Journal of proteome research

pubs.acs.org/jpr

Article

Mass Spectral Library Methods for Analysis of Site-Specific N-Glycosylation: Application to Human Milk Proteins

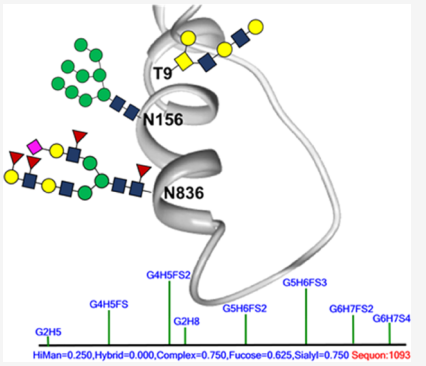
Concepcion A. Remoroza,* Meghan C. Burke, Xiaoyu Yang, Sergey Sheetlin, Yuri Mirokhin, Sanford P. Markey, Dmitrii V. Tchekhovskoi, and Stephen E. Stein

Cite This: https://doi.org/10.1021/acs.jproteome.2c00286

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present a mass spectral library-based method for analyzing site-specific N-linked protein glycosylation. Its operation and utility are illustrated by applying it to both newly measured and available proteomics data of human milk glycoproteins. It generates two varieties of mass spectral libraries. One contains glycopeptide abundance distribution spectra (GADS). The other contains tandem mass spectra of the underlying glycopeptides. Both originate from identified glycopeptides in proteolytic digests of human milk and purified glycoproteins, which include tenascin, lactoferrin, and several antibodies. Analysis was also applied to digests of a NIST human milk standard reference material (SRM), leading to a GADS library of N-glycopeptides, enabling the direct comparison of glycopeptide distributions for individual proteins. Tandem spectra underlying each glycopeptide GADS peak are combined to create a second type of library that contains spectra of the underlying glycopeptide spectra. These were acquired by higher-energy (stepped) collision dissociation fragmentation followed by ion-trap fragmentation. Spectra are annotated using MS_Piano, recently reported annotation software. This data, with extensions of a widely used spectral library search and display software, provides accessible mass spectral libraries.



KEYWORDS: *human milk, glycan distribution, glycopeptides, mass spectrometry, site-specific glycosylation*

INTRODUCTION

Identities and abundances of individual glycans attached to specific sites within a protein can show considerable, often unpredictable variability. Not only can these distributions be highly complex, but they can depend on synthesis conditions and their precise location at the exterior of the protein. For proteins with multiple possible glycosylation sites, this determination requires glycan identification while still connected to their original sites on the protein. This is generally done by identifying individual glycopeptides released through proteolysis using the well-developed methods of bottom-up proteomics.[1,2] However, since glycan fragmentation is generally far faster than cleavage of bonds in peptides and many different glycan compositions are possible, unique analysis methods are required for reliable glycopeptide identification, generally requiring human judgment. For example, retention values are not used in current software used for identification, even though it is well known that glycopeptides with a given sequence and number of sialyl groups elute at very similar times. This work shows that using this and other aspects of glycopeptide analysis along with results of available glycopeptide search software, one may significantly reduce subjectivity and enhance reporting of results, leading to more efficient and reliable analysis as is now available in general proteomics analysis. In this process, libraries are generated for viewing and comparing glycopeptide abundance distributions as described in a previous paper[3] as well as libraries of annotated tandem spectra of the underlying glycopeptides.

The present work focuses on N-linked glycans, located at "sequons" on a protein, which are asparagine residues where threonine or serine, or cysteine[4] residues are located two residues distant in the C-terminal direction. Glycopeptides containing these sequons are first generated proteolytically, then separated by liquid chromatography, and finally identified from their mass spectral fragmentation patterns. Since it is presently not possible to fully characterize glycopeptides containing more than a single N-glycosylated site, various proteases other than trypsin may be required to isolate each sequon on a single peptide.

Identification requires the determination of both the peptide sequence and glycan. The first relies on a predefined set of possible sequences and the latter on a set of potential glycans. Examination of specific glycan fragment ions can aid in determining glycan connectivity.[5] Since peptide sequence information is derived from relatively high energy fragmentation

Received: May 12, 2022

and glycan fragmentation occurs at lower energies, reliable identification requires fragmentation over a range of collision energies.

The objective of this paper is to extend and illustrate methods described in a previous paper for representing distributions of glycans on a sequon in the form of mass spectral-like glycopeptide abundance distribution spectra (GADS). These serve to aid in identifying, displaying, and comparing site-specific glycopeptide results. In the present work, we create GADS libraries from in-house spectral measurements of a variety of different milk-related proteins as well as milk itself. We also describe the creation of a searchable mass spectral library created from the tandem spectra that serve to make the glycopeptide identification that underlies each GADS peak. GADS libraries provide distributions that depend on protein expression and analysis details, although they do not directly depend on fragmentation details. Spectra in the tandem library do depend on fragmentation conditions, which are described below. These spectra serve to confirm the identification of each peak in a GADS as well as to enable the reidentification of glycopeptides using similar fragmentation conditions. Peaks in this library are annotated using the recently reported MS_Piano[6] program, which allows the labeling of assigned peaks as well as the determination of various details of glycan connectivity.

In this work, fragmentation uses the higher-energy collision dissociation (HCD) "stepped" energy options of 15, 25, and 35% normalized collision energy (NCE). When a characteristic oxonium ion is generated, that fragmentation is followed by relatively low-energy ion-trap fragmentation, which preserves initially formed glycopeptide product ions. The specific milk proteins examined are native lactoferrin (hLf) and tenascin (TNC),[7,8] immunoglobulins A1 (IgA1), A2 (IgA2), J (IgJ), and polymeric immunoglobulin receptor (pIgR).[9,10] These proteins contain between 1 and 23 N-glycosylation sites, many of which have significant N-glycan heterogeneity. They were selected not only for their inherent interest as predominant milk proteins but also to represent a wide range of glycoproteins to develop and test current methods. Applications of these methods were applied to a NIST SRM 1953 Organic Contaminants in Non-Fortified Human Milk, as well as to results reported in the literature for whole milk samples.

MATERIALS AND METHODS

Materials

Proteolytically digested glycoproteins included human tenascin (Sigma CC065, UniProt Accession P24821), human lactoferrin (Abcam 90354, UniProt Accession P02788), human immunoglobulins A1, A2 (Abcam ab91021, UniProt Accessions P01876, P01877), polymeric immunoglobulin receptor (Abcam ab174011, UniProt Accession P01833), and thyroglobulin (Sigma 609312, UniProt Accession P01266). Also, digested glycoproteins of NIST SRM 1953 Organic Contaminants in Non-Fortified Human Milk[11] were analyzed. Proteases used include sequencing grade (Promega) trypsin, GluC, LysC, AspN, chymotrypsin, and α-lytic (Sigma).

Digestion and Hydrophilic Interaction Liquid Chromatography (HILIC) Enrichment of Glycoproteins in Human Milk

An aliquot of 200 μL of SRM 1953 was centrifuged (15,000g, 1 h) at 4 °C as in a previously reported method.[12] The upper-fat layer was discarded. The remaining milk was further ultra-centrifuged at 15,000g for 1 h at 4 °C to remove the insoluble pellet. The proteins from supernatant milk were subsequently concentrated using a 10 kDa filter (Amicon, Millipore). The protein concentration was determined using the colorimetric measurement (Pierce BCA Protein Assay Kit) at an absorbance of 580 nm in a microplate reader (Infinite 200 Pro, Tecan). Milk proteins (1 mg) were prepared for glycopeptide characterization by first resuspending the proteins in 8 mol/L urea in 100 mmol/L Tris−HCl (pH 8.5) and 10 mmol/L diothiothreitol (DTT). The proteins were reduced for 1 h at room temperature with agitation. Alkylation of cysteines was performed in the presence of 55 mmol/L iodoacetamide (IAM) for 1 h at room temperature in the dark. Excess IAM was quenched by bringing the final concentration of DTT to 10 mmol/L. Next, the urea concentration was reduced to 1 mol/L with the addition of 100 mmol/L Tris−HCl (pH 8.5). Various proteases, trypsin, chymotrypsin, GluC, LysC, or α-lytic, were used for in-solution digestion. The proteases were added at a ratio of 1:20 (w/w) enzyme to a substrate and incubated at 37 °C for 18 h. A separate chymotrypsin digestion condition (25 °C, 18 h), as suggested by the manufacturer, was performed for milk proteins. Digestion was terminated by heating the mixture for 10 min at 90 °C, then cooling at room temperature, and drying in a rotary evaporator.

Glycopeptide enrichment was performed using 1 mL HILIC-based solid phase extraction cartridges (iSPE-HILIC SPE Cartridge, 100 mg, 50 μm particle, Hilicon). The HILIC columns were first washed with 1 mL of (v/v %) formic acid (FA) and equilibrated with 1 mL of loading buffer (80 v/v % acetonitrile (ACN)/0.5 v/v % trifluoroacetic acid (TFA)). The peptide digests were reconstituted with 1 mL of loading buffer and loaded onto the column. The cartridges were washed with 1 mL of loading buffer, and the glycopeptides were step-eluted with 250 μL of 75 v/v % ACN/0.5 v/v % TFA, 250 μL of 70 v/v % ACN/0.5 v/v % TFA, 250 μL of 65 v/v % ACN/ 0.5 v/v % TFA, and 250 μL of 0.5 v/v % FA as previously described.[12] All eluents were collected, dried down, and then resuspended in 0.1% FA in water prior to liquid chromatography−mass spectrometry (LC-MS) analysis.

Digestion of Individual Proteins

Individual glycoproteins purified from human milk were digested using multiple proteases. Detailed sample preparation of these glycoproteins is documented in the Supporting Section S1. Multiple proteases are generally needed for reliable identification of glycoforms of all sequons since abundant peptides that can be separated by LC are required. Differing and sometimes unpredictable protease selectivity can make this challenging. We find that α-lytic, for example, while very useful, has an unpredictable selectivity. Agreement of GADS determination from different proteases is desirable due to the many sources of false identifications in glycopeptide analysis.

