Comparative Analysis of Glycoproteomic Software Using a Tailored Glycan Database

by
Reuben Aaron Hogan

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Nevan Krogan
071F8FAE3EA8493...

Nevan Krogan
Chair

DocuSigned by:
Martin Kampmann
[illegible]4AF...

Martin Kampmann

DocuSigned by:
Robert Chalkley
[illegible]46A...

Robert Chalkley

DocuSigned by:
Michael Grabe
[illegible]414...

Michael Grabe

DocuSigned by:
Ken Nakamura
01DE401715D648A...

Ken Nakamura
Committee Members

ACKNOWLEDGEMENTS

There are many people to give thanks to for this work. I will start with those who offered help outside the science. First, I would like to thank God for seeing me through this experience because there were many times that I could not fathom it. Next, I would like to thank my mother, father, sisters, and brother-in-law for all their support while I endured. I am the first person in my family to obtain a PhD, and I know I worried them at times with my intense focus. I have to thank every teacher that has taught me anything. You all were helping mold me into the person I am now. I want to thank administrators such as Nicole Flowers and Tiffani Quan. Nicole was a regular source of support and her administrative experience saved me on multiple occasions. Tiffani was similarly helpful in giving me the realist perspective that I needed to get my work done. I would especially like to thank D'Anne Duncan, PhD, whose office I sat in on more than one occasion to bemoan the pains of graduate school.

For the science, I would like to thank the members of my thesis committee for their time and input. My process through this PhD was untraditional, and I am appreciative that they were there when I needed them most. In particular, I would like to thank Robert Chalkley, PhD, for offering resources and support that was instrumental for the namesake work of this dissertation. I would like to thank Danielle Swaney, PhD. The only reason that I was able to touch a mass spectrometer was because she allowed it, and that single decision changed the course of my entire PhD. I would like to thank Monita Muralidharan, PhD, for her encouragement to study unusual things. When I first stated an interest in glycoproteomics, she sent me a 50-page-long review that she had found. I would like to thank Justin McKetney, PhD, who answered my questions about mass spectrometry and ultimately introduced me to Nicholas Riley, PhD, whom I would also like to thank and who helped me establish a glycoproteomic workflow at UCSF. I would like to thank

Lauren Pepi, PhD, for her collaboration and execution of the glycomic profiling done in this work. I would like to thank Kelsey Haas, PhD, and Snow Naing, PhD, for their support when things were rough and for the check-ins. I would like to thank Zeynep Cakir, PhD, for her start to the APOE project that paved the way for me to end up with a project like this. I would like to thank Erica Stevenson, who helped complete the anti-biotin antibody purifications. I would like to thank Robert Raffai, PhD, for a rotation that solidified my interest in (glyco)lipoproteins. I would like to thank Aashish Manglik, MD, PhD, who gave me advice about what really mattered during my PhD.

To those whose names I have not mentioned but who played any part in this work, I thank you as well. Finally, I want to thank myself. I do not mean this in arrogance. I hope the reader will understand this. I am appreciative for the way that I, at times, willed myself through the failure and disappointment and pain of different events during my PhD. I never lost sight of the end, I viewed every failure as an opportunity, and I always remained CURIOUS. I thank myself for that.

“Our deepest fear is not that we are inadequate. Our deepest fear is that we are powerful beyond measure. It is our light, not our darkness that most frightens us.”

-Marianne Williamson

Comparative Analysis of Glycoproteomic Software Using a Tailored Glycan Database

Reuben Aaron Hogan

ABSTRACT

Glycoproteomics is a rapidly developing field, and data analysis has been stimulated by several technological innovations. As a result, there are many software tools from which to choose; and each comes with unique features that can be difficult to compare. This work presents a head-to-head comparison of five modern analytical software: Byonic, Protein Prospector, MSFraggerGlyco, pGlyco3, and GlycoDecipher. To enable a meaningful comparison, parameter variables were minimized. One potential confounding variable is the glycan database that informs glycoproteomic searches. We performed glycomic profiling of the samples and used the output to construct matched glycan databases for each software. Up to 19,000 glycopeptide spectra were identified across three replicates of wild-type SH-SY5Y cells. There was substantial overlap among most software for glycoproteins identified, locations of glycosites, and glycans, although Byonic reported a suspiciously large number of glycoproteins and glycosites of questionable reliability. We show that Protein Prospector identified the most glycopeptide spectrum matches with high agreement to known glycosites in UniProt. Overall, our results indicate that glycoproteomic searches should involve more than one software to generate confidence. It may be useful to consider software with peptide-first approaches and with glycan-first approaches.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: ALZHEIMER'S, APOE, AND GLYCOSYLATION

Introduction

Alzheimer's Disease (AD) is a neurodegenerative condition that is currently the 5th leading cause of death for individuals aged 65 years and older.[1] In 2022, it was estimated that 6.5 million individuals were living with AD.[2] Prevalence is expected to rise in proportion to the median age in the United States if novel treatments are not developed.[3] By 2060, it is projected that 13.8 million people will be living with AD.[2,3] The sheer burden of cases will place stress on the United States in a number of ways. First, medical care will need to be handled by geriatricians, of which there is currently a shortage.[3,4] Second, other care will be displaced upon the family. It is estimated that $271.6 billion worth of care is provided by caregivers, who are usually family members.[3,5] Third, medical costs due to AD for the country will total $321 billion in 2022.[3,5] Fourth, AD disproportionately affects individuals who identify as Black or Latino.[6–8] AD is a matter of urgent concern.

Medical treatment for AD has been incomplete. Before 2021, the treatment strategy was to mitigate symptoms of the disease.[3] With the approval of aducanumab and recently lecanemab, new options have become available. Unfortunately, results of clinical trials for aducanumab have been mixed, and lecanemab only delays disease progression by 27%.[9,10] This is true even though both monoclonal antibodies successfully clear amyloid beta plaques, which have long been considered one of the two defining histopathological signs of this disease. We need a better understanding of the molecular pathogenesis of AD in order to design better treatments.

APOE has emerged as one of the strongest genetic risk factors for AD.[11–15] There are three allelic variants of this gene that are abundant in the population: ε2, ε3, and ε4. Each carries a differential risk for AD. The ε4 allele has been shown to have a dose-dependent relationship with both the likelihood of developing AD and the age of onset.[11,16,17] Individuals who have one copy of the ε4 allele have a threefold increased risk of developing the condition; individuals who have two copies have a 12-fold increased risk.[16,18–20] From a different angle, it is estimated that up to 60% of individuals with AD have at least one copy of ε4.[17,21,22] These numbers are striking, yet how ε4 influences this disease is unknown. Moreover, ε2 is slightly protective against AD, and ε3 is the common "null" allelic variant in the population.[11,16,23] That each allele has a differential effect on the disease suggests that the ApoE protein might be involved in the pathogenesis of AD. Therefore, it is necessary to understand ApoE biology.

Interestingly, each APOE allele only results in one to two amino acid changes in the protein.[24] The ε4 is the ancestral allele and encodes arginines at sites 112 and 158. The ε3 encodes an arginine at site 112 but substitutes the arginine at 158 for a cysteine. The ε2 encodes for cysteines at both sites 112 and 158 and is thought to represent the most recently evolved variant. Studies have indicated that these single amino acid substitutions lead to differential effects in how the ApoE protein sequesters lipids[25], how dynamically it behaves[26,27], how tightly it binds to receptors[28,29], how it affects mitochondrial biology[30–32], and how it affects metabolic biology[33,34]. Proteomic analysis in the brain of AD patients complemented with transcriptomics analysis also points to an effect at the protein level, not the RNA level.[35] Therefore, how different APOE alleles modulate the biology is also important.

ApoE biology is quite complex. This is for two reasons. First, ApoE in the brain has multiple sources. Multiple reports have discovered that ApoE is produced by microglia[36,37], astrocytes[25,30,34,36–39], and neurons[39–41]. Second, ApoE has multiple sinks in the brain. It has been reported that ApoE can be uptaken by all the cell types listed above with the addition of oligodendrocytes.[42–44] It is reasonable to question which source, sink, or combination thereof is the important axis along which ApoE mediates its effects on AD.[38,44,44,45] It has been reported in the past and reaffirmed recently that ApoE can have a neurotoxic effect at the level of the neuron.[46,47] Considering that AD is a neurodegenerative condition, it merits consideration of ApoE at the level of the neuron.

Even when narrowed down to the cell type, there is still not consensus on which protein-protein interactions (PPIs) mediate the process of ApoE secretion or uptake. This motivated our lab to apply proteomics by mass spectrometry to generate PPI networks for ApoE. To do this, we fused the proximity labeling enzyme APEX to the three common alleles of APOE. We then performed affinity purification-mass spectrometry (AP-MS) to identify interacting and nearby proteins to ApoE.

Results

Expression of APOE Proximity Labeling Constructs

To generate PPI networks of ApoE, different APOE alleles were fused N- and C-terminally to the proximity labeler APEX2. APEX2 is a peroxidase that catalyzes the formation of radicals from biotin phenol in the presence of hydrogen peroxide.[48,49] Radicals are short-lived and will covalently bind to electron-rich amino acids in proteins within a 20 nm radius. Labeled proteins can then be purified from the cell lysate for mass spectrometry. In all plasmids, the APEX-APOE fusion was placed downstream of a doxycycline-inducible promoter and contained a FLAG tag directly preceding APEX2. N2a cells, a mouse neuroblastoma cell line, were transfected with one of the APOE constructs and treated with doxycycline for 24 hours. Cells were harvested at 1 day, 3 days, and 5 days after treatment. **Figure 1.1** indicates that N-terminally fused constructs of APOE had better expression than C-terminally fused APOE as shown by both an anti-FLAG antibody and an anti-ApoE antibody.

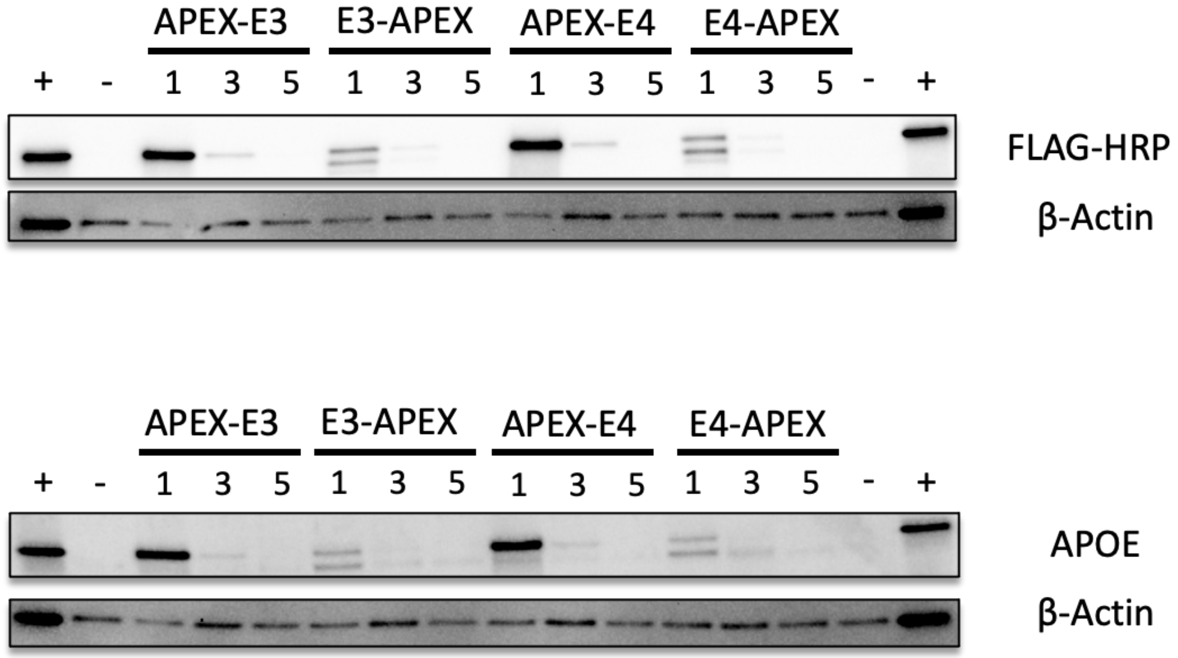


Figure 1.1: Western Blot Confirms Expression of APEX-APOE Constructs. 10µg of protein was loaded into each lane. Positive control is previously validated APEX-APOE protein. Negative control is untransfected N2a lysate.

APOE PPI Networks Are Enriched in Glycosylation-Related Proteins

Having confirmed that N-terminally fused APEX-APOE constructs expressed in cells, we then proceeded to perform proximity labeling. For this, we switched to using SH-SY5Y cells, a human neuroblastoma cell line. N2a cells grow quickly and will change in morphology if left unpassaged for more than two days. Conversely, SH-SY5Y cells grow more slowly and will reach confluence within a week. The time to passage for SH-SY5Y cells allows one more leisure to plan experiments before tissue culture. Moreover, given that both options are neuroblastoma cells, it was preferable to use human cells over mouse to study a human disease. For proximity labeling, SH-SY5Y cells were plated into five 15cm dishes for each APOE allele. All dishes were transfected with the respective APOE allele. Three of the five dishes were incubated with biotin phenol for one hour before adding hydrogen peroxide to initiate proximity labeling. The remaining two dishes were incubated with fresh complete media before addition of hydrogen peroxide and served as negative controls. APEX labeling was carried out for 1 minute before quenching the reaction and harvesting cells. Cell pellets were lysed. Biotinylated proteins were purified with streptavidin coated beads. Proteins were reduced, alkylated, and then digested using trypsin. Digested peptides were collected and prepared for analysis by mass spectrometry. Data was analyzed using MaxQuant.[50] PPIs were scored using SAINTexpress.[51] Proteins were considered PPIs if they were scored by SAINTexpress both when using spectral counts and when using intensities. Proteins were entered into STRING to generate PPI networks. **Figure 1.2** displays the networks by APOE allele.

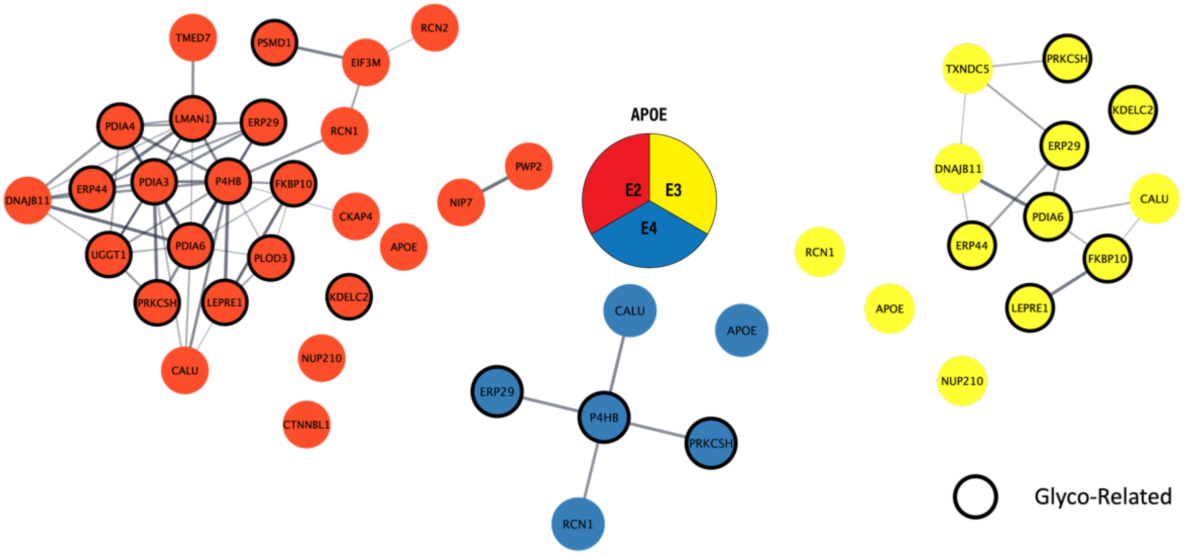
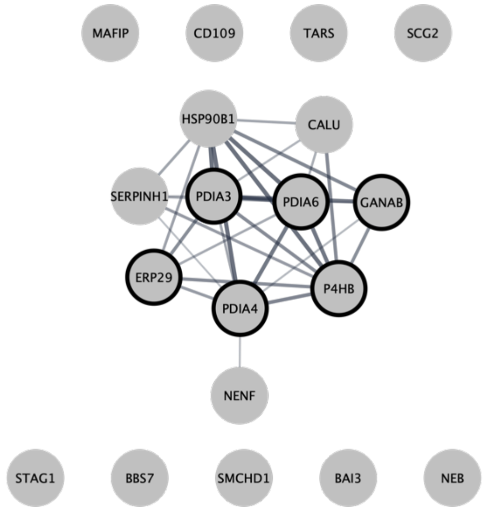
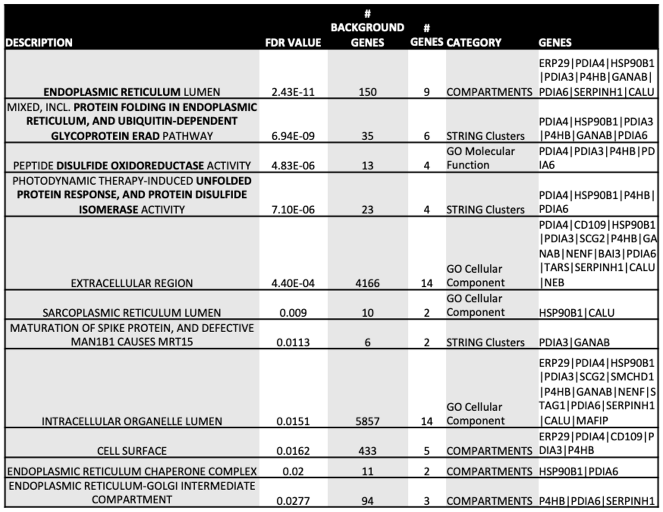


Figure 1.2: Glycosylation accounts for at least 50% of PPIs regardless of APOE allele. Proximity labeling with APEX N-terminally fused to APOE was used to biotinylate proximal or interacting proteins for AP-MS. Biotinylated proteins were purified with Streptavidin and digested. Peptides were used for MS. Identified proteins are showed as nodes. Colors denote the fused APOE allele (Red = APOE2, Yellow = APOE3, Blue = APOE4). Edges were produced by STRING database. Manual annotation was performed to indicate proteins with recorded relationships to glycosylation. This information was taken from *Essentials of Glycobiology* or details provided in STRING.

