WESTERN UNIVERSITY
LONDON, ONTARIO, CANADA

Recent Advances in N-linked Glycoproteomics

Sitong Li, sli2332@uwo.ca

Abstract

As one of the most important and abundant post-translational modifications (PTMs), glycosylation is involved in various biological events. Glycoproteomics is the study of the glycosylation of proteins, which includes the identification of the peptide sequence, glycan chain sequence, and prediction of glycosylation site. The properties of glycosylation, such as macro-/micro-heterogeneity, low abundance and ionization efficiency, make it challenging to develop a high-throughput and accurate tool for glycoprotein identification. In an effort to surmount existed difficulties, various mass spectrometry (MS) strategies and corresponding bioinformatic tools have been developed. This review focuses on recent progresses in N-linked glycoproteomics mainly from the past three years.

1 INTRODUCTION

PTMs are covalent modification of proteins involving addition of modifying groups or proteolytic cleavages. PTMs extensively affect protein function and physiological processes, such as immune response, protein trafficking and interactions with other proteins. Also, in the context of personalized medicine and diagnostics, glycoproteins are considered as interesting candidates for biomarker discovery, as changes in protein glycosylation are associated with disease states [1]. Glycosylation of proteins is a common PTM by enzymatically attaching glycans to proteins. Glycoproteomics, as an emerging subfield of proteomics, providing qualitative and quantitative information of glycoproteins. Over the last decade, tremendous progress in glycoproteomics have paved the way for better understanding of biological attributes of glycans and potential applications such as discovery of biomarkers [39].

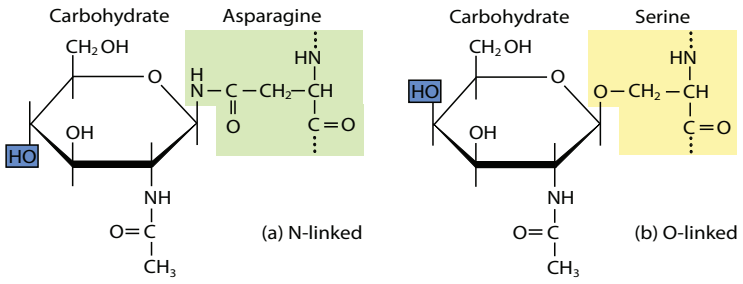

Figure 1.1: N-linked(a) linked to asparagine and O-linked(b) linked to serine glycoproteins.

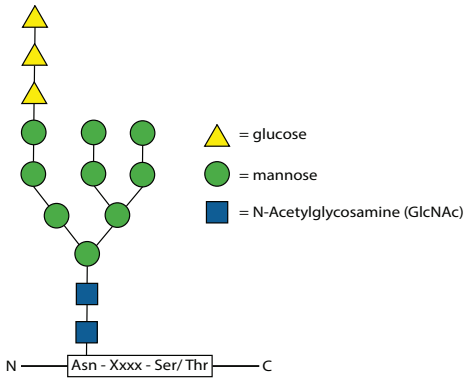

Figure 1.2: The N-linked core oligosaccharide. N-linked glycans are added to proteins in the endoplasmic reticulum as “core oligosaccharides” with the above structure.

Based on the mode of attachment, glycosylation can be divided into two types: N-linked and O-linked Figure 1.1. N-linked glycans are attaching to asparagine residue in a tripeptide consensus sequence Asn-Xxxx-Ser/Thr (where X can be any amino acid except proline), while O-linked glycans are attached to serine or threonine or hydroxylysine residues [16, 15, 37]. N-linked glycans have five-monosaccharide core structure Figure 1.2, while O-linked glycans have more varied core structures. Most strategies in this review were developed for and tested on N-glycoproteomics, though some of them have shown promising potential for O-glycoproteomics.

Depending on the general pipeline, glycoproteomics strategies can be categorized into different groups: the “bottom-up”, the “top-down”, and a combination of the two. The “bottom-up” approach is peptide-centric, where glycoproteins are proteolytically cleaved to generate glycopeptides, typically using trypsin. Then the peptides are separated from their unglycosylated counterparts using liquid chromatography and analyzed by MS/MS in a manner similar to peptide sequencing [47]. By contrast, intact glycoproteins are subjected to MS analysis in the “top-down” approaches [19]. Both two strategies have their respective advantages and drawbacks.

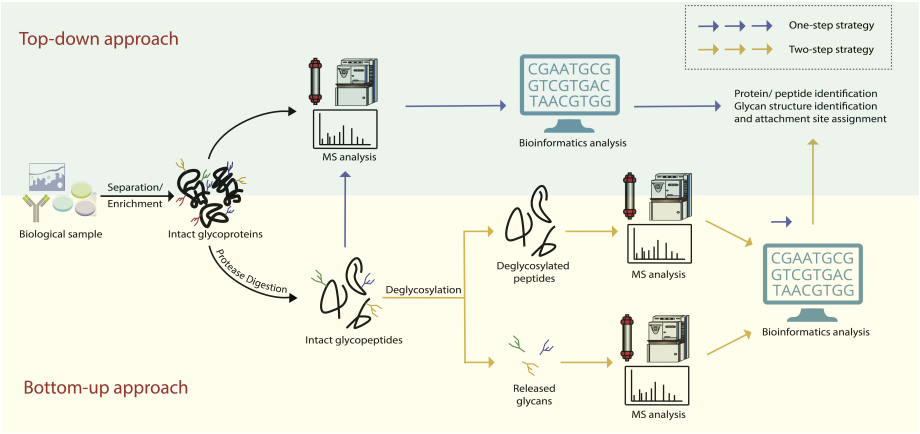

Figure 1.3: Schemed workflows of “top-down” and “bottom-up” techniques in glycoproteomics.

It is easier to identify proteins using the “bottom-up” approach since separation and analysis of peptides is considerably easier than the parent proteins. However, the “bottom-up” approach is suboptimal for determining glycosylation and alternative splicing isoforms. For the “top-down” approach, complete description of protein sequence and all modifications can be provided if adequate fragmentation information was obtained. Yet, the complexity of the analyte is obvious [40]. Moreover, pipeline integrating two approaches such as the so called “middle-down” approach have been developed, trying to achieve a balance between high-throughput and in-depth analysis [55, 13]. From another perspective, glycoproteomics strategies can be categorized into two main genres: “two-step” and “one-step”, depending on whether or not glycopeptide/glycoprotein is deglycosylated prior to MS/MS. Since “two-step” approach usually requires significant human interference, it cannot be developed into fully automated protocol. Thus, we pay more attention to “one-step” approaches in this review.

Similar to conventional proteomics, glycoproteomics also uses with two different strategies to automatically interpret MS data, which are database searching and de novo sequencing. Database searching, as its name suggests, is matching the theoretical spectrum from the database to the experimental result [36, 20]. Various scoring schemes for matching have been adopted in previous studies based on the common assumption that the higher the matching score, the more similar a match is and the higher probability that the spectrum is generated from the matched target. De novo sequencing predicts structure from MS by adopting algorithms like dynamic programming algorithm and heuristic algorithm [48, 29]. This approach is database independent, enabling identification of novel or unknown glycans.