LC-MS/MS Analysis

Nanoflow liquid chromatography was performed on an UltiMate 3000 (Thermo Fisher Scientific) in line with an Orbitrap Fusion Lumos (Thermo Fisher Scientific) mass spectrometer. Chromatographic separation was performed on an Acclaim PepMap RSLC 75 μm × 15 cm column (Thermo Fisher Scientific). The composition of the two mobile phases was 0.1 v/v % formic acid in water (A) and 0.1 v/v % FA in 99.9 v/v % ACN (B). The flow rate was 300 nL/min, with Solvent B increasing from 0 to 35% over 85 min, followed by an increase from 35 to 98% over 5 min and held at 98% for 15 min before decreasing to 2% over 5 min. The gradient was then decreased to

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

0% Solvent B over 10 min and held for an additional 30 min. Full scan mass spectra acquisition was performed with a scan range of m/z 380 to m/z 2000 and detection in the Orbitrap at a resolution of 120,000. The maximum ion injection time was 50 ms with automatic gain control (AGC) target of 400,000 and an RF lens value of 40%. Filters included charge states +2 to +8, a minimum intensity threshold of 5.0×10^4, a maximum intensity threshold of 1.0×10^{20}, a dynamic exclusion of 15 s, and a total cycle time of 5 s. Precursor priority for the data-dependent acquisition was filtered according to the highest charge state and lowest m/z. Quadrupole isolation mode was used with an isolation window of 2 ms for tandem mass spectra acquisition. After a number of trials using thyroglobulin digests, optimal stepped energy settings were found to be 15, 25, and 35% normalized collision energy (NCE, Thermo Fisher Scientific) (Supporting Section 3A). Beam-type collision-cell spectra (HCD), followed by GlcNAc oxonium ion (m/z 204.087)-triggered ion-trap fragmentation, were acquired with a starting m/z of 120. Also, we used a maximum injection time of 60 ms and an AGC target of 50,000. Fragment ions and intensities were determined in the Orbitrap at a resolution of 30,000.

Data Processing

Glycopeptide identifications and abundances were derived from Byonic and Byologic software (version 3.10 Protein Metrics, Inc.) for both HCD/CID and EThcD spectra. Parameters were set to fully specific with up to three missed cleavages, fixed cysteine carbamidomethylation, and allowed possible methionine oxidation and N-term glutamine deamination with a library of 444 human N-glycans (Supporting Section 3B). Precursor and product ion m/z tolerance settings were 5 μg/g (5 ppm) and 20 μg/g (20 ppm), respectively. We note that with appropriate adjustments in thresholds, all of the methods described here are applicable to any glycopeptide identification methods that generate XIC abundances and retention times. It has recently been fully applied to MSfragger-glyco.[13] We also note that any localization of the glycan is done by the search software, and in general, HCD cannot reliably localize glycans on peptides containing multiple potentially occupied residues.

GADS Library

Since the construction of this library is described in a previous paper, only an overview is provided here. Individual GADS were derived from individual peptide ions and combined-charge state sequences identified in a single digest.[3] The mass of each peak is that of the glycan. The abundance is that of the identified glycopeptide ion (derived from its extracted ion chromatogram, XIC) and then normalized to the largest peak in the GADS. Each peak in a spectrum contains various annotation details, including glycan composition, using a compact representation described later. This format enables direct comparison of two GADS in a manner similar to the comparison of conventional mass spectra since the same glycan in two GADS have identical masses.

Retention Time Tolerances

As described previously,[3] retention times for a given glycopeptide depend primarily on the number of sialyl groups and provide necessary confirmation of the number of sialyl groups in a glycopeptide. GADS peaks for glycopeptides whose relative retention times are out of an expected range are labeled as uncertain and shown in red.[14] Expected retention values and deviations are derived for each peptide sequence. This is done by computing retention time distributions for all glycopeptides with a given number of sialyl groups (up to four residues). The median is taken as the expected value and the interquartile distance (time difference between the first and third quartile of distribution sorted by retention time). Current software labels glycans whose peaks are out of range of the expected retention by more than twice the interquartile distance as out-of-range, and their deviation in minutes is shown along with the glycan identity. Peaks whose scores are not greater than a set threshold (a value of 100 is used for the present Byonic identifications) but within retention tolerances are also labeled as uncertain and displayed in blue. Peaks that fail both tests are discarded. Peaks that satisfy score and retention criteria, the vast majority, are shown as green. Glycopeptides containing multiple sequons were excluded. Each glycopeptide peak in GADS is accompanied by a number of relevant values, as previously described.[3] For each GADS, a set of three retention characteristics are reported for each number of sialyl groups. This includes the median time, number of glycans, and the measured interquartile time. If inconsistencies such as lower retention time for a higher number of sialyl groups or excessive interquartile distances are found, an error is reported after the GADS name. Note that the details of these methods are expected to be refined in the future as more data is analyzed.

Search and User Interface Software

NISTMS.exe software,[15] widely distributed for use with mass spectral reference collections from NIST and other sources, has been adapted to import, view, search, and compare GADS as well as glycopeptide spectra. The version provided with this paper adds new display and search capabilities for the glycopeptides, and the libraries underlie the present work. It also displays glycopeptide fragment ion annotation from MS_Piano.[6] All glycopeptide spectra used for validation in this paper are provided in these libraries. Software and libraries listed (Tables 1 and 2) may be freely downloaded.[16] Note that glycan "cartoons" shown in some figures were manually added to the program output. This program can create libraries from

Table 1. HCD MS^2 Libraries of Annotated Glycopeptides for Human Milk[a]

library	protein	number of MS^2 spectra	unique glycopeptides
hLf_nist	lactoferrin	6295	448
iga1_nist	immunoglobulin A1	22,551	594
iga2_nist	immunoglobulin A2	6308	421
igj_nist	immunoglobulin J, NIST SRM1953	596	91
pigr_nist	polymeric immunoglobulin receptor	29,498	2305
tenascin_nist	tenascin	44,837	2695
milk-srm1953_nist	Human Milk Standard Reference Material 1953	3653	962
milk_wuhan[b]	Human Milk, Wuhan Virology Laboratory	257	155
milk-donor1_utrecht	Human Milk, Utrecht University	2671	200
milk-donor2_utrecht	Human Milk, Utrecht University	4678	204

[a]#Public data repositories: ProteomeXchange Consortium (PXD), mass spectrometry interactive virtual environment (MSV), integrated proteome resources (IPX). Tryp, trypsin; Chymo, chymotrypsin; and alP, α-lytic. [b]TMT-labeled (IPX0002443001);[27] Utrecht University (PXD01417).[26]

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 2. GADS Libraries of Purified Glycoproteins and Milk Samples Used in the Study

GADS library	no. of protein (s)	nonredundant glycosylated peptides	number of N-glycans	fragmentation technique
hLf-nist	1	56	69	HCD, IT-FT
iga1-nist	1	51	107	HCD, IT-FT
igA2-nist	1	50	113	HCD, IT-FT
Igj-nist	1	22	41	HCD, IT-FT
pigr-nist	1	109	212	HCD, IT-FT
tenascin-nist	1	344	164	HCD, IT-FT
[a]milk-oregon	54	93	91	HCD, EThcD
milk-nist_srm1953	111	288	71	HCD, IT-FT
[b]milk-donor1-utrecht	30	48	61	HCD
[b]milk-donor2-utrecht	22	44	56	HCD
[b]milk-utrecht	150	346	381	EThcD

[a]Oregon State University.[29] [b]University of Utrecht, PXD01417,[30] MSV000083710;[30] Public Data Repositories: ProteomeXchange Consortium (PXD), Mass Spectrometry Interactive Virtual Environment (MSV), National Institute of Standards and Technology (NIST), Standard Reference Material (SRM) 1953. Tryp, trypsin; Chymo, chymotrypsin; alP, α-lytic; and Byonic score filter > 100.

GADS that can be searched in various ways. The software generates multiple scores comparing each library hit to a query spectrum. These include: "Score," an overall match employed for tandem spectra; "Dot-Score," a simple cosine comparison; Rev-Dot, a cosine comparison that only considers peaks in common; and a new partial spectrum score (PSS) that ignores nonmatching peaks in library spectra. These scores are normalized to a base peak of 999 (for historical reasons) and are the same as those long used by this software for tandem spectral comparison. The "Score" was derived from the Dot-Score after correction for spectra where all or almost all of the intensity arose from a few abundant peaks. For the GADS spectra described here, "Score" and "Dot Product" agree with a few percent. The "Names" tab enables browsing of different varieties of GADS in various ways by name or "synonym," and the "Comparison" tab allows the mass (glycan) aligned comparison of multiple GADS or tandem spectra. Further details are described in the documentation provided with the program and in the GADS paper.[1]

Annotation of Glycopeptide Fragment Ions

Raw LC-MS data and Byonic output files of human proteins were processed automatically using multiple in-house programs developed to extract the MS^2 spectra, followed by peak annotation.[6] Representations for the six common components of mammalian glycans are represented as one or two-character abbreviations as follows: HexNAc (G), hexose (H), sialic acid (S), fucose (F), SO_3 (So), and HPO_3 (Po). An example of a full glycopeptide representation is TAGWNV-PIGTLRPFLnWTGPPEPIEAAVAR/4_1(15,N,G:G5H6F4S). This begins with the usual representation of a peptide, with the asparagine point of glycan attachment (sequon) represented with a lower-case letter (n). Here, /4 denotes a charge state of +4 (4 protons attached), with a single modification (_1) attached to an asparagine (N) at position 15 (starting from the N terminus at position 0). G: denotes a glycan modification, which is G5H6F4S. This format is an extension of representations long used in NIST peptide libraries.[17] As shown in Figure 2, different colors are used for three types of ions: oxonium ions (red), peptide sequence ions (green), and glycopeptide ions (blue). Some glycopeptide ions are represented as Yn, which have glycan compositions consistent with the decomposition of high-mannose glycans, including a possible fucose attachment. The integer n denotes up to two GlcNAc units, with the remainder being Hex units. Examples are Y1 (contains one G and is often predominant at higher energies), Y2 (2 G's), Y1F (GF), and Y3 (G2H). Other glycans are represented as losses from the precursor glycopeptide, p, p-GHS, and p-GHFS. As for conventional peptide fragmentation, amino acid sequences are represented as y_n and b_n, where n is the number of amino acids beginning at the C and N terminus, respectively. Sequences containing the sequon are labeled with a tilde (~) superscript.

Criteria for Rejecting Tandem Spectra of Glycopeptides

We first annotated all tandem spectra using MS_Piano[6] to remove erroneous glycopeptide identifications. Fragment ions that did not match within 10 ppm were labeled unassigned. Any of the following criteria led to the rejection of spectra: (1) >50% of total abundance was unassigned, (2) >25% of total peaks were unassigned, and (3) no Y1 ion was present. Also, when different identifications were made for associated HCD and contingent ion-trap spectra, the identifications were rejected.

Further, nonsialylated spectra were rejected when significant (>5%) amounts of any of the followed sialylated oxonium ions were found as fragment ions: m/z 274.092 (S-H2O), m/z 292.103 (S), and m/z 657.234 (GHS). Likewise, sialylated glycans were rejected when no such ions were present. Fucosylation was confirmed by the presence of m/z 512.196 (GHF) or Y1F (peptide + GF). Any of the above inconsistencies were reported by software and then verified by manual inspection.

