Multiattribute Glycan Identification and FDR Control for Glycoproteomics

Authors

Daniel A. Polasky, Daniel J. Geiszler, Fengchao Yu, and Alexey I. Nesvizhskii

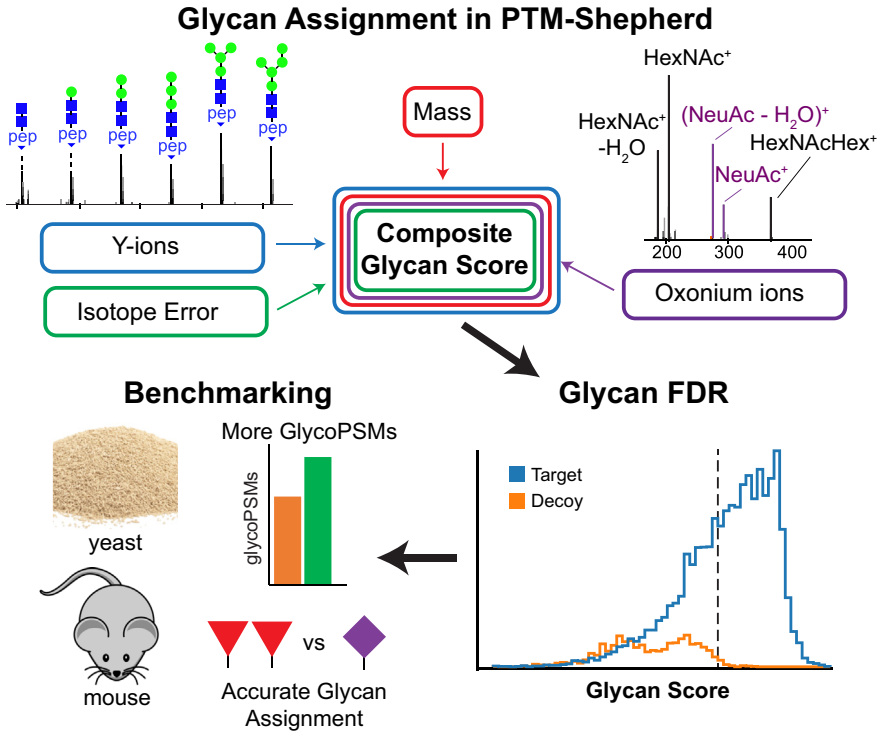
Correspondence

nesvi@med.umich.edu

In Brief

Glycoproteomics has seen rapid advances in methods for identifying glycopeptides, but challenges remain confidently determining the composition and structure of the attached glycan. We have developed a method using multiple sources of information from the mass spectrum to assign the composition of N-linked glycopeptides and an associated method for false discovery rate control. We show that this method is able to identify more glycopeptide spectra while also providing more accurate composition assignments than existing tools.

Graphical Abstract



Highlights

- Identifying the glycan on intact glycopeptides remains difficult in glycoproteomics.
- We developed a method to assign glycan compositions in N-glycoproteomics searches.
- We demonstrate well-controlled glycan FDR in multiple sample types.
- The method annotates more glycopeptide spectra than competing tools.
- The method is included PTM-Shepherd for a full glycoproteomics workflow in FragPipe.

2022, Mol Cell Proteomics *21*(*3*), 100205

https://doi.org/10.1016/j.mcpro.2022.100205

Multiattribute Glycan Identification and FDR Control for Glycoproteomics

Daniel A. Polasky[1], Daniel J. Geiszler[2], Fengchao Yu[1], and Alexey I. Nesvizhskii[1,2,*]

Rapidly improving methods for glycoproteomics have enabled increasingly large-scale analyses of complex glycopeptide samples, but annotating the resulting mass spectrometry data with high confidence remains a major bottleneck. We recently introduced a fast and sensitive glycoproteomics search method in our MSFragger search engine, which reports glycopeptides as a combination of a peptide sequence and the mass of the attached glycan. In samples with complex glycosylation patterns, converting this mass to a specific glycan composition is not straightforward; however, as many glycans have similar or identical masses. Here, we have developed a new method for determining the glycan composition of N-linked glycopeptides fragmented by collisional or hybrid activation that uses multiple sources of information from the spectrum, including observed glycan B-type (oxonium) and Y-type ions and mass and precursor monoisotopic selection errors to discriminate between possible glycan candidates. Combined with false discovery rate estimation for the glycan assignment, we show that this method is capable of specifically and sensitively identifying glycans in complex glycopeptide analyses and effectively controls the rate of false glycan assignments. The new method has been incorporated into the PTM-Shepherd modification analysis tool to work directly with the MSFragger glyco search in the FragPipe graphical user interface, providing a complete computational pipeline for annotation of N-glycopeptide spectra with false discovery rate control of both peptide and glycan components that is both sensitive and robust against false identifications.

Glycosylation is one of the most common post-translational modifications (PTMs) of proteins, involved in a vast array of biological processes and implicated in numerous diseases (1–4). Because of the analytical challenges resulting from the heterogeneity of glycosylation, both in sites occupied and glycans present at a given site, analysis of the glycoproteome has generally lagged behind other omics fields (5). Improvements to methods for enriching, separating, and analyzing glycopeptides by mass spectrometry have been accelerating in recent years (5–7), however, resulting in increasingly large and complex glycoproteomics data being generated. Analysis of these data has represented a significant bottleneck in glycoproteomics (8), particularly for proteome-scale analysis of intact glycopeptides. A rapid expansion of software tools is underway in this area, with many new methods capable of this type of analysis being reported recently (9–18).

Statistical control of the results reported by these new tools has fallen behind, however, in large part to the extra challenges of correctly identifying intact glycopeptide spectra (19). As a result, despite the accelerating development of these tools, it remains common practice to manually validate and/or empirically filter search results to remove incorrect glycan composition assignments (20), presenting a major bottleneck for large-scale glycoproteomics studies. Many tools for glycoproteomics data analysis adapt methods from proteomics for glycoproteomics by treating glycans similarly to other PTMs or chemical modifications of peptides. Some search tools provide additional capabilities that can assist in controlling the false discovery rate (FDR) of modified peptides, such as the use of the extended mass model of PeptideProphet (21) to model distinct probabilities for modifications with different masses used with MSFragger (17), or distinguishing between rare and common modifications in Byonic (12). These tools and many others have increasingly been applied to large-scale glycoproteomics analyses (14, 18, 22–25) utilizing peptide-focused FDR methods, often in conjunction with a second empirical filtering or manual validation step.

Several studies have pointed out, however, that peptide-focused FDR approaches can fall short for glycoproteomics analyses because of the complexity and heterogeneity of glycans (8, 19, 26, 27). There can be hundreds of different glycans present in an individual glycoproteomics analysis with frequencies that vary over multiple orders of magnitude. Furthermore, N-glycans are comprised of a common core and various extensions, often containing repeating carbohydrate units. There are some residue combinations that are isomeric (*e.g.*, *N*-glycolyl neuraminic acid [NeuGc] plus fucose has the same atomic composition and exact mass as *N*-acetyl neuraminic acid [NeuAc] plus hexose) and several more that are

From the [1]Department of Pathology, and [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

*For correspondence: Alexey I. Nesvizhskii, nesvi@med.umich.edu.

https://doi.org/10.1016/j.mcpro.2022.100205

very similar in mass to other combinations or to peptide modifications (28–30). Errors in assigning the monoisotopic mass of the precursor, also called "off-by-X" or peakpicking errors, result in a glycan mass that is off by 1 (or several) Da, and there are several additional combinations of common carbohydrate residues or peptide modifications that are very similar in mass to another combination plus such an isotope error (20). Treating glycopeptides as modified peptides and using adapted proteomics methods may thus be sufficient for analyses in which the glycan population is relatively simple and well characterized, and any overlapping glycan masses can be watched for and resolved manually if needed. This manual verification remains a major bottleneck of glycoproteomics analyses, however, and often precludes large-scale analysis of more complex or uncharacterized glycan populations where such overlaps are common.

Several recent methods have been proposed in which a target-decoy analysis is performed specifically on the glycan-matching portion of glycopeptide-spectrum matching to generate a glycan-specific FDR or "glycan FDR" (10, 13, 31). In more recent examples, this has been combined with a "peptide FDR" typical to modern proteomics methods to evaluate the quality of match between the spectrum and both the proposed peptide sequence and glycan composition (9, 11, 32, 33). This approach has the potential to enable automated analysis of glycoproteomics data with complex and uncharacterized glycan populations and remove the manual validation bottleneck. The proposed methods thus far have generally used a "glycan-first" approach, in which possible glycan candidates are first identified by matching the Y-ion series from the spectrum, then the determined glycan mass is subtracted from the observed precursor to determine the peptide mass, and finally the spectrum is searched for matching peptide fragment ions from precursors matching the determined peptide mass. While this method has been shown to be effective at controlling glycan FDR, it is limited to data in which abundant Y-ions are produced, making it challenging to adapt for O-glycoproteomics and potentially its reducing sensitivity when fragmentation conditions are not optimal for producing Y-ions.

Here, we propose a new approach for identifying the glycan component of an N-linked glycopeptide and an associated glycan FDR estimation method with two major differences from the existing methods. We first identify the peptide sequence, using a mass offset-style glyco search from MSFragger (17, 34), then match the mass difference between the peptide sequence mass and the observed precursor mass to candidate glycans to determine the composition. This "peptide-first" approach leverages the well-developed capabilities of modern proteomics methods to solve the peptide portion of glycopeptide identification first, reducing the glycan identification portion to distinguishing between a few glycans that match the mass difference from the determined peptide sequence, rather than distinguishing between the complete search list of up to hundreds of possible glycan compositions. Second, we generate a composite glycan score from a variety of spectral evidence, including Y-ions, oxonium ions, and the observed mass and precursor isotope errors, rather than just Y-ions alone. We demonstrate that by simplifying the problem by matching the peptide first using our MSFragger search tool and then maximizing the glycan-specific information gleaned from spectra, we are able to annotate many more glycopeptide spectra at the same FDR as existing glycan-first methods. Unlike our previous peptide-only FDR approach, we show that this method controls glycan FDR across many search scenarios, including when searching for entrapment glycans known not to be present in the sample, while maintaining the high sensitivity of the MSFragger glyco search. The method has been implemented in the open-source tool PTM-Shepherd (35), version1.2 and has been incorporated into the FragPipe graphical interface and pipeline to provide a complete solution for glycoproteomics analyses.

