ANALYSIS AND SIMULATION OF TANDEM MASS SPECTROMETRY DATA

Dennis Goldfarb

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2019

Approved by:

Michael B. Major

Wei Wang

Leonard McMillan

Jack Snoeyink

Martin Styner

ABSTRACT

DENNIS GOLDFARB: ANALYSIS AND SIMULATION OF TANDEM MASS SPECTROMETRY DATA
(Under the direction of Michael B. Major and Wei Wang)

This dissertation focuses on improvements to data analysis in mass spectrometry-based proteomics, which is the study of an organism's full complement of proteins. One of the biggest surprises from the Human Genome Project was the relatively small number of genes (∼20,000) encoded in our DNA. Since genes code for proteins, scientists expected more genes would be necessary to produce a diverse set of proteins to cover the many functions that support the complexity of life. Thus, there is intense interest in studying proteomics, including post-translational modifications (how proteins change after translation from their genes), and their interactions (e.g. proteins binding together to form complex molecular machines) to fill the void in molecular diversity.

The goal of mass spectrometry in proteomics is to determine the abundance and amino acid sequence of every protein in a biological sample. A mass spectrometer can determine mass/charge ratios and abundance for fragments of short peptides (which are subsequences of a protein); sequencing algorithms determine which peptides are most likely to have generated the fragmentation patterns observed in the mass spectrum, and protein identity is inferred from the peptides. My work improves the computational tools for mass spectrometry by removing limitations on present algorithms, simulating mass spectroscopy instruments to facilitate algorithm development, and creating algorithms that approximate isotope distributions, deconvolve chimeric spectra, and predict protein-protein interactions.

While most sequencing algorithms attempt to identify a single peptide per mass spectrum, multiple peptides are often fragmented together. Here, I present a method to deconvolve these chimeric mass spectra into their individual peptide components by examining the isotopic distributions of their fragments. First, I derived the equation to calculate the theoretical isotope distribution of a peptide fragment. Next, for cases where elemental compositions are not known, I developed methods to approximate the isotope distributions. Ultimately, I created a non-negative least squares model that deconvolved chimeric spectra and increased peptide-spectrum-matches by 15-30%.

To improve the operation of mass spectrometer instruments, I developed software that simulates liquid chromatography-mass spectrometry data and the subsequent execution of custom data acquisition algorithms. The software provides an opportunity for researchers to test, refine, and evaluate novel algorithms prior to implementation on a mass spectrometer.

Finally, I created a logistic regression classifier for predicting protein-protein interactions defined by affinity purification and mass spectrometry (APMS). The classifier increased the area under the receiver operating characteristic curve by 16% compared to previous methods. Furthermore, I created a web application to facilitate APMS data scoring within the scientific community.

To my parents, Alek and Galina, and to the friends who have kept me smiling and laughing.. . .

ACKNOWLEDGEMENTS

I'd like to thank both my advisors, Ben Major and Wei Wang, for supporting me and giving me the freedom to follow my research interests even when they didn't perfectly align with their own. I am grateful that Wei continued to mentor and meet with me every week even after she had moved to a new university and for never letting me forget deadlines and requirements. To Ben, thank you for trusting a computer scientist who's never worked with his hands to operate and maintain expensive mass spectrometry and liquid chromatography equipment, as well as financially supporting a long graduate career that allowed me to focus on science instead of finances.

I would also like to thank my Committee for their comments and advice, and specifically Leonard McMillan for reminding me not to get too lost in subjects outside the Department of Computer Science.

I am immensely indebted to my labmates throughout the years, Priscila Siesser, Bridgid Hast, Kelly Orgel, Kathryn Mackinnon, Feng Yan, Matthew Walker, Kathleen Mulvaney, Alex Rabinowitz, Erica Cloer, Tigist Tamir, Abby Winn, Emily Cousins, Nalini Peres-da-silva, Megan Agajanian, Brittany Bowman, Ryan Murphy, Trent Stohrer, Yuko Nakamichi, Stephen Serafin, Selin Altinok, and Seth Weir for giving me a crash course in biology and chemistry, assisting me with wet lab work, and their patience when listening to my ideas on topics I had nearly no knowledge of. You made coming into work fun and exciting.

A big thank you to my collaborators: Buddy Weissman, Cyrus Vaziri, Joe Harrison, Gary Johnson, Laura Herring, Emily Wilkerson, and many others. These opportunities helped convert textbook knowledge into practical experience. My appreciation goes out to the Thermo Scientific field service engineers and technical support staff who helped me through some "uh oh" moments: Josh Strickland, Brian Fulp, Jeff King, Paulina Reddy, and Mike Henry. Thank you to the departmental staff, especially Jodie Turnbull who often appeared to be more invested in my graduation than myself.

To my college and high school family, Mike Kimlat, William Chow, Elisa Lee, Ryan Orfi, Kat Tam, Kate Chamberlain, Liz Brule, Lorelei Schreiber, Kyle Allen, Jeremy Schuh, Bekah Bolton, Tim Heiman, Steve Spellman, Kira Henderson, Sara Laprade, and Maxx and Samantha Garrison, to whom I retreat to when I need encouragement or a simple distraction. You provide the love and happiness I need to keep going.

Finally, my parents Alek and Galina, my brother Max, sister-in-law Marina, and my grandmother Paulina. Nothing I've done would be possible without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA	Amino acid
API	Application programming interface
APMS	Affinity purification-mass spectrometry
AUC	Area under the receiver operating characteristic curve
CPU	Central processing unit
Da	Dalton
DDA	Deoxyribonucleic acid
DIA	Data-dependent acquisition
DNA	Data-independent acquisition
EGH	Exponential gaussian hybrid
ESI	Electrospray ionization
FFT	Fast-Fourier transform
IUPAC	International Union of Pure and Applied Chemistry
LASSO	Least absolute shrinkage and selection operator
LC	Liquid chromatography
m/z	Mass / charge
MS	Mass spectrometry or mass spectrometer (depending on context)
mRNA	Messenger ribonucleic acid
NNLS	Non-negative least squares
PCC	Pearson correlation coefficient
PIF	Precursor ion fraction
PSM	Peptide-spectrum-match
PTM	Post-translational modification
RAM	Random-access memory
ROC	Receiver operating characteristic
SVM	Support vector machines
TSC	Total spectral count

CHAPTER 1: INTRODUCTION

1.1 Mass spectrometry-based proteomics

One the biggest surprises from the Human Genome Project was the small number of genes encoded in our DNA. Our genome's limited repertoire of $\sim$20,000 genes did not reflect the complexity of life, so scientists have turned to the study of the proteome–an organism's full protein complement—to begin filling the void in molecular diversity. Proteins are comprised of amino acid (AA) chains (Definition 2.2) and can be represented as a sequence of characters from a 20 letter alphabet, one for each amino acid. While the genome provides the blueprints for protein AA sequences, proteins are subject to a myriad of post-transcriptional and post-translational regulatory mechanisms. Proteins are spliced into isoform variants, chemically modified, cleaved, and degraded. Estimates for the number of these distinct protein versions range from one to six million per cell type (Aebersold et al., 2018). Furthermore, proteins physically bind each other to form protein complexes (Definition 2.2) that function as microscopic machines. It is the protein molecules that perform most processes of life, determine our traits, and have major ramifications on health and disease.

The goal of mass spectrometry in proteomics is to determine the abundance and amino acid sequence of every protein in a biological sample. A typical workflow begins by cleaving the proteins into non-overlapping amino acid subsequences called peptides. The resulting peptide mixture is separated by a technique called liquid chromatography (Definition 2.3) and injected into a mass spectrometer over the course of minutes to hours. At any given time, peptides are simultaneously entering the mass spectrometer. Using the standard data acquisition strategy (Definition 2.4.6), the mass spectrometer performs an MS1 scan; it measures and records the mass-to-charge ratios (*m/z*) and signal intensities of the present molecules—also known as a mass spectrum. Next, an MS2 scan is performed; the peptides are fragmented and the mass spectrum of the peptide's fragments is recorded. Using the measured *m/z* of the peptides and their fragments, a sequencing algorithm then determines which peptides most likely generated the observed fragmentation patterns. Afterwards, proteins are identified by mapping the peptide AA sequences to a reference database of protein AA sequences. Finally, post-processing steps infer biological implications.

MS-based proteomics finds itself at a similar stage as DNA sequencing was in the late 90's and early 2000's; mass spectrometers have the capability to generate vast quantities of complex data that as of yet cannot be fully analyzed, let alone interpreted for biological meaning. While mass spectrometry has been a scientific discipline since the end of the 19th century, high-throughput MS-based proteomics only became feasible following the success of the Human Genome Project. DNA sequences of genes were translated into protein AA sequences, and a reference database for the human proteome was created. The availability of a reference database spurred algorithmic development and led to computer-automated interpretation of MS data. Although data analysis tools now receive considerable attention, improvements in instrument engineering and experimental design continue to outpace algorithmic advances. In spite of these limitations, MS-based proteomics is still immensely powerful. As a result, MS techniques have been deployed in the search for disease biomarkers, the elucidation of molecular signaling pathways and protein-protein interactions (Definition 2.2), drug-target discovery, the determination of cellular spatial organization, and the real-time detection of cancerous tissues during surgery (Crutchfield et al., 2016; Kosako and Nagano, 2011; Huttlin et al., 2017; Klaeger et al., 2017; Fatou et al., 2016; Marx, 2015).

Three of the main challenges faced by mass spectrometry are the complexity of the proteome, limitations of mass spectrometry instrumentation, and inadequate data analysis methods.

First, even for short peptides the number of possible sequences is huge, so sequencing algorithms must identify sequences from a vast search space. With only the 20 naturally abundant amino acids, a peptide of length n already has 20^n possible sequences. Chemical modification of individual amino acids such as the phosphorylation of serines, threonines, and tyrosines, compound this number. One can restrict the search to a reference database, but these are understandably incomplete: new chemical modifications continue to be discovered, and even for known chemical modifications most of their sites on a protein have not yet been observed. Furthermore, a protein's unmodified AA sequence is subject to variation due to splicing, in which different regions of a gene are used as the blueprint for a protein. Many splice variants also remain undiscovered. Mutations in a gene lead to amino acid substitutions, and the mutations found in a biological specimen are usually not known. Moreover, DNA sequences encoding peptides <100 amino acids long are typically not annotated as genes and are therefore absent from reference databases (Frith et al., 2006). As a result, sequencing algorithms that use reference databases are limited, and *de novo* sequencing algorithms, which don't use reference databases, are computationally intensive.

Second, data can be incorrect, missing, or confounded. Incorrect data is introduced by measurement error, bias, and electronic noise. Data are frequently missing due to a combination of the instrumentation's limit of detection and restricted dynamic range. The limit of detection is a frequent problem in MS-based proteomics because proteins cannot be copied and amplified like DNA, and is therefore constrained by its starting material. The restricted dynamic range is an issue because the dynamic range of protein abundance is much large than the instrumentation's. For example, the distribution of protein abundance in human blood spans more than 10 orders of magnitude, while contemporary instrumentation is capable of detecting $\sim$4 within an experiment, and $\sim$3 within a single mass spectrum (Anderson and Anderson, 2002; Wu and Han, 2006). Confounded data is often due to the finite resolution of mass analyzers. Molecules with similar *m/z* cannot always be distinguished from each other. However, and even with adequate resolution, space-charge effects can cause two nearby *m/z* peaks (Definition 2.4.4) to coalesce into one, or a homogeneous *m/z* population can create two nearby peaks each with the wrong *m/z* (Kaufmann and Walker, 2018). Third, an incomplete understanding of peptide fragmentation pathways (Definition 2.4.5.2), multiple signals per molecular species, and the simultaneous fragmentation of multiple peptides inhibits sequencing efforts. In this stage, observed mass spectra are compared to expected mass spectra for a given peptide candidate. Though scientists have described many fragmentation pathways, their frequencies are not yet predictable and uncommon pathways are ignored. Most sequencing algorithms expect the common pathways will be observed with uniform signal intensity, yet in reality these intensities vary. The uncommon peptide fragments are not yet incorporated into sequencing algorithms (Medzihradszky and Chalkley, 2013; Verheggen et al., 2017). Multiple signals exist for almost all molecular species due to their distribution of isotopes (Definition 2.1) and charge states (Definition 2.4.5.2). Chimeric mass spectra, which result from the co-fragmentation of multiple distinct peptides, contradict the prevailing approach of matching a single peptide per spectrum. Though many peptides and proteins are identified using mass spectrometry, most mass spectra are not matched to the reference sequence database, and for those that are matched, much of their signals are unexplained.

Computer science is well-suited to improve MS-based proteomics. Better models of signal patterns are needed, and these models can be developed through mathematical modeling, data mining, and machine learning techniques. Due to improvements in instrument engineering, researchers are also exploring novel data acquisition strategies (Bilbao et al., 2015), calling for analysis tools and algorithms tailored to the new types of acquired data. The integration of data acquisition and data analysis in real-time holds great potential, but is relatively untapped. Currently, many instrument settings are defined prior to the start of an experiment and remain static throughout its execution, but on-line data analysis can optimize parameters in real-time. This requires a shift to on-line and extremely efficient algorithms. Finally, post-processing methods for biological interpretation should be integrated with other fields including genomics and transcriptomics.

1.2 Thesis statement

In this thesis, I show that chimeric spectra deconvolution, machine learning, and improvements to data acquisition lead to increased identifications of proteins and protein-protein interactions from mass spectrometry-based proteomics experiments.

Mass spectrometry data analysis is in its infancy and this dissertation contributes to three levels of the computational workflow: 1) analysis of mass spectra, 2) decision-making during data acquisition, and 3) scoring candidate protein-protein interactions. Beginning with the analysis of mass spectra, Chapter 3 derives methods to calculate theoretical and approximate isotope distributions for fragments of peptides and proteins. Using these methods, Chapter 4 describes a non-negative least squares regression (NNLS) model to deconvolve chimeric mass spectra by attributing a fragment's isotopic distribution to different peptides. Application of the NNLS model increased peptide-spectrum-matches by 15-30% and protein identifications by 5-9%. Next, Chapter 5 describes software I developed to simulate the data acquisition process of a mass spectrometer and to assist the evaluation of novel acquisition and analysis algorithms. Finally, to improve the prediction of protein-protein interactions, Chapter 6 describes a logistic regression model that integrates orthogonal sources of biological information with mass spectrometry data and increased the area under the receiver operating characteristic curve compared to previous methods by an average of 16%.

1.3 Contributions

1.3.1 Isotope distributions

In MS-based proteomics, only 20-50% of MS2 mass spectra are reliably matched to peptides. Interpretation of these data is hindered by the inability to analyze fundamental signal patterns found in all spectra. To alleviate this, I present methods to compute theoretical and approximate isotope distributions of peptide fragments. The major contributions are:

1. **Equations for the theoretical isotope distribution of a fragment ion.** Fragments have different isotope distributions than precursors (Definition 2.4.5.2); they depend on the set of precursor isotopes isolated during an MS2 scan. I derive the equation to determine the theoretical probabilities of each fragment isotope. The equation requires elemental compositions of the molecules to be given as input.

2. **Approximation of fragment isotope distributions by approximate elemental compositions.** Typically, elemental compositions are not known *a priori*. For these cases, I developed an approximation method that uses observed masses to approximate elemental compositions and then calculates isotope probabilities. It matches observed isotopic distributions with chi-squared scores within 2% of the theoretical fragment isotope distributions.

3. **Alternative approximation method using splines.** I show that isotope probabilities follow a tight non-linear pattern that depends on peptide mass. Cubic splines fit to these probabilities provide approximations with accuracy equal to the current state-of-the-art approach, but cubic splines are 20 times faster to evaluate.

4. **Approximation of isotope distributions with sulfur-specific models.** The unique isotope distribution of sulfur atoms causes inaccuracy in the approximations of small sulfur-containing peptides. I achieved more accurate approximations using a modified model and sulfur-specific splines that account for the expected number of sulfur atoms in the fragment and peptide.

I validated these methods experimentally through direct infusion experiments of angiotensin I peptide and a shotgun whole-cell HeLa lysate experiment on an Orbitrap Fusion Lumos mass spectrometer. Furthermore, I added the methods to the OpenMS software library, allowing the mass spectrometry community to use them to develop novel approaches to process MS2 spectra.

1.3.2 Deconvolution of chimeric spectra

Chimeric spectra are MS2 spectra that contain fragments from multiple distinct peptides. In complex samples containing millions of distinct peptide species, most MS2 spectra are chimeric. Sequencing algorithms, however, are designed for and perform best on spectra from a single peptide. Fortunately, fragments have isotope distributions that depend on their precursor. Leveraging this dependency, I developed a non-negative least squares regression (NNLS) model that can determine each fragmented precursor's contribution to an

MS2 spectrum. Using this model, I present a method to deconvolve a chimeric spectrum by decomposing it into separate spectra for each precursor peptide, and removing non-monoisotopic (Definition 2.1) peaks:

1. **NNLS model for deconvolution.** Approximate fragment isotope distributions are used as fuzzy basis templates and define the design matrix of the model described here. An observed MS2 spectrum is assumed to be a sparse linear combination of these templates whose coefficients need to be determined by solving a convex optimization problem. To promote the desired sparsity, the NNLS model is regularized with a sparse group lasso penalty. Separate spectra for each precursor are constructed from groups of templates with positive coefficients. The resulting spectra have fewer contaminating peaks from other peptides.

2. **De-isotoping and determination of monoisotopic mass.** Sequencing algorithms expect only monoisotopic peaks to be present in the spectrum. In practice, other isotopic peaks are present and prevalent. The model described here determines the isotopic state (Definition 3.2.1) and monoisotopic mass of each peak. Using this information, I developed a method that removes non-monoisotopic peaks, and adds their signal intensity to their corresponding monoisotopic peak. This makes the spectrum ideal for analysis by standard sequencing algorithms.

3. **Decoupling from sequencing algorithm.** Instead of creating a new sequencing algorithm that accounts for fragment isotopic distributions, which would add to an already bloated ecosystem, my deconvolution method is a pre-processing step independent of any sequencing algorithm. Consequently, the method is compatible with any sequencing algorithm. Furthermore, it is also independent and complementary to previous approaches to process chimeric spectra, which allows this approach and others to be integrated together into a more powerful pipeline.

This work describes the first application of approximate fragment isotope distributions. Using four different data sets from three different laboratories, the deconvolution procedure increased peptide-spectrum-matches (PSMs) by $>$17.7% on average (from 25,472 to 31,043), unique peptide identifications by 5% (from 15,237 to 16,203), and proteins identifications by $>$6% (from 3,455 to 3,699).

1.3.3 Simulation of data-dependent acquisition

The incredible complexity of biological samples and the insufficient scan speed of mass spectrometers necessitates judicious use of the instrument's duty time to maximize sequencing depth. Novel acquisition algorithms are difficult to test because few vendors provide an application programming interface for custom control of the instrument. To this end, I developed MSAcquisitionSimulator, a collection of C++ programs to simulate data-dependent acquisition (Definition 2.4.6) algorithms on *in silico* generated liquid chromatography-mass spectrometry (LC-MS) proteomics data. This software has the following novel features:

1. **A model to generate realistic peptide-spectrum-matches.** This allows for evaluation of acquisition algorithms with respect to the number of peptides and proteins identified. Existing simulation software attempt to simulate MS2 spectra, but these programs are inadequate for use with peptide sequencing software.

2. **Efficient pruning of low-abundant ions.** This results in better scaling with respect to runtime and memory for larger data sets. A case study with simulations containing over 45,000 proteins is provided. Existing simulators for LC-MS data require massive amounts of RAM and/or CPU time, making simulations of this size infeasible.

3. **Decoupled generation of ground truth data and the simulation of data acquisition.** Separate programs for these two tasks allows for the comparison of different algorithms on identical data without the computational cost of re-generating the ground truth.

MSAcquisitionSimulator fills a gap left by existing simulators. It provides an opportunity for additional research in an area critical to MS-based proteomics.

1.3.4 Prediction of protein-protein interactions

Decreasing instrumentation costs and improved technologies have made affinity purification-mass spectrometry (APMS) (Definition 2.4.8) approaches commonplace in academic science. With the vast amounts of data being produced, data quality control measures have become critically important, particularly for APMS technologies which suffer high false positive discovery rates. To address these concerns, I developed Spotlite, a machine learning classifier and web interface for scoring APMS data. The achievements of this work are:

1. **Comparative analyses of existing APMS scoring approaches.** Three popular and fundamentally different APMS scoring approaches (CompPASS, HGSCore, and SAINT) were evaluated on five disparate APMS data sets, revealing complementarity in identifying genuine protein-protein interactions from contaminants.

2. **Logistic regression model to predict protein-protein interactions.** To improve the scoring performance of CompPASS, HGSCore, and SAINT, a variety of non-MS data were integrated using a logistic regression model. Inclusion of these non-MS data improved APMS data classification by an average of 16% relative to the APMS scoring methods alone, as determined by the area under the curve (AUC) of the receiver operator characteristic (ROC) curves.

3. **User-friendly web application.** Because implementation of existing APMS scoring methods requires computational expertise beyond many laboratories, I developed a web application for APMS data scoring, analysis, annotation, and network visualization.

4. **Case study on the KEAP1 E3 ubiquitin ligase.** Through APMS analysis of KEAP1, Spotlite was employed to reveal true KEAP1 protein interactions as well as to annotate the interacting proteins for a variety of functional and disease-relevant characteristics.

The improved scoring performance of Spotlite combined with Spotlite's user-friendly, fast, and open-access web interface provides an invaluable resource for researchers to analyze and interpret APMS data.

1.4 Organization

The remainder of this dissertation is organized into the following chapters:

- **Chapter 2** introduces the fundamentals of mass spectrometry, proteomics, and relevant basic chemistry and biology. It includes definitions and nomenclature of terms used throughout the dissertation.

- **Chapter 3** presents equations to calculate the theoretical isotope distributions of peptide fragments. Several methods to approximate isotope distributions of both precursors and fragments are developed and described for cases when elemental compositions are not known *a priori*.

- **Chapter 4** provides an application for the methods outlined in Chapter 3. A non-negative least squares model is developed to deconvolve chimeric MS2 spectra into separate spectra for each precursor peptide. The model's performance is evaluated on experimental data from three different laboratories.

- **Chapter 5** describes software to simulate the data acquisition process of a mass spectrometer. It provides methods to generate ground truth data and to execute and evaluate user-developed acquisition algorithms for the interrogation of an *in silico* peptide mixture.

- **Chapter 6** details a logistic regression model to classify candidate protein-protein interactions from APMS data as *bona fide* interactors or contaminants. It describes a web application developed for researchers with less computational expertise.

- **Chapter 7** concludes this dissertation and discusses plans for future research.

CHAPTER 2: BACKGROUND

2.1 Atomic elements

Atomic elements are comprised of three subatomic particles: *protons*, *electrons*, and *neutrons*. Protons and electrons have positive and negative charges, respectively, of equal magnitude. An atomic element is defined by the number of protons in its nucleus, and elements with different numbers of neutrons are called *isotopes*. While neutrons have no charge, they have a mass of 1.008645 Daltons (Da), and cause isotopes to have different masses (Figure 2.1). Many atomic elements have multiple stable isotopes and their relative terrestrial abundances have been determined (Table 2.1). The natural abundances of an element can be thought of as a discrete probability distribution, or *isotope distribution*, for its isotopic state. These isotope distributions play an important role in mass spectrometry because a population of a single molecular entity is detected as multiple sub-populations with distinct masses. A molecule whose elements are all in their smallest stable isotopic state is said to be *monoisotopic*.

2.2 Protein structure

Amino acids are the building blocks of proteins and are composed of the following organic elements: hydrogen (H), carbon (C), nitrogen (N), oxygen (O), and sulfur (S). While the backbone of each amino acid is the same, they have different R groups that makes them unique (Figure 2.2). Both individual and chains of amino acids are written with the *N-terminal* side (NH_2) on the left and the *C-terminal* side (COOH) on the right. Chains of amino acids are created through the formation of *peptide bonds* (Figure 2.3). During the chemical reaction that forms a peptide bond, amino acids lose a water molecule and are then called amino acid residues. The remainder of this dissertation deals with amino acid residues and will refer to them simply as amino acids. *Proteins* are long chains of amino acids (Figure 2.4). There are 20 common amino acids and in proteomics they are often referred to by their single letter codes (Table 2.2). Except for leucine (L) and isoleucine (I), each amino acid has a different mass. A protein's AA sequence is its primary structure, and can be represented by a string of characters from a 20 letter alphabet. Certain permutations of

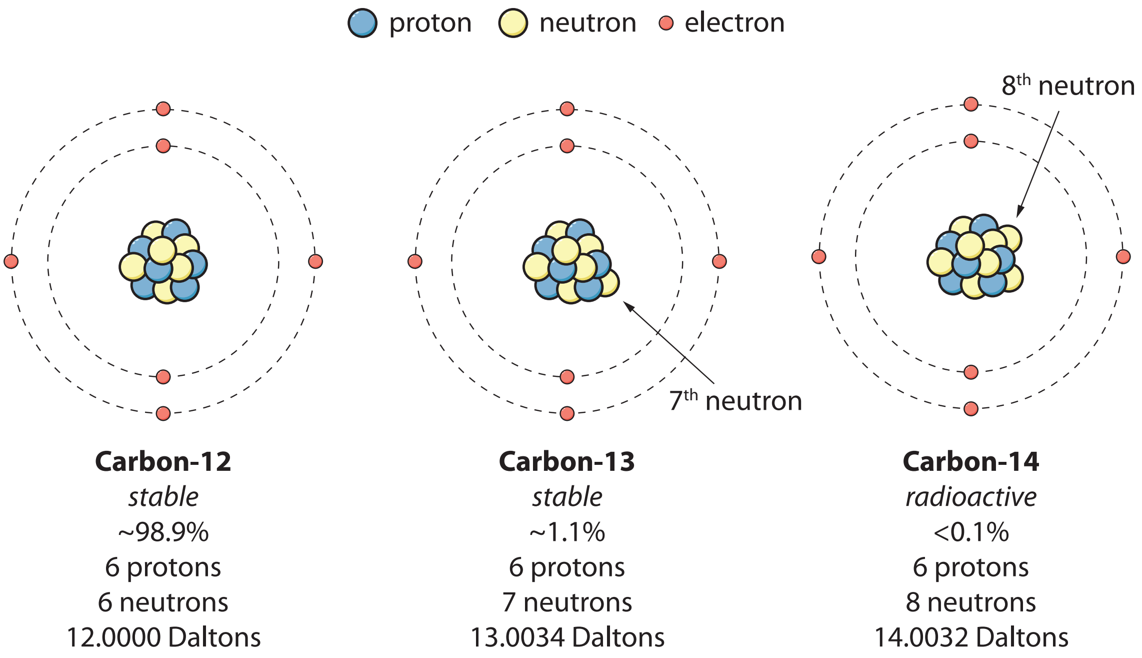


Figure 2.1: Schematic of three carbon isotopes. C^{12} and C^{13} are carbon's only stable isotopes. C^{14} is one of 13 known radioactive carbon isotopes. For each isotope, the numbers of protons and electrons remain constant while the number of neutrons varies. The change in mass between isotopes is not equal to the mass of the additional neutrons because some of the mass is converted into nuclear energy.

Table 2.1: Stable isotopes of common organic elements found in biology

Element	Symbol	Nominal Mass	Exact Mass (Da)	% Natural Abundance[1]
Hydrogen	H	1	1.0078	99.99
	^{2}H or D	2	2.0141	0.01
Carbon	^{12}C	12	12.0000	98.93
	^{13}C	13	13.0034	1.07
Nitrogen	^{14}N	14	14.0031	99.64
	^{15}N	15	15.0001	0.36
Oxygen	^{16}O	16	15.9949	99.76
	^{17}O	17	16.9991	0.04
	^{18}O	18	17.9992	0.20
Phosphorus	P	31	30.9738	100.00
Sulfur	^{32}S	32	31.9721	94.99
	^{33}S	33	32.9715	0.75
	^{34}S	34	33.9679	4.25
	^{36}S	36	35.9671	<0.01

[1] Natural abundances provided by IUPAC.

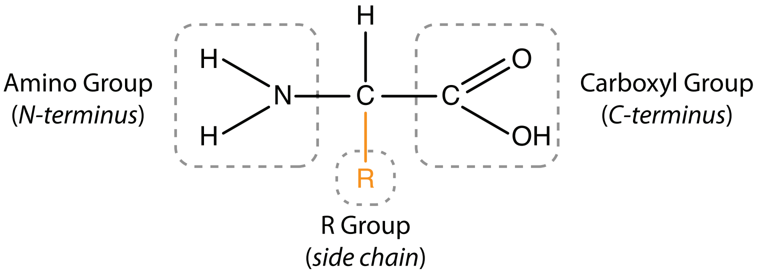


Figure 2.2: Schematic of an amino acid. All amino acids contain an *amino group* (NH_2, usually depicted on the left side), a *carboxyl group* (COOH, usually depicted on the right side), and the variable *R group*, all bound to a *central carbon* molecule. The R group is specific to each amino acid and determines its chemical properties.

amino acids form hydrogen bonds with nearby amino acids to create secondary structures called α helices and β sheets. These structures in turn fold onto each other to form a three dimensional tertiary structure. Finally, multiple proteins will physically bind to each other to create what are called *protein complexes* and referred to as a quaternary structure. These protein complexes function as small molecular machines. Two proteins that part of the same complex are said to have a *protein-protein interaction* and are *co-complexed*. If co-complexed proteins are physically touching each other, then they have a *direct interaction*, otherwise they have an *indirect interaction*.

2.2.1 Digestion

A typical sample preparation step in MS-based proteomics is to enzymatically digested proteins into peptides. *Peptides* are AA subsequences of the protein AA sequences (Figure 2.5). This type of workflow is known as *shotgun proteomics*. The most common enzyme utilized for digestion is trypsin, which cleaves at the C-terminal side of arginines (R) and lysines (K). Due to the distribution of R's and K's in the human proteome, this creates peptides that are predominantly 7-20 amino acids long. The length of a peptide or protein is

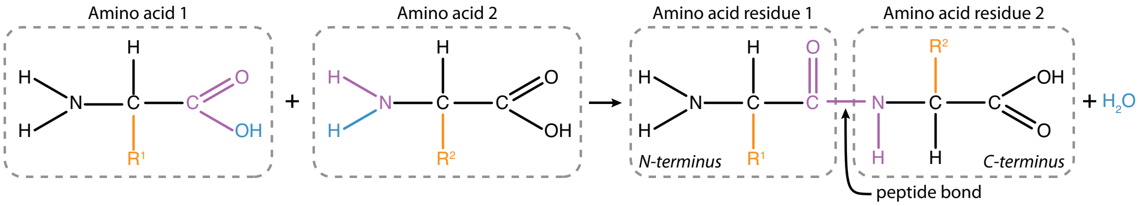
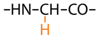
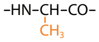
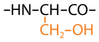
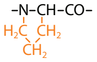
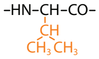
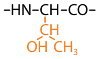
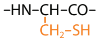
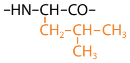
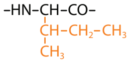
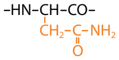
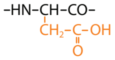
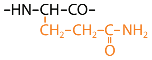
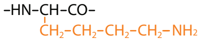
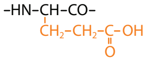
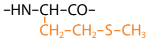
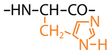
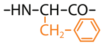
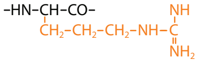
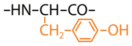
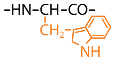
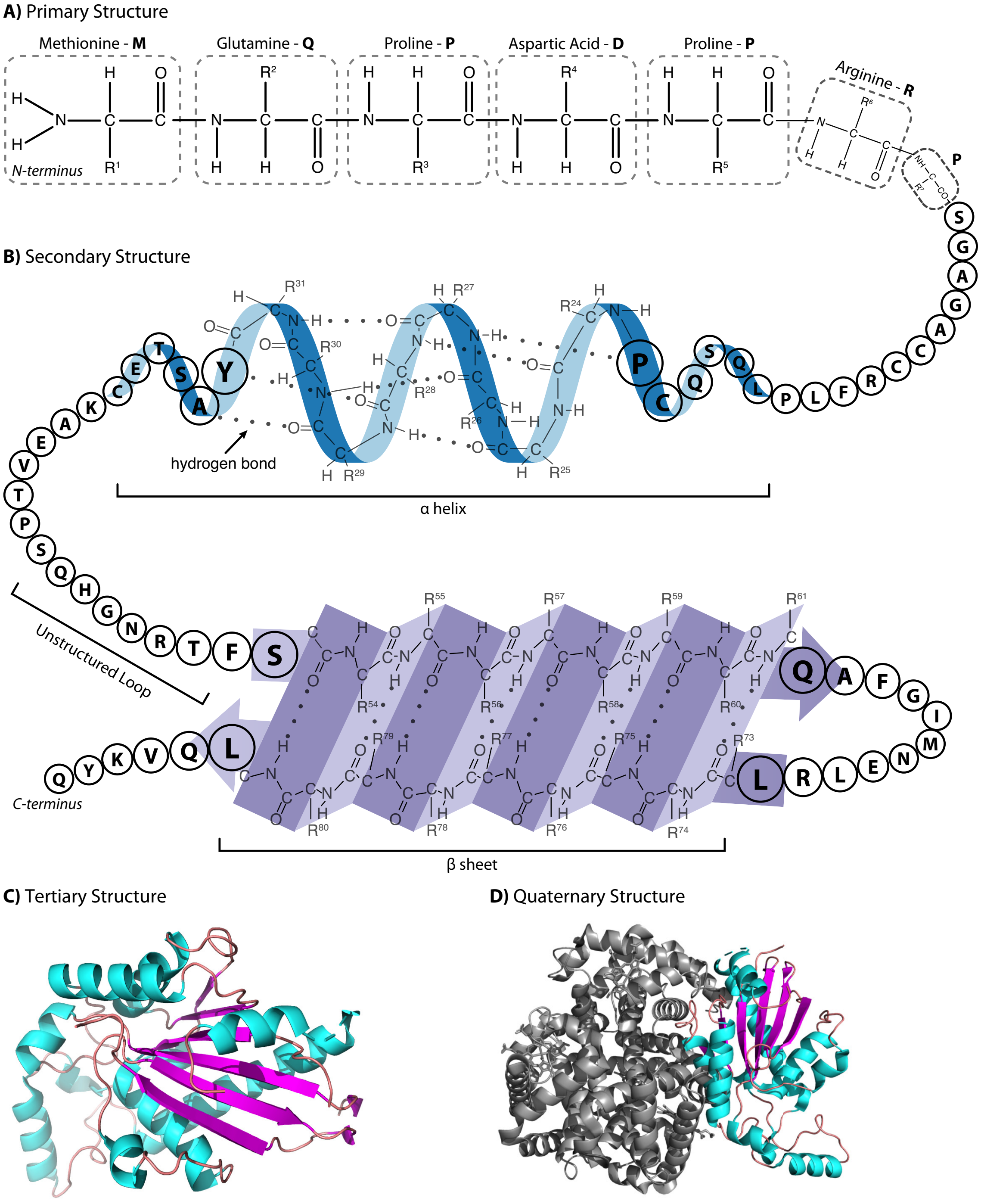


Figure 2.3: Formation of a peptide bond. Two amino acids form a peptide bond when the carboxyl group of one amino acid reacts with the amino group of the other amino acid. This chemical reaction results in the release of a water molecule and the joining of the two amino acids.

Table 2.2: Amino acid residues

Name	Symbols		Elemental Composition	Monoisotopic Mass	Structure
Glycine	Gly	**G**	C_2H_3NO	57.021464	
Alanine	Ala	**A**	C_3H_5NO	71.037114	
Serine	Ser	**S**	$C_3H_5NO_2$	87.032029	
Proline	Pro	**P**	C_5H_7NO	97.052764	
Valine	Val	**V**	C_5H_9NO	99.068414	
Threonine	Thr	**T**	$C_4H_7NO_2$	101.04768	
Cysteine	Cys	**C**	C_3H_5NOS	103.00919	
Leucine	Leu	**L**	$C_6H_{11}NO$	113.08406	
Isoleucine	Ile	**I**	$C_6H_{11}NO$	113.08406	
Asparagine	Asn	**N**	$C_4H_6N_2O_2$	114.04293	
Aspartic Acid	Asp	**D**	$C_4H_5NO_3$	115.02694	
Glutamine	Gln	**Q**	$C_5H_8N_2O_2$	128.05858	
Lysine	Lys	**K**	$C_6H_{12}N_2O$	128.09496	
Glutamic Acid	Glu	**E**	$C_5H_7NO_3$	129.04259	
Methionine	Met	**M**	C_5H_9NOS	131.04048	
Histidine	His	**H**	$C_6H_7N_3O$	137.05891	
Phenylalanine	Phe	**F**	C_9H_9NO	147.06841	
Arginine	Arg	**R**	$C_6H_{12}N_4O$	156.10111	
Tyrosine	Tyr	**Y**	$C_9H_9NO_2$	163.06333	
Tryptophan	Trp	**W**	$C_{11}H_{10}N_2O$	186.07931	

Compositions, masses, and structures are for internal amino acid residues and therefore exclude both an N-terminal H and a C-terminal OH group that would be removed during the creation of a peptide bond on each terminus.



A) Primary Structure

B) Secondary Structure

C) Tertiary Structure

D) Quaternary Structure

Figure 2.4: Four levels of protein structure. (A) The primary structure of a protein is the sequence of amino acids that form the protein, and is usually represented as a string of characters. Each character represents a single amino acid. (B) The secondary structure of a protein consists of the local three-dimensional folds of nearby amino acids. The common secondary structures are α-helices, β-sheets, and unstructured loops. (C) The tertiary structure of a protein is the overall three-dimensional shape that a protein takes after its secondary structures are folded together. (D) The quaternary structure of a group of proteins is their physical arrangement when bound together in a protein complex.

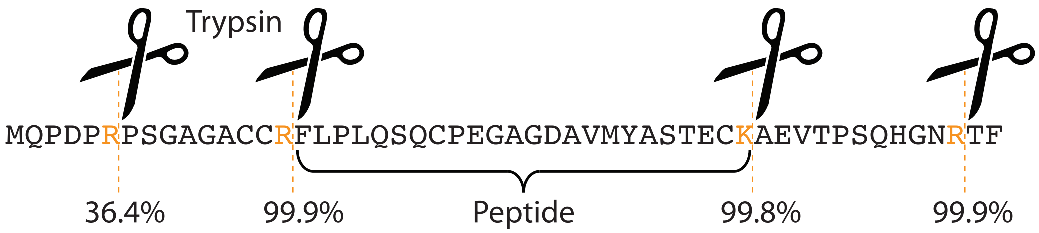


Figure 2.5: Enzymatic digestion of a protein into peptides. Trypsin preferentially cleaves peptide bonds at the C-terminal side of R's and K's (orange). The neighboring amino acids affect digestion efficiency, with C-terminal P's having the most significant inhibitory effect. The peptide AA sequences will later be identified using mass spectrometry.

the number of amino acids in its sequence. While peptides of this length are amenable for measurement by mass spectrometry, they also complicate protein identification because some peptides belong to multiple proteins. Due to the low occurrence of missed cleavages and non-tryptic digestion, overlapping peptides are low-abundant and therefore often not identified. The probability of tryptic cleavage is affected by the amino acids surrounding the candidate site. For example, prolines (P) at the C-terminal side of RK's inhibits trypsin activity. However, even under ideal conditions, 100% cleavage is unlikely. Moreover, other enzymes present in a cell cause cleavages at unexpected locations. These missed cleavages and non-tryptic digestion sites create many different peptides in a sample. Additionally, the amino acids can be chemically modified which changes a peptide's mass. Hundreds of chemical modifications exist and can occur in combinations on a single peptide. Therefore, a complex peptide mixture may contain millions to billions of distinct peptides, most of which are low-abundant.

