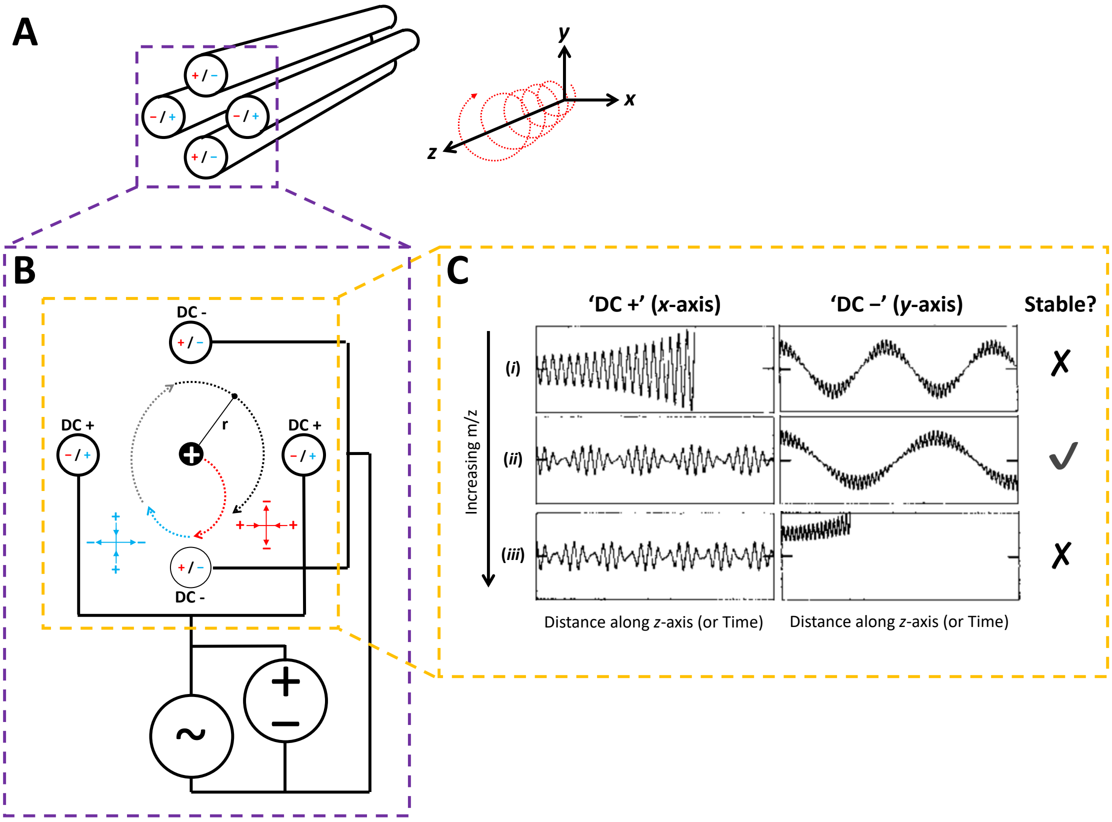


Figure 1-V. *Schematic of Ion Motion along a Quadrupole Mass Analyser.* ***A:*** *A Quadrupole consists of four electrodes, paired in left-right (x) and top-bottom (y) dimensions through which ions travel orthogonally to both electrode pairings (z).* ***B:*** *An electrical current comprising an AC component and a DC component is applied across the electrodes. The AC component inversions (shown in blue/red) cause ions to adopt a corkscrew motion.* ***C:*** *Ion trajectory through the analyser is only stable for a narrow range of m/z; the boundaries of this range change as a function of the AC and DC voltages (or the AC frequency). Ion motion traces adapted from Steel, C., and Henchman, M., 1998 (63).*

The cross-sectional radius of the helical path is dependent on ion *m/z*, with lower mass ions describing a larger radius, closer to the radius of the channel itself. Ions with very low *m/z* cannot stabilise at a path where *r* is less than the radius of the channel (the stability threshold is lower than the physical channel radius, as the strength of the field increases with proximity to the electrode). For these ions, their trajectory will move ever closer to the electrodes and eventually intersect the edge of the channel (row (*i*), Figure 1-V-C). This applies a lower bound on the *m/z* of ions able to traverse the quadrupole.

For higher *m/z* ions, the AC component has a stabilising effect of 'nudging' their trajectory into a path with *r* less than the radius of the channel (row (*ii*), Figure 1-V-C). As *m/z* increases further, this effect is reduced and the ion trajectories become dominated by the constant DC component, deviating from the central channel and eventually intersecting the edge (row (*i*), Figure 1-V-C). This applies an upper bound on the *m/z* of ions able to traverse the quadrupole.

By altering the parameters (AC/DC voltages or AC frequency) of the composite field the quadrupole may be 'tuned' to act as a filter for ions in a particular range of *m/z*. The width of this range may be varied by changing the parameters, but is also constrained by the fabrication tolerances of the quadrupole, including the symmetry of field potentials, the range of voltages and AC frequencies that can be applied, and the range of initial velocities and angular momenta of ions entering the channel.

Varying the parameters over time to 'scan' the region of stable traversal through an *m/z* range allows a beam of ions to be effectively separated by *m/z* in a time dependent manner. When coupled to a method of detecting the abundance of the filtered ion sub-population exiting the quadrupole this can be used to generate a mass spectrum.

Quadrupoles may also act as a simple 'ion guide' by setting the voltage of the DC component to zero; in this mode, all ions with *m/z* high enough to be stabilised by the AC component will be transmitted. Other 'multipole' devices with three or four pairs of electrode rods (hexapoles and octupoles) are often used in this role. The additional electric fields provide a shallower gradient of field strength across the central channel (with a much steeper gradient close to the electrodes). This allows wider mass ranges of ions to be efficiently contained (and thus transmitted with reduced loss of signal). The trade-off for this is that selected transmission of a narrow mass range of ions (i.e. filtering) is much less efficient, so these devices are generally only used as ion guides (60).

1.1.2.5.2.2 Ion Traps

An ion trap is a device that uses a combination of electric and magnetic fields to contain a population of ions. Electric field-based ion traps are historically classified into two types: the 3D ion trap or the 2D ion trap (Figure 1-VI). The first ion traps used as mass analysers were 3D quadrupole ion traps or 'Paul traps' (64), made up of a circular electrode, with two ellipsoid caps on the top and the bottom to create a 3D quadrupolar field. A conceptually simpler design which was developed later is the 2D ion trap, which may be thought of as a quadrupole mass analyser with the ends capped by lenses that reflect ions forwards and backwards within the quadrupole, such that they are contained radially by the quadrupolar field (by the mechanism described above), and axially by electric fields generated from end caps. Modern terminology generally refers to this design as the Linear Ion Trap (LIT) or Linear Trapping Quadrupole (LTQ). The linear ion trap design lends itself to a larger trapping volume than the Paul trap, which reduces undesirable interactions/collisions between trapped particles. Introduction (injection) of ions into the trap and ejection of ions from a linear trap is *via* slots in one pair of quadrupoles. The presence of the slots causes a perturbation of the RF field which reduces containment precision compared to a quadrupole of the same length. This can be somewhat mitigated by slightly stretching the quadrupole, increasing the distance between the cut rods.

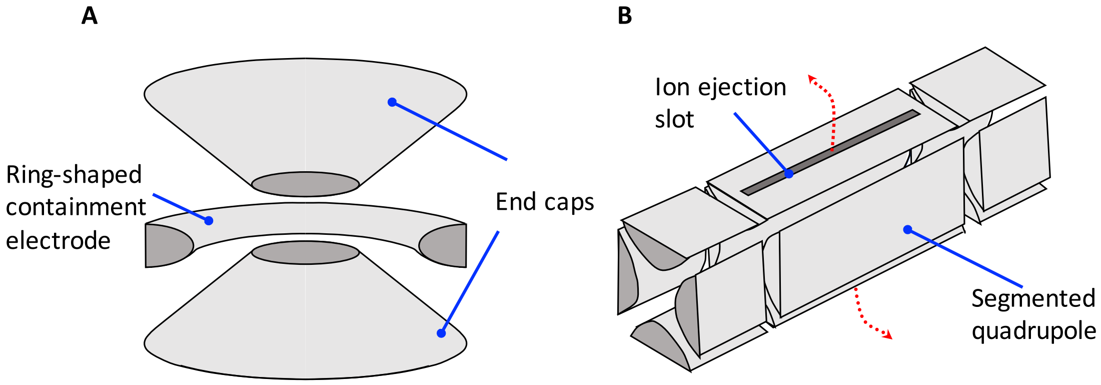


Figure 1-VI. *Ion Trap Layouts.* ***A:*** *Three-dimensional (Ring) Ion Trap.* ***B:*** *Two-dimensional (Linear) Ion Trap with hyperbolic electrodes (modern designs are typically segmented to optimise the distribution of the radiofrequency field).*

In quadrupole instruments, the potentials are adjusted so that a constant flow of ions is serially filtered to allow only ions with a selected *m/z* to pass through. In ion traps, a discrete population of ions with various masses is initially contained together within the trap. A spectrum is generated by expelling ions according to their *m/z*. In both cases the intensity recorded from the transiting (quadrupole) or expelled (ion trap) ions is correlated with the filter or expulsion settings to generate the spectrum.

1.1.2.5.2.3 Orbitraps

The Orbitrap (Figure 1-VII) is an alternative design of ion trap mass analyser proposed by Makarov (65), albeit based on a much older design for ion containment by Kingdon (66). The Orbitrap design consists of a central spindle with opposite charge to the ion population, inside a larger shell at ground potential with a split halfway along the long axis of the device (Figure 1-VII). Neither magnetic nor radiofrequency fields are applied, instead a static quadrupole field is applied in combination with a logarithmic field. Ions are first ‘cooled’ to low kinetic energy before injections, so that the spread of the distribution of individual ion energies is narrow and they can be injected into the Orbitrap as a tight packet. Once injected, ions oscillate in spirals around the central spindle. The frequency of axial oscillation is proportional to the square root of the *m/z* ratio (60) and is independent of their kinetic energy (radial oscillation and rotation are not independent). The population of oscillating ions induce a differential current between the two halves of the outer shell, which when amplified yields an ‘image current’ (composite waveform) that may be de-convoluted by Fourier transform into a spectrum of frequencies and thence scaled to yield a *m/z*-intensity spectrum. The Orbitrap offers ‘high’ performance in terms of resolution (>1 million FWHM) and mass accuracy (<2 ppm with internal calibrants) and has garnered substantial popularity in the field since its introduction.

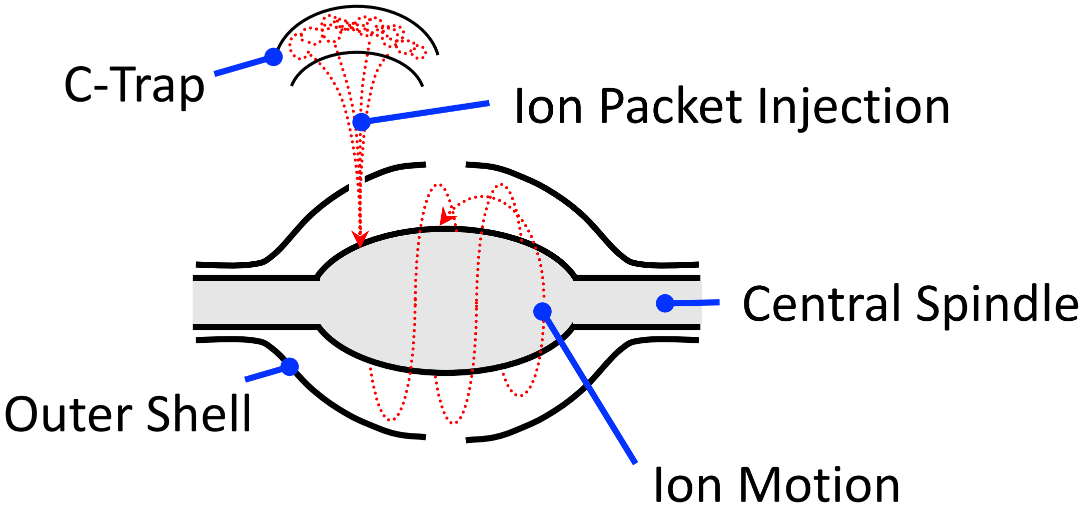


Figure 1-VII. *Orbitrap Layout. Ions are 'cooled' to low kinetic energy in the C-trap, then injected as a single packet into the Orbitrap. There, contained by the quadrupolar and logarithmic fields, ions oscillate back-and-forth along the central spindle in a spiral motion. The frequency of the axial component of this motion is proportional to the square root of their m/z; the induced current waveform generated between halves of the outer shell is a composite of the frequencies generated by ions of different m/z and may be de-convoluted to a spectrum by Fourier transform. Injection of ions as a single initial packet minimises the time over which the current waveform must be observed in order to de-convolute with precision, maximising achievable resolution.*

1.1.2.5.2.4 Ion Cyclotron Resonance Devices

The ion cyclotron (or 'Penning trap') is an ion trap where the ions are contained axially in a quadrupolar field but radially by a homogenous and static magnetic field (60, 67). Ions travel in a circular trajectory within the magnetic field. Under resonant excitation, by an electromagnetic wave of specific frequency, ions of particular *m/z* are excited, and their kinetic energy is increased, which results in an increase in velocity and thus a larger diameter of circular motion. The 'image current' that is induced by the ions circulating in the analyser wall perpendicular to the trajectory of the ions can be measured by the difference in induced current between two opposing detection plates and converted to a spectrum by Fourier transform in a similar manner to the Orbitrap.

Fourier-Transform Ion Cyclotron Resonance (FT-ICR) resolution depends on the strength of the magnetic field and the quality of the cell. For low mass ions, the maximal achievable resolution is higher than current-generation Orbitraps. However, while the resolution of FT-ICR is inversely proportional to *m/z*, the resolution of the Orbitrap is inversely proportional to the square root of *m/z*. Orbitraps are therefore able to offer high resolution across a wider mass range.

1.1.2.5.2.5 Time-of-Flight Mass Analysers

The Time-of Flight (ToF) mass analyser is essentially a long vacuum drift tube along which a packet of ions is fired (Figure 1-VIII). The packet is collected at the start of the tube and kinetic energy is imparted by an electric field of known strength. Each ion acquires the same kinetic energy and is propelled down the tube with velocity inversely proportional to the square root of their *m/z* (60). In combination with a detector at the far end of the tube to record the ion signal over time following the initial dispatch of the ion packet, the time taken for an ion to traverse a drift tube of a given length can thus be converted to *m/z*. The longer the tube, the further the ions of different *m/z* will be separated over their journey, increasing the resolution.

ToF performance can be substantially improved by the addition of an ion reflecting device (‘reflectron’) at the far end of the tube, which helps correct for starting differences in kinetic energy between ions of the same species. More energetic ions will penetrate further into the reflectron, which slightly increases their journey length and helps to normalise their flight time against less energetic particles of the same *m/z* which are reflected at a shallower depth. Additionally, by reflecting the flight path back in a V-shape to the starting end of the tube, the total journey length is doubled, which reduces the amount of physical space the ToF analyser must occupy in order to reach a certain level of resolution (additional reflectrons may also be used to further increase path length) (68).

ToF analysers offer superior resolution and mass accuracy to the quadrupole scanning technique. Ion-scanning based methods generally have a small speed advantage when generating spectra for small *m/z* ranges but are slower than ToF devices in generating a full-range spectrum. ToF-based mass spectrometers were essentially unchallenged for high-resolution work in proteomics until the advent of Orbitrap based designs which are able to offer greater resolution and mass range (and, due to the absence of the drift tube. a smaller physical footprint).

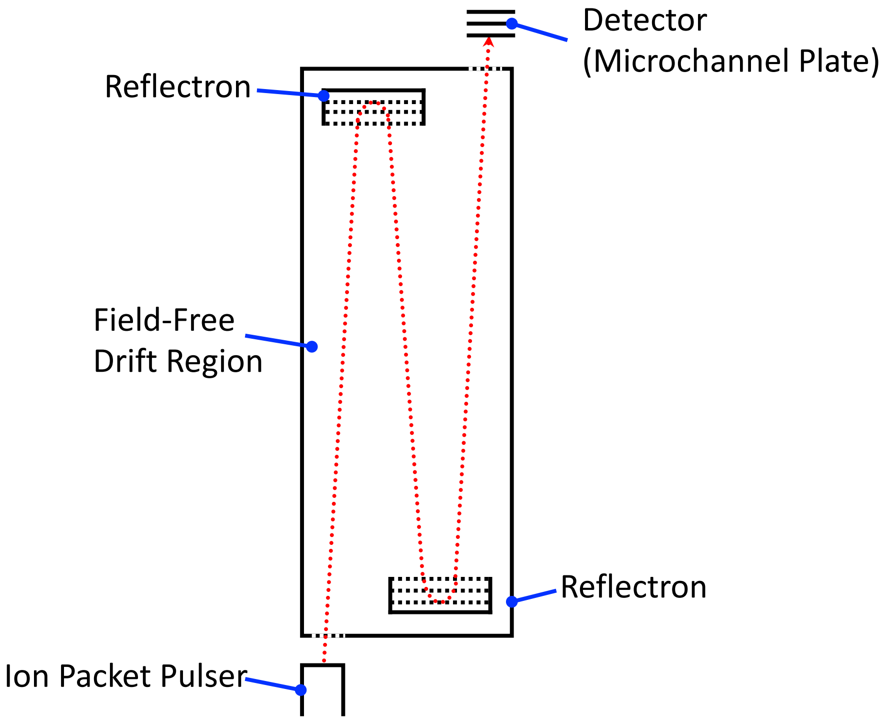


Figure 1-VIII. *Time-of-Flight Drift Tube. Ions are collected at the start of the drift tube and 'pulsed in' as a packet by an electric field. They travel down the tube with velocity inversely proportional to the square root of m/z. The ion path may be reflected back and forth several times by reflectrons to maximise path length and thus the separation of ions by m/z, improving resolution. Ions arriving at the end of the tube are detected by a microchannel plate; the change in signal over time since the ion packet was pulsed in yields the m/z spectrum.*

1.1.2.6 Fragmentation

Determining peptide mass alone is insufficient to derive complete peptide or protein sequence information. Notwithstanding knowledge of the exact mass of a species, it is impossible to differentiate residue sequence variants or other configurations resulting in molecular isomers. Various combinations of residues may result in the same net elemental contribution to the peptide total; cysteine + valine and alanine + methionine both contribute $C_8H_{14}N_2O_2S_1$ to the total elemental composition of a peptide, so two peptides which differ only in having one or other cysteine + valine / alanine + methionine pair will be indistinguishable by mass. Many other combinations of amino acids produce elemental contributions which are so close in mass as to be effectively indistinguishable within feasible reasonable mass precision tolerances. In addition, complicating factors such as mutations and post-translational modifications exponentially inflate the number of potential matches to an observed mass value. Using the LC Retention Time (RT) of the species to estimate hydrophobicity of the species, and limiting the search space by assuming a set of possible peptides (based on prior knowledge of the sample), one can reduce the number of possible peptide matches. However, even in cases when the list is much smaller than a typical biological sample (for instance, a digest of a purified protein, or a mixture of synthetic peptides) it may not always be possible to differentiate between peptides from the list on the basis of mass alone.

1.1.2.6.1 MS/MS Analysis

One method to generate such information is to deliberately induce ion fragmentation. The fragments generated will depend on the fragmented peptide sequence, its modifications and charge state (a Peptide-Modification-Charge entity; PMC) and thus the pattern of fragment ions observed may be used (in conjunction with the precursor ion *m/z*) to deduce or infer sequence information. Fragmentation in proteomics is usually performed as part of a multistage mass analysis pathway whereby fragment ion spectra (or whole or selected parts of the mass range) are collected interspersed with spectra of the non-fragmented ion population. The most ubiquitous example of such a strategy is to identify, isolate and then fragment a single ion mass from the eluting population (referred to as the ‘precursor’). In practice, ‘isolation’ of a mass means imposing a narrow *m/z* filter centred on the desired value – the shape and minimum width of this window are dependent on the mass analyser used for selection. Any other masses sufficiently close to the target mass to fall within the selection window will also be co-selected, so an assumption that the selection window isolates a single ion species will not always be correct. The isolated ions are fragmented, and a mass spectrum of the resulting fragments is collected. This process thus involves the collection of two spectra, first the non-fragmented eluting population (denoted as ‘MS^1’), followed by the fragments from a single selected ion species from that population (denoted as ‘MS^2’). This two-stage method is generally referred to as ‘MS/MS’ analysis (or ‘Tandem MS’ analysis). The archetypical bottom-up proteomic instrument setup involving a liquid chromatography system connected in-line with an MS configured for two-stage analysis is frequently summarised as ‘LC-MS/MS’.

For certain tasks, particularly the identification of post-translational modifications and in cases where complete fragmentation of the precursor ion is desirable, ions produced after fragmentation may themselves be isolated and subjected to further fragmentation (‘MS/MS/MS’ analysis, producing an MS^3 spectrum). This approach is particularly useful when a single round of fragmentation is not expected to yield sufficient information about the analyte to make a firm identification (for example, if one round of fragmentation merely results in the loss of simple neutral molecules such as water or phosphate), or if the MS^2 spectrum is too complicated (for example, if the analyte can fragment by multiple pathways). This process of selection and fragmentation may be repeated for multiple additional rounds, referred to as multistage mass spectrometry (denoted by MS^n). Interpretation of peptide ion MS^2 data in the context of proteomics is discussed further in Section 1.1.2.10, below.

1.1.2.6.2 Collision-Induced Dissociation

The most commonly employed technique for fragmentation is Collision-Induced Dissociation (CID). Analyte ions are accelerated into collisions with atoms of a neutral gas (e.g. nitrogen). Depending on the instrument configuration, ions may be accelerated by electric potential (e.g. in a quadrupole), or by resonant excitation (e.g. in an ion trap) (69). The kinetic energy of the impact is transferred to internal vibrations within the analyte, which lead to fragmentation of the peptide molecule along the backbone. In CID, this fragmentation predominantly occurs at peptide amide bonds; for each possible cleavage site, the two fragments that may be produced are sub-sequences of the original amino acid sequence from the N and C terminals to the point of cleavage. Under the nomenclature of fragment ions (Figure 1-IX) originally proposed by Roepstorff and Fohlman (70), the most frequently observed species in CID are b- and y-ions (corresponding to the fragments containing the original N-terminal and C-terminal respectively). Each cleavage event will typically produce only one charged (and therefore observable fragment).

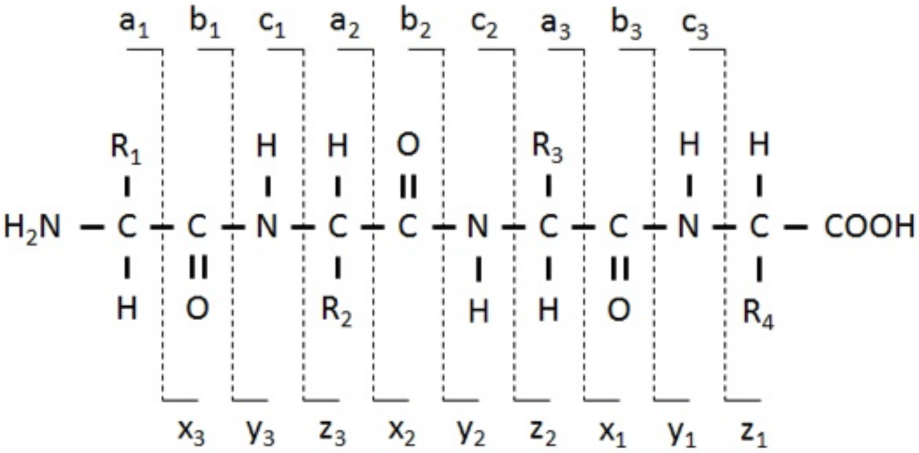


Figure 1-IX. *Notation of Peptide Fragmentation using the Roepstorff-Fohlman Scheme.*

In CID cleavage of tryptic peptides, y-ion fragments are typically observed at a higher intensity, being more stable due to the presence of the guaranteed basic residue (arginine or lysine) and are thus more easily defined in the fragment spectrum (71). However, the relative intensities of individual fragment ions are difficult to predict. The current framework of understanding is the mobile proton theory (72), which assumes that for protonated peptides formed by soft ionization methods such as ESI, the positive charges (ionising protons) are initially localized to the most basic sites, i.e. the N-terminus and the side chains of basic residues (arginine, lysine or histidine), and further that most fragmentation of protonated peptides is charge-directed, i.e. requires the involvement of a proton at the cleavage site. Given these assumptions, the theory postulates that when the peptide ion becomes energised during fragmentation (e.g. by collision in CID), the ionising protons are 'mobilised' and move from the basic sites to other locations in the peptide that would not normally be energetically favourable. Some possible relocations may provide a mechanistic route to cleavage of the (normally non-labile) backbone that yields a charged backbone fragment as the final product after decomposition. The locations where the mobilised proton is likely to move to, the possible cleavage mechanisms these enable and the likelihood of producing a charged background fragment, as well as the energy required to initially mobilise the ionising protons, are all characteristics specific to a particular PMC. The fragmentation pattern of a particular

PMC is thus the synthesis of many competing mechanistic effects (72). Furthermore, the dominance of particular mechanisms may change with fragmentation conditions (the kinetic energy of the analyte ions, neutral gas composition and pressure), as well as instrument-specific factors such as the method by which the ion population is subjected to fragmentation and the duration of the process (especially with regard to the potential for an ion to undergo multiple fragmentation events), as well as the mass range of the spectrum that is acquired from the final fragment population (73). Deterministic prediction of the relative intensities of the ion signals resulting from fragmentation of a particular sequence is thus extremely complex, but the same PMC analysed under similar conditions will generally give the same result. Within the context of the mobile proton theory it is, however, possible to form generalisations (for example, particularly abundant y-ions tend to be observed N-terminal to proline residues), and the problem is particularly well suited to the application of machine-learning prediction tools (74-77).

1.1.2.6.3 Alternative Dissociation Methods

The most popular alternative to CID involves addition of electrons to analyte to generate analyte-radical ions that fragment. The initial approach, Electron Capture Dissociation (78) required an FT-ICR instrument, but was later refined to a more generally applicable technique, Electron Transfer Dissociation (ETD) (79). ETD has some limits in peptide fragmentation, as it does not work well for 2+ charged species (80), but at higher charge states (z >2) it is very efficient, so is most commonly used for protein fragmentation or fragmentation of peptides carrying additional charges due to post-translational modifications. ETD results in a complementary fragmentation pattern to CID (c- and z-ions rather than y- and b-; Figure 1-IX); the results may also be combined for higher confidence identification.

1.1.2.7 Hybrid Instrument Designs

Most modern mass spectrometers are hybrid designs, wherein multiple mass analysers are coupled together to leverage their various advantages and thus allow parallelised analysis strategies, particularly those involving fragmentation. Ion fragmentation is usually performed in a quadrupole or linear ion trap-type analyser, while the analysis of non-fragmented analytes is optimally performed in a mass analyser with higher resolution.

1.1.2.7.1 Triple Quadrupole

The simplest design in regular use in proteomics, the ‘Triple-Quadrupole’ design (Figure 1-X) involves three multipole mass analysers, generally referred to as Q1, q2 and Q3 for purposes of describing their configuration, followed by an EMT detector. Q1 and Q3 are generally quadrupole devices whereas q2 may also be a hexapole or octupole analyser, especially when primarily used as a collision cell for fragmentation. The instrument configuration is often abbreviated as ‘QqQ’ to denote the different role of q2. The QqQ operates on a beam of ions from the source at all times. To collect an unfragmented spectrum (MS^1 mode), the mass range is scanned through in Q1, and q2 and Q3 operate as ion guides to transmit selected ions through to the detector. To collect a fragmented spectrum from a selected precursor (MS^2 mode), Q1 is used to select the precursor ion of interest, which is then fragmented in q2 and the fragment ion masses scanned through Q3 to generate the fragment ion spectrum.

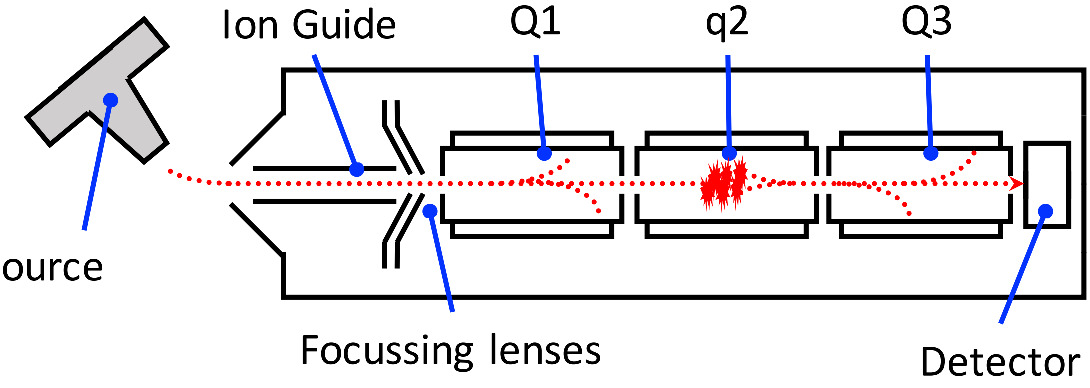


Figure 1-X. *Triple Quadrupole Schematic. In **MS^2** mode, ions are selected in Q1 (fixed filter), fragmented in q2 and scanned out to the detector (scanning filter) in Q3.*

1.1.2.7.2 Quadrupole-ToF

Before the advent of the Orbitrap, ToF mass analysers were the pre-eminent option for collecting high resolution mass spectra. Most commonly, a ToF analyser replaces what would be Q3 in a Triple Quad configuration, typically performing the role of spectrum collection in both MS^1 and MS^2 modes. The preceding quadrupoles are used for precursor selection and fragmentation in MS^2 mode. A modern example of a quadrupole-ToF hybrid system is the 'TripleTOF' line of instruments produced by ABSciex (Figure 1-XI).

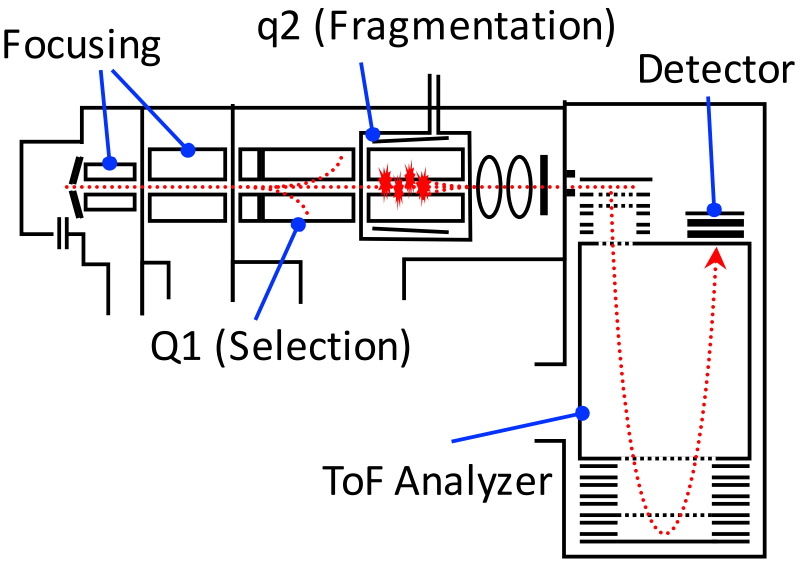


Figure 1-XI. *ABSciex TripleTOF 5600 Quadrupole-ToF Schematic. Based on ABSciex product literature.*

1.1.2.7.3 Orbitrap Hybrids

Hybrid instruments which combine quadrupole and linear ion trap mass analysers with an Orbitrap are popular in proteomics research. The coupling of an Orbitrap to a linear ion trap was the first reported configuration for a commercially available Orbitrap instrument (81). In this configuration, a linear ion trap was placed in series with the Orbitrap, connected by an ion guide quadrupole. This allowed the speed and sensitivity of the ion trap for ion selection and fragmentation (and spectrum collection for MS^2 spectra) to be combined with the resolution and mass accuracy of the Orbitrap for collection of MS^1 spectra. Various advances in Orbitrap and LTQ speed sensitivity have led to a line of successive instruments following this design with increased resolving power and sensitivity. To date, these are (in release order) the original LTQ-Orbitrap Classic and the

LTQ-Orbitrap XL (Figure 1-XII), the LTQ-Orbitrap Velos (with an improved dual-pressure ion trap and collision cell) and most recently the LTQ-Orbitrap Elite, LTQ Orbitrap Fusion and LTQ-Orbitrap Fusion Lumos, with second-generation high field Orbitrap analysers and a new signal processing method, which together approximately quadruple the resolving power.

Advances in the scan speed of Orbitraps have led to a second design in which the Orbitrap and an associated collision cell effectively replaces q2/Q3 in a triple quad setup, in which the Orbitrap is used to collect both MS^1 spectra (with Q1 transmitting all ions directly to the Orbitrap) and MS^2 spectra (with Q1 selecting a precursor, passing it to the collision cell for fragmentation which then passes to the Orbitrap for spectra collection). This instrument line consists of (in release order) the Q-Exactive (Figure 1-XIII), the Q-Exactive Plus (with an improved, segmented quadrupole), the Q-Exactive HF (with the second-generation Orbitrap and signal processing advances described above) and most recently the Q-Exactive HF-X (with improved ion optics).

The two instrument schematics shown on the next page are the instruments used to collect the datasets discussed in Chapters 3 and 4, respectively.

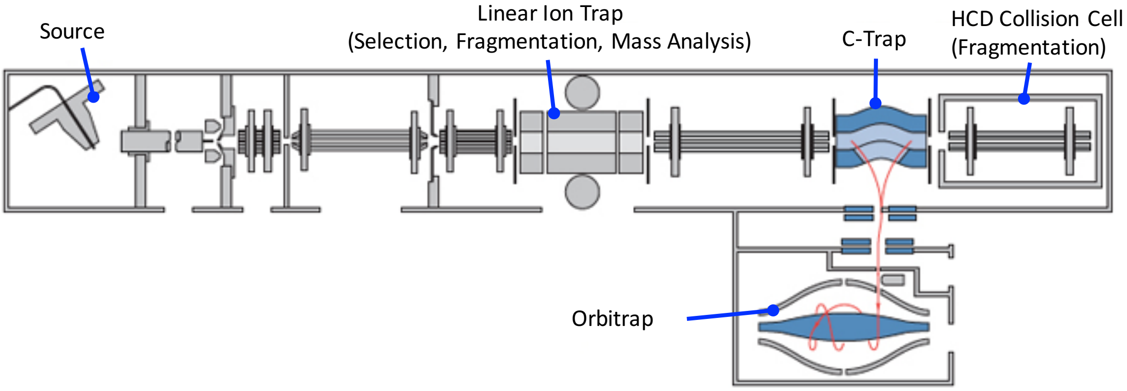


Figure 1-XII. *Thermo Fisher LTQ-Orbitrap XL Schematic. Adapted from Thermo Fisher product literature.*

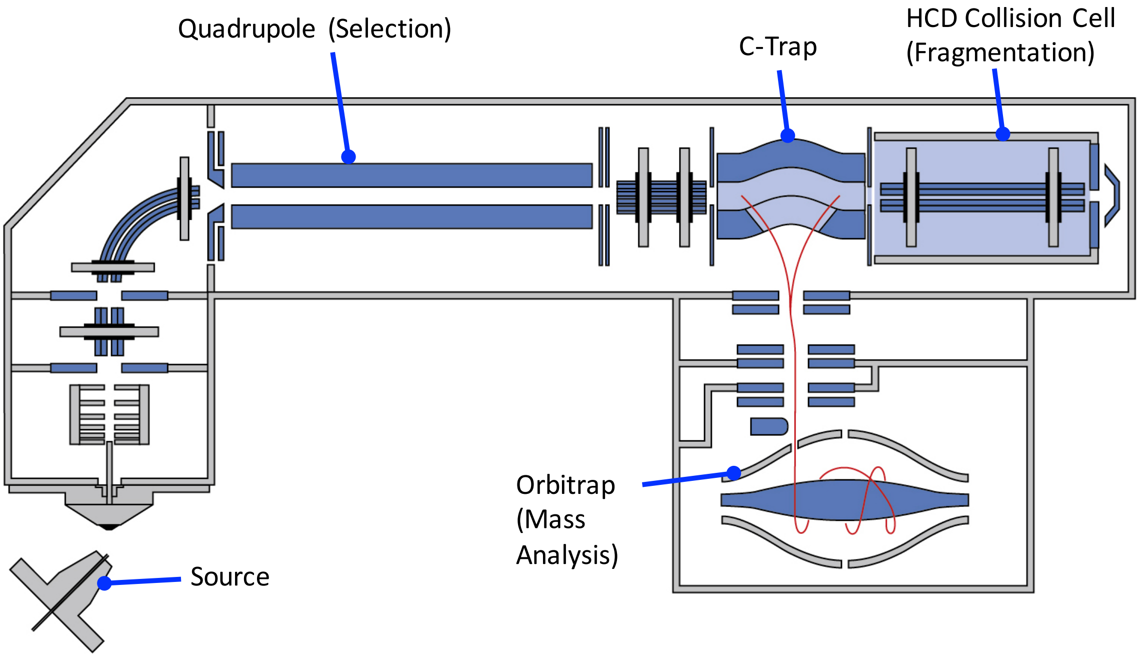


Figure 1-XIII. *Thermo Fisher Q-Exactive Schematic. Adapted from Thermo Fisher product literature.*

1.1.2.8 Data-Dependent Acquisition

The most common experimental paradigm in bottom-up proteomics is Data-Dependent Acquisition (DDA). Peptides are eluted from an on-line RP-LC system into a mass spectrometer which repeatedly samples the eluting peptides to collect full-range MS^1 spectra. In real time, the most recent MS^1 spectrum is analysed to identify potential peptide signals of sufficient intensity to warrant further analysis, excluding background noise signals and likely non-peptide contaminant species. If species of potential interest are identified, they are ranked in order of descending interest and the top candidate on the list is analysed further. Ions of this *m/z* value are selectively isolated, fragmented and an MS^2 spectrum is then collected. Additional candidates may then be selected in turn, moving down the list, until a predefined limit (typically 5-50) is reached, at which point the instrument returns to MS^1 mode, collecting a new spectrum from which new candidates may be selected for MS^2. Selected candidates are generally excluded from subsequent rounds of selection for a period of time in order to prevent repeated selection of the highest abundance species. This concept is referred to as Data-Dependent Analysis, since the *m/z* values selected for fragmentation and MS^2 spectrum collection are dependent on the ion intensities observed in MS^1 (Figure 1-XIV).

In most MS instruments, operation in MS^1 and MS^2 modes is mutually exclusive due either to configuration (e.g. in a Triple-Quad, the first quadrupole may either scan through *m/z* values to produce an MS^1 scan or select a single *m/z* for fragmentation) or control logic (firmware) limitations. Recently, a degree of parallelisation has been implemented in some instruments, such as the later models of the LTQ-Orbitrap series, which allow simultaneous use of the ion trap detector for MS^2 scans whilst high resolution MS^1 scans are acquired in the Orbitrap, with a potential for more involved schedules to minimise the idle time of both detectors (41).

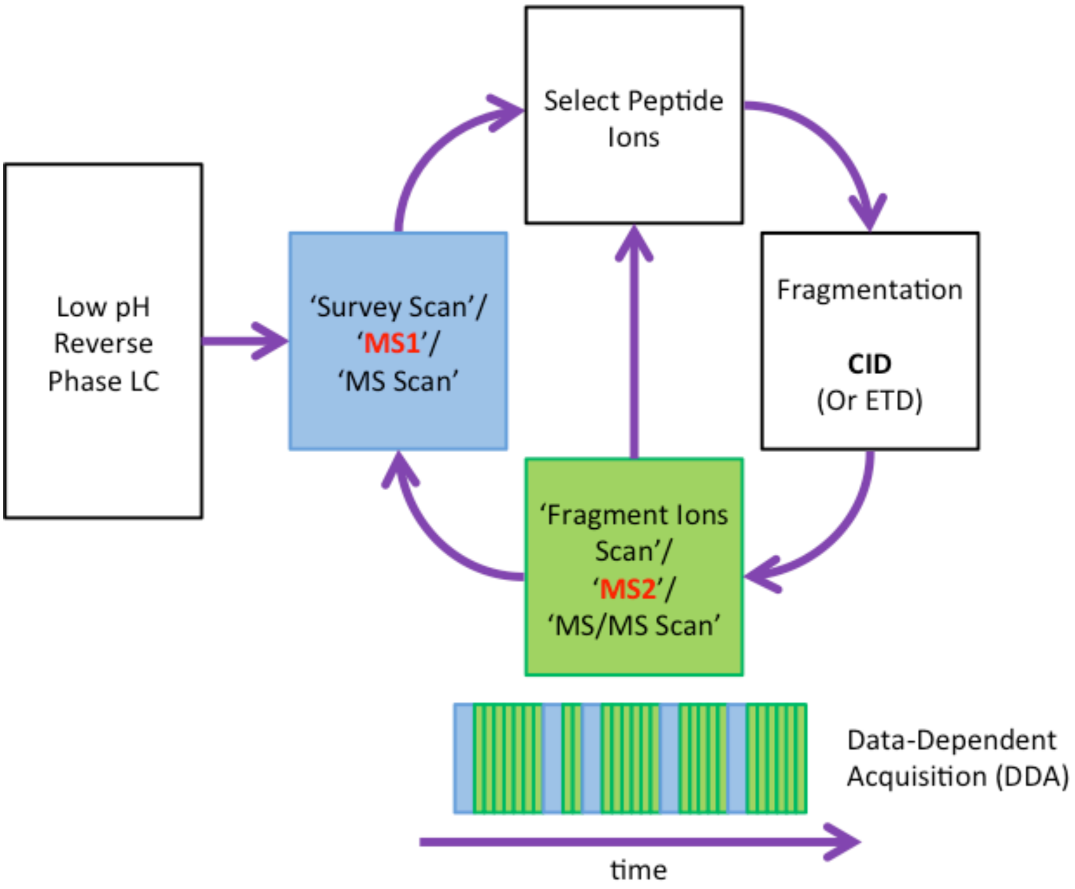

Select Peptide Ions
Low pH Reverse Phase LC
'Survey Scan'/ 'MS1'/ 'MS Scan'
Fragmentation
CID
(Or ETD)
'Fragment Ions Scan'/ 'MS2'/ 'MS/MS Scan'
Data-Dependent Acquisition (DDA)
time

Figure 1-XIV. *Data-Dependent Acquisition Process Flow.*

As depicted in Figure 1-XIV, the resulting data from a typical DDA experiment will be a series of MS^1 spectra showing the intensities of ions eluting from the LC gradient over a time course. Interspersed among these scans will be the MS^2 scans that have been acquired based on the preceding MS^1 spectra, each of which show the result of fragmentation of a selected precursor ion. The number of MS^2 events in between each MS^1 spectrum may lie anywhere between zero and the predefined limit, as the maximum selectable number of candidates that are both intense enough and not currently excluded from consideration may not be found in every MS^1 scan. The times between the collection points of sequential MS^1 scans are thus irregular, although certain examples of newer instruments such as the LTQ-Orbitrap Fusion Lumos do have the ability to enforce MS^1 spectra collection at fixed intervals. The elution profiles of *m/z* ions are thus sampled in a non-uniform manner, which has ramifications for quantification approaches using MS^1 data (as discussed in detail under Section 1.1.2.11.1). Furthermore, there is no guarantee that a particular PMC (Peptide-Modification-Charge entity; see Section 1.1.2.6.1) will be selected for MS. If, following each MS^1 scan during the elution period of a particular PMC, the PMC signal is never ranked high enough in the list of potential candidates to be selected (generally because the signal or signal-to-noise value at the time of each MS^1 is low) then no MS^2 will be collected for that particular PMC during that MS run. On mass spectrometers with a slow overall cycle time this can result in variable visibility of PMCs with very short, sharp elution peaks (even if their intensity, if measured at peak apex, would be high) if they elute between MS^1 scans. With modern instruments, this consideration is relatively minor due to the speed of acquisition although it is relevant when designing a method with a large number of allowed MS^2 events between each MS^1. More common is the problem of low abundance (but still detectable) PMCs, or higher abundance PMCs with poor chromatographic resolution, resulting in broad elution peaks where the signal is spread over a relatively long period of elution time (e.g. several minutes) where even though the integral of the signal may be high, the maximum intensity is not. In both cases, in a series of MS analyses of similar samples (or even technical repeats of the same sample), all detectable PMCs may not be found in every analysis. This stochastic sampling of the detectable PMC space, with a bias against low abundance PMCs, poses challenges for statistical analysis as inconsistent visibility leads to missing values. Various methods have been developed to allow identifications to be inferred between runs of similar samples for PMCs which are visible in MS^1 but not necessarily identified in a particular sample run. Typically, this is done by matching MS^1 retention-time dependent peak clusters across runs, either on an individual feature basis, or more often for all features. Such strategies are particularly important for Label-Free Quantification (see Section 1.1.2.11.3), a prominent example being the Match-Between-Runs feature in *MaxQuant* (82).

1.1.2.9 Data-Independent Acquisition

While DDA is the typical paradigm for bottom-up proteomic mass spectrometry, several Data-Independent Acquisition (DIA) approaches have recently been described. Rather than selecting a single *m/z* value for fragmentation, a conceptually similar non-dependent approach is to perform fragmentation on either the full mass range, termed MS^E (83), or on a series of consecutive intervals (e.g. 25 Da windows) across the full *m/z* range. This latter concept is commonly referred to as SWATH (Sequential Window Acquisition of all THeoretical spectra) (84) or SWATH-like, although such strategies pre-date this term (85). These approaches rely on computational algorithms to de-convolute the resulting data based on matching elution profiles of fragment ion spectra with ions in the MS^1 scan.

A more targeted approach that may be used if the peptides of interest are known in advance is to predefine the monitoring of both a required precursor and a particular fragment from that precursor, rather than collecting the whole MS^2 spectrum. Since certain combinations of a selected MS^1 precursor and an observed MS^2 fragment 'daughter' (known as a 'transition') ion are unique to a particular analyte, this method, known in various incarnations as Selected Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM), can allow for rapid targeted identification and quantification of pre-selected peptides (86). SRM is particularly suited to Triple-Quadrupole design mass spectrometers as it maximises a strength (the three consequent quadrupoles can be used for selection of a precursor, fragmentation, and selection of a fragment ion) while avoiding a weakness (quadrupole mass analyser-based detection is slower to collect a spectrum across an *m/z* range). In Orbitrap hybrid designs, the Orbitrap mass analyser cannot be set to detect only a single MS^2 fragment as it measures all fragments simultaneously. A complementary technique to SRM on such machines involves selecting a particular MS^1 precursor (pre-defined rather than based on a preceding spectrum), fragmenting the precursor but then analysing all fragments simultaneously. This technique is termed Parallel Reaction Monitoring (PRM) (87).

1.1.2.10 Interpretation of MS^2 Spectra in Proteomics

MS^2 spectra collected in proteomics may be the result of fragmenting a single precursor mass (with the caveat that multiple masses may be co-selected, see 1.1.2.6.1), which applies to both DDA-style experiments and DIA experiments with a fixed isolation mass lists (e.g. PRM). In the case of DIA experiments that isolate larger mass window ranges (as opposed to a single target mass) interpretation of the fragmentation spectra is considerably more complex. Discussion in this section will be limited to the former case (single precursor isolation) as being most relevant to this work.

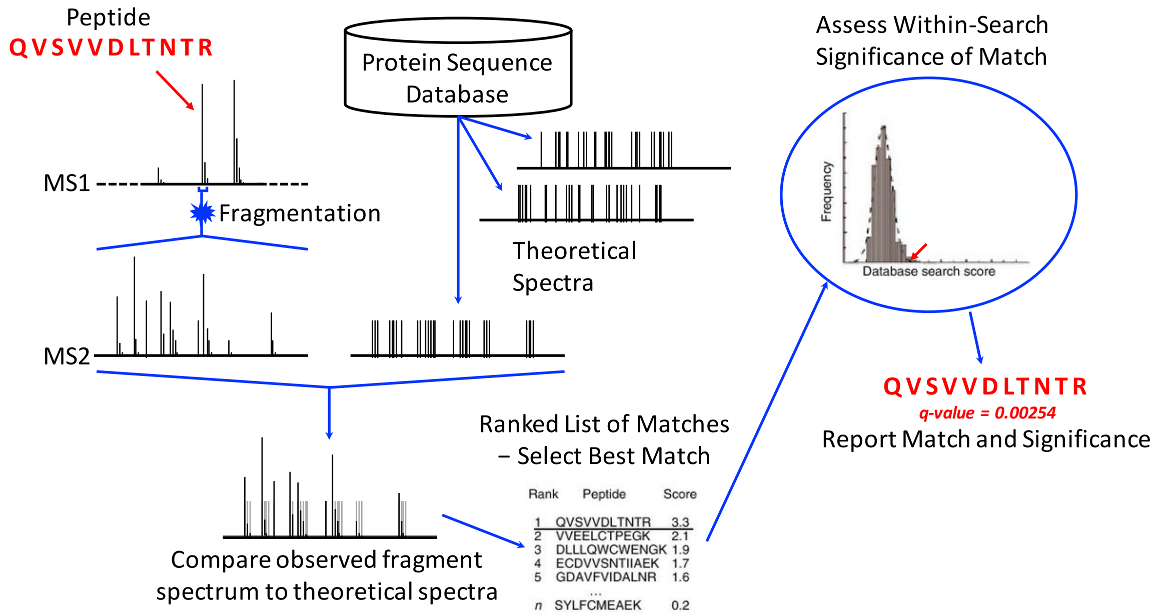
In the case of bottom-up proteomics, where the precursors are peptides, a 'perfect' MS^2 spectrum would be one in which all possible fragment ions from at least one side of the fragmentation location are observed (a complete 'ion series', see Figure 1-IX). This would be sufficient to identify the amino acid sequence of any non-post-translationally modified precursor, and to presume with high confidence the sequence of any precursor with typical post-translational modifications. The assignment of peptide sequence, modification state and charge to an MS^2 spectrum is referred to as a Peptide-Spectrum Match (PSM), although technically 'PMC-Spectrum Match' would be a more accurate definition of the acronym.

Fragment spectra may be interpreted by manual assignation of the fragment ions. In the case of peptides, one can assume that each fragment peak observed will be from the proportion of the total precursor species population undergoing fragmentation at a particular point on the backbone or side chains. The likely fragmentation locations may be predicted from the method of fragmentation. CID, for example, is expected to result primarily in cleavages at amide bonds in the amino acid chain. For most experiments, the number of spectra collected render manual assignation ion-by-ion to be impractical, and computational strategies are employed for bulk analysis of acquired spectra.

1.1.2.10.1 Automated Peptide Identification

It is possible to predict some or all of the amino acid sequence based solely on the observed data in the same manner as manual sequence assignation, commonly referred to as *'de novo'* sequencing. However, it can be challenging to distinguish the two complementary ion series (e.g. b- and y-ions) from each other, and from background noise or contaminating co-selected species due to non-specificity in the precursor selection window. Alternatively, a putative identity may be produced by comparing the observed spectrum to either theoretically generated fragmentation patterns, or a spectral library generated from previous analyses.

Theoretical fragmentation patterns are generated by applying the expected fragmentation behaviour to peptides generated by *in silico* digestion of a proteome reference (generally derived from genomic data) according to the known cleavage specificity of the protease used. In order to reduce the complexity of the problem and thus the number of comparisons to be made, a number of constraints are typically employed. The *m/z* of the precursor ion selected prior to fragmentation is used to limit the number of theoretical peptides considered to only those with a corresponding mass. Possible modifications are pre-specified as either fixed (assumed to be present on all corresponding possible sites) or variable (theoretical peptides will be considered with and without the modification on each possible site); the number of variable modifications is generally limited to avoid exponential increases in the number of theoretical peptides considered. Error tolerances for precursor and fragment ion mass are generally set to account for the expected resolution and mass accuracy of the mass analyser used for spectrum collection. Commonly used commercial and open-source search tools implementing this approach include *Mascot* (88), *SEQUEST/Crux* (89, 90), *X!Tandem* (91), *OMSSA* (92) and *Andromeda* (93). A generalised workflow for such tools is illustrated in Figure 1-XV.


Peptide
QVSVVDLTNTR
MS1
Fragmentation
MS2
Protein Sequence Database
Theoretical Spectra
Compare observed fragment spectrum to theoretical spectra
Ranked List of Matches – Select Best Match
Assess Within-Search Significance of Match
Frequency
Database search score
QVSVVDLTNTR
q-value = 0.00254
Report Match and Significance

Rank	Peptide	Score
1	QVSVVDLTNTR	3.3
2	VVEELCTPEGK	2.1
3	DLLLQWCWENGK	1.9
4	ECDVVSNTIIAEK	1.7
5	GDAVFVIDALNR	1.6
	...	
n	SYLFCMEAEK	0.2

Figure 1-XV. *Peptide Identification by **MS**^{2} Spectrum Database Searching. Adapted from Nesvizhskii* et al.*, 2007 (94).*

In silico fragmentation patterns are generally compared on the basis of expected *m/z* value only, rather than taking intensities into account. Prediction of relative fragment ion abundance is considerably more challenging than just predicting the *m/z* values of the ions generated. As an alternative to theoretical peptides, one may use a library of high confidence PSMs from previous analyses to identify new PSMs in the current experiment by similarity. This has the advantage of allowing fragment ion relative abundances to be included in the matching process, which improves selectivity and can reduce false positives. This approach is most useful in contexts where there is a large amount of existing data available. Error tolerances for precursor and fragment ion mass are used, as above, to limit the search space for each spectrum to be matched so that only plausible contenders are scored. Examples of open-source spectral library search tools include *X! Hunter* (95), *SpectraST* (96) and *BiblioSpec* (97).

