Journal of proteome research

pubs.acs.org/jpr

Article

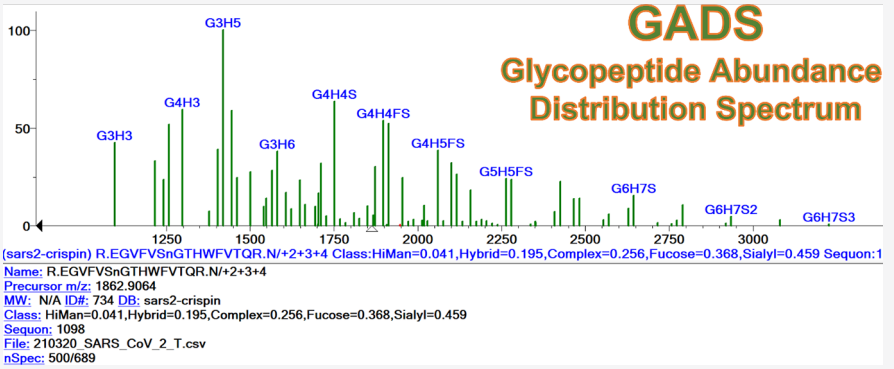
Representing and Comparing Site-Specific Glycan Abundance Distributions of Glycoproteins

Concepcion A. Remoroza, Meghan C. Burke, Yi Liu, Yuri A. Mirokhin, Dmitrii V. Tchekhovskoi, Xiaoyu Yang, and Stephen E. Stein*

Cite This: https://doi.org/10.1021/acs.jproteome.1c00442

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: A method for representing and comparing distributions of N-linked glycans located at specific sites on proteins is presented. The representation takes the form of a simple mass spectrum for a given peptide sequence, with each peak corresponding to a different glycopeptide. The mass (in place of m/z) of each peak is that of the glycan mass, and its abundance corresponds to its relative abundance in the electrospray MS^1 spectrum. This provides a facile means of representing all identifiable glycopeptides arising from a single protein "sequon" on a specific sequence, thereby enabling the comparison and searching of these distributions as routinely done for mass spectra. Likewise, these reference glycopeptide abundance distribution spectra (GADS) can be stored in searchable libraries. A set of such libraries created from available data is provided along with an adapted version of the widely used NIST-MS library-search software. Since GADS contain only MS^1 abundances and identifications, they are equally suitable for expressing collision-induced fragmentation and electron-transfer dissociation determinations of glycopeptide identity. Comparisons of GADS for N-glycosylated sites on several proteins, especially the SARS-CoV-2 spike protein, demonstrate the potential reproducibility of GADS and their utility for comparing site-specific distributions.

KEYWORDS: *N-glycopeptides, site-specific glycosylation, NIST-MS search, SARS*

INTRODUCTION

The variable and complex distribution of N-linked glycans at specific protein sites has been of interest for some time.[1,2] With the increasing sensitivity and diversity of mass spectral measurement methods and the recognized importance of the complex glycan shields on external viral proteins, they have been the subject of numerous recent studies.[3−5] Of current special concern is the spike protein on the SARS-CoV-2 virus, where each of its 22 N-glycosylated site carries a specific distribution.[6,7] A complete characterization of these distributions is required to fully characterize the composition of any glycoprotein and compare samples since these profiles can depend on protein origin and processing.

At present, these site-specific glycan distributions cannot be readily compared. They are currently represented as groups of defined classes[8] or simply as listings of the identified glycans for each site.[6] With over 150 different known compositions of human N-linked glycans, each with a characteristic relative abundance and dozens of different glycans detectable on some sites, this representation is inadequate for an accurate description or comparison of these distributions. This paper reports a means of representing and comparing such distributions in a mass spectral-like format, termed here glycopeptide abundance distribution spectra (GADS). We note that N-glycans are attached to the side chain of an asparagine separated at the C-terminus from serine or