2.3 Liquid chromatography

Due to the large number of distinct peptides in a sample, the peptides need to be sorted, separated, and introduced into a mass spectrometer in an orderly fashion so that the instrument can analyze fewer peptides at a time. To achieve this, *liquid chromatography* (LC) is the most common separation technique employed (Figure 2.6). Beads coated with chains of carbon (stationary phase) are densely packed into columns. A peptide mixture is loaded on to the column and a liquid (mobile phase) is flowed through it. Initially, hydrophilic peptides begin to travel down the column. Over time, the concentration of an organic liquid is increased, causing gradually more hydrophobic peptides to start moving. These chromatographic separations typically use gradients lasting several minutes to a few hours long. As the peptides elute from the column, they are introduced to the mass spectrometer ion source. Typically, copies of a particular peptide

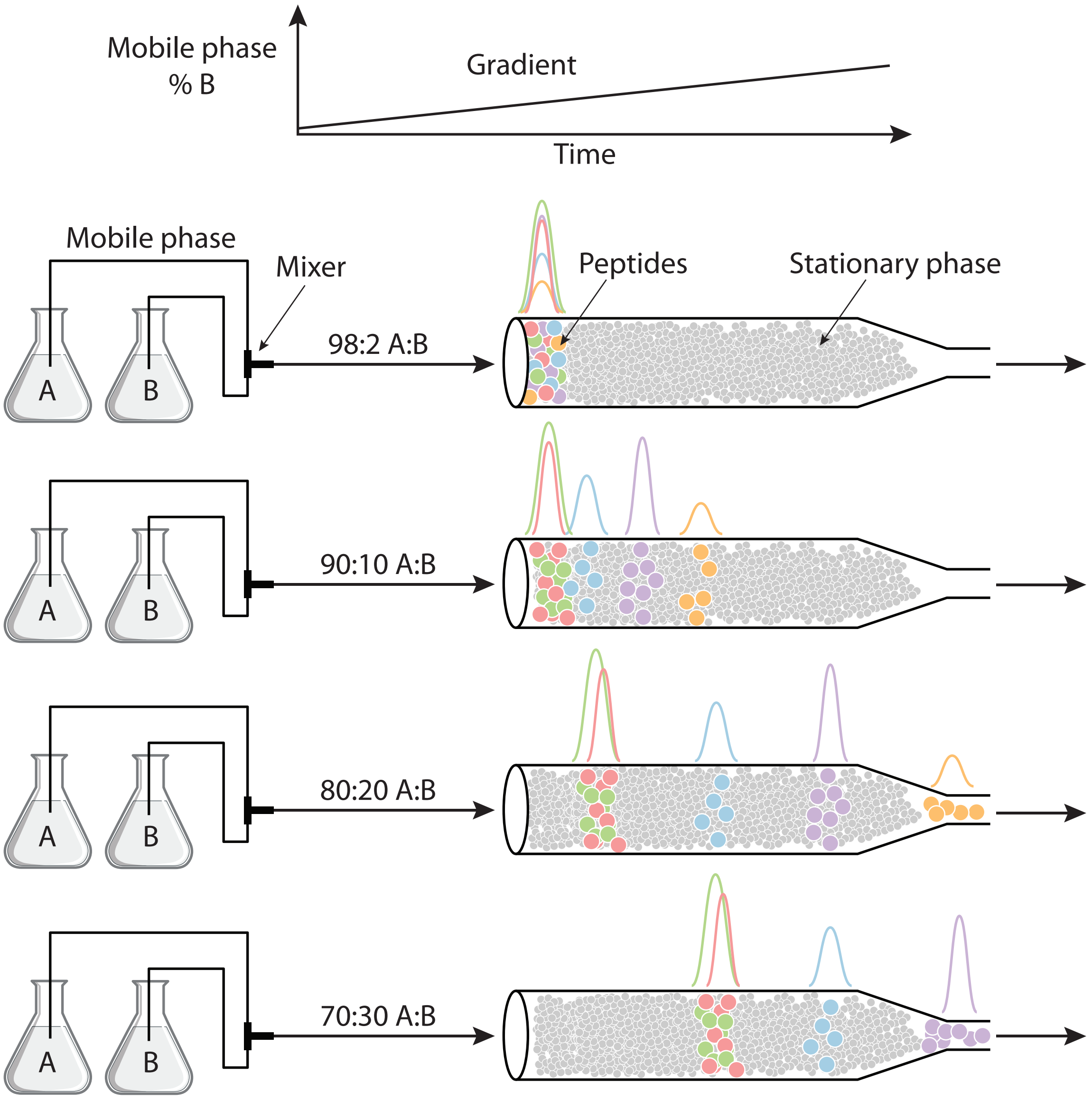


Figure 2.6: Schematic of liquid chromatography. Liquid chromatography is used to separate peptides over time. Peptides (colored) bind the stationary phase of the column (grey), but also bind the molecules of the mobile phase that are flowing through the column and cause the peptides to travel in the direction of flow. The ratio of an organic mobile phase B relative to the aqueous mobile phase A is increased over time in order to elute increasingly hydrophobic peptides (red and green) in an orderly fashion. This separation keeps downstream detectors from being overwhelmed by too many different peptides.

will elute from a column over a span of several seconds. Due to the complexity of some peptide mixtures and imperfect separations, many peptides simultaneously elute from the column.

2.4 Mass spectrometry

Mass spectrometers measure the mass-to-charge ratios (*m/z*) and signal intensities of ion populations. A *signal* is a specific (*m/z*, intensity) pair. *Ions* are charged molecules, and they are necessary for analysis by mass spectrometry because electromagnetic fields are used to influence their trajectories. The trajectory of an ion depends on its *m/z* value and it is this relationship that is used to determine the *m/z*. A basic mass spectrometer contains an ion source to generate ions, a mass analyzer to separate the ions by *m/z*, and a detector to measure the electrical charge or current produced by the ions. Detectors are not sensitive enough to detect a single ion, so populations of ions are required. Most modern mass spectrometers also contain collision cells to fragment the ions, which is necessary for AA sequence identification.

2.4.1 Ionization

To create ions from the peptides eluting from a liquid chromatography column, the end of the column is fitted with a needle (Figure 2.7). A high voltage is applied to the needle to provide a source of protons. As the mixture of peptides and liquid sprays out from the needle tip, droplets are formed containing both peptides and protons. Under atmospheric conditions, the droplets begin to evaporate until the liquid is gone and the protons are transferred to a neighboring peptide. This technique creates multiply-charged peptide ions and is called electrospray ionization. The newly-formed ions make their way into the mass spectrometer where their *m/z*'s and signal intensities will be measured. It is important to distinguish signal intensity from a molecule's abundance. *Signal intensity* refers to the measured electrical output due to ions detected by a mass spectrometer, while a molecule's *abundance* refers to the actual number of copies present in the biological specimen under investigation. Signal intensity and abundance are loosely correlated.

2.4.2 Mass analyzers

All *mass analyzers* use electromagnetic fields to influence the trajectory of an ion in order to determine its *m/z*. The path of an ion depends on its *m/z* and known instrument parameters. Figure 2.8 shows a schematic of an Orbitrap mass analyzer. To start mass analysis with an Orbitrap, a tightly packed population of ions

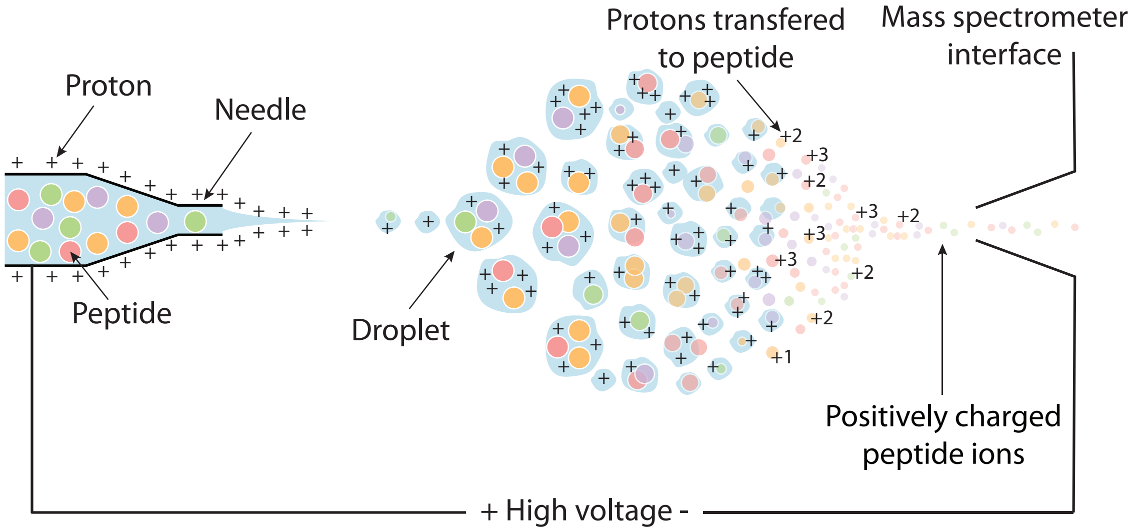

Proton
Needle
Peptide
Droplet
Protons transfered to peptide
Mass spectrometer interface
Positively charged peptide ions
+ High voltage -
+1
+2
+3

Figure 2.7: Schematic of electrospray ionization. Peptides embedded in a liquid (blue) emerge as a droplet at the end of a needle. A high voltage applied to the needle supplies protons, and these protons are transferred to peptides as the droplets evaporate. The resulting peptide ions (denoted by their charges: +1, +2, and +3) enter the mass spectrometer interface.

is injected into the mass analyzer. The ions begin to revolve around the central electrode and separate into distinct bands for each *m/z*. The bands periodically travel back and forth along the length of the electrode in a sinusoidal manner. A signal amplifier and detector is located in the middle of the top and bottom electrodes. They detect the electrical current produced as ions pass the center of the mass analyzer. The detected signal is recorded over time and a fast Fourier transform is performed to recover the *m/z*'s and signal intensities of the ions in the Orbitrap mass analyzer.

2.4.3 Isotope distributions

A population of a single molecular entity is a mixture of sub-populations with different isotope compositions and therefore different masses. A molecule's natural (or *precursor*) *theoretical isotope distribution* can be determined from its elemental composition using polynomial expansion. For example, let A_x represent the probability of a specific isotopic element, then the isotope distribution of $C_{34}H_{53}N_7O_{15}$ can computed by expansion of the polynomial expression:

$$(A_{12C} + A_{13C})^{34} \cdot (A_H + A_{2H})^{53} \cdot (A_{14N} + A_{15N})^7 \cdot (A_{16O} + A_{17O} + A_{18O})^{15}$$

After expansion, each product term corresponds to a specific combination of elemental isotopes. The exponent signifies the count of the corresponding elemental isotope in the molecule. Evaluating the product term gives the probability of that combination. This approach determines a *fine isotope distribution*: terms

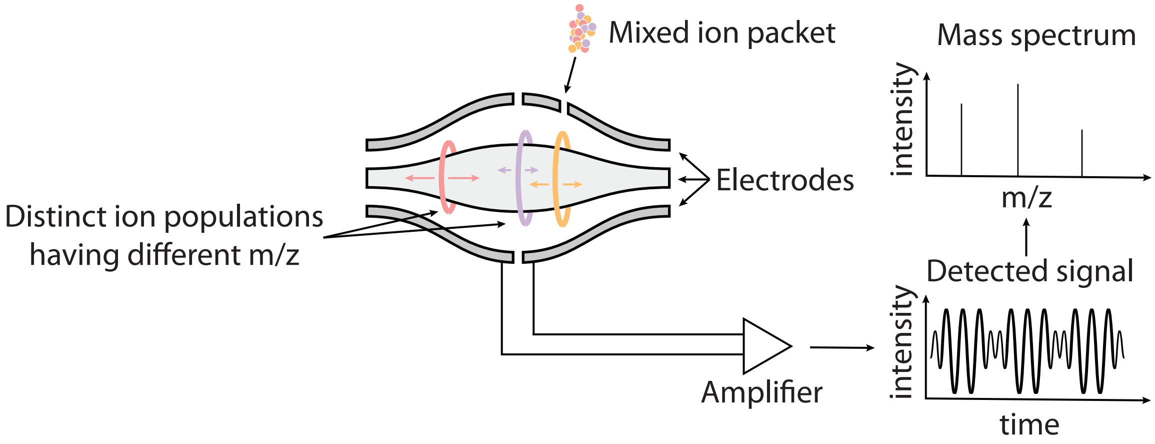


Figure 2.8: Schematic of an Orbitrap mass analyzer. A central electrode is surrounded by two outer electrodes. Ions oscillate back and forth along the central electrode at frequencies that depend on their *m/z*. As the ions move, they create an electric current, and amplifiers connected to the outer electrodes increase the signal before being measured by a detector. Ions are allowed to oscillate for 16-256ms, followed by the application of an enhanced Fourier transform to determine the *m/z*'s and signal intensities of the individual components that created the detected signals.

with the same number of extra neutrons but different isotopic elements are treated separately because they have slightly different masses. To compute a *nominal isotope distribution*, the terms with the same number of neutrons must be added together. Nominal isotope distributions are more common in mass spectrometry due to the inability to differentiate between extremely similar masses. The monoisotopic population of a molecule is denoted as M, and subsequent isotopes are denoted as $M + i$ where i is the number neutrons greater than the monoisotope present in the molecule.

2.4.4 MS1 scans

An *MS1 scan* measures the *m/z* and signal intensity of ions currently entering the mass spectrometer (Figure 2.9). These scans are acquired periodically over the course of an experiment. Consecutive MS1 scans have similar signals because they are performed faster and more often than the time it takes for a peptide species to completely elute from the LC column. A peptide species will first elute from the column at low abundance, then gradually increase in abundance until reaching an apex, and then decrease until it is finally no longer detected. Therefore, copies of a particular peptide will be observed in multiple consecutive MS1 scans and its elution profile can be stitched together from the MS1 data. When referring to data in a mass spectrum that are related to isotopes, the adjective *isotopic* will used to differentiate it from the concept of a probability distribution. Furthermore, the isotopic data can be *theoretical* is it based on known elemental

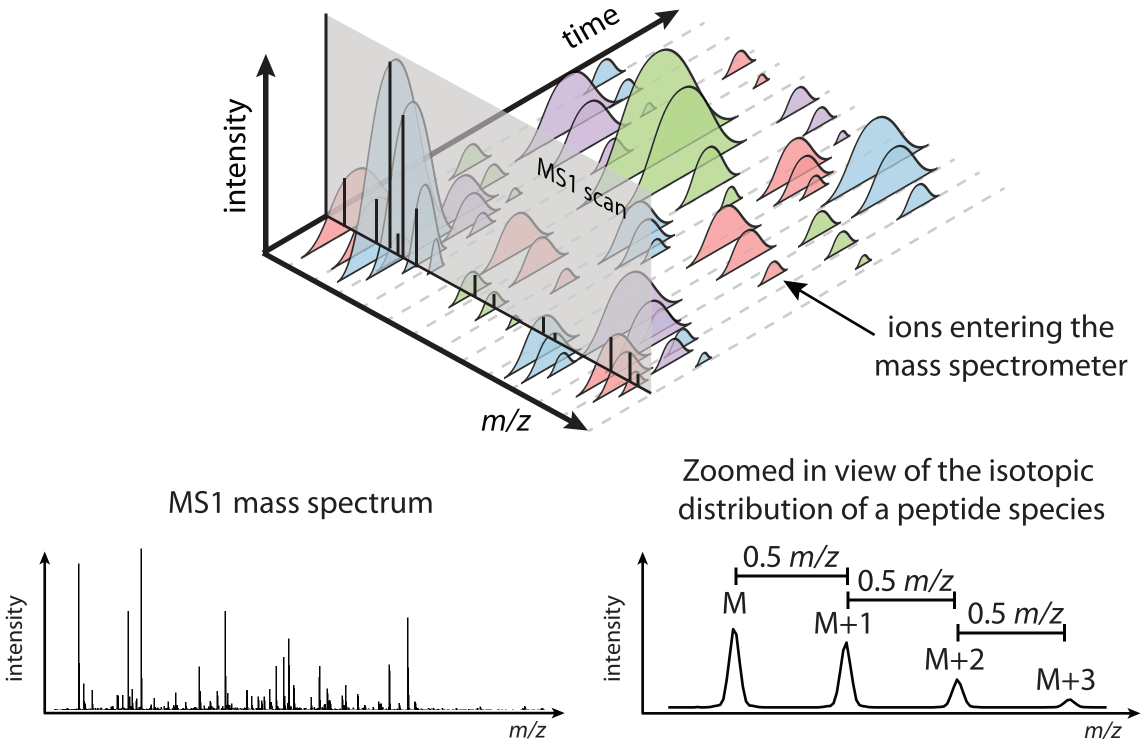

time
intensity
MS1 scan
ions entering the mass spectrometer
m/z
MS1 mass spectrum
intensity
m/z
Zoomed in view of the isotopic distribution of a peptide species
0.5 m/z
0.5 m/z
0.5 m/z
M
M+1
M+2
M+3
intensity
m/z

Figure 2.9: Liquid chromatography–mass spectrometry data consists of three dimensions: time, *m/z*, and signal intensity (top). During an MS1 scan, the mass spectrometer allows all ions into the mass analyzer and determines their *m/z*'s and intensities. This results in an MS1 spectrum: the mass spectrum of the ions in the mass spectrometer at a particular time (bottom left). With a *high-resolution* mass analyzer, nominal isotopes will create separate peaks and can be used to determine the charge and mass of an ion (bottom right).

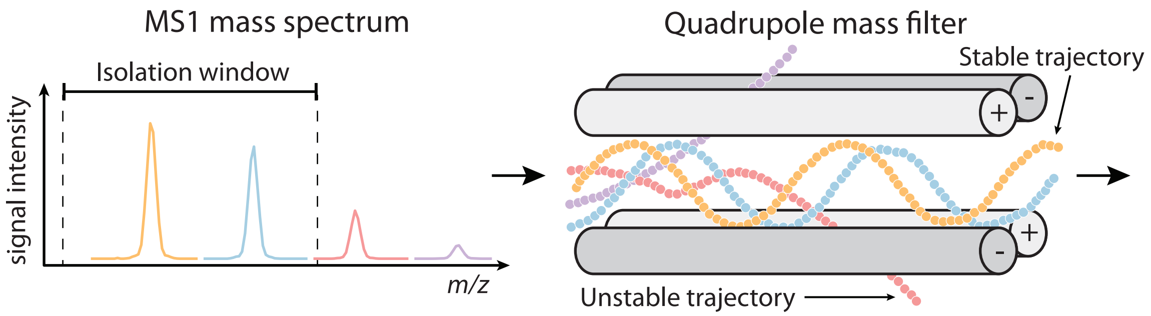


Figure 2.10: Ion isolation with a quadrupole mass filter. An isolation *m/z* center and width is specified to capture a set of ions (left). To achieve the isolation of only the ions whose *m/z* falls within the bounds, a quadrupole mass filter consisting of 2 pairs of cylindrical poles will toggle their voltage polarity at a precise frequency and voltage strength such that only the desired ions will have stable trajectories when moving through the mass filter (right). Ions with stable trajectories will make it through to the next stage of the mass spectrometer, while ions with unstable trajectories will collide with the quadrupole's housing, lose their charge, and fall to the bottom of the chamber.

compositions, *approximate* if it is based on unknown elemental compositions, or *observed* if it is referring to experimental data measured by an actual mass analyzer. In an *MS1 mass spectrum* (or MS1 spectrum for short), isotopic distributions of peptides are clearly visible (Figure 2.9). Each nominal isotope will create an isotopic *peak*, which is a collection of signals that appear to represent the same molecular entity. The *m/z* spacing between isotopic peaks is used to determine the charge and mass of the observed peptide. Since an extra neutron adds ∼1.003 Da to a molecule, if the *m/z* difference between two isotopic peaks is ∼0.5, then z must equal 2. The mass can then be computed by solving for m. An accurate mass measurement is not sufficient to identify a peptide, however, because permutations containing the same combination of amino acids will have the identical mass. To elucidate the AA sequence, further mass analysis is necessary.

2.4.5 Tandem mass spectrometry

Tandem mass spectrometry refers to the use of two rounds of mass analysis. After an MS1 scan determines the mass of a peptide, the peptide is isolated, fragmented, and the mass spectrum of its fragments is measured. This type of scan is called an *MS2 scan*. MS2 spectra contain the *m/z* and signal intensity of fragments which are often indicative of a peptide's AA sequence.

2.4.5.1 Ion isolation

Ion isolation refers to the accumulation of desired ions, or conversely, the filtering of undesired ions. It is used to accumulate an ion population of interest for further study. An *isolation window* is defined by its *m/z* center and *m/z* width. Ions whose *m/z* values are within the bounds of the window are isolated, and those outside are filtered out. The amount of time spent accumulating ions is called *injection time*. An example of a quadrupole mass filter is shown in Figure 2.10. Voltages and radio frequency fields are adjusted so that only ions with the proper *m/z* will have stable trajectories. Ions with *m/z* values outside the bounds will not travel to the end of the quadrupole.

Typically, an isolation window is centered on either on a peptide's monoisotopic peak or the isotopic peak with the strongest signal intensity. Narrower isolation windows centered on the desired ions have three important effects. First, advantageously, they increase the proportion of desired ions that are isolated relative to other ions because it's less likely other ions are within the window's bounds. Second, however, narrower windows also have decreased isolation efficiency because even ions within the window's bounds start to have unstable trajectories if their initial velocities are outside an ever decreasing acceptable range. Longer injection times are necessary to counter the decreased isolation efficiency. Third, narrower windows will isolate a *partial isotopic distribution* where some isotopic peaks will fall outside the isolation window. Isolation of a partial isotopic distribution leads to different isotopic distributions in subsequent mass spectra and must be accounted for during data analysis. Therefore, isolation window widths balance specificity and sensitivity, and change observed isotopic distributions.

2.4.5.2 Fragmentation

Once a population of ions is isolated, they can be fragmented in a collision cell inside the mass spectrometer (Figure 2.11). Prior to fragmentation, the ions are called *precursors*. The precursors enter the collision cell at high velocity and collide with gas particles. The collisions cause the peptides to vibrate violently until a bond is broken and two complementary B/Y fragments are created. The fragments are then sent to a mass analyzer to measure their *m/z* and signal intensity. *Sequencing algorithms* determine the most likely AA sequence to have generated the observed fragmentation pattern.

Fragmentation pathways refer to the location and relative frequency of a break at a particular chemical bond compared to other chemical bonds. There are three bonds on each amino acid that can be broken to

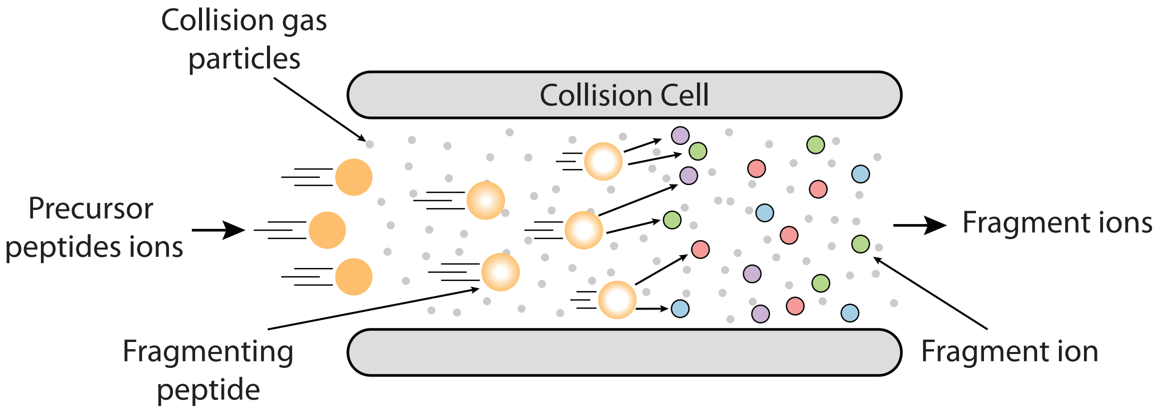


Figure 2.11: Schematic of collision-induced dissociation. A collision cell is populated by gas particles such as helium, nitrogen, or argon. Peptide ions are propelled into the collision cell at high energy and collide with the gas particles causing the peptides to dissociate into smaller fragments. Some fragments will not have a charge because the protons remained associated with the complementary fragment. These neutral particles will not be affected by electromagnetic forces and therefore cannot be measured by the mass analyzer. Fragment ions, however, continue on to the next stage of the mass spectrometer.

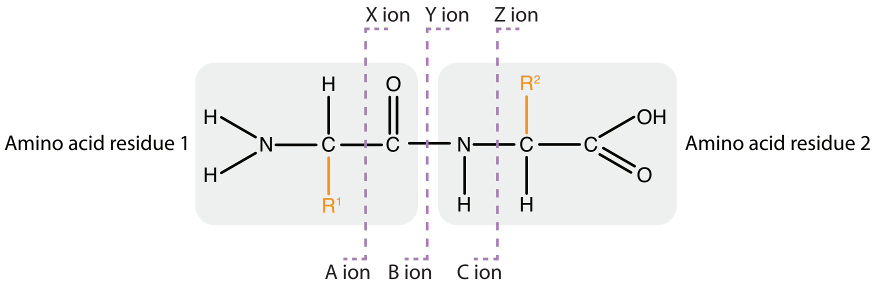


Figure 2.12: Common fragmentation pathways. Two amino acids are shown connected by a peptide bond. Dashed lines indicate the three bonds that, when individually broken, will dissociate the two amino acids and provide sequence information. For notation, a break between the central carbon and the carbon of the carboxyl group can generate A and X ions. An ion that contains atoms from the peptide N-terminus to the central carbon is an A ion. An ion that contains atoms from the peptide C-terminus to the carboxyl carbon is an X ion. B ions contain atoms from the peptide N-terminus to the carboxyl carbon. Y ions contain atoms from the peptide C-terminus to the amino nitrogen. C ions contain atoms from the peptide N-terminus to the amino nitrogen. Z ions contain atoms from the peptide C-terminus to the central carbon.

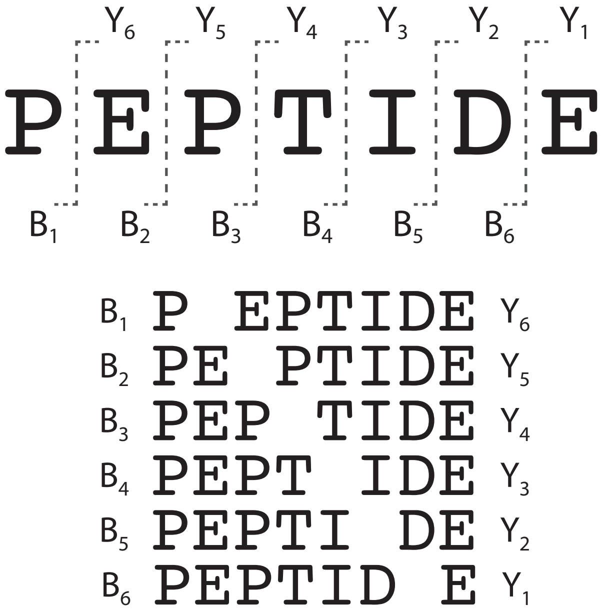


Figure 2.13: A peptide consisting of seven amino acids and its possible B/Y fragments are shown. The fragment notation extends to sequences of any length n. The index for a fragment ion represents its amino acid length, and therefore ranges from 1 to $n - 1$.

dissociate chains of amino acids from each other, and fragment ions have different notation depending on which bond was broken (Figure 2.12). Collision-induced dissociation predominantly causes breaks at peptide bonds because they are the weakest bond among the three options. This dissociation results in B and Y fragment ions. Fragments containing the amino acids from the N-terminus up to the broken peptide bond are called B ions, while fragments starting at the peptide bond and ending at the C-terminus are called Y ions. The fragments are further indexed based on the number of amino acids in the fragment (Figure 2.13). After fragmentation, the ions are sent to a mass analyzer. An example of an MS2 mass spectrum of a peptide annotated with matching B/Y fragment ions is shown in Figure 2.14. The number of pluses following an ion label indicate its *charge state*: the molecule's number of protons minus its number of electrons.

2.4.6 Data acquisition

The algorithm used by a mass spectrometer to decide which scans to perform is called a *data acquisition strategy*. When the strategy makes decisions based on the data it is observing, it is called *data-dependent acquisition* (DDA). Conversely, when the observed data is not used to drive the decision-making process, the strategy is called *data-independent acquisition* (DIA). The most common data acquisition strategy is TopN

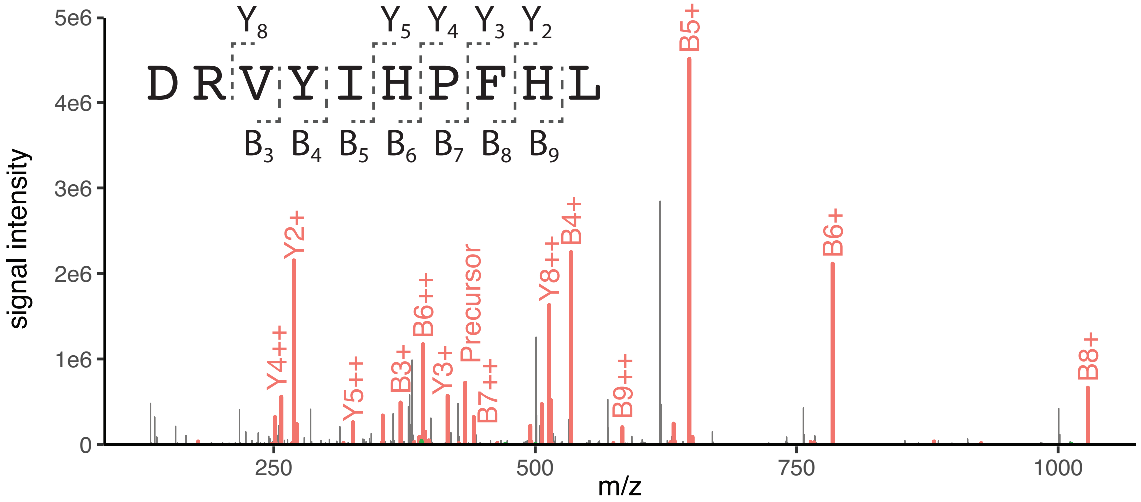


Figure 2.14: MS2 mass spectrum of angiotensin I peptide analyzed on an Orbitrap analyzer after CID fragmentation. Peaks that match expected B/Y ion *m/z*'s are shown in red and their location are annotated on the peptide's AA sequence. Some complementary B/Y ions are not observed because they either did not retain any protons, underwent further fragmentation, or a different fragment ion was responsible for the other peak. Unmatched peaks are colored black.

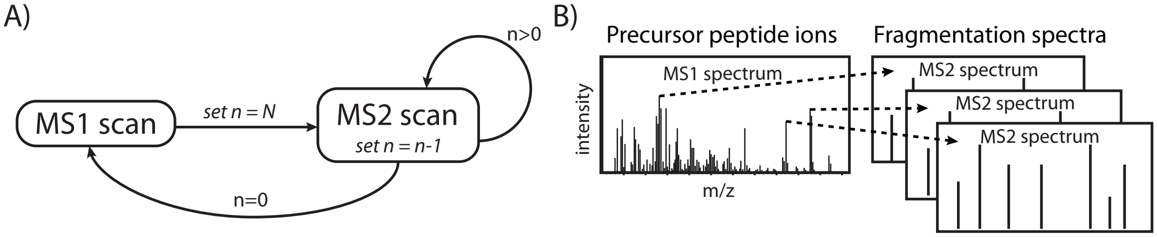


Figure 2.15: (A). Schematic of the TopN data-dependent acquisition algorithm. The mass spectrometer begins with an MS1 scan, performs up to N MS2 scans if there are enough peaks to target, and then repeats this process until the end of the experiment. (B) The N most intense peaks from an MS1 spectrum are chosen as targets to fragment in the following MS2 scans.

data-dependent acquisition. In TopN, MS1 scans are followed by N MS2 scans (Figure 2.15 A). Each MS2 scan isolates and fragments the next most intense monoisotopic peak from the previous MS1 scan (Figure 2.15 B). This process is repeated until the experiment is complete. In order to avoid redundant MS2 scans of the same peptide, a feature called dynamic exclusion is used: once a peak is targeted for an MS2 scan it is not allowed to be the target of another scan for a period of time defined by the user. At the end of the experiment, a file containing the collection of MS1 and MS2 spectra is written and later computationally processed to determine the identity and quantities of peptides and proteins.

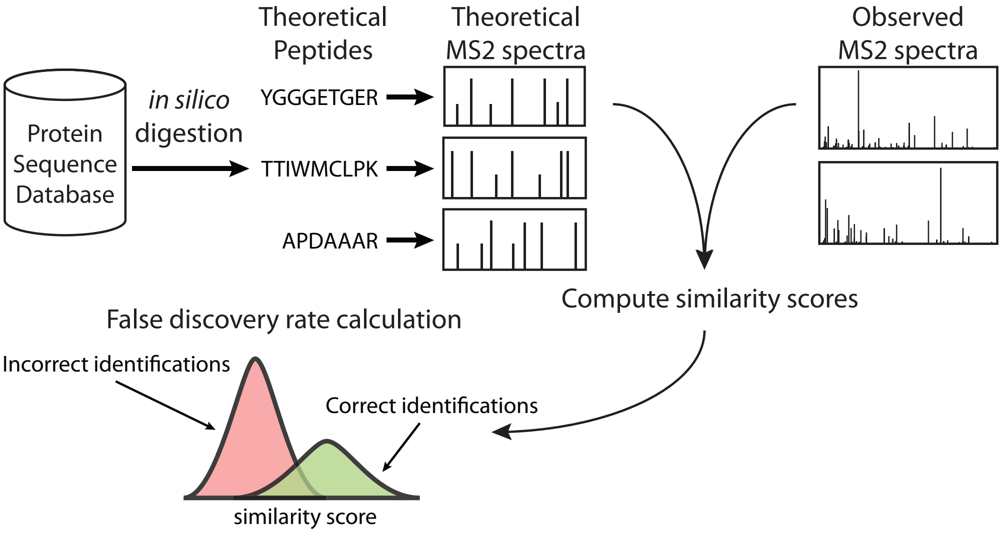


Figure 2.16: Schematic of a database search algorithm. Theoretical MS2 spectra are generated for peptides from a reference database of protein AA sequences. Theoretical MS2 spectra are compared to observed MS2 spectra using one of many developed similarity scores. A high score is indicative of a correct peptide-spectrum-match and therefore peptide identification. The distribution of similarity scores can be modeled by a two-component mixture model in which the distribution with higher scores (green) contains correct identifications, and the distribution with lower scores (red) stems from incorrect identifications. False discovery rates and p-values are computed based on these distributions.

2.4.7 Peptide identification

When the experiment is performed on a peptide mixture with a known proteome, peptide AA sequences can be identified using a database search algorithm (Figure 2.16). In a database search, a reference database of protein AA sequences database is digested *in silico* to create a list of possible peptide AA sequences. For each peptide, a theoretical MS2 spectrum is computed using simple assumptions about expected fragmentation pathways: all monoisotopic B and Y ions are present with equal intensities, and the additional loss of H_2O and NH_3 are present with lower intensities. Each observed MS2 spectrum is then compared to theoretical spectra whose precursor peptide mass is equal to the peptide mass of the MS2 target (± a small error tolerance). A similarity score is computed and the best peptide-spectrum-matches (PSMs) are recorded. Finally, PSMs passing a 1% false discovery rate (FDR) are accepted as confident identifications. Alternatively, sequences can be determined without a reference database using *de novo sequencing* algorithms, but these have lower success rates and are not the focus of this dissertation.

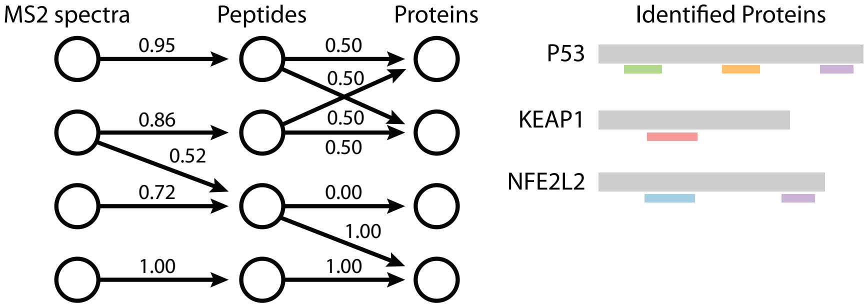


Figure 2.17: Schematic of the protein inference problem. (Left) Protein identifications are determined from a collection of observed MS2 spectra. Each MS2 spectrum is typically matched to a single peptide, however they can also match to multiple peptides, creating a many-to-many relationship. Each peptide may belong to one or more proteins from a reference database for another many-to-many relationship. Probabilities are computed for each peptide-spectrum-match and peptide-to-protein mapping. (Right) Proteins are accepted as identified (grey) based on identified and mapped peptides (colored).

2.4.8 Protein inference

The probability of a protein identification is calculated after mapping identified peptide AA sequences to protein AA sequences from a reference database (Figure 2.17). Though many peptide AA sequences are unique to a particular protein and are therefore trivial to assign, some are shared between multiple proteins and their assignment is therefore ambiguous. In ambiguous cases, probabilities are computed for each candidate protein assignment by taking into account the presence of other observed peptides that map to the same proteins. For example, a shared peptide will be assigned to a protein with many unique peptides instead of a protein with zero unique peptides. The probability of a protein's presence is computed from the identification scores of the PSMs and the weights of their assignment to the protein. Finally, a FDR is computed and controlled at the protein level. In *affinity purification-mass spectrometry* (APMS) experiments, protein complexes that contain a protein of interested are first extracted from a biological specimen followed by the identification of the co-complexed proteins by mass spectrometry.

CHAPTER 3: ISOTOPE DISTRIBUTIONS OF FRAGMENT IONS

3.1 Introduction

In mass spectrometry (MS)-based proteomics, peptide AA sequences are determined by performing MS2 scans, which isolate and subsequently fragment precursor ions. Frequently, only part of a precursor's isotopic distribution is captured due to isolation windows that are too narrow or are offset relative to the precursor. Experiments using data-dependent acquisition typically use isolation windows that are 1.4-4 *m/z* wide (Michalski et al., 2011; Scheltema et al., 2014). With a 1.4 *m/z* wide isolation window, only one to three isotopic peaks of a charge +2 peptide can fit within its boundaries. For example, if the window is centered $>$0.2 *m/z* below the monoisotopic peak, then only the monoisotopic peak would be isolated. This can occur for co-eluting peptides that were not the intended target of an MS2 scan because their *m/z* position relative to the isolation window is random. Since co-fragmentation is encountered in as many as 50% of MS2 spectra of complex samples, isolation of unexpected isotopes from co-eluting peptides is common (Houel et al., 2010). The isolation of only some isotopes leads to fragments with complex isotope distributions; these distributions depend on the subset of isolated precursor isotopes and the elemental compositions of both the precursor and the fragment of interest. While a general method to calculate the theoretical isotope distribution of a fragment has been developed, this method requires exact knowledge of those inputs (Rockwood et al., 2003). Typically, peptide AA sequences and elemental compositions are unknown *a priori*. Therefore, computational tasks that occur prior to sequence determination, including MS2 de-isotoping, monoisotopic mass calculation, charge assignment of fragment peaks, and chimeric spectra deconvolution, do not take full advantage of fragment isotopic distributions. In order to improve these pre-processing endeavors and to increase protein and peptide identifications, an efficient method is needed to approximate theoretical fragment isotope distributions based on observed peaks and isolation window parameters.

The isotope distribution of a molecule arises from the varying number of neutrons in its individual elements. In mass spectrometry, there are two types of isotope distributions to consider: precursor (or natural) isotope distributions and fragment isotope distributions. A molecule's precursor isotope distribution is its distribution of isotope abundances prior to fragmentation. After fragmentation, however, the isotope distribution of a particular fragment molecule is called a fragment isotope distribution. When computing theoretical isotope distributions for either type, there are two further scenarios: either the elemental composition of the molecule is known, or it is not. When the elemental composition is known, its theoretical precursor isotope distribution can be computed using methods such as polynomial expansion, multinomial expansion or the fast Fourier transform (FFT) (Brownawell and Filippo, 1982; Yergey, 1983; Rockwood et al., 1995).

However, if a molecule's elemental composition is not known, but is comprised of similar structural units such as amino acids or nucleotides, then its theoretical precursor isotope distribution can be approximated in one of two ways. The most common method is to first approximate the elemental composition using the Averagine model, which represents the elemental composition of an average amino acid weighted by frequency in the human proteome, and then to compute the corresponding theoretical precursor isotope distribution (Senko et al., 1995). A fractional Averagine model was later developed that allowed continuous values for element counts and therefore avoided discontinuities due to element rounding, but was also more computationally intensive (Renard et al., 2008). The second approximation method utilizes the relationship between mass and isotope ratios. In the case of peptides, approximate precursor isotope distributions are reconstructed by evaluating polynomial functions that are fit to the isotope ratios and masses of peptides generated *in silico* (Valkenborg et al., 2008; Ghavidel et al., 2014). However, because of its unique isotope distribution, the number of sulfur atoms within a peptide creates a divergence in these patterns, particularly for shorter peptides. If the number of sulfurs can be determined, then a more accurate prediction of isotope ratios can be achieved by utilizing models that are fit specifically to peptides with the same sulfur count.

The second type of isotope distribution, fragment isotope distributions, arise from more than just the elemental composition of the molecule. During the isolation and fragmentation of an individual precursor isotopic peak, each precursor in the population has the same number of neutrons, but the locations of the extra neutrons vary. Consequently, the isotope distribution of a fragment depends on the stochastic arrangement of neutrons within the precursor. The isotope distribution of a specific fragment is governed by the probabilities of extra neutrons residing in the given fragment versus its complementary fragment. Isolating multiple precursor isotopic peaks adds further complexity, as the resultant fragment isotope distributions are linear combinations of the fragment isotope distributions stemming from individual precursor peaks. Conveniently, isolation of the complete isotopic distribution creates fragments whose distributions are equivalent to the fragment's natural isotope distribution. For the case where the elemental composition of the precursor and fragment are known, and only a single precursor isotopic peak was fragmented, software has been developed to calculate the theoretical fragment isotope distribution (Ramaley and Herrera, 2008). Unfortunately, utilization of this method has been minimal (Rockwood and Palmblad, 2013). Extending the framework to handle the fragmentation of multiple precursor isotopic peaks, as well as providing a method to approximate fragment isotope distributions will increase its utility and range of applications. Such opportunities exist in the pre-processing of MS2 spectra of unknown elemental compositions, whose methods often rely upon approximate precursor isotope distributions (Carvalho et al., 2009; Xiao et al., 2015; Chen et al., 2006; Horn et al., 2000; Zabrouskov et al., 2005; Liu, 2011; Kou et al., 2014; Yuan et al., 2011; Mechtler, 2016).