This review describes the recent advances in several aspects, including glycopeptides enrichment and fragmentation, glycopeptide-spectrum matching algorithms, false discovery rate (FDR) estimation, machine learning involved approaches, and

potent workflow for comparison between search engines.

2 New Strategies for Glycopeptides Enrichment

The heterogeneity of glycans and the low abundance of many glycoproteins makes the glycoproteomics analytically challenging in complex samples. The procedures of glycoproteomics usually include glycopeptides enrichment prior to MS analysis. Current enrichment approaches can be generally classified into affinity-based and chemical-based. The former includes lectin-based [28], dioxide-based enrichment methods [12], and hydrophilic interaction capture [59, 21], while the latter contains borate chelating [63, 38, 41, 53] and hydrazide chemistry-based [23, 43]. However, most current enrichment strategies suffer from certain drawbacks. The lectin-based methods can only enrich subsets of glycopeptides because the limited ability of each lectin to recognize specific glycans. HILIC-based methods suffer from suboptimal specificity. Therefore, optimal enrichment methods are highly expected.

2.1 Boronic Acid (BA)-based Method: DBA Method

BA-based methods have shown great potential in universally enriching glycopeptides and glycoproteins by forming reversible covalent bonds with sugars. However, the interaction between BA and sugar is relatively weak. Recently, a research group has been engaged in searching a BA derivative to strengthen the interaction. Five different BA derivatives have been tested, and benzoboroxole was demonstrated to be most effective in glycopeptide enrichment. Since one glycan usually contains multiple monosaccharides, multiple benzoboroxole molecules were conjugated to one synthesized dendrimer to enable synergistic benzobotoxole-glycan interaction [58].

This Dendrimer-conjugated Boronic Acid derivative (DBA) method was then compared with some trending lectin- and HILIC (hydrophilic interaction liquid chromatography)-based methods. The results showed that DBA method identified more unique glycopeptides with lower non-glycopeptides rate, which indicated that DBA method outperformed the other two in both sensitivity and specificity.

2.2 HILIC-based Methods: MagG@Mg-MOFs-1C

Various hydrophilic materials have been developed for a higher efficiency of hydrophilic interaction capture, among which graphene derivative hydrophilic composites have shown great potential. Although hydrophilic dendrimer-assisted graphene-silica materials and zwitterion-modified hydrophilic composites have already been developed, their characteristics such as low surface area, no size-exclusion effect and magnetism, rendering suboptimal enrichment performance on complex biological samples [50, 24].

Metal-organic frameworks (MOFs), an emerging class of porous crystalline materials, are widely used in various areas, including enrichment of low abundance

biomolecules [9, 25, 51, 52]. In a recent study, a MOFs-based magnetic graphene composite, Mg-MODs-functionalized magnetic graphene (MagG) composite (MagG@ Mg-MOFs-1C) was synthesized following a three-step scheme [52]. This novel composite exhibited various properties such as high surface area, excellent hydrophilicity, suitable porous structure and abundant amount of affinity sites. Those properties enabled its remarkable performance in glycopeptide enrichment, which has been demonstrated in the experimental results.

3 New Strategies for Fragmentation

Mass spectrometry (MS) is the ideal and general tool for proteomics, glycomics and glycoproteomics analysis. Fragment ions generated by MS enable subsequent sequence analysis. Commonly used dissociation methods in MS include collision induced dissociation (CID), higher-energy collision dissociation (HCD), electron capture dissociation (ECD)/ electron transfer dissociation (ETD). CID and HCD usually breaks the glycosidic bond of glycopeptide and yields abundant B- and Y- ions Figure 3.1. On the other hand, ECD/ETD have extensive peptide backbone cleavage while preserve glycosylation PTMs, yielding c- and z- ions. The productivity and accuracy of glycopeptide analysis is highly dependent on fragmentation information acquired from MS. Therefore, for complete glycopeptide analysis, various pipelines have been developed to obtain richer fragmentation information.

The most commonly used strategies adopt a combination of orthogonal dissociation methods and suitable search engine. For example, Sweet-Heart adopted CID coupled with targeted MS3 [56], HCD-product-dependent-ETD (HCD-pd-ETD) were used in [45], [57] and [42]. In this chapter, two new fragmentation strategies proposed by the same group are introduced, which may provide fresh perspectives on complete spectral information acquisition.

3.1 Combination of Two Parallel Product-dependent Procedures in pGlyco

In the pipeline pGlyco [62], two HCD-product-dependent MS runs were implemented in parallel Figure 3.2:

- HCD-pd-CID-MS/MS: Precursors from full scan was first fragmented by HCD with normalized collision energy (NCE) at 40% [8]. Oxonium ion peaks have been used to distinguish glycopeptide spectra from MS/MS data [27]. The HCD-MS/MS results showed that with the peak 138.055 above the relative intensity of 30%, the spectra had other two glyco-oxonium ions with very high probabilities. Therefore, Therefore, ion 138.055 with at least 30% relative intensity was used as criterion for filtering real glycopeptide precursors. HCD-pd-CID was performed with the same precursor of HCD, and complementary fragments from each HCD/CID-MS/MS spectrum pair were used to filter the candidate Y_1 ion by counting the total number of matched trimannosyl core ions. Then, for each glycan candidate, corresponding theoretical mass of all the

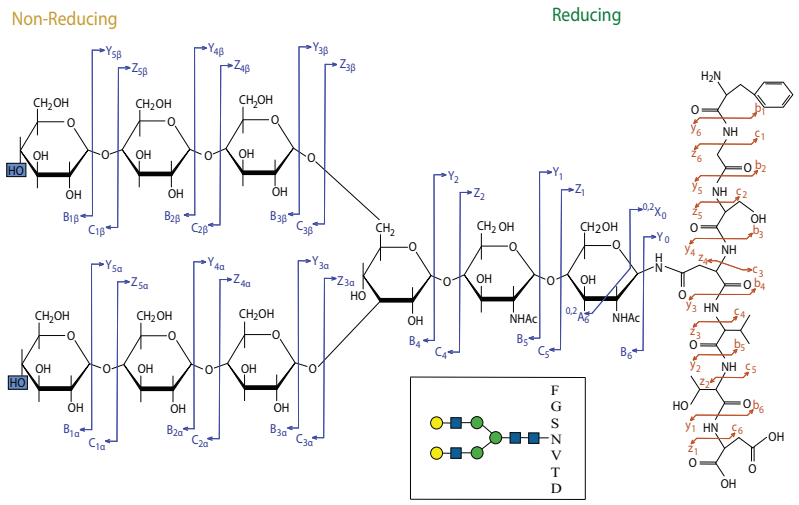

Figure 3.1: Fragmented ions examples in N-linked glycopeptide $FGSNVTD + GlcNAc_4Man_3Gal_2$. Glycan cleavages are shown in blue and yield A-, B-, C-ions and X-, Y-, Z-ions, where A/X- ions generated from cross ring cleavages. Peptide fragmentations are shown in red, generating a-, b-, c-, x-, y-, and z-ions. A simplified diagram of the same glycopeptide is shown in the inset.

