2019

Algorithms for integrated analysis of glycomics and glycoproteomics by LC-MS/MS

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

ALGORITHMS FOR INTEGRATED ANALYSIS OF GLYCOMICS AND GLYCOPROTEOMICS BY LC-MS/MS

by

JOSHUA A. KLEIN

B.A., Vassar College, 2012

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved By:

First Reader: ____
Joseph Zaia, Ph.D.
Professor of Biochemistry

Second Reader: ____
Luis Carvahlo, Ph.D.
Associate Professor of Mathematics and Statistics

Acknowledgments

I have insufficient superlatives to describe my gratitude to the people who supported me during my pre-doctoral training and development. This work would not have been possible otherwise.

First, I must thank my advisors, Professors Joseph Zaia and Luis Carvahlo. I would not have known there was a problem to solve without Joe to explain that there was one in pictures. It was then to Luis to help keep me pointed roughly in the direction of a solution. They provided me with guidance, while letting me have the freedom to explore new problems. They both showed tremendous patience as I worked to translate their respective fields, and were kind enough to keep things from getting lost in the process.

The Program for Bioinformatics has been greatly supportive, from fostering a close community among its graduate students and protecting us from university paperwork deadlines, to providing opportunities to develop new skills by attending (and teaching) workshops. To my classmates, who suffered my lack of ability to discuss anything not on a computer or sufficiently close to reality, I apologize. I would like to thank Professor Gary Benson who, in addition to chairing my thesis committee, had he not responded to my email asking if I was qualified to apply to the Ph.D. program seven years ago, would have saved a lot of people a lot of work.

I am also indebted to my colleagues from the Center for Biomedical Mass Spectrometry. In particular, I am grateful to Dr. Kshitij Khatri for teaching me about mass spectrometry and glycans as I joined the group, and Christian Heckendorf for giving me computational advice when I needed it. Nothing would have been possible without the mountains of data that Dr. Kshitij Khatri, Dr. Kevin Chandler, Dr. Le Meng, Dr. Chun Shao, Deborah Chang, and Rekha Raghunathan provided me.

I must also thank my family. My parents, Joel, whose fault this is, and Andrea

Klein, for being understanding and tolerant of this fact. I am grateful for their love and support.

But all of this is nothing compared to the debt I owe to my wife, Lily Pytel. Against all rationality, she decided to stay in this country and move to Boston with me at the start of this degree, and has kept both of us alive all throughout it. Without her love and support, I would not have been able to survive, let alone finish this degree. I am thankful for her understanding presence during the isolating periods of the degree, and her willingness to work around the constraints I put on us, like my cooking. Her tolerance at times for my aforementioned lack of conversation topics is appreciated, as was her help in developing more of them. I would not be half the human I am today without her help.

Algorithms for Integrated Analysis of Glycomics and Glycoproteomics by LC-MS/MS

Joshua A. Klein

Boston University, Graduate School of Arts and Sciences and College of Engineering, 2019

Major Professor: Joseph Zaia, Ph.D., Professor of Biochemistry

Abstract

The glycoproteome is an intricate and diverse component of a cell, and it plays a key role in the definition of the interface between that cell and the rest of its world. Methods for studying the glycoproteome have been developed for released glycan glycomics and site-localized bottom-up glycoproteomics using liquid chromatography-coupled mass spectrometry and tandem mass spectrometry (LC-MS/MS), which is itself a complex problem.

Algorithms for interpreting these data are necessary to be able to extract biologically meaningful information in a high throughput, automated context. Several existing solutions have been proposed but may be found lacking for larger glycopeptides, for complex samples, different experimental conditions, different instrument vendors, or even because they simply ignore fundamentals of glycobiology. I present a series of open algorithms that approach the problem from an instrument vendor neutral, cross-platform fashion to address these challenges, and integrate key concepts from the underlying biochemical context into the interpretation process.

In this work, I created a suite of deisotoping and charge state deconvolution algorithms for processing raw mass spectra at an LC scale from a variety of instrument types. These tools performed better than previously published algorithms by enforcing the underlying chemical model more strictly, while maintaining a higher degree of signal fidelity. From this summarized, vendor-normalized data, I composed a set of algorithms for interpreting glycan profiling experiments that can be used to quantify glycan expression. From this I constructed a graphical method to model the active biosynthetic pathways of the sample glycome and dig deeper into those signals than would be possible from the raw data alone. Lastly, I created a glycopeptide database search engine from these components which is capable of identifying the widest array of glycosylation types available, and demonstrate a learning algorithm which can be used to tune the model to better understand the process of glycopeptide fragmentation under specific experimental conditions to outperform a simpler model by between 10% and 15%. This approach can be further augmented with sample-wide or site-specific glycome models to increase depth-of-coverage for glycoforms consistent with prior beliefs.

Contents

List of Figures

List of Tables

Acronyms

AGP α-1 Acid Glycoprotein xiv, 42, 70, 71, 119, 152, 186, 192

AUC Area Under the Curve 86

CID Collision Induced Dissociation 15, 16, 28, 29, 35, 134

CS Chondroitin sulfate 10, 11

CSF Cerebrospinal Fluid xv, 154, 158, 160, 162, 165, 167

Da Dalton 70–72, 128, 131, 132, 142, 146

DS Dermatan sulfate 10

ECD Electron Capture Dissociation xvii, 17

ECM Extracellular Matrix 3

ETD Electron Transfer Dissociation xvii, 17, 28, 29

eV electron-volts 15

ExD The class of electron-based dissociation methods of which ETD and ECD are members 17, 18, 120, 194

FDR False Discovery Rate xvi, 27, 123, 132, 150–152, 170, 172, 173, 176, 178–180, 189, 191

FPR False Positive Rate 86

FTICR Fourier Transform Ion Cyclotron Resonance xiii, xxi, 13, 40–43

GAG Glycosaminoglycan 49–51, 155

GLM Generalized Linear Model 145, 163

GPSM Glycopeptide Spectrum Match 132, 151, 175

HCD Higher Energy Collisional Dissociation xv, 16, 18, 28, 29, 38, 65, 71, 120, 121, 134, 135, 137, 142, 144, 150, 157–159, 161, 166, 174, 193, 194

HPLC High Performance Liquid Chromatography 12

HS Heparin and Heparan sulfate 10, 11, 50

IAV Influenza A Virus 2, 40, 132, 133, 187

KS Keratan sulfate 10, 11, 132

LC Liquid Chromatography xviii, 18, 31, 37, 44, 175, 183

LC-MS Liquid Chromatography (LC)-Mass Spectrometry (MS) 30, 31, 37, 70, 119, 176, 183, 187, 189

LFQ Label-Free Quantification 19

m/z mass-to-charge ratio 12, 21–23, 35, 36, 40, 45–47, 50, 52, 56, 57, 61, 63, 68

MALDI matrix-assisted laser desorption/ionization 31

MRM Multiple Reaction Monitoring 33

MS Mass Spectrometry xviii, 12, 29, 31, 175

neV normalized electron volts 15

PEP Posterior Error Probability 27, 151

PMPF peptide mass prediction filter 124–126, 128

PNGase F Peptide *N*-glycosidase F 19, 20, 120

PPM Parts Per Million 22, 132, 141, 170

PSM Peptide-Spectrum Match 25–27, 123, 152

PTM post-translational modification 21, 28, 30, 123–126, 130, 182

Q-TOF Quadrupole Time-of-Flight xiii, 13, 40, 43, 70, 72, 77, 94, 110

ROC Receiver Operator Characteristic 86

SNR signal to noise ratio 23, 41, 43, 65

TDA Target Decoy Analysis xv, 27–29, 151, 179, 181

THRASH Thorough high resolution analysis of spectra by Horn 63, 67, 68

TOF Time-of-Flight 40–42, 65

TPR True Positive Rate 86

Glossary

FASTA A file format for encoding a biological sequence such as DNA, RNA, or Protein, with some descriptive information preceding the sequence on a line delimited by a ">" symbol. May also be written "Fasta" 30, 125

glycoform A particular glycan modifying a peptide, a specialization of peptidoform 21, 126

glycosite Short form of glycosylation site 20, 21, 178

HUPO-PSI The Human Proteome Organization Proteomics Standards Initiative is a steering group and oversight committee responsible for producing and maintaining open community formats and controlled vocabularies. xxi

isotopologue An ion species that differs from the monoisotopic species by one or more neutrons. Also called "isotopic peaks". 24, 43, 49, 64

macroheterogeneity A single glycoconjugate may possess multiple glycosites, each occupied by different populations of glycans. 4, 119

MGF An ASCII based format for representing collections of centroided MS^n spectra with little to no standard structure. Created by Mascot [6], which still holds a sizable market share, and its simplicity makes it very popular. xxi, 66

microheterogeneity A single glycosylation site may be occupied by many different glycans in the molecular population. 5, 119

mzML An open XML-based format for storing mass spectra and chromatograms and associated metadata, provenance, and limited workflow information. It is the current preferred format in community, and it is managed by HUPO-PSI. xxi, 66, 72

mzXML An open XML-based format for storing mass spectra developed at the Institute for Systems Biology in Seattle, Washington. It was superseded by mzML. xxi, 66

Orbitrap A Fourier Transform based mass analyzer with high mass accuracy, but different from an FTICR, using a electrostatic field instead of a high field magnet. xiii, 13, 41, 42, 66, 72, 77, 94, 132

peptide mass prediction filter A filtering technique that limits the set of theoretical glycopeptides to only those for which a peptide mass can be predicted by observing a sequence of $peptide{+}Y$ ions describing the loss of a glycan core motif. xix, 124

peptidoform A particular modified version of a peptide sequence xx, 124, 126

proteoform A single intact protein with each amino acid and its modifications fully specified as defined by [7] 119

raw file A mass spectrometer vendor's proprietary binary format used to hold the raw, unprocessed mass spectra and metadata acquired during an analysis. Each vendor uses their own format, though they periodically share file extensions. The only way to access these data are through the vendor's software, or through a driver library they provide. ProteoWizard is a suite of open source tools which includes a utility, `msconvert` that includes all of the vendors' drivers, letting the program convert raw files into open formats like mzML, mzXML or MGF. 66

reliability The posterior probability of a single peak match given its relationship with other peak matches in the same glycopeptide spectrum match. This procedure is described in Sec. 4.2.4, and the original idea was published in [8]. 156, 164

tryptic Refers to a peptide produced by cleaving a protein with the protease Trypsin 24

Chapter 1

Introduction

Glycobiology is one of the most complex branches of molecular biology, and an critical component to our understanding of modern systems biology [9]. Glycans and glycoconjugates are required for all forms of life and play an essential role in a vast number of physiological functions [9]. Glycosylation is one of the most complex and varied protein post- and co-translational modification found on proteins, modulating folding, trafficking, binding, and function [10]. Glycoconjugates are found in intracellular vesicles, on cell surfaces/glycocalyx, basement membranes, and extracellular matrices. They modulate protein physio-chemical and adhesive properties, enabling binding with lectin domain-containing partners. Glycoconjugate-lectin binding is a key part of signal transduction at many levels, impacting downstream biological processes including immunity, cancer, extracellular architecture, and differentiation [9, 11–15].

1.1 Wherein We Beg You To Care

All living cells, unicellular or otherwise, are covered in a dense and complex coat of glycans, the glycocalyx [16], and secrete glycosylated molecules into their environment. Glycosylation is conserved across whole taxonomic kingdoms [9], and there exist visible common roots between kingdoms [17]. It functions at the interface between the interior and exterior of the cell, supporting the signal transduction that allows cells to communicate with their environment. Protein glycosylation is driven by selection pressure [18], is capable of respond rapidly in response to transcriptional changes [19, 20]. Whole cell types in complex organisms have tightly coupled themselves to a specific set of glycans [21] while others undergo constant shifts in the unending adaption race against pathogens [22]. The loss of epitope patterns or whole monosaccharides in response to selective pressures has striking effects on the speciation process, even in humans [16].

Glycosylation plays many important roles in infectious diseases. Viruses such as Influenza A Virus (IAV) [23] and human immunodeficiency virus (HIV) [24–26] use glycosylation to shield immunogenic protein sites on their capsid proteins. IAV's capsid is covered in trimers of hemagluttinin, a glycoprotein that binds to sialylated glycans on the surface of host cells to initiate membrane fusion [27]. Bacteria use a diverse array of membrane bound glycoproteins, peptidoglycan and glycolipid complexes such as lipopolysaccharides (LPS) [28–30] to adaptively evade the host immune system. The ability to measure and quantify these biomolecules could help lead to better understanding of their mechanisms and how they might be controlled or negated.

Glycosylation is involved in host immunity. Immunoglobulins, or antibodies, come in many varieties. Humans and other mammals produce IgG, IgA, IgM, IgD, and IgE and all are glycoproteins with multiple sites of glycosylation [31–35]. The glycosylation at each site on these proteins influences their physio-chemical properties and binding affinities, which in turn impact their effector functions [31, 32]. This makes tracking and controlling antibody glycosylation a critical part of therapeutic antibody development and quality control [36]. Techniques for interpreting glycosylation are essential for effective antibody engineering applications, such as vaccinating against exogenous infectious diseases [37] or detecting the presence of endogenous malignant molecule or cell populations such as tumors [9, 38].

Glycosylation is altered in cancer, one of the most diverse and complex disease families we face [9, 38, 39]. Cancer is driven by cellular growth and replication, and alterations in glycosylation can dysregulate growth signals and resource consumption [40, 41]. There is a litany of different ways in which aberrant glycosylation is involved in altering the behavior of normal proteins post-translationally to proliferate or protect tumors, it's microenvironment or surrounding Extracellular Matrix (ECM) [12, 42]. To use a computing analogy, if the genetic changes to a tumor cell genome are alterations to the executing instructions of the a cell or tissue, changes to the glycosylation pattern are alterations to the run time state of that cell or tissue. Circulating tumors can be easily detected by screening for these hyper-glycosylated cells [43]. These altered patterns of glycosylation are not static, and change as the tumor cells do [44]. High throughput methods for identifying the types of glycosylation that occur on normal cells and cancer cells would be invaluable for learning how cancer circumvents normal regulatory mechanisms and how it suborns nearby tissue.

Broader assemblages that provide physical support or behaviors such as wound healing and tissue boundaries are governed in part by glycoconjugates [45, 46]. Certain types of tissue differentiation are believed to be governed by proteoglycan isoform expression [47]. The physio-mechanical properties of the extracellular matrix can affect cell organization, migration, and even gene expression [48–50]. There is evidence that proteoglycan glycosylation is involved in learning and synaptic plasticity [51, 52], and the ECM is implicated in a range of disease mechanisms in the brain [53]. The experimental techniques needed to study these molecules are just becoming practical, but still well below what is needed for clinically relevant research, and the computational methods are still to follow [54].

1.2 What Are Glycans?

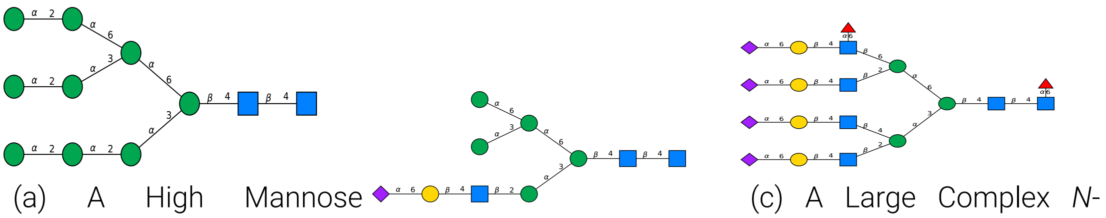
The units of interest in glycobiology are glycans, carbohydrate molecules composed of one or more monosaccharides and substituent groups, and glycoconjugates, molecular complexes with glycan and non-glycan moieties such as proteins or lipids. Glycans are enzymatically synthesized predominantly in the Endoplasmic Reticulum (ER) and the Golgi Apparatus (GA) in eukaryotes, and at the cytoplasmic membrane in prokaryotes and to a limited extent in eukaryotes as well [55]. Their biosynthesis is not template-driven, in that their structure depends upon biosynthetic reactions that are not strictly specified by a template, as is the case with proteins[10]. They are instead assembled and degraded stochastically by a range of glycoenzymes such as glycosyltransferases, which transfer monosaccharides or substituent groups onto a substrate, or glycosidases, which cleave glycosidic bonds connecting monosaccharides to one-another [56]. As mentioned before, glycans may be branched, and the degree of which is also influenced by specific enzymatic reactions. The branching pattern of a glycan influences its binding specificities with lectins and other molecules [5]. This work will discuss *N*-glycans, mucin-type *O*-glycans, and glycosaminoglycans from mammals, though there are other types of glycosylation, and notable differences between the discussed glycosylation classes in even closely related phylogenies. Glycosyltransferases and monosaccharides found active within one type of cell may not be found in other cell types of that same organism [57]. There may be marked changes in glycosylation patterns between closely related species, such as the loss of *N-glycolylneuarminic Acid* (NeuGc) in humans shortly after diverging from chimpanzees [58].

Glycoconjugates often have multiple sites of glycosylation with sub-populations with different sites occupied simultaneously, a phenomenon called macroheterogeneity. Each glycosylation site may be occupied by different glycans with varying frequency further dividing sub-populations by site-specific patterns of glycosylation, a phenomenon called microheterogeneity. The combination of macro- and micro-heterogeneity give rise to diverse and complex populations of molecules derived from a single core molecule or gene product. Small changes in glycosylation can dramatically alter binding properties [59], potentially causing a cascading change in function [42].

In genomics, where the sequence alphabet is made up of four or five nucleotide bases and in proteomics, the alphabet is made up of the twenty standard amino acids. The glycomics alphabet depends on family, genus, and species. Mammalian glycans are constructed from ten monosaccharide precursors, some of which become modified after glycosidic bond formation. Each glycosidic bond has a defined stereoconfiguration and anomericity. Additionally, glycans commonly form branching structures, making representing them as linear sequences in text difficult. Unless otherwise noted, this work will use the IUPAC [60] nomenclature for writing monosaccharide chains. For example, `b-D-Glcp2NAc` denotes β*-n-acetyl-dextro-glucopyranosamine*. Here, `b` refers to the anomericity of the monosaccharide, the orientation of its first carbon, with `b` = β and `a` = α, or `?` if unknown. The `D` refers to the molecule's chiral state which may be in *D*extro or *L*aevo configuration. `Glc` refers to the stereo-centers of the carbon backbone's orientation, where for each carbon atom the hydroxyl group or its substitution may be up or down and each arrangement has its own name, sometimes this is called the "stem" [61]. The `p` following the stem indicates the ring type of the carbon backbone *p*yranose corresponding to a five member ring and *f*uranose to a four member ring, with excess carbons forming a linear chain on either side of the ring. The subsequent notation `2NAc` indicates that at the second carbon there is a substitution of the hydroxyl group with an N-acetyl group. Many features of canonical monosaccharides are known, but most assays we will discuss cannot discriminate between their chiral states or stereo-center configurations, so these may be omitted. A mass spectrometer, no matter how sensitive, cannot determine stereo-centers, chirality, or anomericity from mass directly, and so `b-D-Glcp2NAc` would instead be written `HexNAc`, simplifying glucose to hexose, and removing any positional information from the name.

1.2.1 *N*-glycans

N-glycans are essential parts of the protein folding, quality control, and routing of secretory proteins in eukaryotes[18, 62]. They are also part of the hallmark glycosylation of many antibody and virus proteins, and play many other roles in biology [9]. The *N*-glycan synthesis process begins with a highly conserved construction of a high mannose precursor molecule with a glucose non-reducing end cap in the ER that is co-translationally attached to a protein at a conserved glycosylation sequon `/N[^P][ST]/` or sometimes `/N[^P][STC]/`, forming an amide bond between the reducing end monosaccharide and the free amine of `N`, an example cartoon is shown in Figure 1.2. After the protein transits from the ER to the Golgi apparatus, the high mannose precursor is sequentially digested by mannosidases and glucosidases, removing monosaccharides from the termini inward towards the reducing end. The glycoprotein may leave the Golgi at any point during this enzymatic trimming, or these enzymes may not be able to physically access the glycan, resulting in a “high mannose type” *N*-glycan (See Figure 1.1a). If processing successfully trims one branch of the high mannose precursor, it can begin to be extended by a series of n-acetyl glucosaminyltransferases, galactosyltransferase, fucosyltransferase, sialyltransferase, and other monosaccharide- and substituent-adding enzymes (See Figure 1.1b). These enzymes are organized spatially throughout the Golgi, and can only act on an glycan if they physically accessible and co-localized, depending upon substrate availability and transport mechanisms that are not well understood. This means that stochastic, time- and space-dependent factors are strong determinants of glycan structure. While the glycan is attached and being enzymatically transformed, the protein is simultaneously undergoing folding, which is a complex, stochastic and cooperative process in its own right, governed by thermodynamics and kinetics [63]. The concurrent processes interact with each other [64–66], and it remains a challenging problem in the field to predict how protein structure and glycosylation affect each other [23, 67]. All eukaryotic *N*-glycans share a common reducing end core structure, called the tri-mannosyl core or the chitobiose core, with the structure `a-D-Manp-(1-6)[a-D-Manp-(1-3)]b-D-Manp-(1-4)-b-D-Glcp2NAc-(1-4)b-D-Glcp2NAc`.



(a) A High Mannose *N*-glycan

(b) A Hybrid *N*-glycan

(c) A Large Complex *N*-glycan

Figure 1.1: Common Types of *N*-glycan structures drawn using SNFG [1]

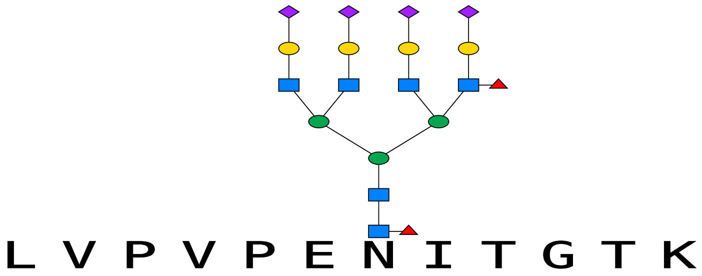


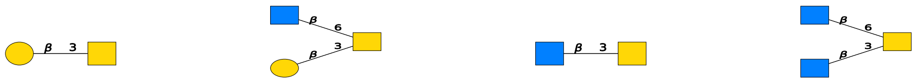
Figure 1.2: An example *N*-linked glycopeptide. The glycan is attached to the asparagine residue (position 7) at the beginning of the sequon `NIT`.

Because of the range of substrate and enzyme reaction pairs is unbounded, the set of structures found in nature is effectively unknown, though the set existing in a given biological context is far smaller than the potential combinatorial space. Many have tried to model *N*-glycan biosynthesis to understand how the glycosylation patterns of a protein or cell population change in response to a stimulus [68, 69]. These models often omit unusual but biologically relevant enzymes in order to remain computationally tractable, but can capture the common glycans for a system well. Others have pursued a broader enumeration in order to define the space of possible structures for exploratory applications [70] by removing the spatial component while adding more enzymes to their model. These models have been targeted at human or mammalian systems out of practical considerations, and it is well known that other phyla express a different panoply of glycoenzymes that still align with the *N*-glycan biosynthetic pathways in mammals [10, 17, 66].

1.2.2 Mucin-type *O*-glycans

O-Glycans are another family of glycosylation that are different than *N*-glycans in two main ways. Firstly, *O*-glycan core structures are more varied, with different classes of *O*-glycan having many different cores and epitopes. Secondly, *O*-glycans do not have a precise sequon, with many targeting any accessible serine or threonine residue, forming a glycosidic bond with a free hydroxyl group. There appears to be a preference towards sites preceded by a proline residue. One prominent class of *O*-glycan is the mucin-type, or the "O-GalNAc" *O*-glycan[71].

Mucin-type *O*-glycans have between 4 [71] and 8 [72] core patterns, depending upon the source and definition of a core motif versus an epitope, and not all are present in all species or tissues [73]. The only conserved component amongst these



(a) *O*-glycan Core 1 (b) *O*-glycan Core 2 (c) *O*-glycan Core 3 (d) *O*-glycan Core 4

Figure 1.3: Common Mucin *O*-glycan cores drawn using SNFG [1]

structures are that they all have a reducing terminal `b-D-Galp2NAc`. The four common cores are shown in Figure 1.3. These types of glycans have more varied branching patterns than *N*-glycans, though they tend not to be as large as many complex and high mannose *N*-glycans.

Mucin-type *O*-glycans get their name from mucins, a family of glycoproteins with hundreds of O-GalNAc glycans attached along their length at dense regions of repeated serine, threonine and prolines created by variable number tandem repeats (VNTR) [74, 75]. These dense regions of glycosylation have marked effects on the glycoprotein's 3D structure and give them a "bottle-brush" like appearance [71, 76]. Many other glycoproteins and proteoglycans exhibit this type of *O*-glycosylation, though to a lesser degree [45, 54].

Mucin-type *O*-glycan synthesis happens entirely in the Golgi, without any component in the ER as in *N*-glycans. There are many glycosyltransferases for attaching `b-D-Galp2NAc` to a serine or threonine, but they each have broader sequence specificities. There are also cooperative effects where one glycosyltransferase will bind to an already attached glycan to increase their efficiency [71]. There are glycosyltransferases dedicated to extending the initial `b-D-Galp2NAc` into the core motifs, and for building up the initial branching patterns, but after this first initial extension, many enzymes share specificity between *N*-glycans and mucin-type *O*-glycans. Just as in *N*-glycans, the degree to which an *O*-glycan is extended depends upon the duration the protein spends in each compartment of the Golgi, the abundance and activity of the localized glycosyltransferases, and the accessibility of the glycosylation sites. As in *N*-glycans, so too is the full range of possible *O*-glycans unbounded and the true number of possible structures unknown. Less work has been done to predict the space of possibilities as well, though some limited work has been done [77].

1.2.3 Glycosaminoglycans

> "Without glycosaminoglycans, we'd all be boring old house plants" - *Joseph Zaia*

The third and final class of glycan to be discussed in this work is glycosaminoglycans (GAGs), formerly called "mucopolysaccharides". As the name suggests, these glycans carry an abundance of amines, though these groups are often modified by acetyl or sulfate groups. GAGs are linear carbohydrate chains which are composed of alternating disaccharide repeats of the form `Hex*-Hex2N*`where the stem type that replaces the `Hex` depends upon the family of GAG [78]. Chondroitin sulfate (CS)'s disaccharide pair is `-(1-4)b-D-Glcp6A-(1-3)b-D-Galp2NAc` with optional sulfate groups at the 4 and 6 positions of the `GalNAc`. Dermatan sulfate (DS) is a variant of CS where the `Glcp6A` may be replaced with `Idop6A`. Heparin and Heparan sulfate (HS)'s pattern is more complex, having the disaccharide `-(1-4)a-D-Idop6A-(1-4)-a-D-Glcp2N` with optional sulfate groups at the 2 position of `IdoA` and optional sulfate groups at the 3 and 6 position of the `GlcN`, and variable acetylation or sulfation of the amine of the `GlcN` at the 2 position. Keratan sulfate (KS)'s disaccharide is `-(1-3)-b-D-Galp-(1-4)b-D-Glcp2NAc`, with optional sulfate groups at the 6 position of both monosaccharides, resembling a variably sulfated lactosamine repeat. Hyaluronan is an unsulfated GAG that is synthesized at the plasma membrane of eukaryotic cells and cytoplasmic membrane of prokaryotic cells [79]. HS and CS/DS are both connected to a protein by a common linker tetrasaccharide ?-?-6-a-Glcp-(1-3)b-D-Galp-(1-3)b-D-Galp-(1-4)b-D-Xylp, with the disaccharide repeat beginning at the non-reducing glucurionic acid, with the CS linker shown in Figure 1.4. KS does not have a linker, instead it occurs on otherwise normal *N*- and *O*-glycans through an as yet poorly understood mechanism where by extended lactosamine units are sulfated in certain tissue types.

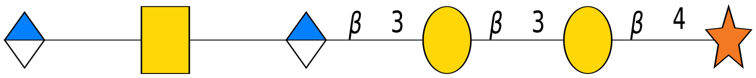


Figure 1.4: Chondroitin Sulfate Linker Saccharide

While there are hundreds if not thousands of glycoproteins in the proteome which carry *N*- and *O*-glycosylation, there are only a few dozen that are GAGylated [78]. HS and CS share the same linker biosynthetic process up to the first GlcA residue. There is some evidence suggesting that GAGylation has a target motif S[GA] and efficiency will depend upon the distribution of properties of nearby amino acids including hydrophobicity, acidity, and nearby glycosylation sites [80]. There is also some evidence that the second residue is strictly required [54]. The GAGylation process starts by transferring a xylose residue onto the serine, forming a glycosidic bond with a free oxygen, and making HS and CS GAGylation another class of *O*-glycosylation.

HS GAGs are polymerized to variable lengths, and depending upon conditions may be as many as 40 monosaccharides long, longer in some cases [78, 79]. There are alternating domains of high, low, and no sulfation which tune the binding specificities of the GAG chain, and the proteoglycan they are attached to [81]. Techniques for sequencing GAGs are complicated by the ease with which sulfate groups are lost during analysis [82–84]. Because of their size, it is often difficult to study intact GAG chains directly, so they must be summarized by first enzymatically digesting them into smaller pieces prior to analysis. Different enzymes are necessary for different types of GAG such as chondroitinase for CS or heparinase for HS.

1.3 Analytical Chemistry Tools For Studying Glycans and Glycoconjugates

There are many ways to study the structure and function of glycans and glycoconjugates such as glycoproteins. Techniques involving glycan-binding molecules such as lectins, endo- and exoglycosidases produce low throughput measures of a glycan or glycoconjugate's form or function [85]. High throughput methods for studying these topics predominantly involve Mass Spectrometry (MS), and High Performance Liquid Chromatography (HPLC) and their related technologies. This work will focus on applications of mass spectrometry and tandem mass spectrometry coupled with liquid chromatography.

1.3.1 Mass Spectrometry

A mass spectrometer is a device that measures the exact mass-to-charge ratio (m/z) of molecules in a sample, and report on their relative abundances. A mass spectrometer is composed, abstractly, of an ionization source, a mass analyzers, and a detector.

The ionization source influences the type of ionization the sample undergoes and the types of molecules that can be ionized. For the purposes of this work we will deal entirely with electrospray or nanoelectrospray ionization. Both methods are effective at ionizing polar molecules, and they can be tuned for smaller or larger analytes [86]. The mass analyzer is responsible for resolving ion m/z, and operates by scanning along a m/z interval, effectively allowing ions in a narrow m/z interval pass through to the detector. Different mass analyzers use very different mechanisms to accomplish this task, and have differing mass accuracy and resolution. Mass accuracy measures the fidelity of the mass measured compared to the true mass of the molecule. The resolution R of a mass analyzer is given by Eq. (1.1) where m is the m/z of a peak and Δm is the m/z difference between m and a second peak of equal height where 10% of the peak overlap [4, 87].

$$R = \frac{m}{\Delta m} \tag{1.1}$$

Naturally, for the same Δm at higher mass, greater resolution is required. With high resolution, a mass analyzer is able measure ions which have a different number of neutrons but share the same elemental composition, or "isotopologue", forming an isotopic distribution for that ion species. When isotopologues are resolvable, it is possible to infer the charge state z of an ion and convert m/z to mass by deconvolving the isotopic pattern from the m/z compression caused by increased charge state. Given an ion with mass m and a charge z and a charge carrier with mass c, the m/z is calculated using Eq. (m/z), and the original mass can be recovered given that we know the charge by Eq. (neutral mass).

$$\mathtt{m/z} = \frac{m + zc}{|z|} \tag{m/z}$$

$$m = \mathtt{m/z} * |z| - zc \tag{neutral mass}$$

The charge carrier is usually a proton, written H^+, with mass 1.00728, though metallic cations such as Na^+ and Ca^{2+}, and other charged molecules may also play this role. Charge may be positive or negative, depending upon the polarity of the instrument.

This work deals only with high resolution mass spectra from Quadrupole Time-of-Flight (Q-TOF), Orbitrap, and Fourier Transform Ion Cyclotron Resonance (FTICR) mass analyzers. The detector responds to the ion beam selected by the mass ana-

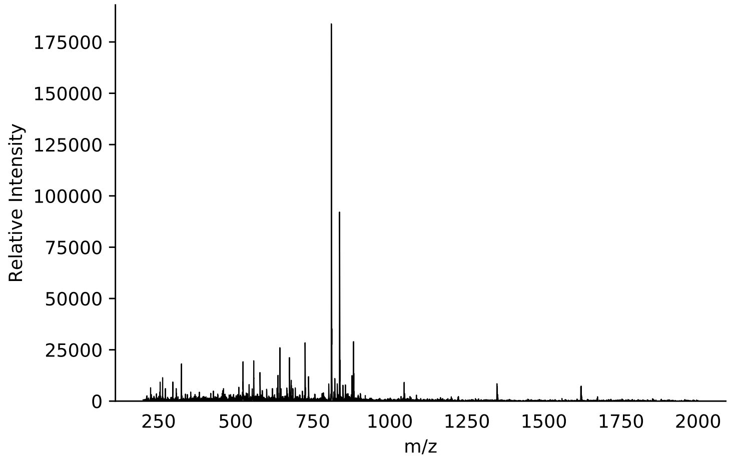
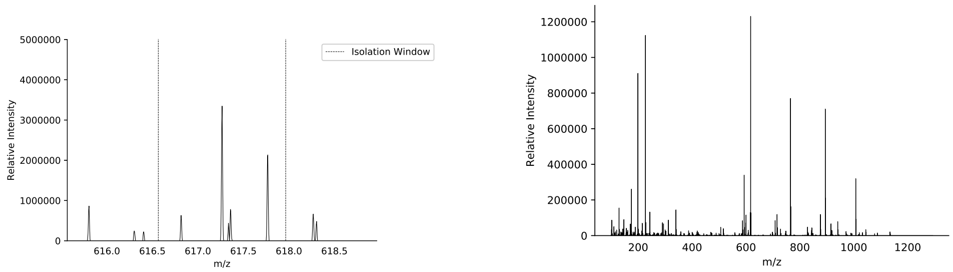


Figure 1.5: A Mass Spectrum

lyzer, measuring the abundance of that m/z interval. This measurement procedure is usually displayed as a plot of m/z by intensity, as shown in Figure 1.5. A mass spectrum may be a profile with continuous data points or centroided with discrete peaks. The process of converting from profile to centroid is called peak picking or centroiding, discussed later in this chapter.

A mass spectrometer may contain multiple mass analyzers, and they can be used individually or in tandem to isolate specific ions or m/z intervals. The selected ion may be "activated" in some way, causing it to dissociate into a new population of product ions, and then have measure the product ions. This is called a tandem mass spectrum, and an example derived is shown in Figure 1.6, derived from Figure 1.5. The selected ion is called the "precursor ion" and the spectrum the precursor ion was isolated from is called the "precursor spectrum", and the spectrum the product ions are measured in is called the "product ion spectrum". This fragmentation process is called "tandem mass spectrometry", or "MS/MS", or "MS^2". Higher degrees of exponentiation are possible and collectively they are called "MS^n".



(a) The isolation window around the ion at 617.264 m/z shown in Figure 1.5

(b) The product ions captured from the selected isolation window.

Figure 1.6: The tandem mass spectrum from the ions captured in the isolation window

1.3.2 Fragmentation Techniques

There are several common dissociation methods used for peptides, glycans and glycopeptides. Collision Induced Dissociation (CID) operates by colliding the precursor ion with a neutral gas with a certain energy, usually measured in electron-volts (eV) or normalized electron volts (neV). CID fragments peptides primarily at the amide bond junctions between C(=O)^1-N(H)^2, producing N-terminal fragments called b ions and C-terminal fragments called y ions as shown in Figure 1.7 [2, 86]. These ion types induce constant gain or loss of mass, giving them a distinct signature in m/z space . CID fragments glycans primarily at glycosidic bonds, analogously to peptides, producing B ions from non-reducing end fragments and Y ions from reducing end fragments as shown in Figure 1.8 [3]. B and Y ions may be able to define which monosaccharides make up a glycan, but they cannot determine the positions at which those monosaccharides are attached to each other, nor where on the carbon backbone a substituent group is attached. Low mass b and B ions are called *immonium* and *oxonium* ions, respectively. Under CID, the glycan component of a glycopeptide fragments preferentially, producing B ions from the non-reducing termini of the glycan, producing abundant oxonium ions and Y ions with the intact peptide attached, producing limited information about what the peptide is and where on the peptide the glycan is attached. The intact peptide-containing Y ions are sometimes called $peptide + Y$ or "stub glycopeptides".

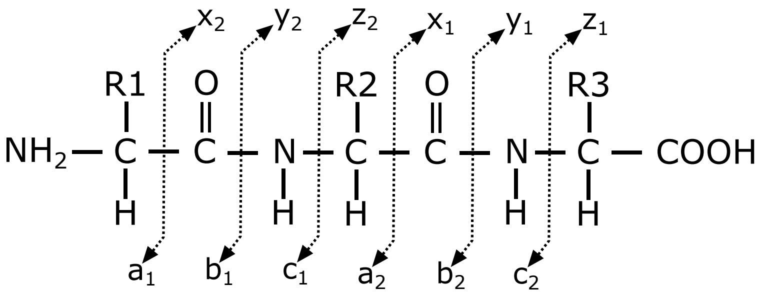


Figure 1.7: Nomenclature for Peptide Fragments [2]

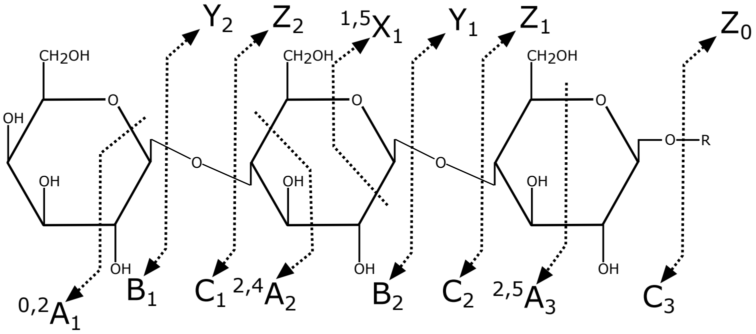


Figure 1.8: Domon and Costello Nomenclature [3] for Glycan Fragments

Higher Energy Collisional Dissociation (HCD) is similar to CID, though more energy is used. It produces the same ion types as CID, though HCD is more effective at fragmenting larger molecules. HCD fragmentation of glycopeptides more effectively dissociates the glycan moiety, and any remaining energy produces peptide backbone b and y ions with or without a small piece of the glycan reducing end still attached. These peptide backbone fragments can be used to identify the peptide sequence, and may be able to localize the site of glycosylation. The identity of the glycan at that site cannot be determined because these site-localizing fragments only retain a small component of the glycan reducing end, making identifying multiply glycosylated peptides difficult [88].

Electron Transfer Dissociation (ETD) and Electron Capture Dissociation (ECD) are techniques that destabilize a molecule by adding or removing electrons, collectively referred to as ExD. Both of these methods produce c and z ions from peptides, which are complementary to b and y ions, respectively. For glycans, ExD methods produce C and Z ions, complementary to B and Y ions, as well as abundant cross-ring fragments called A and X ions [3], depending upon whether or not a non-reducing or reducing terminal is included. These cross-ring fragments can localize bonds on the monosaccharide's carbon backbone, determining linkage. For glycopeptides, ExD preferentially produces peptide c and z ions without dissociating the glycan, allowing for site specific localization of multiple glycans on the same peptide [86, 88, 89]. ExD methods are much slower than collisional dissociation, and require a higher charge state ion to produce abundant fragment ions [90]. The reaction may lead to a bond breaking but the fragment ions remaining associated in the gas phase, called "ETnoD". This can be avoided by mixing ExD activation with supplemental collisional activation, going by monikers including "Hot ECD" or "EThcD". For glycopeptides, these methods produce b, c, y, z, B, $peptide + Y$ and $b/y/c/z + Y$ ions in varying proportions depending upon ExD reaction time and supplemental collision energy [88]. Glycans dissociated with these techniques yield A, B, C, X, Y, and Z ions, as well as many neutral losses thereof [91]. ExD also can also produce additional radical ions which further split signal from the same bond cleavage over multiple peaks. These extra radical series are expressed by adding $\cdot$ to their name i.e. $z\cdot$.

There are other ion series that these dissociation methods produce, such as a, x, d, v, and w peptide fragments which may have diagnostic value but occur less frequently [86]. Other less common fragmentation pathways involve neutral losses from canonical fragments, such as the loss of NH_3 from a C-terminal fragment or the loss of H_2O from an N-terminal peptide fragment or from any monosaccharide fragment. There are also other dissociation methods, such as infrared multi-photon dissociation (IRMPD) and ultraviolet photon dissociation (UVPD) which produce the common fragment ions as well as other uncommon fragment series, but the required instrumentation is not commonly available.

Each fragmentation method has strengths and weaknesses, and the method used still must be calibrated for a given problem. This work will deal primarily with HCD-type dissociation methods with some variation for glycopeptides in Chapter 4, and limited coverage of HCD and ExD methods for glycans in Chapter 3.

1.3.3 Chromatography

A mass spectrometer is often coupled to a Liquid Chromatography (LC) system, or another separation system like Capillary Electrophoretic device (CE) [92]. These tools cause materials passed through them to travel at different rates depending upon one or more physical properties such as hydrophobicity for Reverse Phase Chromatography (RPC), hydrophilicity for HILIC or molecule size for Size Exclusion Chromatography (SEC), separating them in time. When combined with a mass spectrometer, the chromatography system introduces only a small fraction of the sample mixture into the mass analyzer at a time, yielding a granular view of the sample [93]. This requires the mass spectrometer operate at a speed compatible with the rate at which analytes elute, limiting the time the mass analyzer may spend on any single scan but increasing the amount of time spent analyzing a sample compared to a case where no chromatography system is used.

An analyte's abundance over chromatographic elution time is the chromatogram of that analyte, and it is often visualized as a smoothed curve, with the area under that curve being proportional to the total abundance of the analyte. This is a form of Label-Free Quantification (LFQ). The shape of this curve is called the "chromatographic peak shape", and can be used to measure whether an analyte's signal is distinguishable from noise [94–99], which can be useful when assessing experimental data where fragmentation data is unavailable or in targeted experiments where even product ions have chromatograms [98, 100].

The time of peak elution from the separation device can also be used diagnostically with known standards or a library of references [101–104]. A common technique for profiling glycans by LC-MS involves spiking in reference polymers of glucose which can be used to predict whether a another glycan identified by mass is eluting at the correct time relative to these references [105]. An equivalent technique is used in proteomics by spiking in a set of known peptides to transform retention time into an approximate normalized retention time, called `iRT` peptides [106].

1.3.4 Sample Preparation, Transformation, and Simplification

There are a multitude of different ways that glycan, peptide and glycopeptide samples can be prepared for analysis. This work will highlight a few steps that will become relevant later.

Glycan Digestion and Release

In order to measure the glycome of a glycoprotein sample, a release step is required. For *N*-glycans, there is a single endoglycosidase enzyme that works on all mammalian *N*-glycans Peptide *N*-glycosidase F (PNGase F). This enzyme cleaves the bond between the *N*-glycan reducing end and the asparagine, resulting in a deamidation of the peptide or protein, and a free *N*-glycan. PNGase F cleaves α-(1-3) fucosylated *N*-glycans such as those found in invertebrates and plants inefficiently. For *O*-glycans, there is no one enzyme that can release everything, even within the subset of mucin-type *O*-glycans. The only universal methods involve harsher procedures using reductive β-elimination [107] or hydrazinolysis [108] which chemically alter the glycan and protein.

Once a glycan is enzymatically released, its free reducing end may be modified to improve its analytical properties. These improvements may range from better chromatographic retention, fluorescence for optical detection, increased charge for better ionization, MS^n quantification and multiplexing, or asymmetric MS^n fragmentation. It may also be desirable to derivatize glycans for similar effects with common treatments, such as permethylation [109].

Both free and attached glycans may be simplified by applying exoglycosidases, such as sialidases or fucosidases to collapse multiple glycoforms into a single core glycan. This may be advantageous when the sample material is limited or too complex to interpret all individual forms simultaneously.

Analyzing released glycans costs losing site-specific information, but concentrates signal from all glycosites carrying each glycan, and allows one to fragment a glycan to get linkage-defining fragments. It can also be used to define the sample's glycome, narrowing the range of compositions one must consider when analyzing the glycoproteome. This topic will be returned to in Chapters 3 and 4.

Protein Digestion

Intact proteins are challenging to study in a "top-down" fashion [110] and glycoproteins are even more difficult due to the heterogeneity of glycosylation, so it is standard practice to digest protein mixtures with one or more proteases to produce peptide mixtures instead. The most commonly used protease is trypsin [93, 111], which cleaves at the C-terminus of arginine and lysine, or more precisely, cleaves the protein at every match to the regular expression `/[KR](?=[^P])/` [112], though there are many other popular candidates such as chymotrypsin `/(?<=[FYWL])(?!P)/`, or glutamyl endopeptidase `/E/` commonly called "Glu-C". Before digestion, it is common to alkylate cysteines to prevent them from cross-linking, usually using iodoacetamide to carbamidomethylate the cysteine side chain. Other chemical modifications may be added such as stable isotope labels or Tandem Mass Tags (TMT) for multiplexing and quantification.

If glycans were not released, the sample may contain a mixture of peptides and glycopeptides. Glycopeptides do not ionize as well as peptides [113], and can be hard to reliably detect against the background complexity of a complete proteome because microheterogeneity splits the signal for a single glycosite across multiple glycoforms. To obtain deeper coverage of the glycoproteome, an online or offline glycopeptide enrichment method can separate the glycopeptides from the peptides [114].

If glycans were released, the deglycosylated peptides can be analyzed using a traditional peptide and protein identification tool to identify the proteins and other PTMs present in the sample. This makes a future glycopeptide identification process more accurate, though potentially much more complex. This topic will be returned to in Chapter 4.

1.4 Computational Methods

At its core, computational methods for interpreting or annotating mass spectra involve transforming a mass spectrum from m/z space into neutral mass space (neutral mass), or a database of intact molecules or building blocks from theoretical neutral mass space to theoretical m/z space (m/z), and looking for close by values within an error tolerance window. There are two types of error tolerances in common use, absolute error tolerance shown in Eq. 1.2

$$|e - t| < k \tag{1.2}$$

$$\frac{|e - t|}{t} * 10^6 < k \tag{1.3}$$

where e is the experimental measure, t is the theoretical measure, and k is the error limit, and relative error tolerance such as Parts Per Million (PPM) is shown in Eq. 1.3 where the same absolute error may be accepted when t is larger. PPM is preferred for higher resolution instruments to reflect the mass-dependent nature of their measurements [4]. Common mass analyzer mass accuracies are shown in Table 1.1.

Mass Analyzer	Mass Accuracy in PPM	Resolution
Quadrupole	100	2000
Ion Trap	100	4000
TOF	200	5000
TOF (reflectron)	<10	20 000
Magnetic Sector	<10	10 000
FTICR	<5	500 000
Orbitrap	<5	100 000

Table 1.1: Mass Analyzer Accuracy and Resolution [4]

1.4.1 Signal Processing

Prior to performing any mass comparison, a profile mode spectrum must be centroided before any other comparisons are done. Peak picking converts continuous profiles into a set of discrete centroids or “peak lists”, shown in Figure 1.9. These

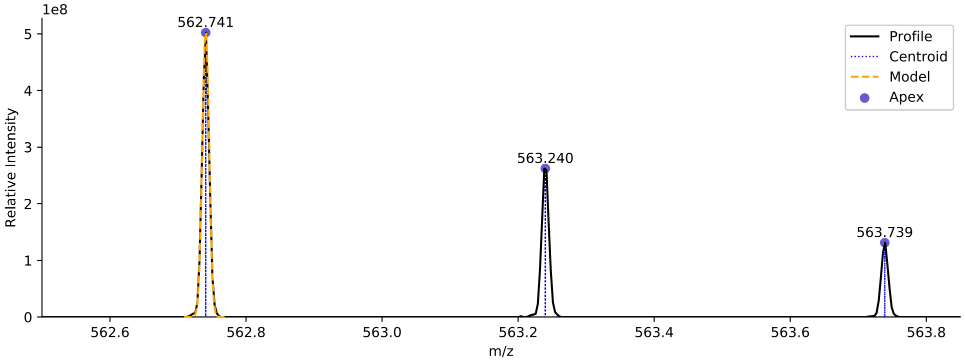


Figure 1.9: The peak picking process converting a profile mass spectrum into a centroided list of peaks with discrete m/z and intensity values. The model used was a Gaussian peak shape with $\sigma = 0.005035$

peaks represent scalar m/z values or masses which correspond to the abundance-weighted average measured value for each observed ion. Peak picking may be done in many ways, starting from simple apex selection, with intensity or signal-to-noise ratio (SNR) thresholds [115], to wavelet based methods [116, 117] and peak shape model based methods [116, 118, 119]; however many mass spectrometry-related tools assume that spectra are already centroided by instrument vendor software [120]. Other signal processing techniques are often applied concurrently, such as background noise reduction, interpolation and smoothing [103, 120, 121]. Another, more complex problem is often addressed simultaneously, which is deisotoping and charge state deconvolution, and is often referred to as peak picking, incorrectly[1] [122, 123].

1.4.2 Deisotoping and Charge State Deconvolution

It is not uncommon to assume the charge state of a small molecule or fragment ion is 1. Most high resolution instruments report the charge state of selected ions when acquiring MS^n spectra, obviating the need to estimate it directly. Instrument control

[1] People making this mistake will have the PSI-MS Controlled Vocabulary thrown at them. It is 326 pages at 12px font as of 4.1.13.

software do not guarantee that they select the monoisotopic peak, and they do not determine the charge states of all other peaks in each spectrum. A high resolution raw or merely centroided mass spectrum contains resolved isotopologues forming an isotopic pattern. For biomolecules, the most abundant isotopologues differ by approximately $1.0033/z$ m/z apart from one another, the mass difference between C^{13} and C^{12} divided by the charge state of the ion.

Simplistic deisotoping and charge state deconvolution techniques may try to determine the peak spacing with the highest charge state with the fewest missing peaks [124–126]. While this technique works for high mass accuracy instruments it is still error-prone for complex spectra. Each peak in an isotopic pattern has an abundance value based upon the isotopic abundance ratios of each element and the amount of those elements in the ion's composition. Given a chemical composition, it is possible to predict the isotopic pattern using a variety of methods for different mass resolutions [127–131], and this theoretical pattern can be compared to the data to determine both whether a peak is a member of a particular isotopic pattern at a particular charge and whether a peak is the monoisotopic peak of an isotopic pattern. Methods for measuring goodness-of-fit for isotopic patterns are discussed in greater detail in Chapter 2.

While precursor ions for typical tryptic peptides and glycans are multiply charged, their product ions are usually singly charged. By contrast, glycopeptide MS^2 spectra are often dominated by multiply charged peaks as shown in Figure 1.10, making it an essential component for glycopeptide identification [54, 132, 133].

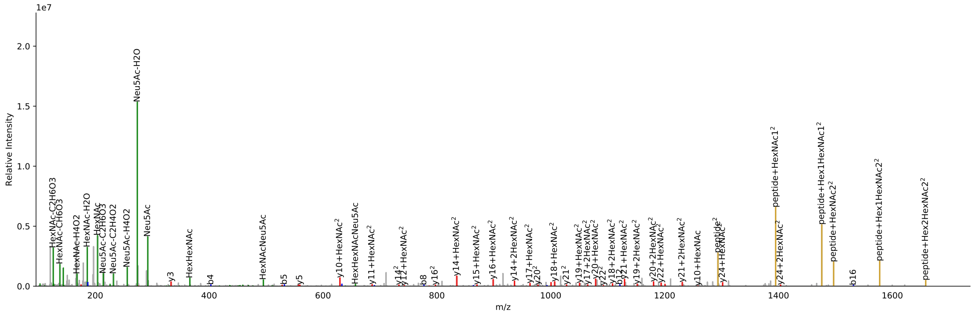


Figure 1.10: An example annotated glycopeptide MS^2 spectrum for the glycopeptide `GESEETGSSEGAPS(O-Glycosylation)LLPAT(O-Glycosylation)RAPEGTR{Fuc:1; Hex:3; HexNAc:2; Neu5Ac:3}` with many doubly charged peaks

1.4.3 Database Search Methods

Search Engines

One of the first computational tools created to try to interpret peptide mass spectra was SEQUEST [134] in 1994. It was purportedly the first peptide database search engine, and it used a heuristic theoretical spectrum alignment algorithm to evaluate a Peptide-Spectrum Match (PSM) based on the scoring function Xcorr which counted the number of common peaks in the alignment. The proteomics database search engines have diversified and multiplied since then [6, 125, 135–138], spawning new classes of search engine but they all share some common features. First, they all begin with a protein database and a set of mass spectra, either raw or converted into simplified peak lists. Second, they convert the protein database into peptides by emulating a protease, such as by splitting on a regular expression, or assume non-specific digestion to produce a set of peptides derived from those proteins, optionally modified using fixed or variable modifications from either a database of modifications like UNIMOD [139] or a user-defined list of rules. Finally, they evaluate PSMs by computing a score relating how well theoretical structure P_i matched spectrum

S_j within a specified mass error tolerance where scores are orderable between all $P_i \in P$ within the scope of S_j such that the best scoring PSM for S_j can be selected exclusively. The database search paradigm may be applied to glycans [140, 141], metabolites [142, 143], lipids [144], intact proteins [145, 146], RNA [147], cross-linked peptides [148, 149], and of course, glycopeptides [54, 132, 133, 150–153]. There are some analogs between the database search process and the more common sequence alignment problem in bioinformatics [154, 155].

The scoring functions commonly used by these algorithms are often posed as heuristic measures of goodness of fit like Xcorr [134], dot product [136] or Hyperscore [125, 138, 156, 157] as shown in Eq. 1.5

$$\texttt{dot} = \sum_i F_i I_i \tag{1.4}$$

$$\texttt{hyperscore} = \log\left(N_b! N_y! \sum_{i=1}^{N_b} I_{b,i} \sum_{i=1}^{N_y} I_{y,i}\right) \tag{1.5}$$

or take the form of a probability model expressing the probability of the spectrum S given the structure p and some likelihood or indicator functions $F(p, i)$ to denote one or more fragment ion matches [158]. In many cases, a heuristic will be posed with a probabilistic interpretation as a foundation, but the final metric will not be meaningfully interpretable as a probability [145, 159–162] such as a binomial tail probability shown in Eq. 1.6.

$$B(n, k, p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{1.6}$$

Intensity information may or may not be used directly, depending upon the scoring function. Few probability models exist for accurately predicting the probability of a peak being of a certain intensity or relative intensity, though some more general predictive models do exist [163, 164]. Many models approach the problem by instead modeling the frequency with which particular amino acids from different ion series dissociate [136, 145, 160, 165, 166].

Very few of these scoring functions are characterizable in a sufficiently rigorous manner that an error model is known *a priori*, leading to a variety of different strategies for estimating the null model at an experiment or spectrum level [136, 167–170]. Many search engines attempted to estimate the relative uncertainty of a PSM by calculating an expectation-value for each PSM's score from the distribution of scores measured for that search [6, 156]. OMSSA [159] modeled its expectation approach after the method used by BLAST [171]. PeptideProphet [169, 172] used a large corpus of training data with known correct ($T_i = 1$) and incorrect ($T_i = 0$) labels and k scoring functions evaluated on each PSM, the charge of the precursor ion, and a set of structure features s for each sequence to learn a mixture model-based discriminant function $F(x_1, ..., x_k)$ that combines these scores in a way. This model could then be used to estimate the Posterior Error Probability (PEP), and in turn False Discovery Rate (FDR). While this technique was successful, it had to be trained anew for each new class of scoring function or dissociation technique, and as with many ensemble methods, as $k \rightarrow \infty$ it grows less interpretable. Target Decoy Analysis (TDA) [170] obviated the need for labeled data by searching both a "target" database of the proteins of interest and a "decoy" database of false but valid sequences that are known to not be in the sample, and used the proportion of decoys above a score threshold to estimate the FDR, though only under asymptotic conditions with many, many PSMs. Percolator went on to incorporate a semi-supervised learning strategy for using additional arbitrary sequence-level features into TDA. Many approaches treated TDA as a black box and violate its assumptions, limiting its efficacy as a rigorous FDR control [173], and it remains an experiment-wide feature. In MS-GF+ [136], a brute-force dynamic programming approach to efficiently calculated the scores of every "reasonable" candidate for a single spectrum to estimate a spectrum level E-value, though the community still prefers to use it in conjunction with TDA [124].