The number of detected PPIs differed by APOE allele. The ε2 allele recovered the most PPIs followed by ε3 and then ε4. Interestingly, we observed that many of these proteins were mentioned by name or by function in Chapter 39: "Glycans in Glycoprotein Quality Control" of the *Essentials of Glycobiology* textbook.[52] The network designates these proteins with a bold outline. We discovered that at least 50% of the protein interactions, regardless of APOE allele, were related to glycosylation in this way.

Streptavidin purification is useful for purifying biotinylated proteins but is difficult for validating that every protein identified was biotinylated. Streptavidin has such a high binding affinity for biotin such that the binding event is irreversible. Therefore, it is challenging to recover direct proof of biotinylation by identifying the biotinylated peptide itself with mass spectrometry. Therefore, we performed an alternate purification of the same biological samples using an anti-biotin antibody.[53] Anti-biotin antibodies can be denatured, which releases the biotinylated peptide for detection. With this workflow, we reanalyzed the samples by mass spectrometry. MaxQuant was used to search for the biotinylated peptides.





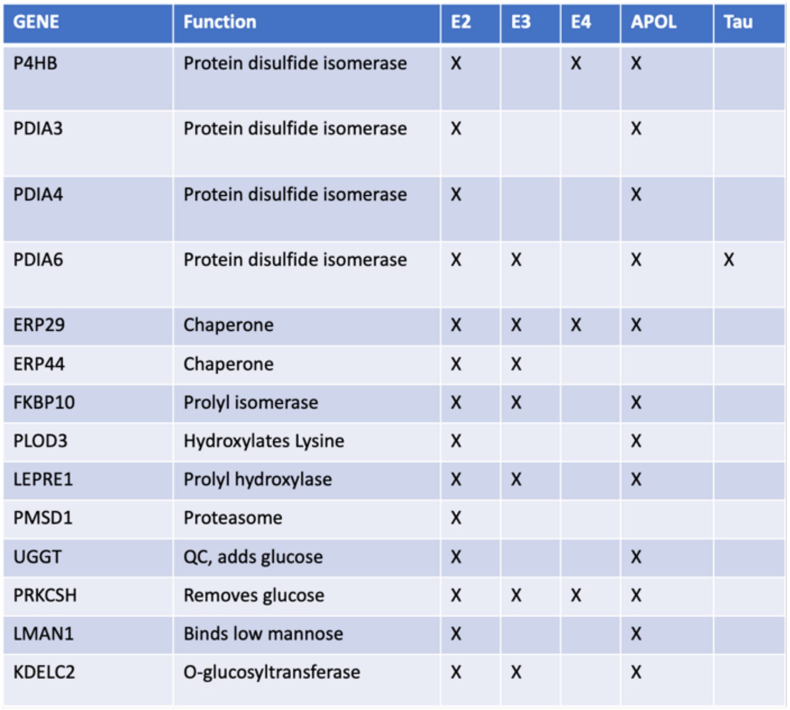
DESCRIPTION	FDR VALUE	# BACKGROUND GENES	# GENES	CATEGORY	GENES
ENDOPLASMIC RETICULUM LUMEN	2.43E-11	150	9	COMPARTMENTS	ERP29|PDIA4|HSP90B1|PDIA3|P4HB|GANAB|PDIA6|SERPINH1|CALU
MIXED, INCL. **PROTEIN FOLDING IN ENDOPLASMIC RETICULUM, AND UBIQUITIN-DEPENDENT GLYCOPROTEIN ERAD** PATHWAY	6.94E-09	35	6	STRING Clusters	PDIA4|HSP90B1|PDIA3|P4HB|GANAB|PDIA6
PEPTIDE **DISULFIDE OXIDOREDUCTASE** ACTIVITY	4.83E-06	13	4	GO Molecular Function	PDIA4|PDIA3|P4HB|PDIA6
PHOTODYNAMIC THERAPY-INDUCED **UNFOLDED PROTEIN RESPONSE, AND PROTEIN DISULFIDE ISOMERASE** ACTIVITY	7.10E-06	23	4	STRING Clusters	PDIA4|HSP90B1|P4HB|PDIA6
EXTRACELLULAR REGION	4.40E-04	4166	14	GO Cellular Component	PDIA4|CD109|HSP90B1|PDIA3|SCG2|P4HB|GANAB|NENF|BAI3|PDIA6|TARS|SERPINH1|CALU|NEB
SARCOPLASMIC RETICULUM LUMEN	0.009	10	2	GO Cellular Component	HSP90B1|CALU
MATURATION OF SPIKE PROTEIN, AND DEFECTIVE MAN1B1 CAUSES MRT15	0.0113	6	2	STRING Clusters	PDIA3|GANAB
INTRACELLULAR ORGANELLE LUMEN	0.0151	5857	14	GO Cellular Component	ERP29|PDIA4|HSP90B1|PDIA3|SCG2|SMCHD1|P4HB|GANAB|NENF|STAG1|PDIA6|SERPINH1|CALU|MAFIP
CELL SURFACE	0.0162	433	5	COMPARTMENTS	ERP29|PDIA4|CD109|PDIA3|P4HB
ENDOPLASMIC RETICULUM CHAPERONE COMPLEX	0.02	11	2	COMPARTMENTS	HSP90B1|PDIA6
ENDOPLASMIC RETICULUM-GOLGI INTERMEDIATE COMPARTMENT	0.0277	94	3	COMPARTMENTS	P4HB|PDIA6|SERPINH1

Figure 1.3: Glycosylation Remains A Significant Theme of Anti-Biotin Antibody Workflow. Anti-biotin purification of biotinylated peptides was used on the same samples collected in Figure 1.2. Biotinylated peptides were searched using MaxQuant. Table on the left displays results from functional enrichment of the network on the left.

Figure 1.3 shows the results from each allele. Using the anti-biotin antibody, we discovered that there were no differences amongst the alleles in regards to biotinylated peptides. Importantly, there were several proteins that overlapped with the networks shown in **Figure 2**. Again, manual annotation revealed that there were several proteins mentioned by name or function in the *Essentials of Glycobiology* textbook. Functional enrichment of these proteins also revealed that the "Protein Folding in Endoplasmic Reticulum and Ubiquitin-Dependent Glycoprotein ERAD" Pathway was one of the enriched themes in this network. Together, these data suggest that at least a subset of the ApoE PPIs are centered in a glycosylation pathway.

It is possible that these results could be an artefact of the expression system that we used. Therefore, we compared the results from the ε2, ε3, and ε4 alleles with APEX datasets collected in the Krogan Lab from APOL and MAPT. APOL is another apolipoprotein like APOE while MAPT is an intracellular cytoskeletal protein. MAPT, commonly referred to as Tau, is related to one of the hallmark pathologies of AD. **Table 1.1** demonstrates that many of the glycosylation-related PPIs are common to both apolipoproteins while only one is shared with Tau. These data underscore that glycosylation seems to be a theme common to apolipoprotein biology but not to any given APEX-fused protein.

Table 1.1: Tabular Overlap of Protein-Protein Interactions from APOE, APOL, and Tau. PPI networks were all collected using N-terminal fused APEX. Subset of proteins with glycosylation ties were used to show overlap across different proteins.



GENE	Function	E2	E3	E4	APOL	Tau
P4HB	Protein disulfide isomerase	X		X	X	
PDIA3	Protein disulfide isomerase	X			X	
PDIA4	Protein disulfide isomerase	X			X	
PDIA6	Protein disulfide isomerase	X	X		X	X
ERP29	Chaperone	X	X	X	X	
ERP44	Chaperone	X	X			
FKBP10	Prolyl isomerase	X	X		X	
PLOD3	Hydroxylates Lysine	X			X	
LEPRE1	Prolyl hydroxylase	X	X		X	
PMSD1	Proteasome	X				
UGGT	QC, adds glucose	X			X	
PRKCSH	Removes glucose	X	X	X	X	
LMAN1	Binds low mannose	X			X	
KDELC2	O-glucosyltransferase	X	X		X	

Glycosylation as a Unifying Link Between AD and APOE

A large-scale proteomics analysis of AD post-mortem brains identified that “Glycosylation/ER” was among the top 3 themes for AD-associated proteins with differential abundance between AD and controls.[35] Of note, increased abundance of Glycosylation/ER proteins was associated with better cognitive performance.[35] To further illustrate the connection between AD and glycosylation, many proteins implicated with AD pathogenesis are themselves glycosylated with evidence for AD-specific N- and O-glycosylation sites (glycosites).[54,55] For example, Tau protein, which forms hyperphosphorylated aggregates in AD, exclusively develops a N-glycosite in AD brains.[55–57] Tau is traditionally a cytosolic protein; but formation of an N-glycosite would enable Tau secretion, which has been reported multiple times in AD patients.[58–60] The existence of AD-specific glycosites suggests that many AD-associated proteins are susceptible to changes in glycosylation.

ApoE is a known metabolic protein, primarily for its role in lipid trafficking; but there is evidence that APOE also has allele-specific effects on glucose metabolism.[34,61–64] Individuals with the ε4 allele have blunted glucose uptake by PET scan in similar regions of the brain to suspected AD patients, even at a young age.[61–63] ApoE4 protein retains the insulin receptor in the endosomes, delaying its recycling to the membrane and mimicking insulin resistance.[64] Most importantly for glycosylation, ε2, ε3, and ε4 all differentially affect from which carbon sources astrocytes construct UDP-hexoses and hexosamines, which are critical substrates for glycosylation.[34] Moreover, a targeted glycoproteomic study of plasma proteins in AD patients found that presence of the ε4 allele was a confounding variable for multiple glycosites, suggesting that the APOE allele can affect glycosylation.[65]

How ApoE is able to cause such far-reaching effects is unknown. Yet, ApoE has been reported to localize to the nucleus, where it can bind double-stranded DNA with ~3nM K_D at a sequence motif that is located in the promoter region of up to 3080 genes.[66,67] Transcriptional changes in response to ApoE could have a wide range of effects that could affect glucose metabolism and glycosylation, specifically at the level of glycan synthesis (**Figure 1.4**).[68,69] In proteomic data collected by Swaney et al. in N2a mouse neuroblastoma cells stably expressing ε3 or ε4, we can show that changes in abundance of enzymes in glycolysis, glycogen & galactose metabolism, and the pentose phosphate pathway suggest unique metabolic programs depending on the allele (**Figure 1.5**).

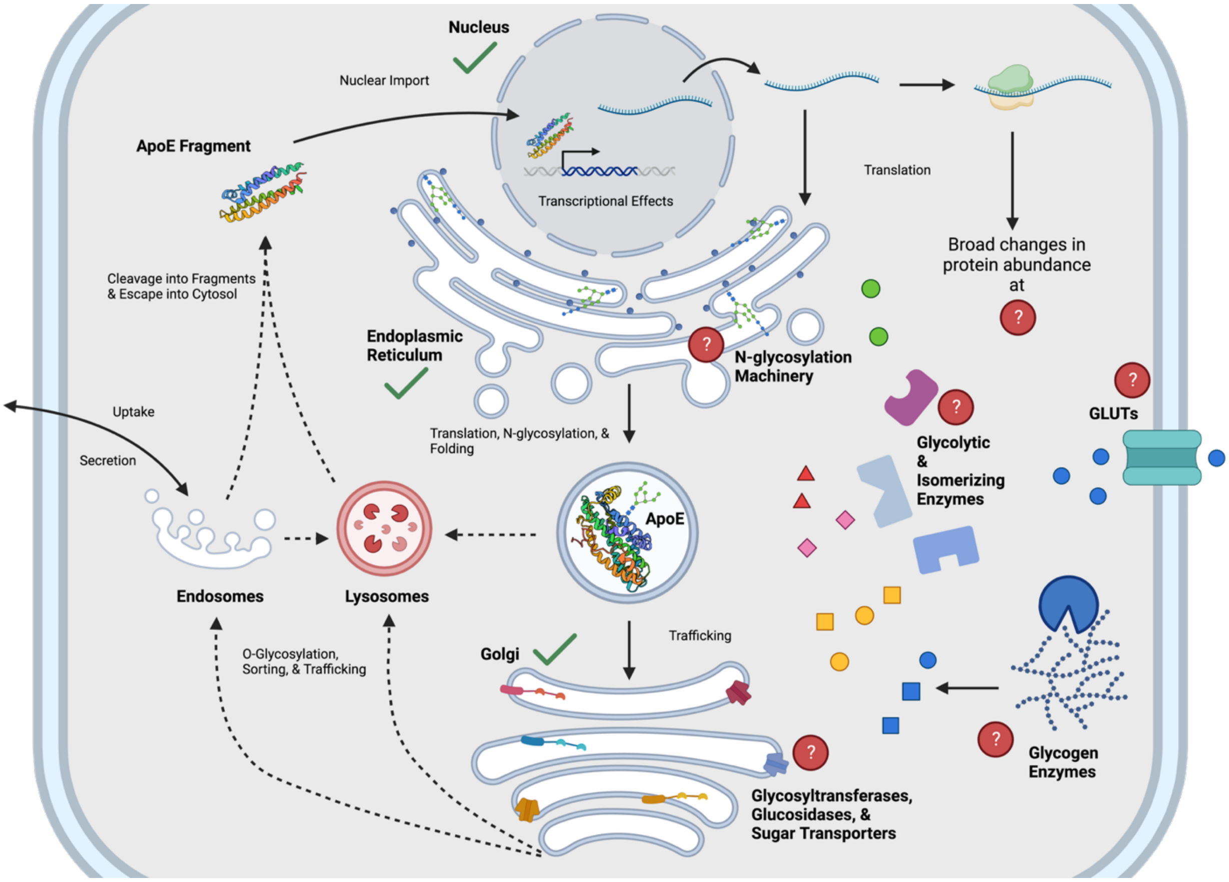

Nucleus
Nuclear Import
ApoE Fragment
Transcriptional Effects
Translation
Broad changes in protein abundance at
?
Cleavage into Fragments & Escape into Cytosol
Endoplasmic Reticulum
?
N-glycosylation Machinery
?
Glycolytic & Isomerizing Enzymes
?
GLUTs
Uptake
Secretion
Translation, N-glycosylation, & Folding
ApoE
Endosomes
Lysosomes
Trafficking
Golgi
O-Glycosylation, Sorting, & Trafficking
?
Glycosyltransferases, Glucosidases, & Sugar Transporters
?
Glycogen Enzymes

Figure 1.4: ApoE as a Transcriptional Regulator of Glycosylation at Multiple Levels. ApoE is synthesized and trafficked from the ER to the Golgi. Following exit from the Golgi, it can undertake one of multiple routes. Dashed lines represent those paths that have not been demonstrated experimentally but can be inferred based on the literature. Solid lines represent paths with experimentally demonstrated trafficking. Green check marks represent locations that preliminary proximity labeling data in our lab can confirm. ApoE fragment imports into the nucleus and functions as a transcription factor. Transcriptional changes can affect any location with a question mark, which would in turn affect glycosylation.

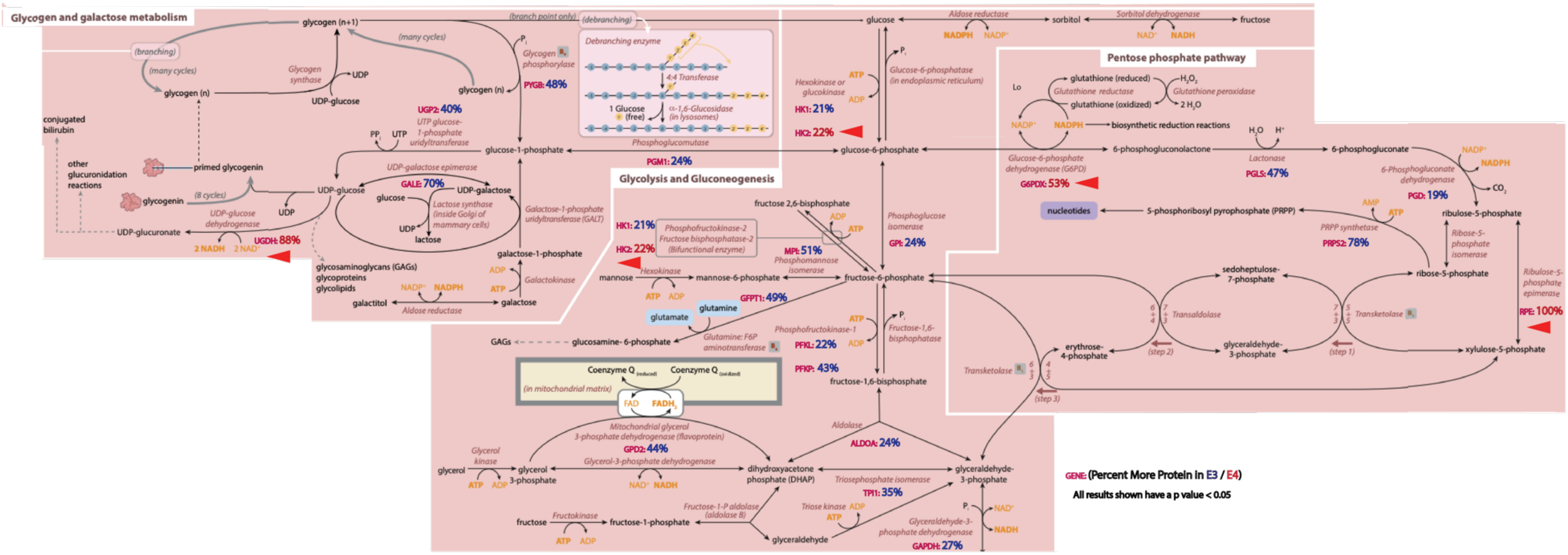


Figure 1.5: APOE alleles differentially alter enzyme abundance at critical points of glucose metabolism, suggesting unique metabolic programs. Glucose is taken up inside the cells and shunted to a variety of different biochemical pathways. Enzyme abundance at individual steps can bias the ability of certain pathways to happen relative to others due to competition for the same substrate. Enzymes with differential abundance between ε3 and ε4 at a p value > 0.05 are shown by gene name near to the step catalyzed by that protein. The percentage reflects the relative amount more protein in the ε3 or ε4 condition. Blue indicates that the protein is more abundant in the ε3 condition; red indicates that the protein is more abundant in the ε4 condition. Biochemical map is adapted from the Stanford School of Medicine *Pathways of Human Metabolism* Version 10.18

Discussion

In this chapter, we demonstrated that we could generate PPI networks by APEX-mediated proximity labeling followed by affinity purification-mass spectrometry for each of the three common alleles of APOE: ε2, ε3, and ε4. PPI networks were generated by using streptavidin or an anti-biotin antibody during purification. The antibody workflow grants higher confidence in reported proteins because the biotinylated peptide of nearby proteins can be detected by mass spectrometry whereas in streptavidin purification it must be inferred as any protein that is detected by the mass spectrometer. Manual annotation of the networks from both workflows demonstrated that several proteins were mentioned by name or by function in a chapter of *Essentials of Glycobiology* that is dedicated to glycoprotein folding quality control. We confirmed that these proteins were not an artefact of the APEX fused to the protein. In fact, comparison of APOE's PPI network to APOL's demonstrated that many of the proteins in this pathway were shared by both apolipoproteins. This potentially points to a common part of apolipoprotein biology.