Y ions can be easily calculated, and were matched with the HCD/CID-MS/MS spectrum pair. The match score for each glycan candidate could be used for ranking.

- HCD-pd-MS3: In this run, it was assumed that Y_1 ion must be one of the three most intense peaks with mass range above 700 m/z in HCD-MS/MS spectrum. Therefore, only these three ions were subjected to MS3, which were then used to identify peptide backbone using search engine for protein identification.

Based on the typical retention time window for a cluster of glycopeptides with the same peptide backbone, the glycan and peptide backbone identified by the two runs could be aligned based on the peptide backbone masses and the retention time. This pipeline allows a data-dependent automated identification of both peptide backbone and glycan, though the criteria to filter Y_1 ions for HCD-pd-MS3 may omit some glycopeptides due to the different fragmentation behaviors of different glycopeptides.

3.2 Stepped-energy Fragmentation in pGlyco 2.0

The accessibility, data quality, and overall throughput of complete glycopeptide identification is highly limited compared to proteomic analyses. Aiming to overcome the barriers, pGlyco 2.0 was released with the capability of precise high-throughput identification of intact N-glycopeptides at the proteome scale. Previous MS/MS analysis, like HCD-pd-CID-MS/MS or HCD-pd-MS3, requires information from several

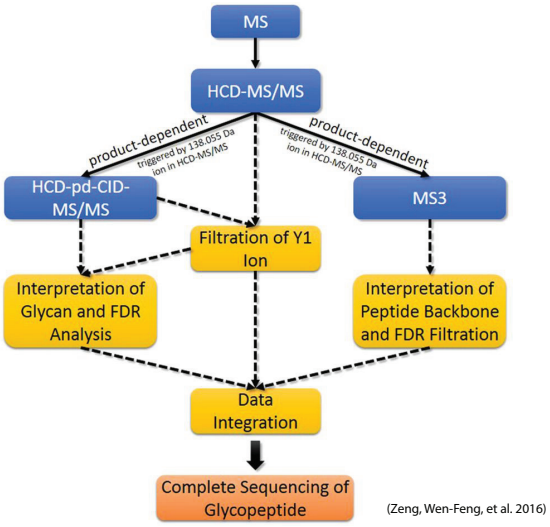

Figure 3.2: The overall workflow of pGlyco. The product-dependent CID-MS/MS and data-dependent MS3 analyses are performed separately. A solid line refers to data acquisition, and a dotted line refers to data interpretation.

spectrum to deduce glycan and peptide backbone. This style strategies compromise the overall throughput and quality. To obtain ideal MS/MS data that contains most comprehensive fragmentation in a single spectrum, a stepped-energy fragmentation strategy was proposed in pGlyco 2.0.

The fine-tuning of the fragmentation conditions includes two steps: To analyze a mixture of five standard glycoproteins, a series of LC-MS/MS with various dissociation parameters (CID, HCD, CTD coupled with CID (ETciD) and ETD coupled with HCD (EThcD)) and different energies were conducted and compared. It was found that HCD with multiple energies could produce complementary fragments of the glycan and the peptide. In the second step, 16 configurations of stepped collision energies (SCE) were simulated and compared, and the preferred SCE condition was selected. The explicit interpretation of the rich information generated by this style stepped-energy fragmentation will be introduced in section 4.1.

4 New Searching and Scoring Algorithm

Score of each theoretical glycan-, peptide-, or glycopeptide-spectrum matching is need to determine ranking. The scoring scheme could be cross-correlation score (Xcorr), probability-based score, or some other measures [22]. An ideal scoring scheme should be both sensitive and specific [26]. To achieve this goal, different scoring algorithms have been customized to interpret MS/MS data acquired from various platforms.

4.1 THREE LEVELS SCORING SCHEME IN PGLYCO 2.0

To interpret abundant information in SCE-HCD-MS/MS, searching engine pGlyco 2.0 was developed to do score against complete glycoproteome database by incorporating glycan and peptide scorings.

For each glycan in the database, the associated peptide backbone mass is the precursor mass minus the glycan mass, then the corresponding Y ions mass could be calculated as following:

$$M(Y\ ion) = M(peptide\ backbone) + M(reducing-terminal\ fragment\ of\ glycan) \tag{4.1}$$

Then score for each glycan candidate could be measured by matching Y ion masses against the HCD/CID-MS/MS spectrum, as shown below:

$$\begin{aligned} Score_G &= \sum_i \log(inten_i)\left(1-\left|\frac{merr_i}{tol_i}\right|^4\right)(ratio_{ion})^{\alpha}(ratio_{core})^{\beta} \\ ratio_{core} &= \frac{\#matched\ trimannosyl\ core\ ions}{\#theoretical\ trimannosyl\ core\ ions} \\ ratio_{ion} &= \frac{\#matched\ ions}{\#theoretical\ ions} \end{aligned} \tag{4.2}$$

The term $inten_i$ is the absolute intensity of a matched peak. The term tol_i refers to the matching mass tolerance of fragment ions, e.g. 20 ppm, and $merr_i$ referes to the matching mass error ranging from $-tol_i$ to tol_i. The score of each matched peak is weighted by a quartic polynomial function, $\left(1-\left|\frac{merr_i}{tol_i}\right|^4\right)$, which aims to penalize the larger mass errors with heavier penalties.

The top-ranked glycan candidates and corresponding peptide backbone masses for each spectrum pair were kept.

The peptide scoring was developed in a similar manner to glycan as following:

$$Score_P = \sum_i \log(inten_i)\left(1-\left|\frac{merr_i}{tol_i}\right|^4\right)(ratio_{ion})^{\gamma} \tag{4.3}$$

Then, the score of the glycopeptide was calculated as:

$$Score_{GP} = \omega * Score_G + (1-\omega) * Score_P \tag{4.4}$$

All four parameters α, β, γ, and ω can be fine-tuned by Ranking SVM, based on the principle of ranking as many correct matches onto top-1 as possible. In the FDR validation test, this scoring method have been proved to be more reliable with much lower FDRs in the glycan part than Byonic, which is also a search engine that could perform a generic database search against a complete glycoproteome database.