Many database search engines and scoring techniques have been designed for emphasis on PTM discovery, and relax or alter the requirements for identifying modifications [137, 138, 161, 174, 175]. Some augment an existing database search algorithm with a modification localization auxiliary scoring criterion [176–179]. Others introduce multiple rounds of search to refine the modifications used [125, 137, 175, 180]. Another group, called "open search engines" [138, 174], allow unknown or arbitrary mass differences between the precursor ion mass and the candidate structure mass, as well as between theoretical fragment masses, provided that the solution is bounded by the precursor ion mass in some way. Many of these techniques were developed to localize biologically relevant modifications such as phosphorylation [161, 177, 178], though they may be used to report any sort of peptide-variant PTM or amino acid substitution, depending upon configuration. While some claims have been made that open search algorithms can be used to identify glycosylation, little evidence has been presented to validate these claims.

Glycopeptide identification by database search has gone through many stages of development. Early methods attempted to identify glycopeptides using CID fragmentation as in GlycopeptideSearch [181], where the precursor mass was used to constrain the possible peptide + glycan combinations, and the peptide + Y ions were used to assign the glycan topology which in turn constrains the peptide, with a small amount of peptide backbone fragmentation. This method did not guarantee that the peptide sequence could be identified, and could not be easily adapted to existing null models from proteomics [182]. Later methods attempted to use HCD directly [54, 153, 183], or combinations of CID, HCD, and/or ETD [132, 184–186]. HCD alone could identify the peptide sequence and localize *N*-glycan sites, but could not reliably characterize the glycan composition, and faced challenges localizing *O*-glycan sites. Multiple dissociation techniques provided complementary information, but had technical challenges integrating data from multiple sources. GlycoFragWork [184] used linear discriminant analysis (LDA) to combine CID and ETD scoring systems into a single score for a TDA-like evaluation in two dimensions. pGlyco [132, 186] used a mixture model over combinations of target peptides, target glycans, decoy peptides and decoy glycans to combine stepped collision energy HCD scans acquired in the same run, only possible on more recent instrument models. These methods ensured that both the peptide and the glycan components are well characterized, but they require substantially greater analysis time [132, 187] as the mass spectrometer must cycle through multiple configurations on the same precursor ion, leading to longer experiment runs. Chapter 4 will discuss these methods in greater detail.

There are many commercially available database search engines, including MASCOT [6] (http://www.matrixscience.com/), PEAKS-DB [166] (http://www.bioinfor.com/peaksdb/), Byonic (https://www.proteinmetrics.com/products/byonic/). We will revisit Byonic in Chapter 4, as it is the only commercial search engine that attempts to properly cover glycopeptides. Few companies have branched out into the other MS-based search engine markets, notably Premier BioSoft (http://www.premierbiosoft.com/products/products.html) markets tools for identifying lipids, glycans [188], and metabolites, but the company's databases of structures remain proprietary. These tools see regular use in industrial and academic settings, but remain closed and charge a yearly licensing fee for each installation. Little is known about the inner workings of their algorithms and implementations, beyond what they appear to report. Most if not all are associated with a heavy desktop client interface, a suite of signal processing tools, and integrated project management systems for working with many samples simultaneously.

Search Databases

A proteomics database search engine naturally requires a *database* to search. This is usually derived from a list of proteins provided by the user in FASTA format, but the user has to get those sequences from somewhere. The two most common sources for model organism databases are UniProt [189] and GenBank [190], where curated both predicted and curated protein sequences can be found. Though it is not uncommon for users to search against large portions of these databases from many taxa in order to avoid mis-identification [191], it costs time and statistical power [192]. A list of common contaminants from sample preparation is usually recommended for inclusion in any search database [193, 194].

A FASTA file might specify the protein sequence, and some database-specific metadata to identify the protein. It contains nothing else useful for defining what kinds of endogenous biological processes might do to that protein, whether modifying the sequence or cleaving it at certain sites. G-PTM [195] worked around this issue by instead using an XML file to build its database, tailored to parse the UniProt XML schema. This feature has since been integrated into MetaMorpheus [137]. This issue has been recognized by the community, and a file format is undergoing standardization, the "PSI Extended Fasta Format" (PEFF) [111, 123, 196, 197]. PEFF aims to augment FASTA files with consistent mechanisms for multiple databases to encode information regarding PTMs, sequence cleavage and processing, genetic variants, and other instructions.

Mass Fingerprinting and Intact Profiling

LC-MS data may be acquired without fragmentation, and identification may be done based upon only the intact mass of the measured molecules alone. This method is referred to by a variation on the phrase “mass fingerprinting” or “intact profiling”. This is often done as a first pass or the sample is simple, and may not even include a separation component as in MALDI [198], especially common for glycans [199, 200]. Because of the enormous range of possible sequences from even moderately long peptides this technique is of limited use in proteomics, though some propose that multiple enzyme digests combined with physical property assays can perform competitively with LC-MS² [201].

This approach sees common use in glycomics where there are few distinct masses of building blocks [82, 94, 97, 99, 141, 199, 202]. These algorithms combine the MS dimension with the LC time dimension to evaluate whether a particular signal is differentiable from noise, and attempt to use the mass of one or more analytes to assign composition-level identifications to LC-MS features from a database. When LC conditions are held constant, some algorithms use a reference chromatographic peak time when assigning identities [202, 203]. This topic will be returned to in greater detail in Chapter 3.

1.4.4 *De Novo* Sequencing

A complementary approach to database search methods is to attempt to reconstruct the sequence of the structure from a mass spectrum *de novo* using just known structural building blocks like amino acid and modification masses. This approach is used widely in genomics and other areas, but was first formalized for peptide sequencing in 1999 with SHERENGA [204]. It, like database search, has proliferated, yielding many new approaches to the problem [8, 163, 205–207]. It has been combined with database search to create hybrid search [136, 166], or tag-based methods [208, 209]. It sees common use in other molecule classes like glycans [210–212], where a comprehensive database is often not available [70, 213, 214]. The major drawback to *de novo* sequencing is the need for high quality mass spectra [215], having few missing product ion peaks and separable noise.

De novo sequencing methods often approach topics like sequence coverage differently than database search engines, and so use different kinds of scoring functions, though the two may be combined to produce a stronger identification [166]. As open search strategies grow more prominent, we may begin to see more sequencing-like criteria in how they select the best solution [207, 216]. PepNovo [163] was one of the first approaches that attempted to incorporate product ion intensity prediction into its scoring process to put more weight on solutions which were consistent with its expectations without absolutely biasing towards the most abundant solution. The recent popularity of deep learning has shown this approach can produce very consistent predictions [164, 206]. While powerful, these methods are sensitive to instrument- and acquisition parameter-specific properties, making them less portable [163, 164].

1.4.5 Spectral Library Search

If we can reliably predict what a structure's mass spectrum would look like it could greatly improve the sensitivity of the identification process as previously mentioned. Even with millions of examples, however, these methods still fail to assign the abundances of all product ions correctly, and do not generalize well to new dissociation conditions, and acquiring many more millions of spectra is often impractical. Spectral clustering [217] and spectral library search attempt to avoid this issue by using the experimentally collected mass spectrum identified by database search or another technique [218]. This method works well for peptides, as well as for small molecules like metabolites where there are no canonical fragmentation pathways generalizable to all molecules [143]. It has also been used for glycopeptides in GPQuest [219], and just clustering has been used for spectral networks in SweetNET [133]. Spectral library search suffers from the same portability problem that intensity prediction methods do, stemming from the same changes in fragmentation process, making it less sensitive the more different acquisition conditions are from the original library construction conditions. Spectral libraries have seen extensions for clustering both identified and unidentified spectra [220], and for Multiple Reaction Monitoring (MRM) identification and quantification with SWATH [98].

1.5 Aims

Mass spectrometry-based glycomics and glycoproteomics data are complex and noisy, with many interconnected components. While they share many related signal-level characteristics, these two problems operate at different scales, calling for distinct approaches for molecule identification. These solutions require methods different from those previously published for unmodified peptides. Many tools have been proposed, but few provide an acceptable solution to the problems faced, either because of neglect of fundamental properties of the data or because of restricted scope or scale.

Within this dissertation, Chapter 2 introduces the problems related to spectral processing and transformation that must be done to interpret mass spectra, and algorithmic approaches to solve them. Chapter 3 expands on the problem of glycan profiling by LC-MS, including introducing an algorithm for sharing information amongst related glycan compositions through biosynthetic network smoothing, separately covered in [94]. Chapter 4 introduces the glycopeptide identification problem, and a set of algorithms partially covered in [23] and [54]. Chapter 5 will summarize the future direction of the described work, and the field of glycoproteomics.

Chapter 2

Preprocessing Complex Raw Mass Spectra

Mass spectrometry data contain complex structure, technical noise, and instrument-component dependent behavior [4]. As previously noted, mass spectrometers measure the mass-to-charge ratio (m/z) of ionized molecules, as well as their abundances. This work will deal with high resolution mass spectrometry where the isotopic distribution of these ions becomes resolvable. In order to compare a mass spectrum to predicted molecules, the mass spectrum must be converted from m/z-space to neutral mass space. When the monoisotopic peak of an ion and its charge state are known, this is trivial using Eq. (neutral mass). As is usually the case for any non-trivial molecule, these are not known *a priori*, and must be estimated from the data.

This topic is relevant for glycan and glycopeptide identification because most of these molecules are large, multiply charged, and produce complex mass spectra. The average monosaccharide mass is in the range of 176.68 Da, while the average amino acid mass is 127.95 Da. They also form mass clusters where two distinct molecule compositions may have very similar masses, requiring accurate and precise monoisotopic mass determination to discriminate them. This is especially important when fragmentation is not used or when the type of fragmentation used cannot produce discriminating fragments, commonly seen with CID fragmentation.

When MS^2 scans are acquired, the instrument records information about which ion it selected to be the "precursor", and the m/z window it isolated. Each instrument vendor's acquisition software tracks this information differently and depending upon configuration, reports different things in the scan metadata. The acquisition software reports the selected ion m/z and optionally, the predicted charge state(s) of that ion. The selected m/z is derived from a peak, but different vendors and different configurations for the same vendor choose which peak to report differently. In some cases, the acquisition software will report it's estimated monoisotopic peak, in other cases it will report the most abundant peak in the isolation window, in some cases doing both at different points in the same acquisition. This reported m/z tells downstream analyses what the precursor's mass is, or at least where to start looking for it, and it is important to assign this value accurately for database search algorithms to work well.

It is because of this failure that many algorithm designers decided to instead include "isotope errors" [136–138, 219] into their search engines to account for this problem. This linear probing scheme makes it possible to resolve off-by-1 neutron errors when the fragments can discriminate the two candidate precursor masses, but this does not always occur. For example, the mass difference of a neutron (1.0033) is close to the mass difference of a amino acid deamidation artefact (0.9840), or the substitution of `2Fuc` for `NeuAc` (1.0204), all of which can occur regularly in the same sample. Therefore, precursor correction is essential for glycan and glycopeptide MS^n experiments which do not produce discriminating ions consistently. For example, if a glycan composition with at least 2 `NeuAc` and 1 `Fuc` is incorrectly extracted, it may be mis-matched to a glycan with -1 `NeuAc` and +2 `Fuc` in the absence of a high quality mass spectrum. Worse, both forms may be present and their signal in the MS^1 spectrum will overlap and be summed, creating an isotopic pattern that would be very difficult to deconvolve correctly unless one knew of it *a priori*. Deconvolution algorithms including features like Hardklör's [221] have been devised to address this overlapping issue. Nonetheless, determining when an isotopic pattern is the sum of two well formed isotopic patterns versus a single, noise distorted pattern remains an unsolved problem.

Three glycan LC-MS tools have been published, each built directly or indirectly on top of Decon2LS [119]: GlycReSoft [97], Multiglycan-ESI [96], and GlyQ-IQ [99]. This is likely due to Decon2LS being only LC-scale deisotoping algorithm that was published openly and freely at the time, though others have since been published, including Dinosaur [222], and the FeatureFinder programs in OpenMS [115]. These tools used the deisotoping process as an opaque quality filter as well as for neutral mass determination.

Glycopeptides inherit the mass scaling problems of both glycans and peptides, and can cause precursor charge states to vary over a larger range than for bare peptides, from 2-9+ for tryptic glycopeptides versus peptides' 2-4+. This translates into higher charge states on glycopeptide product ions as shown in Figure 1.10, but the problem may be much worse depending upon the dissociation method used as shown in Figure 2.1 when the product ion is a mostly intact precursor glycopeptide. SweetNet [133] used MS-DeconV [223] to perform its deconvolution, including precursor recalculation, though due to its polyglot nature, also depends upon Mascot's handling of charge determination. GPQuest used an undisclosed isotopic pattern fitting method and a spectral averaging method for precursor recalculation [219], but they do not appear to examine product ion charge states [183]. pGlyco [132] used pParse [224], developed by the same group for both precursor and product ion deconvolution. GlycoPAT [225] deconvolves precursors but not the product ions. glyXtoolMS [226] uses an OpenMS [115] FeatureFinder to recalculate precursor masses, though the runtime of such an algorithm may be impractical for large datasets.

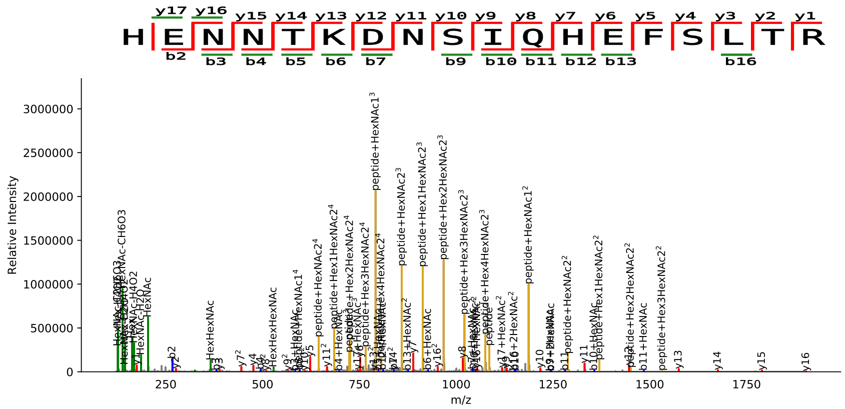


Figure 2.1: A highly charged spread of product ions from an *N*-glycopeptide using stepped HCD fragmentation with a precursor charge state of 4+ with abundant 4+, 3+, and 2+ product ions dominating the 1+ ions in the spectrum. Spectrum from data published with pGlyco2 [132], annotation produced by my algorithm.

Early signal processing and charge state deconvolution decisions can make the difference between whether the identification process returns correct or spurious assignments. I implemented these steps from scratch to be able to take full control over how the process is handled, to prevent the deconvolution from being opaque, and to be able to detect failures downstream. The building blocks of the deconvolution process are described in this chapter, with various alternatives for testing goodness of fit (Section 2.3.2) and solution extraction (Section 2.3.3). Additionally, because the previous implementations of these tasks were not as part of re-usable libraries but tightly coupled to an executable, I first implemented the algorithms I considered in a set of Python libraries with a high level interface package, `ms_deisotope`, and used that library to build a separate deconvolution tool, with the overly-optimistic hope that some of those features could be reused elsewhere.

The sequence of steps involved in this process is described in Algorithm 1, and this chapter will walk through the sub-problems in each of these steps. The degree of signal cleaning depends upon the characteristic signal properties of the mass spectrometer used, described in Section 2.1, and how those properties interact with peak picking. Peak picking in turn influences what the deconvolution process sees, and how accurate it can be.

```
Algorithm 1: Spectrum Preprocessing
  Data: Raw Spectrum $S$
  Result: Deconvoluted Neutral Mass Peak List $D$ from $S$
  $D \leftarrow \emptyset$;
  $S' \leftarrow$ CleanSignal($S$);
  $P \leftarrow$ PickPeaks($S$');
  $i \leftarrow 0$;
  maxiter $\leftarrow 10$;
  while HasRemainingPeaks($P$) and $i <$ maxiter do
      solutions $\leftarrow$ FitIsotopicPatterns($P$);
      $P \leftarrow$ SubtractIsotopicPatterns($P$, solutions);
      $D \leftarrow D+$ solutions;
      $i \leftarrow i + 1$;
  end
  return $D$
```

2.1 Instrument Types and Data Characteristics

A mass spectrum's properties depend heavily upon the instrument type that was used to produce it. A mass analyzer's resolution plays a major role in determining how a mass spectrum looks, influencing peak width, and the mass analyzer sets which physical principles are used to measure the the ion's properties, which in turn defines a set of baseline characteristics or noise patterns.

2.1.1 TOF Mass Spectrometers

A Time-of-Flight (TOF) instrument, such as a Q-TOF is a high resolution instrument, data from which consists of unfiltered detector signal, including electronic noise. Each spectrum has a dense layer of random noise in the low abundance range as shown in Figure 2.2. This also shows jagged, asymmetric peaks which may introduce complications during peak picking.

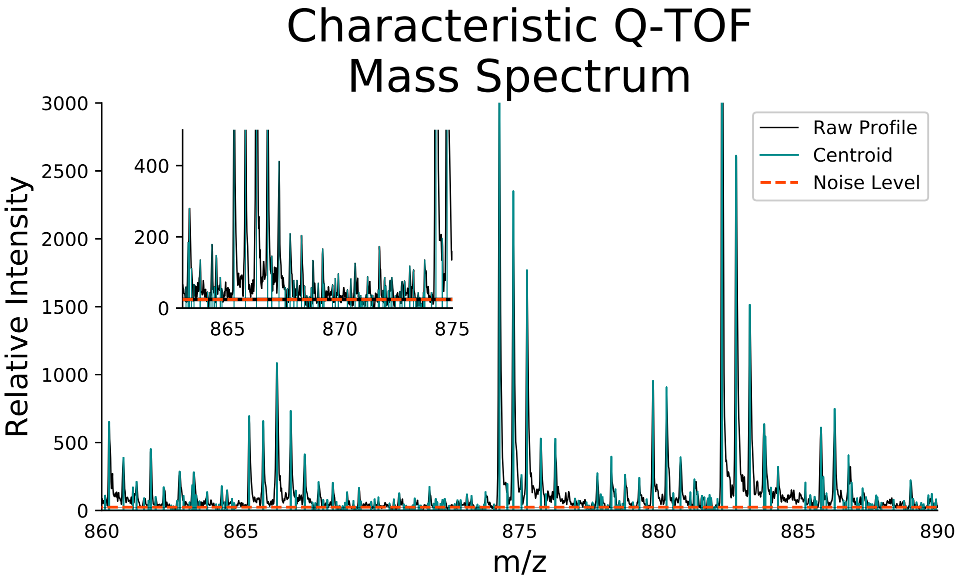


Figure 2.2: An example Q-TOF mass spectrum, showing the dense noise in the low abundance range and jagged peaks. This spectrum is from an Agilent 6550 model instrument, from data derived from IAV released glycans [23]

2.1.2 FTICR Mass Spectrometers

A Fourier Transform Ion Cyclotron Resonance (FTICR) instrument produces an ion image current that is Fourier transformed from frequency domain to m/z domain. The FT removes some of the electronic noise, but may introduce other artefacts. The FTICR has much greater resolving power and mass accuracy than a TOF instrument (See Table 1.1), and this is reflected in the sharpness of the peak shape shown in Figure 2.3. Similar to TOF spectra, any algorithms written to handle FTICR spectra must be able to take into account the dense, very low abundance signal which may have varying SNR between 1 and 3. The Fourier Transform also introduces some artifact peaks at either shoulder of abundant peaks, shown in the inset of the Figure 2.3. These shoulders may be accounted for by using a smoothing technique, such as Savitzky-Golay or Gaussian smoothing, or another more Fourier Transform specific apodization technique.

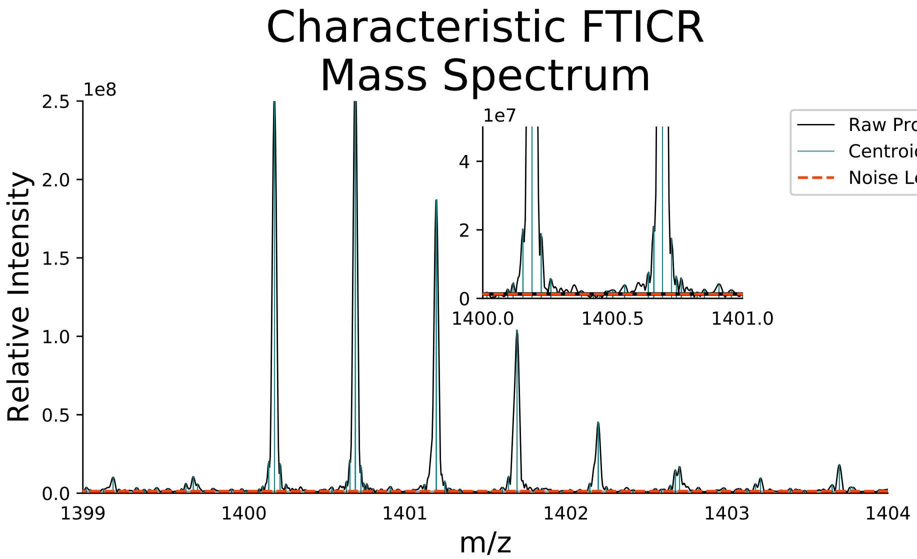


Figure 2.3: An example FTICR mass spectrum, showing the Fourier Transform baseline in the very low abundance range and sharp, narrow peaks. This spectrum is from a Bruker SolariX model instrument, from data derived from a commercially available synthetic deuteroreudced and permethylated *N*-glycan [210]

2.1.3 Orbitrap Mass Spectrometers

The Orbitrap instrument series from Thermo Fisher Scientific show resolution and mass accuracy similar to the FTICR as they are also based on a Fourier Transform. Spectra acquired with an Orbitrap undergo substantially more processing during acquisition and do not appear to have a constant noise baseline as in TOF or FTICR spectra, as shown in Figure 2.4. This preprocessing can also substantially alter or delete peaks which are either low abundance or not expected by the instrument's proprietary processing method. This "simplification" of the data can introduce extra complexity when attempting to recover some of that lost information, either by requiring a stabilizing method like scan averaging to try to fill back in lost peaks, or by forcing any modeling technique to deal with missing data.

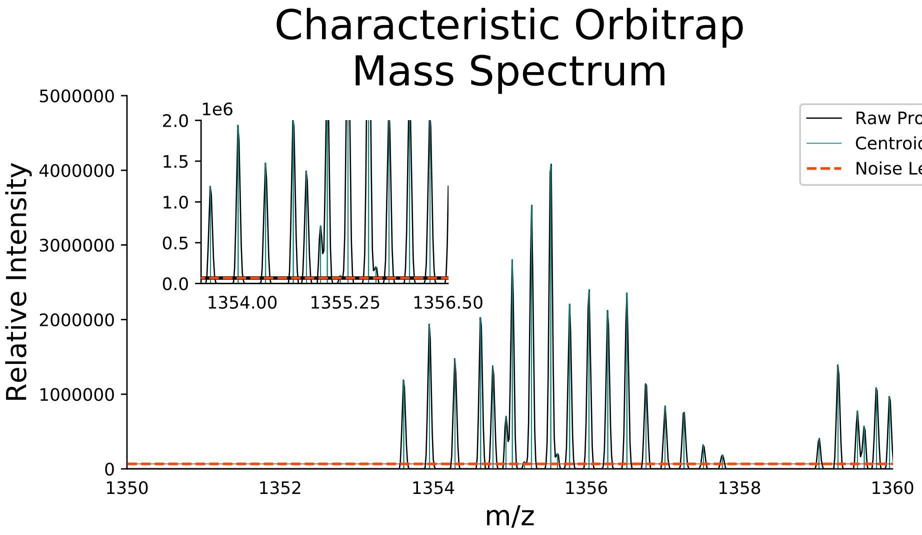


Figure 2.4: An example Orbitrap mass spectrum, showing the general absence of noise peaks, and irregular peak heights caused by the acquisition software. This spectrum is from a Thermo-Fisher Scientific QExactive-HF model instrument, from data derived from a tryptic digest of commercially available purified α-1 Acid Glycoprotein (AGP).

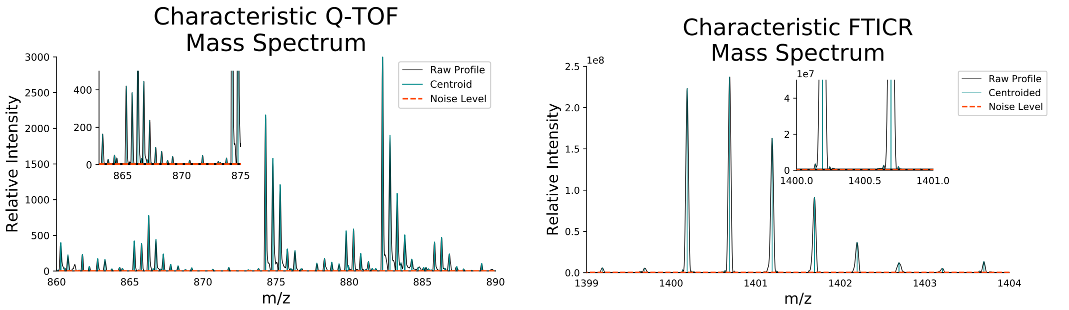
In each of the instrument characteristic examples shown earlier, we are able to observe isotopic peaks, as all of these are considered high resolution mass analyzers. These isotopic peaks can be used to estimate the monoisotopic peak of a molecule and the charge state of the ions forming that isotopic pattern. Because the isotopic pattern is “mixed” with the charge state, this is called “charge state deconvolution”. The process of collapsing isotopologues is called “deisotoping”, which will be the main topic of this chapter.

2.2 Signal Transformations

Deisotoping and charge state deconvolution depend upon the peak picking method used, and the peak picking method depends upon the quality of the raw signal it is interpreting. The raw signal is usually perturbed in some way by the instrument specific properties mentioned earlier, and cleaning the raw signal may improve the resulting centroids.

2.2.1 Background Reduction

The background noise of a mass spectrum may easily introduce spurious peaks during peak picking, and even with a global intensity or SNR threshold can drastically increase the computational workload. Some common signal filters like Fourier transform, auto-correlation, moving average, high pass, and low pass filters [227, 228] can be used to eliminate periodic noise. The background reduction techniques use an estimate of the local SNR [229, 230] to determine which data points can be used for peak fitting and what can be ignored. Most of these techniques rely on profile spectra, and cannot be used directly on centroid spectra. When this work refers to background reduction, it means the method described in [229] followed by Savitsky-Golay smoothing. This method was applied to the Q-TOF and FTICR spectra shown above, and the change can be seen in Figure 2.5.



(a) The background reduced version of Figure 2.2

(b) The background reduced version of Figure 2.3

Figure 2.5: A demonstration of the effects of background reduction on previously shown mass spectra with dense noise

2.2.2 Scan Programs and Spectrum Averaging

A mass spectrometer may acquire only MS^1 spectra, or it may acquire MS^1 spectra intermixed with MS$^\mathrm{n}$ spectra when certain conditions are satisfied, or only MS$^\mathrm{n}$ spectra of a pre-specified ion list. When MS^2 spectra are acquired, the instrument must first isolate the precursor ion to be fragmented, as shown in Figure 1.6a, recording information about the selected ion in the metadata of the product ion scan. These scan "programs" may be combined with a separation device like an LC column, inducing to a temporal ordering relationship, or be done on a simple direct infusion of the sample.

When no MS$^\mathrm{n}$ scans are acquired and multiple MS^1 scans are acquired as when using an LC system, the kth scan should contain very similar ions to the $k-1$th scan and $k+1$th scan. In this case, we can stabilize the signal and reduce the noise level of the kth scan by averaging it with its preceding and following scans. When MS$^\mathrm{n}$ scans are acquired, the previous MS^1 scan is separated in time by the time spent acquiring the interceding MS$^\mathrm{n}$ scans. Depending upon the speed of the separation device, may no longer be as similar, though the width of an LC peak under common conditions

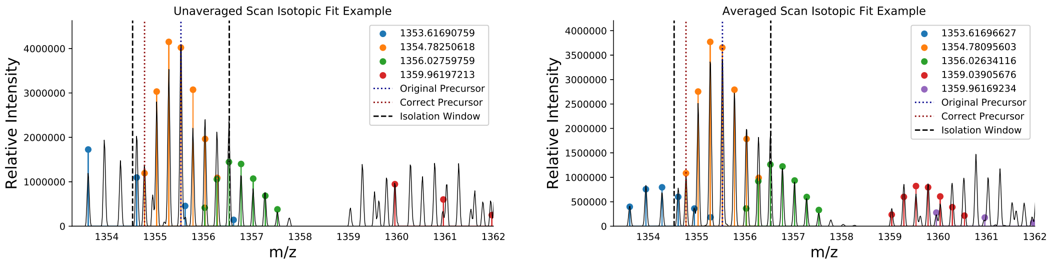
(1-2 minutes) tend to still be broad enough that averaging is useful. Scan averaging must be handled differently for profile spectra and centroid spectra. Averaging a set of profile spectra can be done using a linear interpolation as described in Eq. 2.1, where I_i is the averaged intensity at the ith point, mz_i is the m/z at the ith point in the averaged scan, K is the number of scans to average, $mz_{k,i}$ is the m/z in the kth scan closest to mz_i, $mz_{k,i+1}$ is the next closest point, and their respective intensities $I_{k,i}$ and $I_{k,i+1}$.

$$I_i = \frac{1}{K}\sum_{k}^{K}\frac{I_{k,i+1}(mz_{k,i+1} - mz_i) + I_{k,i}(mz_i - mz_{k,i})}{mz_{k,i+1} - mz_{k,i}} \tag{2.1}$$

Averaging centroid spectra is less well defined, with methods for producing “consensus spectra” described in [220]. It is also simple to convert centroid spectra with known or assumed peak widths back to a profile spectrum using a theoretical peak shape model and average the “re-profiled” spectra. This work will only deal with averaged MS^1 spectra as averaged MS^2 spectra cannot be used to localize variable modifications. An example of the effect spectrum averaging has is shown in Figure 2.6, demonstrating improved isotopic pattern fitting, even in the presence of noise

2.2.3 Peak Picking

Peak picking can be made as complex or as simple as needed, though given that it is the first step to getting accurate masses, it must be done right, or else all other steps are moot. As mentioned in Section 1.4.1, there are a variety of methods for approaching the problem. The model-fitting-based approaches like those described in [116, 118, 229, 230] were too computationally expensive as they required a first pass to identify where to attempt peak fitting prior to executing a non-linear optimization for each peak. The method used in [119] combined peak detection and peak fitting



(a) The deconvolution result annotated Orbitrap spectrum without averaging. Pearson Correlation of $\rho = 0.8713$ between the theoretical and experimental isotopic pattern for the ion in the isolation window.

(b) The deconvolution result annotated Orbitrap spectrum averaged with four adjacent MS^1 scans. Pearson Correlation of $\rho = 0.9525$ between the theoretical and experimental isotopic pattern for the ion in the isolation window.

Figure 2.6: A demonstration of the effects of spectrum averaging for stabilizing isotopic patterns for complex Orbitrap spectra, using the spectrum shown in Figure 2.4 as an example. The averaged spectrum shows better correlation with the theoretical pattern, and shows more accurate isotopic pattern fitting among overlapping isotopic clusters.

into a single process, and had a convenient closed form quadratic estimation procedure for Gaussian peak shapes. Peak shape selection may be critical for detecting overlapping peaks, but also requires making strong assumptions about the instrument used. The peak picking algorithm implemented in this work consumes a paired m/z array and intensity array, and produces a peak list object corresponding to the picked centroids in the input arrays.

Why Implement Peak Picking

As previously mentioned, most software written for interpreting mass spectrometry data expects to receive centroided peak lists. We, as a field, have left the problem for the instrument vendor to solve, and let each vendor do as good a job or as bad a job as they wish. Additionally, as previously mentioned, many noise reducing or stabilizing transformations of signal only work on profile mode spectra, and cannot be used in conjunction with vendor peak picking through ProteoWizard [120] or the vendor's commercial software. I implemented peak picking to be able to use a wider range of signal transformations and to be able to handle data from different vendors and formats more consistently.

2.3 Deisotoping and Charge State Deconvolution

Once a peak list has been produced, we can identify putative isotopic patterns by looking for peaks which are within $\frac{1.0033}{z}$ m/z away from each other, the mass difference between C^{13} and C^{12} divided by the supposed charge, as is done in simple deconvolution schemes [124, 126]. This method cannot discriminate between the serendipitous alignment of peaks and a real isotopic pattern, and such peak alignments may happen regularly, as shown in Figure 2.4. Others attempt to be more specific by adding some constant proportional rule such as that the monoisotopic peak is always the most intense [125, 225], but these cannot account for the way that isotopic patterns change as mass increases [231].

2.3.1 Isotopic Patterns

To be able to discriminate these examples, we can look for isotopic patterns which use the known abundances of different isotopes to predict the abundance of the component peaks. To do this, we must have compositions, drawn either from known molecular formulae or interpolated from a linear scaling composition model called an "averagine" [232]. An averagine, as published in 1995, was a biologically weighted average amino acid composition. An average peptide with mass m would have therefore have been made up of k averagine residues, where k can be computed by dividing m by the mass of the averagine, rounding k to an integer, and then deducting hydrogens from $k \times$ averagine's composition. This approach assumes that as mass increases, the molecule's composition changes the same way, but it can produce an average elemental composition for any non-zero mass.

Generating Theoretical Isotopic Patterns

Given a composition, there are many algorithms for generating isotopic patterns [127–131, 233, 234] at coarse or fine scale. When fine isotopic structure is desired, an algorithm like IsoSpec [130] takes into account the minute differences in mass that the extra neutrons of isotopes different elements have to produce multiple peaks for each additional neutron with abundance corresponding to the abundance of that elemental isotope. These peaks are only distinguishable at very high resolution which means that they are seldom observed when running an instrument in a high throughput mode for molecules larger than small metabolites. When resolution is insufficient to tease apart isotopic fine structure, it may still influence the shape of the aggregated peak. As resolution increases, peaks become asymmetric in an elemental composition dependent manner before deforming into sub-peaks, which in turn split into sub-peaks until each isotopologue is perfectly resolved. This process is shown Figure 2.7. For a coarse isotopic pattern which collapses all isotopologues with the same number of extra neutrons into a single peak, algorithms such as BRAIN [127, 128, 234], Mercury [235], or other methods based on Fourier transforms [233] are more appropriate, and far faster. There are minor differences in the end product of these isotopic pattern generators, but the choice of which to use is more a matter of compute time, required memory for large molecules, and stopping conditions.

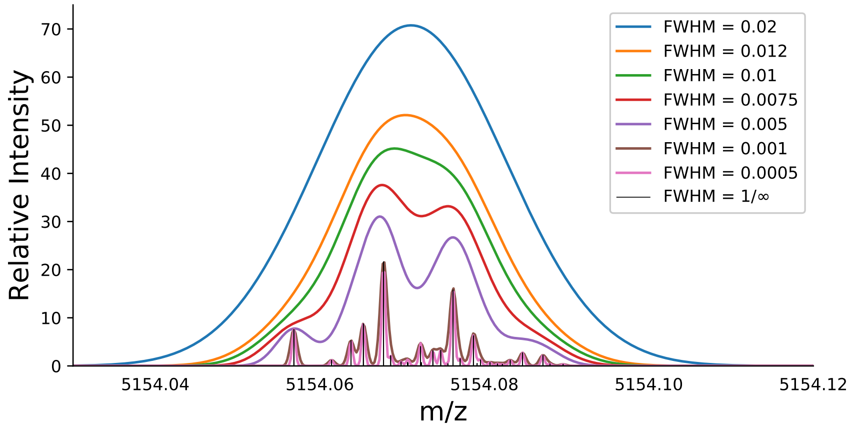


Figure 2.7: A demonstration of the effect of resolution on peak shape in the presence of isotopic fine structure for a large GAG, `{@sulfate:16; HexN:4; a-Hex:10; HexN(S):6}` ($C_{120}H_{192}N_{10}O_{167}S_{22}$). As resolution increases, the peak shape changes from seemingly symmetric (blue) to slightly asymmetric (orange), deformed (green), splitting (red), sub-peaks (purple), and repeating this process again (brown and pink) until each isotopologue is perfectly isolated (black). Because resolution depends upon the mass being measured, I instead use the peak full width at half maximum (FWHM) definition of Δm which can be used to find the required resolution by Eq. 1.1.

Previous work in the Zaia lab included an implementation of BRAIN in C++03 and Boost 1.43 [211, 234], and I have since moved that code into C 89, without a dependency on Boost, and include a Python binding. This library, `brainpy` [236], has comparable speed to state of the art C++ implementations [115], and has been used in multiple publications by members of the Zaia lab [54, 83, 94, 237] and by others

[238].

Alternative Averagine Models

There are flaws to the averagine model that arise from the assumption that the composition of a molecule scales uniformly with its mass. This was evident in the case of HS, where the degree of sulfation may increase, or the chain length may increase, changing the isotopic distribution in different ways [211]. In the case of glycopeptides, the glycan moiety may grow, or the peptide moiety may grow, skewing the isotopic distribution more towards the average for those molecule types. For HS and other sulfated GAGs, a solution is to have a high sulfation averagine and a low sulfation averagine, and to try both on every candidate mass and take whichever is better. For glycopeptides, this would mean a glycan heavy, peptide heavy, and a "balanced" glycopeptide averagine model. Using the correct averagine model for the data given is crucial, especially at higher masses. For the purposes of this study, I used the formulae shown in Table 2.1 for the glycan, peptide, and glycopeptide averagines. A comparison of their performance at a fixed theoretical m/z and z is shown in Figure 2.8. The Peptide model is notably more back-heavy compared to the Glycan and Glycopeptide because of the proportional abundance of ^{15}N to ^{14}N compared to the abundance of ^{18}O to ^{16}O. It is worth noting that this Glycan model is based upon the pattern common to *N*-glycans which span multiple sub-groups with variable amounts of nitrogen, for example high mannose *N*-glycans will contain fewer nitrogens than sialylated complex type *N*-glycans.

Another alternative approach to this problem, when the set of possible molecules to consider is tractable, is to search for the isotopic pattern of every molecule in the search space in the spectrum, at every charge. This method steps around the problem of selecting an appropriate averagine model, and should theoretically be better

Model Name	Formula	Monomer Mass
Peptide	$H_{7.76}C_{4.94}S_{0.04}O_{1.48}N_{1.36}$	111.054 31
Glycan	$H_{11.83}C_{7.00}O_{5.17}N_{0.50}$	185.5677
Glycopeptide	$H_{15.75}C_{10.93}S_{0.02}O_{6.48}N_{1.66}$	274.5067

Table 2.1: Averagine formulae used in this chapter. Note that the Peptide model is the original Senko Averagine [232], while the Glycan Averagine was manually estimated from a representative subset of *N*-glycans. The Glycopeptide Averagine ≈ Peptide Averagine + $\frac{(\texttt{HexNAc}+\texttt{Hexose})}{2}$ respectively.

than an averagine for high fidelity signal. In previous work, we used this approach for various types of GAG deconvolution [83, 237]. This significantly changes the time complexity and interpretation of the results, and is not tractable for direct use with whole glycoproteomes.

Isotopic Pattern Width

As molecules grow, the number of isotopic peaks abundant isotopic peaks grows as well, the trend shown in Figure 2.9. As the isotopic pattern grows more complex, the number of trivial low abundance peaks increases, first in the right tail, and later, as the heavier peaks of the isotopic pattern shift towards the center, the left tail begins to become trivial as well. Searching for ever peak in an isotopic pattern is not useful as most low abundance peaks do not help us find the correct pattern. The number of peaks to include in a theoretical isotopic pattern has been debated in [224, 233, 234] discuss number of peaks to use when constructing a theoretical isotopic pattern, but no consensus is reached.When an analyte is not abundant, many of its later isotopic peaks are indistinguishable from noise, so looking beyond the first few peaks is impractical. In this work when I consider MS^1 spectra I use the first 95% of the isotopic pattern signal, and for MS^2 spectra I use the first 80% of the isotopic pattern signal. This produces a stronger fit for the wide, complex isotopic patterns in MS^1,

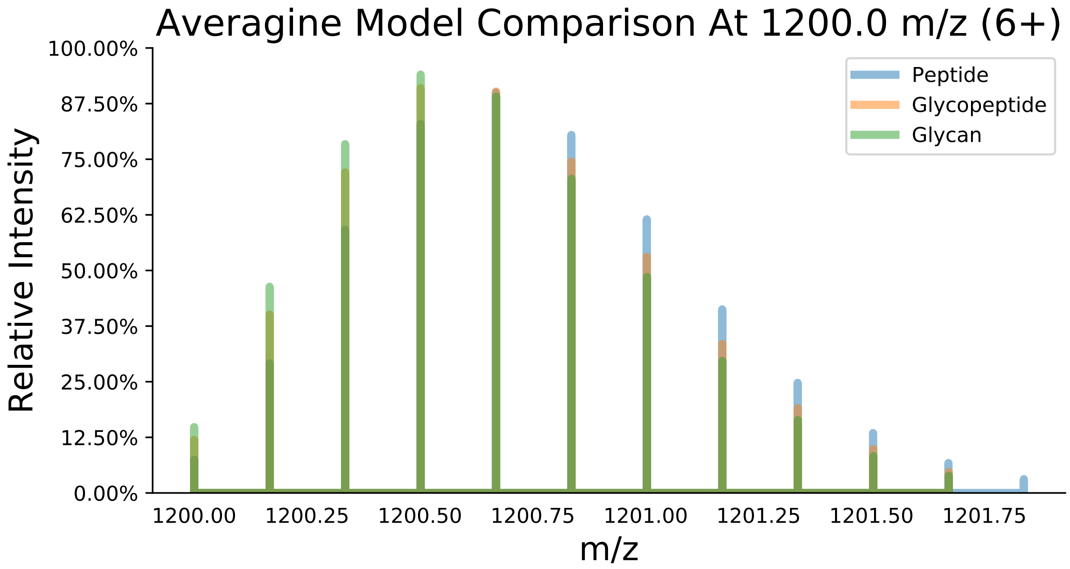


Figure 2.8: A comparison of three averagine models at $m/z = 1200$ and $z = 6+$. Note the Glycan model is front-heavy, and the Peptide model is back-heavy, while the Glycopeptide model is balanced between them, as desired.

but avoids having many missed peaks in low abundance MS^2.

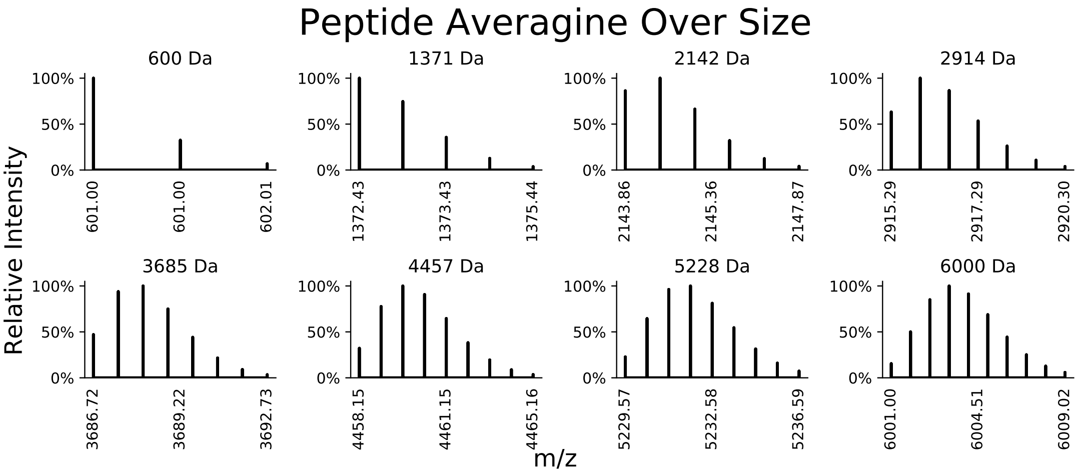


Figure 2.9: Each panel shows a theoretical peptide isotopic pattern at a given mass. As mass grows, the number of non-trivial peaks grows. Generating a theoretical isotopic pattern without a threshold produces many trivial peaks. All patterns shown have been truncated after 99% of their signal has been displayed.

2.3.2 Goodness of Fit

Given a theoretical isotopic peak intensities T and a set of matched observed peak intensities O, we need a function $f(O,T)$ to determine how well T fits O. First, to make the comparison fair, we must scale T such that $\sum_i^k t_i \approx \sum_i^k o_i$. The simplest normalization is to begin with the precondition that $\sum_i^k t_i = 1.0$, and then simply set $\hat{T} = T \sum_i^k o_i$, though there are many other techniques. One example is non-negative least squares [118, 230], another is scaling every peak by a constant c such that $cT_{argmax(O)} = max(O)$, or a weighted average over possible values of c, though each introduces its own complications when deconvolving overlaps [221].

Minimizing Functions and Fit Constraints

We have an abundance of goodness-of-fit scores to choose from, with some methods growing more optimal as $f(O,T) \rightarrow 0$, minimizing functions, and other methods grow more optimal as $f(O,T) \rightarrow \infty$, maximizing functions. Minimizing functions come naturally from common statistical tests. The χ^2-test (Eq. 2.2) is one such example used in foundational work by Senko [232] and it is used as a reference for others [119, 239].

$$\chi^2(O,T) = \sum_i^k \frac{(\hat{t}_i - o_i)^2}{o_i} \tag{2.2}$$

When $\hat{T} = O$, $\chi^2 = 0$, and it reflects a perfect fit, though the error is on the scale of the absolute intensity, making selecting a maximum error threshold problematic. A related measure used by Decon2LS [119] which they refer to as a “AreaFit” works around this scale issue by instead operating on the max-normalized versions of both

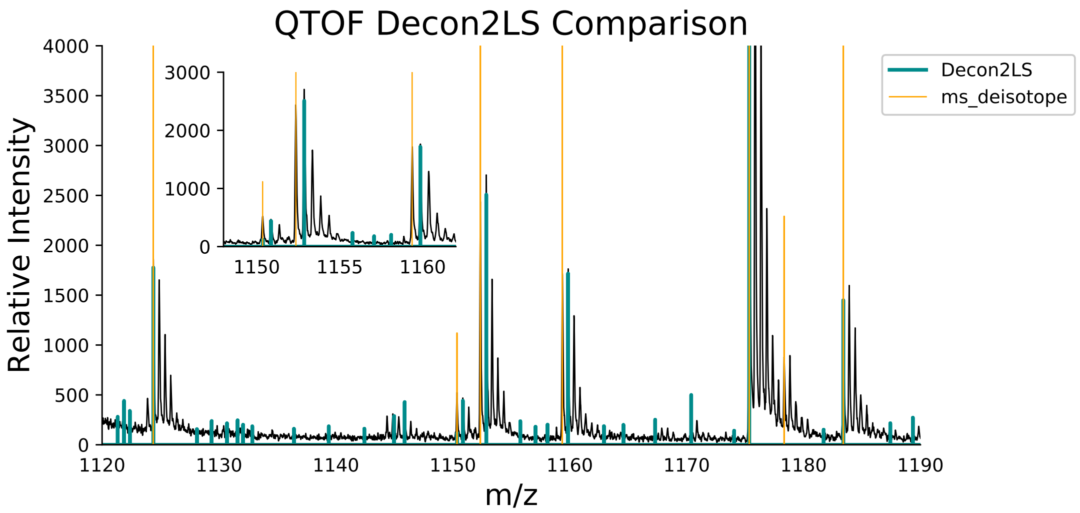


Figure 2.10: A comparison of Decon2LS [119] with ms_deisotopederived deconvolution. Decon2LS produces many spurious peaks, and does not reliably select the monoisotopic peak.

T and O in Eq. 2.3.

$$AreaFit(O,T) = \frac{\sum_i^k ((t_i/max(T)) - (o_i/max(O)))^2}{\sum_i^k (t_i/max(T))^2} \tag{2.3}$$

This lets us set a goodness-of-fit threshold that is independent of the magnitude of O, and we can instead employ the same scaling method on the χ^2 function to step around this problem too. Previous work by the Zaia lab had used Decon2LS [119] for glycomics [97] and attempted to use it for glycoproteomics. The goodness-of-fit function used by Decon2LS is the AreaFit shown in Eq. 2.3 and does not reliably select the monoisotopic peak, as shown in Figure 2.10.

A third function that behaves similarly is the G-test shown in Eq. 2.4. The G-test is more stringent than χ^2, though the two tests are related.

$$G(O,T) = 2\sum_i^k o_i \log\left(\frac{o_i}{\hat{t}_i}\right) \tag{2.4}$$

Maximizing Functions and Signal Usage

These minimizers are good at discriminating the real patterns from noise patterns, and the scaled variants all work report proportional errors, in that they treat errors of large peaks the same as the errors of small peaks. However, this means that they will seriously consider every small peak, which makes the prior noise filtering steps leading to peak picking more important. Additionally, these error minimizing functions are biased against larger masses, as they require more points to follow the same pattern, preferring fits that depend upon the fewest peaks. A maximizing criterion like the Pearson correlation coefficient will have the same problem, as it will penalize including additional sub-optimal peaks in a fit. Hardklör [221] uses a normalized dot product between the theoretical and experimental isotopic pattern to select the best fit shown in Eq. 2.5.

$$d(O,T) = \frac{O \cdot \hat{T}}{\sqrt{\hat{T} \cdot \hat{T}}\sqrt{O \cdot O}} \tag{2.5}$$

Using Hardklör on glycopeptide MS^1 spectra produced some correct monoisotopic peaks, though it missed many isotopic patterns, shown in Figure 2.11

A goodness-of-fit score that uses an additive model which does not directly reduce the quality by adding a sub-optimal peak would be biased towards solutions using more peaks. An unscaled dot product score for example would have this pathology. The MS-Deconv [223] scoring function is more complex, shown in Eq. 2.6, uses both intensity and m/z information, where $mz(o)$ corresponds to the observed m/z

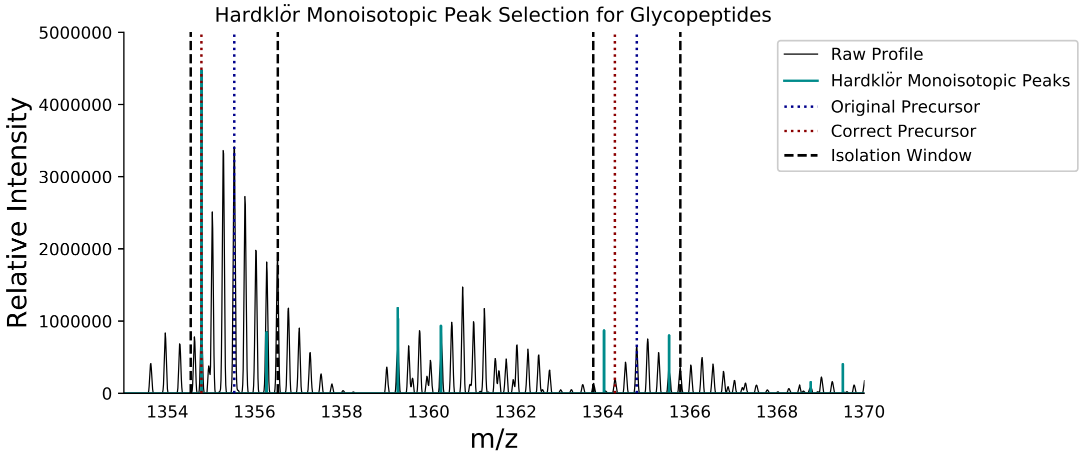


Figure 2.11: A visual comparison of the published Hardklör's [221] selected monoisotopic peaks and the manually selected correct precursor peaks.

for the observed peak and $mz(t)$ corresponds to the m/z of the theoretical peak.

$$s_{mz}(o,t) = \begin{cases} 1 - \frac{\left|mz(o) - mz(\hat{t})\right|}{d} & \left|mz(o) - mz(\hat{t})\right| \leq d \\ 0 & \text{Otherwise} \end{cases} \tag{2.6}$$

$$s_i(o,t) = \begin{cases} 1 - \frac{\hat{t}-o}{o} & \text{if } o < \hat{t} \text{ and } \frac{\hat{t}-o}{o} \leq 1 \\ \sqrt{1 - \frac{o-\hat{t}}{o}} & \text{if } o \geq \hat{t} \text{ and } \frac{o-\hat{t}}{o} \leq 1 \\ 0 & \text{Otherwise} \end{cases} \tag{2.7}$$

$$\text{MS-Deconv}(O,T) = \sum_i^k \sqrt{\hat{t}_i} \times s_{mz}(o_i,\hat{t}_i) \times s_i(o_i,\hat{t}_i) \tag{2.8}$$

This goodness-of-fit score is on the order of $\sqrt{\hat{T}}$ when optimal, and the addition of an extra peak does not improve the fit, but does not penalize it either. The s_i function penalizes deviation from the theoretical distribution, but it penalizes exceeding the expected intensity less, presumably because it can be caused by overlapping peaks pooling signal intensity. The authors of MS-Deconv recognized that their goodness-of-fit criterion did not reliably select the monoisotopic peak, so they added post-hoc monoisotopic peak recalculation that tests to determine whether the best alignment between O and $\hat{T}$ arises at the selected monoisotopic peak, or at the peak at $mz(o_1) - \frac{1}{z}$. In practice, these errors may be larger than a single neutron error for overlapped isotopic patterns, and the overlapping pattern may be more complicated. While MS-Deconv is freely available, its source code is not, and it is not possible to adjust the averagine model it uses. The published executable was not successful when applied to complex glycopeptide MS^1 data, shown in Figure 2.12a. Its goodness-of-fit score is simple to implement, and performs reasonably well. The algorithm performed well on glycopeptide MS^2 data, particularly for the smaller peptide backbone product ions, but did not achieve full coverage of the high mass range fragments shown in Figure 2.12d. My re-implementation of the scoring function in ms_deisotope was able to capture those missing isotopic patterns, with a spectral similarity of 0.98 with the MS-Deconv result.