In both approaches, the observed spectrum is compared to fragment ion patterns of eligible theoretical/library peptides and the correspondence in each case is scored. Scoring models differ between search algorithms, and there is a wide variety of approaches. The general aim is two-fold, to identify the closest theoretical match (and thus assign an identity) and to somehow represent how close a match this was, for comparison with other identifications. Comparative methods always seek to return the 'best' match *via* theoretical spectra, spectra library or *de novo* analysis, even in the case of MS^2 spectra for which there is no 'correct' answer (for example, the spectrum was overwhelmed by noise, or a non-peptide precursor was selected, or the peptide selected does not appear in the proteome sequence database or spectral library). Repeat analyses cannot be guaranteed to produce identical MS^2 spectra due to, among other factors, LC variability and stochastic selection of potential peptide precursors (see Data-Dependent Acquisition, Section 1.1.2.8) and variance in sample processing. It is therefore necessary to consider each PSM in the context of all PSMs produced within each analysis, to determine an appropriate score threshold in order to control the rate of incorrect PSM assignments (usually referred to, slightly misleadingly, as the False Discovery Rate (FDR) of the search).

FDR methodologies rely on assessment of the comparison score distributions among all PSMs in an analysis (Figure 1-XVI). It is possible to differentiate the distribution of the scores of genuine PSMs from those of incorrect PSMs as the 'correct' median score may be assumed to be higher than incorrect PSMs (otherwise, the comparison metric would have no selective power at all). For proteome database searching (the most widely used identification method), two approaches to FDR control have predominated (94). The first approach is to model the correct and incorrect PSMs as a mixture of two distributions (Figure 1-XVI-A). This 'mixture model' approach works best when there are a very large number of PSMs within the analysis, so that the distribution function is well characterised. The alternative 'target-decoy' approach, which is reasonably robust even at a relatively low number of PSMs, is to include an approximately equal number of 'decoy' sequences within the proteome to be searched. Such sequences should be equivalent in amino acid relative composition and peptide length to the proteome sequences; the easiest way to achieve this is simply to reverse the sequences of the proteome. Spectra producing incorrect matches (i.e. that match effectively at random) may be assumed to match in approximately equal numbers (and with equal scoring distribution) to both proteome and decoy sequences. The number and scoring distribution of PSMs matching decoy sequences is thus an estimate of half the underlying incorrect PSM distribution (Figure 1-XVI-B).

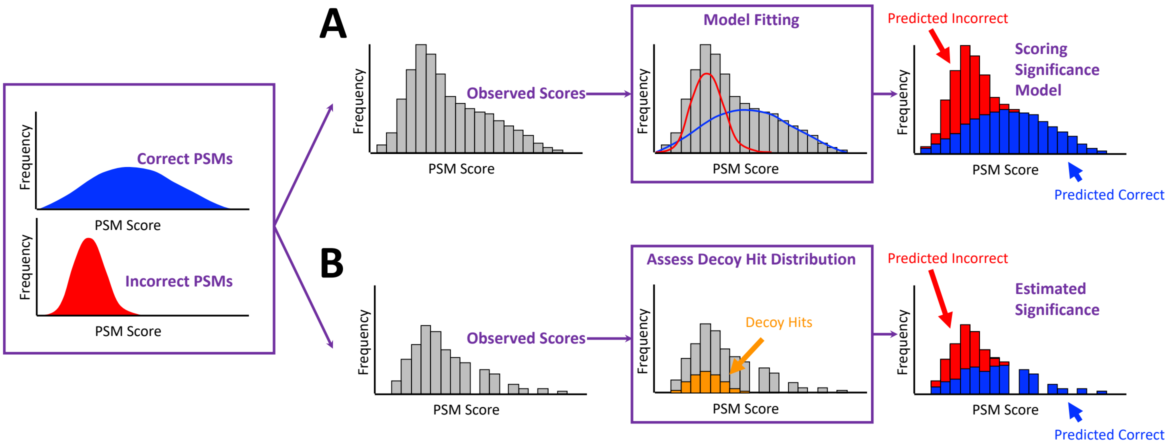


Figure 1-XVI. *Statistical Assessment of PSM Scores. The red and blue distributions represent incorrectly and correctly assigned PSMs respectively. On the left side of the figure, these indicate the true distributions in both sets of PSMs while on the right side, the predicted distributions derived by the two approaches (that may be then used to infer FDR values).* ***A****: Mixture model approach on large set of PSMs.* ***B****: Target-decoy approach on smaller set of PSMs. The orange bars indicate the distribution of PSMs assigned to a decoy database sequence.*

If the genome or proteome of the organism being studied is poorly characterised, then the proteome database or spectral libraries available may be unacceptably incomplete. Furthermore, if the analysed peptide mixture contains a large number of PMCs with novel sequence, point mutations or unusual or highly complex post-translational modifications, these will be outside the space typically considered by a database search under standard parameters. In such cases it may be necessary to fall back on purely *de novo* analysis tools such as *PepNovo* (98). More recently, so-called 'second generation' search engines such as *Peaks* (99) and *Byonic* (100) combine *de novo* and database searches. In these engines, the identification of common PMCs by database search is supplemented with the identification of mutations and unusual modifications by the more flexible *de novo* approach.

1.1.2.10.2 Protein Inference

Bottom-up proteomics produces information regarding the identities (sequence and modification state) of peptides in a sample. This information may be collated to infer the identity of the protein compositions prior to digestion, but the direct connection between peptides and proteins is broken by the digestion of all proteins simultaneously. As discussed in the comparison with top-down proteomic methodology, it is usually impossible to derive the exclusive set of parent proteins from peptide-level results, as many proteins (particularly isoforms and splice variants) share considerable sequence homology and may therefore produce the same peptide upon digestion. This issue in bottom-up proteomics has been recognised in the field since inception (22, 101); even before the widespread adoption of LC-MS it was recognised that unresolved proteoform 'spots' in 2DGE would not be differentiable by mass spectrometry of a peptide digest from the excised spot. The historical 'rule of two' solution, in which proteins with less than two uniquely assignable peptides were discounted has largely been supplanted by protein grouping solutions either built into search engines (e.g. *Mascot*, *Andromeda*) or stand-alone tools for re-analysis of identified PSM lists (e.g. *ProteinProphet* (102)). Protein grouping approaches attempt to find a minimal number of 'groups' containing one or more parent proteins which would explain the set of observed PSMs. Such approaches resolve problems where overwhelming evidence has been observed for a group of two or more proteoforms, but all relevant detected peptides map to more than one proteoform and thus none are strictly 'unique'. Under the two-peptide rule, no evidence for any of the proteoforms is admissible; it is more representative of the observed data to say that the group of proteoforms was observed but cannot be further separated.

Whilst protein grouping resolves many protein inference problems with regard to identification, basing protein quantification on peptide identifications has the additional challenge that it is unknown how much of the observed peptide intensity is potentially contributed by each potential parent protein in a protein group. Either quantification must be qualified as applying to the protein group as a whole, or it must be restricted to unique peptides, or assigned proportionately to group members based on a statistical model of the likelihood that they were actually observed.

1.1.2.11 Quantification Techniques

In bottom-up proteomics, quantification of proteins is achieved by measuring ion current derived from surrogate peptides or their derived fragments. Comparisons of abundance may be made directly between MS runs, or between species with comparable ionisation kinetics such as the same peptide differentially labelled so as to have a resolvable mass shift. If the concentration on one side of the comparison is known, the concentration and thus absolute abundance of the other side may be deduced; i.e. quantification is absolute. This is generally only the case with internal standards that have been pre-quantified, either by amino acid analysis or by quantitative nuclear magnetic resonance (103-105), and requires the identity of the peptides to be quantified by MS to be known in advance. In most cases, quantification is relative and can only be expressed as fold-changes between samples.

Labelling strategies are designed to allow separation of samples by mass in either MS^1 or MS^2 spectra, although there are examples of experimental designs in which a combination of MS^1 and MS^2 labelling strategies is used to address specific biological problems (106). This allows the simultaneous analysis of multiple samples, or of one sample and a shared standard against which samples acquired over a series of runs may be normalised (thus allowing sample quantification across multiple runs). Such an internal standard is generally constructed so as to be comparable against the most extreme samples. One way to achieve this is to generate a ‘pooled’ standard from equal aliquots of all samples.

Protein and peptide mass labels may be introduced metabolically, i.e. by providing them to the organism or cell/tissue culture over a period of time in nutrient sources, such that they are taken up and incorporated naturally into biological components. Alternatively, labels may be introduced after protein extraction by means of a chemical reaction.

Metabolic labelling allows samples to be combined and processed together as early as possible in the MS analysis workflow and facilitates analysis of metabolic processes governing the incorporation or depletion of a label over time. Metabolic labelling strategies include Stable Isotope Labelling by Amino acids in Cell culture (SILAC) and Elemental Metabolic Labelling (EML) such as ^{15}N.

Chemical labelling strategies have fewer limitations in terms of biological impediments to labelling chemistry and may be applied in cases where metabolic labelling is not possible, for example tissue biopsies. Chemical labelling strategies are, however, exposed to more technical variability.

Variation in labelling efficiency correlates with the technical variation introduced by all intermediate sample extraction and processing steps prior to labelling and combination of samples (107). Metabolic labelling allows combination after the fewest steps (Figure 1-XVII). Since there are many potential targets for chemical labelling in the polypeptide chain, a large number of chemical labelling strategies have been reported (108). Many such strategies are limited in practicality due to incomplete or nonspecific labelling, which complicates interpretations. Chemical labelling strategies that have been widely employed include enzyme-catalysed ^{16}O to ^{18}O exchange, Isotope-Coded Affinity Tagging (ICAT), Dimethyl labelling, and Isobaric Tagging (Isobaric Tags for Relative and Absolute Quantification; iTRAQ™, and Tandem Mass Tags; TMT™); these are discussed in more detail in Section 1.1.2.11.1.3.

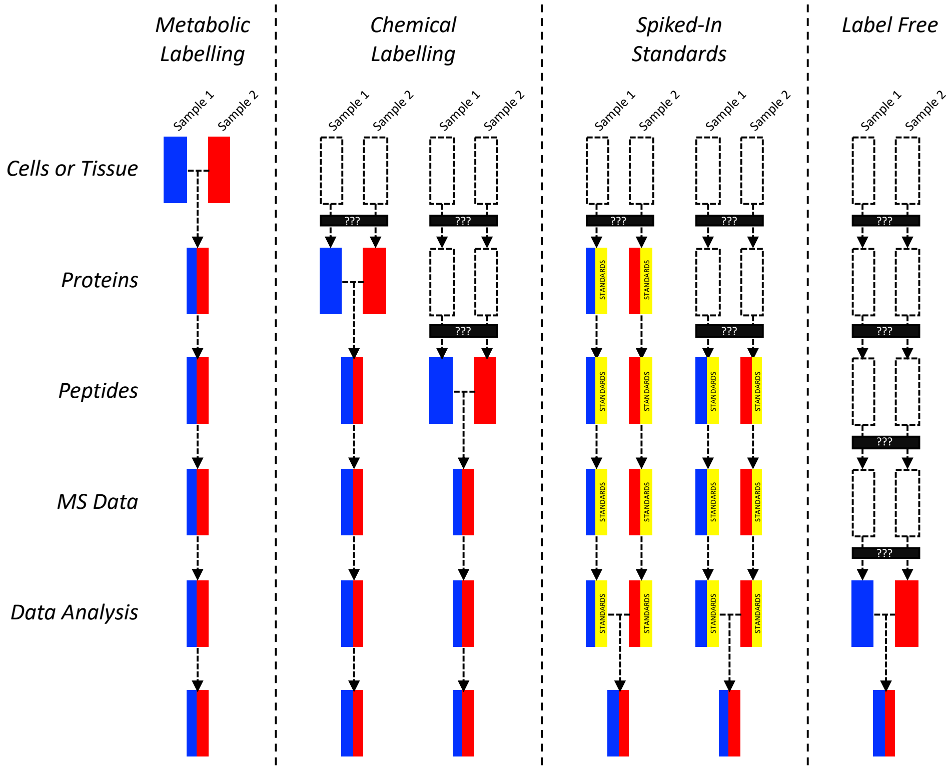


Figure 1-XVII. *Summary of Strategies for the Comparison of Proteomic Samples. Red and blue blocks indicate differentially labelled samples. Yellow blocks indicate a differentially labelled reference standard. Black bars with three question marks ('???')*

indicate steps where technical variance may be introduced due to unaccounted differences in processing or measurement. Adapted from Bantscheff et al., *2012 (109).*

1.1.2.11.1 MS^1-Based-Quantification

Separation of samples by a mass shift at the peptide level allows multiple samples to be quantified simultaneously from MS^1 spectra. Ideally the method for inducing this mass shift should not also affect the behaviour of the labelled peptides on LC gradients, so that differently labelled peptides still co-elute and are measured in the same context. A simple way to induce mass shifts without (substantially) changing LC characteristics is to use stable isotopes, such that differentially labelled peptides retain the same elemental composition but with some number of atoms having differential isotopic enrichment and thus overall different masses. Within the Born-Oppenheimer approximation, the increased mass of isotope labels will lead to a reduction in nuclear vibrational wave function amplitude and thus reduce the average volume and polarizability of bonds involving the labelled atom, potentially reducing the hydrophobicity of the molecule, depending on the intramolecular location of the label (110). The retention time effects of replacing ^{14}N with ^{15}N and ^{12}C with ^{13}C are usually small enough to disregard, but the effect of hydrogen-deuterium labelling, i.e. replacing ^{1}H with ^{2}H, is more substantial. Nevertheless, hydrogen-deuterium replacement still has applications, especially where the number of deuterium replacements is small (e.g. Dimethyl labelling - see Section 1.1.2.11.1.3).

For each peptide, quantification is achieved by comparison of the differently labelled versions of that peptide within each of the MS^1 scans across the joint elution window. Note that herein and henceforth I will use ‘signal’ to refer to the intensity corresponding to (or at least presumed to correspond to) a specific differently labelled version of a PMC in the MS. When multiple PMCs correspond to the same peptide, there are various methods for inferring a peptide-level value from PMC quantification results. For a peptide with a single PMC instance, it is generally assumed that the signal intensity is a proxy for peptide abundance, although the gradient of the linear correlation between the two is not consistent between PMCs, or between the same PMC measured in different sample contexts.

Intensity signal-to-noise can be maximised by comparing the elution peak apex intensities, or the integral of intensities over RT. Whilst there is no guarantee that the apex of the elution peak will be measured due to MS^1 sampling intervals, modern instruments sample with sufficient rapidity that uncertainty in peak shape is minimal. For label-to-label comparisons within a single MS run, peak uncertainty will in any case be equivalent between labels so long as labelling does not introduce a substantial chromatographic shift.

1.1.2.11.1.1 Stable Isotope Labelling of Amino Acids in Cell culture

Stable Isotope Labelling of Amino Acids In Cell culture (SILAC) labelling (111) is achieved by metabolically introducing isotopically labelled amino acids with a fixed mass shift to proteins. When labelling by particular amino acids is paired with use of an appropriately specific protease (e.g. trypsin with arginine/lysine labelling) to guarantee at least one labelled amino acid per peptide, the resulting labelled peptides will have fixed mass shift that is a multiple (allowing for missed cleavage) of the labelled amino acid shift. This allows for straightforward quantification as the labelled signal in the MS will always be shifted by a predictable mass from the unlabelled signal. Widely used arginine/lysine labelling labels induce mass shifts of +6 Da ($^{13}C_6$ arginine and $^{13}C_6$ lysine) or +10/+8 ($^{13}C_6$, $^{15}N_4$ arginine and $^{13}C_6$, $^{15}N_2$ lysine). The property of quantitative interest is the ratio between the unlabelled and labelled signal intensities, which is a function of both the ratio between the unlabelled and labelled samples, and the incorporation level of the SILAC label in the samples (Figure 1-XVIII-A).

SILAC has proved an extremely popular technique in the field for a number of reasons. The relatively simple nature of the quantitative data analysis and robust support for the technique by polished, relatively easy to operate quantification platforms such as *MaxQuant* (82), *Mascot Distiller* (112) (Matrix Science) and *Proteome Discoverer* (Thermo Fisher) have facilitated use of the technique without requiring heavy bioinformatic support. A high level of label incorporation is achievable. Indeed complete labelling of model organisms such as *Drosophila melanogaster* (113) and *Mus musculus* (114) has been described, and Geiger *et al* demonstrated the labelling of multiple human cell lines that could be combined to effectively approximate a generalised human tumour proteome for the purposes of an internal standard for quantification (115).

Not all organisms are amenable to SILAC-style *in vivo* labelling, however. Ideally, organisms must be auxotrophic for the labelled amino acids (116). It is also recommended that they be unable to convert the labelled amino acids into other amino acids (causing unintended secondary labelling), although there are strategies that can somewhat mitigate this second issue by experimental design (117-120) or bioinformatic post-processing (121-123). While labelling in green algae has been successful (124), the ability of higher plants to easily convert between amino acids (particularly arginine to proline) makes SILAC labelling difficult (125), although successes have been reported in plant cell culture (126) and even seedlings (127).

1.1.2.11.1.2 Elemental Metabolic Labelling

An alternative approach, rather than labelling specific amino acids, is to label all atoms of a particular element. This is frequently referred to as metabolic labelling, although I refer here specifically to Elemental Metabolic Labelling (EML) to distinguish this strategy from SILAC-style (amino-acid centric) labelling techniques, as both introduce the label by metabolism. Isotopes of ^{18}O, ^{2}H, ^{13}C and ^{15}N have all been used for such studies (128-132). The mass shift thus induced is proportional to the number of atoms of the labelled element, although the labelled signal will appear as a distribution of different masses in the mass spectrum if the incorporation level of the labelled element is less than 100%. Since EML leads to stable isotope labelling of both the proteome, metabolome and all other biological components, it has extensive history (133-137) and considerable ongoing interest (138-141) in proteomics and metabolomics as well as applicability to questions involving nucleic acids (142). There are two properties of quantitative interest, the ratio between the unlabelled and labelled signal intensities, and (when labelling is not performed to completion) the incorporation level of the labelled element observed in each labelled signal; unlike SILAC these may be measured independently (Figure 1-XVIII-B).

The automated quantification of EML data is more complicated than that of SILAC. Whereas incomplete labelling in SILAC affects only the ratio between the intensity of the labelled and unlabelled signal, in EML, the proportion of incorporation also affects both the masses of the major isotopologues of the labelled signal, and their relative proportions. Realities of experimental design with regard to controlling all metabolic intake of a particular element, and the financial limitations on maximum reagent isotopic purity mean it is usually impractical to achieve complete incorporation of an elemental label (143). Procedures for quantitative EML data analysis are discussed below in further detail.

1.1.2.11.1.3 Chemical Labelling

An early example of chemical labelling in order to introduce a mass shift in labelled peptides is the enzyme-catalysed exchange of two ^{16}O for two ^{18}O (supplied *via* $H_2{}^{18}O$) during digestion. First described in 1981 (144), this technique was used for MS-based quantification of peptides as early as 1983 (145) and applied in various ways for quantitative proteomic studies from 2000 onwards (128, 146, 147). An alternative approach by Münchbach *et al.* (148), also published in 2000, proposed labelling peptide N-termini using H_4 and $^2H_4$ (four hydrogen versus four deuterium, with a mass shift of approximately +4 Da) versions of nicotinoyloxy succinimide esters. In this example, primary amines on lysine side chains were blocked *via* succinylation to ensure only N-terminal amino groups were labelled.

The first methodology used broadly in the field was Isotope-Coded Affinity Tagging (ICAT) (149). ICAT involves labelling peptides with reagent consisting of a cysteine-directed reactive group, an isotopically coded linker (originally using ^2H but more commonly now using ^{13}C), and a biotin group for labelled species recovery. The labelled signal is separated from unlabelled signal in MS^1 according to the total isotopic mass delta of the linker component, which thus allows simultaneous analysis of multiple samples. Samples are labelled at the protein level, allowing pooling prior to digestion and thus avoiding the introduction of sample bias at subsequent steps. Whilst considered a prototypical example of chemical labelling, ICAT is limited to cysteine-containing peptides by design. In recent years it has been mostly supplanted by strategies which target primary amine groups on lysine side chains and amino termini, and thus present at least one labelling site in all peptides with unblocked N-termini. A popular example of such a technique is Dimethyl labelling, which labels the primary amines of peptides with ^2H and ^{13}C isotopically-labelled formaldehyde and cyanoborohydride (150). This approach has the advantages of low reagent costs and few experimental limitations, although analysis must be robust to potential RT shifts introduced by ^2H isotope labels and, unlike ICAT, samples must be labelled post-digestion. For all chemical labelling strategies, the property of quantitative interest is the ratio between the unlabelled and labelled signal intensities (Figure 1-XVIII-C). Incorporation level in the labelled sample may generally be assumed to be close to 100%, unless the labelling chemistry is disrupted.

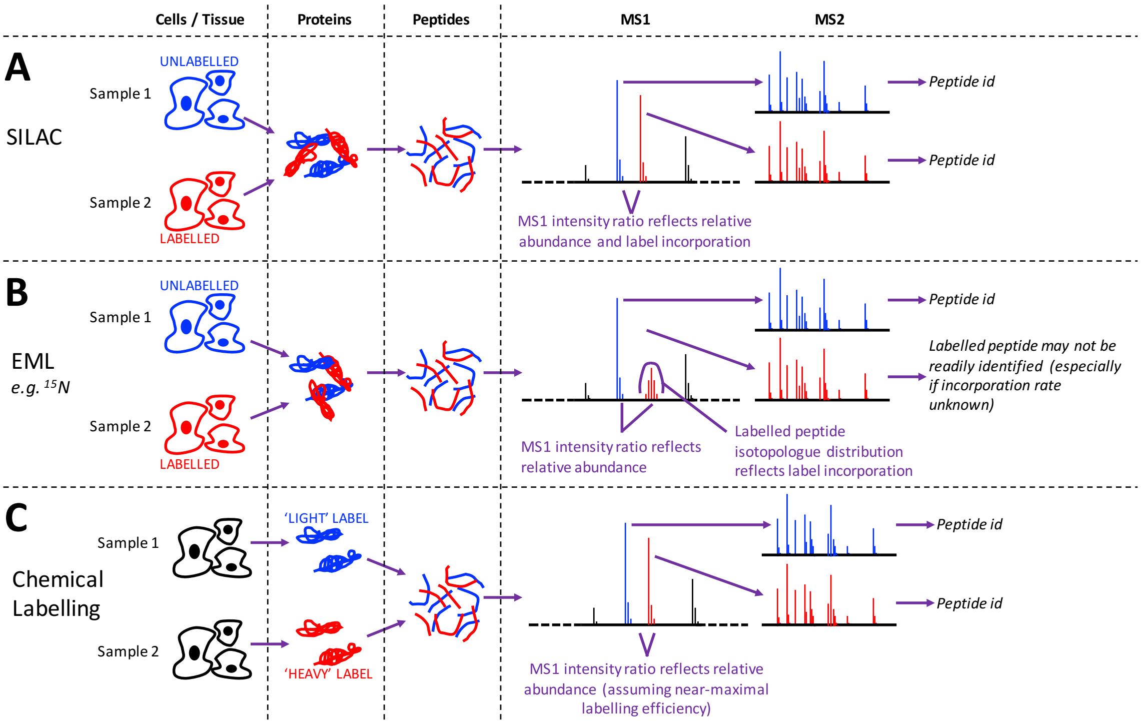

Cells / Tissue
Proteins
Peptides
MS1
MS2
A
SILAC
UNLABELLED
Sample 1
Sample 2
LABELLED
Peptide id
Peptide id
MS1 intensity ratio reflects relative abundance and label incorporation
B
EML
e.g. ¹⁵N
UNLABELLED
Sample 1
Sample 2
LABELLED
Peptide id
Labelled peptide may not be readily identified (especially if incorporation rate unknown)
MS1 intensity ratio reflects relative abundance
Labelled peptide isotopologue distribution reflects label incorporation
C
Chemical Labelling
Sample 1
Sample 2
'LIGHT' LABEL
'HEAVY' LABEL
Peptide id
Peptide id
MS1 intensity ratio reflects relative abundance (assuming near-maximal labelling efficiency)

Figure 1-XVIII. *Summary of* ***MS¹*** *Labelling Approaches.*

1.1.2.11.2 MS^2-Based-Quantification by Chemical Labelling

In MS^2-based labelling, samples have the same mass prior to fragmentation, so that they may be selected as a single precursor mass. This is achieved by labelling with a two-component reagent where the isotopic mass differences of individual ‘reporters’ (labels) are offset by corresponding balance groups. This strategy is employed by two related labelling systems - Isobaric Tags for Relative and Absolute Quantification (iTRAQ) (151) and Tandem Mass Tags (TMT) (152).

In both approaches, an isobaric tag is attached to peptide primary amines *via* N-succinimide ester chemistry. The tag (consisting of reporter and balance components) contributes an equal mass shift to the labelled signal regardless of label (this has the advantage of not increasing the complexity of the MS^1 spectrum). All labels are co-selected, and during fragmentation both the reporter and balance are cleaved from the peptide. The reporter group retains a charge and is observed in the MS^2 spectrum. Separated from their balance groups, reporter ions corresponding to each label now have different masses and thus may be quantified individually yielding the relative proportion of each label in the selected peptide precursor (Figure 1-XIX). The reporter ions are designed to have masses in a low-mass area of the spectrum with minimal potentially similar-mass ions derived from the peptide backbone fragmentation that might hinder quantification. By directly linking peptide identification with relative quantification across labels, a quantitative reading for each label is guaranteed for all identified peptides, even if the abundance of the peptide in some samples would otherwise be too low to allow consistent identification (i.e. sufficient intensity both to trigger selection of the precursor for MS^2, and for the resulting fragment ion spectrum to yield a highly scored PSM) of that peptide in those samples. An issue arises with the co-selection of peptides that are below the background detection limit, whose fragment ions remain spread across the mass range at background levels, but whose reporter ions, being all of the same mass as the reporter ions of the selected peptide, contribute to quantitative error. This has led to refinements in isobaric tagging analysis whereby the unfragmented precursor from the MS^2 scan is subject to a second round of fragmentation to yield a MS^3 spectrum with the majority of co-selected background excluded (153, 154). Isobaric tagging approaches are available in 2- to 11-plex formats, allowing for a wide range of experimental designs. However, as there is an upper limit on the number of ions that may be simultaneously selected and fragmented from a single precursor mass, high-multiplex formats risk spreading the available signal too ‘thinly’ resulting in higher variance of individual label measurements.

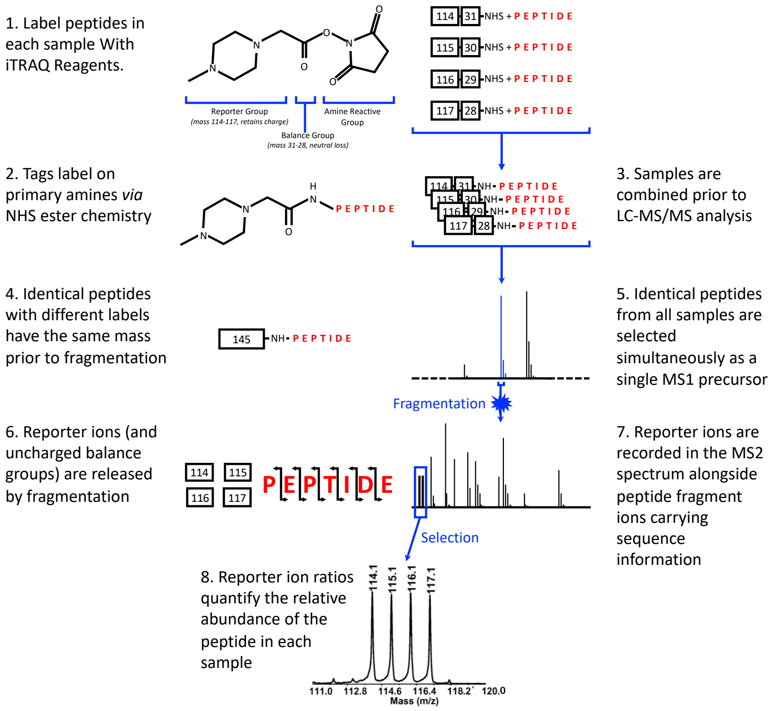

1. Label peptides in each sample With iTRAQ Reagents.
Reporter Group (mass 114-117, retains charge)
Balance Group (mass 31-28, neutral loss)
Amine Reactive Group
114 31 NHS + PEPTIDE
115 30 NHS + PEPTIDE
116 29 NHS + PEPTIDE
117 28 NHS + PEPTIDE
2. Tags label on primary amines via NHS ester chemistry
PEPTIDE
3. Samples are combined prior to LC-MS/MS analysis
114 31 NH- PEPTIDE
115 30 NH- PEPTIDE
116 29 NH- PEPTIDE
117 28 NH- PEPTIDE
4. Identical peptides with different labels have the same mass prior to fragmentation
145 NH- PEPTIDE
5. Identical peptides from all samples are selected simultaneously as a single MS1 precursor
Fragmentation
6. Reporter ions (and uncharged balance groups) are released by fragmentation
114 115 116 117
PEPTIDE
7. Reporter ions are recorded in the MS2 spectrum alongside peptide fragment ions carrying sequence information
Selection
8. Reporter ion ratios quantify the relative abundance of the peptide in each sample
114.1 115.1 116.1 117.1
111.0 112.8 114.6 116.4 118.2 120.0
Mass (m/z)

Figure 1-XIX. *Isobaric Tagging. Example of isobaric-tag labelling using the iTRAQ 4-plex labelling scheme. Adapted from Ross, P. L.* et al.*, 2004 (151).*

1.1.2.11.3 Label-Free Quantification

The primary purpose of MS^1 and MS^2-based labelling is to allow all comparisons of peak intensity to be performed within a single run, comparing between runs only *via* an internal standard included as one of the labels. Historically, direct comparison of a peptide signal between runs without a labelled internal standard was frustrated by chromatographic variation, and label-free approaches were limited to semi-quantitative approaches such as spectral counting.

Spectral counting is based on the observation that, in a typical DDA experiment, the less abundant the parent protein, the fewer peptides are expected to be detected, therefore the ratio of PSMs corresponding to a protein between two runs is an approximate indicator of relative abundance. This base metric may be refined by normalising for protein length to give the Normalised Spectral Abundance Factor (NSAF) value (155) and theoretical predictions of detectable peptide numbers to give the Exponentially Modified Protein Abundance Index (emPAI) value (156). A more recent refinement is to account for fragment intensity in the 'counted' spectra to give the Normalised Spectral Index Quantitation (SINQ) value (157). Nevertheless, these metrics are hamstrung by incomplete modelling of the relationship between protein abundance and spectral counts, and are generally only reliable for fold changes greater than an order of magnitude (158).

More recent advances in peak-picking and retention-time alignment have enabled robust direct comparison of ion signals integrated from MS^1 spectra in the same manner as MS^1 label-based quantification, generally referred to as Label-Free Quantification (LFQ). Software packages offering this functionality include *Progenesis QIP* (Nonlinear Dynamics), *Census* (121) and *MaxQuant* (82).

The great advantage of LFQ strategies is the absence of a need for any labelling, which applies no constraints on sample compatibility and reduces per-sample preparation costs. Each sample must be analysed separately by MS, however, increasing the required MS instrument time, and chromatographic variation must be minimised, which requires samples to be processed in a single batch, and a robust front-end chromatography system with minimal run-to-run variation, especially for large sample cohorts where tens (or even hundreds) of samples are to be run sequentially.

1.2 MS^1-Based-Quantification of ^{15}N EML

This thesis will concentrate on the quantitative analysis of MS^1 data from EML experiments in which the labelling is achieved by increasing the ^{15}N:^{14}N ratio in labelled samples. Labelling with ^{15}N is widely used in turnover studies (particularly in plants, where factors such as arginine-proline conversion would complicate SILAC approaches), and ^{15}N salts are comparatively inexpensive and easily sourced (141). Labelling with ^{15}N has also been used for proteomic studies in algae (159, 160) and yeast (85, 161-164), and, by employing near-completely labelled algae or yeast as food sources, higher eukaryote model organisms such as *C. elegans, D. melanogaster* (138), *M. musculus* (165) and *A. thaliana* (107, 141).

1.2.1 The Analytical Challenge

A typical DDA experiment generates what may be thought of as a three-dimensional dataset, with each signal in each MS^1 spectrum corresponding to a point in the dimensions of RT (from in-line low pH RP-HPLC), mass-to-charge ratio and intensity (Figure 1-XX).

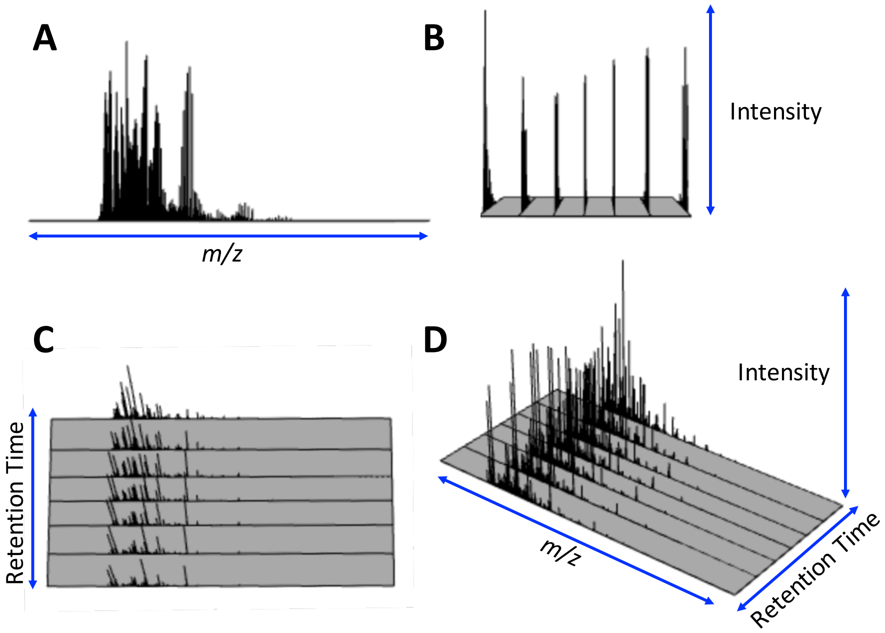


Figure 1-XX. *The Three Dimensions of Data Recorded in a DDA Experiment. Analyte signals recorded are defined in terms of* ***A:*** *m/z,* ***B:*** *Intensity and* ***C:*** *the RT of the* ***MS^1*** *scan.* ***D:*** *Together, they form a three-dimensional data space.*

Our ability to accurately characterise this dataset is limited by the capabilities of the instrument setup. The *m/z* axis is limited by the mass resolution and accuracy of the mass spectrometer, and the intensity axis by the minimum and maximum recordable signal of the detector, as well as detector-specific effects such as ion saturation and intensity-dependent (heteroscedastic) variance. Comparing run-to-run, chromatographic differences will result in RT shifts and variance in the elution peak width of the same species. This dimension is also sampled at discrete intervals rather than as a continuum, and these intervals are not guaranteed to be consistent.

Each PMC yields a separate signal in this data-space, eluting from the LC column over a particular RT window and consisting of distributions of ion species (each consisting of multiple *m/z*-intensity signals – see below) corresponding to the unlabelled and labelled forms of the PMC. The goal in EML quantification is twofold; to identify the ^{15}N incorporation level in the labelled form of the PMC, and to quantify the total signal for both unlabelled and labelled forms so that peptide- level (and protein-level) labelled to unlabelled ratios can be calculated.

1.2.1.1 Peptide Isotopologue Patterns

Due to the existence of elemental isotopes, the signal produced by a PMC (in each labelling state) is further split into several discrete peaks with increasing *m/z*. At a superficial level (fine detail being obscured by limits of instrument resolution), this appears as a series of peaks in a mass spectrum with *m/z* differences at a consistent fraction of 1 that tail off (in terms of relative signal proportion) until they are no longer observable. These peaks correspond to the different isotopologues (peptides with the same elemental composition but different isotopic composition) found in the PMC population, with the fractional difference between each peak corresponding to the charge state of the PMC (since the *x*-axis in a MS is *mass-to-charge* rather than just mass). The first peak corresponds to the isotopologue where all atoms in the peptide are the lowest mass isotope of their element, and thus has a *m/z* value of the monoisotopic molecular mass (M) divided by peptide charge (z_P). This peak may be referred to as the monoisotopologue. The next peak corresponds to the isotopologue where one atom (anywhere in the molecule) has an additional neutron, thus the total mass is greater than the monoisotopic mass by the approximate mass of one neutron, i.e. a mass increase of approximately +1 (due to binding energy mass loss, the mass differences between isotopes of different elements are not exactly 1). This peak thus has an *m/z* value of (M+1)/z_p. The next peak corresponds to the isotopologue where two atoms (anywhere in the molecule) have additional neutrons, thus a mass increase of +2 and an *m/z* value of (M+2)/z_p and so on (Figure 1-XXI).

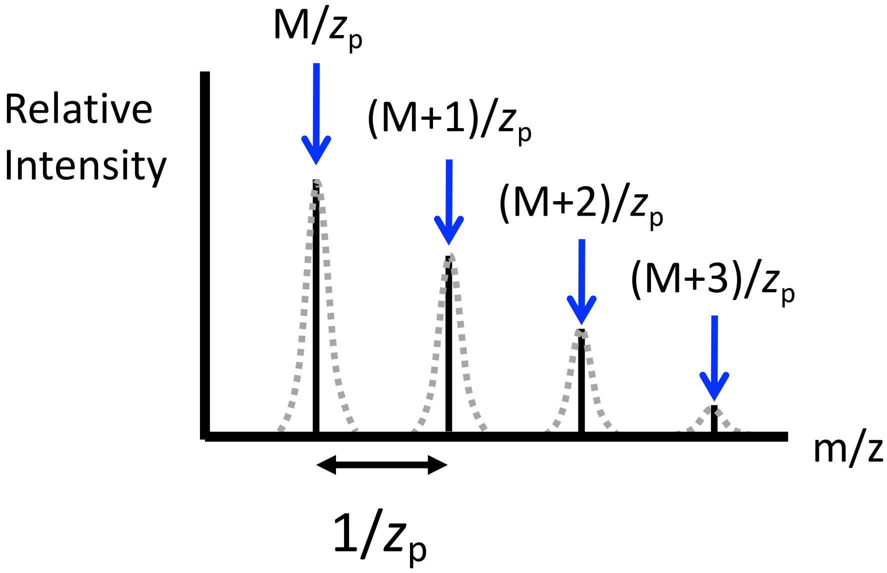


Figure 1-XXI. *Peptide Isotopologue Distribution. Peaks are observed at m/z values corresponding to the monoisotopic molecular mass (M) with zero, 1, 2… additional neutrons. The charge state (*z_p*) of the PMC may be deduced from the m/z distance between peaks; a distance of 1/2 means a charge of 2, a distance of 1/3 means a charge of 3.*

The total signal for a PMC is distributed between the isotopologues according to the underlying probability distribution that a given molecule in the population of the PMC will have 0, 1, 2 and so on extra neutrons. The natural relative isotopic abundances for the five elements (C, H, N, O, S) comprising the standard 20 amino acids (i.e. disregarding Se in selenocysteine) are all disproportionately found in the lowest mass form (see Table 1-III), so for short peptides the monoisotopologue peak is reliably the largest proportion of the total signal intensity for the PMC. As peptide length increases, the total number of atoms in the molecule increases along with the probability that at least one atom somewhere in a given molecule will not be a monoisotope. Thus, for longer peptides the second or third isotopologue peak may be the most probable scenario and will therefore be observed with the highest relative abundance. Peptides in the observed mass range in proteomics (typically 0 to 2 kDa), are generally not long enough that more than 6-7 isotopologues of the unlabelled signal are observed even if the peptide is very abundant.

Element	Isotope Mass	Relative Abundance (%)
Carbon (C)	12.0000000	98.930
	13.0033554	1.070
Hydrogen (H)	1.0078246	99.985
	2.0141021	0.015
Nitrogen (N)	14.0030732	99.632
	15.0001088	0.368
Oxygen (O)	15.9949141	99.757
	16.9991322	0.038
	17.9991616	0.205
Sulphur (S)	31.9720700	95.020
	32.9720700	0.750
	33.9678660	4.210
	35.9670800	0.020

Table 1-III. *Elemental Isotopes in Standard Proteinogenic Amino Acids. The precise values listed are those used in Fan* et al., *2016 (166).*

Since the mass shifts engendered by each isotope are not identical, then (for example) the *m/z* of a PMC with a single ^{15}N or ^{2}H isotope is not precisely the same as the *m/z* of a PMC with a single ^{13}C isotope. The isotopologue peaks are not a single signal from ions of identical *m/z* but rather a composite signal of all ions with *m/z* values produced by combinations of isotopes that yield a certain net number of additional neutrons. The instrument resolution required to reliably separate these signals is considerably higher than typical operating parameters for proteomic analysis, and so for practical purposes these may be considered as a single peak. In an unlabelled signal, the probability distribution is overwhelmingly dominated by the effect of carbon due to a proportionately high percentage of the atomic composition and large relative abundance of the ^{13}C isotope.

It is relevant to consider, however, that the observed centroid *m/z* values of isotopologue peaks (particular those corresponding to a net neutron increase of more than 2) are comprised of contributions from many isotopic composition permutations. As well as engendering a difference in intensity, the centroid *m/z* of the +5 neutron isotopologue peak of an unlabelled signal will not be exactly the same as the centroid *m/z* of the +5 neutron peak of a ^{15}N labelled signal. The *m/z* values of the components of the peak do not change, but the relative probability of the component corresponding to five ^{15}N will be increased against the previously dominant component corresponding to five ^{13}C, and so a weighted centroid of all component *m/z* values will be different.

1.2.1.1.1 Prediction of Peptide Isotopologue Patterns

In SILAC and (ICAT/Dimethyl labelling), the isotopologues of both the unlabelled and labelled signals occur at predictable *m/z* and with effectively equivalent distributions. The addition of a few atoms in the SILAC label is not enough to substantially change the expected isotopologue proportions. The labelled distribution is always shifted in mass by a fixed amount multiplied by the number of labelled amino acids. It is not necessary to predict the expected isotopologue distribution of either unlabelled or labelled signal (although this can be used for noise assessment and quality control) since labelled signal intensity may be calculated by whatever method was used for the unlabelled signal intensity, with an appropriate mass-shift to account for the label. Furthermore, a label incorporation level of less than 100% merely alters the expected ratio of unlabelled to labelled signal intensity, as a single molecule may either have a label (and thus contribute to the labelled signal intensity) or not (and thus contribute to the unlabelled signal intensity).

In EML only the isotopologue distribution of the unlabelled signal is predictable without knowledge of the label incorporation level. Labelling effectively increases the relative abundance of a particular non-monoisotope relative to the monoisotope (so, for example, labelling with ^{15}N increases the ^{15}N isotope abundance versus the ^{14}N isotope abundance). This has the effect of increasing the mass of the labelled peptide and, for incorporation levels less than 100%, also changing the isotopologue distribution of the corresponding signal, as a single peptide molecule can have a variable number of heavier isotope atoms. The incorporation level (and thus relative abundance) of the labelled peptide must be determined by identifying the incorporation level at which optimum fitting of the observed spectrum of the labelled signal is achieved (Figure 1-XXII).

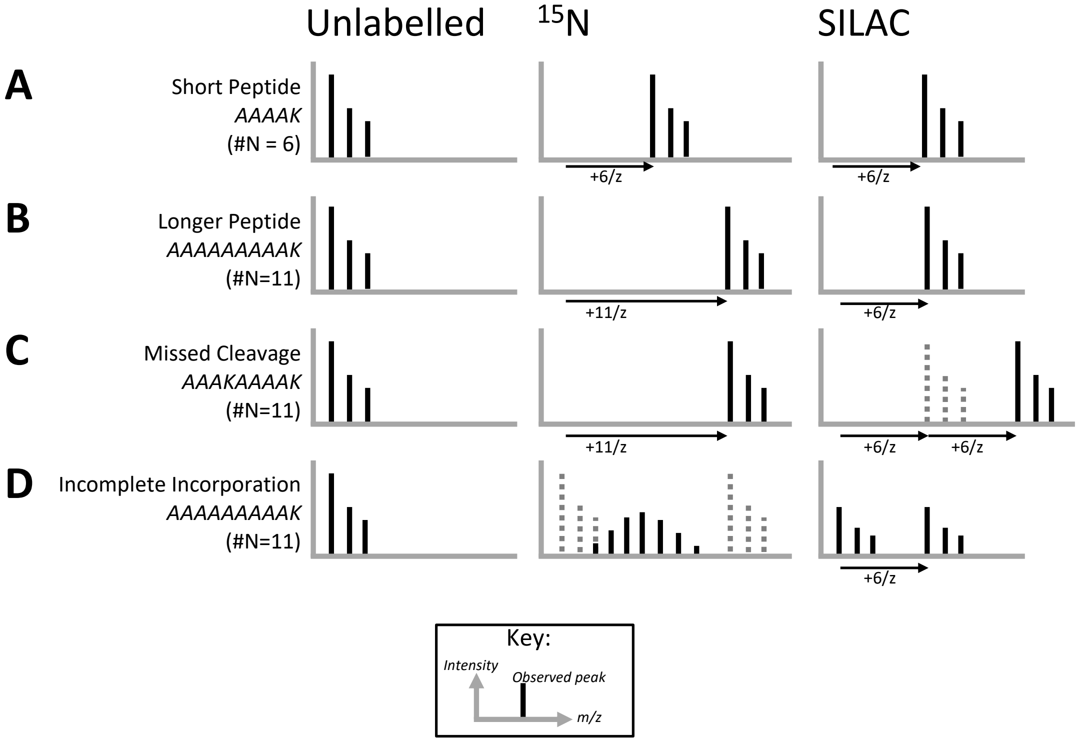
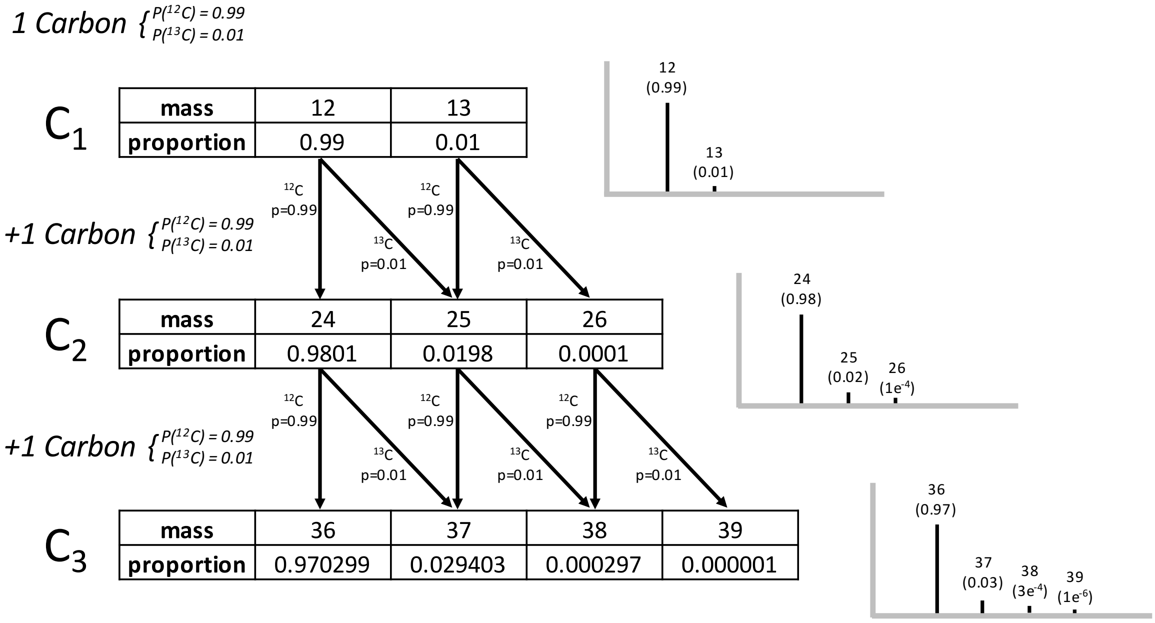


Figure 1-XXII. *Effect of Labelling on Peptide Isotopologues in ^{15}N EML* vs. *SILAC.* ***A & B:*** *The mass shift induced by SILAC is proportional to the number of labelled amino acids whereas the shift induced by ^{15}N EML is dependent on the number of labelled nitrogens.* ***C:*** *If labelling occurs prior to digestion (as in SILAC but not in ICAT/Dimethyl labelling) then any missed cleavages will mean the resultant peptide will have multiple SILAC-labelled amino acids and thus the mass will be increased by a multiple of the base shift – here, the dotted lines for SILAC indicate where the distribution would be located without this shift.* ***D:*** *The labelling methods are particularly distinguished in cases of incomplete labelling – here, the dotted lines for ^{15}N indicate where the unlabelled (left) and fully labelled (right) distributions would be located. In ^{15}N EML the distribution shape of the whole sample is altered, whereas in SILAC the signal is split into fully labelled and unlabelled populations. Unlabelled peptides from a SILAC labelled sample with incomplete incorporation produce a signal indistinguishable from peptides that have not been labelled at all.*

Isotopologue distributions may be predicted *in silico* from the chemical composition of the peptide. Distribution calculation algorithms described by Kubinyi (167) and Rockwood *et al.* (168-171) have been widely used for this purpose. Development of these algorithms has largely been with an eye towards the more complicated case of intact protein masses and these algorithms are optimised for speed. Distribution calculations for peptides are thus extremely fast on modern computers, allowing rapid computation of many possibilities. Peptide distributions may be found step-wise by repeated convolution of the isotopic probability distribution of each element (C, H, N, O, S) with itself to find the isotopic probability distributions for a molecule consisting only of a number of atoms of that element equal to that in the peptide composition (Figure 1-XXIII), then convoluting those elemental molecule probability distributions together to find the distribution for the complete molecule.


1 Carbon {$P(^{12}C) = 0.99$, $P(^{13}C) = 0.01$
C_1
^{12}C p=0.99
^{13}C p=0.01
+1 Carbon {$P(^{12}C) = 0.99$, $P(^{13}C) = 0.01$
C_2
+1 Carbon {$P(^{12}C) = 0.99$, $P(^{13}C) = 0.01$
C_3
12 (0.99)
13 (0.01)
24 (0.98)
25 (0.02)
26 (1e-4)
36 (0.97)
37 (0.03)
38 (3e-4)
39 (1e-6)

mass	12	13
proportion	0.99	0.01

mass	24	25	26
proportion	0.9801	0.0198	0.0001

mass	36	37	38	39
proportion	0.970299	0.029403	0.000297	0.000001

Figure 1-XXIII. *Stepwise Prediction of an Isotopologue Distribution. For simplicity, the molecule is assumed to consist only of three carbon atoms.*

The approach described by Kubinyi is an efficient shortcut to this process. Rather than convoluting each 'running total' elemental distribution with the atomic distribution *n-1* times (where *n* is the number of atoms of that element), this approach represents *n* as a binary number. The atomic distribution is convoluted with itself, then the resulting distribution with itself and so on, until it represents the distribution corresponding to the highest power of two in the addends of the binary representation of *n*. Finally, each distribution corresponding to an addend of the binary representation of *n* together is convoluted.

For example, to find the distribution for C50 the stepwise approach is to convolute C1 and C1 giving C2, then convolute C2 with C1 giving C3 *etc*. This involves 49 convolution operations to get to C50. Alternatively, by the Kubinyi method, the binary addends of 50 are 32, 16 and 2; once these three stages are calculated, one can then can convolute C32 with C16 (giving C48) then the result with C2 to get C50. This optimisation already reduces the number of operations required to 32 (to get to C32) + 2 = 34. But optimisation can be taken further since the binary addends, as powers of two, can be reached quickly by convoluting C1 with itself to get C2, then C2 with itself to get C4 and so on. The largest addend required is C32 i.e. 2^5, which requires only 5 operations in total counting the initial convolution of C1 with C1 to get C2, producing C16 along the way. Calculation of C50 thus involves a total of only 5 (to get C2, C16 and C32) + 2 (to combine the addends at the end) = 7 convolution operations, many fewer than the step-wise method (Figure 1-XXIV).