4.2 ENSEMBLE SCORE INTEGRATING MULTIPLE ANALYSIS

In a comprehensive, open-source, modular software for glycopreoteomics data analysis called GlycoPAT (GlycoProteomics Analysis Toolbox), a novel scoring scheme

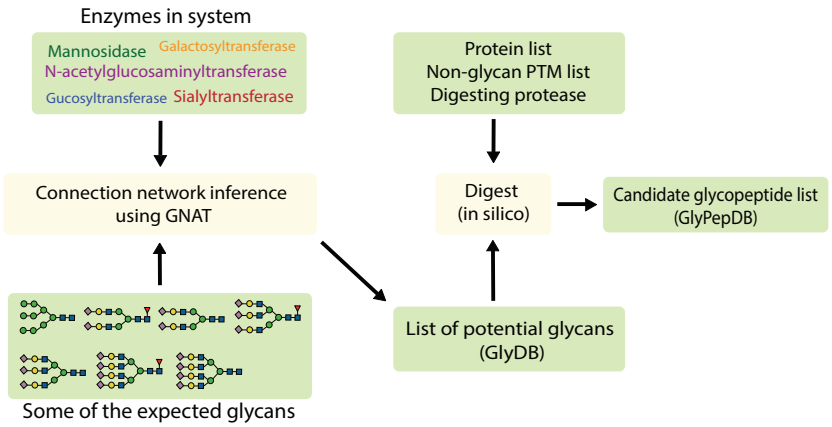

Figure 4.1: A schematic illustration of GlyPepDB generation. Connection inference routine described in [34] was adopted to generate the glycan search library (GlyDB), a list of potential O- and N-glycan PTMs to search for. The theoretical proteins was decorated with these glycans and additional fixed/variable nonglycan PTMs. Digestion of this glycoprotein by proteases results in the theoretical glycopeptide database (GlyPepDB).

was developed to calculate the so-called ensemble score (ES). ES integrated multiple statistical parameters including cross-correlation and probability-based scores [33].

The theoretical glycopeptide database (GlyPepDB) was generated in a workflow described in Figure 4.1. Then GlycoPAT scored in a two-step manner: First, a list of (glyco)peptide candidates were obtained by searching theoretical database with MS1 precursor mass. Second, the ES was calculated for each GPSM with noise reduction methods customized for different fragmentation modes:

- Xcorr analysis: The intensity of each peak in the theoretical spectra was set to 50. By calculating the Pearson correlation coefficient ($X_{corr}(\tau)$) between theoretical and experimental spectrum with parameter of lag τ, height center (HC_corr) was quantified as the normalized height of $X_{corr}(\tau = 0)$ with respect to the mean $X_{corr}(\tau)$. Then a scoring parameter s_1 was used to capture "good" match, which should possess small τ and large HC_{corr} as well.
- % ion match: % ion match was calculated as the percentage of theoretical peaks that have corresponding experimental matches. Then score s_2 was calculated based on the percentage.
- Top 10 peaks: s_3 is the portion of matched peaks among the 10 most intense experimental peaks.
- Poisson probability: To determine whether the predicted match was a chance event, twenty-five decoys were generated for each glycopeptide candidate in the strategy described in 5.3. Then, the p value for each candidate was calculated, based on which s_4 was obtained.

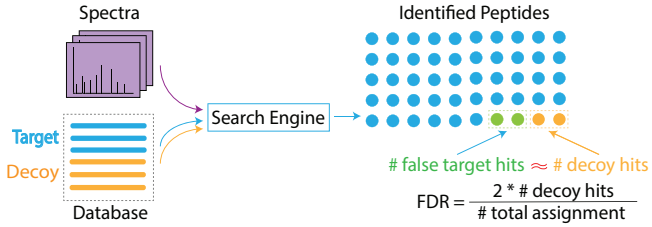

Figure 5.1: Fig. 22 Target-decoy database searching for FDR estimation.

ES was calculated as the weighted sum of the four parameters, where the individual weights could be adjusted on different fragmentation modes. Comparing to a popular commercial software Byonic, GlycoPAT considers the extensive fragmentation of glycans during scoring, whereas Byonic primarily only considers the oxonium ions, nonglycosylated peptide, peptide plus core HexNax and loss of sialic acid. These differences are implied in their experimental results, where GlycoPAT yielded higher scores in CID data analysis.

5 New FDR Estimation Methods

FDR is defined as the "expected" percentage of incorrect assignment among the accepted assignments [4]. It is extensively used for validation and quality control of the glycoprotein identification. Target-decoy strategy is a popular approach to estimate FDR, as Figure 5.1 shows.

This approach is based on two assumptions: (1) There is no overlap between target and decoy database; (2) The distribution of incorrect matches to target sequences is identical to that of matches to decoy peptides.

Since target-decoy strategy is considered to be both effective and robust, various strategies based on this idea have been proposed. The key step in target-decoy strategy is the generation of decoy database, for which different strategies have been developed, including Markov chain model, de novo generation [18, 65]. It is commonly thought that the decoy generation for glycan is more difficult than peptide, since glycan is of tree structure instead of a linear sequence as peptide. This makes evaluation of glycan identification more complicated than peptide. Actually, severely underestimation of FDR has been reported before [57]. To address this problem, new FDR strategies have been proposed in recent studies.

5.1 Generate Abundant Decoying Glycopeptides De Novo

The number of spectra that can be verified from glycopeptides is generally much less than 1000. As a result, when using the conventional approach for FDR estimation, the distribution of decoy glycopeptide matches may not accurately reflect that of incorrect matches to target glycopeptides [46, 17, 6]. In this situation, a research group argued that generating "decoys" based on the target protein sequence database

is not optimal in measuring the confidence of N-linked glycopeptide matches. To tackle this problem, they developed a tool named GPE (GlycoPep Evaluater), and later a more advanced tool called GDG (Glycopeptide Decoy Generator). Both of them are able to generate multiple glycopeptide decoys de novo for each target glycopeptide, and the later one can provide more user-defined features [65, 30]. The FDR could be estimated as follows:

$$FDR = \frac{\#\ decoy\ assignments}{\#\ total\ assignments}\left(1+\frac{\#\ targets}{\#\ decoys}\right) \tag{5.1}$$

To generate decoys for each target glycopeptide in GDG, a glycan was randomly selected from a library containing over 300 biologically relevant N-linked glycans and a peptide mass was generated with respect to several specified rules. The high decoy-to-target ratio guaranteed that sufficient decoy glycopeptides were available even for a small number of target glycopeptides. A curated set of 100 expert-assigned CID spectra of glycopeptides was adopted to evaluate the accuracy of a CID scoring algorithm: Glycopep Grader [54, 30]. The experimental result showed GDG could help to improve Glycopep Grader's accuracy in scoring spectra of fucosylated glycopeptide compositions. Furthermore, they provided evidence that abundant decoys were required for accurate assessment of tools that assign glycopeptides to MS/MS spectra [65].

Later, the same group proposed DecoyDeveloper, offering an interface for on-demand, de novo decoy glycopeptide generation [44]. For each target glycopeptide, a decoy was created in four steps: 1) selection of a decoy glycan from a database of 245 N-linked, biologically relevant glycans; 2) random generation of peptide sequences with various lengths; 3) selection of a tetrapeptide closest to the remaining mass; and 4) checking that the mass of the decoy glycopeptide was within the specified tolerance of the target. This workflow can be incorporated into any new or existing glycoproteomics analysis tool to efficiently produce decoys with high decoy: target ratio for further FDR estimation.