The score used in MetaMorpheus's mzLib [137] is less modular, not being constrained by MS-Deconv's design requirements for dynamic programming. The full expression is shown in Eq. 2.12.

$$r_i = \frac{\hat{t}_i}{o_i} \tag{2.9}$$

$$\sigma(O,T) = \sqrt{\frac{1}{k-1}\sum_i^k \left(r_i - \left(\frac{1}{k}\sum_i^k r_i\right)\right)^2} \tag{2.10}$$

$$z(O,T) = \left\lfloor \frac{1.0033}{mz(o_2) - mz(o_1)} \right\rfloor \tag{2.11}$$

$$M(O,T) = \frac{\sum_i^k o_i}{\sigma(O,T)^{0.13}} \times \frac{|O|^{0.4}}{z(O,T)^{0.06}} \tag{2.12}$$

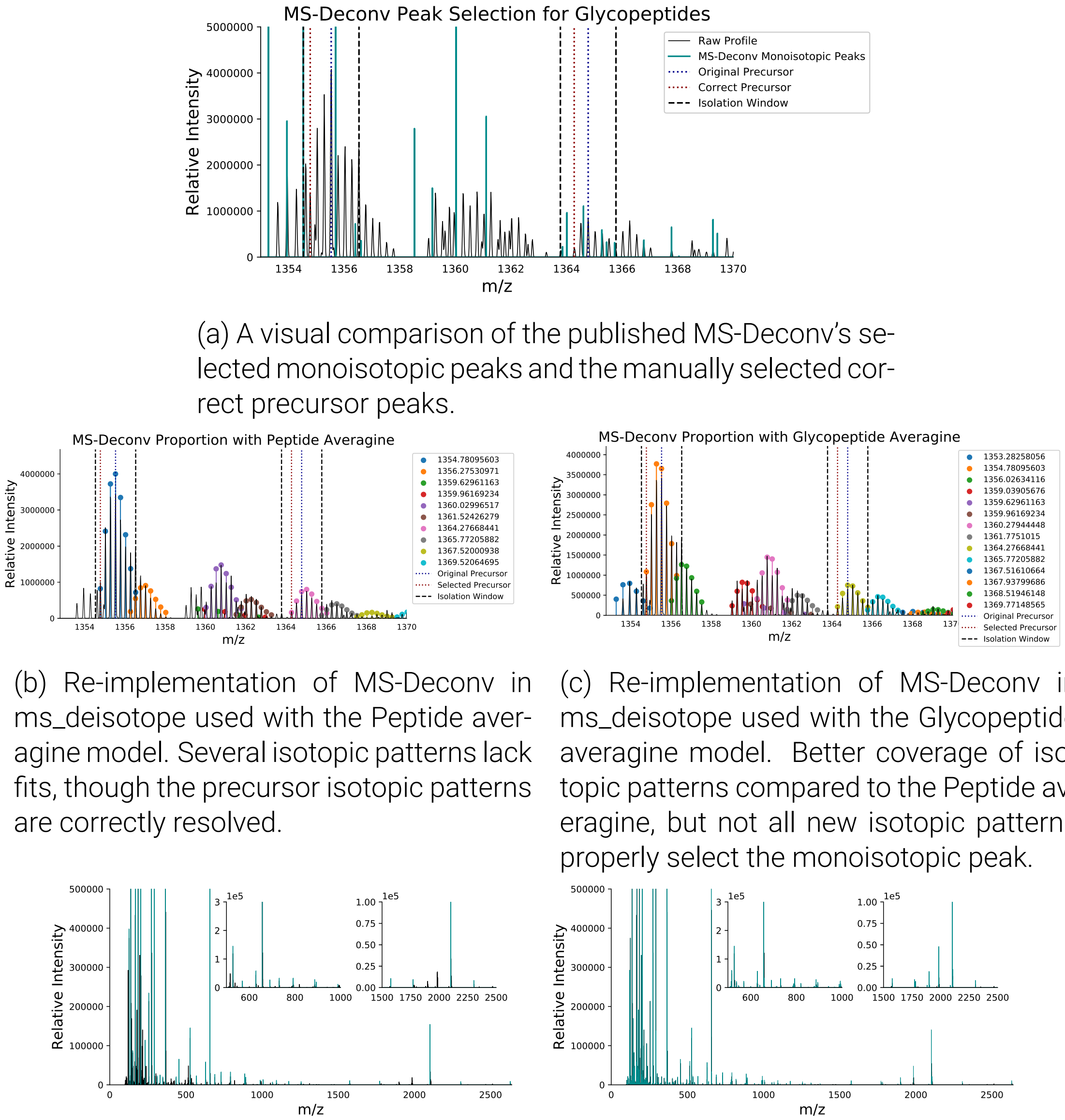
This score places an explicit preference for the fit which uses the most peaks and the smallest charge state (Eq. 2.11). The three constant exponents may be tunable parameters to adjust the biases, though the method as implemented uses them as constants. The constraint on the isotopic pattern fit is vulnerable to being washed out abundance much faster than in MS-Deconv's due the use of intensity rather than its square root as in MS-Deconv.

I propose a variant of MS-Deconv shown in Eq. 2.15 which is scaled down by a scaled version of the G test statistic shown in Eq. 2.14.

$$\begin{aligned}
\hat{o}_i =& \frac{o_i}{\sum_j^k o_j} && (2.13)\\
\hat{G}(O,T) =& 2\sum_i^k \hat{o}_i \log\left(\frac{\hat{o}_i}{t_i}\right) && (2.14)\\
\text{P-MS-Deconv}(O,T) =& \text{MS-Deconv}(O,T) \times (1 - \gamma\left|\hat{G}(O,T)\right|) && (2.15)
\end{aligned}$$

This method augments the parameterization of MS-Deconv with a second term γ which controls how strong the penalty is for deviating from the expected isotopic pattern. This method is not monotonic, as adding peaks to the fit can lead to a worse fit, but when it is applied, it can prevent non-optimal patterns from being proposed unless no other solution is found. A partitioning of hypothetical score spaces is shown in Figure 2.13.

All additive scores must be thresholded at some value to select only real fits and omit spurious ones, like non-additive or or scaled scores. While non-additive or scaled additive scores may not be on the order of the intensity, all of the maximizing additive scores shown here are. Selecting a threshold for such scores depends upon the the magnitude of the intensity measure used, which can vary wildly from instrument to instrument. Another advantage of the penalized MS-Deconv score is that while it does not put an upper limit on the score, it does make many poor fits have a score less 0, proposing a natural threshold. Because the penalty is multiplicative, it cannot be reproduced by first filtering by a scaled G-test followed by MS-Deconv.



(a) A visual comparison of the published MS-Deconv's selected monoisotopic peaks and the manually selected correct precursor peaks.

(b) Re-implementation of MS-Deconv in ms_deisotope used with the Peptide averagine model. Several isotopic patterns lack fits, though the precursor isotopic patterns are correctly resolved.

(c) Re-implementation of MS-Deconv in ms_deisotope used with the Glycopeptide averagine model. Better coverage of isotopic patterns compared to the Peptide averagine, but not all new isotopic patterns properly select the monoisotopic peak.

(d) The published MS-Deconv's selected monoisotopic peaks from a subsequent glycopeptide MS^2 spectrum.

(e) Re-implementation of MS-Deconv in ms_deisotope used with the Peptide averagine model on the same subsequent glycopeptide MS^2 spectrum.

Figure 2.12: Application of MS-Deconv to the complex glycopeptide spectrum shown in Figure 2.6b, and the precursor isolated at 1354.78 m/z

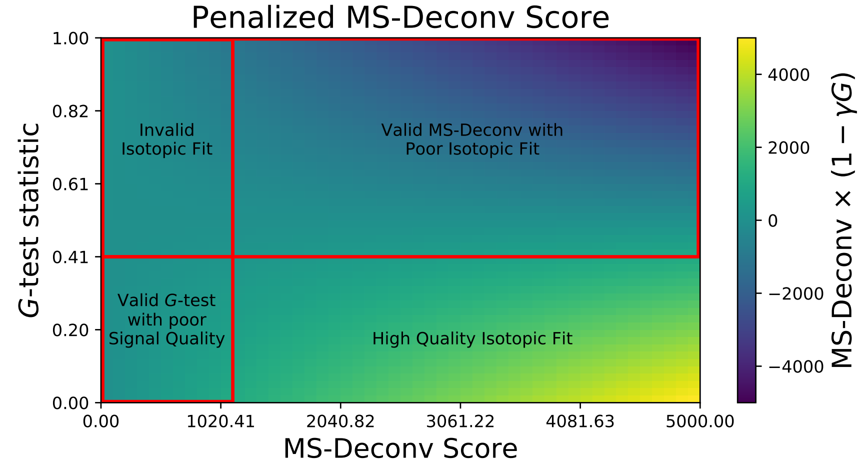


Figure 2.13: Describes the different domains in which the G-test and MS-Deconv fail to produce the desired result, but that the combination can be used to select only high quality patterns.

2.3.3 Solution Search Strategy

The strategy that uses one or more of these goodness-of-fit scores to choose amongst a set of possible deconvolution solutions is just as important as the goodness-of-fit metric. There are two broad classes of deconvolution strategy with representative implementations, THRASH [240] and dependency tracking [223].

THRASH

Thorough high resolution analysis of spectra by Horn (THRASH) [240], was the first successful automated mass spectrum deconvolution algorithm, laid the foundation for much later work [119, 221, 229]. THRASH works by aligning a theoretical isotopic pattern at every $\pm\frac{1.0033}{z}$ m/z increment from a reference peak in a 2 m/z interval around that peak, or until the score stops improving from steps either direction. This search finds a locally optimal alignment, which is then recorded and subtracted from the spectrum, and the search repeats until there are no more solutions to find. Because THRASH's solution is only a local optimum, it does not take into account nearby peaks which may be part of an isotopic pattern that shares one or more peaks with its local solution, leading to whichever reference peak that is chosen as the first reference peak having an arbitrary advantage.

Dependency Graphs

This prompted the development of a set of related dependency tracking approaches, where all possible solutions that depend on a subset of peaks are solved simultaneously. MS-Deconv [223] first presented this approach, where it constructed a dependency graph, and then evaluated every combination of isotopic pattern to peak assignments that allowed either binary or partial but equal peak ownership sharing.

The solution could be found in a reasonable amount of time by constraining the combinatorial search of peak ownership to just those fits that were not disjoint and by using an additive score that permitted a dynamic programming solution. A feature based deisotoping scheme [241] expanded on this idea by incorporating information about related isotopic fits into the isotopic fit's score. These approaches are good for complex, overlapping mass spectra where the local greedy solution may not be the globally optimal solution. It is worth noting that MS-Deconv does not use iterative signal removal as THRASH does, so all overlapped peaks are taken entirely or shared evenly between all claiming isotopic fits in a solution, they cannot account for unequal sharing when scoring isotopic patterns, trickling down as lower scores due to Eq. 2.7.

Another approach to solving this problem was to try to solve the whole spectrum simultaneously, as done by IPPD [230]. IPPD estimates an abundance coefficient for each possible pattern using ℓ_1-regularized non-negative least squares. The regularization forces poor fits to have a coefficient zero, but this constrains the type of isotopic fit score used, requiring multiple postprocessing steps to select the real fits from spurious ones. ℓ_1-regularized non-negative least squares also has the advantage of simultaneously determining both the placement and the scale of isotopic patterns, as well as allowing unequal peak sharing between multiple isotopic patterns.

When the set of possible compositions is known and tractable, other solutions may be used to deconvolve complex mixtures. MassTodon [242] approaches the problem by first computing every fragment from a query structure, and then estimates an isotopic pattern for each product ion based on the theoretical composition of each fragment, maps each isotopologue to experimental signal, and casts the problem of assigning signal abundance as a constrained Max-Flow problem with euclidean distance penalties for imperfect fits. This operates on top-down protein spectra from ultra-high resolution instruments, where isotopic fine structure is present [130], which also reduces the efficiency of an averagine model.

2.3.4 Why Implement Deisotoping and Charge State Deconvolution?

With so many well made deconvolution algorithms already published, implementing another does not seem like productive venture. This may be true in theory, but design decisions make re-use of the *implementations* of these ideas difficult.

Specialization

As shown in Section 2.3.1, glycans and glycopeptides follow different average elemental trends compared to bare peptides, but not all tools allow users to change their averagine models, nor do they allow the averagine model to differ between MS^1 and MS^2 spectra. For example, a glycopeptide, when intact, has a glycopeptide-like isotopic pattern, but under HCD dissociation it's product ions may be bare peptide backbone ions following a peptide-like isotopic pattern, partially intact glycan on an intact peptide be glycopeptide-like. Additionally, the noise types of each implementation assume that the same model is applicable to both MS^1 and MS^2, and that the noise level is global [119, 223] when it is known it may vary with location [229, 230]. For example, applying a global noise level estimation according to the mean signal in a spectrum with abundant oxonium ions may skew the threshold too high to reliably assign glycopeptide backbone product ion monoisotopic peaks from TOF data. Likewise, even with a high quality scoring function, attempting to interpret every jagged peak in a TOF spectrum will be impractically slow, even with a SNR threshold those non-zero points need to be visited, a pathology suffered by Decon2LS [119]. Other tools like Dinosaur [222] explicitly do not attempt to defend against peak noise because Orbitrap instruments do not exhibit it.

File Formats and Runtimes

Each tool also reads data in mass spectra in different formats, and writes data out in its own ad hoc text file format. MS-Deconv [223] (Java) can read MGF and mzXML, and can write out MGF or a tabular format of its own devising, stripped of metadata. Decon2LS [119] (MSVC++ and .NET) can read from raw files, mzML, and mzXML formats using a CLR binding for ProteoWizard [120], and writes its output in a set of CSV files, also stripped of metadata. Hardklör [221] (C++) reads mzML and mzXML format files, and writes its output in a tabular text file. IPPD [230] is an R library, and MasSpike/BUDA [229] and the feature-based deisotoping technique described in [241] are MATLAB libraries, but require their own run-times environments as well as metadata and I/O management. pParse [224] (.NET/C++) can only read Thermo Fisher's raw file format or MGF files, and writes its output in binary files to be read by other tools of the pFind suite. This makes it difficult to write software to read modern, metadata-rich mzML files when the legacy system being used for deconvolution is unable to read them. One alternative is to use `msconvert` [120] to convert from mzML to the desired legacy format, run the deconvolution, and then map the deconvoluted results back to the original file, but this may involve copying tens to hundreds of gigabytes of data, and the deconvolution process remains a black box. Additionally, it would make it difficult to add new features to the neutral mass determination process such as for co-isolation detection [137] or make decisions based upon metadata not available in the chosen format.

2.4 Design of ms_deisotope

2.4.1 Signal Processing Pipeline

Prior to peak picking of MS^1 spectra, I included a signal averaging step as described in Section 2.2.2. Next, I employed a background noise reduction method based upon [229], followed by a Savitsky-Golay smoother, as described in Section 2.2.1. Following this denoising step, I used a fast least-squares peak picking method derived from [119] to centroid profile peaks. Unless explicitly requested, no denoising is done to MS^n spectra which do not have the same noise characteristics. Additional signal transforms may be specified by the user.

2.4.2 Pattern Search Algorithm

From the development of dependency graphs described in 2.3.3, I knew the shortcomings of THRASH, but that the dependency graph method did not reliably work on complex glycopeptide spectra. To this end, I used the dependency graph method to determine which isotopic fits share peaks, and then extract greedy solutions. For MS^1 spectra, which are far more complex than MS^2 spectra, the Penalized MS-Deconv score worked best for enforcing the selection of valid isotopic patterns. As shown previously in Figures 2.12 and 2.10, MS-Deconv and Decon2LS were insufficient. I combined ideas from THRASH and dependency graphs by constructing the dependency graph using a normal traversal of the spectrum shown in Alg. 2, but used an iterative monoisotopic peak recalibration from THRASH to generate alternative starting points described in Alg. 3. Because I cannot guarantee the scoring function will behave as MS-Deconv's does, I solve each subgraph greedily. This prevents arbitrary traversal order from favoring one isotopic pattern over another, as in traditional

Algorithm 2: FitIsotopicPatterns

```
Data: Peak List $P$
Data: Charge Range $C$
Data: Isotopic Fit Graph $G$
Data: Isotopic Fit Score Threshold $t$
Result: Solution Set $D$
foreach $p_{seed}$ in $P$ do
    foreach $c$ in $C$ do
        // Recalibrate $p$ from other nearby peaks given $c$
        $M$ ←MonoisotopicPeakRecalibration($P, p_i, c$);
        foreach $p$ in $M$ do
            $tid$ ← CreateIsotopicPattern($p, c$);
            // Match experimental peaks and construct an
            // Isotopic Fit
            $fit$ ← MatchIsotopicPattern($P, tid, 10ppm$);
            $fit.score$ ← ScoreIsotopicFit($fit$);
            if $fit.score > t$ then
                put $fit$ in $G$;
            end
        end
    end
end
$D \leftarrow \emptyset$;
foreach $g$ in FindDisjointSubgraphs($G$) do
    $D \leftarrow D \cup$ SolveSubgraph($g$);
end
return $D$
```

THRASH. I subtract the used signal and repeat the process until either the peak list stops changing significantly from subtraction or after $n = 10$ iterations, letting the algorithm remove overlaps and uncover distorted patterns.

Because precursor recalculation was an important part of the deconvolution process, they are queried first, and with greater stringency. When the deconvolution process fails to locate an acceptable solution that spans the provided isolation window, matches the vendor selected charge state if given, and includes the centroid peak nearest to the vendor reported m/z, the precursor is marked as "defaulted", signaling

```
Algorithm 3: MonoisotopicPeakRecalibration
    Data: Peak List $P$
    Data: Starting Peak $p_{start}$
    Data: Putative Charge State $c$
    Data: Max Step Count $k$
    Result: Monoisotopic Peak Candidates $M$
    $M \leftarrow \{(p_{start}, c)\}$;
    $i \leftarrow 1$;
    while $i < k$ do
        // Search for better monoisotopic peaks in the lower mass
        // range to the left of $p_{start}$
        $mz_i \leftarrow p_{start}.mz - (1.0033/c) * i$;
        $P_{mz_i} \leftarrow$ AllPeaksFor($P, mz_i, 10ppm$);
        $M \leftarrow M \cup \{$PlaceholderPeak($p_j.mz - (1.0033/c) * i$) : $p_j \in P_{mz_i}\}$;
        $i \leftarrow i + 1$;
    end
    return $M$
```

to downstream processes that they should not trust the reported mass as shown in Alg. 4. Additionally, the isolation window is queried for overlap with other abundant isotopic patterns, which introduce co-isolated fragmentation products to the associated MS^2 spectrum.

2.4.3 Test and Comparison

I tested the algorithm on several datasets used in [23, 243] to select which features should be used for streamlined processing. The Penalized MS-Deconv scoring function worked well for MS^1 spectra where the ions are abundant and isotopic patterns are complex, and produced fewer spurious solutions than the original MS-Deconv scoring function. In MS^2 spectra, I found the MS-Deconv scoring function worked better under less ideal circumstances common to glycopeptide spectra, while still being able to reliably recover the charge state and monoisotopic peak when an isotopic pattern was observed. Compared to Decon2LS, ms_deisotope's implementation was

Algorithm 4: PrecursorIsotopicFitExtraction

```
Data: Isotopic Fit Graph $G$
Data: Query Peak $p$
Result: Best Spanning Isotopic Fit $fit$
$fit \leftarrow$ NULL;
$fits \leftarrow \texttt{FindDependentSolutions}(G, p)$;
if $fits \neq \emptyset$ then
    $fit \leftarrow \arg\max_{fit} \{fit.score : fits\}$;
else
    $fits \leftarrow \texttt{FindDisjointInterval}(G, p.mz)$;
    if $fits \neq \emptyset$ then
        $fit \leftarrow \arg\min_{fit} \left\{\frac{|fit.monoisotopic\_mz - p.mz|}{fit.score} : fits\right\}$;
    end
end
return $fit$
```

substantially better at selecting the monoisotopic peak for the glycan LC-MS samples, all of which were ran on a Q-TOF-based instrument, and Decon2LS was not successful when deconvolving product ion peaks for glycopeptides. The consequences of this difference is partially described in [94], discussed in more detail in Chapter 3. It is not possible to do a fair side-by-side comparison of deconvolution results between Decon2LS and the ms_deisotope deconvoluter using GlycReSoft because the downstream method uses information passed along by ms_deisotope that would be absent in the Decon2LS workflow.

To demonstrate the performance of the efficacy of ms_deisotope on glycopeptide data when compared to a separately published deconvoluter, I used MS-Deconv [223] and ms_deisotope on a tryptic AGP digest sample ran on a Thermo Fisher QE-HF used for internal calibration. As a crude first measure I compared the precursor masses reported by both tools for each MS^2 spectrum, shown in Figure 2.14. The nature of ms_deisotope's precursor recalculation policy stipulates that it cannot be more than 3 Da away from the vendor reported precursor peak, which accounts for

12.5% of the differences in precursor masses reported. Of the remaining spectra, 20.3% have a mass difference between 0.2 to 3 Da, which may account for differences in isotopic peak selection or precursor peak. An error of 0.2 Da is sufficient to violate a 10 ppm mass accuracy constraint even over 12,000 Da.

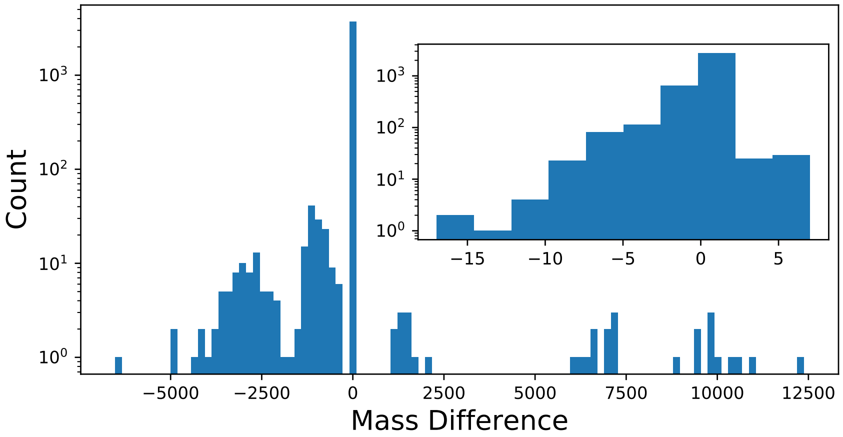


Figure 2.14: A histogram of the difference between MS-Deconv [223] vs ms_deisotope for a tryptic AGP digest. The counts are shown in log-scale, showing that most reported precursor masses were very close between the two tools. The plot however shows many large errors on the scale of hundreds to thousands of Da. These errors are the result of MS-Deconv failing to find a satisfactory solution for the instrument reported peak and charge, and either reporting a different charge or finding the nearest fitted peak to report. The inset plot shows the breakdown of mass differences close to zero, which still shows substantial deviations.

While MS^1 errors are substantial, they are considerably more complex than what MS-Deconv was originally intended to solve, and are governed by different isotopic models. Glycopeptide product ions produced by HCD tend to be dominated by peptide-like ions with a smaller number of monosaccharides present. To measure MS^2 similarity, I extracted each scan and calculated the cosine similarity between MS-Deconv's interpretation and ms_deisotope's, using precision to the second decimal place. The resulting similarity trend is shown in Figure 2.15. This method of comparison remains crude as well, but an error of 0.001 Da will not be missed at 20 ppm mass accuracy for a mass greater than 500 Da.

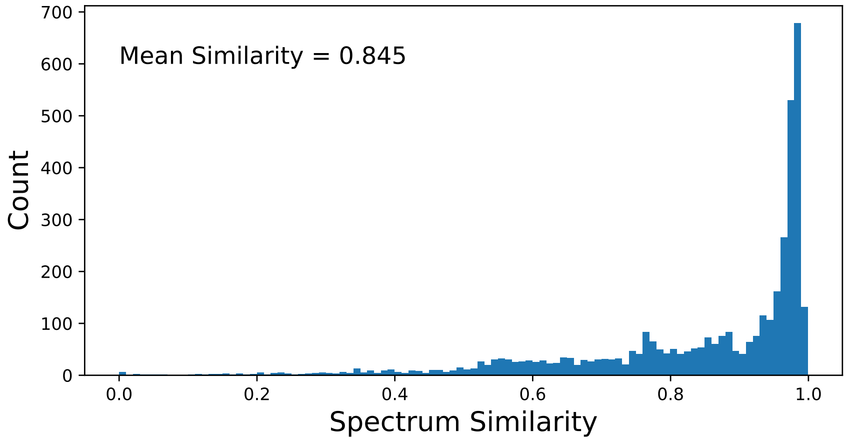


Figure 2.15: A histogram of the cosine similarity between MS-Deconv and ms_deisotope MS^2 spectra.

2.4.4 Conclusion

These comparisons demonstrate that the deisotoping and charge state deconvolution method selected suitable for glycans and glycopeptides. The integrated signal processing tools make it possible to interpret both Q-TOF and Orbitrap data, enabling downstream tools to be written with fewer mass analyzer specific concerns. Additionally, the spectrum representation chosen for the implementation was built directly on the metadata rich mzML standard [244], and I use a streaming mzML serializer in the deconvolution pipeline to write the results, including source metadata and additional annotations acquired during processing to a standard compliant file.

Chapter 3

Glycan Identification and Glycome Inference from LC-MS and LC-MS/MS Experiments

This work was originally published in "Klein, J., Carvalho, L., & Zaia, J. (2018). Application of network smoothing to glycan LC-MS profiling. Bioinformatics, 34(20), 3511–3518. https://doi.org/10.1093/bioinformatics/bty397"

Glycosylation modulates the structures and functions of proteins and lipids in a broad class of biological processes [9]. Accurate mass measurement defines monosaccharide composition given assumptions regarding glycan class and biosynthesis [245]. For unseparated mixtures, mass spectrometry analysis determines the mass-to-charge ratio values for only the most abundant glycans; dynamic range for detection of glycans is poor because of ion suppression [246]. By contrast, online separations coupled with mass spectrometry improve dynamic range and reproducibility of glycan analysis, at the cost of increased analysis time and workflow complexity.

There are many tools for interpreting glycan mass spectral datasets [96, 97, 99, 141, 199, 246, 247] for both unseparated and separated experimental protocols. These programs address instrument-specific signal processing requirements. For example SysBioWare [247] performs sophisticated baseline removal prior to fitting peaks, while GlyQ-IQ [99] was written for cleaner Fourier Transform MS (FTMS) that does not require such a baseline removal step. Tools that build on the THRASH implementation from Decon2LS [96, 97, 119] are unable to deal with variable baseline noise or extreme dynamic range.

Each tool also has its own format for defining glycan structures or compositions, some even bundling a large database with their software to remove the burden from the user to build a list of candidates themselves [96, 99, 199] while others define methods for building glycan databases as part of the program [97, 141]. Many of these tools are designed for specific glycan subclass such as *N*-glycans or glycosaminoglycans and/or organisms, limiting their vocabulary of possible monosaccharides to just those commonly found in that subgroup [96, 99, 199, 246]. Often, these tools are tailored for analysis of a particular derivatization state, adduction conditions, or neutral loss pattern [96, 97, 246]. Work has been done to construct a standardized namespace and representation for glycans, glySpace including both structures and compositions [213, 214]. This data is publicly accessible, including a programmatic query interface using SPARQL over HTTPS [248]. Tools that can communicate with these services have the potential to lead researchers to find deeper connections from cross-referenced information, and other researchers can more readily find and use their work.

These spectral processing and glycan library properties are reflected in the scoring function that each program uses to discriminate glycan signal from the background noise and contaminants. Several methods have been developed using different facets of the observed data. [96] used the isotopic pattern goodness-of-fit while [246] used intensity features of associated MS^2 scans to evaluate partial structure and composition match quality. [99] combined several features of the MS^1 evidence, including elution profile peak shape goodness-of-fit, isotopic fit, mass accuracy, scan count, and in-source fragmentation correlation. Some of these methods are well-defined and invariant from instrument to instrument in this era of high resolution mass spectrometry, but others are tightly coupled to the experimental equipment. Missing from this list are methods to target a glycan's intrinsic properties, such as charge state distribution or facility in acquiring adducts, which can increase the number of spurious assignments if not considered. We propose a new scoring function which is able to combine those properties which are independent of experimental setup with these glycan-aware features.

As observed by [199], there is also value in including related glycan composition identifications in how much confidence one assigns to a given glycan composition assignment. They used a method to exploit the known biosynthetic rules of *N*-glycans to connect peaks in a MALDI mass spectrum assigned to a particular *N*-glycan by intact mass alone. Their method using the maximum weighted subgraph of the biosynthetic network had demonstrably better performance than chance with their expert system annotation method. [99] considered a similar idea with more emphasis on handling in-source fragmentation observed in LC-MS and LC-MS/MS experiments.

We extend this notion of a glycan family to cover more sectors of the biosynthetic landscape which we term "neighborhoods", and present an algorithm for learning the importance of each neighborhood from observed data, which can in turn be used to improve glycan composition assignment performance. We also apply our method using three different glycan composition search spaces to show how the underlying database can influence results. We present our method on typical *N*-glycans in humans, though our method can be applied to any variety of glycan composition whose monosaccharides can be described using IUPAC trivial names or whose components can be described in terms of chemical formulae.

3.1 Methods

3.1.1 Glycan Hypothesis Generation

In eukaryotes, a 14 monosaccharide *N*-glycan of composition `HexNAc2 Hex12` is transferred to a newly synthesized protein in the endoplasmic reticulum by the oligosaccharyl transferase protein complex. This glycan is trimmed to `HexNAc2 Hex9` during protein folding and quality control. As the glycoprotein transits the Golgi apparatus, *N*-glycans are trimmed to `HexNAc2 Hex5` before being elaborated into hybrid and complex *N*-glycan classes [5]. Glycan structures are refined by a series of reactions that yield over a million possible *N*-glycan topologies, as shown in [70]. These topologies define the glycan's geometry and protein binding properties. Neither MS^1 nor collisional tandem MS of glycans can capture the full tree or graph structure of an *N*-glycan, so we reduced the topology to a count of each type of residue, a composition.

Starting with the core motif `HexNAc2 Hex3`, we generated all combinations of monosaccharides ranging between the limits in Table 3.1 to build a human *N*-glycan composition database, which produced 1240 distinct compositions. These rules are able to efficiently generate all glycan compositions from canonical branching patterns and lactosamine extensions, as well as rarer constructs such as LacdiNAc [199] at the cost of including some wholly improbable compositions. To perform a side-by-side comparison we also extracted the glycan list from [96]derived from the biosynthetic rules in [249] with 319 compositions, and another database using all curated *N*-glycans from glySpace via GlyTouCan [213] containing only `[Hex, HexNAc, Fuc, Neu5Ac, sulfate]`, with 275 distinct compositions. As previous analysis of Influenza A virus samples detected sulfated *N*-glycans [23], we also created a com-

Monosaccharide	Lower Limit	Upper Limit	Constraints
HexNAc	2	9	
Hex	3	10	
Fuc	0	4	$\texttt{HexNAc} > \texttt{Fuc}$
NeuAc	0	5	$(\texttt{HexNAc} - 1) > \texttt{NeuAc}$

Table 3.1: Human *N*-glycan Composition Bounds [5]

binatorial database with up to one sulfate included, for a total of 2480 compositions. As our algorithm treats HexNAc and HexNAc(S) as distinct entities, for all monosaccharides with post-attachment substituents such as sulfate and phosphate, we detached the substituent from the core monosaccharide. Our implementation is able to interpret IUPAC trivial names and compositions thereof with standard substituent and unambiguous backbone modifications, permitting a wide range of possible glycan compositions.

3.1.2 LC-MS Data Preprocessing

We analyzed samples from several sources, including both Q-TOF and Orbitrap instruments as shown in Table 3.6. For details on sample preparation and data acquisition, please see the source citations in the referenced table. We converted all datasets to mzML format [244] using Proteowizard [120] without any data transforming filters. We applied the deconvolution procedure described in Chapter 2 using an averagine [232] formula appropriate to the molecule under study. For native glycans, the formula was H1.690C1.0O0.738N0.071, for permethylated glycans, the formula was H1.819C1.0O0.431N0.042.

3.1.3 Chromatogram Aggregation

We clustered peaks whose neutral masses were within $\delta_{mass} = 15$ parts-per-million error (PPM) of each other. When there were multiple candidate clusters for a single peak, we used the cluster with the lowest mass error. Next, we sorted each cluster by time, creating a list of aggregated chromatograms. To account for small mass differences, we found all chromatograms which were within $\delta_{mass} = 10$ PPM of each other and which overlap in time and merge them. These mass tolerances were selected empirically, and can be adjusted as needed by the user.

3.1.4 Glycan Composition Matching

For each chromatogram, we searched each glycan database for compositions whose masses were within $\delta_{mass} = 10$ PPM for QTOF data, 5 PPM for FTMS data. These values are commonly used for data from these instruments based upon information from their manufacturers. We merged all chromatograms matching the same composition. Then, for each mass shift combination expected for each sample, we searched each glycan database for compositions whose neutral mass were within δ_{mass} of the observed neutral mass - mass shift combination mass, followed by another round of merging chromatograms with the same assigned composition. We reduced the data by splitting each feature where the time between sequential observation was greater than $\delta_{rt} = 0.25$ minutes and removed chromatograms with fewer than $k = 5$ data points. The same chromatogram may be given multiple assignments and designated multiple mass shifts, and chromatograms without glycan assignments may use chromatograms with glycan assignments as mass shifted components. This ambiguity information was propagated through each merge and split step. We termed these remaining assigned and unassigned chromatograms *candidate features*.

3.1.5 Feature Evaluation

We computed several metrics to estimate how distinguishable each candidate feature was from random noise. The metrics are mentioned in List 3.1, but for more information see Sec. 3.5.3.

1. Goodness-of-fit of chromatographic peak shape to a model function [99, 250].
2. Goodness-of-fit of isotopic pattern to glycan composition weighted by peak abundance [97].
3. Observed charge states with respect to glycan composition and mass.
4. Time gap between MS^1 observations detecting missing peaks and interference.
5. Adduction states with respect to glycan composition and mass.

List 3.1: Chromatographic Feature Metrics

These metrics are bounded in $(-\infty, 1)$. Any observation for which any metric was observed below a feature specific threshold was discarded as having insufficient evidence for consideration. The observed score s for each candidate feature is the sum of the logit-transformation of these metrics. This produces a single value bounded in $(-\infty, \infty)$, whose distribution we assume is asymptotically normal. A value of $s < 8$ reflects a low confidence match, with confidence increasing as s does. As these metrics are tied to reliable detection of the the glycan by the mass spectrometer, they depend upon glycan abundance, sample quality and mass spectrometer resolution.

3.1.6 Glycan Composition Network Smoothing

Ideally, each glycan present in a sample under analysis would produce sufficient experimental evidence that they can be identified. In practice, glycan compositions with lower abundances may not present strong evidence, leading to those glycan compositions being discarded. Others have demonstrated that it is advantageous to use relationships between glycans based on biosynthetic or structural rules to adjust the score of a single glycan assignment [99, 199]. To improve performance, we propose a method based on Laplacian regularized least squares [251] to use evidence from glycan compositions related over a network to smooth its evaluation of glycan composition feature matching.

Previous approaches to using information regarding identification of one glycan composition to increase the confidence in another have been proposed by [199] and [99] using different techniques. [199] used random walks along the biosynthetic network between identified glycan compositions to increase the confidence of those connected compositions. This method works well but requires that the parameters of the random walk be properly tuned for the biosynthetic network being used. Laplacian regularized least squares is more robust to small changes to the network and is able to use the entire network. [99] included a term in their criterion for detection requiring the presence of another glycan composition with one more or one less monosaccharide to permit identification. This puts substantial weight on a boolean term, giving it the ability to overrule other experimental evidence. Similar methods could be devised using methods like ant colony optimization to traverse the biosynthetic graph, or a a database-specific belief network, but these methods would require considerable manual tuning for each new database to be tested.

Glycan Composition Graph

For each database of theoretical glycan compositions we create, we define each composition to be a coordinate vector in a $\mathcal{Z}^{+c}$ space where c is the number of components in any glycan composition, and represented by a node in an undirected glycan composition graph $\mathcal{G}$. Under this interpretation, we can compute the L_1-distance between two glycan compositions, representing the biosynthetic distance between the two compositions, an analog for the number of enzymatic steps needed to go from one glycan to the other. For any two glycan compositions g_u, g_v, if $L_1(g_u, g_v) = 1$ we add an edge connecting g_u and g_v to $\mathcal{G}$ with weight $w = 1$.

Neighborhood Definition

Our definition of distance connects glycan compositions which differ by a single monosaccharide, but we can assert how larger collections of glycan compositions are related. To this end, we extend the definition of neighborhoods for *N*-glycansusing intervals over monosaccharide counts shown in Table 3.2. These neighborhoods are arranged to span particular epitopes or biosynthetically related subtypes of *N*-glycans, such as sialylation state or branching pattern. Neighborhoods overlap sets of glycan compositions which are also biosynthetically related. Each neighborhood spans the eponymous class of glycan compositions, as well as the preceding class and proceeding class. For example, the Tri-Antennary neighborhood spans Bi-Antennary and Tetra-Antennary compositions. This helps to channel the estimation of τ among related groups. The Hybrid, Bi-Antennary and Asialo-Bi-Antennary neighborhoods introduce complications because they are biosynthetically close to each other. For the simplicity, we chose to include all of Hybrid in Asialo-Bi-Antennary and permit up to one NeuAc in its members.

Glycan compositions may belong to zero or more neighborhoods, as there are unusual glycan compositions which do not satisfy any neighborhood's rules, and several neighborhoods intentionally overlap to express broad relationships between groups.

We define a matrix $\mathbf{A}$ as an $n \times k$ matrix where $A_{i,k}$ is the degree to which g_i

	HexNAC		Hex		NeuAc		Size
Name	Min	Max	Min	Max	Min	Max	
High Mannose	2	2	3	10	0	0	16
Hybrid	2	4	2	6	0	2	80
Bi-Antennary	3	5	3	6	1	3	104
Asialo-Bi-Antennary	3	5	3	6	0	1	96
Tri-Antennary	4	6	4	7	1	4	172
Asialo-Tri-Antennary	4	6	4	7	0	0	56
Tetra-Antennary	5	7	5	8	1	5	240
Asialo-Tetra-Antennary	5	7	5	8	0	0	60
Penta-Antennary	6	8	6	9	1	5	280
Asialo-Penta-Antennary	6	8	6	9	0	0	60
Hexa-Antennary	7	9	7	10	1	6	300
Asialo-Hexa-Antennary	7	9	7	10	0	0	60
Hepta-Antennary	8	10	8	11	1	7	150
Asialo-Hepta-Antennary	8	10	8	11	0	0	30

Table 3.2: N-Glycan Neighborhood Definitions. These define the ranges of monosaccharides which will be used to classify a glycan composition as being a member of each neighborhood, and the number of *combinatorial* *N*-glycan compositions in each neighborhood.

belongs kth neighborhood:

$$A_{i,k} = \frac{1}{|\text{neighborhood}_k|} \sum_{g^* \in \text{neighborhood}_k} L_1(g_i, g^*) \tag{3.1}$$

To reduce the impact of neighborhood size on the elements of $\mathbf{A}$, the columns of $\mathbf{A}$ are first normalized to sum to 1, and then the rows of $\mathbf{A}$ are normalized to sum to 1. We assume that members of the same neighborhood will share a central tendency τ.

Laplacian Regularization

To accomplish our goal, we can use Laplacian regularized least squares to find a new score ϕ, based upon s and relationships among the observed glycans described by our biosynthetic graph $\mathcal{G}$. These relationships can be directed to move towards some central tendency τ using the Laplacian of $\mathcal{G}$ and some definitions of broad groups in $\mathcal{G}$.

We combine the observed score $\mathbf{s}$ and the structure of $\mathcal{G}$ to estimate a smoothed score ϕ that combines the evidence for each individual glycan composition as well as its relatives. As $\mathbf{s}$ is the size of the set of observed glycan composition p while ϕ is of size n, we partition ϕ into a block vector $\begin{bmatrix} \phi_o \\ \phi_m \end{bmatrix}$ with dimensions $\begin{bmatrix} p \\ n-p \end{bmatrix}$.

Let $\mathbf{L}$ be the weighted Laplacian matrix of $\mathcal{G}$, which is an $n \times n$ matrix. To ensure $\mathbf{L}$ is invertible, we add $\mathbf{I}_n$ to $\mathbf{L}$. We partition $\mathbf{L}$ into blocks $\begin{bmatrix} \mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}} \end{bmatrix}$. We also partition $\mathbf{A}$ into $\begin{bmatrix} \mathbf{A}_o \\ \mathbf{A}_m \end{bmatrix}$ and $\tau_o = \mathbf{A}_o\tau$, $\tau_m = \mathbf{A}_m\tau$.

We find the ϕ that minimizes the expression

$$\ell = (\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda\mathcal{S}(\mathbf{L}, \phi, \tau) \tag{3.2}$$

$$\mathcal{S}(\mathbf{L}, \phi, \tau) = \begin{bmatrix} \phi_o - \tau_o \\ \phi_m - \tau_m \end{bmatrix}^t \begin{bmatrix} \mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}} \end{bmatrix} \begin{bmatrix} \phi_o - \tau_o \\ \phi_m - \tau_m \end{bmatrix} \tag{3.3}$$

$$\tag{3.4}$$

where λ controls how much weight is placed on the network structure and τ.

To obtain the optimal ϕ, we take the partial derivative of ℓ w.r.t ϕ_m:

$$0 = \frac{\partial \ell}{\partial \phi_m}\Big((\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda\mathcal{S}(\mathbf{L}, \phi, \tau)\Big) \tag{3.5}$$

$$\hat{\phi}_m = -\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o - \tau_o) + \tau_m \tag{3.6}$$

and w.r.t. ϕ_o

$$0 = \frac{\partial \ell}{\partial \phi_o}\left((\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda \mathcal{S}(\mathbf{L}, \phi, \tau)\right) \tag{3.7}$$

$$\hat{\phi}_o = \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right]^{-1}(\mathbf{s} - \tau_o) + \tau_o \tag{3.8}$$

To use this method, we must provide values for λ and τ. While these values could be chosen based on the expectations of the user for a given experiment, we provide an algorithm for selecting their values in Section 3.5.5. These methods use the topology of the glycan composition graph and the distribution of observed scores, and cannot fully capture boundary cases or related but disconnected parts of the graph.

3.2 Results

We demonstrated the performance of our algorithm using released influenza hemagglutinin data set *20141103-02-Phil-BS* and a serum glycan data set *Perm-BS-070111-04-Serum*. Please refer to section 3.5.7 for all other datasets. For each comparison, the unregularized case is not smoothed, effectively $\lambda = 0$, the partially regularized case uses the grid search fitted values of τ but uses a fixed $\lambda = 0.2$, and the fully regularized case uses the grid search fitted values of both τ and λ.

3.2.1 Chromatogram Assignment Performance for *20141103-02-Phil-BS*

The fitted parameters for the network constructed for *20141103-02-Phil-BS* are shown in Table 3.3. The assigned chromatograms are shown in Figure 3.1. We observe up

Neighborhood τ	Phil-BS Combinatorial + Sulfate	 glySpace	 Krambeck	Serum Combinatorial	 glySpace	 Krambeck
high-mannose	18.008	15.061	17.089	20.328	19.392	19.720
hybrid	13.440	12.435	12.503	20.997	18.610	20.056
bi-antennary	0.000	0.000	0.000	15.901	16.826	17.593
asialo-bi-antennary	14.078	10.916	13.591	22.585	21.563	21.827
tri-antennary	0.000	0.000	0.000	26.420	19.605	23.644
asialo-tri-antennary	14.538	6.565	11.952	20.025	21.128	19.764
tetra-antennary	0.000	0.000	0.000	19.508	18.542	17.674
asialo-tetra-antennary	14.331	4.842	12.373	2.472	7.180	2.568
penta-antennary	0.000	0.000	0.000	11.878	15.035	11.682
asialo-penta-antennary	11.588	1.255	9.784	0.000	0.000	0.000
hexa-antennary	0.000	0.000	0.000	0.000	0.000	0.000
asialo-hexa-antennary	11.094	3.883	13.223	0.000	0.000	0.000
hepta-antennary	0.000	0.000	0.000	0.000	0.000	0.000
asialo-hepta-antennary	3.117	1.529	2.703	0.000	0.000	0.000
$\hat{\lambda}$	0.99	0.69	0.99	0.99	0.99	0.99
$\hat{\gamma}$	11.39	14.60	10.42	20.57	18.42	20.72

Table 3.3: Estimated values of smoothing parameters τ, λ, and γ for each dataset and database

to seven branch structures in this sample, consistent with these *N*-glycans being derived from an avian context (5, 23).

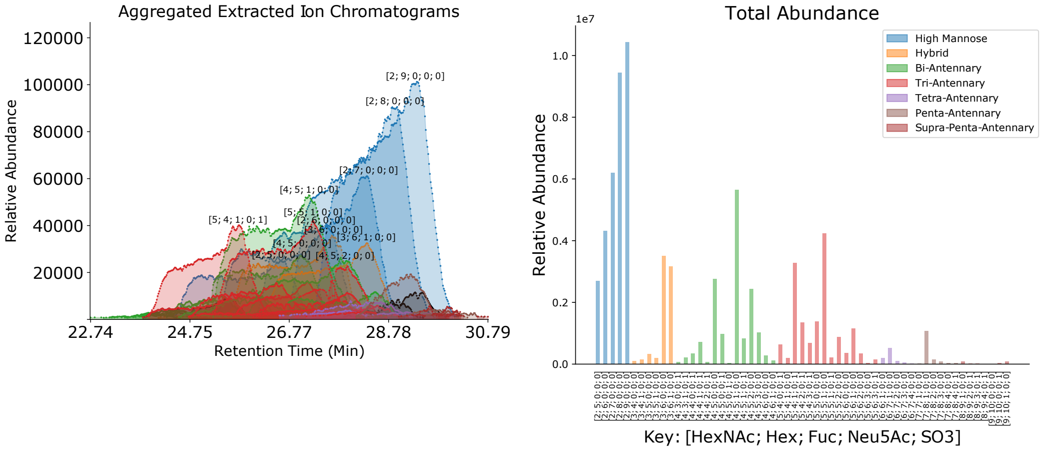


Figure 3.1: Chromatogram Assignments and Quantification for *20141103-02-Phil-BS*Using the *Combinatorial + Sulfate* database. The Retention Time (Min) axis shows the experimental retention time in minutes, and the Relative Abundance axis shows the intensity of the signal from each aggregated ion species. The identified glycan compositions are labeled with a tuple describing the number of each component of the form [`HexNAc`, `Hex`, `Fuc`, `NeuAc`, `SO3`]

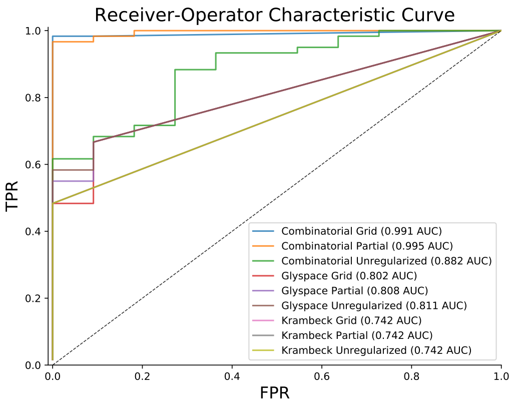


Figure 3.2: Performance Comparison with and without Network Smoothing for *20141103-02-Phil-BS*. The ROC comparing TPR to FPR shows how each database performed under different regularization conditions, summarized with the AUC in the legend. The *Combinatorial + Sulfate* database showed the best performance, and improved with regularization.

The comparison of assignment performance with differing degrees of smoothing for each database are shown in Figure 3.2 and Table 3.4. We used the Receiver Operator Characteristic (ROC) Area Under the Curve (AUC) to measure performance, using manually validated compositions as ground truth. We observed the greatest number of assignments using the *Combinatorial + Sulfate* database, and the greatest ROC AUC in the partially regularized condition.

Name	ROC AUC	True Matches[a]
Combinatorial Unregularized	0.882	56
Combinatorial Partial	0.995	57
Combinatorial Grid	0.991	57
GlySpace Unregularized	0.811	40
GlySpace Partial	0.808	38
GlySpace Grid	0.802	31
Krambeck Unregularized	0.742	28
Krambeck Partial	0.742	29
Krambeck Grid	0.742	29
[23]	-	46

[a] Selected at $\phi_o > 5.0$

Table 3.4: Performance Comparison for *20141103-02-Phil-BS* using ROC AUC. The Combinatorial Partial Regularization approach performed best.

3.2.2 Chromatogram Assignment Performance for *Perm-BS-070111-04-Serum*

The fitted parameters for the network constructed for *Perm-BS-070111-04-Serum* are shown in Table 3.3. The assigned chromatograms are shown in Figure 3.4.

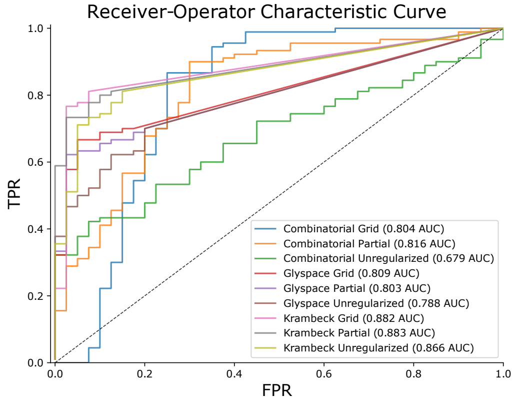


Figure 3.3: Performance Comparison with and without Network Smoothing for *Perm-BS-070111-04-Serum* . The Receiver Operator Characteristic Curve (ROC) comparing True Positive Rate (TPR) to False Positive Rate (FPR) shows how each database performed under different regularization conditions, summarized with the Area Under the Curve (AUC) in the legend

The comparison of assignment performance with differing degrees of smoothing is shown in Figure 3.3. We observe the greatest number of total true identifications using the partially regularized Combinatorial database. However, the Combinatorial database also has many more false positives, with a ROC AUC of 0.816. These false positives do not appear in the biosynthetically constrained Krambeck database which maximizes its ROC AUC in the partially regularized condition at 0.883. After removing all ambiguous matches, the Krambeck database also has nearly the same number of true matches as the Combinatorial database.

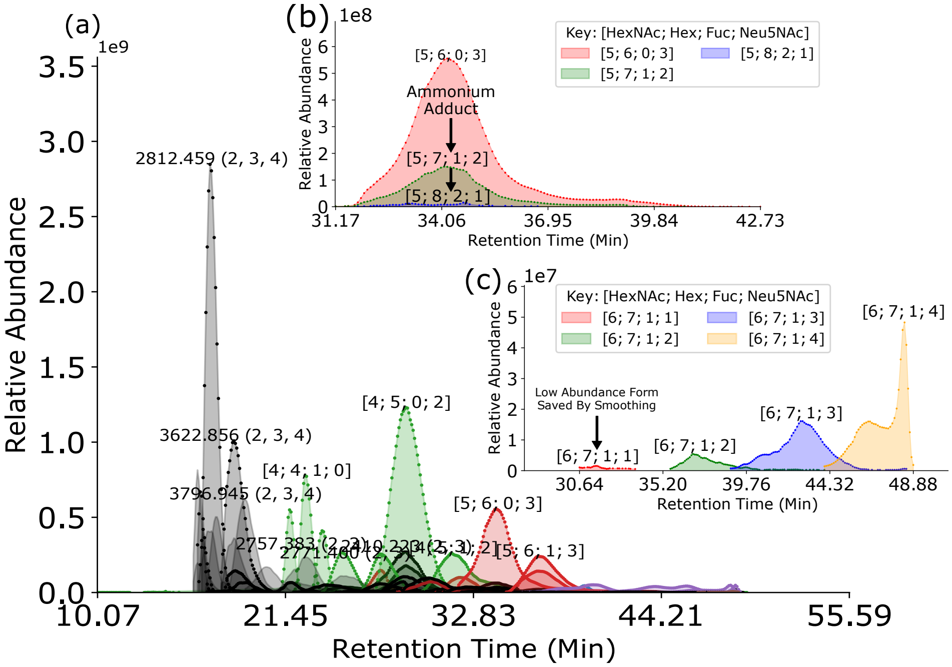


Figure 3.4: Chromatogram Assignments for *Perm-BS-070111-04-Serum*. In all panels, the Retention Time (Min) axis shows the experimental retention time in minutes, and the Relative Abundance axis shows the intensity of the signal from each aggregated ion species. The identified glycan compositions are labeled with a tuple describing the number of each component of the form [HexNAc, Hex, Fuc, NeuAc] (a) Features Assigned After Grid Regularization of *Perm-BS-070111-04-Serum* (b) This sample contains heavy ammonium adduction which introduces ambiguity in intact mass based assignments (c) Low scoring features which may be discarded based on individual evidence alone may be more reasonable to accept given evidence from related composition, such as our network smoothing method

Name	ROCAUC	True Matches[a]	Non-Ambiguous Matches
Combinatorial Unregularized	0.679	86	61
Combinatorial Partial	0.816	87	62
Combinatorial Grid	0.804	86	61
GlySpace Unregularized	0.788	59	51
GlySpace Partial	0.803	60	52
GlySpace Grid	0.809	60	52
Krambeck Unregularized	0.866	70	60
Krambeck Partial	0.883	70	60
Krambeck Grid	0.882	69	59
[96]	-	72[b]	59

[a] Selected at $\phi_o > 5.0$

[b] Only includes cases with sufficient MS1 scans available for comparison

Table 3.5: Performance Comparison for *Perm-BS-070111-04-Serum* using Receiver Operator Characteristic (ROC) Area Under the Curve (AUC) and number of non-ambiguous matches. While the Krambeck database had a better ROC AUC, the Combinatorial database had more true matches.

3.3 Discussion

We demonstrated that the regularization method improved the sensitivity and specificity of glycan composition assignment for LC-MS based experiments. The method used similar assumptions about the importance of common sub-structural elements of *N*-glycans to [199], but we extend this concept with the addition of a procedure for learning the relationship strengths and use broader groups of structures.

The experimental results from the original analysis of *20141103-02-Phil-BS* and *20141031-07-Phil-82*82 demonstrated that while both strains expressed predominantly high-mannose glycosylation, *20141103-02-Phil-BS* expressed more larger complex-type structures [23]. In our findings shown in Figure 3.1, we recapitulate these results while reducing the number of false assignments, Table 3.4. There are substantial differences in both the mass spectral processing and scoring schemes which contribute to these results, but the regularization procedure is responsible for recovering many low abundance features from this comparison. As these samples are derived from chicken eggs, we have observed larger branching patterns than are observed in normal mammalian tissue [5]. There is evidence for this in the *20141103-02-Phil-BS* with `HexNAc9 Hex10`-based compositions suggesting a seven branch pattern, though this cannot be determined without high quality MS^2 data. The τ fit for Phil-BS (shown) and Phil-82 (supplement) have smaller values in the neighborhoods of their largest glycan compositions as these features tended to be low in abundance and not high scoring in their own right, but were partially supported by the overlap with the next largest neighborhood, as expected. We observed the best performance with the *Combinatorial + Sulfate* database, which produced more than half-again as many true matches than the other two databases. It produced several false matches as well, but the smoothing process removed these while boosting the score of other low abundance matches which were consistent with higher scoring matches.

The Krambeck database performed identically in all smoothing conditions as it was only able to match the common species, not including cases that were multiply fucosylated or sulfated. It had no false matches ranked alongside its true matches so smoothing could not change its performance. The glySpace-derived database produced more true matches, but also lacked some of these more fucosylated and complex compositions. Some of the compositions included by the glySpace-derived database were lower scoring, but the chosen value of γ for that database was greater than 18, causing the fitted values of τ to omit the larger, less abundant complex-type *N*-glycans. This caused smoothing to lower the scores of these real matches rather than raise them, as with the *Combinatorial + Sulfate* database.

As we show in Figure 3.3, regularization improves the predictive performance of the identification algorithm on *Perm-BS-070111-04-Serum* for all databases. We reproduce the majority of the glycan assignments from [96], but the ambiguity caused by ammonium adduction as shown in Figure 3.4 makes a direct comparison of composition assignment lists difficult. Our algorithm requires a minimum amount of MS1 information in order to compute a score, which some of the assignments in the original published results do not possess, and are omitted from the count in Table 3.5. After accounting for ambiguity, we were able to assign all of the compositions previously reported using the Krambeck database, which was used by [96], and with the combinatorial database. The glySpace-derived database did not contain all of these compositions, but performed competitively with the combinatorial database's ROC AUC. The combinatorial database matched a small number of glycan compositions which were not in Krambeck but which were consistent with other glycan compositions observed nearby in retention time. The combinatorial database also benefited most substantially from smoothing, discarding many false positives while retaining many more true positives at the same false positive rate compared to the other databases. These invalid glycan compositions can match LC-MS features at any point in the elution profile, though in this dataset the majority of these matches appear to be in the time range between 10 and 22 minutes, and similar glycan compositions that are biosynthetically valid elute later on in the experiment. Therefore a for a retention-time aware approach to evaluating glycan composition assignments, as described in [203] could also be useful, but this is likely dependent upon the experimental workup and separation technique used.

While the biosynthetically constrained Krambeck database performed better on *Perm-BS-070111-04-Serum*, it did not contain all of the reasonably assignable glycan compositions, and it performed poorly on *20141103-02-Phil-BS* with a false negative rate of 50% compared to the combinatorial database. This is because the necessary enzymatic pathways were either not considered in the original authors' model because either the enzyme was excluded for simplicity [68] or because the particular enzymes used were not within the scope of the model used [252, 253]. This highlights the importance of selecting a good reference database, though a post-processing step such as the we described here can help mitigate using too large a database, but not a too small one.