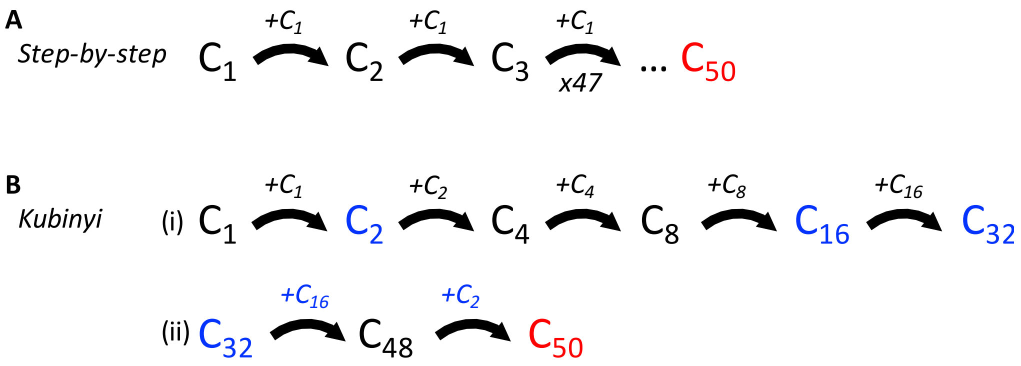


Figure 1-XXIV. *Comparison of Isotopologue Distribution Prediction Algorithms.* ***A:*** *The step-by-step method for prediction of an n multi-atom isotopologue distribution (in this case, C50) involves iterative convolution of a single atom with the result of the previous operation n times.* ***B:*** *The Kubinyi method reduces the number of convolutions required by calculating only the binary addends of n. (i) These may be found quickly by convoluting of each result with itself (thus finding the distributions representing powers-of-two). (ii) Once the highest binary addend required has been found, the distribution representing n can be calculated with a minimal number of further operations.*

SILAC/ICAT/Dimethyl labelled distributions may be found by applying the mass shift of the label to the isotopologues of the unlabelled distribution. For EML labelling, due to the change in isotope proportions, the isotopologue distribution should ideally be fully re-calculated. An alternative ‘short-cut’ method for EML distributions used by some existing approaches is to derive the expected masses of the labelled isotopologues by adding incremental multiples of the heavier isotope mass delta (e.g. for ^{15}N-^{14}N this is ~0.9970348932) to the molecular weight of the monoisotopic isotopologue. The advantage of this approach is that it saves computation time by avoiding re-generation of the distribution from scratch for each increment of incorporation in the label incorporation range to be tested. This method, however, is vulnerable to cumulative error as it does not account for the changing relative contribution to each isotopologue centroid *m/z* by the EML isotope (see above). For example, in ^{15}N EML, the proportionally weighted centroid mass of each isotopologue decreases as the contribution from ^{15}N rises, because the shift due to ^{15}N (~0.997 Da heavier than ^{14}N) is less than that due to ^{13}C (~1.003 Da heavier than ^{12}C). At high incorporation, therefore, the ‘short-cut’ method is relatively accurate, but at low incorporation, when the isotopologue centroid is dominated by the ^{13}C mass difference, the effect is to substantially underestimate the centroid *m/z*. While this inaccuracy appears to have little effect when viewed at a scale across the full mass range of the possible distributions (Figure 1-XXV-A), applying error windows appropriate for modern instruments to individual isotopologues is more revealing. The short-cut method results in substantial accumulated errors compared to the weighted isotopologue centroids from fully calculated theoretical distributions. Taking the peptide VVISAPSK as an example, this effect results in *m/z* errors of > 10 ppm (Figure 1-XXV-B). Mass accuracy in an MS^1 scan on an Orbitrap instrument (~60-120k resolution) is typically below 5 ppm, so errors of more than 10 ppm are quite substantial.

Since the labelled isotopologue distributions must be predicted, this also means any quantification is reliant on first identifying the PMCs present (in order for the prediction to be accurate). Identification of EML-labelled peptides presents additional difficulties as current peptide identification search engines generally do not adequately handle EML peptides of unknown incorporation (172). By allowing a range of incorporations, pre-filtering of the search space by precursor masses would become much less selective, substantially increasing the number of potentially matching possible sequences and thus greatly increasing the complexity of the search. In contrast, SILAC/ICAT/Dimethyl labelling, by introducing a known mass shift, can be handled as a routine post-translational modification. The identification problem is commonly addressed by using reference analyses of fully unlabelled but compositionally identical samples run under the same chromatographic conditions, so that any peptides identified can be matched against the mixed-labelling dataset by RT.

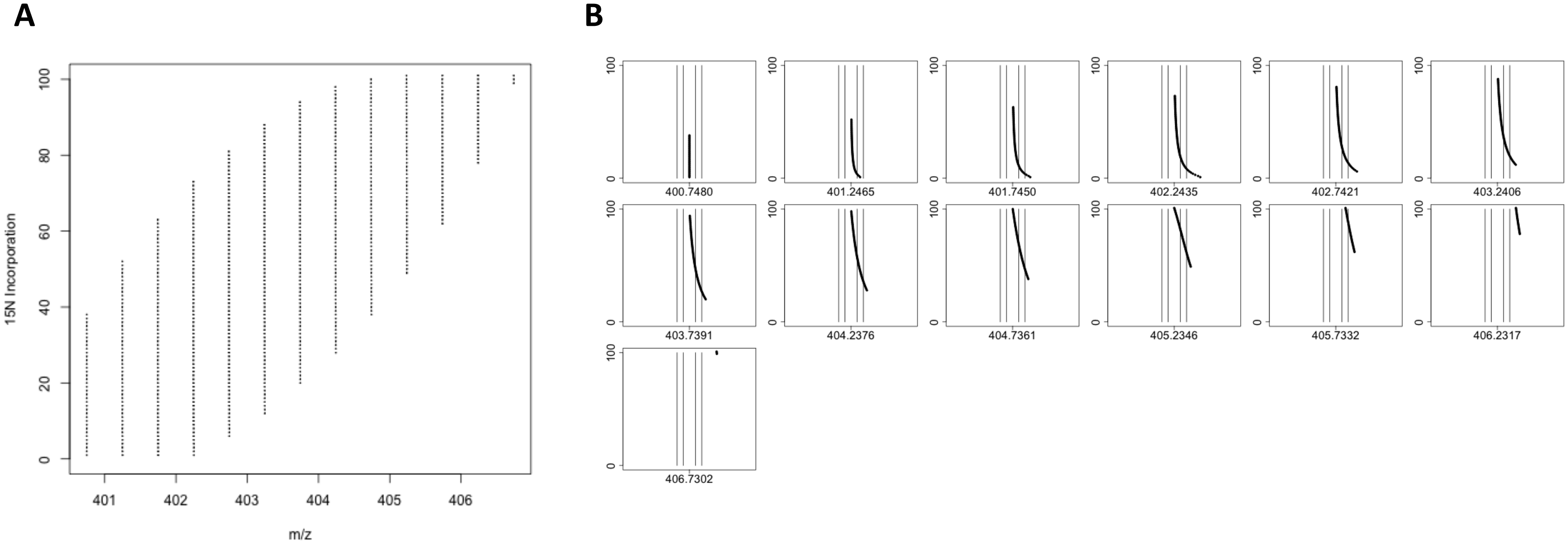


Figure 1-XXV. *Mass Errors Associated with the 'Short-Cut' Method. The 'short-cut method' estimates the labelled distribution by applying fixed mass shifts to the unlabelled distribution.* ***A****: At the isotopologue-window scale, the isotopologues of example peptide VVISAPSK appear to be calculated correctly.* ***B****: Closer examination reveals that at low incorporation, the higher-neutron number isotopologue centroid m/z values are underestimated by more than 10 ppm. Each box shows a 'zoomed-in' snapshot of an isotopologue mass from A on the same axes (showing m/z in a limited window, y-axis showing increasing* ^{15}*N incorporation). Boxes left-to-right top-to-bottom correspond to isotopologues in increasing mass order. Vertical thresholds indicate 5 and 10 ppm error windows either side of the centroid m/z.*

1.2.2 Analysis Methodologies

The *MSQuant* framework first referenced in publications in 2003 (173, 174) was capable of quantifying ^{15}N labelled signal given a fixed, pre-specified level of ^{15}N incorporation. However, the framework was geared towards SILAC labelling and ^{15}N support was somewhat awkward, requiring additional scripts to pre-patch search result input before quantification. In 2006, Andreev *et al.* (175) described a dedicated ^{15}N quantification algorithm, again assuming a fixed level of incorporation. Although the algorithm theoretically supports any level of ^{15}N incorporation, the authors only demonstrated performance in the case of complete labelling, where the analytical challenge is reduced since there is no need to account for differences in isotopologue distribution shape. Subsequently, Palmblad *et al.* (132) described a similar analytical pipeline for quantification of fully incorporated ^{15}N labelling using Bruker instrument software (*DataAnalysis* and *CompassXport*) in combination with peak extraction tools in the *Trans-Proteomic Pipeline* (176). Both methods demonstrated good performance on high incorporation biological samples, in the process indicating a degree of robustness to small deviations from 'idealised' complete labelling, which in reality is generally impractical to achieve (159). Another similar algorithm (*Peakardt*) was robustly validated by orthogonal quantification *via* 2DGE using the Difference-In-Gel Electrophoresis approach with fluorescent CyDyes (177).

Two later software tools support ^{15}N quantification (with pre-specified incorporation levels) as part of a broader MS^1 quantification offering. *Census* (121, 178) is a free tool with ongoing support (179) in which peak extraction tolerances are highly customisable, allowing optimisation for both low and high resolution data. *Mascot Distiller* (Matrix Science) is a commercial quantification solution which integrates closely with the *Mascot Server* search engine from the same vendor. A complete pipeline for automated analysis using *Mascot Distiller* (Matrix Science) was described by Bindschedler *et al.* (112). Further optimisations to a *Mascot Distiller*-based approach *via* optimisation of chromatographic alignment and cross-run matching (by artificial insertion of PSM MS^2 spectra into aligned runs at matching RTs) were described by Russell and Lilley (107).

The approaches described above require the ^{15}N incorporation level to be pre-specified and assume the level to be the same for all peptide/proteins. This narrows the range of feasible experimental designs (requiring incorporation to reach a stable level) and precludes use of ^{15}N for pulse-chase experiments (particularly useful in the analysis of protein turnover), where the level of incorporation not only varies from peptide-to-peptide but is the metric of experimental interest (as opposed to the labelled/unlabelled ratio, indeed the sample may not even contain unlabelled peptide). Accounting for variable incorporation introduces considerable challenge, in particular controlling the potential for incorrect estimation of peptide incorporation (due to noise, or co-eluting species within the mass error window of integrated peaks) leading to mis-quantification.

There have been two paradigms described for incorporation-agnostic quantification of ^{15}N data, which may be classified as to whether or not ^{15}N incorporation in the labelled signal is estimated in the course of determining the unlabelled to labelled ratio. In many experimental designs the relative amount of unlabelled to labelled (but unknown incorporation) peptide is of more interest than the incorporation level itself. In this ‘ratio-only’ approach, ^{15}N quantification is analogous to SILAC quantification and provides an alternative *in vivo* methodology when SILAC labelling is complicated by other factors. It is not necessary to determine an exact incorporation level to quantify the unlabelled to labelled ratio as the total intensities of the unlabelled and labelled signals across all their isotopologues can simply be compared. A downside to not estimating ^{15}N incorporation is that co-eluting interference by species of coincident mass with certain isotopologues can only be detected by the effect on the calculated ratio, rather than during calculation of the respective unlabelled and labelled signal intensities (where it may be corrected).

An early published example of this approach is the work by Zhang *et al.* (180) whose *ProTurnyzer* tool refines the broad approach described above to optimise for cases of very low incorporation where the unlabelled/labelled mass coincidence is substantial. Subsequently Lyon *et al.* (181) described a further refinement to the ‘ratio-only’ approach. Their *Protover* tool processes samples in order of expected decreasing ^{15}N label incorporation level, for example by processing a time course of MS sample analyses in order of expected decreasing incorporation (so, for example, a time course of incorporation on labelled media would be analysed in reverse). For suitable experimental designs this allows the extraction of label masses at each time point to be refined based on the masses observed in previously processed data files (as the maximum isotopologue mass found for a peptide in each file may be taken as an upper bound of the possible isotopologue masses for that peptide in subsequent files).

The alternative paradigm for ^{15}N quantification may be summarised as the 'theoretical-distribution-matching' approach in which labelled distributions are characterised by comparing the intensity ratios of the isotopologues of each ^{15}N labelled signal to a range of theoretical distributions using a scoring system, in order to find the best fit. Knowledge of the incorporation level may then be applied to optimise calculation of the unlabelled to labelled ratio by estimating the total area of the labelled signal based on the most intense isotopologues only (increasing the effective signal to noise).

Over the course of two papers (131, 182), Snijders *et al*. described a manual implementation of this paradigm, in which ion intensities for the unlabelled and labelled distributions of each peptide were extracted using the peak integration capabilities in the instrument vendor software *Analyst Qs* (Applied Biosystems). For the labelled distribution, incorporation was then manually characterised by comparison to theoretical distributions generated using IsoPro (https://sites.google.com/site/isoproms/home) to generate theoretical spectra.

Contemporaneously, MacCoss *et al.* (183) described an automated incorporation determination approach to theoretical distribution fitting, although this lacked the ability to track across the RT dimension to find the peak apex or integral, or match by RT between runs.

The work by MacCoss *et al.* was later adapted by Huttlin *et al.* (184) into a quantification workflow which combines ion chromatogram extraction with automated incorporation determination to produce an incorporation-agnostic automated approach. In the Huttlin *et al.* paper, the approach was benchmarked at both near-complete labelling, and also at very low (<10%) partial incorporation. The MacCoss and Huttlin studies use Pearson correlation as their isotopologue distribution matching metric.

A similar workflow was described by Price *et al.* (165), detailed methodology for which was later elaborated by Guan *et al.* (185). These methods use a non-negative least squares algorithm for distribution matching. These analytical workflows involved a series of scripts rather than a single integrated software tool, which may limit their uptake by the wider community. Later iterations of the 'two-feature' approach have presented ^{15}N quantification tools that are a single program or script, which make further refinements to the theoretical-distribution approach. *Protein TurnStILE* (160) allows the list of peptides for quantification to be generated separately from the quantification runs, so that peptide identifications can be obtained in separate dedicated runs using only unlabelled sample for optimum peptide-spectrum matching. The isotopologue distribution matching metric used by *Protein TurnStILE* is least-squares rather than Pearson correlation.

The recent work by Fan *et al.* (166) describes a hybrid approach, *ProteinTurnover*. After ion extraction, an amalgam of the unlabelled and labelled distributions represented by a composite of two beta-binomial distributions is fitted to the data using maximum likelihood estimation. Theoretical ^{15}N incorporation distributions are not matched directly to the data, but the composite distribution includes a shape parameter representing ^{15}N incorporation in the labelled signal, so the parameters of best fit do include an estimate of incorporation level. This particular tool will be discussed in more detail in Chapter 4 in the context of a comparison with the work of this thesis.

1.2.3 *HeavyMetL*

Despite steady development of quantification approaches, no tool described thus far for EML (or specifically ^{15}N) quantification has seen substantial adoption by the proteomics community for this purpose (although there are some widely used tools such as *Mascot Distiller* which support rudimentary ^{15}N quantification among other features). This may be attributable to two factors. Firstly, usability; published approaches are typically scripts run in an environment such as R or Python, which are intimidating and inaccessible to many researchers. A hallmark of most tools widely used by the field is that they have at least a rudimentary graphical user interface (for local programs e.g. *MaxQuant* (82)) or a web interface for remotely hosted services such as *UniProt* (186) or *Panther* (187); this is not the case for existing options. Secondly, the tool must be sensitive and accurate for general use (not just for very high-quality data). For the tool to be useful on a large scale, it must produce robust results for a majority of PMCs identified. For reference data, results must be consistent with expected values, and for experimental data must compare favourably to results obtained by alternative quantification algorithms. Critically, it must not require extensive manual inspection of spectra to corroborate findings and must have acceptable performance in cases of lower than average signal-to-noise. While heretofore published solutions provide broadly accurate results, the incidence of substandard quantification is difficult to gauge as publications typically present little benchmarking and either include no comparisons to earlier work or report only cherry-picked example peptides. This may also relate back to the first factor, in that even bioinformaticians working in the same field are not confident in setting up and configuring other published programs in a way that will allow fair benchmarking!

The aims of this work were four-fold. Firstly, to develop and present a tool with an accessible interface front-end to a sound quantitative algorithm. Secondly, to show that the design choices made in the algorithm structure and calculations support a claim of robust and accurate quantification, and to explore the limitations of the algorithm as data quality decreases. Thirdly, to show that the results produced compare favourably to existing modern approaches. Lastly, to argue that the tool, supported by this analysis, is justifiably an improvement on existing work, and a practical addition to analysis resources in the field of proteomics.

In the chapters below, I present, a software tool, *HeavyMetL* for analysis of ^{15}N MS data that addresses the issues above, namely usability/accessibility, accuracy of quantification and robust handling of suboptimal quality data without extensive manual supervision. I have compared the performance of several isotopologue distribution matching metrics in the context of this tool to assess spectral matching performance over a range of incorporation levels, and to vindicate the selected metric as the keystone of my algorithmic approach. Finally, I have benchmarked my tool by comparison against a recently published hybrid quantification algorithm (*ProteinTurnover*) that has used a degree of orthogonality in its approach (166), and have concluded with an argument for its utility.

Chapter 2: *HeavyMetL*: A Program to Analyse ^{15}N-Labelled Proteomic MS Data

2.1 Introduction

This chapter describes the design and implementation of a software package for determination of relative abundance and label incorporation in ^{15}N-labelled samples analysed by LC-MS/MS. The package, *HeavyMetL*, is written in JavaScript and runs on the latest versions of popular freely available web browsers (Mozilla Firefox and Google Chrome), enabling graphical analysis of ^{15}N data without the necessity for command-line interaction. The quantification algorithm is an evolution of several approaches that have been previously described for ^{15}N, with refinements to improve signal-to-noise and to handle sample-to-sample chromatographic differences, implemented in a multi-threaded system. Chapters 3 and 4 expand upon the work detailed herein, the former describing analysis of various possibilities for the spectral matching scoring system used in *HeavyMetL*, and the latter analysing a benchmark comparison of this software to a pre-existing analysis tool.

2.2 Proteomic Software Design Considerations

Researchers working with proteomics data are drawn from a wide array of backgrounds and cannot be expected to be experts in either proteomic mass spectrometry or bioinformatics. A recurrent issue in all ‘big data’ -omics fields is a disconnect between the expectations of the bioinformaticians and computer scientists who build new tools for analysis and the experience of biological researchers who ultimately form the bulk of the software user base after release.

A recent report (188) on the experiences of end users of MS software highlighted this disconnect succinctly:

> *"Interviewees commonly complained about the lack of user-friendly software. Parameter setting and manual interaction is a significant time consumption for mass spectrometry scientists. Twenty-seven interviewees specifically mentioned how manual intervention required a significant amount of their time, with percentages ranging 10−50%. Users also spend a significant amount of time in learning how to use software. User complaints about software included hard to learn interfaces, inefficient interfaces, and broken features."*

'Manual intervention', in this case, refers to manually selecting parameters on a 'dataset-by-dataset' basis (or even on individual elements within a dataset) because the default parameters (whether global defaults or automatically selected based on dataset characteristics) do not yield acceptable algorithmic decision making (with regard to peak selection, spectral matching *etc.*) for most elements within the dataset.

Furthermore, the report goes on to observe that while developers (interviewees who spent a significant portion of their time coding) consider the majority (92%) of the remaining unsolved problems in computational MS analysis to be minor issues, among non-developers the reverse was true; they considered the majority (86%) of the remaining unsolved problems to be major issues. While subjective, this report could be interpreted to suggest that although many problems have been addressed by theoretical algorithm development, the rate of translation of these solutions into broadly accessible 'complete products' (or at least, software that is seen by non-developers as complete products!) is low.

Prior to the implementation of *HeavyMetL* as a software package, extensive consideration was given to the characteristics of the user base, and a number of design decisions were made with these characteristics in mind. My target user is a researcher looking to analyse data from a proteomic experiment. I assumed that the researcher would have sufficient involvement in the experiment to be familiar with the experimental design, sample processing and MS run configuration, but would not necessarily be the proteomic mass spectrometrist who performed the MS part of the analysis. They could also be a primary researcher who had entrusted the MS work to a colleague or service facility, or they could be a bioinformatician, either working with the primary researcher or re-analysing a public dataset. The target user could therefore possess a range of skill levels with regard to the relevant steps in a proteomic experiment (Table 2-I), although for practicality a lower bound had to be assumed for most skills, including basic familiarity with proteomic techniques (digestion of sample to peptides, identification of peptides by MS^2 and database or spectral library searches, and inference of protein parents) and mass spectrometry concepts (spectra of mass-to-charge ratios *vs.* intensities, and total/extracted ion chromatograms). From personal experience it also seemed reasonable to assume the user would have some prior experience with other graphical user interface (GUI)-based identification and quantification software (such as *Mascot* and *MaxQuant*) and spreadsheet programs such as *Microsoft Excel*, but may not be comfortable working on the command line (navigating the file system, performing file operations, installing/running command-line based programs, installing dependencies *via* package managers or compiling code) or with dedicated statistical environments, whether command-line based (such as R) or GUI-based (such as *PRISM*).

	Researcher Demographic		
	Biologist / Biochemist	Proteomic Mass Spectrometrist	Bioinformatician
Sample Preparation			
Biological Background	Expertise	Familiarity	Acquaintance
Experimental Design	Expertise	Expertise	Expertise
In-vivo labelling	Proficiency	Expertise	Familiarity
Proteomic Sample Preparation			
Protein Extraction	Familiarity	Expertise	Familiarity
Sample Cleaning	Familiarity	Expertise	Familiarity
Digestion	Familiarity	Expertise	Familiarity
Bottom-up Proteomic MS Analysis			
LC Separation	Familiarity	Expertise	Familiarity
Data-Dependent MS Analysis	Acquaintance	Expertise	Familiarity
Data Processing			
Peptide-spectrum matching	Acquaintance	Expertise	Proficiency
MS data types (peaklist, rawfile etc) and formats (mgf, mzML etc)	Acquaintance	Proficiency	Proficiency
DDA data concepts (retention times, ion chromatograms, spectra etc)	Familiarity	Expertise	Familiarity
Popular GUI quantitation software	Acquaintance	Expertise	Proficiency
Command-line navigation	Acquaintance	Acquaintance	Expertise
Statistical Analysis			
Basic statistical concepts	Familiarity	Familiarity	Expertise
MS Data-specific statistical concepts	Acquaintance	Proficiency	Proficiency
Spreadsheet software	Proficiency	Proficiency	Expertise
Dedicated statistical environments e.g R, MATLAB	Acquaintance	Acquaintance	Expertise

Table 2-I. *Anticipated Minimum Proficiencies for End-User Demographics.*

Assumed proficiency is ranked as Acquaintance < Familiarity < Proficiency < Expertise.

In order for the software to be an attractive solution for an average such researcher, I considered the following themes to constrain design decisions for the software.

1. **System Requirements.** The software must be usable on an average contemporary personal computer (broadly, purchased within the last 5 years, with major system updates having been applied and with sufficient hard disk space to hold the data to be analysed locally). Ideally, system requirements such as operating system and prerequisite libraries are to be minimised so that initial setup is as straightforward as possible. Experience has shown that the average non-bioinformatician user does not have limitless patience to play around installing software, particularly if it requires interaction *via* the command line.

2. **User interaction.** Further to the above, general operation of the program must also be accessible to users. The number of parameters that **must** be manually configured should be minimised, but advanced options should be available to expert users to allow for optimisation. The program should have a graphical user interface to avoid command-line interaction. The results should be output in a simple text table format to maximise compatibility with statistical analysis and spreadsheet software. Furthermore, the results should also be displayed graphically (and at a quality suitable for publication) to allow easy evaluation of quantitative performance and assess the effect of configuration changes. Graphical views of spectra are offered in many existing MS quantification packages such as *Progenesis*, *Proteome Discoverer*, *Skyline (189), Spectronaut and Mascot Distiller*. *MaxQuant* originally lacked any such display, and the later introduction of this feature was (in my personal experience) well received in the field.

3. **Speed.** It is acceptable, taking contemporary quantification software such as *MaxQuant*, *Progenesis or Proteome Discoverer* as examples, for a large analysis involving several raw data files to take several hours on a desktop computer. It is considered typical that quantification analyses may need to be run overnight, even for modestly-sized experiments (for example up to 5000 confident peptide IDs, 6 raw data files). However, a single run should be analysable well within a working day (ideally no more than 1-2 hours). Quantification for a single PMC in a single run should therefore take no more than a few seconds to compute, given an expectation of thousands or even millions of quantifications to be performed when

accounting for a large dataset with many runs. 'Lag' or unresponsiveness of the interface should be avoided (a problem even in commercial software, particularly in *Proteome Discoverer* and *Progenesis* when navigating results) and quantification progress for time-consuming tasks should be updated frequently to avoid giving the impression that the interface has 'frozen'.

4. **Memory usage.** Analysis of MS data involves data file sizes ranging from a few hundred megabytes up to several gigabytes. A contemporary midrange personal computer may be assumed to have between 8 and 64 gigabytes of memory. Analysis will require the reading of large amounts of data from each file. It is more efficient to store this data in memory where possible but loading whole raw data files into memory will quickly use up all available space, resulting in the system swapping memory with disk storage (a very slow operation). Loading data only as required reduces memory load but introduces a throughput bottleneck of disk-reading speed, therefore disk operations must be arranged such that reading from disk is performed before the data are required for quantification. A technique used by *MaxQuant* (among others) is pre-indexing of raw data files before quantification starts so that disk reads during quantification are limited to the exact place within the file where each spectrum is stored.

5. **Quantification quality.** Robustness of quantification is critical. If manual analysis of the data consistently out-performs a quantification algorithm it will be seen as not reliable. This requirement sets a deceptively high bar, as experienced researchers are very good at evaluating spectral quality 'by eye' and setting integration limits so as to maximise signal-to-noise. The quality of manual quantification produced by an experienced researcher will typically be very high. The theoretical advantages of computational approaches (aside from the feasibility of analysing many PMCs) are granular quality evaluation (rather than pass/fail), allowing information from poor quality quantification to be used with minimal bias, and consistent quality independent of researcher expertise. In order for the output of such approaches to be useful, however, they must compare robustly to experienced manual quantification.

2.2.1 Language and Distribution

Open source bioinformatics projects are usually implemented in one of several ways. They may be made available as a downloadable program which is compiled against a target operating system (OS) (e.g. *Skyline*). The major advantage of directly compiled code is speed, particularly in the case of extensive numerical calculations (which does apply to quantification tasks). As discussed above regarding memory usage, however, a second bottleneck is the speed at which data may be read into memory from raw data files stored on disk (since the size of such files precludes keeping the entire file set in memory in the case of many analyses). Even with appropriate indexing of raw data files to minimise the data read, and scheduling of disk reads to ensure no wasted disk-read time, disk read time still applies an effective minimum bound on quantification, thus the speed advantage of compiled code is limited. The disadvantage to compiled programs is that they induce additional testing workload on development as some bugs may be platform-specific (assuming that versions of the program are compiled for more than one OS).

Alternatively, the program may be written in an interpreted language (e.g. Perl, Python), or compiled to virtual machine bytecode (e.g. Java), both of which result in a platform-agnostic distribution. This approach is quite popular for proteomics tools, as it is somewhat easier to resolve bug issues, and the loss of speed *vs*. compiled code is, in most cases, minimal. Some operating system-related issues are quite common, often relating to differences in the interpretation of file paths and text file line endings, or use of non-OS agnostic code libraries, but these common pitfalls may be avoided. A trickier issue lies in installation. While an OS-agnostic distributed program may be run on any system with correct setup, it is common for non-bioinformaticians to run into problems during installation such as incorrect execution environment version (especially in the case of Java, where version updates are frequent and may not be backwards compatible without adjusting default security policies), mis-configured system variables or failure to install dependencies not bundled with the program download (for example, in scripting languages such as Perl, Python and R, users are often expected to be familiar with installation of modules from centralised repositories). Such foibles are minor annoyances for experienced computer users but are very off-putting for users expecting things to 'just work'. Existing ^{15}N quantification tools as described in Section 1.2.2 have a poor track record in this regard. The four most recently published approaches implemented as distributable software (*ProTurnyzer* (180), *Protein TurnStILE* (160), *Protover* (181) and *ProteinTurnover* (166) are all command-line based approaches which run *via* an interpreting runtime (*ProTurnyzer* and *Protover* in Python, *Protein TurnStILE* in Perl and *ProteinTurnover* in R), a compatible version of which (not necessarily the latest!) must be installed and configured correctly to be callable from a command line console. Indeed, a prototype design of the *HeavyMetL* core quantification algorithm was initially written and tested in Perl. Ultimately, users reporting issues and deployment inconveniencies I experienced myself led me to discard this strategy before the final implementation.

Thirdly, projects may be hosted on a remote server and accessed *via* a web interface. This approach is arguably the most operating system-agnostic; great efforts have been made to standardise behaviour of websites in browsers across operating systems.

Usually, this approach places the onus of computation on the hosting server rather than on the local machine. The actual program may be written in a compiled or interpreted form, or even some combination thereof. This is invisible to the user and does not increase setup complexity, so may be engineered for optimal speed or maintenance. For quantification, this may on first glance seem ideal, as the servers are likely to be far more powerful than a personal computer. However, the server must be maintained (requiring a larger on-going commitment to the tool) and may become overloaded if usage is greatly increased. Furthermore, in the case of quantification, a large amount of data must be sent to the server, which, unless the server is on a local network, is time consuming and risks time-out of the connection. This is likely to offset any speed benefits of computation server-side unless the quantification to be performed is extremely complex. Transferring data for remote processing also raises issues of privacy; many researchers are understandably reluctant to send large amounts of data to third parties for analysis. In some cases (e.g. medical data containing identifiable patient information) submitting data to a third party for processing presents ethical and legal difficulties.

Another possibility, which is the route I have taken for *HeavyMetL*, is to provide a web-based tool in which the processing is performed locally within the browser. This takes full advantage of the highly standardised behaviour of popular multiple operating system-compatible browsers, particularly *Mozilla Firefox* and *Google Chrome,* and a ready-made GUI platform *via* the HTML document object model, while avoiding the need for server maintenance or long data transfer times. This option has only recently become feasible as a possibility following large improvements in JavaScript speed in-browser (particularly driven by competition between Mozilla's *SpiderMonkey* JavaScript engine and Google's *V8* JavaScript engine) and the adoption of several key JavaScript standard features including local file access and multi-threaded execution. Running in a browser still has limitations; the maximum usable memory is limited, and raw computation speed will never reach the limits of compiled, platform-optimised code. These limitations may be mitigated somewhat by careful program design. Memory footprint is to be minimised in any case - see Design Constraint 4, Section 2.2, and computation time is only one bottleneck on the overall analysis speed; disk access rates will also constrain the analysis of large files and, at a higher level, the ease with which the user can supply the necessary input data and evaluate results also contributes to the total time for the analysis.

2.2.2 Program Requirements

2.2.2.1 Input

Extensive pre-processing of raw data is both time consuming and compares unfavourably to similar modern quantification platforms such as *MaxQuant* that are able to read raw data directly. The program must therefore be able to accept a set of raw data files with minimal pre-processing. Data files produced by MS instruments from different vendors are each in different, proprietary binary formats. While most vendors supply code libraries that allow third-party programs to read these formats, they are almost universally Windows-only, and thus would require the quantification software to also be limited to Windows if it is to directly read these proprietary formats. Conveniently, a conversion tool (*msconvert* from the *ProteoWizard* suite (190)) which reads the vast majority of MS data formats is available, and widely used within proteomics. This may be used to convert raw data files into open standards for raw MS data that are widely used in proteomics, namely the mzML format (191) and the older mzXML format (192). For maximum compatibility, it makes sense to target these open standards as raw data files from almost all instruments can be rapidly and easily converted to this format either directly from the acquisition software or with *msconvert*. Raw data in mzmL/mzXML format takes the form of the three-dimensional data space described in the Introduction (see Figure 1-XX.), specifically, a series of spectra (*m/z vs.* intensity) stored in order of acquisition, and thus corresponding to an LC retention time. The information regarding each consecutive scan is stored as a list of scan parameters (the ‘scan header’) followed by a record of the ‘spectrum data’ in terms of *m/z* values and intensity pairs, in a space minimising structure (and possibly further compressed, depending on file format). Various items of additional metadata, such as instrument hardware and configuration information, are also stored at the start of the file, and at the end there may be (depending on the conversion process) a byte-indexed record of every scan entry.

To accompany the raw data, information regarding the PMCs to be quantified is required. Peptide-spectrum matching and identification of PMCs for quantification is a significant computational challenge that is outside the scope of this work, and (at least for non-^{15}N labelled peptides) a wide array of software solutions exist. It is assumed for the purposes herein that such analysis will be performed separately. Unfortunately, there is no standard peptide identification output format that is consistently supported across search engines; the closest example would be mzIdentML (193), but as an XML-based format this is not readily produced from the output of search engines that do not export it directly. It is therefore necessary that for inputting PMCs the program should accept a simple text-table file which can be generated from the tabular output of any search engine output requiring only minimal rearrangement of the data. Direct support for the outputs of popular search engines is something to consider as a secondary goal, but this would impose a large burden in terms of potential input bugs and would also add an ongoing requirement to maintain compatibility for these formats over time.

Finally, the analysis of generalised ^{15}N data from various experimental designs, with a wide variety of sample preparation approaches and analysis setups will entail some pre-configuration of analytical parameters, at least if the algorithm is to yield optimal results. While it will be possible to estimate (or extract from raw data file metadata) many characteristics of the data during processing, it will likely speed processing and reduce anomalous results if certain parameters can be constrained beforehand. Examples of such constraints would include the expected range of ^{15}N incorporations and the expected range of RT shifts for a PMC between runs. Other parameters must be set by the user, such as the definition of any expected post-translational modifications on the peptides. For simplicity, the default settings should be expected to produce reasonable quantification for common arrangements of experiment, sample preparation and analysis.

2.2.2.2 Output

Given the input data, the program should calculate quantification information for each PMC. For ^{15}N data this information consists of the incorporation level of the ^{15}N labelled signal and the relative abundance of the labelled and unlabelled signals. Various further statistics for each quantitative result will also be of interest, including the RTs of the labelled and unlabelled peaks and a level of uncertainty for the quantification. The program should also calculate quantitative information at the protein level since this will be the most relevant output for many researchers. The requirement for both text-table and graphical type outputs are necessary for easy user interaction (see Design Constraint 2, Section 2.2).

2.3 Program Implementation

2.3.1 Overview

HeavyMetL is a quantification tool for the analysis of ^{15}N-labelled samples by mass spectrometry. Given a set of raw MS data files and a list of PMCs, spectra are extracted from raw data files and fitted to a range of potential ^{15}N incorporations to generate values for both abundance ratio and incorporation level of peptides and proteins. User interaction is *via* a graphical interface that lists the files and peptides/proteins specified, allowing users to 'browse' the results of quantification graphically at both the peptide and protein levels, alter quantification settings from within the program and view the results of such changes. Once satisfied, the user can export both the peptide and protein level results to a text-table format for further analysis in a spreadsheet program (e.g. *Microsoft Excel*) or statistical analysis suites such as *Perseus*, R or *MATLAB*.

Details of access to the program code and a web address for a live implementation of *HeavyMetL* are given in Appendix I.

2.4 Program Operation

HeavyMetL takes three types of input data and produces output in the form of text tables and graphics. A program schema is shown in Figure 2-I. The first type of input data required is raw data, in one of the open standard formats (mzML or mzXML) as described above in Section 2.2.2.1. Secondly, a list of peptides to be quantified is supplied (in tab- or comma-separated format), listing each identified instance of PMC individually, as they represent different *m/z* and RT locations in specific raw data files. The columns required in this list are given in Table 2-II. These data are commonly available in standard exports from popular programs. For example, the “evidence.txt” result file in *MaxQuant*, the tab-delimited export of the Peptide table in *Proteome Discoverer*, the spectrum report from the *Scaffold* data aggregation tool (http://www.proteomesoftware.com/products/scaffold/), and the mzTab export from supporting programs (including *Mascot*) all contain the necessary data without the need to splice together multiple search engine exported output files. Thirdly, various quantification configuration parameters may be set by the user, although the default settings are selected to give reliable quantification in most circumstances. An overview of these parameters is given in Table 2-III.

After peptides have been quantified, results may be exported at the peptide or protein level. Protein level quantification is inferred from constituent peptides (according to protein grouping information supplied in the list of PMCs), with a filter applied to exclude a percentile of peptides with the worst labelled signal Similarity Scores in each run (default 5%). The labelled/unlabelled ratio and labelled signal incorporation are calculated separately as the corresponding median quantification result across all the remaining quantified peptides for the protein in that run. The columns included in the output are also listed in Table 2-II

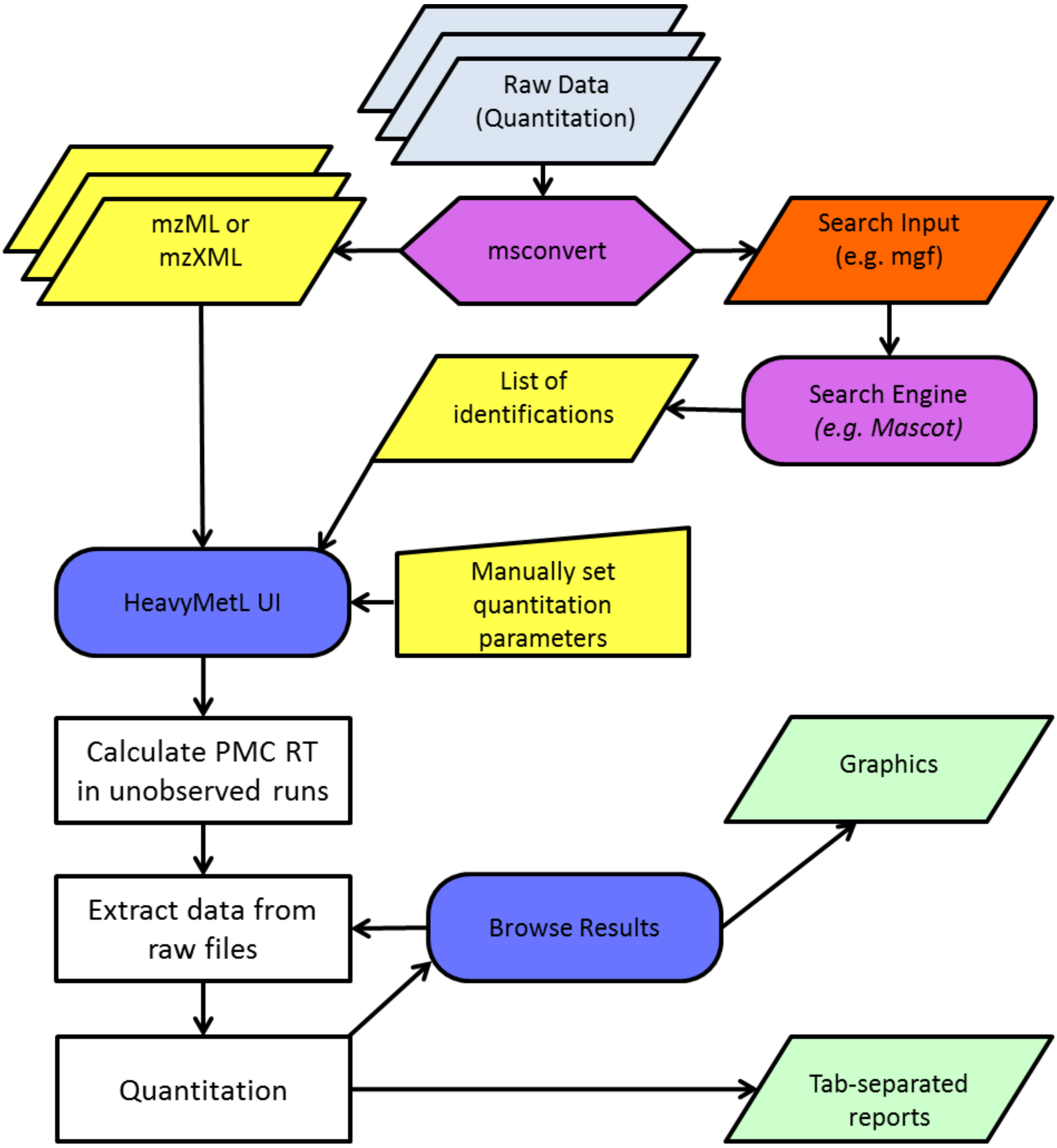


Figure 2-I. HeavyMetL *Program Schema / Processing Workflow. Yellow and green boxes are input provided by and output provided to the user respectively. The grey box is the starting raw data file. The purple boxes indicate pre-processing performed prior to* HeavyMetL *analysis; the orange box indicates intermediate data typically generated during pre-processing. The blue boxes indicate user interaction with* HeavyMetL *via the GUI, while the white boxes indicate processing performed by* HeavyMetL *without user interaction.*

	Type	Parameter	Table Column Name	Notes
INPUT	Raw Data	.mzML	-	Any combination accepted.
	Raw Data	.mzXML	-	
	PMC List	Protein Group	PROTEIN	Used to group peptides for protein-level quantitation.
	PMC List	Protein Description	PROTEIN_DESCRIPTION	For data pass-through and display only.
	PMC List	Protein ID Score/Likelihood	PROTEIN_SCORE	For data pass-through and display only.
	PMC List	Peptide Sequence	PEPTIDE_SEQUENCE	Given without preceeding/trailing cleavage site indicator or neighbouring residues. Case is ignored.
	PMC List	Peptide ID Score/Likelihood	PEPTIDE_SCORE	For data pass-through and display only.
	PMC List	Peptide contributes to protein quantitation?	CONTRIBUTES_TO_PROTEIN	Whether peptide should be considered for when calulating protein-level quantitiation data (this will generally be true). Accepts synonyms of yes/no and true/false and ignores case.
	PMC List	File Name where PMC was detected	FILE_NAME	For files where there a particular PMC was not detected, the RT is estimated using the mean RT from all files where it was detected.
	PMC List	Scan Number of PMC detection	SCAN_NUMBER	Either may be supplied. Retention time is used by preference if both are present.No assumption is made regarding if the Scan/RT referrs to the MS^1 event in which the precursor was observed or the MS^2 event which gave rise to the PSM - both are calculated by reference to the corresponding raw file.
	PMC List	Retention Time of PMC detection	RETENTION_TIME	
	PMC List	Modifications	MODIFICATIONS	*Scaffold* -style modification format i.e. [ResidueLetterCode][Position]: ModificationName
	PMC List	Charge State	CHARGE	All values are assumed to be positive charge (sign is ignored).
OUTPUT	Peptide Table	Protein	PROTEIN	Pass-through from input PMC list.
	Peptide Table	Peptide	PEPTIDE_SEQUENCE	
	Peptide Table	Is Unique?	CONTRIBUTES_TO_PROTEIN	
	Peptide Table	Charge	CHARGE	
	Peptide Table	Modifications	MODIFICATIONS	
	Peptide Table	Unlabelled Match Score	[file]_SCORE_UNLABELLED	Repeated for each raw data input.
	Peptide Table	Unlabelled Intensity	[file]_INTENSITY_UNLABELLED	
	Peptide Table	Labelled Match Score	[file]_SCORE_LABELLED	
	Peptide Table	Labelled Intensity	[file]_INTENSITY_LABELLED	
	Peptide Table	Labelled Incorporation %	[file]_INCORP_LABELLED	
	Protein Table	Protein Accession	PROTEIN	Pass-through from input PMC list.
	Protein Table	Protein Description	PROTEIN_DESCRIPTION	
	Protein Table	Unlabelled/Labelled Ratio	[file]_RATIO	Repeated for each raw data input.
	Protein Table	Label Incorporation %	[file]_INCORP	

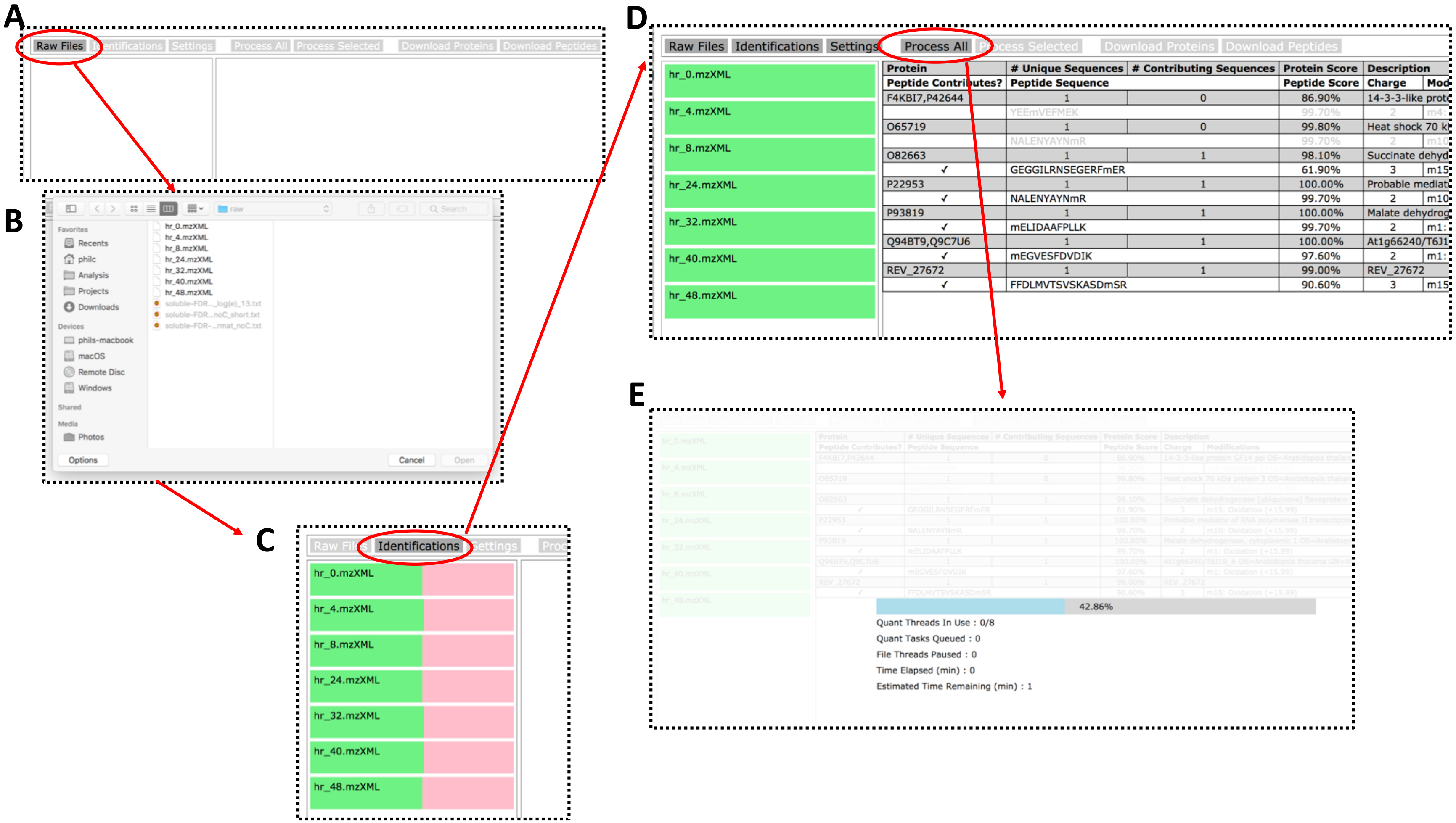
Table 2-II. HeavyMetL *Inputs and Outputs.*

Parameter	Description	Default Value
Fixed Modifications	List of fixed modifications (comma or semicolon-separated). Format is "Modification Name (Residue Single Letter)"	Carbamidomethyl (C)
m/z Error Tolerance (ppm)	Window in *m/z* units (Thompsons) around theoretical *m/z* values from which intensity is to be retrieved.	10
Retention Time Window (min)	Retention time window about observed/estimated peptide identification time in which to retrieve spectra.	0.5 *(i.e. 30 s)*
Maximum Peak Apex Shift (min)	Maximum retention time shift to allow when searching for labelled signal apex relative to unlabelled signal apex (if one was found).	0.2 *(i.e. 12 s)*
Do Not Quantify Unlabelled Signal	Do not quantify unlabelled signal, or correct for unlabelled signal presence when quantifying labelled peptides.	FALSE
Minimum Label Incorporation %	Minimum incorporation percentage tested when quantifying labelled signal.	10
Maximum Label Incorporation %	Maximum incorporation percentage tested when quantifying labelled signal.	95
Peptide Match Score Threshold	Do not use peptides with Similarity Score below this value for protein-level quantitation.	0.85
Show XICs on Log10 Scale	Show extracted ion chromatogram (XIC) *y-* axis on log scale to accentuate chromatographic variation at peak boundaries. Affects displayed graphics only.	FALSE
Y -Axis Precision	Number of decimal places to show. Affects displayed graphics only.	2

Table 2-III. *User Configurable Processing and Display Parameters.*

2.4.1 Interface

HeavyMetL presents an empty table to the user on launch, with a series of buttons along a top menu bar, all of which are initially greyed out except “Raw Data” (Figure 2-II-A). Clicking on this button presents a standard file selection dialog (Figure 2-II-B) allowing the user to select the raw data to be processed. Submitting this dialog begins a process of raw data file pre-indexing (see Section 2.4.2) to optimise later data read rates. Indexing progress is shown for each file (Figure 2-II-C). During and after the pre-indexing process, the user may click on the “Identifications” button to select a list of PMCs for analysis *via* another file selection dialog. After pre-indexing, *HeavyMetL* iterates through the list of PMCs, where necessary predicting a suitable retention time for extraction if there is no direct MS^2 evidence in that file (also see Section 2.4.2). Having predicted any missing retention times, a table of the proteins represented by the PMC list input is presented by the *HeavyMetL* interface (Figure 2-II-D). The “Settings” and both quantification processing buttons are now available. The “Settings” button presents users with a screen that allows various configuration parameters to be changed (see Table 2-III). The two quantification processing buttons, “Process All” and Process Selected” allow users to choose to analyse all defined PMCs at once, or alternatively select a specific PMC or protein and perform quantification only on the selected PMCs for a faster result. This latter option also graphically reports additional metadata regarding the quantification (relative ratio, the chromatographic profile of the extracted signal and final integrated spectra for the unlabelled and labelled signals) to allow the user to browse through and assess quantification performance visually (see Section 0). In either case, although most relevant for the Process All option, initiating quantification processing displays a processing overlay with a progress bar and (for advanced users) the current distribution of CPU effort between data extraction *via* the file worker threads and quantification *via* the quantification worker threads (Figure 2-II-E). Larger sized screenshots may be found in Appendix II.



Protein	# Unique Sequences	# Contributing Sequences	Protein Score	Description	
Peptide Contributes?	Peptide Sequence		Peptide Score	Charge	Mod
F4KBI7,P42644	1	0	86.90%	14-3-3-like prot	
	YEEmVEFMEK		99.70%	2	m4:
O65719	1	0	99.80%	Heat shock 70 k	
	NALENYAYNmR		99.70%	2	m10
O82663	1	1	98.10%	Succinate dehyd	
✓	GEGGILRNSEGERFmER		61.90%	3	m15
P22953	1	1	100.00%	Probable mediat	
✓	NALENYAYNmR		99.70%	2	m10
P93819	1	1	100.00%	Malate dehydrog	
✓	mELIDAAFPLLK		99.70%	2	m1:
Q94BT9,Q9C7U6	1	1	100.00%	At1g66240/T6J1	
✓	mEGVESFDVDIK		97.60%	2	m1:
REV_27672	1	1	99.00%	REV_27672	
✓	FFDLMVTSVSKASDmSR		90.60%	3	m15

42.86%

Quant Threads In Use : 0/8

Quant Tasks Queued : 0

File Threads Paused : 0

Time Elapsed (min) : 0

Estimated Time Remaining (min) : 1

Figure 2-II. *User Interface Screenshots.* ***A****: Initially presented blank table.* ***B****: Raw file selection dialog.* ***C****: Raw file pre-indexing progress.* ***D****: Ready to start quantification.* ***E****: Quantification in progress. For expanded versions of these screenshots see Appendix II.*

2.4.2 Loading Data and Pre-Processing

Raw spectral data files are pre-indexed before quantification begins to improve spectrum retrieval speed during the quantification process. *HeavyMetL* loads the scan header data (see Section 2.2.2.1) into memory along with a byte reference to the within-file location of the spectral data so that information such as scan number, MS level (e.g. MS^1) and RT are quickly accessible, while the much larger spectral data structure is only loaded if necessary. If there is no byte index at the end of the raw data file, one is constructed by mapping all scan headers within the file – while this delays indexing, it is necessary for efficient navigation of the file during quantification. *HeavyMetL* further cross-references MS^1 headers as a linked list, such that each header records the index of the previous and next headers (raw data file indices, while consecutive, are not guaranteed to be continuous if, for example, intervening MS^2 data were removed during conversion of the raw data file to an open format). This allows processing to quickly iterate through consecutive spectra within a RT window without having to advance scan number by scan number, checking to see if each is present, or having to unnecessarily load non-relevant MS^2 (or MS level) data.

After file indexing is complete, the list of PMCs is loaded from the user-supplied text table. This file is cross-referenced to the selected raw data files while loading to calculate a RT for each PMC in each file based on what on information (MS^1 or MS^2 Scan number or RT) is provided in the PMC list. *HeavyMetL* will attempt to quantify PMCs across all input raw data files; allowing for chromatographic RT shift within a window (default 30 s); it is assumed that files will have comparable gradient conditions and chromatographic performance (necessary both for PMC matching between raw data files and to minimise differences in performance of the quantification algorithm). This assumption is not unreasonable; it is generally considered good practice to analyse runs from the same dataset sequentially and chromatographic differences in the hands of experienced operators are minimal. Even in the case of technical issues mid-sequence (e.g. replacement of part of the LC system between runs), if the chromatographic run parameters are unchanged then a shift of more than 5 minutes would be very unusual.

