Analytical and Bioanalytical Chemistry (2021) 413:7215–7227
https://doi.org/10.1007/s00216-021-03621-z

PAPER IN FOREFRONT

The opportunity cost of automated glycopeptide analysis: case study profiling the SARS-CoV-2 S glycoprotein

Eden P. Go[1] · Shijian Zhang[2,3] · Haitao Ding[4] · John C. Kappes[4,5] · Joseph Sodroski[2,3,6] · Heather Desaire[1]

Received: 28 March 2021 / Revised: 29 June 2021 / Accepted: 16 August 2021 / Published online: 27 August 2021

Abstract

Glycosylation analysis of viral glycoproteins contributes significantly to vaccine design and development. Among other benefits, glycosylation analysis allows vaccine developers to assess the impact of construct design or producer cell line choices for vaccine production, and it is a key measure by which glycoproteins that are produced for use in vaccination can be compared to their native viral forms. Because many viral glycoproteins are multiply glycosylated, glycopeptide analysis is a preferrable approach for mapping the glycans, yet the analysis of glycopeptide data can be cumbersome and requires the expertise of an experienced analyst. In recent years, a commercial software product, Byonic, has been implemented in several instances to facilitate glycopeptide analysis on viral glycoproteins and other glycoproteomics data sets, and the purpose of the study herein is to determine the strengths and limitations of using this software, particularly in cases relevant to vaccine development. The glycopeptides from a recombinantly expressed trimeric S glycoprotein of the SARS-CoV-2 virus were first analyzed using an expert-based analysis strategy; subsequently, analysis of the same data set was completed using Byonic. Careful assessment of instances where the two methods produced different results revealed that the glycopeptide assignments from Byonic contained more false positives than true positives, even when the data were assessed using a 1% false discovery rate. The work herein provides a roadmap for removing the spurious assignments that Byonic generates, and it provides an assessment of the opportunity cost for relying on automated assignments for glycopeptide data sets from viral glycoproteins.

Keywords SARS-CoV-2 · Glycopeptide · Glycoprotein · Mass spectrometry

Introduction

The study of glycosylation on viral glycoproteins has been an important aspect of vaccine development for more than three decades [1, 2]. In recent years, the HIV-1 field has taken full advantage of this technology, harnessing glycopeptide analysis to advance vaccine discovery and development efforts on several fronts. Multiple laboratories have shown that native, trimeric HIV-1 Env glycoproteins, either isolated from cell surfaces or generated in ways that stabilize the trimeric conformation, have distinctly different glycosylation profiles than monomeric or even trimeric forms of the same protein that are poorly folded [3–5]. Moreover, differences in glycosylation between membrane-anchored HIV-1 Env glycoproteins and even well-folded soluble, stabilized Env trimers have been observed [6, 7]. Thus, the glycosylation profile of the protein provides useful information about the recombinant protein's ability to mimic its viral counterpart. Glycopeptide analysis also has been used to verify that vaccine immunogens with selected Env glycosylation sites removed can still retain the

Published in the topical collection *Analytical Characterization of Viruses* with guest editor Joseph Zaia.

✉ Heather Desaire
hdesaire@ku.edu

[1] Department of Chemistry, University of Kansas, Lawrence, KS 66049, USA

[2] Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA

[3] Department of Microbiology, Harvard Medical School, Boston, MA 02215, USA

[4] Department of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[5] Birmingham Veterans Affairs Medical Center, Research Service, Birmingham, AL 35233, USA

[6] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA

native glycosylation profile at non-deleted sites [8, 9]. Finally, glycopeptide analysis studies are a cornerstone method used to assess new protein production platforms that may be amenable to generating large quantities of recombinantly expressed protein [10, 11]; these advances in scale are a necessary prerequisite for mass vaccination programs that rely on recombinant proteins.

While the analysis of glycosylation on proteins can be conducted by either releasing the glycans or analyzing glycopeptides [12–14], the glycopeptide-based approach is most desirable for multiply glycosylated viral proteins, since analyzing glycopeptides affords the opportunity to identify the glycosylation pattern at individual sites. In both the established field of research in support of HIV-1 vaccine development, where the Env glycoprotein is the target immunogen [3, 4, 8, 9, 11, 15], and the emerging field of SARS-CoV-2 studies, where the spike (S) glycoprotein is the primary immunogen of interest [16–19], researchers predominantly use the glycopeptide-based analysis approach. While consensus is emerging on the overall optimal platform of glycopeptide analysis, the field is still undecided about the best way to assign the LC-MS data acquired in this experiment.

Over the last 13 years, we have used the same workflow for glycopeptide analysis on multiple projects, both involving HIV-1 Env and other glycosylated proteins [3, 4, 20–23]. After acquiring LC-MS/MS data on tryptic digests of glycoproteins, data files are assigned with the help of database tools, but most of the work is completed through expert manual assignment of the high-resolution MS and MS/MS data. This approach, while effective at providing a thorough analysis of the data, is cumbersome by today's standards and is not easily transmittable to laboratories with limited expertise in manual assignment of MS/MS data of glycopeptides. While caveats exist, manual analysis is still considered the gold standard in the field, and it has been used in a few instances in SARS-CoV-2 S glycoprotein studies to provide deep glycosylation coverage [24, 25].

In the work described herein, we assessed the possibility of replacing the expert-based analysis strategy with one that is becoming more widely adopted: the use of the automated analysis tool, Byonic. This software product is the most common choice for the analysis of glycopeptides from the SARS-CoV-2 S glycoprotein; at least five different examples are already published where Byonic was used to analyze the data [18, 25–28]. The use of such a product is surely alluring: a complete protein analysis can be done in hours instead of days or weeks. The product allows users to select a 1% false discovery rate (FDR), providing assurance that the output has believability. Finally, the human expertise component is removed from the equation, so laboratories without a specific focus in glycopeptide analysis could use it. The potential benefits are high. But what remains unaddressed in the literature is: What is the opportunity cost? What is given up by implementing this strategy? The study described herein was designed specifically to answer these questions.

Experimental

The protein expression and purification have been described previously, [29], and the detailed protocol is included here for the reader's convenience.

SARS-CoV-2 S glycoprotein expression vector Inducible expression of the wild-type SARS-CoV-2 S glycoprotein (with Asp 614) was achieved using a self-inactivating lentivirus vector comprising TRE3g-SARS-CoV-2-Spike-6His.IRS6A.Puro-T2A-GFP (K5650). In this vector, the codon-optimized *S* gene is under the control of a tetracycline response element (TRE) promoter and encodes the wild-type S glycoprotein with a carboxy-terminal 2xStrep-Tag II sequence. The internal ribosome entry site (IRES) allows expression of puro. T2A.EGFP, in which puromycin N-acetyltransferase and enhanced green fluorescent protein (eGFP) are produced by self-cleavage at the Thosea asigna 2A (T2A) sequence.

Cell lines The wild-type SARS-CoV-2 S glycoprotein, with Asp 614, was inducibly expressed in Lenti-x-293T human female kidney cells from Takara Bio (Catalog #: 632180). Lenti-x-293T cells were grown in DMEM with 10% heat-inactivated FBS (purchased from Gibco; Amarillo, TX) supplemented with L-glutamine and Pen-Strep. Lenti-x-293T cells constitutively expressing the reverse tetracycline-responsive transcriptional activator (rtTA) (Lenti-x-293T-rtTa cells (D1317)) [30] were used as the parental cells for the 293T-S cell line. The 293T-S (D1483) cells inducibly expressing the wild-type SARS-CoV-2 S glycoprotein with a carboxy-terminal 2xStrep-Tag II sequence were produced by transduction of Lenti-x-293T-rtTA cells with the K5650 recombinant lentivirus vector described above. The packaged K5650 lentivirus vector (60 μl volume) was incubated with 2 × 10^5 Lenti-x-293T-rtTA cells in DMEM, tumbling at 37 °C overnight. The cells were then transferred to a 6-well plate in 3 mL DMEM/10% FBS/Pen-Strep and subsequently selected with 10 μg/mL puromycin.