In this work, we used the same network neighborhood imposed over different underlying sets of composition nodes, and the connectivity of those networks did not take into account the constraints of the biosynthetic process. It may be possible to obtain better performance by defining network connectivity according to concrete enzymatic relationships. This may also alter how the neighborhoods are defined and how $\mathbf{A}$ is parameterized, and in turn how τ is learned. Similarly, this procedure depends upon the scoring functions used, so selecting another set of functions for the data to fit may lead to different parameter values.

Lastly, while these case studies have demonstrated the algorithm's ability to learn network parameters from the data, an expert can define τ and $\mathbf{A}$ themselves or obtain a model fitted on related data and apply it directly without a fitting step. An expert could use this model specification to impose prior beliefs on the evaluation process, and adjust λ to control the importance of the these beliefs. Similarly, one could also use the derivation of $\hat{\phi}_m$ to estimate the score for an unobserved glycan composition, given $\mathbf{A}$ and τ.

We used our glycoinformatics toolkit to produce a richer abstraction of glycans and monosaccharides, including producing standard-compliant textual representations of these structures and compositions. We produced a text file containing all of the glycan compositions found in the Krambeck and Combinatorial database but not the glySpace-derived database in the above samples (see Sec. 3.5.9), and have submit it to GlyTouCan [213] for registration so that future researchers can use these structures.

3.4 Conclusions

In this study, we demonstrated the advantages of our application of Laplacian Regularization to smooth LC-MS assignments of glycan compositions across multiple experimental protocols [23, 254]. Our algorithm's performance is competitive with existing tools for analyzing the same type of data, with the added benefit of more flexible evaluation process and broader range of understood monosaccharides. Our tools integrate with glySpace and allows users to leverage existing glycomics repositories to build databases where applicable.

All of the methods demonstrated in this paper are available as part of the open source, cross-platform glycomics and glycoproteomics software `GlycReSoft`, freely available at http://www.bumc.bu.edu/msr/glycresoft/.

3.5 Supplemental Materials

3.5.1 Experimental Samples Used

We demonstrate our algorithm on several samples from a variety of instruments and conditions described in Table 3.6. We present two samples in the main text, Q-TOF analysis of Native, Formate adducted *N*-glycans from Influenza strain Phil-BS virions *20141103-02-Phil-BS*, and Orbitrap analysis of Permethylated and Reduced Ammonium adducted *N*-glycans from human serum *Perm-BS-070111-04-Serum*.

Sample Name	Instrument	Derivatization	Adduction	Source	Taxon
20150930-06-AGP	Q-TOF	Native	Formate (1)	[23]	Human
20141031-07-Phil-82	Q-TOF	Native	Formate (3)	[23]	Human Virus in Avian Tissue
20141103-02-Phil-BS	Q-TOF	Native	Formate (3)	[23]	Human Virus in Avian Tissue
20151002-02-IGG	Q-TOF	Native	Formate (2)	[243]	Human
20141128-11-Phil-82[1]	Q-TOF	Deutero-reduced, Permethylated	Ammonium (3)	[23]	Human Virus in Avian Tissue
AGP-DR-Perm-glycans-1[1]	Orbitrap	Deutero-reduced, Permethylated	Ammonium (3)	[23]	Human
AGP-permethylated-2ul-inj-55-SLens[1]	Orbitrap	Reduced, Permethylated	Ammonium (3)	[23]	Human
Perm-BS-070111-04-Serum[1]	Orbitrap	Reduced, Permethylated	Ammonium (3)	[96, 254]	Human

[1] Included MS^n Scans

Table 3.6: Samples Used

As Table 3.6 describes, we analyze data from several different combinations of configurations of instrument, derivatization, and reduction.

3.5.2 Database Comparison

The three databases we used were overlapping but distinct. The size of these overlaps is shown in Figure 3.5.

3.5.3 Chromatographic Feature Evaluation

For each candidate feature, we computed several metrics to estimate how distinguishable the observed signal was from random noise. We use the quantities de-

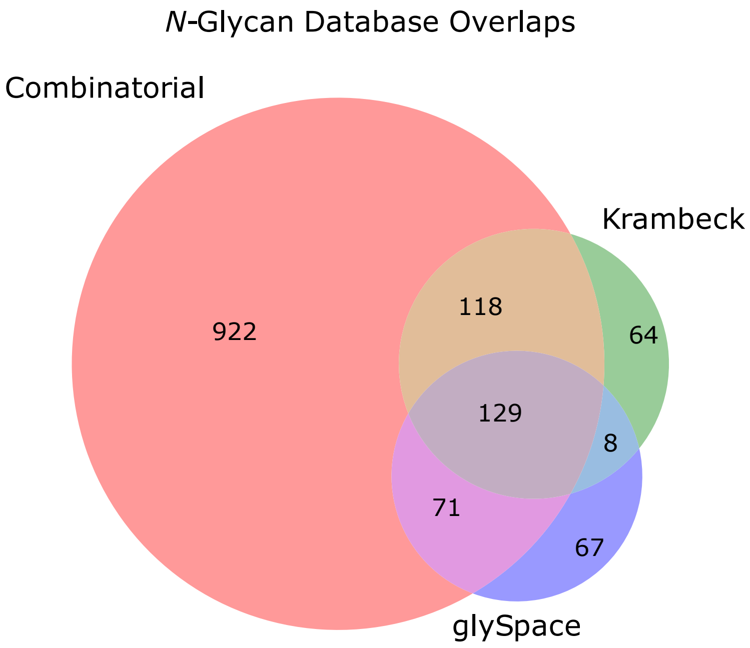


Figure 3.5: The overlap of the source databases used. As expected, the combinatorial database contains an enormous number of compositions not found in either other database, many of which are not biosynthetically feasible for humans. Those found in the Krambeck database but not the combinatorial or glySpace database are derived from lactosamine extensions run to the limit of the biosynthetic process covered in the original simulation [249]. The glySpace database contained composition units not found in the other two databases, such as Xylose, Sulfate, and Phosphate.

scribed in Table 3.7 from each LC-MS feature.

All metrics are penalized by an $\epsilon = 1e{-}6$ to prevent scores from actually achieving a value of 1.0 which would make the logit value infinite. If a metric's value would be less than $0 + \epsilon$, it is given a value of ϵ instead to prevent the logit value from being undefined.

Table 3.7: Chromatogram Feature Definitions

$\mathcal{M}_i$	The neutral mass of the ith chromatogram
$\mathcal{I}_i$	The total intensity array assigned to the ith chromatogram
$\mathcal{I}_{i,j}$	The sum of all peak intensities for peaks observed in the jth scan for the ith chromatogram
$\mathcal{I}_{i,j,k}$	The intensity assigned to the kth peak at the jth scan for the ith chromatogram
$\mathbf{c}_i$	The set of charge states observed for the ith chromatogram
$\mathcal{I}_{i,c=j}$	The total intensity assigned to the ith chromatogram with charge state j
$\mathbf{t}_{i,j}$	The time of the jth scan of the ith chromatogram
T_j	The time of the jth scan of the experiment
$\textbf{env}_{i,j,k}$	The normalized experimental isotopic envelope composing the kth peak of the jth scan of the ith chromatogram, whose members sum to 1
$\mathbf{a}_i$	The set of adduction states observed for the ith chromatogram
$\mathcal{I}_{i,a=j}$	The total intensity assigned to the ith chromatogram with adduct j
$\hat{g}_i$	The glycan composition assigned to the ith chromatogram, or Ø if there was no matched glycan composition

Chromatographic Peak Shape

An LC-MS elution profile should be composed of one or more peak-like components, each following a bi-Gaussian peak shape model [250] or in less ideal chromatographic circumstances, a skewed Gaussian peak shape model. We fit these models using non-linear least squares (NLS). As measures of goodness of fit are not generally available for NLS, we use the following criterion:

$$\begin{aligned}
\hat{y}_i &= NLS(\mathcal{I}_i, \mathbf{t}_i) \\
e_{i,NLS} &= \mathcal{I}_i - \hat{y}_i \\
\bar{y}_i &= \mathbf{t}_i \left(\left(\mathbf{t}_i^t \mathbf{t}_i \right)^{-1} \mathbf{t}_i \mathcal{I}_i \right) \\
e_{i,null} &= \mathcal{I}_i - \bar{y}_i \\
\mathscr{L}_i &= 1 - \frac{\sum e_{i,NLS}^2}{\sum e_{i,null}^2} \qquad (3.9)
\end{aligned}$$

where line score describes how much the peak shape fit improves on a ordinary least squares regression linear model.

We apply two competitive peak fitting strategies to address distorted, overlapping, or multimodal elution profiles. The first works iteratively by finding a best-matching peak shape using non-linear least squares, subtracting the fitted signal and checks if there is another peak with at least half as tall as the removed peak, if so repeating the process until no peak can be found, saving each peak model so constructed. The second approach starts by locating local minima between putative peaks, and partitioning the chromatogram into sub-groups which would are fit independently. This method generates a candidate list of minima, and selects the case which has the greatest difference between the minimum and its pair of maxima to split the feature at. The strategy which produces the maximum $\mathscr{L}_i$ is chosen. $\mathscr{L}_i$ is bounded in $(-\infty, 1]$, where 1 corresponds to a perfect fit, and 0 would correspond to the peak shape fit being no better than the OLS straight line fit. This metric is thresholded at 0.15, with any chromatogram scoring below 0.15 being discarded as having insufficient peak shape evidence to interpret.

Composition Dependent Charge State Distribution

As the number of monosaccharides composing a glycan increases, the number of possible sites for charge localization increases. This relationship is visualized in Figure 3.6. Under normal conditions, we would expect to observe the same molecule in multiple charge states [97] . Which charge states are expected would depend upon the size of the molecule and it's constituent units' electronegativity. In it's native state, NeuAc's acidic group causes glycans with one or more NeuAc to have a propensity for higher negative charge states [10] . To capture this relationship, we modeled the probability of observing a glycan composition for sialylated and unsialylated compositions separately. For permethylated glycans, charge is carried by protons or metallic cation adducts like sodium, the relationship between acidic monosaccharides and charge state propensities is weaker.

$$
\begin{aligned}
m_i &= (\lfloor(\mathcal{M}_i/w)/10\rfloor + 1) * 10 \\
\mathcal{H}_{i,j} &= \frac{\mathcal{I}_{i,c=j}}{\mathcal{I}_i} \\
P(c, m) &= \frac{\sum_{m_i \in m} \mathcal{H}_{i,j}}{\sum_j \sum_{m_i \in m} \mathcal{H}_{i,j}} \\
\mathscr{C}_i &= \sum_{c_{i,j} \in \mathbf{c}_i} P(c_{i,j}, m_i)
\end{aligned} \tag{3.10}
$$

where w is the width of the mass bin divided by 10 and $P(c, m)$ is defined as part of the model estimation procedure. If the model is complete, then this metric is bounded within $[0, 1]$ where 0 corresponds to having no observed charge states and 1 corresponds to all expected charge states being observed. In practice, the model is not complete, where an existing mass range may be missing a charge state in which case $P(c, m)$ is the average over all known values of c in m. When a mass range is required but missing from the model, the model will fall back to a naive model where

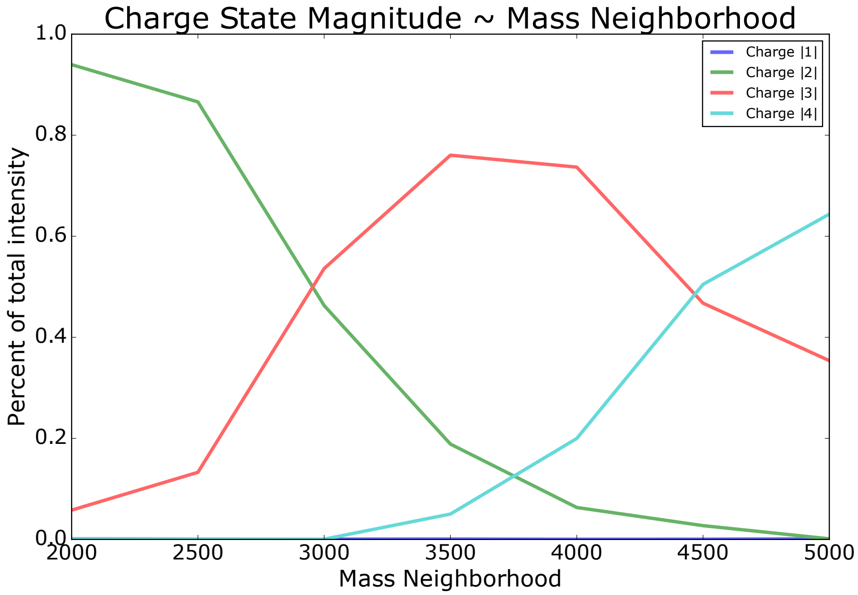


Figure 3.6: The trend of charge state relative abundance for acidic glycans

$P(c, m) = 0.4 \,\forall\, c$ and as such this metric must be clamped to not exceed 1.0. This metric has an exceptional threshold of 0.05 instead of 0.15.

Adduction Frequency

For the datasets *AGP-permethylated-2ul-inj-55-SLens* and *Perm-BS-070111-04-Human-Serum* we also include an adduction frequency model score $\mathscr{A}_i$, following the same pattern as the charge state distribution, with the same extension of justification from [97]. We use one mass scaling model for all glycan compositions as ammonium adduction is not expected to be composition dependent.

$$\begin{aligned} m_i &= (\lfloor(\mathcal{M}_i/w)/10\rfloor + 1) * 10 \\ \mathcal{H}_{i,j} &= \frac{\mathcal{I}_{i,a=j}}{\mathcal{I}_i} \\ P(a, m) &= \frac{\sum_{m_i \in m} \mathcal{H}_{i,j}}{\sum_j \sum_{m_i \in m} \mathcal{H}_{i,j}} \\ \mathscr{A}_i &= \sum_{a_{i,j} \in \mathbf{a}_i} P(a_{i,j}, m_i) \end{aligned} \tag{3.11}$$

We fit an ammonium adduction model on *AGP-permethylated-2ul-inj-55-SLens* in order to make our comparison to third-party data less biased given limited sample data. This metric is bounded within $[0, 1]$ where 0 corresponds to having no observed adduction states within the model and 1 corresponds to all observing all adduction states in the model. This metric follows the same behavior as the charge state distribution metric w.r.t. missing information within the model, but will reject chromatograms when this metric score is below 0.15.

We fit a sialylation-aware formate adduction model on a collection of sialylated and unsialylated native *N*-glycan samples from replicates of the *20150930-06-AGP*, *20151002-02-IGG*, and *20141031-07-Phil-82* datasets. This model was used for *20150930-06-AGP*, *20151002-02-IGG*, *20141031-07-Phil-82* and *20141103-02-Phil-BS*. This model had its upper limit set to 0.7, so it could not contribute a large positive number to the score of a match after logit transformation. This is desirable because we want to be able to eliminate matches which are made with improbable formate adducts when no reasonable adduction state is present.

Isotopic Pattern Consistency

Our ahead-of-time deconvolution procedure uses an averagine isotopic model and does not capture the consistency of the isotopic pattern that was fit with the isotopic pattern of the glycan composition that matched that peak. The criterion

$$\mathscr{I}_i = 1 - 2\mathcal{I}_i^{-t}\mathbf{I}_i \sum_j^J \sum_k^K \mathcal{I}_{i,j,k} \textbf{env}_{i,j,k}^t \left(\ln \textbf{env}_{i,j,k} - \ln \textbf{tid}_i\right) \tag{3.12}$$

where **tid** is the theoretical isotopic pattern derived from either $\hat{g}_i$ or an averagine interpolated for $\mathcal{M}_i$ if $\hat{g}_i = \emptyset$and any mass shifting molecular adduct or neutral loss for the matched peak. This computes a per-peak intensity weighted mean G-test comparing the goodness of fit between the experimental envelope and the theoretical isotopic pattern. This metric is bounded within $(-\infty, \infty)$ as the G-test achieves its optimal value at 0, and can take on extreme values towards either signed ∞, however because of the previous deconvolution process, in practice it cannot take on such extreme values and is bounded within $(-\infty, 1]$. This metric is thresholded at 0.15, with any chromatogram scoring below 0.15 being discarded as having insufficient isotopic consistency to interpret.

Observation Spacing Score

The less time between observations of a glycan composition the less likely the chromatogram is to contain peaks missing or caused by isotopic pattern interference or missing information.

$$\begin{aligned} d_i &= \frac{1}{T_i - T_j} \\ \mathscr{T}_i &= 1 - 2\mathcal{I}_i^{-t}\mathbf{I}_i \sum_{j=1}^J \mathcal{I}_{i,j} f\left(d_i\left(\mathbf{t}_{i,j} - \mathbf{t}_{i,j-1}\right)\right) \end{aligned} \tag{3.13}$$

Table 3.8: Score Thresholds

Chromatographic Peak Shape	0.15
Charge State Distribution	0.05
Adduction Frequency	0.15
Isotopic Pattern Consistency	0.15
Observation Spacing	0.15

As this metric depends heavily on the speed of the mass spectrometer, a scaling function f must be estimated from the total ion chromatogram to reduce the penalty on slower instruments. When $\frac{1}{J}\sum_j^J T_j - T_{j-1} > 0.2$,

$$f(x) = \frac{x}{\left(\frac{1}{J}\sum_j^J T_j - T_{j-1} \times 15\right)} \tag{3.14}$$

Otherwise, $f(x) = x$. This metric is bounded within $(-\infty, 1]$ as $(\mathbf{t}_{i,j} - \mathbf{t}_{i,j-1})$ is always positive. This metric is thresholded at 0.15, with any chromatogram scoring below 0.15 being discarded as having insufficient detection consistency to interpret.

Summarization Score

Each scoring metric $\in [\mathscr{L}_i, \mathscr{C}_i, \mathscr{I}_i, \mathscr{T}_i, \mathscr{A}_i]$ is penalized by $\epsilon = 1\text{e}{-6}$ bounded in the range $[0, 1)$, with values below 0 set to ϵ.

$$s_i = \sum_{f_{i,j} \in \text{features}_i} \ln \frac{f_{i,j}}{1 - f_{i,j}} \tag{3.15}$$

producing a value between $(-\infty, \infty)$. $s_i < 8$ reflects multiple poor scores and is unexpected to be real, while $s_i > 15$ is consistent with model expectations.

3.5.4 A more complete derivation of $\hat{\phi}$

To obtain the optimal ϕ, we take the partial derivative of ℓ w.r.t ϕ_m

$$
\begin{aligned}
\mathcal{S} &= \lambda \begin{bmatrix} \phi_o - \tau_o, & \phi_m - \tau_m \end{bmatrix} \begin{bmatrix} \mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}} \end{bmatrix} \begin{bmatrix} \phi_o - \tau_o \\ \phi_m - \tau_m \end{bmatrix} \\
0 &= \frac{\partial \ell}{\partial \phi_m} \Big((\mathbf{s} - \phi_\mathbf{o})^t (\mathbf{s} - \phi_\mathbf{o}) + \mathcal{S} \Big) \qquad (3.16) \\
&= \lambda(\phi_o - \tau_o)^t \mathbf{L_{om}} + \lambda \mathbf{L_{mo}} (\phi_o - \tau_o) + \lambda(\phi_m - \tau_m)^t ({\mathbf{L_{mm}}}^t + \mathbf{L_{mm}}) \\
&= 2\lambda \mathbf{L_{mo}} (\phi_o - \tau_o) + 2\lambda \mathbf{L_{mm}} (\phi_m - \tau_m) \\
-\mathbf{L_{mm}} (\phi_m - \tau_m) &= \mathbf{L_{mo}} (\phi_o - \tau_o) \\
(\phi_m - \tau_m) &= -{\mathbf{L_{mm}}}^{-1} \mathbf{L_{mo}} (\phi_o - \tau_o) \\
\hat{\phi}_m &= -{\mathbf{L_{mm}}}^{-1} \mathbf{L_{mo}} (\phi_o - \tau_o) + \tau_m \qquad (3.17)
\end{aligned}
$$

and w.r.t. ϕ_o

$$\begin{align*}
\mathcal{S} &= \lambda \left[\phi_o - \tau_o, \quad \phi_m - \tau_m\right] \begin{bmatrix} \mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}} \end{bmatrix} \begin{bmatrix} \phi_o - \tau_o \\ \phi_m - \tau_m \end{bmatrix} \\
0 &= \frac{\partial \ell}{\partial \phi_o} \Big((\mathbf{s} - \phi_{\mathbf{o}})^t (\mathbf{s} - \phi_{\mathbf{o}}) + \mathcal{S} \Big) \tag{3.18} \\
&= -2\mathbf{s} + 2\phi_o + \lambda \left(\mathbf{L_{oo}} + \mathbf{L_{oo}}^t \right) (\phi_o - \tau_o) + \lambda \mathbf{L_{om}} (\phi_m - \tau_m) + \lambda \mathbf{L_{mo}}^t (\phi_m - \tau_m) \\
&= -2\mathbf{s} + 2\phi_o + 2\lambda \mathbf{L_{oo}} (\phi_o - \tau_o) + 2\lambda \mathbf{L_{om}} (\phi_m - \tau_m) \\
\mathbf{s} &= \phi_o + \lambda \left(\mathbf{L_{oo}} (\phi_o - \tau_o) + \mathbf{L_{om}} (\phi_m - \tau_m) \right) \\
&= \phi_o + \lambda \Big(\mathbf{L_{oo}} (\phi_o - \tau_o) + \mathbf{L_{om}} (-\mathbf{L_{mm}}^{-1} \mathbf{L_{mo}} (\phi_o - \tau_o) + \tau_m - \tau_m) \Big) \\
&= \phi_o + \lambda \Big(\mathbf{L_{oo}} (\phi_o - \tau_o) - \mathbf{L_{om}} \mathbf{L_{mm}^{-1}} \mathbf{L_{mo}} (\phi_o - \tau_o) \Big) \\
\mathbf{s} - \tau_o &= \phi_o - \tau_o + \lambda \Big(\mathbf{L_{oo}} (\phi_o - \tau_o) - \mathbf{L_{om}} \mathbf{L_{mm}^{-1}} \mathbf{L_{mo}} (\phi_o - \tau_o) \Big) \\
&= \mathbf{I} (\phi_o - \tau_o) + \lambda \Big(\mathbf{L_{oo}} (\phi_o - \tau_o) - \mathbf{L_{om}} \mathbf{L_{mm}^{-1}} \mathbf{L_{mo}} (\phi_o - \tau_o) \Big) \\
&= \Big[\mathbf{I} + \lambda \Big(\mathbf{L_{oo}} - \mathbf{L_{om}} \mathbf{L_{mm}^{-1}} \mathbf{L_{mo}} \Big) \Big] (\phi_o - \tau_o) \\
(\phi_o - \tau_o) &= \Big[\mathbf{I} + \lambda \Big(\mathbf{L_{oo}} - \mathbf{L_{om}} \mathbf{L_{mm}^{-1}} \mathbf{L_{mo}} \Big) \Big]^{-1} (\mathbf{s} - \tau_o) \\
\hat{\phi}_o &= \Big[\mathbf{I} + \lambda \Big(\mathbf{L_{oo}} - \mathbf{L_{om}} \mathbf{L_{mm}^{-1}} \mathbf{L_{mo}} \Big) \Big]^{-1} (\mathbf{s} - \tau_o) + \tau_o \tag{3.19}
\end{align*}$$

3.5.5 Estimation of Laplacian Regularization Parameters

We model the relationship between $\mathbf{s}$, $\phi_{\mathbf{o}}$, and τ as a multivariate Gaussian distribution.

$$\begin{align*}
(\mathbf{s} | \phi_{\mathbf{o}}, \tau) &\sim \mathcal{N}(\phi_{\mathbf{o}}, \Sigma) \tag{3.20} \\
\Sigma &= \rho \mathbf{I} \tag{3.21}
\end{align*}$$

$$\left(\begin{bmatrix}\phi_{\mathbf{o}}\\ \phi_{\mathbf{m}}\end{bmatrix}\middle|\tau\right) \sim \mathcal{N}(\mathbf{A}\tau, \lambda^{-1}\mathbf{L}^{-}) \tag{3.22}$$

$$(\phi_{\mathbf{o}}|\tau) \sim \mathcal{N}\left(\mathbf{A_o}\tau, \Sigma_{\phi_o}\right) \tag{3.23}$$

$$\Sigma_{\phi_o} = \lambda^{-1}\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)^{-1} \tag{3.24}$$

$$\tau \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right) \tag{3.25}$$

Fully expanded, this becomes

$$\begin{bmatrix}\mathbf{s}\\ \phi_{\mathbf{o}}\\ \tau\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\Sigma + \Sigma_{\phi_o} + \sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o} + \sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \Sigma_{\phi_o} + \sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o} + \sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{I}\end{bmatrix}\right) \tag{3.26}$$

We can form the conditional distribution $\tau|\mathbf{s}$ which has a mean

$$\mu_{\tau|\mathbf{s}} = 0 + (\sigma^2\mathbf{A_o}^t)\left(\Sigma + \Sigma_{\phi_o} + \sigma^2\mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \tag{3.27}$$

$$= \mathbf{A_o}^t\left(\tilde{\rho}\mathbf{I} + \frac{1}{\tilde{\lambda}}\mathbf{L_{oo}^{-}} + \mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \tag{3.28}$$

We assume that $\sigma^2 \gg 1$, and treat λ and ρ as relative to σ^2, as $\tilde{\rho}$ and $\tilde{\lambda}$. This model gives us an estimate for τ given a value for ρ and λ. As ρ has no direct role in the central tendency of ϕ or $\mathbf{s}$, we choose to fix the value of $\tilde{\rho} = 0.1$, which leaves only $\tilde{\lambda}$. We estimate the optimal $\tilde{\lambda}$ by grid search, minimizing the predicted residual error sum of squares (PRESS) statistic.

$$\mathbf{e} = \mathbf{s} - \hat{\phi}_{\mathbf{o}} \tag{3.29}$$

$$\mathbf{H} = \left(\mathbf{I} + \tilde{\lambda}\mathbf{L}\right)^{-1} \tag{3.30}$$

$$\arg\min_{\tilde{\lambda}} \sum_{i}^{n} \left(\frac{e_i}{1 - h_{i,i}}\right)^2 \tag{3.31}$$

This formulation depends upon the value of **s** and is sensitive to low scoring matches, which can lead to incorrect estimates of τ and PRESS. We therefore perform a grid search over both $\tilde{\lambda}$ and a minimum threshold for **s**, γ.

As we increase γ we remodel the graph $\mathcal{G}$, removing nodes whose score is below γ. For each pair of neighbors of removed node g_m, (g_u, g_v), if $L_1(g_u, g_v) > L_1(g_u, g_m) + L_1(g_m, g_v)$, we add an edge from g_u to g_v with weight $\frac{1}{L_1(g_u,g_m)+L_1(g_m,g_v)}$, up to a limit of $L_1(g_k, g_m) < 5$. We give the result of this grid search the name $\mathbf{r}$. At each point, on the grid, we save the value of τ in $r_{\lambda_i,\gamma_j,\tau}$ and the PRESS in $r_{\lambda_i,\gamma_j,PRESS}$. To select the optimal parameters, we traverse the grid along γ, computing τ_γ:

$$\bar{\lambda}_j = \arg\min_{\lambda_i} r_{\lambda_i,\gamma_j,PRESS} \tag{3.32}$$

$$\tau_{\gamma_j} = |r_{\bar{\lambda}_j,\gamma_j,\tau}| * \left(\frac{\gamma_j}{b} + (1 - \frac{1}{b})\right) \tag{3.33}$$

where b is a bias factor defining how much weight to give to higher values of γ which correspond to networks made up of higher confidence assignments. We chose $b = 4$. We define $\bar{\tau}_\gamma = \max \tau_\gamma$ and define the vector $\bar{\gamma} = \left[\gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.95\right]$. This favors values of γ where large values of τ are selected, meaning that the neighborhoods are well populated, while also giving an estimate for $\tilde{\lambda}$ that is non-zero. We term the values of γ in $\bar{\gamma}$ the *target thresholds* of **s**.

To estimate $\tilde{\lambda}$ and τ from these results, we select the columns of the grid $\mathbf{r}$ at each $\gamma_j \in \bar{\gamma}$ and applied the following procedure:

$$\bar{\tau}_\gamma = \max \tau_\gamma \tag{3.34}$$

$$\bar{\gamma} = \left\{ \gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.9 \right\} \tag{3.35}$$

$$\bar{\lambda} = \left\{ \bar{\lambda}_j \leftarrow \gamma_j \in \bar{\gamma} \right\} \tag{3.36}$$

$$\mathbf{s}_{\gamma_\mathbf{j}} = \{ s_i \leftarrow s_i > \gamma_j \} \tag{3.37}$$

$$\bar{\tau}_\mathbf{j} = \mu_{\tau | \mathbf{s}_{\gamma_j}, \bar{\lambda}_j} \tag{3.38}$$

$$\hat{\lambda} = \frac{1}{|\bar{\lambda}|} \sum_j \bar{\lambda}_j \tag{3.39}$$

$$\hat{\tau} = \frac{1}{|\bar{\tau}|} \sum_j \bar{\tau}_\mathbf{j} \tag{3.40}$$

$$\hat{\gamma} = \frac{1}{|\bar{\gamma}|} \sum_j \bar{\gamma}_j \tag{3.41}$$

where $\mathbf{s}_{\gamma_j}$ is the set of observed scores which are greater than γ_j, but where the estimation of is carried out with the complete Laplacian $\mathbf{L}$, not the reduced network used to compute $\mathbf{r}$. This set of averaged estimates of $\hat{\lambda}$ and $\hat{\tau}$ are then used to estimate $\hat{\phi}_o$ by 3.19, labeled 3.8 in the main text.

3.5.6 MS^n Signature Ion Criterion

This feature was not used in the main article in order to make the comparison between our results and previously published work more straight forward.

When MS$^\text{n}$ scans are present, it may be useful to consider only those MS^1 features which are associated with MS$^\text{n}$ scans that contain glycan-like signature ions.

We include an algorithm for classifying an MS^n scan as being "glycan-like":

$$I = max(intensity(p)) \tag{3.42}$$

$$t = I * 0.01 \tag{3.43}$$

$$\begin{aligned} p_{oxonium} = \{ & p_i \leftarrow |ppmerror(mass(p_j), mass(f_g))| < e, \\ & f_g \in oxonium(g), f_g \neq \text{Fucose}, intensity(p_i) > t\} \end{aligned} \tag{3.44}$$

$$\begin{aligned} p_{edges} = \{ & (p_i, p_j) \leftarrow |ppmerror(mass(p_j) - mass(p_i), mass(f_g))| < e, \\ & oxonium(f_g) \in g, intensity(p_i) > t, intensity(p_j) > t\} \end{aligned} \tag{3.45}$$

$$s_{oxonium} = \frac{1}{|p_{oxonium}|} \sum_{p_i}^{p_{oxonium}} \left(\frac{intensity(p_i)}{I} \right) * min(log_4 |p_{oxonium}|, 1) \tag{3.46}$$

$$s_{edges} = \frac{1}{|p_{edges}|} \sum_{p_i, p_j}^{p_{edges}} \left(\frac{intensity(p_i) + intensity(p_j)}{I} \right) * min(log_4 |p_{edges}|, 1) \tag{3.47}$$

$$s_g = max(s_{oxonium}, s_{edges}) \tag{3.48}$$

$$\tag{3.49}$$

Where p is the set of peaks in the scan, g is the glycan composition, e the required parts-per-million mass accuracy. $oxonium()$ is a function that given a glycan composition g, produces fragments f_g of g composed of between one and three monosaccharides, commonly observed as oxonium ions alone, or as the mass difference between two peaks formed from consecutive fragmentation of a glycosidic bond. This method is not intended to identify a glycan structure, just detect patterns in the signal peaks of the MS^n scan that could indicate the fragmentation of a glycan.

3.5.7 Algorithmic Performance on All Datasets

For more details on each sample, please see Table 3.6.

Results for AGP

We analyzed three different sample workups of *N*-glycans released from Alpha 1 Acid Glycoprotein. See Table 3.9 for a comparison of estimated τ values for each sample. For *AGP-DR-Perm-glycans-1* and *AGP-permethylated-2ul-inj-55-SLens*, we used an MS^n Signature Ion Criterion threshold of 0.17 to filter out large contaminants that may be introduced by permethylation reagents.

The estimate of γ for *20150930-06-AGP* was larger than the score for the larger penta-antennary

τ_i	*20150930-06-AGP*	*AGP-DR-Perm-glycans-1*	*AGP-permethylated-2ul-inj-55-SLens*
high-mannose	0.000	0.000	0.000
hybrid	11.520	7.240	21.092
bi-antennary	15.691	12.859	20.627
asialo-bi-antennary	0.000	0.000	13.253
tri-antennary	21.752	21.693	21.550
asialo-tri-antennary	0.000	0.000	6.792
tetra-antennary	15.993	15.276	17.452
asialo-tetra-antennary	0.000	0.000	0.000
penta-antennary	11.446	10.127	7.282
asialo-penta-antennary	0.000	0.000	0.000
hexa-antennary	2.211	0.000	0.000
asialo-hexa-antennary	0.000	0.000	0.000
hepta-antennary	0.000	0.000	0.000
asialo-hepta-antennary	0.000	0.000	0.000
$\hat{\lambda}$	0.99	0.99	0.99
$\hat{\gamma}$	15.74	16.22	17.64

Table 3.9: Estimated values of smoothing parameters τ, λ, and γ for each AGP-based dataset and using a combinatorial database

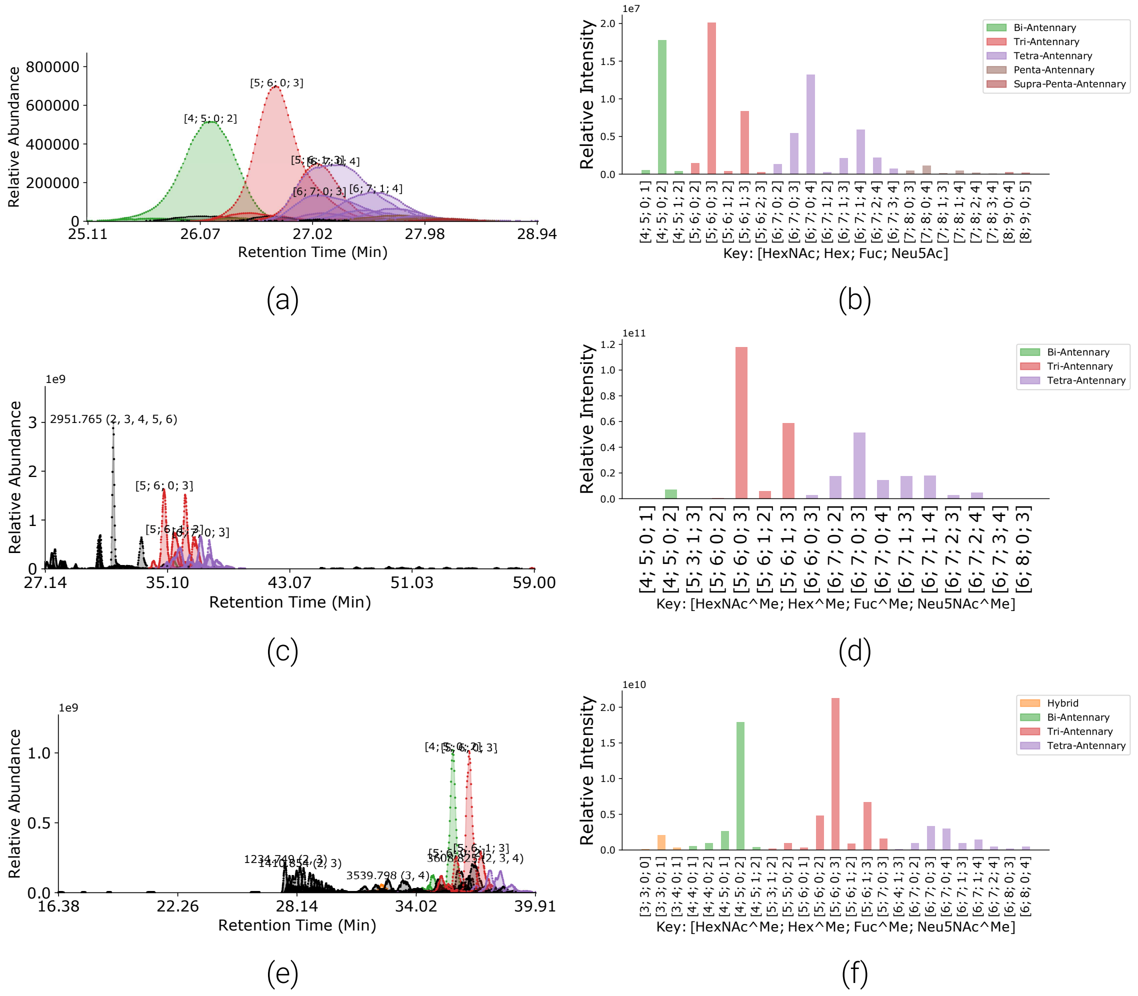


Figure 3.7: Chromatogram Assignments for *20150930-06-AGP*(a, b), *AGP-DR-Perm-glycans-1*(c, d) and *AGP-permethylated-2ul-inj-55-SLens*(e, f)

Results for Phil-82

We analyzed native and deutero-reduced and permethylated *N*-glycans released from virions of Influenza-A Virus strain Phillipines 1982, both samples acquired on a Q-TOFmass spectrometer. See Table 3.10 for a comparison of estimated τ values for each sample. In the case of *20141128-11-Phil-82*, MS^n scans were acquired, resulting in lower resolution chromatographic peaks. We observed little ammonium adduction in *20141128-11-Phil-82*. As expected, we observed abundant formate adduction in *20141031-07-Phil-82*, particularly on the high mannose glycans. *20141128-11-Phil-82* also displays considerable in-source fragmentation of the high mannose series, defined by the multimodal chromatographic peaks of smaller high mannose glycans appearing in lower abundance directly under larger peaks for high mannose glycans. This fragmentation, combined with permethylation altering the ionization efficiency of these analytes, makes a direct comparison of glycan composition abundance between *20141031-07-Phil-82*and *20141128-11-Phil-82*inadvisable. We observe markedly different peak shapes between *20141031-07-Phil-82* and *20141128-11-Phil-82* but the relative order of elution is preserved, with the largest high mannose glycans eluting later than the largest observed complex type.

τ_i	*20141031-07-Phil-82*	*20141128-11-Phil-82*
high-mannose	17.070	19.395
hybrid	14.039	17.147
bi-antennary	0.000	0.000
asialo-bi-antennary	16.287	17.689
tri-antennary	0.000	0.000
asialo-tri-antennary	15.220	18.865
tetra-antennary	0.000	0.000
asialo-tetra-antennary	7.103	7.660
penta-antennary	0.000	0.000
asialo-penta-antennary	0.000	3.365
hexa-antennary	0.000	0.000
asialo-hexa-antennary	0.000	0.000
hepta-antennary	0.000	0.000
asialo-hepta-antennary	0.000	0.000
$\hat{\lambda}$	0.99	0.99
$\hat{\gamma}$	16.51	15.50

Table 3.10: Estimated values of smoothing parameters τ, λ, and γ for each Phil-82-based dataset and using a combinatorial database

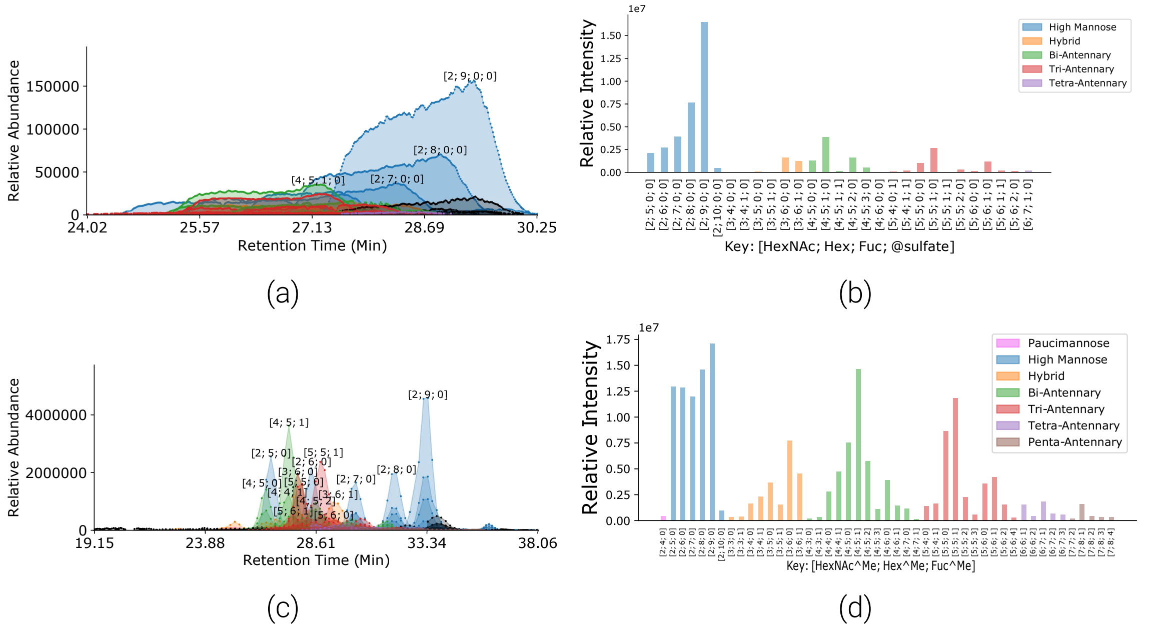


Figure 3.8: Chromatogram Assignments for *20141031-07-Phil-82*(a, b) and *20141128-11-Phil-82*(c, d)

Results for IGG

We analyzed native *N*-glycansreleased from IgG. The estimated τ values shown in Table 3.11are consistent with the expectation that IgG glycans will be either hybrid or small complex-type structures. These findings are consistent with the results from [246], though their study used different sample preparation and instrumentation, and their data were not available for side-by-side comparison. The EICs and integrated abundances for this sample are shown in Figure 3.9.

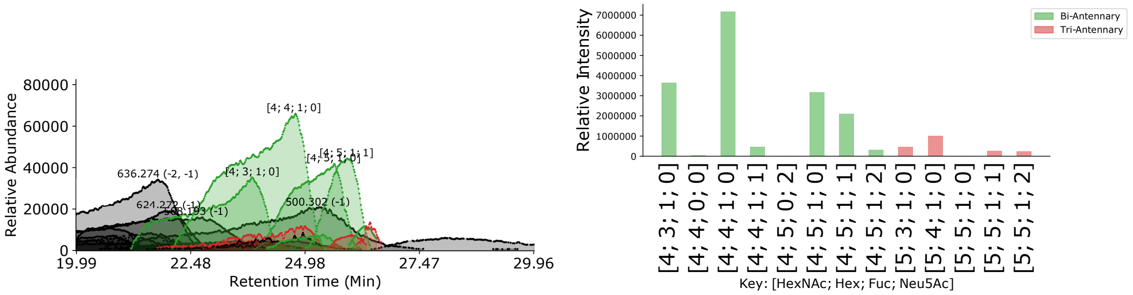
3.5.8 Differences in Assigned Glycans for *Perm-BS-070111-04-Serum*

Of the compositions assigned by our algorithm that were not mentioned in [96] but were annotated in the original publication of this dataset in [254] include HexNAc3

τ_i	*20151002-02-IGG*
high-mannose	0.000
hybrid	15.737
bi-antennary	12.594
asialo-bi-antennary	13.614
tri-antennary	7.657
asialo-tri-antennary	15.724
tetra-antennary	4.252
asialo-tetra-antennary	0.000
penta-antennary	0.000
asialo-penta-antennary	0.000
hexa-antennary	0.000
asialo-hexa-antennary	0.000
hepta-antennary	0.000
asialo-hepta-antennary	0.000
$\hat{\lambda}$	0.99
$\hat{\gamma}$	14.12

Table 3.11: Estimated values of smoothing parameters τ, λ, and γ for IGG using a combinatorial database

Hex4, HexNAc3 Hex4 NeuAc1, and HexNAc5 Hex3. Because our database was constructed based on combinatorial rules that did not take into account all biosynthetic constraints, we include infeasible compositions in our search space, such as HexNAc2 Hex10 Fuc1 and HexNAc5 Hex3 Fuc1 NeuAc2. Future work could be done to restrict the database to only biosynthetically feasible glycan compositions. This would also have benefits for the construction of the composition network where only those compositions which have an enzymatic reaction to from one to the other would have an edge connecting them, such that HexNAc5 Hex6 NeuAc2 would not have an edge to HexNAc5 Hex7 NeuAc2 as in our current model.



(a)

(b)

Figure 3.9: Chromatogram Assignments for *20151002-02-IGG*

3.5.9 glySpace Integration and Upload

We extracted *N*-glycanstructures from GlyTouCan Query Endpoint (http://ts.glytoucan.org/sparql) using the SPARQL query

```
PREFIX glycan: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX glycoinfo: <http://rdf.glycoinfo.org/glycan/>

SELECT DISTINCT ?saccharide ?glycoct ?motif WHERE {
    ?saccharide a glycan:saccharide .
    ?saccharide glycan:has_glycosequence ?sequence .
    ?saccharide skos:exactMatch ?gdb .
    ?gdb glycan:has_reference ?ref .
    ?ref glycan:is_from_source ?source .
    ?source glycan:has_taxon ?taxon
    FILTER CONTAINS(str(?sequence), "glycoct") .
    ?sequence glycan:has_sequence ?glycoct .
    ?saccharide glycan:has_motif ?motif .
```

```
        FILTER(?motif in (glycoinfo:G00026MO))
    }
```

and converted each structure into a glycan composition, followed by substituent separation for sulfated and phosphorylated monosaccharides, and filtering out compositions containing units not in **[Hex, HexNAc, Fuc, Neu5Ac, sulfate]**. This procedure is implemented in Python in the included "glyspace_extract_nglycans.py" script.

Note that lines 4-7 restricts the query to only compositions which were in GlycomeDB which came from externally curated sources with taxonomic information, though it is not limited to human *N*-glycansspecifically. If these lines are omitted, the query will return over 800 compositions, compared to the expected 275, but the additional compositions will not have been curated. The precise number of compositions returned by this modified query is not fixed as GlyTouCan is a living database, accepting new submissions.

We converted our *N*-glycancompositions into partially determined topologies assuming that the chitobios core was present to ensure that they were classified as *N*-glycans.

From *Perm-BS-070111-04-Serum*

```
{Fuc:1; Hex:5; HexNAc:3; Neu5Ac:1}
{Fuc:2; Hex:5; HexNAc:4; Neu5Ac:2}
{Fuc:2; Hex:6; HexNAc:5; Neu5Ac:3}
{Fuc:2; Hex:7; HexNAc:6; Neu5Ac:3}
{Fuc:2; Hex:7; HexNAc:6; Neu5Ac:4}
{Hex:7; HexNAc:6; Neu5Ac:2}
{Hex:7; HexNAc:6; Neu5Ac:3}
```

```
{Hex:8; HexNAc:7; Neu5Ac:3}

{Hex:8; HexNAc:7; Neu5Ac:4}

{Hex:9; HexNAc:8; Neu5Ac:2}
```

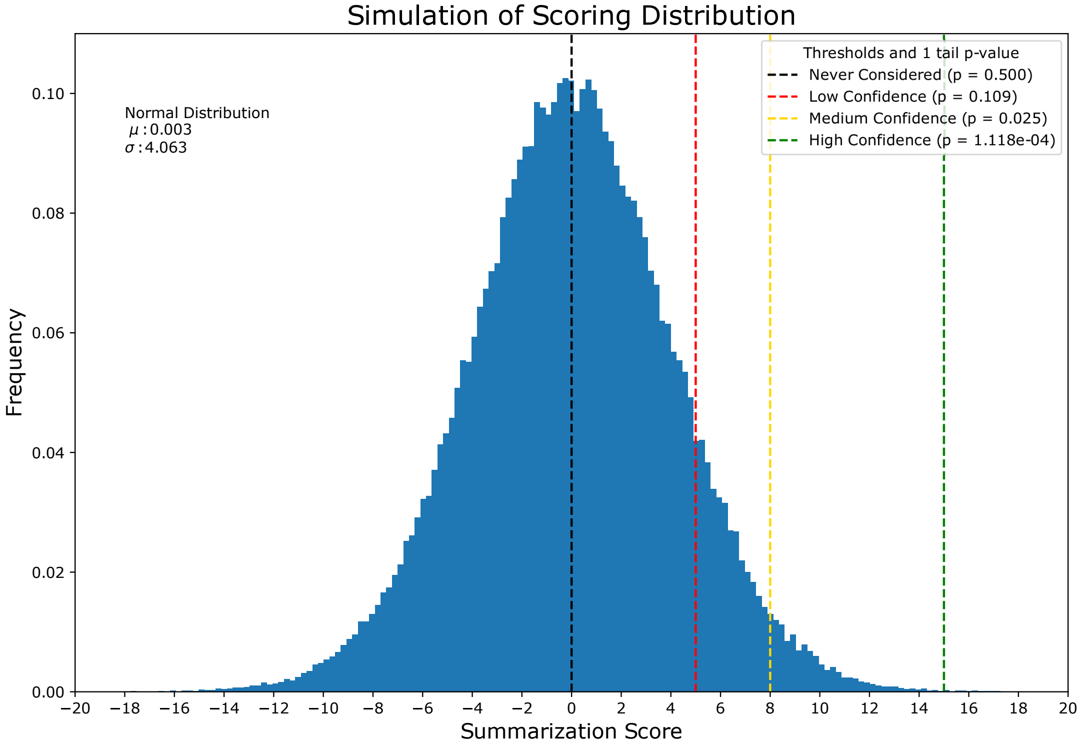
From *20141103-02-Phil-BS*

```
{@sulfate:1; Fuc:1; Hex:4; HexNAc:5}

{@sulfate:1; Fuc:1; Hex:5; HexNAc:4}

{@sulfate:1; Fuc:1; Hex:5; HexNAc:5}

{@sulfate:1; Fuc:2; Hex:4; HexNAc:5}

{@sulfate:1; Fuc:2; Hex:6; HexNAc:5}

{@sulfate:1; Fuc:2; Hex:9; HexNAc:8}

{@sulfate:1; Fuc:3; Hex:4; HexNAc:5}

{@sulfate:1; Fuc:3; Hex:6; HexNAc:5}

{@sulfate:1; Fuc:3; Hex:9; HexNAc:8}

{@sulfate:1; Fuc:4; Hex:6; HexNAc:5}

{@sulfate:1; Fuc:4; Hex:8; HexNAc:7}

{@sulfate:1; Fuc:4; Hex:9; HexNAc:8}

{@sulfate:1; Hex:10; HexNAc:9}

{@sulfate:1; Hex:4; HexNAc:5}

{@sulfate:1; Hex:5; HexNAc:4}

{Fuc:2; Hex:8; HexNAc:7}

{Fuc:3; Hex:7; HexNAc:6}

{Fuc:3; Hex:8; HexNAc:7}

{Fuc:4; Hex:8; HexNAc:7}

{Hex:10; HexNAc:9}
```

3.5.10 Simulation of Summarization Score

To simulate the summarization score, we assume that each component scoring feature is drawn from an independent uniform distribution. We sample these five distributions 100,000 times, and for each set of five feature scores compute $\sum_j \text{logit}(f_{i,j})$. According to the central limit theorem, the distribution of the summarization score should be normal, with a mean at approximately 0. We noted two score thresholds, 8 and 15 for lower confidence and high confidence matches. We connect these score thresholds to p values from one-sided tests for significance from thes simulated distribution. The threshold of 8 has a p value of ~ 0.025, while 15 has a p value of $\sim 1.1 \times 10^{-4}$. This distribution is visualized in Figure 3.10.

Figure 3.10: Simulation of Summarization Score



Chapter 4

Integrated Glycopeptide Identification from LC-MS/MS Experiments

4.1 Introduction

Glycosylation is one of the most pervasive co- and post-translational protein modifications in nature [9, 255]. At least a third of the human proteome is believed to be secreted or contain a transmembrane region [256], and 80% of proteins passing through the secretory pathway have at least one *N*-glycan sequon [18], and often these proteins carry multiple glycosylation sites. Measuring the released glycans from a sample provides a crude measure of the broad range of different glycosylation pathways that are active in that sample. To capture where those glycans are localized and infer more specific behavior, we must analyze an intact glycoprotein or the glycopeptides that cover its glycosites. This can tell us substantially more about the state of each protein in a sample, at the cost of substantially greater complexity compared to released glycomics and unglycosylated-peptide proteomics [257]. This added complexity is necessary in order to study common and important classes of molecules, like many classes of antibody [32, 34–36] and molecules they recognize [14, 24, 42]. They are also needed to study other fundamental components of the extracellular matrix including proteoglycans [45] and other secreted and membrane-bound glycoproteins [41].

Glycoproteins are proteins with one or more glycosylation sites. Each glycosylation site may be occupied by a glycan drawn from a population of distinct structures, microheterogeneity. Each glycosylation site is generally able to vary independently, macroheterogeneity [59]. The glycans at each site influence the physical properties of the protein, and in turn modulate its function through a number of channels [9]. These include cell-cell and cell-matrix adhesion [48], receptor/ligand recognition [12], and more through complex binding interactions with other messengers [39, 42].

LC-MS is a high-throughput and relatively precise for studying protein glycosylation, as compared to broad physical property tests such as binding assays [23, 258, 259]. Studying intact glycoproteins is challenging because each distinct proteoform [7] is large and complex, potentially too large or too complex for current instrumentation to properly detect [260]. To make the problem tractable, we use proteases like those introduced in 1.3.4 to reduce the total space of proteoforms to glycopeptides. A glycopeptide is much simpler than a glycoprotein, with tryptic glycopeptides usually only containing a single *N*-glycosylation sequon, or a few *O*-glycosylation sites, though if missed cleavages occur it is not uncommon to see these numbers grow quickly as shown in Figure 4.1.

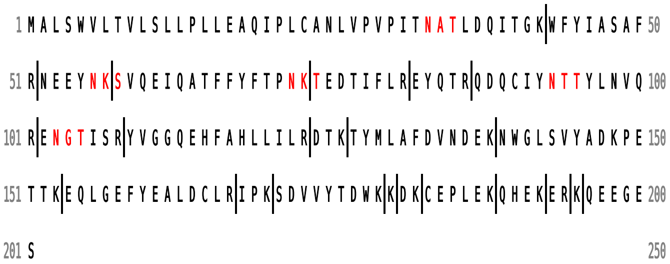

1 MALSWVLTVLSLLPLLEAQIPLCANLVPVPITNATLDQITGKWFYIASAF 50
51 RNEEYNKSVQEIQATFFYFTPNKTEDTIFLREYQTRQDQCIYNTTYLNVQ 100
101 RENGTISRYVGGQEHFAHLLILRDTKTYMLAFDVNDEKNWGLSVYADKPE 150
151 TTKEQLGEFYEALDCLRIPKSDVVYTDWKKDKCEPLEKQHEKERKQEEGE 200
201 S 250

Figure 4.1: AGP isoform 1 with *N*-glycosites denoted in red, and tryptic cleavage sites denoted by the black bars. If a cleavage site were missed, an tryptic glycopeptide could readily contain multiple glycosites.

There are a variety of methods and techniques for studying glycopeptides by LC-MS/MS, at different levels of detail. There are several approaches that partially or completely remove the glycan while leaving a marker on the glycosylation site, such as PNGase F induced deamidation [23, 261] or acid hydrolysis [262], that can be detected with traditional proteomics search algorithms to measure site occupancy. These methods can determine if a site is glycosylated or not, but can't determine what kind of glycan was at any site. Collisional dissociation applied to intact glycopeptides, as described in 1.3.2. These types of spectra can only be interpreted by search engines that have been designed to take the dissociation of labile modifications into account [54, 132, 133, 153, 181, 183–186, 263]. Though ExD and EThcD methods have been shown to be valuable for characterizing complex, multiply glycosylated peptides, they are not widely available, and there are few published datasets to draw on. This work will focus primarily on HCD and Stepped HCD.