Here, I developed methods that approximate fragment isotope distributions when elemental compositions are not known. I re-derived the existing general framework for fragment isotope distributions of individual precursor isotopic peaks and then extended it for subsets of isotopes. Next, I incorporated the Averagine model within this framework in order to support biomolecules of unknown elemental compositions. Given that sulfurs have a large effect on the isotope distributions of small peptides, which are abundant in MS2 spectra, a I developed a sulfur-specific Averagine method and evaluated it on both precursors and fragments. Furthermore, I observed that individual precursor isotope probabilities followed a smooth non-linear pattern and summarized them with splines and used those splines in place of the Averagine model. I evaluated the accuracy and speed of these on *in silico* digested peptides, mass spectrometry experiments utilizing the angiotensin I peptide, and in complex peptide mixtures from HeLa cells lysate.

3.2 Methods

3.2.1 Probabilistic model for fragment isotope distributions

The nominal isotope probabilities of a fragment after the isolation and fragmentation of a single precursor ion were modeled. A molecule's nominal isotopic state is its difference in neutrons relative to its monoisotopic form. In the remaining chapter, nominal isotopes are referred to simply as isotopes. For notation: random variables are represented with capital letters; specific values are represented with lowercase letters; a collection of specific values is denoted by bold lowercase letters; and unions represent logical "ors". Model variables are defined in Table 3.1. The five assumptions underlying our model are as follows:

1. **Mutual exclusivity of isotopic states.** A single molecule cannot simultaneously exist in multiple isotopic states.

2. **Independence of isotopic states between a fragment and its complementary fragment.** The isotopic state of a fragment does not influence the isotopic state of its complementary fragment when not conditioned upon another event.

3. **Non-negativity of isotopic states.** A molecule cannot have fewer neutrons than its monoisotopic form; therefore, the probability of having fewer neutrons than the monoisotope is zero.

4. **Uniform isolation efficiency within the isolation window boundaries.** All ions whose *m/z* values fall within the boundaries of the isolation window are isolated with equal efficiency. Thus, the relative abundance of the permitted isotopes is identical prior to and after isolation.

5. **Zero isolation outside the isolation window boundaries.** Ions whose *m/z* values fall outside the boundaries of the isolation window are not isolated. The isolation window is assumed to perform as a perfect square.

Table 3.1: Variable descriptions for isotope probability model

Symbol	Description
P	Random variable for the nominal isotopic state of precursor with known elemental composition
F	Random variable for the nominal isotopic state of fragment with known elemental composition
C	Random variable for the nominal isotopic state of a complementary fragment, whose elemental composition is that of the precursor minus the fragment
p	Specific value for the precursor's nominal isotopic state
f	Specific value for the fragment's nominal isotopic state
$\boldsymbol{p}$	Subset of precursor isotopes that can be isolated by the isolation window

An equation to compute the probability that a fragment will exist as a specific isotope given that its precursor belonged to one of the permitted isotopes was derived utilizing the assumptions stated above. Applying Bayes' theorem results in equation 3.1, and the mutual exclusivity assumption leads to equation 3.2. A precursor's isotopic state must be the sum of the isotopic states of its fragment and the corresponding complementary fragment. Consequently, if a fragment's isotopic state equals f, then the following two events are identical: 1) the precursor's isotopic state equals p, and 2) the complementary fragment's isotopic state equals $p - f$. These descriptions lead to equation 3.3, which is then simplified to equation 3.4 because conditioning the complementary fragment's isotopic state on the fragment's isotopic state has no effect due to the assumption of independence described above. Finally, in equation 3.5, the denominator is substituted with an equivalent expression that avoids computing the precursor's isotope distribution by reusing the calculations of the numerator. This final equation is composed exclusively of unconditional events whose probabilities can be computed by methods for precursor isotope distributions.

$$\Pr\bigl(F = f \mid \bigcup_{p \in \boldsymbol{p}} P = p\bigr) = \frac{\Pr\bigl(F = f\bigr) \cdot \Pr\bigl(\bigcup_{p \in \boldsymbol{p}} P = p \mid F = f\bigr)}{\Pr\bigl(\bigcup_{p \in \boldsymbol{p}} P = p\bigr)} \tag{3.1}$$

$$= \frac{\Pr\bigl(F = f\bigr) \cdot \sum_{p \in \boldsymbol{p}} \Pr\bigl(P = p \mid F = f\bigr)}{\sum_{p \in \boldsymbol{p}} \Pr\bigl(P = p\bigr)} \tag{3.2}$$

$$= \frac{\Pr\bigl(F = f\bigr) \cdot \sum_{p \in \boldsymbol{p}} \Pr\bigl(C = p - f \mid F = f\bigr)}{\sum_{p \in \boldsymbol{p}} \Pr\bigl(P = p\bigr)} \tag{3.3}$$

$$= \frac{\Pr\bigl(F = f\bigr) \cdot \sum_{p \in \boldsymbol{p}} \Pr\bigl(C = p - f\bigr)}{\sum_{p \in \boldsymbol{p}} \Pr\bigl(P = p\bigr)} \tag{3.4}$$

$$= \frac{\Pr\bigl(F = f\bigr) \cdot \sum_{p \in \boldsymbol{p}} \Pr\bigl(C = p - f\bigr)}{\sum_{f=0}^{\max(\boldsymbol{p})} \left(\Pr\bigl(F = f\bigr) \cdot \sum_{p \in \boldsymbol{p}} \Pr\bigl(C = p - f\bigr)\right)} \tag{3.5}$$

3.2.2 Sulfur-specific Averagine model

A modified Averagine method can be used to approximate the elemental composition of a peptide when its composition of sulfur atoms is known. The average mass of the sulfurs is subtracted from the observed average mass of the molecule. The elemental composition of the remaining mass is then approximated using a modified Averagine model that does not contain sulfurs: $C_{4.9384}H_{7.7583}O_{1.4773}N_{1.3577}$. Following the standard Averagine method, the remaining mass is divided by the average mass of the modified Averagine model, and the result is multiplied by the model's elemental composition. Finally, the element counts are rounded, and hydrogens are added or subtracted to compensate for any error in nominal mass.

3.2.3 Averagine model incorporation

Both the Averagine and sulfur-specific Averagine models were incorporated into the general framework in order to approximate fragment isotope distributions. Our evaluations were performed with *a priori* AA sequence information; thus, average masses and sulfur counts for each fragment and its complementary fragment were calculated based on AA sequence information rather than from observed peak intensities and *m/z* values. For each average mass, both the Averagine and sulfur-specific Averagine methods were used to approximate its elemental composition. Precursor isotope distributions were computed for the approximate elemental compositions up to the largest isolated precursor isotope using the fast Fourier transform method implemented in OpenMS (Röst et al., 2016). The approximate isotope distributions were then used as substitutes for their exact counterparts in equation 3.5.

3.2.4 Approximation using splines

Splines were fitted to theoretical isotope probabilities of *in silico*-generated AA sequences. Each isotope had a training data set consisting of average masses and corresponding isotope probabilities for simulated sequences. The simulated sequences were varied in length from one to 1000 amino acids with a maximum mass of 100 kDa. For each sequence length, 1000 sequences were generated by choosing a random amino acid for each position. To mimic the distribution of amino acid combinations observed in nature, the amino acids were sampled from a probability distribution corresponding to the observed frequencies for the 20 most common amino acids found in the human canonical SwissProt database (downloaded 11/28/16) (Uniprot Consortium, 2018). After sequence generation, theoretical isotope distributions were computed up to the first 100 isotopes. Sulfur-specific training sets containing zero to five sulfurs were generated separately for each case. The construction of the sulfur-containing sequences was identical to the procedure described above, except that once the sequence contained the desired number of sulfurs (from methionine or cysteine amino acids), the rest of the sequence was derived from the remaining 18 amino acids. Random sequences were chosen over *in silico* proteome digests in order to minimize gaps and biases in mass coverage.

Individual cubic splines were fit for each isotope (M to M+100) and sulfur count (0-5, all) combinations using MATLAB's Curve Fitting Toolbox (version R2016a). Initially, knots were uniformly spaced along the mass axis at 2 kDa intervals with the first and last knots repeated four times to force the splines to have two continuous derivatives. Next, cubic B-splines were fit using a least-squares approximant and the initial knot sequence. Knot selections were then adjusted to uniformly distribute the errors of the least-squares approximant, and the B-splines were re-fit. The final B-splines were converted to piece-wise polynomial format and written to an XML file.

3.2.5 Chemicals and standards

Angiotensin I was purchased from Sigma (St. Louis, MO; catalog number A9650) and reconstituted to a final concentration of $1\,\mathrm{pmol}/\mu\mathrm{l}$ in a solution of 50:50 (methanol:water) containing 0.1% acetic acid. Pierce$^{\mathrm{TM}}$ HeLa Protein Digest Standard was purchased from Thermo Fisher Scientific (Waltham, MA; catalog number 88328) and diluted to a final concentration of $200\,\mathrm{ng}/\mu\mathrm{l}$ in a solution of 98:2 (water/acetonitrile) containing 0.1% formic acid.

3.2.6 Mass spectrometry

Angiotensin I peptide was analyzed by direct infusion into an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). The syringe pump was operated at a flow rate of $3\,\mu\mathrm{l}/\mathrm{min}$. The Heated Electrospray Ionization (HESI) ion source voltage was $3.5\,\mathrm{kV}$; the RF lens was set to 30%; and the ion transfer tube was maintained at $300\,^{\circ}\mathrm{C}$. MS2 scans were acquired by the Orbitrap analyzer at 15k resolution using a 5e4 AGC target, $30\,\mathrm{ms}$ max injection time, and collision-induced dissociation (CID) at 30% collision energy. The MS2 scans were performed in a targeted manner using an inclusion list to isolate and fragment varying isotopes of the precursor in the +3 charge state. The inclusion list consisted of isolation windows with widths ranging from 0.4 to 2.4 *m/z* at 0.1 *m/z* intervals and isolation window offsets ranging from -1.2 to 1.2 *m/z* at 0.05 *m/z* intervals relative to the +3 precursor monoisotope (*m/z* = 432.9).

Trypsinized peptides ($200\,\mathrm{ng}$) from HeLa cell lysate were separated via reverse-phase chromatography using a nanoACQUITY UPLC system (Waters Corporation; Milford, MA) and analyzed by an Orbitrap Fusion Lumos. Peptides were trapped on a $2\,\mathrm{cm}$ column (Pepmap 100, $3\,\mu\mathrm{m}$ particle size, $100\,\mathrm{\AA}$ pore size) and separated in a $25\,\mathrm{cm}$ EASY-spray analytical column ($75\,\mu\mathrm{mol}$ ID, $2.0\,\mu\mathrm{m}$ C18 particle size, $100\,\mathrm{\AA}$ pore size) at $300\,\mathrm{nl}/\mathrm{min}$ and $35\,^{\circ}\mathrm{C}$ using a $180\,\mathrm{min}$ gradient from 2-25% buffer B (0.1% formic acid in acetonitrile). The EASY-spray ion source voltage was set to $1.95\,\mathrm{kV}$; the RF lens was set to 30%; and the transfer tube was maintained at $275\,^{\circ}\mathrm{C}$. The mass spectrometer was operated in data-dependent acquisition mode with a $3\,\mathrm{s}$ cycle time (TopSpeed). Full MS scans were obtained at 60k resolution by the Orbitrap mass analyzer, with a 400-1550 *m/z* scan range, 4e5 AGC target, and $50\,\mathrm{ms}$ maximum injection time. For MS2 selection, peptide monoisotopic peak determination was enabled, and dynamic exclusion was set to $60\,\mathrm{s}$ with a 10 ppm mass tolerance. Further MS2 selection criteria included a 5e4 intensity threshold and inclusion of charges 2-7. Isolation was performed by a quadrupole using isolation windows of 1.6 *m/z* width and centered on the monoisotopic peak. MS2 scans were obtained by the Orbitrap mass analyzer at 15k resolution using a 5e4 AGC target, $50\,\mathrm{ms}$ maximum injection time, and 25% CID collision energy.

3.2.7 Data analysis

Angiotensin I data were processed via custom programs utilizing the OpenMS library. Prior to analysis, raw data were converted twice into mzML format using ProteoWizard's MSConvert (Chambers et al., 2012). In one conversion, the profile data were centroided; in the second conversion, the profile data were preserved for plotting purposes. Scans that isolated contiguous subsets of the first four precursor isotopes were identified based on isolation window parameters, and isotopic peaks from the two most intense fragment ions (B_5^+ DRVYI and B_9^{++} DRVYIHPFH) were extracted. The extraction process consisted of searching the centroided data for the monoisotopic fragment peak up to the largest isolated isotope using a 10 ppm mass tolerance. Observed isotopic peak intensities for each fragment within each scan were normalized to a sum of one. Theoretical and approximate isotope distributions were computed using the OpenMS implementations of the previously described methods. When calculating precursor isotope distributions, the first seven isotope probabilities were computed, and isotopes were removed if both of the following were true: 1) their abundance was less than 10% of the most abundant isotope, and 2) the isotope was greater than the maximum isolated isotope. After filtering, the isotope probabilities were re-normalized such that they sum to one. To evaluate goodness of fit between observed and computed distributions, chi-squared (χ^2) statistics were calculated using the computed distributions as the expected values.

HeLa cell lysate data were analyzed by database search within an OpenMS workflow, followed by a custom program to evaluate the fits of approximated isotope distributions. After conversion to mzML, a database search was performed using MSGF+ (Kim and Pevzner, 2014) against the human canonical SwissProt database (downloaded 11/28/16) appended with reversed decoy sequences. Search parameters included a static Carbamidomethyl (C) modification, variable Oxidation (M) modification, maximum of two modifications, 10 ppm precursor mass tolerance, fully tryptic digestion, 6-40 amino acid length, charge states of 2-4, no isotope error, and the Q-Exactive instrument parameter. Peptide-spectrum-matches (PSMs) were scored using Percolator (version 3.0) (The et al., 2016) and filtered for a 1% false discovery rate (FDR).

The custom program then extracted MS2 spectra for each PSM and calculated the *m/z* for each B and Y ion of charge +1 up to one less than the precursor charge. Using the same procedure as described above for angiotensin I peptide, the fragment isotopes were found in the spectrum; their theoretical and approximate fragment isotope distributions were computed; and chi-squared statistics were calculated. The source code used to generate all figures in this manuscript is available at www.github.com/MajorLab/Fragment-Isotope-Distribution-Paper/.

3.3 Results and Discussion

3.3.1 Derivation of a model for fragment isotope distributions

When performing an MS2 scan in mass spectrometry, it is common that only part of an isotopic distribution is isolated and fragmented, which results in fragments with complex isotope distributions. To predict these distributions, I modeled the probabilities of a fragment's isotopic state given its elemental composition, the elemental composition of its precursor, and the boundaries of the employed isolation window. This model requires explicitly stating the five assumptions employed, of which the first three are self-evident: 1) a single molecule cannot simultaneously be in multiple isotopic states; 2) the molecule cannot have fewer neutrons than its monoisotopic state; and 3) when no other information is available, the number of neutrons in a fragment and its complementary fragment are independent of each other. However, the other two assumptions are not entirely accurate: 4) there is uniform isolation efficiency within the isolation window, and 5) there is zero isolation outside of the isolation window's boundaries. Current mass spectrometry instrumentation does not achieve the perfect box shape for an isolation window; isolation efficiency often decreases near the edges of the window and is non-zero just outside of its boundaries (Scheltema et al., 2014; Lawson et al., 2017). Therefore, our model reflects an idealized scenario. For the fragmentation of a single precursor isotopic peak, the model is equivalent to the framework by Rockwood. The method described here for determining the theoretical fragment isotope distributions was added to the OpenMS library along with unit tests to ensure correctness.

3.3.2 Averagine model incorporation

Typically, the identities and elemental compositions of the molecules in each MS2 scan are unknown. For such cases, the model for calculating fragment isotopic distributions cannot be used directly, and an approximation approach must be used instead. The approach described here uses the Averagine method to approximate the elemental composition of a fragment and its complementary fragment from their average masses. An alternative approach is to first approximate the compositions of the precursor and fragment and then subtract the composition of the fragment from the precursor to determine the composition of the complementary fragment. This is the more computationally efficient approach when approximating multiple fragment isotopic distributions for the same precursor. However, this approach will often lead to negative hydrogen counts for the complementary fragment due to rounding and hydrogen compensation performed by the Averagine method. For example, for a mass of 1340 Da the Averagine method approximates an elemental composition of $C_{60}H_{76}O_{18}N_{16}S_1$, and $C_{54}H_{106}O_{16}N_{15}S_0$ for a fragment mass of 1220 Da. Subtracting the two compositions leads to an approximate elemental composition of $C_6H_{-30}O_2N_1S_1$ for the complementary fragment of mass 120 Da, which is not compatible with software that calculates isotope distributions.

The standard Averagine method uses the average mass calculated from observed peaks; however, when only part of an isotopic distribution is isolated, a fragment's isotopic distribution is no longer representative of its average mass. Furthermore, difficulty arises for low-intensity ions where the monoisotopic peak may not be observed due to low abundance. For the evaluations performed in this work, elemental compositions were known, and the correct average masses were used. When average masses are not known, a method based on observed peaks will be necessary and will result in some mass error; however, the effect on approximate isotopic distributions due to inaccuracy of a few Daltons is negligible. The approximation methods for fragment isotope distributions using the Averagine and sulfur-specific Averagine models have also been added to the OpenMS library.

3.3.3 Spline construction

While the Averagine model combined with the FFT has successfully been used to approximate isotope distributions, they have two undesirable properties. The Averagine model has discontinuities due to the rounding of element counts, with the largest effect due to sulfurs as demonstrated by the vertical jumps within the blue lines of Figure 3.1. Additionally, the FFT is often replaced with a pre-computed lookup table at several Dalton intervals when extremely fast computation is necessary. The fractional Averagine method avoids discontinuities, but requires five additional convolutions and is therefore slower to compute. As an alternative, I used splines to model isotope probabilities in a compact and efficient data structure (Figures 3.1-3.5). Although the probabilities follow a consistent pattern, divergence is present due to the distinct

Table 3.2: Goodness of fit statistics for splines

Sulfurs	Isotope	RMSD	Mean deviation	R^2
all	M	0.00870	0.00506	0.99858
all	M+1	0.00882	0.00648	0.99378
all	M+2	0.00759	0.00571	0.98928
all	M+3	0.00619	0.00472	0.99021
all	M+4	0.00504	0.00370	0.99385
0	M	0.00673	0.00398	0.99916
0	M+1	0.00478	0.00329	0.99824
0	M+2	0.00330	0.00248	0.99802
0	M+3	0.00265	0.00191	0.99834
0	M+4	0.00206	0.00142	0.99907
1	M	0.00439	0.00185	0.99931
1	M+1	0.00306	0.00161	0.99922
1	M+2	0.00225	0.00142	0.99935
1	M+3	0.00253	0.00185	0.99820
1	M+4	0.00200	0.00140	0.99902
2	M	0.00389	0.00165	0.99934
2	M+1	0.00274	0.00146	0.99932
2	M+2	0.00206	0.00130	0.99945
2	M+3	0.00243	0.00181	0.99803
2	M+4	0.00194	0.00137	0.99896

isotopic distribution of sulfur-containing peptides (Figure 3.1). To address this, sulfur-specific splines were fitted separately to peptides containing the matching number of sulfurs (Figures 3.1, 3.6-3.9). Both the sulfur-specific and average splines showed excellent goodness of fit with >0.99 R^2 values (Table 3.2), and the best fits were exhibited by the sulfur-specific models. Computing approximate isotope distributions with splines is nearly 20 times faster than the Averagine and FFT method (Figure 3.10). The disadvantage of splines is that the requested mass must be within the mass range to which that spline was fitted. This can be mitigated by training the model to the anticipated range of queries or by defaulting to the Averagine and FFT method when the requested mass is out of range. A sample Java program to parse and compute approximate isotope distributions using the spline models is available at our GitHub repository.

3.3.4 *in silico* evaluation

To determine how well approximate fragment isotope distributions matched to theoretical fragment isotope distributions, I calculated chi-squared statistics between approximate and theoretical distributions for each B and Y fragment from all tryptic peptides in the human proteome (Fig. 3.11). The precursor Averagine method was included as a baseline and to demonstrate that it is inappropriate for fragment isotopes except when most of the precursor isotopic distribution is isolated. As shown in the first row of Figure 3.11, the precursor Averagine approximation improves as more isotopes are isolated. For the fragment methods, the

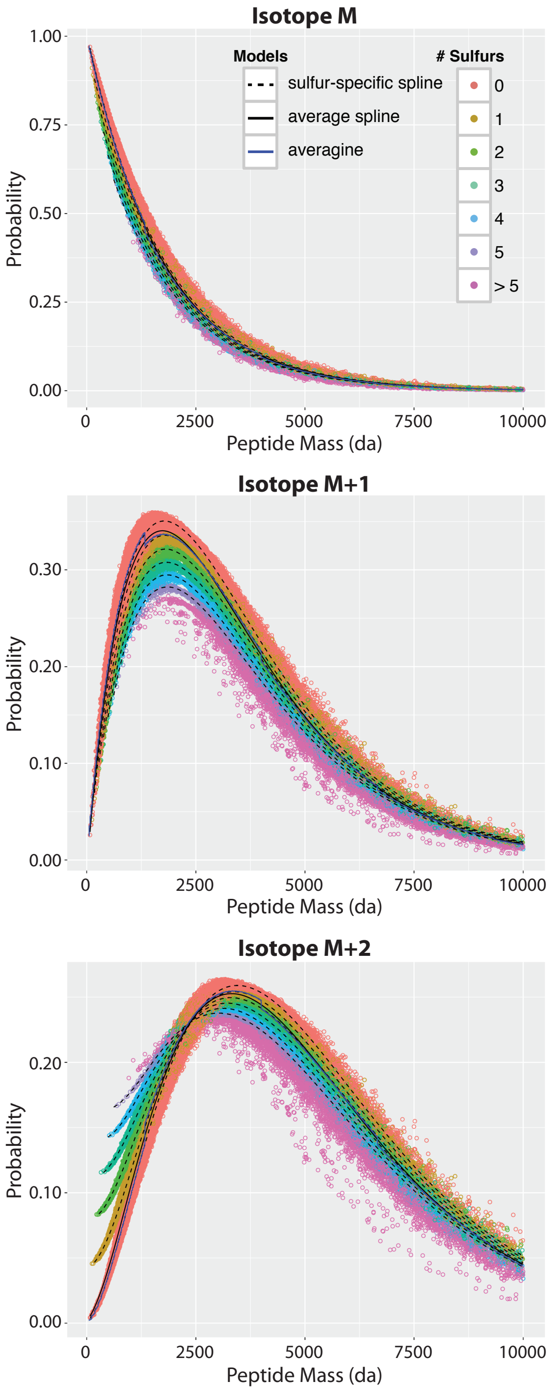


Figure 3.1: Splines were fit to the isotope probabilities of *in silico* generated tryptic peptides. Theoretical precursor isotope probabilities (circles) of human tryptic peptides were overlaid with predictions by the Averagine model, average splines, and sulfur-specific (0-5 sulfurs) splines.

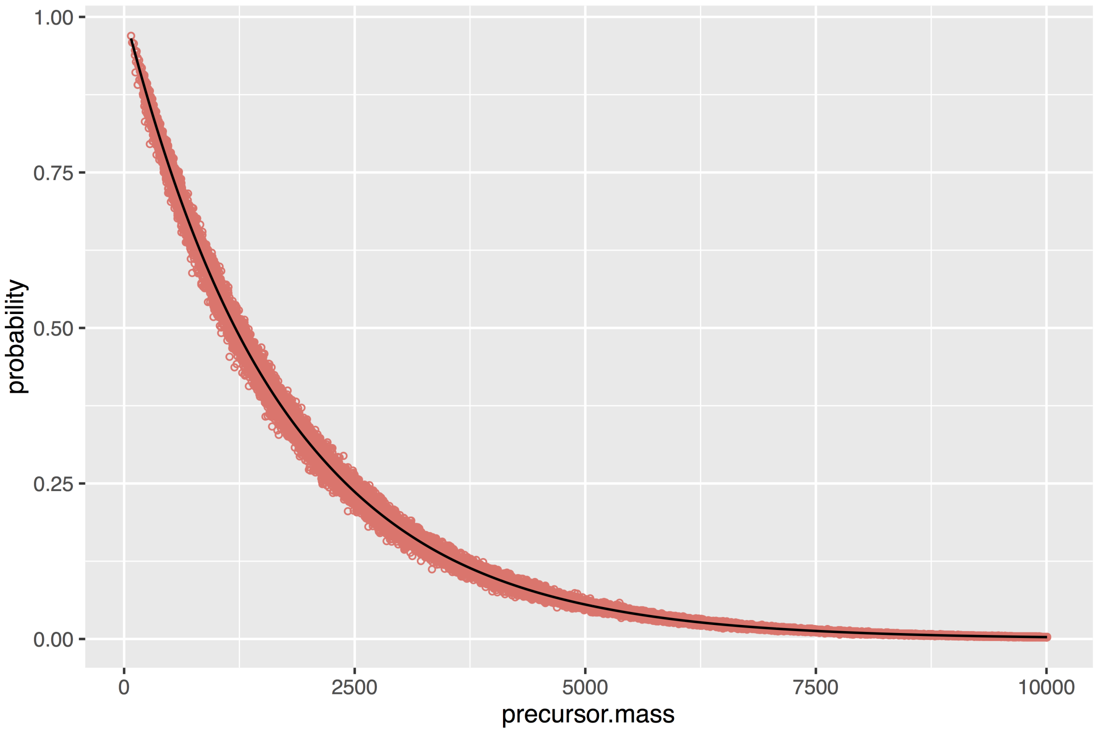


Figure 3.2: Spline (black line) fitted to the probabilities of the monoisotope (M) for simulated peptides.

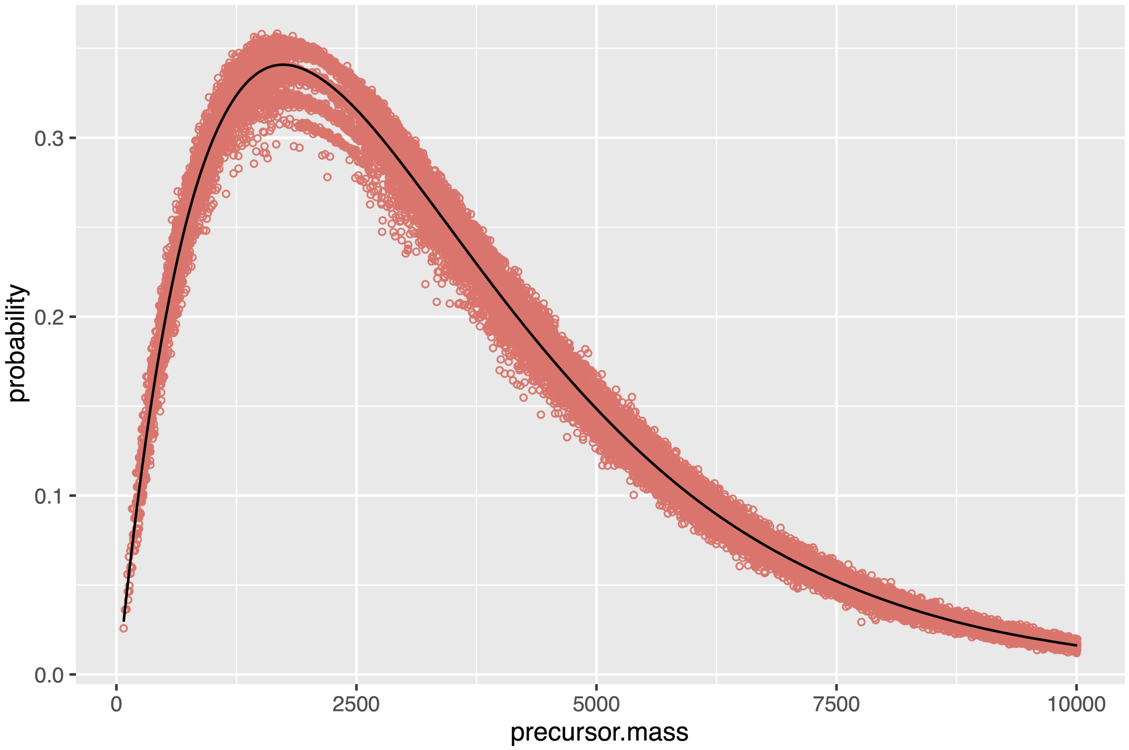


Figure 3.3: Spline (black line) fitted to the probabilities of isotope M+1 for simulated peptides.

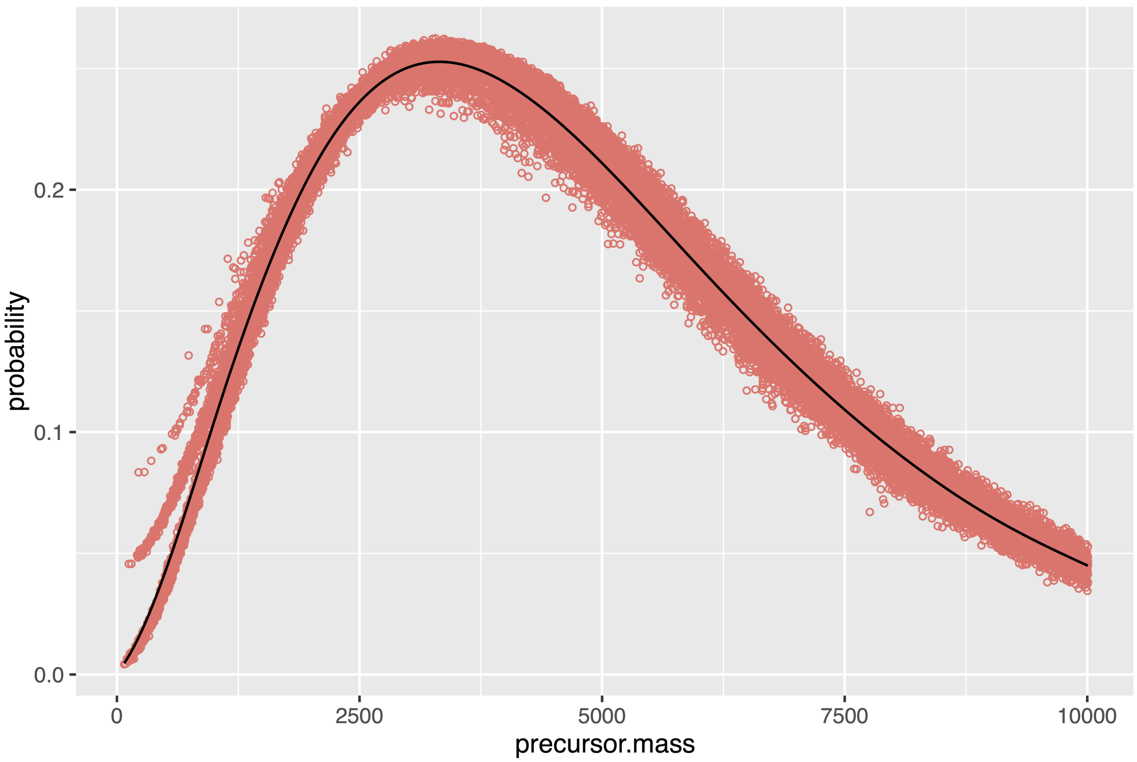


Figure 3.4: Spline (black line) fitted to the probabilities of isotope M+2 for simulated peptides.

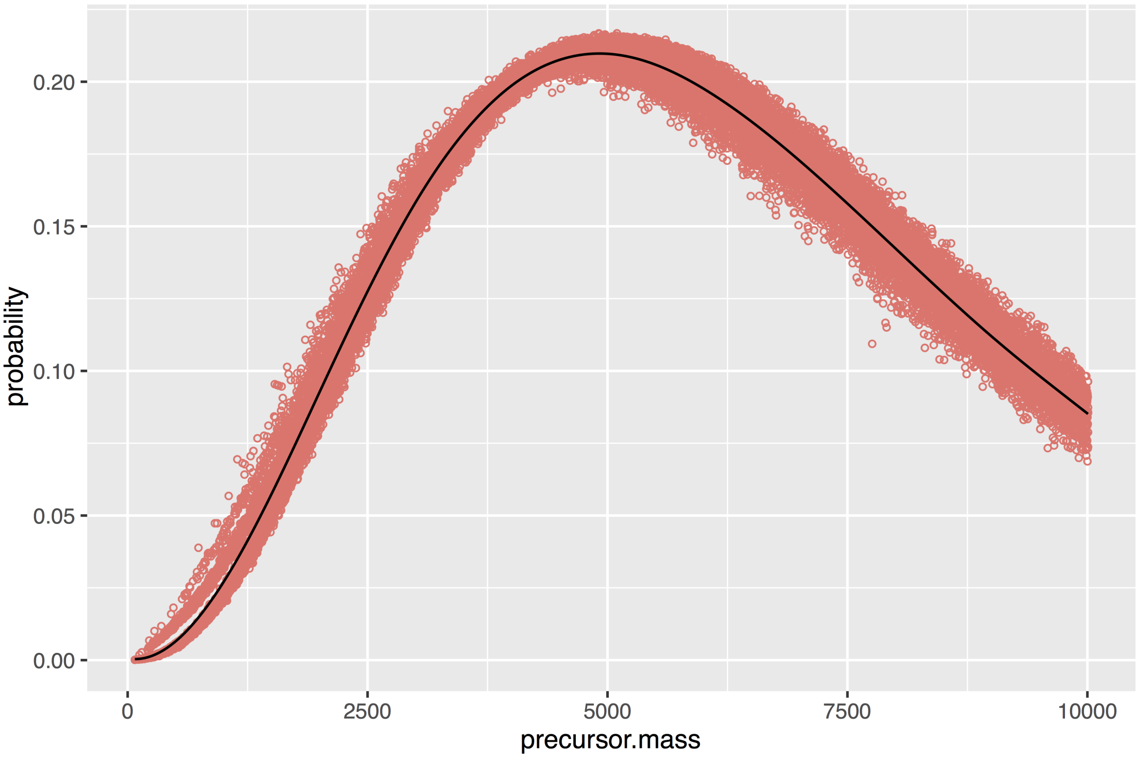


Figure 3.5: Spline (black line) fitted to the probabilities of isotope M+3 for simulated peptides.

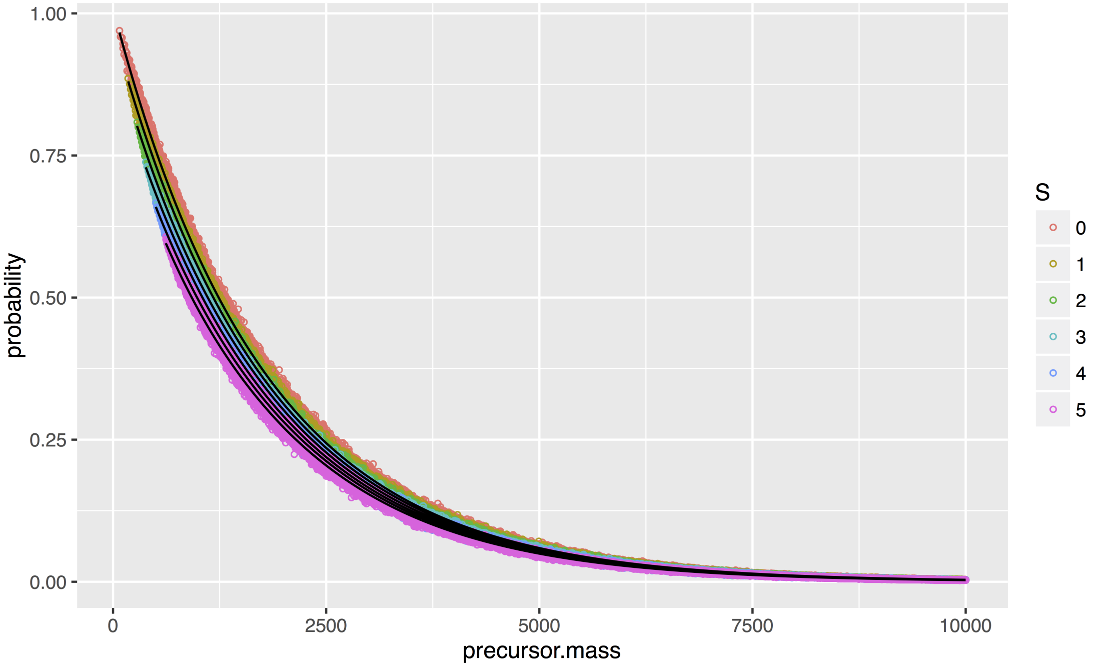


Figure 3.6: Splines (black lines) fitted to the probabilities of the monoisotope (M) for simulated peptides with specific numbers of sulfurs.

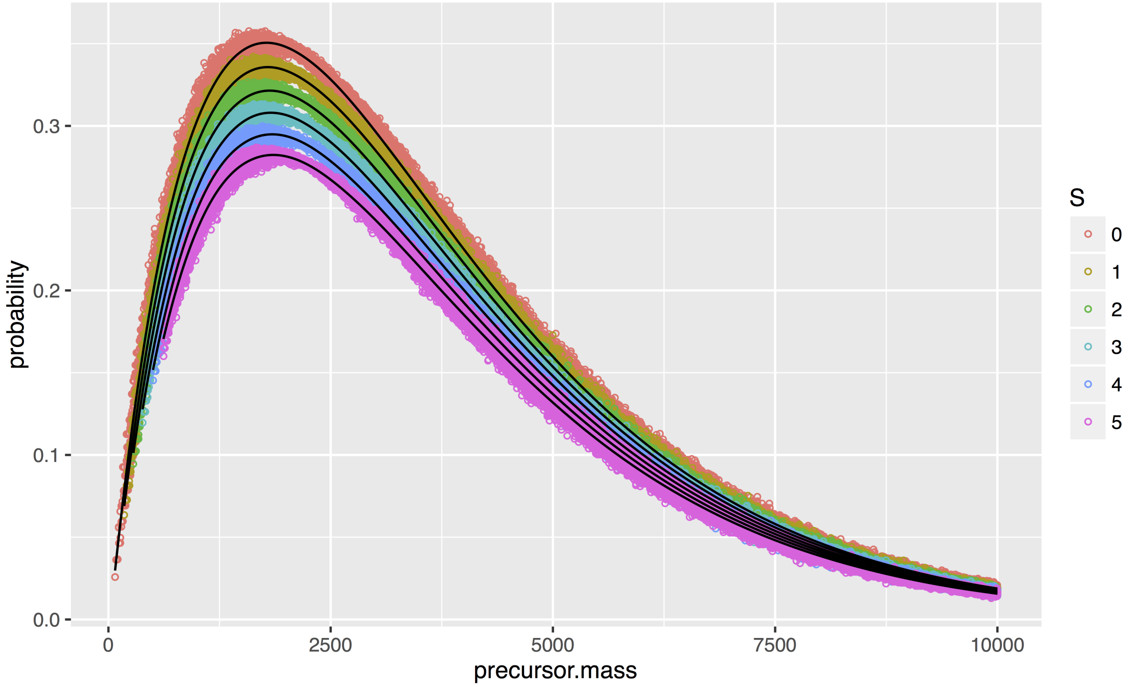


Figure 3.7: Splines (black lines) fitted to the probabilities of isotope M+1 for simulated peptides with specific numbers of sulfurs.

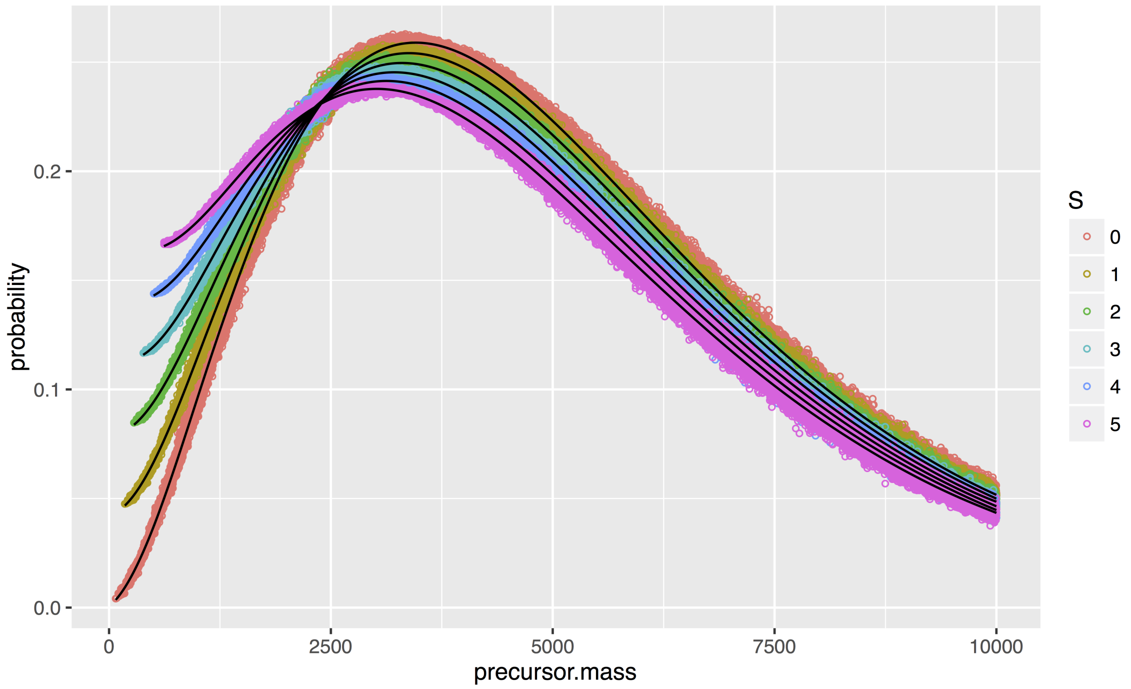


Figure 3.8: Splines (black lines) fitted to the probabilities of isotope M+2 for simulated peptides with specific numbers of sulfurs.