If the samples are pre-fractionated, corresponding fractions across samples can be analysed together but analysis of multiple fractions across multiple samples is currently unsupported, as there is no method for defining which runs correspond to each pre-fraction (and thus limiting RT-matching to within a fraction group). If the PMC list is based on identification runs separate from the main quantitative dataset, these raw data files must also be included in the input.

For raw data files in which the PMC does not have a directly corresponding MS^2 spectrum (and thus MS^1 RT), the RT must be predicted. RT prediction has a wide range of possible approaches (Table 2-IV) but most quantification software uses some variant of cross-run matching with various levels of sophistication. In the case of *HeavyMetL*, a simple cross-run matching approach is implemented using the average of RTs from raw data files in the analysis where RT data is known.

Method	Required prior knowledge	Accuracy
In-run MS2 data	None	Exact
Cross-run comparison (same LC setup)	MS2 data collected from previous runs of sample	Good (variance depends on sample and LC reproducibility)
Computational based on existing data / Machine Learning	Requires comparable dataset as a starting point	Good-Fair
Published data (differences in LC setup)	Published data	Moderate-Poor
De novo prediction	Published RT constants, for example Meek, 1980	Poor

Table 2-IV. *Options for Matching PMCs Between MS Runs.*

2.4.3 Definition of Analysis Parameters

Before processing, the user may also define parameters relevant to the quantification *via* the Settings dialog: any fixed modifications (by default, carbamidomethylation of cysteine), the windows for mass error (default 10 ppm) and RT (default 30 s), the maximum unlabelled/labelled apex RT difference (default 12 s), and the expected range of the label incorporation percentage (default 10%-95% ^{15}N). *HeavyMetL* does not consider 96%+ incorporation by default (although the user can change this if they anticipate a very high level of incorporation) in order to avoid confusion with unlabelled co-eluting peptides in the expected mass range of the fully labelled target peptide. The default settings are expected to give good performance in most scenarios, assuming relatively modern instrumentation (2005 onwards).

2.4.4 Quantification Processing

HeavyMetL breaks down the processing of each PMC in the supplied list (and corresponding RTs) into two stages. Firstly, a mini-dataset for the PMC in each raw data file is assembled, comprising a set of theoretical distributions across the range of potential ^{15}N incorporations, and a minimal set of MS^1 data extracted from the raw data. Secondly, each mini-dataset is analysed to quantify the labelled and unlabelled ^{15}N isotopologue envelopes. For the labelled distribution, the process of quantification also involves identification of the ^{15}N incorporation whose corresponding distribution gives the best match to the data. This allows processing capability to be divided between the retrieval of spectra from the raw data files (which involves a lot of disk activity), and the calculation of quantification results (which is processor intensive).

HeavyMetL is multi-threaded, in that multiple sequences of operations are executed simultaneously, to make optimal use of modern computer processor capabilities. Threads are implemented *via* the WebWorker HTML standard (https://html.spec.whatwg.org/multipage/workers.html), which is effectively a JavaScript in-browser implementation of multi-threading. Processing is split between a ‘main thread’ and two groups of processing threads, firstly a set of ‘file workers’, each assigned to handle disk access for a particular raw data file, and the second a pool of ‘quantification workers’ which handle the calculations to quantify individual PMCs within a set of spectra.

The main thread is responsible for displaying the user interface, co-ordinating access to raw data files *via* file workers, loading the list of PMCs to be analysed and coordinating the exchange of data between the file workers and the quantification workers. After processing, the main thread also handles generation of graphics for display of results and synthesis of peptide/protein export tables.

To begin processing, the main thread constructs a list of the PMCs to be analysed in each file. PMCs are sorted in order of ascending RT. Theoretical distributions are generated for each PMC for both the unlabelled and possible labelled incorporation levels using the algorithm described by Kubinyi (see Figure 1-XXIV) (167). The input peptide sequence is parsed to gather information about the frequency of each amino acid, then the total number of carbon, hydrogen, nitrogen, oxygen and sulphur atoms are calculated. For each PMC, a set of 'extractions' is generated, where one extraction defines the theoretical *m/z* values (and associated expected proportionate intensities) for a particular ^{15}N incorporation level (including the unlabelled case) and thus, in combination with the *m/z* Error Tolerance parameter, defines *m/z* windows from which signal intensity should be 'extracted' from the observed spectral data. To save memory, the extraction data are only retained as long as is necessary to process the associated PMC in every file for which it is to be quantified. Each extraction has a corresponding *uses* counter which is initially set to the number of files in which the PMC is to be quantified and decrements by one every time the extraction is applied for a new file; when zero, the extraction data are deleted. The theoretical distribution code was tested for bugs by comparison with *IDCalc*, an existing implementation of the Kubinyi approach (http://proteome.gs.washington.edu/software/IDCalc/).

A 'quantification task' dataset is created for each PMC to be quantified in each file. A quantification task comprises all the elements required to perform quantification on a single PMC in a single file. The main thread fills in the starting information which comprises the full PMC-to-be-quantified definition - the protein, peptide, charge, modifications, file name, closest RT, and RT extraction window (first and last scan number). While some of this information (such as protein name) is not necessary for quantification itself or is implied by the handling file worker (e.g. the raw data file name), it is required later to combine the results of individually processed tasks into peptide- and protein-level quantification across files.

The set of tasks is then passed to the file workers which, for each PMC, pair this data with the set of extractions generated by the main thread, and then retrieve and add the spectral data (Figure 2-III). For each quantification task, the set of MS^1 spectra falling within the RT window around the RT of the PMC must be retrieved, which involves both reading the data from the raw data file and then decoding the data into a set of *m/z*-intensity pairs. The complete spectra are cropped to a relevant mass range according to the extractions. By delegating access to each file to a separate thread, *HeavyMetL* ensures that disk access (data can only be loaded from one disk location at a time) is not wasted while spectra are being decoded from the raw data. To avoid performing many time consuming short reads of the raw data file to extract individual spectra, the RT windows of multiple consecutively queued PMCs are combined to form a single large RT window corresponding to a large continuous block of raw data (since the PMCs are sorted by retention time), which can be loaded from disk in a single read and then accessed in memory as each PMC is processed in turn.

These optimisations help to prevent file access becoming a bottleneck for quantification by minimising the time that the disk is idle (not loading data from a file). Furthermore, uncropped full spectra from previous quantification tasks are not discarded until the current quantification task RT window has moved past them (in order to avoid repeating the work of decoding the data to *m/z*-intensity pairs). Since the PMCs are sorted in order of ascending RT, once the extraction window has ‘moved on’, spectra outside the window can be safely discarded.

Once the spectral data have been added, the quantification task is ready to be passed to a quantification worker for the calculation of a result. Due to the optimisations described above, the time taken by the file workers to retrieve all the spectra necessary for each quantification task is quite variable. If the file workers were to wait for each task to be passed to a free quantification worker before beginning the retrieval process for the spectra needed by the next task, a lot of disk access time would be wasted. If the file workers retrieve spectra and generate quantification tasks as quickly as possible, however, a different problem arises. When a series of tasks can be processed by the file workers very quickly (due to substantial RT overlap minimising the retrieval of new MS^1 spectra from each raw data file), or the tasks currently being processed by the quantification pool are particularly time consuming, then a large number of quantification tasks may ‘pile up’ waiting for a free worker thread. Storing even the cropped spectra for many tasks can cause memory usage to rapidly balloon.

This problem is resolved by a buffered scheduling system (Figure 2-IV). When a file worker completes preparation of a quantification task, if there is a free quantification worker, the scheduler in the main thread immediately assigns the task for processing. If all quantification workers are busy, the task is held in a first-in-first-out queue (of equal length to the quantification worker pool) allowing the file workers to proceed with retrieval of the spectral data for the next PMC. The ‘ballooning memory’ issue is mitigated by applying a maximum length to this queue; file workers attempting to add to the task queue when it is already full are instead held in a second queue (also first-in-first-out); the file worker does not proceed with further spectral retrieval until the quantification task can be moved to the queue proper.

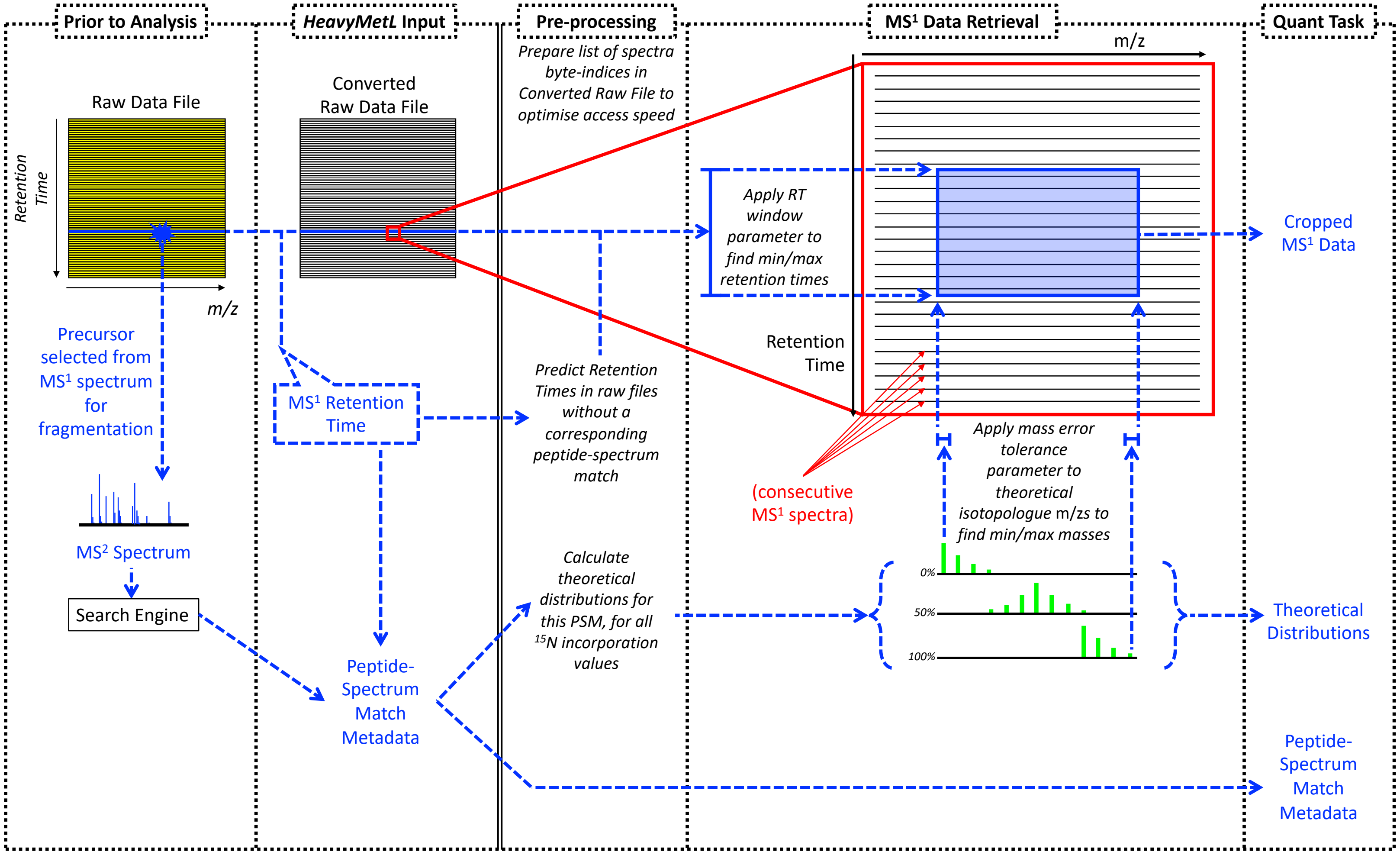

Prior to Analysis
HeavyMetL Input
Pre-processing
MS¹ Data Retrieval
Quant Task
Raw Data File
Retention Time
m/z
Converted Raw Data File
Prepare list of spectra byte-indices in Converted Raw File to optimise access speed
m/z
Apply RT window parameter to find min/max retention times
Cropped MS¹ Data
Retention Time
Precursor selected from MS¹ spectrum for fragmentation
MS¹ Retention Time
Predict Retention Times in raw files without a corresponding peptide-spectrum match
(consecutive MS¹ spectra)
Apply mass error tolerance parameter to theoretical isotopologue m/zs to find min/max masses
MS² Spectrum
Search Engine
Calculate theoretical distributions for this PSM, for all ¹⁵N incorporation values
0%
50%
100%
Theoretical Distributions
Peptide-Spectrum Match Metadata
Peptide-Spectrum Match Metadata

Figure 2-III. HeavyMetL *Pre-processing and Data Extraction. When quantification is started (rightwards from the double solid border after* 'HeavyMetL *Input'), the 'Pre-processing' section is handled by the main thread while 'MS*^{1} *Data Retrieval' is handled by the file workers.*

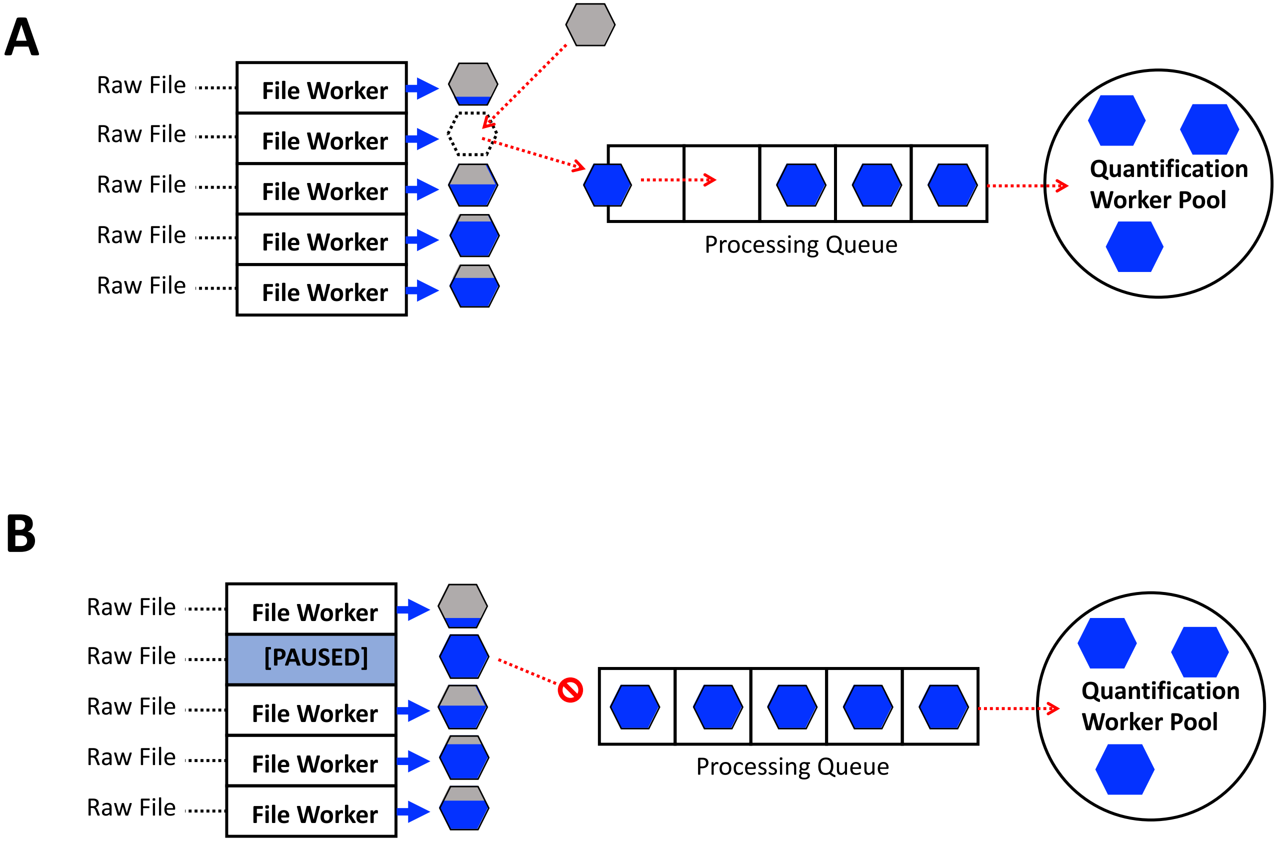


Figure 2-IV. *The Buffered Scheduling Queue.* ***A:*** *If there are remaining slots in the task queue, quantification tasks (blue hexagons) can wait for a worker in the quantification pool to become free while the file worker that created them is free to continue with further extraction work.* ***B:*** *if the task queue is full, both quantification task and file worker must wait for a free slot in the queue before the file worker can proceed with further extractions, this allows flexibility while preventing the file worker from filling up memory with prepared quantification tasks.*

The quantification workers, on receiving a task, perform a series of steps for each possible theoretical distribution, for each labelling state; first for the unlabelled signal (a single theoretical distribution), then for the labelled signal (a range of theoretical distributions).

In each cropped MS^1 spectrum the signal intensity is summed within a mass error window (default 10 ppm) around each isotopologue. If this is the labelled signal, it is possible that some of the unlabelled signal isotopologues may have *m/z* values very close to those of the theoretical labelled signal isotopologues for low ^{15}N incorporation values. To prevent the matching algorithm accidentally matching the isotopologue 'tail' of the unlabelled distribution, the mass range of the matched unlabelled isotopologues (step 2) is excluded from further matching. This avoids a mis-matching of the labelled distribution to parts of the unlabelled signal. The search window for the labelled signal is also restricted to a smaller retention window around the recognized unlabelled apex (the maximum unlabelled/labelled apex RT difference, by default 12 s).

Next, the extracted spectra are matched to the theoretical distributions. The first approach I tried at this point was to use every spectrum in the retention time window, and for each spectrum to try matching without one of the isotopologue masses ('leave-one-out') on the assumption that the other isotopologues were unaffected. This approach generated results with a number of problems. The algorithm frequently returned a match that was not part of the elution peak of the heavy signal but, instead, was a noise or interference mismatch at one or other extreme of the retention time window. Additionally, for lower ^{15}N incorporation levels where there are many isotopologues to be monitored, an interfering co-incident peptide *m/z* will be part of an isotopologue distribution that will also interfere with many other isotopologues in the matched distribution. The leave-one-out approach scales poorly to leave-many-out since at this point one is discarding most of the isotopologue distribution and matching to theoretical spectra with only a few data points.

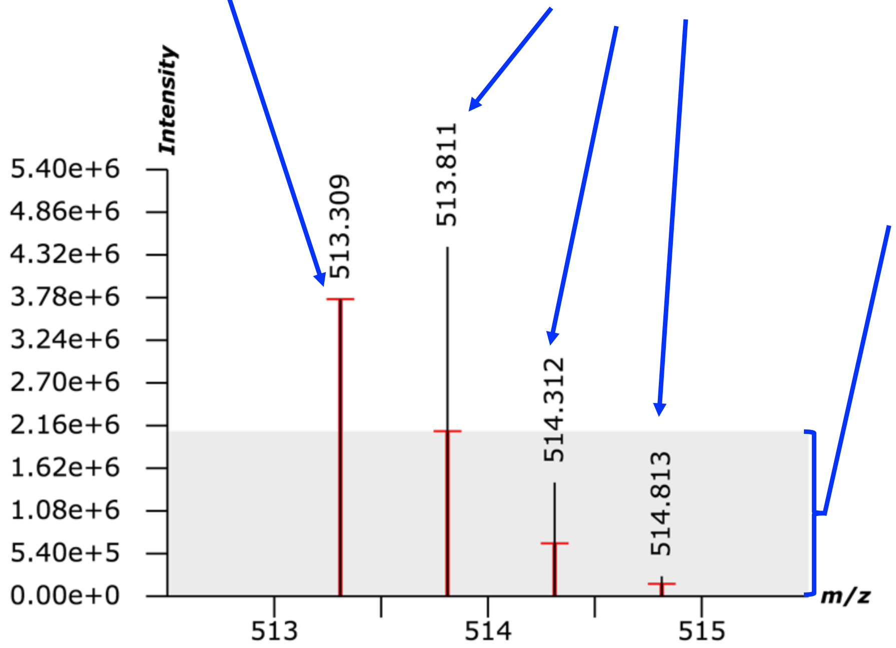
I found that a better solution was to ignore the lowest theoretical intensity isotopologues, ranking highest to lowest intensity and discarding any past a certain percentage (default 70%) of cumulative total intensity. I then generated a 'scaled spectrum' by finding the isotopologue with the smallest ratio between observed intensity and predicted proportion, and scaled up the theoretical distribution by this ratio. This produces a spectrum with the same proportions as the theoretical distribution where at least one isotopologue is the same intensity as in the observed spectrum, while the other isotopologues may be more intense but they can never be less intense (following the assumption that co-eluting interfering peaks may add to the intensity of some isotopologues but will never subtract; Figure 2-V). In essence, this is an estimate of the minimum possible intensity associated with the labelling state currently under consideration (assuming the current theoretical distribution is the right one) at all points over the retention time window. The total intensities of the scaled spectra are used to locate chromatographic intensity maxima within the retention time range, but not for actual matching to the theoretical spectra. Instead, having located maxima, the corresponding non-scaled spectrum at each local maximum, and one neighbouring spectrum on each side are summed (to enhance the signal-to-noise ratio) to give a 'signal-enhanced' maximum spectrum, and these 'signal-enhanced' spectra are matched against the current theoretical distribution. This ensures that matching is performed using all isotopologues, but only at the points during elution when the signal was highest. Under the assumption that the noise specific to single isotopologues will be primarily additive (i.e. overlap of co-eluting signals) rather than multiplicative (e.g. variable ionisation efficiency), then this strategy aims to minimise such noise by using the least affected isotopologue to locate the point at which the true signal is strongest.

The actual difference between the 'signal-enhanced' maximum spectrum and the corresponding theoretical distribution is measured by a Similarity Score based on the Kullback–Leibler Divergence between the observed and theoretical isotopologue distributions (194). The choice of Similarity Score was a key element in the development of the *HeavyMetL* algorithm; the score used was selected from among a number of potential candidates based on experimental results; this work is discussed in detail in Chapter 3.

Across all compared theoretical distributions at all local maxima, the highest scoring maximum-theoretical distribution pair is taken to be the elution peak apex for the current label state. The reported incorporation level and intensity for this unlabelled or labelled apex then defines the incorporation level of the matched theoretical distribution, and the total intensity of the scaled maximum spectrum (without neighbour summing) respectively (Figure 2-VI). The unlabelled and labelled peak apexes are located independently, which ensures the only assumption made regarding the degree of co-elution of labelled and unlabelled peptides is the parameter defining a maximum allowed difference between the two apexes.

Scaling isotopologue (black line and red line same height) – has lowest ratio to corresponding theoretical proportion

Observed intensity higher than expected for other isotopologues suggests co-eluting noise.



Isotopologues that do not contribute a substantial amount to total intensity (below shaded area) are not considered for the scaling isotopologue, so are not used to locate chromatographic maxima, or for fitting the final scaled spectrum (but are used for incorporation estimation)

Figure 2-V. *Fitting a Scaled Theoretical Spectrum to Observed Data. The use of a scaled spectrum reduces the effects of co-eluting co-incident mass species.*

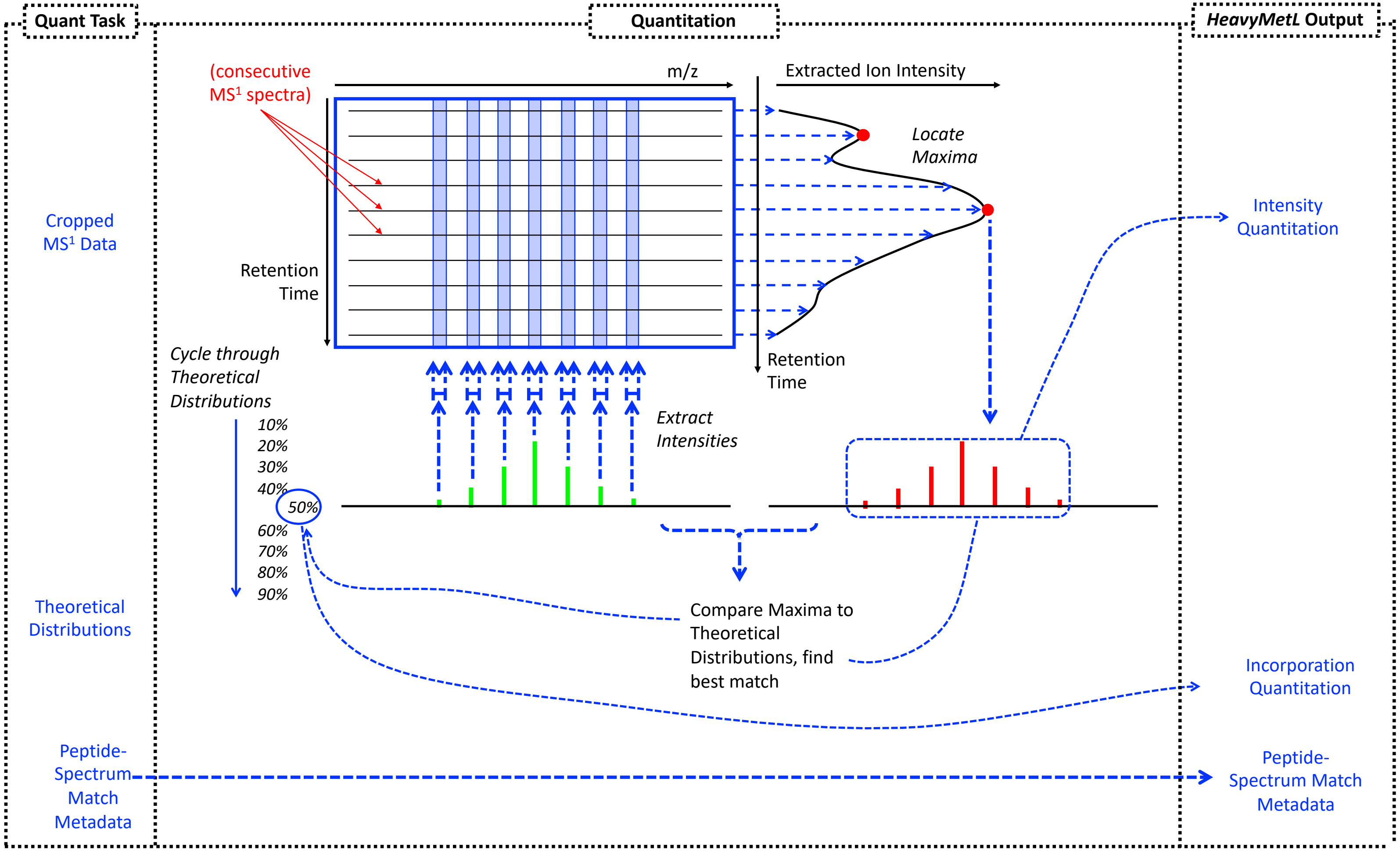

Quant Task
Quantitation
HeavyMetL Output
Cropped MS¹ Data
Theoretical Distributions
Peptide-Spectrum Match Metadata
(consecutive MS¹ spectra)
m/z
Retention Time
Extracted Ion Intensity
Locate Maxima
Retention Time
Cycle through Theoretical Distributions
10%
20%
30%
40%
50%
60%
70%
80%
90%
Extract Intensities
Compare Maxima to Theoretical Distributions, find best match
Intensity Quantitation
Incorporation Quantitation
Peptide-Spectrum Match Metadata

Figure 2-VI. *HeavyMetL Quantification. The 'Quantification' section is handled by quantification workers, then the data are passed back to the main thread for collation and output.*

Protein-level results are calculated based on only those PMCs flagged as ‘contributing’ to their parent protein in the input file (see Table 2-II). In each file, all the PMCs contributing to each protein are collated. For each protein, the ratio reported is the median ratio of contributing PMCs, and likewise the label incorporation level reported is the median label incorporation percentage of contributing PMCs. Taking the median values rather than the mean avoids skewed quantification in the case of severe mis-quantification of a single PMC.

The total run time for quantification of all PMCs mainly depends on computer hardware configuration, dataset size, quantification parameters and processor and memory pressure from running processes during analysis (there will also be some differences in performance due to the choice of web browser). Since files are processed in parallel, analysis time does not grow linearly with the number of raw files, but in general an analysis of two conditions in triplicate (six raw data files) on relatively modern hardware will generally be completed in about an hour or less. For a ‘real world’ example, see Section 4.2.2 below.

To further examine the relationship between dataset size and analysis time, I analysed an increasing number of clones of a single 1.5 GB raw file (the ‘unlabelled’ *Ostreococcus tauri* sample from Chapter 3; see Section 3.4) against a list of 1042 unique, high-confidence PMCs (for details, again see Section 3.4), and timed how long quantification processing took with default settings (this does not include the time taken to pre-index the files, typically less than a minute).These runs were conducted on an Apple MacBook Pro running macOS 10.14.6; 2.7 GHz Intel Core i7 with 8 logical processors; 16 GB RAM, in Firefox v. 69.0.1. The results are shown in Figure 2-VII. The data highlight the effectiveness of the measures taken to avoid duplication of effort for multiple files; although calculation of many theoretical spectra for a number of PMCs is computationally expensive, the additional overhead from processing additional files is minimal. The processing time increases linearly with file number, taking approximately 2 extra minutes for each additional file. Despite the MacBook having only 16 GB RAM, 30 raw files (~45 GB data) were processed without issue.

This analysis is admittedly somewhat artificial. The single raw file that was ‘cloned’ (to ensure differences in processing time were purely due to the number of files analysed) for this analysis comes from a genuine dataset. In truly ‘real’ datasets, however, the files will be more diverse, and file-to-file differences may affect processing time. Such issues are not considered by this analysis. For an alternative example of timing on a ‘real’ dataset, see Section 4.2.2 below.

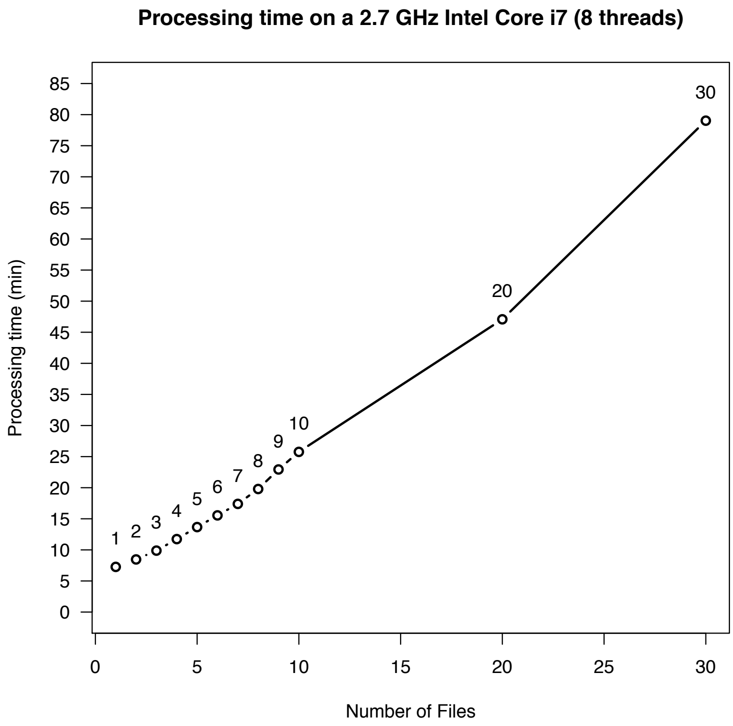


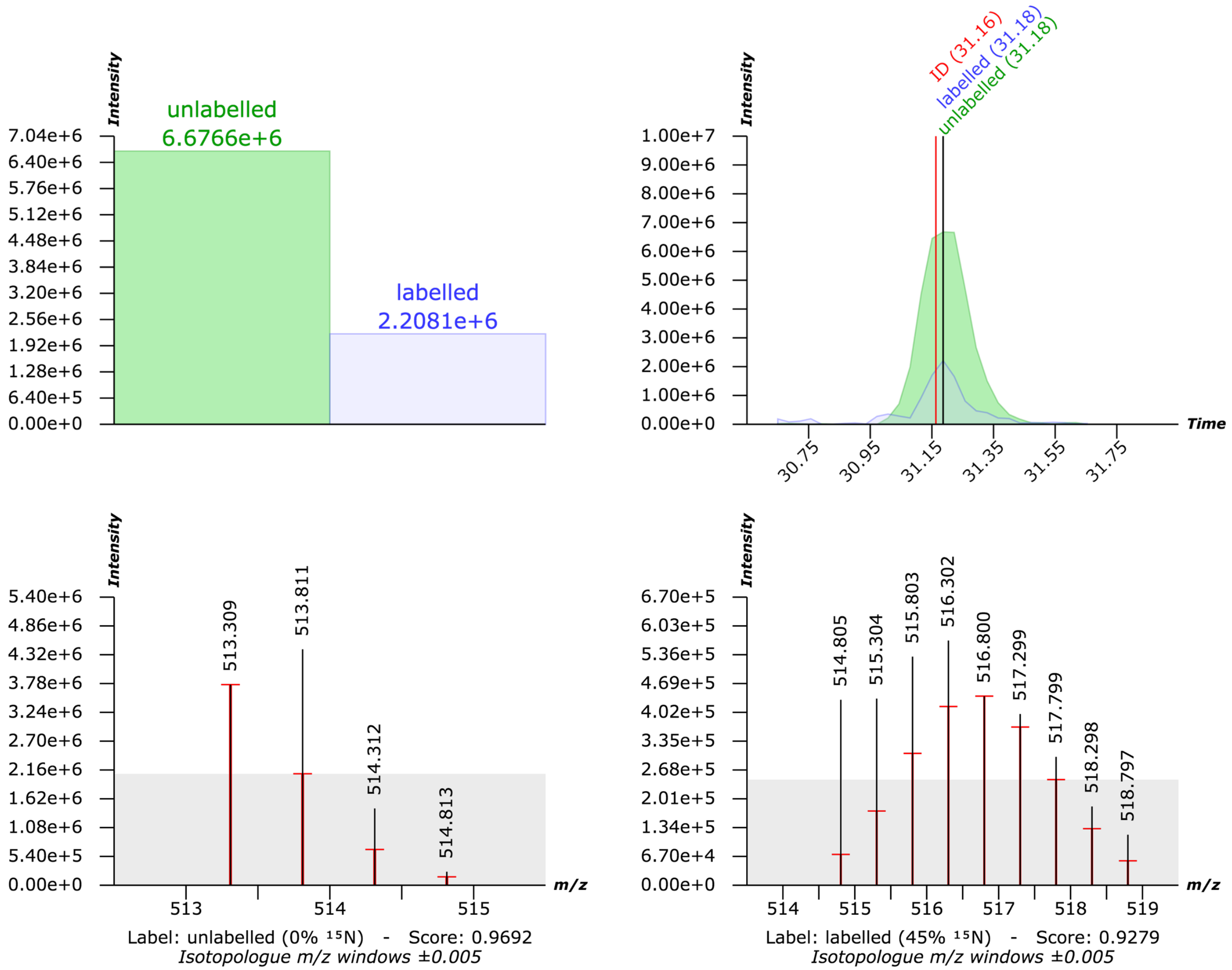
Figure 2-VII. *Processing Time versus Number of Files. Various numbers of clones of a raw EML MS data file (see text for details) were analysed by* HeavyMetL *(x-axis) and the time taken to complete processing was measured (y-axis; times in minutes). The number of files at each point is also shown above the point.*

2.4.5 Result Display

Graphics are drawn dynamically using the canvas HTML element. A third-party JavaScript library (fabric.js; http://fabricjs.com/) was used to abstract much of the underlying complexity for ease of use. Overview figures are displayed for both single PMC level quantification and protein level quantification summary according to which row in the protein/PMC table is selected. The PMC level graphic shows (Figure 2-VIII L-R top row) the relative quantified intensity of the unlabelled and labelled signals and the corresponding extracted intensity chromatograms and matched apex spectra (Figure 2-VIII bottom row). The protein level graphic shows a scatter plot for the incorporation level and unlabelled:labelled ratio across all files in the analysis. The median value taken as the protein-level statistic is shown as a black diamond while the first few letters of the sequence of individual PMCs are shown in light grey to illustrate the distribution of results (Figure 2-IX). The design of both overview displays is to highlight when quantification has produced a good or poor result, assessment of which is not easily reduced to a single ‘quality value’.

Figure 2-VIII. *Graphical display of Results at the PMC-Level (Next Page). The PMC identity (sequence, modification state, charge) is shown above the figures.* ***Top Row:*** *Relative label intensity (**left**, showing each sample as a bar on the x-axis* vs. *intensity on the y-axis) and extracted chromatogram (**right**, with retention time on the x-axis* vs. *intensity on the y-axis), showing the time of each label peak maximum finally selected by HeavyMetL. The red line indicates the time of the MS^2 scan that led to PMC identification (when applicable).* ***Bottom Row:*** *The peak maximum spectrum used for quantification of the unlabelled (**left**) and labelled (**right**) signal, showing the m/z value on the x-axis* vs. *intensity on the y-axis. Red lines indicate the distribution of the scaled fitted theoretical distribution and where the observed intensity was greater or less than expected. The width of the red lines (ignoring the T-piece at the top, which just serves to highlight the end of the line) indicate the m/z extraction windows based on the user-configurable ppm error. Only isotopologues equal or greater in height than the grey shaded areas are used for spectrum scaling and chromatographic maxima detection (see also Figure 2-V).*

IGGIGTVPVGR (unmodified), 2+



P25858

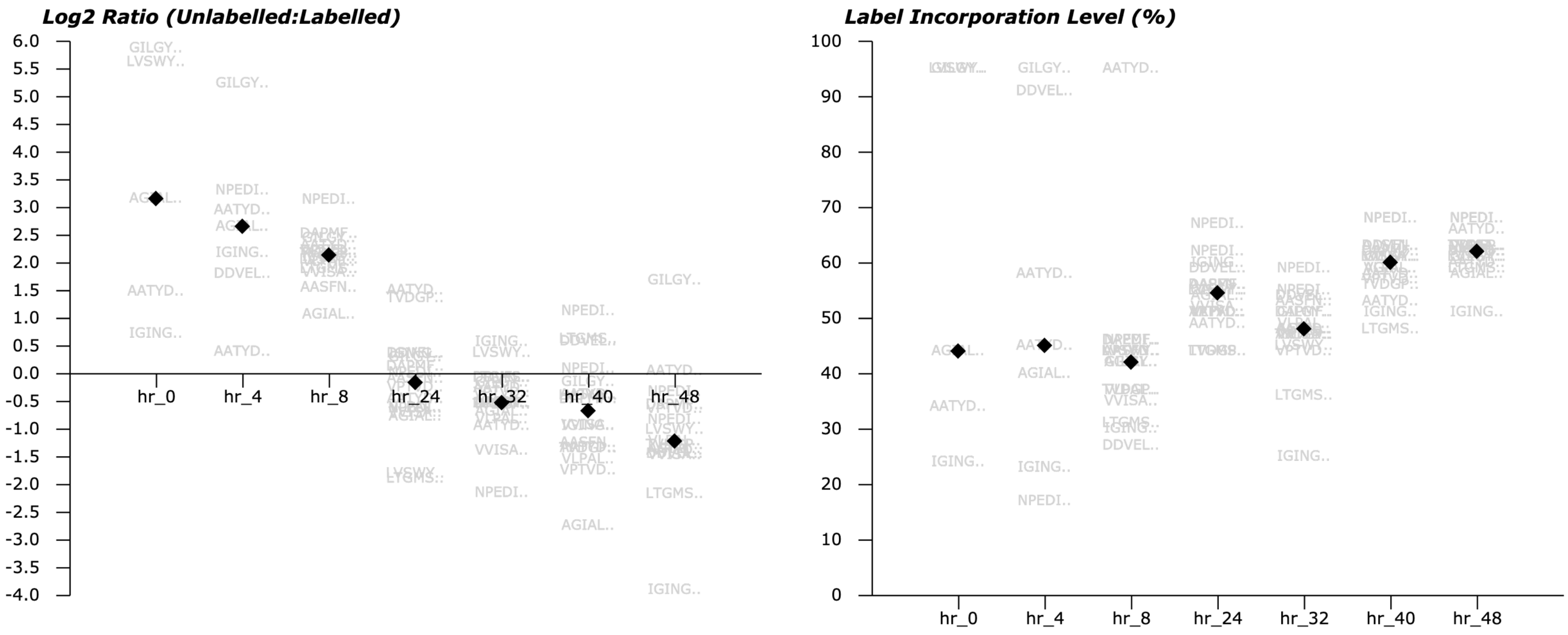


Figure 2-IX. *Graphical display of Results at the Protein-Level.* ***Left side:*** *Protein Log2 Ratio (Unlabelled/Labelled) indicated by black diamonds.* ***Right side:*** *Label Incorporation Level (%) indicated by black diamonds. In both figures, protein values are taken as the median of all contributing peptides - these are shown in light grey to illustrate the spread of the data.*

2.5 Conclusions

The *HeavyMetL* quantification approach is similar to the strategy employed by other 'theoretical-distribution matching' approaches, particularly *Protein TurnStILE*, but is substantially different from the 'ratio-only' approach (e.g. *Protover*) or the hybrid strategy employed in *ProteinTurnover*. As described throughout this chapter, in addition to interface usability and robust quantification, specific consideration was given to algorithm memory usage and speed

While the choice of a browser platform resolved the question of installation difficulty and facilitated implementation of a 'clean' interface that is consistent across operating systems, this choice also applied limitations in terms of speed and, especially, memory. A number of careful optimisations such as the task-scheduling queue system were implemented to minimise memory usage, but as the number of raw data files in an analysis increases, memory pressure will inevitably grow. The memory optimisations ensure this is unlikely to be relevant for the analysis of a modestly sized dataset in the case of an average researcher. High-throughput proteomic specialists, however, who frequently analyse 200+ raw data files simultaneously, may run into browser-imposed limitations (either on memory space or number of concurrent WebWorker threads) when attempting to perform similarly scaled experiments in *HeavyMetL*. These limitations may usually be avoided by configuration of browser internal settings (typically accessed *via* the about:config address), but precisely how these limitations may be overridden is subject to frequent change in both Mozilla Firefox and Google Chrome. In some cases, it may be necessary to break down the analysis in order to limit the number of simultaneously analysed files.

Nonetheless it is important to stress that this issue is not unique to *HeavyMetL*. Large scale analyses in proteomic software packages (and indeed analyses of very large data files in general) frequently require some re-configuration of the program to increase memory limits. Java-based programs, for example, often require that the memory space assigned to the Java runtime at program start (the 'heap size') is manually re-configured for large data sets. Ultimately, assigning a substantial proportion of system resources to any program will impact the performance of other programs and system responsiveness. Most datasets will be modestly sized, and it is reasonable that for general use, a quantification package should be expected to ‘play nice’ with other programs. If the user wishes to analyse a very large dataset or improve processing speed at a cost to other running processes on the computer, a requirement for specific manual assignation of extra resources (e.g. memory, CPU time *etc*.) is a sensible precaution, ensuring that standard operation of the program does not degrade system performance. In this regard, *HeavyMetL* is no different to other proteomic analysis platforms.

It is reasonable to conclude that *HeavyMetL* is a successful implementation of a solution to the issues described in Section 1.2.3 and further formalised as requirements in Section 2.2.2. Analysis of a typical modestly-sized dataset is acceptably fast on modern computers, provides reasonable analysis times, and the in-browser nature of the interface makes access and use of the tool straightforward.

Chapter 3: A Comparison of Spectral Similarity Assessment Methods

3.1 Introduction

The *HeavyMetL* algorithm described in Chapter 2 relies on a measure of spectral similarity in order to identify the theoretical distribution that best matches the isotopologue pattern observed in the raw data, a process I will henceforth refer to as ‘incorporation assessment’. Spectral similarity comparison methods are often formulated in terms of a measure of ‘distance’ between a target entity and one or more of comparison candidates– the most similar comparison is the one with the smallest distance. Alternatively, similarity may be formulated in terms of the likelihood that two entities are the same, in which case the goal is to find the comparison with the highest likelihood. Examples of both approaches are common in regression analysis (fitting a statistical model to a data set); common methods include ‘Least Squares’ (minimising the sum of squared differences) and ‘Maximum Likelihood’ (maximising a likelihood function). From a numerical optimisation view both comparisons are equivalent problems, for example maximum likelihood may be, and often is, computed by finding the lowest negative (log-)likelihood. As well as different ‘optimum’ values in typical formulation, comparison methods may have different ranges (e.g. 0 to 1, 0 to infinity *etc.*). For consistency throughout this chapter I will discuss spectral comparison in terms of spectral Similarity Score (SS) with a range of 0 to 1, with 1 being the optimum value achieved with perfect similarity. The various comparison methods discussed will be transformed to a corresponding Similarity Score when necessary.

In the context of the *HeavyMetL* algorithm, incorporation assessment involves creating a series of theoretical isotopologue distributions representing the peptide molecule with a range of ^{15}N incorporation percentages (henceforth the theoretical distribution, “*T*”, set). The predicted isotopologue *m/z* values of each *T* distribution are used to retrieve corresponding intensities from the observed data within an RT window. Specifically, *m/z*-intensity data distributions at each local total intensity maximum form the observed distribution, the "*O*", set, for that maximum. Each theoretical distribution of *T* is thus compared against multiple corresponding *O* extractions.

Spectral comparison may be symmetrical or one-way. To calculate a Similarity Score, one may compare the relative intensities of all *m/z* values observed in either spectrum (symmetrical), or just the intersection (symmetrical), or consider one spectrum to define the valid *m/z* values for comparison and ignore non-matching *m/z* values in the other spectrum. For symmetrical comparison, the ordering of the two spectra (*T vs. O* or *O vs. T*) is irrelevant. In the case of ^{15}N distribution matching, a one-way comparison is more appropriate since additional *m/z* values detected in the observed spectrum should not detract from match quality as they may originate from a co-eluting species or background noise. Failure to observe intensity at *m/z* values in the observed spectrum that are predicted to be present in the theoretical distribution, however, is an indication of less than perfect similarity. Furthermore, in the case of *HeavyMetL*, the theoretical distribution *m/z* values are the reference used for extraction of the *m/z*-intensity pair data from the observed spectrum, so it is most appropriate that the theoretical spectrum defines which *m/z* values are to be compared using the chosen SS.

A number of mass spectrometry spectral similarity scoring methods have been defined in the literature; most successfully applied previous methods were compared by Toprak *et al* (195), in the context of fragment ion spectral comparisons for quality assessment in PRM and SWATH-style DIA analyses. Any spectral similarity measure will have advantages and disadvantages in the context of a particular MS application as the sources and distribution of interference from background noise, overlapping signals and transformation artefacts will differ. While, *prima facie*, matching an observed fragment ion spectrum to a theoretical distribution *T* or previously acquired spectral library entry is similar to matching an observed precursor isotopologue distribution to a predicted isotopologue distribution (as herein), in practice there are many differences (Table 3-I).

Spectral matching of MS^2 fragment ion spectra heavily penalises large differences between *O-T/L* pairs, particularly with regard to absence of high intensity peaks or unexpectedly high intensity peaks predicted/previously observed at low intensity, as these are the best indications of an incorrect match. They are not assessed in the context of finding the ’best’ match from a number of very similar comparisons, all of which are derived from the ‘correct’ PMC definition but at incremental percentage steps (e.g. 48%, 49%, 50% *etc*.) of ^{15}N around the optimum match. A ‘good’ SS in our case should ideally have a clear maximum at the correct ^{15}N incorporation (Figure 3-I) so that minor effects of noise do not substantially change the ‘best’ match.

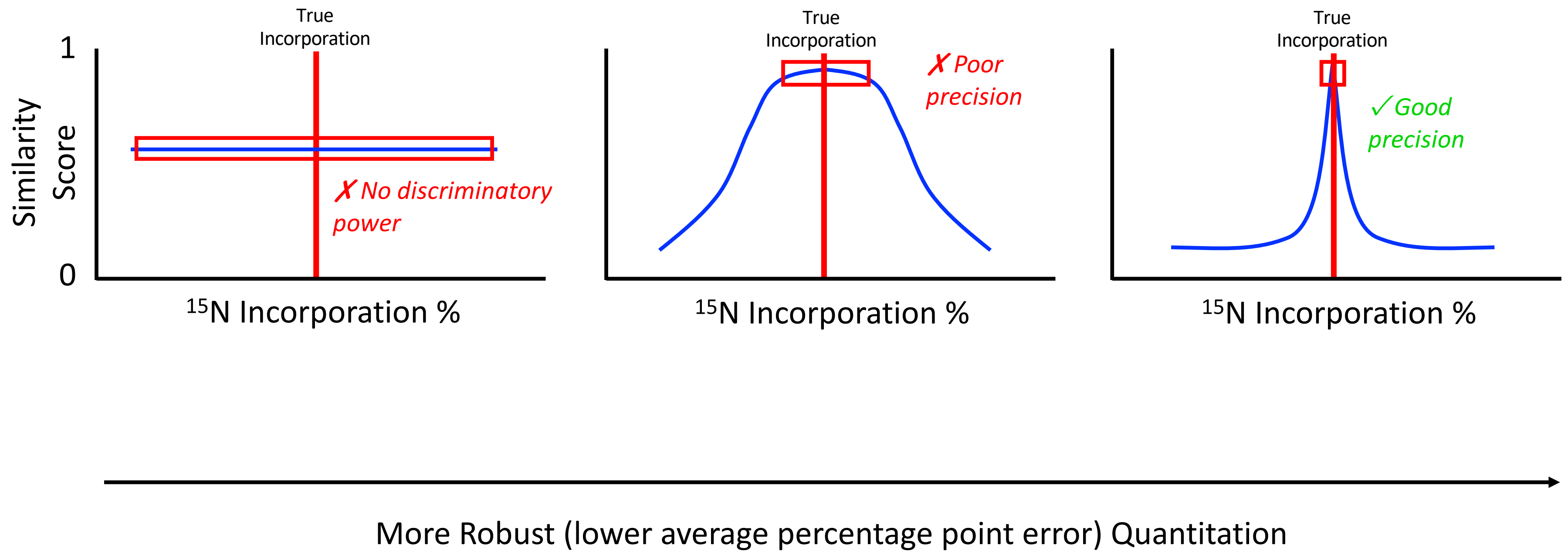


Figure 3-I. *Idealised Behaviour of a Similarity Score. For accurate and sensitive assessment of incorporation level the score should be maximal when and only when the spectrum is compared to a theoretical spectrum at the correct incorporation level. The smaller the relative increase in score when matching correctly, the greater the opportunity for noisy spectra to be mis-assessed, and the greater likelihood that mis-assessed incorporations will be further from the correct value (ranges represented by the red boxes).*

Context	Contemporary Examples	Observed Data	Comparator	Highly weighted elements	Considerations	Number of ions in observed data	Proportional impact of each ion-ion comparison
Peptide-Spectrum Matching	Mascot (Pappin *et al.,* 1993) Andromeda (Cox *et al.* , 2011)	Fragment Ion Spectrum	Predicted Fragment Ion Series	Presence/absence of ion masses corresponding to predicted fragment masses	The relative proportions of predicted fragment masses are difficult to predict and thus relative intensities of ions are generally not considered, beyond applying a background noise cutoff	More than 20	Low
Peptide-Spectrum Matching	X! Hunter (Craig *et al.,* 2006) BiblioSpec (Frewen *et al.* , 2006) SpectraST (Lam *et al.,* 2007)	Fragment Ion Spectrum	Fragment Ion Spectrum from Spectral Library	Presence/absence and relative intensities of observed ion masses corresponding to masses in spectral library spectrum	There are likely to be low intensity noise (co-selected precursor or background) ions in both spectra which should be ignored or given very low weighting	More than 20	Low
PRM/SWATH Quality Assessment	Skyline (MacLean *et al.* , 2010) See also De Graaf *et al* ., 2011 and Toprak *et al* ., 2014	Fragment Ion Spectrum Subset	Fragment Ion Spectrum from Spectral Library	Relative intensities of observed ion masses corresponding to masses in spectral library spectrum	The relative intensities of observed ion masses should be very close to the spectral library spectrum, so large differences should be heavily penalized	Less than 20	High
MS1 Quantitation / Feature Detection	MaxQuant (Cox *et al.* , 2008) Progenesis QI for Proteomics (commercial; Nonlinear Dynamics, Newcastle upon Tyne, UK) Proteome Discoverer (commercial; Thermo Fisher Scientific, Waltham, MA, USA)	MS1 Precusor Isotopologue Ion Distribution	Predicted Isotopologue Ion Distribution	Relative proportions of observed ions corresponding to predicted ion proportions	Low proportional intensity ions are more affected by noise and should be given less weight	Less than 10	High

Table 3-I. *Considerations of Spectral Distance Functions in Proteomic MS Contexts.*

For *HeavyMetL*, the SS must value the proportional differences between precursor isotopologue ions very highly, in order to be able to determine the subtle changes induced by fractional increases in ^{15}N incorporation, and to reject cases where the isotopologue distribution is highly contaminated by interference. In practice, the approach will have to cope with varying levels of background noise, levels of co-eluting interference, mass accuracy, effective mass resolution and intensity measurement precision. It is impractical to model these various parameters to create a theoretical dataset with any confidence of real-world applicability. While various comparisons between spectral similarity calculations have previously been made in other proteomic MS contexts (195, 196), the criteria on which they are judged were primarily designed to assess match plausibility, rather than picking a 'best' match out of a series of theoretical spectra, with a substantial proportion of very similar candidates corresponding to ^{15}N incorporation levels close to the true value. Using these previous approaches and earlier ^{15}N quantification work as a guide, I selected several representative options that seemed likely to be appropriate for the *HeavyMetL* algorithm and compared them *via* an experimental approach.