4.1.1 How Glycopeptides Fragment Under HCD

When collisionally activated, a glycopeptide breaks down in an energy dependent manner. The first bonds to break are the weakest, the glycosidic bonds [88, 187], leading to B ions from attached glycans and complementary, often multiply charged, $peptide{+}Y$ ions. If a peptide is multiply glycosylated, bonds may break in all glycans simultaneously, leading to complex ion ladders. At higher energies, the $peptide{+}Y$ ions are further dissociated, yielding smaller $peptide{+}Y$ ions with lower charge states, and the peptide bonds begin to break, producing b and y ions as described in 1.3.2. This means that by the point at which we begin to observe substantial peptide backbone fragmentation, most of the structural information about the attached glycans has been destroyed. Therefore, we can only localize the glycosylation by observation of the reducing end monosaccharide still attached to a peptide backbone fragment, rather than the intact glycan. As a consequence, HCD cannot be used to disambiguate multiply glycosylated peptides.

4.1.2 Glycopeptide Representation

It is important to note that topological information is not necessary to predict the masses of many of these fragments, particularly those from higher energies, which means we can continue to use glycan compositions rather than fully specified structures. Additionally, because we do not attempt to assign a glycan composition to a specific site, but just track an aggregate composition over the entire glycopeptide and denote which sites may carry glycosylation markers. This leads to a simplified representation of a glycopeptide, which we compare possible renderings in Table 4.1. I implemented a set of data structures for all three levels of representation, and how they can be used to generate theoretical fragments for MS^n matching.

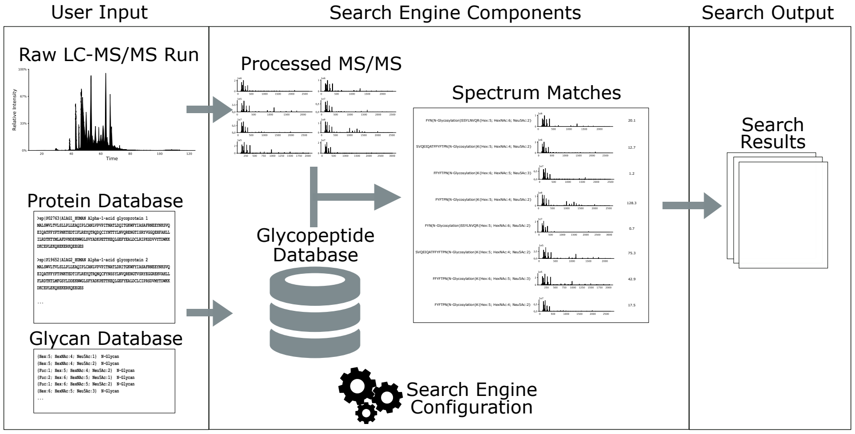


Figure 4.2: A schematic diagram of a glycoproteomics search engine, marking inputs, outputs, and intermediate steps.

A database search engine has four basic components, a database or “search space” construction and traversal procedure, a mass spectrum preprocessor, a scor-

Fully Specified	Localized Composition	Simplified
LSVQN(#:iupac_simple,glycosylation_type=n_linked:Fuc(a1-6)[Neu5Ac(a2-6)Gal(b1-4)Glc2NAc(b1-2)Man(a1-6)[Neu5Ac(a2-3)Gal(b1-4)Glc2NAc(b1-2)Man(a1-3)]Man(b1-4)Glc2NAc(b1-4)]Glc2NAc)ETLADR	LSVQN(#:glycosylation_type=n_linked:{Fuc:1; Hex:5; HexNAc:4; Neu5Ac:2})ETLADR	LSVQN(N-Glycosylation)ETLADR{Fuc:1; Hex:5; HexNAc:4; Neu5Ac:2}

Table 4.1: Glycopeptide Textual Representations. A fully specified glycopeptide has the complete glycan topology localized on the peptide sequence, along with any other modifications. Some metadata must be conserved in order to communicate the type of glycan at a site and what the glycan's encoding format is, IUPAC with "simple linkages" in this case. A topology can be collapsed into a composition, with metadata describing conserved structure, but still localized on the peptide sequence. Finally, the glycosylation can be aggregated over the entire sequence, but the glycosites labeled with enough metadata to denote which motif was found there.

ing model that is used to evaluate the quality of the match between a spectrum and a structure, and a model for evaluating the uncertainty of an identification used to estimate the FDR for reported PSMs. This schematic is described in Figure 4.2. In Chapter 2, I discussed the tasks that a mass spectrum preprocessor must address, and Chapter 3 introduced some topics related to the creation of a glycan search space. This chapter will go into more detail on the database construction and traversal topic, as well as on the statistical modeling and machine learning methods used for building a scoring model and for estimating identification confidence.

4.2 Methods

4.2.1 Glycopeptide Search Space Construction

The database construction and traversal process defines how the search procedure enumerates theoretical glycopeptides and determines which should be compared against which spectra. While this notion may be viewed simply as an interval search over the full cross-product between each peptide and the combination of glycans which it can host, this representation is impractical for large search spaces. Given the 4762 proteins in the human proteome that are annotated as glycoproteins on UniProt [189], there are 597,219 tryptic peptides, and a lower bound of 448 *N*-glycans. Of these peptides, 90,273 carry *N*-glycosites, which translates to $90,273 \times 448 = 40,442,304$ *N*-glycopeptides assuming exactly one glycosite per peptide. The actual number for 448 glycan compositions is 46,405,632 denoting several thousand peptides with multiple glycosites. This is impractical to store in memory, and does not take into account any other PTMs. This problem grows far worse when *O*-glycans are considered. There are 22,252 *N*-glycosites in this space, but there are 387,165 *O*-glycosites in it as well, and *O*-glycosites tend to cluster together, making it common to have many sites per tryptic peptide. This implies two types of alternative approaches, streaming generation of theoretical glycopeptides, and preconstruction and indexed traversal of theoretical glycopeptides.

Construction and Traversal Strategies

Historically, streaming generation has been used because it is simpler to implement and easier to optimize [136, 264], with only limited use of intermediary disk-stored indices. Streaming generation works by iterating over the search space from start to finish over a subset of the input spectra, and to optionally traverse multiple subsets in multiple threads of execution. If spectra are sortable by precursor mass, then the theoretical space to traverse can be constrained by not generating peptidoforms of peptides whose mass does not fall within the mass range of the current subset. Byonic [263] constrains the combinatorial expansion of modifications including glycosylation by giving each PTM a probability, and setting a probability threshold below which no joint set of modifications will be considered. This prevents catastrophic expansion of mucin domains. Methods like pGlyco2 [132] takes this process one step further and only generates glycopeptides for which a theoretical peptide mass can be inferred from one or more queried MS^2 spectra, using the complementarity of the $peptide{+}Y$ ions and the precursor to backsolve putative peptide masses, ranked by their "coarse score". This approach was appropriate because they used a stepped collision energy approach which ensured $peptide{+}Y$ would be present. This type of filtering approach has been proposed previously [150, 265], and I will refer to this procedure as peptide mass prediction filter (PMPF). In all published, cases they require complete glycan *topologies* rather than *compositions*. Under a simple biosynthetic model, there are 448 distinct *N*-glycan compositions derived are distinct 19,194 topologies. If we expand this to accommodate NeuGc and Gal(α1-3)Gal commonly found in mammals, we instead have 1,766 distinct *N*-glycan compositions and 296,514 distinct topologies.

Pre-construction involves a time-consuming initial enumeration of all possible candidates and stores them on disk with an index for extracting structures by their neutral mass. During runtime, when a precursor ion is considered, an interval around the precursor ion mass is read from disk. Extensive use of caching and appropriate batch management can be used to spread work out across multiple threads of execution to cover subgraphs of the spectrum-to-structure space while minimizing disk traffic. If the cost to enumerate at runtime exceeds the cost of reading just the required structures from disk, then if the same search space is used multiple times, the argument can be made that pre-construction strategy is optimal. In practice this is difficult to measure because disk performance can vary considerably depending upon system load and hardware. Because the pre-constructed search space must be saved to disk, a fast-to-search, indexed storage format is necessary. Several groups propose their own binary formats [136, 264, 266]. During an indexed traversal of the search space at runtime can be combined with other filtering methods like PMPF, though care must be taken when combining it with caching.

Search Space Components

As described in Sec. 1.4.3, the search space is comprised of an *in-silico* digestion of an input protein list, combined with a list of constant and variable modification rules, with the addition of glycosylation as a "variable modification". The input protein list may be derived from a FASTA file or an annotated protein sequence format, such as PEFF [196] or UniProt XML [195] in the case of more general PTM proteomics. It can also be advantageous to use existing proteomics search results on a closely related sample for the basis of a glycoproteome as I showed in [243] using mzIdentML [267]. A deglycosylated proteome from the a sample is a closer match to the glycoproteome of that sample than by simply guessing which glycoproteins may be present. By using more sophisticated methods for identifying smaller PTMss and non-specific cleavage sites, the peptidoforms can be specified exactly to reduce combinatorial expansion of modified peptides prior to generating glycoforms. These modifications can add several additional peptidoforms with glycosites that would otherwise be unavailable to a naive search space construction, and would either go unreported or mis-assigned.

When generating theoretical glycopeptides, each glycosite is combined with each glycan of an appropriate type, *N*-glycans to *N*-glycosites, *O*-glycans to *O*-glycosites, just as with other PTMs, though it may be far fewer if PMPF is used. If the correct glycan is missing from the database, spectra from that glycan may be mis-assigned, so the choice of glycan source is important. As discussed in Chapter 3, databases for glycans are not nearly as well developed as those for proteins. GlyTouCan [213] provides a database of reported glycan structures, the successor to Glycome-DB [268], while UnicarbKB [214] and GlyConnect [269] annotate their presence on known proteins. Much of the space of glycan structures remain unexplored, and a comprehensive database could be impractical to search against directly [70]. Approaches using biosynthetic simulation [185, 249], "expert curation" [132, 150, 263], and combinatorial expansion [23, 54] have been proposed. A combinatorial space spans the same region as *de novo* sequencing [270], though with many energies *de novo* sequencing of the glycan is impossible. An alternative approach was proposed in SweetNet [133], which used a small combinatorial list of starting compositions to extrapolate the remaining space of *N*-glycans, *O*-glycans, and GAG linker saccharides by using spectral networks to connect similar spectra and infer monosaccharide gain/loss from the difference between precursor masses. Additionally, a glycan space defined over compositions is much more compact than one over structures, but lacks much of the information required to reason about biosynthetic properties without making broad assumptions about potential motifs. These problems can be partially side-stepped by constructing a glycan structure space first biosynthetically, and then reducing it to compositions later, but it trades the assumption of biosynthetic pathways for the assumption of biosynthetic enzymes.

Implementation

I implemented a pre-construction based procedure using `SQLite3` [271] to store all theoretical peptides, glycans, glycan combinations, and glycopeptides. This makes my application's performance sensitive to disk speed, disk page size, the `SQLite3` query plan, and the average result set size. This method worked well for small databases and for large ones where the number of potential glycopeptides for a given query was relatively small but the space to traverse was large which held for Human data, but not necessarily for more general mammalian data. The schema is shown in Figure 4.3. The query that all precursor mass searches execute is shown in Listing 1.

As there is considerable overhead to simply initiate a query against the database

```sql
SELECT Glycopeptide.id, Glycopeptide.calculated_mass,
       Glycopeptide.glycopeptide_sequence, Glycopeptide.protein_id,
       Peptide.start_position, Peptide.end_position,
       Peptide.calculated_mass as peptide_mass,
       Glycopeptide.hypothesis_id
FROM Glycopeptide JOIN Peptide on Glycopeptide.peptide_id = Peptide.id
WHERE Glycopeptide.hypothesis_id = :hypothesis_id AND
      Glycopeptide.calculated_mass BETWEEN :lower_mass AND :upper_mass;
```

Listing 1: The mass search query used to extract theoretical glycopeptides from the disk. This uses the index over `Glycopeptide.calculated_mass` to quickly filter out invalid Glycopeptides, and uses a covering index over `Peptide` to retrieve only the relevant columns. The cost of the `JOIN` is trivial ($\leq 5\%$) compared to the cost of traversing the `Glycopeptide` index and table.

on disk, I assume that precursor masses will be densely clustered around similar locations, I extract intervals of 1 Da around the queried mass, not the exact mass accuracy interval requested, and cache these glycopeptides in memory. Successive queries check to see if they are fully contained in the cached interval and if so they do not require going to the disk. Partially overlapped interval requests are serviced using only the uncovered region of the mass interval, and include the same dense assumption of the query mass interval. For low performance disks, this may be prohibitively slow for very dense search spaces, like those of whole proteome glycoproteomes. In those cases, at least a partial runtime traversal of the search space is necessary. An on disk index of peptides by mass and an in memory index of glycans would perform well in both cases, particularly if combined with PMPF or other branch-and-bound constrained traversal.

A Simple Integration of Glycomics While my overarching goal to integrate glycomics will be discussed later in Chapter 5, because my suite of tools included glycan composition and structure-related components, I provide several ways to generate a glycan database. The first and simplest method, from the user's perspective, is an implementation of biosynthesis simulation. Using a set of taxonomy- and glycan class-specific enzymes, I simulate the action of glycosidases and glycosyltransferases, building up a graph of source-product relationships. This method was inspired by work done by Krambeck [68, 249] and Liu [69, 77, 225], though to my knowledge, neither attempted to deal with some common mammalian patterns like Galα-Gal and multiple fucosylation. After the simulation, all structures are collapsed into compositions, forming a multi-graph where each edge corresponds to an enzyme, allowing the user to choose to opt out of a specific enzyme easily. The complete simulation is saved to disk and common human and mammalian net-

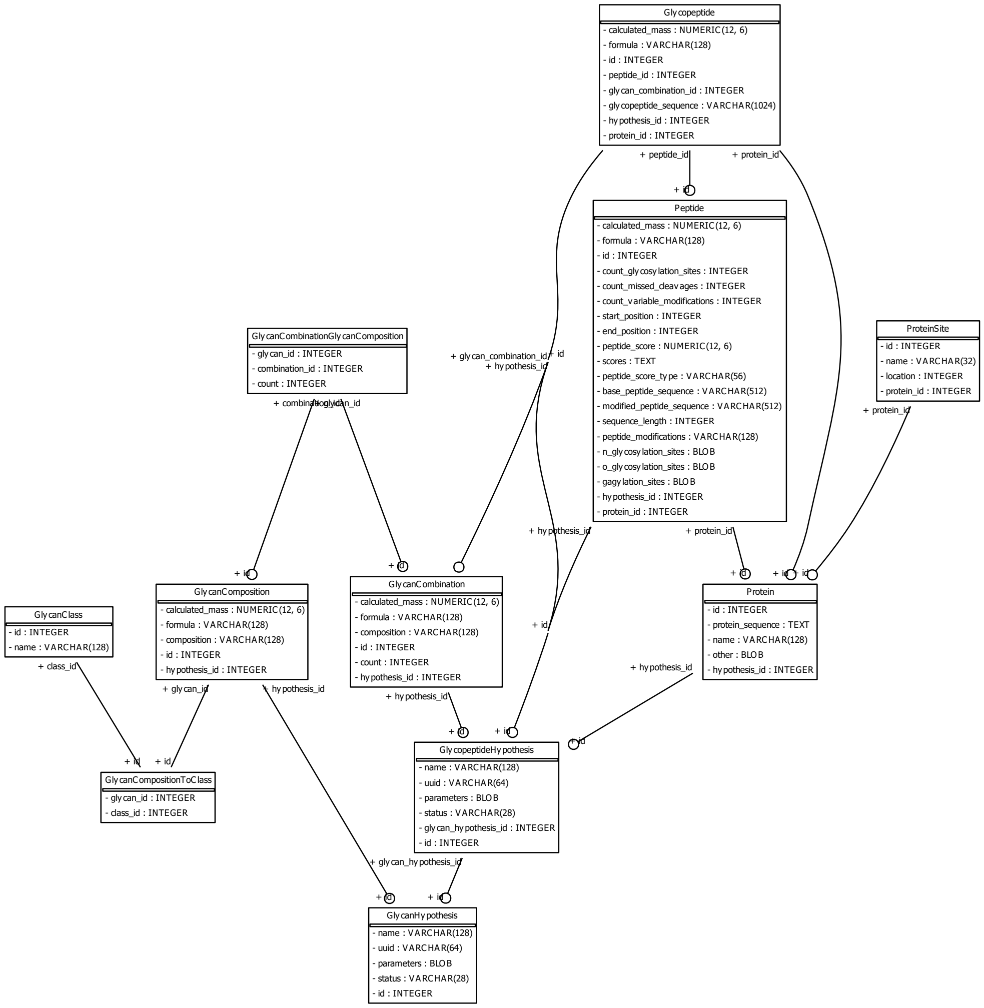


Figure 4.3: The schema of the database for describing theoretical glycopeptide search spaces I used.

works are pre-defined for convenience. The second method is through integration with the glycan structure repository GlyTouCan [213, 248] using SPARQL queries to pull down specific glycan classes and optionally taxonomic annotations carried over from Glycome-DB [268]. Lastly, if the user has glycomics data from their sample of interest, and a broader or unstructured glycan search space, my implementation can identify glycan compositions using the algorithm discussed in Chapter 3 and use them as the basis for the glycan combinations to construct. When no information is known, I offer a simple combinatorial expansion method, or generation from a user-provided text file.

Integration of Proteomics While proteomics database search engines have made the concept of performing a combinatorial expansion of variable modifications over an in-silico digest of the protein database almost pedestrian, there are still many problems to address here. A single variable PTM can expand the glycopeptide search space exponentially when it has more than one site on a glycopeptide sequence, making it challenging to include more than a single variable modification when searching a large glycoproteome [132, 263]. Including many missed cleavages, or even semi-specific digests becomes intractable, making the direct identification of glycopeptides near to signal peptide cleavage sites or non-tryptic cleavage sites impractical. I implemented both a feature extraction procedure to read cleavage sites from UniProt [189] inspired by G-PTM [195], and the ability to build a glycopeptide search space from the identified peptides, including modifications, and baseline unmodified peptides defined by an mzIdentML file [267]. This allows my method to build peptides with a wide array of modification states based upon a measured proteome, though it performs best with a deglycosylated proteome sample. This work was discussed in [54] and [243].

4.2.2 Complications

Glycopeptides are complicated enough on their own, but certain experimental conditions can make them worse. Glycopeptides be adducted, adding to their intact mass and changing the way they fragment. One common adduct found in several published sources is ammonium, which replaces a proton for a net gain of NH_3. As previously mentioned in Sec. 4.2.3, this composition shift can introduce incorrect glycan composition assignment, `NeuAc + NH`$_3$ into `Hex + Fuc`. The occurrence of ammonium adduction appears to be rarely acknowledged, with the only other glycoproteomics-specific discussion appearing in [133], where it was used as a modification for adding edges between spectral clusters. In [219], the authors comment upon the presence of an unidentified 17 Da mass shift observed on several high mannose *N*-glycans they observed. Ammonium adducts do not induce a retention time shift, and can be observed perfectly tracking with their unadducted parent species, as shown in Figure 4.4 displaying the trace for `TITNDQIEVTN(N-Glycosylation)ATELVQSSSTGR{Hex:7; HexNAc:2}`.

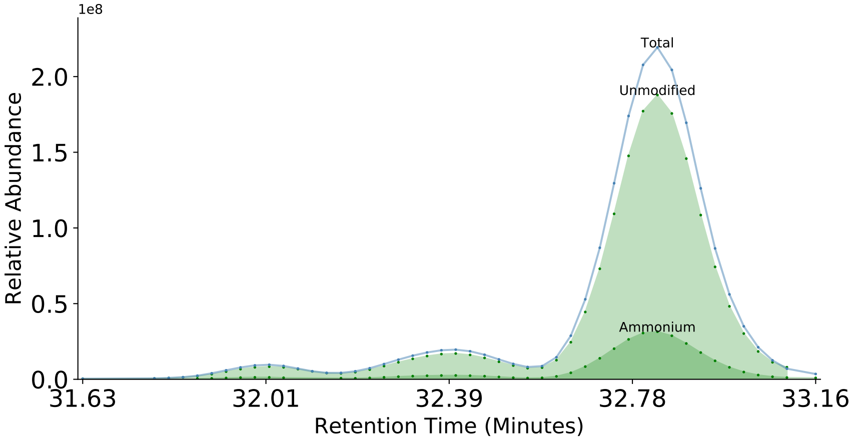


Figure 4.4: The extracted ion chromatogram of demonstrating the identical elution profiles of the adducted and unadducted forms of `TITNDQIEVTN(N-Glycosylation)ATELVQSSSTGR{Hex:7; HexNAc:2}`.

Another mass similarity is Hex$_3$ and HexNAc$_2$ + SO$_3$, which occurs readily on KS glycans like those reported from proteoglycans in [54, 133] or on viral glycans reported [94] and Ch. 3. Depending upon collision energy, HexNAc(S) may appear as a low abundance oxonium ion, and the ambiguous high mannose-like glycan might also be expected to produce an abundant Hex$_2$ peak.

In the Orbitrap Phil-BS glycopeptide samples from [23], I applied the CovBinom (Eq. 4.19) [54] scoring model to identify glycopeptides with sulfated glycans. An example spectrum is shown in Figure 4.5. This glycopeptide, N(N-Glycosylation)C(Carbamidomethyl)TLIDALLGDPHC(Carbamidomethyl)DGFQNEK{@sulfate:1;Hex:4;HexNAc:4}, would be ambiguous with N(N-Glycosylation)C(Carbamidomethyl)TLIDALLGDPHC(Carbamidomethyl)DGFQNEK{Hex:7;HexNAc:2}, and they are within 0.05 Da of each other, just over 10 PPM, though the sulfated glycopeptide has a mass accuracy of 0.8 PPM while the alternative is 10.4 PPM. This is of particular relevance because sulfated glycans are found on IAV that are actively circulating and included in vaccines [272] and appears to impact viral infection severity [253]. Of the 622 GPSMs found in Phil-BS-tryp-GP-1.raw at [23] 5% FDR, 17.6% were identified with either NH$_3$ or Na adducts, and 24.5% had sulfated glycans. In addition to sodium, other metallic cations can be found in attached to glycopeptides. The sample AGP-tryp-GP-1.raw from the same PRIDE repository shows several adducts on abundant species in Figure 4.6. These metallic cation adducts change the fragmentation pattern of the glycopeptide, reducing the fragmentation efficiency of the glycan, and splitting the signal between peaks with and without the cation adduct.

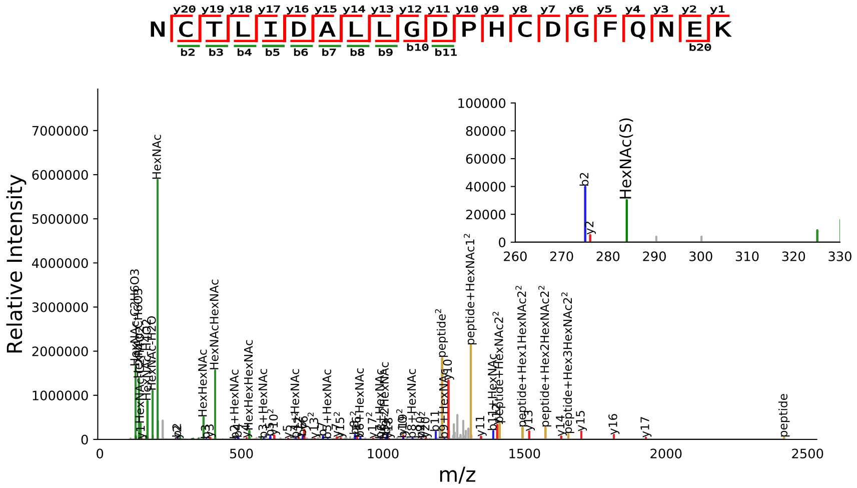
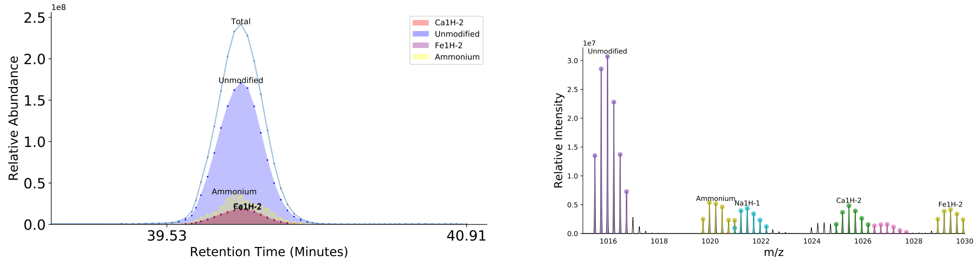


Figure 4.5: An example spectrum for a sulfated glycopeptide from IAV Hemagluttinin, `N(N-Glycosylation)C(Carbamidomethyl)TLIDALLGDPHC(Carbamidomethyl)DGFQNEK{@sulfate:1;Hex:4;HexNAc:4}`. Note the `HexNAc(S)` oxonium ion in the inset is low in abundance.



(a) The adducted extracted ion chromatograms for `LVPVPITN(N-Glycosylation)ATLDQITGK{Hex:6;HexNAc:5;Neu5Ac:1}`, showing the trace for the unmodified species, ammonium adducted species (NH_3), iron adducted species (FeH-2), and calcium adducted species (CaH-2). Not shown is the trace for the sodium adduct.

(b) The MS^1 scan corresponding to the apex of the chromatographic peak shown in Figure 4.6a, showing distinct isotopic patterns for each adduct and the parent species. Notice the NH_3 adduct and the NaH-1 adduct overlap, introducing isotopic interference, and potentially contaminating isolation windows.

Figure 4.6: An AGP glycopeptide, `LVPVPITN(N-Glycosylation)ATLDQITGK{Hex:6;HexNAc:5; Neu5Ac:1}` with many adducts.

4.2.3 Glycopeptide MS/MS Scoring Models

The design of the scoring models used by these algorithms depends upon the information expected to be contained in a mass spectrum, which in turn depends upon the collision energy or energies used [132], and the size of the glycopeptide [23]. This means that there is no one "best" model for collisional dissociation spectra. Additionally, there are electron-based dissociation techniques [88] which produce other types of fragments that many of these search engines would not be able to assign, and for which the technologies are not widely available. This work will focus on HCD- and Stepped HCD-based models. Because of the assumptions made when designing scoring functions, it is difficult to construct a fair comparison between two models when they were designed for different collision energies. For example, a model designed for lower energy HCD or CID would not be comparable to a model for higher energy HCD data. Within the same energy range, there are many ways to approach identification, and define an optimal identification.

Bond Coverage

A structure is matched against an MS^2 spectrum by mapping experimental peaks onto theoretical fragments. If we assume that only one precursor was fragmented, each experimental peak corresponds to a fragment from that precursor. Each fragment f_i corresponds to a bond in the original structure breaking, and matching a fragment implies that there is a bond in the precursor which connects a substructure with mass m_i to the remainder of the structure. If all bonds in a theoretical structure are observed, and the theoretical structure matches the precursor ion's mass, then the structure would be fully specified by the spectrum, and would constitute perfect identification.

This "coverage" model of identification makes two broad assumptions, the first is that every bond can be enumerated and observed, the second is that peaks do not match at random. The first assumption can be relaxed by defining what bonds can be observed to break and whether they are sufficiently enumerable. For most HCD spectra, these correspond to the peptide backbone bonds, which fragment to form b and y ions, of which there are $n_p - 1$ bonds where n_p is the number of amino acid residues in the peptide sequence of the query structure q. For notational brevity, the spectrum is denoted s, b_j, y_j, and Y_j are each indicator variables with value 1 if the indicated fragment was observed and 0 otherwise.

$$C_p(s, q) = \frac{1}{2(n_p - 1)} \sum_{i}^{n_p} (b_i + y_{n_p - i}) \tag{4.1}$$

$$C_p(s, q, k) = \frac{1}{n_p - 1} \sum_{i}^{n_p} \log_k(b_i + y_{n_p - i} + (k - 2)) \tag{4.2}$$

Using Eq. 4.1 creates a balanced coverage, observing the either the b ion for the ith bond or the corresponding y ion ($y_{n_p - i}$) is worth the same amount of coverage information, while Eq. 4.2 creates a weighted scheme placing more value on the first fragment, proportional to $\log_k(k-2)$ using an arbitrary base k s.t. $k > 2$. Glycan fragments may still be observable, but not consistently among all peptide length and glycan sizes, and if a glycan composition is used, there are no bonds to enumerate. A conserved core motif might be generated for glycan classes for which they are expected, as is the case for the three classes I consider here, *N*-glycans, mucin *O*-glycans and GAG linker tetrasaccharides, but these cannot be used to completely cover more elaborated structures. If all of these "pseudo-bonds" are expected to be observed, then glycan coverage can be described over these bonds alone, however, this assumes that larger glycan fragments do not convey additional information.

$$C_{g,s}(s,q,h) = \frac{min(\sum_i^{n_g} Y_i, h)}{h} \tag{4.3}$$

$$\tag{4.4}$$

Under this type of uncertainty, Eq. 4.3 defines a simple model of glycan coverage where Y_i corresponds to observing the ith glycan fragment and h is the number of $peptide{+}Y$ fragments observed on average in high confidence spectra.

Coverage may be further extended to peptide backbone fragments which are expected to carry glycosylation. For each occupied glycosylation site a fragment spans, it may appear unmodified or carry a remnant of the glycan reducing end (`HexNAc` for *N*-glycans and *O*-glycans, or `Xyl` for GAG linkers) cumulatively.

$$C_{gp}(s,q) = \frac{\sum_i^{n_p} b_{i,g} + y_{n_p-i,g}}{\sum_i^{n_p} b'_{i,g} + y'_{n_p-i,g}} \tag{4.5}$$

Using Eq. 4.5, I express glycosylated backbone coverage as the sum over each peptide bond position where the b_i or the y_{n_p-i} were observed with a glycan remnant, divided by the sum of the number of bond positions b'_i and y'_{n_p-i}which were could have produced a glycan carrying fragment. This expression will favor the solution which has the most fragments supporting a particular localization of a glycan.

We can express an aggregate notion of coverage by mixing these concepts. Combining Eqs. 4.2 and 4.5, we get a solution that favors peptide backbone coverage but puts more weight on the observation of glycosylated backbone fragments shown in Eq. 4.6. This construction uses a mixture parameter $\alpha \in [0, 1]$ which expresses how much weight to put on the peptide coverage portion and puts the remaining weight on the glycosylated coverage portion. When $\alpha = 0.5$, each component is weighted equally, which is the value I choose to use.

$$C_{\text{backbone}}(s, q, k, \alpha) = \alpha C_p(s, q, k) + (1 - \alpha) C_{gp}(s, q) \tag{4.6}$$

There has been little discussion in the literature of whether a glycosylation remnant peptide backbone fragment such as $b_4 + \texttt{HexNAc}$ should be considered as evidence for that the 4th peptide backbone bond was observed to break, if b_4 is not observed. From the limited description available [132, 153, 183, 185, 265] suggests that there is no consensus, and that not all methods search for these localizing ions. Additionally, for longer peptide sequences, the position of the glycosite becomes harder to reliably determine using these ions. They are often lower in abundance than their unglycosylated form and this can drive them below the limit of detection. Such pathological cases are difficult to model without a notion of whether or not an ion is expected to be abundant, and even then it is challenging to determine what the threshold would be for omitting such fragments from the enumeration the denominator in Eq. 4.5.

For higher energy HCD data, a complete coverage model can be specified by further mixing peptide coverage with the simple model of glycan coverage as shown in Eq. 4.7. The parameter γ may be adjusted to put more or less weight on $peptide{+}Y$ fragments. I chose to use $\gamma = 0.7$ for most HCD datasets where the $peptide{+}Y$ fragments were not abundant. Independently, pGlyco2 chose a similar mixture model with very close mixing parameters in [132, 186].

$$C(s, q, k, h, \alpha, \gamma) = \gamma C_{\text{backbone}}(s, q, k, \alpha) + (1 - \gamma) C_{g,s}(s, q, h) \tag{4.7}$$

A coverage score alone has many shortcomings. The first is that it does not take into account charge state at all. Observing the same fragment in a single charge state is worth the same as observing it in two, three, or ten charge states, which means it discards valuable information. It also makes no difference whether a fragment matches a high intensity peak or a low intensity peak. Lastly, it places a heavier burden of proof on longer sequences and larger structures as they have more bonds to cover to achieve the same score a smaller one would from a spectrum with the same information content.

Fragment Match Counting

Coverage may be difficult to define because it assumes that all sequences can be evenly covered independent of sequence length. An alternative is to simply count fragment matches in some way.

$$Count(s, q) = \sum_{i}^{n_p} b_i + y_{n-i} + \sum_{i}^{n_g} Y_i \tag{4.8}$$

$$FragBinom(s, q) = \sum_{i=k}^{n_p} \binom{n_p}{i} p^i (1-p)^{n_p - i} \tag{4.9}$$

In Eq. 4.8, the number of fragments matched is just counted. This simple score was first used by Morpheus [126], and later adapted for glycopeptides by GPQuest [183]. Another common approach shown in Eq. 4.9 is instead model the event that k fragments matched out of n_p as a binomial event with some probability p which is a function of the mass error tolerance used and the precursor mass. This can be used to compute a probability of occurring by random chance using an upper tail test [133, 160, 266].

Both of these methods are less constrained than coverage, but they can both still misuse noisy spectra to assign many low abundance peaks to inflate a the score of a low quality match over one that uses more high abundance peaks but matches fewer fragments overall.

Intensity Utilization

Many algorithms attempt to utilize intensity in some way to improve the score of a glycopeptide. There are many ways to do this, and each comes with its own caveats. For notational brevity, I is a vector of peak intensities from the matched spectrum of size n and w_i is an arbitrary weight for each matched peak and is 0 for all unmatched peaks.

$$LogInt(s,q) = \sum_{i}^{n} \log_{10} I_i w_i \tag{4.10}$$

$$PercInt(s,q) = I^{-1} \cdot \sum_{i}^{n} I_i w_i \tag{4.11}$$

$$BinomInt(s,q) = \prod_{a=1}^{4} \sum_{i=s(m_a)}^{s(m_{a-1})} \binom{s(m_{a-1})}{i} p^i (1-p)^{n-i} \tag{4.12}$$

In Eq. 4.10, the total score is simply a weighted sum of the $\log_{10}$ scaled intensities. This is good in that it puts more weight on more intense peaks, but it makes the overall intensity scale matter when comparing matches between spectra, as when calculating an FDR. By using this construct, we implicitly trust matches to more abundant precursors more. This formulation is used in parts of pGlyco2's scoring functions [132] where $w_i = \left(1 - \left|\frac{e_i}{e_{\text{tol}}}\right|^4\right)$, e_i is the mass error of the ith peak, and e_{tol} is the maximum mass error tolerated. This allows them to use a wider than expected error tolerance window while managing score inflation from low accuracy fragment matches, though these matches still contribute to the coverage components equally to other fragments.

In Eq. 4.11, if $w_i = 1$ for all matched peaks, we obtain the percentage of intensity matched. This construct is used in SweetNET [133] to build their Validation Score of $FragBinom(s,q,n,k,p) \times PercInt(s,q)$ (Eq. 4.9). The final score of GPQuest2 constitutes $PercInt(s,q) + Count(s,q)$

The construction in Eq. 4.12 uses $s(m_a)$ to denote the number of peaks in s where their intensity is >= the ath median intensity, with $p = 0.5$ from Peppy, described in [160]. Successive medians are taken from only those peaks above the previous median. This is also combined with $FragBinom$ in Peppy to form a log-transformed p-value score. Unlike other features discussed so far, $BinomInt$ is not monotonic w.r.t. new peak matches, in that adding a new fragment match to a low abundance peak can result in a lower score overall. In that sense, $BinomInt$ enforces a parsimony of matched fragments.

Auxiliary Features

Some components could not be used as scoring functions in their own right, but serve to more gently guide other scoring functions or to work as pre-scoring filters. For example, GPQuest2 [183] fit an SVM to predict whether a spectrum was *N*-glycan-like or *O*-glycan-like by looking at the ratio of the intensity of each oxonium ion with the oxonium ion for `HexNAc`. SweetNet [133] used a similar intensity ratio of four neutral losses of `HexNAc` to classify whether they were derived from `GalNAc` or `GlcNAc`, which implied whether a spectrum were *N*-glycan-like or *O*-glycan-like through a different route. SweetNet also formalized a simple filter for classifying a spectrum as being glycopeptide-like or not, the G-score shown in Eq. 4.14, unrelated to the G-test from 2.3.2. The heuristic can be used to quickly filter out spectra from a sample containing a mix of glycosylated and unglycosylated peptides. It uses several oxonium ions derived from HexNAc.

$$H = \Big\{203.079, 185.0688, 167.0582, 143.0582, 125.0476, 137.0476\Big\} \tag{4.13}$$

$$GScore(s) = \frac{1}{|H|}\sum_{o \in H} \frac{I_o}{\max(I)} * 100 \tag{4.14}$$

Some monosaccharides, such as NeuAc and NeuGc have distinguishing oxonium ions which must be present for a match with those monosaccharides included to be believable. This problem is exacerbated by the fact that $\texttt{NeuAc} + \texttt{Hex} = \texttt{NeuGc} + \texttt{Fuc}$ exactly in both mass and elemental composition, requiring signature ions to discriminate them. Byonic [263] includes a validation step where it seeks out these signature ions, but how this influences the score is unclear. I implement my own signature ion score penalty described in Eq. 4.17. The notation I_o refers to the intensity of the oxonium ion for the monosaccharide o, which may be 0 if missing, and $q[o]$ refers to the number of occurrences of monosaccharide o in the glycan composition of q, which may be 0.

$$UnexpSigIon(s, o) = 10\log_{10}\left(1 - \frac{I_o}{\max(I)}\right) \tag{4.15}$$

$$MissSigIon(s, q, o) = 10\log_{10}\left(1 - \min(q[o] * 0.5, 0.99)\right) \tag{4.16}$$

$$SigIonScore(s, q) = \sum_{o \in \{\texttt{NeuAc}, \texttt{NeuGc}\}} \begin{cases} UnexpSigIon(s, o) & q[o] = 0 \\ MissSigIon(s, q, o) & q[o] > 0 \;\&\; \frac{I_o}{\max I} \leq 0.01 \\ 0 & \text{otherwise} \end{cases} \tag{4.17}$$

There are cases where multiple glycan compositions are within 10 PPM of each other at higher precursor masses, such as $|(\texttt{Hex} \times 3) - (\texttt{HexNAc} \times 2 + \text{SO}_3)| = 0.04291$

at 4291.0357 Da, or $|(\texttt{NeuAc} + \mathrm{NH}_3) - (\texttt{Hex} + \texttt{Fuc})| = 0.01123$ at 1123.3 Da. It may not be possible to diagnose the presence of this type of ambiguity from the product ions directly. If we assume that precursor mass errors are normally distributed [145, 158], then matches with smaller precursor mass errors are more likely. Its only function is to discriminate between two cases with identical scores, so its scale may be kept small.

$$MassAccScore(s,q) = -10\log_{10}\left(1 - \exp - \frac{(e_{precursor} - \mu_{precursor})^2}{2\sigma^2_{precursor}}\right) \tag{4.18}$$

Complete Models

Several models can be described by composing one or more of the features described above into new or useful structures. Several published models are direct combinations as already described. The model used in [54] is shown in Eq. 4.19. It scales the binomial fragment matching (Eq. 4.9) and intensity utilization (Eq. 4.12) by the total coverage for higher energy HCD (Eq. 4.7), shifted by signature ion errors (Eq. 4.17) and mass accuracy (Eq. 4.18)

$$\begin{aligned} CovBinom(s,q) =& (-10\log_{10}(FragBinom(s,q)) + -10\log_{10}(BinomInt(s,q)))\times \\ & C(s,q,3,3,0.5,0.7) + SigIonScore(s,q) + MassAccScore(s,q) \end{aligned} \tag{4.19}$$

This score works reasonably well for spectra dominated by peptide backbone fragmentation, favoring solutions which make the best use of of the spectrum's intense peaks while covering most of the sequence. It also does not have a natural decomposition for evaluating just the peptide or just the glycan component of the match.

pGlyco2's scoring function partitions peaks into peptide- and glycan-matching groups, and proceeds to combine them with a log-intensity (Eq. 4.10). There, $I_{i,p}$ or $I_{i,g}$ are the intensity of the ith matched peak if they correspond to a peptide or glycan fragment, 0 otherwise. This differs from a directly scaling log-intensity by full coverage (Eq. 4.7) because it partitions the value of the intensity used for the peptide matches and glycan matches, so that coverage of one component does not affect the other component. They fit model parameters $\alpha = 0.56$, $\beta = 0.42$, $\gamma = 0.94$ and $w = 0.35$. They use a formulation of peptide coverage which reduces to C_p (Eq. 4.1), but their definition of glycan coverage requires a glycan topology, not a composition.

$$PepScore(s,q,\gamma) = \sum_{i}^{n} \log_{10} I_{i,p} \left(1 - \frac{|e_i|}{e_{tol}}^{4}\right) \times C_p(s,q)^{\gamma} \tag{4.20}$$

$$GlyScore(s,q,\alpha,\beta) = \sum_{i}^{n} \log_{10} I_{i,g} \left(1 - \frac{|e_i|}{e_{tol}}^{4}\right) \times \text{ratio}_{\text{ion}}^{\alpha} \times \text{ratio}_{\text{core}}^{\beta} \tag{4.21}$$

$$TotalScore(s,q,\alpha,\beta,\gamma,w) = w \times GlyScore(s,q,\alpha,\beta) + (1-w) \times (PepScore(s,q,\gamma)) \tag{4.22}$$

A topology can be precisely enumerated using a tree or graph traversal method, with the knowledge that every fragment theoretically produced might appear if that topology is the correct one. With a composition, fragments cannot be enumerated precisely. For *N*-glycans, the conserved chitobiose core can be enumerated, and if side groups such as `Fuc` or `Xyl` are present in the composition they may be on the core. A procedure for generating these fragments from a composition is described in Alg. 5, which first produces the conserved motif fragments which may be guaranteed, and then produces extended fragments beyond the conserved motif. It also attempts to add a `Fuc` to the fragments as soon as possible because it cannot know whether that residue is located on the core or on an antenna.

For the structure `Fuc(a1-6)[Neu5Ac(a2-6)Gal(b1-4)Glc2NAc(b1-2)Man(a1-`

`6)[Neu5Ac(a2-3)Gal(b1-4)Glc2NAc(b1-2)Man(a1-3)]Man(b1-4)Glc2NAc(b1-4)]Glc2NAc`, with the composition `{Fuc:1; Hex:5; HexNAc:4; Neu5Ac:2}`, the composition fragments produce six fragments that are not possible from this structure, such as `{Hex:5; HexNAc:3}` and `{Hex:5; HexNAc:2}`, because it cannot know that those residues are not directly attached to the chitobiose core if the the structure is fully complex-type. Additionally, the structure fragmentation process produces fragments that may explicitly contain `NeuAc`, which the composition fragment generation method expressly ignore because HCD breaks those bonds first, making them unlikely to be observed in any case. Furthermore, a full enumeration of glycan fragments would also include $peptide{+}Y$ fragments with glycan fragments containing in excess of 10 monosaccharides, which are unlikely to appear, even in lower collision energy data, which ensures the burden of proof weighs down the score for large glycans where many of those extended fragments will not be observed. This is made worse for heavily fucosylated structures where the total number of fragments to observe is multiplied by 1 + the number of `Fuc` residues, which may also fall off easily under HCD [273]. It is arguable that this exacerbates the burden of proof on larger structures beyond what is reasonable to expect from the data.

I propose an approximation of pGlyco2's glycan coverage for glycan compositions in Eq. 4.25. The number of theoretical fragments produced by Alg. 5 is larger than the set of real fragments for the same number of bond cleavages, so a without a known topology, an approximated normalization factor $B(n_g)$ is used in place of the exact count of expected theoretical fragments.

$$C_{g,e}(s,q) = \frac{1}{B(n_g)} \sum_{i}^{n_g} Y_i \tag{4.23}$$

$$C_{g,c}(s,q) = \frac{1}{n_{\text{core}}} \sum_{i}^{n_{\text{core}}} Y_{i,\text{core}} \tag{4.24}$$

$$C_g(s,q) = C_{g,c}(s,q) \times \min\left(C_{g,e}(s,q), 1\right) \tag{4.25}$$

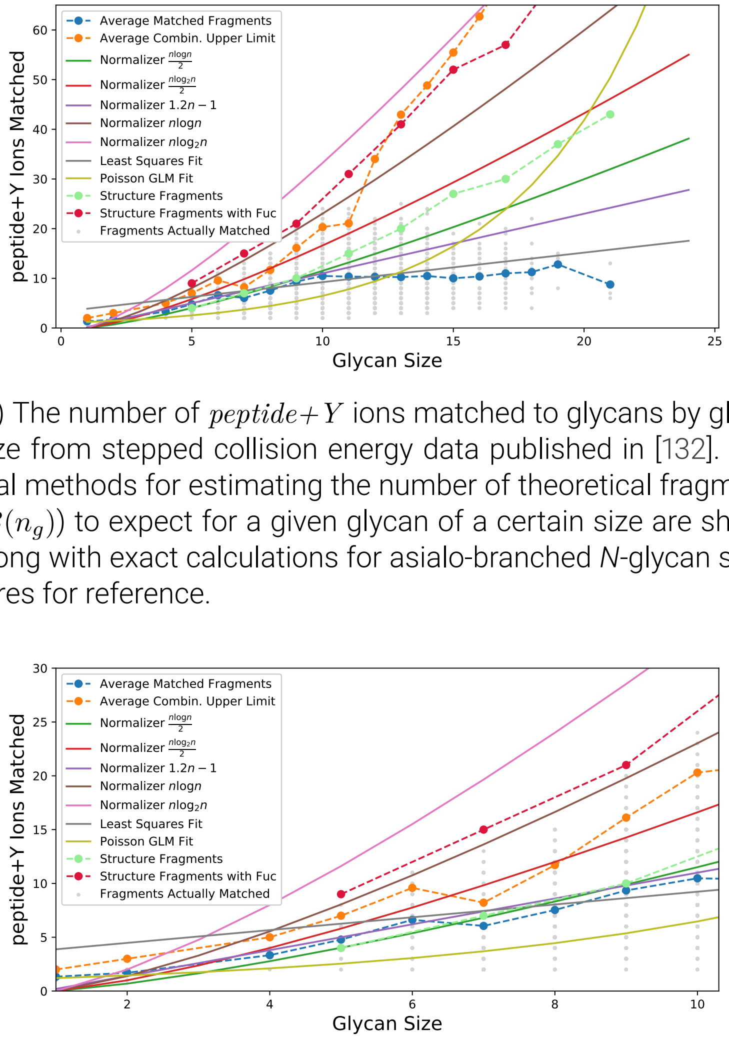
$$B(n_g) = \frac{n_g \log n_g}{2} \tag{4.26}$$

There are many ways to define B, as shown in Fig. 4.7. Using data published in [132], I counted the number of $peptide{+}Y$ ions matched for each glycan size, along with theoretical bounds calculated using the number of Y fragments produced from asialo-*N*-glycan structures with and without core fucosylation with up to five glycosidic cleavages. Using techniques to learn the number of fragments to expect by using the number of counts observed alone failed to produce adequate results because, particularly for large glycans, the observed fragmentation was rarely complete, impacting both the linear regression and the Poisson Generalized Linear Model (GLM) fits.

N-glycans are branching structures, similar to binary trees. The height of a balanced binary tree with n nodes is $\log_2 n$, which is the length of a single branch. Because $peptide{+}Y$ ions include cleavage events in multiple branches, it is not possible to simply reduce this to an upper limit of number of branches × length of branch^2. This introduces a combinatorial component on the order of the branching of the glycan. For most unfucosylated canonical *N*-glycans, the degree of branching is small, less than six, often much less. It follows then that the upper limit may be closer to $n \log_2 n$. I compared $n \log_2 n$ and $n \log n$ to the theoretical structural counts within the branching intervals of interest, and observed the natural log under-estimated while

$\log_2$ overestimated the fucosylated series and dividing by 2 similarly enclosed the unfucosylated series, with close tracking at small sizes. Given the desirability of a measure that would be less severe than the exact count for large glycans while still being accurate for small glycans, $B(n_g) = \frac{n \log n}{2}$ is an appropriate choice. Accounting for multiple `Fuc` would be more difficult due to spontaneous gas phase re-arrangement [273], and that fucosylation might be either branch or core bound rather than assuming the canonical core bound fucosylation as the default. It also is reasonable to truncate the glycan composition to remove labile units like sialic acids (`Neu`, `NeuAc`, `NeuGc`) and ignore excess `Fuc`. This lets us substitute $C_g(s, q)$ for ratio$_{ion}$ in Eq 4.21, while using glycan compositions, which can substantially expand the range of possible structures we can consider. This approximation is not without its weaknesses. The glycan composition `{Hex:11; HexNAc:2}` and `{Hex:5; HexNac:4; NeuAc:2}` are within 16 Da of each other. Both compositions contain the subset `{Hex:5; HexNAc:2}`, but while it is expected as part of any high mannose-type *N*-glycan's fragment ladder, that particular fragment is only possible if the glycan is a hybrid glycan with either bisecting `GlcNAc` or `LacDiNAc` with multiple `NeuAc` on the complex arm. While this is theoretically possible, it is highly unlikely, and that glycan composition usually represents the canonical bi-antennary complex type *N*-glycan, but the composition has no way to express this. The mass difference could be explained by a deconvolution error paired with an ammonium adduct.

This scoring model still lacks the ability to select glycosylation sites, given multiple options with the same number of observed ions. Given the glycopeptides `VTLIT{O-Glycosylation}SE`, `VTLITS{O-Glycosylation}E`, and `VT{O-Glycosylation}LITSE`, and only $y4{+}HexNAc$ is observed with high confidence, it can rule out `VT{O-Glycosylation}LITSE`, but it cannot make a value statement regarding the other localizations. Parsimony would dictate that the glycosylation site is located at the Threonine



(a) The number of $peptide{+}Y$ ions matched to glycans by glycan size from stepped collision energy data published in [132]. Several methods for estimating the number of theoretical fragments ($B(n_g)$) to expect for a given glycan of a certain size are shown, along with exact calculations for asialo-branched *N*-glycan structures for reference.

(b) As in Fig 4.7a, zoomed in on the small glycan range to visualize the more common cases prior to substantial divergence. In particular, the alignment of $\frac{n\log n}{2}$ and $1.2n-1$ with the unfucosylated structure series and $n\log n$ with the fucosylated series is noteworthy.

Figure 4.7: Approximating $peptide{+}Y$ ion matches expected given glycan size

at position 5, though it remains possible the glycosylation is located at position 6. The addition of a glycosylated coverage-specific term to the total score would least impact existing performance while introducing a slight bias towards more localized solutions at the top level. This can be accomplished by defining a localization bonus based on Eq. 4.5 with some small weight λ. When combined with Eq. 4.22 at $\lambda = 10$, SigIonScore (Eq. 4.17), and MassAccScore (Eq. 4.18). I define a modified version of the pGlyco2 scoring model, which I will refer to later as the "naive scoring model", shown in Eq. 4.27.

$$\begin{aligned} NaiveScore(s, q, \alpha, \beta, \gamma, w, \lambda) = w \times GlyScore(s, q, \alpha, \beta) + \\ (1 - w) \times (PepScore(s, q, \gamma)) + \\ SigIonScore(s, q) + \lambda C_{gp}(s, q) + \\ MassAccScore(s, q) \end{aligned} \tag{4.27}$$

Algorithm 5: *N*-Glycan Composition Fragment Generation

```
Data: Glycan Composition Mapping Monosaccharide To Count $G$
Result: List of Glycan Fragments $F$ Possible From $G$
$F \leftarrow \emptyset$;
$HexNAcInAggregate \leftarrow G[$HexNAc$]$;
$HexInAggregate \leftarrow G[$Hex$]$;
$FucInAggregate \leftarrow G[$Fuc$]$;
$BaseHexNAc \leftarrow \min(HexNAcInAggregate + 1, 3)$;
$BaseHex \leftarrow \min(HexInAggregate + 1, 4)$;
for $HexNAcCount \in [0, BaseHexNAc)$ do
    if $HexNAcCount = 0$ then
        Append($F$, $\{\}$);
    else if $HexNAcCount = 1$ then
        $f \leftarrow \{$HexNAc $: HexNAcCount\}$;
        Append($F$, $f$);
        // If there are any Fucose in the composition, copy the current
        // fragment and add a Fucose to it and add that to the set of fragments
        if $FucInAggregate > 0$ then
            Append($F$, FucosylateShift(Copy($f$)));
        end
    // At this point, there are 2 HexNAc, and Hexose can begin to appear
    // from the chitobiose core
    else
        $f \leftarrow \{$HexNAc $: HexNAcCount\}$;
        Append($F$, $f$);
        if $FucInAggregate > 0$ then
            Append($F$, FucosylateShift(Copy($f$)));
        end
        for $HexCount \in [1, BaseHex)$ do
            $f \leftarrow \{$HexNAc $: HexNAcCount,$ Hex $: HexCount\}$;
            Append($F$, $f$);
            if $FucInAggregate > 0$ then
                Append($F$, FucosylateShift(Copy($f$)));
            end
            // Begin generation of extended fragments beyond the conserved core
            if $HexCount = 3$ && $HexNAcInAggregate > 2$ then
                for $ExtraHexNAc \in [0, HexNAcInAggregate - HexNAcCount]$ do
                    $f \leftarrow \{$HexNAc $: HexNAcCount + ExtraHexNAc,$ Hex $: HexCount\}$;
                    Append($F$, $f$);
                    if $FucInAggregate > 0$ then
                        Append($F$, FucosylateShift(Copy($f$)));
                    end
                    if $HexInAggregate > 3$ then
                        for $ExtraHex \in [1, HexInAggregate - HexCount]$ do
                            $f \leftarrow \{$HexNAc $: HexNAcCount + ExtraHexNAc,$ Hex $: HexCount + ExtraHex\}$;
                            Append($F$, $f$);
                            if $FucInAggregate > 0$ then
                                Append($F$, FucosylateShift(Copy($f$)));
                            end
                        end
                    end
                end
            end
        end
    end
end
return $F$
```

False Discovery Rate Estimation

Just as in bare peptides, glycopeptide FDR estimation is not approached using a single model, or a strong distribution assumptions [54, 132, 153, 169, 173, 184]. Instead, one of a variety of empirical models are used, which only provide an accurate estimate asymptotically. The prevailing methods are based on Target-Decoy [170]. As in proteomics, spectra are searched against decoy proteins alongside the target proteins, and matches must compete for ownership of spectra. Some methods attempt to exploit structural properties of their scoring model to accomplish better separation [153, 184, 186] or employ hierarchical filtering [132, 133, 183, 185] to combine several weak filters into a strong filter.

Care must be used when defining the criteria by which a glycopeptide is graded and later filtered. If the score used is based on both peptide and glycan evidence, this can produce identifications with "reasonable" scores while leaving either the peptide or the glycan poorly characterized as shown in Figure 4.8. This is often the case for HCD glycopeptide identification where only a single energy is used, which cannot guarantee to fragment all glycopeptides equally well [187, 274–276]. Stepped collision energies like those described in [132, 187, 277] more consistently produce large, multiply charged $peptide{+}Y$, smaller $peptide{+}Y$, b and y fragments which can be used to characterize both the glycan and the peptide component independently. This is the reasoning behind Eqs. 4.22, 4.21, and 4.20, where the total score reflects both components, but there remains a separate score for the peptide evidence and the glycan evidence to be assessed on independently.