It is important to note that due to the very different energetics of glycan and peptide fragmentation, tandem spectra of glycopeptides depend greatly on fragmentation conditions. In this work, we chose to use "stepped" fragmentation to generate both low-energy glycopeptide ions as well as higher-energy sequence ions. The lower-energy fragments enable confirmation of glycan identity, and higher-energy ions enable sequence identification. Conditions used here began at the low NCE value of 15, which was found to yield the most complete glycopeptide spectra while not having any measurable effect on the number of glycopeptides identified. Moreover, contingent ion-trap fragmentation served as the softest possible fragmentation where most product ions were generally created by a single bond-breaking event for glycan identity confirmation. Both stepped HCD and ion-trap spectra are included in tandem libraries available with this work. These libraries are intended to enable manual confirmation of any questionable glycopeptide identifications. They can be searched against glycopeptide libraries using fragmentation conditions similar to those in the library using conventional (nonpeptide) scoring. Methods for com-

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

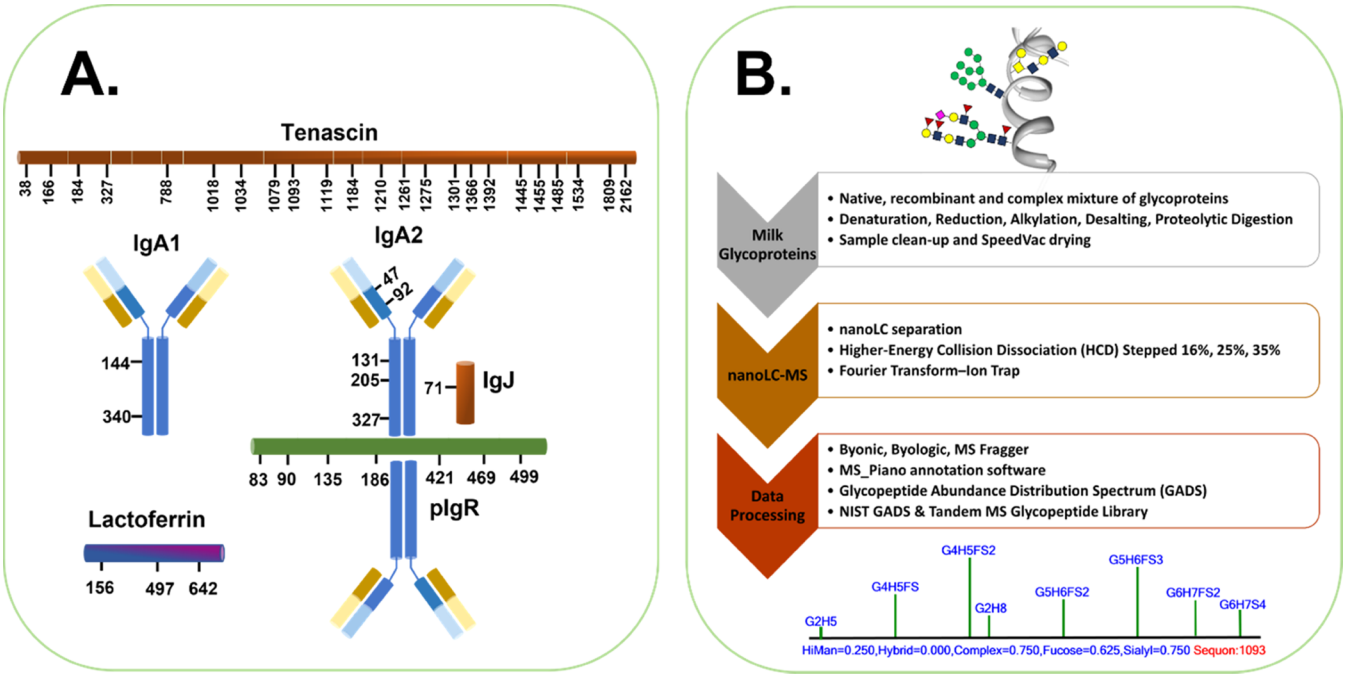


Figure 1. (A) Glycosylation of human milk proteins and their corresponding sequons based on UniProt[18] nomenclature. These proteins include tenascin (TNC), lactoferrin, and immunoglobulin A1 (IgA1). Immunoglobulin A2 (IgA2) complex, which includes the immunoglobulin J (IgJ) chain and polymeric immunoglobulin receptor (pIgR).[19,20] (B) Schematic representation of the present glycopeptide analysis.

puter validation using more detailed analysis are under development.

RESULTS

Tandem MS Spectra and Libraries of Selected Milk Glycoproteins

For the individual glycoproteins examined in this study (Figure 1A), glycopeptide-based workflows involved proteolytic enzyme digestion of glycoproteins, nanoLC-MS analysis, followed by data processing (Figure 1B). The stepped NCE option of recent Orbitrap mass spectrometers generates product ions at three collision energies, covering the range needed to generate significant ions from both glycan and peptide cleavage. Byonic software then served to identify glycopeptides. Next, MS_Piano[6] was employed to annotate N-linked glycopeptides in tandem MS library format (.msp files).[16]

Figures 2−4 show the glycosylation sites identified and verified through annotation of the MS^2 spectra of glycopeptides derived from lactoferrin, tenascin, IgA1, IgA2, IgJ, and pIgR.

Lactoferrin. Glycosylation patterns of human lactoferrin in colostrum and mature milk are diverse.[10,21−24] A total of 448 different N-linked lactoferrin glycopeptides (Table 1) have been identified from three known glycosylation sites (Figure 1A) in this work.[16]

A series of oxonium ions, amino acid sequence, and glycopeptide fragment ions are required for the confident characterization of glycopeptides. An example of a stepped HCD MS^2 spectrum of a complex glycopeptide derived from human lactoferrin, as annotated by MS_Piano,[6] is presented in Figure 2A. The most intense peaks are oxonium ions (G, S, GH, GHF, GHF2, GHFS, G2H2F3). These provide information on the glycan sequence containing multiple fucose residues. Identified *y* and *b* ions confirm the peptide sequence. The glycopeptide ion peaks, i.e., Y1, Y1F, p-G2H2F3S, p-GHS, are consistent with the structure of the presumed attached glycan, G5H6F4S, and oxonium ions. However, we note that all connectivity details generally cannot be deduced from this information. Shown in Figure 2B is the corresponding ion-trap collision-induced dissociation (CID) (NCE = 30%) MS^2 spectrum, which yields primarily glycopeptide ions produced by the loss of singly charged oxonium ions. Note that such cleavages result in a reduction in ion charge, which inhibits further decomposition since lower charge state ions are less prone to dissociation than higher charge state ions.

Tenascin. Human tenascin is a significant constituent of human milk.[7,8] In the present work, 23 N-linked glycosites were identified and yielded 2695 distinct glycopeptides with charge states ranging from +2 to +6. Figure 3 shows the annotation of ion fragments of two glycopeptide ions. One confirms the connectivity of complex glycan containing multiple sialic acid residues (Figure 3A), which is also consistent with its retention time. Annotation of a glycopeptide containing a high-mannose glycan (Figure 3B) is also illustrated.

As an example of a complex glycan, MS_Piano annotated a triantennary-sialofucosylated glycan (G5H6FS3) on the peptide sequence R.LnWTAADQAYEHFIIQVQEANK.V (Figure 3A). A series of *y* ions (y_2, y_4 to y_9), oxonium ions (G, S, GH, HS, GHS), and Y*n* (Y1 to Y4, Y1F, Y4F), as well as peptide-bound glycosidic cleavages (i.e., p-GHS2, p-S), were observed.

As an illustration of a high-mannose glycan, the MS^2 spectrum (Figure 3B) revealed nearly complete coverage of glycopeptide sequence of G2H5 on R.LnYSLPTGQWVGVQLPR.N. This includes multiple glycan fragments (G, GH, GH2, GH3), peptide-bound glycosidic Y-type ions (Y0 to Y6), and the peptide sequence *y* ions (y_2, y_3, y_5 to y_{10}, y_{12} to y_{15}). Abundant y ions are expected for these peptides, especially those with a basic amino acid at the C-terminus that can carry a proton.[2]

Major glycosylated proteins in milk are immunoglobulins IgA1 and IgA2 and secretory components IgJ and pIgR. Site-

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

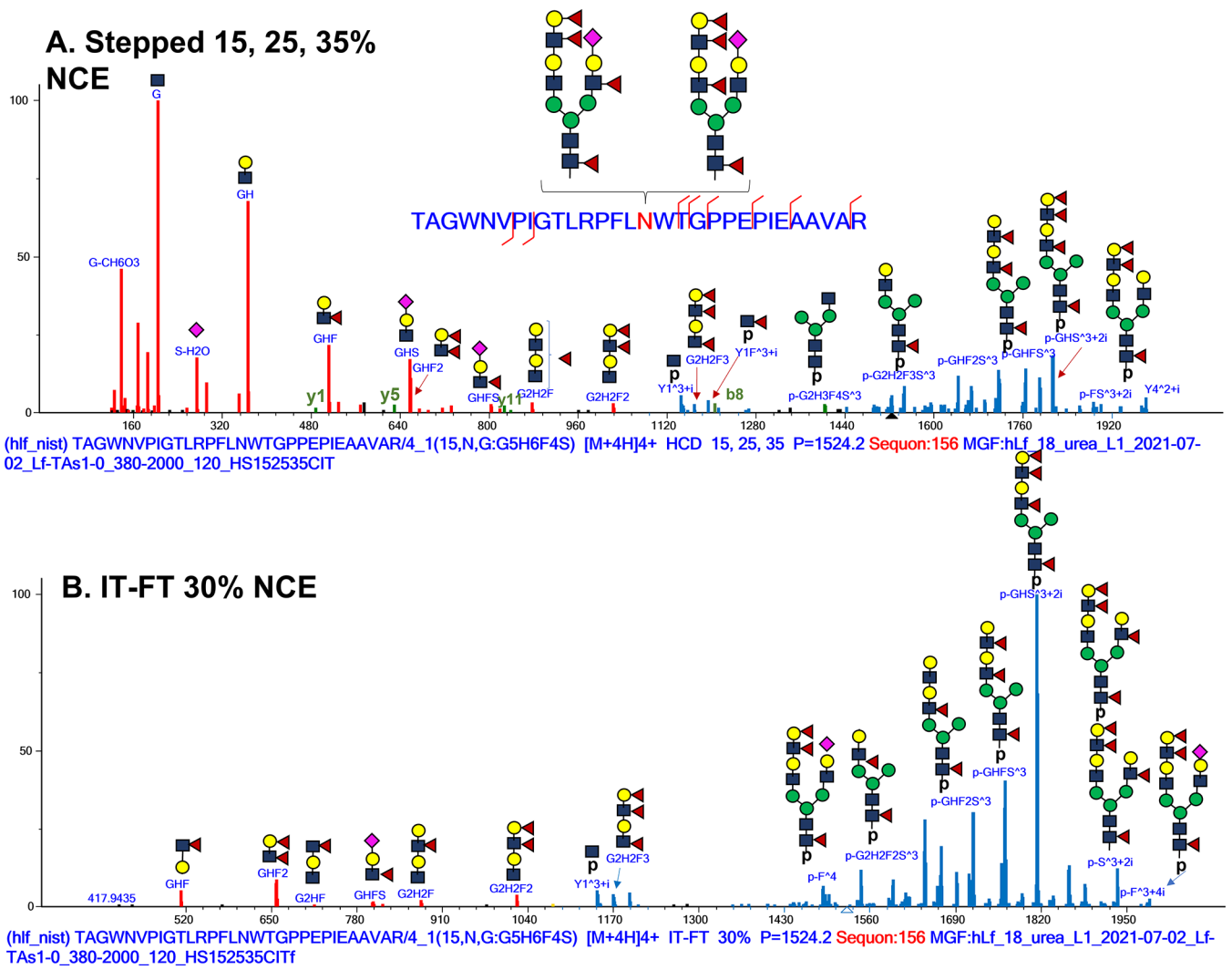


Figure 2. HCD and CID MS^2 spectra of a sialofucosylated diantennary glycopeptide (m/z 1524.196, z = 4+) derived from human lactoferrin. Spectra are shown for stepped HCD (2A) and ion-trap fragmentation (2B), both with Orbitrap detection. Annotation, Y*n*, and *p* (precursor, blue letter) denote glycopeptide ions. Glycans are composed of galactose (H, yellow circle), mannose (H, green circle), fucose (F, red triangle), N-acetylglucosamine (G, blue square), N-acetylneuraminic acid (S, pink diamond), peptide (p, black letter), and glycan "cartoons" represent glycan−peptide cleavage ions. Note that the glycan "cartoons" have been manually added to the program output.