EXPERIMENTAL PROCEDURES

Dataset Details

Raw data were downloaded from ProteomeXchange (36) repositories and converted to mzML with MSConvert (version 3.0.19296-ebe17a86f) (37). The "yeast" dataset from PXD005565 contains glycopeptides enriched by ZIC-hydrophilic interaction chromatography from fission yeast (*Schizosaccharomyces pombe*) analyzed by stepped-energy higher energy collisional dissociation (HCD) on an Orbitrap Fusion mass spectrometer (9). The "Riley" dataset from PXD011533 contains glycopeptides enriched by lectin affinity chromatography from mouse brain tissue analyzed by HCD and activated ion–electron transfer dissociation (AI–ETD) fragmentation on an Orbitrap Fusion Lumos mass spectrometer (38). The "mouse" 5-tissue dataset contains glycopeptides enriched from mouse brain (PXD005411), kidney (PXD005412), heart (PXD005413), liver (PXD005553), and lung (PXD005555) analyzed by stepped-energy HCD on an Orbitrap Fusion mass spectrometer (9).

Glycoproteomics Searches and Peptide FDR

Analysis of the data was performed in two parts: first, glycopeptide search using MSFragger's glyco mode with peptide validation and FDR filtering in Philosopher, and second, glycan assignment and glycan FDR filtering in PTM-Shepherd. MSFragger glyco search, described previously (17), produces a list of peptide-spectrum matches (PSMs) for both nonglyco and glycopeptides found, in which glycopeptides are matched as a peptide and a delta mass corresponding to the mass of the glycan (Fig. 1*A*).

"Yeast" data from PXD005565 were converted to mzML format using MSConvert (37, 39) using default (vendor) peakpicking and searched against a combined yeast and mouse proteome database with common contaminant proteins and decoys appended in Philosopher (downloaded February 8, 2021; 22,307 nondecoy entries). Searches were performed against two glycan lists: a "yeast" list containing only glycans with two HexNAc residues and 4 to 20 Hex residues (17 total compositions) and a combined yeast and mouse glycan list, equivalent to the "pGlyco-N-mouse-large" list (containing 1670 unique compositions). This combined yeast plus mouse list was the same as used in the analysis of the yeast dataset in the articles describing the pGlyco2 (9) and pGlyco3 (11) software packages. One

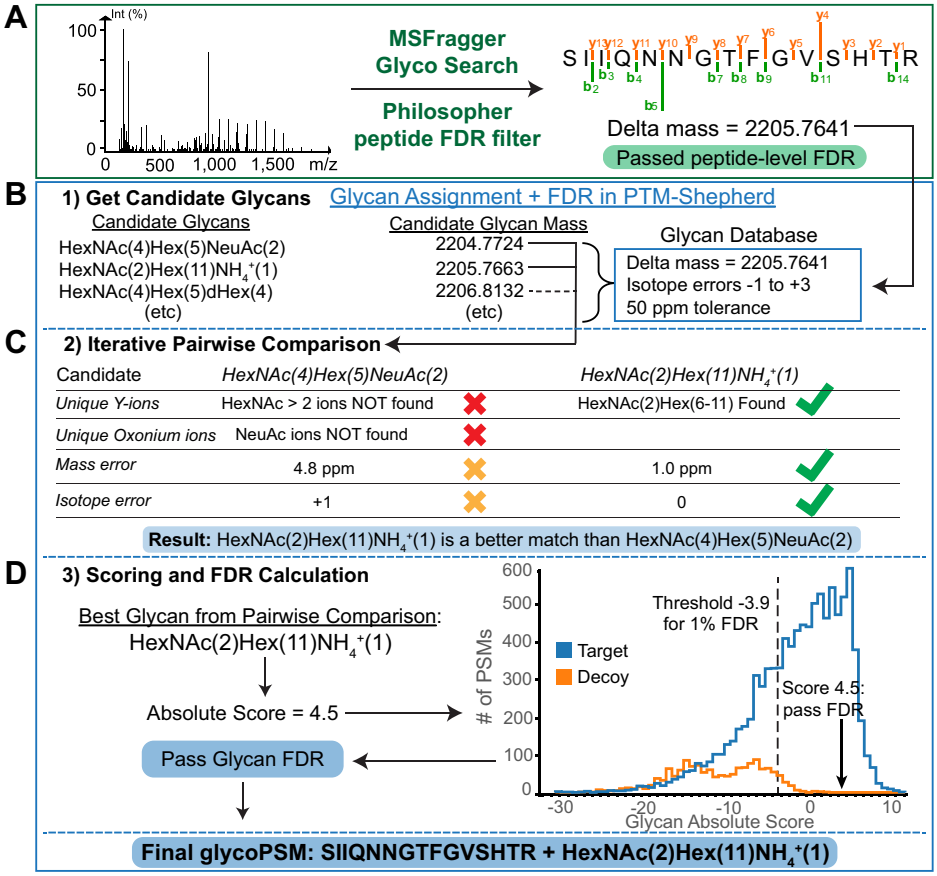


FIG. 1. **Glycan assignment workflow in PTM-Shepherd.** *A*, glyco search in MSFragger and peptide FDR filtering in Philosopher annotates a spectrum with a peptide and a delta mass. *B*, possible glycan candidates with masses similar to the provided delta mass are gathered from the internal glycan list in PTM-Shepherd. *C*, pairwise comparison of all candidates determines the best match to the spectrum based on unique fragment ions for each candidate and mass and isotope errors. Note that mass error is calculated after first correcting any isotope error. *Green check marks* in a candidate's column indicate positive evidence for that candidate, *red* and *orange* x's indicate negative or slightly negative evidence, respectively. *D*, the best candidate is rescored to generate the absolute score, and FDR is computed using the score distribution of target and decoy glycans. FDR, false discovery rate.

ammonium adduct was allowed, increasing the total number of unique mass offsets searched in MSFragger to 34 for the yeast-only search or 2325 for the yeast–mouse combined search (redundant masses [rounded to two decimal places] were removed). The glycans represented in each glycan list can be found in supplemental Data 1. MSFragger (version 3.4) searches were performed against the combined yeast–mouse database, allowing two missed cleavages by trypsin, fixed carbamidomethylation of Cys, and variable Met oxidation and protein N-terminal acetylation, in n-glycan mode with precursor and fragment mass tolerances of 20 and 10 ppm, respectively, precursor isotope error correction enabled using the built-in correction algorithm, *b*, *y*, *b* + HexNAc, *y* + HexNAc, and Y-ions considered, and oxonium ion filtering enabled (minimum 10% summed oxonium ion intensity relative to the base peak of the spectrum to search for glycopeptides) with default oxonium ion masses. Following MSFragger search, Philosopher (version 4.0.0) (40) was used to perform peptide FDR filtering. PeptideProphet (21) with extended mass model (mass width of 4000) was used to model PSM probabilities in semiparametric mode with n-glycan motif modeling enabled and cLevel set to 0. ProteinProphet (41) protein inference was performed with default parameters except maxppmdiff set to 20000000 to prevent exclusion of glycopeptides with large delta masses. The final PSMs, peptides, and proteins were filtered to 1% PSM and protein FDR, using a sequential filtering step to remove PSMs and peptides from proteins that did not pass FDR. Raw entrapment rates were calculated as the number of glycoPSMs assigned to any glycan other than the 17 yeast glycan compositions out of the total number of glycoPSMs. Adjusted entrapment rates were calculated by multiplying the raw entrapment rate by the ratio of the relative size of the true glycan list (17) to the total glycan list (1670).

Results from a pGlyco3 primary search (without ammonium adducts) of the yeast dataset shown in Figure 2 performed as part of (11) were downloaded from the Massive repository MSV000086771. pGlyco3 (build 20210615) was used to perform an equivalent search with one ammonium adduct. The same combined yeast–mouse proteome database as the MSFragger/PTM-Shepherd search (aforementioned) was used and the "p-Glyco-N-mouse-Large" glycan database. Raw files were parsed with pParse using default parameters and peakpicking. Parameters for protein digestion, peptide modifications, and mass tolerances were set to the same values as in the MSFragger search. One "aH" variable modification was allowed on glycans (ammonium adduct). Default FDR options were employed (1% peptide and glycan) using the built-in peptide FDR method.

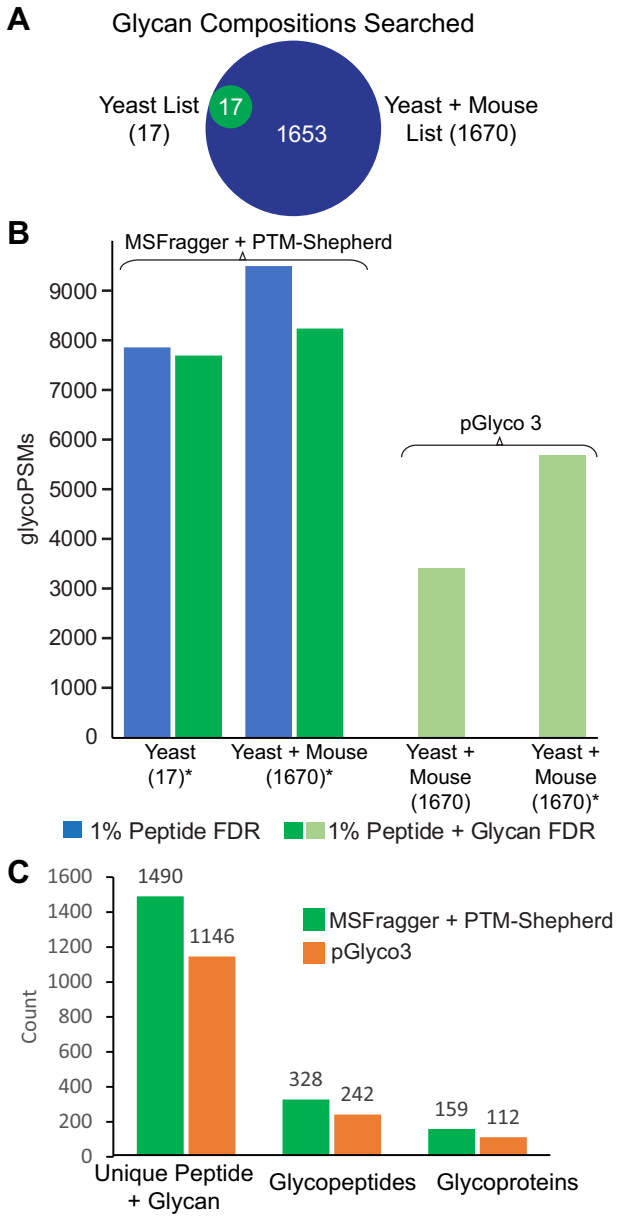


FIG. 2. **Results of entrapment searches of yeast dataset.** *A*, Venn diagram of the two glycan lists used in MSFragger searches. *B*, GlycoPSMs annotated at 1% peptide FDR (*blue*) and combined 1% peptide and glycan FDR (*green*) for MSFragger searches of the glycan lists with PTM-Shepherd glycan FDR filtering. *Indicates ammonium adduction was allowed. Comparison with pGlyco 3 reported results for the same peptide database, glycan list, and FDR levels is at right either with or without allowing one ammonium adduct per glycan (encoded by pGlyco3 as "aH" residues). *C*, unique glycoproteins, glycopeptide sequences, and glycan–peptide combinations (each glycan composition on each unique peptide sequence counts as a separate entry) detected in MSFragger + PTM-Shepherd (*green*) and pGlyco3 (*orange*) results at 1% peptide and glycan FDR. FDR, false discovery rate; PSM, peptide-spectrum match.