Received: May 24, 2021

Table 1. GADS Library Names, Glycoproteins, and Mass Spectral Datasets Used in the Study[a]

library name	protein[b]	proteases	fragmentation technique	data source[c]	reference
ace2-ccrc	ACE2_HUMAN Angiotensin-converting enzyme 2	Tryp, chymo, alP	HCD	PXD019937	18
ace2-parastoo	ACE2_HUMAN Angiotensin-converting enzyme 2	Try, chymo, Glu-C	HCD, IT-FT	GPST0001110	19
ebv-seattle	Epstein−Barr virus, EBV-gp42, EBV-gh, EBV-gl	Tryp, chymo	EThcD	PXD006403	20
ha-5B8-zaia	hemagglutinin SWZHA_5B8lA/Switzerland	Tryp, chymo	HCD	PXD018920	21
ha-phil82gal-zaia	hemagglutinin AFG99160lA/Phil/2/1982	Tryp, chymo	HCD	PXD018920	21
ha-phil82-zaia	hemagglutinin AFG99160lA/Phil/2/1982	Tryp, chymo	HCD	PXD018920	21
ha-swz-zaia	hemagglutinin SWZHAlA/Switzerland	Tryp, chymo	HCD	PXD018920	21
hkui-s-crispin	HKU1-S	Tryp, Chymo	HCD	MSV000084993	5
hku1-s-melbourne	HCoV-HKU1-S	Tryp	HCD	PXD019163	22
hku1-s-utrecht	HKU1-S HE F11	Tryp, chymo,alP	EThcD	PXD017545	23
mers-cdc	K0BRG7 9BETC spike glycoprotein	Tryp, Chymo, Asp-N, Lys-C	EThcD		12
mers-seattle	MERS Hek293T	Tryp, chymo,alP, Asp-N	EThcD	PDX010494	24
nl63-seattle	HCoV-NL63	Tryp, chymo	EThcD	PXD004557	25
rbd-sars1−2-utrecht	SARS-CoV-1&2 GCN4 TEV mOrange2 Twinstrep	Tryp, Glu-C	EThcD	PXD021825	26
rbd-sars2-chengdu	SARS-CoV spike glycoprotein	Tryp, Glu-C	HCD	PXD018506	27
rbd-sars2-insect-chengdu	P2019 CoV spike RBD insect	Tryp, Glu-C	HCD	PXD018506	27
s1s2-parastoo	SARS-CoV-2 spike Subunits 1 &2	Tryp, chymo	HCD, IT-FT	GPST000067	8
s1-sars2-cdc	2019 nCoV S-Protein (PDBID:6VSB, McLellan lab)	Tryp, Chymo, Asp-N, Lys-C	EThcD		12
s1-sars2-chengdu	P2019 CoV spike glycoprotein	Tryp, Glu-C	HCD	PXD018506	27
sars1-crispin	SARSCOV1	Tryp, Chymo, alP	HCD	MSV000084993	5
sars1-insect-cdc	SARS-CoV spike (S) protein-recombinant from baculovirus NR-686	Tryp, Chymo, Asp-N, Lys-C	EThcD		12
sars1-seattle	SARS-S	Tryp, chymo, alP, Asp-N	EThcD	PXD010494	24
sars2-beijing	P0DTC2 SPIKE_SARS2 spike glycoprotein	Tryp, chymo, alP, Glu-C, Lys-C	EThcD	PXD023346	28
sars2-ccrc	SARS-CoV-2 spike	Tryp, chymo, alP, Asp-N	HCD	PXD019937	18
sars2-crispin	SARSCOV2	Tryp, Chymo, alP	HCD	MSV000085202	6
sars2-georgetown	SARS-CoV-2 spike	Tryp, Glu-C	EThcD	JPST001013	13
sars2-insect-cdc	QHD43416.1 surface glycoprotein SARS-CoV-2	Tryp, Chymo, Asp-N, Lys-C	EThcD		12
sars2-insect-chengdu	P2019_CoV spike glycoprotein insect	Tryp, GluC	HCD	PXD018506	27
sars2-insect-zhou	P2019 SPIKE COVID spike glycoprotein	Tryp, chymo	HCD	IPX0002212001	7
sars2-melbourne	SARS-CoV-2 spike	Tryp	HCD	PXD019163	22
sars2-n282q-insect-beijing	P0DTC2 SPIKE_SARS2 spike glycoprotein insect	Tryp, chymo, alP, Glu-C, Lys-C	EThcD	PXD023346	28
sars2-s2p-cdc	2019 nCoV S-Protein (PDBID:6VSB, McLellan lab)	Tryp, Chymo, Asp-N, Lys-C	EThcD		12
sars2-wt-beijing	P0DTC2 SPIKE_SARS2 spike glycoprotein	Tryp, chymo, alP, Glu-C, Lys-C	EThcD	PXD023346	28