3.2 Spectral Similarity Scores

Given a set of mass values from a theoretical distribution (assuming a particular ^{15}N incorporation level) *T*, each SS described below compares an observed distribution of intensities across those *m/z* values in *O,* and the predicted intensities of those *m/z* values in *T*. These are all one-way comparisons, predicated on the *m/z* values in *T*. Ion intensity is considered as a fraction of the total intensity (or sometimes, the most intense ion) so as to standardise differences in signal strength. Let *n* be the number of isotopologues with a non-negligible fraction of total intensity in *T*. Let T_i be the fractional intensity of the *i*th-isotopologue in *T*, and O_i the fractional intensity of the *i*th-isotopologue in *O*.

Previous 'theoretical-distribution-matching' ^{15}N quantification approaches (see Introduction, Section 1.2.2) have relied on various formulations of a 'Least Squares' approach, which in essence is the minimisation of a spectral distance value based on the Euclidean Distance (between the fractional intensities of each T_i <-> O_i ion pairing (see Equation 1, below). An obvious modification to this method to emphasise the role of high-intensity ions is to weight the contribution of each squared difference by the intensity of the observed ion (Equation 2).

$$\mathrm{SS_{EUC}} = 1 - \sqrt{\sum_{i=1}^{n}(O_i - T_i)^2} \qquad \text{(Equation 1)}$$

$$\mathrm{SS_{WEUC}} = 1 - \sqrt{\sum_{i=1}^{n} O_i(O_i - T_i)^2} \qquad \text{(Equation 2)}$$

The work of Toprak *et al.* (195) in assessing distance calculations for use when comparing fragment ion spectra suggests that angle-based (dot-product type) approaches perform well for analysis of MS^2 data, particularly a normalised version of the Spectral Contrast Angle (197). This is a measure specifically designed to be sensitive to differences in relative ion intensity and thus is an attractive option for this work. The formulation given in Toprak *et al.* (Equation 3) is already normalised to the (0,1) range.

$$\mathrm{SS_{SCA}} = 1 - \frac{2\cos^{-1}\sum_{i=1}^{n} O_i T_i}{\pi} \qquad \text{(Equation 3)}$$

The hybrid ^{15}N quantification approach of Fan *et al.*, while not making use of a spectral distance calculation directly, suggests a further approach. Their algorithm matches the composite of the unlabelled and ^{15}N labelled distributions simultaneously using a maximum likelihood approach to locate the optimum composite distribution. This could also be re-formulated as minimising the Kullback-Leibler Divergence (194, 198). This measure is frequently used in information theory to represent the relative entropy between two probability distributions. In this case, it could be thought of as the information lost when a particular observed distribution *O* is used to approximate the theoretical distribution *T whose* *m/z* values were used to extract the isotopologue intensities in *O* (199). If *T* was close to the actual incorporation level of the signal, then the resulting *O* should be a good approximation of *T* with a low amount of information lost. The Kullback-Leibler Divergence itself ranges from 0 to infinity, so it is necessary to use a logistic transformation to yield a score in the desired (0,1) range (Equation 4).

$$\mathrm{SS_{KL}} = 2 - \frac{2}{1+e^{-\sum_{i=1}^{n} O_i \ln\frac{O_i}{T_i}}} \qquad \text{(Equation 4)}$$

Note that in Equation 4, where O_i is zero, the i-th term as a whole is also zero (the limit of x log(x) as x approaches zero is zero). T_i can never be zero as T only contains isotopologues with non-negligible fractional intensity.

The candidates considered are representative of three general approaches for assessing similarity; scalar distance (SS$_{\mathrm{EUC}}$/SS$_{\mathrm{WEUC}}$), vector angle (SS$_{\mathrm{SCA}}$) and information entropy

(SS_{KL}). A fourth category of correlation-based candidates were considered, such as Pearson's correlation coefficient (r) or the alternative nonparametric Spearman's correlation coefficient (rho). They were not considered in this case, based on Toprak *et al.*'s demonstration that measures of correlation performed poorly for spectral comparisons in general.

While transformation of the various comparisons to a consistent range does not guarantee that the scaling will be consistent within those ranges (one SS might tend to report values in the range 0.9 to 1 unless spectra were wildly different, while another might value the same set of spectra in the range 0.1 to 1), transformation ensures that the maximum and minimum values are always consistent which makes implementation easier (so, for example, the *HeavyMetL* code does not have to allow for an SS to be a negative value, or have infinite magnitude).

3.3 Comparison of Similarity Scores Using Real-World Data

I wished to compare SS performance to determine which produced the 'best' incorporation assessment results when *HeavyMetL* was run using each SS algorithm in turn. Such optimisation might be done using theoretically generated data, in this case generating theoretical distributions for a range of peptides at different ^{15}N incorporation levels then applying a noise function to simulate real-world interference. However, this would be biased by any assumptions of the noise function (e.g. maximum relative noise to signal, degree of interference across masses, degree of noise uniformity). The characteristics of spectral noise are not well defined (195), and this is particularly the case here, where the precursor isotopologue distribution to be matched is an isolated *m/z* range divorced from the context of the full MS^1 spectrum, given that noise is variable across typical proteomic MS *m/z* ranges of 0-2000 Th. I reasoned that a robust comparison could only be performed with real-world acquired data.

There is a downside to such an approach, when compared to theoretical data, in that in the latter case the exact 'true' incorporation is known. In a real-world context, even if cells are cultured to a target ^{15}N incorporation level, there are a number of challenges. Firstly, 100% pure ^{15}N salts (for the growth media) are impractically expensive; the purity of typically available salts is in the range of 98-99%. This means the true incorporation level of peptides will be slightly lower than the experimental target (even assuming the growth media has been mixed to a precise level of ^{15}N incorporation without pipetting error). Secondly, peptides in incompletely labelled samples have been shown to display considerable variance even when efforts have been made to ensure that labelling time was sufficient to ensure a stable label incorporation level (200). It is likely that much of this variance is technical in nature (lower intensity peptides experiencing higher signal-to-noise, higher mass peptides splitting their signal among more isotopologues giving more data points for incorporation estimation). The possibility of biological effects such as differences in parent protein synthesis rates resulting in different unlabelled/labelled elemental incorporation bias cannot be entirely discounted, however. It is therefore necessary to use a robust calculation of the calculated peptide incorporation distribution centroid as an estimator of the actual ^{15}N incorporation level. All of these challenges apply equally to any genuine ^{15}N labelling experiment; however, it is reasonable to assume that a SS which performs well under these conditions will (generally) perform well in real incorporation studies.

In collaboration with Dr. Sarah Martin at the University of Edinburgh, an experiment was designed to assess the performances of each the four SS measures. Dr Martin performed the sample preparation and MS analysis work (see Section 3.4, below). The data processing, quantification with *HeavyMetL* and subsequent analysis of the results are my own work.

Dr. Martin and I designed an experimental *Ostreococcus tauri* dataset consisting of an unlabelled sample and 3 labelled samples with different target levels of ^{15}N labelling at 40%, 50% and 60% ^{15}N (Figure 3-II), henceforth known as Samples A, B and C respectively. *O.tauri* is a green algae and one of the smallest (in physical size) known eukaryotes (201). It is frequently studied as a model organism for metabolic cycles such as circadian rhythm. It is relatively easy to culture and label by EML. Furthermore, proteomic MS time-courses using ^{15}N labelling have been extensively described (159, 160, 202).

The levels of labelling were selected such that the resulting precursor ion distributions would be maximally dissimilar to any unlabelled contaminating precursors that might originate from sample preparation (such as Human Keratins), providing the best assessment of spectral distance measure performance. As incorporation approaches unlabelled (isotopologue distributions at <20% ^{15}N) or fully labelled (isotopologue distributions at >80% ^{15}N) the precursor distributions would be very similar to the distribution of (unlabelled) contaminant precursors with the same mass, which could be selected instead by the matching algorithm (see Figure 1-XXII-D). The apparent performance at those incorporation levels could thus be artificially inflated by matches to contaminant peptides. A step range of 40, 50, 60% was an acceptable compromise to minimise this risk while examining performance over an incorporation range.

The PSM list for analysis was obtained by contemporaneously analysing an unlabelled *O. tauri* sample with the same sample preparation step and LC gradient conditions, so that features could be matched across runs by RT. There were three reasons for this approach. Firstly, this avoided differences in PSM identification from different samples, so that the same PMC list was analysed in each case. Secondly, incomplete EML labelling of peptides substantially increases the complexity of peptide-spectrum matching (even in cases where the actual incorporation level is known in advance, which is only possible in the case of complete labelling with very high purity ^{15}N sources) and produces fewer identification results than searching unlabelled data. Thirdly, the expected usage scenarios for *HeavyMetL* involve either samples containing only labelled peptides (at unknown incorporation) with a contemporaneously analysed unlabelled PMC identity reference sample as is the case here, or the presence of fully unlabelled peptides in each of the samples to be analysed (a typical abundance-comparison design). The latter design would likely produce more easily quantified data; the unlabelled signal RT apex is easier to locate (there is only a single possible incorporation level to be tested), and the presence of an unlabelled signal RT apex in-run may be used to narrow the search window for the labelled signal apex. It is necessary, however, for the Similarity Score to perform well in both scenarios. I chose to refine my selection of Similarity Score on the more challenging experimental scenario to hopefully accentuate any performance differences between the candidates.

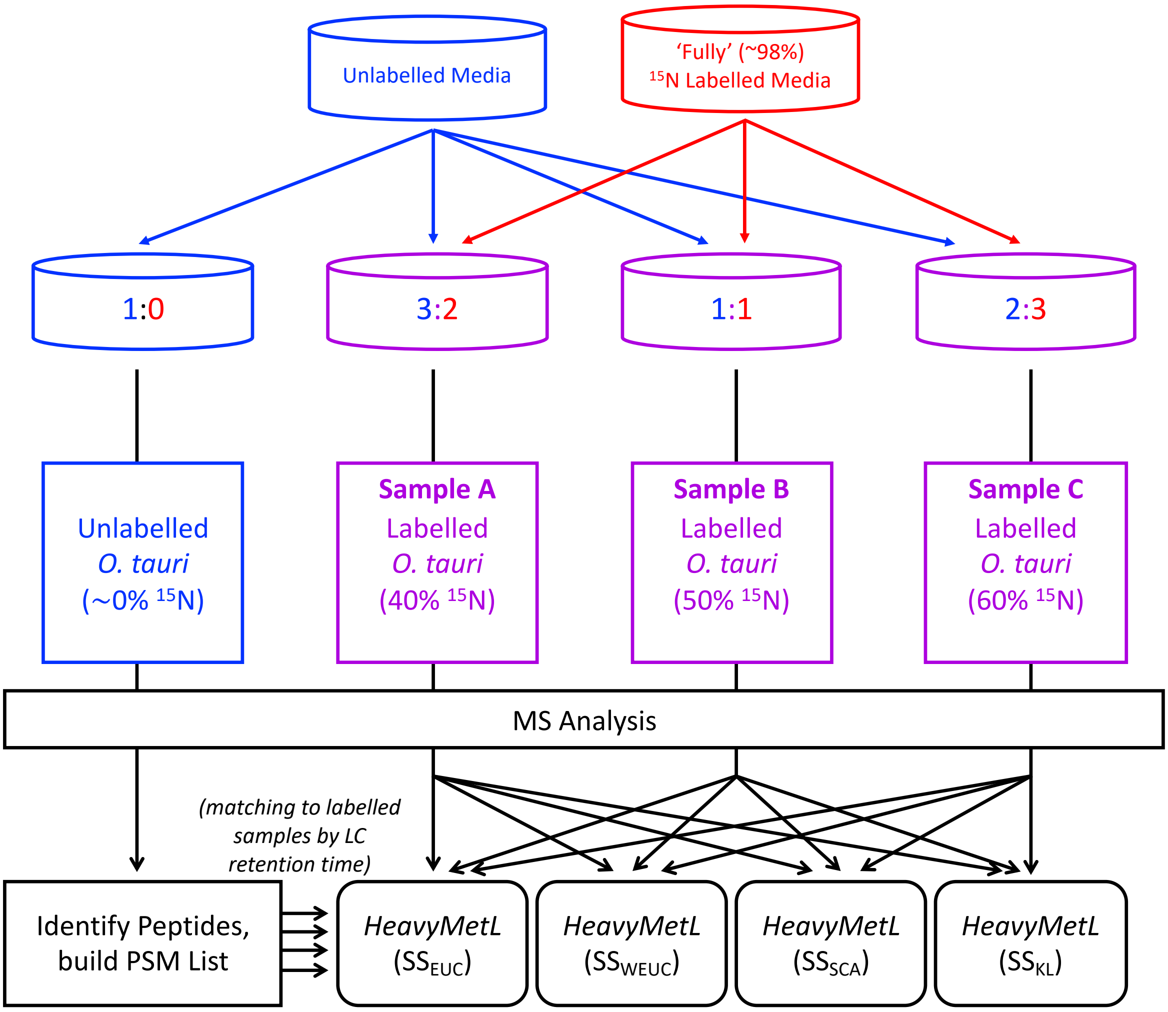


Figure 3-II. *Experimental Design for Similarity Score Assessment. The unlabelled sample (in blue) was not used directly in the analysis of ^{15}N incorporations but rather to generate the list of peptides for quantification (since direct identification of ^{15}N labelled peptides is not robust unless the incorporation level is 100%).*

3.4 Experimental Dataset Methods

N.B. The Subsections 3.4.1 to 3.4.3 (inclusive) of this methods section are included for information but do not represent work undertaken by myself. They were performed by Dr. Sarah Martin, a collaborator at the University of Edinburgh. All data analysis subsequent to MS acquisition was performed by me, including post-processing (Subsection 3.4.4).

3.4.1 Cell Culture

Ostreococcus tauri OTTH059543 were cultured in 0.22 μm filter sterilized artificial sea water (Instant Ocean powder) at a salinity of 30 parts per thousand as described in Le Bihan *et al.*, 2011 (159). Briefly, cultures were split weekly to 1 part in 50 to ensure continuous growth. In preparation for the experiment, cultures were passaged twice 1:50 into media containing a mix of ^{14}N sodium nitrate and ^{14}N ammonium chloride (both from Sigma Aldrich, U.K.) and ^{15}N sodium nitrate (98% pure) and ^{15}N ammonium chloride (99% pure) (both from Cambridge Isotope Laboratories) mixed in appropriate ratios to give final combinations with 40, 50 and 60% ^{15}N incorporation, and an additional unlabelled sample. Samples were cultured under a 12-hour daylight/ 12-hour darkness cycle at a constant 20 °C in a vertical environmental test chamber (MLR-350, Sanyo). A light intensity of 17.5 μEm^{2} s^{-1} was maintained using 724 Ocean Blue, Lee filter. Cells were grown for 8 days to an optical density of ~0.1 mm^{-1} at 600 nm in parallel with FACS (Fluorescence-Activated Cell Sorting) analysis (equivalent to approximately ~10 k cells per μL, or 700 μg protein per 100 mL). Whole cell lysate was sampled from each culture by centrifuging 30 mL culture (3200 g, 10 min) and washing pellets with 1 mL PBS before centrifuging again (12000 g, 5 min). Full pellet resuspension and cell lysis was achieved by pipetting up and down with 200 μL 2M urea. Samples were stored at 20 °C before digestion.

3.4.2 Sample Preparation

Samples were reduced with 12.5 μL each of 200 mM dithiothreithol and 1M ammonium bicarbonate for 30 min at room temperature. 12.5 μL of 500 mM iodoacetamide and 5 μg sequencing grade porcine trypsin (Roche, UK) were added for alkylation and digestion overnight. 10 μL digest were diluted in 20 μL buffer A (97.5% HPLC grade water, 2.5% HPLC grade acetonitrile (both Fisher, U.K.), 0.1% formic acid (Suprapure Merck, Germany), cleaned on Stagetips, eluted in 10 μL buffer B (90% acetonitrile, 10% water, 0.1% formic acid, 0.025% trifluoroacetic acid (sequencing grade, Sigma, U.K.)), vacuum-dried (RC 10-10, Thermo Fisher, U.K.) and stored at -20 °C.

3.4.3 MS Analysis

Dried samples were re-suspended in 11 μL buffer A and analysed on a capillary-HPLC-MS/MS system (1200 binary HPLC, Agilent, U.K., coupled to a hybrid LTQ-Orbitrap XL mass spectrometer, controlled by XCalibur v. 2.0.7, Thermo Fisher, U.K.) in 140 min. gradients. Capillary Picotip columns (10 cm x 360 μm o.d. x 75 μm i.d.) with a 15 mm tip opening and fitted with a borosilicate frit were purchased from New Objective (Presearch, UK). Fused-silica tubing was purchased from Composite Metal (UK). The reversed-phase bulk material used was 5mm Pursuit C18 obtained from Varian (UK).

With buffer A as 97.5% water, 2.5% acetonitrile, 0.1% formic acid, and buffer B as 90% acetonitrile, 10% water, 0.025% trifluoroacetic acid, 0.1% formic acid, the solvent gradient program was as follows: 0% buffer B (0–12 min.), 0–5% buffer B (12–16 min.), 5–15% buffer B (16–36 min.), 15–35% buffer B (36–80 min.), 35–100% buffer B (80–96 min.), followed by 100% buffer B for 18 min. and back to 0% buffer B for 6 min. Prior to the analysis, a column/pre-column wash and conditioning step was performed consisting of a 1 h gradient of 0–100% buffer B over 20 min. followed by an isocratic conditioning step at 0% buffer B over 40 min.

Data-dependent acquisition was performed with one profile-mode MS^1 scan at 60 k resolution in the Orbitrap followed by five MS^2 scans in the LTQ.

3.4.4 Post-Processing

The raw data were converted to mzXML format using the *msconvert* tool in the *ProteoWizard* suite (http://proteowizard.sourceforge.net/). MS^2 peak-lists in MGF format were also generated from all samples using the same tool. The unlabelled sample peak-list was searched using *Mascot Server* (v. 2.5.1, Matrix Science) with the following parameters: fixed modifications = carbamidomethyl (C); variable modifications =

oxidation (M), acetyl (N-term); mass tolerance = 10 ppm; fragment mass tolerance = 0.1 Da; max missed cleavages = 2; search database = *Ostreococcus tauri* UniProt reference proteome (retrieved 24/08/2016). The list of identified peptides was exported from *Mascot* in mzTab format and re-formatted for compatibility with *HeavyMetL*. Briefly, a new table was created based on the "PSM" rows in the mzTab file. The following columns were directly copied across (mzTab column names from the "PSH" row given first, *HeavyMetL* input names given second; see Table 2-II for details): "sequence" as "PEPTIDE_SEQUENCE", "accession" as "PROTEIN"; "search_engine_score[1]" as "PEPTIDE_SCORE"; "charge" as "CHARGE". The column "RETENTION_TIME" was added based on the mzTab "PSM" column "retention_time" divided by 60, the column "MODIFICATIONS" was added based on the mzTab PSM column "modifications" reformatted to Scaffold-style definitions, and the column "PROTEIN_DESCRIPTION" was added cross-referencing the protein accessions in "PROTEIN" to the "description" column in the mzTab "PRT" rows. Finally, the column FILE_NAME was added containing the file name of the unlabelled run, and the column "CONTRIBUTE_TO_PROTEIN" was added with a value set to TRUE for every row (irrelevant for analysis of this dataset as the data were considered at the peptide level only). Finally, the list was filtered to include only peptides with PEPTIDE_SCORE (*Mascot* Expect value) less than 0.001 to minimise the effects of mis-quantification due to incorrect sequences, yielding 1061 PMC entries, 1042 of which were unique. The list was analysed against all four raw data files (including the unlabelled sample for RT matching) four times with *HeavyMetL*, in which the SS was implemented as each of the four candidates shown in Equations 5-8 in turn. Default settings were used elsewhere, except for the 'Maximum Peak Apex Shift' parameter, which was set to the full extraction window of 30 s (samples A, B and C contained no unlabelled signals to be matched within-runs). This included the 'Do Not Quantify Unlabelled Signal' parameter being left as 'false" despite there being no unlabelled signal in samples A, B and C; it is not necessary to change this parameter to true unless there is both no unlabelled signal and an extremely low level of ^{15}N incorporation (<10%) is expected for the labelled signal. Quantifying any apparent unlabelled signal (e.g. from mis-matched co-eluting peptides) also prevents such signals being incorrectly matched as a labelled signal.

Subsequent statistical analysis of results and generation of all figures was performed in R (v. 3.5.1).

3.5 Estimation of Average Sample Incorporation Levels

While the labelled samples A, B and C were grown to fixed target incorporations, the 'true' peptide incorporations were expected to be a distribution around an *a priori* unknown median, rather than a fixed point, due to ^{15}N salt impurity, pipetting variance and, potentially, incorporation kinetics (see Section 3.3). Before analysing the data with *HeavyMetL*, I investigated a previously described method for estimating the average peptide ^{15}N incorporation level in each sample without using cross-run matching of peptide identifications, based on a previously observed relationship between peptide mass and peptide Mass Decimal Residual (i.e. the fractional part of the mass value, henceforth MDR). This phenomenon was first described by Mann (203) and subsequently expanded into the Half Decimal Place Rule (HDPR) (204, 205). Informally, the HDPR observes that the first digit of the MDR is near the half of the first digit of mass values between 500 and 999, near the half of the first two digits of mass values between 1000 and 1999, and near the half of the first digit of the mass values between 2000 to 3000, with various papers defining 'mass' as molar mass, molecular mass or $(M+H)^+$; the relationship is observable on any scale and I use molecular mass in Da hereafter. When plotting MDR *vs*. mass , a characteristic series of diagonal bands are observed corresponding to a banded 'wrapping' of the linear relationship across the (0,1) MDR scale (see Figure 3-III-A for an example).

HDPR has typically been used for quality control in MALDI-ToF analyses of peptide mass fingerprinting (identifying proteins based solely on peptide mass rather than using peptide fragmentation) studies (204, 205), although recently the approach has been applied to modern LC-MS/MS (206). Of particular interest to me was a 2010 study by Fetzer *et al.* in which the HDPR relationship was used to predict partial ^{13}C incorporation in bacterial peptides (207). The Fetzer *et al.* approach first involves transformation of the banded MDR:Mass relationship to a linear Corrected MDR:mass relationship by identifying points associated with the 'wrapped' bands and correcting their MDR by adding a +1, +2, *etc*. correction factor across the bands. The data are transformed by a scalar rotation such that the bands are vertical in the *y*-axis. Specifically, peptide masses are rescaled as

$$\text{Transformed Mass} = \text{Mass} - 1800 \cdot \text{MDR}$$

followed by k-means clustering to assign each point to a band. Using unlabelled and completely labelled standards, Fetzer *et al.* demonstrated that as the percentage of ^{13}C incorporation increases from ~1% to 100%, the gradient of the Corrected MDR:mass relationship also increases. Using unlabelled and fully labelled standard samples for calibration, they were able to predict ^{13}C incorporation based on the peptide masses of any similar sample.

I applied a similar approach to two mass datasets. I first developed an analysis workflow on using a (partly) theoretical dataset based on 1042 unique, high-confidence (*Mascot* Expect value <0.001) PMCs identified in the unlabelled sample. Using the *HeavyMetL* isotopologue prediction algorithm, I calculated the expected masses of the most intense isotopologue for each PMC at natural ^{15}N incorporation (~0.368%) and at 10-100% ^{15}N incorporation in step sizes of 10. This resulted in 10 sets of mass values between 500 and 3500 Da. There was one mass value greater than 3500 Da which would fall into a separate cluster, which was ignored to simplify analysis. I replicated the Fetzer analysis on these masses (and their corresponding MDRs), with some modifications. To avoid manually specifying a rotation scalar (chosen by Fezter *et al.* by eye) I instead transformed the data by Principal Component Analysis, which neatly separates the bands along one principal component (usually the first, although in the case of particularly noisy data it may be the second principal component) (Figure 3-III-B,C). Fetzer *et al.* then used k-means clustering to separate the rotated clusters. I found this approach to be insufficiently robust, frequently incorrectly splitting the clusters and that a more robust approach was to simply estimate a cluster separation threshold by density analysis of the principal component. I estimated the density along the axis with a cosine model to locate the density maxima, then took the half-way point between the maxima as a clustering cut-off (Figure 3-III-D,E,F). MDRs from the higher mass cluster were then corrected by adding +1 as in the Fetzer analysis (Figure 3-III-G) and I then calculated a gradient by linear fit for each ^{15}N% step (Figure 3-IV). As in the Fetzer work, I observed a strong linear relationship between ^{15}N incorporation and the Corrected MDR:mass gradient, although in the case of ^{15}N this correlation is negative while Fetzer *et al.* observed a positive correlation. On reflection, this is to be expected. The mass delta of ^{12}C to ^{13}C is ~1.0033 so increasing the percentage of ^{13}C will, on average, raise the MDR. In contrast the mass delta of ^{14}N to ^{15}N is ~0.9970 so increasing the percentage of ^{15}N will decrease the MDR.

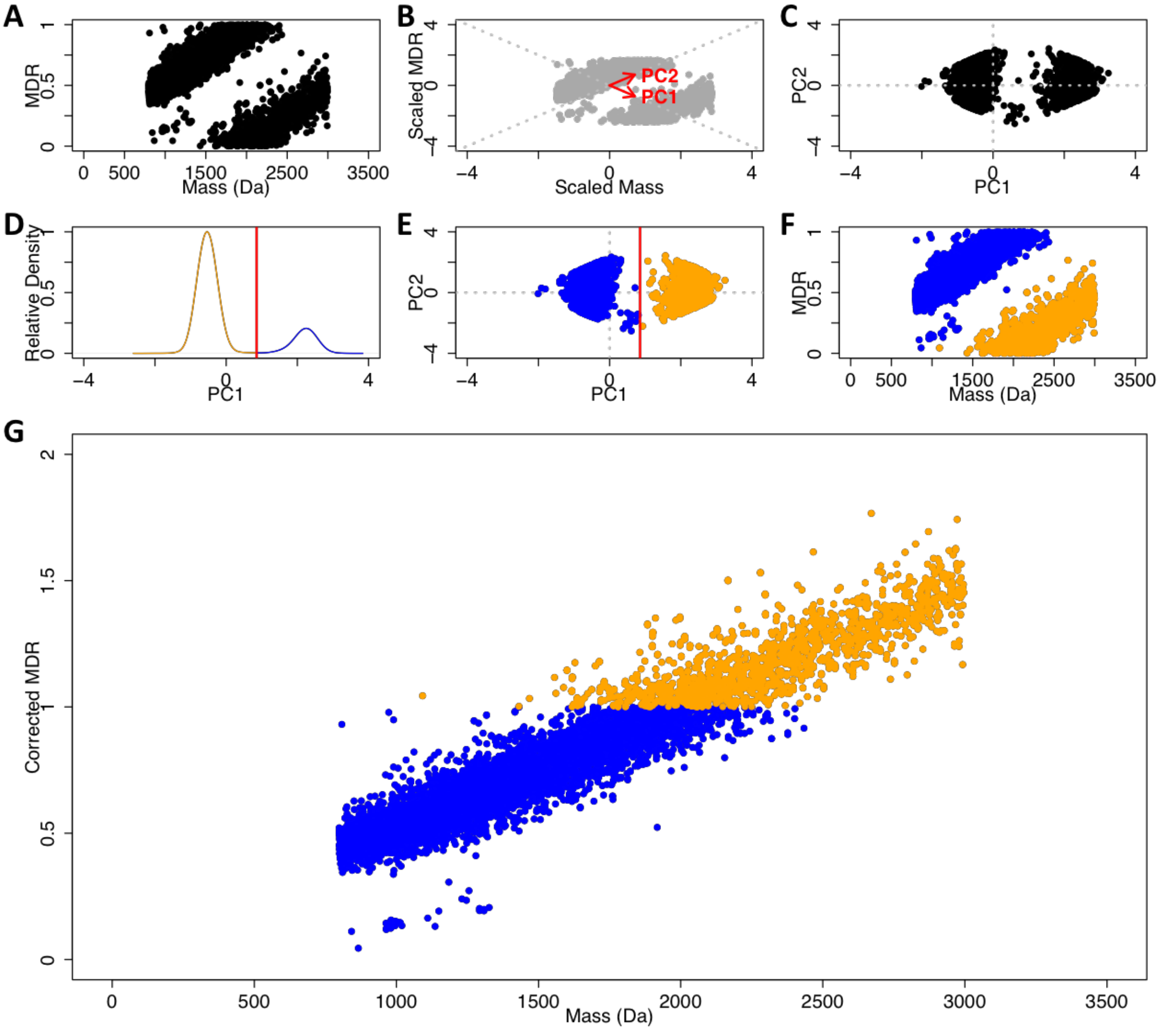


Figure 3-III. *Transformation of Mass Decimal Residual Values. MDRs were transformed to allow linear analysis of their relationship with peptide mass. The data shown are the masses selected for MS^2 in the unlabelled sample.* ***A****: Initial distribution of MDR vs. peptide mass for all masses less than 3500.* ***B:*** *Data centred and scaled (to unit variance) showing the calculated principal components.* ***C:*** *Data plotted on principal component axes.* ***D:*** *Density analysis of points along principal component 1. The red line indicates a cut-off point half-way between the two density maxima, partitioning the data by 'band' (orange and blue).* ***E:*** *Cluster assignment by density cut-off on principal component axes.* ***F:*** *Cluster assignment by density cut-off on original axes.* ***G:*** *Corrected MDR vs. mass (applying +1 to all MDR in orange cluster).*

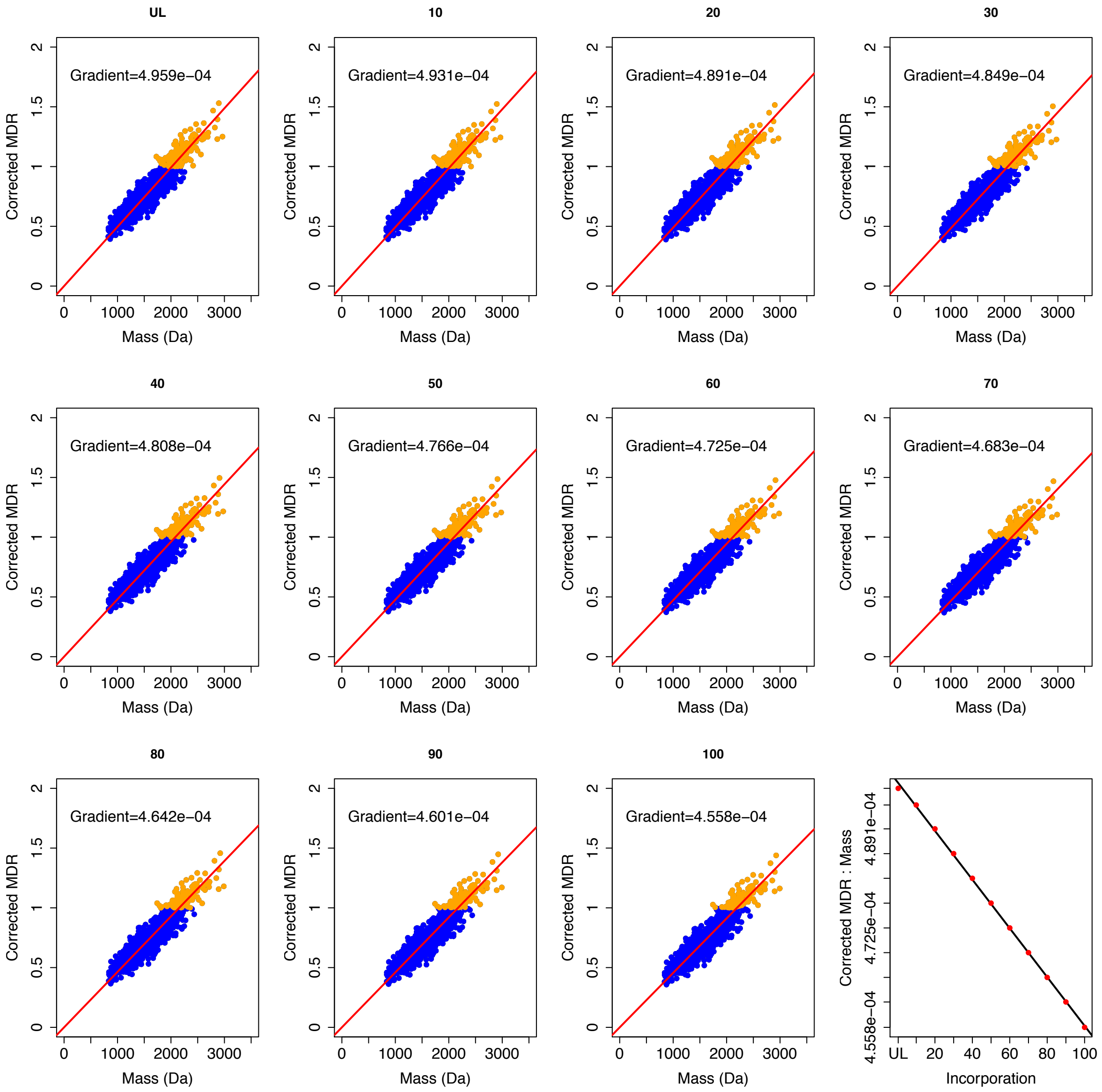
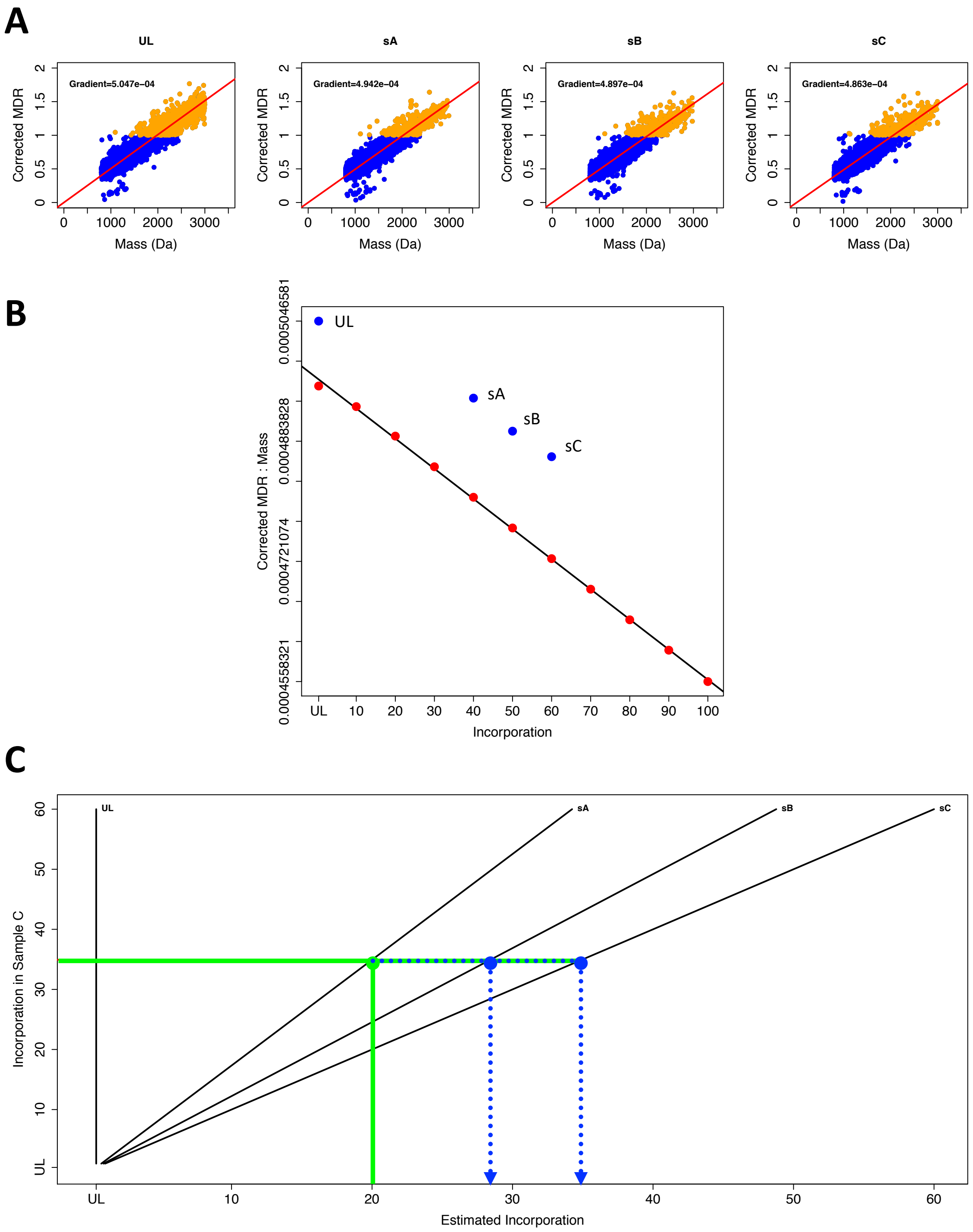


Figure 3-IV. *Mass to Corrected MDR Gradients across a Theoretical Dataset. Data shown are the masses of the highest intensity predicted isotopologue (at the percentage of ^{15}N incorporation shown above each panel) for the set of 1042 unique PMCs identified from the unlabelled sample. UL=Unlabelled sample, corresponding to a natural abundance of ^{15}N (about 0.368%). A linear fit of gradient to incorporation was performed (lower right corner) demonstrating a linear relationship between the expected gradient and incorporation.*

I then applied the same analysis to my experimental unlabelled and labelled data. I took the list of all precursor masses (i.e. those selected for MS^2) reported in mgf peak-list files generated from each raw data file for the unlabelled sample and labelled samples A, B and C*,* transforming the mgf PEPMASS value for each spectral entry according to corresponding reported CHARGE. I made the assumption here that even in the case of partial ^{15}N labelling, the vast majority of species selected for MS^2 will still be sample-derived peptides (as opposed to environmental protein contamination such as skin keratins, or non-peptide contaminant ions falsely recognised as peptides), even if the fragment ion spectra are not easily identifiable without knowing their incorporation. For consistency of analysis with the theoretical data above I also ignored mass values greater than 3500 Da.

Figure 3-V. *Corrected MDR to Mass Gradients using MS^2 Precursor Masses. (Next Page)* ***A:*** *Precursor masses are shown for each sample: UL=Unlabelled sample, sA, sB, sC = labelled samples A,B and C.* ***B:*** *The Corrected MDR:mass gradients for unlabelled sample and samples A, B and C (labelled as above) are shown as blue dots on the same plot shown in the lower right hand corner of Figure 3-IV. While the gradients do apparently decrease linearly and (nearly) parallel the gradient predicted with theoretical data, the y-axis intercept is clearly different.* ***C:*** *Inferred relationship between UL and labelled samples A, B, and C, y-axis scaled for sample C. For example, if I estimate the true average ^{15}N incorporation of sample A to be 20% (green vertical line), then as the incorporations values for B and C on the x-axis intersect their respective gradient lines at the same y-axis value (green horizontal line) this predicts samples B and C to be at ~28% and ~35% incorporation respectively (blue dotted lines).*


A
UL
sA
sB
sC
Gradient=5.047e−04
Gradient=4.942e−04
Gradient=4.897e−04
Gradient=4.863e−04
Corrected MDR
0
0.5
1
1.5
2
0
1000
2000
3000
Mass (Da)
B
UL
sA
sB
sC
Corrected MDR : Mass
0.0004558321
0.0004721074
0.0004883828
0.0005046581
UL
10
20
30
40
50
60
70
80
90
100
Incorporation
C
UL
sA
sB
sC
Incorporation in Sample C
UL
10
20
30
40
50
60
Estimated Incorporation
UL
10
20
30
40
50
60

The precursor mass data also showed an apparently linear decrease in Corrected MDR:mass gradient relative to the expected (i.e. target) incorporation in each sample (Figure 3-V-A). The ratios were not in the same range of those attained with the theoretical dataset (Figure 3-V-B), but plotting the Corrected MDR:mass gradients vs. Incorporation calculated from the theoretical prediction (based on only well-characterised PSMs from the unlabelled sample) showed that they appeared to be approximately parallel to the same gradient obtained using all precursor masses from the unlabelled sample and samples A, B, and C, the intercept of the Corrected MDR:mass gradient axis intercept was clearly different, and it seemed unwise to assume that the gradient would nevertheless be the same.

I felt it would therefore be inaccurate to attempt to directly predict the incorporation of labelled samples A, B and C assuming a linear relationship with the same parameters as that observed for the theoretical data, but it did not seem unreasonable to assume that the relationship was still linear, and therefore the inter-sample ratios of Corrected MDR:mass gradients between samples A, B and C should predict the corresponding inter-sample ratios of ^{15}N incorporation.

For example, the HDPR approach predicts ratios of 0.73 and 0.88:1 for A and B relative to C, thus in the case of sample C incorporation being 1% ^{15}N, this predicts the ^{15}N incorporation levels of samples A and B would be 0.73% and 0.88% respectively. This principle can be extended for any value of C (Figure 3-V-C).

Since the original ‘target’ incorporation values would instead yield a ratio of 0.67:0.83:1 for A:B:C, the HDPR-predicted pairwise ratios between all three incorporations (0.73:0.88:1) are larger than expected. Assuming that the true incorporation levels can be lower than or equal to the target, but not higher, then the model has to be constrained by sample C<=60 since taking sample A or B as the baseline for the ratio instead results in predicts incorporations for sample C that are greater than the target. Under this constraint, the HDPR estimation predicts that samples C and B will show a similar incorporation percentage point deficit while the percentage point deficit for sample A will be larger.

3.6 Similarity Score Evaluation

I then analysed the list of 1042 unique, high-confidence (*Mascot* Expect value <0.001) PMCs identified from across Samples A, B and C using *HeavyMetL*. The analysis was repeated four times, using each of the candidate SS described above (Equations 5-8) in turn.

The estimated peptide incorporations produced using each analysis are shown in Figure 3-VI (top four rows). For comparison purposes, I also generated a random guessing baseline (SS_{RAND}) consisting of an equal-length population of simulated incorporation quantification results. For each result I simulated the effect of the SS maximisation approach by generating a random number (between 5 and 20) of dummy chromatographic maxima each with a random SS uniformly distributed between 0 and 1, then taking the highest score value. This was then paired with a random incorporation estimate uniformly distributed between the minimum and maximum assignable incorporation levels (10% to 95%) (Figure 3-VI, bottom row).

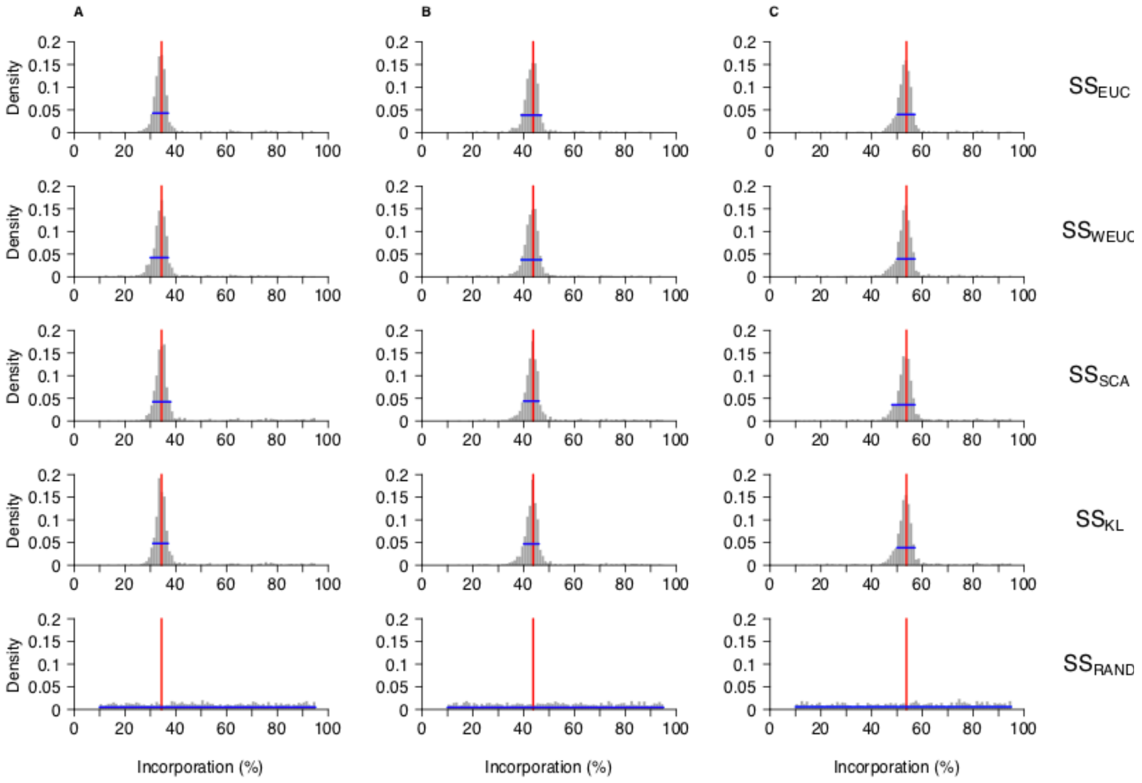


Figure 3-VI. *Performance of Incorporation Level Estimators. Histograms of reported peptide incorporation levels are shown after calculation using four measures of spectral similarity (rows 1-4). For comparison purposes, a baseline result generated by random guessing (using a uniform distribution) is shown in row 5 (SS_{RAND}). In each case the range of possible reported incorporations was 10-95%. Comparisons are shown for three different samples of* O. tauri *(**A**, **B** and **C**; left-to-right across grid columns) grown to target incorporation levels of 40%, 50% and 60% respectively. The actual attained incorporation levels were estimated by excluding the lowest quartile of reported incorporations by frequency (below the blue lines, these lines therefore also indicate the spread in terms of 'full width at 25% maximum') and taking the mean across all four Similarity Scores for each sample, shown as the red line overlaid through each column of the grid.*

From the HDPR estimation I expected the true incorporation level for each sample to be several percentage points lower than the target due to impurities in labelled media, so in each sample I estimated a true rate as the mean of all measured labelled signal incorporations in the upper 75% of incorporations when ranked by relative frequency (to exclude the background of randomly distributed mis-quantifications, following a similar rationale to the FWHM estimation of peak width, see Figure 1-IV). To estimate the true mean, for each sample I took the mean of all four SS means, yielding estimates for labelled samples A, B and C of 34.4% 43.7% and 53.7% respectively. These estimates are shown for each sample as the red vertical lines in Figure 3-VI. The spread of reported incorporations around the estimated true incorporations was not substantially different between SS candidates when taking the corresponding standard deviations of the upper 75% of peptide incorporations (as for the mean above), leaving little to choose between the candidates in terms of incorporation accuracy, although SS_{KL} consistently gave the tightest spread (Figure 3-VII).

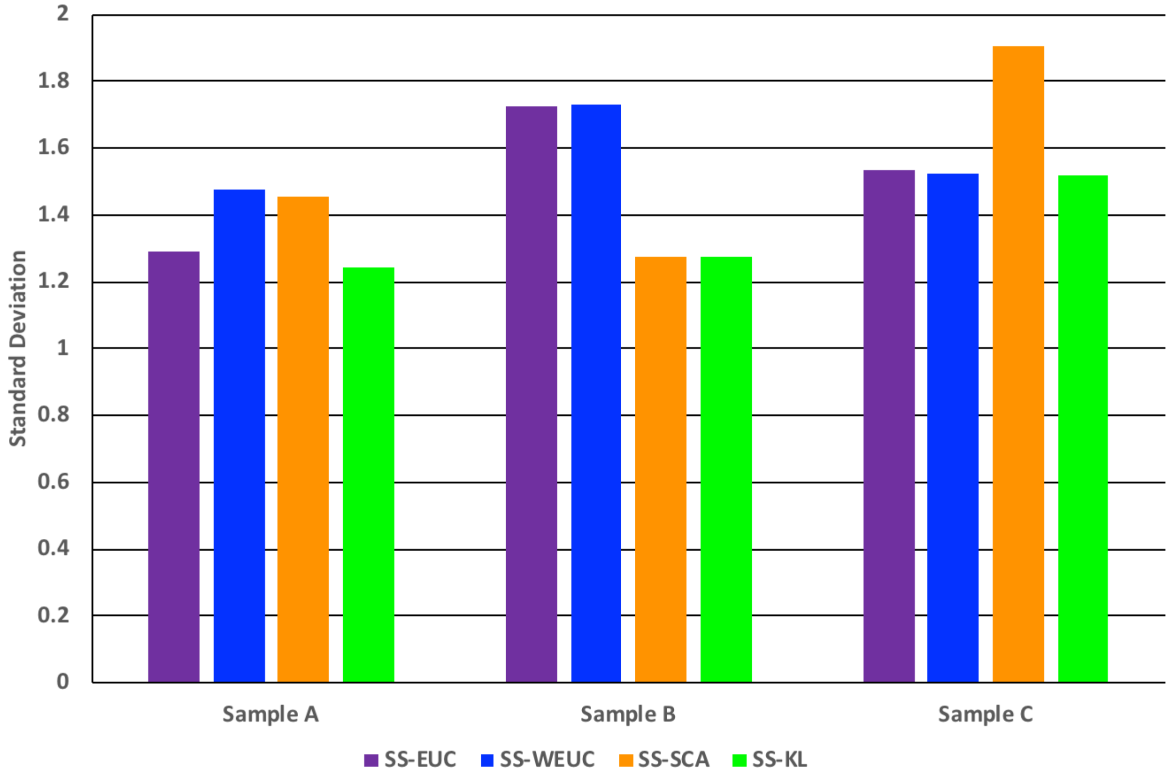


Figure 3-VII. *Standard Deviations of Reported ^{15}N Percentage Incorporations. The coloured bars show standard deviations for the upper 75% of incorporation results when ranked by relative frequency.*

Assuming the estimates for samples B and C were accurate, the incorporation estimate for sample A was several percentage points higher than the relationship between A,B and C predicted by the HDPR approach described in Section 3.5 (Figure 3-VIII); but closer to what would be expected with a constant deficit across all 3 samples of approximately 6 percentage points. The simplest interpretation would be that this was indeed the case, in which case the HDPR-based estimation for sample A was not particularly robust given the sizeable error.

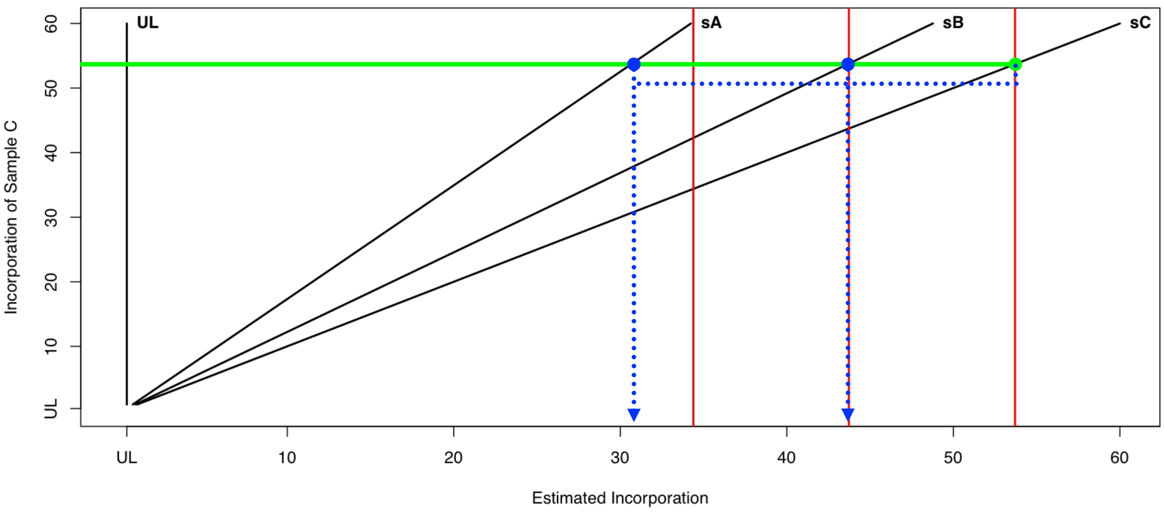


Figure 3-VIII. *Comparison of Incorporation Estimates by* HeavyMetL *and HDPR. The black line for each sample (UL=Unlabelled sample, sA, sB, sC = labelled samples A, B and C) indicates the relationship inferred by the HDPR approach. The y axis is scaled for labelled sample C (although the data could also be plotted scaled to sample A or B), such that if sample C takes a particular incorporation value, the x-axis value corresponding to that value on the y-axis indicates (for each sample) what the expected incorporation of that sample will be (thus the sample C line lies on y=x). The estimates obtained by* HeavyMetL *quantification are shown as vertical red lines, illustrating that assuming the sample C incorporation to be 53.7% as measured by* HeavyMetL *(green line), the HDPR approach (blue dotted lines) predicts that the sample B incorporation to be ~44% (in close agreement with the* HeavyMetL *result) but the sample A incorporation to be ~31% rather than closer to ~34-35% as predicted by* HeavyMetL.