Purification of the S glycoprotein To express the SARS-CoV-2 S glycoprotein for purification, 293T-S cells were induced with 1 μg/mL doxycycline for 2 days. The cells were resuspended in 1× PBS and spun at 4500 × g for 15 min at 4 °C. Cell pellets were collected and lysed by incubating in lysis buffer (20 mM Tris HCl (pH 8.0), 150 mM NaCl, 1% Cymal-5, 1× protease inhibitor cocktail (Roche)) on ice for 10 min. Cell lysates were spun at 10,000 × g for 20 min at 4 °C, and the supernatant was incubated with Strep-Tactin XT Superflow resin (IBA # 2-4030-010) by rocking end over end at room temperature for 1.5 h in a 50-mL conical tube. After incubation, the supernatant-resin suspension was applied to a Biorad column allowing flowthrough by gravity, followed by

washing with 20 bed volumes of washing buffer (IBA # 2-1003-100 containing 0.5% Cymal-5), and elution with 10 bed volumes of elution buffer (IBA # 2-1042-025 containing 0.5% Cymal-5 and 1× protease inhibitor cocktail). For the second step of purification, the eluate was incubated with AAL-agarose resin (Vector Laboratories # AL-1393-2) at room temperature for 1 h in a 10-mL conical tube. The eluate-AAL resin suspension was applied to a Biorad eco-column for gravity flowthrough. The column was washed with 20 bed volumes of washing buffer (20 mM Tris HCl (pH 8.0), 150 mM NaCl, 0.5% Cymal-5, 1× protease inhibitor cocktail (Roche)), after which the sample was eluted with 10 bed volumes of elution buffer (9 parts elution buffer (Vector Laboratories # ES-3100-100), 0.5 parts 1 M Tris HCl (pH 8.0), 0.5 parts 10% Cymal-5). The eluate was buffer exchanged by ultrafiltration three times to remove fucose; this was accomplished using a 15-mL ultrafiltration tube (Thermo Fisher Scientific # UFC903024) at 4000 × g at room temperature with a buffer consisting of 20 mM Tris HCl (pH 8.0), 150 mM NaCl and 0.5% Cymal-5.

Glycopeptide analysis

Materials Trizma@ hydrochloride, Trizma@ base, ammonium bicarbonate, urea, tris(2-carboxyethyl) phosphine hydrochloride (TCEP), iodoacetamide (IAM), and glacial acetic acid were purchased from Sigma. Other reagents used in this study included optima LC/MS-grade acetonitrile, water, formic acid (Fisher Scientific), sequencing grade trypsin (Promega), chymotrypsin (Promega), and glycerol-free peptidyl-N-glycosidase F (PNGase F, New England BioLabs). All reagents and buffers were prepared with deionized water purified with a Millipore Direct-Q3 (Billerica, MA) water purification system.

Proteolytic digestion of the SARS-CoV-2 S glycoproteins The purified SARS-CoV-2 glycoprotein samples (30 μg) at a concentration of ~ 0.03 mg/mL were denatured with 7 M urea in 100 mM Tris buffer (pH 8.5), reduced at room temperature for one hour with TCEP (5 mM), and alkylated with 20 mM IAM at room temperature for another hour in the dark. The reduced and alkylated samples were buffer exchanged with 50 mM ammonium bicarbonate (pH 8) using a 50-kDa MWCO filter (Millipore) prior to trypsin digestion. The resulting buffer-exchanged sample (80 μL) was aliquoted into two portions—one digested with trypsin and the other with chymotrypsin. Trypsin digestion was performed at a 30:1 protein to enzyme ratio and was incubated overnight at 37 °C. A 15-μL aliquot of the tryptic digest was treated with chymotrypsin (20:1 protein to enzyme ratio) and was incubated at 37 °C in the dark for eight hours. A 10-μL aliquot of each trypsin and chymotrypsin digest was deglycosylated with PNGase F and was incubated overnight at 37 °C. The digests were either directly analyzed or stored at − 20 °C until further analysis.

Chromatography and mass spectrometry High-resolution LC/MS experiments were performed using an Orbitrap Fusion Lumos Tribrid (Thermo Scientific) mass spectrometer equipped with ETD that is coupled to an Acquity UPLC M-Class system (Waters). Mobile phases consisted of solvent A: 99.9% deionized H_2O + 0.1% formic acid and solvent B: 99.9% CH_3CN + 0.1% formic acid. Three microliters of the sample were injected onto a C18 PepMap™ 300 column (300 μm i.d. × 15 cm, 300 Å, Thermo Fisher Scientific) at a flow rate of 3 μL/min. The following CH_3CN/H_2O multistep gradient was used: 3% B for 3 min, followed by a linear increase to 45% B in 50 min then a linear increase to 90% B in 15 min. The column was held at 90% B for 10 min before re-equilibration. All mass spectrometric analysis was performed in the positive ion mode using data-dependent acquisition with the instrument set to run in 3-s cycles for the survey and two consecutive MS/MS scans with CID and ETxxD (EThcD or ETciD). The full MS survey scans were acquired in the Orbitrap in the mass range 400–1800 m/z at a resolution of 120,000 at m/z 200 with an AGC target of 4×10^5. Following a survey scan, MS/MS scans were performed on the most intense ions with charge states ranging from 2 to 6 and with intensity greater than 5000. CID was carried out with a collision energy of 30% while ETD was performed using the calibrated charge-dependent reaction time. Resulting fragments were detected using rapid scan rate in the ion trap.

Glycopeptide identification Glycopeptide compositional analysis was performed using (a) the methods that have been described previously [3, 20, 31, 32] and (b) the Byonic software (v 3.10.2. Protein Metrics). Briefly, glycopeptide compositions were manually determined from both MS and tandem MS data in a glycopeptide-rich region of the LC/MS data in a well-established expert-based analysis strategy that is elaborated upon in detail here. The first step in this strategy is to locate the time domains in the LC-MS data where glycopeptides of any type are eluting. These regions are identifiable by observing clusters of peaks in the high-resolution MS data whose mass difference corresponds to the masses of monosaccharide units (Hex, HexNAc, NeuAc, Fuc) and from observing the presence of characteristic oxonium ions in the tandem MS data during the same timeframe. The second step of the process includes identifying the elution range of each "glycopeptide cluster," where a given cluster has the same peptide portion but different glycans. All the glycopeptide cluster members elute within a short timespan during LC-MS on a reverse-phase column. To identify each cluster's elution range, CID data are further examined: Each cluster is identified by the frequent and abundant detection of its

common Y_1 ion, observed in numerous CID data within a given glycopeptide-rich fraction. In cases where the elution region of a glycopeptide cluster is either ambiguous or difficult to identify using this strategy, a secondary experiment could be run where the elution times of the formerly glycosylated peptides could be identified from LC/MS data of PNGase F–treated samples. When using this alternative strategy, the elution region for the glycopeptide is considered to fall within a 3-min retention time window of the maximum abundance of the eluting deglycosylated glycopeptide, acquired under the same LC-MS conditions. Once the elution region for a particular glycopeptide cluster is determined, plausible glycopeptide compositions are searched for using the extracted *m*/*z*'s in the high-resolution MS data within the chromatographic region of interest. These data are compared against a custom glycan database containing 393 biologically relevant *N*- and *O*-linked glycans using GlycoPep DB [32]. The complete list of glycans used in this project is provided in Supplemental Fig. 1. The following parameters are used for the identification: charge state values from 2 to 7, mass tolerance of 5 ppm, carbamidomethyl for cysteine modification, variable modification if any, and peptide sequence. Once the data are submitted, GlycoPep DB generates a list of plausible glycopeptide compositions that are subsequently examined: The correct monoisotopic peak, charge state, and ion intensity (> 10^3) in the high-resolution MS data are manually verified. After the list of plausible glycopeptides, based on high-resolution mass and retention time, is generated, each identified glycopeptide is further confirmed from the glycosidic bond cleavages resulting from the losses of the monosaccharide units, the Y_1 ion, and b/y fragment ions in the CID data; for ETxxD data, c/z fragment ions, oxonium ions, and the charged-reduced species must be identified among the abundant ions in the spectrum. A correct assignment entails the ability to identify at least half of the ions above 20% relative abundance with plausible product ions for the glycopeptide of interest.