[a]HA = Hemagglutinin; SARS = severe acute respiratory syndrome. [b]Protein name used in the FASTA files provided by the authors. [c]Public Data Repositories: ProteomeXchange Consortium (PXD), Mass Spectrometry Interactive Virtual Environment (MSV), GlycoPost (GPST), Integrated Proteome Resources (IPX), Japan Proteome Standard (JPST). Tryp = Trypsin; Chymo-Chymotrypsin; alP-alpha-lytic; MERS = Middle East Respiratory Syndrome; RBD = Receptor Binding Domain.

threonine by one amino acid; the sequence of three amino acids is termed a "sequon".

The proposed format is a simple modification of a widely used mass spectral format for representing individual mass spectra. Instead of each peak representing a fragment ion by its m/z and abundance, the new format represents each glycopeptide by the mass of its N-glycan and the abundance of its precursor glycopeptide ion. Since each glycan composition can be assigned a unique mass, comparison of GADS by mass will inherently compare glycans of matching composition, regardless of the glycopeptide sequence or charge state involved. Importantly, this data format is identical in form to that used by library-search programs,[9] which can be adapted to efficiently process and compare GADS.

Such "spectra" enable the full representation of an experimentally determined distribution for individual glycopeptides containing a given peptide sequence. Distributions may be combined, and differences in distributions are examined with the same methods and scoring presently used for chemical identification by mass spectral matching. Mirror plots clearly show differences in relative abundances of each glycopeptide. Comparing N-linked glycan distributions at different locations in a protein and the same sequon on a protein from other sources and variants is also straightforward.

In the present work, GADS were derived from raw data files held in repositories,[10] contributed by others, and measured in our own laboratories. Glycopeptide identifications, abundances, and retention times were extracted from Byonic (Protein Metrics, Cupertino, CA) results and converted to an extended

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

".msp" format using in-house software. We note that we have found that other software tools for glycopeptide identification are equally compatible with the methods described in the paper.

This paper provides a link to a modified version of a widely used mass spectral library-search program developed for use in mass spectral matching of electron ionization, small molecule tandem, and peptide spectra.[9] A spectral comparison function for comparing "incomplete" (partial) spectra was added. The software aids the comparison of GADS with low coverage to higher quality library representations. New data fields and peak (glycopeptide) labeling and coloring are used to distinguish lower from higher confidence peaks. This software is provided with 33 GADS libraries, each containing one protein.

METHODS

Data

Specific sources of mass spectra used to generate libraries are listed in Table 1. All were measured at other laboratories, each using their own methods and protein sources. All were acquired by dynamic data acquisition (DDA) of digests of recombinant proteins by electrospray liquid chromatography−mass spectrometry (LC−MS) analysis. Glycopeptide identifications were made from their product ion spectra, generated by either collision-induced fragmentation (CID/HCD)[11] alone or in combination with electron-transfer dissociation (ETD) using the electron-transfer/higher-energy collision dissociation (EThcD)[12,13] protocol. In all cases, abundances of glycopeptide ions are acquired from extracted ion chromatograms of their precursor electrospray generated ions.

Some illustrative results from digests of human proteins tenascin and alpha-1-acid glycoprotein from Sigma-Aldrich (St. Louis, MO, USA) and the SARS-CoV-2 RBD segment (RayBiotech, Atlanta, GA) were obtained at NIST on an Orbitrap mass spectrometer (Thermo Scientific Orbitrap Fusion Lumos) using a C18 column (Acclaim PepMap RSLC 75 μm × 150 mm, Thermo Fisher Scientific) on a nanoflow LC for separation. Detailed sample preparation and sequences of these glycoproteins can be found in Supplementary Section S1, Figures S1 and S5.

Glycopeptide Identification

Due to the diversity of glycans potentially attached to a site, special methods are required for glycopeptide identification. In this paper, identification, retention time, and abundance values were extracted from the output of Byonic and Byologic software (version 3.10 Protein Metrics, Inc.). The Byonic parameters were set to full specificity with up to three missed cleavages, fixed cysteine carbamidomethylation, and allowed possible methionine oxidation and N-term glutamine deamination with a library of 173 human N-glycans (no multiple fucosylation). Precursor and product ion m/z tolerance settings were 10 and 20 ppm, respectively.