3.7 Standardisation of Incorporation Quantification Errors

In order to evaluate the quality of results produced by each SS candidate, it is necessary to define a framework for comparison. The outcome of interest is the accuracy of the incorporation assessment reported by *HeavyMetL* when using each SS candidate, comparing each quantification result to the 'true' sample mean to get an error in percentage points. The size of the error is of interest, while the sign (positive/negative) is not; a positive error is no better or worse than a negative one. Since the 'true' incorporation values are not known (see discussion in Sections 3.3 and 3.5 above), the approximation I will use is the robust mean based on the upper 75% of the distribution of incorporation levels across all four SS (see Figure 3-VI).

In order to compare errors between samples with different ^{15}N incorporation levels, a method of error standardisation is necessary. Consider an incorporation range X_{min} to X_{max} and an incorporation quantification result x for a particular peptide. If x is produced by random draw between X_{min} to X_{max} with equal probability, the population X of x will be uniformly distributed. However, the incorporation percentage point error of x compared to the true (or estimated true) incorporation T is not necessary uniform because T may not be equidistant from both X_{min} and X_{max}. Furthermore, in other samples where T is different but X_{min} and X_{max} remain the same, the range of possibly quantification results either side of T is different. A standardised score allows combination of the results from labelled samples A, B, and C to compare SS candidate performance across all samples in aggregate (Figure 3-IX). This requires that the percentage point errors from comparing the *HeavyMetL* quantification incorporation values to the estimated 'true' incorporations are transformed to a Standardised Incorporation Quantification Error (SIQE).

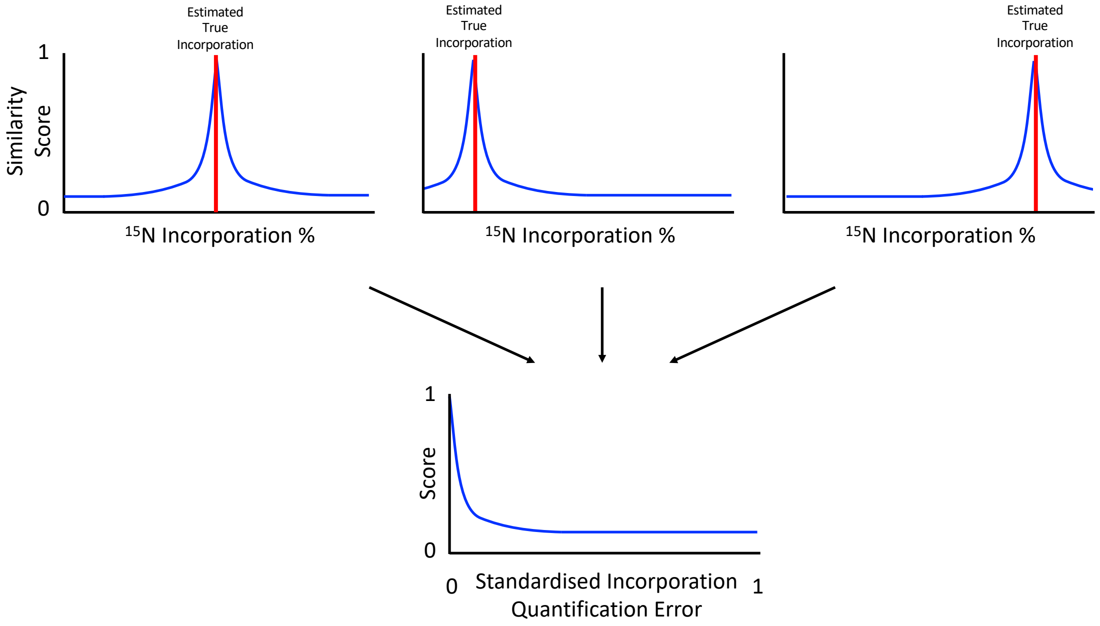


Figure 3-IX. *A Standardised Incorporation Quantification Error. Such a score is required to directly compare incorporation percentage point errors between Samples A, B and C.*

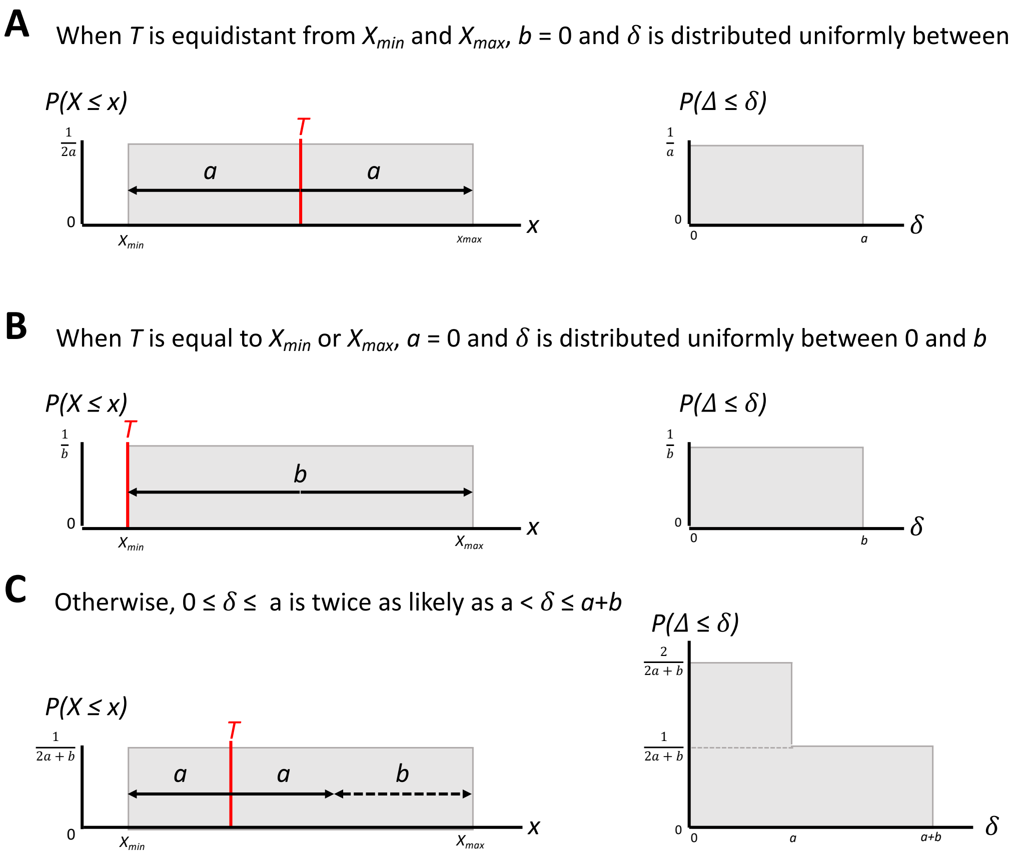
Let δ be defined as the absolute (i.e. unsigned) percentage point error between the true incorporation and an observed incorporation, i.e. $\delta = |T - x|$. Let a be the smaller of the two distances from true incorporation T to the limits of the incorporation range (Equation 5). Let b be the difference between a and the larger of the two distances from the true incorporation to the limits of the incorporation range (Equation 6).

$$a = \min\left((T - X_{min}), (X_{max} - T)\right) \qquad \text{(Equation5)}$$

$$b = \max\left((T - X_{min}), (X_{max} - T)\right) - a \qquad \text{(Equation 6)}$$

When T is equidistant from X_{min} and X_{max}, $b = 0$ and a uniformly distributed population of random incorporations X will yield a corresponding distribution of absolute percentage point errors Δ that is uniformly distributed between 0 and a (Figure 3-X-A). When T is equal to X_{min} or X_{max}, $a = 0$ and a uniformly distributed population of random incorporations X will yield a distribution of δ (Δ) that is uniformly distributed between 0 and b (Figure 3-X-B). In every other case T will be closer to either X_{min} or X_{max}, therefore a uniformly distributed population of random incorporations X will yield a distribution of signed errors uniformly distributed between either -a and a+b, or -(a+b) and a. Because δ

is an absolute error, a random x yielding $0 \leq \delta \leq a$ is thus twice as likely as $a < \delta \leq a{+}b$ and the distribution of Δ is therefore not uniform (Figure 3-X-C).



A When T is equidistant from X_{min} and X_{max}, b = 0 and δ is distributed uniformly between 0 and a

B When T is equal to X_{min} or X_{max}, a = 0 and δ is distributed uniformly between 0 and b

C Otherwise, $0 \leq \delta \leq$ a is twice as likely as $a < \delta \leq a{+}b$

Figure 3-X. *Distribution of Errors from a Random Estimator. Let δ be the absolute difference between the true incorporation T and an estimate of the incorporation x, i.e. $\delta = |T - x|$ If a population of random estimates X has uniform distribution, then the corresponding distribution of the absolute error Δ, depends on the position of T within the range of possible estimations X_{min} to X_{max}. Let a be the smallest difference between T and X_{min} to X_{max}. Let b be the difference between a and the largest difference between T and X_{min} to X_{max} (See equations 9 and 10, main text).* ***A & B:*** *If T is equidistant from X_{min} and X_{max} or equal to one or the other, then the range of absolute deviations is also uniform.* ***C:*** *In all other cases the range $0 \leq \delta \leq a$ is twice as likely as $a < \delta \leq a{+}b$.*

From Figure 3-X I derive a probability distribution function for *Δ*:

$$f(\delta) = \begin{cases} \frac{2}{2a+b}, & 0 \le \delta \le a \\ \frac{1}{2a+b}, & a < \delta \le a+b \\ 0, & \text{otherwise} \end{cases} \qquad \text{(Equation 7)}$$

To obtain a SIQE I calculate the probability of achieving an error less than or equal in magnitude to that observed if one were to randomly select an incorporation estimate. This is the cumulative distribution function for *Δ*, which for a given error *δ* is the integral of the probability distribution function *f(δ)* (Equation 7) from 0 to *δ*:

$$\mathrm{SIQE} = \mathrm{P}(\Delta \le \delta) = \begin{cases} 0, & \delta < 0 \\ \frac{2\delta}{2a+b}, & 0 \le \delta \le a \\ \frac{a+\delta}{2a+b}, & a < \delta \le a+b \\ 1, & \delta > a+b \end{cases} \qquad \text{(Equation 8)}$$

3.8 Comparison of SIQEs Between Similarity Scores

To further differentiate the performance of the four SS candidates, I compared the ability of the methods to consistently assign a higher score to matches with low-error compared to matches with high-error. I combined the data from samples A, B and C, and removed zero-scored PMCs, for a total of 3108 score-estimate pairs for each of the four SS candidates (and a corresponding dataset of 3108 random score-estimate pairs in $\mathrm{SS}_{\mathrm{RAND}}$).

I first compared SS–SIQE relationships directly (Figure 3-XI-A). Regardless of SIQE, most score values were generally close to 1. This is a result of the scoring process selecting the ‘best’ (highest) score and is also reflected in the distribution of $\mathrm{SS}_{\mathrm{RAND}}$. A mathematical explanation for this is given in Appendix III. The relationship between score validity and score, while monotonic for all four SS, would not necessarily have the same gradient shape. To enable a fair comparison, I transformed each score to a score rank (assigning rank 1 as the highest score) (Figure 3-XI-B,C). Comparing SIQE across score ranks, all four methods clearly outperformed random guessing, with a greater weighting of high scores towards low SIQE. Many apparent differences (such as higher general scores for $\mathrm{SS}_{\mathrm{WEUC}}$) were removed by rank normalisation, showing that both Euclidean distance candidates ($\mathrm{SS}_{\mathrm{EUC}}$, $\mathrm{SS}_{\mathrm{WEUC}}$) gave almost identical performance to $\mathrm{SS}_{\mathrm{SCA}}$ and $\mathrm{SS}_{\mathrm{KL}}$.

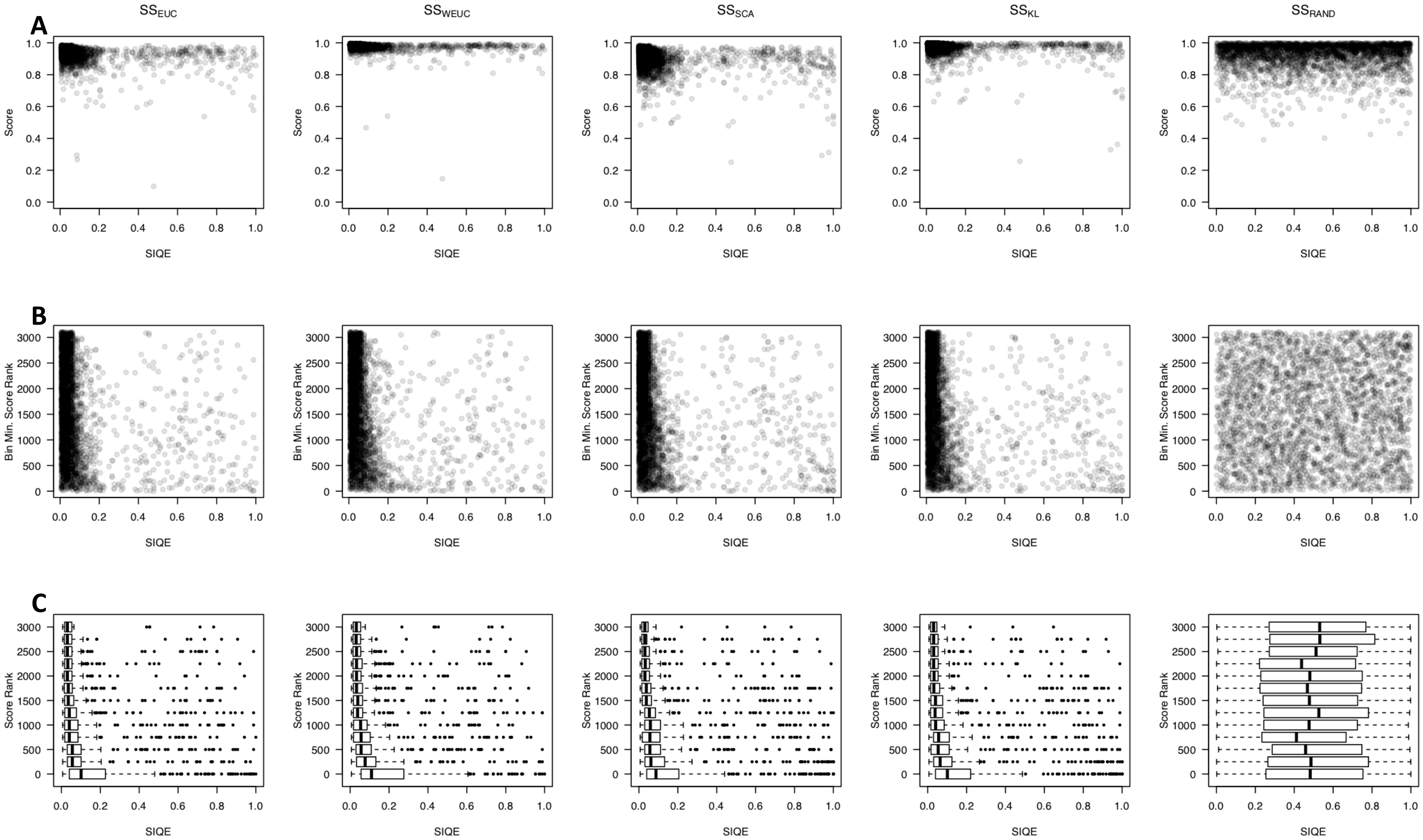

A
B
C
SS_{EUC}
SS_{WEUC}
SS_{SCA}
SS_{KL}
SS_{RAND}
Score
Bin Min. Score Rank
Score Rank
SIQE
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
3000

Figure 3-XI. *Comparison of Relationship between Similarity Score and SIQE. (Previous Page)* ***A:*** *Similarity Score (y-axis) vs. SIQE (x-axis).* ***B:*** *Similarity Score Rank (y-axis) vs. SIQE (x-axis). C: The distribution of SIQEs (shown via horizontal boxplots) of SS ordered by rank in bins of 25. The 'boxes' cover the 25th-75th percentile range (i.e. the interquartile range) in each case, and the whiskers extend 1.5 times the interquartile range in either direction, bounded by the plot limits. Each column of the figure corresponds to results for a particular SS candidate (or the simulated random baseline score) as indicated.*

I also examined the data for peptide composition-related differences in SIQE. The peptides analysed were drawn from a list of peptide-spectrum matches and therefore their characteristics are heavily biased towards features that favour identification by standard proteomics MS and are further constrained by the processing prior to MS analysis. For example, the peptides were produced by tryptic digest, so the vast majority will end in either arginine or lysine. I reasoned that any significant composition-related difference in SIQE would affect the relationship between SIQE and peptide length and/or the number of nitrogens the peptide contained, regardless of sequence. I did not observe any such patterns (Figure 3-XII), which suggested that (as expected) all the Similarity Score candidate scores were unaffected by peptide composition.

Looking at the whole dataset it was difficult to tease out any further differences, but closer inspection of the binned data revealed both Euclidean distance methods showed a spike in mean SIQE at the highest score, suggesting that in some circumstances SS_{EUC} and SS_{WEUC} assign scores very close to maximum to mis-fitted data (Figure 3-XIII).

Ultimately, I felt that the poor mean SIQE of both Euclidean distance measures at very high score rank weighed against their use in *HeavyMetL*, as applying increasingly conservative score thresholds should (ideally) always yield better quality data. Although the observed difference in performance was small, I selected the Kullback-Leibler Divergence-based score SS_{KL}, as the better-performing of the remaining two SS candidates for use in the incorporation assessment part of the *HeavyMetL* algorithm

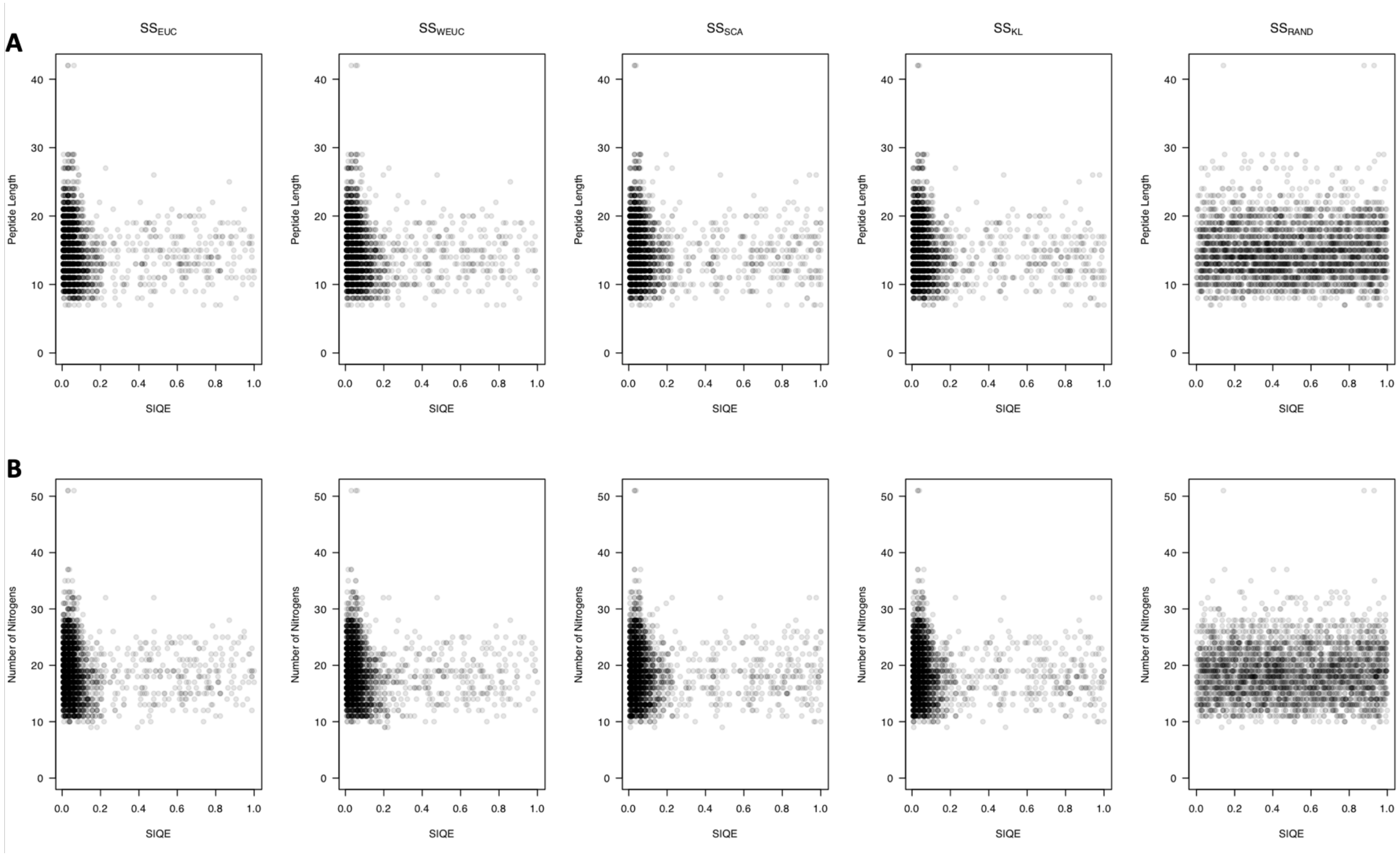


Figure 3-XII. *Comparison of Relationship between Peptide Characteristics and SIQE.* ***A:*** *Peptide length (y-axis) vs. SIQE (x-axis).* ***B:*** *Number of nitrogen atoms in the peptide (y-axis) vs SIQE (x-axis). Each column of the figure corresponds to results for a particular SS candidate (or the simulated random baseline score) as indicated.*

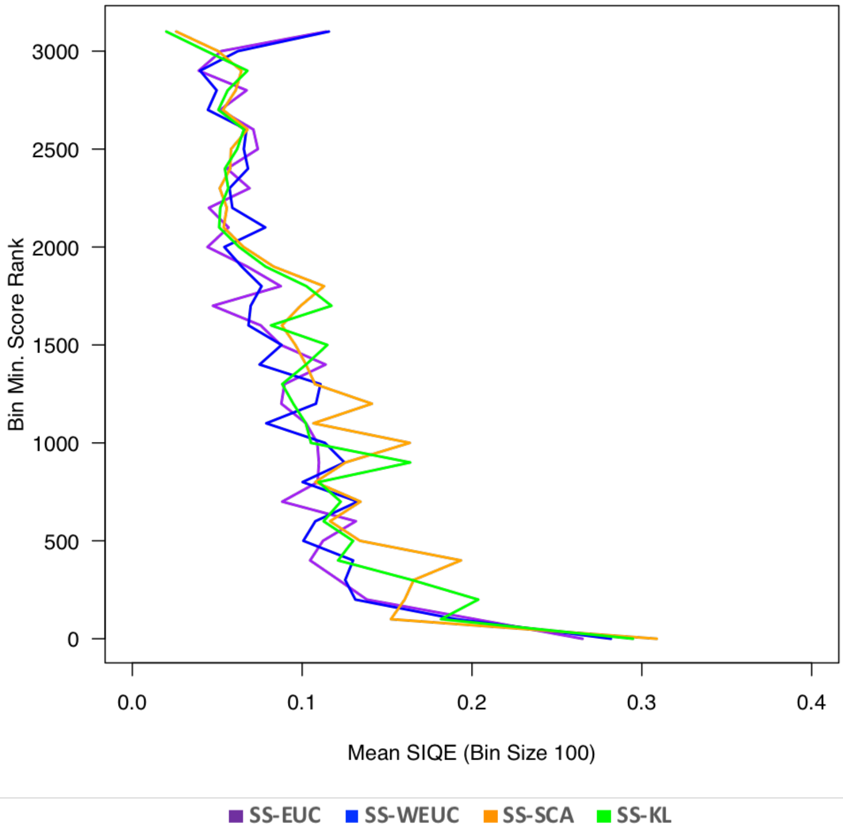


Figure 3-XIII. *SIQE as a Function of Score Rank. The SIQE is shown as the mean SIQE value over bins of 100 scores in ascending order. Note in particular the 'spike' in mean SIQE at very high score rank for the Euclidean based methods.*

3.9 Incorporation Assessment Performance at High Spectral Noise

Having observed how *HeavyMetL* incorporation assessment (using the selected Similarity Score, SS_{KL}) performed on a 'real-world' dataset. I was also interested as to how the incorporation assessment performance would hold up as the data quality decreased. To explore this, I created a dataset of 10000 random peptides by concatenating several random protein sequences (generated using the ExPASy random protein sequence generator at https://web.expasy.org/randseq/ - the generator has a maximum residue length per protein) then applying standard tryptic digestion rules. Using R with the v8 package to call *HeavyMetL* code where appropriate, for each peptide, I assigned a random incorporation level between 0 and 100 (restricted to multiples of 2 in order to halve processing time) and used the *HeavyMetL* isotopologue prediction algorithm to generate a theoretical spectrum at the assigned incorporation level (thus yielding approximately 200 peptides at each incorporation level). From the true theoretical spectrum, I then generated ‘simulated observations’ with increasing amounts of log-normal noise applied to each isotopologue intensity. To generate a ‘simulated observation’ spectrum with a given noise level, I modelled the intensity of each isotopologue as a log-normal distribution with log mean equal to the log theoretical intensity of that isotopologue and log standard deviation equal to the noise level, then drew a replacement ‘noisy’ intensity value from this distribution. For example, for a noise level (in log standard deviation units) of 0, each isotopologue ‘noisy’ intensity is drawn from a log normal distribution with log mean = log theoretical intensity and log standard deviation = 0, i.e. the ‘simulated observation’ spectrum for noise level 0 is identical to the theoretical distribution. I generated 101 ‘simulated observation’ spectra for noise levels (in log standard deviation units) from 0 to 1 in increments of 0.01.

For each simulated observation I then calculated the best matching theoretical distribution (again for incorporation values between 0 and 100 in multiples of 2) and the associated SIQE. Note that in this analysis (due to the parameterisation of the log-normal distribution in R), ‘log’ means the natural log. A log standard deviation of 1 corresponds to a fold change of approximately 2.7, which would be well above what is generally considered to be the lower bound of detectable genuine fold changes using EML (208) and therefore considerably ‘noisier’ than typical real-world data.

As Figure 3-XIV shows, the performance of the incorporation assessment decreases as spectral noise (in log standard deviation units) increases. The relationship between Similarity Score and noise appears to have an inverse nonlinear component, as both the median value and the lower limit of spread (1.5 times the interquartile range) decrease more rapidly as the log standard deviation increases (Figure 3-XIV-A). However, the relationship between inaccuracy of incorporation assessment (as measured by SIQE) and noise is closer to a positive linear one, judging by the 90% density threshold in Figure 3-XIV-B (90% of the density for each level of log standard deviation noise is below the red line). It is possible to still correctly assign the incorporation level even with a low Similarity Score, so long as it is still the highest among all tested incorporation levels, so it follows that decreasing Similarity Score does not necessarily mean a proportionate increase in SIQE.

The results suggest that the optimum Similarity Score threshold for any given dataset would ideally be derived empirically. However, this is contrary to the stated goal of setting default parameters that are robust in most scenarios. Following the same logic as above, since a 1.5-fold change has traditionally been the minimum threshold to observe a genuine change (208), estimate of a standard deviation of 1.5 as an upper noise limit in typical data seems appropriately conservative; this corresponds to a log standard deviation of approximately 0.4 (see purple vertical lines in Figure 3-XIV) which, from the simulated dataset, suggests a possible Similarity Score threshold of 0.85 as this would encompass nearly all genuine matches (i.e. when there is measurable signal present) and yield 90% of incorporation assessments with a SIQE less than 0.1.

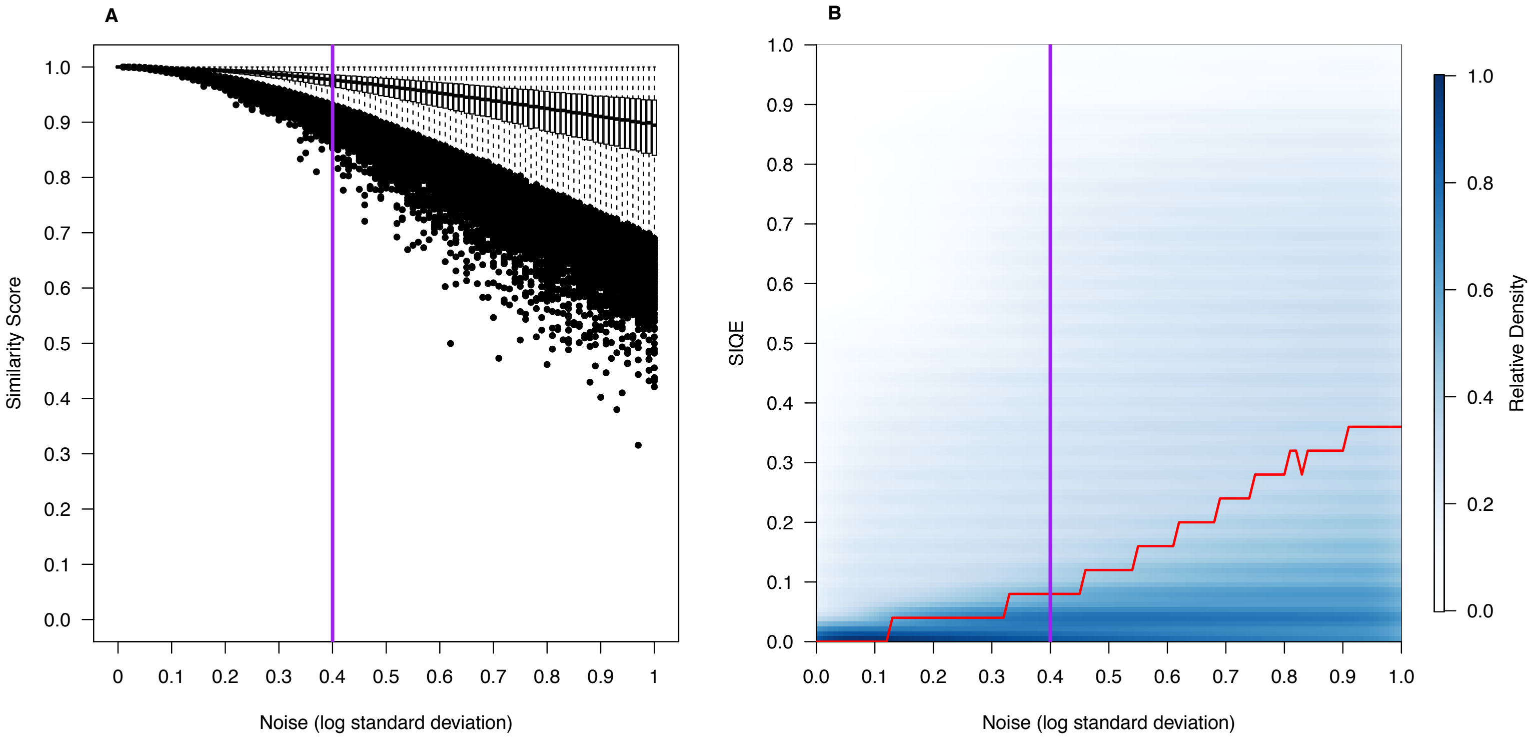


Figure 3-XIV. *Incorporation Assessment Performance Versus Simulated Noise.* ***A:*** *Peptide length (y-axis) vs. Noise (in log standard deviation units; x-axis), shown as box plots for each simulated log standard deviation value. The 'boxes' cover the 25th-75th percentile range (i.e. the interquartile range) in each case, and the whiskers extend 1.5 times the interquartile range in either direction, bounded by the plot limits.* ***B:*** *SIQE (y-axis) vs Noise (in log standard deviation units; x-axis), plotted as a density estimate (where darker blue = greater density). The red line indicates a cut-off underneath which 90% of the density lies. The purple line indicates a (conservative) estimate of the typical upper bound for spectral noise in EML data.*

3.10 Conclusions

HeavyMetL is dependent upon spectral similarity in order to identify the theoretical distribution that optimally matches the isotopologue pattern observed in the raw data. I thus considered several Similarity Score candidates, described above (Equations 1-4), each transformed for consistency into a result between 0 (no similarity) and 1 (perfect similarity). All four SS candidates were tested against three different labelled *O. tauri* samples (A, B and C) generated with 'target' incorporations of 40, 50 and 60% ^{15}N to see which provided the most robust assessment of incorporation.

To estimate the 'true' mean peptide ^{15}N incorporation for each sample A, B and C, I first tried a HDPR estimation approach, with mixed success. It was clear that the HDPR analysis yielded a linear relationship between the Corrected MDR:mass gradients sampled, but my attempt to interpret these relationships in terms of ^{15}N incorporation by comparison of the resulting gradients with a set of standard gradients derived from predicted masses for various levels of ^{15}N incorporation proved unsuccessful, as the theoretical results clearly covered a different range of gradient values. A combination of three factors may explain the discrepancy; first, the theoretical mass lists were all inferred from a limited subset of the unlabelled precursor masses with an inherent bias for identifiability (no co-eluting peptide resulting in a chimeric MS^2 spectrum, higher intensity precursor); secondly, for the theoretical data this subset of precursors was the same across all incorporation levels, whereas for the experimental data the same peptide species would not always have been selected; thirdly, the theoretical data assumed that the most intense isotopologue of the peptide species would always be the selected mass, which may not always be the case in the complete data.

I was, however, able to infer an expected ratio between the ^{15}N incorporation levels of samples A, B and C, which was approximately consistent with the expected ratios between the ^{15}N incorporation values for samples B and C, but predicted a lower incorporation for sample A relative to B and C.

I then analysed the data with *HeavyMetL* using all four SS candidates. All four measures outperformed random guessing and, overall, showed very similar performance. Taking the mean incorporation across the four SS candidates for each sample yielded estimated sample incorporation levels of 34.4% 43.7% and 53.7% for sample A, B and C respectively. This result was much closer to the consistent ^{15}N incorporation percentage point drop initially expected, rather than the lower sample A incorporation relative to A and B predicted by HDPR, suggesting that the HDPR analysis was not particularly accurate.

Comparing the performance of the four SS candidates, the Kullback-Leibler Divergence based score SS_{KL} yielded the lowest standard deviations in ^{15}N incorporation across the three samples (when background mis-quantification was removed). After transformation of the absolute ^{15}N incorporation error (deviation from the predicted mean) to a SIQE, I compared SS candidate SIQEs against SS ranks, and against peptide characteristics which might affect incorporation assessment. These comparisons revealed no marked differences between any of the Similarity Scores. However, closer examination of mean SIQE showed some evidence of inconsistency of performance in SS_{EUC} and SS_{WEUC} at very high score ranking, suggesting there may be spectral characteristics that they overvalue. This last observation was also consistent with manual evaluation of results, in that the Euclidean based methods assigned more ‘obviously’ (i.e. by eye) incorrect incorporation estimates. When taken as a whole, there was much less difference in performance between the Similarity Scores than I had expected based on *ad hoc* manual evaluation of the SS while testing the algorithm code, demonstrating the danger of bias in a manual evaluation approach. On the other hand, making ‘obviously’ incorrect incorporation estimates (even if a particular score performs similarly in the aggregate) contributes to a user perception of poor quantification, so there could be an argument to admit (effectively anecdotal) manual evaluation into the overall consideration.

A further observation was that, when comparing across SS rankings, the SIQE increased with decreasing score as expected but the average SIQE for each of the 4 SS at low score was still much lower than the baseline ‘random guessing’ SIQE (compare Figure 3-XI-C across columns). The fact that lower ranked scores are still generally associated with an SIQE lower than obtained by random matching suggests that the majority of the time in this dataset, these may be genuine low-intensity labelled signals with an associated greater uncertainty around the incorporation estimate, rather than a complete mismatch, which speaks to the robustness of the algorithm using any of the four SS candidates. One explanation could be that even if there is only a weak signal with 1 or two isotopologues, if a particular incorporation value is the only one that yields intensity when its *m/z* extraction windows are applied, that will be the best matching distribution even if the SS is low. If this were the case, a re-analysis of the sample on an older instrument with poorer resolution and mass accuracy, or a different dataset composition with a more complex sample might be expected to result in a substantial increase in SIQE at low SS rank. Alternatively, in a sample with a wide spread of expected peptide incorporation values, even the increase in SIQE observed in this analysis at low score rank might be unacceptable. In all of these cases, it would be advisable to apply a conservative SS threshold to ensure only good matches are considered further.

There was very little to differentiate the remaining two scores, but SS_{KL} had consistently (if by a very small margin) outperformed SS_{SCA} in terms of a smaller spread of incorporation assessments, so I chose the Kullback-Leibler Divergence based Similarity Score for further use in *HeavyMetL*. Finally, I examined the performance of incorporation assessment (using the chosen Similarity Score) as data quality decreased, using a simulated dataset to which I introduced increasing amounts of log-normal noise. The incorporation assessment showed robust performance even when noise was increased substantially beyond what is typically encountered in EML, with more than 90% of results having SIQE < 0.1 even when the added noise (in log standard deviation) corresponded to an average of 1.5-fold change in intensity. This analysis also suggested that a default Similarity Score threshold of 0.85 could be used to separate incorporation assessments based on true-but-noisy signal from those based on false signals (such as incorporation assessments based on ‘signal’ that is just background noise).

Chapter 4: Benchmark of *HeavyMetL* Performance *vs.* an Orthogonal Approach

4.1 Introduction

To assess quantification performance, *HeavyMetL* was benchmarked against an existing public dataset. The recent publication of the *ProteinTurnover* algorithm (166) describes a cohort of *Arabidopsis* seedling samples with increasing ^{15}N incorporation over a series of time points (4, 8, 24, 32, 40, 48 h). The dataset also includes a 0 h time point which effectively contains no labelled sample. The authors have evaluated the MS^2 spectra manually to further filter the list of peptide identifications. These data provide a useful public benchmark across a range of ^{15}N incorporation percentages in which the number of mis-sequenced peptide-spectrum matches may be assumed to be relatively low.

In contrast to the dataset used in Chapter 3, this dataset is a ‘SILAC-style’ dual labelling containing both unlabelled peptides and labelled peptides at a range of ^{15}N incorporations.

ProteinTurnover is an ideal candidate for a benchmarking comparison since the approaches of *HeavyMetL* and *ProteinTurnover* are partially orthogonal, with respect to the actual quantification of the unlabelled and labelled peak. *HeavyMetL* locates the unlabelled and labelled peak apexes within the RT window independently. The process of fitting a range of labelled distributions simultaneously determines both labelled peak apex RT and incorporation percentage for the distribution with the best fit. The unlabelled/labelled ratio is then calculated from the ratio intensity at each label apex. In contrast, *ProteinTurnover* does not attempt to define the limits or apex of either label elution peak directly, but instead measures the extracted ion chromatogram of a series of *m/z* windows corresponding to the approximate location of possible isotopologues. The gradients of the linear correlations between the highest intensity isotopologue and the other isotopologues over all observations within an RT window are used to determine a combined unlabelled and labelled spectrum. The unlabelled and labelled distributions are modelled by fitting a mixture of two beta-binomial distributions using maximum likelihood estimation; the ratio of unlabelled to labelled is then derived from the area under the curve of the two fitted distributions, and the labelled incorporation percentage from the parameters of the rightmost fitted beta-binomial curve. The use of a beta-binomial distribution is interesting; essentially the authors propose that since incorporation into different amino acids may not proceed at the same rate, for a given peptide there is not an equal probability of every nitrogen in the peptide being labelled (this probability would also be the overall incorporation level). To account for this, rather than fitting a binomial distribution (which would assume equal probability), they fit a beta-binomial distribution where the probability for each nitrogen atom being labelled follows a beta distribution, which varies between 0 and 1 with mean π equal to the overall incorporation level.

The effect of this adjustment is to slightly broaden the distribution shape from a standard binomial distribution. *HeavyMetL*, by comparison, assumes an equal probability of every nitrogen in the peptide being labelled (as per the standard binomial distribution). However, since the Kullback-Leibler Divergence based Similarity Score used in *HeavyMetL* weights in favour of high intensity isotopologues which will typically not be substantially affected by this broadening, there may be very little practical difference. One effect of the peak broadening however is that taking into account more, low intensity isotopologue masses during matching may increase susceptibility of the matching to background noise, increasing variance when quantifying low intensity spectra.

Since the extraction of ion intensity data is performed before the determination of label incorporation, there are higher practical limits on the mass error tolerance for extraction in *ProteinTurnover*. However, by calculating relative isotopologue proportions from the gradients of their correlations across the full RT extraction window rather than just using just the spectra closest to the peak apex, *ProteinTurnover* is potentially more robust in scenarios where the apex signal for a label state is heavily contaminated by noise while the rest of the elution is unaffected. For the best comparison, therefore, I assessed the performance of both algorithms over the time course as a whole, which includes scenarios of low signal-to-noise for the labelled signal in the earlier time points (where ^{15}N incorporation is low), as well as stronger signal-to-noise in the later time points where incorporation is higher.

Through the graphical user interface, *HeavyMetL* allows quantification parameters to be easily optimised for a particular dataset, as the result of changing settings can be quickly re-calculated for individual proteins and peptides and assessed visually. For the comparison to *ProteinTurnover,* I applied no manual optimisation to this dataset, relying on default parameters, to avoid any optimisation bias. In addition, to ensure the best comparison, the same isotopic mass and abundance constants used in the *ProteinTurnover* algorithm were used in *HeavyMetL* (copied directly from the *ProteinTurnover* source code).

4.2 Benchmark Dataset Materials and Methods

N.B. This chapter describes a re-analysis of publicly available data. Sample preparation, mass spectrometry analysis and processing of the dataset using ProteinTurnover *were performed by Fan et al. (the authors of the* ProteinTurnover *paper) and details on these steps are included here for information purposes only. The re-analysis of the data using* HeavyMetL*, the comparison of the results to the* ProteinTurnover *quantification data, and the evaluation of the benchmark conclusions are my own work.*

4.2.1 Dataset Details

The *ProteinTurnover* paper describes a number of datasets in which they investigate both label incorporation and label dilution (166). This analysis used the label incorporation time-course dataset, for which both raw data and quantification values obtained by the authors are publicly available on MassIVE *via* the accession number MSV000079223). The files relevant to this analysis were:

1. The MS data in mzXML format (raw/hr_2/4/8/24/32/40/48.mzXML). At the time of writing, the 16 h time point file (hr_16.mzXML) hosted on MassIVE was truncated and thus excluded from the analysis.

2. The list of identified peptides in Spectrum report file format (other/soluble-FDR-Scaffold spectrum report.csv).

3. *ProteinTurnover* quantification values (other/ Root-soluble_Turnover-many.rar/ fits-many_soluble.csv). The labelled signal incorporation percentage and relative abundance used in this analysis are derived from the π and alpha columns for each time point in the

spreadsheet: labelled signal incorporation percentage = 100 π; relative abundance = alpha/(1-alpha).

4.2.1.1 Dataset Method Summary

A full description of this dataset is given in the paper of Fan et al. *(166). Pertinent details on the generation of this dataset are given here for information purposes (and do not represent my own work).*

Arabidopsis seedling root tissue samples were homogenized by grinding in an ice-cold grinding buffer consisting of 290 mM sucrose, 250 mM Tris-HCl (pH 7.6), 25 mM EDTA, 5 mM DTT, 1 mM PMSF, 0.5 × protease inhibitor cocktail (Roche, Indianapolis, IN), filtered through Miracloth, and the soluble protein fraction separated by centrifugation. Soluble proteins were recovered by acetone precipitation, pelleted by centrifugation and resuspended at a concentration of 8 μg/μL in 1 M urea/1 mM DTT which was then diluted to 1 μg/μL with 1 M urea/1 mM DTT/50 mM ammonium bicarbonate before digestion with trypsin (Promega, Madison, WI). The peptides were analysed using a Q-Exactive MS (Thermo Fisher Scientific, San Jose CA) with an ACQUITY UPLC BEH C18 column (1.0 mm × 150 mm, 1.7 μm particle size, Waters, Milford, MA). With buffer A as 99.9% water, 0.1% formic acid and buffer B as 99.9% acetonitrile, 0.1% formic acid, the solvent gradient program was as follows: 2%–10% buffer B (0–2 min.), 10–40% buffer B (2-62 min.), 40%–85% buffer B (62-63 min, then maintained for 10 min.). The column was equilibrated for 15 min. with 2% B prior to the next run. Data were acquired in DDA mode with MS^1 scans (range 350−1800 *m/z*) acquired at 70 k resolution and a target value based on predictive automatic gain control of 1×10^6 with 20 ms of maximum injection time. Based on an ion selection threshold of 1×10^4 counts, the 12 most intense precursor ions ($z \geq 2$) were isolated (2.0 *m/z* isolation width) and sequentially fragmented in the HCD collision cell with normalized collision energy of 30%. MS^2 scans were acquired with 35k resolution and a target value of 2×105 with 120 ms of maximum injection time. Selected precursor ion *m/z* values were dynamically excluded from further selection for 15 s.

4.2.2 Analysis of Dataset with *HeavyMetL*

Since the sample was not alkylated before digestion, the validity of the few cysteine-containing hits that were reported was difficult to evaluate. For simplicity, the list of identified peptides list was filtered to remove the 16 cysteine-containing entries and any peptides identified solely in the 16 h time point for which data were not available. This left a total of 1903 PMCs to be analysed across 6 time points - 0, 4, 8, 24, 32, 40 and 48 h. The list of peptides was re-formatted for compatibility with *HeavyMetL* input in accordance with the required column headers. Briefly, a new table was created based on the Scaffold report (item 2 in the list above). The following columns were directly copied across (Scaffold report column names given first, *HeavyMetL* input names given second; see Table 2-II for details): "Peptide sequence" as "PEPTIDE_SEQUENCE", "Protein accession" as "PROTEIN"; "Protein name" as "PROTEIN_DESCRIPTION"; "Peptide identification probability" as "PEPTIDE_SCORE"; "Protein identification probability" as "PROTEIN_SCORE"; "Variable modifications" as ”MODIFICATIONS"; "Spectrum charge" as "CHARGE"; "Exclusive" as "CONTRIBUTE_TO_PROTEIN". The columns "FILE_NAME" and "RETENTION_TIME" were added based on the data extracted from the Scaffold report column "Spectrum name".

The data were then analysed with *HeavyMetL* using the default settings (see Table 2-III). Analysis of the 1903 PMCs across 6 files (8GB in mzXML format) took 16 minutes for complete processing (Apple MacBook Pro running macOS 10.13.6; 2.7 GHz Intel Core i7 with 8 logical processors; 16 GB RAM; Firefox v. 62.0).

Subsequent statistical analysis of results and generation of all figures was performed in R (v. 3.5.1).

4.3 Benchmark Results

4.3.1 Overview

HeavyMetL reports both the label incorporation level and the apex intensity of the unlabelled and labelled peaks for each peptide and associated protein, using the protein grouping information provided in the input peptide list. Depending on the experimental design, either label incorporation or relative abundance or both may be used to infer protein turnover. I evaluated both measures at both peptide and protein level, comparing to the published quantification results from *ProteinTurnover* as a benchmark, taken from the output files published alongside the raw data.

Quantification results were filtered to a single result per unique peptide sequence to allow straightforward comparison against the *ProteinTurnover* results. In cases with multiple PMC instances sharing the same sequence (but different modification/charge states), the PMC with the highest mean intensity across all unlabelled and labelled signals was used. The final dataset used for comparison thus contained no duplicate peptide sequences, consisting of 1284 peptides with at least one successfully quantified time point. Although, in Section 3.9, I established that a Similarity Score threshold of 0.85 would likely be an effective way to further filter *HeavyMetL* results to improve dataset quality, I did not have a corresponding way to filter the *ProteinTurnover* dataset. For the fairest comparison, therefore, no Similarity Score filter was applied.

4.3.2 Peptide-Level Quantification Comparison

Since the data represent a biological system in flux, it cannot be assumed that all proteins in a time point sample will necessarily have the same level of incorporation. Assuming a similar level of incorporation estimation accuracy, the resulting distributions of labelled signal incorporations reported by different quantification algorithms should have similar properties. Comparing the peptide incorporation results reported by *HeavyMetL* to the *ProteinTurnover* benchmark, at timepoints after 8 h, there was good agreement, especially using only the top 75th percentile of data by frequency (in both datasets) in order to exclude the most obvious mis-quantifications (Pearson's $r = 0.57, 0.68, 0.76, 0.72$ for 24 h, 32 h, 40 h, 48 h respectively; Figure 4-I). In contrast, at 4 h and 8 h the data were not well correlated (Pearson's $r = 0.43$ and -0.11 for 4 h, 8h respectively), with both datasets showing clear evidence of a large amount of mis-quantification based on the wide spread of incorporations reported. Since the incorporation level of ^{15}N is expected to rise over time, these time points have the lowest ^{15}N incorporation and consequently the weakest labelled signal (and highest signal-to-noise); later time points have higher ^{15}N incorporation and thus stronger labelled signal and larger signal-to-noise ratio. Agreement between programs is therefore correlated with the expected signal-to-noise ratio, which is a logical outcome. Since the SS_{KL} method used by *HeavyMetL* to assess spectral matching is mathematically equivalent to distribution fitting using maximum likelihood estimation as performed by *ProteinTurnover,* then the algorithms may be expected to have similar weaknesses when confronted with low intensity or noisy data.

Nevertheless, there are still significant differences between the two algorithms, so it does not necessarily follow that peptides with a poorly assigned incorporation level in one method will also have a poorly assigned rate in the other. I therefore compared the distribution of incorporation levels and the unlabelled/labelled intensity ratios between the two methods (Figure 4-II). The distribution of incorporation (Figure 4-II-A,B) in the 4 h and 8 h time points indicate a greater degree of agreement between the two than suggested by the individual value comparisons in Figure 4-I, with a similar trend across median values despite the much greater spread of data.

Since an accurate determination of the incorporation is necessary for accurate determination of the labelled peak intensity, the unlabelled/labelled ratio quantification measures are, to an extent, dependent on the quality of the labelled incorporation level result. This is reflected in the distributions of reported unlabelled/labelled ratios across time points (Figure 4-II-C,D), which follow the same trend observed in the incorporation levels (Figure 4-II-A,B), and the median labelled signal intensity, as an approximation of signal-to-noise (Figure 4-II-E).

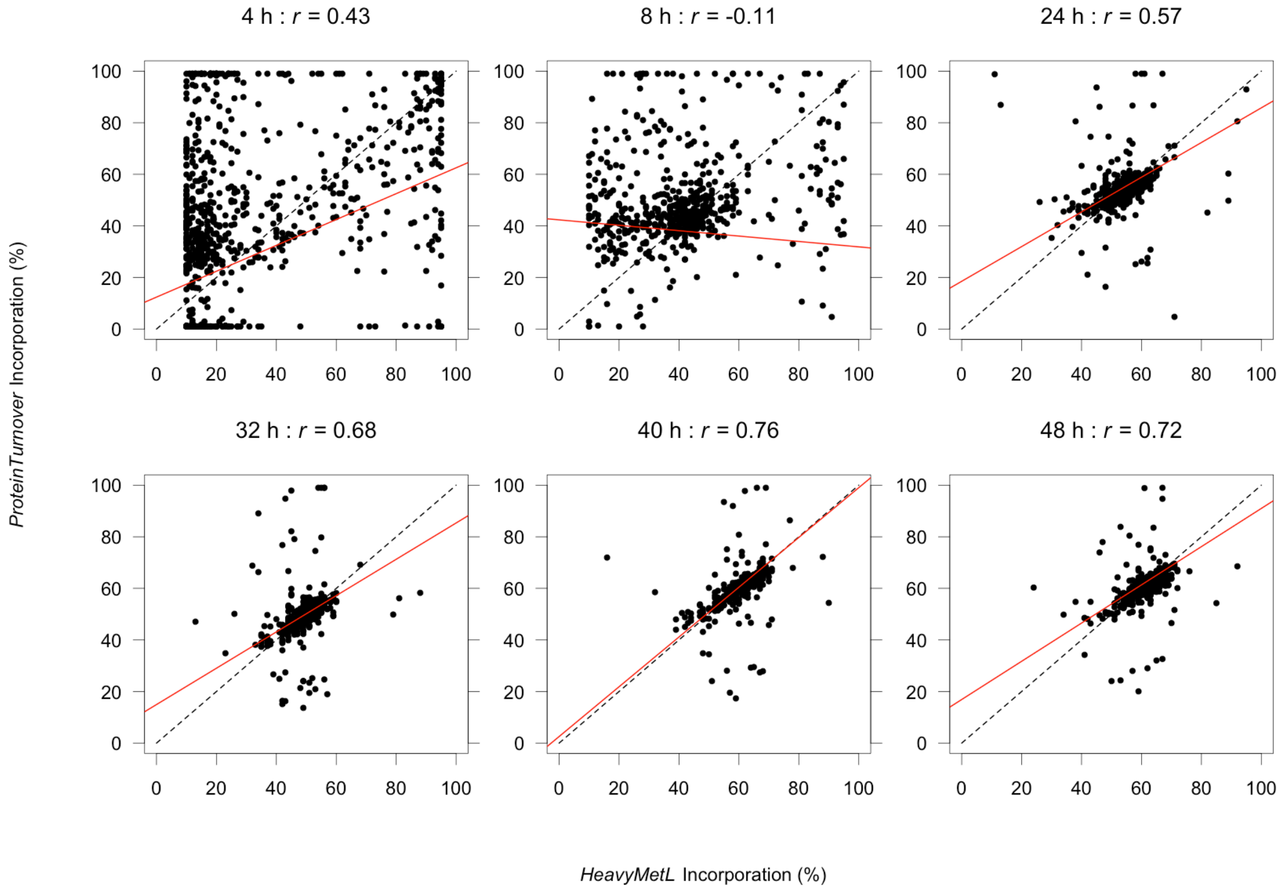


Figure 4-I. *Correlation of Labelled Signal Incorporation Values. Dotted lines indicate where perfect correlated data would lie. Pearson's* r *values given in plot titles. Solid red lines indicate a linear fit to values in the top 75% of data by frequency in both* HeavyMetL *and* ProteinTurnover.