The "Riley" dataset from PXD011533 was searched and FDR filtered with the same method and parameters except for the following differences. HCD and AI–ETD scans were extracted to separate mzML files with default (vendor) peakpicking in MSConvert and searched separately before results were combined for validation and FDR filtering. Spectra were searched against mouse glycoprotein-focused database from Riley *et al.* (38) with decoys appended in Philosopher (3574 nondecoy entries), allowing up to three missed cleavages, and the mouse glycan list (182 unique masses). AI–ETD search considered *b*, *y*, *c*, *z*, and Y-ions and reduced required oxonium ion intensity to 2.5%.

"Mouse" data from PXD005411, PXD005412, PXD005413, PXD005553, and PXD005555 were converted to mzML with MSConvert using default (vendor) peakpicking and searched against a full reviewed mouse proteome with common contaminant proteins and decoys appended in Philosopher (downloaded September 24, 2019; 17,019 nondecoy entries). The same list of glycans was used as in the yeast analysis, equivalent to the "pGlyco-N-Mouse-Large" glycan list with one ammonium adduct allowed. All other MSFragger parameters were the same as in the yeast searches. An entrapment search of the mouse data was also performed, with 248 entrapment glycan compositions added to the MSFragger search and PTM-Shepherd glycan list (see supplemental Data 1 for the list of entrapment glycans). About 110 additional mouse glycans were added to correct missing compositions in the original 1670 list. Raw and adjusted entrapment glycan rates were calculated as in the yeast data, using 1780 *versus* 2028 (1780 + 248) glycans as the true and total glycan list sizes.

Comparative searches of the mouse data were performed with pGlyco3 (build 20210615) using the same mouse protein database and glycan list with one "aH" variable glycan modification (ammonium adduct) allowed. Raw data were read with pParse using default parameters and peakpicking. Variable peptide modifications were set to the same as MSFragger search (protein N-terminal acetylation and Met oxidation), and precursor and fragment mass tolerances were set to 20 and 10 ppm, respectively, as in the MSFragger search. Default FDR options were used (1% peptide and glycan FDR) using the built-in peptide FDR method.

Converting Delta Mass to Glycan Composition in PTM-Shepherd

PTM-Shepherd (35) reads PSM results, including the identified peptide and delta mass, from the output of MSFragger and Philosopher. Determining the identity of the glycan from the delta mass begins by determining possible glycans with intact masses near the observed delta mass from a provided list of glycan compositions or PTM-Shepherd's internal glycan list (Fig. 1*B*). This internal list is constructed from reported N-glycans from several glycoproteomics analyses and databases with the intent of including most known mammalian glycan compositions so that it can be used without modification for a range of glycoproteomics analyses (the list can be found in supplemental Data 1). Custom lists of glycan compositions can also be supplied by the user and should be used when analyzing samples with glycosylation that differs markedly from mammalian N-glycosylation. If specified by the user, adducted forms of glycans can be considered, for example, the replacement of a proton by an ammonium adduct, which does not change the expected fragment ions as the noncovalent adduct is not expected to be retained. The current version of the method supports only peptides with a single glycosylation event, though glycopeptides with multiple potential glycosylation sites are not explicitly excluded from analysis and can thus be matched if only one site is in fact glycosylated.

To allow for FDR control, a decoy glycan is appended to the list for each target (and each target adduct, if specified). A decoy glycan candidate is generated by shifting the intact mass of a target glycan by a random value within the provided glycan mass error tolerance and assigning a randomly chosen isotope error from the set of such errors being considered in the analysis. This allows us to distinguish decoy glycans from targets on the basis of their different mass and isotope

error distributions in addition to Y and oxonium ions, while ensuring that decoys are not being shifted to masses that would exclude them from consideration alongside their target glycans. A decoy glycan candidate has the same nominal composition as the target glycan from which it was generated, but its fragment (Y and oxonium) ions are also each randomly shifted by a unique value between 1 and 20 Da. As a result, the decoy has the same number of fragment ions of each type as its corresponding target, but with randomly shifted masses. Target and decoy glycans within a user-specified tolerance of the observed delta mass (50 ppm used for all analyses here) and with allowed precursor isotope errors (–1, 0, +1, +2, or +3 for all analyses here) are considered as possible candidates (Fig. 1*B*). The observed isotope error for a candidate is determined by subtracting the candidate mass from the observed mass and rounding to the nearest integer. The resulting isotope error mass, which is the determined isotope integer error times an average peptide isotope spacing of 1.00235 Da, is removed from the candidate mass prior to computing the mass error score.

For each spectrum with a mass shift potentially corresponding to a glycan, pairwise comparisons are then performed to determine the best candidate glycan from the list of possible candidates within tolerance of the observed delta mass from the available evidence (Fig. 1*C*). Starting from two arbitrary candidates, the current best candidate is compared with other candidates until all candidates have been considered, with current best candidate being updated any time a compared candidate generates a higher score. Candidates are compared using four components that are ultimately combined into a single score: Y-ions, oxonium ions, mass error, and isotope error. For Y-ions and oxonium ions, scoring is based on ions that are unique to one candidate or the other; ions that can be generated by both candidates or neither candidate have no impact on the score. The pairwise scoring function is based on summed log likelihood estimation of the impact of each piece of evidence (each Y-ion or oxonium ion and the observed mass and isotope errors) on the likelihood of the spectrum representing one of the candidates rather than the other (Equations 1–3). For Y-ions and oxonium ions unique to one of the candidates, an empirically determined probability ratio is used to express the effect of observing the ion on the likelihood of the spectrum representing that candidate rather than the other, determined from manual inspection of spectra (42) and results. Currently, Y-ions are divided into two categories based on whether they contain fucose, and all ions within each category are given the same score. The Y-ion score can thus be expressed as a constant alpha, representing the log of the hit or miss probability ratio, times the number of unique Y-ions found or not found for each candidate for each of nonfucose and fucose-containing Y-ion sets (Equation 2). Because the number of Y-ions observed tends to be less than the number of possible Y-ions, particularly for larger glycans, Y-ion counts are square root normalized to avoid overpenalizing large glycans with more possible Y-ions than are typically observed. Probability ratios for oxonium ions are encoded separately for several composition categories (currently NeuAc/NeuGc (43–45), fucose (46, 47), phosphate (48), and sulfate (49, 50) are supported based on our empirical observations and oxonium ions reported elsewhere (51, 52), resulting in the ratios shown in supplemental Table S1. The oxonium ion score is thus computed similarly to the Y-ion score, except with potentially different probability ratios for each ion type and without square-root normalization of ion counts (Equation 3). Following the observation that oxonium ions resulting from cofragmentation of glycopeptides had a negative impact on assignment quality, an intensity weighting factor was added to the oxonium fragment score. The hit probability ratio is multiplied by the ratio of observed divided by expected intensity, such that low-intensity oxonium ions will result in a smaller increase in score than more intense ones. The adjustment is capped so that finding a hit with extremely low intensity cannot negatively impact the score: in these cases, the low-intensity hit results in no change to the score. The mass error score is the log of the ratio of observed mass errors for candidates 1 and 2, such that a lower mass error for candidate 1 than candidate 2 results in a positive score, with a weight factor β (set to 1 by default). The isotope error score is the log of the probability ratios (alpha) for the observed isotope errors of each candidate (see supplemental Table S2 for the values used). The equation for the pairwise score of glycan candidate 1 *versus* candidate 2 is thus:

$$S_{pairwise} = S_Y + S_{oxo} + \beta_{mass} \log \left| \frac{\Delta m_2}{\Delta m_1} \right| + \alpha_{isotope} \quad (1)$$

where scores for Y-ions and oxonium ions (S_Y and S_{oxo}) are defined below.

$$S_Y = \sum_{type\ t}^{2} \left(\alpha_{Y_{t_{hit}}} \left(\sqrt{U_1} - \sqrt{U_2} \right) + \alpha_{Y_{t_{miss}}} \left(\sqrt{V_1} - \sqrt{V_2} \right) \right), \quad t\ \varepsilon \begin{cases} HexNAc,\ Hex\ only \\ Fucose\ containing \end{cases} \quad (2)$$

$$S_{oxo} = \sum_{type\ t}^{5} \left(\alpha_{oxo_{t_{hit}}} \left(\frac{I_{observed}}{I_{expected}} \right) (U_1 - U_2) + \alpha_{oxo_{t_{miss}}} (V_1 - V_2) \right), \quad t\ \varepsilon \begin{cases} NeuAc \\ NeuGc \\ Fucose \\ Phosphate \\ Sulfate \end{cases} \quad (3)$$

U_1 is the number of unique fragment ions (Y-ions in Equation 2 or oxonium ions in Equation 3) to glycan candidate 1 found in the spectrum (unique hits), and V_1 is the number of unique fragment ions from glycan 1 not found in the spectrum (unique misses), and U_2 and V_2 are analogous for glycan candidate 2. *I* is the intensity of a fragment ion observed in the spectrum or "expected" (provided as a parameter). The ratio of observed to expected intensity has a minimum set such that α cannot be negative for very low observed intensities; in these cases, α is set to 0. For Y-ions and oxonium ions, probability ratios for hits are always greater than 1 (leading to $\alpha > 0$) and always less than 1 for misses ($\alpha < 0$), so that hits for candidate 1 increase the score and misses decrease it, whereas hits for candidate 2 decrease the score and misses increase it. Probability ratios and expected intensities used for each ion type and isotope error can be found in supplemental Tables S1 and S2.