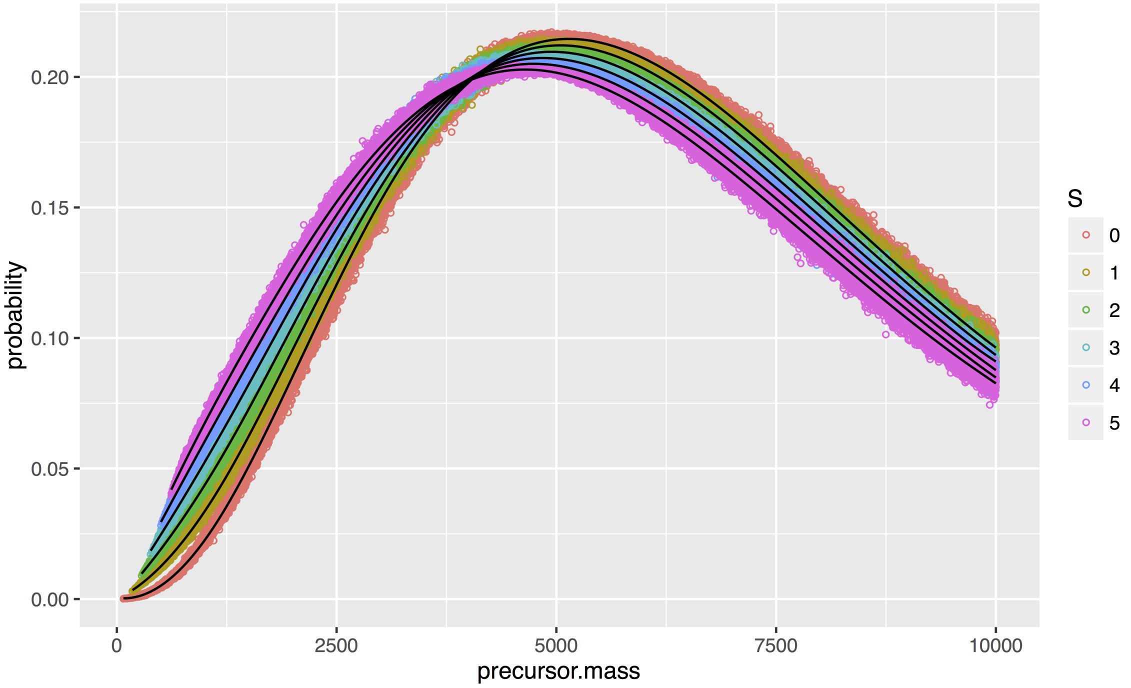


Figure 3.9: Splines (black lines) fitted to the probabilities of isotope M+3 for simulated peptides with specific numbers of sulfurs.

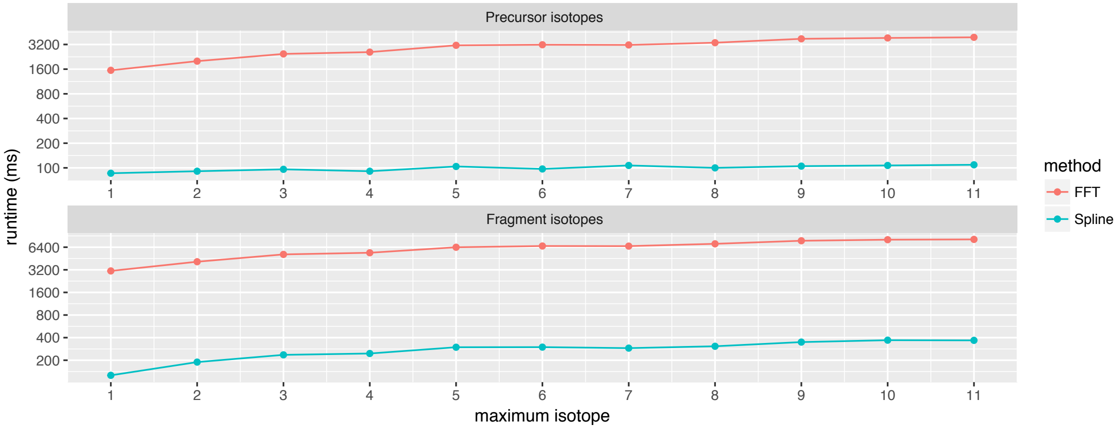


Figure 3.10: Runtime comparison between splines and the Averagine followed by fast-Fourier transform (FFT) method for precursor and fragment isotope distributions. For the precursor comparisons, 10,000 masses were randomly sampled between 400-9500 Da and their approximate isotopic distributions were calculated up to the designated number of isotopes (x-axis). For the fragment comparisons, 10,000 masses were randomly sampled in the same manner, and then the fragment isotope distribution was approximated with the fragment mass being equal to the i^{th} mass and the precursor mass equal to the $i + (i + 1)$ sampled masses.

sulfur-specific Averagine and sulfur-specific splines were the best matches. The sulfur-specific splines were slightly better, having a 10% smaller median χ^2 score and 7% smaller mean (Table 3.3). The fragment Averagine and splines were nearly identical to the sulfur-specific methods when isotopes less than M+2 were isolated. Interestingly, the fragment Averagine method has a 37% smaller median χ^2 score than the splines, but it has a 23% larger mean. The fragment Averagine method has a better best case because it can sometimes approximate a peptide's exact or near exact elemental composition, but in rare situations the compositions are very inaccurate and negatively skew the mean. Overall, the sulfur-specific methods are the best matches to theoretical fragment isotope distributions, but the sulfur-specific methods require that the number of sulfur atoms be known. Furthermore, the fragment Averagine method is a better match than the splines in most cases.

3.3.5 Angiotensin I evaluation

To experimentally validate the theoretical calculations and approximation methods, I directly infused angiotensin I peptide into the mass spectrometer and isolated and fragmented different subsets of precursor isotopes (Figure 3.12). The two most intense fragment ions, B_5^+ and B_9^{++}, displayed minor deviation from the theoretical distributions at least partially due to sample sizes and non-uniform isolation efficiency within and beyond the isolation window boundaries. Evidence for isolation outside of the isolation window are

Table 3.3: χ^2 statistics between theoretical and approximate fragment isotopic distributions

Isotopes	Method	mean	min	Q1	median	Q3	max
M+1	Precursor Averagine	0.31166	1.82e-17	0.02362	0.13105	0.44372	4.58303
M+1	Fragment Averagine	0.00182	0	0.00013	0.00066	0.00214	0.06662
M+1	Splines	0.00186	0	0.00013	0.00065	0.00212	0.08643
M+1	Sulfur-specific Averagine	0.00181	0	0.00014	0.00067	0.00214	0.07448
M+1	Sulfur-specific splines	0.00184	0	0.00013	0.00063	0.00207	0.08605
M+1—M+2	Precursor Averagine	0.31318	1.52e-10	0.06149	0.15200	0.40871	10.4607
M+1—M+2	Fragment Averagine	0.01063	0	0.00042	0.00150	0.00480	2.25174
M+1—M+2	Splines	0.00712	7.76e-11	0.00070	0.00172	0.00427	1.1017
M+1—M+2	Sulfur-specific Averagine	0.00179	0	0.00018	0.00071	0.00210	0.05280
M+1—M+2	Sulfur-specific splines	0.00162	7.05e-12	0.00016	0.00062	0.00184	0.06577
M+1—M+3	Precursor Averagine	0.32850	3.10e-05	0.06400	0.15873	0.42464	12.6815
M+1—M+3	FragmentAveragine	0.01493	0	0.00055	0.00200	0.00636	2.37977
M+1—M+3	Splines	0.00990	9.16e-11	0.00110	0.00240	0.00561	1.1796
M+1—M+3	Sulfur-specific Averagine	0.00183	0	0.00020	0.00073	0.00213	0.05315
M+1—M+3	Sulfur-specific splines	0.00164	4.64e-12	0.00018	0.00064	0.00185	0.06542
M+1—M+4	Precursor Averagine	0.33619	3.42e-05	0.05785	0.16062	0.43790	13.5404
M+1—M+4	Fragment Averagine	0.01790	0	0.00063	0.00236	0.00770	2.52145
M+1—M+4	Splines	0.01166	4.34e-11	0.00147	0.00294	0.00657	1.39113
M+1—M+4	Sulfur-specific Averagine	0.00187	0	0.00022	0.00075	0.00217	0.05393
M+1—M+4	Sulfur-specific splines	0.00167	4.14e-12	0.00020	0.00068	0.00191	0.06537
M+2	Precursor Averagine	1.69740	6.72e-10	0.11598	0.58786	1.91653	120.986
M+2	Fragment Averagine	0.02561	0	0.00081	0.00290	0.01039	3.8416
M+2	Splines	0.01979	4.75e-11	0.00168	0.00469	0.01349	2.07881
M+2	Sulfur-specific Averagine	0.00278	0	0.00041	0.00129	0.00326	0.18362
M+2	Sulfur-specific splines	0.00268	2.80e-11	0.00040	0.00121	0.00311	0.06401
M+2—M+3	Precursor Averagine	1.73207	5.65e-05	0.17054	0.63613	1.92412	121.38
M+2—M+3	Fragment Averagine	0.02852	0	0.00097	0.00340	0.01256	3.71784
M+2—M+3	Splines	0.02146	7.14e-10	0.00188	0.00492	0.01382	2.04985
M+2—M+3	Sulfur-specific Averagine	0.00267	0	0.00035	0.00113	0.00296	0.17695
M+2—M+3	Sulfur-specific splines	0.00239	3.54e-12	0.00034	0.00104	0.00271	0.06285
M+2—M+4	Precursor Averagine	1.79561	8.61e-05	0.19664	0.67603	1.99525	126.088
M+2—M+4	Fragment Averagine	0.03161	0	0.00113	0.00396	0.01430	3.85248
M+2—M+4	Splines	0.02326	2.16e-10	0.00224	0.00554	0.01516	2.02176
M+2—M+4	Sulfur-specific Averagine	0.00267	0	0.00035	0.00109	0.00289	0.17877
M+2—M+4	Sulfur-specific splines	0.00237	1.56e-10	0.00034	0.00102	0.00266	0.06291
M+3	Precursor Averagine	6.97127	9.59e-05	0.41309	1.73223	6.30168	352.666
M+3	Fragment Averagine	0.03663	0	0.00142	0.00504	0.02011	3.7248
M+3	Splines	0.03010	2.16e-10	0.00279	0.00718	0.01933	2.32176
M+3	Sulfur-specific Averagine	0.00370	0	0.00051	0.00157	0.00398	0.21951
M+3	Sulfur-specific splines	0.00377	2.00e-10	0.00052	0.00150	0.00390	0.10934
M+3—M+4	Precursor Averagine	7.36454	7.21e-05	0.51901	1.89298	6.66907	478.385
M+3—M+4	Fragment Averagine	0.04067	0	0.00154	0.00559	0.02235	4.1501
M+3—M+4	Splines	0.03195	1.33e-09	0.00292	0.00733	0.02008	2.21615
M+3—M+4	Sulfur-specific Averagine	0.00363	0	0.00045	0.00139	0.00366	0.23376
M+3—M+4	Sulfur-specific splines	0.00345	6.15e-11	0.00045	0.00132	0.00354	0.10041
M+4	Precursor Averagine	37.11730	5.68e-05	1.06872	4.58227	21.35562	10648.4
M+4	Fragment Averagine	0.05331	0	0.00201	0.00765	0.03189	5.98558
M+4	Splines	0.04510	5.45e-10	0.00425	0.01198	0.03447	7.36276
M+4	Sulfur-specific Averagine	0.00508	0	0.00058	0.00179	0.00481	0.38125
M+4	Sulfur-specific splines	0.00547	1.87e-10	0.00060	0.00179	0.00524	0.34323
M—M+1	Precursor Averagine	0.04382	1.25e-14	0.00789	0.02155	0.05513	0.55427
M—M+1	Fragment Averagine	0.00079	0	7.20e-05	0.00035	0.00108	0.01568
M—M+1	Splines	0.00074	9.28e-18	6.69e-05	0.00031	0.00094	0.01766
M—M+1	Sulfur-specific Averagine	0.00078	0	6.81e-05	0.00034	0.00104	0.01595
M—M+1	Sulfur-specific splines	0.00073	6.41e-20	6.45e-05	0.00030	0.00092	0.01680
M—M+2	Precursor Averagine	0.03243	4.77e-09	0.00336	0.01086	0.03463	0.67269
M—M+2	Fragment Averagine	0.00504	0	0.00027	0.00100	0.00286	0.54412
M—M+2	Splines	0.00353	8.78e-11	0.00048	0.00103	0.00246	0.38792
M—M+2	Sulfur-specific Averagine	0.00105	0	0.00012	0.00048	0.00142	0.02462
M—M+2	Sulfur-specific splines	0.00096	1.68e-12	0.00010	0.00041	0.00123	0.02411
M—M+3	Precursor Averagine	0.02222	9.68e-07	0.00180	0.00585	0.02105	0.65202
M—M+3	Fragment Averagine	0.00796	0	0.00036	0.00138	0.00414	0.67376
M—M+3	Splines	0.00547	1.01e-11	0.00078	0.00149	0.00347	0.48207
M—M+3	Sulfur-specific Averagine	0.00118	0	0.00015	0.00055	0.00160	0.03092
M—M+3	Sulfur-specific splines	0.00108	1.61e-12	0.00012	0.00047	0.00138	0.03060
M—M+4	Precursor Averagine	0.016780	5.81e-08	0.00112	0.00387	0.01475	1.07931
M—M+4	Fragment Averagine	0.01006	0	0.00041	0.00166	0.00527	1.12703
M—M+4	Splines	0.00679	2.58e-11	0.00102	0.00184	0.00429	0.75066
M—M+4	Sulfur-specific Averagine	0.00124	0	0.00016	0.00059	0.00169	0.04040
M—M+4	Sulfur-specific splines	0.00115	5.08e-13	0.00014	0.00050	0.00148	0.03919

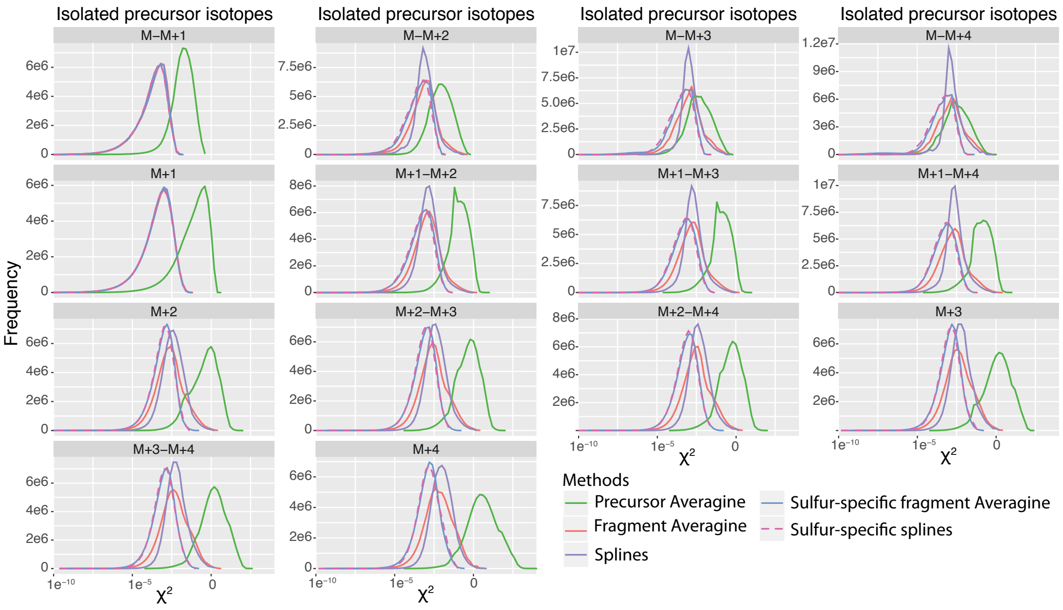


Figure 3.11: The match quality of approximation methods to theoretical isotope distributions was assessed by the chi-squared statistic. The distribution of chi-squared statistics is shown for each approximation method. Every B and Y ion from human tryptic peptides was tested, and each contiguous subset of precursor isotopes between M and M+4 was evaluated separately.

the small M+1 peaks observed for both fragments when only the monoisotopic precursor should have been isolated. Once again, the precursor Averagine approximation was only appropriate when most of the precursor isotopic distribution was isolated. Conversely, all the fragment methods recapitulated the observed isotopic distributions. It is notable that many of the isotopic distributions are visibly distinguishable from each other except when the only difference is the isolation of a low-intensity precursor isotope. This implies that the set of isolated precursor isotopes that created a fragment could be inferred from the fragment's isotopic distribution, and can potentially be used to deconvolve chimeric spectra generated by the co-isolation of multiple precursors with different sets of isotopes.

3.3.6 Whole-cell lysate evaluation

To test the accuracy of these methods on complex samples utilizing typical instrument settings, I performed a shotgun proteomics experiment with whole-cell lysate from HeLa cells using data-dependent acquisition. After a database search to identify peptide-spectrum-matches, fragment isotopic distributions were compared to theoretical and approximate fragment isotope distributions (Figure 3.13 and Table 3.4). The multimodal nature of the chi-squared scores is due to separate, but overlapping, distributions that correspond

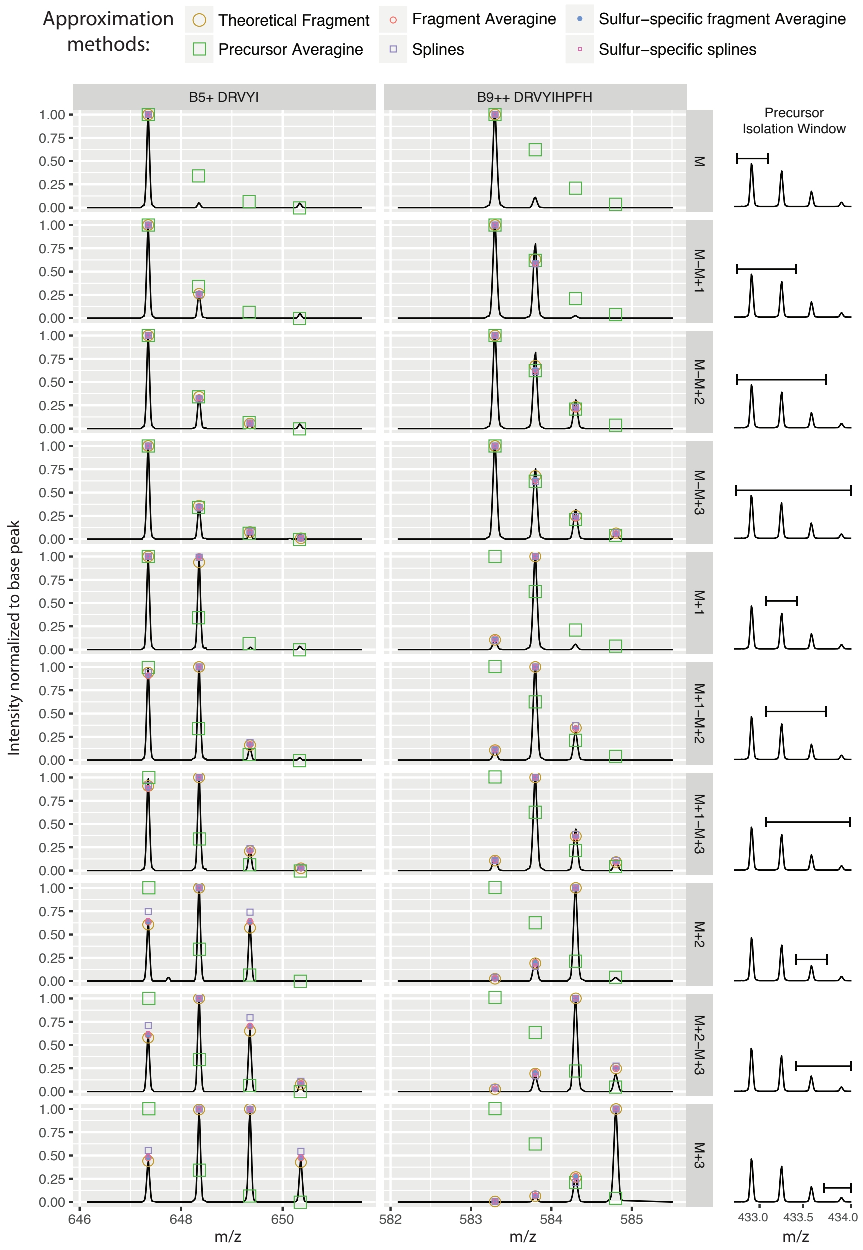


Figure 3.12: MS2 scans were performed on directly infused angiotensin I peptide using various isolation windows. Different sets of precursor isotopes were captured in each scan (right axis labels and diagrams). Profile data is displayed of the two most intense fragments of angiotensin I after CID fragmentation: B_5^+ and B_9^{++}. All signals within 1 *m/z* of a fragments isotopic distribution were extracted from the profile data, and computed distributions were scaled to the extracted base peak. Circles and squares represent the predicted intensities.

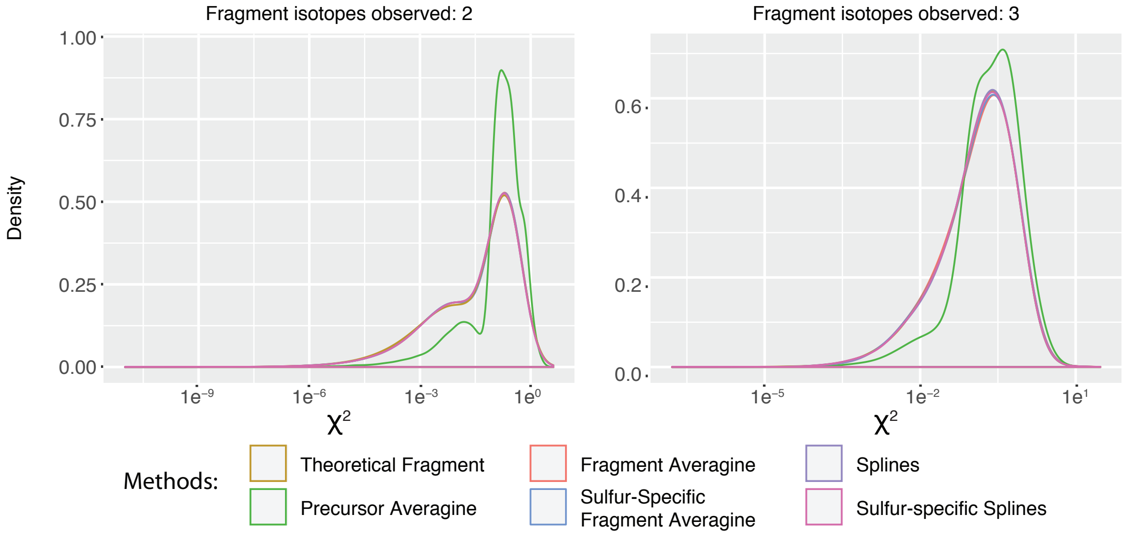


Figure 3.13: Match quality of theoretical and approximate isotope distributions compared to observed fragment isotopic distributions. Distributions of chi-squared statistics between each method and observed fragment isotopic distributions from a shotgun proteomics experiment on trypsin-digested HeLa cell lysate are shown. Isotopic distributions of B and Y fragment ions were only tested if the first two or three isotope peaks were detected. Isolation windows were centered on the monoisotope with a 1.6 *m/z* isolation width.

to the number of missing fragment isotopes. The leftmost distributions have no missing isotopes, while more undetected isotopes result in greater chi-squared scores. The precursor Averagine method had 34% and 74% higher chi-squared scores on average compared to all the other approximation methods. The fragment methods' average chi-squared scores were all within 2% of each other (including the sulfur-specific methods), suggesting that in a high-throughput and complex setting, experimental sources of variance, bias, and interference outweigh the theoretical impact of sulfurs.

3.4 Conclusion

Theoretical fragment isotope distributions can be computed and approximated and accurately match observed fragment isotopic distributions despite the inability of current mass spectrometers to employ perfect box-shaped isolation windows. Taking a probabilistic approach, I re-derived the equations for theoretical fragment isotope distributions and expanded the model to handle the isolation of multiple precursor isotopes. I developed two approximation methods: one using the Averagine model and the other using splines. Although the spline models can be slightly less accurate than the Averagine model when compared to theoretical distributions, in a high-throughput shotgun experiment the splines were equally accurate. Therefore, the spline models are a viable alternative, especially when speed is a top priority. Furthermore, I introduced sulfur-

Table 3.4: Summary of chi-squared statistics from HeLa cell lysate experiment

method	median	mean	sample size	isotope count
Theoretical Fragment	0.0931	0.1463	69027	2
Precursor Averagine	0.1711	0.2586	69027	2
Fragment Averagine	0.0957	0.1486	69027	2
Splines	0.0911	**0.1459**	69027	2
Sulfur-specific Averagine	0.0956	0.1488	69027	2
Sulfur-specific splines	**0.0909**	**0.1459**	69027	2
Theoretical Fragment	0.1679	0.3008	20131	3
Precursor Averagine	0.2527	0.4121	20131	3
Fragment Averagine	0.1710	0.3064	20131	3
Splines	0.1685	**0.3017**	20131	3
Sulfur-specific Averagine	0.1695	0.3064	20131	3
Sulfur-specific splines	**0.1671**	0.3021	20131	3

specific variants for both methods, but neither improved matches to observed fragment isotopic distributions. The worst performing method was the precursor Averagine method, which is only appropriate for calculating fragment isotope distributions when a precursor's entire isotopic distribution is isolated. I contributed to the OpenMS library the methods to calculate theoretical and approximate fragment isotope distributions using the Averagine and sulfur-specific Averagine models so that they can be utilized by future approaches to process MS2 spectra.

CHAPTER 4: DECONVOLUTION OF CHIMERIC SPECTRA

4.1 Introduction

In shotgun proteomics, proteins are enzymatically digested into peptides and separated by liquid chromatography. As peptides elute from the chromatography column, they are ionized and introduced into a mass spectrometer. To determine a peptide's AA sequence, a subset of peptide ions–or precursors–is first isolated and then dissociated into fragment ions during MS2 scans. Using the fragments' measured mass-to-charge (*m/z*) ratios and signal intensities, algorithms can identify the AA sequence that was most likely to have generated the observed mass spectrum. Ideally, isolated precursor populations are homogeneous, as interfering signals from other peptides result in diminished identification rates (Michalski et al., 2011; Houel et al., 2010; Gorshkov et al., 2016; Hebert et al., 2018). Unfortunately, in complex samples, the large number of distinct peptide species means that co-elution of different peptides is unavoidable, and the likelihood of co-eluting peptides of similar *m/z* is high (Michalski et al., 2011). Therefore, it is common for multiple peptides to reside within the bounds of an isolation window used to isolate a precursor ion population. The simultaneous isolation and fragmentation of multiple peptides results in chimeric spectra (Figure 4.1). During data-independent acquisition (DIA), where wide isolation windows are typically utilized, nearly every MS2 spectrum is chimeric by design (Chapman et al., 2013). However, even the narrower windows used during data-dependent acquisition (DDA) methods inadvertently result in $>$50% chimeric spectra (Houel et al., 2010). Thus, chimeric spectra are the norm, rather than the exception, and strategies must be developed to appropriately analyze these spectra.

Historically, peptide sequencing algorithms have been developed assuming that each spectrum is generated by a single peptide. These peptide sequencing algorithms fall into two main categories: database searching and *de novo* sequencing. In a database search, MS2 spectra are scored for how well they match to theoretical spectra from a reference database of protein AA sequences, or to previously observed and identified spectra from a spectral library. While many scoring approaches have been developed, they tend to share three fundamental characteristics: 1) reward matching peaks, 2) penalize missing peaks, and 3) penalize

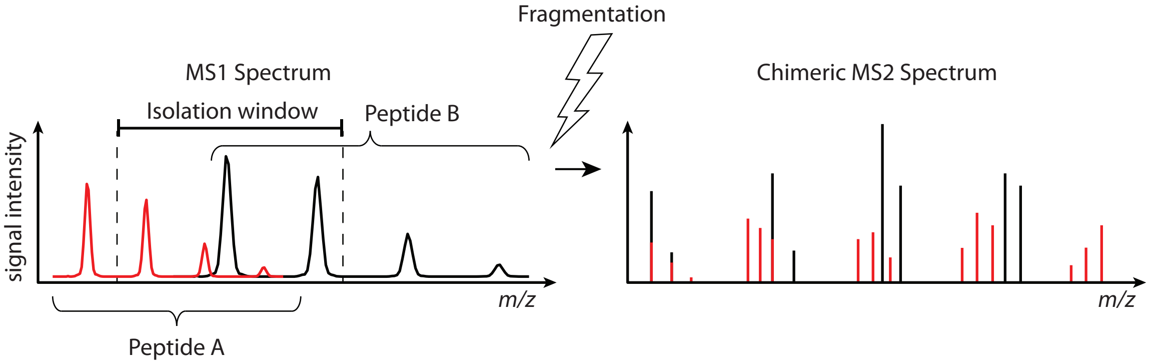


Figure 4.1: Formation of a chimeric MS2 spectrum. Overlapping isotopic distributions from two peptides are co-isolated. For peptide A, isotopes M+1, M+2, and M+3 are isolated. For peptide B, isotopes M and M+1 are isolated. After fragmentation, the resulting MS2 spectrum contains fragments from both peptides.

unexplained peaks (Verheggen et al., 2017). Chimeric spectra are difficult to sequence with a database search because they contain a large number of unexplained peaks for any single peptide match (Houel et al., 2010). Additionally, many post-processing methods reward peptides that match significantly better than all other candidates, but this difference decreases when two peptides have many matching peaks (The et al., 2016). These challenges lead to increased false negatives, yet they have little effect on false positives (Houel et al., 2010). The second approach—*de novo* sequencing—converts MS2 spectra into spectrum graphs in which vertices represent masses and the vertices are connected by edges if their mass difference corresponds to an amino acid (Yan et al., 2015). The highest-scoring path through the graph is then identified. The advantage of *de novo* sequencing is that peptides missing from a reference database can be identified. Their disadvantage is that they require the presence of a more complete fragmentation pattern and fewer interfering peaks. Chimeric spectra hinder these calculations because they have many peaks and therefore introduce more paths in the graph. This leads to many possible AA sequences with similar probabilities without a clear winner. Peptide identification rates on chimeric spectra using *de novo* sequencing are even lower than chimeric identification rates using database searches (Gorshkov et al., 2016). For both database search and *de novo* sequencing, cleaner spectra produce superior sequencing results.

Three main strategies have been developed to identify peptide AA sequences from chimeric spectra. The first strategy is to use specialized scoring methods that maintain high accuracy in the presence of unexplained peaks by decreasing their associated penalty. This is a common approach in DIA analysis. DIA methods also match relative signal intensities of fragment peaks by querying against entries in a spectral library rather than comparing to theoretical spectra (Gillet et al., 2012). Using this strategy, multiple peptides can be identified per spectrum by individually testing each candidate peptide AA sequence. Unfortunately, spectral libraries are experimentally time consuming to create, and they limit future identifications to previously observed peptides. Similarly, for DDA, a robust scoring method was used on chimeric spectra that were submitted multiple times to a database search, where each submission included a different precursor mass corresponding to one of the co-isolated precursors (Zhang et al., 2014). While these scoring methods have been moderately successful, they have lower identification rates than the standard methods that were developed for the one peptide per spectrum model.

The second strategy is to use standard scoring methods and to simultaneously test multiple peptides from a spectral library (Wang et al., 2010) or sequence database (Wang et al., 2011) to find the best combination of peptides that match the chimeric spectrum. Despite aggressive pruning, this strategy is time consuming due to the large number of possible peptide combinations. Alternatively, peptides can be identified one at a time through iterations of peptide identification followed by elimination or attenuation of the matched peaks until no high scoring peptides remain (Zhang et al., 2005; Jurgen et al., 2011; Shteynberg et al., 2015). Nevertheless, subsequent identifications have lower scores because only one peptide will be assigned a fragment peak that is shared by multiple peptides.

In contrast to the strategies described above, the third strategy attempts to deconvolve chimeric spectra into their individual components prior to sequencing. These methods aim to create a distinct spectrum for each co-fragmented precursor peptide. In DIA experiments, elution profiles of fragments that correlate with a precursor's elution profile can be grouped together (Li et al., 2001; Plumb et al., 2006; Geiger et al., 2010; Weisbrod et al., 2012). However, many peptides have similar elution profiles; therefore, chimeric spectra will be created. In another DIA strategy, Egertson and co-authors use a variation of Hadamard multiplexing, and deconvolved spectra are recovered by NNLS (Egertson et al., 2013). Nevertheless, these spectra are still highly chimeric because they correspond to medium-sized isolation windows of 4 m/z. For DDA, deconvolution has been achieved by extracting complementary fragment pairs that together sum to the expected precursor mass (Ledvina et al., 2011; Kryuchkov et al., 2013; Gorshkov et al., 2015). Many of these approaches require novel sequencing algorithms that add to an already bloated collection, while other approaches utilize pre- and post- processing steps independent of the sequencing approach, allowing their integration with any method.

An unexploited feature for deconvolution is the isotope distribution of fragment ions. Isotope distributions arise from the natural abundances of element isotopes that make up a molecule. For a precursor peptide, the theoretical isotope distribution can be computed from its elemental composition and approximated if its AA sequence is not known (Rockwood et al., 1995; Senko et al., 1995). Precursor isotopic distributions are used ubiquitously in the interpretation of MS1 scans. Overlapping distributions are identified, separated, and then used to determine monoisotopic masses and targets for MS2 (Renard et al., 2008; Samuelsson et al., 2004; Slawski et al., 2012). Conversely, use of fragment isotopes has been minimal, and most MS2 isotope processing methods assume that the fragments follow a precursor isotope distribution (Rockwood and Palmblad, 2013; Carvalho et al., 2009; Xiao et al., 2015; Chen et al., 2006; Horn et al., 2000; Zabrouskov et al., 2005; Liu, 2011; Kou et al., 2014; Yuan et al., 2011; Mechtler, 2016). However, fragment isotope distributions depend on its elemental composition, the precursor's elemental composition, and the set of precursor isotopes that were fragmented (Rockwood et al., 2003). If a chimeric spectrum was the result of multiple peptides for which different sets of isotopes were isolated, then their fragments will have distinct isotopic distributions. These isotopic distributions can be identified and attributed back to their precursor.

Here, I propose a novel strategy to deconvolve chimeric spectra by leveraging the dependence of fragment isotope distributions on their isolated precursor isotopes. For each MS2 spectrum, precursors with isotopes with *m/z* that are within the isolation window are extracted from neighboring MS1 spectra. For each peak within the MS2 spectrum, a basis template is created by computing an approximate fragment isotope distribution. Next, I solve a NNLS regression model that is regularized with a sparse group lasso (Simon et al., 2013). Deconvolved spectra are then created by utilizing the basis templates and coefficients corresponding to distinct precursors. Importantly, this is a pre-processing step that is independent of the search algorithm used, and is therefore compatible with all search algorithms.

4.2 Methods

4.2.1 Mass spectrometry

An angiotensin I and neurotensin peptide mixture (Sigma, St. Louis, MO; catalog numbers A9650 and N6383) was analyzed by direct infusion into an Orbitrap Fusion Lumos mass spectrometer. Instrument parameters were identical to section 3.2.6 except for the following: MS2 scans were acquired with higher-energy collision dissociation (HCD) at 30% collision energy. Multiplexed MS2 scans were performed using an inclusion list to isolate and fragment both angiotensin I and neurotensin peptides in the +3 charge state.

The isolation window parameters were adjusted for each scan in order to target all combinations of contiguous isotopes for each peptide.

Trypsinized peptides ($1\,\mu\mathrm{g}$) from HeLa cell lysate from Thermo Fisher Scientific (Waltham, MA; catalog number 88328) were separated and analyzed by an Orbitrap Fusion Lumos using the same parameters as section 3.2.6 except for the following: the separation gradient was $60\,\mathrm{min}$ long; MS1 scans were obtained at 120k resolution with a 300-2000 *m/z* scan range; and MS2 scans were acquired with 30k resolution using a 2e5 AGC target, $54\,\mathrm{ms}$ maximum injection time, and 30% HCD collision energy.

4.2.2 Data sets

Four whole-cell lysate experiments were used for evaluation and training. Two data sets were generated from in-house experiments: a $200\,\mathrm{ng}$ HeLa cell lysate from section 3.2.6 and the $1\,\mu\mathrm{g}$ sample described above. The third data set was generated from a $400\,\mathrm{ng}$ HeLa cell lysate experiment completed by the UNC Proteomics Core Facility using a Q Exactive HF mass spectrometer (Thermo Fisher) with a 1.6 *m/z* isolation window and $90\,\mathrm{min}$ gradient. The fourth data set was downloaded from the PRIDE repository and encompassed data from an SW480 cell lysate (Vizcano et al., 2016). The SW480 lysate experiment was performed on a Q Exactive mass spectrometer using a 2.0 *m/z* isolation window and a $180\,\mathrm{min}$ separation gradient.

4.2.3 Mixture model

A mass spectrum is represented as a sequence of pairs $\left\{\left((m/z)_i,\ y_i\right)\right\}_{i=1}^{n}$, where $(m/z)_i$ is a mass (m) divided by a charge (z), and y_i is the corresponding signal intensity observed at $(m/z)_i$ for $i = 1,\ \ldots,\ n$.

In an ideal scenario, the observed intensities $\boldsymbol{y} = (y_i)_{i=1}^{n}$ of an MS2 mass spectrum could be written as a linear combination of fragment isotope distribution templates. Let $\boldsymbol{A}^*$ be a non-negative matrix of templates and let $\boldsymbol{x}$ be a non-negative vector of coefficients:

$$\boldsymbol{y} = \boldsymbol{A}^*\boldsymbol{x} \tag{4.1}$$

Both $\boldsymbol{A}^*$ and $\boldsymbol{x}$ can be grouped according to their corresponding precursors $p \in \boldsymbol{p}^*$ that were isolated and fragmented to create sub-matrices $\boldsymbol{A}_p^*$ and sub-vectors $\boldsymbol{x}_p$. Each $p \in \boldsymbol{p}^*$ is a unique combination of a peptide's AA sequence and a set of isolated isotopes. Each sub-matrix $\boldsymbol{A}_p^*$ can then be divided into columns

$\boldsymbol{t}^*_{p,1},\ \ldots,\ \boldsymbol{t}^*_{p,T_p}$ representing the theoretical isotopic distribution of each fragment-charge pair that dissociated from precursor p. The entries in template $\boldsymbol{t}^*_{p,j}$ represent the discrete probability of an isotope peak at the corresponding $(m/z)_i$. This gives a more detailed model:

$$\boldsymbol{A}^*\boldsymbol{x} = \sum_{p \in \boldsymbol{p}^*} \boldsymbol{A}^*_p \boldsymbol{x}_p,\ \boldsymbol{A}^*_p = [\boldsymbol{t}^*_{p,1},\ \ldots,\ \boldsymbol{t}^*_{p,T_p}] \tag{4.2}$$

In practice, however, there is measurement noise and error, limited sample sizes, imperfect isolation efficiency, unknown elemental compositions, and uncertainty with respect to the isolated precursors. Therefore, I attempt to approximate the observed spectrum as a sparse linear combination of fuzzy templates. Let $\boldsymbol{A}$ be a non-negative matrix of fuzzy templates. The entries in template $\boldsymbol{t}_{p,j}$ represent a fragment's approximate isotope probability at the corresponding $(m/z)_i$ using the fragment Averagine approach described in section 3.2.3. Each element of $\boldsymbol{p}$ is a unique combination of a precursor mass and a set of isolated isotopes. Accordingly, the approximate version of 4.2 is given as:

$$\boldsymbol{y} \approx \boldsymbol{A}\boldsymbol{x} = \sum_{p \in \boldsymbol{p}} \boldsymbol{A}_p \boldsymbol{x}_p,\ \boldsymbol{A}_p = [\boldsymbol{t}_{p,1},\ \ldots,\ \boldsymbol{t}_{p,T_p}] \tag{4.3}$$

4.2.4 Identification of isolated precursors

To use the fragment Averagine method, each precursor's monoisotopic mass and set of isolated isotopes must be determined. To this end, the Hardklör algorithm from the Crux toolkit (version 3.1) was used to identify precursor isotopic distributions and their monoisotopic masses in all MS1 scans (Hoopmann et al., 2007; Park et al., 2008). Default parameters were used except for the following: Hardklör-algorithm=version2, instrument=orbitrap, and resolution=85000. Precursors identified in the nearest MS1 scans before and after each MS2 scan were examined for any isotopes that were within the bounds of the MS2 scan's isolation window. Only charge states ≥ 2 were considered. If an isotope $\leq$ M+3 was isolated, its corresponding precursor was added to $\boldsymbol{p}$ for this MS2 scan. The single exception occurred when only the monoisotopic peak was isolated, in which case the precursor was excluded. In cases where multiple precursors had the same isolated isotopes and charge, the precursor with the greatest total signal intensity of its isolated isotopes was added, and the other precursors were excluded. Finally, precursors with total signal intensity $\leq 20\%$ of the most intense precursor were removed.