For glycopeptide analysis using the Byonic software, raw files from LC/MS/MS data were search against the combination of two Byonic glycan databases consisting of 287 mammalian (no sodium) and 167 human (no multiple fucose) *N*-glycans and 78 mammalian *O*-glycans for *O*-linked glycosylation analysis using the following search parameters: SARS-CoV-2 S glycoprotein sequence; enzyme: trypsin and chymotrypsin; a maximum miscleavage of 2 per peptide, mass tolerance of 10 ppm for precursor and ± 0.8 Da (IT) /20 ppm (FT) for fragment ions; amino acid modifications that were used:

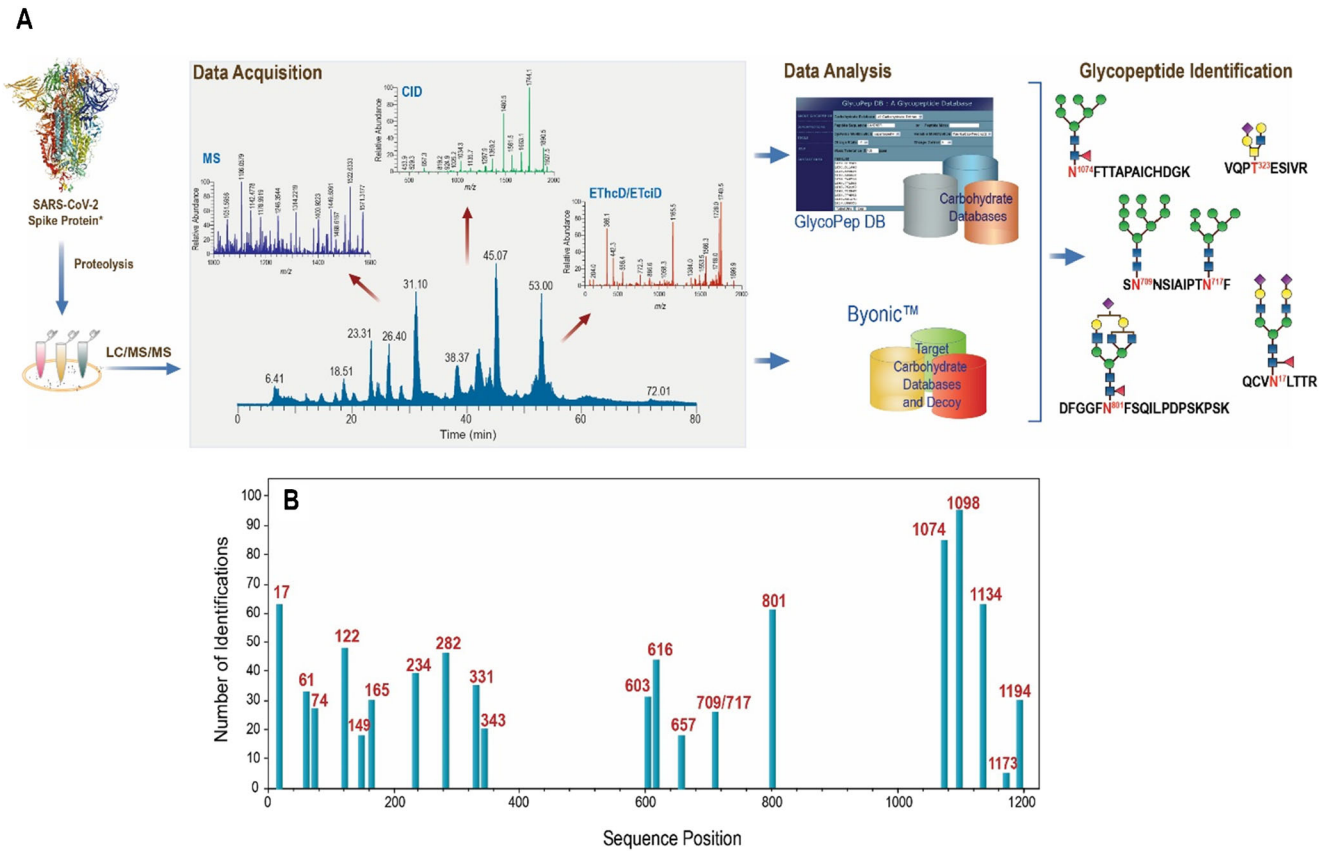


Fig. 1 Overview and summary of the glycopeptide analysis. **A** Workflow for collecting LC-MS data on the SARS-CoV-2 S protein glycopeptides and two analysis strategies compared herein; **B** graphical depiction of the number of glycoforms identified in the expert analysis strategy at each N-linked glycosylation site. (*SARS-CoV-2 S protein PDB ID: 6ZGE)

fixed: carbamidomethyl (C); variable: deamidation (N/Q) and oxidation (M), phospho (S,T,Y), Gln->pyro-Glu (Nterm Q), and palmitoyl (C). All other parameters were kept at default values. A 1% false discovery rate was used for the search. All Byonic results were manually validated with the following criteria: presence of glycopeptide characteristic oxonium and Y_1 ions in the MS/MS data, retention time matching within a 3-min window of the known retention time of the glycopeptides determined from the manual analysis using the elution profile of deamidated peptides from the deglycosylated digests, and the monoisotopic peak and charge state of the high-resolution MS ions being correctly assigned. Results were further filtered with a confidence identification threshold of Byonic score > 100.

Results and discussion

The fundamental question this study asked was: What is the opportunity cost for obtaining glycosylation profiles on viral spike proteins using software that assigns the glycosylation data automatically? In other words, what is traded, in terms of scientific knowledge, for the ability to rapidly identify the glycosylation profile of this protein or any other glycoprotein? To answer this question, a thorough expert-based analysis of a trimeric SARS-CoV-2 spike glycoprotein was performed to establish a ground truth set of the protein's glycosylation. These data were then used to assess the results from the most widely implemented glycosylation analysis software, Byonic, which has been used previously in multiple labs to analyze the glycosylation on this protein [13, 25–28]. By using a single data set to compare the ground truth assignments to those generated by Byonic, the true reliability of this tool is obtainable, and the opportunity cost can be determined.

Figure 1A shows the overall workflow for the sample preparation, data collection, and analysis strategy. This approach has been used for over 10 years in mapping the glycopeptides of numerous variants of a similar trimeric, viral spike protein, HIV-1 Env, which has a similar number of glycosylation sites [3, 4, 21, 31]. Briefly, after reduction, alkylation, and protease digestion, LC-MS data is acquired in a data-dependent fashion. Both high-resolution MS data, to confirm the mass of the glycopeptide, and MS/MS data, including both CID and ETD data, are acquired, to confirm the glycan composition and peptide sequence of the glycopeptides. In sum, 828 different glycopeptides were identified on the SARS-CoV-2 S protein. A summary of the number of glycoforms identified at each site by this analysis workflow is found in Fig. 1B, with the complete glycosylation data found in Supplemental Table 1.

After the glycopeptide analyses were complete using the expert-based strategy, the same LC-MS files used for the assignments in Fig. 1B were subjected to automated analysis by Byonic. The analysis parameters for this study are reported in Supplemental Table 2. Every attempt was made to choose the most optimal analysis parameters in advance, based on extensive experience in analyzing LC-MS data of viral glycoproteins. For example, the glycan libraries chosen contained only mammalian glycans and did not contain sodium adducts, since these adducts are rarely seen in similar LC-MS experiments when data are acquired in the positive ion mode. The mass tolerance was set to 10 ppm on the precursor ion: a larger value would include too many incorrect matches; a smaller value would be too restrictive for some peaks and therefore lead to missed assignments. A fixed modification of carbamidomethyl was set at C, since the samples were reduced and alkylated with iodoacetamide; variable modifications known to be present in these types of samples, such as phosphorylation and methionine oxidation, or known to occur during the sample preparation steps, such as modification of the N-terminal glutamine to pyroglutamic acid, were also included in the search.

After the LC-MS files were analyzed by Byonic, the underlying MS data for every glycopeptide that had been reported was manually inspected to determine if each assignment was correct. Criteria for accepting or rejecting an assignment are described in the "Experimental" section. Through this process, many Byonic-assigned glycopeptides were clearly inaccurate and rejected. This manual curation process revealed several recurring problems with the software, and these are described next in more detail. After a description of the issues uncovered, a comparison of the overall results of the two workflows is presented.