After the best candidate glycan is determined by pairwise comparison for each PSM, FDR estimation is performed. Because the candidate's score in pairwise comparison depends both on the candidate and the identity of the next-best candidate, it is not optimal for distinguishing between targets and decoys. Instead, an "absolute" score is computed for the top-ranked glycan from pairwise comparison for each PSM (Equation 4). This absolute score is similar to the pairwise comparison score but treats all fragment ions from the best candidate glycan as unique and compares against typical mass and isotope errors rather than those of another candidate. Intuitively, the absolute score can be viewed as the total weight of evidence for and against the best candidate.

$$S_{absolute} = S_{abs.\ Y} + S_{abs.\ oxo} + \beta_{mass} \log \left| \frac{\Delta m}{\sigma_{unmodified}} \right| + \alpha_{isotope} \quad (4)$$

Because there is no second candidate to compare against, probability ratios for the mass and isotope errors of the candidate are

compared with typical values. The average mass error of all unmodified peptides in the analysis ($\sigma_{unmodified}$) is used as a typical value for mass error, and no isotope error is used as the typical value for isotope error. Y-ion and oxonium ion scores are computed in the same fashion as in the pairwise score but without a second candidate, treating all possible fragment ions from the chosen candidate as unique (and thus contributing to the score):

$$S_{abs.\ Y} = \sum_{type\ t}^{2} \left(\alpha_{Y_{t_{hit}}} \sqrt{U} + \alpha_{Y_{t_{miss}}} \sqrt{V} \right),\ t \in \begin{cases} HexNAc,\ Hex\ only \\ Fucose\ containing \end{cases} \quad (5)$$

$$S_{oxo} = \sum_{type\ t}^{5} \left(\alpha_{oxo_{t_{hit}}} \left(\frac{I_{observed}}{I_{expected}} \right)(U) + \alpha_{oxo_{t_{miss}}}(V) \right),\ t \in \begin{cases} NeuAc \\ NeuGc \\ Fucose \\ Phosphate \\ Sulfate \end{cases} \quad (6)$$

Where U is the number of theoretical fragment ions from the candidate found in the spectrum, analogous to the unique hits from the pairwise score, and V is the number of theoretical fragment ions not found, analogous to the unique misses from the pairwise score. I is the intensity of the fragment (expected and observed) as in the pairwise scoring and has the same minimum value to prevent α from turning negative for very low observed intensities.

FDR is computed by collecting absolute scores of all target and decoy best candidates and determining the score threshold necessary to achieve the desired decoy-target ratio (Fig. 1*D*). FDR is computed as the ratio of decoys to targets at a given score value. Results are reported in the PSM results table with the q value and identity of the matched candidate. By default, PSMs that do not pass glycan FDR are still reported but can be removed by filtering by the appropriate q-value cutoff (*e.g.*, less than 0.01 for 1% glycan FDR). PSMs in which a decoy glycan was assigned instead have the best target glycan reported but with a q value of 1. An option is available to instead print decoy glycan assignments directly for diagnostics.

All datasets tested used the default probability ratios for fragment ions and mass and isotope errors displayed in supplemental Tables S1 and S2. Glycans with noncovalent ammonium (NH_4^+) adduct(s) were appended to the internal PTM-Shepherd database in the "yeast" and "mouse" analyses as well after testing revealed high prevalence of these adducts. The "Riley" dataset was searched with no adducts.

Experimental Design and Statistical Rationale

The yeast dataset from PXD005565 contains a single sample analyzed in technical triplicate. The Riley dataset from PXD011533 also contains a single sample analyzed in technical triplicate (and fractionated into 12 fractions per replicate). The mouse dataset contains a single sample from each tissue type analyzed in five technical replicates. All data analyzed are being used for method development, and no biological conclusions are drawn. The output of MSFragger search is deterministic, as such, each MSFragger search was performed only once. PTM-Shepherd glycan assignment scoring is deterministic, but the method of generating decoys by randomly shifting the masses of decoy glycans and their fragments would result in different decoy scoring and thus different FDR thresholds, in replicate runs. As generating different results for repeated analyses of the same input data and parameters is not desirable, we have opted to fix the random seed used to generate decoys, as is done in other software tools employing this decoy generation strategy. Thus, the same decoys will always be generated for the same input raw data and parameters but will be different (randomly) if the input raw data, glycan database, or search parameters are changed.

RESULTS AND DISCUSSION

The method presented here identifies the glycan composition represented by the mass shift reported by MSFragger glyco search and uses the target-decoy approach to enable FDR filtering of the identified glycans to a defined confidence level. Unlike many PTMs, the heterogeneity of glycosylation and abundance of monosaccharide combinations with very similar or identical masses can make it challenging to determine the glycan composition represented by a given mass shift, particularly in the analysis of complex glycosylation profiles. Glycan compositions can also correspond to multiple structures resulting from various connection points between monosaccharides and different branching; however, directly determining glycan structure from glycopeptide fragmentation data is extremely challenging and typically requires specialized MS methods. All results reported here aim to confidently identify a glycan composition, that is, the identity and count of all monosaccharide classes (*e.g.*, hexose rather than specific monosaccharides like glucose) present, from typical glycoproteomics MS data without implying specific connectivity or branching information. To assess the performance of the glycan assignment method in PTM-Shepherd, we first turned to a well-characterized fission yeast dataset (PXD005565) (9). The fission yeast analyzed therein has a relatively simple glycosylation profile, with the vast majority of glycans consisting of HexNAc(2)Hex(n) structures, with n ranging from 4 to approximately 20. Following the example set by several other studies that use these data for benchmarking (9, 11, 32, 53), we performed an entrapment analysis by searching the yeast data against both yeast and mouse proteomes and glycomes to evaluate the accuracy of both peptide and glycan assignment in the presence of peptides and glycans not expected to be present in the sample. MSFragger searches were performed with one of two glycan mass lists: a "yeast-only" list and a yeast–mouse combined list equivalent to the large mouse glycan list used to perform an entrapment search in these data by pGlyco3 (11) (Fig. 2*A* and supplemental Data 1). In all cases, the number of mouse peptides matched was well controlled by the MSFragger pipeline (supplemental Table S4). The output of the MSFragger search, a peptide and mass shift for each spectrum, was then analyzed using PTM-Shepherd to assign the glycan corresponding to the mass shift for each PSM using the combined glycan list containing both yeast and entrapment (mouse) glycans for both MSFragger searches.

The search results from the different glycan lists illustrate several important factors in glycan assignment and the performance of our method in a variety of situations. When searching the yeast–mouse combined protein database with yeast-specific glycans, MSFragger matches 7855 potential glycoPSMs, and because the list of glycans searched closely matches the actual glycan population, glycan FDR filtering in

PTM-Shepherd removes only a small fraction of the matched spectra, yielding 7689 total glycoPSMs at 1% peptide and 1% glycan FDR (Fig. 2*B*). Because the PTM-Shepherd glycan analysis is considering the combined yeast–mouse glycan list, there are multiple possible compositions for each glycan mass found by MSFragger. The rate of matching nonyeast glycans is very well controlled in this analysis, with one PSM containing NeuAc, seven containing fucose, and 38 with other nonyeast glycans (mostly glycans with more than two HexNAcs but no sialic acid or fucose) passing the glycan FDR in PTM-Shepherd, accounting for 0.6% of all glycoPSMs (Table 1). Given the difference in the number of true (yeast) and entrapment (mouse) glycans in the search list, we also compute an adjusted entrapment rate using the ratio of true yeast (17) to total (1670) glycans in the search list (Table 1). We provide both the raw and adjusted entrapment rates in Table 1 as high and low estimates of the actual glycan FDR, but note that these are approximations intended to provide a comparison metric between searches and software methods rather than an absolute measure of the glycan FDR.

Expanding the MSFragger search to the combined yeast–mouse glycan list, with 1670 glycan masses, results in a moderate increase in potential glycoPSMs from MSFragger, most of which do not pass glycan FDR filtering in PTM-Shepherd, ultimately yielding 8234 glycoPSMs at 1% peptide and glycan FDR (Fig. 2*B*). Crucially, despite the nearly 100-fold increase in glycan search space, the number of PSMs matched to entrapment glycans remains controlled, with 311 total entrapment PSMs matched for an adjusted entrapment rate of 0.038% (Table 1). Notably, only two entrapment PSMs are reported for any glycan containing a sialic acid or fucose, indicating excellent control of glycan matching when comparing across categories of glycans that generate distinct fragment ion types (Table 1). Distinguishing between glycans in the same category (in this case, containing only HexNAc and Hex in varying amounts) presents a much greater challenge; however, our method is still able to maintain a reasonable entrapment rate for these glycans given the disparity between the sizes of the true and entrapment glycan lists.

Overall, both the yeast-specific and combined glycan list searches had a moderate number of spectra removed by glycan FDR filtering, lower in the yeast-specific search. The majority of these removals are due to low-quality fragmentation failing to provide sufficient fragment-ion evidence for the correct composition. There were a few cases where an incorrect peptide assignment or missing peptide modification resulted in a low-quality glycan match that was removed by filtering. Some peptides with two potential N-glycosylation sites were assigned glycan compositions that appeared to be the combination of two glycans, which generated very low glycan assignment scores and were also removed. Unlike the peptide-only FDR control of the original MSFragger glyco search, which provided reasonable assignments only when the searched glycans are well matched to the actual glycans present in the data, the method presented here was able to control the rate of entrapment matches regardless of the glycan list being searched.