There is a plethora of ties in the literature that connects AD to glycosylation. There is a plethora of ties in the literature that connects AD to APOE. There is currently no work elucidating APOE's ties to glycosylation although there are several works that provide data suggestive of a link between the two. Moreover, the proximity labeling data showed here indicates that there may indeed be a physical relationship between ApoE protein and proteins involved in glycosylation. We also posit a metabolic link between ApoE and glycosylation. This theory is laid out as follows. ApoE is produced at the endoplasmic reticulum. It escapes at some point during the secretory pathway or during its uptake by another cell. A fragment of ApoE localizes to the nucleus. Alternatively, it is possible that there may be a direct line of transport from the endoplasmic reticulum to the nucleus although this has yet to be shown. Once in the nucleus, ApoE acts as a transcription factor where it effects the transcription and translation of other proteins. In favor of this hypothesis, we can recover PPIs in our network for APOE that localize to the endoplasmic reticulum, the Golgi, and the nucleus. Proteomic data previously collected by Swaney et al but reanalyzed here and overlaid onto a biochemical pathway map suggest that expression of ε3 or ε4 lead to changes in glycogen and galactose metabolism, glycolysis and gluconeogenesis, and the pentose phosphate pathway. These data support the idea that ApoE expression leads to metabolic changes in the cell that are allele specific. In particular, enzymes with significantly altered abundance are changing the abundance of glucose and its isomers as well as substrates for nucleotide synthesis. These data agree with the requirement for activated sugar-nucleotide conjugates as synthetic building blocks for glycans. Taken altogether, these data outline a clear and direct motivation for performing glycoproteomics in regards to APOE alleles.

CHAPTER 2: GLYCOPROTEOMIC ANALYSIS

Introduction

Protein glycosylation, the enzymatic attachment of sugars to proteins, is a very heterogeneous yet common post-translational modification (PTM).[70] *N*-glycosylation, which occurs on asparagine residues within a defined motif, is estimated to occur on over half of the proteins encoded in our genome.[71] Not every potential glycosylation site (glycosite) is always occupied. In fact, the specific pattern of occupied glycosites on a given protein, referred to as macroheterogeneity, can indicate its function.[72–74] To make this process even more complex, the sugar chains (glycans) attached at a given glycosite can have different monosaccharide compositions, linkages, and branching structures, referred to collectively as microheterogeneity. This multifaceted heterogeneity makes glycoproteomics, the large-scale study of glycoproteins, challenging.

Modern glycoproteomics is mostly performed using mass spectrometry and requires careful considerations.[75,76] First, one must identify a method to enrich glycopeptides from the background of mostly unmodified peptides.[77] Second, one must select a method of fragmentation that generates informative ions about both peptide and glycan parts of the molecule.[78] Third, one must analyze the spectra in a way that generates confident, reproducible results.[79]

For this final consideration, there are a host of software suites that are available.[70,75] A community evaluation study was published in 2021 to understand how researchers analyzed their data and how consistent their results were.[80] The study showed that manual analysis was able to improve on the best raw search engine outputs, highlighting the room for improvement. Since then, several new software have been developed, and some existing tools have been improved.

A major difference among glycopeptide analysis software is the way in which each software interprets a glycopeptide spectrum. For example, Byonic attempts to identify peptide and glycan components in one step, treating the glycan like a large variable modification.[81,82] It generates a theoretical complete glycopeptide spectrum for all peptide and glycan permutations supplied from user databases and then scores spectra based on their match. Other software, such as Protein Prospector or MSFraggerGlyco, rely on the mass offset approach.[83–85] In this method, masses of potential glycopeptides are calculated in the same way as Byonic, but the fragmentation spectrum is initially only compared to theoretical peptide fragments from the glycopeptide. Having identified the peptide, if there are multiple potential glycans of similar mass, a scoring system is applied to determine the best assignment among these. Newer software, such as pGlyco3 and GlycoDecipher, initially make use of Y-ions.[86,87] Y-ions contain the peptide backbone and fragments of the glycan. Peptide assignment is performed only after the glycan has been identified. Both pGlyco3 and GlycoDecipher flex an ability to identify modified monosaccharides. GlycoDecipher can perform glycan database-independent identification, an effective *de novo* glycan construction. MSFraggerGlyco, pGlyco3, and GlycoDecipher employ false discovery rate (FDR) calculations for both the peptide and the glycan. This could be a great advance in the field over traditional confidence scores, which are difficult to translate into a probability.

Because of the variety of analytical innovations, it is imperative that another benchmark be performed.[88] Publications of the newer software each conducted a comparison to competing tools.[84,86,87,89] However, these only focused on the total number of spectra identified. They did not discuss overlap or agreement and did not break down the results in terms of relevant information such as unique glycopeptides or glycosites. There was also no comparison to the literature to determine agreement with known information.

Herein, we report a comparative benchmarking of five software: Byonic, Protein Prospector, pGlyco3, MSFraggerGlyco, and GlycoDecipher. For this benchmark, we acquired a novel glycoproteomic dataset from SH-SY5Y cells, a human neuroblastoma derived cell line, using strong anion exchange –electrostatic repulsion liquid chromatography (SAX-ERLIC) for glycopeptide enrichment followed by high pH reverse phase fractionation (HpH-RPF). Data was acquired using stepped collisional energy higher energy collisional dissociation (sceHCD), which balances fragmentation quality and acquisition speed to maximize the number of highly quality spectra for *N*-glycopeptide analysis.[78] Of highlight, we performed glycomic profiling of the SH-SY5Y cells to identify the glycans present and constructed matched glycan databases for all searches. Downstream analysis compared search engine results based on multiple criteria, including agreement with reported glycosites.

Methods

Cell Culture

SH-SY5Y cells were cultured in 15 cm dishes using DMEM:F12 + 10% FBS with no antibiotics and placed in an incubator at 5% CO_2 and 98% humidity at 37°C. For subculturing, media was aspirated from the cells. Cells were washed once with PBS before adding accutase. Cells were incubated at 37°C for 5 minutes. Detached cells were collected and centrifuged at 500g for 5 min to pellet in a 1.5 mL microcentrifuge tube. The supernatant was aspirated from the pellet. Pellets were frozen on dry ice and stored at -80°C until lysis. Each replicate used here represents a cell pellet from separate passages.

Glycomic Profiling

For *N*-glycome profiling, samples were homogenized in an SDS lysis buffer. Proteins were denatured using 10 mM DTT. *N*-glycans were then loaded onto an S-trap plate (Protifi) and incubated with PNGase F at 37°C overnight using the manufacturer's protocol with minor adjustments for glycomics.[90] Briefly, samples are loaded onto S-trap column using S-trap binding buffer (90% methanol, 100mM TEAB). Samples were eluted from S-trap column with two aliquots of 60 µL 0.1% TFA. Following incubation, *N*-glycans were cleaned by Hypercarb column (ThermoFisher). Hypercarb was conditioned with 2 column volumes (CVs) of 99.9% acetonitrile and 0.1% TFA, followed by 2 CVs of 0.1% TFA. Samples were then loaded onto the columns and washed with 4 CVs of 0.1% TFA. Samples were eluted in 50% acetonitrile with 0.1% TFA. *N*-glycans were analyzed by PGC-LC-MS/MS on a ThermoFisher TSQ Altis Mass Spectrometer coupled to a Vanquish LC system. A targeted *N*-glycan method utilizing over 200 *N*-glycan standards was used for the analysis, and samples were run over an 80 min gradient. Collision energies were previously optimized for each standard. A Dextran ladder was used to normalize retention times across runs.[91] Fragmentation pattern and elution order were compared to the standard library to make glycan assignments. Data was analyzed using ThermoFisher Freestyle software, GlycoWorkbench and Skyline.[92,93]

Tryptic Digestion

A buffer was made of 8M Urea, 100mM Tris-HCl, 10mM TCEP, 40mM 2-choloroacetamide buffer at pH of 8.0. Frozen cell pellets were lysed in approximately 1mL. Mixture was pipetted up and down until homogenous. Lysate underwent two freeze-thaw cycles. Then, the lysate was sonicated twice by probe tip for 10s at 20% magnitude. Additional cycles were used if the mixture remained viscous. The lysate was incubated at 37°C for 5 min on a Thermomixer at 1500 rpm to reduce cysteines. Lysate was diluted to 1.5M Urea with 100mM Tris-HCl at pH 8.0. Tryspin and Lys-C were added to the lysate at a 1:100 ratio. Digestion ran overnight at 37°C and 1200 rpm.

Desalting

Following digestion, samples were brought to 1% trifluoroacetic acid (TFA). A Waters Sep-Pac Vac tC18 3cc cartridge was conditioned with three CVs of acetonitrile (ACN) followed by one column volume of 40% ACN / 0.1% TFA. The column was equilibrated with three column volumes of 0.1% TFA. The sample was then loaded onto the column until all the tryptic digest had flown through once. The column was washed with three column volumes of 0.1% TFA. Then, the sample was eluted with 2 mL of 40% ACN / 0.1% TFA followed by 2 mL of 80% ACN / 0.1% TFA. The eluate was lyophilized by SpeedVac.

SAX-ERLIC Enrichment

This protocol was taken from Bermudez and Pitteri 2021[94]. In brief, lyophilized tryptic peptides were resuspended in 1 mL of 50 mM ammonium bicarbonate. The SOLA SAXE SPE cartridge was washed with 3mL of ACN. The column was activated using 3mL of 100mM triethylammonium acetate. The column was conditioned with 3mL of 1% TFA. The column was equilibrated with 3mL of 95% ACN / 1% TFA. Sample was loaded onto the column twice. The column was washed using 6mL of 95% ACN / 1%TFA. Enriched glycopeptides were eluted from the column in two steps: first with two volumes of 850µL of 50% ACN / 1% TFA and second with two volumes of 850uL of 5% ACN / 1% TFA. The two fractions were lyophilized using a SpeedVac.

High pH-Reverse Phase Fractionation (HpH-RPF)

For HpH-RPF, eight fractions were collected in the following manner. Buffers containing 50%, 20%, 17.5%, 15%, 12.5%, 10%, 7.5%, and 5% ACN in 0.1% triethylamine in water were made. A C18 NEST tip was washed with 200µL of ACN followed by 400µL of 0.1% formic acid. Glycopeptides were resuspended in 100µL of 0.1% formic acid. From 5% ACN to 50% ACN, 100µL of each buffer was added and then eluted by microcentrifuge. All fractions were collected and then lyophilized by SpeedVac. Samples were resuspended in 0.1% formic acid before analysis by mass spectrometry.

Mass Spectrometry Acquisition

The liquid chromatography gradient was 120 min at constant flow of 600nL/min. Buffer A was 0.1% formic acid. Buffer B was 80% ACN / 0.1% formic acid. Buffer B gradient from 0 to 50% was 113 min followed by a short 5 min 95% Buffer B phase before ending at 0% at 120min.

For Orbitrap Lumos, MS1 resolution was set to 120K. Scan range was 400 to 1800 *m/z*. RF lens was 60%. AGC target was set to 100%. Maximum injection time was 50ms. Dynamic exclusion was set to exclude peaks for 60s after first appearance. Only ion charge states 2-8 were selected. For MS2, fragmentation was performed with sceHCD 20, 30, 40 nce. Resolution was

30K. Scan range was 120 – 2000 *m/z*. AGC target was set to 200%. Maximum injection time was 200ms.

Data Analysis

All analyses except Protein Prospector were performed on a computer with 1TB RAM and an Intel Xeon 2.40GHz CPU. Protein Prospector was submitted as a job to a server for processing using its web-based interface. All protein databases used the UP000005640_9606 human proteome with one FASTA sequence per protein from UniProt. All glycan databases were informed by glycomic profiling results. For a more detailed explanation, see Glycan Database Conversion. Only specific tryptic peptides with a maximum of 3 missed cleavages were allowed in all searches. All searches allowed for carbamidomethylation as a fixed modification. All searches allowed for oxidation of methionine and protein N-term acetylation as variable modifications.

Protein Prospector Parameters

Protein Prospector (version 6.5.0) was used for the analysis. Raw data was converted into .mgf format peak list files using in-house software 'PAVA', which makes use of Monocle for improved monoisotopic peak selection.[95] These were filtered for the presence of a HexNAc oxonium ion at *m/z* 204.087 (+/- 20ppm) in MSMS scans. The filtered peaklist was searched in Batch-Tag using the same parameters as for other software, other than Gln->pyro-Glu (N-term) was additionally allowed as a variable modification. Also, to adjust for a calibration error in the data, precursor ion mass tolerance considered was a systematic error of 8ppm, then +/- 8 ppm tolerance. Fragment tolerance was 20 ppm. The list of identified glycosylated peptides was then input into MS-Filter to score glycan assignments and find additional glycoforms. Minimum peptide and glycan scores of 0 and 3 were employed, then the best scoring glycan result for each spectrum was reported.

pGlyco3 Parameters

pGlyco3.1 was run in N-glycan mode. Initial search did not include variable modifications on the glycan and allowed for a glycan database size of 1e5. Subsequent searches allowed for 2 max variable modifications on the glycan with a glycan database size of 1e6. Only peptides between 6 and 40 amino acids long were allowed. Minimum peptide weight was 600Da. Maximum peptide weight was 4000. Carbamidomethylation was allowed as a fixed modification on cysteines. Two max modifications were allowed on the peptide. Precursor tolerance was 10ppm. Glycan and Peptide FDR thresholds were set to 1%.

MSFraggerGlyco Parameters

FragPipe (v21.1) with "N-glyco-HCD" workflow was loaded as default. Only peptides between 7 and 50 amino acids long were allowed. Three max modifications were allowed on the peptide. Peptide charges between 1 and 4 were considered. Precursor tolerance was 20 ppm. Glycan and Peptide FDR thresholds were set to 1%.

GlycoDecipher Parameters

GlycoDecipher (v1.0.4) was used. Three max modifications were allowed on the peptide. Peptide length was between 6 and 40 amino acids long. Precursor tolerance was set to 5ppm. Peptide charges between 2 and 6 were considered. Minimum peptide mass was 600Da. Maximum peptide mass was 4500Da.

Byonic Parameters

Byonic from Protein Metrics (v5.4.52) was used. One max modification was allowed on the peptide. Precursor tolerance was set to 20ppm. Maximum precursor mass was 10000Da.

MaxQuant Parameters

Peptides were allowed to be between 7 and 40 amino acids long. Maximum peptide mass was 4600Da. Protein FDR was set to 1%.

Glycan Database Conversion

GlyTouCan Accession Numbers from glycomic profiling were used to import 53 glycan structures into a GlycoWorkbench file. The GlycoWorkbench file was converted into a glycan database file for pGlyco3 (.gdb) using the Convert Glycoworkbench script available in the software. The logic of this script creates all the unique fragments from structures in GlycoWorkbench. This generated a list of 200 unique glycan structures. Custom scripts were created by dictating the logic of conversion to ChatGPT-4 and allowing for it to generate Python scripts that could be run in terminal.[96] Results were manually inspected for accuracy. This approach created glycan databases for Protein Prospector, Byonic, and MSFraggerGlyco. GlycoDecipher uniquely uses the GlyTouCan Accession Numbers. In this case, the contents of the existing "database.csv" file were replaced with only the GlyTouCan Accession Numbers and accessory information.

Preparation of Results of Glycoproteomic Softwares

Each glycoproteomic software generates results in a different format. Therefore, to perform a comparison, the output required manipulation. For Byonic and GlycoDecipher, ChatGPT-4 was used to create a bash script that would concatenate all the results into a single file.[96] For MSFraggerGlyco, all processing used the "psm.tsv" output. For pGlyco3, the pre-supplied "protein_site_analysis.py" script was used to generate a file that contained all information for downstream processing. For Protein Prospector, a tab-delimited output of the MS-Filter results was used for downstream processing.

Analysis in R

Once results from all fractions and replicates could be uploaded in single files, analysis was performed completely in R. Logic for processing and specific functions were dictated to ChatGPT-4, which generated scripts that greatly expedited the analysis.[96] For Byonic, results were filtered for presence of a glycan and a peptide score above 200. For MSFraggerGlyco, results were filtered for presence of a single HexNAc as a modification on the peptide. Protein Prospector, pGlyco3 and GlycoDecipher required no additional processing because their outputs only included glycopeptides.

Cytoscape

UniProt IDs for the glycoproteins identified in all software was searched using the STRING DB plug-in of Cytoscape. Edge information was set to physical protein interactions with a score filter of 0.4. Functional enrichment was performed using the network of proteins against the whole genome.

Results

SAX-ERLIC Enrichment With HpH-RPF Produced High Quality Glycopeptide Spectra By All Software

RAW files were searched with MaxQuant to identify non-glycosylated peptides. Although peptide-first search engines identify modified and unmodified peptides, we chose MaxQuant as an independent search engine to determine the quality of the data. It detected 348,623 spectra in total. Then, RAW files were searched using Byonic, Protein Prospector, pGlyco3, MSFraggerGlyco, and GlycoDecipher (**Figure 2.1, Top**).The same FASTA protein database and glycan databases were used in each software to minimize software variability. To provide a glycan database specific to the sample, glycomic profiling of SH-SY5Y cells was performed (**Figure 2.1, Bottom**). There were 53 *N*-glycan structures identified (**Supplementary Table 2.1**). High mannose glycans were the most abundant, which has been a reported feature of neuronal tissue[97].