specific glycosylation analysis has been done for each, which has also been the subject of study in various biological fluids.[14,20,25]

Immunoglobulin A (IgA1 & IgA2). A total of 594 and 421 different nonredundant N-glycopeptides were identified from the IgA1 (two sequons) and IgA2 (five sequons) digests, respectively. MS_Piano was applied to annotate the intense oxonium, b/y, and glycopeptide (expressed as Y*n* or *p*-glycan) peaks of IgA1 and IgA2 glycopeptide ions (Figure 4A,B).

Together with the precursor mass and the G, S, GH, and GHS ions, the y_4, y_5, y_7/b_7, y_8/b_8 ions, glycosidic cleavage residues were confirmed, thereby providing assignment of the glycopeptide R.LAGKPTHVnVSVVMAE.V and G4H5S (Figure 4A). Similarly, the 100% sequence coverage of glycan residues (G, S, GHS) and the Y0 to Y6 ions plus additional *p*-HS and *p*-S peptide-glycosidic cleavage sites confirmed a hybrid-type N-linked glycopeptide (Figure 4B). The identified glycopeptide contained a series of peptide cleavage type *y* (y_3−y_8) and b (b_2−b_7) ions in the spectrum, confirming the sequence E.DLLLGSEAnLTCTLTGLR.D.

Secretory Component Proteins (pIgR & IgJ). Immunoglobulin A2 (Figure 1) is a polypeptide complex consisting of two IgA monomers connecting the J chain and the polymeric immunoglobulin receptor.[14,19,25] Protein IgJ has one and protein pIgR has seven known glycosylation sites.

Since a source of human IgJ could not be found, the glycan distribution of its single sequon was derived from an analysis of the NIST human milk SRM 1953. From this data, 596 MS^2 high-quality spectra for 120 unique IgJ glycopeptides were identified. These were confirmed from three different peptides as reported in the Supporting Section 3C. An example is shown in Figure 4C, which confirms the assignment of a disialo-diantennary N-linked glycan on the tryptic peptide R.EnISDPTSPLR.T.

Digests of polymeric immunoglobulin receptor protein yielded 2305 different MS^2 glycopeptides originated from its seven identified glycosylation sites. To illustrate the importance of annotation by MS_Piano, Figure 4D shows a case of the MS^2 spectrum where positions of multiple fucose residues can be deduced. The abundant peaks for Y1F, Y4F, and Y5F indicate that a fucose residue is attached to the core HexNAc of a G5H6F2 glycan. In addition, an oxonium ion (GHF) and peptide-bound glycosidic cleavage site type of ions were also

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

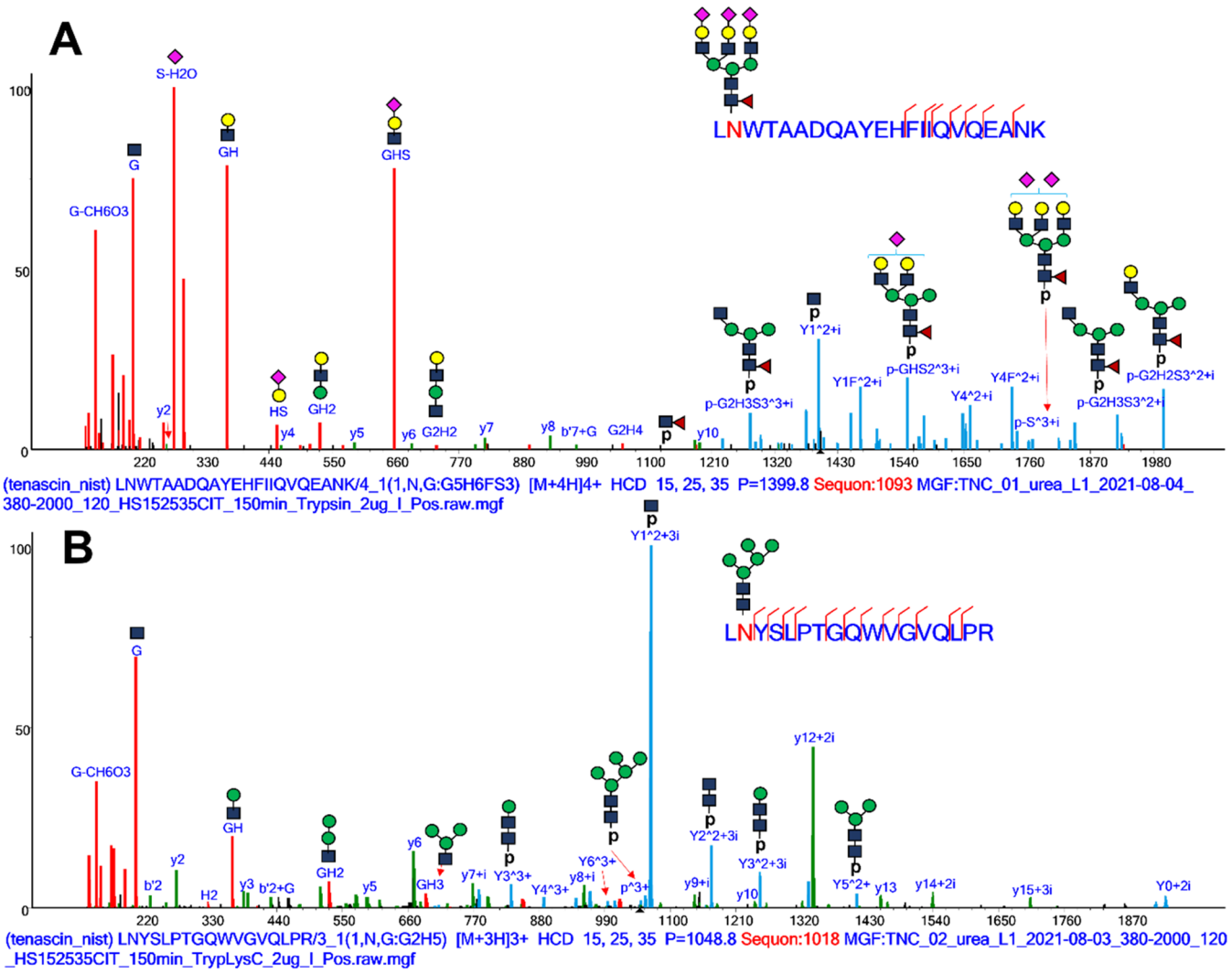


Figure 3. MS^2 product ion annotation of glycopeptide spectra at glycosylation sites (3A) N1093, m/z 1399.838 and (3B) N1018, m/z 1048.822 from digests of human tenascin.

detected and served as diagnostic ions for the second fucose residue located on the antenna of the G5H6F2S glycan.

Annotated Spectra of Human Milk Glycoproteins

Human Milk SRM 1953. The analysis of Standard Reference Material illustrates the application of the present method to N-glycosylation analysis of biological fluids. Digests were enriched in glycopeptides by chromatography—hydrophilic interaction liquid chromatography solid-phase extraction (HILIC SPE). This was followed by glycopeptide identification after LC-MS analysis using the three-energy stepped fragmentation method described earlier. Data were processed and converted into a library of MS^2 spectra of glycopeptides with annotation by MS_Piano. Glycopeptides from 111 identified glycoproteins are listed in the Supporting Section 3D. The most abundant glycoproteins identified in the human reference milk were pIgR, lactoferrin, IgA1, IgA2, (IgJ), (α-S1-casein), and (tenascin). Overall, 3653 MS^2 spectra of intact glycopeptides and 373 unique glycosylation sites were found (Tables 1 and 2). A tandem MS library for all glycopeptides identified is included with other libraries for downloading,[16] with selected illustrations shown in Figures 2−5. Results show a high degree of spectral similarity to the same peptides from other sources of proteins in both recombinant and different milk samples (Supporting Figures 1S−3S).

Figure 5 demonstrates the ability of the present method to compare MS^2 spectra acquired from multiple sources. Figure 5A,B shows the annotation of MS^2 spectra of a pIgR glycopeptide derived from a purified protein and the reference milk. The combination of low and high energies in the stepped method enhanced the signal of peaks confirming the glycan and peptide identification in a given sequon. A diagnostic ion (GHF) was observable among the fragment ions, confirming that the fucose residue is attached to the antenna of G4H5FS for the peptide. The completeness of the present fragmentation method aids the confident identification of glycan identity regardless of search engine scores.