Retention Time Constraints

It has been well documented that different glycopeptides having the same peptide sequence and a fixed number of sialic acids elute over a narrow retention time range in reversed-phase LC columns.[14−16] Each sialic acid group increases retention by substantial amounts (typically by 10 min in a 90-min gradient). The present analysis used this behavior to mark as suspect glycopeptides that elute well outside of expected retention times. Glycopeptides for a given sequence and run are divided into groups according to the number of sialic acid groups, ranging from 0 to 4. If the retention time deviation of a glycopeptide from the median retention time of its group is greater than a specified amount, that GADS peak is annotated with a symbol > or < followed by the deviation in minutes. The allowed deviation in the present work is 4 min. Furthermore, all such out-of-range identifications with a score below a set threshold value (currently a Byonic score of 100) are rejected. This may be further refined in the future, especially for well-characterized distributions. Further improvements may be made in the future by narrowing tolerances by using the clear but small trend of decreasing retention with an increasing glycan mass.

Glycan Abundance Distribution Spectra (GADS)

Each peak in a GADS represents an identified glycopeptide containing a single occupied N-glycan "sequon" in a specific charge state in a single data file. GADS are also derived by combining charge states for individual GADS in a single file. Consensus spectra for each peptide ion at a given charge state in a library are also derived. The mass for each peak is that of the glycan, and the abundance is that of the identified glycopeptide ion (derived from its extracted ion chromatogram, XIC, and normalized to the largest peak in each GADS). Each peak in a spectrum contains annotation details, including glycan composition using a compact representation, which will be described later. Peaks for glycopeptides whose retention times are out of range (see the previous section) are labeled as uncertain and shown in red. Peaks whose scores are not greater than a set threshold (a value of 30 is used for 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.

The following simple quality score was found to be a useful measure for rejecting low quality GADS.

$$\text{Quality} = \text{Ng} \times (\text{Ng}/\text{Nt}) \quad (1)$$

It simply reduces the number of good glycopeptide peaks by their fraction of all peaks.

Here, Ng is the number of identified good quality glycopeptides (score above threshold and within retention tolerance), and Nt is the total number of identified glycopeptides (peaks). Presently, a quality of 3 is required for inclusion in a library. A robust measure of depth of coverage, the maximum-to-median abundance, is reported.

Each GADS or each peak is accompanied with a number of relevant values, including abundance fractions of common subclasses of glycans as described in Supplementary Section S2. Glycopeptides containing multiple sequons are excluded.

Data Analysis Software

A program was written in "C" to generate GADS ".msp" files as input for library creation using the freely available Lib2NIST program.[17] This program extracts data from Byologic output .csv files (abundances are taken as "XIC area summed," and retention time "Apex Time (Posit)." GADS generation is done entirely algorithmically).

Search and User Interface Software

NISTMS.exe software, widely distributed for use with mass spectral reference collections from NIST and other sources, has been adapted to import, view, search, and compare GADS. Without modification, it can create libraries from GADS that