Issue #1. Incorrect glycopeptide assignments based on retention time Researchers with expertise in glycopeptide analysis

Table 1 Examples of misassigned glycopeptides from Byonic based on inaccurate retention times

Peptide	Glycan	Score	Scan time
DLPQGFSALEPLVDLPIGINITR	HexNAc(2)Hex(7)	377.97	58.9123
DLPQGFSALEPLVDLPIGINITR	HexNAc(2)Hex(8)	218.4	58.9114
DLPQGFSALEPLVDLPIGINITR	HexNAc(3)Hex(4)	103.27	59.2252
DLPQGFSALEPLVDLPIGINITR	HexNAc(3)Hex(6)	173.61	59.0913
DLPQGFSALEPLVDLPIGINITR	HexNAc(5)Hex(3)Fuc(1)NeuAc(1)	139.97	47.2818
DLPQGFSALEPLVDLPIGINITR	HexNAc(5)Hex(6)Fuc(1)NeuAc(2)	136.14	6.3723

have long known that a glycopeptide interacts with the stationary phase of a C18 column through its peptide portion. Consequently, all the glycoforms from the same peptide will have about the same retention time on a reverse-phase column, with small shifts potentially occurring due to differences in the number of sialic acids [12, 33, 34]. We capitalized on this information to identify incorrectly assigned glycopeptides in the Byonic-assigned data. Table 1 shows an example of two easily identifiable incorrect assignments. Six glycopeptides are assigned by Byonic for the peptide sequence, DLPQGFSALEPLVDLPIGINITR. Each of them has a score of > 100, indicating that these are supposed to be high-confidence identifications. Four of them elute within 0.3 min of each other, at around 59 min, but two of the assigned glycopeptides have radically different retention times. One of them elutes 12 min earlier, and the other supposedly elutes 53 min earlier. These assignments are most certainly wrong, as there is no conceivable explanation for such a large shift in retention time. A more reliable software product would make use of the retention time information to automatically rule out obviously spurious results; until that exists, users should carefully curate their data to remove these false assignments.

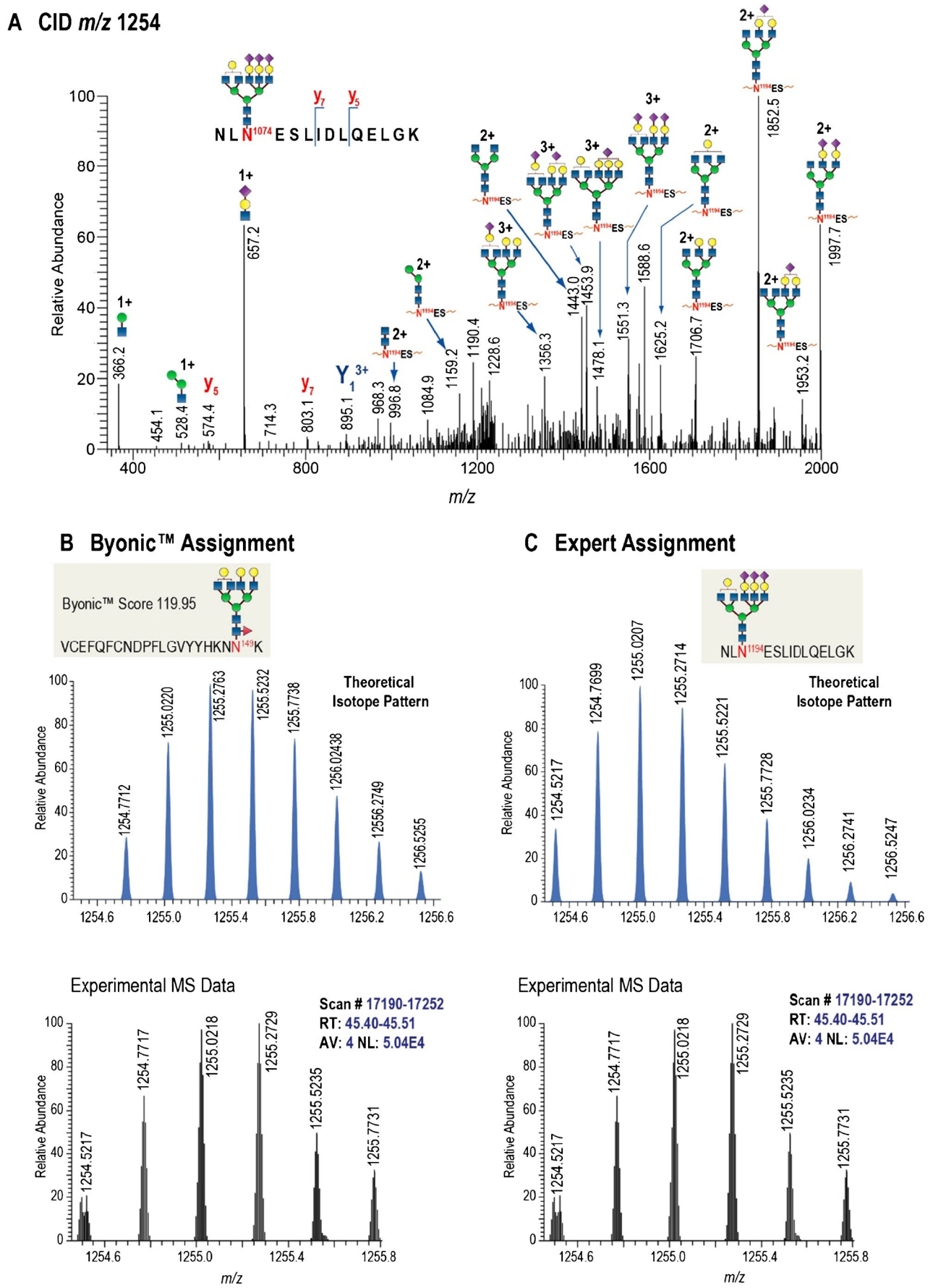
Issue #2. Incorrect glycopeptide assignments based on exact mass Mass spectrometry analysis software has a particularly hard time determining the monoisotopic mass of glycopeptides. Many times, the first ^{13}C isotope is misassigned as the monoisotopic mass. Sometimes, the opposite problem occurs, and the peak corresponding to the monoisotopic mass is misassigned as a ^{13}C isotope. If the monoisotopic mass is misassigned, the software has no chance of correctly assigning the glycopeptide because it will not consider the correct assignment as a possible match for the spectrum. This is such a significant problem, Byonic has included a clever option to reduce the number of missed assignments: It allows the software to consider assignments that are off by exactly 1 or 2 Da, and this option is set "on" by default. The advantage of leaving this default option on is that the software is more likely to find correct matches for the ions when good fragmentation spectra exist, even when the correct monoisotopic precursor ion mass is unassignable by the software. However, if the underlying MS/MS scoring algorithms are deficient, incorrectly assigned glycopeptides, with incorrect masses and incorrectly assigned MS/MS data, can result. These spurious assignments must be ruled out by careful manual inspection.

Figure 2 shows an example where Byonic's assignment for a glycopeptide peak can be determined to be inaccurate based on the monoisotopic mass of the precursor ion. In Fig. 2 A, the CID data appear, and Fig. 2 B and C contain the high-resolution data, along with theoretical isotopic distributions for the Byonic-assigned glycopeptide (2B) and the correct assignment (2C). The correct glycopeptide composition was assigned by considering both the high-resolution data and the CID spectrum. It is a tri-sialylated glycan attached to the peptide NLNESIDLQELGK. Virtually every major product ion in the CID spectrum was assignable to a logical glycosidic or peptide cleavage ion from the precursor, as shown in Fig. 2A. The assignment was further confirmed based on its high-resolution mass, which was within 2 ppm of the theoretical mass for this ion, and the retention time; this glycopeptide eluted with others of the same peptide component. Byonic produced a very different assignment for this glycopeptide; both the peptide *and* the glycan were different (see Fig. 2B). The Byonic assignment is clearly wrong. Had manual assignment of this glycopeptide not been done in advance, one could verify that the Byonic assignment was inaccurate by simply checking the isotopic distribution of the precursor ion. Byonic's assignment does not match the data, as shown in Fig. 2B. Figure 2C compares the isotopic distribution of the precursor ion with the correctly assigned glycopeptide, and one can clearly see that the correctly assigned species is a better match.

Issue #3. Algorithmic deficiencies for glycan-based cleavage ions Perhaps the most troubling aspect of the incorrect assignment in Fig. 2 is that Byonic's wrong assignment did not resemble anything remotely similar to the correct assignment, yet this wrong structure received a score of 120, clearly indicating that it is a "high-confident" assignment. We postulate that this high score is an indicator that the algorithm is not scoring the CID data effectively. The glycopeptide that Byonic assigned to these data had no sialic acids in it, while the CID spectrum is dominated with losses of sialic acid. The ion, *m/z* 657, which corresponds to a trisaccharide containing sialic acid, is clearly abundant and readily signifies the presence of this monosaccharide. Additionally, the glycan in Byonic's assigned structure contains a fucose, a residue that is typically labile and usually produces product ions resulting from loss of fucose, when this monosaccharide is present. Yet, even considering these facts, Byonic's inaccurate, nonsialylated, fucosylated glycopeptide still received a high score. Clearly, a scoring algorithm that gives such a high score to such an inaccurate composition needs some retooling.