5.2 Spectrum-based Decoy Method and FDR Estimation for Glycan Identification

A so called spectrum-based decoy method was developed in pGlyco for the evaluation of glycan identification. This method is inspired by the ideal that generating a theoretical decoy spectrum against the experimental glycopeptide spectrum is easier than generating tree-based structure for a glycan structure [62]. Roughly, this method consisted of three steps:

1. A list of peptide backbone masses was deduced from an experimental spectrum. The corresponding Y ion masses could be calculated by the peptide backbone mass plus the masses of the reducing-terminal glycan fragments as described in 4.1. Then, the theoretical target glycopeptide spectrum could be generated.

2. A theoretical decoy spectrum could be generated by adding a random mass ranging from 1-30 Da to each deduced Y ion.

3. Matching both theoretical target and decoy spectra against the experimental spectrum, with possible bias adjusted by finite mixture model (FMM) [64], the distributions of correct and incorrect scores were drawn, and subsequently, FDR was estimated.

Since the assumption of target-decoy method might not hold when using the spectrum-based decoy method, FMM was adopted in the 3rd step to estimate the density functions of the correct ($f(x|+)$) and incorrect ($f(x|-)$) score distributions. This is realized as following steps:

1. $f(x|-)$ was modeled by the finite gamma-mixture model, which used several gamma distributions to fit decoy scores by the expectation- maximization (EM) algorithm, and the number of gamma components was determined by the Bayesian information criterion (BIC).
2. $f(x|+)$ was then estimated by finite Gaussian-mixture models and EM algorithm.
3. According to the Bayes rule, given a score x, the PEP can be expressed as:

$$PEP(x) = \pi_0 f(x|-)/[\pi_0 f(x|-) + (1-\pi_0) f(x|+)] \tag{5.2}$$

Where π_0 is mixture probability of incorrect identifications. The FDR at threshold x can be estimated by the equation: $FDR(x) = E[PEP(x)]$.

5.3 Combination of Random Scrambling of Peptide Backbone and Disturbance of Monosaccharide Mass

In the software GlycoPAT mentioned in 4.2, twenty-five decoys were generated for each glycopeptide candidate. Each decoy was obtained in two steps: (1) The peptide backbone sequence in the glycopeptide was randomly scrambled; (2) The molecular mass of each monosaccharide in the glycopeptide was added or subtracted by an arbitrary value between -50 to +50, while the total glycan mass was kept unaltered [33].

5.4 FDR Evaluation at Three Levels

In pGlyco 2.0, FDR estimation was estimated by the following formula:

$$\begin{aligned}\widehat{FDR}(x) =& p(G = false \cup P = false | X \geq x) \\ =& p(G = false | X \geq x) + p(P = false | X \geq x) - p(G = false \cap P = false | X \geq x) \\ =& \widehat{FDR}_G(x) + \widehat{FDR}_P(x) - \widehat{FDR}_{G\cap P}(x)\end{aligned} \tag{5.3}$$

For a glycopeptide spectrum match (GSPM), a false identification refers to an incorrect identification of either glycan or peptide. In the formula, $\widehat{FDR}_G(x)$ was calculated as previously described in 5.2. $\widehat{FDR}_P(x)$ and $\widehat{FDR}_{G\cap P}(x)$ were estimated in conventional way. This novel FDR estimation enabled accuracy of intact glycopeptide identification by integrating three levels of matches to glycans, peptides and glycopeptides.

6 Machine Learning Algorithms Applied to Glycoproteomics

Machine learning, as a branch of artificial intelligence, aims at developing algorithms to summarize large and complex data sets. With more and more related applications available for prediction problems, including glycoproteomics research, machine learning shows promising potential for a better explanation of glycoproteomics data. Here, we describe some of the recent glycoproteomics research involving machine learning algorithms.

6.1 Random Forest Model in Sweet-Heart

In the modular computational suite named Sweet-Heart, more than five machine learning algorithms have been evaluated for the performance of glycopeptide prediction, including support vector machine (SVM) and random forest (RF). Thirty six normalized spectral features were used for machine learning and model learning. RF was finally adopted due to its better performance, simplicity in parameter setting, and stability with respect to the feature selection [56].

Later, this group built a workflow to enable adaptive model optimization with respect to different sampling strategies, training sample sizes and feature sets [31]. The automated workflow could be realized in three steps:

1. An initial RF model on 106 manually validated spectra using 36 normalized spectral features was incorporated in Sweet-Heart to facilitate data collection for future model improvement.

2. A separate automated workflow was implemented with all spectra data and flexible specifications in parameters, including feature sets, sampling methods for training dataset, training sample sizes, number of trees in RF model, and k-fold cross-validation.

3. The workflow could dynamically evaluate the performance of RF model with specified parameters and enable optimization of parameters.

The training of the model revealed that factors including feature set, type of glycoforms, training sample size and sampling method contributed to model performance.

6.2 Tuning Scoring Function Using Structured SVMs with Latent Variables in De Novo Sequencing of Glycans

Spectrum graphs were commonly used in de novo peptide sequencing, and could also be employed in de novo glycan sequencing [3, 29, 7, 14, 2]. In a recent work, spectrum graph is represented by $G = (V, E)$, where each vertex $v \in V$ is specified by tuple (m, r, b), where m is the prefix residue mass (PRM), r is the residue type of the root of the corresponding tree, and b is the linkage type. [29] Given the MS/MS spectrum, the problem of recovering a glycan tree was formulated as integer programming (IP), which solved assignments of each observed peak to vertex in the spectrum graph.

A glycan structure was represented by $g = (x, y, z)$ and a peak assignment was q. A series of constraints were defined to satisfy the consistency of the glycan structure and the peak assignment, under which the objective function $S(g, q; P)$ was to be maximized:

$$\begin{aligned} S_\lambda(g, q; P) &= \sum_{i=1}^{|V|} \{\sigma(v_i)x_i + \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \tau(e_{ij})y_{ij} + \sum_{i=1}^{|V|} \rho(v_i)z_i \\ \sigma(v_i) &= \sum_{p_j \in (SP(v_j))} \phi(r_i, b_i, CP(v_i, p_i))q_{ji} \end{aligned} \tag{6.1}$$

Direct optimization with these constraints is difficult, therefore Lagrangian relaxation was used, then the primal problem was substituted by minimizing dual objective function with the subgradient method [5].