Another state-of-the-art software package, pGlyco3, recently analyzed the same data (11), allowing for a direct comparison. At the same nominal FDR (1% peptide and 1% glycan), pGlyco3 reported 3405 glycoPSMs when not considering ammonium adducts or 5684 glycoPSMs when allowing an ammonium adduct, compared with the 8234 glycoPSMs from MSFragger/PTM-Shepherd when searching the same peptide database and glycan list and allowing an ammonium adduct. pGlyco3 reported no PSMs containing NeuAc or NeuGc, compared with one PSM with NeuAc or

TABLE 1
Results from entrapment searches of yeast data at 1% FDR

Search engine	Glycans searched	GlycoPSMs		NeuAc PSMs	NeuGc PSMs	Fucose PSMs	Other nonyeast PSMs	Raw entrapment rate (%)	Adjusted entrapment rate (%)
		1% peptide FDR	1% peptide & glycan FDR						
MSFragger + PTM-Shepherd	Yeast[a]	7855	7689	1	0	7	38	0.6	0.006
	Yeast + mouse[a]	9493	8234	1	0	1	309	3.8	0.038
pGlyco3	Yeast + mouse	N/A	3405	0	0	114	135	7.3	0.074
	Yeast + mouse[a]	N/A	5684	0	0	259	169	7.5	0.077

GlycoPSMs annotated at 1% peptide and combined 1% peptide and glycan FDR are shown for searches of yeast-only or combined yeast–mouse glycan lists.

Abbreviation: N/A, not available.

[a]Indicates search was performed allowing ammonium adduction (all searches except first row of pGlyco3). PSMs matched to each type of entrapment glycan and raw and adjusted entrapment rates are shown at *right*. Other nonyeast PSMs refer to entrapment glycans containing only HexNAc and Hex residues that are not in the 17-yeast glycan list. *Top two rows* of results are from MSFragger and PTM-Shepherd; the *bottom two rows* contain results from pGlyco3 (11).

NeuGc reported by MSFragger/PTM-Shepherd, indicating good FDR control for sialylated glycans by both tools. However, pGlyco3 also reported 259 PSMs containing fucose (compared with one such PSM from PTM-Shepherd) and 169 other nonyeast glycans containing only HexNAc and Hex (3% of all glycoPSMs), for an overall entrapment rate that is twice as high as that of our method (Table 1). Our method is thus able to annotate roughly 45% more glycoPSMs while also providing improved FDR control relative to pGlyco3, particularly with regard to fucosylated glycans. Ultimately, this enables confident annotation of 30 to 40% more glycopeptides and glycoproteins from the same data (Fig. 2*C*). We attribute the improved performance of our method compared with pGlyco3 primarily to two factors, namely, the peptide-first search approach resulting in many more possible glycoPSMs being considered in glycan assignment and incorporation of multiple types of information to glycan scoring in PTM-Shepherd. In particular, we found that inclusion of a weak oxonium ion filter for fucosylated glycans (in addition to the Y-ions considered by both our method and pGlyco3) provided greatly improved control of erroneous matches.

Additional tests confirm the robustness and sensitivity of the glycan FDR estimation in PTM-Shepherd. Results from searches with various glycan FDRs (1%, 0.5%, and 0.1%) are shown in Table 2. At 0.1% glycan FDR, no entrapment glycoPSMs containing monosaccharides other than HexNAc and Hex are reported with only 33 total entrapment PSMs, but over 6700 glycoPSMs are still obtained, more than pGlyco3 at 1% glycan FDR. Finally, we tested our method without prior peptide FDR filtering, providing over 33,000 potential glycopeptide spectra from the MSFragger search to PTM-Shepherd instead of only the 9493 glycopeptide spectra that passed peptide FDR. These additional ~23,000 spectra with low-confidence peptide assignments represent an increased challenge for glycan FDR filtering, as any incorrectly assigned peptide sequences may have an incorrect delta mass used to determine possible glycan candidates. The number of spectra matched to nonyeast glycans tracks neatly with the provided FDR, and adjusted entrapment rates remained within the given FDR in all cases (Table 3). As in the peptide FDR-controlled data, very few matches were made to entrapment glycans containing monosaccharides that generate distinct fragment ions (NeuAc, NeuGc, and fucose), but distinguishing between real and entrapment glycans containing only HexNAc and Hex was more challenging.

Compared with existing glycan FDR estimation approaches that rely only or primarily on the Y-ion series, our method uses several additional components to evaluate potential glycan candidates. To evaluate the individual contribution of each of these components, we performed the assignment and FDR procedure sequentially with a single score component removed and assessed the number of PSMs passing glycan FDR and any changes in the entrapment rates of different glycan types (Fig. 3). Of the four components, isotope error provided the least valuable contribution, indicated by its removal not substantially changing the number of glycoPSMs passing FDR and only slightly increasing the number of entrapment glycans matched (Fig. 3). This is perhaps unsurprising given that MSFragger was set to attempt to correct any errors in monoisotopic peak selection prior to PTM-Shepherd analysis. In cases where such correction is not performed, the isotope error component may provide greater benefit. Removal of the oxonium ion score also resulted in only a small change in the number of PSMs reported but caused uncontrolled matching to compositions containing sialic acids and, to a lesser extent, fucose. Since the sialic acids are typically dissociated from the glycan in the HCD fragmentation employed in this dataset, they generally do not affect the Y-ions produced and are thus reliant on oxonium ion scores for appropriate scoring and filtering. Removal of the mass error score resulted in a moderate decrease in the number of glycoPSMs matched, indicating that it provides a valuable contribution in distinguishing true matches but did not cause an increase in entrapment glycans matched. Finally, removal of Y-ions resulted in the largest decrease in glycoPSMs, indicating the central role they play in obtaining confident matches. These results illustrate the importance of including multiple sources of information when evaluating glycans, as Y-ions, oxonium ions, and mass error all substantially improved the ability to distinguish between glycan compositions. The relative contributions of the categories may vary in different analyses, however, as a result of different fragmentation methods or settings changing the likelihood of observing different types of fragment ions, or different instruments or instrument settings changing the distribution of mass and isotope errors compared with the data analyzed here, for example.

TABLE 2
Comparison of search results of yeast data with various glycan FDR levels

Glycan FDR (%)	Peptide & glycan 1% FDR glycoPSMs	NeuAc PSMs	NeuGc PSMs	Fucose PSMs	Other nonyeast PSMs	Raw entrapment rate (%)	Adjusted entrapment rate (%)
1	8234	1	0	1	309	3.8	0.04
0.5	8022	1	0	1	232	2.9	0.03
0.1	6741	0	0	0	35	0.5	0.005

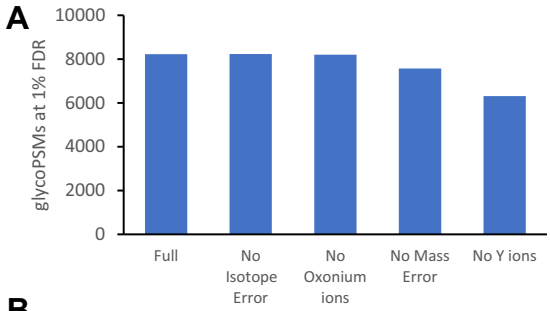
The combined yeast and mouse (1670) glycan list was used for the MSFragger search, resulting in 9493 potential glycoPSMs passing 1% peptide FDR. Entrapment glycans matched are shown at *right* for each search, along with raw and adjusted entrapment rates.

TABLE 3
Summary of yeast search results without prior peptide FDR filtering at various glycan FDR levels

Glycan FDR (%)	Glycan FDR only glycoPSMs	NeuAc PSMs	NeuGc PSMs	Fucose PSMs	Other non-east PSMs	Raw entrapment rate (%)	Adjusted entrapment rate (%)
1	14,052	1	0	50	1178	8.7	0.09
0.5	11,291	1	0	13	424	3.9	0.04
0.1	5171	0	0	0	41	0.8	0.008

The combined yeast and mouse (1670) glycan list was used for the MSFragger search. In total, 33,450 potential glycoPSMs were supplied to PTM-Shepherd with no peptide FDR filter. PSMs matched to entrapment glycans are shown for each FDR level along with raw and adjusted entrapment rates.

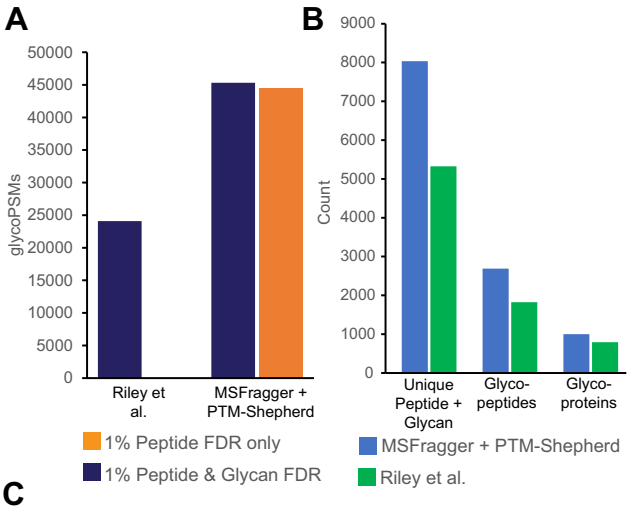
Having validated our glycan assignment and FDR methods with entrapment searches, we set out to characterize the performance of our method in more typical, nonentrapment, glycoproteomics data. We reanalyzed data from Riley *et al.* (38) to compare with our previous analysis by MSFragger-Glyco (17). The original analysis, performed with Byonic (12), identified 24,099 glycoPSMs after applying Byonic's peptide FDR filtering and additional empirically determined score filters. Applying the same mouse glycoprotein database and glycan list, MSFragger identified 45,318 glycoPSMs at 1% peptide FDR, of which 44,781 glycoPSMs passed 1% glycan FDR in PTM-Shepherd (Fig. 4*A*). Unlike the entrapment searches of fission yeast, the glycan FDR filtering removed relatively few PSMs, indicating the high quality of the initial matches from MSFragger when searching only for glycopeptides that are likely to be present in the data. However, the glycan assignment method is still critical to distinguish the correct glycan out of multiple possibilities, even when only expected glycans are included in the search. The additional glycoPSMs annotated result in identification of additional glycopeptides and glycoproteins, as we have previously noted, and additional high confidence pairings of peptide sequences with specific glycan compositions (Fig. 4*B*). We find similarly small proportions of glycan compositions containing fucose and sialic acids as in the original analysis of Riley *et al.* Mouse brain tissue has been observed to have a high proportion of glycans lacking fucose or sialic acids (9, 54), and the concanavalin A lectin enrichment performed specifically enriches for oligomannose-type glycans over others, both of which lend support to our assignment of the majority of glycoPSMs to compositions containing only HexNAc and Hex-type residues. While our method assigned different glycans than Byonic to several hundred glycoPSMs, the differences were largely between similar compositions that differ by isotope errors. For example, our method frequently assigned the composition HexNAc(2)Hex(8) to spectra that Byonic assigned as HexNAc(6)Hex(3), often when the precursor was misassigned as the +2 isotope peak, as the two compositions differ in mass by 2.05 Da (supplemental Table S5). Comparisons of assigned glycans are necessarily limited, however, as Byonic does not perform a glycan-specific FDR calculation, controlling peptide FDR only. While we did not perform a deliberate entrapment search on these data, two alternative methods for confirming our glycan assignments are available since mouse brain tissue is expected to contain only trace amounts of NeuGc or sulfated glycans, and the dataset contains paired HCD and AI–ETD scans of the same precursor. The MSFragger search performed included only 182 glycan compositions, but for the PTM-Shepherd analysis, we used a large mammalian N-glycan and O-glycan list that included NeuGc and sulfate-containing glycans (supplemental Data 1), many of which have similar or identical masses to the glycans included in the MSFragger search. Less than 0.1% of all glycoPSMs were matched to NeuGc or sulfated glycans(Fig. 4*C*), providing evidence that our FDR control is functioning well for these data. In addition, the paired HCD and AI–ETD scans should have same peptide and glycan assigned, since they are of the same glycopeptide precursor fragmented using different activation methods. Of the 28,970 glycoPSMs that had a paired scan that passed both peptide and glycan FDR, 99.4% were independently assigned to the same peptide and glycan by our method (Fig. 4*D*), as would be expected given a 1% peptide and