Additional examples of deficiencies in the scoring algorithm The underlying scoring algorithm in Byonic appears to be more effective at scoring peptide-based fragmentation, and ETD data, than glycan-based fragmentation, and CID data;

Fig. 2 Demonstration of a case where a Byonic-assigned glycopeptide was determined to be misassigned due to its high-resolution mass and CID data. **A** CID spectrum for the expert-assigned glycopeptide at m/z 1254, with major product ions assigned; **B** theoretical and experimental high-resolution MS data for the Byonic-assigned glycopeptide of the CID spectrum in **A**; **C** theoretical and experimental high-resolution MS data for the expert-assigned glycopeptide


A CID m/z 1254
NLN1074ESLIDLQELGK
Relative Abundance
m/z
B Byonic™ Assignment
Byonic™ Score 119.95
VCEFQFCNDPFLGVYYHKNN149K
Theoretical Isotope Pattern
C Expert Assignment
NLN1194ESLIDLQELGK
Theoretical Isotope Pattern
Experimental MS Data
Scan # 17190-17252
RT: 45.40-45.51
AV: 4 NL: 5.04E4
Experimental MS Data
Scan # 17190-17252
RT: 45.40-45.51
AV: 4 NL: 5.04E4

consequently, spectra with glyco-centric fragmentation can be misassigned (as in Fig. 2) or unassigned, as described below. Figure 3 shows an example of how two very similar glycopeptides, with similar spectral quality, receive very different scores depending on whether or not rare, peptide-based product ions are present in the CID data. The glycan component of these two species are identical, and the peptide portion in the top example is just a miscleaved version of the glycopeptide in the bottom panel. Both CID spectra have similar signal-to-noise, and the vast majority of the ions can be manually assigned in both cases. Yet, the glycopeptide in the top panel received a very high score of 398, while the glycopeptide in the bottom panel received a score of 30, thirteen times smaller. The latter score is low enough that the assignment would have been disregarded in many investigations, since an emerging common practice is to ignore any glycopeptide assignments with a score below a particular threshold, say, 100 for example. We identified numerous cases where the expert analysis workflow confidently assigned a glycopeptide based on its retention time, its high-resolution mass, and a well-matching CID spectrum that the Byonic analysis did not find. These missing assignments appear to be directly related to the fact that the scoring algorithm does not have sufficient scoring rules for species that undergo glyco-centric fragmentation.

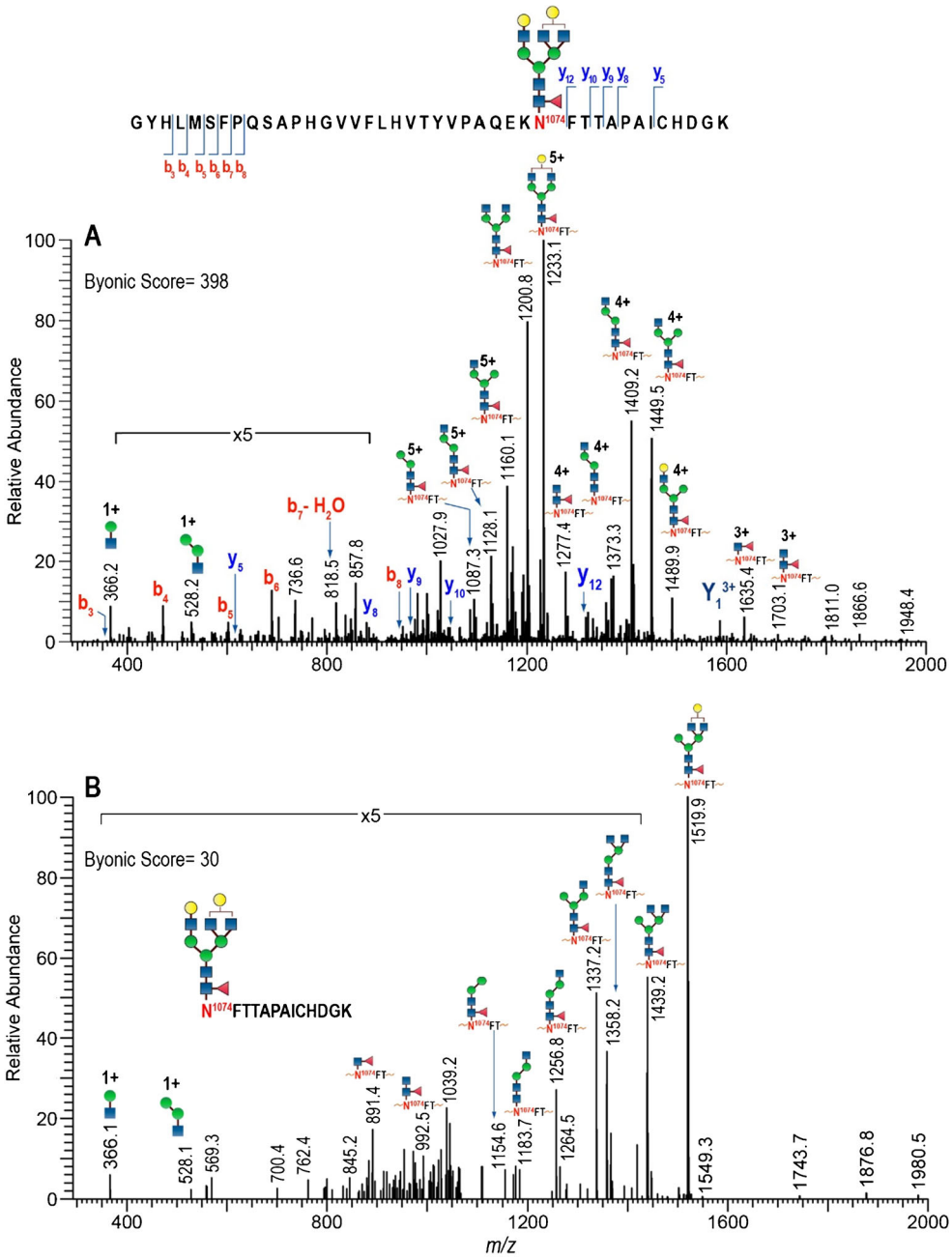


Fig. 3 CID data for two very similar glycopeptides highlighting the weaknesses in the Byonic scoring algorithm when glyco-centric fragmentation dominates. **A** Example of a high-scoring glycopeptide, where peptide-based fragmentation is abundant. **B** Example of a low-scoring glycopeptide, where glycan-based fragmentation dominates the spectrum. In both panels **A** and **B**, the same glycan is attached to the same glycosylation site; the only difference is that the top panel includes a missed tryptic cleavage site

This aspect is highly problematic, as CID data of glycopeptides are well known to *mainly* include abundant glycan-based fragmentation with very minimal peptide-based fragmentation [35, 36]. Without good scoring rules for glycan-based fragmentation, the implementation of Byonic on MS data containing only CID spectra could result in a high false positive rate with many missed true positives.

A final example showing that the underlying scoring algorithms are deficient and can lead to missed assignments and wrong assignments is shown in Fig. 4. In this case, the same glycopeptide, a high-mannose glycan attached to a long peptide with a miscleavage site, was identified in two different product ion spectra. This type of glycopeptide is rather straightforward to assign because a series of hexose losses is commonly observed in the CID data, as shown in Fig. 4A, and the long peptide backbone provides numerous sites to generate c and z ions during EThcD, as shown in Fig. 4B. Both CID and EThcD produced high-quality spectra with numerous assignable product ions. While both spectra are clearly assignable to the same glycopeptide, the software generated a much higher score from the EThcD spectrum compared to the score generated by the CID spectrum. In this particular instance, the score for the CID spectrum was still good enough to beat out the score of the decoy candidates, but this is likely because peptide-based b and y ions are still observable, albeit at low abundance. Had the spectrum been even slightly noisier, this assignment would have likely received a score under 100, and possibly been ignored by some investigators. Furthermore, these peptide-centric fragmentation ions are generally rare in CID data of glycopeptides [35, 36]. Again, this example reiterates that the underlying scoring algorithm in Byonic is less able to provide high, confident scores to the correctly assigned glycopeptide based on CID data. This issue is particularly worrisome because many laboratories rely *solely* on CID data for glycopeptide assignments, because relatively few instruments have ETD capabilities.