To tune the score functions $\phi(r, b, c)$ (cleavage), $\tau(e_{ij})$ (bond) and $\rho(v_i)$ (branching), structured SVM was adopted with latent variables q, under the assumption that peaks will not be assigned to substructures [61]. Given the training dataset with experimental MS/MS spectrum and its glycan structure, the goal is to find λ to minimize the following objective function:

$$f(\lambda) = \sum_{(P,g)\in\mathcal{D}} \max_{\hat{g},\hat{q}} [S_\lambda(\hat{g}, \hat{q}; P) + \Delta(g, \hat{g})] - \max_q S_\lambda(g, q; p) + C\|\lambda\|_1) \tag{6.2}$$

This could be realized by the stochastic subgradient descent or the forward-backward splitting (FOBOS) [11]. Their experiments on both N-linked glycans and O-linked glycans showed that tuned parameters significantly improved the accuracy of de novo sequencing in three level measures: the composition of monosaccharides, glycosidic bonds, and the glycan tree structure. This suggests that training structured SVM could learn the preference of internal cleavages as well as that of glycan structures that might have some sort of dependence on enzymatic pathways that synthesize glycoprotein.

6.3 Ranking SVM in pGlyco 2.0

In pGlyco 2.0, ranking SVM is adopted to fine-tune four parameters α, β, γ and ω in the scoring function of peptide and glycopeptide matches in 4.1 [35]. The ideal scoring scheme is able to always rank the correct match as top-1. The goal of the parameter tuning was to rank as many correct matches onto top-1 as possible. The logarithm of $Score_G$ could transfer the parameters from exponential to linear form.

7 Potent workflow for comparison between search engines

To evaluate the accuracy of pipelines, a new quantitative analysis method was proposed in pGlyco 2.0 [35]. Using a mixed sample of unlabeled and $^{15}N/^{13}C$ metabolically labeled glycoproteome, this method can validate the FDR estimation of an engine. Two FDR methods were developed as following:

- Isotope-based FDR: Given a GPSM, a quantification software tool pQuant was employed to find the signal pair of unlabeled and labeled glycopeptide precursors in full MS [32]. pQuant output a NaN ratio if it failed to find a pair or the ratio was far from 1:1. Then the isotope-based FDR could be calculated as the portion of GPSMs associated with NaN ratios in target GPSMs scaled by the rate of NaN ratios in all decoy GPSMs, which should all be false positive.

$$isotope-based\ FDR = \frac{\frac{\#\ NaN\ in\ all\ target\ GPSMs}{\#\ all\ target\ GPSMs}}{\frac{\#\ NaN\ in\ all\ decoy\ GPSMs}{\#\ all\ decoy\ GPSMs}} \tag{7.1}$$

- Entrapment-based FDR: A combination of mouse glycome and proteome database was used as the entrapment database, then the entrapment-based FDR could be calculated as the percentage of GPSMs with a mouse-only peptide or mouse-only glycan:

$$entrapment-based\ FDR = \frac{\#\ GPSMs\ with\ mouse-only\ peptide\ or\ glycan}{\#\ all\ GPSMs} \tag{7.2}$$

This workflow was tested on Byonic and pGlyco 2.0, both of which could perform generic search against complete glycoproteome database. The result showed that pGlyco 2.0 returned lower FDR than preset FDR value, while FDR value of Byonic was severely underestimated

8 Conclusions

Glycoproteomics approaches have seen a continual evolution with new experimental technologies and analysis strategies in recent years. In this survey we have reviewed the most significant improvements mainly in database-searching-based strategies. Although de novo sequencing is usually considered to be slow and inaccurate, it shows superiority in some aspects. So far, a considerable number of techniques have been developed for de novo sequencing, including hidden Markov models [22], machine learning and deep learning [22, 10, 60, 49]. With more experimental and analytical improvements in these subfields, it is natural to speculate that the determination of glycosylation profiles will become more accurate and efficient.

References

[1] Hyun Joo An, Scott R Kronewitter, Maria Lorna A de Leoz, and Carlito B Lebrilla. Glycomics and disease markers. *Current opinion in chemical biology*, 13(5-6):601–607, 2009.

[2] Sandro Andreotti, Gunnar W Klau, and Knut Reinert. AntilopeâĂŤa lagrangian relaxation approach to the de novo peptide sequencing problem. *IEEE/ACM transactions on computational biology and bioinformatics*, 9(2):385–394, 2011.

[3] Christian Bartels. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical & environmental mass spectrometry*, 19(6):363–368, 1990.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[5] Korte Bernhard and J Vygen. Combinatorial optimization: Theory and algorithms. *Springer, Third Edition, 2005.*, 2008.

[6] Robert J Chalkley. When target–decoy false discovery rate estimations are inaccurate and how to spot instances. *Journal of proteome research*, 12(2):1062–1064, 2013.

[7] Vlado Dančík, Theresa A Addona, Karl R Clauser, James E Vath, and Pavel A Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology*, 6(3-4):327–342, 1999.

[8] Loïc Dayon and Jean-Charles Sanchez. Relative protein quantification by ms/ms using the tandem mass tag technology. In *Quantitative Methods in Proteomics*, pages 115–127. Springer, 2012.

[9] Pravas Deria, Diego A GÃşmez-GualdrÃşn, Idan Hod, Randall Q Snurr, Joseph T Hupp, and Omar K Farha. Framework-topology-dependent catalytic activity of zirconium-based (porphinato) zinc (ii) mofs. *Journal of the American Chemical Society*, 138(43):14449–14457, 2016.

[10] Arun Devabhaktuni and Joshua E Elias. Application of de novo sequencing to large-scale complex proteomics data sets. *Journal of proteome research*, 15(3):732–742, 2016.

[11] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[12] Kasper Engholm-Keller and Martin R Larsen. Titanium dioxide as chemo-affinity chromatographic sorbent of biomolecular compoundsâĂŤapplications in acidic modification-specific proteomics. *Journal of proteomics*, 75(2):317–328, 2011.

[13] Luca Fornelli, Daniel Ayoub, Konstantin Aizikov, Alain Beck, and Yury O Tsybin. Middle-down analysis of monoclonal antibodies with electron transfer dissociation orbitrap fourier transform mass spectrometry. *Analytical chemistry*, 86(6):3005–3012, 2014.

[14] Ari Frank and Pavel Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4):964–973, 2005.

[15] RH Garrett and Ch M Grisham. Biochemistry, 1999. *SaunderâĂŹs College Publishing*.

[16] Ylva Gavel and Gunnar von Heijne. Sequence differences between glycosylated and non-glycosylated asn-x-thr/ser acceptor sites: implications for protein engineering. *Protein Engineering, Design and Selection*, 3(5):433–442, 1990.

[17] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry*, 22(7):1111–1120, 2011.

[18] Wilhelm Haas, Brendan K Faherty, Scott A Gerber, Joshua E Elias, Sean A Beausoleil, Corey E Bakalarski, Xue Li, Judit Villen, and Steven P Gygi. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Molecular & Cellular Proteomics*, 5(7):1326–1337, 2006.

[19] Franz-Georg Hanisch. O-glycoproteomics: Site-specifi c o-glycoprotein analysis by cid/etd electrospray ionization tandem mass spectrometry and top-down glycoprotein sequencing by in-source decay maldi mass spectrometry. In *Mucins*, pages 179–189. Springer, 2012.