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

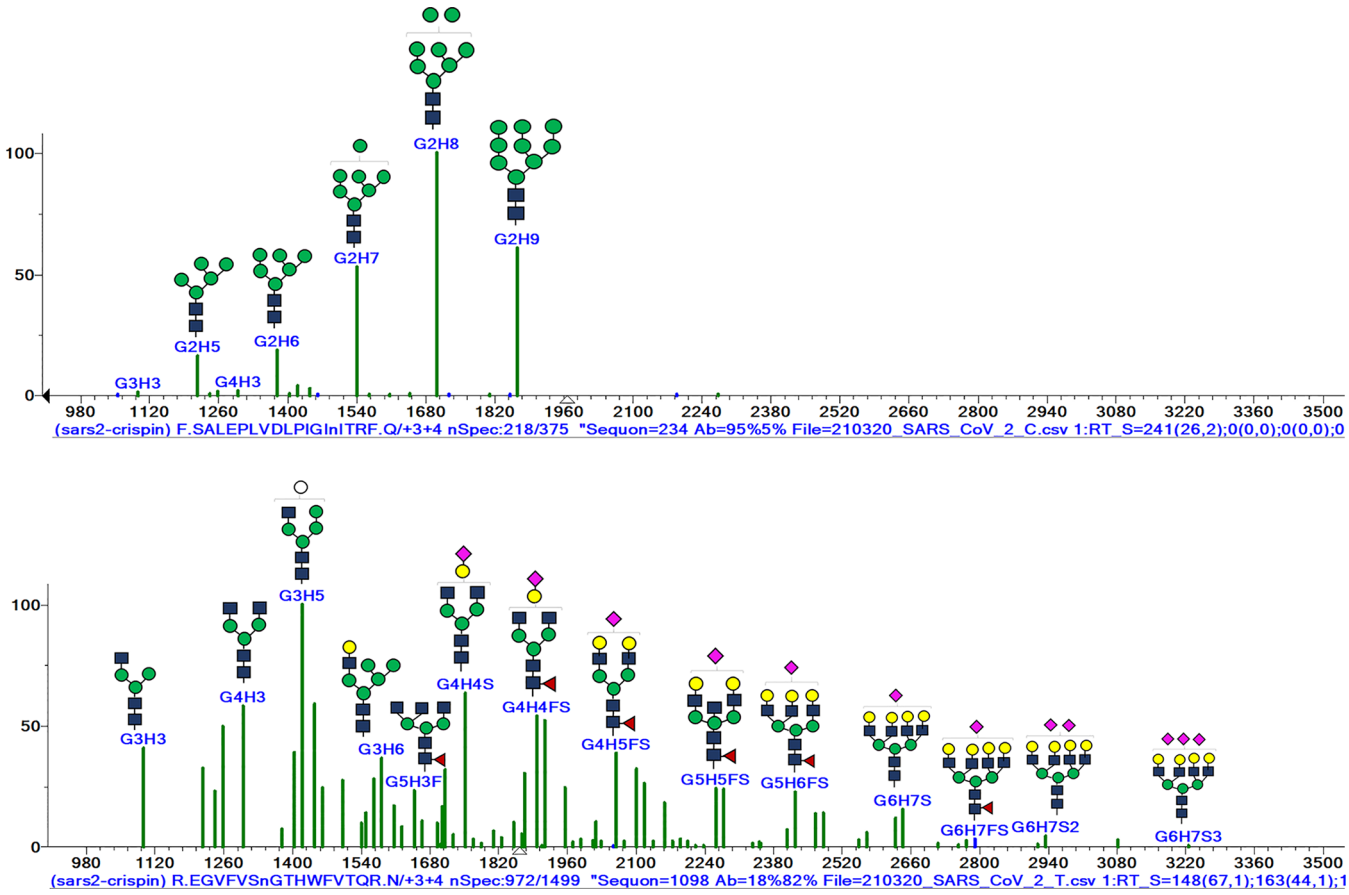


Figure 1. GADS illustration: high-mannose (top) and complex (bottom) N-linked glycans on SARS-CoV-2 N234 and N1098 sequons.[6] Green peaks have scores and retention times within thresholds (see the text). Glycan representation: G = GlcNAc (dark blue square), H = Hexose (circle, mannose-green, and galactose-yellow), F = Fucose (red triangle), and S = Sialic acid (NeuAc, pink diamond). Library name, sequences, and charges are given in the caption below GADS, along with numbers of identifed spectra, the sequon site number, and the file name. Ab = percentage abundance of tryptic peptide with two different charge states, 18% (+3) and 82% (+4). The remaining text is the beginning of retention annotation. The "cartoon" glycan representations above labeled peaks were manually entered. All GADS and text representations of glycans in this paper are sceen captures. In the program, glycan identities can be viewed by expanding the mass region near the peak. They are also shown in a text representation associated with each GADS.

can be searched and displayed in various ways. Using a range of constraints, the software can also generate a variety of comparison scores. For example, one can find similar GADS in many libraries, or matching peptide sequences, sequons, or subsequences. The "Names" tab enables browsing different varieties of GADS in various ways by name or "synonym", and the "Comparison" tab allows the mass (glycan) aligned comparison of multiple GADS. The "Librarian" tab allows library building and editing.

A new comparison function was developed to compare GADS acquired at a lower level of sensitivity, hence with fewer peaks, to a library entry with more peaks (a greater level of coverage). The "partial spectrum score" (PSS) provided by this function ignores peaks in library entries that are not present in query GADS (examples given in the Results section).