4.2.5 Template selection and construction

Templates are created for every plausible monoisotopic fragment mass, fragment charge, and precursor p. For each $(m/z)_i$ having a positive y_i, fuzzy templates are created using $(m/z)_i$ as the monoisotopic fragment m/z. Templates are created for each precursor $p \in \boldsymbol{p}$ and fragment $z = 1, \ldots, Z_p - 1$, where Z_p is the precursor charge state determined by Hardklör. A template is computed by approximating the fragment isotope distribution with the following inputs: the monoisotopic mass of precursor p, the set of isolated isotopes of p, and the monoisotopic fragment mass $m = (m/z)_i \cdot z$. To handle the cases where the monoisotopic peak was not detected due to low abundance, a template is created for an $(m/z)_i$ with $y_i = 0$. Specifically, if the approximated probability of a monoisotopic fragment is less than the M+1 isotope probability, and $y = 0$ at the m/z of the M-1 isotope, then a template is created at this M-1 isotope m/z.

When constructing $\boldsymbol{A}$ and $\boldsymbol{y}$, all $(m/z)_i$ entries with $y_i = 0$ are initially removed; therefore, $\boldsymbol{y}$ only contains positive values. With each new template, if one of its isotope m/z values is not within 20 ppm of a m/z having a corresponding $y \in \boldsymbol{y}$, then a 0 intensity is inserted at the appropriate location.

4.2.6 Non-negative least squares model

Many more templates are created than are expected to represent a given spectrum. Their creation is due to the uncertainty of their correct placement and identity; thus, extra templates are generated in order to cover all possible cases. With this overabundance of templates, there are an infinite number of solutions for Equation 4.3. In these situations, some form of regularization is used to favor specific types of solutions. To promote sparsity, the *least absolute shrinkage and selection operator* (LASSO) is a common penalty used for regularization. The LASSO shrinks coefficients to exactly zero and promotes overall sparsity, though the ideal solution for the chimeric spectrum deconvolution model is group-level sparsity. In most cases, each isotopic peak should be represented by a single template. To achieve this type of sparsity, the sparse group LASSO penalty along with its two penalty parameters λ_1 and λ_2, were added to the model. The goal is compute $\hat{\boldsymbol{x}}$, the value of $\boldsymbol{x}$ which minimizes the regularized equation:

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 + \lambda_1 \|\boldsymbol{x}\|_1 + \lambda_2 \sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{A}_t \boldsymbol{x}_t\|_2 \tag{4.4}$$

For this convex optimization problem, λ_1 and λ_2 must be properly selected through a grid-based search as described in section 4.2.9. After computing the value of $\hat{\boldsymbol{x}}$, a deconvolved spectrum for each precursor, can be calculated by:

$$\hat{\boldsymbol{y}}_{\boldsymbol{p}} = \boldsymbol{A}_p \hat{\boldsymbol{x}}_p, \; p \in \boldsymbol{p} \tag{4.5}$$

Furthermore, a deconvolved monoisotopic spectra, $\hat{\boldsymbol{y}}_{\boldsymbol{p}}^{\text{mono}}$, can also be calculated. Let $\boldsymbol{A}_p^{\text{mono}}$ be a matrix consisting of monoisotopic templates $[\boldsymbol{t}_{p,1}^{\text{mono}}, \; \dots, \; \boldsymbol{t}_{p,T_p}^{\text{mono}}]$, which have a value of one at their monoisotopic m/z and zeros in the remaining entries:

$$\hat{\boldsymbol{y}}_{\boldsymbol{p}}^{\text{mono}} = \boldsymbol{A}_p^{\text{mono}} \hat{\boldsymbol{x}}_p, \; \boldsymbol{A}_p^{\text{mono}} = [\boldsymbol{t}_{p,1}^{\text{mono}}, \; \dots, \; \boldsymbol{t}_{p,T_p}^{\text{mono}}] \tag{4.6}$$

Monoisotopic spectra are necessary because most peptide sequencing algorithms expect monoisotopic masses. A visual example of the model is provided in Figure 4.2.

4.2.7 Model solution

To find solutions to Equation 4.4, I utilized CVX (Version 2.0), a package for specifying and solving convex programs (Grant and Boyd, 2008, 2014). CVX decides whether to solve the specified problem or the dual problem, chooses a convex optimization solver, and handles conversion to the proper input format. In this case, SDPT3, an algorithm for semidefinite-quadratic-linear programs was selected to solve the dual problem (TüTüNcü et al., 2003).

4.2.8 Evaluation

The number of peptide-spectrum-matches, unique peptides, and proteins following a database search were compared on three versions of each data set: the original spectra, deconvolved spectra, and monoisotopic deconvolved spectra. For each case, a database search was performed using the Crux library (Mcilwain et al., 2014) against the human canonical SwissProt database (UniProt Consortium, 2012) (downloaded 11/28/16) appended with reversed decoy sequences. Search parameters included a static carbamidomethyl (C) modification, variable oxidation (M) modification, fully tryptic digestion, a 3 Da precursor mass tolerance, and a 0.02 mz-bin-width. Peptide-spectrum-matches were scored using Percolator (version 3.0) (The et al.,

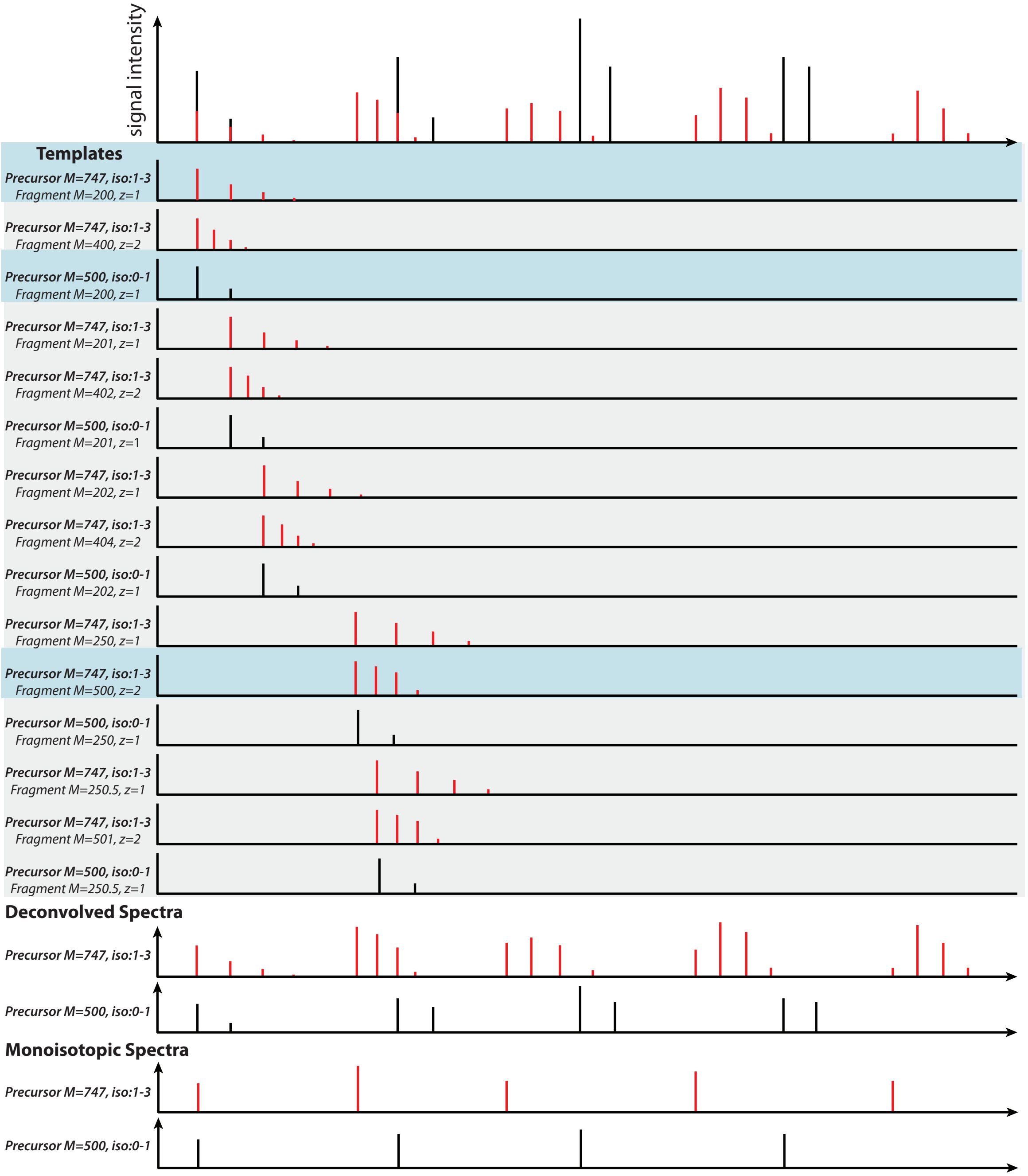


Figure 4.2: Visual schematic of NNLS model on an example chimeric spectrum. Two precursors are co-isolated and co-fragmented as shown in Figure 4.1. After solving the NNLS problem, templates highlighted in blue are assigned positive coefficients. Templates belonging to the same precursor are merged to create deconvolved spectra.

2016) and filtered for a 1% false discovery rate. Protein inference was performed by Fido with default parameters (Serang et al., 2010).

4.2.9 Parameter optimization

To optimize the values for the λ_1 and λ_2 parameters, a grid-based search was performed for all combinations of values in $\{1, 0.1, 0.001, 0.0001\}$ on the multiplexed angiotensin I and neurotensin data. Parameters were excluded if they resulted in deconvolved spectra that were visibly identical or accounted for less than 90% of the original total signal intensity. Then, a finer-grained search was performed on the $200\,\mathrm{ng}$ HeLa experiment. Both λ parameters were set equal to each other. The grid-based search tested λ in $\{0.1, 0.09, \ldots, 0.02, 0.01\}$.

4.3 Results

4.3.1 Angiotensin I and neurotensin

To test and refine the described NNLS model, single-peptide and chimeric spectra were created during a direct infusion experiment of a two-peptide mixture comprised of angiotensin I and neurotensin. To create single-peptide MS2 spectra, separate MS2 scans were performed with an isolation window that isolated and fragmented the monoisotopic peak of each peptide (Figure 4.3). Expected fragments dominate the spectra, however many unmatched peaks are still present. The unmatched peaks are likely due to unexpected fragmentation pathways. Additionally, a few low-intensity peaks match the *m/z* of fragments from the non-targeted peptide because the two peptides can create fragments with similar or identical mass. These spectra represent the best case result of chimeric spectra deconvolution.

Next, chimeric spectra were generated by performing multiplexed MS2 scans. Multiplexed MS2 scans perform multiple isolations to capture a mixed ion population followed by a single round of fragmentation and mass analysis. Here, ions were sequentially accumulated using two isolation windows; one window for each peptide, with half the injection time dedicated to each isolation. Then, the mixed ion population was fragmented and scanned in an Orbitrap mass analyzer. The isolation window parameters were adjusted to capture different subsets of contiguous precursor isotopes for each peptide. All combinations of isotopes were generated up to isotopes M+3. Figure 4.4 contains an example where the first two isotopes of angiotensin I and the second isotopes of neurotensin were co-isolated and co-fragmented. This combination of co-isolated

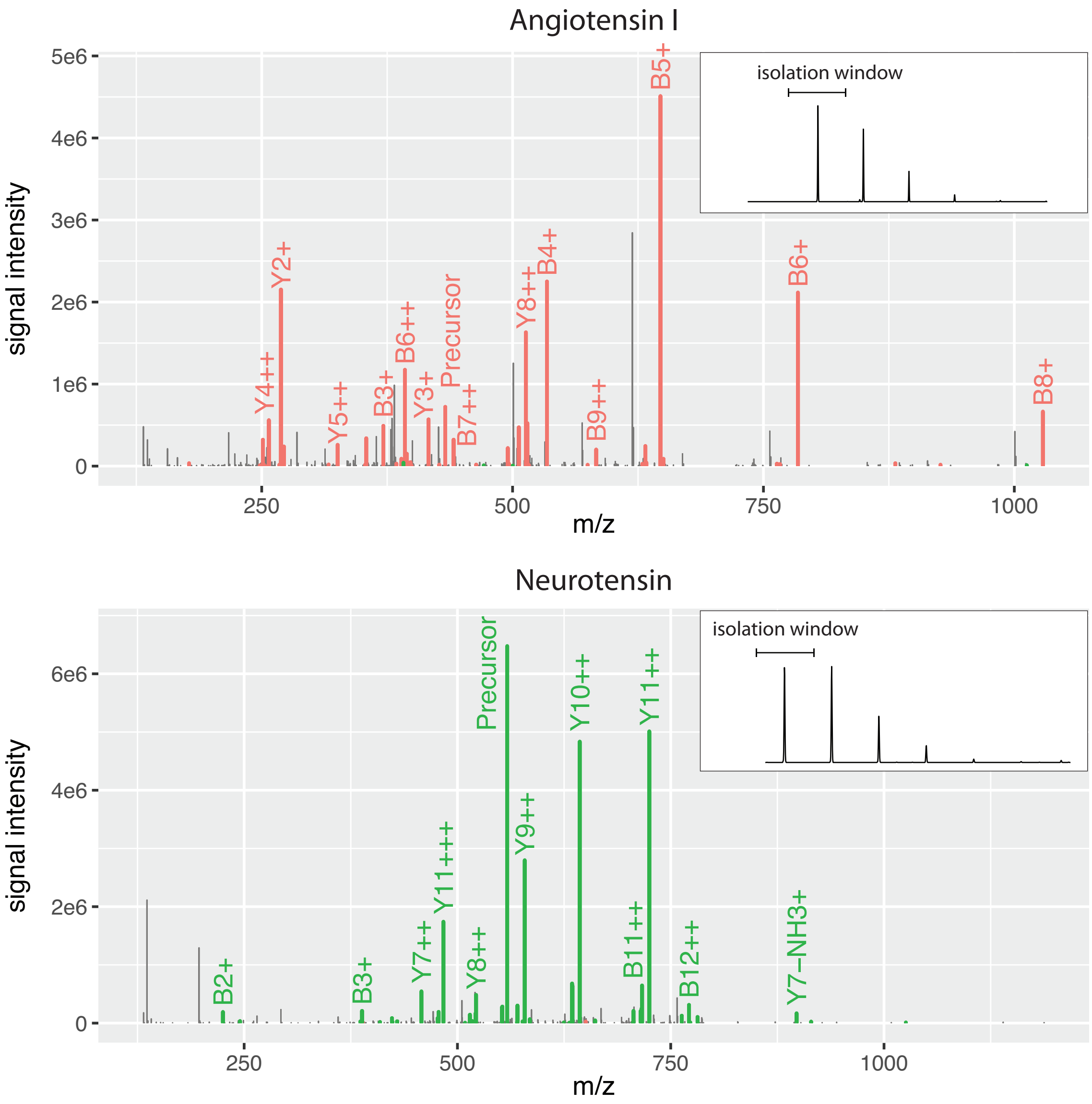


Figure 4.3: MS2 spectra of angiotensin I and neurotensin from a direct infusion experiment. Peaks matching angiotensin I and neurotensin monoisotopic fragments colored red and green, respectfully. Unmatched peaks are colored grey. Inset: Observed MS1 spectrum of the corresponding peptide and the employed isolation window.

isotopes is common in shotgun proteomics experiments that use a 1.6 *m/z* wide isolation window. Although the signal intensities of the isolated isotopes from angiotensin I and neurotensin were nearly equal, most of the matching monoisotopic fragments are from angiotensin I. As expected due to the isolation of the monoisotopic peak of angiotensin I and the lack of the monoisotopic peak from neurotensin, the fragments have significantly different isotopic distributions. Specifically, monoisotopic fragments have high intensity for angiotensin I, and low intensity for neurotensin. The high-intensity unmatched peaks in the chimeric spectrum are M+1 and M+2 isotopes of neurotensin precursors and fragments.

An NNLS model was created and solved for the chimeric spectrum, which resulted in two deconvolved spectra. The deconvolved spectrum corresponding to the angiotensin I peptide is dominated by matches to angiotensin I fragments and the high-intensity unmatched peaks are also found in the single-peptide spectrum from Figure 4.3. However, more neurotensin fragments are observed in the deconvolved spectrum than in the single-peptide angiotensin I spectrum. Similarly, the deconvolved neurotensin spectrum primarily consists of neurotensin fragments, though the most intense peaks are not monoisotopic and would hinder sequencing algorithms. Likewise, 20 low-intensity angiotensin I fragments were incorrectly assigned to the neurotensin spectrum.

Monoisotopic versions of the deconvolved spectra were created using Equation 4.6. These de-isotoped spectra lack the M+1 and M+2 fragment peaks from their corresponding deconvolved spectra. As expected, the angiotensin I spectrum is relatively unchanged as the high-intensity fragments were already monoisotopic. Conversely, the high-intensity unmatched peaks in the neurotensin spectrum were removed, and the high-intensity peaks now match neurotensin fragments. Furthermore, the number of peaks matching angiotensin I fragments decreased to 12. Though these monoisotopic spectra still contain noise and interference from the other peptide, they are extremely similar to their single-peptide versions.

A more difficult example is shown in Figure 4.5. Here, the second two isotopes of angiotensin I and the first three isotopes of neurotensin are isolated. Due to the relatively low abundance of the second two isotopes, the isolated ion populated consisted $<20\%$ of angiotensin I peptides. Additionally, the monoisotopic fragments are rare based on their theoretical fragment isotope distributions. The combination of these two factors resulted in many missing monoisotopic peaks and few matches to angiotensin I fragments in the chimeric spectrum. The deconvolved angiotensin I spectrum is of low quality; it contains few matches to angiotensin I fragments, few peaks overall, and the most intense matching fragment belongs to neurotensin. On the other hand, most peaks from the chimeric spectrum were assigned to the deconvolved neurotensin

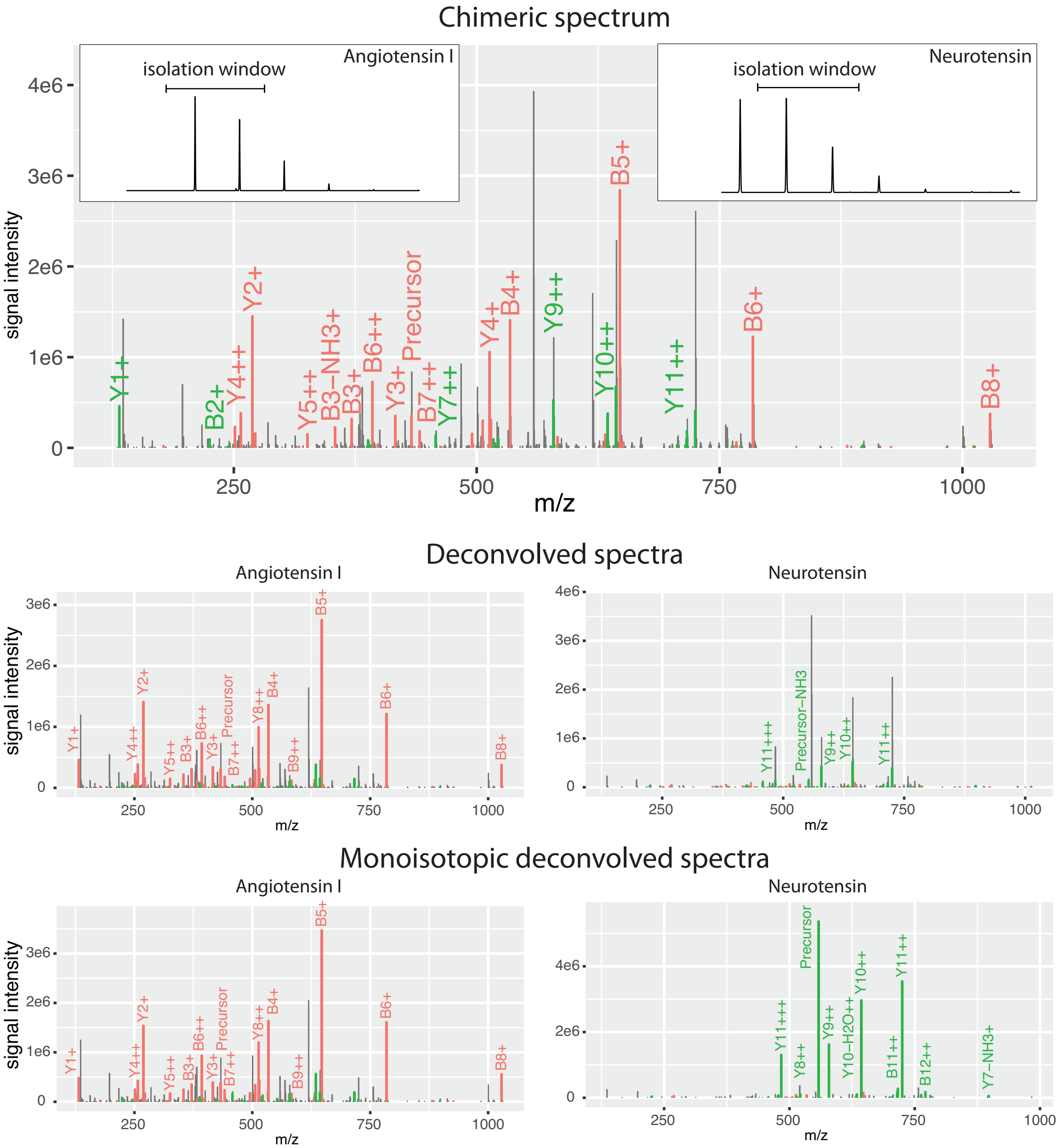


Figure 4.4: Chimeric and deconvolved spectra of angiotensin I and neurotensin from a direct infusion experiment. A multiplexed MS2 scan was performed using two separate isolation windows. Isotopes M and M+1 were isolated for angiotensin I, and isotopes M+1 and M+2 were isolated for neurotensin. A NNLS model was solved to create deconvolved and monoisotopic spectra. Peaks matching angiotensin I and neurotensin monoisotopic fragments colored red and green, respectfully. Unmatched peaks are colored grey. Inset: Observed MS1 spectrum of the corresponding peptide and the employed isolation window.

spectrum, which mimicked the single-peptide version. The monoisotopic angiotensin I spectrum was of higher-quality than the deconvolved-only spectrum; it matched the same angiotensin I fragments, but with higher intensity due to de-isotoping, and the high-intensity neurotensin fragments were no longer matches.

4.3.2 Whole-cell lysate evaluation

Table 4.1: Whole-cell lysate identification statistics

Data set	Method	PSMs	Unique peptides	Proteins
HELA_200ng_Major	original	21619	10656	2915
	deconvolved	26184 (+21.1%)	10883 (+2.13%)	3082 (+5.73%)
	mono + deconvolved	**28117 (+30.05%)**	**11441 (+7.36%)**	**3182 (+9.16%)**
HELA_1ug_Major	original	24616	17738	3722
	deconvolved	30614 (+24.36%)	18746 (+5.68%)	3972 (+6.71%)
	mono + deconvolved	**29818 (+21.13%)**	**18644 (+5.10%)**	**3931 (+5.61%)**
SW480_1ug_PS	original	30182	17319	3728
	deconvolved	34484 (+14.25%)	18047 (+4.20%)	3920 (+5.15%)
	mono + deconvolved	**35194 (+16.60%)**	**18524 (+6.95%)**	**3985 (+6.89%)**
HELA_400ng_Core	original	20147	15788	4221
	deconvolved	20349 (+1.00%)	15513 (-1.74%)	4345 (+2.93%)
	mono + deconvolved	**20760 (+3.04%)**	**15868 (0.50%)**	**4394 (+4.09%)**

To test the NNLS model on more complex, high-throughput MS-based proteomics data, four whole-cell lysate experiments were obtained from three separate laboratories. Each experiment was performed using different sample concentrations and instrument settings, and a different mass spectrometer was used by each laboratory. For each experiment, analyses were performed on three data sets: 1) the original spectra, 2) deconvolved spectra, and 3) monoisotopic deconvolved spectra. A database search was performed on the data sets to obtain peptide and protein identifications. deconvolved spectra increased peptide-spectrum-matches, unique peptides, and proteins in all cases except the HELA_400ng_Core experiment (Table 4.1). The monoisotopic deconvolved spectra further increased every category for every experiment, except for the HELA_1ug_Major experiment. The greatest improvement was observed for the HELA_200ng_Major experiment, which is likely because the λ penalty optimization was performed on that data set. The HELA_400ng_Core experiment had the smallest improvements. A likely cause of the poorer results was the low automatic-gain-control (AGC) setting of 2e4. This value controls the number of ions isolated per MS2 scan, and a low value increases scan speed, but results in poor isotopic distribution statistics and missing peaks. On the other hand, the HELA_1ug_Major experiment used the highest recommended AGC target of

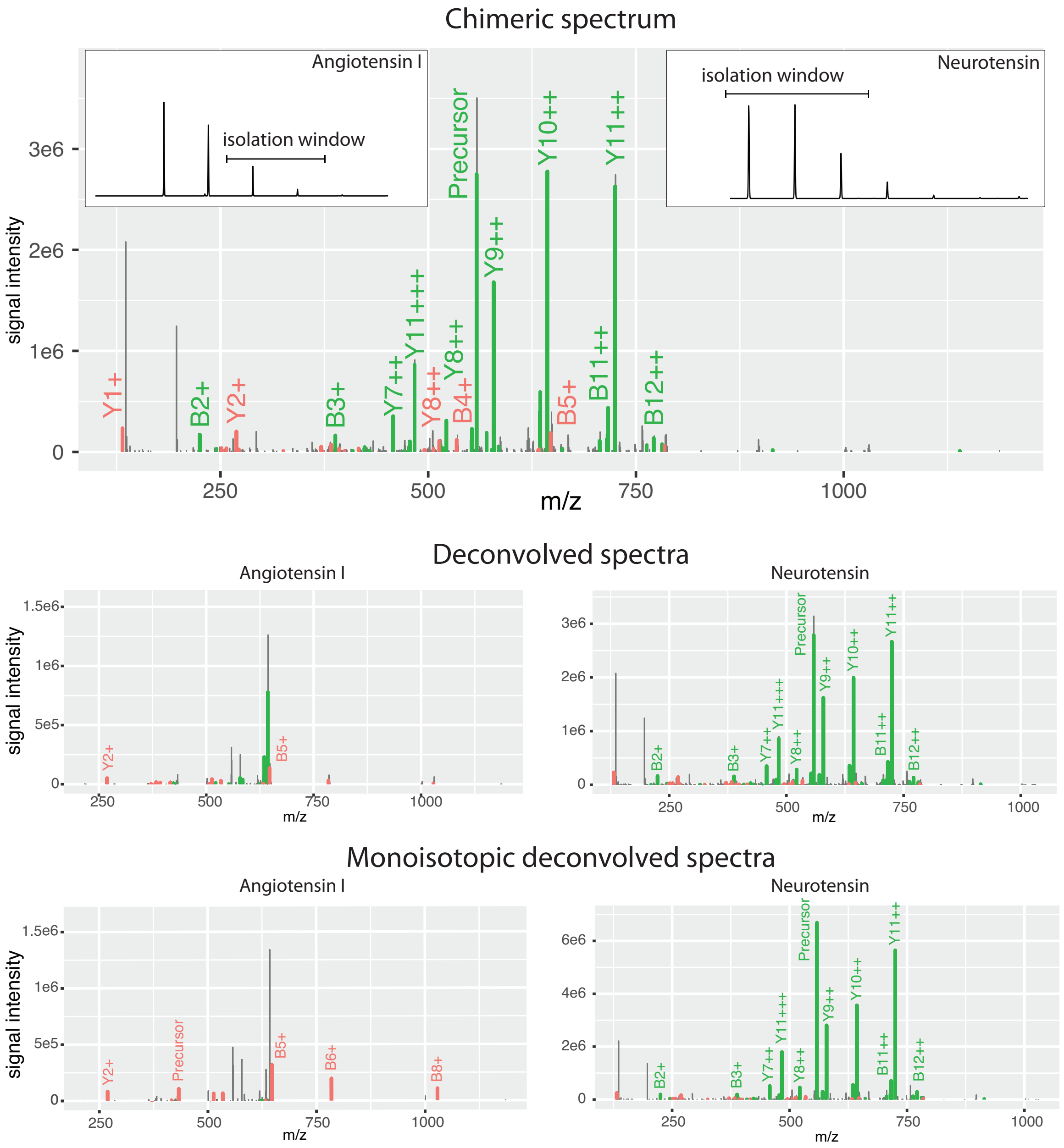


Figure 4.5: An alternative chimeric and deconvolved spectra of angiotensin I and neurotensin from a direct infusion experiment. As opposed to Figure 4.4, isotopes M+2 and M+3 were isolated for angiotensin I, and isotopes M, M+1, and M+2 were isolated for neurotensin.

2e5 and concomitantly had nearly eight times more chimeric spectra with three peptide identifications than the SW480_1ug_PS experiment (Table 4.2).

Table 4.2: PSMs per spectrum

Data set	1 PSM	2 PSMs	3 PSMs
HELA_200ng_Major	25224	1152	2
HELA_1ug_Major	22034	3802	60
SW480_1ug_PS	28290	3440	8
HELA_400ng_Core	19154	800	2

Though the increase in PSMs was substantial, less than half resulted in new peptide identifications. The remaining peptides were already identified in other scans from the original spectra and are therefore redundant. Serendipitously, these redundant spectra provide an opportunity to further confirm the quality of the NNLS model and the subsequent PSMs. Given the chromatographic conditions used in the whole-cell lysate experiments, a peptide's elution profile spans only several seconds to a minute. Therefore, correctly identified redundant peptides should come from spectra that were acquired at similar times. However, since the PSM false discovery rate was controlled to 1%, some deviations due to incorrect peptide identifications are expected. To test this, the scan numbers for deconvolved peptides were compared to their corresponding scan numbers in the original data (Figure 4.6). Identifications with the same scan number were excluded. For all data sets, a nearly one-to-one correlation was observed, which strengthens the confidence of deconvolved identifications.

Finally, a detailed inspection was performed on a chimeric spectrum that resulted in three distinct peptide identifications (Figure 4.7). The MS1 spectrum acquired prior to the MS2 spectrum contained four identified precursors, and four other unexplained peaks within the bounds of the isolation window. The monoisotopic peak of the green precursor is located at the edge of the isolation window which causes poor isolation efficiency that is unaccounted for in the template generation process. After deconvoluting the spectra, AA sequences for the three intense peptides were identified, while the low-intensity precursor remained unidentified. The deconvolved spectra still contained interfering peaks from the other peptides, particularly in the spectrum for VMLMASPSMEDLYHK, which contained most of the peaks from the original chimeric spectrum. Regardless, the three peptide-spectrum-matches were high scoring and all passed a 0.1% FDR.

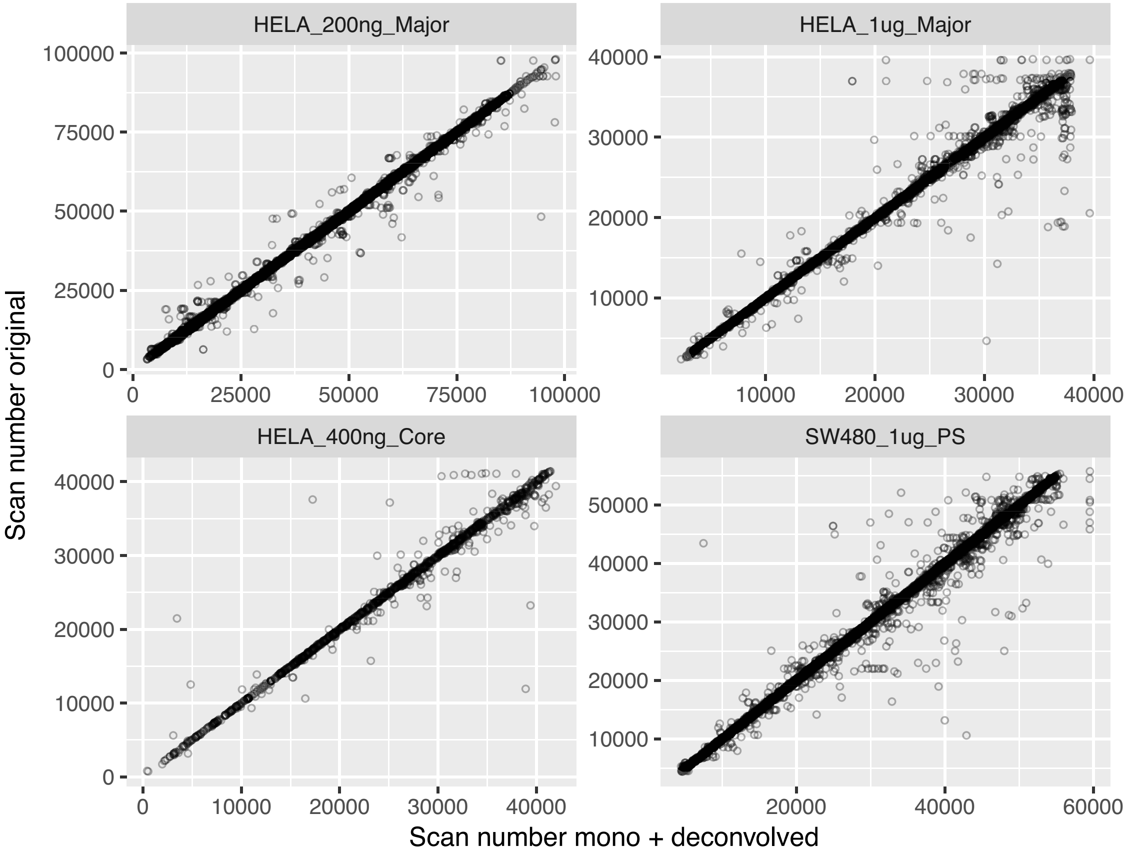


Figure 4.6: Scan number correlations. Scan numbers for peptides identified from deconvolved spectra were plotted against scan numbers from the original spectra that identified the same peptide. Only identifications with different scan numbers between the two experiments were included.

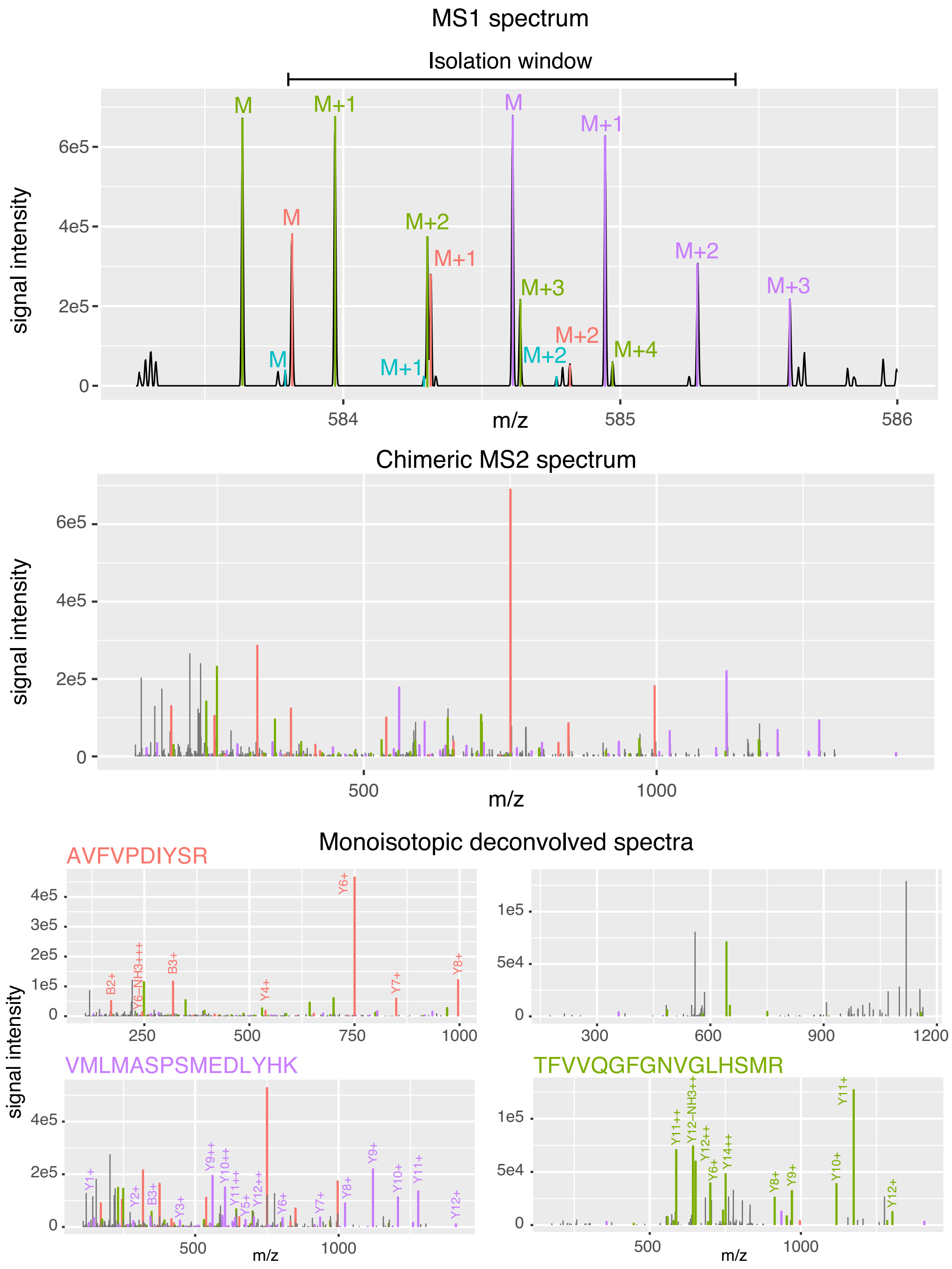


Figure 4.7: MS2 scan number 20990 from the HELA_1ug_Major experiment and its prior MS1 scan. Four precursor peptides are in the bounds of the isolation window used for the subsequent MS2 scan. The chimeric MS2 spectrum was deconvolved and de-isotoped, and three distinct peptides were identified. Non-grey colors correspond to peaks from distinct peptide species.

4.4 Discussion

Due to the complexity of the human proteome, and the limitations of peptide chromatography, chimeric spectra are often unavoidable. In fact, in many experiments they are more common than single-peptide spectra. The NNLS model developed in this work provides a novel method to deconvolve these spectra. To use this method, MS2 spectra must be acquired on a high-resolution, accurate-mass instrument. Low-resolution instruments, such as ion traps, are unable to resolve signals from different nominal isotopes and therefore fragment isotopic distributions cannot be examined. Additionally, some instrument settings have a large impact on the quality of deconvolution. Low AGC targets create spectra with highly variable isotopic distributions, and should be avoided. Isolation widths should be kept narrow to avoid isolating a precursor's entire isotopic distribution, otherwise the relationship between a fragment and precursor encoded by the fragment isotopic distributions would be eliminated. Lastly, some instruments automatically choose the *m/z* scan range of an MS2 scan based on the mass of the targeted precursor. Using this functionality, some fragments from an inadvertently co-isolated precursor with a larger mass will be missed. Nevertheless, the improvement observed in this study on four varying experiments demonstrates that the deconvolution method is robust to vastly different instrument settings.

Though the increase in PSMs following deconvolution was substantial, there are opportunities for further improvement. Penalty parameters should be optimized for each data set, and possibly for each spectrum. Efficient optimization procedures need to be developed, as the one used here involved a computationally intensive grid-based search of possible penalty values evaluated on the entire data set. Two improvements can be made to increase the quality of fragment isotopic distribution templates. First, the non-uniform isolation efficiency during MS2 scans is currently ignored, but can be taken into account when calculating approximate fragment isotope distributions. Second, the distribution of heavy isotopes along a peptide may be determined from deviations in the fragment isotopic distributions. It may be possible to iterate between generating templates, solving the NNLS model, and adjust the templates to compensate for the deviations.

Deconvolution of chimeric spectra through fragment isotopic distributions creates possibilities for new analysis pipelines and data acquisition strategies. From the analysis side, the deconvolution method presented here is complementary to previous methods to sequence chimeric spectra and they can be integrated together. For example, iterations of database searches and the subsequent subtraction of matching peaks can be performed after deconvolution. When co-fragmented peptides have similar masses and the same set of isotopes were isolated, subsequent subtraction may be the only option to identify multiple peptides as their fragments would have nearly identical isotopic distributions.

In the context of data acquisition strategies, improvements can be made to both data-dependent and data-independent acquisition. For data-dependent acquisition, instead of using the same isolation window parameters throughout an experiment, isolation windows can be adjusted in real-time to maximize the difference between fragment isotopic distributions. By shifting the isolation width and offset, the set of isolated precursor isotopes can be controlled and chosen to improve deconvolution. Furthermore, the isolation of only the monoisotopic peak can be avoided, which is a problematic case for the NNLS model.

For data-independent acquisition, narrower isolation windows can used as opposed to the current practice of wide isolation. Though the limited scan speed of contemporary mass spectrometers will force a small *m/z* range to be interrogated per experiment, the MS2 spectra are more amenable to analysis. Moreover, the utilization of fragment isotopic distributions will be greater because a precursor's isotopes will be split between adjacent scans and create complementary fragment isotopic distributions that can be identified. This additional constraint may significantly improve chimeric spectrum deconvolution.

Finally, for any acquisition strategy the chromatographic gradient can be shortened. This will require less instrument time, but concomitantly increase peptide co-elution and therefore generate more chimeric spectra. With successful chimeric spectra deconvolution, the disadvantages of short experiments may no longer out-weight the benefits.