I implemented the finite mixture model FDR estimation procedure described in [132, 186] to assess performance when benchmarking against stepped collision energy data. This method compares each spectrum to a forward-protein peptide with

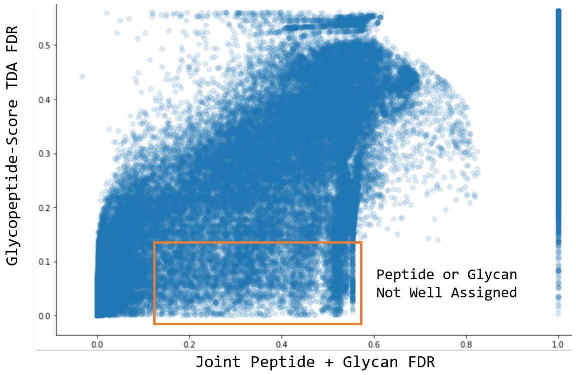


Figure 4.8: A comparison of intact glycopeptide scoring function filtering with joint peptide and glycan scoring functions filtered together. A large population of glycopeptides identified using just an intact aggregate score reports many glycopeptides with either low quality peptide or glycan identification.

correct $peptide{+}Y$ fragment masses (TT), a forward-protein peptide with randomly shifted $peptide{+}Y$ fragment masses (TD), a reverse-protein peptide with correct $peptide{+}Y$ masses (DT), and lastly a reverse-protein peptide with randomly shifted $peptide{+}Y$ masses. The GlyScore of TD is used to fit a Gamma mixture model Γ and the GlyScore of TT are used to fit a Gaussian mixture model $\mathcal{N}$ which contains as an additional component a weighted version of Γ with weight π_d. The PEP of each observation is given by $\frac{\pi_d \Gamma(m_g)}{\mathcal{N}(m_g)}$ where m_g is a vector of glycan scores, which when averaged yields the glycan-level FDR at m_g. The peptide-level FDR is calculated using traditional TDA [170] over TT and DT where the PepScore is used to rank GPSMs. A glycopeptide FDR is also computed from the TotalScore of TT and DD using the same finite mixture model method for the GlyScore component. The total FDR is then given by $\max(\text{fdr}_{peptide}, \text{fdr}_{glycan}, \text{fdr}_{peptide} + \text{fdr}_{glycan} - \text{fdr}_{glycopeptide})$.

Earlier Work Previously published work used intact-glycopeptide score based TDA using reverse-peptide decoys instead of reverse protein decoys. Reverse peptide decoys share the $peptide{+}Y$ ion ladder, forcing all discrimination to be based on b and y

ions but allowing the $peptide+Y$ ions to contribute to the selection of the best matching candidate overall. Because a single energy is insufficient for all glycopeptides, especially short peptides, it appeared useful to calculate the FDR of glycopeptides whose amino acid sequence was less than 10 residues long separately from longer sequences. Some of this perceived utility may have been caused by the abundance of short, poorly fragmented glycopeptides in AGP, from which much of the intuition behind those earlier scoring models were derived.

Naive Results

The naive scoring model should behave similarly to the pGlyco2 scoring model, save that it can use a much wider range of glycan definitions. Using a combination of the fixed glycan structure database published with the tool, and a biosynthetic simulation using mammalian glycosyltransferase rules, I identified glycopeptides in the brain tissue samples from [132]. I compared the identification performance using the same finite mixture model FDR estimation procedure, shown in Figure 4.9. The naive model performs better at all FDR levels, though this is likely due to the expanded search space covering more glycans as shown in Figure 4.10. The addition of the localization was observed to reduce the number of PSMs at higher FDR, but the difference is negligible.

4.2.4 Learning Fragmentation Processes

Peak Relationships

A fragment ion match in isolation may happen due to random chance. Several probabilistic approaches have been proposed for modeling whether a peak match is likely to be real or not. This concept is implicit in the binomial fragment matching model

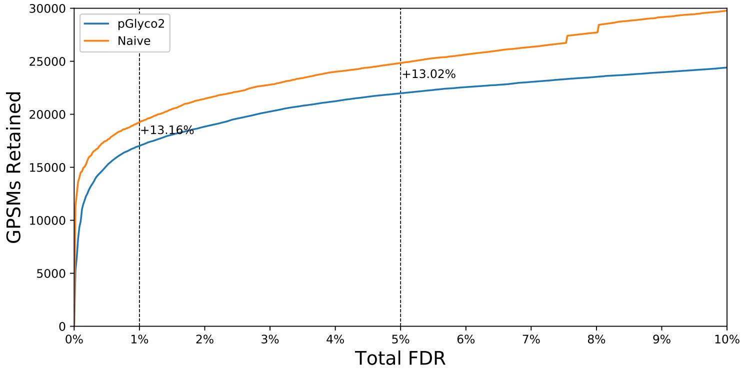


Figure 4.9: Comparison of originally published results for pGlyco2 [132] with my naive model using the aggregated composition glycan database. The naive model performs better at all FDR levels, though this is likely due to the expanded search space.

(Eq. 4.9). This approach was first proposed in SHERENGA [204], which learned the probability of b and y ion series for *de novo* peptide sequencing, and was later included in several other methods from the same group [8, 136, 163], as well as in other groups [205, 278] and implicitly this appears in several other scoring algorithms which treat b and y ions differently [138, 160]. Among these is these tools, UniNovo [8] is distinctive in that it codifies the process of learning arbitrary inter-peak relationships, rather than a fixed set of common ones, especially for a more complex structure like a glycopeptide.

As previously mentioned, glycopeptides produce a wider array of product ion charge states, and have more product ion types than bare peptides, and have more dimensions over which to partition the model than bare peptides. By estimating parameters by peptide length, glycan size, and proton mobility index [279] it is possible to measure how the glycan influences the efficiency of each fragmentation process. I fit these models on two large cohorts with different collision energies. The first was a stepped collision energy dataset using 20, 30, and 40 neV published with [132] derived from five different mouse tissues, shown in Figure 4.11. The second was a single, high collision energy dataset using 36 neV from TMT10plex tagged

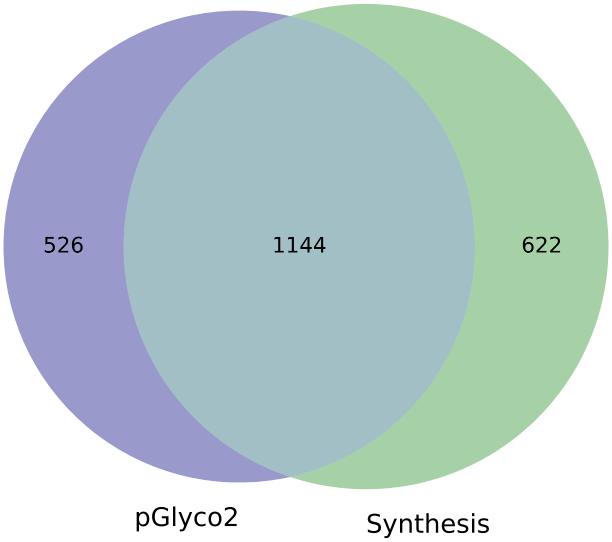


Figure 4.10: Comparison of the *N*-glycan compositions found in the database bundled with [132], from Glycome-DB [268] and those produced by biosynthesis simulation from common mammalian glycoenzymes. [132] included all *N*-glycan structures from humans, mice, and some from yeast. It also includes some structures that are ostensibly *O*-glycans. I used the union of these two databases to identify glycopeptides.

glycopeptides derived from human Cerebrospinal Fluid (CSF), enriched for `NeuAc`, acquired through a collaborator, shown in Figure 4.12. The MS^2 spectra in these training datasets were identified using CovBinom (4.19) during an initial survey.

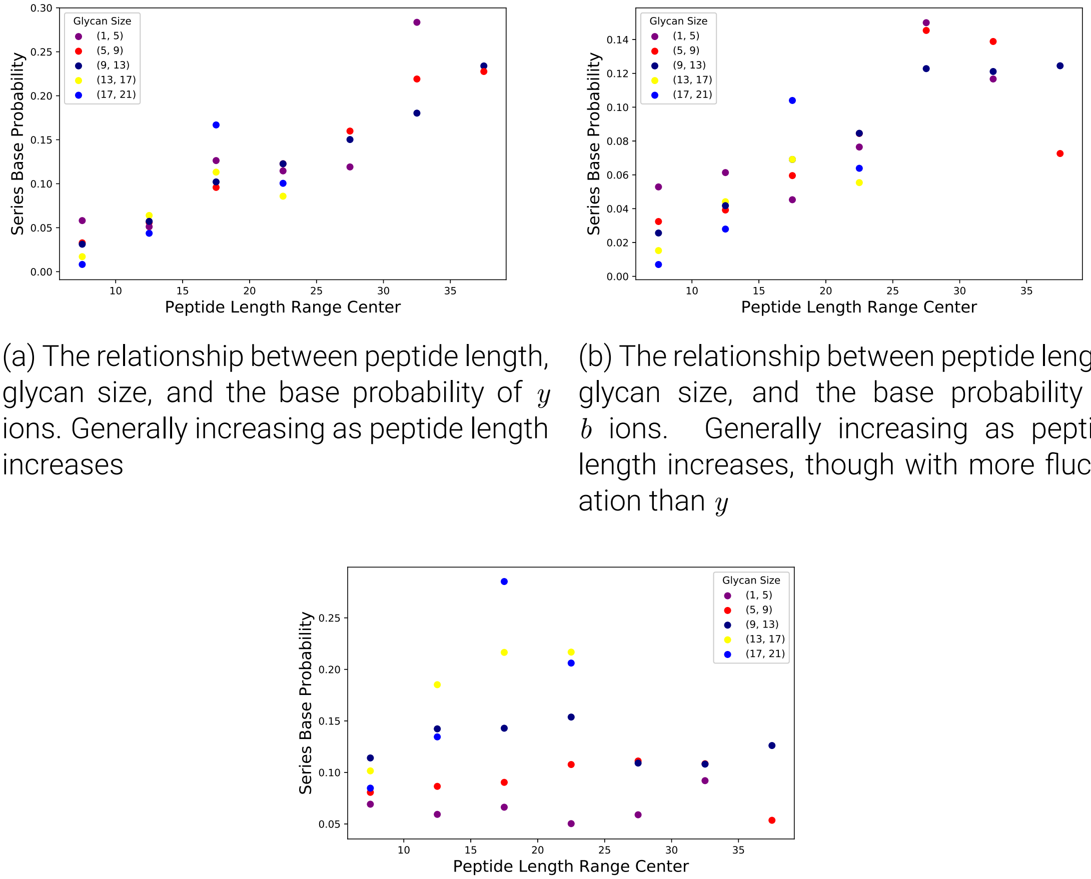
The same general trends are visible for b and y ions for both datasets, though not for $peptide{+}Y$ ions. This suggests that a component of those common trends is driven by the higher energy steps on the ramp. The differences in the $peptide{+}Y$ ion trends are likewise expected because the larger examples of that series are produced by the low energy step, meaning only large glycans can reliably produce them under higher energies. Because of the intact mass dependent nature of neV, if an energy is selected to dissociate both the glycan and the peptide, and the glycan dissociates more easily than the peptide, it follows that the glycan will be dissociated faster as more energy is applied. This may be exacerbated by the coincidence that the addition of a single glycosidic bond adds more mass on average than the addition of a single amide bond.

Because of the larger charge state ranges common to glycopeptides, I looked at the relationship between partition and the posterior probability of the difference-of-charge feature in Figures 4.13 and 4.14. These violin plots depict the distribution of the posterior probability for different specializations of the charge-difference feature function as defined by UniNovo [8]. Each specialization corresponds to an intensity ratio and a charge state pair. The relationship between aggregate size and charge state appears to be conserved between the two experimental conditions, in keeping with the intuition that larger molecules tend to carry higher charge states and produce more highly charged fragments.

In addition to producing multiply charged peptide backbone fragments, glycopeptides can also produce peptide backbone fragments with small amounts of intact glycosylation, usually only a single `HexNAc`, though occasionally multiple `HexNAc` may be present. This also serves as a common mass shift in $peptide{+}Y$ fragments. Again, in both datasets, as peptide size increased, the propensity for these fragments increased for b and y ions, and remained consistently high in reliability for $peptide{+}Y$ ions.

Beyond these two features, I also included `Hex`, `Fuc`, as well all common amino acids as link functions as described in the original publication of UniNovo, as well as Carbamidomethylation of Cysteine Oxidation of Methionine as well as the reducing end monosaccharide masses from *N*-glycopeptides, *O*-glycopeptides, and GAG linker glycopeptides.

For each matched peak, the posterior probability of that peak match was computed as described in [8], though no restrictions were made on the number of features that could contribute to a peak's posterior beyond that a single peak pair can only be counted once for each terminal. I omitted the feature selection step because, unlike in the original publication, I did not look for amino acid neutral losses like NH_3 or H_2O, nor peaks from complementary termini, and I operate on spectra which do not have isotopic peaks, having been removed during spectral preprocessing as described in Chapter 2. Future work may eventually add sufficient features that this filtering process is necessary. The posterior probability of each peak match is referred to as the reliability of that peak match in subsequent sections, denoted by ψ. The reliability may be "padded" to start from 0.5, $\psi_{0.5} = 0.5 + (1 - 0.5) \times \psi$ to not completely remove peaks which do not have many supporting features.

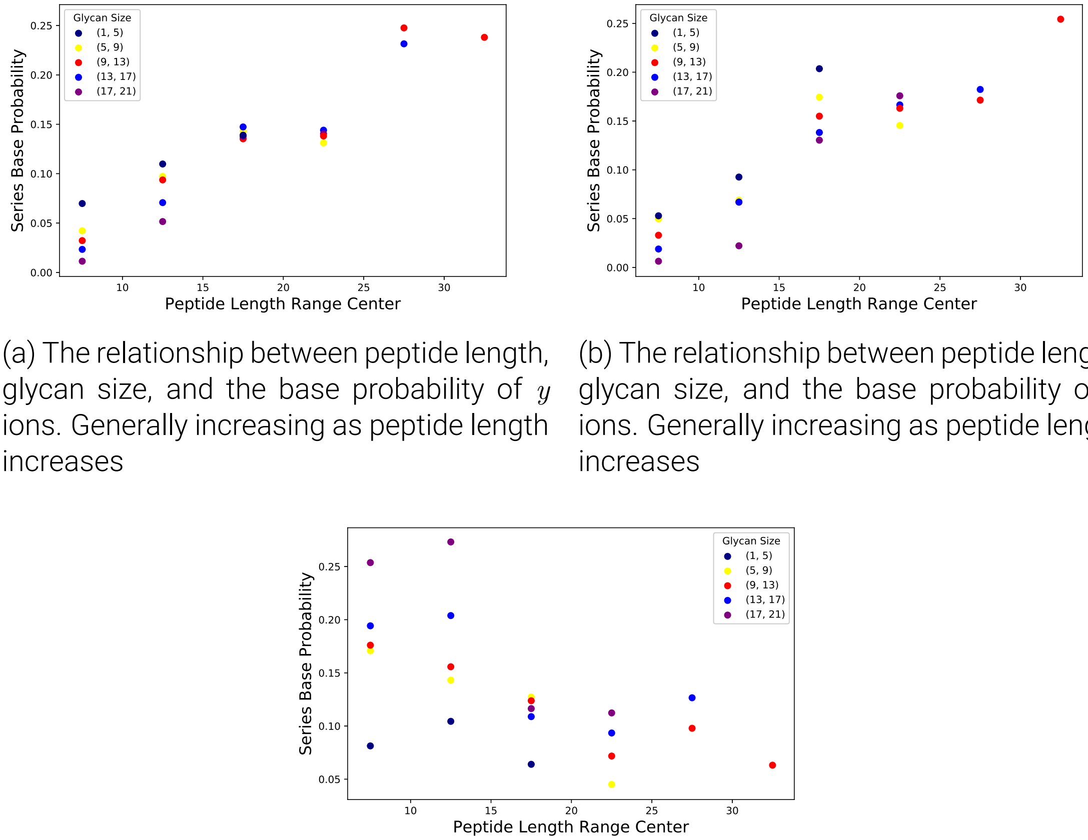


(a) The relationship between peptide length, glycan size, and the base probability of y ions. Generally increasing as peptide length increases

(b) The relationship between peptide length, glycan size, and the base probability of b ions. Generally increasing as peptide length increases, though with more fluctuation than y

(c) The relationship between peptide length, glycan size, and the base probability of $peptide{+}Y$ ions. There appears to be a weak interaction between peptide length and glycan size, but the effect is still dominated by glycan size

Figure 4.11: The estimated series probability for b, y, and $peptide{+}Y$ ions in the stepped collision energy HCD dataset from [132]

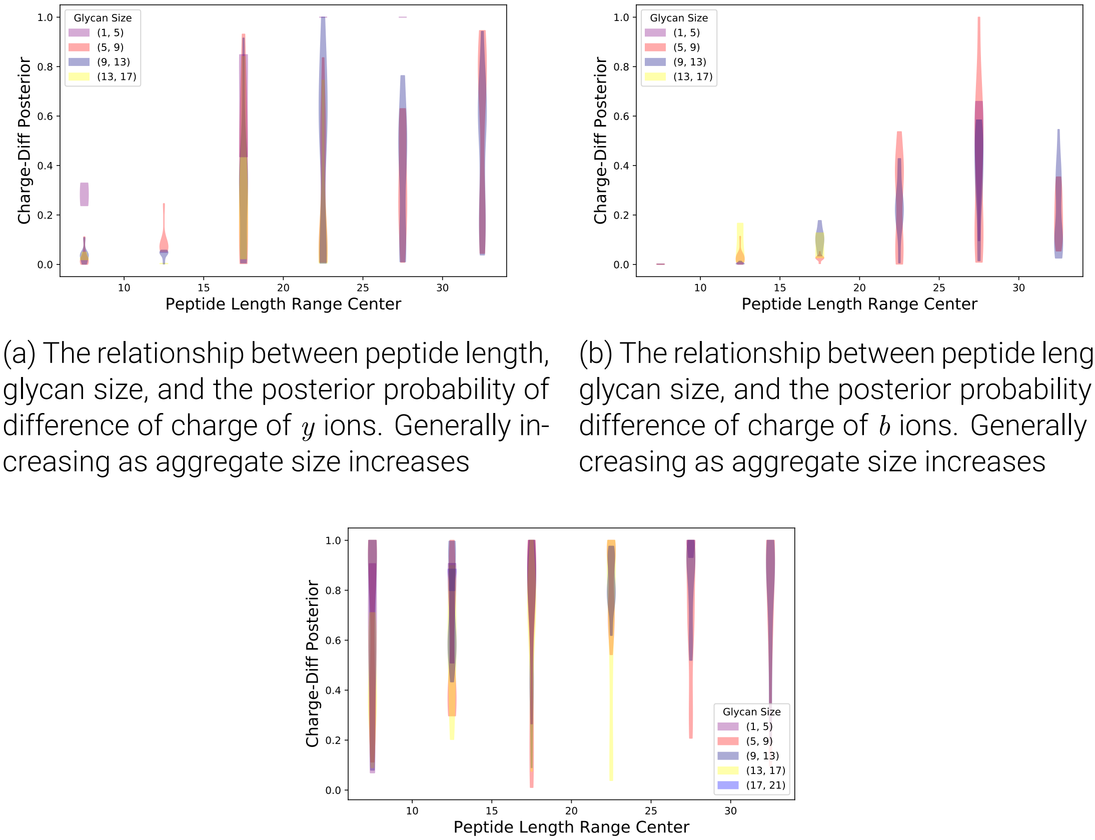


(a) The relationship between peptide length, glycan size, and the base probability of y ions. Generally increasing as peptide length increases

(b) The relationship between peptide length, glycan size, and the base probability of b ions. Generally increasing as peptide length increases

(c) The relationship between peptide length, glycan size, and the base probability of $peptide{+}Y$ ions. There appears to be a strong interaction between peptide length and glycan size as the single energy cannot work equally well on all combinations of peptide and glycan size

Figure 4.12: The estimated series probability for b, y, and $peptide{+}Y$ ions in single energy HCD data aggregated from an enriched, TMT10plex tagged *N*-glycopeptide dataset acquired from human CSF with 36 neV.

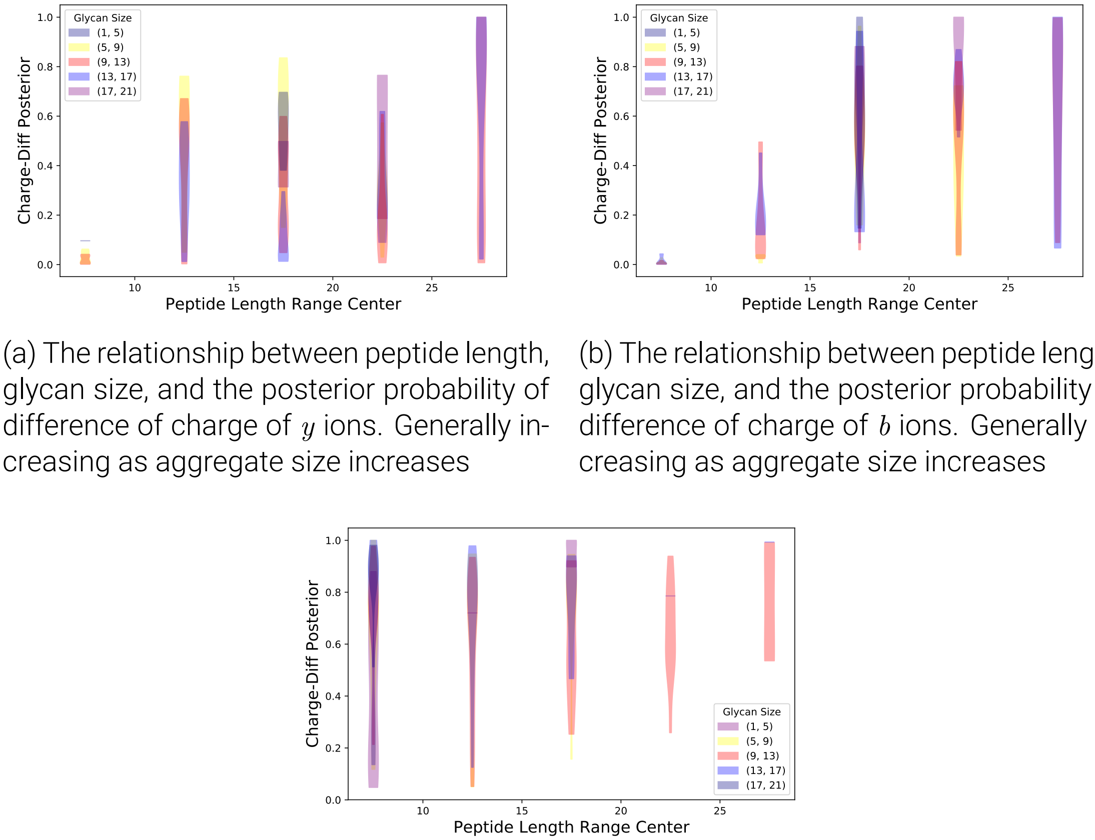


(a) The relationship between peptide length, glycan size, and the posterior probability of difference of charge of y ions. Generally increasing as aggregate size increases

(b) The relationship between peptide length, glycan size, and the posterior probability of difference of charge of b ions. Generally increasing as aggregate size increases

(c) The relationship between peptide length, glycan size, and the posterior probability of difference of charge of $peptide{+}Y$ ions. Appears to be important at all size ranges.

Figure 4.13: The posterior probability estimated for observing the same fragment under two different charge states for b, y, and $peptide{+}Y$ ions in the stepped collision energy HCD dataset from [132]

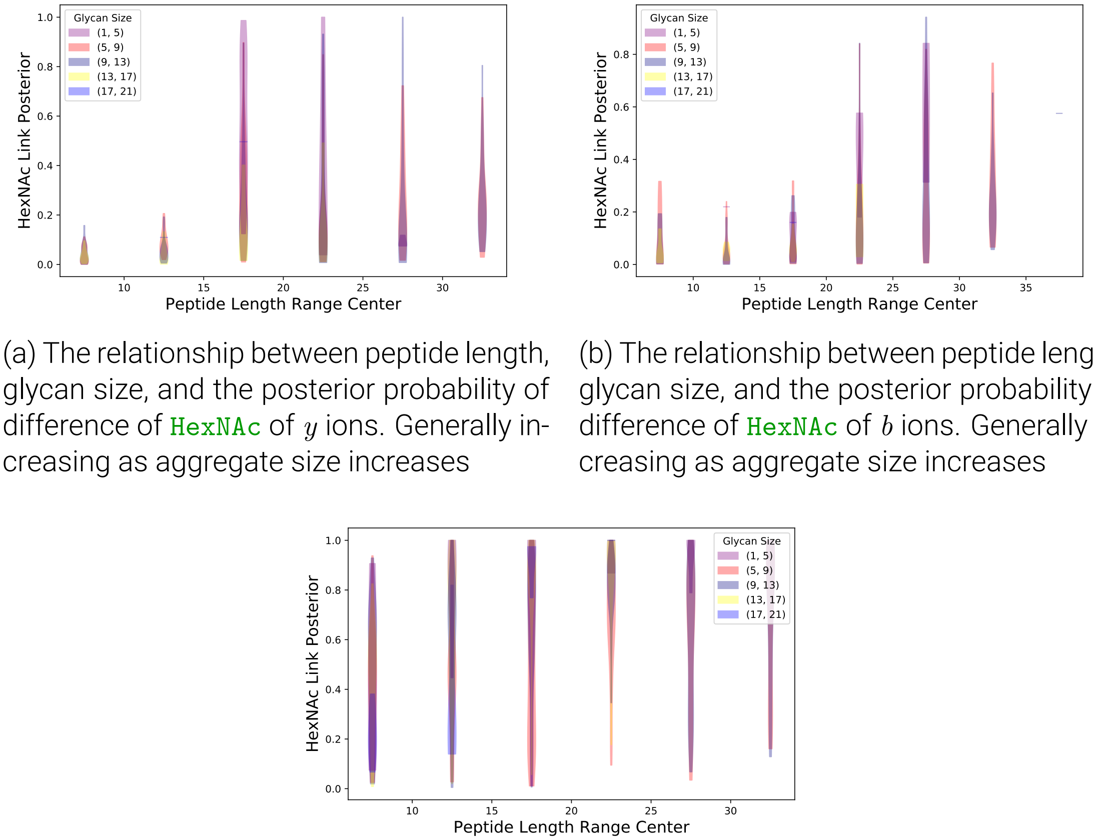


(a) The relationship between peptide length, glycan size, and the posterior probability of difference of charge of y ions. Generally increasing as aggregate size increases

(b) The relationship between peptide length, glycan size, and the posterior probability of difference of charge of b ions. Generally increasing as aggregate size increases

(c) The relationship between peptide length, glycan size, and the posterior probability of difference of charge of $peptide{+}Y$ ions. Appears to be important at all size ranges.

Figure 4.14: The posterior probability estimated for observing the same fragment under two different charge states for b, y, and $peptide{+}Y$ ions in the single collision energy human CSF

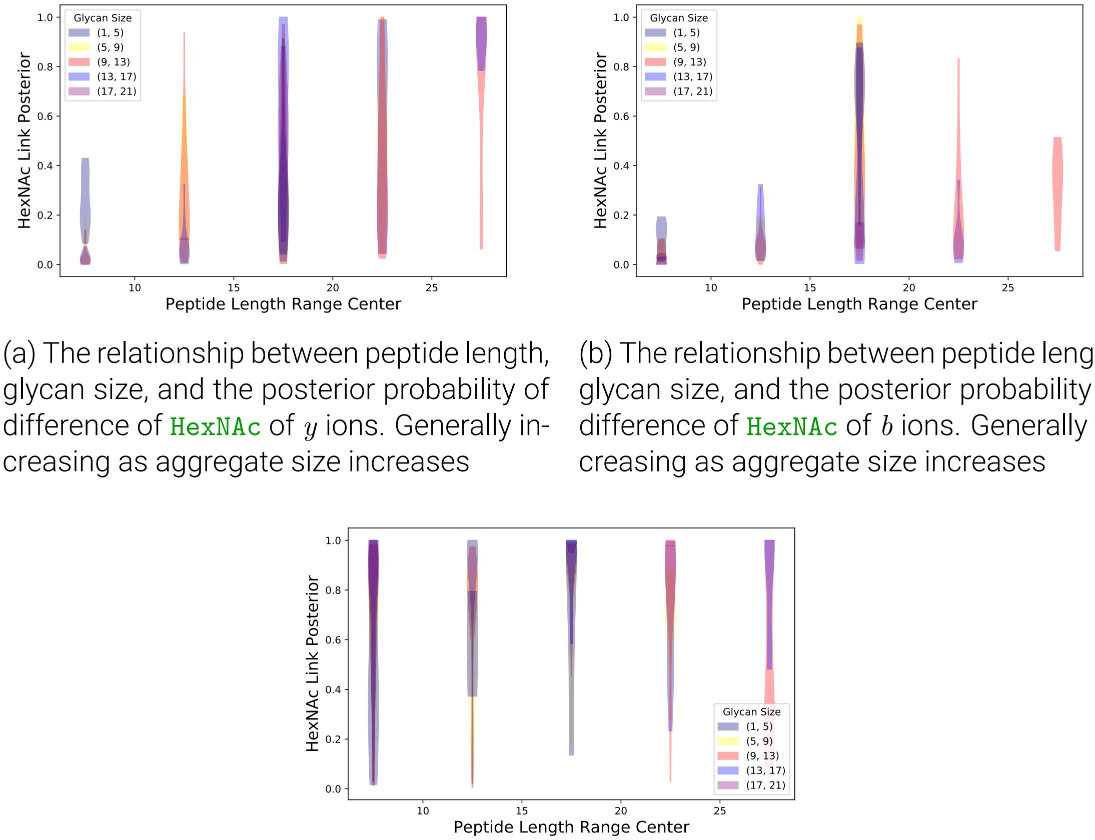


(a) The relationship between peptide length, glycan size, and the posterior probability of difference of HexNAc of y ions. Generally increasing as aggregate size increases

(b) The relationship between peptide length, glycan size, and the posterior probability of difference of HexNAc of b ions. Generally increasing as aggregate size increases

(c) The relationship between peptide length, glycan size, and the posterior probability of difference of HexNAc of $peptide+Y$ ions. Appears to be important at all size ranges.

Figure 4.15: The posterior probability estimated for observing the same fragment with and without a mass shift corresponding to a HexNAc for b, y, and $peptide+Y$ ions in the stepped collision energy HCD dataset from [132]



(a) The relationship between peptide length, glycan size, and the posterior probability of difference of HexNAc of y ions. Generally increasing as aggregate size increases

(b) The relationship between peptide length, glycan size, and the posterior probability of difference of HexNAc of b ions. Generally increasing as aggregate size increases

(c) The relationship between peptide length, glycan size, and the posterior probability of difference of HexNAc of $peptide{+}Y$ ions. Appears to be important at all size ranges.

Figure 4.16: The posterior probability estimated for observing the same fragment with and without a mass shift corresponding to a HexNAc for b, y, and $peptide{+}Y$ ions in the single collision energy human CSF

Intensity Prediction

In order to further understand the fragmentation process of glycopeptides, I attempted to model the intensity of a product ion as a function of local sequence characteristics. For example, the well known “Proline Rule”, that a Proline at the C-terminus of an amide bond will make a bond much more likely to break and that if it is at the N-terminus of the bond it will be unlikely to break, can be inferred from experimental observation [215, 280]. There have been many efforts using a wide range of different techniques [163, 164, 206, 242, 281, 282] to computationally model these phenomena, and the underlying physiochemical processes are still not fully understood [275, 283–285]. No solution is completely accurate, even those deep learning cases which were trained on hundreds of thousands of examples. Additionally, most if not all of these models are instrument and acquisition parameter dependent. The method of dissociation and the energy used for dissociation are implicit parameters of these models, making it difficult to apply the model on new instruments or under different acquisition conditions. Recognizing that there is not sufficient data publicly available for a single acquisition method to fit a high quality model, I intended for this to only provide guidelines for estimating match quality and should not be a dominating factor in the scoring process.

I used a multinomial GLM to model the fragmentation process. I considered a sequence of nested models, starting with a simple model of just ion series and adjacent amide bonds, up to a more complex model considering several neighboring bonds, charge state, fragment size, and peptide backbone composition. I used three-fold cross-validation to select the optimal model specification, using Pearson’s Correlation Coefficient (ρ) of prediction with experimental data as a measure of goodness of fit for its comparability across spectra. The selected model was the fourth most complex design. A peak's intensity is a function of 232 factors with 102 governing b and y ion fragmentation and 130 governing $peptide{+}Y$ ion fragmentation. Each spectrum match was decomposed into a factor matrix of size $k_i \times n_f$ where k_i is the number of peak matches + 1 and n_f is the number of features, with a observation level weight equal to the total signal of the spectrum, and a peak-level weight equal to the peak's reliability. In addition to the partitioning rules used from the peak relationship modeling, I also partitioned over precursor charge state, which interacts with proton mobility index. After cross-validation chose the optimal model, I refit each partition, and for partitions which contained matches with a correlation below 0.5 with the fitted model, I fit a second "mismatch" model on these poorly characterized spectra. Subsequently, I use a weighted mixture of the predictions of the full model and the "mismatch" model, weighted by a function (Eq. 4.30) of the inverse of the Pearson residual of the prediction (Eq. 4.28).

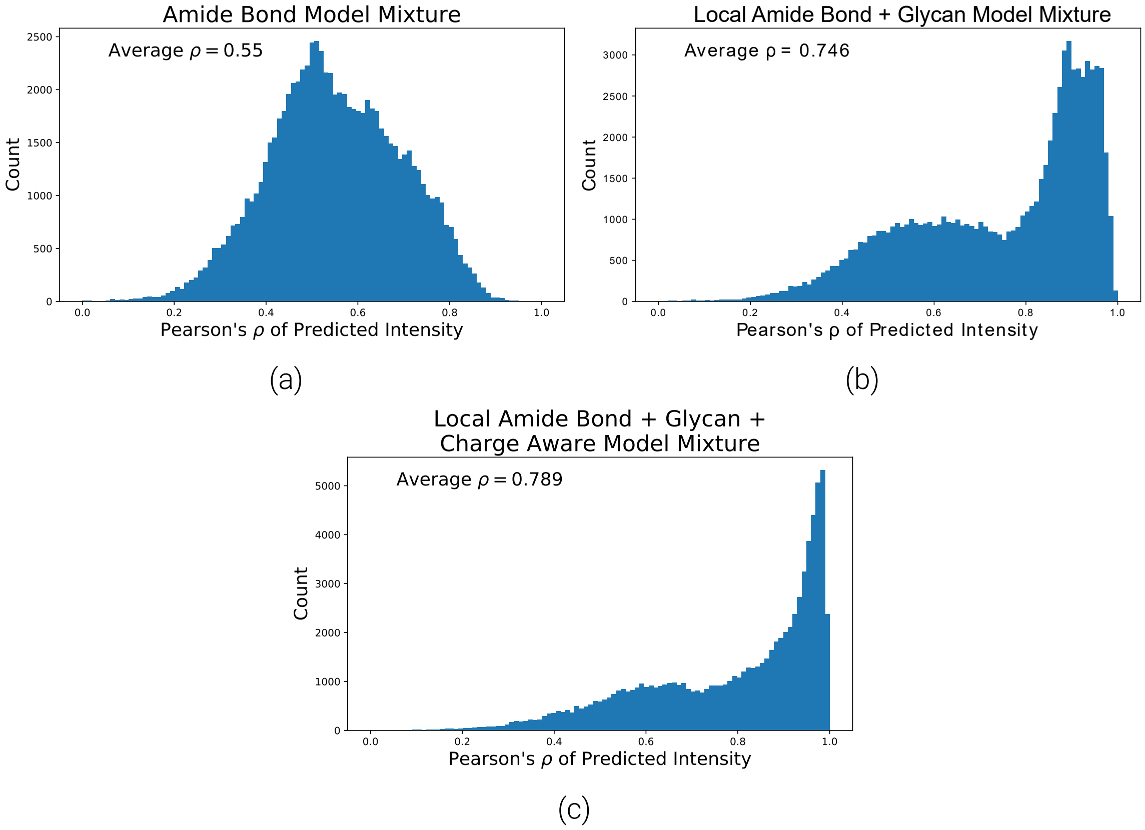
$$r_p = \frac{w(I - \hat{I})^2}{\hat{I}(1 - \hat{I})\psi_{0.5}} \tag{4.28}$$

$$w = \begin{cases} 0.5 & I \geq \hat{I} \\ 1 & \text{otherwise} \end{cases} \tag{4.29}$$

$$\pi_i = \frac{\frac{1}{(r_{p,i} \cdot \mathbf{I})^4}}{\sum_j \frac{1}{(r_{p,j} \cdot \mathbf{I})^4}} \tag{4.30}$$

Here, I is the experimental peak intensity vector, normalized to sum to 1, $\hat{I}$ is the predicted peak intensity vector of the regression model, given the feature matrix of matched peaks from query structure q against the spectrum s. The weight w is used to impose the belief that predicting a peak to be less abundant than it is should not be considered as bad as the reverse, reflecting that any model would be incomplete and that this should just be used as a guide. In the expression for π_i, $r_{p,i}$ corresponds to the Pearson residuals for the ith component of the mixture.

The resulting model fits chosen by cross-validation for the stepped collision energy mouse dataset are shown in Figure 4.17, and the fits chosen for the single energy human CSF data are shown in Figure 4.18. The naive amide bond model already produces a mean ρ of around 0.5, which is better than 0. As additional features are added to look beyond the immediate amide bond, capture portions of the peptide composition and glycan fragment size, the model approaches a mean ρ of 0.8. This is not close to the median ρ of 0.98 [164] reports for deep learning on bare peptides, but I have a mere fraction of the training data they used. The second model definition adds some information about the $peptide{+}Y$ ions depending upon the peptide backbone composition and whether or not the fragment contains Fuc, while the final model fits each charge state of $peptide{+}Y$ separately. The gain in performance from the second to the third model for the stepped collision energy data reflects the abundance of those types of fragments, while the change in the single energy data is minor as those fragments are much less common in that data.

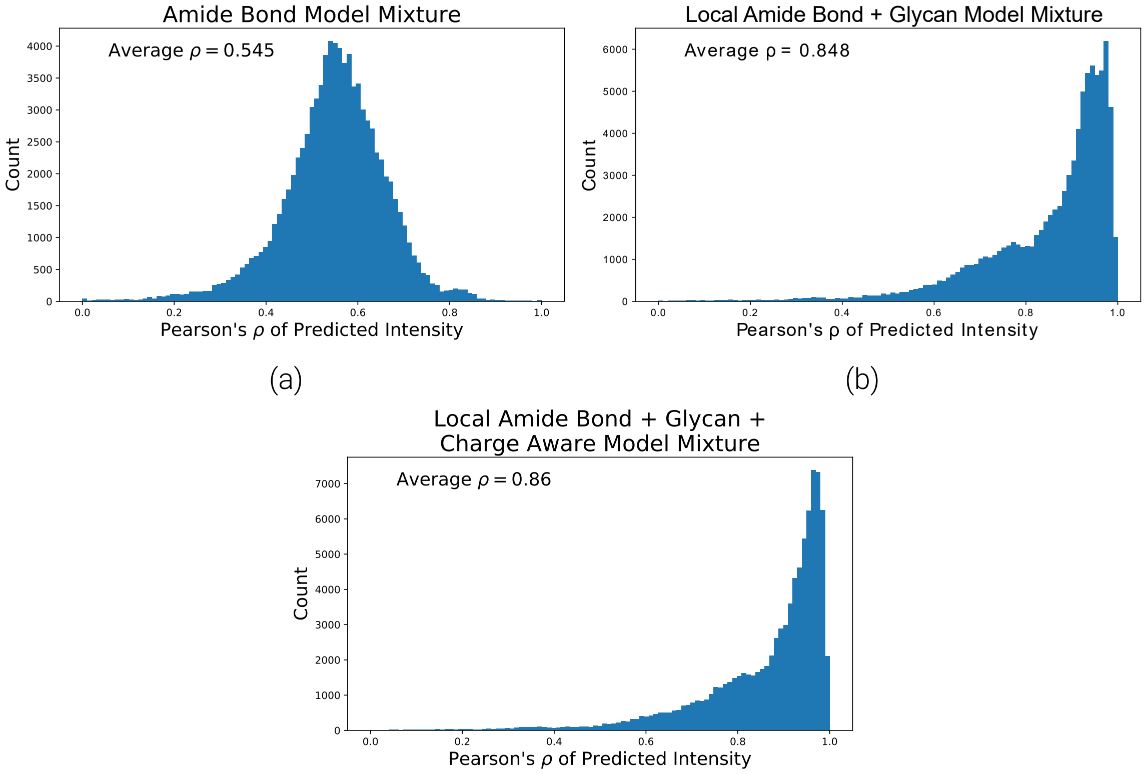


(a)

(b)

(c)

Figure 4.17: Peak intensity prediction correlation in the stepped collision energy HCD dataset



(a)

(b)

(c)

Figure 4.18: Peak intensity prediction correlation in the single collision energy human CSF

Integrating Predictions with Scoring

In order to to use these learned models to identify glycopeptides, I augmented the existing PepScore (Eq. 4.20) and GlyScore (Eq. 4.21) with model predictions. Because the partitions form a tree-like structure, I refer to this as a “multinomial regression mixture tree” (MRMT). The model to be used is selected by traversing the tree along the branch leading to the matching peptide length range, glycan size range, precursor charge range, and proton mobility index. If no matching model was found, the nearest model would be selected. The new scoring model incorporates both the reliability ψ and a transformation of the Pearson residual through its log CDF shown in Figure 4.19b given by Eq. 4.31.

$$d_i = -\frac{1}{6}\log_{10}\left[\text{CDF}\left(\frac{w_i(I_i - \hat{I}_i)^2}{\hat{I}_i(1-\hat{I}_i)\psi_{0.5,i}}\right) + 10^{-6}\right] \tag{4.31}$$

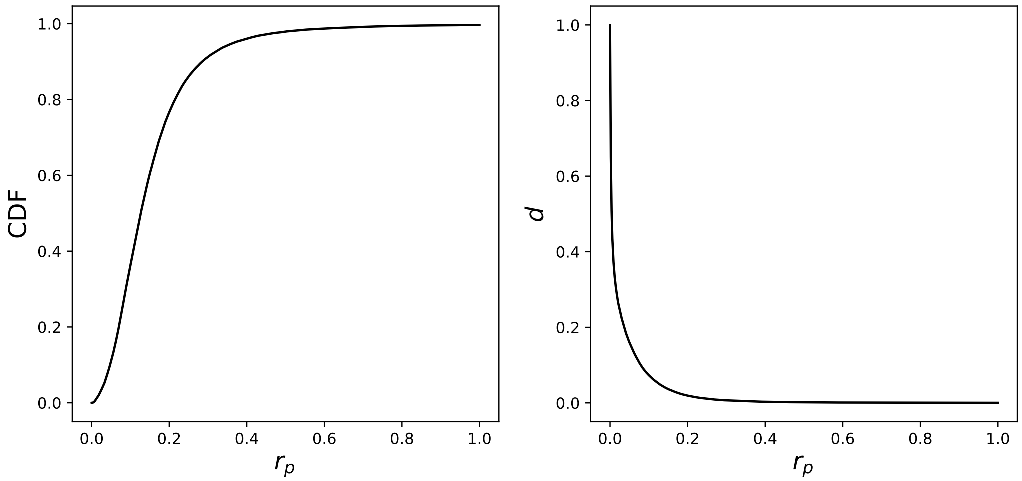
$$\text{pcc}(I, \hat{I}) = 2\frac{1+\rho(I,\hat{I})}{2} \times \log_{10}|I| \tag{4.32}$$

$$\begin{aligned} PepScore_{MRMT}(s,q,\gamma,\delta) = &\left[\sum_{i}^{n}\left(\log_{10} I_{i,p}\left(1 - \frac{|e_i|}{e_{tol}}^4\right)(\delta+\psi_i)d_i\right) + \text{pcc}(I_p, \hat{I}_p)\right] \\ &\times C_p(s,q)^\gamma \end{aligned} \tag{4.33}$$

$$\begin{aligned} GlyScore_{MRMT}(s,q,\alpha,\beta,\delta) = &\sum_{i}^{n}\left(\log_{10} I_{i,g}\left(1 - \frac{|e_i|}{e_{tol}}^4\right)(\delta+\psi_i)d_i\right) \\ &\times C_g(s,q)^\alpha \times \text{ratio}_{\text{core}}^{\beta} + SigIonScore(s,q) \end{aligned} \tag{4.34}$$

$$\begin{aligned} TotalScore_{MRMT}(s,q,\alpha,\beta,\gamma,\delta,w,\lambda) = {} & w \times GlyScore_{MRMT}(s,q,\alpha,\beta,\delta) + \\ & (1-w) \times PepScore_{MRMT}(s,q,\gamma,\delta) + \\ & SigIonScore(s,q) + \lambda C_{gp}(s,q) + \\ & MassAccScore(s,q) \end{aligned} \tag{4.35}$$

This differs from the original scores in that the log-scaled intensity is now weighted by the relative mass accuracy, an up-shifted but unpadded ψ, and a measure of the intensity goodness of fit through multinomial model's Pearson residuals d_i (Eq. 4.31). The pcc function works around the problem that $\rho(y, \hat{y})$ can easily be nearly 1 when there are few peak matches, causing it to favor poor matches and decoy matches over longer matches which are less consistent with the model, but not necessarily wrong. If more weight is placed on pcc, the number of targets retained increases, but this increases the bias towards longer peptides.



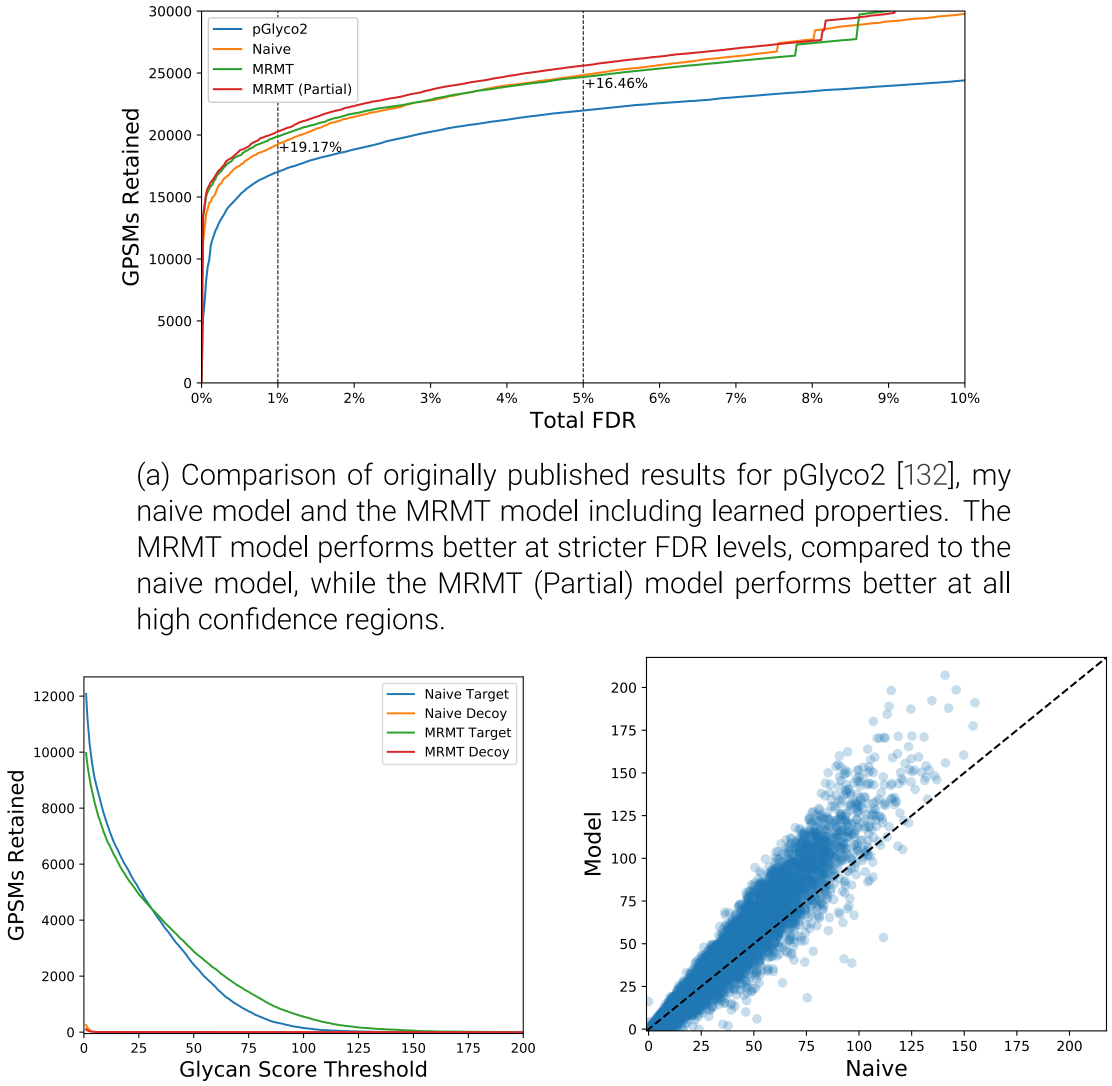
(a) An empirically estimated CDF of the Pearson residual

(b) The $\log_{10}$ transformation of the Pearson residual CDF shown in Figure 4.19a, normalized between 0 and 1.

Figure 4.19: The Pearson residual r_p and its empirical CDF transform d

4.3 Results

I compared the performance of this MRMT model, trained on the non-brain tissues from [132], with the naive model shown in Figure 4.20a. The parameters used were $\alpha = 0.5, \beta = 0.4, \gamma = 1, \delta = 0.75, w = 0.35, \lambda = 10$ with a 20 PPM mass error tolerance on product ions. The MRMT model performed better at stricter FDR thresholds, though it performs slightly worse than the naive model at the more permissive 5% threshold.



(a) Comparison of originally published results for pGlyco2 [132], my naive model and the MRMT model including learned properties. The MRMT model performs better at stricter FDR levels, compared to the naive model, while the MRMT (Partial) model performs better at all high confidence regions.

(b) A comparison of model $GlyScore_{MRMT}$ and naive $GlyScore$ for Glycan Targets vs. Glycan Decoys

(c) A comparison of model $GlyScore_{MRMT}$ and naive $GlyScore$

Figure 4.20: A comparison of the MRMT Model with the Naive Model

4.4 Discussion

The MRMT model performing worse at higher thresholds is most likely due to the prediction model being harsher on poorer glycan matches, as shown in Figure 4.20b. The prediction of $peptide{+}Y$ peak intensities is a function of the entire peptide backbone composition, the ion's charge, and the glycan fragment's size with respect to the total glycan structure. It does not explicitly specify the per-bond trends as done for the peptide where each amide bond participant is considered. Even so, we obtain a 3.3% improvement over the naive model at 1% FDR, and the ratio continues to improve as the threshold approaches 0. One can argue that at this level of specificity the FDR is based on too few observations to be meaningful. If I remove the d_i term from $GlyScore_{MRMT}$, the performance improves at all levels, creating the series shown in Figure 4.20a as MRMT (Partial). This partial $GlyScore_{MRMT}$ is still more stringent than the original, reflecting the gain of specificity from ψ.

This type of scoring model does have disadvantages. Since it takes into account sequence-specific properties, two different glycopeptides with the same length and glycan composition may get different scores while matching the same peaks, due to differences in peptide backbone composition or in some extreme cases proton mobility index. This can lead to a sub-optimal match being selected simply because that match is more consistent with the model's expectations, even if the experimental evidence is weaker. Most such weaker matches would be eliminated during FDR thresholding, but it is still possible for these spectra to be correctly identified. This is particularly true spectra where there are multiple glycopeptides that can match the abundant $peptide{+}Y$ with differing expectations, but the discriminating peptide backbone fragments are too low in abundance and inconsistent with model expectations to overcome the model's bias towards the other solution. Such trivial cases could be ruled out quickly by adding a post-match filtering step which discards matches with peptide scores below some arbitrary threshold like 1.0, where on average the 5% FDR threshold for the peptide score was around 4.

4.5 Conclusion

In this chapter, I described my work to model how to identify glycopeptides from MS^2 spectra and to increase the number of spectra identified per experiment. This work demonstrated a model building pipeline that could be used to learn to predict glycopeptide fragmentation from multiple types of glycopeptide fragmentation protocols. The fitted model was able to augment an existing glycopeptide identification procedure and reject more decoy matches while retaining 6% more target matches at the same 1% FDR threshold, consistent with manual inspection of the spectrum, for a total of 19% over the original study [132]. These methods are not without their limitations, as discussed, but they may be addressed by the addition of more examples or by augmenting the linear mixture approach used to compute the TotalScore. Such work might depend upon placing a weaker, non-uniform assumption of prior probability of low proton mobility [279], or by employing a non-linear mixing effect.

Deeper coverage of the sample is advantageous because it permits you to consider more potential glycopeptides in your downstream analysis. This is another opportunity to observe a new pattern, or to reinforce an existing one. Quantitative analysis of site-specific glycosylation across a single glycoprotein requires many more identification events per *site* than the standard Top3 peptide or iBAQ methods of protein quantification [286–288]. With a better model, we can learn more from the same data. This can in turn lead to a better model, as I will discuss in Chapter 5, but more importantly, it leads to a better understanding of glycobiology.

Despite the potential returns, little effort has been made to make glycopeptide search engines more intelligent beyond reducing contextually unlikely comparisons [133, 183] and tuning of glycan structural information utilization [132, 153, 181, 185]. None the less, the scoring model described here was only possible because of the original work done by [132, 186], whose model served as the foundation for my own. Part of this shortage of modeling of the finer details may be related to the innate complexity of the domain, which inherits all of the issues of mass spectrometry of large molecules along with less regular fragmentation than bare peptides, expanded search spaces, and lack of standardization of methods. Another component is that there is limited information available to study how glycans fragment, particularly in a high throughput context where precise experimental controls on cation adducts cannot be employed [91], exacerbated by the proton mobility interaction with the peptide [275]. Finally, most glycopeptide search engines built "from scratch" must create each piece of the engine [135] themselves, leading to substantial effort just to get to ground level.

As data acquisition methods evolve, new models must be developed. Even with the wealth of data published in [132], and other articles like [277], there is not yet a large corpus of data acquired with consistent methods. Future work will depend upon what kinds of activations grow to prominence, whether stepped HCD will be adopted more widely, or if EThcD will become available to more researchers. It remains to be seen whether these methods can even be applied to shotgun proteomics datasets where the glycopeptides represent a small fraction of the overall signal, and still obtain useful, reproducible results.

Chapter 5

Extending Glycopeptide Identification beyond the Spectrum Match

5.1 Introduction

In the previous chapter, I discussed methods for improving our ability to identify glycopeptide spectrum matches GPSMs. Spectrum-wise structural identification is a fundamental part of computational mass spectrometry, and it serves as the basis for many techniques [135] across all types of "-omics" studies done with MS. However, there is a great deal of information missing from the spectrum. Some information is only available through associations between multiple spectrum matches [133, 166, 217, 220]. Some information is only available by examining the larger picture of the experiment from the perspective of the LC or collective set of runs [289]. Some information is only expressed through the beliefs of the spectrum's interpreter [145, 175, 195]. In this chapter, I will discuss methods for recognizing and tracking external consistency.

5.2 Glycoproteome-Wide Site-Specific Glycome Network Smoothing

In Chapter 3, I introduced my recently published work on network smoothing at the glycome level, and in Chapter 4 I presented work done to identify glycopeptides in a complex sample with the scope of the complete *N*-glycoproteome. Here, I will discuss how these two techniques can be combined to increase depth of coverage. This method uses the model's bias towards previously identified glycans and its relatives to adjust identification confidence, information external to the spectrum match.

5.2.1 Methods

For this work, I reused the test data from Chapter 4, five LC-MS runs of Mouse brain tissue. I used the first pass identifications using CovBinom across all five samples to identify glycopeptides, and mapped those passing a 5% FDR threshold onto LC-MS features.

Model Specification

To begin, I extracted identified glycopeptides across several samples in a cohort assumed to share the same glycoproteome. I scored them using the same MS^1 feature model described in Chapter 3, save that the charge state model was replaced with a constant and no adducts were considered. Next, I aggregated glycopeptides around glycosylation sites across samples, producing a list of glycopeptide features for each site to fit ϕ and τ as discussed in Chapter 3.