Overall comparison After rigorously assessing each assigned glycopeptide generated from Byonic, the assignments were either verified as true positives or false positives. Figure 5A shows the overall results of this assessment and compares it to the expert analysis workflow described above. When

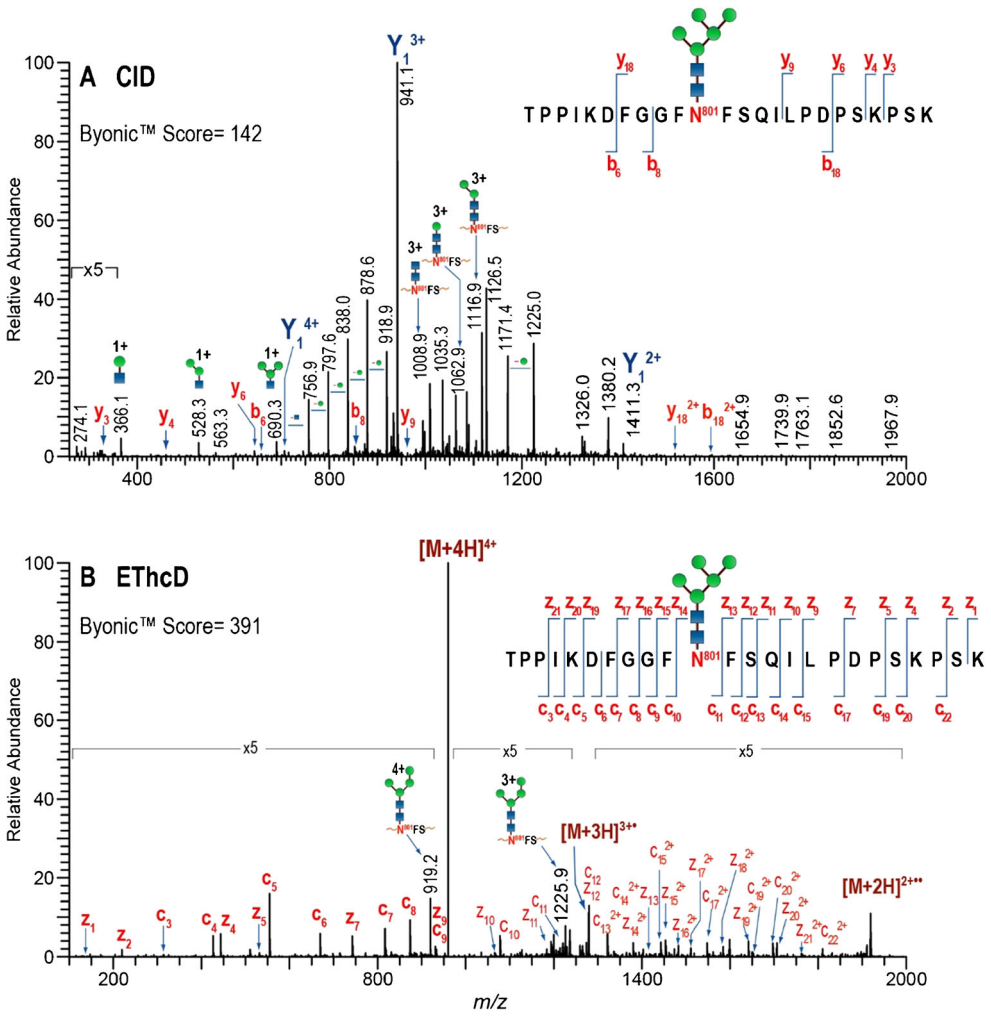


Fig. 4 MS/MS data for the same glycopeptide using two different dissociation methods. **A** CID data; **B** EThcD data; the Byonic scoring algorithm for CID is not as effective at assigning the glycopeptide with a high-confidence score

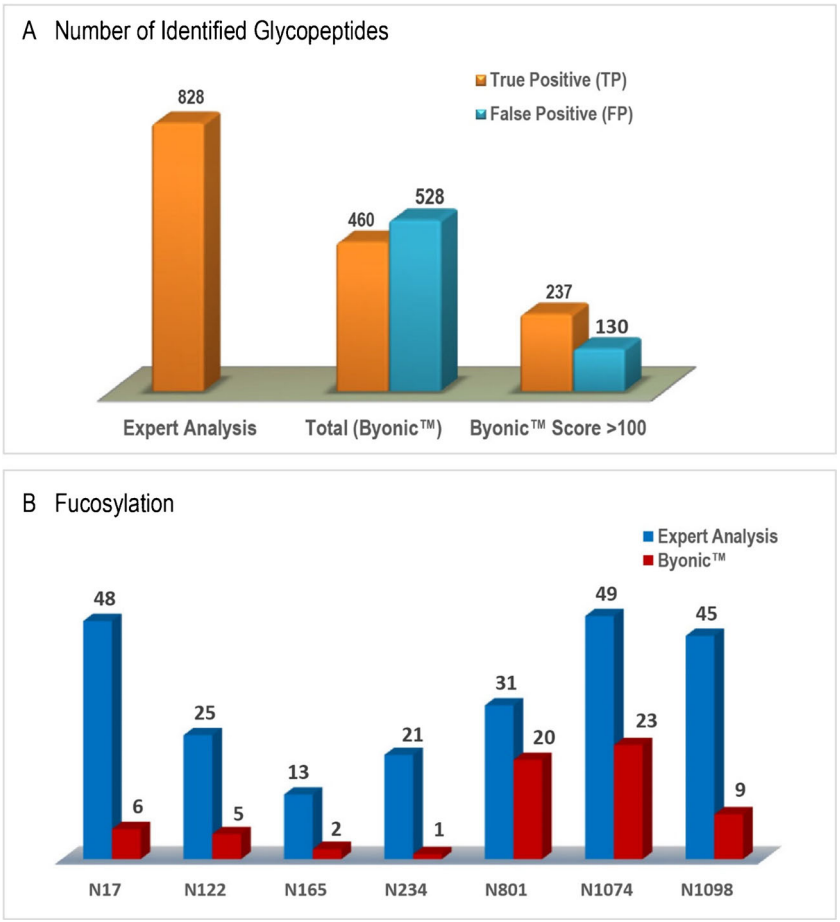


Fig. 5 Results summary of the two different analysis approaches for the SARS-CoV-2 S data set. **A** Tally of the number of correctly identified SARS-CoV-2 S glycopeptides (orange) and false positives (blue), based on the expert-based assignment criteria. **B** Assessment of fucosylation at different glycosylation sites: Depending on the approach used to assign the data, researchers would draw different conclusions about the fucosylation profile of the SARS-CoV-2 glycoprotein

considering all the SARS-CoV-2 S glycopeptides that had been identified as above a 1% FDR threshold, a set of glycopeptides that is supposedly 99% accurate, more false positives than true positives were present: 460 correct assignments were verified, while 528 were rejected as false positives. In other words, for any given assignment generated by Byonic, the glycopeptide was more likely *not* to be present in the sample than to be present!

Many users of the Byonic software are aware of the dangers of accepting the 1% FDR results at face value and additionally filter down their assigned glycopeptides to accept only those assignments that reach a certain threshold score [18, 37–39]. We therefore also compared the overall true positives and false positives for the assignments with scores > 100 (right-most comparison in Fig. 5A). Here, at least the true positives (237) outnumbered the false positives (130). Yet, even at this higher level of rigor, the output from Byonic was wrong about one third of the time. And, potentially more importantly, about 70% of the glycopeptides present in the data set were left unassigned. More than 550 assignable glycopeptides were missed by Byonic. This last point, that ~ 70% of the glycopeptides were left unassigned, underscores the point that users under-assign their data when they run the Byonic analysis and then filter out the wrong answers. Even if one were to use a careful curation strategy, the end result would be a dramatic underrepresentation of the true glycosylation profile of the protein.