4.5 Conclusion

I developed a method to deconvolve chimeric spectra into separate components for each co-fragmented peptide. I treated an observed spectrum as a linear combination of possible fragment isotopic distributions and solved a non-negative least squares model (NNLS) regularized with a sparse group lasso. Using the positive coefficients, individual spectra were created for each peptide. Furthermore, the deconvolved spectra were de-isotoped and converted into monoisotopic versions. The resultant single-peptide spectra are compatible with any sequencing algorithm. The NNLS model was tested on whole-cell lysate mass spectrometry experiments obtained from multiple laboratories. The deconvolved spectra increased peptide-spectrum-matches, unique peptides, and proteins. This algorithm describes the first application of approximate fragment isotope distributions in the literature.

CHAPTER 5: SIMULATION OF DATA ACQUISITION

5.1 Introduction

Mass spectrometry is an analytical technique used in proteomics to identify and quantify proteins. Many of its applications share the underlying goal of uncovering the full protein complement in a biological sample. However, for complex samples this is rarely achieved partly because the number of ion populations exceeds that which contemporary instruments can individually target for AA sequence analysis with an MS2 scan (Jurgen et al., 2011). Acquisition algorithms are necessary to control the data acquisition process and manage the limited scan speed.

Data-dependent acquisition (DDA) constitutes a major class of data acquisition algorithms. DDA algorithms perform an MS1 scan to determine the mass-to-charge ratio (*m/z*) and signal intensity of ions currently entering the mass spectrometer, followed by sequence determining MS2 scans on ions from a subset of detected peaks. The standard DDA algorithm, TopN, selects ions for MS2 scans that contributed to peaks of greatest signal intensity from the latest MS1 spectrum. Several other approaches, as well as adjustments to TopN, have been proposed in order to increase peptide and protein identifications (Zerck et al., 2013; Graumann et al., 2012; Liu et al., 2011; Rudomin et al., 2009; Scherl et al., 2004). However, TopN continues to be the dominant choice for acquisition control despite its bias towards abundant proteins and relatively poor reproducibility.

Currently, evaluation of novel acquisition strategies requires access to both a mass spectrometer and its application programming interface (API). Unfortunately, few instrument vendors provide an API, and therefore the pool of researchers with the necessary tools to explore this field is extremely limited. An alternative is to evaluate algorithms with *in silico* simulations. Existing simulator software for mass spectrometry proteomics has focused on generating ground truth data and realistic signals for MS1 and MS2 spectra (Smith and Prince, 2015; Noyce et al., 2013; Schulz-Trieglaff et al., 2008; Bielow et al., 2011). However, the size of the simulations is limited due to speed, memory, and/or disk space requirements. Most importantly, the simulated MS2 spectra lead to nearly perfect peptide-spectrum-matches (PSMs) when analyzed with existing

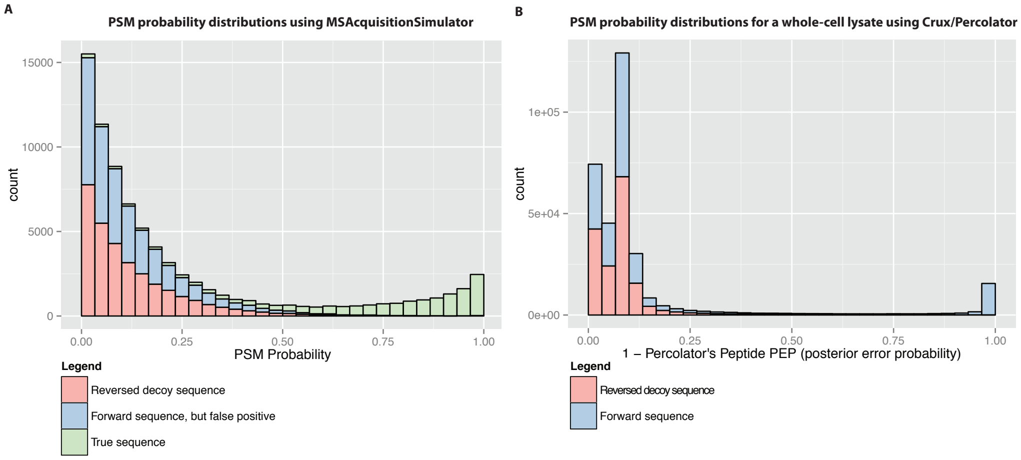


Figure 5.1: **A**. Distribution of PSM probabilities from AcquisitionSimulator using TopN. Both false positive forward PSMs and reversed decoy PSMs are generated. False positive and decoy PSM probabilities are sampled from an exponential distribution with a user-specified lambda parameter. Lambda=8 was used for this simulation. Both peptide and protein-level false discovery rates can be estimated using the decoys. **B**. Peptide probability distributions from a real whole-cell lysate experiment using Crux and Percolator (Park et al., 2008).

database search algorithms. The reasons for this include the absence of co-fragmentation from neighboring ions within an isolation window, the difficulty of predicting fragmentation patterns, and potentially other not yet understood phenomena. This limitation makes evaluating any acquisition strategy impractical as the metrics for success are based on the number of confident peptide and protein identifications.

Here, I present an acquisition simulator that produces PSMs with realistic peptide AA sequences and probability assignments for DDA strategies by foregoing fragmentation simulation and instead directly generating PSMs based on the precursor ion fractions of MS2 scans. It builds upon previous work on LC-MS simulations and scales to larger data sets due to probabilistic models of ion generation and subsequent pruning of rare ions. This allows for an increased number of proteins, peptides, and post-translational modifications that can be simulated.

5.2 Methods

MSAcquisitionSimulator consists of three standalone command line programs. The first, FASTASampler, assists users in creating a FASTA file containing the proteins to be included in the simulation. Users select a protein FASTA file, the distribution of protein abundance and the number of proteins to be sampled. The

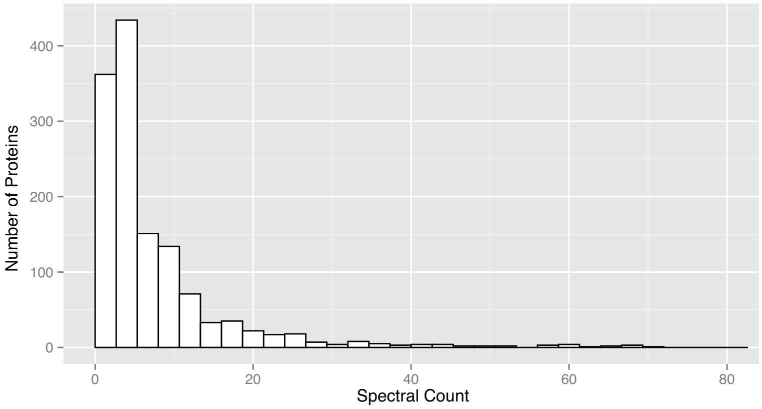


Figure 5.2: Histogram of spectral counts per protein. Spectral counts represent the number of peptide-spectrum-matches that were assigned to a protein. Titin, the largest human protein with 34,350 amino acids was omitted for visual clarity. It was observed with 307 spectra. Only true positive proteins and PSMs identified at their respective 1% FDRs were included.

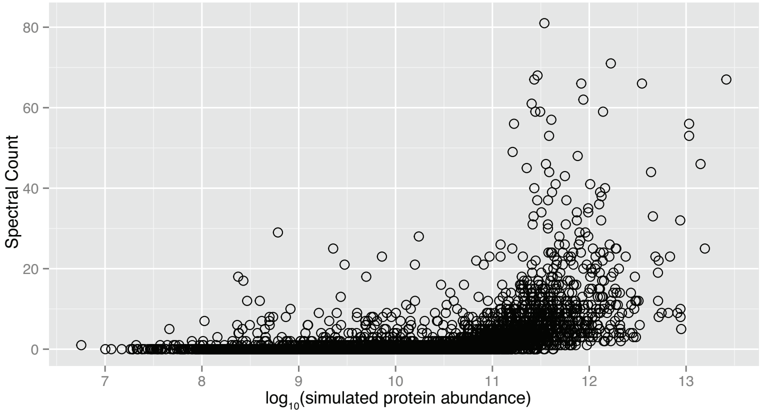


Figure 5.3: Spectral counts vs simulated abundance plotted for simulated proteins. Titin, the largest human protein with 34,350 amino acids was omitted for visual clarity. It was observed with 307 spectra. Only true positive proteins and PSMs were included. Only the PSMs were required to be identified at a 1% FDR.

output contains a random subset of proteins with each header appended with a ”#” followed by the protein’s abundance.

The second program, GroundTruthSimulator, uses the previously created FASTA file and a configuration file containing simulation parameters to simulate digestion, AA modifications, elution time, elution profiles, ionization efficiency, and charge and isotope distributions. It outputs the tab delimited ground truth data on the generated ions, which will be used for testing acquisition algorithms. In contrast to previous simulators, there are no limits on missed cleavages, enzymatic termini, or modifications. A probabilistic approach is taken instead and rare peptides are efficiently pruned.

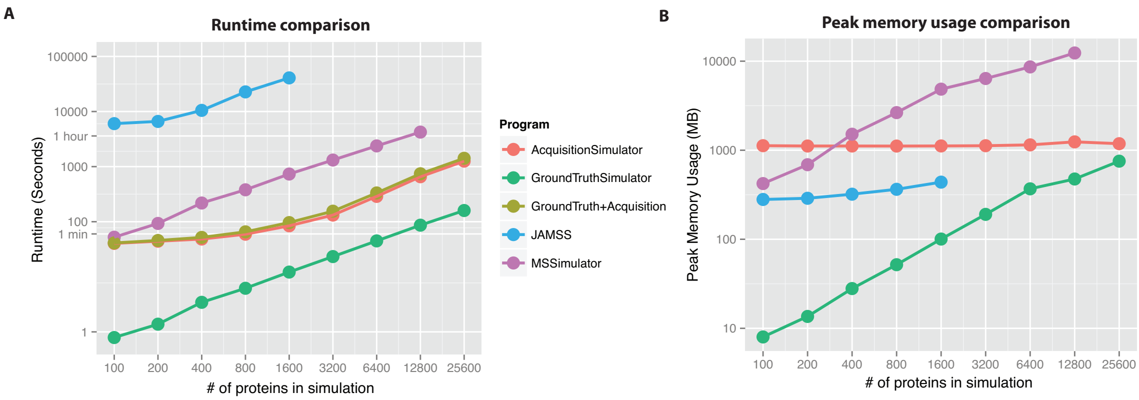


Figure 5.4: Performance comparisons were performed on an Apple MacBook Pro with 16 GB 1600 MHz DDR3 RAM and two quad-core 2.8 GHz Intel Core i7s using the GNU time command. FASTASampler was used to generate increasingly larger data sets for simulation. In order to achieve reliable comparisons, similar parameters were used for each program, namely strict trypsin cleavage, no missed cleavages, no post-translational modifications, a 1 hour gradient, and 1 MS scan per second. JAMSS and MSSimulator generate MS2 spectra in a very different fashion from MSAcquisitionSimulator, so this feature was turned off for JAMSS and MSSimulator, but still enabled for AcquisitionSimulator. JAMSS simulations were stopped after 1,600 proteins due to the long runtime, and MSSimulator was stopped after 12,800 proteins due to memory constraints.

Finally, the third program, AcquisitionSimulator, takes as input the previously generated ground truth file and another configuration file. This program takes a data-dependent acquisition algorithm developed by a user and simulates it on the ground truth data. It models ion accumulation, MS1 spectra, scan time duration, and database search PSMs (Figure 5.1). Currently, MS2 spectra are not simulated. Instead, the PSMs are generated by sampling from the list of precursors isolated in an MS2 scan and candidate peptides from a database search. This approach leads to true positives, false positives, and reversed decoy matches similar to real experiments (Figures 5.2, 5.3). Ion accumulation is simulated by numerically integrating an ion's elution profile. Ion isolation efficiency is not modeled and assumed to be 100%. The elapsed time for a scan is equal to the scan overhead time plus the larger of either the injection time or the transient time. This models the scan time for a QExactive-like instrument. The output includes an mzML file and a PSM graph file, which is used as input for the Fido protein inference algorithm (Serang et al., 2010). Speed and memory usage comparisons with existing simulation software are provided in (Figure 5.4).

Model parameters are described in Table 5.1. To model digestion, all enzymes specify a probability of cleavage given the AA present at the N-terminal side of the cleavage site, N, and the AA present at C-terminal side of cleavage site, C, and are assumed to be independent of each other. That is:

Table 5.1: Variable definitions

Variable	Description
$\boldsymbol{e}$	Array of enzymes in the simulation
ion	An ion and all its attributes (sequence, charge, neutrons relative to the monoisotopic isotope, PTMs, etc)
ion_i	The AA at position i of the peptide represented by *ion*
N_{ion}	Random variable for the index of the N-terminus of *ion*
C_{ion}	Random variable for the index of the C-terminus of *ion*
Z_{ion}	Random variable for the charge state of *ion*
$IE_{peptide}$	Random variable for the ionization efficiency of *peptide*
M_{ion}	Random variable for the number of neutrons greater than the monoisotopic form of *ion*
cleavage	Random boolean variable for a peptide cleavage event
$ionization_{ion}$	Random boolean variable for an ionization event for the peptide corresponding to this *ion*
PTM_{ion}	Array of the modification state for each AA on *ion*. This includes the state of no modification.
$\boldsymbol{P}_{ion}$	Array of proteins that can generate *ion*

$$\Pr(cleavage|N = n, C = c, e) \tag{5.1}$$

is given for all $e \in \boldsymbol{e}$. The probability of a peptide's cleavage from a single copy of its protein is then equal to the probability of cleavage at the peptide's N-terminus, C-terminus, and no cleavages in between, conditional on the enzymes in the simulation:

$$\begin{aligned}\Pr(N_{ion} = n_{ion}, C_{ion} = c_{ion}|\boldsymbol{e}) = &\Big(1 - \prod_{e \in \mathbf{e}} 1 - \Pr(cleavage|N = ion_{n_{ion}-1}, C = ion_{n_{ion}}, e)\Big)\\ &\cdot\Big(1 - \prod_{e \in \mathbf{e}} 1 - \Pr(cleavage|N = ion_{c_{ion}}, C = ion_{c_{ion}+1}, e)\Big)\\ &\cdot\Big(\prod_{i=n_{ion}}^{c_{ion}-1} \prod_{e \in \mathbf{e}} 1 - \Pr(cleavage|N = ion_i, C = ion_{i+1}, e)\Big)\end{aligned} \tag{5.2}$$

Each modification has a probability of occupying a particular site (e.g. there are probably serines that are never phosphorylated), and a percentage of that AA that will be modified (e.g. if a particular serine *is* chosen to be phosphorylated, maybe only 1% of it ever exists in that form at any given time), given as $\Pr(PTM)$. Candidate modification sites are first randomly assigned to each protein based on each PTM's probability of occupying a particular site. Afterwards, for each peptide created during the digestion process, modification combinations are created and their probability of existing is calculated as:

$$\Pr(\mathit{PTM}_{ion}) = \prod_{i=n_{ion}}^{c_{ion}} \Pr(\mathit{PTM}_{ion_i}) \tag{5.3}$$

For the case where PTM_i is the state of no modification:

$$\Pr(\mathit{PTM}_{ion_i} = \text{no mod}) = 1 - \sum_{\text{possible PTMs assigned to this AA}} \Pr(\mathit{PTM}) \tag{5.4}$$

If the sum of probabilities for PTMs assigned to a particular AA is greater than 1, then their probabilities for that AA are normalized to sum to 1.

During simulation of the ground truth data, the ionization efficiency of each ion is randomly selected from a uniform random distribution. All ions of the same peptide have the same ionization efficiency.

$$\Pr(\mathit{ionization}_{ion}) = \mathit{IE}_{peptide} \tag{5.5}$$

$$\mathit{IE}_{peptide} \sim \text{uniform}(0, 1) \tag{5.6}$$

The probability of an ion having a particular charge is modeled as a binomial distribution, with the binomial distribution's probability of success chosen randomly for each peptide in the simulation.

$$\Pr(Z_{ion} = z_{ion}) = \text{binomial}(n, z_{ion}, p) \tag{5.7}$$

$$n = 1 \text{ + number of basic AAs in peptide} \tag{5.8}$$

$$p = 0.7 + r \tag{5.9}$$

$$r \sim \text{uniform}(0, 0.3) \tag{5.10}$$

The isotopic distribution for a molecule is computed using Mercury++ which uses a fast-Fourier transform to convolve the various element isotopes and their probabilities (Rockwood and Haimi, 2006). For the purposes of our model, it is assumed the value of $\Pr(M_{ion} = m_{ion})$ is given.

The probability of a particular ion's existence is assumed to be based on all the previous equations, and that they are all independent of each other. Therefore the probability of an ion is given by:

$$\begin{aligned}\Pr(\mathit{ion}|\boldsymbol{e}) = \Pr(N_{\mathit{ion}} = n_{\mathit{ion}}, C_{\mathit{ion}} = c_{\mathit{ion}}|\boldsymbol{e})\\ \cdot \Pr(\mathit{PTM}_{\mathit{ion}}) \cdot \Pr(\mathit{ionization}_{\mathit{ion}})\\ \cdot \Pr(Z_{\mathit{ion}} = z_{\mathit{ion}}) \cdot \Pr(M_{\mathit{ion}} = m_{\mathit{ion}})\end{aligned} \tag{5.11}$$

Furthermore, the number of copies of a particular ion A_{ion} is its probability of existence from a single protein, times the total protein abundance of proteins that can generate that ion $\boldsymbol{P}_{\mathit{ion}}$:

$$A_{\mathit{ion}} = \Pr(\mathit{ion}|\boldsymbol{e}) \cdot \sum_{\mathit{protein} \in \boldsymbol{P}_{\mathit{ion}}} A_{\mathit{protein}} \tag{5.12}$$

The elution time for each peptide is determined by the BioLCCC (Liquid Chromatography of Biomacromolecules at Critical Conditions) library (Gorshkov et al., 2006). BioLCCC models the adsorption of peptides on porous media and can calculate the expected elution time for a particular molecule given the dimensions of column length and diameter, pore size, solvent concentrations, gradients, and flow rates. The effects of post-translational modifications can be modeled by specifying estimates of binding energy, which are user-specified parameters found in the ground truth configuration file.

The shape of an ion's elution profile due to liquid chromatography is modeled by an Exponential Gaussian Hybrid (EGH) function (Lan and Jorgenson, 2001). The EGH takes into account the tailing typically observed at the end of a elution profile. The two parameters of the EGH are elution width (σ) and the amount of tailing (τ) (Figure 5.5). These elution profiles are used to determine the number of ions reaching the ion detector by numerically integrating the profiles using Simpson's method at one-millisecond intervals for every ion present at the current time and *m/z* constraints. This integration continues until either a desired total ion count is reached, or the maximum injection time is reached.

To generate realistic PSMs, the precursor ion fraction (PIF) is first calculated for each peptide in an MS2 scan. The PIF is defined as the sum of ion intensities for a given peptide (i.e., the sum of all its isotope intensities) divided by the total ion intensity of the scan. Next, one peptide is randomly selected from these peptides—weighted by their PIF. If the peptide is contained in the peptide database (including reversed decoy sequences) that is being used to mimic a database search, and within the precursor mass tolerance, then the

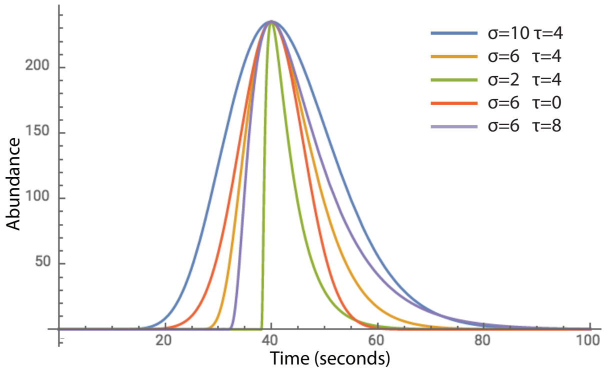


Figure 5.5: Elution profile modeled by an Exponential Gaussian Hybrid function. The two parameters, σ and τ, control the width of the distribution and the amount of tailing, respectively. Default parameters of $\sigma = 6$ and $\tau = 4$ are shown in orange.

PSM AA sequence is set to this peptide and the PSM probability is set to the corresponding PIF. If peptide is not in the peptide database, then a peptide is chosen in a uniform random fashion from the peptides in the database search within the mass tolerance of the targeted precursor *m/z*, and given a PSM probability sampled from a truncated exponential distribution.

5.3 Case study

To demonstrate the utility of MSAcquisitionSimulator, I evaluated three simple DDA algorithms–TopN, RandomN, and WeightedN. TopN selects the most intense peaks for MS2 scans, RandomN samples from a uniform random distribution of the observed peaks, and WeightedN samples from a random distribution weighted by observed peak intensity. Dynamic exclusion was enabled, and MS1 spectra were de-isotoped prior to MS2 selection decisions. FASTASampler was executed on the human proteome provided by UniProtKB using a log-normal abundance distribution and 50% of the proteins, resulting in 45,809 protein AA sequences. Default configuration files were used with both GroundTruthSimulator and AcquisitionSimulator. TopN resulted in the greatest number of confident protein and peptide identifications, closely followed by WeightedN, and RandomN provided far fewer protein identifications (Figure 5.6). RandomN's poor performance stemmed from the challenges in targeting low-intensity ions. The wide MS2 isolation window of 2 *m/z* captured neighboring intense ions and created spectra dominated by peptides whose monoisotopic mass fell outside the small precursor mass tolerance used to simulate the database search. Additionally, fewer scans were performed due to increased injection time.

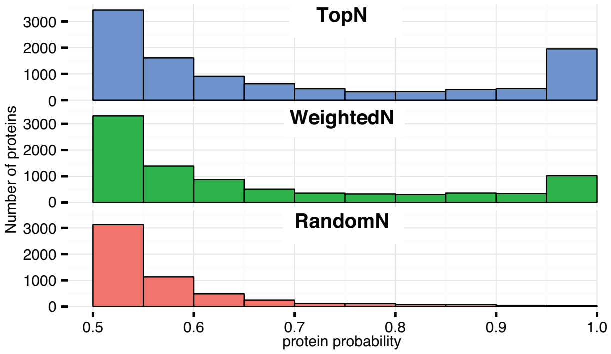


Figure 5.6: Distribution of protein probabilities using three different DDA algorithms on simulated data. Fido was used for protein inference with parameters alpha=0.1, beta=0.01, and gamma=0.5.

5.4 Conclusion

Data-dependent acquisition simulation will assist in the development and assessment of novel algorithms. The next generation of algorithms will likely further integrate data generation with data analysis, such as real-time peptide sequencing and protein inference. They may also become more goal-oriented, seeking to identify subsets of proteins, specific modifications, or to improve quantification. Their sophistication may also come at a computational cost too great for their implementation on contemporary mass spectrometers. Previous simulation software only support TopN data-data dependent acquisition and do not provide a software architecture for convenient modification of the acquisition algorithm. Furthermore, previous simulators are prohibitively slow and memory intensive for simulations of realistically complex peptide mixtures. Finally, they require re-generation of ground truth data for each simulation, which adds unnecessary time when testing multiple acquisition algorithms on the identical data set. MSAcquisitionSimulator overcomes these limitations and is the only practical tool for the simulation and evaluation of novel data-dependent acquisition algorithms.

CHAPTER 6: PREDICTING PROTEIN-PROTEIN INTERACTIONS

6.1 Introduction

Mapping the global protein-protein interaction network and defining its dynamic reorganization during specific cell state changes will provide an invaluable and transformative knowledgebase for many scientific disciplines. Recent advancements in two-hybrid technologies and affinity purification-mass spectrometry (APMS) have dramatically increased protein connectivity information, and therefore a high-coverage proteome-wide interaction map may be realized in the not-so-distant future. Specifically, technological and computational advancements in MS-based proteomics have increased sample throughput, detection sensitivity and mass accuracy, all with decreasing instrumentation costs. Consequently, to date $\sim$2,400 human proteins have been analyzed by APMS, as estimated through BioGRID and data presented herein (Stark et al., 2011). Similarly, the generation of arrayed human clone sets has revealed binary interactions among approximately 13,000 proteins (HI-2012 Human Interactome, Center for Cancer Systems Biology) (Rolland et al., 2014). While both approaches detect direct protein interactions, only APMS can detect indirect interactions—though with limited ability to distinguish between the two types.

In general, APMS-based protein interaction experiments are performed by selectively purifying a specific protein, termed the bait, along with its associated proteins from a cell or tissue lysate. Mass spectrometry is then used to identify and more recently quantify the bait and associated proteins within the affinity-purified protein complex, collectively termed the prey. Though a prey's presence supports its existence within a complex, high numbers of non-specific contaminants—owing largely to technical artifacts during the biochemical purification—lead to false protein complex identifications and therefore significantly hamper data interpretation. As such, numerous computational methods have been developed to differentiate between genuine APMS protein complex interactions and false-positive discoveries.

These methods can be broadly grouped along two axes: the type of quantitative data used, and which connectivity model is adopted. The first axis contains two categories: whether they use binary or quantitative APMS data. The binary presence of the protein is used as evidence for an interaction by methods such as

SAI, Hart, Purification Enrichment scores and Dice Coefficients (Gavin et al., 2006; Collins et al., 2007; Bader and Hogue, 2002; Bader et al., 2004; Gilchrist et al., 2004; Hart et al., 2007; Zhang et al., 2008). By ignoring the quantitative aspects of APMS data, many candidate interactions are treated equally even though there is more evidence for one over the other. More recently, computational approaches employed by SAINT (Choi et al., 2011; Teo et al., 2014), MiST (Jäger et al., 2011), CompPASS (Sowa et al., 2009) and HGSCore (Guruharsha et al., 2011) achieved improved scoring accuracy by taking advantage of label-free quantification using spectral counts, a semi-quantitative reflection of the abundance of a protein after purification. Additionally, SAINT-MS1 is an extension of SAINT that uses label-free MS1 intensities for quantification, which is better suited for low-abundant interactors (Choi et al., 2012). Along the second axis, there are also two categories: whether a spoke or matrix model is used to represent protein connectivity. The spoke model represents only bait-prey interactions, while the matrix model—used by the Hart and HGSCore methods—additionally represents all prey-prey interactions, resulting in a quadratic number of candidate interactions per experiment instead of linear, and therefore contain an order of magnitude more interactions to test. Though the matrix model has the potential to detect more true complex co-memberships, it not only has to determine whether either of the two prey proteins are contaminants, but also whether pairs of prey are in the same or distinct complexes with the bait—leading to more false positives. Each method has its merits and has been successfully applied in APMS experiments; however, their widespread utilization has been limited.

In addition to using features from APMS experiments to predict the validity of putative protein-protein interactions, success in the *de novo* prediction of protein interactions has been achieved through the analysis of indirect data (Beyer et al., 2007; Myers and Troyanskaya, 2007; Qiu and Noble, 2008; Qi et al., 2006). Specifically, mRNA co-expression has been shown to positively correlate with co-complexed proteins, and the Gene Ontology's (GO) biological process and cellular component annotations have proven to be useful for interaction prediction by utilizing semantic similarity (Resnik, 1995; Jain and Bader, 2010; Yang et al., 2012). Both co-expression and GO co-annotation are also commonly used metrics for evaluating the quality of predicted interactions. AA sequence and structural homology at the domain and whole-protein levels have established themselves as powerful predictors as well (Deng et al., 2002; Ben-Hur and Noble, 2005). Though individually useful, integration of these indirect sources using machine learning techniques such as support vector machines (Koike and Takagi, 2004), Random Forests (Lin et al., 2004), naïve Bayes (Jansen et al., 2003), and logistic regression (Bader et al., 2004) have further increased prediction accuracy. APMS data have also been used as a discriminative feature, once as a binary value representing an interaction's presence—far less powerful than the sophisticated APMS scoring methods now available (Qi et al., 2006), and once using a novel method that lacked rigorous comparison to other methods (Havugimana et al., 2012).

Among the label-free methods, only SAINT's software is available for public use. It can be executed as a standalone program, or through two separate web applications—Prohits (Liu et al., 2010) and the CRAPome (Mellacheruvu et al., 2013). CompPASS provides a public web interface to search its data, but no option to employ the algorithm on private data sets. Aside from APMS scoring methods, numerous web applications are available for *de novo* protein-protein interaction prediction (Franceschini et al., 2013; McDowall et al., 2009). These methods do not incorporate new APMS data, and therefore provide an insufficient resource for researchers wishing to integrate their own experiments into the predictions.

Given the independent successes of using direct and indirect data to predict protein-protein interactions, I enhanced HGSCore, CompPASS, and SAINT by incorporating a variety of indirect data using logistic regression classification models to identify genuine interactions from human APMS experiments. To foster its use within the proteomics community, I developed Spotlite, a web application for executing both the enhanced and original APMS scoring methods on novel data sets. In addition to providing an integrated scoring tool, the resulting protein interactions are annotated for function, model organism phenotype and human disease relevance.

6.2 Methods

6.2.1 Data collection

To develop a classification strategy capable of efficiently segregating false positive protein interactions from true interactions within APMS-derived data, I collected five publicly available and well-diversified APMS data sets (Table I). These data were received directly from the authors or from their respective publications, whose sequencing parameters and filtering criteria are described in their methods. The data contained spectral counts, baits, and preys for each experiment. For the purposes of establishing a classifier, I defined known protein-protein interactions as those deposited in iRefWeb (Turner et al., 2010) (`http://wodaklab.org/iRefWeb/` Release 4.1), physical interactions from BioGRID (`http://thebiogrid.org/` Release 3.2.105), and the HI-2012 Human Interactome project's two-hybrid data from the Center for Cancer Systems Biology at the Dana-Farber Cancer Institute (`http://interactome.dfci.harvard.edu/`) (Rolland et al., 2014).

Protein AA sequences and cross database accession mappings were downloaded from IPI (Kersey et al., 2004) (http://www.ebi.ac.uk/IPI/ final releases) and UniProt/SwissProt (UniProt Consortium, 2012) (http://www.uniprot.org/ Release 09/2013). Protein domains were determined with PfamScan (Punta et al., 2012) (http://pfam.sanger.ac.uk/ Release 26.0) using an e-value threshold of 0.05. Entrez Gene IDs, official symbols, aliases, and gene types were extracted from NCBI Gene's FTP site, http://www.ncbi.nlm.nih.gov/gene (gene_history.gz and gene_info.gz - downloaded 10/05/13).

Gene homolog data was downloaded from NCBI's Homologene (http://www.ncbi.nlm.nih.gov/homologene Build 66). Pearson correlation coefficients for co-expression data were downloaded from COXPRESdb (Obayashi and Kinoshita, 2011) (http://coxpresdb.jp/) for Homo sapiens (version c4.1), Mus musculus (version c3.1), Caenorhabditis elegans (version c2.0), Gallus gallus (version c2.0) Macaca mulatta (version c1.0), Rattus norvegicus (version c3.0), and Danio rerio (version c2.0). Ontology hierarchies and annotations were downloaded on 10/05/13. The Gene Ontology supplied the biological process and cellular component ontology hierarchies, where the annotations were downloaded from NCBI Gene's FTP site (Ashburner et al., 2000). The Mammalian Phenotype Ontology (relevant organism: Mus Musculus) hierarchy and annotations were downloaded from Mouse Genome Informatics (Smith and Eppig, 2009) (http://www.informatics.jax.org/). The Human Phenotype Ontology's hierarchy and annotations were downloaded from www.humanphenotype-ontology.org (Robinson et al., 2008). The Disease Ontology annotations were taken from its associated publication's supplemental data (http://projects.bioinformatics.northwestern.edu/do_rif/) and the hierarchy from the OBO Foundry (Osborne et al., 2009) (http://obofoundry.org/).

6.2.2 Feature calculation

For classification, all putative APMS-derived protein-protein interactions were characterized by one APMS scoring method feature and several indirect features. The APMS feature is the negative natural log p-value of either the HGSCore, CompPASS WD-score, or SAINT probability. The HGSCore is capable of testing matrix model interactions, however, for implementation within Spotlite, I restricted it to spoke model interactions for consistency with the other methods and computational efficiency. SAINT scores were computed using the spectral count version of SAINTexpress (Teo et al., 2014) version 3.1. I modified this version to output the full precision of probability calculations, as opposed to the default 2 digits. Only the TIP49 and HDAC data sets were applicable, as the SAINTexpress model requires control experiments.

The number of virtual controls and replicates were set to the number of controls and maximum number of replicates for each data set. For CompPASS, in cases where both proteins of a candidate interaction were tested as baits, the smaller p-value was chosen.

Listing 6.1: Pseudo code for permuting an APMS dataset

```
input  : $mean\_prey\_per\_exp$, $mean\_TSC\_per\_exp$, $bait\_TSC$, $prey2TSC$
         a vector of length $p$, $exp2bait$ a vector of length $e$
output: $e \times p$ matrix representing spectral counts of a permuted dataset.
$permuted\_dataset$ = new $e \times p$ matrix;
// Ensure each prey has at least one experiment
for $prey$ = 1 to $p$ do
    $exp$ = random integer from 1 to $e$;
    $permuted\_dataset[exp, prey] = 1$;
    $prey2TSC[prey]$ -= 1;
// Fill each experiment with prey
for $exp$ = 1 to $e$ do
    for $i$ = 1 to $mean\_prey\_per\_exp$ do
        $prey$ = sample without replacement from prey2TSC, excluding
          prey already in experiment $exp$;
        $permuted\_dataset[exp, prey] = 1$;
// Fill each experiment with spectral counts
for $exp$ = 1 to $e$ do
    for $i$ = 1 to $mean\_TSC\_per\_exp$ do
        $prey$ = sample without replacement from prey2TSC, including
          only prey already in experiment $exp$;
        $permuted\_dataset[exp, prey]$ += 1;
// Distribute bait spectral counts
for $i$ = 1 to $bait\_TSC$ do
    $exp$ = random integer from 1 to $e$, excluding experiments that were
      controls;
    $bait = exp2bait[exp]$;
    $permuted\_dataset[exp, bait]$ += 1;
```

Sampling without replacement decrements $prey2TSC$ to a minimum of 1 to ensure sampling is never performed on an empty set

The p-values for APMS scores in each data set were computed by generating simulated data sets via permutation of spectral counts and protein identifications (Listing 6.1) and is similar to a previously described approach (Sowa et al., 2009). First, each prey protein was represented by its total spectral count (TSC) in the original data set, excluding instances where it was the bait. Simulated experiments were generated by randomly sampling without replacement from this weighted set of prey until each experiment contained the average number of proteins per experiment of the original data set. Sampling without replacement then continued until each experiment had a TSC equal to the average experiment TSC (excluding the bait) of

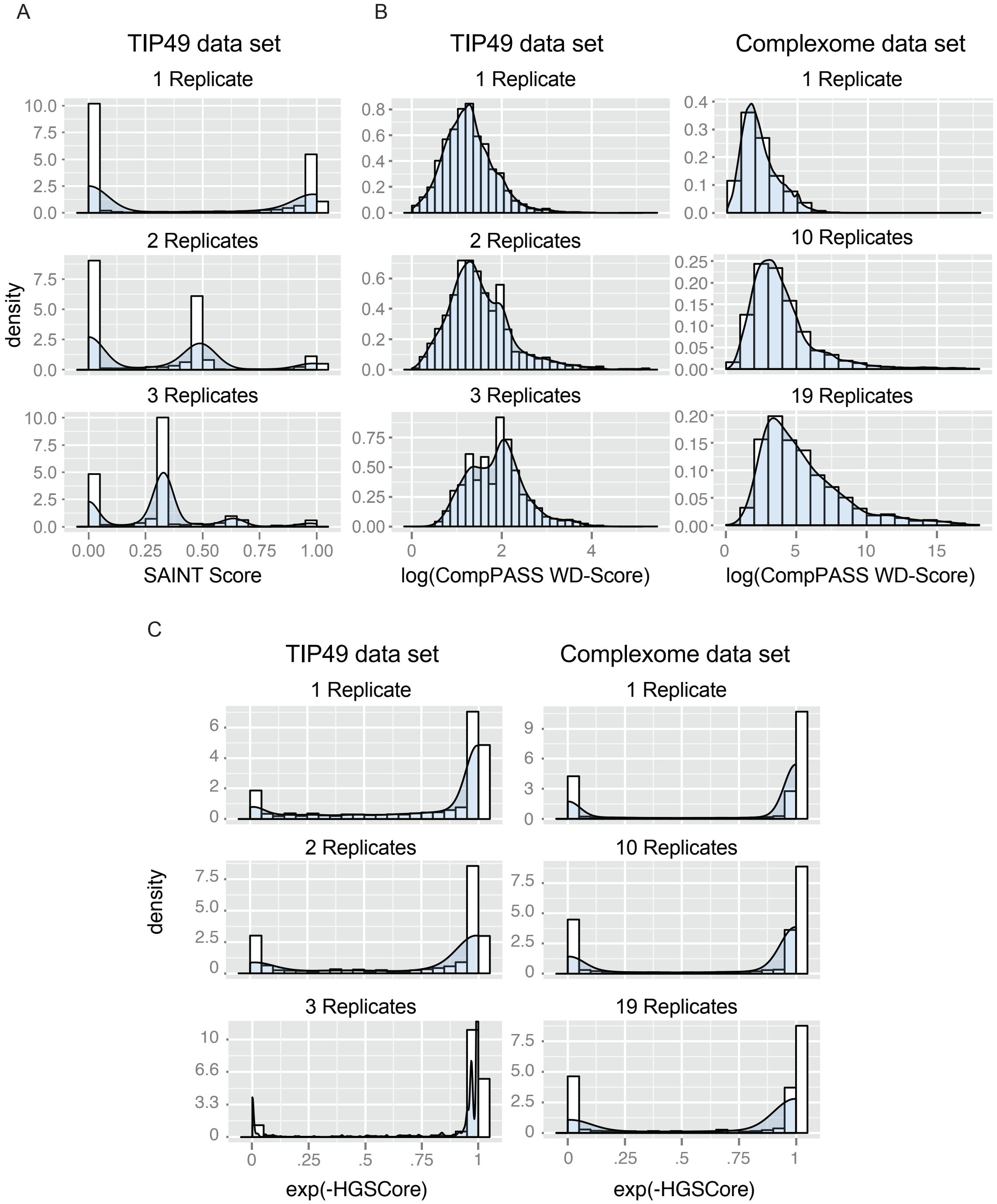


Figure 6.1: Distributions of SAINT, HGSCore, and CompPASS scores on data sets having different numbers of replicates per bait experiment (TIP49 and Complexome). SAINT was not tested on the Complexome data set because it did not contain the necessary control experiments.

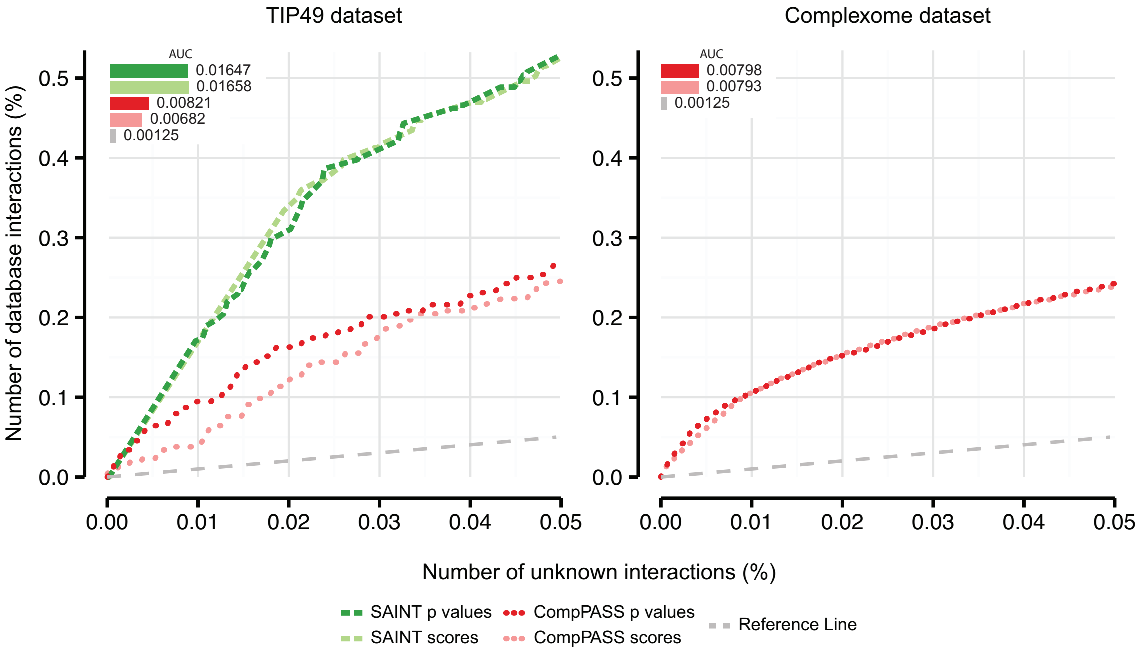


Figure 6.2: ROC curves of SAINT and CompPASS p-values and scores on data sets having different numbers of replicates per bait experiment (TIP49 and Complexome). SAINT was not tested on the Complexome data set because it did not contain the necessary control experiments.

the original data set. Finally, experiments were randomly sampled and given one bait spectral count at a time until the TSC of all baits in the simulated data set equaled that of the original. Replicate and control experiments went through the identical process, except controls were not given bait spectral counts. For the HGSCore, the simulated data sets were generated until the number of simulated interactions was 200 times the number of unique interactions in the original data set. However, for CompPASS and SAINT, since the distribution of scores depends on the number of replicates for a particular bait (Figure 6.1), the simulations were continued until the number of simulated interactions for each replicate number was equal to 200 times the number of total unique interactions in the original data set. Sorting interactions based on these conditional p-values had a slight increase in classification accuracy compared to raw scores on data sets with variable number of replicates (Figure 6.2).