B

Analysis	1% FDR Filtered glycoPSMs	NeuAc-containing PSMs	NeuGc-containing PSMs	Fucose-containing PSMs
Full	8234	1	0	1
No Isotope Error	8233	2	0	3
No Oxonium ions	8210	567	301	49
No Mass Error	7558	0	0	0
No Y ions	6250	1	0	5

FIG. 3. **Impact of individual score components on glycan assignment performance.** *A*, glycoPSMs passing 1% glycan (and peptide) FDR with full score (all components) or one component removed. *B*, table of total glycoPSMs and entrapment glycoPSMs of various types for each analysis presented in *A*. FDR, false discovery rate; PSM, peptide-spectrum match.



C

glycoPSMs	HexNAc, Hex only PSMs	Fucose PSMs	NeuAc PSMs	NeuGc PSMs	Sulfate PSMs
44781	38025	6440	967	22	0

D

Total glycoPSMs	paired glycoPSMs	Paired PSMs with different peptide	Paired PSMs with different glycan	Paired PSMs with same peptide & glycan
44781	28970	52	124	28794
		0.2%	0.4%	99.4%

FIG. 4. **Comparison of glycoPSMs annotated from mouse brain tissue from Riley *et al.*** *A*, glycoPSMs annotated by the original analysis in Byonic and our reanalysis with MSFragger and PTM-Shepherd. The improvement in annotated spectra from MSFragger search is maintained after applying 1% glycan FDR filtering. *B*, unique glycoproteins, glycopeptide sequences, and glycan–peptide sequence combinations reported by MSFragger/PTM-Shepherd or Riley *et al.* *C*, glycan composition categories observed sorted by residue type(s) contained, showing few unexpected (NeuGc or sulfate-containing) glycans. *D*, analysis of paired HCD and AI-ETD scans of the same precursor from MSFragger/PTM-Shepherd search, assessed for whether the same peptide and glycan were matched in both of the paired scans. AI–ETD, activated ion–electron transfer dissociation; FDR, false discovery rate; HCD, higher energy collisional dissociation; PSM, peptide-spectrum match.

glycan FDR. Notably, while our empirical fragment ion probability ratios were developed using HCD data, performance in glycan assignment was comparable in the hybrid AI–ETD data included here.

Finally, while the entrapment glycan searches performed in yeast establish a strong foundation for FDR control in N-glycoproteomics analyses, the yeast glycoproteome contains few to no true glycan compositions that have very similar masses. To establish the performance of our method in more complex data, in which there are many examples of mass shifts that could be assigned to multiple possible glycan compositions expected to be present in the data, we reanalyzed N-glycopeptide data from five mouse tissue types originally presented in the description of the pGlyco2 software (9). In the MSFragger and PTM-Shepherd searches, we used the same glycan list as the combined yeast–mouse list from the yeast analysis, equivalent to the "pGlyco-N-mouse-large" glycan list from pGlyco3, allowing a single ammonium adduct. In all five tissue types, we find 40 to 66% more glycoPSMs at 1% peptide and glycan FDR than pGlyco3 when searching the same peptide and glycan lists, which in turn obtained a similar advantage over pGlyco2 (Fig. 5*A*). The pGlyco2 search (results as reported (9)) did not include ammonium adducts as this feature was not supported in pGlyco2. As in the yeast data, we attribute the similarly sized increase in glycoPSMs annotated by MSFragger/PTM-Shepherd over pGlyco3 to the peptide-first search strategy and incorporation of additional information in glycan scoring as compared with pGlyco3. As in the other datasets, the increase in glycoPSMs translated to an increase in unique glycopeptide sequences in all tissue types and of glycoproteins in all tissues except liver and lung, in which glycoprotein counts were similar (Fig. 5*B*). We observed broadly similar trends in glycan compositions across tissue types, including the lack of NeuGc in brain tissue and predominance of high-mannose glycans in liver (Fig. 5*C*). However, when we directly compare the frequency at which compositions containing various glycan residues were assigned, a clear pattern emerges of much higher rates of fucose-containing glycans assigned by pGlyco3 (*orange bars*) with roughly opposite increases in sialic acid–containing glycans assigned by PTM-Shepherd (*green* and *yellow bars*) (Fig. 5*D*). Given the high frequency of entrapment matches to fucosylated glycans by pGlyco3 in yeast, we suspect a similar bias toward fucose may be occurring in these mouse data, and many of the fucosylated glycans assigned by pGlyco3 may in fact be sialylated, particularly since the substitution of two fucoses for one NeuAc is a common glycan assignment mistake (30). An example spectrum with such a substitution is shown in supplemental Fig. S1, in which a series of abundant NeuAc-containing oxonium ions provides strong evidence for our composition assignment (HexNAc-4_Hex-5_NeuAc-1) over pGlyco3's (HexNAc-4_Hex-5_Fuc-2). However, without a ground truth of known glycopeptides to compare against, we cannot say for certain which tool's assignments are more frequently correct.

Evaluating the accuracy of our method in these data is more challenging than in yeast, as there is not as straightforward a filter for entrapment compositions since glycans containing fucose, NeuAc, and NeuGc are all expected to be present in various tissues. However, we were able to generate a list of 248 entrapment glycans that have very similar masses (within 0.05 Da) of true mouse glycans using several substitutions of glycan residues (see supplemental Data 1, "mouse entrapment" for the list of compositions). We generated NeuAc (27), NeuGc (17), Fuc (24), and phosphate (180)-containing entrapment glycans with similar masses to many of the most commonly observed mouse glycans (supplemental Table S5). Unlike in the yeast entrapment search, these had to be manually checked against the mouse glycan list to ensure that true mouse glycans were not included as potential

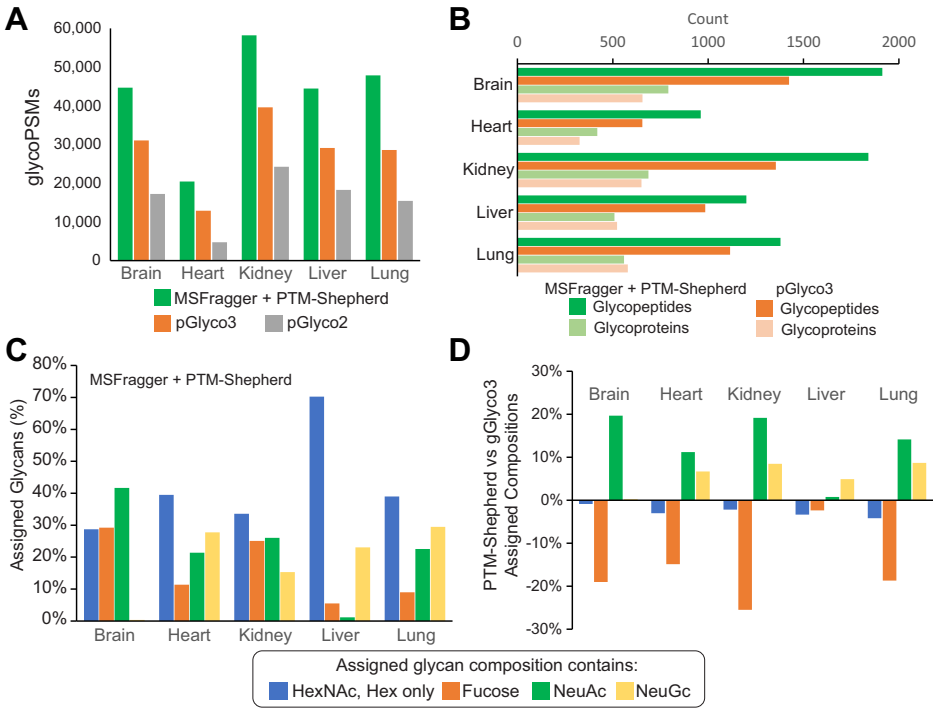


FIG. 5. **Comparison of N-glycan search in mouse tissue dataset between MSFragger + PTM-Shepherd and pGlyco3.** *A*, glycoPSMs passing 1% peptide and glycan FDR from MSFragger/PTM-Shepherd, pGlyco3, or pGlyco2 searches in each of five tissue types. *B*, comparison of the number of unique glycoproteins and glycopeptide sequences (1% peptide and glycan FDR) from searches presented in *A*. *C*, PTM-Shepherd assigned compositions by residues included in the glycan. Note that glycoPSMs containing multiple residue types (*e.g.*, fucose and NeuAc) will be counted in multiple categories. *D*, comparison of PTM-Shepherd assigned glycan compositions with pGlyco3 compositions. Positive values indicate more of a residue type assigned by PTM-Shepherd, and negative values indicate more of a residue type assigned by pGlyco3. In all tissues, PTM-Shepherd assigns more sialic acid compositions, whereas pGlyco3 assigned more fucose. FDR, false discovery rate; PSM, peptide-spectrum match.

entrapments, resulting in smaller numbers of these entrapment glycans overall and an imbalance between the phosphate-containing and other types. Nevertheless, by targeting these entrapment glycans to have similar masses to, and contain the same types of residues as, the most commonly observed real glycans, we provide an intensive test of whether our method can distinguish between real and entrapment glycans in a complex analysis with many overlapping masses.