Unlike in Chapter 3, I may have multiple observations for the same glycan, so I need to summarize them before I can apply Eq. 3.8. To do this, I construct a matrix $\mathbf{E}$ that maps each observation of the same glycan g_i to the same entry in ϕ. I next multiply both sides by left-inverse of $\mathbf{E}$, $\mathbf{E}^{-}$ which removes the transformation from ϕ_o and makes $\mathbf{E}^{-}S_o$, which in effect averages all observations for the glycan g_i in S_o to form a single summarized value, and applies a similar reduction on the variance matrix Σ such that the variance for g_i is $\frac{1}{k}$ where k is the number of times g_i was observed. The altered model specification is shown in Eq. 5.2. When each glycan is observed exactly once $\mathbf{E} = \mathbf{I}$. I also augment the glycan graph with "decoy" glycan compositions matching each original “target” glycan composition, which shared edges with other decoys but not with targets, but otherwise occupied the same neighborhoods, and thus controlled by the same $\mathbf{A}\tau$.

$$S_o|\phi_o,\tau \sim \mathcal{N}(\mathbf{E}\phi_o,\Sigma) \tag{5.1}$$

$$\mathbf{E}^{-}S_o|\phi_o,\tau \sim \mathcal{N}(\phi_o,\mathbf{E}^{-}\Sigma\mathbf{E}^{-\mathbf{t}}) \tag{5.2}$$

$$\tilde{S} = \mathbf{E}^{-}S_o \tag{5.3}$$

This adjustment to the model alters the optimization of ϕ_o, culminating in Eq. 5.5.

$$0 = \frac{\partial\ell}{\partial\phi_o}(S_o-\mathbf{E}\phi_o)^t(S_o-\mathbf{E}\phi_o)+\lambda\begin{bmatrix}\phi_o-\tau_o, & \phi_m-\tau_m\end{bmatrix}\mathbf{L}\begin{bmatrix}\phi_o-\tau_o\\ \phi_m-\tau_m\end{bmatrix} \tag{5.4}$$

$$= \frac{\partial\ell}{\partial\phi_o}(\mathbf{E}^{-}S_o-\phi_o)^t\mathbf{E}^t\mathbf{E}(\mathbf{E}^{-}S_o-\phi_o)+\lambda\begin{bmatrix}\phi_o-\tau_o, & \phi_m-\tau_m\end{bmatrix}\mathbf{L}\begin{bmatrix}\phi_o-\tau_o\\ \phi_m-\tau_m\end{bmatrix}$$

Let $\tilde{S}_o = \mathbf{E}^{-}S_o$ and $\mathbf{V}_o = \mathbf{E}^t\mathbf{E}$

$$= \frac{\partial\ell}{\partial\phi_o}(\tilde{S}_o-\phi_o)^t\mathbf{V}_o(\tilde{S}_o-\phi_o)+\lambda\begin{bmatrix}\phi_o-\tau_o, & \phi_m-\tau_m\end{bmatrix}\mathbf{L}\begin{bmatrix}\phi_o-\tau_o\\ \phi_m-\tau_m\end{bmatrix}$$

$$\hat{\phi}_o = \left[\mathbf{I}+\lambda\mathbf{V}_o^{-}(\mathbf{L_{oo}}-\mathbf{L_{om}}\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}})\right]^{-1}(\tilde{S}_o-\tau_o)+\tau_o \tag{5.5}$$

From the adjusted model specification, and $\mathbf{E}$, I fit λ and τ as discussed in Chapter 3. All subsequent fitting of ϕ was done using $\min(\lambda, 0.2)$, as in the partial regularization condition.

Using the estimates of τ and ϕ_o, I also estimate ϕ_m, spreading information from high confidence glycans with experimental evidence to unobserved variants. ϕ_m includes both unobserved target glycans and all decoy glycans.

Integration of Network Smoothing with Glycopeptide Identification

I repeated the glycopeptide identification process with a modified GlyScore as shown in Eqs. 5.6 and 5.7 where ϕ_g is the estimated smoothed ϕ for glycan g in the query structure q. This "prior" weight on the glycan identity is influenced by the glycan coverage, but serves as extra evidence for that glycan, as if more signal were present and explained by that glycan. This biases the scoring model towards glycopeptides whose glycan is similar to those previously identified at that site. Because GlyScore contributes a fraction w of its magnitude to the TotalScore, it impacts both the top match as well as the glycan-level FDR

$$GlyScore_{\text{Smoothed}}(s,q,\alpha,\beta) = \left[\phi_q + \sum_i^n \log_{10} I_{i,g} \left(1 - \frac{|e_i|}{e_{tol}}^4\right)\right] \times C_g(s,q)^\alpha \times \text{ratio}_{\text{core}}^{\beta} \tag{5.6}$$

$$\begin{aligned} GlyScore_{MRMT,\text{Smoothed}}(s,q,\alpha,\beta,\delta) = &\left[\phi_q + \sum_i^n \left(\log_{10} I_{i,g} \left(1 - \frac{|e_i|}{e_{tol}}^4\right) \times (\delta + \psi_i) \times d_i\right)\right] \\ &\times C_g(s,q)^\alpha \times \text{ratio}_{\text{core}}^{\beta} + SigIonScore(s,q) \end{aligned} \tag{5.7}$$

To integrate this method with the finite mixture model approach to estimating FDR, the TT and DT class groups used the target ϕ_q for that glycan composition/glycosite pair, while TD and DD received the decoy ϕ_q for that glycan composition/glycosite pair. For decoy peptides, the protein sequence is reversed prior to *in silico* digestion, however the glycosylation site is preserved and reflected, so the each glycosite of the decoy protein shares the same model with the analogous site on the target protein. When $\lambda > 0$, ϕ_m will be shifted towards $\mathbf{A}_T$, but will usually be less than the equivalent ϕ_o, while when $\lambda \rightarrow 0$, ϕ_m will be close to 0 while ϕ_o will remain unchanged. This means that a site which selected a large λ will contribute several high scoring decoys along with several observed and unobserved targets, while a site which selected a small λ will have only contribute a small number of high scoring observed targets. This may weaken the efficacy of TDA because it explicitly treats targets and decoys differently, but the dependence on C_g and the use of target glycans on DT decoy-peptide decoys prevent the target glycan bias from invalidating its usefulness.

5.2.2 Results

The performance of the smoothing adjusted scoring functions are shown in Figure 5.1, comparing the performance of the models discussed in Chapter 4 without network smoothing. As expected, the addition this external information increased the number of identifications in all cases, but the gain in identifications was larger for the MRMT-based scoring models than the naive model. As observed in Chapter 4, this was likely because the MRMT models were better at discriminating target peptides from decoy peptides, while being more stringent towards the glycan, but with the external evidence to help discriminate the glycan, they were able to more fully utilize their peptide backbones. A histogram of the contribution made by the prior for each glycopeptide shown in the Smoothed MRMT (Partial) series is shown in Figure 5.2, showing that it still functions when the prior is small to non-existent.

FDR	pGlyco2	Naive	MRMT	MRMT (Partial)	Smoothed Naive	Smoothed MRMT	Smoothed MRMT (Partial)
10.0%	24410	29775	31490	31362	31313	33650	31581
5.0%	21992	24855	24673	25611	25513	28238	26756
3.0%	20263	22799	22853	23650	23385	24561	24702
1.0%	17015	19254	19887	20276	19814	21171	21219
0.5%	15165	17531	18364	18758	17827	19392	19596
0.1%	9901	14531	15662	15930	14710	16441	16501

Table 5.1: Number of spectra retained at each FDR threshold, by model

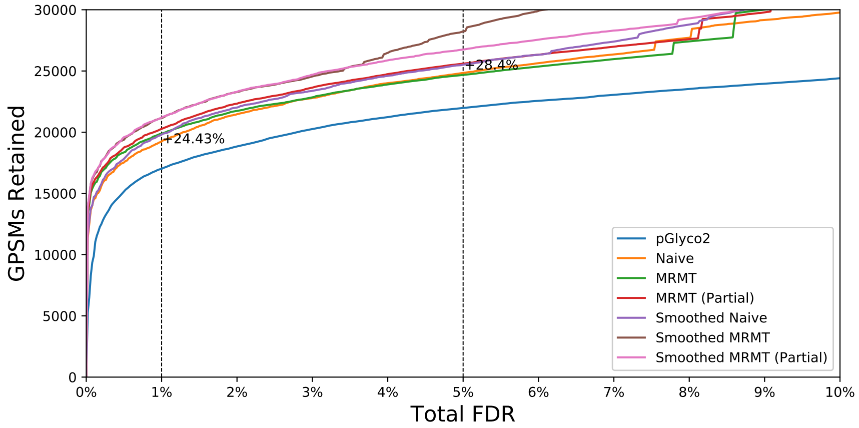


Figure 5.1: Comparison of originally published results for pGlyco2 [132], my naive model, the MRMT model including learned properties, and their network smoothed variants. The MRMT model and MRMT (Partial) model both performed substantially better than their un-smoothed counterparts as well as the smoothed version of the naive model. A few checkpoint counts are shown in Table 5.1

5.2.3 Discussion

The increased depth of identifications produced by the smoothing procedure are dominated by cases which are associated with a strong prior, though it does not automatically drown out those cases which are not associated with a strong prior, as shown in Figure 5.2. The argument can be made that because the feature score used for learning ϕ doesn't use MS^2 information beyond the assigned identity, it does not reflect the identification confidence. It would be possible to add another component to the feature score based on a transformation of the FDR of the identification, but this may make the surface the network smoothing parameter optimization procedure traverse less smooth. It would also make the method more sensitive to the identification procedure's bias. It is already driven by them implicitly, but it doesn't explicitly reinforce them. Additionally, the network neighborhoods used here were identical to the neighborhoods used in Chapter 3, which do not include some of the broader mammalian patterns, specifically Galα-Gal and LacDiNAc, which are known

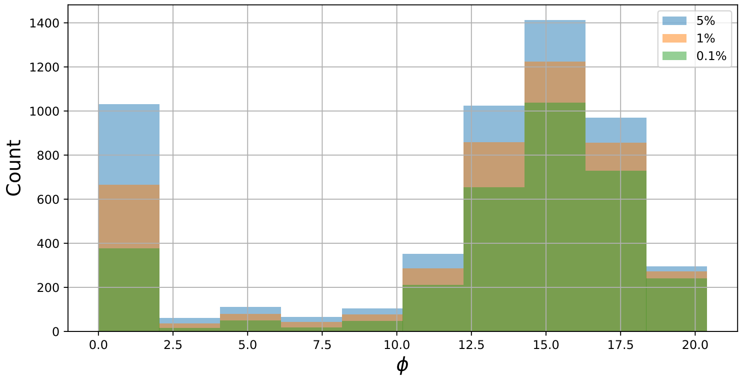


Figure 5.2: A histogram of the contribution made by the glycan prior for each glycopeptide from the Smoothed MRMT (Partial) series in Figure 5.1. The majority of identifications have a large prior component, but many do not.

to occur in mouse tissue [9, 59, 199]. This suggests that were the neighborhood definitions adjusted, those glycans would be better represented.

As noted in Chapter 3, while I included an automated procedure for learning the model parameters for each site, an expert can also explicitly encode their expectations in $\mathbf{A}_{\mathcal{T}}$, or to apply models from different contexts to new samples. It might be desirable to have a standard model for each tissue type which could then be used for consistency between experiments on the same tissue. Such an approach would be less specific than a spectral library spanning all experimental glycoforms, but would also be able to cover structures not found in the library.

This method would be harder to apply to traditional TDA, because it would mean selecting a different type of decoy mapping. With single energy-like fragmentation, if DT decoys are the only decoy type, should DT decoys be treated the same way as shown here? The answer is not clear, semantically. On the one hand, if reverse protein decoys are used, then using the decoy glycan value will be slightly more biased towards targets, but the difference may not be large enough to matter. On the other hand, if reverse peptide decoys are used, the model is already biased against the targets, and using the target glycan for the decoy peptides would only strengthen that bias. It would also depend upon whether the glycan masses for decoys were permuted at all, and whether the scoring function rewarded parsimony [133, 160, 266] over an additive scoring component [132, 153, 182].

Given work discussed in Chapter 4 regarding peptide identification as a precursor to building a theoretical glycopeptide database, it seems natural to attempt to extend this notion of network smoothing to peptides as well. This idea is not far from multi-round search [137, 156] or common protein boosting techniques [166, 175]. It is doable, though the graph structure would be more challenging to construct because the natural distance function between two peptides, edit distance, would not necessarily reflect the desired relationship. For example, a single missed cleavage represents one semantic difference, but introduces a minimum of one edit, with an unbounded maximum number of edits, though on average it might be between three and ten depending upon the protease and the protein domain. Which measure of difference more accurately expresses the degree of relatedness between those two peptides would be a matter of opinion and sample preparation [111, 136, 172, 290]. Another challenge would be the ranking of small molecule PTMs which cannot all be treated in the same fashion. The classification and ranking of PTMs by confidence is explored in [100, 137, 145, 174, 175, 195].

5.3 Retention Time Prediction

Separation characteristics such as retention time are an orthogonal metric to mass measurement for identifying molecules. Models of the behavior of different chromatographic systems for separating peptides have been published [104, 106, 291–293]. One of the challenges with these techniques is that they must be used with known experimental configurations, with specific chemical properties. These are often well studied, standardized methods. The concept has application to glycans [101, 203] and glycopeptides [294, 295], but there is a lack of reproducible computational method that covers the space of common glycan building blocks dealt with in my body of work.

5.3.1 Linear Model for Predicting Retention Time Within Glycoform Group For Internal Consistency

In [23], several LC-MS datasets were acquired on a C18 reverse-phase LC column with a slightly polar characteristic. This induced a monosaccharide-specific retention time shift, aggregated over the entire glycan composition. In order to defend the claim that I had identified sulfated glycans that were discriminable from their ambiguous unsulfated approximate isobars discussed in Section 4.2.2, it was advantageous to complement fragment ions with retention time. I extracted the aggregated ion chromatograms for each glycoform at a specific site and fit a linear model weighted by abundance specified in Eq. 5.8.

$$\begin{aligned}\text{Apex RT} \sim \beta_0 C(\text{Peptide}) + \beta_1 \texttt{Hex} + \beta_2 \texttt{HexNAc} + \beta_3 \texttt{Fuc} \\ + \beta_4 \texttt{NeuAc} + \beta_5 \texttt{SO}_3 + \epsilon\end{aligned} \tag{5.8}$$

Results

When fit over all observed high mannose, asialo-complex type, and sulfated asialo-complex type glycoforms of NCTLIDALLGDPHCDGFQNEK, I obtained a single fit with $R^2 = 0.981$ with coefficients $\beta = [42.816, -0.1, -0.097, -0.159, 0, 1.29]$ A plot comparing intact mass by retention time is shown in Figure 5.3. Using the same data points with `{@sulfate:1, HexNAc:2}` replaced with `{Hex:3}` and refit the model. The produced a fit $R^2 = 0.409$, which suggests that these mass shifts, despite their similarity, are produced by different molecules.

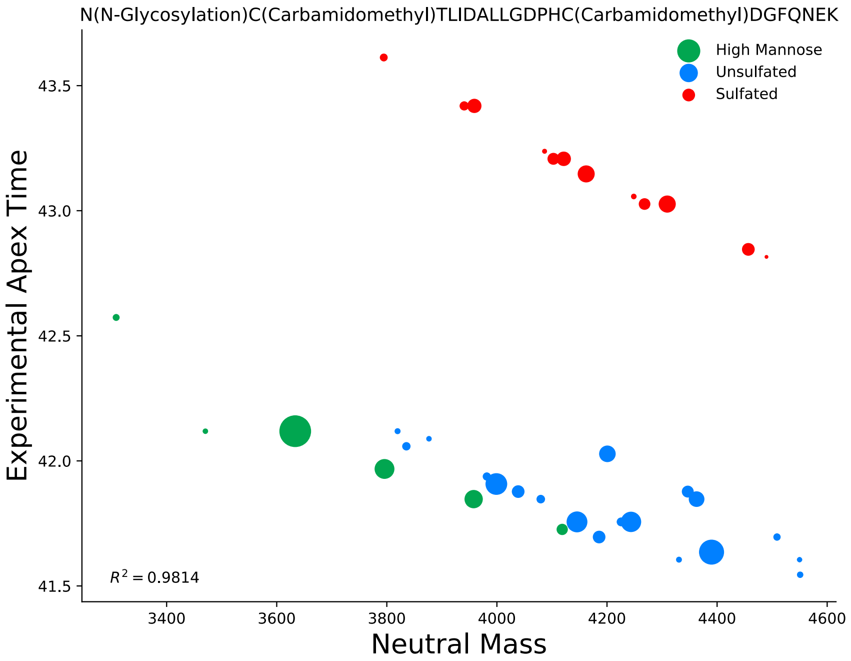


Figure 5.3: A linear model fit for the high mannose, asialo-complex type, and sulfated asialo-complex type glycoforms of `NCTLIDALLGDPHCDGFQNEK` with high accuracy, $R^2 = 0.9814$.

5.3.2 Predicting Relative Retention Times For Different Peptide Glycoforms For Classification

Using the estimated parameters from `NCTLIDALLGDPHCDGFQNEK` to predict the retention time of glycoforms of another peptide would not immediately make sense, because β_0 reflects something about `NCTLIDALLGDPHCDGFQNEK`, which is the only part not shared with another set of similar glycopeptides sharing glycan compositions but not peptide sequences. It would be possible to replace β_0 with a vector of coefficients to absorb each peptide-specific effect while averaging the contribution of the common monosaccharides, but this does not give a sense of how well the known trend for those common monosaccharides learned on NCTLIDALLGDPHCDGFQNEK approximate the trend for that new set of glycopeptides.

Assuming there are multiple observations in that new set, and the monosaccharide trends are assumed to be fixed and should not be re-estimated for the new glycopeptides, the monosaccharide level trends can be used by measuring whether the distance predicted by the fitted model from glycoforms from peptide p_1 for two glycopeptides gp_1 and gp_2 from peptide p_2. This can be measured by predicting the retention time for gp_1 and subtracting it from the predicted time for gp_2 as a proxy for the residual, in the least squares sense. With many glycopeptides from p_2, averaging over all other glycoform the relative difference residual would approach the true error of fit, assuming no peptide-specific interaction with the monosaccharides.

This residual could be converted into a goodness-of-fit value between 0 and 0.5 by passing it through the t distribution survival function with mean 0, degrees of freedom $|gp \in p_2| - |\beta|$ and some standard deviation σ. σ cannot be learned directly from the raw apex retention times because there is by definition only one observation for each glycoform. Instead, it might be assumed to be 1.0 to enforce a relatively lax prediction interval, while it might be estimated from the average relative difference in retention time prediction across all test cases to narrow the window. From the survival function value, multiplying by 2 produces a value between 0 and 1, where 1 is a perfect alignment between experiment and prediction and 0 reflects the worst possible alignment, with the magnitude of the difference depending upon σ. This goodness of fit value is then compatible with a logit-transform for use in an MS^1 feature score as in Chapter 3, or untransformed it can be used to threshold a set of chromatograms for some downstream operation assuming shared identity.

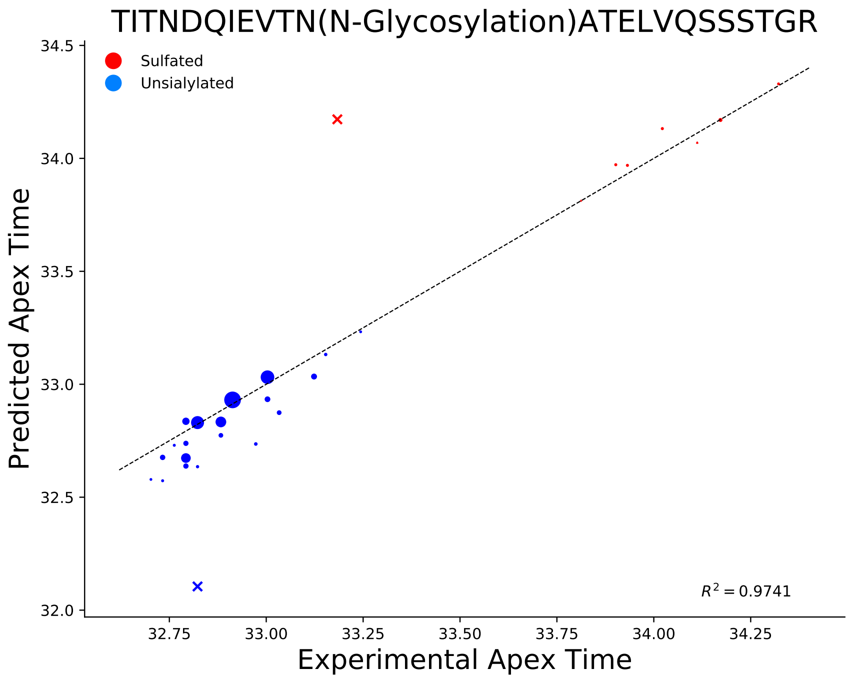


Figure 5.4: The fitted regression model from `NCTLIDALLGDPHCDGFQNEK` predicting retention times for `TITNDQIEVTNATELVQSSSTGR`, with sulfated and unsulfated glycans. Outliers are marked with `X`s, and the R^2 went from 0.76 to 0.97 from the removal of outliers.

Results

Using the fit on `NCTLIDALLGDPHCDGFQNEK` to predict outliers in `TITNDQIEVTNATELVQSSSTGR` with $\sigma = 1$ and a threshold of 0.5 to eliminate outliers, dropped to extreme chromatograms produced the fit shown in Figure 5.4.

5.3.3 Predicting Unfragmented Precursor Identities

Using similarly acquired AGP data with abundant ammonium and metallic cation adducts, I extracted all identified chromatograms for `SVQEIQATFFYFTPNK` and fit a model using only those cases which were identified in their unadducted state, producing a fit with $R^2 = 0.974$ and $\beta = [60.07, 3.39, -3.86, -0.25, 2.9, 0]$. Here, the coefficients for each monosaccharide are considerably larger, reflecting that the solution for the intercept is anchored at a different point than the IAV examples. These glycopeptides are also sialylated but not sulfated. It is unlikely that the monosaccharide coefficients here are accurate on an absolute scale, but they can still be used relatively. Next, I extracted all unassigned chromatograms from the sample, and matched intact masses for other SVQEIQATFFYFTPNK glycoforms. These new cases have no MS^2 spectra to identify them. Using the relative retention time predicted for these matches can be used to rule them in or out. Using the original fit on unadducted glycoforms, I next predicted outliers on the set containing the unadducted, adducted only, and unfragmented chromatograms, and marked those with a score <= 0.1 as outliers. The inlier fit produced an $R^2 = 0.961$, shown in Figure 5.5. In addition to recognizing unfragmented precursors and mis-assigned adducts, the model fit was able to flag deconvolution artefacts where the A+1 peak was chosen for the monoisotopic peak and the structure matched converted NeuAc into Fuc 2. These are often flagged separately by the chromatographic peak shape feature, but this may help to improve detection.

5.3.4 Discussion

Using models like these would be useful for larger scale studies where we cannot reliably fragment every precursor of interest, but would still like to quantify them. This is related to the missing value problem in general, which is endemic to mass spectrometry, especially proteomics [296–299]. Previous solutions to this problem have been to use “Accurate Mass and Time Tagging” or “Mass Tags” [300], along with the similar “Match Between Runs” [266, 301] idea, use previously time-aligned/normalized identifications with high precursor mass accuracy to propagate identities between LC-MS

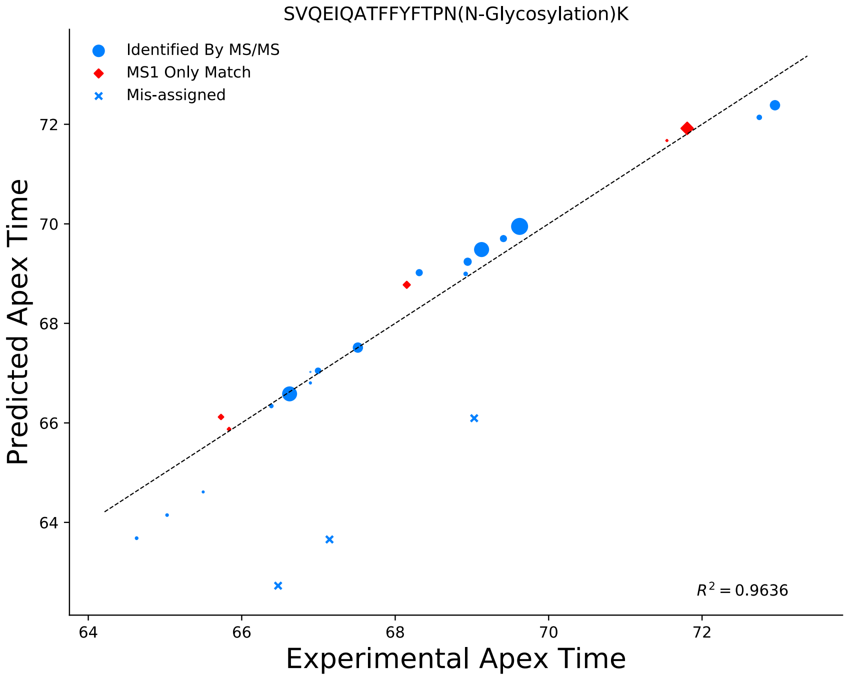


Figure 5.5: Predicting outliers and inliers from regression on `SVQEIQATFFYFTPNK`, including adducted, removal of mis-assigned glycan composition chromatograms and the inclusion unfragmented chromatograms corresponding to new glycoforms of the same peptide sequence.

runs to assign chromatograms which may not be identified in a particular run, but which are identifiable independently from other sources, but these techniques all rely on having previously identified the structure by MS^2. Other methods like [201, 302] use retention time as one dimension of the identification process, either on equal footing with, or as a replacement for fragmentation evidence.

Limitations

While the method I proposed here works well when there are enough cases covering all monosaccharides for a single peptide backbone, this is not always the case, especially for most sparsely covered glycosites common to large, complex samples. Without at least one high confidence point of reference, it is not able to make useful predictions, making it difficult to use as a general purpose predictor of retention time. Further, it assumes constant chromatographic conditions throughout the run, which is not consistent with reality. Therefore, its best utility is as an internal consistency check to complement other methods, like the examples shown in Figures 5.3 and 5.5. This might complement the fitting of network smoothing models to down-rank glycan compositions which are not consistent with other glycoforms of the same peptide.

5.4 Conclusion

Using information shared among different glycan compositions at the same site, I demonstrated how the concept of network smoothing, originally presented for glycomics experiment in [94], shown in Chapter 3, could be extended to glycoproteomics. This technique, when used in concert with a more sensitive scoring model produced 5.25% increase in identifications at 1% FDR over the model alone, which already improved on the baseline naive model by 6%, with a total of 24% improvement on the original analysis [132].

I also discussed a method for using a local estimate of glycan component specific retention time effects. This method is in agreement with the literature [294, 295] and work under way by [104]. While the method discussed is not nearly as efficacious as network smoothing it does provide additional, orthogonal information which could be used to augment an existing approach or for validating the internal consistency of a glycoform group, or to disambiguate similar monosaccharide aggregates. This could be useful for assessing the FDR from another angle [172], or as a higher order pattern to look for when aligning LC-MS runs of glycopeptides [103].

These methods incorporate information into the identification process from out-side the spectrum match, which can be used independent of the fragmentation method used. They depend upon additional chemical and biological features which can be adapted to other types of glycans than *N*-glycans. While the techniques discussed here operate on abstracted and simplified concepts, they do utilize semantic relationships which underlies the glycan structure [9, 85] and how they express themselves through the measurements we use to study them [259].

Chapter 6

Conclusion

6.1 Summary of Work

In this work, I described the development of a layered set of tools for the interpretation of glycopeptide tandem mass spectra, with the intent to show how integrating information from each layer of the problem domain enabled deeper understanding of the data. I introduced how the problem must be solved with appropriate signal processing from the very beginning in Chapter 2. I demonstrated how that signal processing helped to develop a platform for glycan composition profiling, which allowed me to create a test-bed for applying network smoothing over glycan compositions in Chapter 3. This network smoothing procedure allows me to impose external information on the glycan composition identification problem, learning the importance of that pattern from the data. Separately, the signal processing step enabled deeper analysis of glycopeptides from complex samples, exposing more information to study in Chapter 4. By exploiting that information with a set of complementary modeling techniques, I was able to create a scoring model that achieved a 6% improvement over the naive model at 1% FDR. Finally, I showed that the same network smoothing model applied to glycan compositions could be applied to glycopeptides on a per-glycosite basis, gaining another 5.25% improvement at the same 1% FDR, for a total of 24% improvement compared to the original interpretation of these samples

[132].

6.2 Why Is This Important?

The creation of a model building pipeline for glycopeptide fragmentation for learning recurring features is useful because glycopeptide identification isn't solved yet. The scoring models described here will always benefit from more data [8], and they are only for one type of dissociation [88]. There are other methods that might work better. That this method is completely transparent could be useful for another implementer.

Similarly, the signal processing tools developed during this work would be valuable for people attempting to solve similar problems. The deconvolution procedure is a necessary evil in order to handle larger molecules [132, 219, 226], but it only serves the main goal of identifying glycopeptides indirectly. Additionally, `ms_deisotope` is a general-purpose deconvolution toolkit, with application to other biomolecule classes, including peptides and chemically derivatized glycans [54, 94]. Components of it have already been reused by others [236, 238]. Because it can read and write mzML [244], it can operate between other tools which were previously reserved spaces for compiled languages and extensions [115, 120].

More significantly however, identifying glycosylation site-specific patterns is difficult because it can take several spectra to identify just one glycopeptide, with the knowledge that there may be between two and forty glycopeptides for each glycosite, just based upon the identification trends observed in toy examples like AGP [23, 92, 303]. Adding the ability to extract more information from the same sample means that you can quantify more from the same signal is important. It means that more glycoproteins can be characterized from the same sample, which means that there is a greater potential to observe meaningful biological patterns. Alternatively, it acts as a lever by which an expert can exert their beliefs on the analysis.

6.3 Implications

Should the network smoothing model be deemed acceptable by other members of the field, it would be advantageous for practitioners interested in using it to acquire and share glycosite and glycoprotein models. This would be consistent with the efforts underway with the active glycoinformatics databases [213, 214, 269]. There is already ample data to draw on for several tissues available through PRIDE [277, 304]. Because the network smoothing method only depends upon the presence of MS^1 features that have been identified as glycopeptides, this component can be used with or without the MS^2 identification methods I discussed in Chapter 4.

If the stepped HCD method [132, 187, 276] is considered an acceptable trade-off in terms of analysis time and depth, then depending upon the energies that become standardized, the fragmentation models published here may see more general use for large cohort studies. Additional work might be necessary to adapt to different energy ranges or calibrate for different instrument biases.

Should anyone believe that ammonium adducts on glycopeptides are real, are a problem, and should be solved, they may choose to further investigate mass accuracy and retention time tracking methods like those discussed in Chapters 4 and 5.

6.4 Next Steps

My intent in creating these algorithms was to make it easier for people to identify glycans and glycopeptides. I have provided a variety of tools for that purpose, and the model building techniques shown here will be added to them. GlycReSoft (http://www.bumc.bu.edu/msr/glycresoft/) is an open source Python application which encapsulates each of the steps described in this work, providing both a command line and graphical user interface. The more recent work discussed here will be added to the user interface.

One direction to go from the models discussed here is to extend network smoothing to the peptide component of the glycopeptide as well, discussed briefly in Chapter 5. This would provide an interesting challenge, as shown in Chapter 4, that the peptide backbone was much more structured and difficult to assign, as compared to the glycan. It would also be more readily applicable to general protein identification, where the notion of a prior probability of a peptidoform or proteoform has existed for some time [145, 175, 195], but has not been explored thoroughly. Alternatively, the existing methods could also be applied to types of *O*-glycopeptides, which might be benefit more from a combination of these approaches.

Another direction is to refine the alignment, validation, and quantification of glycopeptides. Because of their complicated nature, glycopeptides are not covered by existing quantification software very well. With the signal extraction tools discussed in Chapter 2, it might be possible to continue to integrate domain information into this problem, building on methods like those proposed by [305] and [95].

While the models discussed here were all HCD-related, there would be value in adapting parts to work with ExD or EThcD, specifically for multiply glycosylated peptides common in middle-down [88]. This would also be an interesting extension to network smoothing where multiple sites are combined to support one glycoform. Similarly, being able to incorporate ion mobility into the glycopeptide identification process would introduce additional information into the problem and could be used to address some of the ambiguities discussed in Chapters 4 and 5.

Chapter 7

Bibliography

Bibliography

[1] Ajit Varki, Richard D. Cummings, Markus Aebi, Nicole H. Packer, Peter H. Seeberger, Jeffrey D. Esko, Pamela Stanley, Gerald Hart, Alan Darvill, Taroh Kinoshita, James J. Prestegard, Ronald L. Schnaar, Hudson H. Freeze, Jamey D. Marth, Carolyn R. Bertozzi, Marilynn E. Etzler, Martin Frank, Johannes F.G. Vliegenthart, Thomas Lütteke, Serge Perez, Evan Bolton, Pauline Rudd, James Paulson, Minoru Kanehisa, Philip Toukach, Kiyoko F. Aoki-Kinoshita, Anne Dell, Hisashi Narimatsu, William York, Naoyuki Taniguchi, and Stuart Kornfeld. Symbol nomenclature for graphical representations of glycans. *Glycobiology*, 25(12):1323–1324, 2015.

[2] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biological Mass Spectrometry*, 11(11):601–601, nov 1984.

[3] Bruno Domon and Catherine E. Costello. A Systematic Nomenclature for Carbohydrate Fragmentation in FAB-MS/MS Spectra of Glycoconjugates. *Glycoconjugate Journal*, 5:397–409, 1988.

[4] Edmond De Hoffmann and Vincent Stroobant. *Mass Spectrometry - Priniples and Applications.*, volume 29. 2007.

[5] Pamela Stanley, Harry Schachter, and Naoyuki Taniguchi. N-Glycans. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2009.

[6] D N Perkins, D J C Pappin, D M Creasy, and J S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[7] Lloyd M. Smith and Neil L Kelleher. Proteoform: a single term describing protein complexity Lloyd. *Nat Methods*, 10(3), 2013.

[8] Kyowon Jeong, Sangtae Kim, and Pavel A. Pevzner. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics (Oxford, England)*, 29(16):1953–62, aug 2013.

[9] Ajit Varki. Biological roles of glycans. *Glycobiology*, 27(1):3–49, 2017.

[10] Ajit. Varki, Richard D Cummings, Jeffrey D Esko, Hudson H Freeze, Pamela Stanley, Carolyn R Bertozzi, Gerald W Hart, and Marilynn E Etzler. *Essentials of Glycobiology. 2nd edition*. Cold Spring Harbor Laboratory Press, 2009.

[11] Emanual Maverakis, Kyoungmi Kim, Michiko Shimoda, M. Eric Gershwin, Forum Patel, Reason Wilken, Siba Raychaudhuri, L. Renee Ruhaak, and Carlito B. Lebrilla. Glycans in the immune system and The Altered Glycan Theory of Autoimmunity: A critical review. *Journal of Autoimmunity*, 57:1–13, jan 2015.

[12] Siobhan V. Glavey, Daisy Huynh, Michaela R. Reagan, Salomon Manier, Michele Moschetta, Yawara Kawano, Aldo M. Roccaro, Irene M. Ghobrial, Lokesh Joshi, and Michael E. O'Dwyer. The cancer glycome: Carbohydrates as mediators of metastasis. *Blood Reviews*, 29(4):269–279, 2015.

[13] S R Stowell, T Ju, and R D Cummings. Protein glycosylation in cancer. *Annu Rev Pathol*, 10:473–510, 2015.

[14] Ieva Bagdonaite and Hans H Wandall. Global aspects of viral glycosylation. *Glycobiology*, 28(7):443–467, jul 2018.

[15] Preethi L. Chandran, Emilios K. Dimitriadis, Edward L. Mertz, and Ferenc Horkay. Microscale mapping of extracellular matrix elasticity of mouse joint cartilage: an approach to extracting bulk elasticity of soft matter with surface roughness. *Soft Matter*, 14:2879–2892, 2018.

[16] Ajit Varki. Nothing in glycobiology makes sense, except in the light of evolution. *Cell*, 126(5):841–845, sep 2006.

[17] Harald Nothaft and Christine M. Szymanski. Bacterial protein n-glycosylation: New perspectives and applications. *Journal of Biological Chemistry*, 288(10):6912–6920, 2013.

[18] Máximo Lopez Medus, Gabriela E. Gomez, Lucía F. Zacchi, Paula M. Couto, Carlos A. Labriola, María S. Labanda, Rodrigo Corti Bielsa, Eugenia M. Clérico, Benjamin L. Schulz, and Julio J. Caramelo. N-glycosylation Triggers a Dual Selection Pressure in Eukaryotic Secretory Proteins. *Scientific Reports*, 7(1):1–11, 2017.

[19] Julia Krushkal, Yingdong Zhao, Curtis Hose, Anne Monks, James H. Doroshow, and Richard Simon. Longitudinal Transcriptional Response of Glycosylation-Related Genes, Regulators, and Targets in Cancer Cell Lines Treated With 11 Antitumor Agents. *Cancer Informatics*, 16, 2017.

[20] Gordan Lauc, Jasminka Krištić, and Vlatka Zoldoš. Glycans - the third revolution in evolution. *Frontiers in Genetics*, 5(MAY):1–7, 2014.

[21] John B. Lowe and Jamey D. Marth. A Genetic Approach to Mammalian Glycan Function. *Annual Review of Biochemistry*, 72(1):643–691, 2003.

[22] Ezequiel Valguarnera, Rachel L. Kinsella, and Mario F. Feldman. Sugar and Spice Make Bacteria Not Nice: Protein Glycosylation and Its Influence in Pathogenesis. *Journal of Molecular Biology*, 428(16):3206–3220, 2016.

[23] Kshitij Khatri, Joshua A Klein, Mitchell R. White, Oliver C. Grant, Nancy Lemarie, Robert J. Woods, Kevan L. Hartshorn, Joseph Zaia, Nancy Leymarie, Robert J. Woods, Kevan L. Hartshorn, and Joseph Zaia. Integrated omics and computational glycobiology reveal structural basis for Influenza A virus glycan microheterogeneity and host interactions. *Molecular & cellular proteomics : MCP*, 13975(615), mar 2016.

[24] Liwei Cao, Jolene K. Diedrich, Daniel W. Kulp, Matthias Pauthner, Lin He, Sung Kyu Robin Park, Devin Sok, Ching Yao Su, Claire M. Delahunty, Sergey Menis, Raiees Andrabi, Javier Guenaga, Erik Georgeson, Michael Kubitz, Yumiko Adachi, Dennis R. Burton, William R. Schief, John R. Yates, and James C. Paulson. Global site-specific N-glycosylation analysis of HIV envelope glycoprotein. *Nature Communications*, 8:1–13, 2017.

[25] Eden P. Go, Haitao Ding, Shijian Zhang, Rajesh P. Ringe, Nathan Nicely, David Hua, Robert T. Steinbock, Michael Golabek, James Alin, S. Munir Alam, Albert Cupo, Barton F. Haynes, John C. Kappes, John P. Moore, Joseph G. Sodroski, and Heather Desaire. A glycosylation benchmark profile for HIV-1 envelope glycoprotein production based on eleven Env trimers. *Journal of Virology*, 91(9):JVI.02428–16, 2017.

[26] Jianhui Tian, Cesar A. López, Cynthia A. Derdeyn, Morris S. Jones, Abraham Pinter, Bette Korber, and S. Gnanakaran. Effect of Glycosylation on an Immunodominant Region in the V1V2 Variable Domain of the HIV-1 Envelope gp120 Protein. *PLoS Computational Biology*, 12(10):1–33, 2016.

[27] John J. Skehel and Don C. Wiley. Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. *Annual Review of Biochemistry*, 69(1):531–569, jun 2000.

[28] Ruslan Medzhitov and Charles. A Jr Janeway. Innate immunity: Minireview the virtues of a nonclonal system of recognition. *Cell*, 91(3):295–298, 1997.

[29] Bruce A Beutler. TLRs and innate immunity. *Blood*, 113(7):1399–1407, sep 2008.

[30] Amr El-Hawiet, Elena N. Kitova, and John S. Klassen. Recognition of human milk oligosaccharides by bacterial exotoxins. *Glycobiology*, 25(8):845–854, 2015.

[31] J. N. Arnold, C. M. Radcliffe, M. R. Wormald, L. Royle, D. J. Harvey, M. Crispin, R. A. Dwek, R. B. Sim, and P. M. Rudd. The Glycosylation of Human Serum IgD and IgE and the Accessibility of Identified Oligomannose Structures for Interaction with Mannan-Binding Lectin. *The Journal of Immunology*, 173(11):6831–6840, 2004.

[32] Jake W. Pawlowski, Adriana Bajardi-Taccioli, Damian Houde, Marina Feschenko, Tyler Carlage, and Igor A. Kaltashov. Influence of glycan modification on IgG1 biochemical and biophysical properties. *Journal of Pharmaceutical and Biomedical Analysis*, 151:133–144, 2018.

[33] Edward S.X. Moh, Chi Hung Lin, Morten Thaysen-Andersen, and Nicolle H. Packer. Site-Specific N-Glycosylation of Recombinant Pentameric and Hexameric Human IgM. *Journal of the American Society for Mass Spectrometry*, 27(7):1143–1155, 2016.

[34] Rosina Plomp, Paul J. Hensbergen, Yoann Rombouts, Gerhild Zauner, Irina Dragan, Carolien A M Koeleman, André M. Deelder, and Manfred Wuhrer. Site-specific N-glycosylation analysis of human immunoglobulin e. *Journal of Proteome Research*, 13(2):536–546, 2014.

[35] Jincui Huang, Andres Guerrero, Evan Parker, John S. Strum, Jennifer T. Smilowitz, J. Bruce German, and Carlito B. Lebrilla. Site-specific glycosylation of secretory immunoglobulin a from human colostrum. *Journal of Proteome Research*, 14(3):1335–1349, 2015.

[36] Roy Jefferis. Isotype and glycoform selection for antibody therapeutics. *Archives of Biochemistry and Biophysics*, 526(2):159–166, 2012.

[37] Taia T. Wang, Jad Maamary, Gene S. Tan, Stylianos Bournazos, Carl W. Davis, Florian Krammer, Sarah J. Schlesinger, Peter Palese, Rafi Ahmed, and Jeffrey V. Ravetch. Anti-HA Glycoforms Drive B Cell Affinity Selection and Determine Influenza Vaccine Efficacy. *Cell*, 162(1):160–169, 2015.

[38] Caroline Soliman, Elizabeth Yuriev, and Paul A. Ramsland. Antibody recognition of aberrant glycosylation on the surface of cancer cells. *Current Opinion in Structural Biology*, 44:1–8, 2017.

[39] Mark M. Fuster and Jeffrey D. Esko. The sweet and sour of cancer: Glycans as novel therapeutic targets. *Nature Reviews Cancer*, 5(7):526–542, 2005.

[40] Y.-C. Liu, H.-Y. Yen, C.-Y. Chen, C.-H. Chen, P.-F. Cheng, Y.-H. Juan, C.-H. Chen, K.-H. Khoo, C.-J. Yu, P.-C. Yang, T.-L. Hsu, and C.-H. Wong. Sialylation and fucosylation of epidermal growth factor receptor suppress its dimerization and activation in lung cancer cells. *Proceedings of the National Academy of Sciences*, 108(28):11332–11337, 2011.

[41] Matthew J. Paszek, Christopher C. Dufort, Olivier Rossier, Russell Bainer, Janna K. Mouw, Kamil Godula, Jason E. Hudak, Jonathon N. Lakins, Amanda C. Wijekoon, Luke Cassereau, Matthew G. Rubashkin, Mark J. Magbanua, Kurt S. Thorn, Michael W. Davidson, Hope S. Rugo, John W. Park, Daniel A. Hammer, Grégory Giannone, Carolyn R. Bertozzi, and Valerie M. Weaver. The cancer glycocalyx mechanically primes integrin-mediated growth and survival. *Nature*, 511(7509):319–325, 2014.

[42] Joana G Rodrigues, Meritxell Balmaña, Joana A Macedo, Juliana Poças, Ângela Fernandes, Julio Cesar M De-Freitas-Junior, Salomé S Pinho, Joana Gomes, Ana Magalhães, Catarina Gomes, Stefan Mereiter, and Celso A Reis. Glycosylation in Cancer: Selected Roles in Tumour Progression, Immune Modulation and Metastasis. *Cellular Immunology*, (January), 2018.

[43] Denong Wang, Lisa Wu, and Xiaohe Liu. Glycan markers as potential immunological targets in circulating tumor cells. In Mark Jesus M. Magbanua and John W. Park, editors, *Circulating Tumor Cells as Cancer Biomarkers in the Clinic*, chapter 10, pages 275–284. Springer Nature, Gewerbestrasse 11, 6330 Cham, Switzerland, 2017.

[44] Ajit Varki, Reiji Kannagi, Bryan Toole, and Pamela Stanley. Glycosylation Changes in Cancer. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2015.

[45] Renato V. Iozzo and Liliana Schaefer. Proteoglycan form and function: A comprehensive nomenclature of proteoglycans. *Matrix Biology*, 42:11–55, 2015.

[46] Edward V. Maytin. Hyaluronan: More than just a wrinkle filler. *Glycobiology*, 26(6):553–559, 2016.

[47] Mary C. Farach-Carson, Curtis R. Warren, Daniel A. Harrington, and Daniel D. Carson. Border patrol: Insights into the unique role of perlecan/heparan sulfate proteoglycan 2 at cell and tissue borders. *Matrix Biology*, 34:64–79, 2014.

[48] Andrea Malandrino, Michael Mak, Roger D. Kamm, and Emad Moeendarbary. Complex mechanics of the heterogeneous extracellular matrix in cancer. *Extreme Mechanics Letters*, 21:25–34, 2018.

[49] Jie Zhou, Xuewen Du, Xiaoyi Chen, and Bing Xu. Adaptive Multifunctional Supramolecular Assemblies of Glycopeptides Rapidly Enable Morphogenesis. *Biochemistry*, 57(32):4867–4879, 2018.

[50] Maria A. Gubbiotti, Thomas Neill, and Renato V. Iozzo. A current view of perlecan in physiology and pathology: A mosaic of functions. *Matrix Biology*, 57-58:285–298, 2017.

[51] Elizabeth J. Bradbury, Lawrence D. F. Moon, Reena J. Popat, Von R. King, Gavin S. Bennett, Preena N. Patel, James W. Fawcett, and Stephen B. McMahon. Chondroitinase ABC promotes functional recovery after spinal cord injury. *Nature*, 416(6881):636–640, 2002.

[52] Tracy Laabs, Daniela Carulli, Herbert M. Geller, and James W. Fawcett. Chondroitin sulfate proteoglycans in neural development and regeneration. *Current Opinion in Neurobiology*, 15(1):116–120, 2005.

[53] Lorraine W Lau, Rowena Cua, Michael B Keough, Sarah Haylock-Jacobs, and V Wee Yong. Pathophysiology of the brain extracellular matrix: a new target for remyelination. *Nature reviews. Neuroscience*, 14(10):722–9, 2013.

[54] Joshua A Klein, Le Meng, and Joseph Zaia. Deep sequencing of complex proteoglycans: a novel strategy for high coverage and site- specific identification of glycosaminoglycan-linked peptides. *Molecular & Cellular Proteomics*, 17, 2018.

[55] Karen J. Colley, Ajit Varki, and Taroh Kinoshita. Cellular Organization of Glycosylation. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2017.

[56] James M. Rini and Jeffrey D. Esko. Glycosyltransferases and Glycan-Processing Enzymes. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2017.

[57] R. L. Schnaar, R. Gerardy-Schahn, and H. Hildebrandt. Sialic Acids in the Brain: Gangliosides and Polysialic Acid in Nervous System Development, Stability, Disease, and Regeneration. *Physiological Reviews*, 94(2):461–518, 2014.

[58] Akemi Suzuki. Genetic basis for the lack of N-glycolylneuraminic acid expression in human tissues and its implication to human evolution. *Proceedings of the Japan Academy. Series B, Physical and biological sciences*, 82(3):93–103, 2006.

[59] Richard D. Cummings. The repertoire of glycan determinants in the human glycome. *Molecular bioSystems*, 5(10):1087–104, 2009.

[60] Alan D Mcnaught. NOMENCLATURE OF CARBOHYDRATES (Recommendations 1996) Prepared. *Pure and Applied Chemistry*, 68(10):1919–2008, 1996.

[61] S Herget, R Ranzinger, K Maass, and C-W V D Lieth. GlycoCT-a unifying sequence format for carbohydrates. *Carbohydrate research*, 343(12):2162–71, aug 2008.

[62] Shiteshu Shrimal, Natalia A. Cherepanova, and Reid Gilmore. Cotranslational and posttranslocational N-glycosylation of proteins in the endoplasmic reticulum. *Seminars in Cell and Developmental Biology*, 41:71–78, 2015.

[63] Francesco Mallamace, Carmelo Corsaro, Domenico Mallamace, Sebastiano Vasi, Cirino Vasi, Piero Baglioni, Sergey V. Buldyrev, Sow-Hsin Chen, and H. Eugene Stanley. Energy landscape in protein folding and unfolding. *Proceedings of the National Academy of Sciences*, 113(12):3159–3163, 2016.

[64] D. Shental-Bechor and Y. Levy. Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. *Proceedings of the National Academy of Sciences*, 105(24):8256–8261, 2008.

[65] Andrei J. Petrescu, Adina L. Milac, Stefana M. Petrescu, Raymond A. Dwek, and Mark R. Wormald. Statistical analysis of the protein environment of N-glycosylation sites: Implications for occupancy, structure, and folding. *Glycobiology*, 14(2):103–114, 2004.

[66] Akitsugu Suga, Masamichi Nagae, and Yoshiki Yamaguchi. Analysis of protein landscapes around N-glycosylation sites from the PDB repository for understanding the structural basis of N-glycoprotein processing and maturation. *Glycobiology*, pages 1–12, 2018.

[67] Robert J Woods. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chemical Reviews*, 118(17):8005–8024, sep 2018.

[68] Frederick J. Krambeck, Sandra V. Bennun, Someet Narang, Sean Choi, Kevin J. Yarema, and Michael J. Betenbaugh. A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology*, 19(11):1163–1175, 2009.

[69] Gang Liu and Sriram Neelamegham. A Computational Framework for the Automated Construction of Glycosylation Reaction Networks. *PLoS ONE*, 9(6):e100939, jun 2014.

[70] Yukie Akune, Chi-Hung Lin, Jodie L. Abrahams, Jingyu Zhang, Nicolle H. Packer, Kiyoko F. Aoki-Kinoshita, and Matthew P. Campbell. Comprehensive analysis of the N-glycan biosynthetic pathway using bioinformatics to generate UniCorn: A theoretical N-glycan structure database. *Carbohydrate Research*, 431:56–63, 2016.

[71] Inka Brockhausen and Pamela Stanley. *O-GalNAc Glycans*. Cold Spring Harbor Laboratory Press, 2015.

[72] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, jan 2017.

[73] Inka Brockhausen, Harry Schachter, and Pamela Stanley. O-GalNAc Glycans. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2009.

[74] Franz Georg Hanisch and Stefan Müller. MUC1: The polymorphic appearance of a human mucin. *Glycobiology*, 10(5):439–449, 2000.

[75] Katja Engelmann, Stephan E. Baldus, and Franz Georg Hanisch. Identification and Topology of Variant Sequences within Individual Repeat Domains of the Human Epithelial Tumor Mucin MUC1. *Journal of Biological Chemistry*, 276(30):27764–27769, 2001.

[76] Kirk S.B. Bergstrom and Lijun Xia. Mucin-type O-glycans and their roles in intestinal homeostasis. *Glycobiology*, 23(9):1026–1037, 2013.

[77] Gang Liu, Dhananjay D. Marathe, Khushi L. Matta, and Sriram Neelamegham. Systems-level modeling of cellular glycosylation reaction networks: O-linked glycan formation on natural selectin ligands. *Bioinformatics*, 24(23):2740–2747, 2008.

[78] Ulf Lindahl, John Couchman, Koji Kimata, and Jeffrey D. Esko. Proteoglycans and Sulfated Glycosaminoglycans. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2015.

[79] Vincent Hascall and Jeffrey D. Esko. Hyaluronan. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2015.

[80] Jeffrey D. Esko and Lijuan Zhangt. Influence of core protein sequence on glycosaminoglycan assembly. *Current Opinion in Structural Biology*, 6(5):663–670, 1996.

[81] Joseph R. Bishop, Manuela Schuksz, and Jeffrey D. Esko. Heparan sulphate proteoglycans fine-tune mammalian physiology. *Nature*, 446(7139):1030–1037, 2007.

[82] Yunli Hu, Shiyue Zhou, Chuan-Yih Yih Yu, Haixu Tang, and Yehia Mechref. Automated annotation and quantitation of glycans by liquid chromatography/electrospray ionization mass spectrometric analysis using the MultiGlycan-ESI computational tool. *Rapid Communications in Mass Spectrometry*, 29(1):135–142, jan 2014.

[83] John D Hogan, Joshua A Klein, Jiandong Wu, Pradeep Chopra, Geert-Jan Boons, Luis Carvalho, Cheng Lin, and Joseph Zaia. Software for peak finding and elemental composition assignment for glycosaminoglycan tandem mass spectra. *Molecular & Cellular Proteomics*, 1(617):mcp.RA118.000590, apr 2018.

[84] Jiandong Wu, Juan Wei, John D. Hogan, Pradeep Chopra, Apoorva Joshi, Weigang Lu, Joshua Klein, Geert-Jan Boons, Cheng Lin, and Joseph Zaia. Negative Electron Transfer Dissociation Sequencing of 3-O-Sulfation-Containing Heparan Sulfate Oligosaccharides. *Journal of The American Society for Mass Spectrometry*, pages 1–11, mar 2018.

[85] Ajit Varki and Stuart Kornfeld. Historical Background and Overview. In *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2015.

[86] Vicki H. Wysocki, Katheryn A. Resing, Qingfen Zhang, and Guilong Cheng. Mass spectrometry of peptides and proteins. *Methods*, 35(3 SPEC.ISS.):211–222, 2005.

[87] Kermit K. Murray, Robert K. Boyd, Marcos N. Eberlin, G. John Langley, Liang Li, and Yasuhide Naito. Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure and Applied Chemistry*, 85(7):1515–1609, 2013.

[88] Kshitij Khatri, Yi Pu, Joshua A. Klein, Juan Wei, Catherine E. Costello, Cheng Lin, and Joseph Zaia. Comparison of Collisional and Electron-Based Dissociation Modes for Middle-Down Analysis of Multiply Glycosylated Peptides. *Journal of The American Society for Mass Spectrometry*, 29(6):1075–1085, jun 2018.

[89] K Håkansson, H J Cooper, M R Emmett, C E Costello, A G Marshall, and C L Nilsson. Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptic to yield complementary sequence information. *Analytical chemistry*, 73(18):4530–6, sep 2001.

[90] John E P Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9528–33, jun 2004.

[91] Xiang Yu, Yiqun Huang, Cheng Lin, and Catherine E. Costello. Energy-Dependent Electron Activated Dissociation of Metal-Adducted Permethylated Oligosaccharides. *Analytical Chemistry*, 84(17):7487–7494, sep 2012.

[92] Kshitij Khatri, Joshua A. Klein, John R. J.R. Haserick, Deborah R. D.R. Leon, C.E. Catherine E. Costello, M.E. Mark E. McComb, and Joseph Zaia. Microfluidic Capillary Electrophoresis-Mass Spectrometry for Analysis of Monosaccharides, Oligosaccharides, and Glycopeptides. *Analytical Chemistry*, 89(12):acs.analchem.7b00875, jun 2017.

[93] Uma Kota and Mark L. Stolowitz. *Improving Proteome Coverage by Reducing Sample Complexity via Chromatography*, volume 919 of *Advances in Experimental Medicine and Biology*. Springer International Publishing, Cham, 2016.

[94] Joshua Klein, Luis Carvalho, and Joseph Zaia. Application of network smoothing to glycan LC-MS profiling. *Bioinformatics*, (May), may 2018.

[95] Bas C. Jansen, David Falck, Noortje De Haan, Agnes L. Hipgrave Ederveen, Genadij Razdorov, Gordan Lauc, and Manfred Wuhrer. LaCyTools: A Targeted Liquid Chromatography-Mass Spectrometry Data Processing Package for Relative Quantitation of Glycopeptides. *Journal of Proteome Research*, 15(7):2198–2210, 2016.