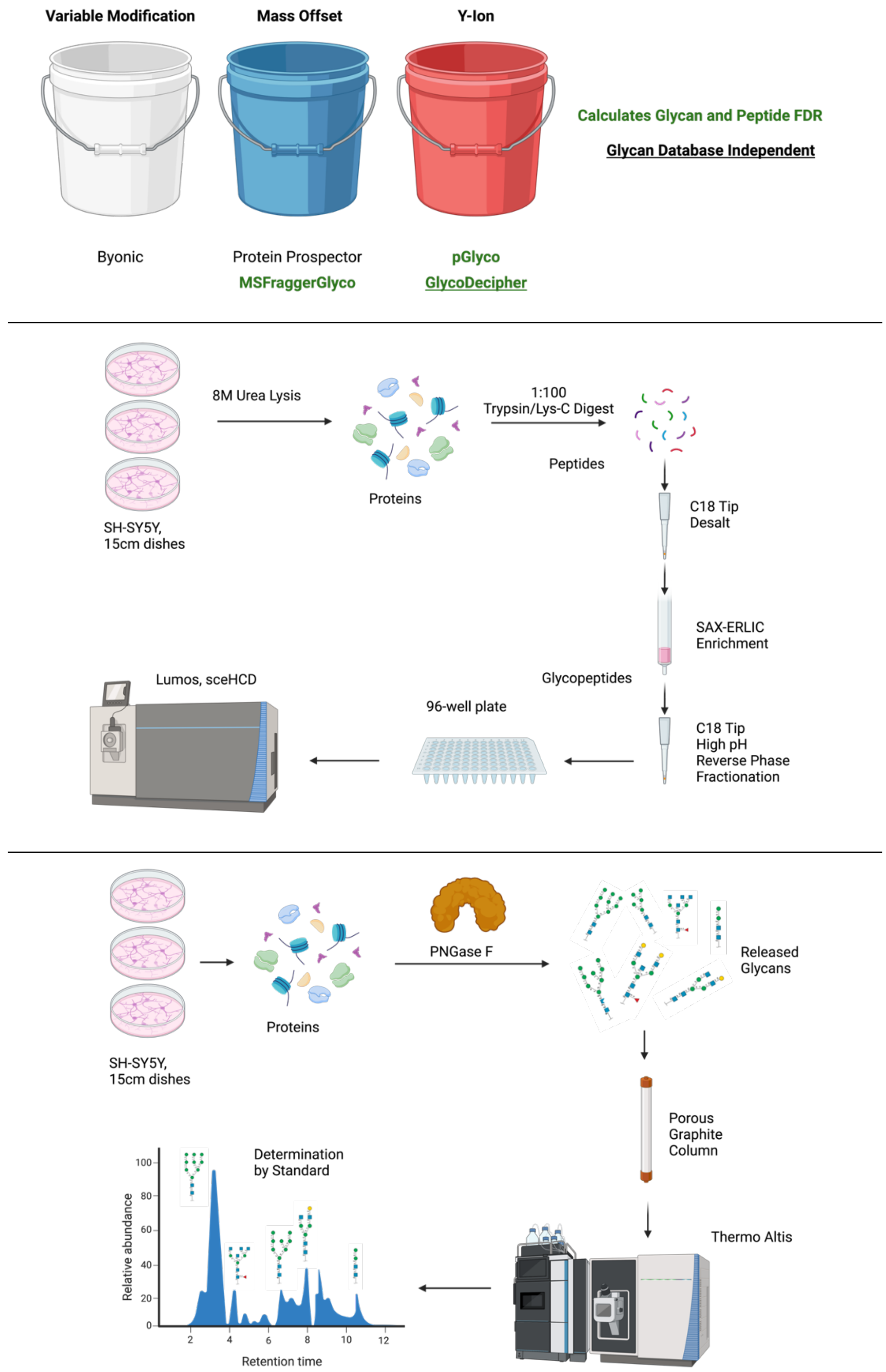


Figure 2.1: Establishing a Head-to-Head Comparison of Glycoproteomic Softwares. (Top) Layout of the different softwares. (Middle) Workflow to generate glycopeptides. (Bottom) Workflow for glycomic profiling. Figure was made with BioRender.com.

While the number of glycopeptides in a single fraction differed across software, the pattern of glycopeptides found across fractions was consistent (**Supplementary Figure 2.1**). The most glycopeptides eluted between 10-15% ACN in 0.1% triethylamine. This underscores the hydrophilic nature of most glycopeptides. These plots also indicate that HpH-RPF is effective at separating glycopeptides into fractions of relatively even complexity.

Figure 2.2 summarizes the number of glycopeptide spectrum matches (GPSMs) reported by each software. In total there were 26,964 unique GPSMs. Protein Prospector reported the most GPSMs (19,717) while pGlyco reported the least (9,482). Other software reported numbers between 11,519 and 14,197. Overall, the mass offset approach software identified the most glycopeptide spectra.

It is interesting that Protein Prospector identified 5,520 more GPSMs than competing software. To investigate further, we analyzed the scan numbers to determine how many of these were unique spectra and how many were detected by another software (**Figure 2.2, Bottom**). 4,430 were spectra that Protein Prospector exclusively identified. The remaining 1,090 were spectra that were also assigned in other software. These data suggest that Protein Prospector was more sensitive at identifying glycopeptide spectra. It is possible that this is because Protein Prospector is not using a glycan FDR threshold like MSFraggerGlyco, pGlyco3, and GlycoDecipher.

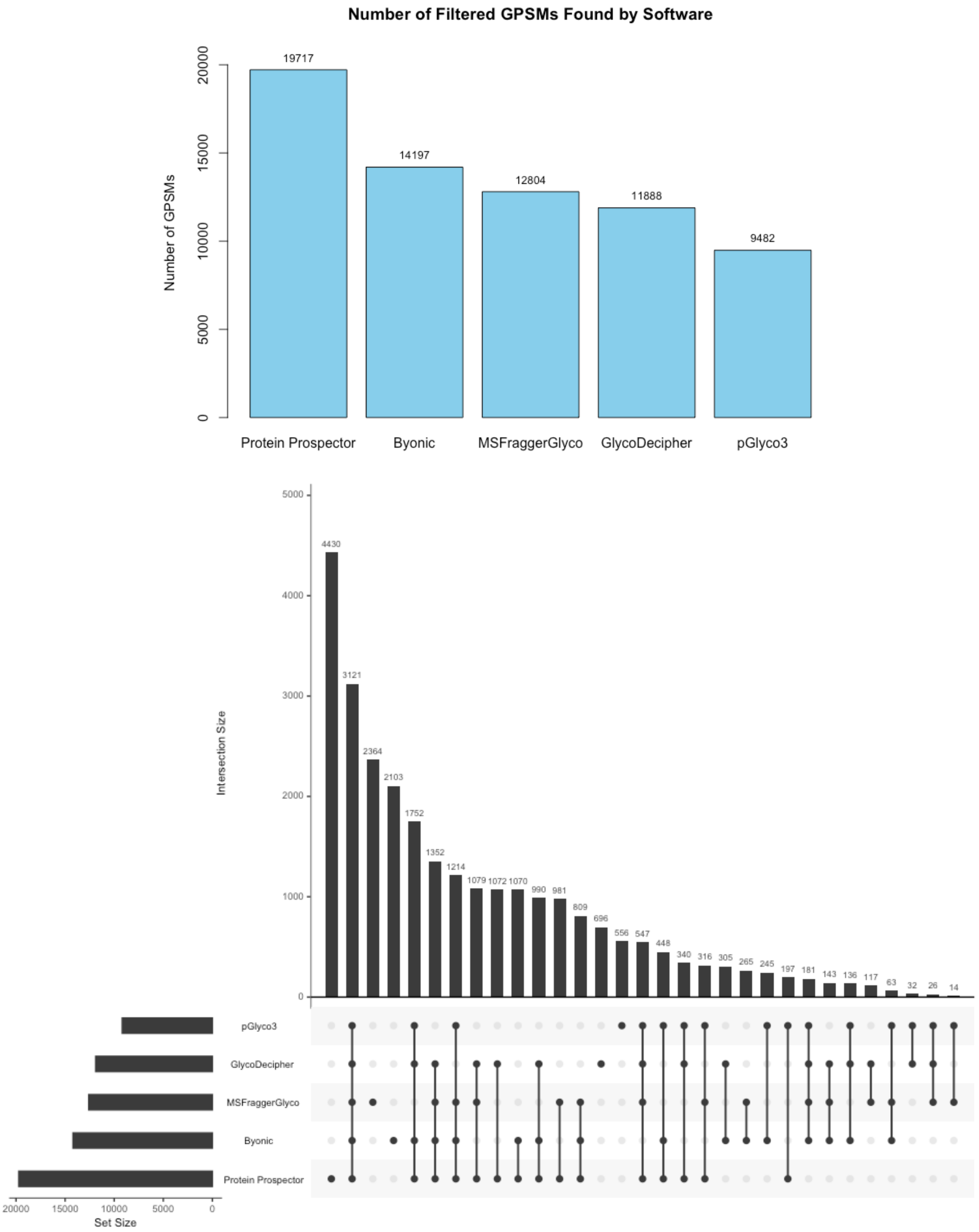


Figure 2.2: Glycopeptide Spectrum Matches (GPSMs) and Their Overlap. (Top) The number of glycopeptide spectral matches that were identified by each software after filtering. (Bottom) UpSet plot of the scan numbers reported for each GPSM.

As an example of what these additional spectra contain, **Table 2.1** summarizes the results for the peptide AGPNGTLFVADAYK from Adipocyte Plasma Membrane-Associated Protein. Protein Prospector identified more spectra to this peptide than competing software. These covered a total of 10 glycoforms, capturing the second most microheterogeneity among software for this peptide. GlycoDecipher reported more glycoforms from many fewer spectra, although some of these were based on *de novo* assignment of extra glycoforms outside the glycan database. Another example where Protein Prospector identifies more glycoforms than other software is shown in **Supplementary Figure 2.2**, which shows annotated spectra for additional glycoforms reported for a peptide from Immunoglobulin superfamily member 3**.**

To understand how many of these spectra corresponded to novel information rather than redundant glycopeptides, we created an identifier for any combination of a protein, site, and glycan (PSG ID). In total, there were 4,383 unique PSG IDs, which represents the total detected diversity of glycopeptides in these samples. **Figure 2.3** shows that Protein Prospector exclusively identified 672 / 4,383 (approximately 15%) PSG IDs. These data indicate that the unique results from Protein Prospector are not redundant information. In fact, Protein Prospector captured the most PSG IDs at 2,559 / 4,383 (approximately 58%) followed by Byonic at 2,490 / 4,383 (approximately 57%).

Table 2.1 : Glycoform Coverage of the Peptide AGPNGTLFVADAYK

Software	Spectra	Glycoforms
Protein Prospector	79	10
pGlyco	62	9
GlycoDecipher	29	11*
MSFragger	4	2
Byonic	-	-

*7 of these are masses that do not correspond to defined glycans.

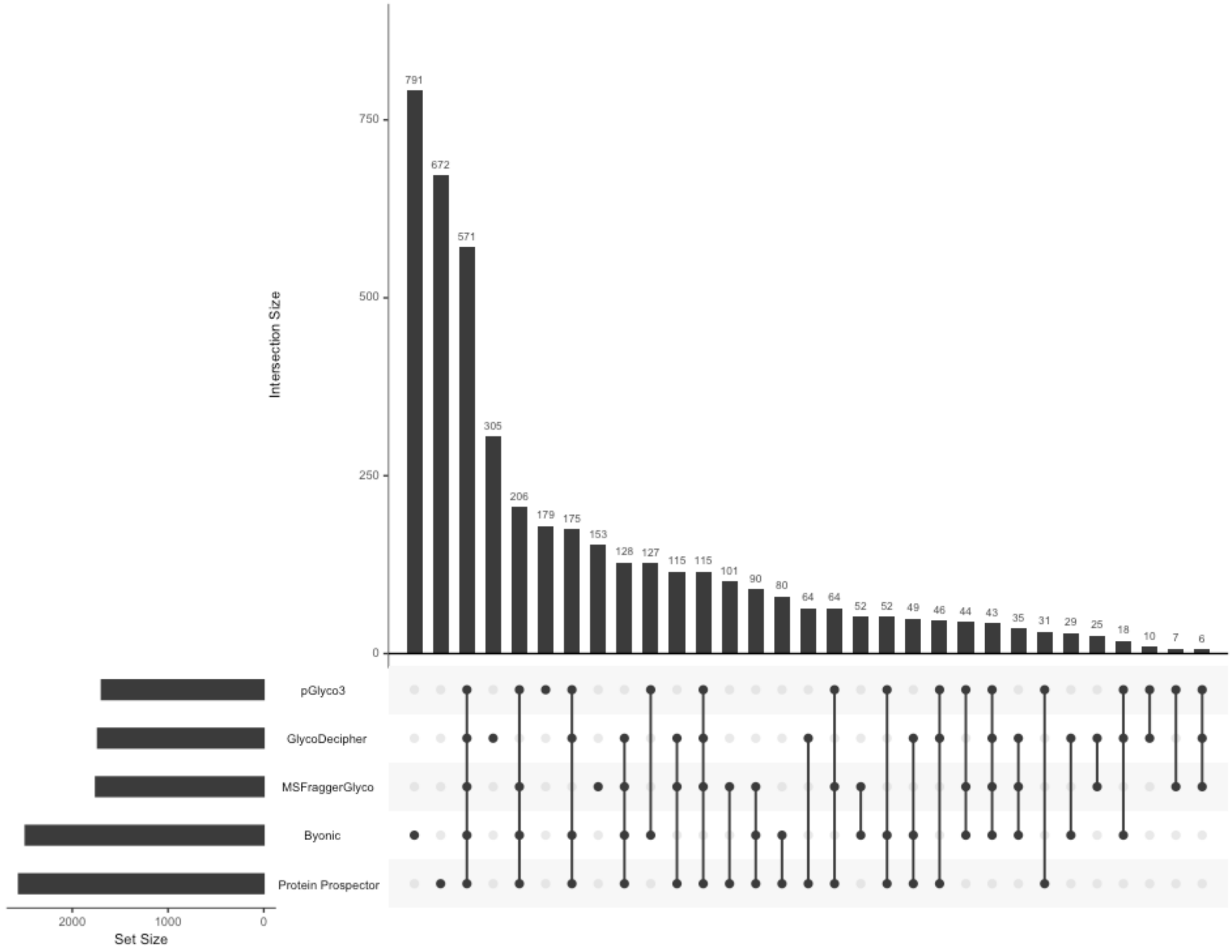


Figure 2.3: Protein Prospector Captures More Information Content from the Data. UpSet plot of the unique protein-site-glycan combination (PSG IDs) by software.

Comparison of the Glycoprotein Identities

To understand whether common glycoproteins were identified and how many, we plotted the overlap in results at the protein level (**Figure 2.4**). In total, there were 947 unique glycoproteins found by at least one software. Of these, 231 glycoproteins (approximately 24%) were found by every software, matching the poor overlap seen in the previous community-wide study.[80] A large portion of the unique glycoproteins came from Byonic, which exclusively identified 382 / 947 proteins (approximately 40%). In fact, Byonic was an outlier because it identified a total of 749 glycoproteins where the nearest competitor, MSFraggerGlyco, found a total of 389. By looking at the spectra uniquely identified in Byonic, we were able to determine that many of these were incorrect assignments (for example, see **Figure 2.4**). On the other hand, despite reporting the most GPSMs, Protein Prospector reported the fewest glycoproteins. Excluding Byonic, using at least two software was sufficient to reproduce at least 254 glycoprotein identities. This would be 65% of hits for MSFraggerGlyco or 86% for Protein Prospector.

Encouragingly, there was good confidence in the commonly reported glycoproteins. Using Cytoscape with STRING, a physical interaction network was generated for 231 common glycoproteins. The results were clustered using a granularity level of 4 (**Figure 2.5**). The network was significantly enriched for protein-protein interactions (PPIs) with a reported PPI value of 1.0E-16 with several visible clusters of proteins. Functional enrichment of the network revealed that the top theme from UniProt Keywords categorization was the term "Glycoprotein" (**Table 2.2**). In comparison, functional enrichment for all peptides identified by MaxQuant did not include "Glycoprotein" at all but rather "Phosphoprotein", "Acetylation", and "Cytoplasm" as top themes (**Supplementary Table 2.1**). These data indicate that for those proteins upon which software agree, there is high confidence in their status as glycoproteins. It is worth noting, however, that the number of results exclusively reported by a single software was considerable (**Figure 2.5**).

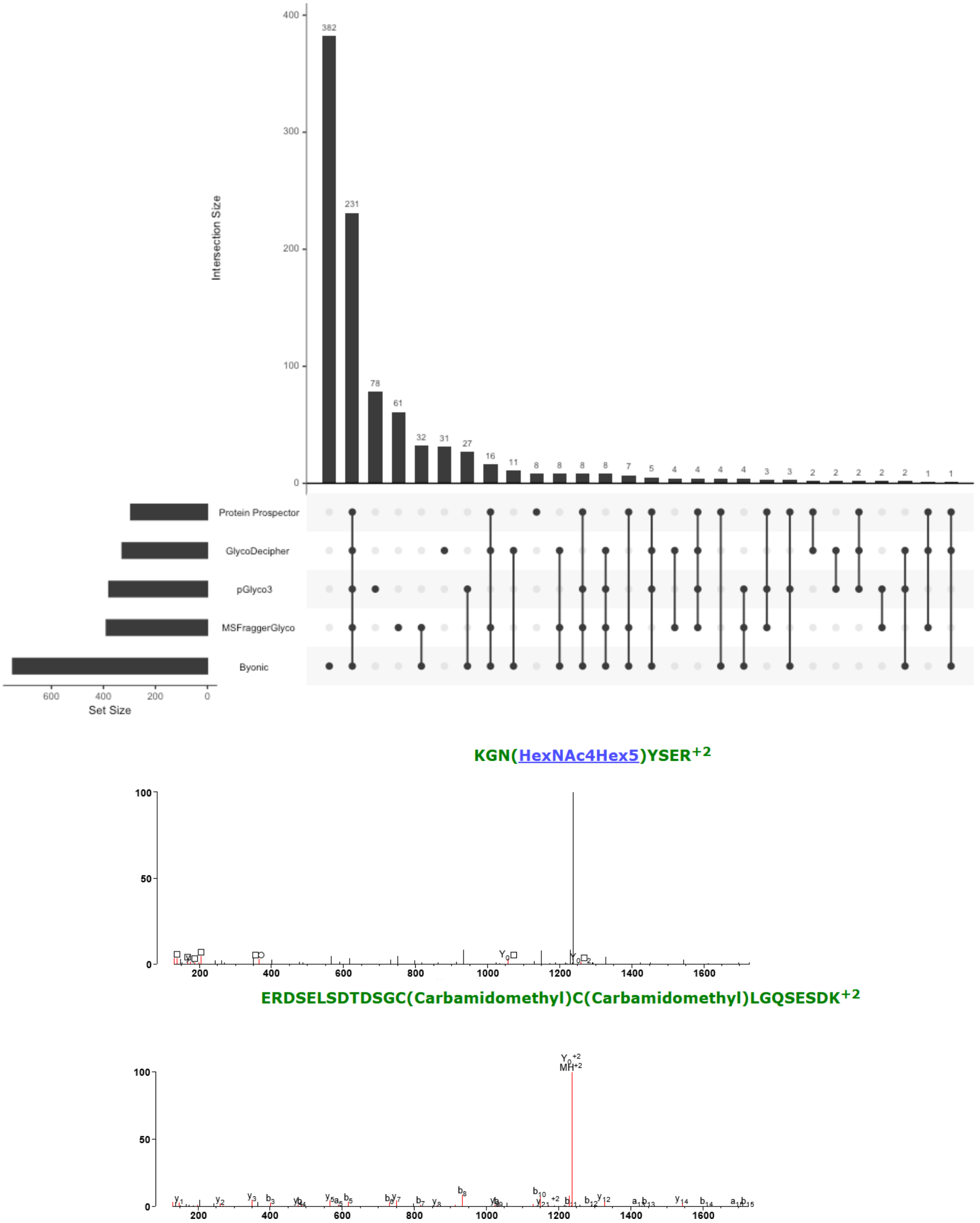


Figure 2.4: Comparison of the Unique Glycoproteins. (Top) UpSet plot of the unique glycoprotein results from each software. (Bottom) This is an example of a unique glycopeptide assignment by Byonic. The peptide Byonic assigns is KGNYSER with glycan HexNAc4Hex5. The peptide comes from HistoneH2A. Protein Prospector assigns this spectrum to an unglycosylated peptide from E3 Ubiquitin-protein ligase UHRF1.