To address the possibility that the poor performance displayed in Fig. 5A is either an artifact of an unnecessarily large mass error threshold or the result of validation criteria that were too strict, additional analyses were conducted. First, the results were re-assessed by considering only the glycopeptides within a 5 ppm threshold from the actual glycopeptide masses. As expected, lowering the threshold from 10 ppm to 5 ppm reduced both the number of true positives and false positives assigned, as shown in Supplemental Fig. 2. In brief, when considering all assignments that are assigned at "1% FDR", both the true positive and false positive assignments are still approximately equal. Further filtering to Byonic scores >100, and filtering the mass to 5 ppm, leaves 232 correct assignments and still 100 false positives. In short, it is unlikely that tuning the mass threshold will offer significant improvements to this tool.

We also considered the implications of requiring the true positive assignments fall within a 6-min. window for each glycopeptide cluster, ± 3 min of the deglycosylated peptide retention time. To determine if this window was too tight, the size of the window was doubled to ± 6 min, and all Byonic-assigned glycopeptides

that fell within this larger window were re-assessed using the manual validation criteria: correct monoisotopic mass and charge state, matching MS/MS data. After doubling the retention time window, the true positives increased by less than 0.5%, confirming that the original window size is an appropriate filter for this particular experiment. We note that other researchers also use retention-time filtering to improve the number of correctly assigned glycopeptides [34]. In this complementary study, most glycosylated glycoforms co-eluted within a ± 3-min window as well; however, sialylated glycoforms typically eluted ~ 6 min after the nonsialylated forms. We note that in this prior study, the glycopeptides eluted over a long gradient, spanning more than 3.5 h, while the gradient used herein was much shorter, with most of the glycoforms eluting prior to 60 min. Both these studies support the utility of retention time filtering for removing spurious assignments; however, when considering the two studies in aggregate, it seems that a single retention time filter will not necessarily be appropriate for every conceivable LC-MS experiment, and one should perhaps consider the gradient length prior to designing the filter.

Figure 5B shows the opportunity cost of using the Byonic software and carefully curating out the inaccurate results vs using a more laborious expert-based analysis. This figure plots the number of fucosylated glycoforms at a selection of the glycosylation sites in the SARS-CoV-2 S glycoprotein. In some cases, such as the N1074 and N801 sites, the Byonic workflow did "relatively well" and identified at least half of the fucosylated glycopeptides present at those sites. In other cases, the number of fucosylated glycoforms was more severely under-represented: Only one of the 21 forms present at N234 was detected; only two of the 13 forms at N165 were detected; only an eighth of the forms at N17 were detected. Because the fucosylated forms were severely under-reported, and the unassigned forms were unevenly distributed, these results could lead biologists to draw inaccurate conclusions about the protein's glycosylation. For example, if one were to see only the Byonic-derived results, s/he may erroneously conclude that N234 is occupied mostly by nonfucosylated forms when in reality, the fucosylated forms dominate this site. Since this site has been suggested from modeling studies to potentially modulate binding to the ACE2 receptor [40], obtaining an accurate glycosylation profile may be important for understanding key interactions that influence SARS-CoV-2 entry into human cells.

Conclusion

We set out to test Byonic, the industry-leading glycopeptide analysis software, on a complex glycopeptide analysis problem from an important viral protein, the SARS-CoV-2 S glycoprotein. The data set contained just a single, recombinantly expressed S glycoprotein precursor with 22 *N*-linked glycosylation sites and the S1 and S2 glycoproteins naturally derived from it by proteolytic cleavage. The goal of the experiments was to determine the opportunity cost for analyzing viral envelope glycoproteins using a fully automated workflow, including the industry-leading commercial analysis software. The outcome is chilling. In short, the software produced more spurious assignments than correct ones, when it claimed a false discovery rate of 1%. Even when choosing a more rigorous standard of only accepting assigned glycopeptides with scores > 100, the software produced wrong assignments a third of the time and missed 70% of the assignable glycopeptides. Furthermore, the missed assignments could lead to errors in understanding of the biological role of glycans on this important protein. Overall, these results provide strong cautionary evidence that the field of automated glycopeptide analysis is still a field in flux, and better tools are needed to support warp speed science. Until such tools are widely available, researchers with significant expertise in glycopeptide analysis should carefully curate any results, provide their curation criteria, and acknowledge that even under the best of circumstances, their software-based analysis will be missing a majority of the assignable glycopeptides in their data set. When the goal is to understand the biological consequences of the glycosylation of the protein, nothing yet replaces a careful study done by researchers with knowledge and experience in glycopeptide analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-021-03621-z.

Acknowledgements This work was supported by grant R35GM103054 to HD and R01 AI125093 to HD, JCK, and JS, and by a gift to JS from the late William F. McCarty-Cooper.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Hansen JES, Clausen H, Nielsen C, Teglbjaerg LS, Hansen LL, Nielsen CM, et al. Inhibition of human-immunodefficiency-virus (HIV) infection invitro by anticarbohydrate monoclonal-antibodies- peripheral glycosylation of HIV envelope glycoprotein-GP120 may be a target for virus neutralization. J Virol. 1990;64(6):2833–40.
2. Leonard CK, Spellman MW, Riddle L, Harris RJ, Thomas JN, Gregory TJ. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type-1 recombinant human immunodeficiency-virus envelope glycoprotein (GP120) expressed in Chinese hamster ovary cells. J Biol Chem. 1990;265(18):10373–82.
3. Go EP, Ding HT, Zhang SJ, Ringe RP, Nicely N, Hua D, et al. Glycosylation benchmark profile for HIV-1 envelope glycoprotein production based on eleven env trimers. J Virol. 2017;91(9):e02428–16.
4. Go EP, Herschhorn A, Gu C, Castillo-Menendez L, Zhang SJ, Mao YD, et al. Comparative analysis of the glycosylation profiles of membrane-anchored HIV-1 envelope glycoprotein trimers and soluble gp140. J Virol. 2015;89(16):8245–57.