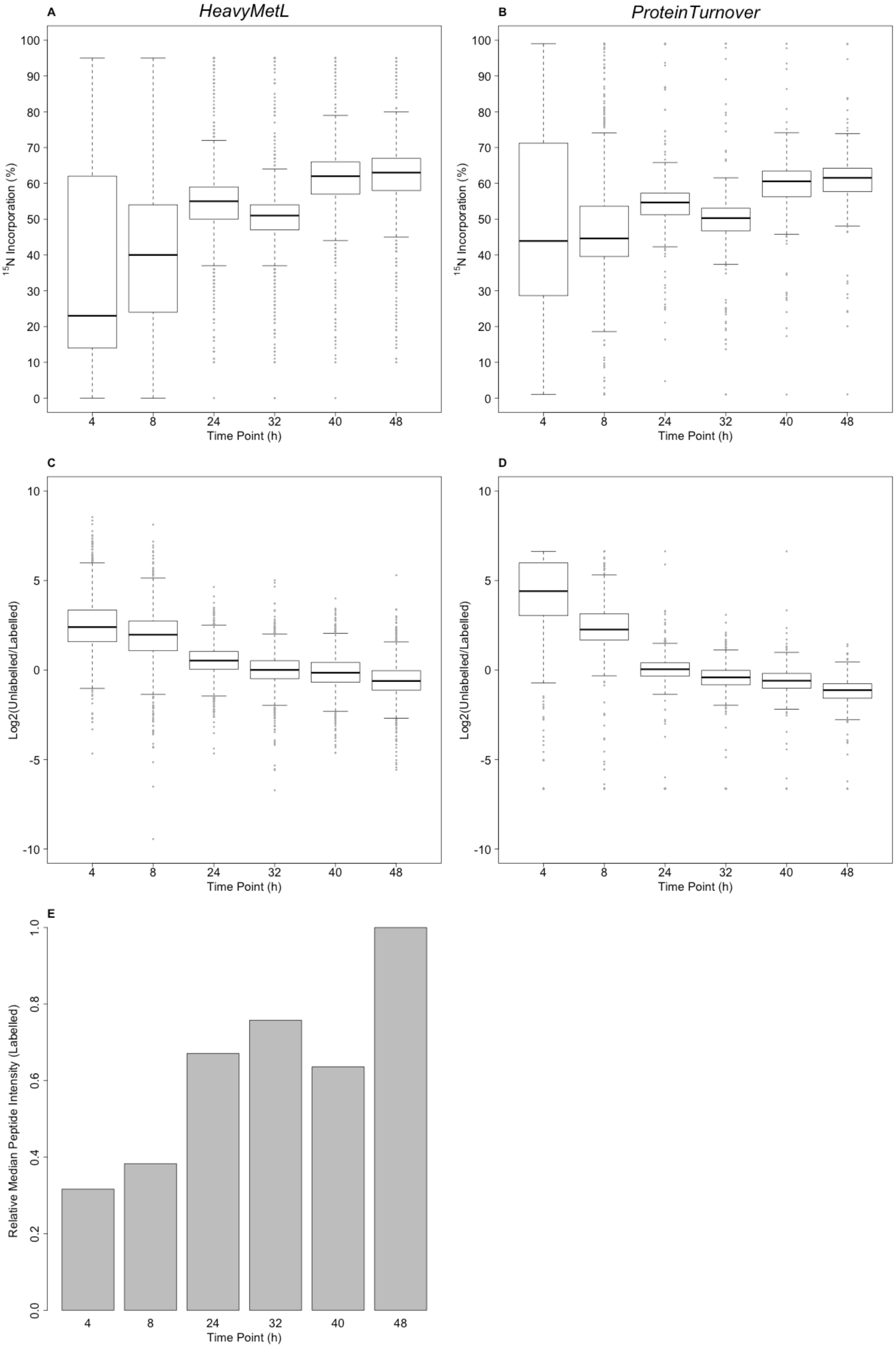


Figure 4-II. *Comparison of Peptide-Level Quantification Result Distributions.* ***A*** *&* ***B:*** *The distributions of labelled signal incorporation reported by* HeavyMetL *and* ProteinTurnover *respectively.* ***C*** *&* ***D:*** *The distributions of unlabelled/labelled signal ratios reported by* HeavyMetL *and* ProteinTurnover *respectively.* ***E****: Relative median intensity of labelled signal (as reported by* HeavyMetL*) in each time point.*

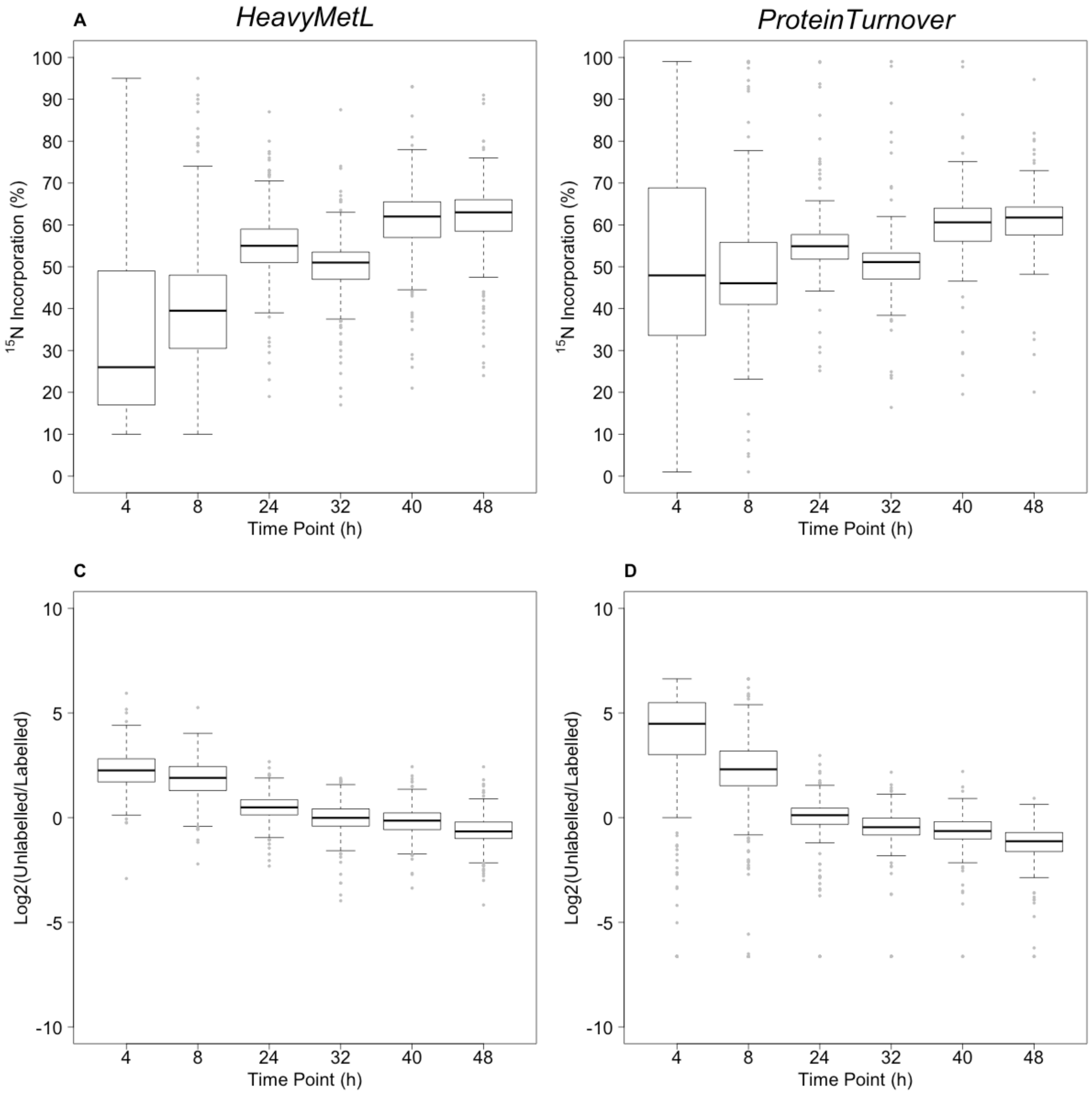


Figure 4-III. *Comparison of Protein-Level Quantification Result Distributions.* ***A & B:*** *The distributions of protein-level labelled signal incorporation reported by* HeavyMetL *and* ProteinTurnover *respectively.* ***C & D:*** *The distributions of protein-level unlabelled/labelled signal ratios reported by* HeavyMetL *and* ProteinTurnover *respectively.*

4.3.3 Protein-Level Quantification Comparison

Finally, the datasets were compared at the protein level. Protein level quantification was derived by taking the median peptide-level value for both metrics independently. As well as the median being a widely used summary statistic, this allowed the fairest comparison to the *ProteinTurnover* results, since *ProteinTurnover* reports only the relative proportion of unlabelled to labelled signal as the alpha parameter which precludes any intensity-weighted approaches. As might be expected, the results closely mirrored the peptide-level comparison, with some reduction in spread (which is to be expected, as taking the median peptide-level values naturally disfavours outlying mis-quantification). While performance of both algorithms was very similar, the incorporation values reported by *HeavyMetL* show a more sustained trend towards zero at low time points (Figure 4-III-A,B), which is consistent with the only *a priori* known data point, that ^{15}N Incorporation at 0 h must equal natural abundance, ~0.37%. This suggests that (at least using a median inference method) the values reported by *HeavyMetL* may, as a whole, be more accurate (although clearly subject to similar levels of variance). The tighter spread of unlabelled/labelled ratios reported by *HeavyMetL* in these time points (Figure 4-III-C,D) is consistent with this conclusion, as accurate determination of ratio relies on accurate determination of the incorporation level.

4.4 Conclusions

Notwithstanding the limitations of no manual optimisation, *HeavyMetL* produced comparable results to *ProteinTurnover*. The lack of correlation between reported labelled signal incorporation values with the benchmark results at low signal-to-noise (despite both approaches producing similar overall distributions) suggests that there is a proportion of these low-intensity peptides whose quantification could be further improved, since they are reported with lower error by *ProteinTurnover* (likewise, there is a similar proportion of peptides whose quantification by *HeavyMetL* has lower error than the corresponding *ProteinTurnover* result). It is interesting that while reported quantification at the low signal-to-noise timepoints (4 h and 8 h) had similarly high levels of variance for both algorithms, extrapolation to protein level results by the median did favour *HeavyMetL*, although there is clearly room for further optimisation of the algorithm to improve discrimination performance and eliminate unreliable peptide quantification at low signal-to-noise. It is possible that the combined labelled and unlabelled distribution matched by the *ProteinTurnover* algorithm tends to overvalue the contribution of the unlabelled isotopologues (whose incorporation and thus distribution is known *a priori*) resulting in an overestimation of the incorporation and intensity of the labelled signal at very low incorporations. The combined distribution matching may allow high intensity, easily quantified unlabelled signal to lend authenticity to what is essentially noise mis-identified as labelled signal.

Chapter 5: Concluding Discussion

In proteomic MS^1-based quantification, there are two methods of introducing a sample label *in vivo*. Labelling *in vivo* allows the earliest possible mixing of samples in a processing pathway and thus minimises the steps where two samples (or a sample and a standard) are processed independently and thus subject to differential technical variance). Labelling can be applied to specific amino acids (SILAC) or to all atoms of a particular element (EML, particularly ^{15}N labelling). SILAC is currently used far more widely; and I propose that this is due to two perceived advantages over ^{15}N:

1. **Ease of data processing.** Several mature, user-friendly software options exist for SILAC quantification, whereas the ^{15}N existing options are limited in capability, produce varied quality of quantification (especially for low intensity signals) and are not very easy to use by non-bioinformaticians. Computationally, EML quantification is more difficult. Since the distribution of isotopologues for the unlabelled and labelled SILAC peptides are nearly identical, the same parameters, mass-shifted, can be used to calculate a chromatographic maximum intensity or integral for both signals However, in EML the total signal in the labelled peptide is spread over a different number of isotopologues than in the unlabelled peptide, thus complicating quantification (as described in the Introduction) and increasing the potential for differential effects due to noise.
2. **A more easily achieved minimum 'quantifiable' incorporation level.** For ^{15}N, it is often suggested that partial labelling results in datasets that are intractable to analysis (particularly when there is differential partial labelling between proteins), and therefore that ^{15}N is only a viable labelling technique when a label incorporation level close to 100% can be achieved. It is technically difficult to achieve near 100% labelling given reagent purity, particularly when under further experimental constraints such as labelling time. This perception has the effect of (apparently) disqualifying ^{15}N labelling from consideration in many experiments.

As a consequence, ^{15}N is relegated to areas where SILAC has important limitations, most prominently in plant proteomics, where partial labelling of other amino acids occurs due to conversion pathways.

I further propose that a robust, easy to use quantification solution that can handle partial, non-uniform ^{15}N incorporation across proteins actually negates not only the first but also the second of the SILAC advantages described above. Although it is more challenging to quantify computationally, if this challenge can be addressed then complete separation of the unlabelled and labelled signal in EML requires only a sufficient incorporation such that the two isotopologue distributions do not substantially overlap for the majority of peptides. This is the case for most peptides at much less than 100% incorporation. Take as an example the set of theoretical peptides *n*AK, i.e. AK, AAK, AAAK and so on, with each increase in *n* adding one more alanine (and thus one more nitrogen atom). Impose a strict separation requirement that less than 1% of the labelled signal overlaps less than 1% of the total unlabelled signal. At *n*=3 (AAAK), 97% ^{15}N incorporation is needed to meet the requirement. However, at *n*=4, only 91% incorporation of ^{15}N is needed, while at *n*=5 (a peptide of 6 total residues, typically the smallest reliably observable by bottom-up proteomic MS), only 85% ^{15}N incorporation is sufficient. If ^{15}N incorporation need only be above 90% to ensure practically no overlap of signal for the majority of peptides, this is a substantially more achievable level of incorporation.

Accordingly, the barrier to wider ^{15}N usage might be argued to be the availability of software with robust performance and usability comparable to that of *MaxQuant* and commercial alternatives. Although there are no mature, user friendly software packages available that offer the ability to reliably quantify ^{15}N data with inconsistent, partial label incorporation, various approaches have been described in the literature (see Section 1.2.2) which suggest that, with further refinement, a robust quantification approach is achievable. The task then would be to implement this as a software package accessible to non-bioinformaticians and capable of performing such quantification at a rate that the users of quantification packages such as *MaxQuant* have come to expect.

At the beginning of Chapter 2, I set out a list of requirements that I considered necessary for such a software package and went on to describe *HeavyMetL*, a browser-based quantification tool which provides robust quantification in a user interface that fulfils the requirement list. It has a fully graphical user interface, allows in-tool visualisation of the matched spectra and chromatograms associated with quantitative results, and on a modern personal computer can process moderately sized MS datasets in the ‘minutes to hours’ timeframe typical of SILAC quantification tools. Unlike existing ^{15}N solutions, *HeavyMetL* is easy to set up (by simply visiting a web page in a supported browser), works across operating systems, has an accessible user interface that does not require use of the command line, and provides graphical output for quantification allowing rapid optimisation of quantification parameters such as *m/z* Error Tolerance.

While the implementation of *HeavyMetL* described in this thesis fulfils the usage requirements as originally set out, there are a number of technical and design aspects which are obvious candidates for future improvement. These fall into three areas:

1. **Improvements and extensions to the quantification algorithm.** Most obviously, the algorithm could be extended to support other types of EML, since the quantification approach is compatible with any elemental isotopic label (in the same way that SILAC-style quantification approaches can support Dimethyl labelling). Since the consideration of non-simultaneous unlabelled/labelled retention times is already supported by *HeavyMetL*, the only change required would be modification of the input to the theoretical distribution prediction algorithm to allow isotopic changes to elements other than nitrogen (the actual algorithm itself is sufficiently generalised to allow this already), and the associated changes to the user interface to allow the type of EML to be specified. A further adaptation would be to follow the proposal advanced by Fan *et al.* in *ProteinTurnover* that incorporation levels may differ between amino acids. Fan *et al.* modelled this process by allowing some variance in the probability of each nitrogen atom being labelled *via* a beta distribution (see Section 4.1), but a more representative way to model this (following the hypothesis to its conclusion) would be to generate the isotope distribution for each population of amino acids in the peptide separately (rather than combining them to a single pool of atoms). A range of theoretical distributions corresponding to different incorporations for each amino acid could be generated, then cross combined to yield peptide-level distributions. This would require generation of a huge number of possible distributions – the number of incorporation levels to be tested, raised to the power of the number of different amino acids in the peptide! Rather than generating all possible combinations, some form of multiple regression would have to be implemented. It would also likely be necessary to apply this optimised fitting only

to the final ‘best fit’ chromatographic maximum determined by the existing method, to avoid very extensive calculations.

2. **Code restructuring to take advantage of new language capabilities.** The JavaScript language standard is frequently updated, and some of these changes lift previous design constraints that necessitated performance concessions. The choice of JavaScript + browser as a platform will always limit the maximum quantification speed (although in many cases disk access speed will still be the bottleneck). Recently, a web standard for a binary instruction format that allows near-native execution speeds in browsers (WebAssembly) was announced. Porting calculation-heavy aspects of *HeavyMetL* into WebAssembly would likely allow further substantial speed improvements (or alternatively facilitate more nuanced quantification as described above without a reduction in performance).
3. **Improvements to management of multiple data files as a single dataset.** From a proteomics perspective, there are also some areas where the ‘gold standard’ in how multiple MS raw data files are managed as a single dataset has substantially advanced since *HeavyMetL* was designed. The approach of matching PMCs between MS raw data files *via* mean sample RT used in *HeavyMetL* is more sophisticated than previous ^{15}N approaches that require the time to be specified explicitly for all files. Even so, this is a relatively unsophisticated approach compared to the algorithms used in prominent label-free/SILAC quantification software such as *MaxQuant* and *Progenesis QIP* (see Section 1.1.2.11.3) and these are also now being outclassed by cutting-edge approaches such as machine learning. The difficulty in the case of ^{15}N is that retention time matching algorithms typically work on the assumption that peptides are represented by a limited range of isotopologue patterns such as found in LFQ and in SILAC, as these techniques do not affect the isotopologue distribution shape, but rather apply a constant mass shift. Thus, potential features can be identified by comparison to a small range of generalised isotopologue distributions representing peptides of increasing mass. In contrast, ^{15}N labelled peptides will not only frequently have a substantially different isotopologue distribution to that of a typical ‘unlabelled’ peptide, but the distribution will potentially differ from run-to-run with changes in ^{15}N incorporation. This makes run-to-run feature matching more difficult without knowing the incorporation level for each peptide in each run prior to quantification (which is what *HeavyMetL* is attempting to calculate; a chicken-and-egg problem).

Using machine learning to identify possible peptides regardless of incorporation and align them between runs may be a plausible solution to this, although would represent considerable development effort to achieve. Any retention time solution should also involve a more comprehensive solution for pre-fractionated experimental designs, as these are now the standard experimental approach for most proteomic investigations. The current necessity to analyse fraction groups individually in *HeavyMetL* is not up to par with, for example the experimental design options in *MaxQuant*.

4. **Improvements to the user interface.** The current interface is functional but there are quality-of-life improvements that could be made. Firstly, the input of the PMC list could be expanded to accept unmodified exports from major search engines (to avoid the currently necessary step of partially re-formatting the various different peptide-level identification tables produced by different search engines to the input format required by *HeavyMetL*. Alternatively, the program could allow the relevant columns in a text table file to be manually defined by the user (or perhaps selected from a list of templates) during import into *HeavyMetL* (rather than explicitly renaming columns according to the *HeavyMetL* specification shown in Table 2-II). This is an approach used by the *Spectronaut* and *SpectroDive* DIA analysis packages available from Biognosys. Secondly, the user configurable settings could be stored as a persistent browser object (a ‘cookie’, or the newer ‘LocalStorage’ specification) thus ensuring that settings were not lost when the browser window was closed.

In Chapter 3, I explored a specific technical aspect of the *HeavyMetL* quantification, namely the choice of Similarity Score used to compare the observed data with the set theoretical distributions I calculated for each PMC. Using an experimental dataset of an unlabelled sample of *O. tauri* and three labelled samples with target incorporations 40%, 50% and 60% ^{15}N, I first explored an orthogonal technique to estimate a global peptide average incorporation for each sample based on the relationship between peptide mass and its MDR, known as the half decimal place rule (HDPR). This method yielded an approximate estimate of the incorporation ratio between the three labelled samples but the results from *HeavyMetL* followed the expected ratio (given the original target values) much more closely. I concluded that the HDPR estimation was likely too sensitive to uncontrollable sources of bias (such as the stochastic selection of precursors in DDA) to provide accurate estimation of global incorporation values. Ultimately, the incorporation estimations were based on very small differences in the gradient of linear fits between mass and MDR, so even small differences due to noise or differential precursor selection could have large effects on the estimate. However, it might be possible to improve HDPR estimation with further data filters (e.g. precursor intensity). The HDPR method still has merit as an unbiased, non-search-result driven estimation of global incorporation in a sample, especially if combined with a peak-picking algorithm to identify probable peptides in MS^1 without the use of MS^2 scans to yield identities. Since peptide identifications are not required, the samples could be analysed on short gradients with MS^1 scans only, reducing MS analysis cost. A scenario where this might be applicable would be estimation of ^{15}N incorporation in cultures being labelled to stability, to monitor whether the incorporation level has stabilised; then at that point a full DDA analysis with identifications could then be performed and analysed with *HeavyMetL* (or another quantification solution).

I then compared four candidate Similarity Scores to see how my quantification workflow from Chapter 2 performed using each in turn as the spectral comparison score. The performances of the four Similarity Scores were very similar, which was not the result I had expected. On further consideration, however, this result was the most likely outcome. Given that the Similarity Scores are only applied for matching purposes to data that has already been extracted using predicted *m/z* values within a very narrow *m/z* window, the chances of extracting intensity associated with another co-eluting PMC are not high. Furthermore, the Similarity Score reported is that of the 'best' matching extraction, so for 'off-target' intensity (e.g. background noise or a co-eluting PMC) to be the bulk of the reported match spectrum, such a 'mostly noise' spectrum would have to yield a higher Similarity Score than the actual labelled target. Completely random matches are therefore relatively unlikely. The *m/z* extraction windows are themselves defined by theoretical prediction for a particular ^{15}N incorporation, so only incorporation values close to the 'true' incorporation will tend to yield substantial extracted intensity and a potentially viable match, regardless of Similarity Score.

With this in mind, it was important that the data from all three levels of target incorporation (samples A, B and C) be compared together, to maximise the sensitivity of the analysis and highlight subtle differences across score rank. The conversion of the incorporation percentage point errors to a Standardised Incorporation Quantification Error (SIQE) was therefore a useful tool to allow the results from all 3 incorporation levels to be directly compared. An experiment with more replicates (or more target incorporations) could have improved the robustness of the Similarity Score comparison by increasing the number of incorporation-score pair data points. Early builds of the *HeavyMetL* algorithm lacked a number of optimisations for speed, and when the experiment was designed it was not clear how many samples it would be practical to analyse, given that the analysis would have to be repeated four times with each of the four different candidate scores. A greater number of samples, with multiple technical replicates and more target incorporation points would in retrospect have been perfectly viable, with the overriding practical limitation being sample preparation time and cost of MS analysis rather than the feasibility of data analysis.

An alternative improvement would be a change of the experimental design. In the Similarity Score comparison, label incorporation was assumed to have stabilised at each target value (less a small amount to account for salt impurity), and thus be distributed tightly (and unimodally) around a mean value in each sample. This was not unreasonable, given previously observed incorporation behaviour over the same labelling period using the same labelling methods (160). Differences in turnover and amino acid composition, however, combined with potential kinetic imbalance between ^{14}N and ^{15}N incorporation in different amino acids, could potentially result in genuine differences in incorporation between peptides. The assumption made here was that this effect would be minor compared to the distribution of errors in the ^{15}N incorporation estimation, but this assumption is difficult to validate. One experimental approach to explore this might be to label until assumed incorporation stability, as above, and then sample across a time course where a pulse of an unlabelled, low-frequency amino acid such as tryptophan (Trp) was briefly introduced. PMCs containing Trp would then exhibit an increase in the variance of ^{15}N incorporation as the unlabelled Trp was incorporated, followed by a decrease as the unlabelled Trp was diluted out. The changes in Trp-containing peptide ^{15}N incorporation variance relative to the variance of incorporation of ^{15}N into peptides that do not contain Trp over the course of incorporation and dilution would yield useful information on the inherent (non-quantification error) incorporation background variance.

At the end of Chapter 3 I concluded that a score based on the Kullback-Leibler Divergence (SS_{KL}) was the optimum choice for the isotopologue distribution matching, in the context of the *HeavyMetL* algorithm. It is worth noting there that while SS_{KL} had the strongest overall performance, the slightly lower cumulative mean SIQE for the Euclidean distance-based scores at lower score ranks (despite their dire performance at high score ranks) suggests a possibility for further optimisation of the Similarity Score by combining both SS_{KL} and SS_{EUC}.

Finally, in Chapter 4, I benchmarked the quantification from *HeavyMetL* (using SS_{KL} for spectral comparison) against a recently published ^{15}N quantification method, *ProteinTurnover*, using the same incorporation time-course study in *Arabidopsis* reported in the *ProteinTurnover* publication. Overall, both methods gave similar results, with good agreement and similar distribution of data points in both peptide and protein-level comparisons in the later stages of the time course (where incorporation levels were higher). In the early stages of the time course (where incorporation levels were low), the performance was notably worse for both approaches, although *HeavyMetL* appeared to have an edge. I concluded that *HeavyMetL* has quantification performance clearly equivalent to or better than *ProteinTurnover,* which is the most recent and sophisticated ^{15}N quantification approach so far proposed.

Given the above-par quantification performance, combined with its advantages in terms of user accessibility and interaction, it does not seem unreasonable to advance the statement that *HeavyMetL* is a general improvement on the existing ^{15}N quantification solutions in the field. While it is not yet as mature as some extensively developed software packages such as *MaxQuant*, a non-bioinformatician could now analyse ^{15}N -labelled data using *HeavyMetL* and expect to generate usable, robust quantitative results, interacting solely *via* a relatively straightforward GUI. They can rely on seeing the same interface and having access to the same functionality whether they are on Windows, macOS or Linux, and can analyse any raw data that can be converted to (or exported as) mzML or mzXML, identifying the PMCs to be quantified by any method of their choice (so long as the result can be coerced into a text-table format with the necessary columns defined in Table 2-II). I therefore submit that *HeavyMetL* is not only a viable analysis tool, but that it potentially enables the usage of ^{15}N labelling outside of speciality niche areas, in more direct competition with SILAC and similar techniques.

Bibliography

1. Sanger, F.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Coulson, A. R.; Fiddes, C. A.; Hutchison, C. A.; Slocombe, P. M.; Smith, M., Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **1977,** 265, (5596), 687-95.

2. Wasinger, V. C.; Cordwell, S. J.; Cerpa-Poljak, A.; Yan, J. X.; Gooley, A. A.; Wilkins, M. R.; Duncan, M. W.; Harris, R.; Williams, K. L.; Humphery-Smith, I., Progress with gene-product mapping of the Mollicutes:Mycoplasma genitalium. *Electrophoresis* **1995,** 16, (1), 1090-1094.

3. Minde, D. P.; Dunker, A. K.; Lilley, K. S., Time, space, and disorder in the expanding proteome universe. *Proteomics* **2017,** 17, (7), 1600399-1600399.

4. Pratt, J. M.; Petty, J.; Riba-Garcia, I.; Robertson, D. H. L.; Gaskell, S. J.; Oliver, S. G.; Beynon, R. J., Dynamics of Protein Turnover, a Missing Dimension in Proteomics. *Molecular & Cellular Proteomics* **2002,** 1, (8), 579-591.

5. Bouvignies, G.; Vallurupalli, P.; Kay, L. E., Visualizing Side Chains of Invisible Protein Conformers by Solution NMR. *Journal of Molecular Biology* **2014,** 426, (3), 763-774.

6. Dunkley, T. P. J.; Watson, R.; Griffin, J. L.; Dupree, P.; Lilley, K. S., Localization of organelle proteins by isotope tagging (LOPIT). *Molecular & cellular proteomics : MCP* **2004,** 3, (11), 1128-34.

7. Gavin, A.-C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dümpelfeld, B.; Edelmann, A.; Heurtier, M.-A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A.-M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J. M.; Kuster, B.; Bork, P.; Russell, R. B.; Superti-Furga, G., Proteome survey reveals modularity of the yeast cell machinery. *Nature* **2006,** 440, (7084), 631-636.

8. Gerhardt, R.; Heldt, H. W., Measurement of Subcellular Metabolite Levels in Leaves by Fractionation of Freeze-Stopped Material in Nonaqueous Media. *Plant Physiology* **1984,** 75, (3), 542-547.

9. Hoermiller, I. I.; Naegele, T.; Augustin, H.; Stutz, S.; Weckwerth, W.; Heyer, A. G., Subcellular reprogramming of metabolism during cold acclimation in Arabidopsis thaliana. *Plant, Cell & Environment* **2017,** 40, (5), 602-610.

10. Burnette, W. N., "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Analytical Biochemistry* **1981,** 112, (2), 195-203.

11. Fields, S.; Song, O., A novel genetic system to detect protein-protein interactions. *Nature* **1989,** 340, (6230), 245-6.

12. Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nature Methods* **2009,** 6, (5), 359-362.

13. Hiemstra, T. F.; Charles, P. D.; Gracia, T.; Hester, S. S.; Gatto, L.; Al-Lamki, R.; Floto, R. A.; Su, Y.; Skepper, J. N.; Lilley, K. S.; Karet Frankl, F. E., Human urinary exosomes as innate immune effectors. *Journal of the American Society of Nephrology : JASN* **2014,** 25, (9), 2017-27.

14. Kaisar, M.; van Dullemen, L. F. A.; Thézénas, M.-L.; Zeeshan Akhtar, M.; Huang, H.; Rendel, S.; Charles, P. D.; Fischer, R.; Ploeg, R. J.; Kessler, B. M., Plasma degradome affected by variable storage of human blood. *Clinical proteomics* **2016,** 13, 26-26.

15. Kaisar, M.; van Dullemen, L. F. A.; Thézénas, M.-L.; Charles, P. D.; Ploeg, R. J.; Kessler, B. M., Plasma Biomarker Profile Alterations during Variable Blood Storage. *Clinical chemistry* **2016,** 62, (9), 1272-4.

16. Chen, F.; Welker, F.; Shen, C.-C.; Bailey, S. E.; Bergmann, I.; Davis, S.; Xia, H.; Wang, H.; Fischer, R.; Freidline, S. E.; Yu, T.-L.; Skinner, M. M.; Stelzer, S.; Dong, G.; Fu, Q.; Dong, G.; Wang, J.; Zhang, D.; Hublin, J.-J., A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **2019,** 569, (7756), 409-412.

17. Kavanaugh, J. S.; Flack, C. E.; Lister, J.; Ricker, E. B.; Ibberson, C. B.; Jenul, C.; Moormeier, D. E.; Delmain, E. A.; Bayles, K. W.; Horswill, A. R., Identification of extracellular DNA-binding proteins in the biofilm matrix. *mBio* **2019,** 10, (3), e01137-19.

18. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C., Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences of the United States of America* **1993,** 90, (11), 5011-5.

19. McLafferty, F. W.; Breuker, K.; Jin, M.; Han, X.; Infusini, G.; Jiang, H.; Kong, X.; Begley, T. P., Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. *FEBS Journal* **2007,** 274, (24), 6256-6268.

20. Chait, B. T., Mass Spectrometry: Bottom-Up or Top-Down? *Science* **2006,** 314, (5796).

21. Yates, J. R.; Ruse, C. I.; Nakorchevsky, A., Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annual Review of Biomedical Engineering* **2009,** 11, (1), 49-79.

22. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & Cellular Proteomics* **2005,** 4, (10), 1419-40.

23. Laemmli, U. K., Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970,** 227, (5259), 680-5.

24. Gundry, R. L.; White, M. Y.; Murray, C. I.; Kane, L. A.; Fu, Q.; Stanley, B. A.; Van Eyk, J. E., Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Current Protocols in Molecular Biology* **2009,** Chapter 10, (SUPPL. 88), Unit10.25-Unit10.25.

25. Roe, M. R.; Griffin, T. J., Gel-free mass spectrometry-based high throughput proteomics: tools for studying biological response of proteins and proteomes. *Proteomics* **2006,** 6, (17), 4678-87.

26. Picotti, P.; Rinner, O.; Stallmach, R.; Dautel, F.; Farrah, T.; Domon, B.; Wenschuh, H.; Aebersold, R., High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nature Methods* **2010,** 7, (1), 43-46.

27. Ren, D.; Pipes, G. D.; Liu, D.; Shih, L.-Y.; Nichols, A. C.; Treuheit, M. J.; Brems, D. N.; Bondarenko, P. V., An improved trypsin digestion method minimizes digestion-induced modifications on proteins. *Analytical Biochemistry* **2009,** 392, (1), 12-21.

28. Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **2019,** 16, (6), 509-518.

29. Burkhart, J. M.; Schumbrutzki, C.; Wortelkamp, S.; Sickmann, A.; Zahedi, R. P., Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *Journal of Proteomics* **2012,** 75, (4), 1454-1462.

30. Swaney, D. L.; Wenger, C. D.; Coon, J. J., Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *Journal of Proteome Research* **2010,** 9, (3), 1323-1329.

31. Ternette, N.; Yang, H.; Partridge, T.; Llano, A.; Cedeño, S.; Fischer, R.; Charles, P. D.; Dudek, N. L.; Mothe, B.; Crespo, M.; Fischer, W. M.; Korber, B. T. M.; Nielsen, M.; Borrow, P.; Purcell, A. W.; Brander, C.; Dorrell, L.; Kessler, B. M.; Hanke, T.,

Defining the HLA class I-associated viral antigen repertoire from HIV-1-infected human cells. *European Journal of Immunology* **2015,** 46, (1), 60-9.

32. Ternette, N.; Block, P. D.; Sánchez-Bernabéu, Á.; Borthwick, N.; Pappalardo, E.; Abdul-Jawad, S.; Ondondo, B.; Charles, P. D.; Dorrell, L.; Kessler, B. M.; Hanke, T., Early Kinetics of the HLA Class I-Associated Peptidome of MVA.HIVconsv-Infected Cells. *Journal of Virology* **2015,** 89, (11), 5760-71.

33. Giansanti, P.; Tsiatsiani, L.; Low, T. Y.; Heck, A. J. R., Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nature Protocols* **2016,** 11, (5), 993-1006.

34. Zubarev, R. A., The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **2013,** 13, (5), 723-726.

35. Di Palma, S.; Hennrich, M. L.; Heck, A. J. R.; Mohammed, S., Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. *Journal of Proteomics* **2012,** 75, (13), 3791-3813.

36. Berg, J. M.; Tymoczko, J. L.; Stryer, L., *Biochemistry*. 5th ed.; W.H. Freeman: New York, 2002.

37. Giddings, J. C., Maximum number of components resolvable by gel filtration and other elution chromatographic methods. *Analytical Chemistry* **1967,** 39, (8), 1027-1028.

38. Alpert, A. J., Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of Chromatography* **1990,** 499, 177-96.

39. Dwivedi, R. C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V., Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Analytical Chemistry* **2008,** 80, (18), 7036-7042.

40. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *Journal of Separation Science* **2005,** 28, (14), 1694-703.

41. Davis, S.; Charles, P. D.; He, L.; Mowlds, P.; Kessler, B. M.; Fischer, R., Expanding Proteome Coverage with CHarge Ordered Parallel Ion aNalysis (CHOPIN) Combined with Broad Specificity Proteolysis. *Journal of Proteome Research* **2017,** 16, (3), 1288-1299.

42. Taylor, G., Disintegration of Water Drops in an Electric Field. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **1964,** 280, (1382), 383-397.

43. Steen, H.; Mann, M., The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology* **2004,** 5, (9), 699-711.

44. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989,** 246, (4926), 64-71.

45. Wilm, M.; Mann, M., Analytical Properties of the Nanoelectrospray Ion Source. *Analytical Chemistry* **1996,** 68, (1), 1-8.

46. Taflin, D. C.; Ward, T. L.; Davis, E. J., Electrified droplet fission and the Rayleigh limit. *Langmuir* **1989,** 5, (2), 376-384.

47. Iribarne, J. V., On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics* **1976,** 64, (6), 2287-2287.

48. Dole, M., Molecular Beams of Macroions. *The Journal of Chemical Physics* **1968,** 49, (5), 2240-2240.

49. Winger, B. E.; Light-Wahl, K. J.; Ogorzalek Loo, R. R.; Udseth, H. R.; Smith, R. D., Observation and implications of high mass-to-charge ratio ions from electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **1993,** 4, (7), 536-545.

50. Fernandez De La Mora, J., Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Analytica Chimica Acta* **2000,** 406, (1), 93-104.

51. Loscertales, I. G.; Fernández de la Mora, J., Experiments on the kinetics of field evaporation of small ions from droplets. *The Journal of Chemical Physics* **1995,** 103, (12), 5041-5060.

52. Nguyen, S.; Fenn, J. B., Gas-phase ions of solute species from charged droplets of solutions. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** 104, (4), 1111-7.

53. Touboul, D.; Jecklin, M. C.; Zenobi, R., Ion internal energy distributions validate the charge residue model for small molecule ion formation by spray methods. *Rapid Communications in Mass Spectrometry* **2008,** 22, (7), 1062-1068.

54. Kebarle, P., A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *Journal of Mass Spectrometry* **2000,** 35, (7), 804-817.

55. Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S., Unraveling the Mechanism of Electrospray Ionization. *Analytical Chemistry* **2013,** 85, (1), 2-9.

56. Karas, M.; Bachmann, D.; Hillenkamp, F., Influence of the Wavelength in High-Irradiance Ultraviolet Laser Desorption Mass Spectrometry of Organic Molecules. *Analytical Chemistry* **1985,** 57, (14), 2935-2939.

57. Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F., Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes* **1987,** 78, (null), 53-68.

58. Wuhrer, M.; Hokke, C. H.; Deelder, A. M., Glycopeptide analysis by matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry reveals novel features of horseradish peroxidase glycosylation. *Rapid Communications in Mass Spectrometry* **2004,** 18, (15), 1741-1748.

59. Uematsu, R.; Furukawa, J.-i.; Nakagawa, H.; Shinohara, Y.; Deguchi, K.; Monde, K.; Nishimura, S.-I., High throughput quantitative glycomics and glycoform-focused proteomics of murine dermis and epidermis. *Molecular & Cellular Proteomics* **2005,** 4, (12), 1977-1989.

60. de Hoffmann, E.; Stroobant, V., *Mass spectrometry : principles and applications*. J. Wiley: 2007; p 489-489.

61. Ladislas Wiza, J., Microchannel plate detectors. *Nuclear Instruments and Methods* **1979,** 162, (1-3), 587-601.

62. McNaught, A. D.; Wilkinson, A., Resolution in Mass Spectroscopy. In Second Edi ed.; Blackwell Scientific Publications: Oxford, 1997.

63. Steel, C.; Henchman, M., Understanding the Quadrupole Mass Filter through Computer Simulation. *Journal of Chemical Education* **1998,** 75, (8), 1049-1049.

64. Paul, W.; Steinwedel, H., Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift für Naturforschung A* **1953,** 8, (7), 448-450.

65. Makarov, A., Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry* **2000,** 72, (6), 1156-1162.

66. Kingdon, K., A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. *Physical Review* **1923,** 21, (4), 408-418.

67. Brown, L. S.; Gabrielse, G., Geonium theory: Physics of a single electron or ion in a Penning trap. *Reviews of Modern Physics* **1986,** 58, (1), 233-311.

68. Mamyrin, B. A.; Karataev, V. I.; Shmikk, D. V.; Zagulin, V. A., The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Soviet Physics JETP* **1973,** 37.

69. Sleno, L.; Volmer, D. A., Ion activation methods for tandem mass spectrometry. *Journal of Mass Spectrometry* **2004,** 39, (10), 1091-1112.

70. Roepstorff, P.; Fohlman, J., Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry* **1984,** 11, (11), 601-601.

71. Cox, J.; Hubner, N. C.; Mann, M., How Much Peptide Sequence Information Is Contained in Ion Trap Tandem Mass Spectra? *Journal of the American Society for Mass Spectrometry* **2008,** 19, (12), 1813-1820.

72. Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A., Mobile and localized protons: A framework for understanding peptide dissociation. *Journal of Mass Spectrometry* **2000,** 35, (12), 1399-1406.

73. Xia, Y.; Liang, X.; McLuckey, S. A., Ion trap versus low-energy beam-type collision-induced dissociation of protonated ubiquitin ions. *Analytical Chemistry* **2006,** 78, (4), 1218-1227.

74. Li, S.; Arnold, R. J.; Tang, H.; Radivojac, P., On the accuracy and limits of peptide fragmentation spectrum prediction. *Analytical Chemistry* **2011,** 83, (3), 790-796.

75. Kelchtermans, P.; Bittremieux, W.; De Grave, K.; Degroeve, S.; Ramon, J.; Laukens, K.; Valkenborg, D.; Barsnes, H.; Martens, L., Machine learning applications in proteomics research: How the past can boost the future. *Proteomics* **2014,** 14, (4-5), 353-366.

76. Zolg, D. P.; Wilhelm, M.; Gessulat, S.; Schmidt, T. K.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Navarro, P.; Delanghe, B.; Huhmer, A.; Kuster, B. In *ProteomeTools: Progress on the generation of reference peptides and spectra for the human proteome*, 66th ASMS Conference on Mass Spectrometry and Allied Topics; San Diego, CA, 2018.

77. Cimermancic, P.; Levy, R.; Palaniappan, K.; Salinas, F.; Tiwary, S.; Gutenbrunner, P.; Cox, J. In *High Quality Peptide MS/MS Spectrum Prediction Using Deep Learning and Its Application in DIA Data Analysis*, 66th ASMS Conference on Mass Spectrometry and Allied Topics; San Diego, CA, 2018.

78. Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W., Electron capture dissociation for

structural characterization of multiply charged protein cations. *Analytical Chemistry* **2000,** 72, (3), 563-73.

79. Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **2004,** 101, (26), 9528-33.

80. Ko, B. J.; Brodbelt, J. S., Enhanced electron transfer dissociation of peptides modified at C-terminus with fixed charges. *Journal of the American Society for Mass Spectrometry* **2012,** 23, (11), 1991-2000.

81. Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S., Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical Chemistry* **2006,** 78, (7), 2113-2120.

82. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **2008,** 26, (12), 1367-72.

83. Plumb, R. S.; Johnson, K. A.; Rainville, P.; Smith, B. W.; Wilson, I. D.; Castro-Perez, J. M.; Nicholson, J. K., UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry* **2006,** 20, (13), 1989-94.

84. Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics* **2012,** 11, (6), O111.016717-O111.016717.

85. Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R., Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods* **2004,** 1, (1), 39-45.

86. Kuhn, E.; Wu, J.; Karl, J.; Liao, H.; Zolg, W.; Guild, B., Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. *Proteomics* **2004,** 4, (4), 1175-1186.

87. Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J., Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Molecular & Cellular Proteomics* **2012,** 11, (11), 1475-1488.

88. Pappin, D. J.; Hojrup, P.; Bleasby, A. J., Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* **1993,** 3, (6), 327-32.

89. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** 5, (11), 976-989.

90. Park, C. Y.; Klammer, A. A.; Käll, L.; MacCoss, M. J.; Noble, W. S., Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research* **2008,** 7, (7), 3022-7.

91. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-7.

92. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search algorithm. *Journal of Proteome Research* **2004,** 3, (5), 958-64.

93. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **2011,** 10, (4), 1794-1805.

94. Nesvizhskii, A. I.; Vitek, O.; Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* **2007,** 4, (10), 787-797.

95. Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C., Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *Journal of Proteome Research* **2006,** 5, (8), 1843-1849.

96. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007,** 7, (5), 655-67.

97. Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J., Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical Chemistry* **2006,** 78, (16), 5678-84.

98. Frank, A.; Pevzner, P., PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry* **2005,** 77, (4), 964-973.

99. Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B., PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics* **2012,** 11, (4), M111.010587-M111.010587.

100. Bern, M.; Kil, Y. J.; Becker, C., Byonic: Advanced peptide and protein identification software. *Current Protocols in Bioinformatics* **2012,** Chapter 13, (SUPPL.40), Unit13.20-Unit13.20.

101. Rappsilber, J.; Mann, M., What does it mean to identify a protein in proteomics? *Trends in Biochemical Sciences* **2002,** 27, (2), 74-8.

102. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry* **2003,** 75, (17), 4646-4658.

103. Huang, T.; Zhang, W.; Dai, X.; Zhang, X.; Quan, C.; Li, H.; Yang, Y., Precise measurement for the purity of amino acid and peptide using quantitative nuclear magnetic resonance. *Talanta* **2014,** 125, 94-101.

104. Josephs, R. D.; Stoppacher, N.; Daireaux, A.; Choteau, T.; Lippa, K. A.; Phinney, K. W.; Westwood, S.; Wielgosz, R. I., State-of-the-art and trends for the SI traceable value assignment of the purity of peptides using the model compound angiotensin I. *TrAC - Trends in Analytical Chemistry* **2018,** 101, 108-119.

105. Melanson, J. E.; Thibeault, M.-P.; Stocks, B. B.; Leek, D. M.; McRae, G.; Meija, J., Purity assignment for peptide certified reference materials by combining qNMR and LC-MS/MS amino acid analysis results: application to angiotensin II. *Analytical and Bioanalytical Chemistry* **2018,** 410, (26), 6719-6731.

106. Welle, K. A.; Zhang, T.; Hryhorenko, J. R.; Shen, S.; Qu, J.; Ghaemmaghami, S., Time-resolved Analysis of Proteome Dynamics by Tandem Mass Tags and Stable Isotope Labeling in Cell Culture (TMT-SILAC) Hyperplexing. *Molecular & Cellular Proteomics* **2016,** 15, (12), 3551-3563.

107. Russell, M. R.; Lilley, K. S., Pipeline to assess the greatest source of technical variance in quantitative proteomics using metabolic labelling. *Journal of Proteomics* **2012,** 77, 441-454.

108. Ong, S.-E.; Mann, M., Mass spectrometry–based proteomics turns quantitative. *Nature Chemical Biology* **2005,** 1, (5), 252-262.

109. Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry* **2012,** 404, (4), 939-965.

110. Turowski, M.; Yamakawa, N.; Meller, J.; Kimata, K.; Ikegami, T.; Hosoya, K.; Tanaka, N.; Thornton, E. R., Deuterium Isotope Effects on Hydrophobic Interactions: The

Importance of Dispersion Interactions in the Hydrophobic Phase. *Journal of the American Chemical Society* **2003,** 125, (45), 13836-13849.

111. Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics* **2002,** 1, (5), 376-86.

112. Bindschedler, L. V.; Cramer, R., Fully automated software solution for protein quantitation by global metabolic labeling with stable isotopes. *Rapid Communications in Mass Spectrometry* **2011,** 25, (11), 1461-1471.

113. Sury, M. D.; Chen, J.-X.; Selbach, M., The SILAC fly allows for accurate protein quantification in vivo. *Molecular & Cellular Proteomics* **2010,** 9, (10), 2173-83.

114. Zanivan, S.; Krueger, M.; Mann, M., In vivo quantitative proteomics: The SILAC mouse. *Methods in Molecular Biology* **2011,** 757, 435-450.

115. Geiger, T.; Cox, J.; Ostasiewicz, P.; Wisniewski, J. R.; Mann, M., Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature Methods* **2010,** 7, (5), 383-385.

116. Hinkson, I. V.; Elias, J. E., The dynamic state of protein turnover: It's about time. *Trends in Cell Biology* **2011,** 21, (5), 293-303.

117. Van Hoof, D.; Pinkse, M. W. H.; Oostwaard, D. W.-V.; Mummery, C. L.; Heck, A. J. R.; Krijgsveld, J., An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics. *Nature Methods* **2007,** 4, (9), 677-678.

118. Bendall, S. C.; Hughes, C.; Stewart, M. H.; Doble, B.; Bhatia, M.; Lajoie, G. A., Prevention of amino acid conversion in SILAC experiments with embryonic stem cells. *Molecular & Cellular Proteomics* **2008,** 7, (9), 1587-97.

119. Lößner, C.; Warnken, U.; Pscherer, A.; Schnölzer, M., Preventing arginine-to-proline conversion in a cell-line-independent manner during cell cultivation under stable isotope labeling by amino acids in cell culture (SILAC) conditions. *Analytical Biochemistry* **2011,** 412, (1), 123-125.

120. Fröhlich, F.; Christiano, R.; Walther, T. C., Native SILAC: metabolic labeling of proteins in prototroph microorganisms based on lysine synthesis regulation. *Molecular & Cellular Proteomics* **2013,** 12, (7), 1995-2005.

121. Park, S. K.; Venable, J. D.; Xu, T.; Yates, J. R., A quantitative analysis software tool for mass spectrometry-based proteomics. *Nature Methods* **2008,** 5, (4), 319-322.

122. van Breukelen, B.; van den Toorn, H. W. P.; Drugan, M. M.; Heck, A. J. R., StatQuant: A post-quantification analysis toolbox for improving quantitative mass spectrometry. *Bioinformatics* **2009,** 25, (11), 1472-1473.

123. Mortensen, P.; Gouw, J. W.; Olsen, J. V.; Ong, S. E.; Rigbolt, K. T. G.; Bunkenborg, J.; Cox, J.; Foster, L. J.; Heck, A. J. R.; Blagoev, B.; Andersen, J. S.; Mann, M., MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *Journal of Proteome Research* **2010,** 9, (1), 393-403.

124. Naumann, B.; Busch, A.; Allmer, J.; Ostendorf, E.; Zeller, M.; Kirchhoff, H.; Hippler, M., Comparative quantitative proteomics to investigate the remodeling of bioenergetic pathways under iron deficiency in Chlamydomonas reinhardtii. *PROTEOMICS* **2007,** 7, (21), 3964-3979.

125. Nelson, C. J.; Li, L.; Millar, A. H., Quantitative analysis of protein turnover in plants. *PROTEOMICS* **2014,** 14, (4-5), 579-592.

126. Gruhler, A.; Schulze, W. X.; Matthiesen, R.; Mann, M.; Jensen, O. N., Stable isotope labeling of Arabidopsis thaliana cells and quantitative proteomics by mass spectrometry. *Molecular & Cellular Proteomics* **2005,** 4, (11), 1697-709.

127. Lewandowska, D.; ten Have, S.; Hodge, K.; Tillemans, V.; Lamond, A. I.; Brown, J. W. S., Plant SILAC: Stable-Isotope Labelling with Amino Acids of Arabidopsis Seedlings for Quantitative Proteomics. *PLoS ONE* **2013,** 8, (8), e72207-e72207.

128. Stewart, I. I.; Thomson, T.; Figeys, D., 18O Labeling: a tool for proteomics. *Rapid Communications in Mass Spectrometry* **2001,** 15, (24), 2456-2465.

129. Ye, X.; Luke, B.; Andresson, T.; Blonder, J., 18O stable isotope labeling in MS-based proteomics. *Briefings in Functional Genomics & Proteomics* **2009,** 8, (2), 136-44.

130. Rachdaoui, N.; Austin, L.; Kramer, E.; Previs, M. J.; Anderson, V. E.; Kasumov, T.; Previs, S. F., Measuring Proteome Dynamics in Vivo: AS EASY AS ADDING WATER? *Molecular & Cellular Proteomics* **2009,** 8, (12), 2653-2663.

131. Snijders, A. P. L.; De Vos, M. G. J.; Wright, P. C., Novel approach for peptide quantitation and sequencing based on 15N and 13C metabolic labeling. *Journal of Proteome Research* **2005,** 4, (2), 578-585.

132. Palmblad, M.; Bindschedler, L. V.; Cramer, R., Quantitative proteomics using uniform15N-labeling, MASCOT, and the trans-proteomic pipeline. *PROTEOMICS* **2007,** 7, (19), 3462-3469.

133. Schoenheimer, R.; Rittenberg, D.; Fox, M.; Keston, A. S.; Ratner, S., The Nitrogen Isotope (N 15) as a Tool In The Study Of The Intermediary Metabolism Of Nitrogenous Compounds. *Journal of the American Chemical Society* **1937,** 59, (9), 1768-1768.

134. Schoenheimer, R.; Rittenberg, D.; Foster, G. L.; Keston, A. S.; Ratner, S., The application of the nitrogen isotope N15 for the study of protein metabolism. *Science* **1938,** 88, (2295), 599-600.