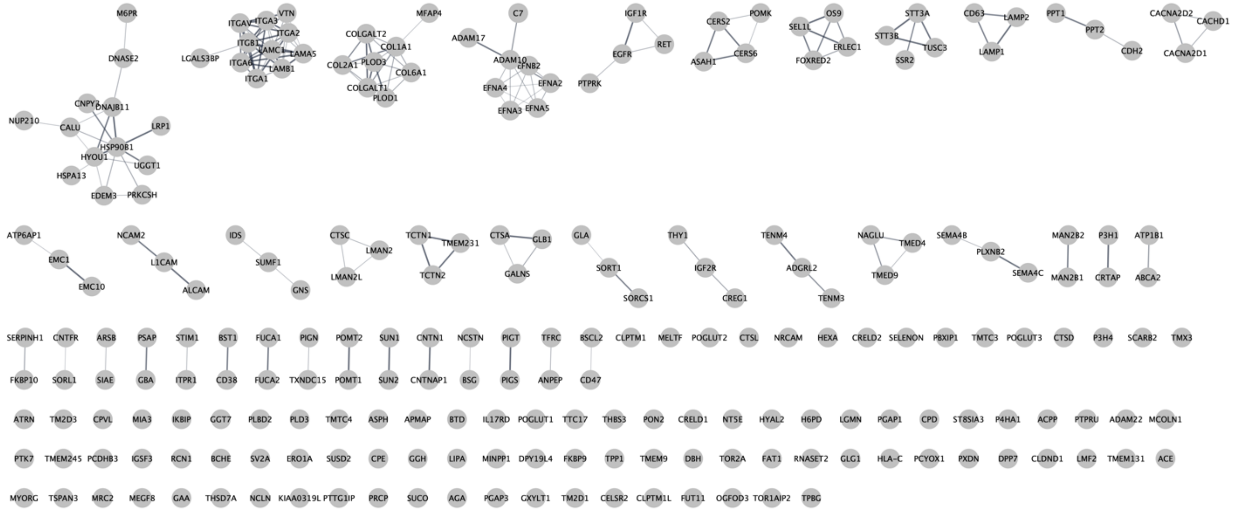


Figure 2.5: 231 Glycoproteins Shared by All Software. UniProt IDs were imported to Cytoscape using the STRING plug-in. Edges denote known physical interactions. Results were clustered with granularity score 4.

Table 2.2: Functional Enrichment of 231 Shared Glycoproteins. Results were filtered for UniProt Keywords.

# background genes	# genes	description	FDR value	p-value
4386	222	Glycoprotein	1.56E-127	2.32E-130
3277	178	Signal	3.18E-87	9.46E-90
3338	126	Disulfide bond	9.86E-36	4.40E-38
1168	73	Endoplasmic reticulum	2.57E-30	1.53E-32
314	44	Lysosome	2.91E-30	2.17E-32
5067	128	Transmembrane	1.34E-19	1.20E-21
5040	127	Transmembrane helix	2.52E-19	2.63E-21
7068	151	Membrane	6.84E-18	8.14E-20
886	41	Calcium	7.49E-12	1.00E-13
1612	56	Hydrolase	1.18E-11	1.76E-13

Comparison of Glycosite Assignments

Figure 2.6 is an UpSet Plot of glycosite assignments reported by software. In total, there were 1,466 unique glycosites discovered in the searches. Of these, 308 (approximately 21%) were common to all software. Again, Byonic was an outlier. It exclusively reported 491 / 1,466 unique glycosites (approximately 33%). 383 of these were because of the glycoproteins it exclusively reported (**Figure 2.4**). Byonic alone reported a total of 1,123 glycosites where the nearest competitor, MSFraggerGlyco, reported 658. As with the glycoprotein-level summary, Protein Prospector reported the fewest glycosites.

There is a sizeable portion of proteins reported in each software to have multiple glycosites. **Figure 2.7** shows that while most proteins have a single *N*-glycosite, somewhere between 24 to 40% have multiple. LRP1 is exceptional in that it has somewhere between 12 and 15 glycosites (**Supplementary Figure 2.3**), as has been observed in prior studies.[98] Overall, there is equivalent consensus on where glycosites are located (approximately 25% of glycosites reported by all softwares) as there is on what proteins are being glycosylated (approximately 24% of glycoproteins reported by all softwares).

A biologically relevant result from glycopeptide search software is the discovery of novel glycosites. Hence, glycosites for each UniProtID identified were compared to the glycosites discovered by the different software (**Figure 2.8**). Not all glycosites reported in UniProt were expected in the results, but our goal was to understand what percent of those identified in our data are also reported in UniProt as a proxy for reliability. Byonic had the lowest agreement with UniProt with 54% of the sites identified also reported by UniProt. While it is possible to conclude from this that Byonic is more sensitive than competitors, a more likely conclusion given its consistently inflated numbers is that it is reporting more spurious results. For reference, Protein

Prospector, which identified the most glycopeptide spectra and the fewest glycosites, had the highest agreement with 82.71% of its sites also being reported in UniProt.

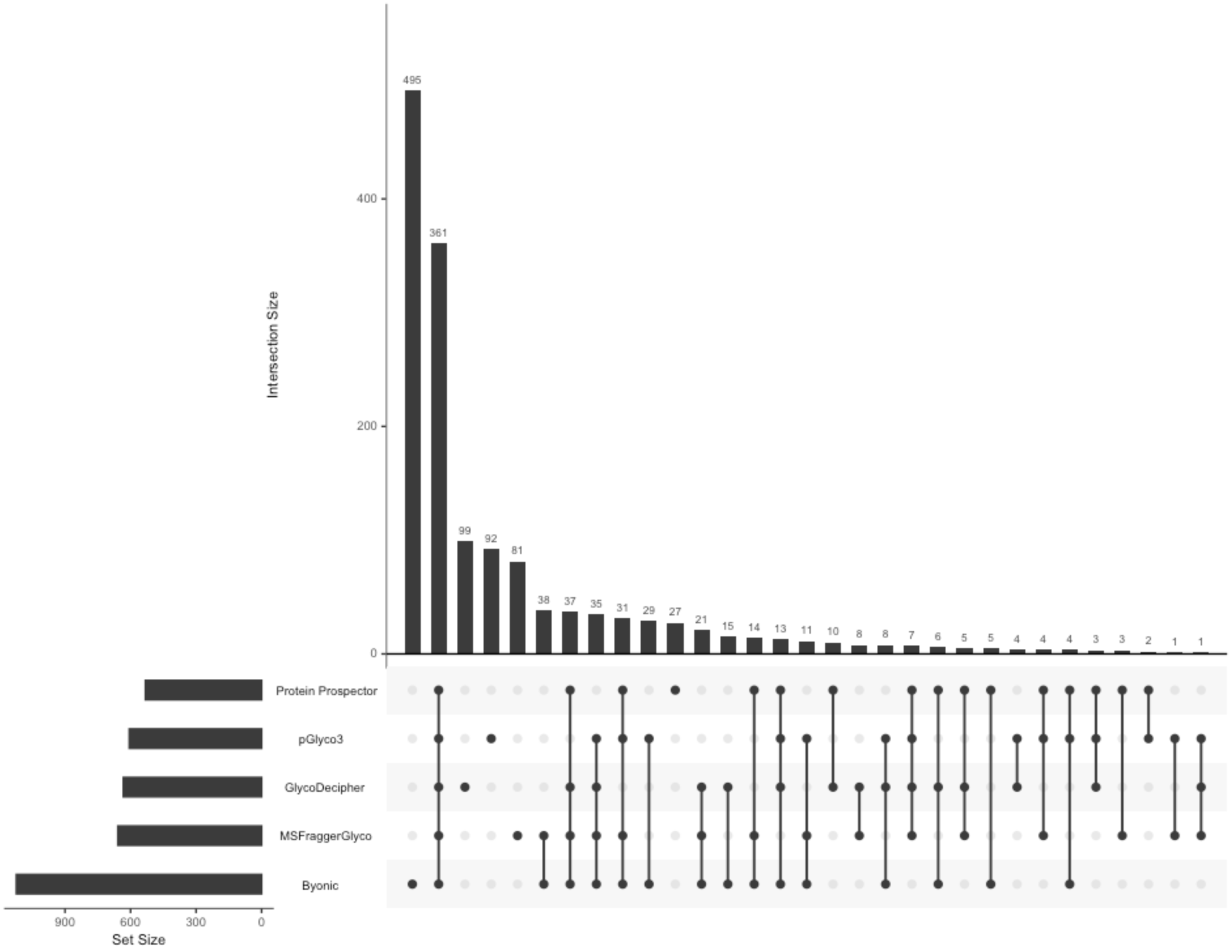


Figure 2.6: Comparison of the Unique Glycosites. UpSet plot of the unique glycosites by software.

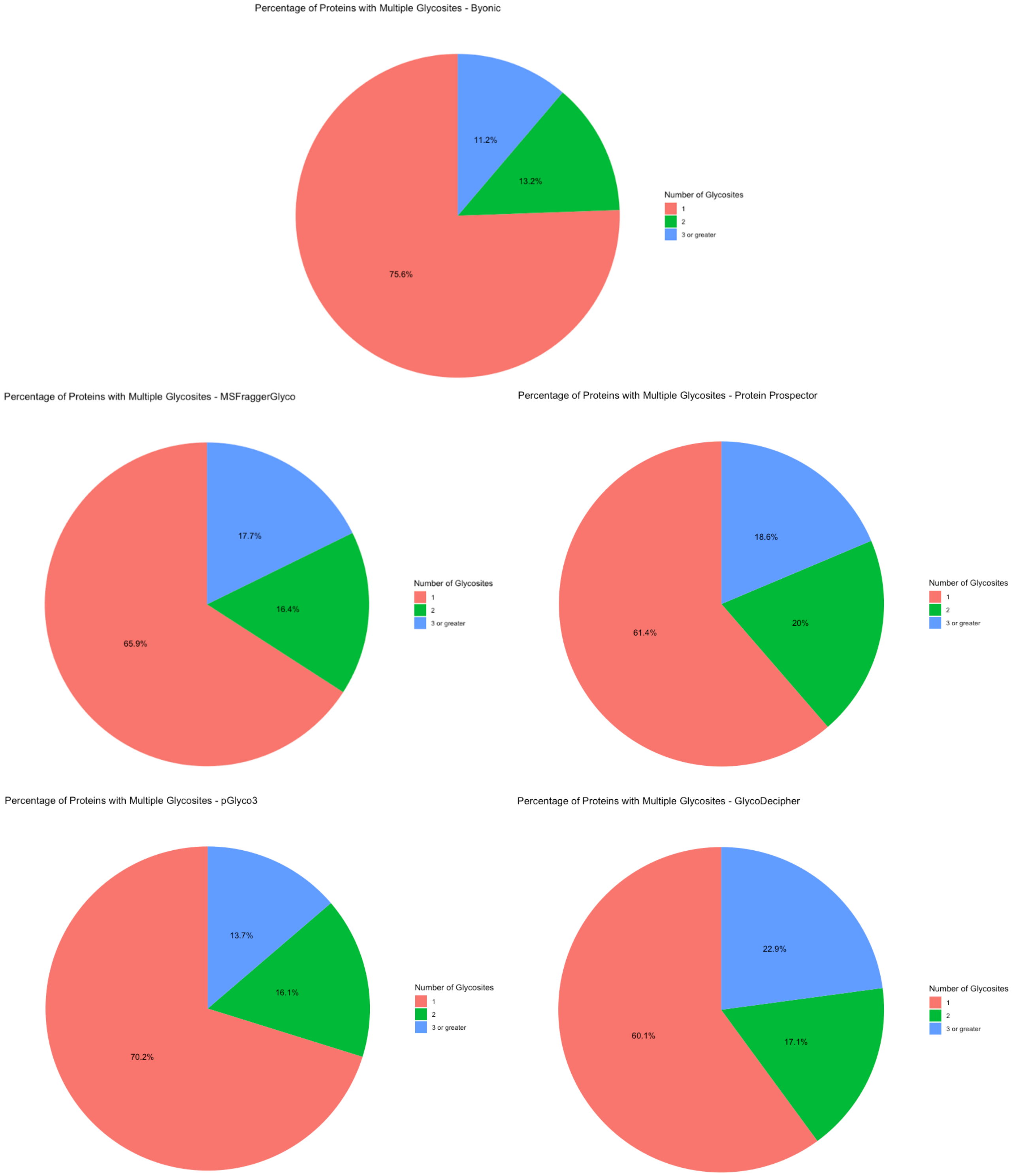


Figure 2.7: Percentages of Proteins with Multiple Glycosites by Software. Pie charts display the percentage of glycoproteins identified that had multiple glycosites detected in this dataset.

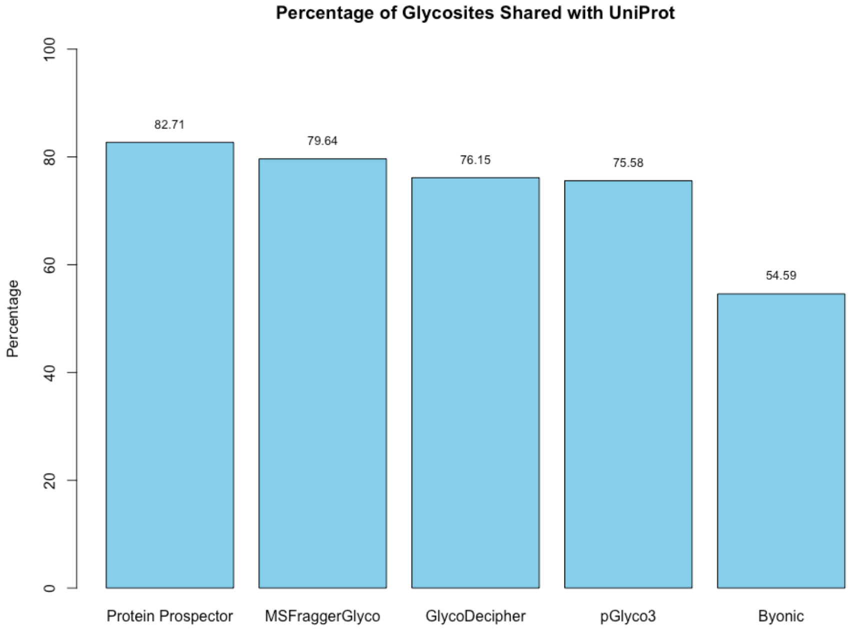


Figure 2.8: Agreement between Glycosites Assigned in the Software and Glycosites Reported on UniProt. For each UniProt ID detected in a software, all glycosites recorded on UniProt were collected. Percentages were calculated as the number of glycosites reported in a software over all the glycosites reported for each UniProt ID.

Comparison of Glycans

Overall, the software identified 99 distinct glycans (**Figure 2.9**). Of these, 25 were common to all software. GlycoDecipher exclusively identified 11. This was expected because GlycoDecipher performs *de novo* “monosaccharide stepping” to report glycans not within the search space of other software. After looking for glycopeptides with glycans provided in its database, GlycoDecipher tries to identify more glycans by allowing for modifications on the monosaccharides to subsequently identify the additional peak/s in a Y-ion series. Most results from GlycoDecipher were identified in this manner but could be translated into defined glycan compositions (**Table 2.3**). In most cases, the additional mass was the result of a misassigned monoisotopic peak that, when corrected, led to a glycan assignment that was within the glycan database considered by other software. However, there were some glycans identified that were not identified in the glycomic analysis, notably those containing phosphate groups. Standards for these glycans were not available for glycomic analysis. These results suggest that GlycoDecipher's *de novo* approach could be useful but is greatly hampered by its reliability.