In addition to these direct APMS-dependent features, indirect characteristics of a putative protein-protein interaction were also included. The correlation between mRNA expression patterns of two genes was quantified using the Pearson correlation coefficient (PCC). In total, seven co-expression features—one for each species discussed in Data Collection—were added to the classification model. The human feature is the PCC for the pair of human genes to be classified. There often exist multiple homologs of a gene within a different species; therefore the co-expression features for genes i and j, in non-human species k, were defined as the maximum PCC among the set of homolog pairs for that species, H_{ijk}:

$$\text{Coex}_{ijk} = \max(\text{PCC}_{mn}); m, n \in H_{ijk} \tag{6.1}$$

A separate feature was used for each of the five ontologies: biological process, cellular component, mouse mutant phenotype, human mutant phenotype, and human disease. Semantic similarity scores were utilized to determine how similar two gene's sets of annotations were to each other. I computed semantic similarity scores using the SimGIC method with downward random walks (Yang et al., 2012; Pesquita et al., 2008). Genes with zero annotations were assigned the root annotation for the corresponding ontology.

I used the Maximum Likelihood Estimation (Deng et al., 2002) method to calculate the probability of each potential domain-domain interaction. This required all interactions for Homo sapiens determined via an experimental method testing for direct interactions—two-hybrid, FRET, co-crystal structure, protein-peptide, and reconstituted complex. During cross-validation, interactions present in the APMS data sets were excluded to avoid training a feature on data I would later test against. A single protein AA sequence was used for each gene, with preference given to the longest UniProt/SwissProt sequence, followed by the longest IPI sequence. A false positive rate of 0.00063 and a false negative rate of 0.7 were used, which are required parameters, and were calculated in the same manner as previously described—assuming 130,000 total direct protein-protein interactions in the human interactome as was previously estimated (Venkatesan et al., 2009). The feature score was the probability of a protein pair interacting and is equal to the probability of at least one of their domains interacting. Computations were performed using the method's original software.

The final feature used was based on database interactions among the homologs of the two proteins in question. It is more likely a pair of proteins will physically interact if their homologs interact, however the extent to which these homolog interactions predict the human interactions depends on a number of factors such as the evolutionary distance of the homolog and the reliability of experimental systems used to determine the interaction. A naïve Bayes model was trained to determine the probability of a human database interaction given the presence or absence of homolog interactions using specific experimental systems. Specifically, I calculated:

$$\Pr(C \mid F_1, \ldots, F_N) \propto \Pr(C) \cdot \prod_{i=1}^{N} \Pr(F_i \mid C) \tag{6.2}$$

$$C = \begin{cases} 1 & \text{co-complexed protein pair} \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

$$F_i = \begin{cases} 1 & \text{co-complexed homolog pair using experimental system } i \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

The model probabilities were estimated from all human protein pairs, except during cross-validation where the test interactions were excluded from training this feature. The prior probability, $P(C)$, is equal to the percentage of all possible protein pairs that are annotated to be co-complexed interactions. Though ideally this would be replaced with an estimation of the true percentage, the predicted number of co-complexed interactions—unlike the predicted number of direct interactions—is an open problem. Fortunately, the true probability of an interaction given homologous interactions is not necessary for our machine learning classifier, but rather a proportional likelihood relative to other proteins. The model did not include evolutionary distance due to very small samples for many combinations of species and experimental systems.

6.2.3 Missing data imputation

Co-expression features are subject to missing values due to lack of microarray probes, and unknown homologs among the various species. Since the chosen species' co-expression patterns are strongly correlated (Obayashi and Kinoshita, 2011), missing values for a specific gene pair were imputed from its available co-expression values. Specifically, a linear regression model was calculated using each species' co-expression values as the response variable and every combination of remaining species' co-expression values as explanatory variables. With seven species, this corresponded to 5,040 models. When imputing a missing value, the model with the best R^2 using available data was applied. If no co-expression values were available for a gene pair, then pre-imputed feature averages were used.

6.2.4 Training data set construction

To segregate false positive protein interactions from true interactions, I trained and tested a two-layer classifier using a supervised learning approach on a subset of the human interactome and five APMS data sets. The first layer was a model for non-APMS features, and was trained on a data set comprising all database interactions as the positive class, while the negative class was a sampled subset of all unknown interactions equaling 20 times the size of the positive set. The negative set is commonly constructed in this manner because a very small percentage of all possible protein pairs are believed to physically interact, and therefore a random sample of all unknown interactions is expected to have few false negatives (Qiu and Noble, 2008; Qi et al., 2006; Jain and Bader, 2010; Ben-Hur and Noble, 2005; Koike and Takagi, 2004). Interactions present in any of the APMS data sets were excluded. The second layer was trained on the probability output of the first layer and the APMS scores of five published human APMS data sets. Each data set was scored with the three APMS scoring approaches—except for SAINT, which was only used on the data sets with controls, TIP49 and HDAC—resulting in five training data sets for each HGSCore and CompPASS, and two for SAINT. When used for training the model, each APMS data set was appended with all unobserved known and unknown interactions with its corresponding baits and given an APMS score of 0. Conversely, when used for testing, only observed interactions were included. Database interactions in the APMS data sets represented by a single publication employing either CompPASS, HGSCore, or SAINT were treated as unknown, as this would create a bias towards one of the methods.

6.2.5 Model training and evaluation

I approached the probabilistic scoring of APMS protein-protein interactions as a binary classification problem in which the two classes are: 1) pairs of proteins that directly or indirectly form a complex together (positive class), and 2) pairs of proteins that are never members of the same complex (negative class). In order to enhance each of the popular APMS scoring methods—HGSCore, CompPASS, and SAINT—a separate model was trained for each of the three, using that particular method as one of the features for the second layer of the classification model. For the first layer, three classification algorithms were evaluated—Random Forest, logistic regression, and SVM. For the second layer, logistic regression was used to combine the predictions of the first layer and one of the APMS scores. For cross-validation, the model of the first layer was trained, then each APMS data set was tested with the second layer classifier trained on the remaining data sets that used the same APMS scoring approach. Some overlap was present among data sets; therefore interactions present in the data set being tested were removed from the training set, avoiding the mistake of testing on trained data. The metric for success was the area under the partial receiver operating characteristic (ROC) curve, up to a false positive rate of 10%, as this region encapsulates the likely interval in which a 5% FDR threshold would lie. For SVM and logistic regression, each feature was centered and standardized by subtracting the feature mean and dividing by the feature standard deviation of all possible protein-protein interactions. For Random Forests—which are unable to extrapolate beyond the range of their training data—features were scaled to have the same range between each data set. Support vector machines were trained using either a linear or Gaussian kernel with no feature interactions. A grid-based search determined optimal cost parameters. Logistic regression was also performed without feature interactions. The Random Forest classifier was trained with 300 decision trees and splitting from a subset of four randomly selected features at each node. Ultimately, a linear kernel SVM and logistic regression were the best performing algorithms for the first layer model on these data, and logistic regression was chosen for its faster calculation speed. Features deemed insignificant by logistic regression were removed from the model and were comprised of the semantic similarity scores for human disease, human mutant phenotype, and mouse mutant phenotype. Many true interactions exist in our set of negative APMS interactions, resulting in a diminished estimate of true interaction prevalence, and therefore an inaccurate estimate of the logistic regression's intercept parameter, β_0. To correct for this, let $\hat{\beta_0}$ be the original intercept, π be the training data set's ratio of known to unknown interactions, and $\hat{\pi}$ be the expected ratio, estimated by accepting interactions with a 5% false discovery rate based on the model's APMS method. The second layer's intercept was then adjusted using the following equation:

$$\beta_0^* = \hat{\beta_0} + log(\frac{\pi}{1-\pi}) - log(\frac{\hat{\pi}}{1-\hat{\pi}}) \tag{6.5}$$

6.2.6 False discovery rate calculation

I currently compute false discovery rates (FDR) for only the APMS scoring algorithm used. First, p-values are calculated for each interaction's two scores by comparing them to their corresponding empirical null distributions determined via the previously mentioned simulation method. The p-value for a particular

Table 6.1: Public dataset statistics

Dataset	AP/IP Method	Experiments	Baits	Controls	Distinct Interactions	Mean Clustering Coefficient[a]
Complexome	Antibody	3,268	1,082	0	253,598	0.1226
DUB	HA	201	101	0	36,066	0.1290
AIN	HA	127	64	0	19,676	0.2013
TIP49	FLAG	35	27	9[b]	5,412	0.3333
HDAC	EGFP	30	10	7	10,175	0.2523

[a] Computed using a protein-protein interaction network comprised of only bait nodes, and edges between them derived from BioGRID using experiments testing direct interactions - reconstituted complex, co-crystal structure, protein-peptide, FRET, and two-hybrid.

[b] Merged from 27 initial control experiments.

score is then equal to 1 plus the number of simulated scores greater than or equal to that score, divided by 1 plus the number of simulated scores. The adjustment by a pseudo count of 1 is necessary because the null distributions were not generated by an exhaustive permutation method (Phipson and Smyth, 2010). Finally, with all p-values calculated, the FDR is controlled by the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). FDRs for the Spotlite classifiers will be the subject of future work.

6.2.7 FLAG affinity purification and western blot analyses

For FLAG affinity purification, HEK293T cells were lysed in 0.1% NP-40 lysis buffer (10% glycerol, 50mM HEPES, 150 mM NaCl, 2mM EDTA, 0.1% NP-40) containing protease inhibitor mixture (1861278, Thermo Scientific, Waltham, MA) and phosphatase inhibitor (78427, Thermo Scientific, Waltham, MA). Cell lysates were cleared by centrifugation and incubated with FLAG resin (F2426, Sigma-Aldrich Corporation, St. Louis, MO) before washing with lysis buffer and eluting with NuPAGE loading buffer (Life Technologies, Carlsbad, CA). Detection of proteins by Western blot was performed using the following antibodies: anti-FLAG M2 monoclonal (Sigma-Aldrich Corporation, St. Louis, MO), anti-MAD2L1 (A300-301A, Bethyl Labs, Montgomery, TX), anti- MCM3 (A300-192A, Bethyl Labs, Montgomery, TX), anti-SLK (A300-499A, Bethyl Labs, Montgomery, TX), anti-actin polyclonal (A2066, SigmaAldrich Corporation, St. Louis, MO), anti-KEAP1 polyclonal (ProteinTech. Chicago, IL), anti-DPP3 polyclonal (97437, Abcam, Cambridge, MA), and anti-VSV polyclonal (A190-131A, Bethyl Labs, Montgomery, TX)

6.3 Results and discussion

6.3.1 Comparisons to existing methods on public data

Existing spectral count-based APMS scoring methods demonstrate a high level of accuracy in predicting protein complex co-membership, thus making them appealing features for classification. I analyzed their performance on five data sets describing protein complexes associated with unique biological functions deubiquitination (DUB) (Sowa et al., 2009), autophagy (AIN) (Behrends et al., 2010), chromatin remodeling (TIP49) (Sardiu et al., 2008), histone modification (HDAC) (Joshi et al., 2013), and transcriptional regulation (Complexome) (Malovannaya et al., 2011) (Table 6.3.1). These data sets range extensively in their number of experiments, interaction network connectivity and purification technique, resulting in a diverse training set capable of testing the generalizability of APMS methods and our classifier. A direct comparison of three popular and fundamentally distinct scoring algorithms-HGSCore, CompPASS, SAINT-revealed overlapping and complementary prediction accuracies (Figure 6.3). Specifically, the three methods were applied separately to each data set, and the top 5% of interactions were accepted as a good and consistent point estimate of a 5% FDR. Although some methods performed better than others, each approach was capable of identifying known protein-protein interactions disjoint from the remaining two. That said, the intersection of the three data sets showed strong enrichment for validated protein interactions. Interestingly, despite the high overlap among known interactions (mean Jaccard coefficient of 0.512), there was large disagreement among the yet-to-be determined interactions (mean Jaccard coefficient of 0.206). As expected, no single method identified all of the previously annotated protein interactions. Each has their own scenarios where they are more appropriate to use than the other. The HGSCore, for example, performs poorly on small data sets such as HDAC (Figure 6.4) and as discussed in the method's original paper. SAINT is limited to data sets with appropriate and comprehensive controls, and CompPASS can have difficulty with data sets comprising of highly interconnected baits such as TIP49 (Figure 6.4). Therefore, I chose to improve each method individually through integration with indirect data to broaden and strengthen the confidence of selected interactions, and to allow users to choose the most suitable APMS method for their data set.

To further improve upon interaction predictions, I chose to include data outside of APMS that had previously been shown to correlate with co-complexed proteins. These indirect sources of evidence were mRNA co-expression patterns among seven species, GO annotation similarity, phenotypic similarity, domain-

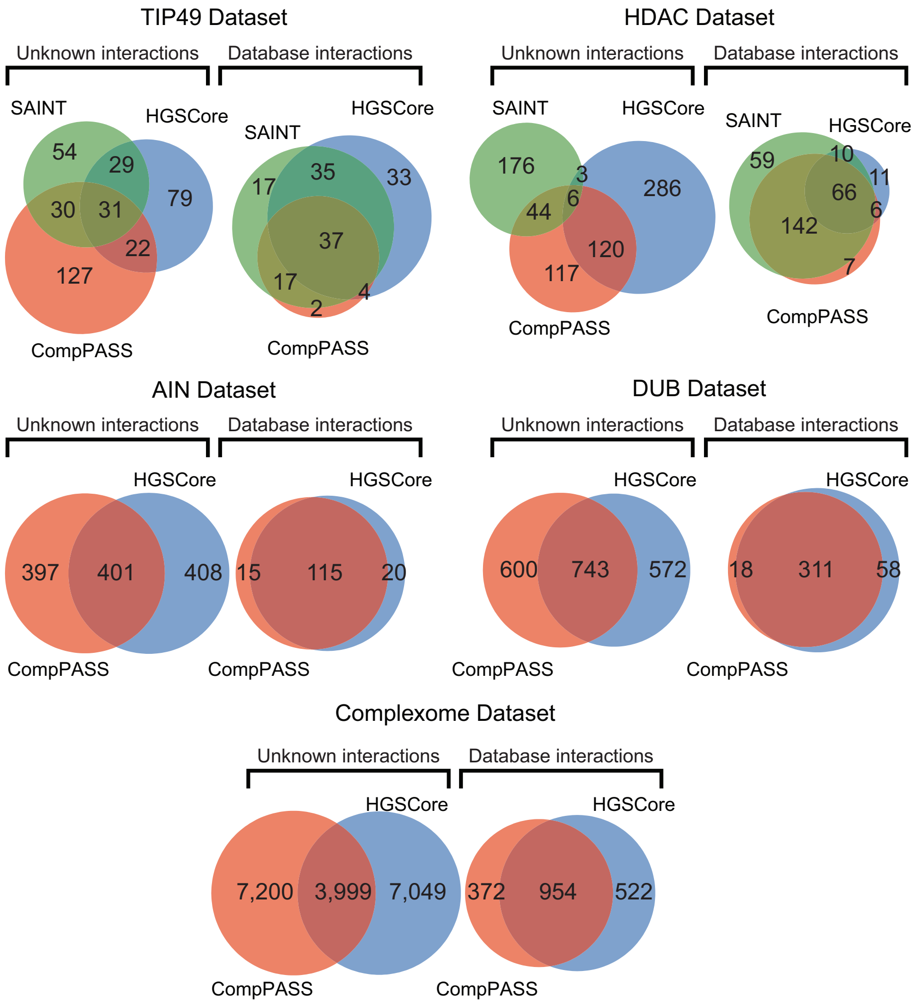


Figure 6.3: Comparison of accepted interactions using various APMS scoring methods. Overlaps of the top 5% of interactions for each APMS scoring method are shown for each data set. Areas are approximately proportional to the total number of interactions within their respective subsets.

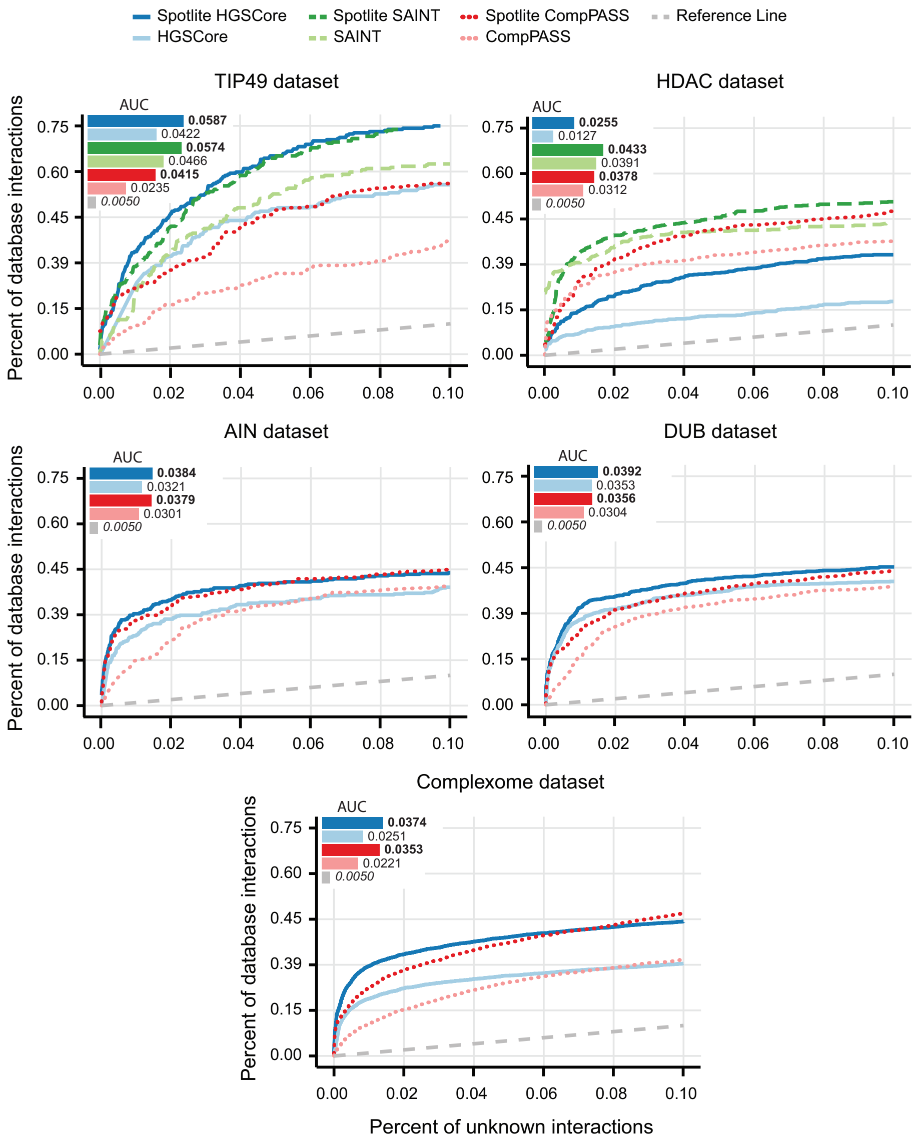


Figure 6.4: Classifier cross-validation and comparison. Receiver operating characteristic curves for each data set. Each scoring method's partial area under the curve is displayed in the graph insets.

domain binding affinities, and homologous interactions. Each was encoded into a feature, and along with the APMS scoring methods, describe a putative pair of interacting proteins. Then, using a two-layer logistic regression classifier, these interactions were predicted to be genuine based on the values of their corresponding features.

In order to benchmark these Spotlite classifiers against the standalone APMS scoring methods, I performed a variation of cross-validation by training our classifier on each combination of data sets, excluding one, and then testing on the remaining data set (Figure 6.4). Spotlite versions consistently outperformed their corresponding APMS only methods based on ROC curve analysis and partial AUC, which demonstrates greater sensitivity and specificity toward previously determined interactions. These data also demonstrate that the discriminatory patterns learned from each data set were generally applicable, as classification accuracy was superior across all cross-validation instances. Mutant phenotype and disease similarity were not selected as significantly discriminating features and were excluded from the model, but remain in the database for annotation purposes. To generate our final classifier for use in the Spotlite web application, all data sets were used for training. Table 6.3.1 shows each feature's coverage within the Spotlite database and its logistic regression log-odds coefficients. As expected, the APMS features were the most important features used for distinguishing between known and false or unknown co-complexed proteins.

6.3.2 Spotlite web application

Spotlite has been made available to the research community through a user-friendly web application that follows a simple workflow (Figure 6.5). Users may upload a tab-delimited file containing each experiment, its bait, prey, and each prey's spectral count. Next, identifier mapping is performed to determine the NCBI entrez gene id of the protein's gene. APMS scores are then calculated, as well as their corresponding p-values by determining the empirical null distribution via permutations of the original data set. Next, the indirect feature data, which has been pre-computed for every potential pair of genes, is retrieved from the database. Unmapped proteins, which have no retrievable indirect data, use raw feature averages to avoid bias towards predicting either true or false interactions. Finally, the data are scored by the logistic regression classifier. The false discovery rates are calculated and users can then explore and visualize their results through the website or export them to a spreadsheet. Users can choose whether use to the logistic regression classifier, or only the APMS methods. This is particularly useful for data sets that are not entirely of human origin

Table 6.2: Feature importances for logistic regression classifiers

				Log-odds coefficients[d]			
Feature	Type[a]	Database Coverage[b]	Training Coverage[c]	First Layer Model	HGSCore	CompPASS	SAINT
-ln(HGSCore p-value)	Direct	11.79%	100.00%	-	0.506	-	-
-ln(CompPASS p-value)	Direct	11.79%	100.00%	-	0.348	-	-
-ln(SAINT p-value)	Direct	11.79%	100.00%	-	-	-	0.496
Non-APMS model	-	-	-	-	0.230	0.230	0.197
Intercept	-	-	-	-2.699	-2.371	-2.370	-2.680
Domain-domain binding affinity	Sequence	70.32%	88.33%	2.693	-	-	-
Homologous interactions	Sequence	85.86%	99.53%	0.585	-	-	-
Cellular localization GO	Functional	61.69%	86.02%	0.324	-	-	-
Chicken co-expression	Expression	29.90%	41.21%	0.266	-	-	-
Mouse co-expression	Expression	53.91%	66.68%	0.210	-	-	-
Biological process GO	Functional	48.66%	84.33%	0.178	-	-	-
Human co-expression	Expression	70.42%	82.04%	0.153	-	-	-
Monkey co-expression	Expression	33.93%	39.33%	0.091	-	-	-
Fish co-expression	Expression	8.51%	15.63%	0.065	-	-	-
Rat co-expression	Expression	33.49%	45.45%	0.022	-	-	-
Worm co-expression	Expression	2.73%	5.23%	0.015	-	-	-

[a] Classification of the type of evidence a feature represents with respect to co-complexed proteins.

[b] Percentage of all potentially co-complexed pairs of genes within the Spotlite database containing values for a feature. APMS score coverages represent the percentage of bait-prey interactions tested, including preys with 0 spectra. Ontology coverages computed by taking the percentage of gene pairs in which both genes have at least 1 annotation. Homologous interactions coverage - both genes must have a known homolog in the same species. Domain-domain binding affinity coverage - both genes must contain a known domain.

[c] Coverages calculated identically to [b] - restricted to the training dataset.

[d] Coefficients are for scaled and centered features in the first layer model, and raw features in the second layers.

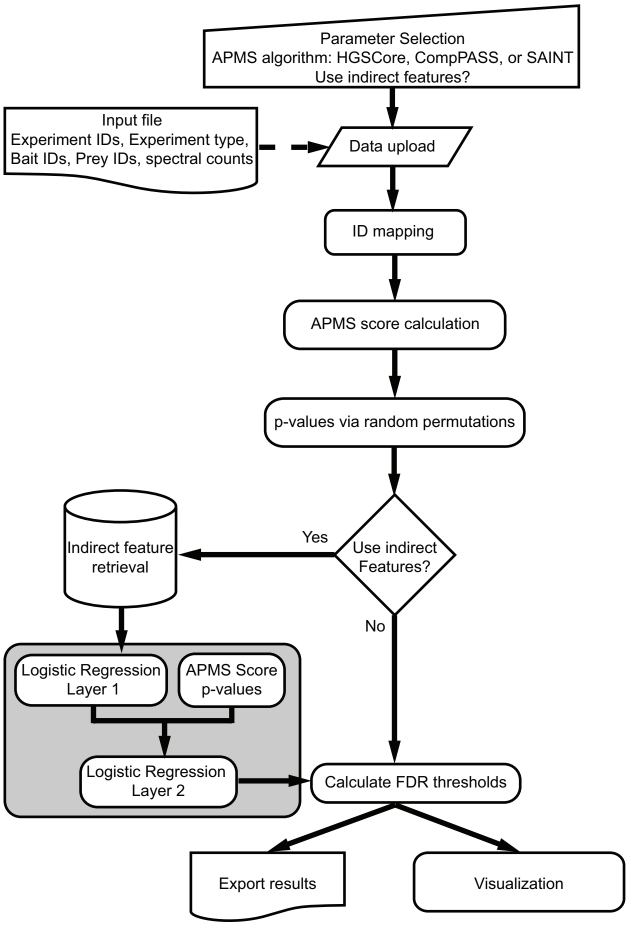

Parameter Selection
APMS algorithm: HGSCore, CompPASS, or SAINT
Use indirect features?
Input file
Experiment IDs, Experiment type,
Bait IDs, Prey IDs, spectral counts
Data upload
ID mapping
APMS score calculation
p-values via random permutations
Indirect feature retrieval
Yes
Use indirect Features?
No
Logistic Regression Layer 1
APMS Score p-values
Logistic Regression Layer 2
Calculate FDR thresholds
Export results
Visualization

Figure 6.5: Schematic of Spotlite workflow. Users perform parameter selection and supply input APMS data. Spotlite parses and scores the candidate protein-protein interactions. Options for visualization and data export are available through the user interface. The grey box represents the two-layer logistic regression classifier.

and therefore do not have indirect features contained within the Spotlite database. To maintain privacy, all uploaded APMS data and results are deleted after 24 hours of upload, or destroyed on command by the user.

6.3.3 Spotlite analysis of KEAP1 APMS data

To demonstrate its utility, performance, and ease in identifying true interacting proteins from APMS data, our previously published data on the KEAP1 E3 ubiquitin ligase affinity purified from HEK293T cells was re-analyzed (Hast et al., 2013). Specifically, cells engineered to stably express FLAG-tagged KEAP1 were detergent solubilized and subjected to FLAG affinity purification and shotgun mass spectrometry. Using biological triplicate KEAP1 APMS experiments and a reference set of an additional 44 FLAG purifications performed on 21 different baits, the KEAP1 protein interaction network was scored and visualized with

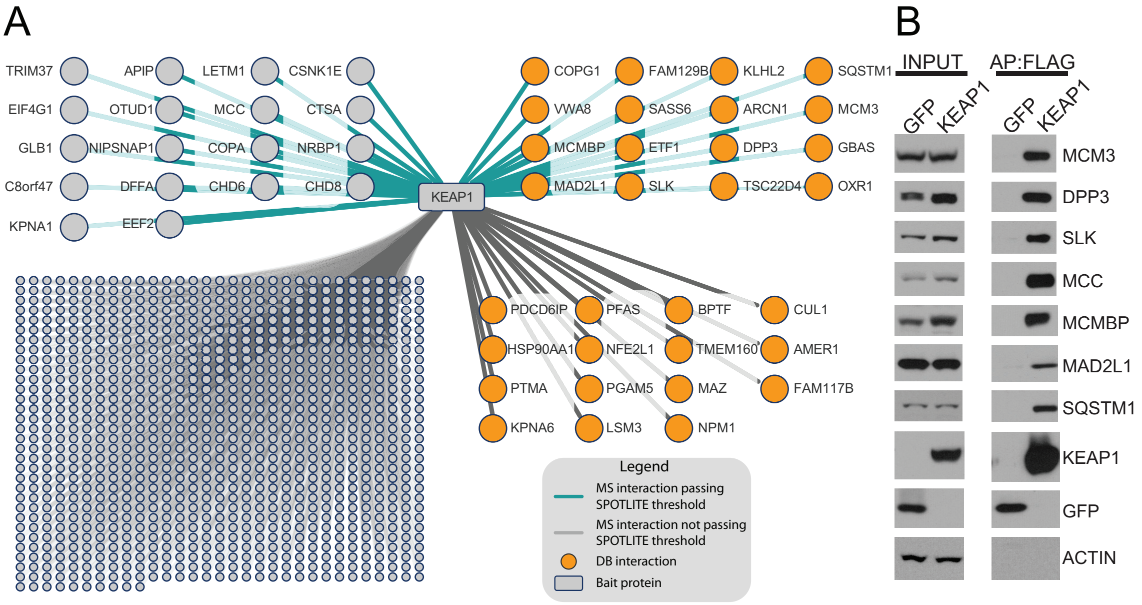
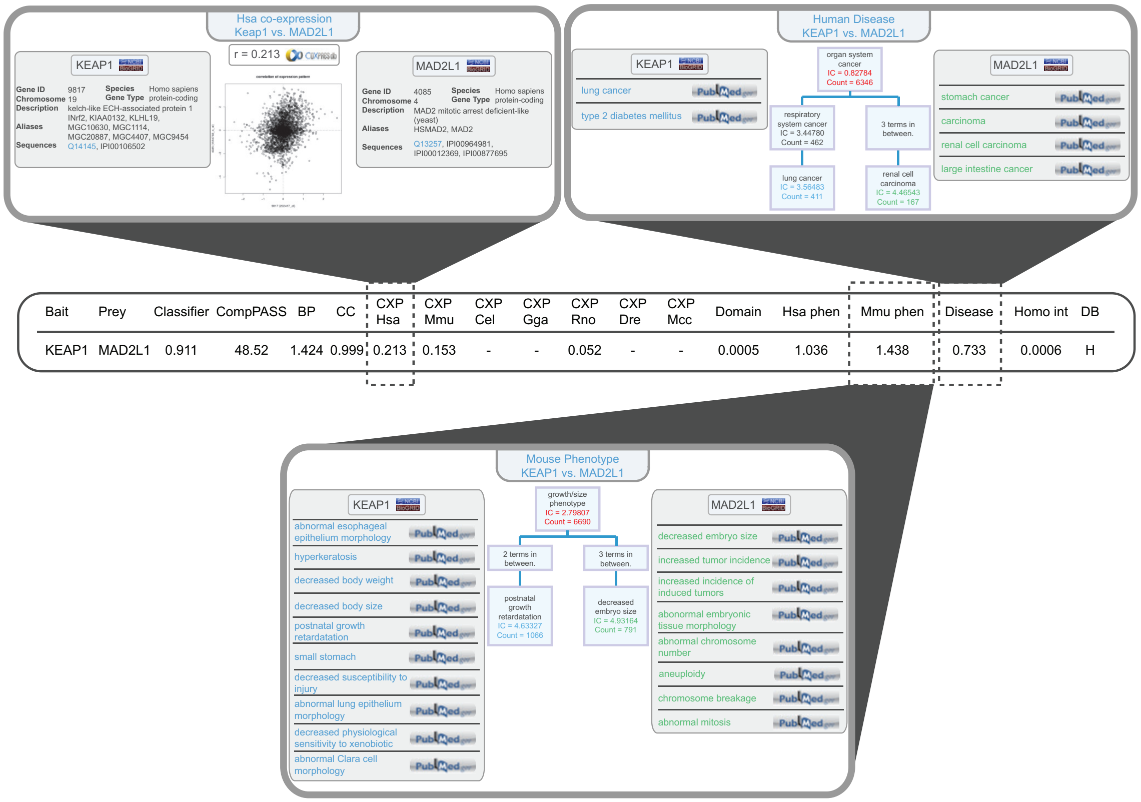


Figure 6.6: Spotlite application to KEAP1 APMS. (A) Spoke model interaction network after Spotlite-CompPASS scoring and accepting the same number of interactions as CompPASS-only with a 5% FDR. (B) FLAG affinity purified protein complexes from HEK293T cells stably expressing FLAG-GFP or FLAG-KEAP1 were analyzed by Western blot for the indicated endogenously expressed proteins.

Spotlite. The unfiltered KEAP1 data set contained 1,010 prey proteins, of which 32 were annotated as being previously identified as KEAP1 interactors (Figure 6.6A). After application of Spotlite-CompPASS and a global 5% FDR threshold based on CompPASS scores, the network reduced to 34 proteins. The same number of proteins were accepted for the Spotlite-CompPASS classifier, of which 16 were database interactions and 18 were putative novel interactors. Next, I selected seven KEAP1 interacting proteins that passed Spotlite thresholding for further validation by immunoprecipitation and Western blot analysis: MCM3, DPP3, SLK, MCC, MCMBP, MAD2L1, SQSTM1. All seven endogenously expressed proteins co-purified with FLAG-tagged KEAP1 (Figure 6.6B).

In addition to providing the logistic regression classification score, the Spotlite web application lists the following individual features for each protein pair: HGSCore, CompPASS, SAINT, gene ontologies for biological process (BP) and cellular component (CC), gene co-expression for seven species (CXP), domain-domain binding score (Domain), homologous interactions (Homo int), shared mutant mouse phenotypes (Phen), shared human diseases (Disease) and whether the proteins have previously been shown to interact (DB). As an example, Spotlite's visualization for the KEAP1-MAD2L1 interaction is provided in Figure 6.7. Both proteins affect growth and size in mice, specifically postnatal growth retardation with KEAP1 and



Hsa co-expression
Keap1 vs. MAD2L1

r = 0.213

KEAP1

Gene ID 9817
Species Homo sapiens
Chromosome 19
Gene Type protein-coding
Description kelch-like ECH-associated protein 1
Aliases INrf2, KIAA0132, KLHL19, MGC10630, MGC1114, MGC20887, MGC4407, MGC9454
Sequences Q14145, IPI00106502

MAD2L1

Gene ID 4085
Species Homo sapiens
Chromosome 4
Gene Type protein-coding
Description MAD2 mitotic arrest deficient-like (yeast)
Aliases HSMAD2, MAD2
Sequences Q13257, IPI00964981, IPI00012369, IPI00877695

Human Disease
KEAP1 vs. MAD2L1

KEAP1
- lung cancer
- type 2 diabetes mellitus

MAD2L1
- stomach cancer
- carcinoma
- renal cell carcinoma
- large intestine cancer

Bait	Prey	Classifier	CompPASS	BP	CC	CXP Hsa	CXP Mmu	CXP Cel	CXP Gga	CXP Rno	CXP Dre	CXP Mcc	Domain	Hsa phen	Mmu phen	Disease	Homo int	DB
KEAP1	MAD2L1	0.911	48.52	1.424	0.999	0.213	0.153	-	-	0.052	-	-	0.0005	1.036	1.438	0.733	0.0006	H

Mouse Phenotype
KEAP1 vs. MAD2L1

KEAP1
- abnormal esophageal epithelium morphology
- hyperkeratosis
- decreased body weight
- decreased body size
- postnatal growth retardatation
- small stomach
- decreased susceptibility to injury
- abnormal lung epithelium morphology
- decreased physiological sensitivity to xenobiotic
- abnormal Clara cell morphology

MAD2L1
- decreased embryo size
- increased tumor incidence
- increased incidence of induced tumors
- abonormal embryonic tissue morphology
- abnormal chromosome number
- aneuploidy
- chromosome breakage
- abnormal mitosis

Figure 6.7: Screenshots of Spotlite visualization for KEAP1-MAD2L1 data. Column headers on the main results screen are the following: Spotlite score (Classifier), APMS score (HGSCore, CompPASS, SAINT), gene ontologies for biological process (BP) and cellular component (CC), gene co-expression for seven species (CXP), domain-domain binding score (Domain), naïve Bayes' homologous interaction classifier (Homo int), shared phenotypes (phen), shared human diseases (Disease) and whether the proteins have previously been shown to interact (DB; H=high throughput, L=low throughput). Transparency is provided through a series of user-triggered pop-up windows which details the information used to generate the Spotlite feature scores.

decreased embryo size with MAD2L1. Additionally, both proteins are encoded by mRNAs, which positively correlate across human tissues, and both proteins are strongly associated with oncogenesis.

6.4 Conclusion

Protein mass spectrometry is quickly becoming a staple technology in academic laboratories. The rapidly decreasing instrumentation costs, often pre-packaged and streamlined bioinformatic pipelines, and enhanced mass accuracy and scan speeds are no doubt driving the recent explosion of protein mass spectrometry data. With similar advances in two-hybrid technologies, it is now economically feasible to pursue, and in fact achieve a fairly comprehensive proteome-wide binary interaction network. A key step in this endeavor is the computational filtering of spurious interactions within the resulting data sets.

After performing hundreds of APMS experiments directed at mapping protein connectivity central to various signal transduction pathways, the Ben Major lab and others quickly found the high rate of false-positive identification rate limiting and exceedingly expensive. Appreciating the need for an accessible and accurate APMS scoring algorithm, I developed Spotlite as a new computational tool capable of discriminating between true interactions and the contaminants within APMS data. Importantly, Spotlite was deployed through a web-based application that provides open access and transparency to any interested scientist. The implementation of popular APMS scoring methods provides researchers the ability to use the most appropriate method for their particular data set. Inclusion of indirect data as features within Spotlite's logistic regression model not only achieves increased prediction accuracy but also yields valuable information regarding shared biological function, phenotype and disease relationships among protein pairs.

Given the success of established scoring approaches employed by CompPASS, HGSCore and SAINT, I initially set out to define their relative performance on various APMS data sets, and by doing so to identify the most accurate approach for implementation within a classification scheme. However, our analyses revealed valuable complementarity between the algorithms, which appeared partially dependent upon the network architecture and size of the analyzed APMS data set, as well as the presence of control experiments. As such, the greatest success was found by providing a separate classification model for HGSCore, CompPASS, and SAINT—allowing the user to choose the most appropriate method for their data set. Though Spotlite's performance shows a marked improvement over existing methods, its success is governed by the small number of known protein interactions (positive data set), the lack of validated non-interactions (negative data set), and mislabeled instances used during training. Furthermore, many indirect features lacked high coverage, resulting in missing values. While these limitations may place a ceiling on current performance, data will continue to pour in and fill the gaps. Spotlite is expected to improve over time due to increased feature coverage, and re-training of the classifiers as larger and more comprehensive interaction networks become available.

A critical aspect of any supervised learning approach is the selection of a gold standard data set containing accurately labeled examples that are representative of the future data to be classified. While many protein-protein interactions are annotated, proteins known not to interact are rare—the Negatome being the sole available resource and of prohibitively small size (Smialowski et al., 2010). The common practice of treating all unknown interactions as false interactions leads to an issue when evaluating the performance of a classifier by ROC curves, because they require accurate knowledge of the ground truth. Though the number of true negatives in the training data sets is expected to greatly exceed the number of false negatives, the number of true positives is likely less than the number of false negatives--as there are many novel interactions still to discover. As I have shown, it is possible to train different classifiers that agree on the already known interactions-resulting in similar ROC curves-but with extremely different predictions for novel interactions. In this case it would be difficult to objectively decide which classifier had superior classification accuracy. An expensive and time-consuming solution would be to update the ROC curves after attempting low-throughput validation of many of the predictions. It would instead be desirable for the research community to generate several well-annotated interaction networks with extremely high accuracy and coverage.

Spotlite currently includes APMS scoring algorithms designed for spectral counting data; however, with the recent accessibility of high-resolution mass spectrometry and its accompanying software, scientists are transitioning to protein quantification based on peptide signal intensity for its superior limits of quantification and linearity. Accordingly, APMS computational methods will also need to support these in the future—as SAINT-MS1 has already accomplished, and Spotlite will as well. Additionally, labeled experiments comparing bait and control purifications within the same sample using SILAC, iTRAQ, or TMT tags are common, but still lack dedicated software for interaction prediction.

Presently, Spotlite classification using indirect features is only available for human APMS data; however, HGSCore, CompPASS, and SAINT themselves can still be used on any data set through the web application. Aside from integrating other species' indirect data using the current workflow, I envision the possibility of using APMS from multiple species to improve predictions through homologous interactions, which is already a powerful feature in our implementation. Along these lines, merging data sets from various laboratories has the potential to further increase accuracy. While this is currently possible with Spotlite, it should be done with great care as contaminants will vary due to differences in cell lines, mass spectrometers and protocols, leading to improperly high APMS feature values for mutually exclusive contaminants which now appear more unique. This combined analysis of data sets is an area of future research. A further limitation is that FDRs are based on the APMS scores instead of the Spotlite classifiers. Machine learning classifiers often use cross-validation to determine a threshold that achieves desired levels of specificity and sensitivity, however this would be far from accurate due to the community's limited knowledge of the true positives. Instead, it is recommended to accept the same number of interactions as the chosen APMS method would at the desired

FDR. This approach is expected to be conservative as the Spotlite classifiers have superior ROC curves. In the future, determining the empirical null distribution of the classifier scores will allow for controlling the FDR directly on the classifier scores.