The results of entrapment glycan searches in each mouse tissue type are shown in Table 4. In each tissue, the adjusted rate of entrapment glycoPSMs matched remains at or below 1%, indicating good FDR control in a much more strenuous test than in the yeast data. Notably, almost no entrapment matches are made to phosphoglycans despite the presence of far more potential entrapment phosphoglycans than the other categories. The majority of entrapment glycoPSMs are to NeuAc in brain and kidney tissues and to NeuGc in heart, lung, and liver, in each case corresponding to the most common category of glycan observed in the respective tissue. These observations, together with the lack of similar entrapment matches in the yeast data, suggest that the primary reason for the matches to entrapment glycans may be cofragmentation of glycopeptides that results in oxonium ions from a glycan other than that of the precursor glycopeptide being including in scoring a given glycoPSM, increasing the rate of erroneous glycan assignments. This hypothesis is strongly supported by the brain tissue data, in which over 90% of all scans contain NeuAc oxonium ion(s) (despite only 40% of glycoPSMs being matched to NeuAc-containing compositions), whereas only ~2% of scans have detectable NeuGc oxonium ion(s). Matches to NeuAc entrapment compositions were the vast majority of entrapment matches observed (89%), and no matches were made to NeuGc entrapment compositions (Table 4), despite being relatively common in the other tissue types where NeuGc oxonium ions occur with some frequency. Distinguishing the correct glycan from spectra in which oxonium ions seem to indicate an alternative composition is indeed a challenging problem, even for expert manual curation of glycan assignments. Maintaining a 1% entrapment glycan rate despite this extensive cofragmentation is thus an impressive achievement for our method. Care should still be exercised in complex analyses, however, particularly when it is clear that oxonium ions for certain compositions are present in a majority of spectra. Narrower isolation windows, ion mobility filtering, and other measures to reduce cofragmentation of

TABLE 4
Results of entrapment glycan search in the mouse dataset

Sample	Total glycoPSMs	Entrapment glycoPSMs	Entrapment glycoPSMs by type				Raw entrapment rate (%)	Adjusted entrapment rate (%)
			NeuAc	NeuGc	Phospho	Fucose only		
Brain	44,931	491	436	0	0	55	1.1	1.0
Heart	20,410	128	43	72	0	13	0.6	0.5
Kidney	65,412	586	358	92	0	136	0.9	0.8
Liver	44,476	24	3	13	2	6	0.05	0.05
Lung	47,292	183	55	114	0	14	0.4	0.3

PSMs matched to entrapment glycans are shown as a count and raw and adjusted rates (%), remaining at or below the FDR used in the searches (1%). Details of the entrapment glycans used can be found in the supplemental Data 1.

glycopeptides would be likely to greatly reduce the incidence of this issue. It is reassuring to note that in all cases without such widespread oxonium presence, entrapment matches were rare to nonexistent.

CONCLUSIONS

We have developed a sensitive and robust method for determining the composition of the glycan component of N-glycopeptides from tandem MS data by combining information from multiple types of fragment ions with mass and isotope errors to distinguish between candidate compositions. We demonstrate that FDR control of the resulting glycan matches performs as expected even in analyses of complex glycan lists and in the presence of entrapment peptide sequences and glycan compositions. As glycoproteomics moves to larger and more complex samples, confident assignment of glycan compositions is critical to move beyond manual validation of glycopeptide spectra. We believe this work represents a promising step in this direction, enabling automated analysis of complex N-glycopeptide samples. Determining FDR for glycan assignment is a challenging problem with many complexities, and there remain several limitations in the method presented here, particularly with regard to distinguishing between highly similar glycans and in extending the method to O-glycopeptides, that we aim to address in future work. The probability ratios used to determine the likelihood of one glycan composition relative to another given the presence (or absence) of a particular fragment ion were determined empirically and are optimized for N-glycan fragmentation by HCD or stepped-energy HCD. While we have confirmed that these parameters still provide valid results for AI–ETD activation of N-glycans, probability parameters will likely need to be optimized specifically for other fragmentation methods to provide similarly high-quality results. Currently, all fragment ions of a given type are given the same probability ratios, despite the actual probabilities to observe various fragments varying greatly, and probabilities of observing the same fragment also depending on the precursor glycan. Using glycan and fragment-specific probabilities would likely improve the performance of glycan assignment, particularly when considering a wider range of fragmentation methods and settings. Recent reports have indicated that structure-specific fragmentation patterns could be used to infer glycan structure in addition to composition (33), a potential further application of fragment-specific probabilities. Finally, while the separation of glycan assignment from peptide sequencing greatly simplifies the problem and enables the high performance of our method, cases in which multiple glycans are present on a single peptide, or a combination of a glycan and other modification(s) found from open searching, are not currently supported and will require development of methods to localize multiple modification sites from these searches prior to glycan assignment. Methods for localizing multiple glycans on a peptide will also be required for the analysis of most O-glycopeptide data, as many O-glycopeptides contain multiple potential glycosites. While the method provided here is readily applicable to O-glycosylation in theory, handling multiply glycosylated peptides and tuning the fragment probabilities for O-glycans is needed to enable accurate O-glycan assignments.

DATA AVAILABILITY

All raw data used can be found in the public repositories noted in the Experimental Procedures section. Processed results (PSM tables) used in the creation of all figures can be found at 10.5281/zenodo.5834131. PTM-Shepherd source code and a standalone JAR executable can be found at https://github.com/Nesvilab/PTM-Shepherd, and it is integrated into the FragPipe graphical user interface (http://fragpipe.nesvilab.org/).

Supplemental data—This article contains supplemental data (11).

Acknowledgments—This work was funded in part by the National Institutes of Health grants R01-GM-094231 and U24-CA210967. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions—D. A. P. and A. I. N. conceptualization; D. J. G. and F. Y. methodology; D. A. P., D. J. G., and F. Y. software; D. A. P. investigation; D. A. P. writing–original draft; D. A. P., D. J. G., F. Y., and A. I. N. writing–review & editing; A. I. N. supervision; A. I. N. funding acquisition.

Conflict of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: AI–ETD, activated ion–electron transfer dissociation; FDR, false discovery rate; HCD, higher energy collisional dissociation; NeuAc, *N*-acetyl neuraminic acid; NeuGc, *N*-glycolyl neuraminic acid; PSM, peptide-spectrum match; PTM, post-translational modification.

Received August 6, 2021, and in revised from, January 10, 2022
Published, MCPRO Papers in Press, January 26, 2022, https://doi.org/10.1016/j.mcpro.2022.100205