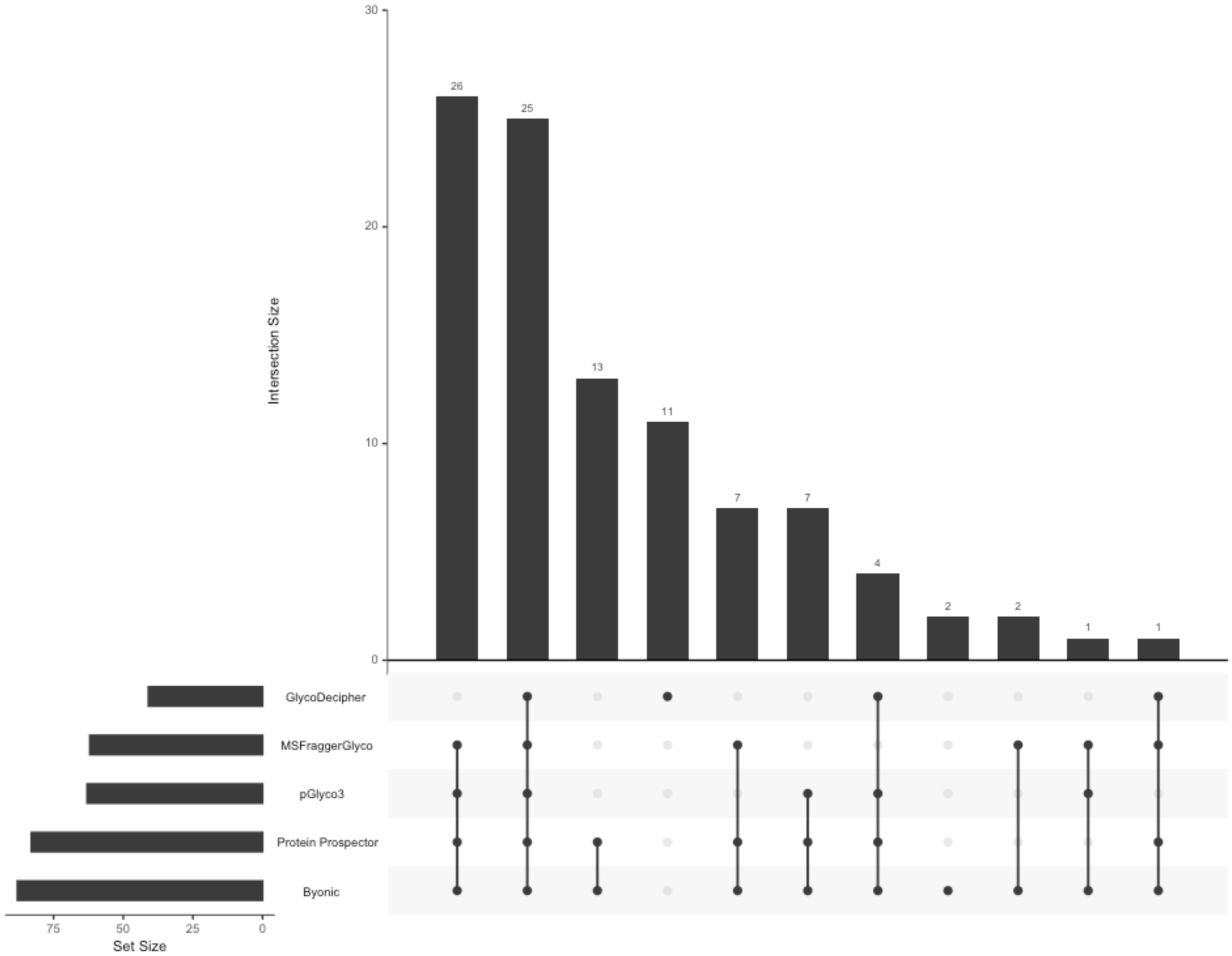


Figure 2.9: Comparison of the Unique Glycans. Upset Plot of the glycans identified by each software.

Finally, we investigated the frequency of each glycan reported. $HexNAc_2Hex_6$ was consistently the most identified glycan (**Figure 2.10**). Interestingly, it was not the most abundant glycan from glycomic profiling, which was $HexNAc_2Hex_8$ (**Supplementary Table 2.2**). The glycoproteomic results were not evaluated for peak intensity, so this difference presumably indicates that $HexNAc_2Hex_6$ glycoforms, although more common on proteins, were typically of lower abundance than $HexNAc_2Hex_8$. An alternate, less likely explanation could also include a small degree of high mannose glycan degradation during sample prep or in the gas-phase. Regardless, $HexNAc_2Hex_8$ was either the second or third most frequent glycan in glycoproteomic searches (**Figure 2.10**). $HexNAc_2Hex_9$ was the second most abundant glycan by glycomic profiling and was the third or fourth most frequent glycan depending on the software. Taken together, all software identified the most common glycans but differed in their detection frequency. Software also differed in the low abundance glycans detected.

We also investigated the percentage fucosylation and sialylation. Very little sialylation was found in this data: 1% or less of IDs by all software. These results matched that of the glycomic profiling (**Figure 2.11**). Fucosylation was more common with up to 20% of glycopeptides reported by a given software. Interestingly, this differed from the glycomic profiling results, where fucosylated glycans were only 3.77% of the glycans identified. A common error in data analysis is to assign two fucose where there is only a sialic acid because two fucose (292.1158 Da) and a sialic acid (291.0954) are close in mass. **Supplementary Figure 2.4** displays the percent of peptides with multiple fucose. Only up to 2% of the glycopeptides in any software contained two fucose, which makes it an uncommon occurrence. The high frequency of fucosylation on glycopeptides relative to the low abundance of fucosylated glycan by glycomic profiling is not due to misassignment. These data suggest that counting the frequency of a glycan on peptides is a poor proxy for its relative abundance; peak intensities need to be taken into consideration.

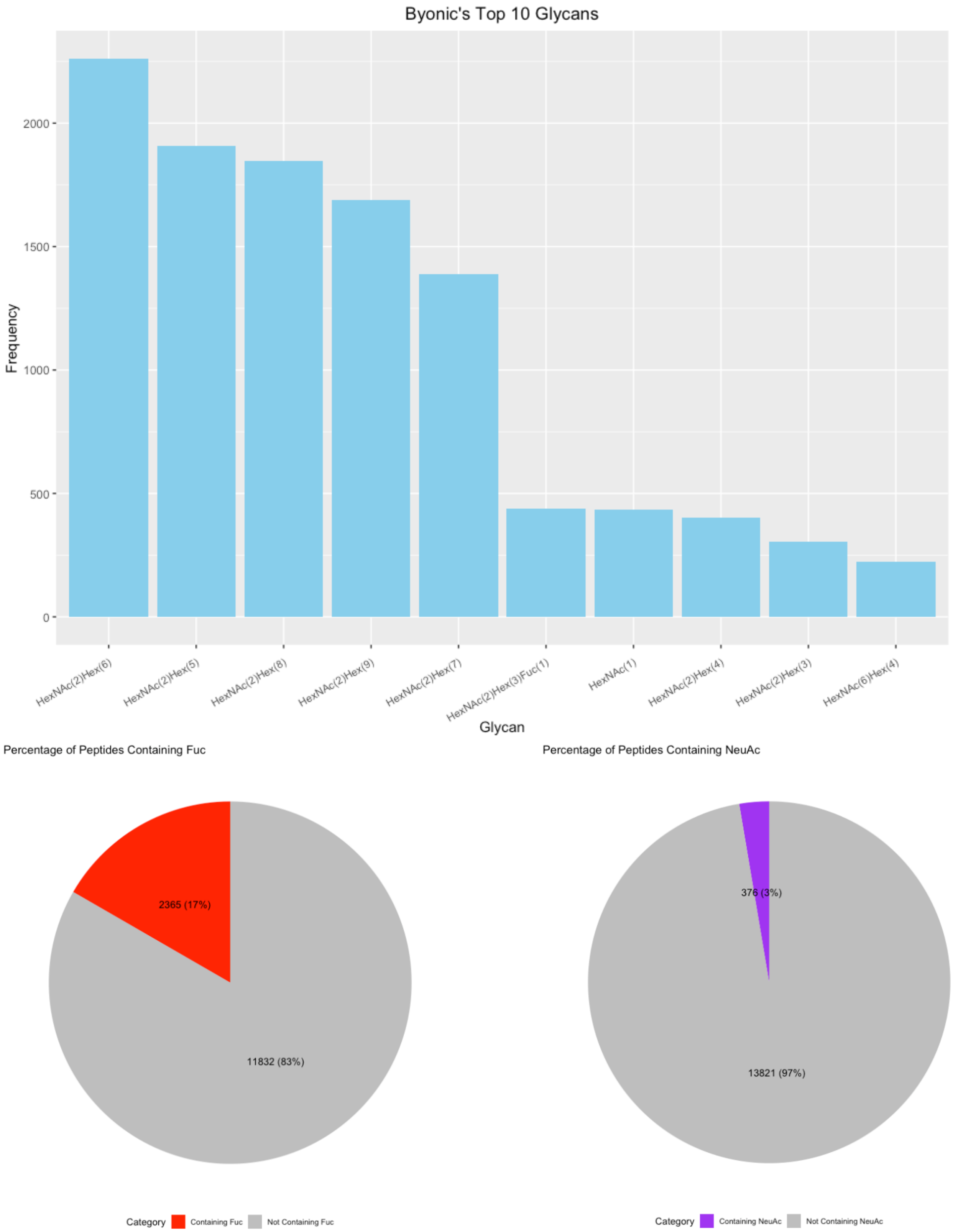


Figure 2.10: Most Frequent Glycans and Distribution of Features. Top 10 glycans identified by each software with pie charts displaying the percent of fucosylated and sialylated peptides.

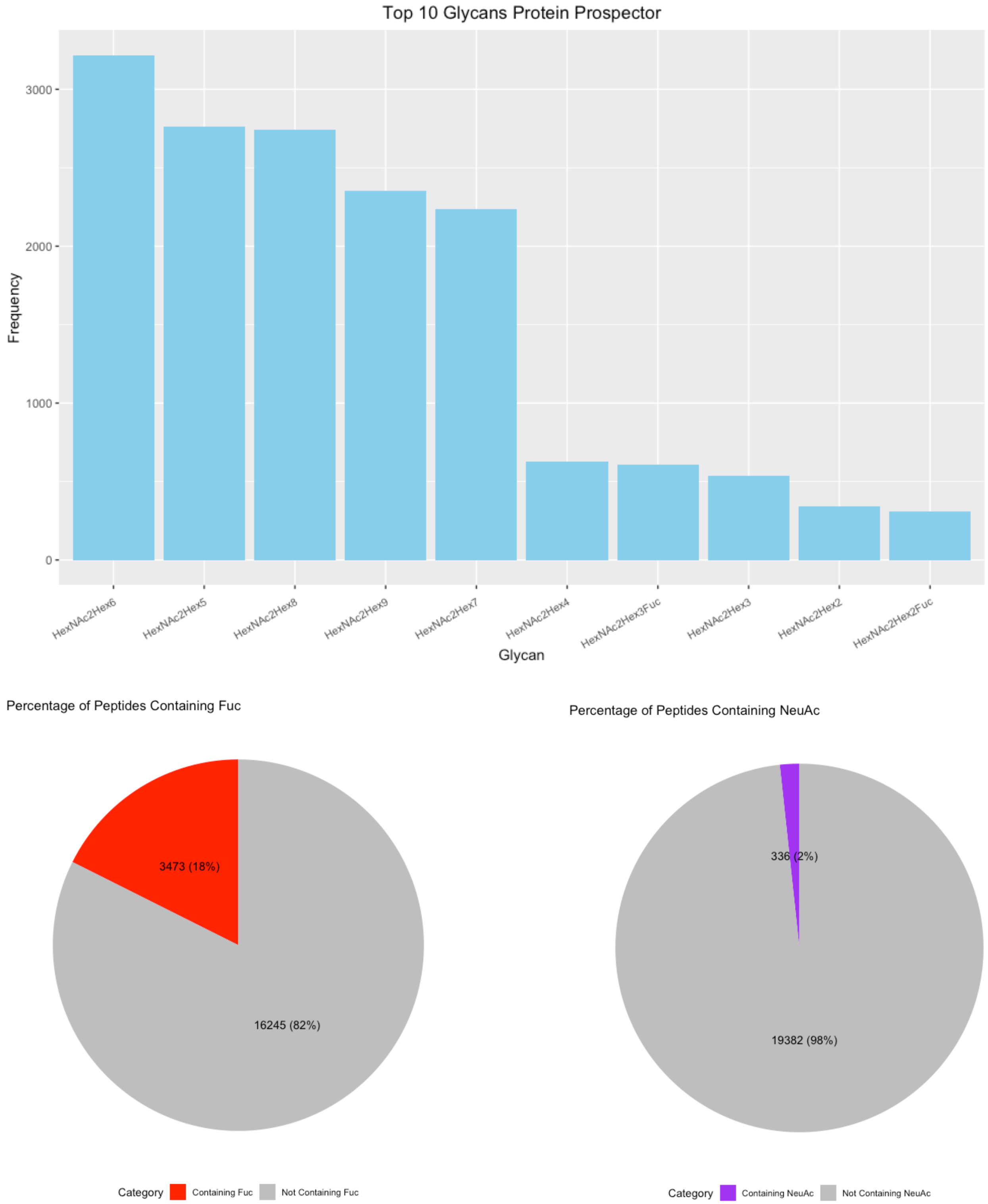


Figure 2.10: Most Frequent Glycans and Distribution of Features. Top 10 glycans identified by each software with pie charts displaying the percent of fucosylated and sialylated peptides.

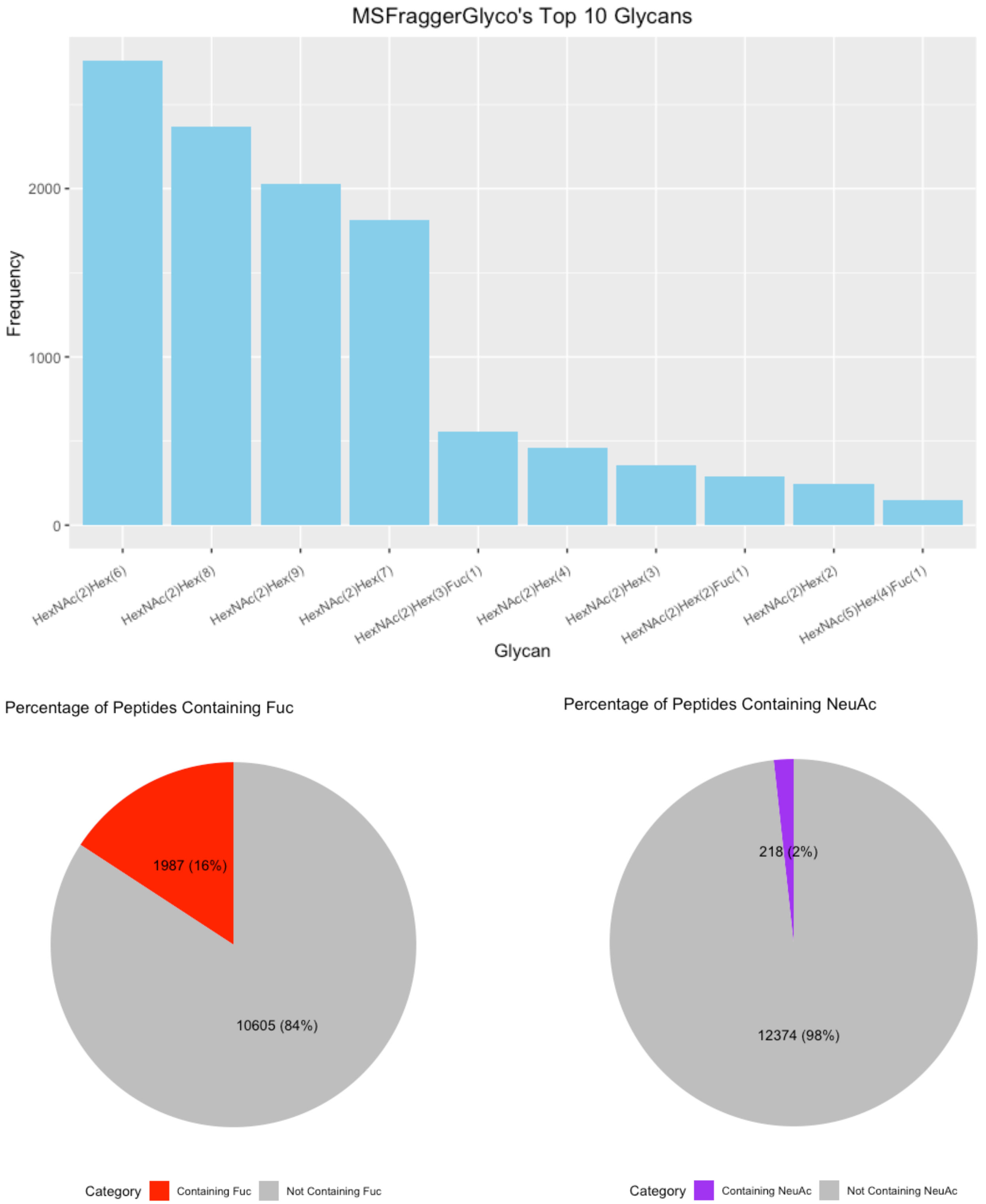


Figure 2.10: Most Frequent Glycans and Distribution of Features. Top 10 glycans identified by each software with pie charts displaying the percent of fucosylated and sialylated peptides.

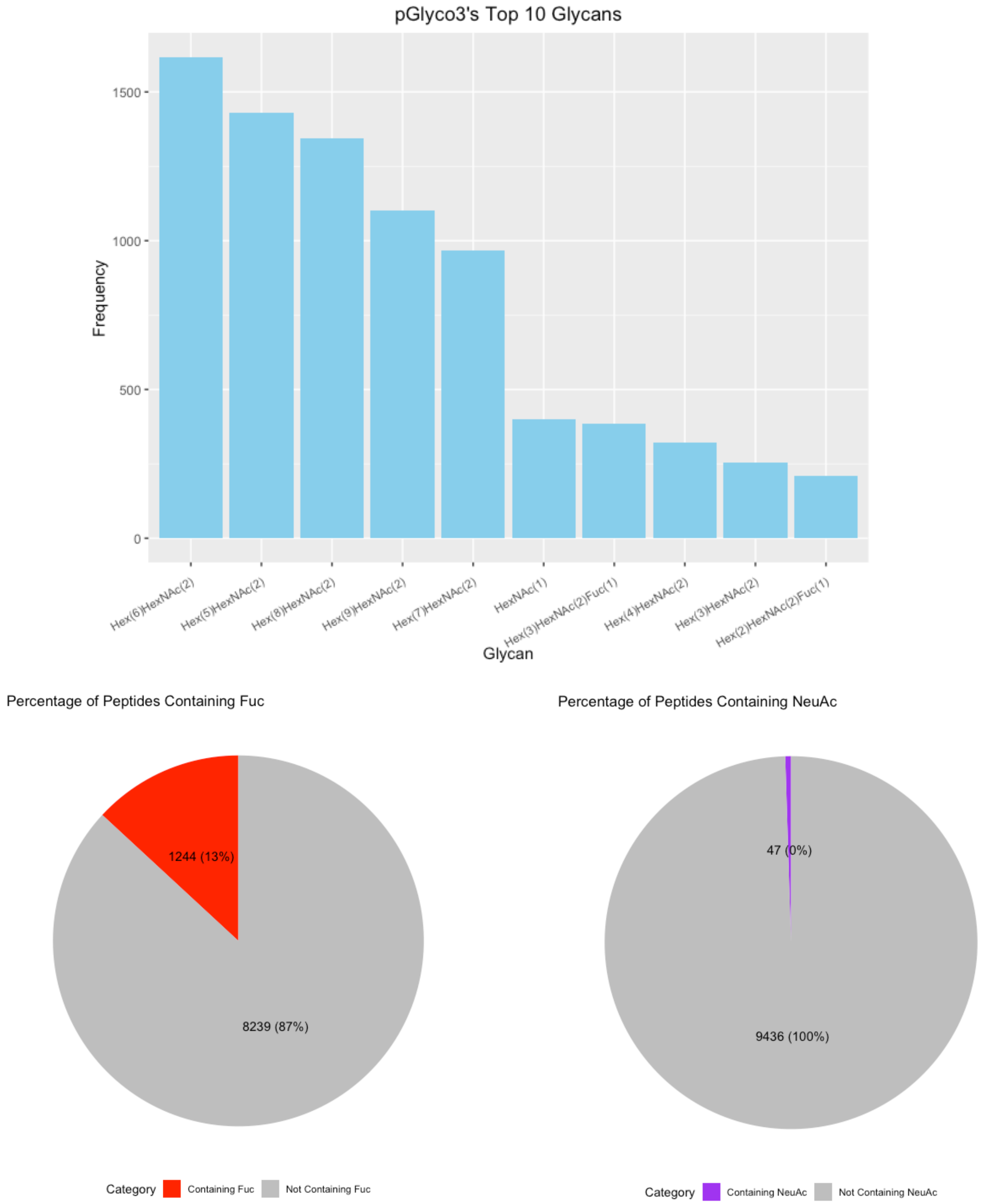


Figure 2.10: Most Frequent Glycans and Distribution of Features. Top 10 glycans identified by each software with pie charts displaying the percent of fucosylated and sialylated peptides.