Comparison to Results of Interlaboratory Studies. The present approach applied to other milk samples is further illustrated using LC-MS HCD datasets obtained from Utrecht,[12,26] and Wuhan[27] laboratories. MS_Piano annotated a total of 2671 (Donor 1, Utrecht), 4678 (Donor 2, Utrecht), and 257 (Wuhan) MS^2 spectra of intact N-linked glycopeptides. Figure 5C,D shows HCD results for a selected glycopeptide from Utrecht[26] (27% NCE) and Wuhan[27] (32% NCE), respectively, using milk samples from individual donors. Since

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

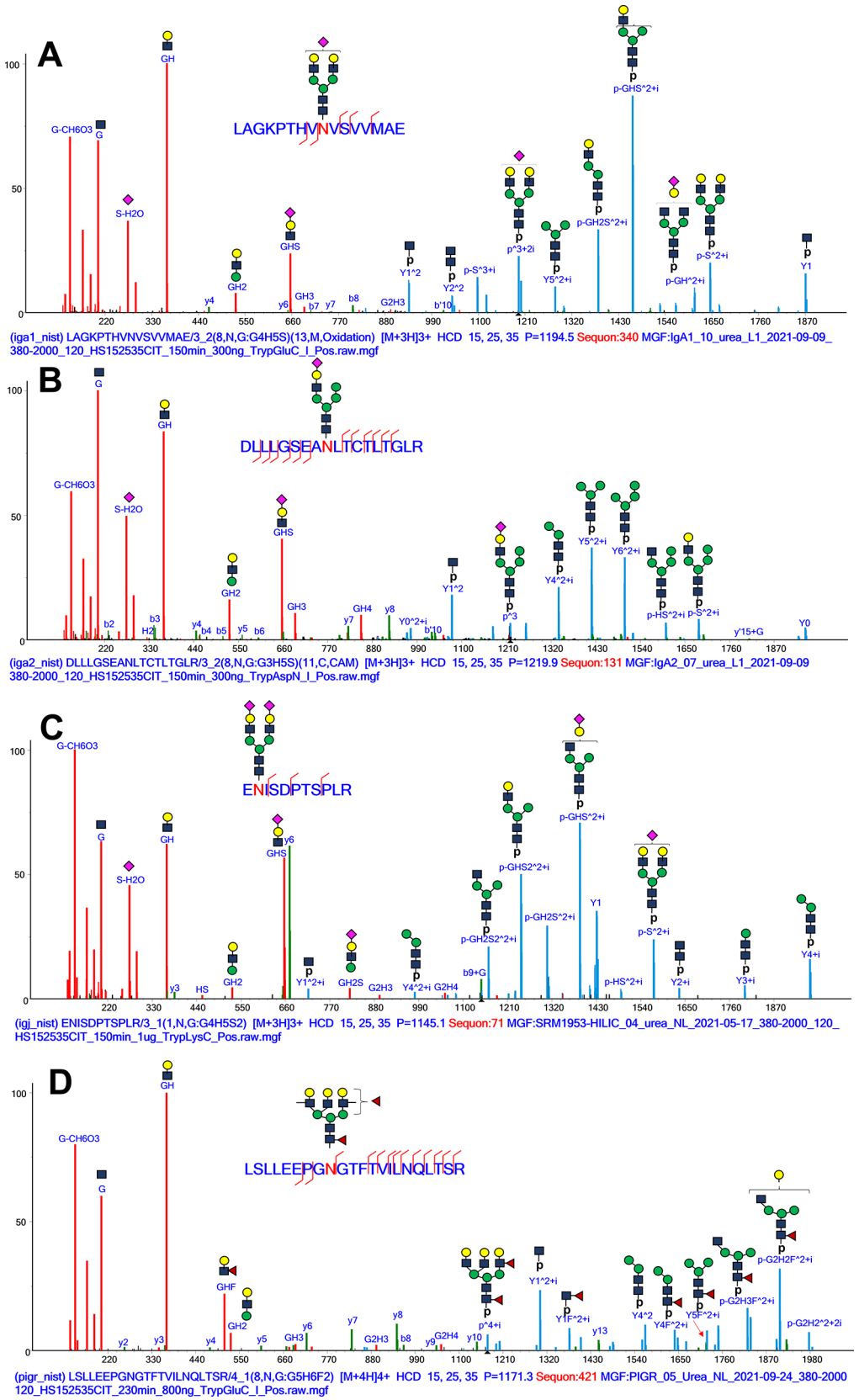


Figure 4. Tandem mass spectral of four immunoglobulin glycopeptides from digests of (A) IgA1, m/z 1291.895; (B) IgA2, m/z 1219.879; (C) IgJ, m/z 1145.132; and (D) pIgR, m/z 1171.287. Here are representative spectra acquired at stepped HCD at NCE 15, 25, and 35%. Annotation by MS_Piano.[6]

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

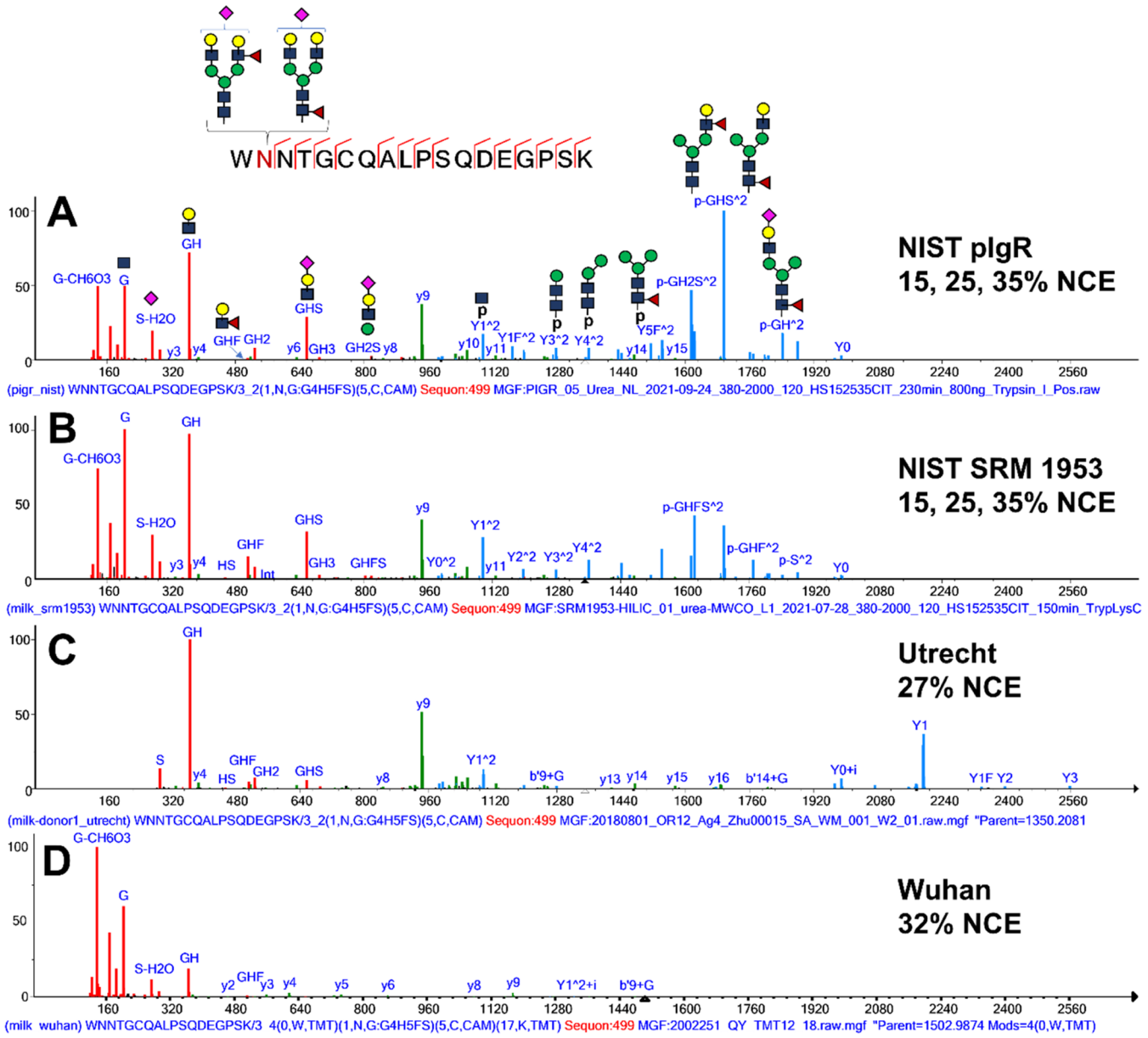


Figure 5. MS^2 product ion annotation of glycopeptide spectra (m/z 1350.209) at glycosylation site N499 of pIgR digests from human milk datasets of (A) NIST pIgR, (B) NIST SRM 1953, (C) Utrecht,[26] and (D) Wuhan.[27] Glycopeptides derived from Wuhan were tagged with a single tandem mass tag (TMT), so the precursor ion has an additional mass of m/z 229.163. Peak annotation marks the oxonium ions (red), peptide product ions (green), and glycopeptide N-glycan ions (blue). Galactose (H, yellow circle), mannose (H, green circle), fucose (F, red triangle), N-acetylglucosamine (G, blue square), and N-acetylneuraminic acid (S, pink diamond). Note that the lower yield of glycopeptide ions in (C) and (D) is due to the use of only higher-energy fragmentation settings.

these studies were done at higher energies, only Y0 and Y1 ions were identified, which, while helpful for confirming glycan masses, do not permit linkage confirmation.

Glycopeptide Abundance Distribution Spectra

We now consider site-specific distributions of glycans, expressed as glycopeptide abundance distribution spectra (GADS)[3] for major glycoproteins in human milk. GADS contain the N-linked glycan distribution, glycan compositions, peptide backbone, sequon, charge state(s), and relative MS^1 abundances. N-glycans are grouped into the three classes[3,28] high-mannose, complex, and hybrid. GADS for selected tenascin peptides are illustrated in Figure 6. Examples for other proteins are presented in the Supporting Figures 4S−7S, and complete documentation is provided with the downloadable software.[16]

GADS of Tenascin. Identification of 23 N-glycosylation sites with 45 K glycopeptide ions and 2695 different glycopeptides (Tables 1 and 2) was made. Figure 6 illustrates the abundance of various glycoforms in tenascin sequons consisting of mannosylated, fucosylated, and sialofucosylated N-glycans.

Figure 6A displays the GADS at site N1093, which holds a wide range of di-, tri, and tetra-antennary N-linked glycans, most of which were sialylated and fucosylated. For sequon N1034, Figure 6B shows a GADS-containing mixture of intermediate-size glycans, almost all of which were fucosylated and the majority sialylated. Abundant high-mannose glycans (G2H5, G2H6) were found at N1018 (Figure 6C). A GADS library for tenascin derived from various proteases for multiple sequences, 23 glycosylation sites, glycoforms, and different charge states is