A major focus of the Major lab's research is on the development of proteomic and functional genomic technologies to define the mechanics and disease contribution of the KEAP1. The KEAP1 protein functions as a CUL3-based E3 ubiquitin ligase, most well-known for its ubiquitination of the NFE2L2 transcription factor (Cullinan et al., 2004; Furukawa and Xiong, 2005; Zhang et al., 2004). Recently, somatic inactivating mutations in KEAP1 have been reported in a variety of solid human tumors, particularly in lung cancer (Padmanabhan et al., 2006; Singh et al., 2006; Ohta et al., 2008; Satoh et al., 2010; Solis et al., 2010; Takahashi et al., 2010; Konstantinopoulos et al., 2011; Li et al., 2011; Muscarella et al., 2011). The leading model posits that KEAP1 inactivation results in constitutive NFE2L2 transcriptional activation of antioxidant and pro-survival genes (Sykiotis and Bohmann, 2010; Ogura et al., 2010). APMS analysis of KEAP1 followed by Spotlite scoring and a 5% FDR filter revealed 34 associated proteins. Of the eight proteins validated to reside within KEAP1 protein complexes by IP/Western blot, the indirect data as visualized through the Spotlite web application drew attention to the KEAP1-MAD2L1 protein association. Specifically, the MAD2L1 protein is known to function pivotally within the spindle assembly checkpoint complex, which holds cells in metaphase until chromosome-spindle attachment is complete (Hoyt et al., 1991; Li and Murray, 1991). Like KEAP1, MAD2L1 is strongly associated with cancer; its over-expression drives chromosomal instability and aneuploidy (Sotillo et al., 2007; Schvartzman et al., 2011). MAD2L2 is also known to be ubiquitinated, although the E3 ubiquitin ligase is unknown (Osmundson et al., 2008; Kim et al., 2011). An intriguing possibility is that KEAP1 ubiquitinates MAD2L1 to control its activity and/or stability. Within cancer systems, somatic mutation of KEAP1 may coincide with elevated MAD2L1 activity, thus driving aneuploidy.

In conclusion, I have provided a user-friendly web application for predicting complex co-membership from APMS data. This web application employs a novel, logistic regression classifier that integrates existing, proven APMS scoring approaches, gene co-expression patterns, functional annotations, protein domains, and homologous interactions, which I have shown outperforms existing APMS scoring methods.

CHAPTER 7: CONCLUSION

The main objective of this dissertation work was to improve mass spectrometry (MS)-based proteomics through computational techniques. This objective was accomplished at three levels: 1) low-level signal analysis through pattern recognition of isotopic distributions, 2) instrument operation via the development of simulation software to evaluate novel data acquisition algorithms, and 3) post-processing data analysis through the creation of a machine learning classifier to predict protein-protein interactions from affinity purification-mass spectrometry (APMS) experiments.

7.1 Summary of results

In the following subsections, the major results of Chapters 3-6 are described.

7.1.1 Isotope distributions

The isotope distribution of a molecule is a fundamental characteristic that contains a wealth of information. The isotopic distribution is one of the few features measured by MS whose patterns can be accurately predicted. For fragment molecules, these patterns have been largely ignored or used inappropriately because the correct mathematical equations have not been available. I derived the equation to compute theoretical isotope distributions of fragment biomolecules in Section 3.2.1. On its own, this equation is of limited utility because it requires *a priori* knowledge of a molecule's elemental composition. However, I used this equation as the basis for two approximation methods developed in Sections 3.2.3 and 3.2.4. One of these outlined methods uses splines that were fit to the isotope probabilities of *in silico* generated peptides. The splines approximate isotope distributions 20x faster compared to the classic approximation approach which relies on the fast-Fourier transform. Importantly, the spline approach is equivalently accurate when compared to the fast-Fourier method.

I evaluated the methods to calculate fragment isotope distributions in three stages: 1) comprehensive *in silico* comparisons to theoretical peptides, 2) low-throughput experiments on a well-characterized peptide, and

3) in high-throughput using a typical MS-proteomics workflow on a whole-cell lysate. The approximations derived from the described algorithms matched the observed fragment isotopic distributions.

To facilitate utilization by the greater MS community, I contributed both methods for calculating theoretical and approximate isotope distributions to the OpenMS software library. Novel tools can now be implemented for tasks such as MS2 de-isotoping and chimeric spectra deconvolution. Additionally, the expected isotopic distributions can be integrated into the scoring approaches used by sequencing algorithms, which currently only reward monoisotopic fragments.

7.1.2 Chimeric spectra deconvolution

Chimeric mass spectra contain fragments from multiple distinct peptides and are therefore difficult for sequencing algorithms to assign matches for. I developed a method to deconvolve the spectra into separate components for each peptide in Chapter 4. I treated an observed spectrum as a linear combination of possible fragment isotopic distributions. Using the the CVX MATLAB package, I then created and solved a non-negative least squares model (NNLS) regularized with a sparse group lasso, and created individual spectrum for each peptide. The resultant single-peptide spectra are compatible with any sequencing algorithm. This algorithm describes the first application of approximate fragment isotope distributions in the literature.

To test and refine the NNLS model in a controlled setting, I created chimeric spectra by purposely co-isolating and co-fragmenting two distinct peptides (angiotensin I and neurotensin). Following the successful deconvolution of this two-peptide mixture, I tested a more complex peptide mixture. I obtained whole-cell lysates from multiple laboratories and deconvolved them with the NNLS model. More peptide-spectrum-matches, unique peptides, and unique proteins were identified following deconvolution. The smallest increases were from an experiment that isolated detrimentally small ion populations during MS2 scans, which leads to high variability in isotopic distributions, and fewer detected peaks. This suggests that it may be best to obtain higher quality data rather than larger quantities of less-refined data.

Although the developed method was successful, opportunities for improvement remain. Parameter optimization was performed on one data set, and these identical parameters were used for all other data sets. Parameter optimization on each data set, or even on each spectrum, could provide even better results if over-fitting can be avoided. Additionally, some peptides are difficult to deconvolve due to having similar fragment isotopic distributions. In these cases, it may be better to not attempt deconvolution.

7.1.3 Data acquisition simulation

Mass spectrometers require data acquisition strategies to perform sample analysis. Data-dependent and data-independent acquisition are the most common approaches used, but many variations exist for each of these strategies. Developing and testing novel strategies requires programmatic control of a mass spectrometer, yet few manufacturers provide an application programming interface (API). The simulator I developed and described in Chapter 5 is the first simulator designed for and capable of simulating custom data acquisition methods. Separate simulation of data acquisition and ground truth generation allows different acquisition strategies to be evaluated on identical data without re-generation of ground truth data. Furthermore, since most ground truth data are only needed once per simulation, its storage on disk and on-demand memory loading provides a low memory footprint. Pruning strategies effectively limit the number of peptides included in the simulation by only including those likely to be detected. A case study showed that a simulated TopN strategy, which targets the most intense peaks for fragmentation, results in similar peptide and protein identifications as a real experiment that used TopN.

The next generation of acquisition algorithms will likely further integrate data generation with data analysis, such as real-time peptide sequencing and protein inference. They may also become more goal-oriented, seeking to identify subsets of proteins, specific modifications, or improved quantification. Though simulations require many assumptions and cannot replace experimental validation, the acquisition simulator will be a great tool for evaluating new methods prior to the laborious implementation on a mass spectrometer.

7.1.4 Protein-protein interaction prediction

A major application of MS-based proteomics is the high-throughput elucidation of protein-protein interactions via APMS. The high rate of false-positive identifications have been recognized as problematic, and the subsequent failures during low-throughput validation experiments are exceedingly expensive with respect to both time and finances. Appreciating the need for an accessible and accurate APMS scoring approach, I developed a new computational tool, Spotlite, capable of classifying candidate protein pairs as *bona fide* interactions or contaminants. Importantly, I deployed Spotlite through a web-based application that provides open access and transparency to any interested scientist. This web application employs novel logistic regression classifiers that integrate existing, proven APMS scoring approaches, gene co-expression patterns, protein domains, homologous interactions, and the semantic similarity of functional annotations.

Spotlite outperforms existing APMS scoring methods by an average AUC of 16%. The implementation and enhancement of popular APMS scoring approaches provides researchers the freedom to use the most appropriate method for their particular study. Inclusion of orthologous data as features within Spotlite's logistic regression model not only increases prediction accuracy but also provides valuable information regarding shared biological function, phenotypic, and disease-relevant relationships among protein pairs. Since its publication, both the Major lab and other research laboratories have successfully identified *bona fide* interactions using Spotlite.

Though cross-validation showed improved sensitivity and specificity over other scoring approaches, Spotlite's classification accuracy is limited by the training data set. The small number of known protein interactions leads to a small positive data set, and the lack of validated non-interactions creates a negative data set with mislabeled instances. Furthermore, many features lacked high coverage, which necessitated error-prone missing value imputation. Spotlite is expected to continuously improve as increased feature coverage and more accurate training data become available to re-train the classifiers.

7.2 Open questions and future work

The work presented in this dissertation lends itself to improvement, extension, and future projects. Here, the possibilities are briefly discussed.

7.2.1 Isotope distributions

As described in Chapters 3 and 4, approximate fragment isotope distributions have multiple applications. However, the work I described in this dissertation was limited to nominal isotope distributions. Nominal isotope distributions assume that isotopes with the same nominal number of neutrons contribute to the same *m/z* peak. While this is true for low- and high-resolution mass analyzers, ultra-high resolution mass analyzers have recently been introduced. These instruments can distinguish between the same nominal isotopes of different elements such as carbon, nitrogen, oxygen, and hydrogen. Due to differences in the nuclear energy necessary to keep an atom nuclei intact, a slightly different amount of mass is converted to energy for each isotope—element combination, resulting in slightly different masses. As a result, these fine isotope distributions contain more information about a molecule's elemental composition. The accuracy and utility of approximate fine isotope distributions has not been evaluated thus far and is an avenue for future research.

Additionally, the equations I developed for fragment isotope distributions are limited to fragments resulting from a single round of isolation and fragmentation. These peaks are observed in MS2 scans. However, modern instruments are capable of MS3 scans in which further rounds of isolation, fragmentation, and mass analysis are performed on peaks found in an MS2 spectrum. These MS3 spectra have different isotopic distributions whose theoretical equations have not yet been derived, though the distributions have the same uses as described in Chapter 3.

Finally, the assumption that isolation efficiency is uniform within the bounds of an isolation window and that nothing is isolated from outside of the window is not entirely accurate. In reality, isolation efficiency depends on isolation width, the isolation *m/z* center, and the *m/z* position relative to the center of isolation window. Modeling and compensating for this non-uniformity will improve the accuracy of fragment isotope distribution predictions.

7.2.2 Dynamic isolation windows

Currently, isolation window parameters for MS2 scans are defined at the beginning of an experiment and do not change over the course of the experiment. Ideally, the isolation window's width and offset would be adjusted in real time to meet user-defined objectives based on observed data. One potential objective is to maximize isolation purity in order to avoid chimeric spectra; this can be thought of as a variation of the maximum sub-array problem. However, as demonstrated in Chapter 4, chimeric spectra can be deconvolved and sequenced when a different subset of isotopes are isolated for distinct peptides. Therefore, an alternative strategy is to make deconvolution of chimeric spectra more effective. Isolation windows can be chosen that maximize the difference between fragment isotope distributions belonging to different peptides by adjusting which precursor isotopes were isolated. Furthermore, the isolation of only the monoisotopic peak can be avoided, which is a problematic case for the NNLS model. Determining dynamic isolation windows requires extremely fast computation due to the real-time constraints of operating on a mass spectrometer.

7.2.3 Chimeric spectra sequencing pipeline

Chimeric spectra deconvolution using fragment isotope distributions is not always possible. When distinct peptides have nearly identical mass and the same isotopes are isolated, the isotopic distributions will be indistinguishable. Unfortunately, this is a common occurrence in complex samples. For such cases, iterations of sequence identification and subsequent subtraction of matching peaks allows multiple peptides to be identified from a single spectrum. This approach is completely independent of the deconvolution approach using the NNLS model, and the two approaches can be integrated into a pipeline. Chimeric spectra can be initially deconvolved using the NNLS model, followed by iterations of identification and subtraction. The deconvolution step may even make the iterative steps more likely to succeed due to having fewer interfering signals that result in lower sequence identification rates.

7.2.4 Data acquisition strategies

The dominant data acquisition strategy is to select the N most intense peaks observed in an MS1 scan for isolation and fragmentation. It has been shown, however, that signal intensity is a poor predictor of successful sequence identification. A better indicator is isolation purity. Instead of targeting the N most intense peaks within an MS1 scan, a combination of intensity and purity can be used to prioritize the ions for fragmentation and identification.

Furthermore, when the goal of an experiment is to maximize protein identifications, many peptides provide redundant data. If a protein is already confidently identified, sequencing another peptide belonging to that protein provides little information gain. Rather than prioritizing peptide targets with the highest likelihood of successful sequence identification, targets can be chosen that have the most mutual information with the current set of protein identifications. Such a strategy will prefer targets that are not only likely to be identified, but that also lead to new protein identifications. This approach would also necessitate the development of real-time, online protein inference algorithms.

Finally, real-time sequencing algorithms can be used to drive data acquisition decisions. Low probability peptide-spectrum-matches can be improved by the isolation and fragmentation of fragments observed in the MS2 spectrum. These MS3 scans can provide the AA sequence identity of the fragment, helping to break ties between two possible peptides or increasing the confidence of a low scoring match. Real-time sequencing algorithms are needed to guide the optimal selection of targets for MS3 scans. Novel sequencing algorithms will need to be developed that can integrate data from multiple scans with different levels of fragmentation.

7.3 Closing remarks

In its current state, MS-based proteomics has many opportunities for improvement, especially via computational techniques. These opportunities range from basic signal analysis to instrument operation and post-processing techniques. Integration of data analysis and data acquisition is a relatively untapped field that will likely play a large role in the near future. The work presented here contributes to the beginning of this transition and to fundamental data analysis techniques.

BIBLIOGRAPHY

Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., Costello, C. E., Cravatt, B. F., Fenselau, C., Garcia, B. A., and et al. (2018). How many human proteoforms are there? *Nature Chemical Biology*, 14(3):206214.

Anderson, N. L. and Anderson, N. G. (2002). The human plasma proteome. *Molecular & Cellular Proteomics*, 1(11):845867.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29.

Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature biotechnology*, 20(10):991–997.

Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nature biotechnology*, 22(1):78–85.

Behrends, C., Sowa, M. E., Gygi, S. P., and Harper, J. W. (2010). Network organization of the human autophagy system. *Nature*, 466(7302):68–76.

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics (Oxford, England)*, 21 Suppl 1(Suppl 1):i38–46.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300.

Beyer, A., Bandyopadhyay, S., and Ideker, T. (2007). Integrating physical and genetic maps: from genomes to interaction networks. *Nature reviews. Genetics*, 8(9):699–710.

Bielow, C., Aiche, S., Andreotti, S., and Reinert, K. (2011). MSSimulator: Simulation of mass spectrometry data. *Journal of proteome research*, 10(7):2922–2929.

Bilbao, A., Varesio, E., Luban, J., Strambio-De-Castillia, C., Hopfgartner, G., Mller, M., and Lisacek, F. (2015). Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics*, 15(5-6):964980.

Brownawell, M. L. and Filippo, J. S. (1982). Simulation of chemical instrumentation. ii: A program for the synthesis of mass spectral isotopic abundances. *Journal of Chemical Education*, 59(8):663.

Carvalho, P. C., Xu, T., Han, X., Cociorva, D., Barbosa, V. C., and Yates, J. R. (2009). Yada: a tool for taking the most out of high-resolution spectra. *Bioinformatics*, 25(20):27342736.

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., and et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10):918920.

Chapman, J. D., Goodlett, D. R., and Masselon, C. D. (2013). Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrometry Reviews*, 33(6):452470.

Chen, L., Sze, S. K., and Yang, H. (2006). Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Analytical Chemistry*, 78(14):50065018.

Choi, H., Glatter, T., Gstaiger, M., and Nesvizhskii, A. I. (2012). SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *Journal of proteome research*, 11(4):2619–2624.

Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z. S., Tyers, M., Gingras, A.-C., and Nesvizhskii, A. I. (2011). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature methods*, 8(1):70–73.

Collins, S. R., Kemmeren, P., Zhao, X.-C., Greenblatt, J. F., Spencer, F., Holstege, F. C. P., Weissman, J. S., and Krogan, N. J. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP*, 6(3):439–450.

Crutchfield, C. A., Thomas, S. N., Sokoll, L. J., and Chan, D. W. (2016). Advances in mass spectrometry-based clinical biomarker discovery. *Clinical Proteomics*, 13(1).

Cullinan, S. B., Gordan, J. D., Jin, J., Harper, J. W., and Diehl, J. A. (2004). The Keap1-BTB protein is an adaptor that bridges Nrf2 to a Cul3-based E3 ligase: oxidative stress sensing by a Cul3-Keap1 ligase. *Molecular and cellular biology*, 24(19):8477–8486.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome research*, 12(10):1540–1548.

Egertson, J. D., Kuehn, A., Merrihew, G. E., Bateman, N. W., Maclean, B. X., Ting, Y. S., Canterbury, J. D., Marsh, D. M., Kellmann, M., Zabrouskov, V., and et al. (2013). Multiplexed ms/ms for improved data-independent acquisition. *Nature Methods*, 10(8):744746.

Fatou, B., Saudemont, P., Leblanc, E., Vinatier, D., Mesdag, V., Wisztorski, M., Focsa, C., Salzet, M., Ziskind, M., Fournier, I., and et al. (2016). In vivo real-time mass spectrometry for guided surgery application. *Scientific Reports*, 6(1).

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue):D808–15.

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L., Grimmond, S. M., and et al. (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genetics*, 2(4).

Furukawa, M. and Xiong, Y. (2005). BTB protein Keap1 targets antioxidant transcription factor Nrf2 for ubiquitination by the Cullin 3-Roc1 ligase. *Molecular and cellular biology*, 25(1):162–171.

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., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636.

Geiger, T., Cox, J., and Mann, M. (2010). Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation. *Molecular & Cellular Proteomics*, 9(10):22522261.

Ghavidel, F. Z., Mertens, I., Baggerman, G., Laukens, K., Burzykowski, T., and Valkenborg, D. (2014). The use of the isotopic distribution as a complementary quality metric to assess tandem mass spectra results. *Journal of Proteomics*, 98:150158.

Gilchrist, M. A., Salter, L. A., and Wagner, A. (2004). A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics (Oxford, England)*, 20(5):689–700.

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

Gorshkov, A. V., Tarasova, I. A., Evreinov, V. V., Savitski, M. M., Nielsen, M. L., Zubarev, R. A., and Gorshkov, M. V. (2006). Liquid chromatography at critical conditions: comprehensive approach to sequence-dependent retention time prediction. *Analytical chemistry*, 78(22):7770–7777.

Gorshkov, V., Hotta, S. Y. K., Verano-Braga, T., and Kjeldsen, F. (2016). Peptide *de novo* sequencing of mixture tandem mass spectra. *Proteomics*, 16(18):24702479.

Gorshkov, V., Verano-Braga, T., and Kjeldsen, F. (2015). Superquant: A data processing approach to increase quantitative proteome coverage. *Analytical Chemistry*, 87(12):63196327.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H., editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited. `http://stanford.edu/~boyd/graph_dcp.html`.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

Graumann, J., Scheltema, R. A., Zhang, Y., Cox, J., and Mann, M. (2012). A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Molecular & cellular proteomics : MCP*, 11(3):M111.013185.

Guruharsha, K. G., Rual, J.-F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D. Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K. H., Yu, C., Parsa, B., Carlson, J. W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S. P., Celniker, S. E., Obar, R. A., and Artavanis-Tsakonas, S. (2011). A protein complex network of Drosophila melanogaster. *Cell*, 147(3):690–703.

Hart, G. T., Lee, I., and Marcotte, E. R. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC bioinformatics*, 8(1):236.

Hast, B. E., Goldfarb, D., Mulvaney, K. M., Hast, M. A., Siesser, P. F., Yan, F., Hayes, D. N., and Major, M. B. (2013). Proteomic analysis of ubiquitin ligase KEAP1 reveals associated proteins that inhibit NRF2 ubiquitination. *Cancer research*, 73(7):2199–2210.

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Fong, V., Phanse, S., Babu, M., Craig, S. A., Hu, P., Wan, C., Vlasblom, J., Dar, V.-u.-N., Bezginov, A., Clark, G. W., Wu, G. C., Wodak, S. J., Tillier, E. R. M., Paccanaro, A., Marcotte, E. M., and Emili, A. (2012). A census of human soluble protein complexes. *Cell*, 150(5):1068–1081.

Hebert, A. S., Thing, C., Riley, N. M., Kwiecien, N. W., Shiskova, E., Huguet, R., Cardasis, H. L., Kuehn, A., Eliuk, S., Zabrouskov, V., and et al. (2018). Improved precursor characterization for data-dependent mass spectrometry. *Analytical Chemistry*, 90(3):23332340.

Hoopmann, M. R., Finney, G. L., and Maccoss, M. J. (2007). High-speed data reduction, feature detection, and ms/ms spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Analytical Chemistry*, 79(15):56205632.

Horn, D. M., Zubarev, R. A., and Mclafferty, F. W. (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11(4):320332.

Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N. G., and Old, W. M. (2010). Quantifying the impact of chimera ms/ms spectra on peptide identification in large-scale proteomics studies. *Journal of Proteome Research*, 9(8):41524160.

Hoyt, M. A., Totis, L., and Roberts, B. T. (1991). S. cerevisiae genes required for cell cycle arrest in response to loss of microtubule function. *Cell*, 66(3):507–517.

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., and et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505509.

Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J. R., McGovern, K. E., Clarke, S. C., Shales, M., Mercenne, G., Pache, L., Li, K., Hernandez, H., Jang, G. M., Roth, S. L., Akiva, E., Marlett, J., Stephens, M., D'Orso, I., Fernandes, J., Fahey, M., Mahon, C., O'Donoghue, A. J., Todorovic, A., Morris, J. H., Maltby, D. A., Alber, T., Cagney, G., Bushman, F. D., Young, J. A., Chanda, S. K., Sundquist, W. I., Kortemme, T., Hernandez, R. D., Craik, C. S., Burlingame, A., Sali, A., Frankel, A. D., and Krogan, N. J. (2011). Global landscape of HIV-human protein complexes. *Nature*, 481(7381):365–370.

Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC bioinformatics*, 11(1):562.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science (New York, N.Y.)*, 302(5644):449–453.

Joshi, P., Greco, T. M., Guise, A. J., Luo, Y., Yu, F., Nesvizhskii, A. I., and Cristea, I. M. (2013). The functional interactome landscape of the human histone deacetylase family. *Molecular systems biology*, 9(1):672–672.

Jurgen, C., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011). Andromeda: A peptide search engine integrated into the maxquant environment. *Journal of Proteome Research*, 10(4):17941805.

Kaufmann, A. and Walker, S. (2018). Coalescence and self-bunching observed in commercial high-resolution mass spectrometry instrumentation. *Rapid Communications in Mass Spectrometry*, 32(6):503515.

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988.

Kim, S. and Pevzner, P. A. (2014). Ms-gf makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5:5277.

Kim, W., Bennett, E. J., Huttlin, E. L., Guo, A., Li, J., Possemato, A., Sowa, M. E., Rad, R., Rush, J., Comb, M. J., Harper, J. W., and Gygi, S. P. (2011). Systematic and quantitative assessment of the ubiquitin-modified proteome. *Molecular cell*, 44(2):325–340.

Klaeger, S., Heinzlmeir, S., Wilhelm, M., Polzer, H., Vick, B., Koenig, P.-A., Reinecke, M., Ruprecht, B., Petzoldt, S., Meng, C., and et al. (2017). The target landscape of clinical kinase drugs. *Science*, 358(6367).

Koike, A. and Takagi, T. (2004). Prediction of protein-protein interaction sites using support vector machines. *Protein engineering, design & selection : PEDS*, 17(2):165–173.

Konstantinopoulos, P. A., Spentzos, D., Fountzilas, E., Francoeur, N., Sanisetty, S., Grammatikos, A. P., Hecht, J. L., and Cannistra, S. A. (2011). Keap1 mutations and Nrf2 pathway activation in epithelial ovarian cancer. *Cancer research*, 71(15):5081–5089.

Kosako, H. and Nagano, K. (2011). Quantitative phosphoproteomics strategies for understanding protein kinase-mediated signal transduction pathways. *Expert Review of Proteomics*, 8(1):8194.

Kou, Q., Wu, S., and Liu, X. (2014). A new scoring function for top-down spectral deconvolution. *BMC Genomics*, 15(1):1140.

Kryuchkov, F., Verano-Braga, T., Hansen, T. A., Sprenger, R. R., and Kjeldsen, F. (2013). Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry. *Journal of Proteome Research*, 12(7):33623371.

Lan, K. and Jorgenson, J. W. (2001). A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks. *Journal of chromatography. A*, 915(1-2):1–13.

Lawson, T. N., Weber, R. J. M., Jones, M. R., Chetwynd, A. J., Giovanny, R.-B., Guida, R. D., Viant, M. R., and Dunn, W. B. (2017). mspurity: Automated evaluation of precursor ion purity for mass spectrometry-based fragmentation in metabolomics. *Analytical Chemistry*, 89(4):24322439.

Ledvina, A. R., Savitski, M. M., Zubarev, A. R., Good, D. M., Coon, J. J., and Zubarev, R. A. (2011). Increased throughput of proteomics analysis by multiplexing high-resolution tandem mass spectra. *Analytical Chemistry*, 83(20):76517656.

Li, L., Masselon, C. D., Anderson, G. A., Paa-Toli, L., Lee, S.-W., Shen, Y., Zhao, R., Lipton, M. S., Conrads, T. P., Toli, N., and et al. (2001). High-throughput peptide identification from protein digests using data-dependent multiplexed tandem fticr mass spectrometry coupled with capillary liquid chromatography. *Analytical Chemistry*, 73(14):33123322.

Li, Q. K., Singh, A., Biswal, S., Askin, F., and Gabrielson, E. (2011). KEAP1 gene mutations and NRF2 activation are common in pulmonary papillary adenocarcinoma. *Journal of human genetics*, 56(3):230–234.

Li, R. and Murray, A. W. (1991). Feedback control of mitosis in budding yeast. *Cell*, 66(3):519–531.

Lin, N., Wu, B., Jansen, R., Gerstein, M., and Zhao, H. (2004). Information assessment on predicting protein-protein interactions. *BMC bioinformatics*, 5(1):154.

Liu, G., Zhang, J., Larsen, B., Stark, C., Breitkreutz, A., Lin, Z.-Y., Breitkreutz, B.-J., Ding, Y., Colwill, K., Pasculescu, A., Pawson, T., Wrana, J. L., Nesvizhskii, A. I., Raught, B., Tyers, M., and Gingras, A.-C. (2010). ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nature biotechnology*, 28(10):1015–1017.

Liu, H., Yang, L., Khainovski, N., Dong, M., Hall, S. C., Fisher, S. J., Biggin, M. D., Jin, J., and Witkowska, H. E. (2011). Automated iterative MS/MS acquisition: a tool for improving efficiency of protein identification using a LC-MALDI MS workflow. *Analytical chemistry*, 83(16):6286–6293.

Liu, X. (2011). Deconvolution and database search of complex tandem mass spectra of intact proteins: A combinatorial approach. *SciVee*.

Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., Kim, B.-J., Li, C., Chen, R., Li, W., Wang, Y., O'Malley, B. W., and Qin, J. (2011). Analysis of the human endogenous coregulator complexome. *Cell*, 145(5):787–799.

Marx, V. (2015). Mapping proteins with spatial proteomics. *Nature Methods*, 12(9):815819.

McDowall, M. D., Scott, M. S., and Barton, G. J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic acids research*, 37(Database issue):D651–6.

Mcilwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C. E., Diament, B., Frewen, B., Howbert, J. J., Hoopmann, M. R., Kll, L., Eng, J. K., and et al. (2014). Crux: Rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research*, 13(10):44884491.

Mechtler, K. (2016). Ms2 spectrum processor. `http://ms.imp.ac.at/?goto=ms2spectrumprocessor`.

Medzihradszky, K. F. and Chalkley, R. J. (2013). Lessons in *de novo* peptide sequencing by tandem mass spectrometry. *Mass Spectrometry Reviews*, 34(1):4363.

Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J.-P., St-Denis, N. A., Li, T., Miteva, Y. V., Hauri, S., Sardiu, M. E., Low, T. Y., Halim, V. A., Bagshaw, R. D., Hubner, N. C., Al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W. H., Goudreault, M., Lin, Z.-Y., Badillo, B. G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A. J. R., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I. M., Bennett, K. L., Washburn, M. P., Raught, B., Ewing, R. M., Gingras, A.-C., and Nesvizhskii, A. I. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature methods*, 10(8):730–736.

Michalski, A., Cox, J., and Mann, M. (2011). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lcms/ms. *Journal of Proteome Research*, 10(4):17851793.

Muscarella, L. A., Parrella, P., D'Alessandro, V., la Torre, A., Barbano, R., Fontana, A., Tancredi, A., Guarnieri, V., Balsamo, T., Coco, M., Copetti, M., Pellegrini, F., De Bonis, P., Bisceglia, M., Scaramuzzi, G., Maiello, E., Valori, V. M., Merla, G., Vendemiale, G., and Fazio, V. M. (2011). Frequent epigenetics inactivation of KEAP1 gene in non-small cell lung cancer. *Epigenetics*, 6(6):710–719.

Myers, C. L. and Troyanskaya, O. G. (2007). Context-sensitive data integration and prediction of biological networks. *Bioinformatics (Oxford, England)*, 23(17):2322–2330.

Noyce, A. B., Smith, R., Dalgleish, J., Taylor, R. M., Erb, K. C., Okuda, N., and Prince, J. T. (2013). Mspire-Simulator: LC-MS shotgun proteomic simulator for creating realistic gold standard data. *Journal of proteome research*, 12(12):5742–5749.

Obayashi, T. and Kinoshita, K. (2011). COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic acids research*, 39(Database issue):D1016–22.

Ogura, T., Tong, K. I., Mio, K., Maruyama, Y., Kurokawa, H., Sato, C., and Yamamoto, M. (2010). Keap1 is a forked-stem dimer structure with two large spheres enclosing the intervening, double glycine repeat, and C-terminal domains. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2842–2847.

Ohta, T., Iijima, K., Miyamoto, M., Nakahara, I., Tanaka, H., Ohtsuji, M., Suzuki, T., Kobayashi, A., Yokota, J., Sakiyama, T., Shibata, T., Yamamoto, M., and Hirohashi, S. (2008). Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth. *Cancer research*, 68(5):1303–1309.

Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L. J., Danila, M. I., Feng, G., and Chisholm, R. L. (2009). Annotating the human genome with Disease Ontology. *BMC genomics*, 10 Suppl 1(Suppl 1):S6.

Osmundson, E. C., Ray, D., Moore, F. E., Gao, Q., Thomsen, G. H., and Kiyokawa, H. (2008). The HECT E3 ligase Smurf2 is required for Mad2-dependent spindle assembly checkpoint. *The Journal of cell biology*, 183(2):267–277.

Padmanabhan, B., Tong, K. I., Ohta, T., Nakamura, Y., Scharlock, M., Ohtsuji, M., Kang, M.-I., Kobayashi, A., Yokoyama, S., and Yamamoto, M. (2006). Structural basis for defects of Keap1 activity provoked by its point mutations in lung cancer. *Molecular cell*, 21(5):689–700.

Park, C. Y., Klammer, A. A., Käll, L., MacCoss, M. J., and Noble, W. S. (2008). Rapid and accurate peptide identification from tandem mass spectra. *Journal of proteome research*, 7(7):3022–3027.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E. N., Falcão, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9 Suppl 5(Suppl 5):S4.

Phipson, B. and Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1):Article39.

Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M., and Nicholson, J. K. (2006). Uplc/mse; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry*, 20(13):19891994.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic acids research*, 40(Database issue):D290–301.

Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500.

Qiu, J. and Noble, W. S. (2008). Predicting co-complexed protein pairs from heterogeneous data. *PLoS computational biology*, 4(4):e1000054.

Ramaley, L. and Herrera, L. C. (2008). Software for the calculation of isotope patterns in tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 22(17):27072714.

Renard, B. Y., Kirchner, M., Steen, H., Steen, J. A., and Hamprecht, F. A. (2008). Nitpick: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9(1):355.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *arXiv.org*.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–615.

Rockwood, A. L. and Haimi, P. (2006). Efficient calculation of accurate masses of isotopic peaks. *Journal of the American Society for Mass Spectrometry*, 17(3):415–419.

Rockwood, A. L., Kushnir, M. M., and Nelson, G. J. (2003). Dissociation of individual isotopic peaks: predicting isotopic distributions of product ions in msn. *Journal of the American Society for Mass Spectrometry*, 14(4):311322.

Rockwood, A. L., Orden, S. L. V., and Smith, R. D. (1995). Rapid calculation of isotope distributions. *Analytical Chemistry*, 67(15):26992704.

Rockwood, A. L. and Palmblad, M. (2013). *Isotopic Distributions*, chapter 3, pages 65–99. Humana Press.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A.-R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., MacWilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J.-C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A.-L., Iakoucheva, L. M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P., and Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226.

Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., and et al. (2016). Openms: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, 13(9):741748.

Rudomin, E. L., Carr, S. A., and Jaffe, J. D. (2009). Directed sample interrogation utilizing an accurate mass exclusion-based data-dependent acquisition strategy (AMEx). *Journal of proteome research*, 8(6):3154–3160.

Samuelsson, J., Dalevi, D., Levander, F., and Rognvaldsson, T. (2004). Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics*, 20(18):36283635.

Sardiu, M. E., Cai, Y., Jin, J., Swanson, S. K., Conaway, R. C., Conaway, J. W., Florens, L., and Washburn, M. P. (2008). Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5):1454–1459.

Satoh, H., Moriguchi, T., Taguchi, K., Takai, J., Maher, J. M., Suzuki, T., Winnard, P. T., Raman, V., Ebina, M., Nukiwa, T., and Yamamoto, M. (2010). Nrf2-deficiency creates a responsive microenvironment for metastasis to the lung. *Carcinogenesis*, 31(10):1833–1843.

Scheltema, R. A., Hauschild, J.-P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., and Mann, M. (2014). The q exactive hf, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field orbitrap analyzer. *Molecular and Cellular Proteomics*, 13(12):36983708.

Scherl, A., Francois, P., Converset, V., Bento, M., Burgess, J. A., Sanchez, J.-C., Hochstrasser, D. F., Schrenzel, J., and Corthals, G. L. (2004). Nonredundant mass spectrometry: a strategy to integrate mass spectrometry acquisition and analysis. *Proteomics*, 4(4):917–927.

Schulz-Trieglaff, O., Pfeifer, N., Gröpl, C., Kohlbacher, O., and Reinert, K. (2008). LC-MSsim–a simulation software for liquid chromatography mass spectrometry data. *BMC bioinformatics*, 9(1):423.

Schvartzman, J.-M., Duijf, P. H. G., Sotillo, R., Coker, C., and Benezra, R. (2011). Mad2 is a critical mediator of the chromosome instability observed upon Rb and p53 pathway inhibition. *Cancer cell*, 19(6):701–714.

Senko, M. W., Beu, S. C., and Mclaffertycor, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229233.

Serang, O., MacCoss, M. J., and Noble, W. S. (2010). Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research*, 9(10):5346–5357.

Shteynberg, D., Mendoza, L., Hoopmann, M. R., Sun, Z., Schmidt, F., Deutsch, E. W., and Moritz, R. L. (2015). respect: Software for identification of high and low abundance ion species in chimeric tandem mass spectra. *Journal of The American Society for Mass Spectrometry*, 26(11):18371847.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231245.

Singh, A., Misra, V., Thimmulappa, R. K., Lee, H., Ames, S., Hoque, M. O., Herman, J. G., Baylin, S. B., Sidransky, D., Gabrielson, E., Brock, M. V., and Biswal, S. (2006). Dysfunctional KEAP1-NRF2 interaction in non-small-cell lung cancer. *PLoS medicine*, 3(10):e420.

Slawski, M., Hussong, R., Tholey, A., Jakoby, T., Gregorius, B., Hildebrandt, A., and Hein, M. (2012). Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinformatics*, 13(1):291.

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D., and Ruepp, A. (2010). The Negatome database: a reference set of non-interacting protein pairs. *Nucleic acids research*, 38(Database issue):D540–4.

Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley interdisciplinary reviews. Systems biology and medicine*, 1(3):390–399.

Smith, R. and Prince, J. T. (2015). JAMSS: proteomics mass spectrometry simulation in Java. *Bioinformatics (Oxford, England)*, 31(5):791–793.

Solis, L. M., Behrens, C., Dong, W., Suraokar, M., Ozburn, N. C., Moran, C. A., Corvalan, A. H., Biswal, S., Swisher, S. G., Bekele, B. N., Minna, J. D., Stewart, D. J., and Wistuba, I. I. (2010). Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 16(14):3743–3753.

Sotillo, R., Hernando, E., Díaz-Rodríguez, E., Teruya-Feldstein, J., Cordón-Cardo, C., Lowe, S. W., and Benezra, R. (2007). Mad2 overexpression promotes aneuploidy and tumorigenesis in mice. *Cancer cell*, 11(1):9–23.

Sowa, M. E., Bennett, E. J., Gygi, S. P., and Harper, J. W. (2009). Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403.

Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., and Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic acids research*, 39(Database issue):D698–704.

Sykiotis, G. P. and Bohmann, D. (2010). Stress-activated cap'n'collar transcription factors in aging and human disease. *Science signaling*, 3(112):re3–re3.

Takahashi, T., Sonobe, M., Menju, T., Nakayama, E., Mino, N., Iwakiri, S., Nagai, S., Sato, K., Miyahara, R., Okubo, K., Hirata, T., Date, H., and Wada, H. (2010). Mutations in Keap1 are a potential prognostic factor in resected non-small cell lung cancer. *Journal of surgical oncology*, 101(6):500–506.

Teo, G., Liu, G., Zhang, J., Nesvizhskii, A. I., Gingras, A.-C., and Choi, H. (2014). SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *Journal of proteomics*, 100:37–43.

The, M., Maccoss, M. J., Noble, W. S., and Käll, L. (2016). Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *Journal of The American Society for Mass Spectrometry*, 27(11):17191727.

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database : the journal of biological databases and curation*, 2010(0):baq023–baq023.

TüTüNcü, R. H., Toh, K. C., and Todd, M. J. (2003). Solving semidefinite-quadratic-linear programs using sdpt3. *Mathematical Programming*, 95(2):189217.

UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, 40(Database issue):D71–5.

Uniprot Consortium, T. (2018). Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 46(5):26992699.

Valkenborg, D., Jansen, I., and Burzykowski, T. (2008). A model-based method for the prediction of the isotopic distribution of peptides. *Journal of the American Society for Mass Spectrometry*, 19(5):703712.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A.-S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabási, A.-L., and Vidal, M. (2009). An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90.

Verheggen, K., Raeder, H., Berven, F. S., Martens, L., Barsnes, H., and Vaudel, M. (2017). Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews*.

Vizcano, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., and et al. (2016). 2016 update of the pride database and its related tools. *Nucleic Acids Research*, 44(22):1103311033.

Wang, J., Bourne, P. E., and Bandeira, N. (2011). Peptide identification by database search of mixture tandem mass spectra. *Molecular & Cellular Proteomics*, 10(12).

Wang, J., Prez-Santiago, J., Katz, J. E., Mallick, P., and Bandeira, N. (2010). Peptide identification from mixture tandem mass spectra. *Molecular & Cellular Proteomics*, 9(7):1476–1485.

Weisbrod, C. R., Eng, J. K., Hoopmann, M. R., Baker, T., and Bruce, J. E. (2012). Accurate peptide fragment mass analysis: Multiplexed peptide identification and quantification. *Journal of Proteome Research*, 11(3):16211632.

Wu, L. and Han, D. K. (2006). Overcoming the dynamic range problem in mass spectrometry-based shotgun proteomics. *Expert Review of Proteomics*, 3(6):611619.

Xiao, K., Yu, F., Fang, H., Xue, B., Liu, Y., and Tian, Z. (2015). Accurate and efficient resolution of overlapping isotopic envelopes in protein tandem mass spectra. *Scientific Reports*, 5:14755.

Yan, Y., Kusalik, A., and Wu, F.-X. (2015). Recent developments in computational methods for *de novo* peptide sequencing from tandem mass spectrometry (ms/ms). *Protein & Peptide Letters*, 22(11):983991.

Yang, H., Nepusz, T., and Paccanaro, A. (2012). Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics (Oxford, England)*, 28(10):1383–1389.

Yergey, J. A. (1983). A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, 52(2-3):337349.

Yuan, Z., Shi, J., Lin, W., Chen, B., and Wu, F.-X. (2011). Features-based deisotoping method for tandem mass spectra. *Advances in Bioinformatics*, 2011:112.

Zabrouskov, V., Senko, M. W., Du, Y., Leduc, R. D., and Kelleher, N. L. (2005). New and automated msn approaches for top-down identification of modified proteins. *Journal of the American Society for Mass Spectrometry*, 16(12):20272038.

Zerck, A., Nordhoff, E., Lehrach, H., and Reinert, K. (2013). Optimal precursor ion selection for LC-MALDI MS/MS. *BMC bioinformatics*, 14(1):56.

Zhang, B., Park, B.-H., Karpinets, T., and Samatova, N. F. (2008). From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics (Oxford, England)*, 24(7):979–986.

Zhang, B., Pirmoradian, M., Chernobrovkin, A., and Zubarev, R. A. (2014). Demix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry. *Molecular & Cellular Proteomics*, 13(11):32113223.

Zhang, D. D., Lo, S.-C., Cross, J. V., Templeton, D. J., and Hannink, M. (2004). Keap1 is a redox-regulated substrate adaptor protein for a Cul3-dependent ubiquitin ligase complex. *Molecular and cellular biology*, 24(24):10941–10953.

Zhang, N., Li, X.-J., Ye, M., Pan, S., Schwikowski, B., and Aebersold, R. (2005). Probidtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics*, 5(16):40964106.