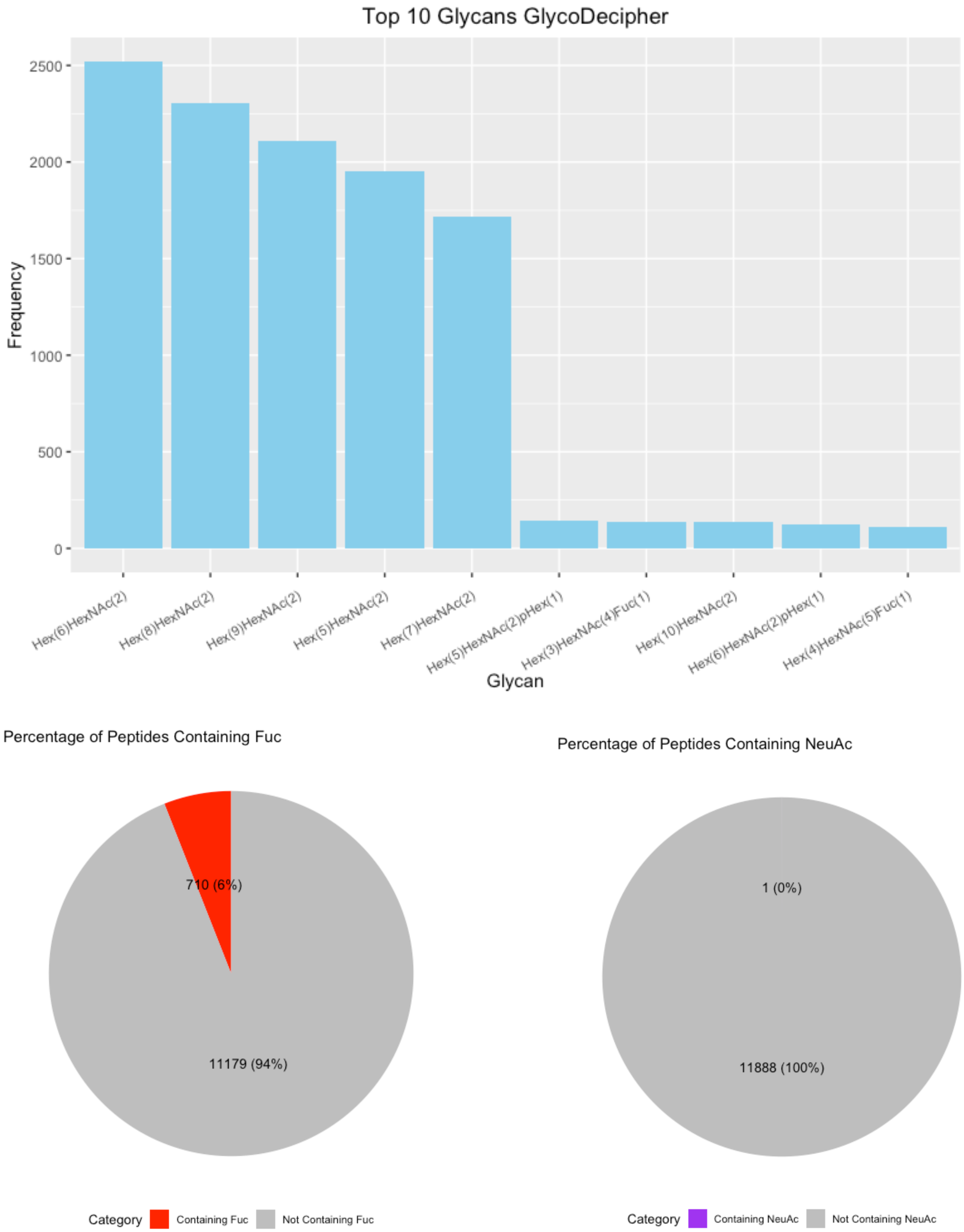


Figure 2.10: Most Frequent Glycans and Distribution of Features. Top 10 glycans identified by each software with pie charts displaying the percent of fucosylated and sialylated peptides.

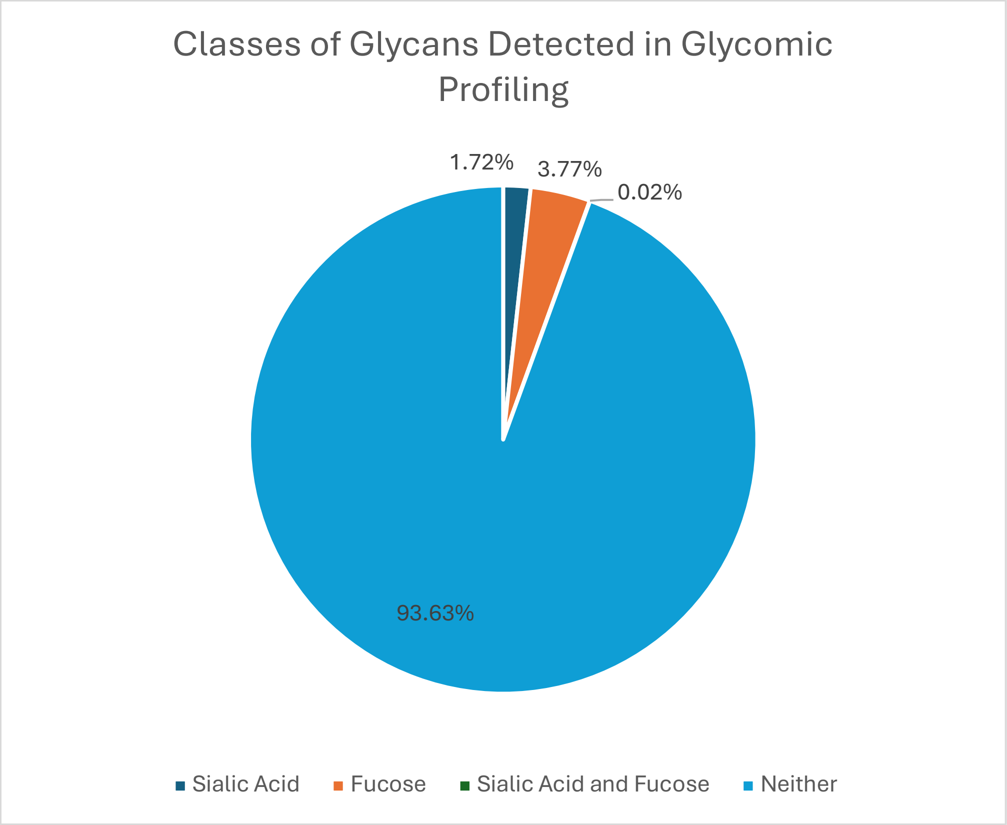


Figure 2.11: Distribution of Features for All Glycans in Glycomic Profiling. Pie chart displays the percent of each of the labelled features by abundance for detected glycans.

Usability of Glycoproteomic Software

Usability of a software greatly influences its longevity with users. A computer with ample RAM is critical for MSFraggerGlyco. All 24 RAW files could not be searched with the 16GB of RAM available on a high-end laptop or typical desktop computer. Using a computer with 1 TB of RAM solved this problem. We did not evaluate the minimum RAM required for MSFragger to complete this search. Provided that there is appropriate RAM, MSFraggerGlyco was the fastest software, completing its search in 91.4 minutes (**Supplementary Figure 2.5**). Byonic was the slowest to complete its search at 364.1 minutes.

The ability to recalibrate the data is useful for a robust glycoproteomic analytical tool. Due to the high mass accuracy of modern mass spectrometers a small systematic error in calibration can impact search performance. The data in this study had a mass error of on average 8 ppm. Every software except Byonic had either a wider default precursor tolerance or automatically adjusted for this systematic error. To get comparable results, Byonic required manual adjustment to a 20 ppm precursor tolerance (**Supplementary Figure 2.6**). Some software, such as MSFraggerGlyco and pGlyco3, have a larger precursor tolerance and did not require adjustment. Protein Prospector and GlycoDecipher had narrower tolerances but could adjust for the systematic error.

Table 2.3: Glycan Misassignments from GlycoDecipher. "Moiety Mass" is the mass of the nontraditional monosaccharide reported by GlycoDecipher. "Corresponds to:" shows our assessment of what the mass corresponds to based on the value and the spectra. "Operation" displays the function applied to correct for the misassignment. Rows with a "#" are major structural errors. Bolded rows are examples of uncommon monosaccharides, such as a phosphorylated hexose.

Moiety Mass	Corresponds to:	Operation
178.06	Hexose + Oxygen	Add 2 Hexose, Subtract Fucose
178.07	Hexose + Oxygen	Add 2 Hexose, Subtract Fucose
178.08	Hexose + Oxygen	Add 2 Hexose, Subtract Fucose
179.07	Hexose + Oxygen, monoisotopic peak misassignment	Add 2 Hexose, Subtract Fucose
179.08	Hexose + Oxygen, monoisotopic peak misassignment	Add 2 Hexose, Subtract Fucose
184.06	Hexose + Na	Add Hexose
187.11	HexNAc - Oxygen	Add HexNAc, Add Fucose, Subtract Hexose
187.12	HexNAc - Oxygen	Add HexNAc, Add Fucose, Subtract Hexose
188.12	HexNAc - Oxygen, monoisotopic peak misassignment	Add HexNAc, Add Fucose, Subtract Hexose
201.02	**2 Hexoses + Phosphate**	**Subtract HexNAc, Add 2 Hexose, Add Phosphate. #**
201.03	**2 Hexoses + Phosphate**	**Subtract HexNAc, Add 2 Hexose, Add Phosphate. #**
203.1	HexNAc	Add HexNAc
203.11	HexNAc	Add HexNAc
203.12	HexNAc	Add HexNAc
204.11	HexNAc, monoisotopic peak misassignment	Add HexNAc
241.08	**Hexose + Phosphate**	**Add Hexose, Add Phosphate**
242.04	**Hexose + Phosphate**	**Add Hexose, Add Phosphate**
242.05	**Hexose + Phosphate**	**Add Hexose, Add Phosphate**
243.04	**Hexose + Phosphate**	**Add Hexose, Add Phosphate**
251.11	**Hexose + 2 Fucose**	**Add Hexose, Add 2 Fucose, Subtract HexNAc. #**

Moiety Mass	Corresponds to:	Operation
267.11	**HexNAc - Oxygen + Phosphate**	**Add HexNAc, Add Fucose, Subtract Hexose, Add Phosphate**
280.99	**5 Hexoses + Phosphate**	**Subtract 3 HexNAc, Add 5 Hexoses, Add Phosphate. #**
283.07	**HexNAc + Phosphate**	**Add HexNAc, Add Phosphate**
307.12	2 Hexoses - Oxygen	Add Hexose, Add Fucose. #
308.13	2 Hexoses - Oxygen	Add Hexose, Add Fucose. #
308.14	2 Hexoses - Oxygen	Add Hexose, Add Fucose. #
308.15	2 Hexoses - Oxygen	Add Hexose, Add Fucose. #
309.15	2 Hexoses - Oxygen	Add Hexose, Add Fucose. #
323.12	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
323.13	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
323.14	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
324.11	2 Hexoses	Add 2 Hexoses
324.12	2 Hexoses	Add 2 Hexoses
324.13	2 Hexoses	Add 2 Hexoses
324.14	2 Hexoses	Add 2 Hexoses
324.15	2 Hexoses	Add 2 Hexoses
324.16	2 Hexoses	Add 2 Hexoses
325.12	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
325.13	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
325.14	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
325.15	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
326.13	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
326.14	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
327.15	2 Hexoses, monoisotopic peak misassignment	Add 2 Hexoses
341.16	2 Hexoses + Ammonia	Add 2 Hexoses

Discussion

In this work, we complete a head-to-head comparison of five contemporary glycoproteomic analytical software: Byonic, Protein Prospector, MSFraggerGlyco, pGlyco3, and GlycoDecipher. These five were chosen because of their blend of features that have different advantages. We note that this study focused on *N*-glycoproteins and did not include tools or analyses more tailored for *O*-glycoproteins.[88,99,100] Byonic searches for a glycan as a variable modification on the peptide. Protein Prospector and MSFraggerGlyco determine the difference in mass between the precursor and the unfragmented peptide backbone and use the mass offset to define the mass and composition of the glycan. pGlyco3 and GlycoDecipher filter spectra for those that contain sufficient Y-ions, which are unfragmented peptides with fragmented glycans, and construct the glycan from these fragments. Additionally, MSFraggerGlyco, pGlyco3, and GlycoDecipher all calculate a false discovery rate for the glycan to generate an estimate of confidence in reported results.

Based on our results, a single winner is not evident; but there are important lessons. One clear finding is that although Byonic was a gold standard in this field and has enabled the development of other tools, it appears to produce results of lower reliability than modern alternatives. It reports the most unique proteins and glycosites but does not identify the most glycopeptide spectra. By spot-checking, we show that some of these are misassignments. Additionally, almost 50% of the glycosites reported by Byonic could not be supported by UniProt. While Uniprot is in no way comprehensive, one can feel more confident in glycosite assignments that have been previously reported. It is possible that Byonic performs poorly because it attempts to identify all glycan and peptide ions at once and reports confidence based on scores for the presence of certain ions. In the example we show, Byonic may score noisy peaks very highly while ignoring critical information, such as the most abundant peaks in the spectra.

Glycoproteomic search engines should be selected based on the goals of the experiment. If one wants to identify the most glycopeptides as in an exploratory experiment, then mass offset approaches would be most favored. Protein Prospector was able to identify the most glycopeptides of any software. Among software calculating a glycan FDR, MSFraggerGlyco reported the most spectra. Whether the reported glycan FDR is accurate is open to question. Although Protein Prospector does not calculate a glycan FDR and instead uses a confidence score, it found the greatest number of unique combinations of protein, site, and glycan (PSG IDs). This did not sacrifice accuracy, at least with metrics used here for glycoprotein and glycosite assignments, since it still reported the least number of glycosites and maintained the highest agreement with UniProt.

If one wants to focus on the high-confidence structure of the glycan at a given site, pGlyco3 is a strong choice. Although it does not report the most glycopeptides, its Y-ion approach and glycan FDR calculation provide greater confidence in the assignment. Put differently, rather than inferring a glycan from its mass, it relies on direct evidence from peaks in the spectrum. Still, it has been reported to have a bias in assigning more fucosylated peptides although recent fixes seem to have solved this problem as other software reported more fucosylation than pGlyco3.[101] GlycoDecipher's *de novo* monosaccharide stepping suffers from problems with accurate glycan assignment. Manual inspection revealed that the overwhelming majority of glycans assigned as novel could be explained by glycans within the supplied glycan database. Despite that, it is possible to correct the results; so there is a clear path forward for this software to improve.

There is clear benefit to users trying more than one software for a given dataset. Although all software agreed on a core set of results, the next largest subsets were the unique results of a single software. At least two software, excluding Byonic, agreed on over half the results. It may prove useful to use one mass offset approach and one Y-ion approach. The mass offset approach will identify the most candidate glycopeptides. The Y-ion approach will provide those glycopeptides with strong Y-ion coverage, which is important for confident glycan assignment. Taken together, these approaches offer complementary information from the peptide and the glycan.

Finally, the use of glycomic profiling to generate a standardized glycan database provided greater confidence in assignments, enabled a rigorous comparison, and illustrated what features of a glycopeptide software are useful. There are currently few options to validate glycopeptide assignments. One can spot check for a specific protein by exo- or endoglycosidase digestion followed by Western blotting, but this is low throughput. One can use lectin microarrays, but these are custom and costly. Glycomic profiling is one way to corroborate the glycan assignments and to limit results to only those glycans which can be identified by other means. However, the GlycoDecipher *de novo* results did uncover that glycans outside of the database are present even after manual correction, so it is important that standards for as many glycans as possible are available during the glycomic analysis. Here, we used the glycomic results to ensure that software were operating in similar search spaces. While this may not benefit every researcher, it allowed a fair comparison of software and was instrumental to discovering the sensitivity of the mass offset approach.

Converting the glycomic results into a database for each glycoproteomic software was challenging. This is partly because some software use input formats that include glycan topology whereas others just use compositions; but even among software that use the same type of input, there are still formatting differences. For example, GlyTouCan provides a universal database for all glycans and can provide unique identifiers to different levels of resolution (monosaccharide composition, isomer composition, topology, and linkage), but using these identifiers would be extremely challenging for comparing software as each software reports different identifiers for the same assignment. A universal glycan database format would be critical to establishing better consistency between software and would eliminate a significant amount of work involved in reformatting databases.