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

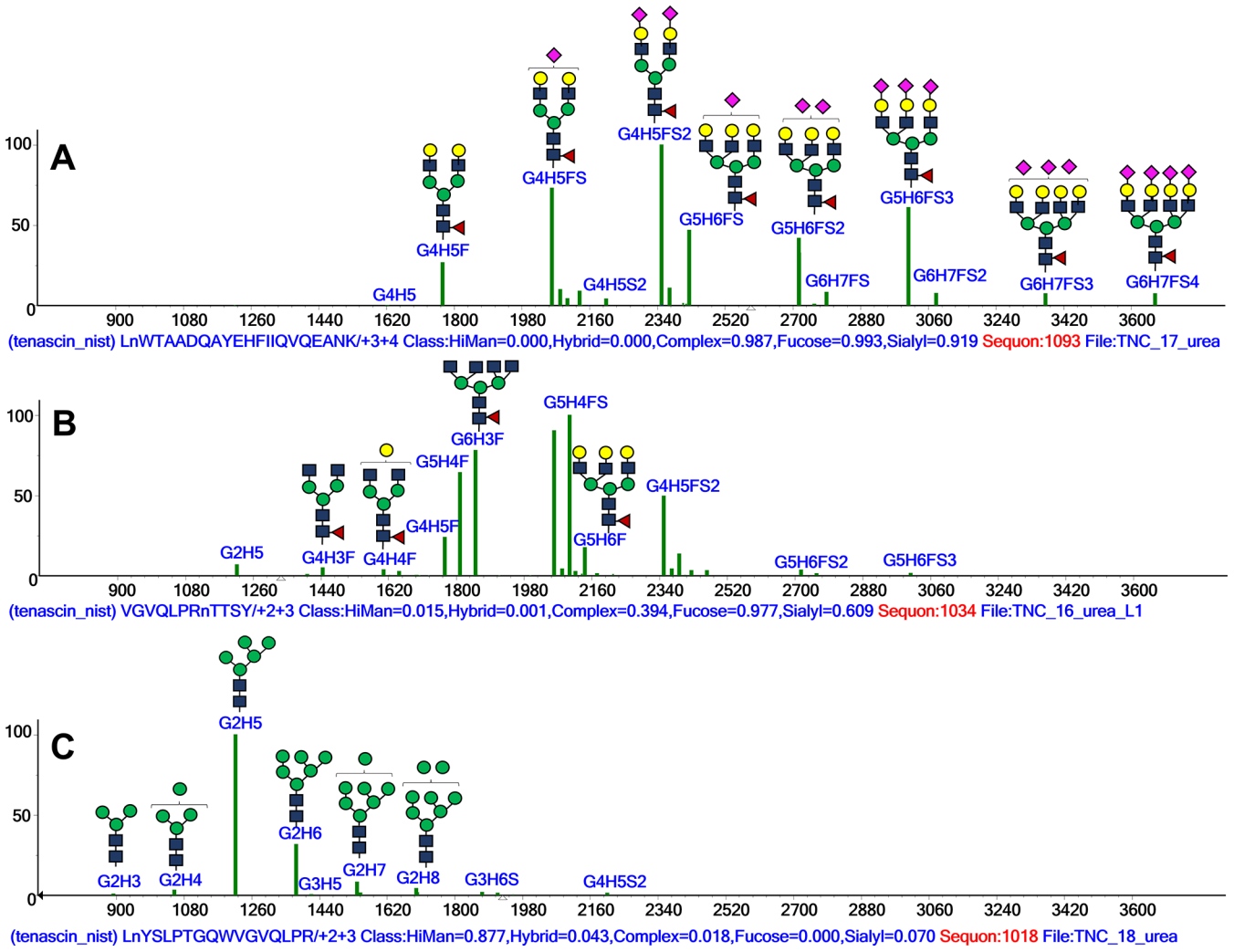


Figure 6. Glycosylation of selected tenascin glycopeptides. (A) GADS sequon N1093 is a highly complex, sialofucosylated glycan; (B) sequon N1034 is a mixture of high-mannose and complex glycans; and (C) sequon N1018 contains primarily high-mannose glycans. Nomenclature: G5H4FS contains five HexNAc (blue square), four hexoses (green or yellow circle), one fucose (red triangle), and one sialic acid (pink diamond) glycans. The "cartoon" illustrations represent presumed N-glycan structures. Green peaks have scores and retention times within thresholds. The x-axis represents the N-glycan mass. Each peak is the identified N-glycopeptide based on its MS^2 fragmentation.

in the Supporting Section 2.[16] Major peptides identified in tenascin digests originated from the cleavages of trypsin, chymotrypsin, α-lytic, Glu-C, and Asp-N is shown in the Supporting Table 1S, along with a number of glycans found for each. This paper is the first that we know of to report the full site-specific glycan distribution for this protein.

GADS of Human Milk SRM 1953 and Interlaboratory Studies. This section describes applications of GADS to two glycoproteins in human milk samples, immunoglobulin A1 (Figure 7) and polymeric immunoglobulin receptor (Figure 8).

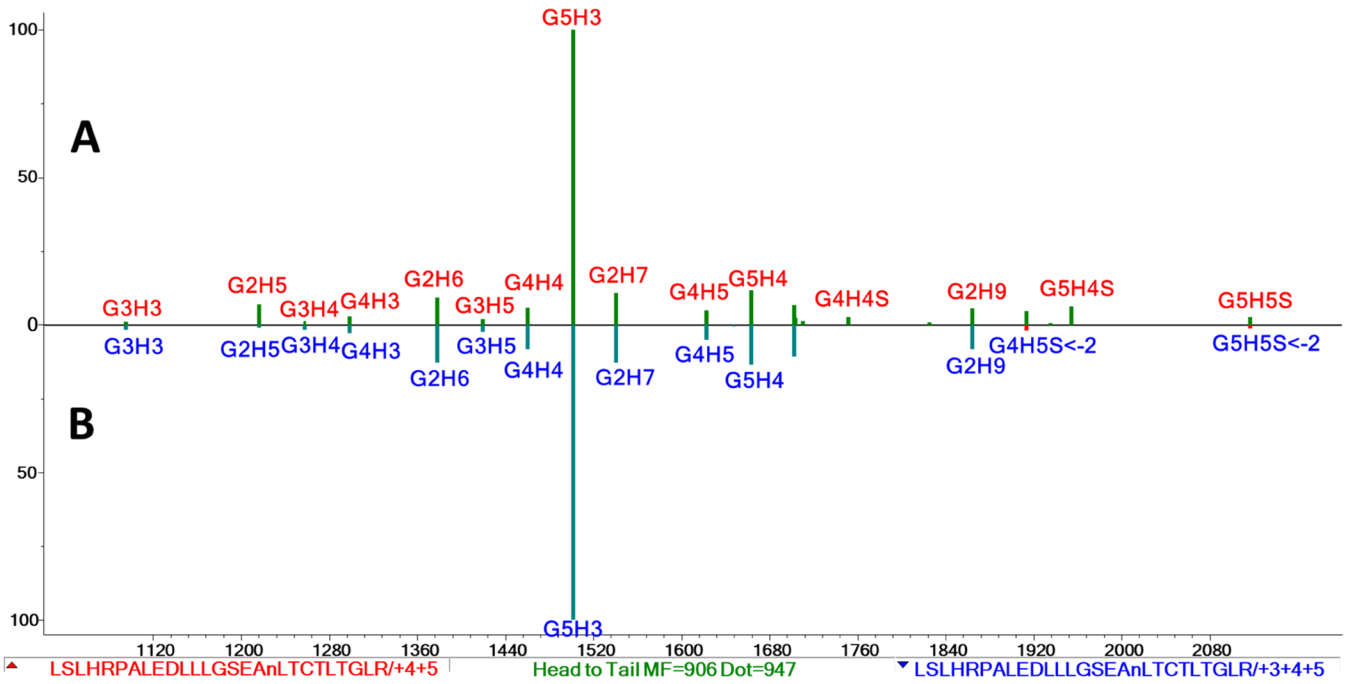
As an illustration of interlab comparisons, GADS for these proteins were also derived from Utrecht University's milk datasets.[26] Figure 7 shows a high degree of similarity between the IgA1 glycosylation site N144 of the reference milk (SRM 1953) and milk from healthy donor two. The most abundant glycopeptide at N144 for the milk SRM and donor two (Utrecht)[12,26] was G5H3. Verification of the spectrum of this glycopeptide is illustrated in Supporting Figure 8S.

This interlaboratory comparison shows the utility of the GADS format for detailed comparisons of datasets acquired for different samples and fragmentation methods such as HCD and EThcD (Figure 8). Human milk samples from other labs prepared and analyzed by LC-MS at Utrecht[12] and Oregon[29] laboratories were compared with the NIST reference milk. Using the same data processing parameters, GADS libraries were generated for the milk datasets listed in Table 2 and shown in Figure 8. GADS of polymeric immunoglobulin receptor glycopeptides at sequon N499 derived from human milk datasets were used for comparison. The comparison DotProd scores of the reference GADS (SRM 1953), from the bottom, are 890 and 851, both from Utrecht University,[12] indicating a high degree of similarity in the profiles of high-mannose, hybrid, and complex N-linked glycopeptides.

■ DISCUSSION

The individual steps described above provide a general path for acquiring and confirming site-specific glycosylation distributions as well as for generating libraries of their underlying glycopeptide spectra. This involves multiple validation steps to deal with inherent uncertainties confronting the reliable determination of the individual glycopeptides involved. As noted by others,[1,32] spectrum search score alone is insufficient

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX



#	Library	Score	Dot-Prod	Rev-Dot	PSS-Dot	Name	Sample
1	**milk-donor2-utrecht**	**906**	**947**	**983**	**933**	**R.LSLHRPALEDLLLGSEAnLTCTLTGLR.D/+3+4+5**	**Week 3**
2	milk-donor2-utrecht	900	949	973	947	R.LSLHRPALEDLLLGSEAnLTCTLTGLR.D/+3+4+5	Week 2
3	milk-donor2-utrecht	898	934	961	924	R.LSLHRPALEDLLLGSEAnLTCTLTGLR.D/+3+4+5	Week 1
4	milk-donor2-utrecht	870	929	939	923	R.LSLHRPALEDLLLGSEAnLTCTLTGLR.D/+3+4+5	Week 6
5	milk-donor2-utrecht	863	906	943	913	R.LSLHRPALEDLLLGSEAnLTCTLTGLR.D/+3+4+5	Week 3
6	milk-donor2-utrecht	860	920	942	912	R.LSLHRPALEDLLLGSEAnLTCTLTGLR.D/+3+4+5	Week 4
7	milk-donor2-utrecht	839	914	981	914	R.LSLHRPALEDLLLGSEAnLTCTLTGLR.D/+4+5	Week 10

Figure 7. GADS comparison of IgA1 glycopeptides at sequon N144 obtained from human milk tryptic digests (A) NIST SRM 1953 and (B) Utrecht dataset at different lactation periods. The top spectrum is an IgA1 glycopeptide derived from the reference material, while the bottom GADS is from human milk in the Utrecht laboratory study.[26] The table shows the hitlist of glycopeptide identifications with corresponding match factor scores for the same peptide from different samples. Score, Dot-Prod, Rev-Dot, and PSS-Dot are comparison scores described in Supporting Figure 9S. A score ≥800 is a good match, and 900 is an excellent match, as previously reported.[31] The small letter "n" is the asparagine connected to the glycan. GADS peaks shown in red are for glycopeptides whose relative retention times are out of an expected range and labeled as uncertain. Annotation <−2 indicates that these peptides eluted earlier than expected by 2 min.

for reliable identification due to the relatively small number of individual monosaccharides involved, the possibility of hundreds of glycans derived from them, and often low signal intensities. The essential components of this process are: (1) tandem spectrum measurement by fragmentation involving high and low collisional energies; (2) identification of glycopeptides by a search engine; (3) annotation of all fragment ions in each identified spectrum, (4) derivation of quality measures for each spectrum; (5) marking or excluding of glycopeptides whose retention times are inconsistent with known sialylation-dependent relative retention behavior; and (6) reproducibility of the data and its independence of specific peptides and proteases. Moreover, tandem spectral libraries may also be used for the general identification of glycopeptides and as a resource for examining fragmentation patterns of glycans as a function of sequence and charge.