135. Arias, I. M.; Doyle, D.; Schimke, R. T., Studies on the synthesis and degradation of proteins of the endoplasmic reticulum of rat liver. *Journal of Biological Chemistry* **1969,** 244, (12), 3303-15.

136. Larrabee, K. L.; Phillips, J. O.; Williams, G. J.; Larrabee, A. R., The relative rates of protein synthesis and degradation in a growing culture of Escherichia coli. *Journal of Biological Chemistry* **1980,** 255, (9), 4125-30.

137. Mosteller, R. D.; Goldstein, R. V.; Nishimoto, K. R., Metabolism of individual proteins in exponentially growing Escherichia coli. *Journal of Biological Chemistry* **1980,** 255, (6), 2524-32.

138. Gouw, J. W.; Tops, B. B. J.; Krijgsveld, J., Metabolic labeling of model organisms using heavy nitrogen (15N). *Methods in molecular biology (Clifton, N.J.)* **2011,** 753, 29-42.

139. Hughes, C.; Krijgsveld, J., Developments in quantitative mass spectrometry for the analysis of proteome dynamics. *Trends in Biotechnology* **2012,** 30, (12), 668-676.

140. Batista Silva, W.; Daloso, D. M.; Fernie, A. R.; Nunes-Nesi, A.; Araújo, W. L., Can stable isotope mass spectrometry replace radiolabelled approaches in metabolic studies? *Plant Science* **2016,** 249, 59-69.

141. Freund, D. M.; Hegeman, A. D., Recent advances in stable isotope-enabled mass spectrometry-based plant metabolomics. *Current Opinion in Biotechnology* **2017,** 43, 41-48.

142. Murrell, J. C.; Radajewski, S.; Ineson, P.; Parekh, N. R., Stable-isotope probing as a tool in microbial ecology. *Nature* **2000,** 403, (6770), 646-649.

143. Gouw, J. W.; Tops, B. B. J.; Mortensen, P.; Heck, A. J. R.; Krijgsveld, J., Optimizing identification and quantitation of 15N-labeled proteins in comparative proteomics. *Analytical Chemistry* **2008,** 80, (20), 7796-7803.

144. Antonov, V. K.; Ginodman, L. M.; Rumsh, L. D.; Kapitannikov, Y. V.; Barshevskaya, T. N.; Yavashev, L. P.; Gurova, A. G.; Volkova, L. I., Studies on the

mechanisms of action of proteolytic enzymes using heavy oxygen exchange. *European Journal of Biochemistry / FEBS* **1981,** 117, (1), 195-200.

145. Desiderio, D. M.; Kai, M., Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. *Biological Mass Spectrometry* **1983,** 10, (8), 471-479.

146. Mirgorodskaya, O. A.; Kozmin, Y. P.; Titov, M. I.; Körner, R.; Sönksen, C. P.; Roepstorff, P., Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using18O-labeled internal standards. *Rapid Communications in Mass Spectrometry* **2000,** 14, (14), 1226-1232.

147. Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C., Proteolytic 18 O Labeling for Comparative Proteomics: Model Studies with Two Serotypes of Adenovirus. *Analytical Chemistry* **2001,** 73, (13), 2836-2842.

148. Münchbach, M.; Quadroni, M.; Miotto, G.; James, P., Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Analytical Chemistry* **2000,** 72, (17), 4047-4057.

149. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* **1999,** 17, (10), 994-999.

150. Hsu, J. L.; Huang, S. Y.; Chow, N. H.; Chen, S. H., Stable-isotope dimethyl labeling for quantitative proteomics. *Analytical Chemistry* **2003,** 75, (24), 6843-52.

151. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics* **2004,** 3, (12), 1154-1169.

152. Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C., Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry* **2003,** 75, (8), 1895-1904.

153. Ting, L.; Rad, R.; Gygi, S. P.; Haas, W., MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods* **2011,** 8, (11), 937-40.

154. McAlister, G. C.; Nusinow, D. P.; Jedrychowski, M. P.; Wühr, M.; Huttlin, E. L.; Erickson, B. K.; Rad, R.; Haas, W.; Gygi, S. P., MultiNotch MS3 Enables Accurate,

Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Analytical Chemistry* **2014,** 86, (14), 7150-8.

155. Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P., Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *Journal of Proteome Research* **2006,** 5, (9), 2339-47.

156. Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M., Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Molecular & Cellular Proteomics* **2005,** 4, (9), 1265-1272.

157. Trudgian, D. C.; Ridlova, G.; Fischer, R.; Mackeen, M. M.; Ternette, N.; Acuto, O.; Kessler, B. M.; Thomas, B., Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics* **2011,** 11, (14), 2790-7.

158. Ahrné, E.; Molzahn, L.; Glatter, T.; Schmidt, A., Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **2013,** 13, (17), 2567-78.

159. Le Bihan, T.; Martin, S. F.; Chirnside, E. S.; van Ooijen, G.; Barrios-Llerena, M. E.; O'Neill, J. S.; Shliaha, P. V.; Kerr, L. E.; Millar, A. J., Shotgun proteomic analysis of the unicellular alga Ostreococcus tauri. *Journal of Proteomics* **2011,** 74, (10), 2060-2070.

160. Martin, S. F.; Munagapati, V. S.; Salvo-Chirnside, E.; Kerr, L. E.; Le Bihan, T., Proteome Turnover in the Green Alga Ostreococcus tauri by Time Course 15 N Metabolic Labeling Mass Spectrometry. *Journal of Proteome Research* **2012,** 11, (1), 476-486.

161. Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T., Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** 96, (12), 6591-6.

162. Zybailov, B.; Coleman, M. K.; Florens, L.; Washburn, M. P., Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Analytical Chemistry* **2005,** 77, (19), 6218-24.

163. Kolkman, A.; Daran-Lapujade, P.; Fullaondo, A.; Olsthoorn, M. M. A.; Pronk, J. T.; Slijper, M.; Heck, A. J. R., Proteome analysis of yeast response to various nutrient limitations. *Molecular Systems Biology* **2006,** 2, 2006.0026-2006.0026.

164. de Groot, M. J. L.; Daran-Lapujade, P.; van Breukelen, B.; Knijnenburg, T. A.; de Hulster, E. A. F.; Reinders, M. J. T.; Pronk, J. T.; Heck, A. J. R.; Slijper, M., Quantitative

proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes. *Microbiology* **2007,** 153, (Pt 11), 3864-78.

165. Price, J. C.; Guan, S.; Burlingame, A.; Prusiner, S. B.; Ghaemmaghami, S., Analysis of proteome dynamics in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** 107, (32), 14508-13.

166. Fan, K.-T.; Rendahl, A. K.; Chen, W.-P.; Freund, D. M.; Gray, W. M.; Cohen, J. D.; Hegeman, A. D., Proteome Scale-Protein Turnover Analysis Using High Resolution Mass Spectrometric Data from Stable-Isotope Labeled Plants. *Journal of Proteome Research* **2016**.

167. Kubinyi, H., Calculation of isotope distributions in mass spectrometry. A trivial solution for a non-trivial problem. *Analytica Chimica Acta* **1991,** 247, (1), 107-119.

168. Rockwood, A. L.; Van Orden, S. L.; Smith, R. D., Rapid Calculation of Isotope Distributions. *Analytical Chemistry* **1995,** 67, (15), 2699-2704.

169. Rockwood, A. L.; Orden, S. L. V., Ultrahigh-Speed Calculation of Isotope Distributions. *Analytical Chemistry* **1996,** 68, (13), 2027-30.

170. Rockwood, A. L.; Van Orman, J. R.; Dearden, D. V., Isotopic compositions and accurate masses of single isotopic peaks. *Journal of the American Society for Mass Spectrometry* **2004,** 15, (1), 12-21.

171. Rockwood, A. L.; Haimi, P., Efficient Calculation of Accurate Masses of Isotopic Peaks. *Journal of the American Society for Mass Spectrometry* **2006,** 17, (3), 415-419.

172. Nelson, C. J.; Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R., Implications of 15N-metabolic labeling for automated peptide identification in Arabidopsis thaliana. *PROTEOMICS* **2007,** 7, (8), 1279-1292.

173. Andersen, J. S.; Wilkinson, C. J.; Mayor, T.; Mortensen, P.; Nigg, E. A.; Mann, M., Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **2003,** 426, (6966), 570-4.

174. Schulze, W. X.; Mann, M., A Novel Proteomic Screen for Peptide-Protein Interactions. *Journal of Biological Chemistry* **2004,** 279, (11), 10756-10764.

175. Andreev, V. P.; Li, L.; Rejtar, T.; Li, Q.; Ferry, J. G.; Karger, B. L., New algorithm for 15N/14N quantitation with LC-ESI-MS using an LTQ-FT mass spectrometer. *Journal of Proteome Research* **2006,** 5, (8), 2039-2045.

176. Keller, A.; Eng, J.; Zhang, N.; Li, X.-j.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology* **2005,** 1, (1), E1-E8.

177. Hebeler, R.; Oeljeklaus, S.; Reidegeld, K. A.; Eisenacher, M.; Stephan, C.; Sitek, B.; Stu, K.; Meyer, H. E.; Sturre, M. J. G.; Dijkwel, P. P.; Warscheid, B.; Stühler, K.; Meyer, H. E.; Sturre, M. J. G.; Dijkwel, P. P.; Warscheid, B., Study of Early Leaf Senescence in Arabidopsis thaliana by Quantitative Proteomics Using Reciprocal 14 N / 15 N Labeling and Difference Gel Electrophoresis. *Molecular & Cellular Proteomics* **2007,** 7, (1), 108-120.

178. Park, S. K.; Yates, J. R., Census for proteome quantification. *Current Protocols in Bioinformatics* **2010,** Chapter 13, Unit 13.12.1-11.

179. Park, S. K. R.; Aslanian, A.; McClatchy, D. B.; Han, X.; Shah, H.; Singh, M.; Rauniyar, N.; Moresco, J. J.; Pinto, A. F. M.; Diedrich, J. K.; Delahunty, C.; Yates, J. R., Census 2: Isobaric labeling data analysis. *Bioinformatics* **2014,** 30, (15), 2208-2209.

180. Zhang, Y.; Reckow, S.; Webhofer, C.; Boehme, M.; Gormanns, P.; Egge-Jacobsen, W. M.; Turck, C. W., Proteome Scale Turnover Analysis in Live Animals Using Stable Isotope Metabolic Labeling. *Analytical Chemistry* **2011,** 83, (5), 1665-1672.

181. Lyon, D.; Castillejo, M. A.; Staudinger, C.; Weckwerth, W.; Wienkoop, S.; Egelhofer, V., Automated Protein Turnover Calculations from 15N Partial Metabolic Labeling LC/MS Shotgun Proteomics Data. *PLoS ONE* **2014,** 9, (4), e94692-e94692.

182. Snijders, A. P. L.; De Koning, B.; Wright, P. C., Perturbation and interpretation of nitrogen isotope distribution patterns in proteomics. *Journal of Proteome Research* **2005,** 4, (6), 2185-2191.

183. MacCoss, M. J.; Wu, C. C.; Matthews, D. E.; Yates, J. R., Measurement of the isotope enrichment of stable isotope-labeled proteins using high-resolution mass spectra of peptides. *Analytical Chemistry* **2005,** 77, (23), 7646-7653.

184. Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R., Comparison of Full Versus Partial Metabolic Labeling for Quantitative Proteomics Analysis in Arabidopsis thaliana. *Molecular & Cellular Proteomics* **2007,** 6, (5), 860-881.

185. Guan, S.; Price, J. C.; Prusiner, S. B.; Ghaemmaghami, S.; Burlingame, A. L., A Data Processing Pipeline for Mammalian Proteome Dynamics Studies Using Stable Isotope Metabolic Labeling. *Molecular & Cellular Proteomics* **2011,** 10, (12), M111.010728-M111.010728.

186. Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Bye-Ajee, H.; Cowley, A.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Fazzini, F.; Castro, L. G.; Figueira, L.; Garmiri, P.; Georghiou, G.; Gonzalez, D.; Hatton-Ellis, E.; Li, W.; Liu, W.; Lopez, R.; Luo, J.; Lussi, Y.; MacDougall, A.; Nightingale, A.; Palka, B.; Pichler, K.; Poggioli, D.; Pundir, S.; Pureza, L.; Qi, G.; Rosanoff, S.; Saidi, R.; Sawford, T.; Shypitsyna, A.; Speretta, E.; Turner, E.; Tyagi, N.; Volynkin, V.; Wardell, T.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Xenarios, I.; Bougueleret, L.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; ArgoudPuy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M. C.; Boeckmann, B.; Bolleman, J.; Boutet, E.; Breuza, L.; Casal-Casas, C.; De Castro, E.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Jungo, F.; Keller, G.; Lara, V.; Lemercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T.; Nouspikel, N.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pozzato, M.; Pruess, M.; Rivoire, C.; Roechert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A. L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L. S.; Zhang, J., UniProt: The universal protein knowledgebase. *Nucleic Acids Research* **2017,** 45, (D1), D158-D169.

187. Thomas, P. D.; Campbell, M. J.; Kejariwal, A.; Mi, H.; Karlak, B.; Daverman, R.; Diemer, K.; Muruganujan, A.; Narechania, A., PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research* **2003,** 13, (9), 2129-2141.

188. Smith, R., Conversations with 100 Scientists in the Field Reveal a Bifurcated Perception of the State of Mass Spectrometry Software. *Journal of Proteome Research* **2018,** 17, (4), 1335-1339.

189. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** 26, (7), 966-968.

190. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* **2008,** 24, (21), 2534-2536.

191. Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W., mzML—a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics* **2011,** 10, (1), R110.000133-R110.000133.

192. Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology* **2004,** 22, (11), 1459-66.

193. Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; Julian, R.; Binz, P.-A.; Deutsch, E. W.; Hermjakob, H.; Reisinger, F.; Griss, J.; Vizcaíno, J. A.; Chambers, M.; Pizarro, A.; Creasy, D., The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Molecular & Cellular Proteomics* **2012,** 11, (7), M111.014381-M111.014381.

194. Kullback, S.; Leibler, R. A., On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951,** 22, (1), 79-86.

195. Toprak, U. H.; Gillet, L. C.; Maiolica, A.; Navarro, P.; Leitner, A.; Aebersold, R., Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Molecular & Cellular Proteomics* **2014,** 13, (8), 2056-71.

196. De Graaf, E. L.; Altelaar, A. F. M.; Van Breukelen, B.; Mohammed, S.; Heck, A. J. R., Improving SRM assay development: A global comparison between triple quadrupole, Ion trap, and higher energy CID peptide fragmentation spectra. *Journal of Proteome Research* **2011,** 10, (9), 4334-4341.

197. Wan, K. X.; Vidavsky, I.; Gross, M. L., Comparing similar spectra: From similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry* **2002,** 13, (1), 85-88.

198. Eguchi, S.; Copas, J., Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma. *Journal of Multivariate Analysis* **2006,** 97, (9), 2034-2040.

199. Burnham, K. P.; Anderson, D. R., *Model Selection and Multimodel Inference*. Springer New York: New York, NY, 2004.

200. Taubert, M.; Von Bergen, M.; Seifert, J., Limitations in detection of 15N incorporation by mass spectrometry in protein-based stable isotope probing (protein-SIP). *Analytical and Bioanalytical Chemistry* **2013,** 405, (12), 3989-3996.

201. Courties, C.; Vaquer, A.; Troussellier, M.; Lautier, J.; Chrétiennot-Dinet, M. J.; Neveux, J.; Machado, C.; Claustre, H., Smallest eukaryotic organism. *Nature* **1994,** 370, (6487), 255-255.

202. Le Bihan, T.; Grima, R.; Martin, S.; Forster, T.; Le Bihan, Y., Quantitative analysis of low-abundance peptides in hela cell cytoplasm by targeted liquid chromatography/ mass spectrometry and stable isotope dilution: Emphasising the distinction between peptide detection and peptide identification. *Rapid Communications in Mass Spectrometry* **2010,** 24, (7), 1093-1104.

203. Mann, M. In *Useful Tables of Possible and Probable Peptide Masses*, 43rd ASMS Conference on Mass Spectrometry and Allied Topics; Atlanta, GA, 1995; pp 639-640.

204. Karty, J. A.; Ireland, M. M. E.; Brun, Y. V.; Reilly, J. P., Artifacts and unassigned masses encountered in peptide mass mapping. *Journal of Chromatography B* **2002,** 782, (1-2), 363-383.

205. Schmidt, F.; Schmid, M.; Jungblut, P. R.; Mattow, J.; Facius, A.; Pleissner, K.-P., Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *Journal of the American Society for Mass Spectrometry* **2003,** 14, (9), 943-956.

206. Koehler, C. J.; Bollineni, R. C.; Thiede, B., Application of the half decimal place rule to increase the peptide identification rate. *Rapid Communications in Mass Spectrometry* **2017,** 31, (2), 227-233.

207. Fetzer, I.; Jehmlich, N.; Vogt, C.; Richnow, H. H.; Seifert, J.; Harms, H.; Von Bergen, M.; Schmidt, F., Calculation of partial isotope incorporation into peptides measured by mass spectrometry. *BMC Research Notes* **2010,** 3, 178-178.

208. Ting, L.; Cowley, M. J.; Hoon, S. L.; Guilhaus, M.; Raftery, M. J.; Cavicchioli, R., Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Molecular and Cellular Proteomics* **2009,** 8, (10), 2227-2242.

209. Jones, M. C., Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology* **2009,** 6, (1), 70-81.

Appendix I: Details of Code Availability

A live version of the *HeavyMetL* program is hosted at

https://pdcharles.github.io/HeavyMetL/HeavyMetL.htm

The source code is released publicly under an MIT license, and is split into two repositories.

HeavyMetL-specific program code can be found at:

https://github.com/pdcharles/HeavyMetL

A generalised library for interaction with MS data, which defines code for handling chromatograms, spectra and scans as JavaScript objects, alongside further object definitions for common proteomic MS file formats (including mzML and mzXML), the isotopologue distribution prediction algorithm, and miscellaneous program flow, thread control and mathematical calculation functions is located separately at:

https://github.com/pdcharles/MSLIB

The following external code projects are also used in *HeavyMetL* code:

Pako (https://github.com/nodeca/pako) is a library for decompression of data compressed in 'zlib' format, which is a compression technique used in some mzML/mzXML files.

Fabric.js (http://fabricjs.com/) is an HTML5 Canvas element abstraction library used in *HeavyMetL* to generate PMC and protein-level graphical output.

Appendix II: Annotated HeavyMetL Screenshots

The following seven pages consist of sequential full-screen screenshots from the *HeavyMetL* interface over the course of setting up and initiating processing on the benchmark dataset presented in Chapter 4.

Page 187 - *Initial view of main screen*

Page 188 - *Selection of raw data files*

Page 189 - *Loading scan headers from raw data files*

Page 190 - *Selection of PMC List in text-table format*

Page 191 - *Main screen with dataset loaded and ready for processing*

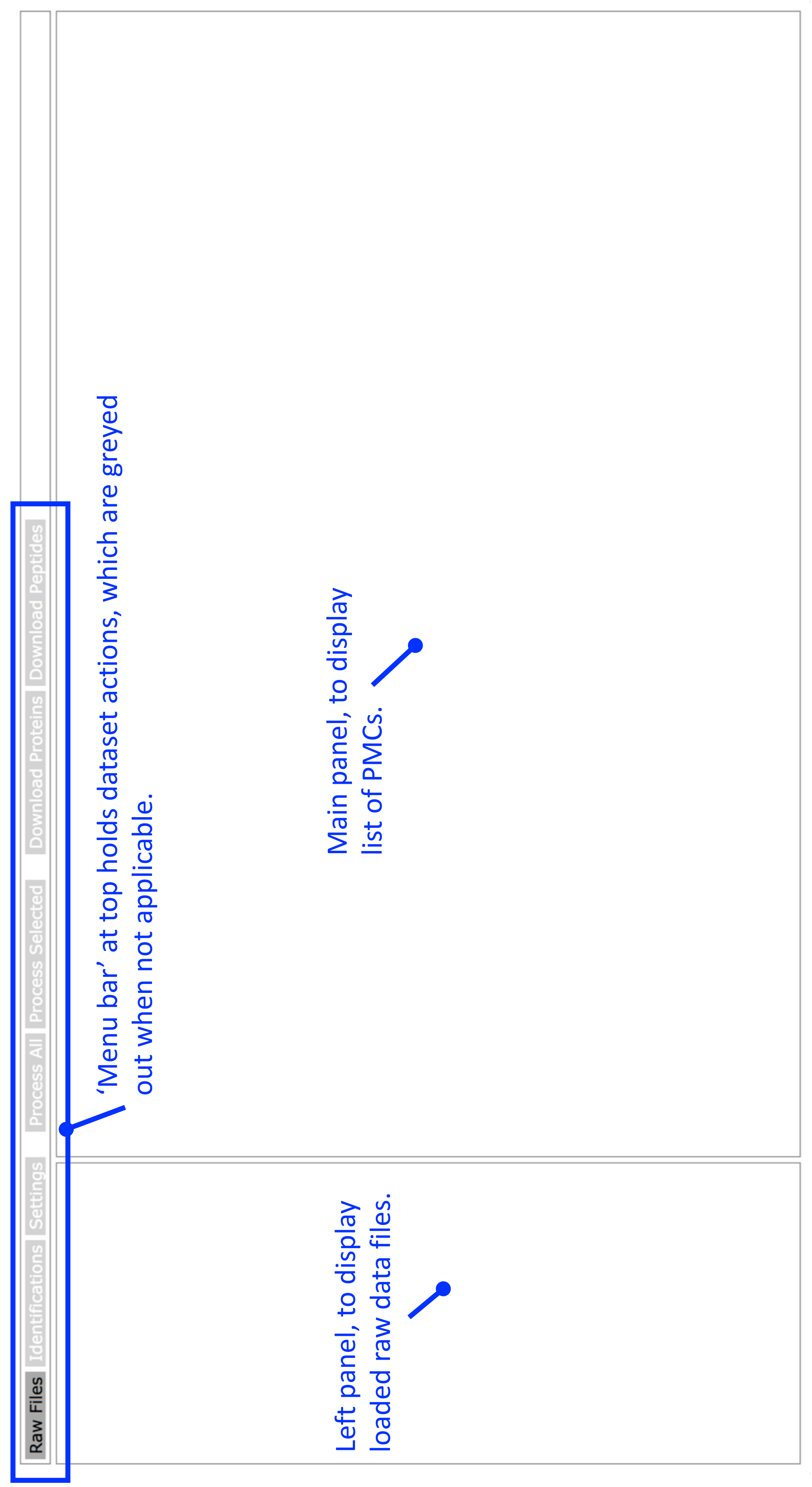
Page 192 - *Settings configuration screen*

Page 193 - *Processing underway*

Annotations are shown in blue text.

Screenshots of the graphical results are already included in Chapter 2; see Figure 2-VIII and Figure 2-IX.

1. *Initial view of main screen*

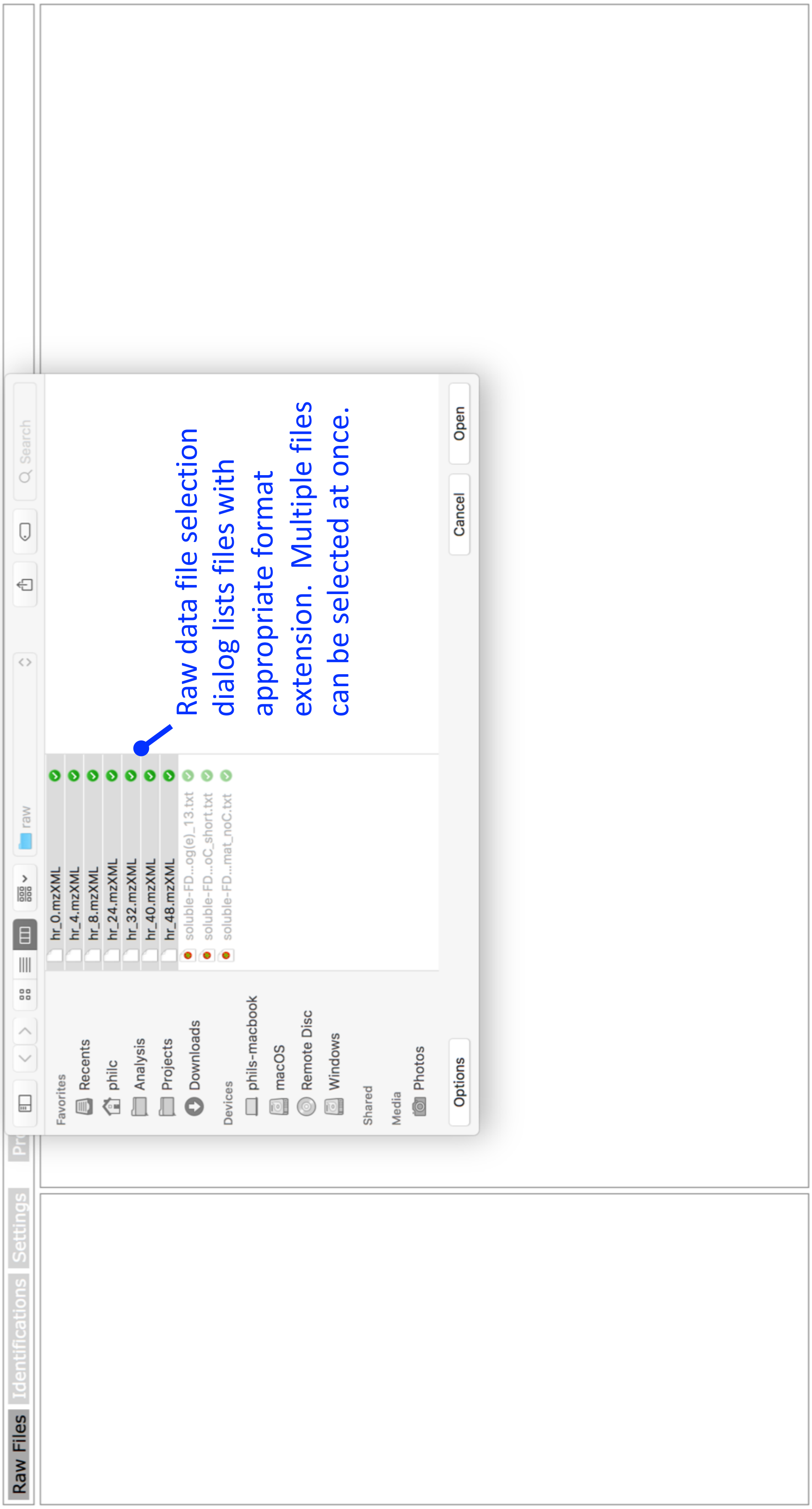


'Menu bar' at top holds dataset actions, which are greyed out when not applicable.

Main panel, to display list of PMCs.

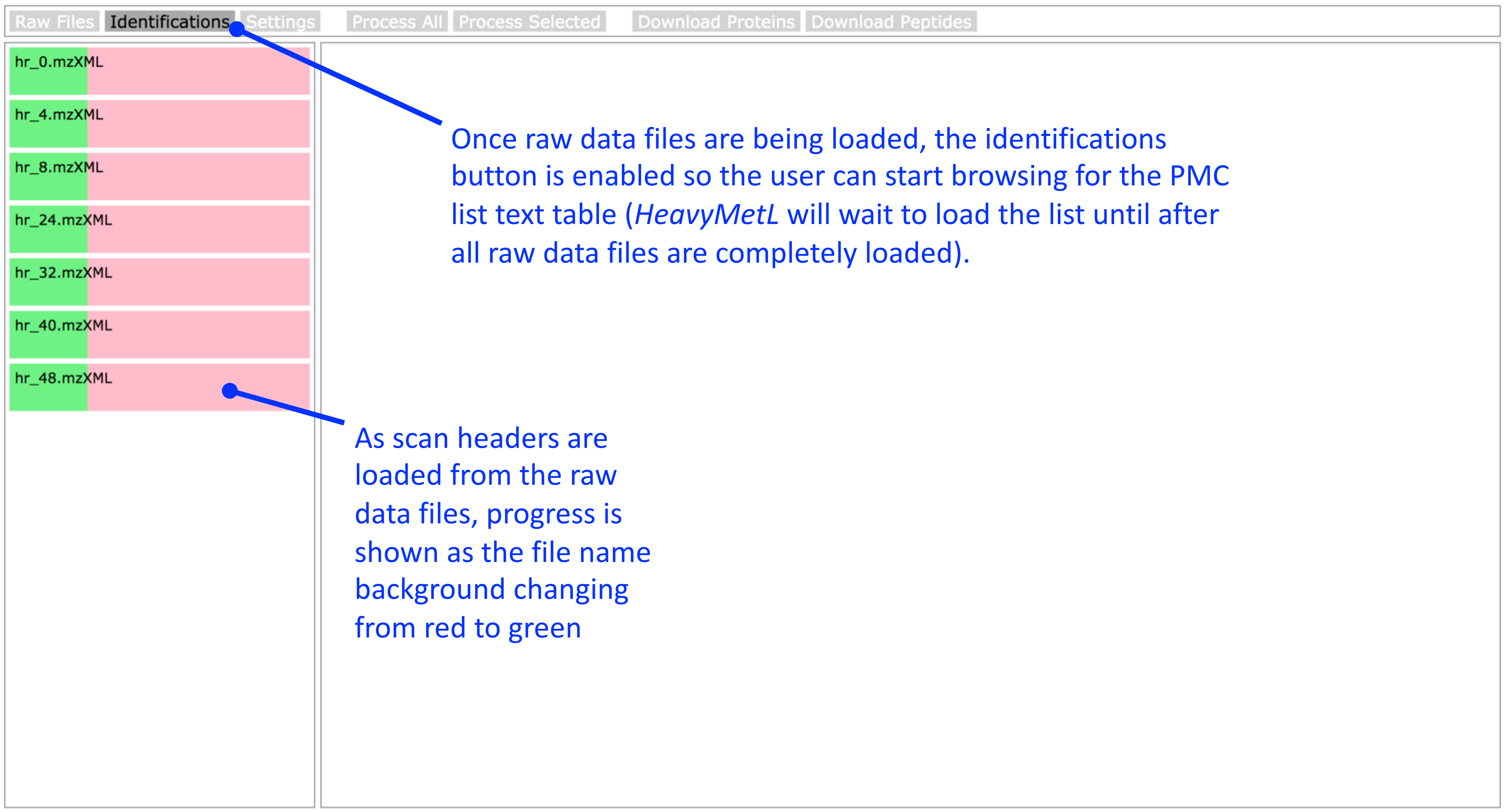
Left panel, to display loaded raw data files.

2. *Selection of raw data files*



Raw data file selection dialog lists files with appropriate format extension. Multiple files can be selected at once.

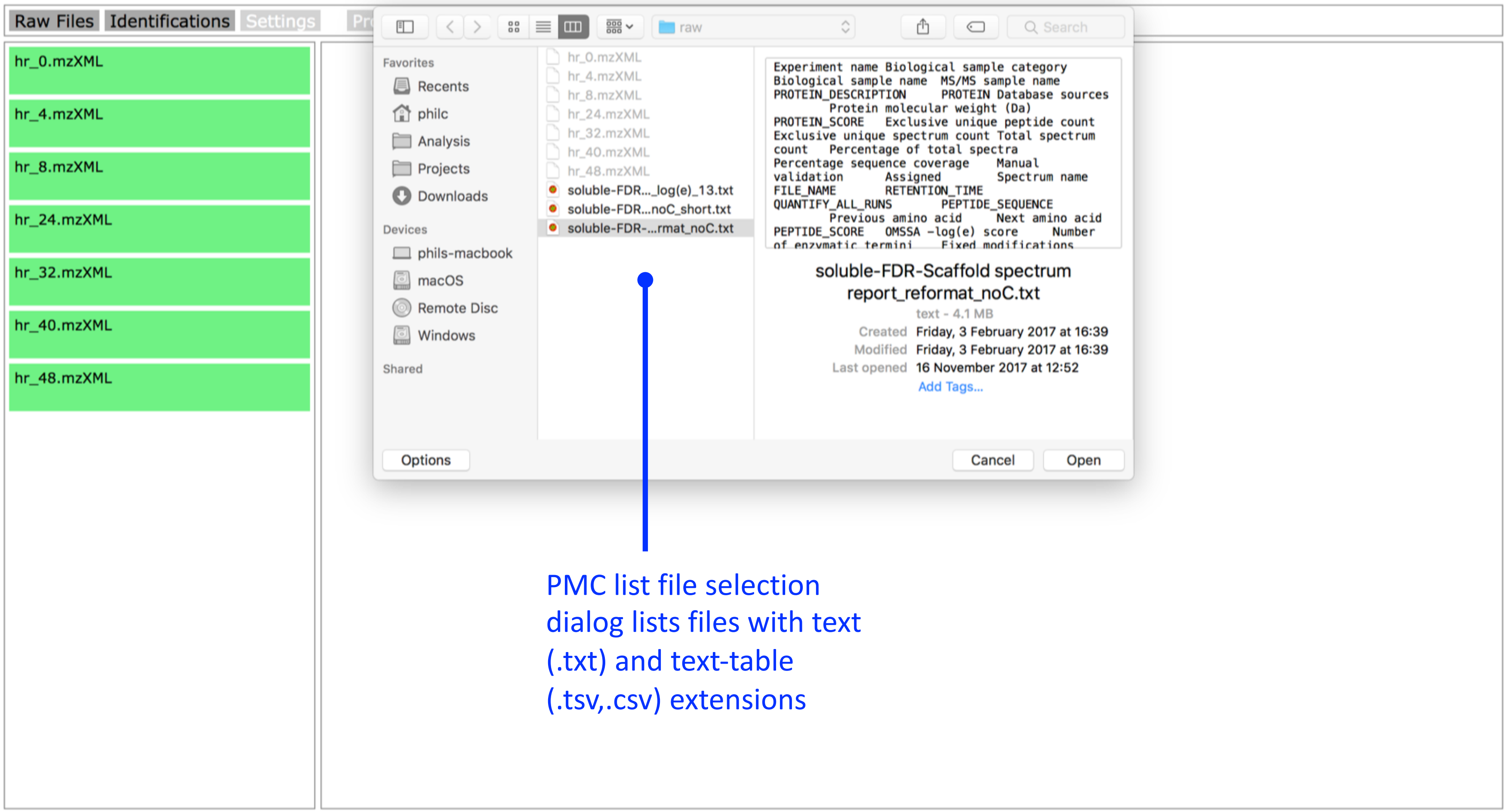
3. *Loading scan headers from raw data files*



Once raw data files are being loaded, the identifications button is enabled so the user can start browsing for the PMC list text table (*HeavyMetL* will wait to load the list until after all raw data files are completely loaded).

As scan headers are loaded from the raw data files, progress is shown as the file name background changing from red to green

4. *Selection of PMC List in text-table format*



```
Experiment name Biological sample category
Biological sample name  MS/MS sample name
PROTEIN_DESCRIPTION     PROTEIN Database sources
        Protein molecular weight (Da)
PROTEIN_SCORE   Exclusive unique peptide count
Exclusive unique spectrum count Total spectrum
count   Percentage of total spectra
Percentage sequence coverage    Manual
validation      Assigned        Spectrum name
FILE_NAME       RETENTION_TIME
QUANTIFY_ALL_RUNS       PEPTIDE_SEQUENCE
        Previous amino acid     Next amino acid
PEPTIDE_SCORE   OMSSA -log(e) score     Number
of enzymatic termini    Fixed modifications
```

soluble-FDR-Scaffold spectrum report_reformat_noC.txt

text - 4.1 MB

Created Friday, 3 February 2017 at 16:39

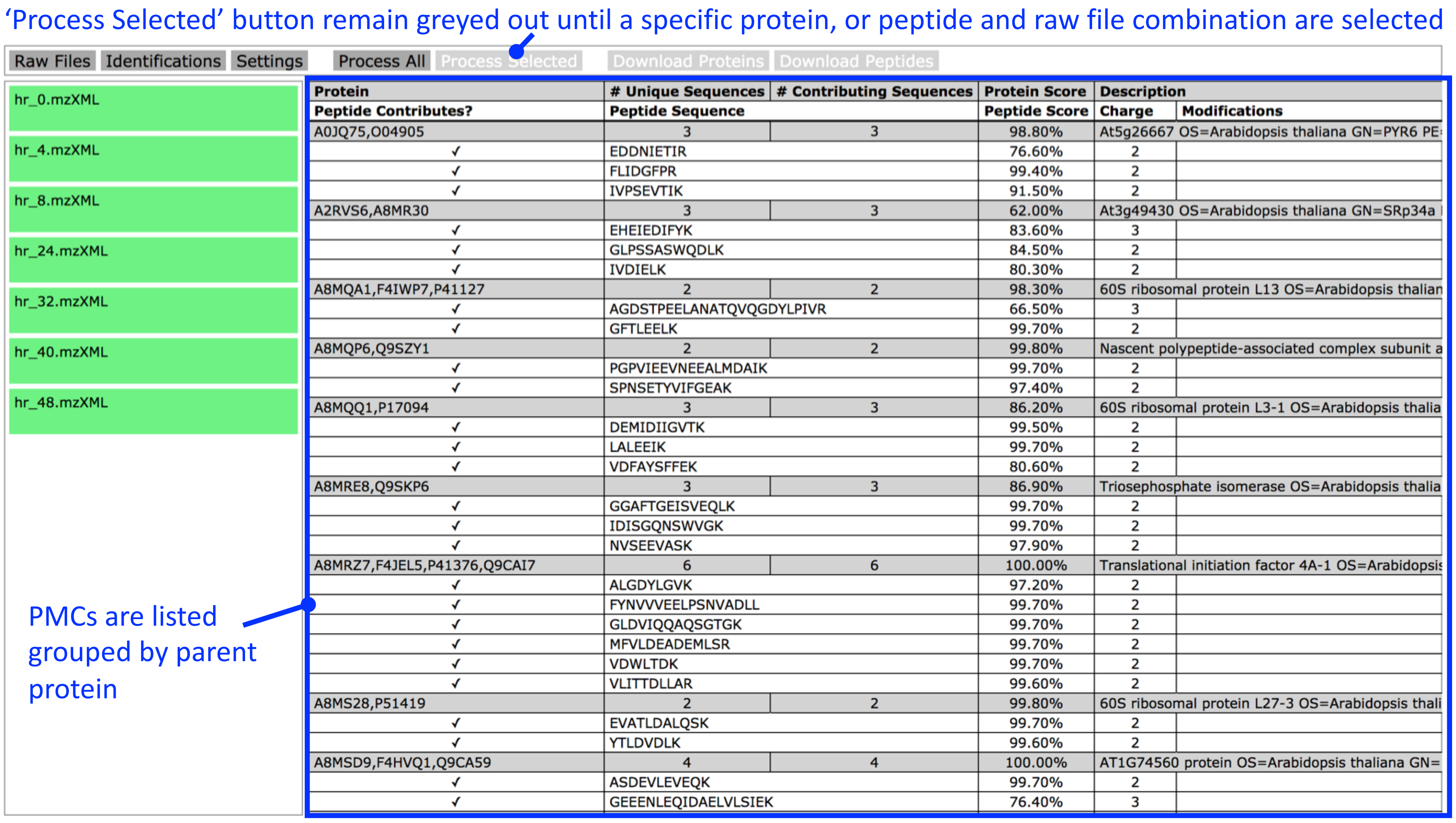
Modified Friday, 3 February 2017 at 16:39

Last opened 16 November 2017 at 12:52

Add Tags...

PMC list file selection dialog lists files with text (.txt) and text-table (.tsv,.csv) extensions

5. *Main screen with dataset loaded and ready for processing*

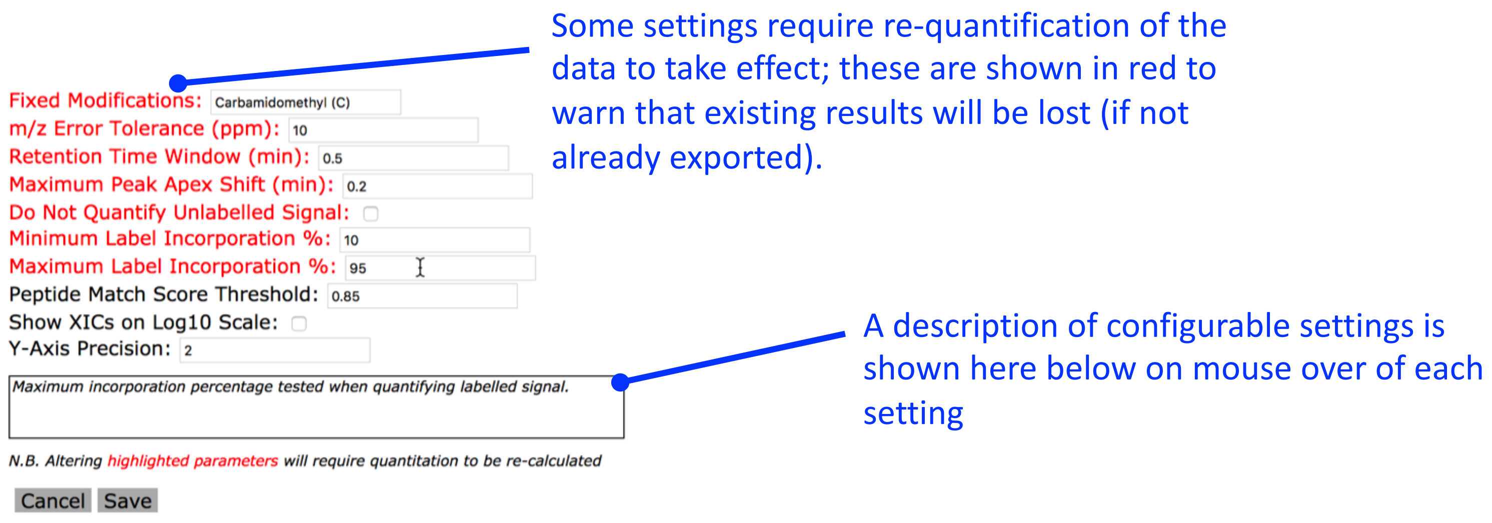


‘Process Selected’ button remain greyed out until a specific protein, or peptide and raw file combination are selected

Protein	# Unique Sequences	# Contributing Sequences	Protein Score	Description	
Peptide Contributes?	**Peptide Sequence**		**Peptide Score**	**Charge**	**Modifications**
A0JQ75,O04905	3	3	98.80%	At5g26667 OS=Arabidopsis thaliana GN=PYR6 PE	
✓	EDDNIETIR		76.60%	2	
✓	FLIDGFPR		99.40%	2	
✓	IVPSEVTIK		91.50%	2	
A2RVS6,A8MR30	3	3	62.00%	At3g49430 OS=Arabidopsis thaliana GN=SRp34a	
✓	EHEIEDIFYK		83.60%	3	
✓	GLPSSASWQDLK		84.50%	2	
✓	IVDIELK		80.30%	2	
A8MQA1,F4IWP7,P41127	2	2	98.30%	60S ribosomal protein L13 OS=Arabidopsis thalian	
✓	AGDSTPEELANATQVQGDYLPIVR		66.50%	3	
✓	GFTLEELK		99.70%	2	
A8MQP6,Q9SZY1	2	2	99.80%	Nascent polypeptide-associated complex subunit a	
✓	PGPVIEEVNEEALMDAIK		99.70%	2	
✓	SPNSETYVIFGEAK		97.40%	2	
A8MQQ1,P17094	3	3	86.20%	60S ribosomal protein L3-1 OS=Arabidopsis thalia	
✓	DEMIDIIGVTK		99.50%	2	
✓	LALEEIK		99.70%	2	
✓	VDFAYSFFEK		80.60%	2	
A8MRE8,Q9SKP6	3	3	86.90%	Triosephosphate isomerase OS=Arabidopsis thalia	
✓	GGAFTGEISVEQLK		99.70%	2	
✓	IDISGQNSWVGK		99.70%	2	
✓	NVSEEVASK		97.90%	2	
A8MRZ7,F4JEL5,P41376,Q9CAI7	6	6	100.00%	Translational initiation factor 4A-1 OS=Arabidopsis	
✓	ALGDYLGVK		97.20%	2	
✓	FYNVVVEELPSNVADLL		99.70%	2	
✓	GLDVIQQAQSGTGK		99.70%	2	
✓	MFVLDEADEMLSR		99.70%	2	
✓	VDWLTDK		99.70%	2	
✓	VLITTDLLAR		99.60%	2	
A8MS28,P51419	2	2	99.80%	60S ribosomal protein L27-3 OS=Arabidopsis thali	
✓	EVATLDALQSK		99.70%	2	
✓	YTLDVDLK		99.60%	2	
A8MSD9,F4HVQ1,Q9CA59	4	4	100.00%	AT1G74560 protein OS=Arabidopsis thaliana GN=	
✓	ASDEVLEVEQK		99.70%	2	
✓	GEEENLEQIDAELVLSIEK		76.40%	3	

PMCs are listed grouped by parent protein

6. *Settings configuration screen*



Fixed Modifications: Carbamidomethyl (C)
m/z Error Tolerance (ppm): 10
Retention Time Window (min): 0.5
Maximum Peak Apex Shift (min): 0.2
Do Not Quantify Unlabelled Signal: □
Minimum Label Incorporation %: 10
Maximum Label Incorporation %: 95
Peptide Match Score Threshold: 0.85
Show XICs on Log10 Scale: □
Y-Axis Precision: 2

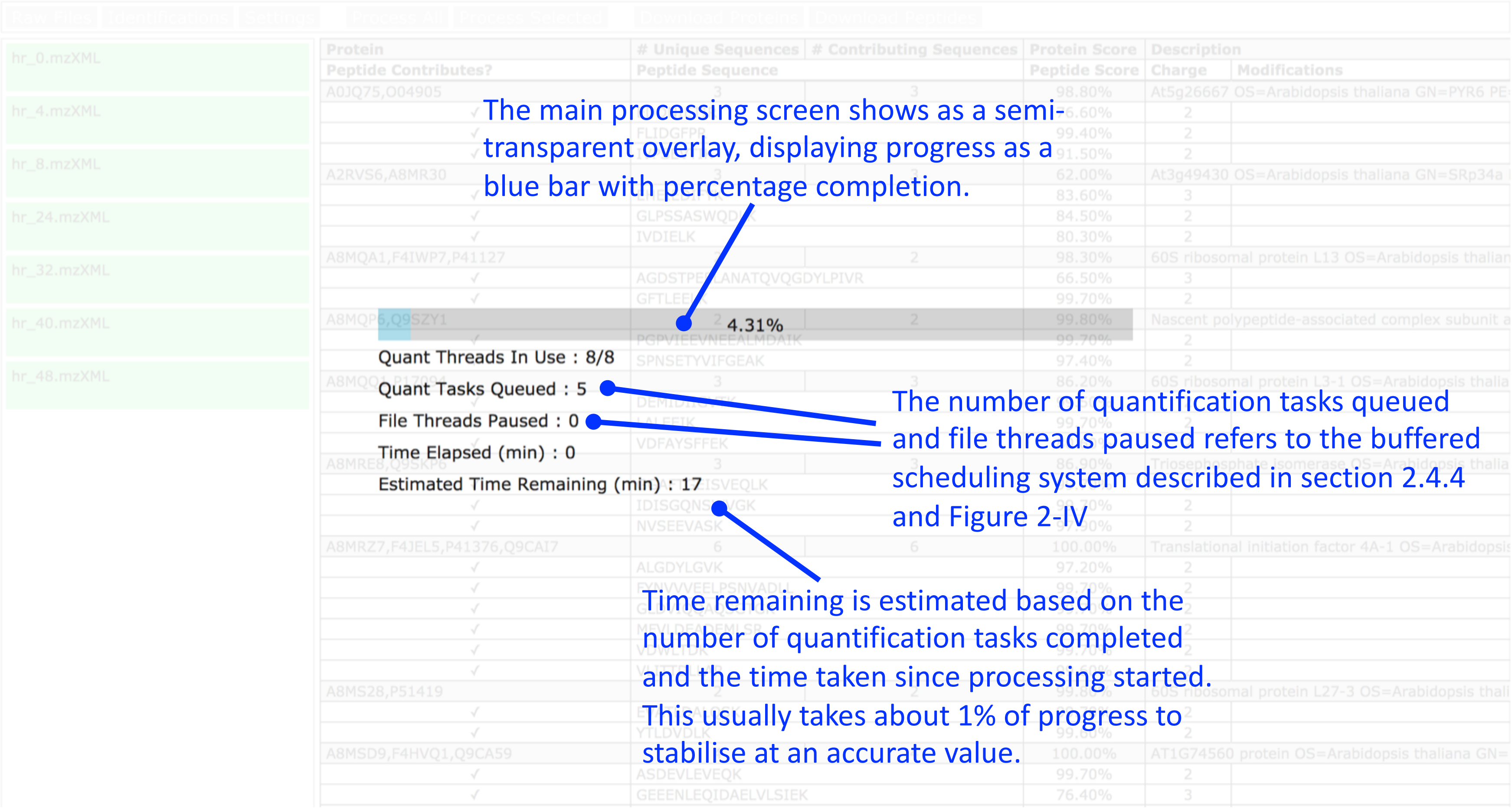
Maximum incorporation percentage tested when quantifying labelled signal.

N.B. Altering highlighted parameters will require quantitation to be re-calculated

Some settings require re-quantification of the data to take effect; these are shown in red to warn that existing results will be lost (if not already exported).

A description of configurable settings is shown here below on mouse over of each setting

7. Processing underway



4.31%

Quant Threads In Use : 8/8

Quant Tasks Queued : 5

File Threads Paused : 0

Time Elapsed (min) : 0

Estimated Time Remaining (min) : 17

The main processing screen shows as a semi-transparent overlay, displaying progress as a blue bar with percentage completion.

The number of quantification tasks queued and file threads paused refers to the buffered scheduling system described in section 2.4.4 and Figure 2-IV

Time remaining is estimated based on the number of quantification tasks completed and the time taken since processing started. This usually takes about 1% of progress to stabilise at an accurate value.

Appendix III: Explanation of Score Distribution in Random Simulation

In Chapter 3, to compare the results of analysis with each of the four candidate Similarity Scores to a baseline (SS_{RAND}), I simulated a dataset where there was no association between Standardised Incorporation Quantification Errors of individual quantification results and their corresponding score assigned to that result (see Section 3.6).

The simulated dataset scores were generated by taking the maximum of a random number *n* (between 5 and 20) of draws from a uniform distribution between 0 and 1.

Let X_n be the maximum of the set of *n* uniformly distributed independent variables in the range (0,1). X_n may be shown to follow a Beta(n,1) distribution (209) (Figure AIII-I).

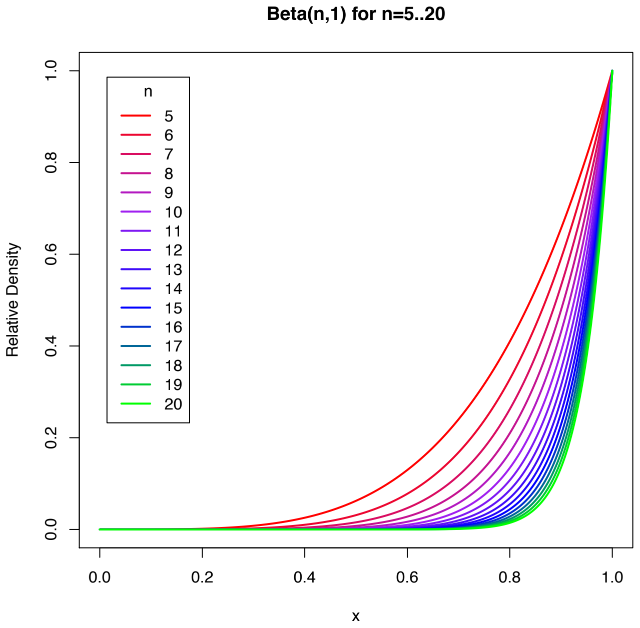


Figure AIII-I. *Densities of the Beta distribution for Beta(n,1) where n=5...20*

The values of SS_{RAND} will be drawn (approximately) equally from each distribution shown in Figure AIII-I, so it is unsurprising that the SS_{RAND} scales in **Error! Reference source not found.**-A are seen to skew heavily towards the upper end of the score range.

Furthermore, the expected value (mean) of $\mathrm{SS}_{\mathrm{RAND}}$ can also be calculated. The expected value for each X_n is:

$$\mathrm{E}(X_n) = \frac{n}{n+1}$$

By the law of total expectation, the expected value for the dataset score is therefore

$$\mathrm{E}(\mathrm{SS}_{\mathrm{RAND}}) = \mathrm{E}(X_5)\mathrm{P}(\mathrm{n} = 5) + \mathrm{E}(X_6)\mathrm{P}(\mathrm{n} = 6) + \cdots + \mathrm{E}(X_{20})\mathrm{P}(\mathrm{n} = 20)$$

Since values of *n* between 5 and 20 are equally likely this is just the mean of $\mathrm{E}(X_n)$ for all *n*

$$\mathrm{E}(\mathrm{SS}_{\mathrm{RAND}}) = \frac{\sum_{n=5}^{20} \frac{n}{n+1}}{20-5} = \frac{\frac{5}{6} + \frac{6}{7} + \cdots + \frac{20}{21}}{15} \approx 0.91$$

A mean value of 0.91 is also consistent with Figure 3-XI-A.