REFERENCES

1. Heron, M. Deaths: leading causes for 2010. *Natl. Vital Stat. Rep. Cent. Dis. Control Prev. Natl. Cent. Health Stat. Natl. Vital Stat. Syst.* **62**, 1–96 (2013).
2. Rajan, K. B. *et al.* Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060). *Alzheimers Dement.* **17**, 1966–1975 (2021).
3. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement.* **18**, 700–789 (2022).
4. Dall, T. M. *et al.* An Aging Population And Growing Disease Burden Will Require ALarge And Specialized Health Care Workforce By 2025. *Health Aff. (Millwood)* **32**, 2013–2020 (2013).
5. Jutkowitz, E. *et al.* Societal and Family Lifetime Cost of Dementia: Implications for Policy. *J. Am. Geriatr. Soc.* **65**, 2169–2175 (2017).
6. Yaffe, K. *et al.* Effect of socioeconomic disparities on incidence of dementia among biracial older adults: prospective study. *BMJ* **347**, f7051 (2013).
7. Rajan, K. B., Weuve, J., Barnes, L. L., Wilson, R. S. & Evans, D. A. Prevalence and incidence of clinically diagnosed Alzheimer's disease dementia from 1994 to 2012 in a population study. *Alzheimers Dement.* **15**, 1–7 (2019).
8. Gurland, B. J. *et al.* Rates of dementia in three ethnoracial groups. *Int. J. Geriatr. Psychiatry* **14**, 481–493 (1999).
9. Budd Haeberlein, S. *et al.* Two Randomized Phase 3 Studies of Aducanumab in Early Alzheimer's Disease. *J. Prev. Alzheimers Dis.* **9**, 197–210 (2022).
10. van Dyck, C. H. *et al.* Lecanemab in Early Alzheimer's Disease. *N. Engl. J. Med.* **388**, 9–21 (2023).

11. Roses, A. D., M. D. Apolipoprotein E Alleles as Risk Factors in Alzheimer's Disease. *Annu. Rev. Med.* **47**, 387–400 (1996).

12. Pericak-Vance, M. A. & Haines, J. L. Genetic susceptibility to Alzheimer disease. *Trends Genet.* **11**, 504–508 (1995).

13. Loy, C. T., Schofield, P. R., Turner, A. M. & Kwok, J. B. Genetics of dementia. *The Lancet* **383**, 828–840 (2014).

14. Holtzman, D., Herz, J. & Bu, G. Apolipoprotein E and Apolipoprotein E Receptors: Normal Biology and Roles in Alzheimer Disease. *Cold Spring Harb. Perspect. Med.* **2**, a006312 (2012).

15. Michaelson, D. M. APOE ε4: The most prevalent yet understudied risk factor for Alzheimer's disease. *Alzheimers Dement.* **10**, 861–868 (2014).

16. Corder, E. H. *et al.* Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. *Science* **261**, 921–923 (1993).

17. Saunders, A. M. *et al.* Association of apolipoprotein E allele ϵ4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467–1467 (1993).

18. Farrer, L. A. *et al.* Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease: A Meta-analysis. *JAMA* **278**, 1349–1356 (1997).

19. Neu, S. C. *et al.* Apolipoprotein E Genotype and Sex Risk Factors for Alzheimer Disease: A Meta-analysis. *JAMA Neurol.* **74**, 1178–1189 (2017).

20. Genin, E. *et al.* APOE and Alzheimer disease: a major gene with semi-dominant inheritance. *Mol. Psychiatry* **16**, 903–907 (2011).

21. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* **39**, 17–23 (2007).

22. Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C. & Bu, G. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nat. Rev. Neurol.* **15**, 501–518 (2019).

23. Corder, E. H. *et al.* Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat. Genet.* **7**, 180–184 (1994).

24. Weisgraber, K. H., Rall, S. C. & Mahley, R. W. Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *J. Biol. Chem.* **256**, 9077–9083 (1981).

25. Lindner, K. *et al.* Isoform- and cell-state-specific lipidation of ApoE in astrocytes. *Cell Rep.* **38**, (2022).

26. Stuchell-Brereton, M. D. *et al.* Apolipoprotein E4 has extensive conformational heterogeneity in lipid-free and lipid-bound forms. *Proc. Natl. Acad. Sci.* **120**, e2215371120 (2023).

27. Hatters, D. M., Budamagunta, M. S., Voss, J. C. & Weisgraber, K. H. Modulation of Apolipoprotein E Structure by Domain Interaction: DIFFERENCES IN LIPID-BOUND AND LIPID-FREE FORMS *. *J. Biol. Chem.* **280**, 34288–34295 (2005).

28. Ruiz, J. *et al.* The apoE isoform binding properties of the VLDL receptor reveal marked differences from LRP and the LDL receptor. *J. Lipid Res.* **46**, 1721–1731 (2005).

29. Chen, J., Li, Q. & Wang, J. Topology of human apolipoprotein E3 uniquely regulates its diverse biological functions. *Proc. Natl. Acad. Sci.* **108**, 14813–14818 (2011).

30. Schmukler, E. *et al.* Altered mitochondrial dynamics and function in APOE4-expressing astrocytes. *Cell Death Dis.* **11**, 1–13 (2020).

31. Yin, J. *et al.* Effect of ApoE isoforms on mitochondria in Alzheimer disease. *Neurology* **94**, e2404–e2411 (2020).

32. Orr, A. L. *et al.* Neuronal Apolipoprotein E4 Expression Results in Proteome-Wide Alterations and Compromises Bioenergetic Capacity by Disrupting Mitochondrial Function. *J. Alzheimers Dis.* **68**, 991–1011 (2019).

33. Wu, L., Zhang, X. & Zhao, L. Human ApoE Isoforms Differentially Modulate Brain Glucose and Ketone Body Metabolism: Implications for Alzheimer's Disease Risk Reduction and Early Intervention. *J. Neurosci.* **38**, 6665–6681 (2018).

34. Williams, H. C. *et al.* APOE alters glucose flux through central carbon pathways in astrocytes. *Neurobiol. Dis.* **136**, 104742 (2020).

35. Johnson, E. C. B. *et al.* Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. *Nat. Neurosci.* **25**, 213–225 (2022).

36. Tcw, J. *et al.* Cholesterol and matrisome pathways dysregulated in astrocytes and microglia. *Cell* **185**, 2213-2233.e25 (2022).

37. Grehan, S., Tse, E. & Taylor, J. M. Two Distal Downstream Enhancers Direct Expression of the Human Apolipoprotein E Gene to Astrocytes in the Brain. *J. Neurosci.* **21**, 812–822 (2001).

38. Guttenplan, K. A. *et al.* Neurotoxic reactive astrocytes induce cell death via saturated lipids. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03960-y.

39. Xu, Q. *et al.* Profile and Regulation of Apolipoprotein E (ApoE) Expression in the CNS in Mice with Targeting of Green Fluorescent Protein Gene to the ApoE Locus. *J. Neurosci.* **26**, 4985–4994 (2006).

40. Xu, P.-T. *et al.* Sialylated Human Apolipoprotein E (apoEs) Is Preferentially Associated with Neuron-Enriched Cultures from APOE Transgenic Mice. *Neurobiol. Dis.* **6**, 63–75 (1999).

41. Qi, G. *et al.* ApoE4 Impairs Neuron-Astrocyte Coupling of Fatty Acid Metabolism. *Cell Rep.* **34**, 108572 (2021).

42. Dekroon, R. M. & Armati, P. J. ENDOCYTOSIS OF apoE-EGFP BY PRIMARY HUMAN BRAIN CULTURES. *Cell Biol. Int.* **26**, 761–770 (2002).

43. Yeh, F. L., Wang, Y., Tom, I., Gonzalez, L. C. & Sheng, M. TREM2 Binds to Apolipoproteins, Including APOE and CLU/APOJ, and Thereby Facilitates Uptake of Amyloid-Beta by Microglia. *Neuron* **91**, 328–340 (2016).

44. Blanchard, J. W. *et al.* APOE4 impairs myelination via cholesterol dysregulation in oligodendrocytes. *Nature* **611**, 769–779 (2022).

45. Shi, Y. *et al.* Microglia drive APOE-dependent neurodegeneration in a tauopathy mouse model. *J. Exp. Med.* **216**, 2546–2561 (2019).

46. Mahley, R. W., Weisgraber, K. H. & Huang, Y. Apolipoprotein E4: A causative factor and therapeutic target in neuropathology, including Alzheimer’s disease. *Proc. Natl. Acad. Sci.* **103**, 5644–5651 (2006).

47. Koutsodendris, N. *et al.* Neuronal APOE4 removal protects against tau-mediated gliosis, neurodegeneration and myelin deficits. *Nat. Aging* 1–22 (2023) doi:10.1038/s43587-023-00368-3.

48. Lam, S. S. *et al.* Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat. Methods* **12**, 51–54 (2015).

49. Tan, B. *et al.* An Optimized Protocol for Proximity Biotinylation in Confluent Epithelial Cell Cultures Using the Peroxidase APEX2. *STAR Protoc.* **1**, 100074 (2020).

50. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

51. SAINTexpress: Improvements and additional features in Significance Analysis of INTeractome software - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S1874391913005381?via%3Dihub.

52. *Essentials of Glycobiology*. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2022).

53. Udeshi, N. D. *et al.* Antibodies to biotin enable large-scale detection of biotinylation sites on proteins. *Nat. Methods* **14**, 1167–1170 (2017).

54. Chen, Z. *et al.* In-depth Site-specific Analysis of N-glycoproteome in Human Cerebrospinal Fluid and Glycosylation Landscape Changes in Alzheimer's Disease. *Mol. Cell. Proteomics MCP* **20**, 100081 (2021).

55. Zhang, Q., Ma, C., Chin, L.-S. & Li, L. Integrative glycoproteomics reveals protein N-glycosylation aberrations and glycoproteomic network alterations in Alzheimer's disease. *Sci. Adv.* **6**, eabc5802 (2020).

56. Wang, J.-Z., Grundke-Iqbal, I. & Iqbal, K. Glycosylation of microtubule–associated protein tau: An abnormal posttranslational modification in Alzheimer's disease. *Nat. Med.* **2**, 871–875 (1996).

57. Losev, Y. *et al.* Novel model of secreted human tau protein reveals the impact of the abnormal N-glycosylation of tau on its aggregation propensity. *Sci. Rep.* **9**, 2254 (2019).

58. Lonati, E. *et al.* Ischemic Conditions Affect Rerouting of Tau Protein Levels: Evidences for Alteration in Tau Processing and Secretion in Hippocampal Neurons. *J. Mol. Neurosci. MN* **66**, 604–616 (2018).

59. Kang, S., Son, S. M., Baik, S. H., Yang, J. & Mook-Jung, I. Autophagy-Mediated Secretory Pathway is Responsible for Both Normal and Pathological Tau in Neurons. *J. Alzheimers Dis. JAD* **70**, 667–680 (2019).

60. Pernègre, C., Duquette, A. & Leclerc, N. Tau Secretion: Good and Bad for Neurons. *Front. Neurosci.* **13**, 649 (2019).

61. Reiman, E. M. *et al.* Preclinical Evidence of Alzheimer's Disease in Persons Homozygous for the ε4 Allele for Apolipoprotein E. *N. Engl. J. Med.* **334**, 752–758 (1996).

62. Reiman, E. M. *et al.* Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 284–289 (2004).

63. Small, G. W. *et al.* Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease. *Proc. Natl. Acad. Sci.* **97**, 6037–6042 (2000).

64. Zhao, N. *et al.* Apolipoprotein E4 Impairs Neuronal Insulin Signaling by Trapping Insulin Receptor in the Endosomes. *Neuron* **96**, 115-129.e5 (2017).

65. Tena, J. *et al.* Glycosylation alterations in serum of Alzheimer's disease patients show widespread changes in N-glycosylation of proteins related to immune function, inflammation, and lipoprotein metabolism. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **14**, e12309 (2022).

66. Theendakara, V. *et al.* Direct Transcriptional Effects of Apolipoprotein E. *J. Neurosci.* **36**, 685–700 (2016).

67. Rohn, T. T. & Moore, Z. D. Nuclear Localization of Apolipoprotein E4: A New Trick for an Old Protein. *Int. J. Neurol. Neurother.* **4**, 067 (2017).

68. Conroy, L. R., Hawkinson, T. R., Young, L. E. A., Gentry, M. S. & Sun, R. C. Emerging roles of N-linked glycosylation in brain physiology and disorders. *Trends Endocrinol. Metab.* **32**, 980–993 (2021).

69. Sun, R. C. *et al.* Brain glycogen serves as a critical glucosamine cache required for protein glycosylation. *Cell Metab.* **33**, 1404-1417.e9 (2021).

70. Bagdonaite, I. *et al.* Glycoproteomics. *Nat. Rev. Methods Primer* **2**, 1–29 (2022).

71. Apweiler, R., Hermjakob, H. & Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database11Dedicated to Prof. Akira Kobata and Prof. Harry Schachter on the occasion of their 65th birthdays. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1473**, 4–8 (1999).

72. Zacchi, L. F. & Schulz, B. L. N-glycoprotein macroheterogeneity: biological implications and proteomic characterization. *Glycoconj. J.* **33**, 359–376 (2016).

73. Castelli, M. *et al.* How aberrant N-glycosylation can alter protein functionality and ligand binding: An atomistic view. *Structure* **31**, 987-1004.e8 (2023).

74. Guay, K. P. *et al.* ER chaperones use a protein folding and quality control glyco-code. *Mol. Cell* **83**, 4524-4537.e5 (2023).

75. Oliveira, T., Thaysen-Andersen, M., Packer, N. H. & Kolarich, D. The Hitchhiker's guide to glycoproteomics. *Biochem. Soc. Trans.* **49**, 1643–1662 (2021).

76. Mariño, K., Bones, J., Kattla, J. J. & Rudd, P. M. A systematic approach to protein glycosylation analysis: a path through the maze. *Nat. Chem. Biol.* **6**, 713–723 (2010).

77. Riley, N. M., Bertozzi, C. R. & Pitteri, S. J. A Pragmatic Guide to Enrichment Strategies for Mass Spectrometry-Based Glycoproteomics. *Mol. Cell. Proteomics MCP* **20**, 100029 (2021).

78. Riley, N. M., Malaker, S. A., Driessen, M. D. & Bertozzi, C. R. Optimal Dissociation Methods Differ for N- and O-Glycopeptides. *J. Proteome Res.* **19**, 3286–3301 (2020).

79. Malaker, S. A. Glycoproteomics: Charting new territory in mass spectrometry and glycobiology. *J. Mass Spectrom.* **59**, e5034 (2024).

80. Kawahara, R. *et al.* Community evaluation of glycoproteomics informatics solutions reveals high-performance search strategies for serum glycopeptide analysis. *Nat. Methods* **18**, 1304–1316 (2021).

81. Bern, M., Kil, Y. J. & Becker, C. Byonic: Advanced Peptide and Protein Identification Software. *Curr. Protoc. Bioinforma.* **40**, 13.20.1-13.20.14 (2012).

82. Roushan, A. *et al.* Peak Filtering, Peak Annotation, and Wildcard Search for Glycoproteomics. *Mol. Cell. Proteomics* **20**, 100011 (2021).

83. Polasky, D. A., Yu, F., Teo, G. C. & Nesvizhskii, A. I. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods* **17**, 1125–1132 (2020).

84. Polasky, D. A., Geiszler, D. J., Yu, F. & Nesvizhskii, A. I. Multiattribute Glycan Identification and FDR Control for Glycoproteomics. *Mol. Cell. Proteomics* **21**, 100205 (2022).

85. Medzihradszky, K. F., Kaasik, K. & Chalkley, R. J. Tissue-Specific Glycosylation at the Glycopeptide Level*. *Mol. Cell. Proteomics* **14**, 2103–2110 (2015).

86. Zeng, W.-F., Cao, W.-Q., Liu, M.-Q., He, S.-M. & Yang, P.-Y. Precise, fast and comprehensive analysis of intact glycopeptides and modified glycans with pGlyco3. *Nat. Methods* **18**, 1515–1523 (2021).

87. Fang, Z. *et al.* Glyco-Decipher enables glycan database-independent peptide matching and in-depth characterization of site-specific N-glycosylation. *Nat. Commun.* **13**, 1900 (2022).

88. Rangel-Angarita, V., Mahoney, K. E., Ince, D. & Malaker, S. A. A systematic comparison of current bioinformatic tools for glycoproteomics data. 2022.03.15.484528 Preprint at https://doi.org/10.1101/2022.03.15.484528 (2022).

89. Shen, J. *et al.* StrucGP: de novo structural sequencing of site-specific N-glycan on glycoproteins using a modularization strategy. *Nat. Methods* **18**, 921–929 (2021).

90. Protocols. *ProtiFi* https://protifi.com/pages/protocols.

91. Ashwood, C., Pratt, B., MacLean, B. X., Gundry, R. L. & Packer, N. H. Standardization of PGC-LC-MS-based glycomics for sample specific glycotyping. *Analyst* **144**, 3601–3612 (2019).

92. Ceroni, A. *et al.* GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans. *J. Proteome Res.* **7**, 1650–1659 (2008).

93. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

94. Bermudez, A. & Pitteri, S. J. Enrichment of Intact Glycopeptides Using Strong Anion Exchange and Electrostatic Repulsion Hydrophilic Interaction Chromatography. in *Mass Spectrometry of Glycoproteins: Methods and Protocols* (ed. Delobel, A.) 107–120 (Springer US, New York, NY, 2021). doi:10.1007/978-1-0716-1241-5_8.

95. Rad, R. *et al.* Improved Monoisotopic Mass Estimation for Deeper Proteome Coverage. *J. Proteome Res.* **20**, 591–598 (2021).

96. OpenAI. ChatGPT (July 4 Version) [Large language model]. (2024).

97. Lee, J. *et al.* Spatial and temporal diversity of glycome expression in mammalian brain. *Proc. Natl. Acad. Sci.* **117**, 28743–28753 (2020).

98. Riley, N. M., Hebert, A. S., Westphall, M. S. & Coon, J. J. Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nat. Commun.* **10**, 1311 (2019).

99. Lu, L., Riley, N. M., Shortreed, M. R., Bertozzi, C. R. & Smith, L. M. O-Pair Search with MetaMorpheus for O-glycopeptide characterization. *Nat. Methods* **17**, 1133–1138 (2020).

100. Polasky, D. A. *et al.* Quantitative proteome-wide O-glycoproteomics analysis with FragPipe. *Anal. Bioanal. Chem.* (2024) doi:10.1007/s00216-024-05382-x.

101. Adams, T. M., Zhao, P., Kong, R. & Wells, L. ppmFixer: a mass error adjustment for pGlyco3.0 to correct near-isobaric mismatches. *Glycobiology* cwae006 (2024) doi:10.1093/glycob/cwae006.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

> I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

Signed by:

8C0109B5E0AF4DD...

Author Signature

8/1/2024

DATE