Annotation and Comparison of the MS^2 Spectra of Glycopeptides

Available software tools such as Sweet-Net,[33] GP-Quest,[34] and other search engines[1,13,35] for glycopeptides deal mainly with identifying N- or O-linked glycosylation. Unlike MS_Piano,[6] such identification programs do not fully annotate experimental spectra or compare them or provide specific measures of spectrum quality. Such annotation using the present methods is illustrated for 11 N-linked glycopeptides from 7 different glycoproteins in Figures 2−5. For example, to our knowledge, this is the first report of an annotated glycopeptide spectrum in lactoferrin digests that contain four fucose residues in sialofucosylated glycans (Figure 2), MS fragmentation, and high-mannose occupancy at Asp 1018 of tenascin glycoprotein (Figures 3 and 6). Varied positions of fucose residues on the same glycosylation site in the MS^2 spectrum of pIgR glycopeptide ion (Figure 4D) were also observed. While none of these findings could have been obtained manually, the present methods provided a facile means of deriving and preserving this information.

As expected for library analysis and its utility, multiple spectra are readily compared to each other or a query spectrum. For example, the Supporting Figures 1S−3S illustrate the comparison of tandem mass spectra between the glycopeptides obtained

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

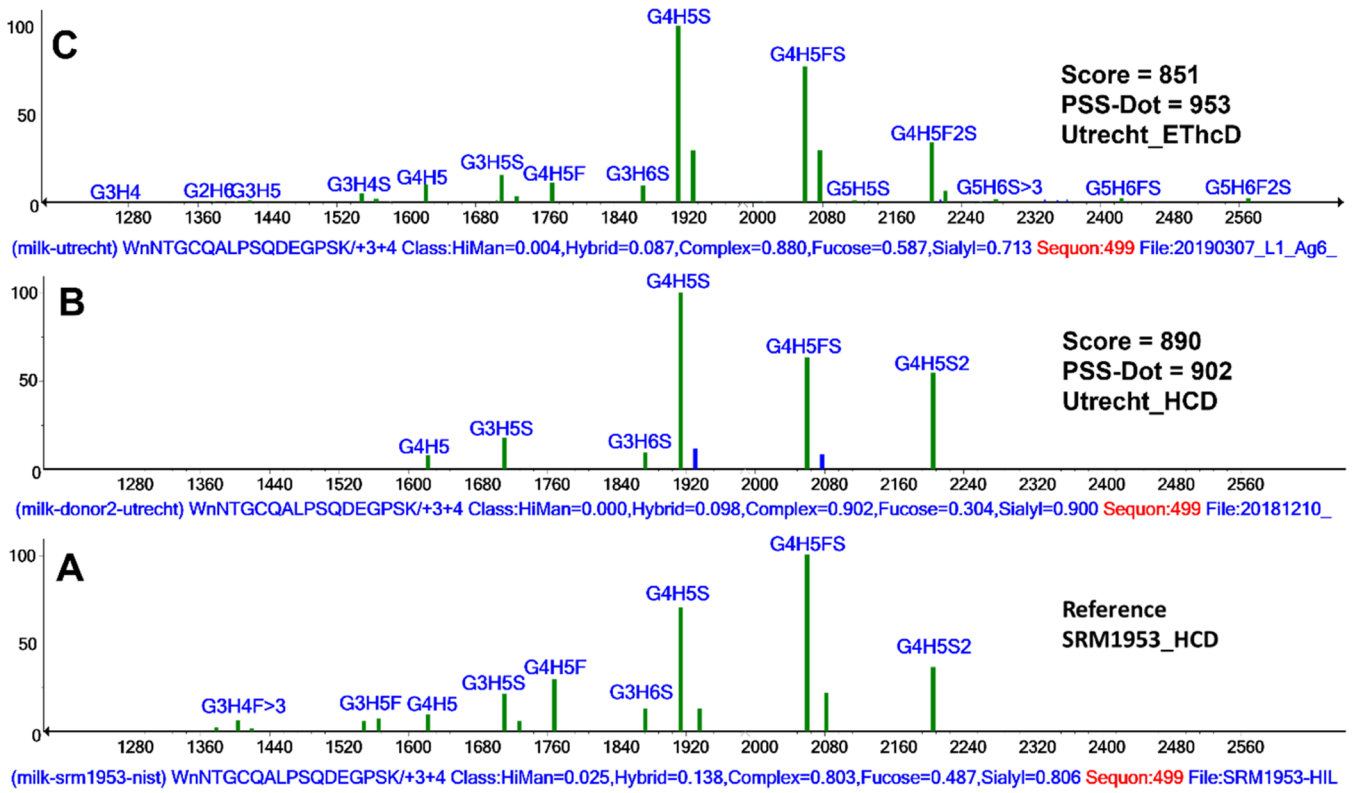


Figure 8. GADS comparison of sequon N499 from polymeric immunoglobulin receptor glycopeptides derived from human milk. Dot product and partial spectrum scores (PSS-Dot) of the (A) reference GADS (SRM 1953), from the bottom, are (B) 890 (Milk-Donor2 Utrecht University)[12] and (C) 851 (Utrecht University).[30]

from the digests of recombinant proteins and human milk reference material. Comparison of site occupancy and relative abundance of various glycans in each sequon of milk glycoprotein can be found in Figures 4S−7S. The current method can be used to compare spectra acquired in different labs or under different conditions to confirm the common origin of an identified glycopeptide (Supporting Figures 8S, 10S−11S). In addition, sulfonated and phosphorylated glycopeptides have not been identified in any milk proteins studied.

The established approach was further illustrated by analyzing LC-MS HCD datasets using single energy from human milk digests at Utrecht[12,26] and Wuhan[27] laboratories.

Comparison of Glycopeptide Distributions

GADS directly displays and compares the relative ion abundances of each glycan on a sequon (Figures 6−8) even for different sequences and charge states. For example, we found that two sequons for lactoferrin are fully occupied and showed a reliable number of identified N-glycans. However, based on peptide abundance, N642 is only 4% occupied. GADS of lactoferrin's three glycosylation sites, N157, N497, and N642, is shown in Supporting Figure 7S. These results agree with previous human milk datasets reported by other laboratories.[22,26,36]

Furthermore, some of the identified N-glycans unique for milk glycoproteins are confirmed present in previous released glycans studies.[37,38] However, glycan release methods do not provide the distributions at each sequon for multiple sequon proteins. Comparisons are illustrated for peptides from multiple protein digests (Figure 6). Presently, distributions are often compared for groups of glycans using bar graphs. Each bar represents groups of glycans, often combining sialylated and nonsialylated glycans and combining glycans containing four or more GlcNAc into a single class. Since GADS intensities are based on MS^1 signal intensities, details of the fragmentation method have no effect on these representations, enabling, for example, a comparison of results from HCD with EThcD fragmentation. Moreover, the present approach enables an automated means of analyzing large datasets with digital storage for reuse in the form of mass spectral libraries.

We note that these methods work equally well with any available glycopeptide identification method that provides ion abundances and retention times after translation to our ".msp" format.[16] We have successfully applied other glycopeptide search methods, namely pGlyco[35] and MSfragger-glyco,[13] and found that the results are in good agreement with the present method (Supporting Figure 12S).

CONCLUSIONS

This paper illustrates a new approach for determining and representing site-specific distributions of N-linked glycopeptides by extending previous methods that generated glycopeptide abundance distribution spectra (GADS). The principal glycoproteins in human milk serve to demonstrate this approach. Validation of the results uses annotation methods, optimized variable energy fragmentation, and confirms all identifications using low-energy ion-trap fragmentation. Further confirmation is done by comparing GADS acquired for different peptides using different proteases. The different underlying peptide sequences and the acquisition of replicate runs are conveniently stored and examined in output libraries. In addition to GADS libraries, annotated tandem libraries are generated to further validate identifications.

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX

The capabilities of these methods are illustrated with human milk glycoproteins. This examined a variety of individual proteins carrying a wide range of glycans and site-specific distributions. Uses of annotation for confirming details of highly complex glycans are shown. Such annotation is essential for validating glycopeptide identities and qualifying spectra for inclusion in glycopeptide libraries. Applications to milk with its large number of glycoproteins are also shown, with tandem identified glycopeptide spectra included in libraries using data from our laboratories and others. In studies in our laboratory, a dual fragmentation approach, starting with beam-type collision cell fragmentation followed by contingent ion-trap fragmentation, enables a high degree of confirmation.

All libraries, both GADS and tandem MS of glycopeptides, discussed in this paper may be freely downloaded along with adapted NIST user interface software.[16] Spectra in the.msp ASCII format as described in the documentation may be imported or exported. This requires the creation of these files by users. MS_Piano peak annotation software is described elsewhere.[6] Different fragmentation and glycopeptide identification methods may be used.

The methods described here provide the basis for developing a fully automated data analysis method for creating validated GADS and their underlying glycopeptide libraries. Development of libraries for other common glycoproteins, such as those found in blood, is underway along with fully automated library-based validation tools.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00286.

Supporting protocol on sample preparation, digestion, and LC-MS/MS analysis; Figures 1S−12S and Table 1S (Section S1); user guide for NIST-MS glycopeptide abundance distribution spectra (GADS) and NIST tandem MS libraries, software, and documentation; software and libraries are available for download (Section S2);[16] PDF)

N-glycan database-optimized stepped HCD energies and list of glycoproteins identified in human milk SRM1953 used in the study (Section S3A−D) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Concepcion A. Remoroza − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States;* orcid.org/0000-0003-1540-1635; Email: concepcion.remoroza@nist.gov

Authors

Meghan C. Burke − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States;* orcid.org/0000-0001-7231-0655

Xiaoyu Yang − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States;* orcid.org/0000-0003-3371-9567

Sergey Sheetlin − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States*

Yuri Mirokhin − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States*

Sanford P. Markey − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States*

Dmitrii V. Tchekhovskoi − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States*

Stephen E. Stein − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States;* orcid.org/0000-0001-9384-3450

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00286

Notes

All commercial instruments, software, and materials used in the study are for experimental purposes only. Such identification does not intend recommendation or endorsement by the National Institute of Standards and Technology, nor does it plan that the materials, software, or instruments used are necessarily the best available.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Drs. Oleg Toropov, Zheng Zhang, Yi Liu, Zachary Goecker, and Yuxue Liang for the technical support.

■ REFERENCES

(1) Bollineni, R. C.; Koehler, C. J.; Gislefoss, R. E.; Anonsen, J. H.; Thiede, B. Large-scale intact glycopeptide identification by Mascot database search. *Sci. Rep.* **2018**, *8*, No. 2117.

(2) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **2007**, *4*, 709−712.