[20] Lin He, Lei Xin, Baozhen Shan, Gilles A Lajoie, and Bin Ma. Glycomaster db: software to assist the automated identification of n-linked glycopeptides by tandem mass spectrometry. *Journal of proteome research*, 13(9):3881–3895, 2014.

[21] Xiao-Mei He, Xi-Chao Liang, Xi Chen, Bi-Feng Yuan, Ping Zhou, Li-Na Zhang, and Yu-Qi Feng. High strength and hydrophilic chitosan microspheres for the selective enrichment of n-glycopeptides. *Analytical chemistry*, 89(18):9712–9721, 2017.

[22] Andrew P Horton, Scott A Robotham, Joe R Cannon, Dustin D Holden, Edward M Marcotte, and Jennifer S Brodbelt. Comprehensive de novo peptide sequencing from ms/ms pairs generated through complementary collision induced dissociation and 351 nm ultraviolet photodissociation. *Analytical chemistry*, 89(6):3747–3753, 2017.

[23] Junfeng Huang, Hao Wan, Yating Yao, Jinan Li, Kai Cheng, Jiawei Mao, Jin Chen, Yan Wang, Hongqiang Qin, Weibing Zhang, et al. Highly efficient release of glycopeptides from hydrazide beads by hydroxylamine assisted pngase f deglycosylation for n-glycoproteome analysis. *Analytical chemistry*, 87(20):10199–10204, 2015.

[24] Bo Jiang, Yu Liang, Qi Wu, Hao Jiang, Kaiguang Yang, Lihua Zhang, Zhen Liang, Xiaojun Peng, and Yukui Zhang. New go–pei–au–l-cys zic-hilic composites: synthesis and selective enrichment of glycopeptides. *Nanoscale*, 6(11):5616–5619, 2014.

[25] Wei Jiao, Jiaxing Zhu, Yun Ling, Mingli Deng, Yaming Zhou, and Pingyun Feng. Photoelectrochemical properties of mof-induced surface-modified tio 2 photoelectrode. *Nanoscale*, 10(43):20339–20346, 2018.

[26] Eugene Kapp and Frédéric Schütz. Overview of tandem mass spectrometry (ms/ms) database search algorithms. *Current protocols in protein science*, 49(1):25–2, 2007.

[27] Ju-Wan Kim, Heeyoun Hwang, Jong-Sun Lim, Hyoung-Joo Lee, Seul-Ki Jeong, Jong Shin Yoo, and Young-Ki Paik. gfinder: a web-based bioinformatics tool for the analysis of n-glycopeptides. *Journal of proteome research*, 15(11):4116–4125, 2016.

[28] Kazutosi Kubota, Yuji Sato, Yusuke Suzuki, Naoko Goto-Inoue, Tosifusa Toda, Minoru Suzuki, Shin-ichi Hisanaga, Akemi Suzuki, and Tamao Endo. Analysis of glycopeptides using lectin affinity chromatography with maldi-tof mass spectrometry. *Analytical chemistry*, 80(10):3693–3698, 2008.

[29] Shotaro Kumozaki, Kengo Sato, and Yasubumi Sakakibara. A machine learning based approach to de novo sequencing of glycans from tandem mass spectrometry spectrum. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(6):1267–1274, 2015.

[30] Jude C Lakbub, Xiaomeng Su, Zhikai Zhu, Milani W Patabandige, David Hua, Eden P Go, and Heather Desaire. Two new tools for glycopeptide analysis researchers: a glycopeptide decoy generator and a large data set of assigned cid spectra of glycopeptides. *Journal of proteome research*, 16(8):3002–3008, 2017.

[31] Suh-Yuen Liang, Sz-Wei Wu, Tsung-Hsien Pu, Fang-Yu Chang, and Kay-Hooi Khoo. An adaptive workflow coupled with random forest algorithm to identify intact n-glycopeptides detected from mass spectrometry. *Bioinformatics*, 30(13):1908–1916, 2014.

[32] Chao Liu, Chun-Qing Song, Zuo-Fei Yuan, Yan Fu, Hao Chi, Le-Heng Wang, Sheng-Bo Fan, Kun Zhang, Wen-Feng Zeng, Si-Min He, et al. pquant improves quantitation by keeping out interfering signals and evaluating the accuracy of calculated ratios. *Analytical chemistry*, 86(11):5286–5294, 2014.

[33] Gang Liu, Kai Cheng, Chi Y Lo, Jun Li, Jun Qu, and Sriram Neelamegham. A comprehensive, open-source platform for mass spectrometry-based glycoproteomics data analysis. *Molecular & Cellular Proteomics*, 16(11):2032–2047, 2017.

[34] Gang Liu and Sriram Neelamegham. A computational framework for the automated construction of glycosylation reaction networks. *PloS one*, 9(6):e100939, 2014.

[35] Ming-Qi Liu, Wen-Feng Zeng, Pan Fang, Wei-Qian Cao, Chao Liu, Guo-Quan Yan, Yang Zhang, Chao Peng, Jian-Qiang Wu, Xiao-Jin Zhang, et al. pglyco 2.0 enables precision n-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nature communications*, 8(1):438, 2017.

[36] Klaus Karl Lohmann and Claus-W von der Lieth. Glycofragment and glycosearchms: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic acids research*, 32(suppl_2):W261–W266, 2004.

[37] Willy Morelle and Jean-Claude Michalski. Analysis of protein glycosylation by mass spectrometry. *Nature protocols*, 2(7):1585, 2007.

[38] Pir Muhammad, Xueying Tu, Jia Liu, Yijia Wang, and Zhen Liu. Molecularly imprinted plasmonic substrates for specific and ultrasensitive immunoassay of trace glycoproteins in biological samples. *ACS applied materials & interfaces*, 9(13):12082–12091, 2017.

[39] Hisashi Narimatsu, Hiromichi Sawaki, Atsushi Kuno, Hiroyuki Kaji, Hiromi Ito, and Yuzuru Ikehara. A strategy for discovery of cancer glyco-biomarkers in serum using newly developed technologies for glycoproteomics. *The FEBS journal*, 277(1):95–105, 2010.

[40] Simone Nicolardi, Linda Switzar, AndrÃl' M Deelder, Magnus Palmblad, and Yuri EM van der Burgt. Top-down maldi-in-source decay-fticr mass spectrometry of isotopically resolved proteins. *Analytical chemistry*, 87(6):3429–3437, 2015.

[41] Yanyan Qu, Jianxi Liu, Kaiguang Yang, Zhen Liang, Lihua Zhang, and Yukui Zhang. Boronic acid functionalized core–shell polymer nanoparticles prepared by distillation precipitation polymerization for glycopeptide enrichment. *Chemistry–A European Journal*, 18(29):9056–9062, 2012.

[42] Julian Saba, Sucharita Dutta, Eric Hemenway, and Rosa Viner. Increasing the productivity of glycopeptides analysis by using higher-energy collision dissociation-accurate mass-product-dependent electron transfer dissociation. *International journal of proteomics*, 2012, 2012.