[96] Chuan-Yih C.-Y. Yu, Anoop Mayampurath, Yunli Hu, Shiyue Zhou, Yehia Mechref, and Haixu Tang. Automated annotation and quantification of glycans using liquid chromatography-mass spectrometry. *Bioinformatics*, 29(13):1706–1707, jul 2013.

[97] Evan Maxwell, Yan Tan, Yuxiang Tan, Han Hu, Gary Benson, Konstantin Aizikov, Shannon Conley, Gregory O Staples, Gordon W Slysz, Richard D Smith, and Joseph Zaia. GlycReSoft: a software package for automated recognition of glycans from LC/MS data. *PloS one*, 7(9):e45474, jan 2012.

[98] Hannes L. Röst, George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T. Schubert, Witold Wolski, Ben C. Collins, Johan Malmström, Lars Malmström, and Ruedi Aebersold. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology*, 32(3):219–223, mar 2014.

[99] Scott R. Kronewitter, Gordon W. Slysz, Ioan Marginean, Clay D. Hagler, Brian L. LaMarche, Rui Zhao, Myanna Y. Harris, Matthew E. Monroe, Christina A. Polyukh, Kevin L. Crowell, Thomas L. Fillmore, Timothy S. Carlson, David G. Camp, Ronald J. Moore, Samuel H. Payne, Gordon a. Anderson, and Richard D. Smith. GlyQ-IQ: Glycomics quintavariate-informed quantification with high-performance computing and glycogrid 4D visualization. *Analytical Chemistry*, 86(13):6268–6276, jul 2014.

[100] George Rosenberger, Yansheng Liu, Hannes L. Röst, Christina Ludwig, Alfonso Buil, Ariel Bensimon, Martin Soste, Tim D. Spector, Emmanouil T. Dermitzakis, Ben C. Collins, Lars Malmström, and Ruedi Aebersold. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nature Biotechnology*, 35(8):781–788, jun 2017.

[101] Sophie Zhao, Ian Walsh, Jodie L Abrahams, Louise Royle, Terry Nguyen-Khuong, Daniel Spencer, Daryl L Fernandes, Nicolle H Packer, Pauline M Rudd, and Matthew P Campbell. GlycoStore: A Database of Retention Properties for Glycan Analysis. *Bioinformatics (Oxford, England)*, (May):0–0, apr 2018.

[102] Yining Huang, Yongxin Nie, Barry Boyes, and Ron Orlando. Resolving isomeric glycopeptide glycoforms with hydrophilic interaction chromatography (HILIC). *Journal of Biomolecular Techniques*, 27(3):98–104, 2016.

[103] Hannes L Röst, Yansheng Liu, Giuseppe D'Agostino, Matteo Zanella, Pedro Navarro, George Rosenberger, Ben C Collins, Ludovic Gillet, Giuseppe Testa, Lars Malmström, and Ruedi Aebersold. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nature Methods*, 13(9):777–783, 2016.

[104] Oleg V. Krokhin. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: Application to 300- and 100-?? pore size C18 sorbents. *Analytical Chemistry*, 78(22):7785–7795, 2006.

[105] L. Veillon, S. Zhou, and Y. Mechref. Quantitative Glycomics: A Combined Analytical and Bioinformatics Approach. In Arun K. Shukla, editor, *Methods in Enzymology*, volume 585, chapter 22, pages 431–477. Academic Press, 2017.

[106] Claudia Escher, Lukas Reiter, Brendan Maclean, Reto Ossola, Franz Herzog, John Chilton, Michael J. Maccoss, and Oliver Rinner. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics*, 12(8):1111–1121, apr 2012.

[107] Adrian M. Taylor, Otto Holst, and Jane Thomas-Oates. Mass spectrometric profiling of O-linked glycans released directly from glycoproteins in gels using in-gel reductive β-elimination. *Proteomics*, 6(10):2936–2946, 2006.

[108] Shin ichi Nakakita, Wataru Sumiyoshi, Nobumitsu Miyanishi, and Jun Hirabayashi. A practical approach to N-glycan production by hydrazinolysis using hydrazine monohydrate. *Biochemical and Biophysical Research Communications*, 362(3):639–645, 2007.

[109] Pilsoo Kang, Yehia Mechref, and Milos V. Novotny. High-throughput solid-phase permethylation of glycans prior to mass spectrometry. *Rapid Communications in Mass Spectrometry*, 22(5):721–734, mar 2008.

[110] Bogdan Bogdanov and Richard D. Smith. Proteomics by fticr mass spectrometry: TOP down and bottom up. *Mass Spectrometry Reviews*, 24(2):168–200, 2005.

[111] Matthias Schittmayer, Katarina Fritz, Laura Liesinger, Johannes Griss, and Ruth Birner-Gruenberger. Cleaning out the Litterbox of Proteomic Scientists' Favorite Pet: Optimized Data Analysis Avoiding Trypsin Artifacts. *Journal of Proteome Research*, 15(4):1222–1229, apr 2016.

[112] Elisabeth Gasteiger, Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D Appel, and Amos Bairoch. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, 31(13):3784–8, jul 2003.

[113] Chris W. Sutton, Jacqui A. O'Neill, and John S. Cottrell. Site-specific characterization of glycoprotein carbohydrates by exoglycosidase digestion and laser desorption mass spectrometry, 1994.

[114] Yu Xue, Juanjuan Xie, Pan Fang, Jun Yao, Guoquan Yan, Huali Shen, and Pengyuan Yang. Study on behaviors and performances of universal N-glycopeptides enrichment methods. *The Analyst*, 2018.

[115] Hannes L Rost, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E Wolski, Oliver Schilling, Jyoti S Choudhary, Lars Malmstrom, Ruedi Aebersold, Knut Reinert,

and Oliver Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Meth*, 13(9):741–748, 2016.

[116] Eva Lange, Clemens Gröpl, Knut Reinert, Oliver Kohlbacher, and Andreas Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 254:243–254, 2006.

[117] William R French, Lisa J Zimmerman, Birgit Schilling, Bradford Wayne Gibson, Christine A Miller, R Reid Townsend, Stacy D Sherrod, Cody R Goodwin, John A Mclean, and David Lee Tabb. Wavelet-Based Peak Detection and a New Charge Inference Procedure for MS / MS Implemented in ProteoWizard ' s msConvert. *J. Proteome Res*, 2014.

[118] Bernhard Y Renard, Marc Kirchner, Hanno Steen, Judith A J Steen, and Fred A Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC bioinformatics*, 9(1):355, jan 2008.

[119] Navdeep Jaitly, Anoop Mayampurath, Kyle Littlefield, Joshua N Adkins, Gordon A Anderson, and Richard D Smith. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC bioinformatics*, 10(1):87, jan 2009.

[120] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, nov 2008.

[121] David P.A. Kilgour, Sam Hughes, Samantha L. Kilgour, C. Logan Mackay, Magnus Palmblad, Bao Quoc Tran, Young Ah Goo, Robert K. Ernst, David J. Clarke, and David R. Goodlett. Autopiquer - a Robust and Reliable Peak Detection Algorithm for Mass Spectrometry. *Journal of the American Society for Mass Spectrometry*, 28(2):253–262, 2017.

[122] Jan Urban, Nils Kristian Afseth, and Dalibor Štys. Fundamental definitions and confusions in mass spectrometry about mass assignment, centroiding and resolution. *TrAC - Trends in Analytical Chemistry*, 53:126–136, 2014.

[123] Gerhard Mayer, Andrew R Jones, Pierre-Alain Binz, Eric W Deutsch, Sandra Orchard, Luisa Montecchi-Palazzi, Juan Antonio Vizcaíno, Henning Hermjakob, David Oveillero, Randall Julian, Christian Stephan, Helmut E Meyer, and Martin Eisenacher. Controlled vocabularies and ontologies in proteomics: overview, principles and practice. *Biochimica et biophysica acta*, 1844(1 Pt A):98–107, jan 2014.

[124] Viktor Granholm, Sangtae Kim, José C F Navarro, Erik Sjölund, Richard D Smith, and Lukas Käll. Fast and accurate database searches with MS-GF+Percolator. *Journal of proteome research*, 13(2):890–7, feb 2014.

[125] Lev I Levitsky, Mark V Ivanov, Anna A Lobas, Julia A Bubis, Irina A Tarasova, Elizaveta M Solovyeva, Marina L Pridatchenko, and Mikhail V Gorshkov. IdentiPy: an extensible search engine for protein identification in shotgun proteomics. *Journal of Proteome Research*, page acs.jproteome.7b00640, 2018.

[126] Craig D Wenger and Joshua J Coon. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *Journal of proteome research*, 12(3):1377–86, mar 2013.

[127] Piotr Dittwald and Dirk Valkenborg. BRAIN 2.0: time and memory complexity improvements in the algorithm for calculating the isotope distribution. *Journal of the American Society for Mass Spectrometry*, 25(4):588–94, apr 2014.

[128] Piotr Dittwald, Jürgen Claesen, Tomasz Burzykowski, Dirk Valkenborg, and Anna Gambin. BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Analytical chemistry*, 85(4):1991–4, feb 2013.

[129] Alan L. Rockwood and Perttu Haimi. Efficient calculation of accurate masses of isotopic peaks. *Journal of the American Society for Mass Spectrometry*, 17(3):415–419, mar 2006.

[130] Mateusz K. Łacki, Michał Startek, Dirk Valkenborg, and Anna Gambin. IsoSpec: Hyperfast Fine Structure Calculator. *Analytical Chemistry*, 89(6):3272–3277, 2017.

[131] Gelio Alves, Aleksey Y. Ogurtsov, and Yi Kuo Yu. Molecular Isotopic Distribution Analysis (MIDAs) with adjustable mass accuracy. *Journal of the American Society for Mass Spectrometry*, 25(1):57–70, 2014.

[132] Ming-Qi Liu, Wen-Feng Zeng, Pan Fang, Wei-Qian Cao, Chao Liu, Guo-Quan Yan, Yang Zhang, Chao Peng, Jian-Qiang Wu, Xiao-Jin Zhang, Hui-Jun Tu, Hao Chi, Rui-Xiang Sun, Yong Cao, Meng-Qiu Dong, Bi-Yun Jiang, Jiang-Ming Huang, Hua-Li Shen, Catherine C. L. Wong, Si-Min He, and Peng-Yuan Yang. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nature Communications*, 8(1):438, 2017.

[133] Waqas Nasir, Alejandro Gomez Toledo, Fredrik Noborn, Jonas Nilsson, Mingxun Wang, Nuno Bandeira, and Göran Larson. SweetNET : A Bioinformatics Workflow for Glycopeptide MS/MS Spectral Analysis. *Journal of Proteome Research*, 15(8):2826–2840, aug 2016.

[134] David L Tabb. The SEQUEST Family Tree. *Journal of The American Society for Mass Spectrometry*, 26(11):1814–1819, nov 2015.

[135] Kenneth Verheggen, Helge Ræder, Frode S. Berven, Lennart Martens, Harald Barsnes, and Marc Vaudel. Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews*, (July):1–15, 2017.

[136] Sangtae Kim and Pavel A Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5:5277, oct 2014.

[137] Stefan K. Solntsev, Michael R. Shortreed, Brian L. Frey, and Lloyd M. Smith. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *Journal of Proteome Research*, page acs.jproteome.7b00873, 2018.

[138] Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. MSFragger : ultrafast and comprehensive peptide identification in mass spectrometry – based proteomics. *Nature Methods*, 14(5), 2017.

[139] David M Creasy and John S Cottrell. Unimod: Protein modifications for mass spectrometry. *Proteomics*, 4(6):1534–6, jun 2004.

[140] Kaijie Xiao, Yue Wang, Yun Shen, Yuyin Han, and Zhixin Tian. Large-scale identification and visualization of N-glycans with primary structures using GlySeeker. *Rapid Communications in Mass Spectrometry*, 32(2):142–148, jan 2018.

[141] Alessio Ceroni, Kai Maass, Hildegard Geyer, Rudolf Geyer, Anne Dell, and Stuart M. Haslam. GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans. *Journal of Proteome Research*, 7(4):1650–1659, apr 2008.

[142] Carlos Guijas, J. Rafael Montenegro-Burke, Xavier Domingo-Almenara, Amelia Palermo, Benedikt Warth, Gerrit Hermann, Gunda Koellensperger, Tao Huan, Winnie Uritboonthai, Aries E. Aisporna, Dennis W. Wolan, Mary E. Spilker, H. Paul Benton, and Gary Siuzdak. METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Analytical Chemistry*, 90(5):3156–3164, 2018.

[143] H P Benton, D M Wong, S A Trauger, and G Siuzdak. XCMS2: Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization. 80(16):6382–6389, 2009.

[144] Michael A. Kochen, Matthew C. Chambers, Jay D. Holman, Alexey I. Nesvizhskii, Susan T. Weintraub, John T. Belisle, M. Nurul Islam, Johannes Griss, and David L. Tabb. Greazy: Open-Source Software for Automated Phospholipid Tandem Mass Spectrometry Identification. *Analytical Chemistry*, 88(11):5733–5741, 2016.

[145] Richard D LeDuc, Ryan T Fellers, Bryan P Early, Joseph B Greer, Paul M Thomas, and Neil L Kelleher. The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. *Journal of proteome research*, 13(7):3231–40, jul 2014.

[146] Zhaorui Zhang, Si Wu, David L. Stenoien, and Ljiljana Paša-Tolić. High-Throughput Proteomics. *Annual Review of Analytical Chemistry*, 7(1):427–454, 2014.

[147] Paul J. Sample, Kirk W. Gaston, Juan D. Alfonzo, and Patrick A. Limbach. RoboOligo: Software for mass spectrometry data to support manual and de novo sequencing of post-transcriptionally modified ribonucleic acids. *Nucleic Acids Research*, 43(10):1–13, 2015.

[148] Jian Wang, Veronica G Anania, Jeff Knott, John Rush, Jennie R Lill, Philip E Bourne, and Nuno Bandeira. Combinatorial approach for large-scale identification of linked peptides from tandem mass spectrometry spectra. *Molecular & cellular proteomics : MCP*, 13(4):1128–36, 2014.

[149] Yi Liu, Weiping Sun, Kaizhong Zhang, Baozhen Shan, and Kaizhong Zhang. DISC: DISulfide linkage Characterization from tandem mass spectra. *Bioinformatics*, 33(December):3861–3870, 2017.

[150] Lin He, Lei Xin, Baozhen Shan, Gilles a Lajoie, and Bin Ma. GlycoMaster DB: Software to Assist the Automated Identification of N-Linked Glycopeptides by Tandem Mass Spectrometry. *Journal of Proteome Research*, 2014.

[151] John S. Strum, Charles C. Nwosu, Serenus Hua, Scott R. Kronewitter, Richard R. Seipert, Robert J. Bachelor, Hyun Joo An, and Carlito B. Lebrilla. Automated assignments of N- and O-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures. *Analytical Chemistry*, 85(12):5666–5675, jun 2013.

[152] Shisheng Sun, Punit Shah, Shadi Toghi Eshghi, Weiming Yang, Namita Trikannad, Shuang Yang, Lijun Chen, Paul Aiyetan, Naseruddin Höti, Zhen Zhang, Daniel W Chan, and Hui Zhang. Comprehensive analysis of protein glycosylation by solid-phase extraction of N-linked glycans and glycosite-containing peptides. *Nature Biotechnology*, 34(1):84–88, 2015.

[153] Gang Liu, Kai Cheng, Chi Y. Lo, Jun Li, Jun Qu, and Sriram Neelamegham. A comprehensive, open-source platform for mass spectrometry based glycoproteomics data analysis. *Molecular & cellular proteomics : MCP*, page mcp.M117.068239, sep 2017.

[154] Xiaowen Liu, Yakov Sirotkin, Yufeng Shen, Gordon Anderson, Yihsuan S. Tsai, Ying S. Ting, David R. Goodlett, Richard D. Smith, Vineet Bafna, and Pavel A. Pevzner. Protein Identification Using Top-Down Spectra. *Molecular & Cellular Proteomics*, 11(6):M111.008524, 2012.

[155] Qiang Kou, Si Wu, Nikola Tolić, Ljiljana Paša-Tolić, Yunlong Liu, and Xiaowen Liu. A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. *Bioinformatics*, 33(December 2016):btw806, 2016.

[156] Robertson Craig and Ronald C Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)*, 20(9):1466–7, jun 2004.

[157] Gelio Alves, Aleksey Y. Ogurtsov, and Yi-Kuo Yu. RAId_aPS: MS/MS Analysis with Multiple Scoring Functions and Spectrum-Specific Statistics. *PLoS ONE*, 5(11):e15438, nov 2010.

[158] John T Halloran, Jeff A Bilmes, and William S Noble. Learning Peptide-Spectrum Alignment Models for Tandem Mass Spectrometry. *Uncertainty in artificial intelligence : proceedings of the ... conference. Conference on Uncertainty in Artificial Intelligence*, 30:320–329, jan 2014.

[159] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *Journal of proteome research*, 3(5):958–64, 2004.

[160] Brian A Risk, Nathan J Edwards, and Morgan C Giddings. A peptide-spectrum scoring system based on ion alignment, intensity, and pair probabilities. *Journal of proteome research*, 12(9):4240–7, sep 2013.

[161] Thomas Taus, Thomas Köcher, Peter Pichler, Carmen Paschke, Andreas Schmidt, Christoph Henrich, and Karl Mechtler. Universal and confident phosphorylation site localization using phosphoRS. *Journal of Proteome Research*, 10(12):5354–5362, 2011.

[162] Chuan-Le Xiao, Xiao-Zhou Chen, Yang-Li Du, Xuesong Sun, Gong Zhang, and Qing-Yu He. Binomial probability distribution model-based protein identification algorithm for tandem mass spectrometry utilizing peak intensity information. *Journal of proteome research*, 12(1):328–35, jan 2013.

[163] Ari Frank and Pavel Pevzner. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77(4):964–973, 2005.

[164] Xie Xuan Zhou, Wen Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si Min He, and Zhifei Zhang. PDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Analytical Chemistry*, 89(23):12690–12697, 2017.

[165] David L Tabb, Christopher G Fernando, and Matthew C Chambers. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research*, 6(2):654–61, feb 2007.

[166] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A. Lajoie, and Bin Ma. PEAKS DB: <i>De Novo</i> Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics*, 11(4):M111.010587, 2012.

[167] Hosein Mohimani, Sangtae Kim, and Pavel a Pevzner. A New Approach to Evaluating Statistical Significance of Spectral Identifications. *Journal of Proteome Research*, 12(4):1560–1568, apr 2013.

[168] Lukas Käll, John D. Storey, and William Stafford Noble. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. In *Bioinformatics*, volume 24, pages i42–i48. Oxford University Press, aug 2008.

[169] Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry*, 74(20):5383–92, oct 2002.

[170] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.

[171] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, 1990.

[172] Kelvin Ma, Olga Vitek, Alexey I Nesvizhskii, J Eng, A McCormack, J Yates, R Craig, R Beavis, B MacLean, J Eng, R Beavis, M McIntosh, A Keller, A Nesvizhskii, E Kolker, R Aebersold, A Nesvizhskii, J Whiteaker, H Zhang, J Eng, R Fang, B Piening, L Feng, T Lorentzen, R Schoenherr, J Keane, T Holzman, M Fitzgibbon, C Lin, H Zhang, K Cooke, T Liu, DC II, L Anderson, J Watts, R Smith, M McIntosh, A Paulovich, H Choi, A Nesvizhskii, J Klimek, J Eddes, L Hohmann, J Jackson, A Peterson, S Letarte, P Gafken, J Katz, P Mallick, H Lee, A Schmidt, R Ossola, J Eng, R Aebersold, D Martin, J Storey, B Efron, L Kall, J Storey, M MacCoss, H Choi, D Ghosh, A Nesvizhskii, Y Ding, H Choi, A Nesvizhskii, A Dempster, N Laird, D Rubin, J Storey, J Elias, S Gygi, L Käll, J Storey, M MacCoss, W Noble, E Deutsch, L Mendoza, D Shteynberg, T Farrah, H Lam, N Tasman, Z Sun, E Nilsson, B Pratt, B Prazen, JK Eng, DB Martin, AI Nesvizhskii, R Aebersold, A Nesvizhskii, A Keller, E Kolker, and R Aebersold. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics*, 13(Suppl 16):S1, 2012.

[173] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry*, 22(7):1111–20, jul 2011.

[174] Seungjin Na, Nuno Bandeira, and Eunok Paek. Fast Multi-blind Modification Search through Tandem Mass Spectrometry. *Molecular & Cellular Proteomics*, 11(4):M111.010199, 2012.

[175] Xi Han, Lin He, Lei Xin, Baozhen Shan, and Bin Ma. PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications. *Journal of Proteome Research*, 10(7):2930–2936, jul 2011.

[176] Peter R Baker, Jonathan C Trinidad, and Robert J Chalkley. Modification site localization scoring integrated into a search engine. *Molecular & cellular proteomics : MCP*, 10(7):M111.008078, 2011.

[177] Jesper V. Olsen, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell*, 127(3):635–648, 2006.

[178] Dave C H Lee, Andrew R Jones, and Simon J Hubbard. Computational phosphoproteomics: From identification to localization. *Proteomics*, 15(5-6):950–63, dec 2014.

[179] Min Sik Kim, Jun Zhong, and Akhilesh Pandey. Common errors in mass spectrometry-based analysis of post-translational modifications. *Proteomics*, 16(5):700–714, 2016.

[180] Attila Kertesz-Farkas, Uri Keich, and William Stafford Noble. Tandem Mass Spectrum Identification via Cascaded Search. *Journal of Proteome Research*, 14(8):3027–3038, aug 2015.

[181] Kevin Brown Chandler, Petr Pompach, Radoslav Goldman, and Nathan Edwards. Exploring site-specific N-glycosylation microheterogeneity of haptoglobin using glycopeptide CID tandem mass spectra and glycan database search. *Journal of Proteome Research*, 12:3652–3666, 2013.

[182] Zhikai Zhu, Xiaomeng Su, Eden P Go, and Heather Desaire. New Glycoproteomics Software, GlycoPep Evaluator, Generates Decoy Glycopeptides de novo and Enables Accurate False Discovery Rate Analysis for Small Data Sets. *Analytical chemistry*, 2014.

[183] Shadi Toghi Eshghi, Weiming Yang, Yingwei Hu, Punit Shah, Shisheng Sun, Xingde Li, and Hui Zhang. Classification of Tandem Mass Spectra for Identification of N- and O-linked Glycopeptides. *Scientific Reports*, 6(October):37189, 2016.

[184] Anoop Mayampurath, Chuan Yih Yu, Ehwang Song, Jagadheshwar Balan, Yehia Mechref, and Haixu Tang. Computational framework for identification of intact glycopeptides in complex samples. *Analytical Chemistry*, 86(1):453–463, 2014.

[185] Gun Wook Park, Jin Young Kim, Heeyoun Hwang, Ju Yeon Lee, Young Hee Ahn, Hyun Kyoung Lee, Eun Sun Ji, Kwang Hoe Kim, Hoi Keun Jeong, Ki Na Yun, Yong-Sam Kim, Jeong-Heon Ko, Hyun Joo An, Jae Han Kim, Young-Ki Paik, and Jong Shin Yoo. Integrated GlycoProteome Analyzer (I-GPA) for Automated Identification and Quantitation of Site-Specific N-Glycosylation. *Scientific reports*, 6:21175, jan 2016.

[186] Wen-Feng Zeng, Ming-Qi Liu, Yang Zhang, Jian-Qiang Wu, Pan Fang, Chao Peng, Aiying Nie, Guoquan Yan, Weiqian Cao, Chao Liu, Hao Chi, Rui-Xiang Sun, Catherine C. L. Wong, Si-Min He, and Pengyuan Yang. pGlyco: a pipeline for the identification of intact N-glycopeptides by using HCD- and CID-MS/MS and MS3. *Scientific Reports*, 6(April):25102, 2016.

[187] Karli R Reiding, Albert Bondt, Vojtech Franc, and Albert JR Heck. The benefits of hybrid fragmentation methods for glycoproteomics. *Trends in Analytical Chemistry*, 2018.

[188] Jenny Albanese, Matthias Glueckmann, and Christof Lenz. SimGlycan™ software: a new predictive carbohydrate analysis tool for MS/MS data . pages 1–7, 2010.

[189] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–212, oct 2014.

[190] Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 44(D1):D67–D72, 2016.

[191] Robert J. Chalkley. When Target:Decoy False Discovery Rate Estimations are Inaccurate and How to Spot Instances. 116(8):1477–1490, 2016.

[192] William Stafford Noble. Mass spectrometrists should search only for peptides they care about. *Nature methods*, 12(7):605–8, jul 2015.

[193] Dattatreya Mellacheruvu, Zachary Wright, Amber L. Couzens, Jean Philippe Lambert, Nicole A. St-Denis, Tuo Li, Yana V. Miteva, Simon Hauri, Mihaela E. Sardiu, Teck Yew Low, Vincentius A. Halim, Richard D. Bagshaw, Nina C. Hubner, Abdallah Al-Hakim, Annie Bouchard, Denis Faubert, Damian Fermin, Wade H. Dunham, Marilyn Goudreault, Zhen Yuan Lin, Beatriz Gonzalez Badillo, Tony Pawson, Daniel Durocher, Benoit Coulombe, Ruedi Aebersold, Giulio Superti-Furga, Jacques Colinge, Albert J R Heck, Hyungwon Choi, Matthias Gstaiger, Shabaz Mohammed, Ileana M. Cristea, Keiryn L. Bennett, Mike P. Washburn, Brian Raught, Rob M. Ewing, Anne Claude Gingras, and Alexey I. Nesvizhskii. The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. *Nature Methods*, 10(8):730–736, 2013.

[194] Avinash K. Shanmugam and Alexey I. Nesvizhskii. Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics. *Journal of Proteome Research*, 14(12):5169–5178, 2015.

[195] Qiyao Li, Michael R. Shortreed, Craig D. Wenger, Brian L. Frey, Leah V. Schaffer, Mark Scalf, and Lloyd M. Smith. Global Post-Translational Modification Discovery. *Journal of Proteome Research*, page acs.jproteome.6b00034, mar 2017.

[196] Eric W. Deutsch, Sandra Orchard, Pierre Alain Binz, Wout Bittremieux, Martin Eisenacher, Henning Hermjakob, Shin Kawano, Henry Lam, Gerhard Mayer, Gerben Menschaert, Yasset Perez-Riverol, Reza M. Salek, David L. Tabb, Stefan Tenzer, Juan Antonio Vizcaíno, Mathias Walzer, and Andrew R. Jones. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *Journal of Proteome Research*, 16(12):4288–4298, 2017.

[197] Tommy Nilsson, Matthias Mann, Ruedi Aebersold, John R Yates, Amos Bairoch, and John J M Bergeron. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature methods*, 7(9):681–685, 2010.

[198] Robin Gras, Markus Müller, Elisabeth Gasteiger, Steven Gay, Pierre-Alain Binz, William Bienvenut, Christine Hoogland, Jean-Charles Sanchez, Amos Bairoch, Denis F. Hochstrasser, and Ron D. Appel. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, 20(18):3535–3550, dec 1999.

[199] David Goldberg, Marshall Bern, Simon J. North, Stuart M. Haslam, and Anne Dell. Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics (Oxford, England)*, 25(3):365–71, feb 2009.

[200] Bas C. Jansen, Karli R. Reiding, Albert Bondt, Agnes L. Hipgrave Ederveen, Magnus Palmblad, David Falck, and Manfred Wuhrer. MassyTools: A High-Throughput Targeted Data Processing Tool for Relative Quantitation and Quality Control Developed for Glycomic and Glycoproteomic MALDI-MS. *Journal of Proteome Research*, 14(12):5088–5098, dec 2015.

[201] Mark V. Ivanov, Irina A. Tarasova, Lev I. Levitsky, Elizaveta M. Solovyeva, Marina L. Pridatchenko, Anna A. Lobas, Julia A. Bubis, and Mikhail V. Gorshkov. MS/MS-Free Protein Identification in Complex Mixtures Using Multiple Enzymes with Complementary Specificity. *Journal of Proteome Research*, page acs.jproteome.7b00365, sep 2017.

[202] Ian Walsh, Terry Nguyen-Khuong, Katherine Wongtrakul-Kish, Shi Jie Tay, Daniel Chew, Tasha Jose, Christopher H. Taron, and Pauline M. Rudd. GlycanAnalyzer: Software for Automated Interpretation of N-Glycan Profiles after Exoglycosidase Digestions. *Bioinformatics*, (May):0–0, 2018.

[203] Yunli Hu, Tarek Shihab, Shiyue Zhou, Kerry Wooding, and Yehia Mechref. LC-MS/MS of permethylated N-glycans derived from model and human blood serum glycoproteins. *ELECTROPHORESIS*, 37(11):1498–1505, jun 2016.

[204] V Dancík, V Dancík, T a Addona, T a Addona, K R Clauser, K R Clauser, J E Vath, J E Vath, P a Pevzner, and P a Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology : a journal of computational molecular cell biology*, 6(3-4):327–42, 1999.

[205] Bin Ma. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of The American Society for Mass Spectrometry*, 26(11):1885–1894, nov 2015.

[206] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31):8247–8252, 2017.

[207] Hao Yang, Hao Chi, Wen Jing Zhou, Wen Feng Zeng, Kun He, Chao Liu, Rui Xiang Sun, and Si Min He. Open-pNovo: De Novo Peptide Sequencing with Thousands of Protein Modifications. *Journal of Proteome Research*, 16(2):645–654, 2017.

[208] Stephen Tanner, Hongjun Shu, Ari Frank, Ling-chi Wang, Ebrahim Zandi, Marc Mumby, Pavel a Pevzner, and Vineet Bafna. InsPecT : Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra is key to understanding various cellular regulatory pro- proved to be very successful in genomics searches . Given selects a small fraction of database D that is gu. *Analytical chemistry*, 77(14):4626–4639, 2005.

[209] Ari Frank, Stephen Tanner, Vineet Bafna, and Pavel Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *Journal of Proteome Research*, 4(4):1287–1295, 2005.

[210] Pengyu Hong, Hui Sun, Long Sha, Yi Pu, Kshitij Khatri, Xiang Yu, Yang Tang, and Cheng Lin. GlycoDeNovo – an Efficient Algorithm for Accurate de novo Glycan Topology Reconstruction from Tandem Mass Spectra. *Journal of the American Society for Mass Spectrometry*, 28(11):2288–2301, 2017.

[211] Han Hu, Yu Huang, Yang Mao, Xiang Yu, Yongmei Xu, Jian Liu, Chengli Zong, Geert-Jan Boons, Cheng Lin, Yu Xia, and Joseph Zaia. A Computational Framework for Heparan Sulfate Sequencing Using High-resolution Tandem Mass Spectra. *Molecular & Cellular Proteomics*, 13(9):2490–2502, 2014.

[212] Oliver Horlacher, Chunsheng Jin, Davide Alocci, Julien Mariethoz, Markus Müller, Niclas G. Karlsson, and Frédérique Lisacek. Glycoforest 1.0. *Analytical Chemistry*, page acs.analchem.7b02754, 2017.

[213] Michael Tiemeyer, Kazuhiro Aoki, James Paulson, Richard D Cummings, William S York, Niclas G Karlsson, Frederique Lisacek, Nicolle H Packer, Matthew P Campbell, Nobuyuki P Aoki, Akihiro Fujita, Masaaki Matsubara, Daisuke Shinmachi, Shinichiro Tsuchiya, Issaku Yamada, Michael Pierce, René Ranzinger, Hisashi Narimatsu, and Kiyoko F Aoki-Kinoshita. GlyTouCan: an accessible glycan structure repository. *Glycobiology*, 27(10):915–919, 2017.

[214] Matthew P Campbell, Robyn Peterson, Julien Mariethoz, Elisabeth Gasteiger, Yukie Akune, Kiyoko F Aoki-Kinoshita, Frederique Lisacek, and Nicolle H Packer. UniCarbKB: Building a knowledge platform for glycoproteomics. *Nucleic Acids Research*, 42(D1):D215–21, jan 2014.

[215] Roman A. Zubarev, Alexander R. Zubarev, and Mikhail M. Savitski. Electron Capture/Transfer versus Collisionally Activated/Induced Dissociations: Solo or Duet? *Journal of the American Society for Mass Spectrometry*, 19(6):753–761, 2008.

[216] Hao Yang, Hao Chi, Wen Jing Zhou, Wen Feng Zeng, Chao Liu, Rui Min Wang, Zhao Wei Wang, Xiu Nan Niu, Zhen Lin Chen, and Si Min He. PSite: Amino Acid Confidence Evaluation for Quality Control of de Novo Peptide Sequencing and Modification Site Localization. *Journal of Proteome Research*, 17(1):119–128, 2018.

[217] Ari M Frank, Nuno Bandeira, Zhouxin Shen, Stephen Tanner, Steven P Briggs, Richard D Smith, and Pavel A Pevzner. Clustering Millions of Tandem Mass Spectra research articles. *J. Proteome Research*, pages 113–122, 2008.

[218] Johannes Griss. Spectral library searching in proteomics, 2016.

[219] Shadi Toghi Eshghi, Punit Shah, Weiming Yang, Xingde Li, and Hui Zhang. GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides. *Analytical Chemistry*, 87(10):5181–5188, 2015.

[220] Ari M Frank, Matthew E Monroe, Anuj R Shah, Jeremy J Carver, Nuno Bandeira, Ronald J Moore, Gordon A Anderson, Richard D Smith, and Pavel A Pevzner. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature methods*, 8(7):587–91, may 2011.

[221] Michael R Hoopmann, Gregory L Finney, Michael J MacCoss, Michael R. Hoopmann, , Gregory L. Finney, Michael J. MacCoss*, Michael R Hoopmann, Gregory L Finney, and Michael J MacCoss. 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):5620–5632, aug 2007.

[222] Johan Teleman, Aakash Chawade, Marianne Sandin, Fredrik Levander, and Johan Malmström. Dinosaur: A Refined Open-Source Peptide MS Feature Detector. *Journal of Proteome Research*, 15(7):2143–2151, jul 2016.

[223] Xiaowen Liu, Yuval Inbar, Pieter C. Dorrestein, Colin Wynne, Nathan Edwards, Puneet Souda, Julian P. Whitelegge, Vineet Bafna, and Pavel A. Pevzner. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Molecular & cellular proteomics : MCP*, 9(12):2772–2782, sep 2010.

[224] Zuo-F ei Yuan, Chao Liu, Hai-Peng Wang, Rui-Xiang Sun, Yan Fu, Jing-Fen Zhang, Le-Heng Wang, Hao Chi, You Li, Li-Yun Xiu, Wen-Ping Wang, and Si-Min He. pParse: A method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *PROTEOMICS*, 12(2):226–235, jan 2012.

[225] Gang Liu and Sriram Neelamegham. Integration of systems glycobiology with bioinformatics toolboxes, glycoinformatics resources, and glycoproteomics data. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(4):163–181, 2015.

[226] Markus Pioch, Marcus Hoffmann, Alexander Pralow, Udo Reichl, and Erdmann Rapp. glyXtool MS : An Open-Source Pipeline for Semiautomated Analysis of Glycopeptide Mass Spectrometry Data. *Analytical Chemistry*, page acs.analchem.8b02087, 2018.

[227] Jürgen Kast, Marc Gentzel, Matthias Wilm, and Keith Richardson. Noise filtering techniques for electrospray quadrupole time of flight mass spectra. *Journal of the American Society for Mass Spectrometry*, 14(7):766–776, 2003.

[228] Jianqiu Zhang, Elias Gonzalez, Travis Hestilow, William Haskins, and Yufei Huang. Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry. *Current Genomics*, 10(6):388–401, 2009.

[229] Parminder Kaur and Peter B O'Connor. Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, 17(3):459–468, mar 2006.

[230] Martin Slawski, Rene Hussong, Andreas Tholey, Thomas Jakoby, Barbara Gregorius, Andreas Hildebrandt, and Matthias Hein. Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC bioinformatics*, 13(1):291, jan 2012.

[231] Piotr Dittwald, Trung Nghia Vu, Glenn A. Harris, Richard M. Caprioli, Raf Van de Plas, Kris Laukens, Anna Gambin, and Dirk Valkenborg. Towards automated discrimination of lipids versus peptides from full scan mass spectra. *EuPA Open Proteomics*, 4:87–100, 2014.

[232] Michael W. Senko, Steven C. Beu, and Fred W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995.

[233] Jorge Fernandez-De-Cossio Diaz and Jorge Fernandez-De-Cossio. Computation of isotopic peak center-mass distribution by fourier transform. *Analytical Chemistry*, 84(16):7052–7056, 2012.

[234] Han Hu, Piotr Dittwald, Joseph Zaia, and Dirk Valkenborg. Comment on "Computation of isotopic peak center-mass distribution by fourier transform". *Analytical chemistry*, 85(24):12189–92, dec 2013.

[235] Alan L. Rockwood and Steven L. Van Orden. Ultrahigh-speed calculation of isotope distributions. *Analytical Chemistry*, 68(13):2027–2030, 1996.

[236] Joshua Klein. mobiusklein/brainpy: Release v1.4.0, October 2018.

[237] Joseph Zaia, Kshitij Khatri, Joshua Klein, Chun Shao, Yuewei Sheng, and Rosa Viner. Complete Molecular Weight Profiling of Low-Molecular Weight Heparins Using Size Exclusion Chromatography-Ion Suppressor-High-Resolution Mass Spectrometry. *Analytical Chemistry*, 88(21):10654–10660, nov 2016.

[238] Andrea Argentini, Ludger J.E. Goeminne, Kenneth Verheggen, Niels Hulstaert, An Staes, Lieven Clement, and Lennart Martens. MoFF: A robust and automated approach to extract peptide ion intensities. *Nature Methods*, 13(12):964–966, 2016.

[239] Fatemeh Zamanzad Ghavidel, Jürgen Claesen, Tomasz Burzykowski, and Dirk Valkenborg. Comparison of the mahalanobis distance and pearson's $\chi 2$ statistic as measures of similarity of isotope patterns. *Journal of the American Society for Mass Spectrometry*, 25(2):293–296, 2014.

[240] David M. Horn, Roman A. Zubarev, and Fred W. McLafferty. Automated reduction and interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *Journal of the American Society for Mass Spectrometry*, 11(4):320–332, apr 2000.

[241] Zheng Yuan, Jinhong Shi, Wenjun Lin, Bolin Chen, and Fang-Xiang Xiang Wu. Features-based deisotoping method for tandem mass spectra. *Advances in Bioinformatics*, 2011:210805, jan 2011.

[242] Mateusz Krzysztof Łącki, Frederik Lermyte, Błażej Miasojedow, Mikołaj Olszański, Michał Startek, Frank Sobott, Dirk Valkenborg, and Anna Gambin. Assigning peaks and modeling ETD in top-down mass spectrometry. 2017.

[243] Kshitij Khatri, Joshua A. Klein, and Joseph Zaia. Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. *Analytical and Bioanalytical Chemistry*, 2016.

[244] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpp, Steffen Neumann, Angel D Pizarro, Luisa Montecchi-Palazzi, Natalie Tasman, Mike Coleman, Florian Reisinger, Puneet Souda, Henning Hermjakob, Pierre-alain Binz, and Eric W Deutsch. mzML–a community standard for mass spectrometry data. *Molecular & cellular proteomics : MCP*, 10(1):R110.000133, 2011.

[245] Joseph Zaia. Mass spectrometry and the emerging field of glycomics. *Chemistry & biology*, 15(9):881–92, sep 2008.

[246] Hannu Peltoniemi, Suvi Natunen, Ilja Ritamo, Leena Valmu, and Jarkko Räbinä. Novel data analysis tool for semiquantitative LC-MS-MS2 profiling of N-glycans. *Glycoconjugate journal*, 30(2):159–70, feb 2013.

[247] Martin Frank and Siegfried Schloissnig. Bioinformatics and molecular modeling in glycobiology. *Cellular and Molecular Life Sciences*, 67(16):2749–2772, aug 2010.

[248] Kiyoko Aoki-Kinoshita, Sanjay Agravat, Nobuyuki P. Aoki, Sena Arpinar, Richard D. Cummings, Akihiro Fujita, Noriaki Fujita, Gerald M. Hart, Stuart M. Haslam, Toshisuke Kawasaki, Masaaki Matsubara, Kelley W. Moreman, Shujiro Okuda, Michael Pierce, René Ranzinger, Toshihide Shikanai, Daisuke Shinmachi, Elena Solovieva, Yoshinori Suzuki, Shinichiro Tsuchiya, Issaku Yamada, William S. York, Joseph Zaia, and Hisashi Narimatsu. GlyTouCan 1.0 – The international glycan structure repository. *Nucleic Acids Research*, 44(D1):D1237–D1242, jan 2016.

[249] Frederick J. Krambeck and Michael J. Betenbaugh. A mathematical model of N-linked glycosylation. *Biotechnology and Bioengineering*, 92(6):711–728, 2005.

[250] Tianwei Yu and Hesen Peng. Quantification and deconvolution of asymmetric LC-MS peaks using the bi-Gaussian mixture model and statistical model selection. *BMC bioinformatics*, 11(1):559, 2010.

[251] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(2006):2399–2434, 2006.

[252] M J Spiro and R G Spiro. Sulfation of the N-linked oligosaccharides of influenza virus hemagglutinin: temporal relationships and localization of sulfotransferases. *Glycobiology*, 10(11):1235–42, 2000.

[253] Tomomi Ichimiya, Shoko Nishihara, Sayaka Takase-Yoden, Hiroshi Kida, and Kiyoko Aoki-Kinoshita. Frequent glycan structure mining of influenza virus data revealed a sulfated glycan motif that increased viral infection. *Bioinformatics*, 30(5):706–711, 2014.

[254] Yunli Hu and Yehia Mechref. Comparing MALDI-MS, RP-LC-MALDI-MS and RP-LC-ESI-MS glycomic profiles of permethylated N-glycans derived from model glycoproteins and human blood serum. *Electrophoresis*, 33(12):1768–1777, 2012.

[255] Jasminka Kristic and Gordan Lauc. Ubiquitous importance of protein glycosylation. In Kiyoko F Aoki-kinoshita, editor, *High Throughput Glycomics and Glycoproteomics*, chapter 1. Springer, 1 edition, 2017.

[256] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten. Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419, jan 2015.

[257] Han Hu, Kshitij Khatri, and Joseph Zaia. Algorithms and design strategies towards automated glycoproteomics analysis. *Mass Spectrometry Reviews*, 36(4):475–498, jul 2017.

[258] G Lauc and M Wuhrer. *High Throughput Glycomics and Glycoproteomics*. Number 1503. Springer, New York, NY, USA, 1 edition, 2017.

[259] Joseph Zaia. Mass spectrometry and glycomics. *Omics : a journal of integrative biology*, 14(4):401–18, aug 2010.

[260] Ruedi Aebersold, Jeffrey N Agar, I Jonathan Amster, Mark S Baker, Carolyn R Bertozzi, Emily S Boja, Catherine E Costello, Benjamin F Cravatt, Catherine Fenselau, Benjamin A Garcia, Ying Ge, Jeremy Gunawardena, Ronald C Hendrickson, Paul J Hergenrother, Christian G Huber, Alexander R Ivanov, Ole N Jensen, Michael C Jewett, Neil L Kelleher, Laura L Kiessling, Nevan J Krogan, Martin R Larsen, Joseph A Loo, Rachel R Ogorzalek Loo, Emma Lundberg, Michael J MacCoss, Parag Mallick, Vamsi K Mootha, Milan Mrksich, Tom W Muir, Steven M Patrie, James J Pesavento, Sharon J Pitteri, Henry Rodriguez, Alan Saghatelian, Wendy Sandoval, Hartmut Schlüter, Salvatore Sechi, Sarah A Slavoff, Lloyd M Smith, Michael P Snyder, Paul M Thomas, Mathias

Uhlén, Jennifer E Van Eyk, Marc Vidal, David R Walt, Forest M White, Evan R Williams, Therese Wohlschlager, Vicki H Wysocki, Nathan A Yates, Nicolas L Young, and Bing Zhang. How many human proteoforms are there? *Nature Chemical Biology*, 14(3):206–214, 2018.

[261] Tatjana Sajic, Yansheng Liu, Eirini Arvaniti, Silvia Surinova, Evan G. Williams, Ralph Schiess, Ruth Hüttenhain, Atul Sethi, Sheng Pan, Teresa A. Brentnall, Ru Chen, Peter Blattmann, Betty Friedrich, Emma Niméus, Susanne Malander, Aurelius Omlin, Silke Gillessen, Manfred Claassen, and Ruedi Aebersold. Similarities and Differences of Blood N-Glycoproteins in Five Solid Carcinomas at Localized Clinical Stage Analyzed by SWATH-MS. *Cell Reports*, pages 2819–2831, 2018.

[262] Cheng Ma, Jingyao Qu, Jeffrey Meisner, Xinyuan Zhao, Xu Li, Zhigang Wu, Hailiang Zhu, Zaikuan Yu, Lei Li, Yuxi Guo, Jing Song, and Peng George Wang. Convenient and Precise Strategy for Mapping N-Glycosylation Sites Using Microwave-Assisted Acid Hydrolysis and Characteristic Ions Recognition. *Analytical chemistry*, 87(15):7833–9, aug 2015.

[263] Ling Y. Lee, Edward S.X. Moh, Benjamin L. Parker, Marshall Bern, Nicolle H. Packer, and Morten Thaysen-Andersen. Toward Automated N-Glycopeptide Identification in Glycoproteomics. *Journal of Proteome Research*, 15(10):3904–3915, 2016.

[264] You Li, Hao Chi, Le-Heng Wang, Hai-Peng Wang, Yan Fu, Zuo-Fei Yuan, Su-Jun Li, Yan-Sheng Liu, Rui-Xiang Sun, Rong Zeng, and Si-Min He. Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing. *Rapid Communications in Mass Spectrometry*, 24(6):807–814, mar 2010.

[265] Weiping Sun, Yi Liu, Gilles Lajoie, Bin Ma, and Kaizhong Zhang. An Improved Approach for N-linked Glycan Structure Identification from HCD MS/MS Spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5963(c):1–1, 2016.

[266] Jürgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research*, 10(4):1794–1805, apr 2011.

[267] Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, Jennifer Siepen, Simon J Hubbard, Julian N Selley, Brian C Searle, James Shofstahl, Sean L Seymour, Randall Julian, Pierre-Alain Binz, Eric W Deutsch, Henning Hermjakob, Florian Reisinger, Johannes Griss, Juan Antonio Vizcaíno, Matthew Chambers, Angel Pizarro, and David Creasy. The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular & cellular proteomics : MCP*, 11(7):M111.014381, jul 2012.

[268] René Ranzinger, Stephan Herget, Claus-Wilhelm C.-W. von der Lieth, and Martin Frank. GlycomeDB–a unified database for carbohydrate structures. *Nucleic acids research*, 39(Database issue):D373–6, jan 2011.

[269] Davide Alocci, Marie Ghraichy, Elena Barletta, Alessandra Gastaldello, Julien Mariethoz, and Frederique Lisacek. Understanding the glycome: an interactive view of glycosylation from glycocompositions to glycoepitopes. *Glycobiology*, (March 2018), mar 2018.

[270] Oliver Serang, John W. Froehlich, Jan Muntel, Gary McDowell, Hanno Steen, Richard S. Lee, and Judith A. Steen. SweetSEQer, Simple <i>de Novo</i> Filtering and Annotation of Glycoconjugate Mass Spectra. *Molecular & Cellular Proteomics*, 12(6):1735–1740, 2013.

[271] The SQLite3 Development Team. SQLite3, 2000–. [Online; accessed 11-03-2018].

[272] Yi-Min She, Aaron Farnsworth, Xuguang Li, and Terry D. Cyr. Topological N-glycosylation and site-specific N-glycan sulfation of influenza proteins in the highly expressed H1N1 candidate vaccines. *Scientific Reports*, 7(1):10232, dec 2017.

[273] Cheng Ma, Jingyao Qu, Xu Li, Xinyuan Zhao, Lei Li, Cong Xiao, Garrett Edmunds, Ebtesam Gashash, Jing Song, and Peng George Wang. Improvement of core-fucosylated glycoproteome coverage via alternating HCD and ETD fragmentation. *Journal of Proteomics*, 146(2):90–98, sep 2016.

[274] Forouzan Aboufazeli and Eric D. Dodds. Precursor Ion Survival Energies of Protonated N-Glycopeptides and their Weak Dependencies on High Mannose N-Glycan Composition in Collision-Induced Dissociation. *The Analyst*, pages 4459–4468, 2018.

[275] Venkata Kolli, Heidi A. Roth, Gabriela De La Cruz, Ganga S. Fernando, and Eric D. Dodds. The role of proton mobility in determining the energy-resolved vibrational activation/dissociation channels of N-glycopeptide ions. *Analytica Chimica Acta*, 896:85–92, 2015.

[276] Hong Yang, Chenxi Yang, and Taolei Sun. Glycopeptides Characterization using a Stepped HCD Energy Approach on Hybrid Quadrupole Orbitrap. *Rapid Communications in Mass Spectrometry*, jun 2018.

[277] Ravi Chand Bollineni, Christian Jeffrey Koehler, Randi Elin Gislefoss, Jan Haug Anonsen, and Bernd Thiede. Large-scale intact glycopeptide identification by Mascot database search. *Scientific Reports*, 8(1):2117, 2018.

[278] Marina Spivak, Michael S. Bereman, Michael J. MacCoss, and William Stafford Noble. Learning score function parameters for improved spectrum identification in tandem mass spectrometry experiments. *Journal of Proteome Research*, 11(9):4499–4508, sep 2012.

[279] Ari M Frank. A ranking-based scoring function for peptide-spectrum matches. *Journal of proteome research*, 8(5):2241–52, may 2009.

[280] Sheila J. Barton, Sylvia Richardson, David N. Perkins, Inga Bellahn, Trevor N. Bryant, and John C. Whittaker. Using statistical models to identify factors that have a role in defining the abundance of ions produced by tandem MS. *Analytical Chemistry*, 79(15):5601–5607, 2007.

[281] Shiwei Sun, Fuquan Yang, Qing Yang, Hong Zhang, Yaojun Wang, Dongbo Bu, and Bin Ma. MS-simulator: Predicting y-ion intensities for peptides with two charges based on the intensity ratio of neighboring ions. *Journal of Proteome Research*, 11(9):4509–4516, sep 2012.

[282] Sven Degroeve, Davy Maddelein, and Lennart Martens. MS2PIP prediction server: Compute and visualize MS2peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research*, 43(W1):W326–W330, 2015.

[283] Vicki H. Wysocki, George Tsaprailis, Lori L. Smith, and Linda A. Breci. Mobile and localized protons: A framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, 35(12):1399–1406, 2000.

[284] Béla Palzs and Sándor Suhal. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*, 24(4):508–548, 2005.

[285] Frederik Lermyte, Dirk Valkenborg, Joseph A. Loo, and Frank Sobott. Radical solutions: Principles and application of electron-based dissociation in mass spectrometry-based analysis of protein structure. *Mass Spectrometry Reviews*, (January):750–771, 2018.

[286] Meena Choi, Ching Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, and Olga Vitek. MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524–2526, 2014.

[287] Jocelyn F Krey, Phillip A Wilmarth, Jung-bum Shin, John Klimek, Nicholas E. Sherman, Erin D Jeffery, Dongseok Choi, Larry L David, and Peter G. Barr-Gillespie. Accurate Label-Free Protein Quantitation with High- and Low-Resolution Mass Spectrometers. *Journal of Proteome Research*, 13(2):1034–1044, feb 2014.

[288] Mhd H.D.Rami Al Shweiki, Susann Mönchgesang, Petra Majovsky, Domenika Thieme, Diana Trutschel, and Wolfgang Hoehenwarter. Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *Journal of Proteome Research*, 16(4):1410–1424, 2017.

[289] David Shteynberg, Eric W Deutsch, Henry Lam, Jimmy K Eng, Zhi Sun, Natalie Tasman, Luis Mendoza, Robert L Moritz, Ruedi Aebersold, and Alexey I Nesvizhskii. iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Molecular & Cellular Proteomics*, 10(12):M111.007690, dec 2011.

[290] Pedro Alves, Randy J. Arnold, David E. Clemmer, Yixue Li, James P. Reilly, Quanhu Sheng, Haixu Tang, Zhiyin Xun, Rong Zeng, and Predrag Radivojac. Fast and accurate identification of semi-tryptic peptides in shotgun proteomics. *Bioinformatics*, 24(1):102–109, 2008.

[291] Irina A. Tarasova, Anton A. Goloborodko, Tatyana Y. Perlova, Marina L. Pridatchenko, Alexander V. Gorshkov, Victor V. Evreinov, Alexander R. Ivanov, and Mikhail V. Gorshkov. Application of Statistical Thermodynamics To Predict the Adsorption Properties of Polypeptides in Reversed-Phase HPLC. *Analytical Chemistry*, 87(13):6562–6569, jul 2015.

[292] Heydar Maboudi Afkham, Xuanbin Qiu, Matthew The, and K Lukas. Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics. 33(October 2016):508–513, 2017.

[293] Sven H. Giese, Yasushi Ishihama, and Juri Rappsilber. Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues. *Analytical Chemistry*, 90(7):4635–4640, 2018.

[294] Petr Kozlik, Radoslav Goldman, and Miloslav Sanda. Study of structure-dependent chromatographic behavior of glycopeptides using reversed phase nanoLC. *Electrophoresis*, 38(17):2193–2199, 2017.

[295] Benlian Wang, Yaroslav Tsybovsky, Krzysztof Palczewski, and Mark R Chance. Reliable determination of site-specific in vivo protein N-glycosylation based on collision-induced MS/MS and chromatographic retention time. *Journal of the American Society for Mass Spectrometry*, 25(5):729–41, may 2014.

[296] Vittoria Matafora, Andrea Corno, Andrea Ciliberto, and Angela Bachi. Missing Value Monitoring Enhances the Robustness in Proteomics Quantitation. *Journal of Proteome Research*, 16(4):1719–1727, 2017.

[297] Wei Wang, Andrew C Sue, and Wilson W B Goh. Feature selection in clinical proteomics : with great power comes great reproducibility. *Drug Discovery Today*, 00(00), 2017.

[298] Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *Journal of proteome research*, 15(4):1116–25, apr 2016.

[299] Zhuxuan Jin, Jian Kang, and Tianwei Yu. Missing value imputation for LC-MS metabolomics data by incorporating metabolic network and adduct ion relations. *Bioinformatics*, 34(December 2017):1–7, 2018.

[300] T P Conrads, G A Anderson, T D Veenstra, L Pasa-Tolić, and R D Smith. Utility of accurate mass tags for proteome-wide protein identification. *Analytical chemistry*, 72(14):3349–54, jul 2000.

[301] Robert J. Millikin, Stefan K. Solntsev, Michael R. Shortreed, and Lloyd M. Smith. Ultrafast Peptide Label-Free Quantification with FlashLFQ. *Journal of Proteome Research*, 17(1):386–391, 2018.

[302] Mark V Ivanov, Lev I Levitsky, Anna A Lobas, Tanja Panic, Ünige A Laskay, Goran Mitulovic, Rainer Schmid, Marina L Pridatchenko, Yury O Tsybin, and Mikhail V Gorshkov. Empirical multidimensional space for scoring peptide spectrum matches in shotgun proteomics. *Journal of proteome research*, 13(4):1911–20, apr 2014.

[303] Yanyan Qu, Liangliang Sun, Zhenbin Zhang, and Norman J Dovichi. Site-Specific Glycan Heterogeneity Characterization by Hydrophilic Interaction Liquid Chromatography Solid-Phase Extraction, Reversed-Phase Liquid Chromatography Fractionation, and Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectro. *Analytical Chemistry*, 90(2):1223–1233, jan 2018.

[304] Yingwei Hu, Punit Shah, David J. Clark, Minghui Ao, and Hui Zhang. Reanalysis of Global Proteomic and Phosphoproteomic Data Identified a Large Number of Glycopeptides. *Analytical Chemistry*, 90(13):8065–8071, 2018.

[305] Anoop Mayampurath, Ehwang Song, Abhinav Mathur, Chuan-Yih Yu, Zane Hammoud, Yehia Mechref, and Haixu Tang. Label-free glycopeptide quantification for biomarker discovery in human sera. *Journal of proteome research*, 13(11):4821–32, nov 2014.

Chapter 8

Vita

[redacted]

[redacted]

[redacted]