(3) Remoroza, C. A.; Burke, M. C.; Liu, Y.; Mirokhin, Y. A.; Tchekhovskoi, D. V.; Yang, X.; Stein, S. E. Representing and Comparing Site-Specific Glycan Abundance Distributions of Glycoproteins. *J. Proteome Res.* **2021**, *20*, 4475−4486.

(4) Lowenthal, M. S.; Davis, K. S.; Formolo, T.; Kilpatrick, L. E.; Phinney, K. W. Identification of Novel N-Glycosylation Sites at Noncanonical Protein Consensus Motifs. *J. Proteome Res.* **2016**, *15*, 2087−2101.

(5) Desaire, H. Glycopeptide Analysis, Recent Developments and Applications*. *Mol. Cell. Proteomics* **2013**, *12*, 893−901.

(6) Yang, X.; Neta, P.; Mirokhin, Y. A.; Tchekhovskoi, D. V.; Remoroza, C. A.; Burke, M. C.; Liang, Y.; Markey, S. P.; Stein, S. E. MS_Piano: A Software Tool for Annotating Peaks in CID Tandem Mass Spectra of Peptides and N-Glycopeptides. *J. Proteome Res.* **2021**, *20*, 4603−4609.

(7) Turula, H.; Wobus, C. E. The Role of the Polymeric Immunoglobulin Receptor and Secretory Immunoglobulins during Mucosal Infection and Immunity. *Viruses* **2018**, *10*, No. 237.

(8) Fouda, G. G.; Jaeger, F. H.; Amos, J. D.; Ho, C.; Kunz, E. L.; Anasti, K.; Stamper, L. W.; Liebl, B. E.; Barbas, K. H.; Ohashi, T.; Moseley, M. A.; Liao, H. X.; Erickson, H. P.; Alam, S. M.; Permar, S. R. Tenascin-C is an innate broad-spectrum, HIV-1-neutralizing protein in breast milk. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 18220−18225.

(9) Picariello, G.; Ferranti, P.; Mamone, G.; Roepstorff, P.; Addeo, F. Identification of N-linked glycoproteins in human milk by hydrophilic

interaction liquid chromatography and mass spectrometry. *Proteomics* **2008**, *8*, 3833−3847.
(10) Yang, Z.; Jiang, R.; Chen, Q.; Wang, J.; Duan, Y.; Pang, X.; Jiang, S.; Bi, Y.; Zhang, H.; Lönnerdal, B.; Lai, J.; Yin, S. Concentration of Lactoferrin in Human Milk and Its Variation during Lactation in Different Chinese Populations. *Nutrients* **2018**, *10*, 1235.
(11) SRM 1953 - Organic Contaminants in Non-Fortified Human Milk. 2017, https://www-s.nist.gov/srmors/view_detail.cfm?srm=1953.
(12) Zhu, J.; Lin, Y.-H.; Dingess, K. A.; Mank, M.; Stahl, B.; Heck, A. J. R. Quantitative Longitudinal Inventory of the N-Glycoproteome of Human Milk from a Single Donor Reveals the Highly Variable Repertoire and Dynamic Site-Specific Changes. *J. Proteome Res.* **2020**, *19*, 1941−1952.
(13) Polasky, D. A.; Yu, F.; Teo, G. C.; Nesvizhskii, A. I. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods* **2020**, *17*, 1125−1132.
(14) Huang, J.; Guerrero, A.; Parker, E.; Strum, J. S.; Smilowitz, J. T.; German, J. B.; Lebrilla, C. B. Site-Specific Glycosylation of Secretory Immunoglobulin A from Human Colostrum. *J. Proteome Res.* **2015**, *14*, 1335−1349.
(15) NIST MS Search Software. 2020, https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:downloads:start#nist_ms_search_program.
(16) NIST GADS & Tandem MS Libraries of Human MIlk Glycopeptides. 2022, https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:human_n_linked_glyco.
(17) NIST Libraries of Peptide Tandem Mass Spectra. 2020, https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:start.
(18) Bateman, A.; Martin, M. J.; Orchard, S.; et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2020**, *49*, D480−D489.
(19) Plomp, R.; de Haan, N.; Bondt, A.; Murli, J.; Dotz, V.; Wuhrer, M. Comparative Glycomics of Immunoglobulin A and G From Saliva and Plasma Reveals Biomarker Potential. *Front. Immunol.* **2018**, *9*, No. 2436.
(20) de Haan, N.; Falck, D.; Wuhrer, M. Monitoring of immunoglobulin N- and O-glycosylation in health and disease. *Glycobiology* **2020**, *30*, 226−240.
(21) Karav, S.; German, J. B.; Rouquié, C.; Le Parc, A.; Barile, D. Studying Lactoferrin N-Glycosylation. *Int. J. Mol. Sci.* **2017**, *18*, No. 870.
(22) Zlatina, K.; Galuska, S. P. The N-glycans of lactoferrin: more than just a sweet decoration. *Biochem. Cell Biol.* **2021**, *99*, 117−127.
(23) Barboza, M.; Pinzon, J.; Wickramasinghe, S.; Froehlich, J. W.; Moeller, I.; Smilowitz, J. T.; Ruhaak, L. R.; Huang, J.; Lönnerdal, B.; German, J. B.; Medrano, J. F.; Weimer, B. C.; Lebrilla, C. B. Glycosylation of human milk lactoferrin exhibits dynamic changes during early lactation enhancing its role in pathogenic bacteria-host interactions. *Mol. Cell. Proteomics* **2012**, *11*, No. M111.015248.
(24) Yu, T.; Guo, C.; Wang, J.; Hao, P.; Sui, S.; Chen, X.; Zhang, R.; Wang, P.; Yu, G.; Zhang, L.; Dai, Y.; Li, N. Comprehensive characterization of the site-specific N-glycosylation of wild-type and recombinant human lactoferrin expressed in the milk of transgenic cloned cattle. *Glycobiology* **2011**, *21*, 206−224.
(25) Zauner, G.; Selman, M. H. J.; Bondt, A.; Rombouts, Y.; Blank, D.; Deelder, A. M.; Wuhrer, M. Glycoproteomic Analysis of Antibodies*. *Mol. Cell. Proteomics* **2013**, *12*, 856−865.
(26) Zhu, J.; Dingess, K. A.; Mank, M.; Stahl, B.; Heck, A. J. R. Personalized Profiling Reveals Donor- and Lactation-Specific Trends in the Human Milk Proteome and Peptidome. *J. Nutr.* **2021**, *151*, 826−839.
(27) Zhao, Y.; Shang, Y.; Ren, Y.; Bie, Y.; Qiu, Y.; Yuan, Y.; Zhao, Y.; Zou, L.; Lin, S.-H.; Zhou, X. Omics study reveals abnormal alterations of breastmilk proteins and metabolites in puerperant women with COVID-19. *Signal Transduction Targeted Ther.* **2020**, *5*, No. 247.
(28) Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **2020**, *369*, 330−333.
(29) Kim, B. J.; Dallas, D. C. Systematic examination of protein extraction, proteolytic glycopeptide enrichment and MS/MS fragmentation techniques for site-specific profiling of human milk N-glycoproteins. *Talanta* **2021**, *224*, No. 121811.
(30) Čaval, T.; Zhu, J.; Heck, A. J. R. Simply Extending the Mass Range in Electron Transfer Higher Energy Collisional Dissociation Increases Confidence in N-Glycopeptide Identification. *Anal. Chem.* **2019**, *91*, 10401−10406.
(31) Wallace, W. E.; Ji, W.; Tchekhovskoi, D. V.; Phinney, K. W.; Stein, S. E. Mass Spectral Library Quality Assurance by Inter-Library Comparison. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 733−738.
(32) Kawahara, R.; Chernykh, A.; Alagesan, K.; Bern, M.; Cao, W.; Chalkley, R. J.; Cheng, K.; Choo, M. S.; Edwards, N.; Goldman, R.; Hoffmann, M.; Hu, Y.; Huang, Y.; Kim, J. Y.; Kletter, D.; Liquet, B.; Liu, M.; Mechref, Y.; Meng, B.; Neelamegham, S.; Nguyen-Khuong, T.; Nilsson, J.; Pap, A.; Park, G. W.; Parker, B. L.; Pegg, C. L.; Penninger, J. M.; Phung, T. K.; Pioch, M.; Rapp, E.; Sakalli, E.; Sanda, M.; Schulz, B. L.; Scott, N. E.; Sofronov, G.; Stadlmann, J.; Vakhrushev, S. Y.; Woo, C. M.; Wu, H.-Y.; Yang, P.; Ying, W.; Zhang, H.; Zhang, Y.; Zhao, J.; Zaia, J.; Haslam, S. M.; Palmisano, G.; Yoo, J. S.; Larson, G.; Khoo, K.-H.; Medzihradszky, K. F.; Kolarich, D.; Packer, N. H.; Thaysen-Andersen, M. Community evaluation of glycoproteomics informatics solutions reveals high-performance search strategies for serum glycopeptide analysis. *Nat. Methods* **2021**, *18*, 1304−1316.
(33) Nasir, W.; Toledo, A. G.; Noborn, F.; Nilsson, J.; Wang, M.; Bandeira, N.; Larson, G. SweetNET: A Bioinformatics Workflow for Glycopeptide MS/MS Spectral Analysis. *J. Proteome Res.* **2016**, *15*, 2826−2840.
(34) Toghi Eshghi, S.; Shah, P.; Yang, W.; Li, X.; Zhang, H. GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides. *Anal. Chem.* **2015**, *87*, 5181−5188.
(35) Zeng, W.-F.; Cao, W.-Q.; Liu, M.-Q.; He, S.-M.; Yang, P.-Y. Precise, Fast and Comprehensive Analysis of Intact Glycopeptides and Monosaccharide-Modifications with pGlyco3. *bioRxiv* **2021**, *0*, No. 430063.
(36) van Veen, H. A.; Geerts, M. E.; van Berkel, P. H.; Nuijens, J. H. The role of N-linked glycosylation in the protection of human and bovine lactoferrin against tryptic proteolysis. *Eur. J. Biochem.* **2004**, *271*, 678−684.
(37) Karav, S.; Casaburi, G.; Arslan, A.; Kaplan, M.; Sucu, B.; Frese, S. N-glycans from human milk glycoproteins are selectively released by an infant gut symbiont in vivo. *J. Funct. Foods* **2019**, *61*, No. 103485.
(38) Dallas, D. C.; Martin, W. F.; Strum, J. S.; Zivkovic, A. M.; Smilowitz, J. T.; Underwood, M. A.; Affolter, M.; Lebrilla, C. B.; German, J. B. N-Linked Glycan Profiling of Mature Human Milk by High-Performance Microfluidic Chip Liquid Chromatography Time-of-Flight Tandem Mass Spectrometry. *J. Agric. Food Chem.* **2011**, *59*, 4255−4263.

https://doi.org/10.1021/acs.jproteome.2c00286
J. Proteome Res. XXXX, XXX, XXX−XXX