REFERENCES

1. Varki, A. (2017) Biological roles of glycans. *Glycobiology* **27**, 3–49
2. Marsico, G., Russo, L., Quondamatteo, F., and Pandit, A. (2018) Glycosylation and integrin regulation in Cancer. *Trends Cancer* **4**, 537–552
3. Schedin-Weiss, S., Winblad, B., and Tjernberg, L. O. (2014) The role of protein glycosylation in Alzheimer disease. *FEBS J.* **281**, 46–62
4. York, I. A., Stevens, J., and Alymova, I. V. (2019) Influenza virus N-linked glycosylation and innate immunity. *Biosci. Rep* **39**. https://doi.org/10.1042/BSR20171505
5. Thaysen-Andersen, M., Packer, N. H., and Schulz, B. L. (2016) Maturing glycoproteomics technologies provide unique structural insights into the N-glycoproteome and its regulation in health and disease. *Mol. Cell. Proteomics* **15**, 1773–1790
6. Suttapitugsakul, S., Sun, F., and Wu, R. (2019) Recent advances in glycoproteomic analysis by mass spectrometry. *Anal. Chem.* **92**, 267–291
7. Reiding, K. R., Bondt, A., Franc, V., and Heck, A. J. R. (2018) The benefits of hybrid fragmentation methods for glycoproteomics. *Trends Anal. Chem* **108**, 260–268
8. Cao, W., Liu, M., Kong, S., Wu, M., Zhang, Y., and Yang, P. (2021) Recent advances in software tools for more generic and precise intact glycopeptide analysis. *Mol. Cell. Proteomics* **20**, 100060
9. Liu, M.-Q., Zeng, W.-F., Fang, P., Cao, W.-Q., Liu, C., Yan, G.-Q., Zhang, Y., Peng, C., Wu, J.-Q., Zhang, X.-J., Tu, H.-J., Chi, H., Sun, R.-X., Cao, Y., Dong, M.-Q., *et al*. (2017) pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nat. Commun.* **8**, 438
10. Zeng, W. F., Liu, M. Q., Zhang, Y., Wu, J. Q., Fang, P., Peng, C., Nie, A., Yan, G., Cao, W., Liu, C., Chi, H., Sun, R. X., Wong, C. C. L., He, S. M., and Yang, P. (2016) pGlyco: A pipeline for the identification of intact N-glycopeptides by using HCD- and CID-MS/MS and MS3. *Sci. Rep.* **6**, 25102
11. Zeng, W.-F., Cao, W.-Q., Liu, M.-Q., He, S.-M., and Yang, P.-Y. (2021) Precise, fast and comprehensive analysis of intact glycopeptides and modified glycans with pGlyco3. *Nat. Methods* **18**, 1515–1523
12. Bern, M., Kil, Y. J., and Becker, C. (2012) Byonic: Advanced peptide and protein identification software. *Curr. Protoc. Bioinformatics*
13. Xiao, K., and Tian, Z. (2019) GPSeeker enables quantitative structural N-glycoproteomics for site- and structure-specific characterization of differentially expressed N-glycosylation in hepatocellular carcinoma. *J. Proteome Res.* **18**, 2885–2895
14. Lu, L., Riley, N. M., Shortreed, M. R., Bertozzi, C. R., and Smith, L. M. (2020) O-pair search with MetaMorpheus for O-glycopeptide characterization. *Nat. Methods* **17**, 1133–1138
15. He, L., Xin, L., Shan, B., Lajoie, G. A., and Ma, B. (2014) GlycoMaster DB: Software to assist the automated identification of N-linked glycopeptides by tandem mass spectrometry. *J. Proteome Res.* **13**, 3881–3895
16. Lynn, K.-S., Chen, C.-C., Lih, T. M., Cheng, C.-W., Su, W.-C., Chang, C.-H., Cheng, C.-Y., Hsu, W.-L., Chen, Y.-J., and Sung, T.-Y. (2015) Magic: An automated N-linked glycoprotein identification tool using a Y1-ion pattern matching algorithm and in silico MS 2 approach. *Anal. Chem.* **87**, 2466–2473
17. Polasky, D. A., Yu, F., Teo, G. C., and Nesvizhskii, A. I. (2020) Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods* **17**, 1125–1132
18. Hu, Y., Shah, P., Clark, D. J., Ao, M., and Zhang, H. (2018) Reanalysis of global proteomic and phosphoproteomic data identified a large number of glycopeptides. *Anal. Chem.* **90**, 8065–8071
19. Hu, H., Khatri, K., Klein, J., Leymarie, N., and Zaia, J. (2016) A review of methods for interpretation of glycopeptide tandem mass spectral data. *Glycoconj. J.* **33**, 285–296
20. [preprint] Kawahara, R., Alagesan, K., Bern, M., Cao, W., Chalkley, R. J., Cheng, K., Choo, M. S., Edwards, N., Goldman, R., Hoffmann, M., Hu, Y., Huang, Y., Kim, J. Y., Kletter, D., Liquet-Weiland, B., *et al*. (2021) Community evaluation of glycoproteomics informatics solutions reveals high-performance search strategies of glycopeptide data. *bioRxiv*. https://doi.org/10.1101/2021.03.14.435332
21. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392
22. Bollineni, R. C., Koehler, C. J., Gislefoss, R. E., Anonsen, J. H., and Thiede, B. (2018) Large-scale intact glycopeptide identification by Mascot database search. *Sci. Rep.* **8**, 2117
23. Fang, P., Xie, J. J., Sang, S., Zhang, L., Liu, M., Yang, L., Xu, Y., Yan, G., Yao, J., Gao, X., Qian, W., Wang, Z., Zhang, Y., Yang, P., and Shen, H. (2020) Multilayered N-glycoproteome profiling reveals highly heterogeneous and dysregulated protein N-glycosylation related to Alzheimer's disease. *Anal. Chem.* **92**, 867–874
24. Blazev, R., Ashwood, C., Abrahams, J. L., Chung, L. H., Francis, D., Yang, P., Watt, K. I., Qian, H., Quaife-Ryan, G. A., Hudson, J. E., Gregorevic, P., Thaysen-Andersen, M., and Parker, B. L. (2021) Integrated glycoproteomics identifies a role of N-glycosylation and galectin-1 on myogenesis and muscle development. *Mol. Cell. Proteomics* **20**, 100030
25. Chen, Z., Yu, Q., Yu, Q., Johnson, J., Shipman, R., Zhong, X., Huang, J., Asthana, S., Carlsson, C., Okonkwo, O., and Li, L. (2021) In-depth site-specific analysis of N-glycoproteome in human cerebrospinal fluid (CSF) and glycosylation landscape changes in Alzheimer's disease (AD). *Mol. Cell. Proteomics* **20**, 100081
26. Hu, H., Khatri, K., and Zaia, J. (2017) Algorithms and design strategies towards automated glycoproteomics analysis. *Mass Spectrom Rev.* **36**, 475–498
27. Hackett, W. E., and Zaia, J. (2021) The need for community standards to enable accurate comparison of glycoproteomics algorithm performance. *Molecules* **26**, 4757
28. Hackett, W., and Zaia, J. (2021) Calculating glycoprotein similarities from mass spectrometric data. *Mol. Cell. Proteomics* **20**, 100028
29. Darula, Z., and Medzihradszky, K. F. (2015) Carbamidomethylation side reactions may lead to glycan misassignments in glycopeptide analysis. *Anal. Chem.* **87**, 6297–6302
30. Lee, L. Y., Moh, E. S. X., Parker, B. L., Bern, M., Packer, N. H., and Thaysen-Andersen, M. (2016) Toward automated N-glycopeptide identification in glycoproteomics. *J. Proteome Res.* **15**, 3904–3915
31. Zhu, Z., Su, X., Go, E. P., and Desaire, H. (2014) New glycoproteomics software, glycopep evaluator, generates decoy glycopeptides de novo and enables accurate false discovery rate analysis for small data sets. *Anal. Chem.* **86**, 9212–9219
32. [preprint] Klein, J., Carvalho, L., and Zaia, J. (2021) Expanding N-glycopeptide identifications by fragmentation prediction and glycome network smoothing. *bioRxiv*. https://doi.org/10.1101/2021.02.14.431154
33. Shen, J., Jia, L., Dang, L., Su, Y., Zhang, J., Xu, Y., Zhu, B., Chen, Z., Wu, J., Lan, R., Hao, Z., Ma, C., Zhao, T., Gao, N., Bai, J., *et al*. (2021) StrucGP: De novo structural sequencing of site-specific N-glycan on glycoproteins using a modularization strategy. *Nat. Methods* **18**, 921–929
34. Yu, F., Teo, G. C., Kong, A. T., Haynes, S. E., Avtonomov, D. M., Geiszler, D. J., and Nesvizhskii, A. I. (2020) Identification of modified peptides using localization-aware open search. *Nat. Commun.* **11**, 4065
35. Geiszler, D. J., Kong, A. T., Avtonomov, D. M., Yu, F., da Veiga Leprevost, F., and Nesvizhskii, A. I. (2021) PTM-shepherd: Analysis and

summarization of post-translational and chemical modifications from open search results. *Mol. Cell. Proteomics* **20**, 100018
36. Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D. S., Bernal-Llinares, M., Okuda, S., Kawano, S., Moritz, R. L., Carver, J. J., Wang, M., Ishihama, Y., Bandeira, N., *et al*. (2017) The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45**, D1100–D1106
37. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536
38. Riley, N. M., Hebert, A. S., Westphall, M. S., and Coon, J. J. (2019) Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nat. Commun.* **10**, 1311
39. Adusumilli, R., and Mallick, P. (2017) Data conversion with ProteoWizard msConvert. *Methods Mol. Biol.* **1550**, 339–368
40. da Veiga Leprevost, F., Haynes, S. E., Avtonomov, D. M., Chang, H. Y., Shanmugam, A. K., Mellacheruvu, D., Kong, A. T., and Nesvizhskii, A. I. (2020) Philosopher: A versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870
41. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658
42. Li, K., Vaudel, M., Zhang, B., Ren, Y., and Wen, B. (2019) PDV: an integrative proteomics data viewer. *Bioinformatics* **35**, 1249–1251
43. Medzihradszky, K. F., Kaasik, K., and Chalkley, R. J. (2015) Characterizing sialic acid variants at the glycopeptide level. *Anal. Chem.* **87**, 3064–3071
44. Halim, A., Westerlind, U., Pett, C., Schorlemer, M., Rüetschi, U., Brinkmalm, G., Sihlbom, C., Lengqvist, J., Larson, G., and Nilsson, J. (2014) Assignment of saccharide identities through analysis of oxonium ion fragmentation profiles in LC-MS/MS of glycopeptides. *J. Proteome Res.* **13**, 6024–6032
45. Pett, C., Nasir, W., Sihlbom, C., Olsson, B. M., Caixeta, V., Schorlemer, M., Zahedi, R. P., Larson, G., Nilsson, J., and Westerlind, U. (2018) Effective assignment of α2,3/α2,6-sialic acid isomers by LC-MS/MS-based glycoproteomics. *Angew. Chem. Int. Ed. Engl.* **57**, 9320–9324
46. Ács, A., Ozohanics, O., Vékey, K., Drahos, L., and Turiák, L. (2018) Distinguishing core and antenna fucosylated glycopeptides based on low-energy tandem mass spectra. *Anal. Chem.* **90**, 12776–12782
47. Lakbub, J. C., Su, X., Hua, D., Go, E. P., and Desaire, H. (2018) Dissecting the dissociation patterns of fucosylated glycopeptides undergoing CID: A case study in improving automated glycopeptide analysis scoring algorithms. *Anal. Methods* **10**, 256–262
48. Caval, T., Zhu, J., Tian, W., Remmelzwaal, S., Yang, Z., Clausen, H., and Heck, A. J. R. (2019) Targeted analysis of lysosomal directed proteins and their sites of mannose-6-phosphate modification. *Mol. Cell. Proteomics* **18**, 16–27
49. Kuo, C. W., Guu, S. Y., and Khoo, K. H. (2018) Distinctive and complementary MS^2 fragmentation characteristics for identification of sulfated sialylated N-glycopeptides by nanoLC-MS/MS workflow. *J. Am. Soc. Mass Spectrom.* **29**, 1166–1178
50. Sanda, M., Benicky, J., and Goldman, R. (2020) Low collision energy fragmentation in structure-specific glycoproteomics analysis. *Anal. Chem.* **92**, 8262–8267
51. Yu, J., Schorlemer, M., Gomez Toledo, A., Pett, C., Sihlbom, C., Larson, G., Westerlind, U., and Nilsson, J. (2016) Distinctive MS/MS fragmentation pathways of glycopeptide-generated oxonium ions provide evidence of the glycan structure. *Chemistry* **22**, 1114–1124
52. Hoffmann, M., Pioch, M., Pralow, A., Hennig, R., Kottler, R., Reichl, U., and Rapp, E. (2018) The fine art of destruction: A guide to in-depth glycoproteomic analyses—exploiting the diagnostic potential of fragment ions. *Proteomics* **18**, e1800282
53. Yang, Y., Yan, G., Kong, S., Wu, M., Yang, P., Cao, W., and Qiao, L. (2021) GproDIA enables data-independent acquisition glycoproteomics with comprehensive statistical control. *Nat. Commun.* **12**, 6073
54. Trinidad, J. C., Schoepfer, R., Burlingame, A. L., and Medzihradszky, K. F. (2013) N- and O-Glycosylation in the murine synaptosome. *Mol. Cell. Proteomics* **12**, 3474–3488