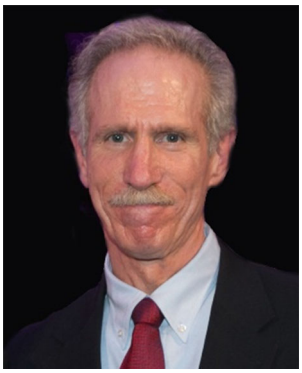
5. Pritchard LK, Vasiljevic S, Ozorowski G, Seabright GE, Cupo A, Ringe R, et al. Structural constraints determine the glycosylation of HIV-1 envelope trimers. Cell Rep. 2015;11(10):1604–13.
6. Cao L, Pauthner M, Andrabi R, Rantalainen K, Berndsen Z, Diedrich JK, et al. Differential processing of HIV envelope glycans on the virus and soluble recombinant trimer. Nat Commun. 2018;9(1):3693.
7. de la Pena A T, Rantalainen K, Cottrell CA, Allen JD, van Gils MJ, Torres JL, et al. Similarities and differences between native HIV-1 envelope glycoprotein trimers and stabilized soluble trimer mimetics. PLoS Pathog. 2019;15(7):e1007920.
8. Saunders KO, Wiehe K, Tian M, Acharya P, Bradley T, Alam SM, et al. Targeted selection of HIV-specific antibody mutations by engineering B cell maturation. Science. 2019;366(6470):eaay7199.
9. Seabright GE, Cottrell CA, van Gils MJ, D'Addabbo A, Harvey DJ, Behrens AJ, et al. Networks of HIV-1 envelope glycans maintain antibody epitopes in the face of glycan additions and deletions. Structure. 2020;28(8):897−909.e6.
10. Dey AK, Cupo A, Ozorowski G, Sharma VK, Behrens AJ, Go EP, et al. cGMP production and analysis of BG505 SOSIP.664, an extensively glycosylated, trimeric HIV-1 envelope glycoprotein vaccine candidate. Biotechnol Bioeng. 2018;115(4):885–99.
11. Zou Z, Wang R, Go EP, Desaire H, Sun PD. High level stable expression of recombinant HIV gp120 in glutamine synthetase gene deficient HEK293T cells. Protein Expr Purif. 2021;181:105837.
12. Patabandige MW, Pfeifer LD, Nguyen HT, Desaire H. Quantitative clinical glycomics strategies: a guide for selecting the best analysis approach. Mass Spectrom Rev. 2021. https://doi.org/10.1002/mas.21688.
13. Pujić I, Perreault H. Recent advancements in glycoproteomic studies: Glycopeptide enrichment and derivatization, characterization of glycosylation in SARS CoV2, and interacting glycoproteins. Mass Spectrom Rev 2021. https://doi.org/10.1002/mas.21679.
14. Ruhaak LR, Xu GG, Li QY, Goonatilleke E, Lebrilla CB. Mass spectrometry approaches to glycomic and glycoproteomic analyses. Chem Rev. 2018;118(17):7886–930.
15. Dubrovskaya V, Tran K, Ozorowski G, Guenaga J, Wilson R, Bale S, et al. Vaccination with glycan-modified HIV NFL envelope trimer-liposomes elicits broadly neutralizing antibodies to multiple sites of vulnerability. Immunity. 2019;51(5):915−929.e7.
16. Bangaru S, Ozorowski G, Turner HL, Antanasijevic A, Huang D, Wang X, et al. Structural analysis of full-length SARS-CoV-2 spike protein from an advanced vaccine candidate. Science. 2020;370(6520):1089–1094.
17. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell. 2020;183(6):1735.
18. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science. 2020;369(6501):330–3.
19. Zhao P, Praissman JL, Grant OC, Cai Y, Xiao T, Rosenbalm KE, et al. Virus-receptor interactions of glycosylated SARS-CoV-2 spike and human ACE2 receptor. Cell Host Microbe. 2020;28(4):586–601 **e6**.
20. Go EP, Chang Q, Liao HX, Sutherland LL, Alam SM, Haynes BF, et al. Glycosylation site-specific analysis of clade C HIV-1 envelope proteins. J Proteome Res. 2009;8(9):4231–42.
21. Go EP, Hewawasam G, Liao HX, Chen HY, Ping LH, Anderson JA, et al. Characterization of glycosylation profiles of HIV-1 transmitted/founder envelopes by mass spectrometry. J Virol. 2011;85(16):8270–84.
22. Go EP, Moon HJ, Mure M, Desaire H. Recombinant human Lysyl oxidase-like 2 secreted from human embryonic kidney cells displays complex and acidic glycans at all three N-linked glycosylation sites. J Proteome Res. 2018;17(5):1826–32.
23. Weber JJ, Kashipathy MM, Battaile KP, Go E, Desaire H, Kanost MR, et al. Structural insight into the novel iron-coordination and domain interactions of transferrin-1 from a model insect, *Manduca sexta*. Protein Science. 2021;30(2):408–22.
24. Antonopoulos A, Broome S, Sharov V, Ziegenfuss C, Easton RL, Panico M, et al. Site-specific characterisation of SARS-CoV-2 spike glycoprotein receptor binding domain. Glycobiology. 2020;31(3):181–187.
25. Shajahan A, Supekar NT, Gleinich AS, Azadi P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology. 2020;30(12):981–8.
26. Sanda M, Morrison L, Goldman R. N- and O-glycosylation of the SARS-CoV-2 spike protein. Anal Chem. 2021;93(4):2003–9.
27. Wang D, Baudys J, Bundy JL, Solano M, Keppel T, Barr JR. Comprehensive analysis of the glycan complement of SARS-CoV-2 spike proteins using signature ions-triggered electron-transfer/higher-energy collisional dissociation (EThcD) mass spectrometry. Anal Chem. 2020;92(21):14730–9.
28. Watanabe Y, Berndsen ZT, Raghwani J, Seabright GE, Allen JD, Pybus OG, et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nat Commun. 2020;11(1):2688.
29. Zhang S, Go EP, Ding H, Anang S, Kappes JC, Desaire H, Sodroski J. Analysis of glycosylation and disulfide bonding of wild-type SARS-CoV-2 spike glycoprotein. bioRxiv. 2021;2021.04.01.438120. https://doi.org/10.1101/2021.04.01.438120.
30. Nguyen HT, Zhang S, Wang Q, Anang S, Wang J, Ding H, et al. Spike glycoprotein and host cell determinants of SARS-CoV-2 entry and cytopathic effects. J Virol. 2020;95(5):e02304–20.
31. Go EP, Irungu J, Zhang Y, Dalpathado DS, Liao HX, Sutherland LL, et al. Glycosylation site-specific analysis of HIV envelope proteins (JR-FL and CON-S) reveals major differences in glycosylation site occupancy, glycoform profiles, and antigenic epitopes' accessibility. J Proteome Res. 2008;7(4):1660–74.
32. Go EP, Rebecchi KR, Dalpathado DS, Bandu ML, Zhang Y, Desaire H. GlycoPep DB: A tool for glycopeptide analysis using a "smart search". Anal Chem. 2007;79:1708–13.
33. Wang BL, Tsybovsky Y, Palczewski K, Chance MR. Reliable determination of site-specific in vivo protein N glycosylation based on collision-induced MS/MS and chromatographic retention time. J Am Soc Mass Spectrom. 2014;25(5):729–41.
34. Klein J, Zaia J. Relative retention time estimation improves N-Glycopeptide identifications by LC−MS/MS. J Proteome Res. 2020;19:2113–21.
35. Mechref Y. Use of CID/ETD mass spectrometry to analyze glycopeptides. Curr Protoc Protein Sci. 2012. https://doi.org/10.1002/0471140864.ps1211s68.
36. Zhu Z, Desaire H. Carbohydrates on proteins: site-specific glycosylation analysis by mass spectrometry. Ann Rev Anal Chem. 2015;6:463–83.
37. Choo MS, Wan C, Rudd PM, Nguyen-Khuong T. GlycopeptideGraphMS: improved glycopeptide detection and identification by exploiting graph theoretical patterns in mass and retention time. Anal Chem. 2019;91(11):7236–44.
38. Lee LY, Moh ES, Parker BL, Bern M, Packer NH, Thaysen-Andersen M. Toward automated N-glycopeptide identification in glycoproteomics. J Proteome Res. 2016;15(10):3904–15.
39. Zhu J, Chen Z, Zhang J, An M, Wu J, Yu Q, et al. Differential quantitative determination of site-specific intact N-Glycopeptides in serum haptoglobin between hepatocellular carcinoma and cirrhosis using LC-EThcD-MS/MS. J Proteome Res. 2019;18(1):359–71.
40. Casalino L, Gaieb Z, Goldsmith JA, Hjorth CK, Dommer AC, Harbison AM, et al. Beyond shielding: the roles of glycans in SARS-CoV-2 Spike Protein. bioRxiv. 2020. https://doi.org/10.1101/2020.06.11.146522.

Eden Go is Research Laboratory Director in the group of Prof. Heather Desaire at the University of Kansas. Her research interests include the application of high-resolution mass spectrometry in the analysis of glycosylation, disulfide bonds in proteins, and small molecules. She received her Ph.D. in Chemical Physics at the University of Maryland, College Park, and worked with Prof. Gary Siuzdak for her post-doctoral training in mass spectrometry at the Scripps Research Institute, in La Jolla, California.

Shijian Zhang is an instructor at the Dana-Farber Cancer Institute and Microbiology and Immunobiology Department, Harvard Medical School. He has been focusing on viral membrane protein studies including HIV-1 Envelop protein and SARS-CoV-2 Spike protein to understand viral membrane protein structures and their antigenicity as well as immunogenicity.

Haitao Ding is Senior Scientist in the Division of Hematology/Oncology, Department of Medicine, University of Alabama at Birmingham. His interests include HIV-1 pathogenesis and vaccine discovery. His research is focused on the development, validation, and application of cell-based HIV assay systems for elucidating correlates of HIV-1 immune protection. He earned an M.S. degree in Agriculture from Nanjing Agriculture University, Jiangsu, China, and a Ph.D. degree in Botany from Peking University, Beijing, China.

John Kappes is Associate Professor in the Department of Medicine at the University of Alabama at Birmingham. His interest is focused on HIV-1 vaccine discovery, including research on the immunogenicity of the trimeric HIV-1 envelope glycoprotein and correlates of immune protection. He received his Ph.D. from the Saint Thomas Institute, Cincinnati, OH.



Joseph G. Sodroski is Professor of Microbiology at the Dana-Farber Cancer Institute and Harvard Medical School. His research has been devoted to understanding the mechanisms by which the human immunodeficiency virus (HIV-1) and, more recently, SARS-CoV-2 enter and kill their host cells. Antibodies and small-molecule inhibitors of these processes are being investigated as potential treatments and preventives.

Heather Desaire is the Dean's Professor of Chemistry at the University of Kansas. Her research interests span the fields of glycobiology, mass spectrometry, and machine learning. She received her Ph.D. from the University of California, Berkeley, and a BA in Chemistry from Grinnell College, Grinnell, IA.