[43] Muhammad Salman Sajid, Fahmida Jabeen, Dilshad Hussain, Muhammad Naeem Ashiq, and Muhammad Najam-ul Haq. Hydrazide-functionalized affinity on conventional support materials for glycopeptide enrichment. *Analytical and bioanalytical chemistry*, 409(12):3135–3143, 2017.

[44] Joshua T Shipman, Xiaomeng Su, David Hua, and Heather Desaire. Decoy-developer: An on-demand, de novo decoy glycopeptide generator. *Journal of proteome research*, 18(7):2896–2902, 2019.

[45] Charandeep Singh, Cleidiane G Zampronio, Andrew J Creese, and Helen J Cooper. Higher energy collision dissociation (hcd) product ion-triggered electron transfer dissociation (etd) mass spectrometry for the analysis of n-linked glycoproteins. *Journal of proteome research*, 11(9):4517–4525, 2012.

[46] John S Strum, Charles C Nwosu, Serenus Hua, Scott R Kronewitter, Richard R Seipert, Robert J Bachelor, Hyun Joo An, and Carlito B Lebrilla. Automated

assignments of n-and o-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures. *Analytical chemistry*, 85(12):5666–5675, 2013.

[47] Bingyun Sun, Jeffrey A Ranish, Angelita G Utleg, James T White, Xiaowei Yan, Biaoyang Lin, and Leroy Hood. Shotgun glycopeptide capture approach coupled with mass spectrometry for comprehensive glycoproteomics. *Molecular & cellular proteomics*, 6(1):141–149, 2007.

[48] Weiping Sun, Gilles A Lajoie, Bin Ma, and Kaizhong Zhang. A novel algorithm for glycan de novo sequencing using tandem mass spectrometry. In *International Symposium on Bioinformatics Research and Applications*, pages 320–330. Springer, 2015.

[49] Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, and Ming Li. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods*, 16:63–66, 2019.

[50] Hao Wan, Junfeng Huang, Zhongshan Liu, Jinan Li, Weibing Zhang, and Hanfa Zou. A dendrimer-assisted magnetic graphene–silica hydrophilic composite for efficient and selective enrichment of glycopeptides from the complex sample. *Chemical Communications*, 51(45):9391–9394, 2015.

[51] Bin Wang, Xiu-Liang Lv, Dawei Feng, Lin-Hua Xie, Jian Zhang, Ming Li, Yabo Xie, Jian-Rong Li, and Hong-Cai Zhou. Highly stable zr (iv)-based metal–organic frameworks for the detection and removal of antibiotics and organic explosives in water. *Journal of the American Chemical Society*, 138(19):6204–6216, 2016.

[52] Jiaxi Wang, Jie Li, Guoquan Yan, Mingxia Gao, and Xiangmin Zhang. Preparation of a thickness-controlled mg-mofs-based magnetic graphene composite as a novel hydrophilic matrix for the effective identification of the glycopeptide in the human urine. *Nanoscale*, 11(8):3701–3709, 2019.

[53] Yanan Wang, Jiaxi Wang, Mingxia Gao, and Xiangmin Zhang. An ultra hydrophilic dendrimer-modified magnetic graphene with a polydopamine coating for the selective enrichment of glycopeptides. *Journal of Materials Chemistry B*, 3(44):8711–8716, 2015.

[54] Carrie L Woodin, David Hua, Morgan Maxon, Kathryn R Rebecchi, Eden P Go, and Heather Desaire. Glycopep grader: a web-based utility for assigning the composition of n-linked glycopeptides. *Analytical chemistry*, 84(11):4821–4829, 2012.

[55] Si Wu, Nikola Tolić, Zhixin Tian, Errol W Robinson, and Ljiljana Paša-Tolić. An integrated top-down and bottom-up strategy for characterization of protein isoforms and modifications. In *Bioinformatics for Comparative Proteomics*, pages 291–304. Springer, 2011.

[56] Sz-Wei Wu, Suh-Yuen Liang, Tsung-Hsien Pu, Fang-Yu Chang, and Kay-Hooi Khoo. Sweet-heartâĂŤan integrated suite of enabling computational tools for automated ms2/ms3 sequencing and identification of glycopeptides. *Journal of proteomics*, 84:1–16, 2013.

[57] Sz-Wei Wu, Tsung-Hsien Pu, Rosa Viner, and Kay-Hooi Khoo. Novel lc-ms2 product dependent parallel data acquisition function and data analysis workflow for sequencing and identification of intact glycopeptides. *Analytical chemistry*, 86(11):5478–5486, 2014.

[58] Haopeng Xiao, Weixuan Chen, Johanna M Smeekens, and Ronghu Wu. An enrichment method based on synergistic and reversible covalent interactions for large-scale analysis of glycoproteins. *Nature communications*, 9(1):1692, 2018.

[59] Zhichao Xiong, Hongqiang Qin, Hao Wan, Guang Huang, Zhang Zhang, Jing Dong, Lingyi Zhang, Weibing Zhang, and Hanfa Zou. Layer-by-layer assembly of multilayer polysaccharide coated magnetic nanoparticles for the selective enrichment of glycopeptides. *Chemical Communications*, 49(81):9284–9286, 2013.

[60] Hao Yang, Hao Chi, Wen-Jing Zhou, Wen-Feng Zeng, Chao Liu, Rui-Min Wang, Zhao-Wei Wang, Xiu-Nan Niu, Zhen-Lin Chen, and Si-Min He. psite: Amino acid confidence evaluation for quality control of de novo peptide sequencing and modification site localization. *Journal of proteome research*, 17(1):119–128, 2017.

[61] Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *ICML*, volume 2, page 5, 2009.

[62] Wen-Feng Zeng, Ming-Qi Liu, Yang Zhang, Jian-Qiang Wu, Pan Fang, Chao Peng, Aiying Nie, Guoquan Yan, Weiqian Cao, Chao Liu, et al. pglyco: a pipeline for the identification of intact n-glycopeptides by using hcd-and cid-ms/ms and ms3. *Scientific reports*, 6:25102, 2016.

[63] Huaiyuan Zhang, Guoping Yao, Chunhui Deng, Haojie Lu, and Pengyuan Yang. Facile synthesis of boronic acid-functionalized magnetic mesoporous silica nanocomposites for highly specific enrichment of glycopeptides. *Chinese Journal of Chemistry*, 29(4):835–839, 2011.

[64] Jiyang Zhang, Jie Ma, Lei Dou, Songfeng Wu, Xiaohong Qian, Hongwei Xie, Yunping Zhu, and Fuchu He. Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics. *Molecular & Cellular Proteomics*, 8(3):547–557, 2009.

[65] Zhikai Zhu, Xiaomeng Su, Eden P Go, and Heather Desaire. New glycoproteomics software, glycopep evaluator, generates decoy glycopeptides de novo and enables accurate false discovery rate analysis for small data sets. *Analytical chemistry*, 86(18):9212–9219, 2014.