RESULTS

GADS Representation

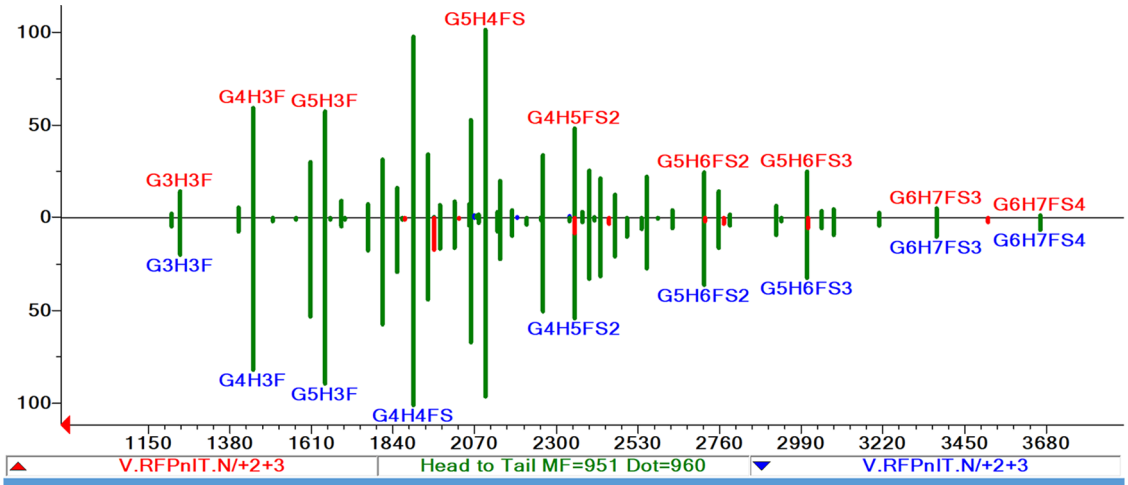
Figure 1 illustrates two GADS, one for a relatively simple and another for a complex distribution, both derived from data provided in ref 6 for the SARS-CoV-2 spike protein. This format employs a compact text representation of glycan composition for concise glycan display as described in the legend. N-linked glycans are attached to asparagine residues represented as "n".

Analytical Variability

In this section, variations in GADS derived from a single protein digest will be examined, the most straightforward measure of which is their variation upon repeated analysis of the same digest solution. In this case, differences arise from stochastic sampling and MS^1 ion intensity variations inherent in electrospray DDA LC experiments. As expected, for abundant peptide sequences, we find a high level of reproducibility. This is illustrated in Supplementary Section S1 Figure S1 for three repeat runs of a single sequence derived from data measured at NIST for a tryptic peptide from tenascin. In general, we find that GADS with quality scores (eq 1) above 3 generate dot product match scores above 950 in repeat runs. These are often derived from over 100 identified MS^2 spectra with a maximum-to-median abundance ratio of 10 or more. For complex cases, over 1000 MS^2 spectra can underlie a single GADS. The numbers are reported in the annotation for each GADS, as are the spectral counts for each glycopeptide "peak".

Separate digestions of the same sample generally produce GADS with variations similar to those of replicate runs of the

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX



#	Library	Score	DotProd	Rev-Dot	PSS-Dot	Name	Denaturant
1	rbd-nist	999	999	999	999	V.RFPnIT.N/+2+3	urea
2	rbd-nist	986	987	998	987	V.RFPnIT.N/+2+3	urea
3	rbd-nist	960	969	971	993	V.RFPnIT.N/+2+3	urea
4	**rbd-nist**	**951**	**960**	**961**	**983**	**V.RFPnIT.N/+2+3**	**SDS-gel**
5	rbd-nist	942	952	955	982	V.RFPnIT.N/+2+3	SDS-gel

Figure 2. GADS mirror-plot comparison for a SARS-CoV-2 RBD sequence generated by alpha-lytic digestion in solution by urea (top) vs SDS gel (bottom). The lower panel shows "hit list" scores for best matching GADS (lower spectrum corresponds to highlighted entry, #4). Score = Overall score developed for tandem spectra, DotProd = simple "dot-product" score, Rev-Dot = dot product ignores nonmatching query (top) peaks, PSS-Dot = partial spectrum score ignores nonmatching library (bottom) peaks. Blue peaks have glycopeptide identification scores below the set threshold described in the text. Red peaks have retention times out of the expected range and may originate as in-source products or retention time errors. Measurements were done at NIST. Glycan compositions shown are screen captures, and depending on screen resolution and font, glycan text display may require expansion of the region of interest.

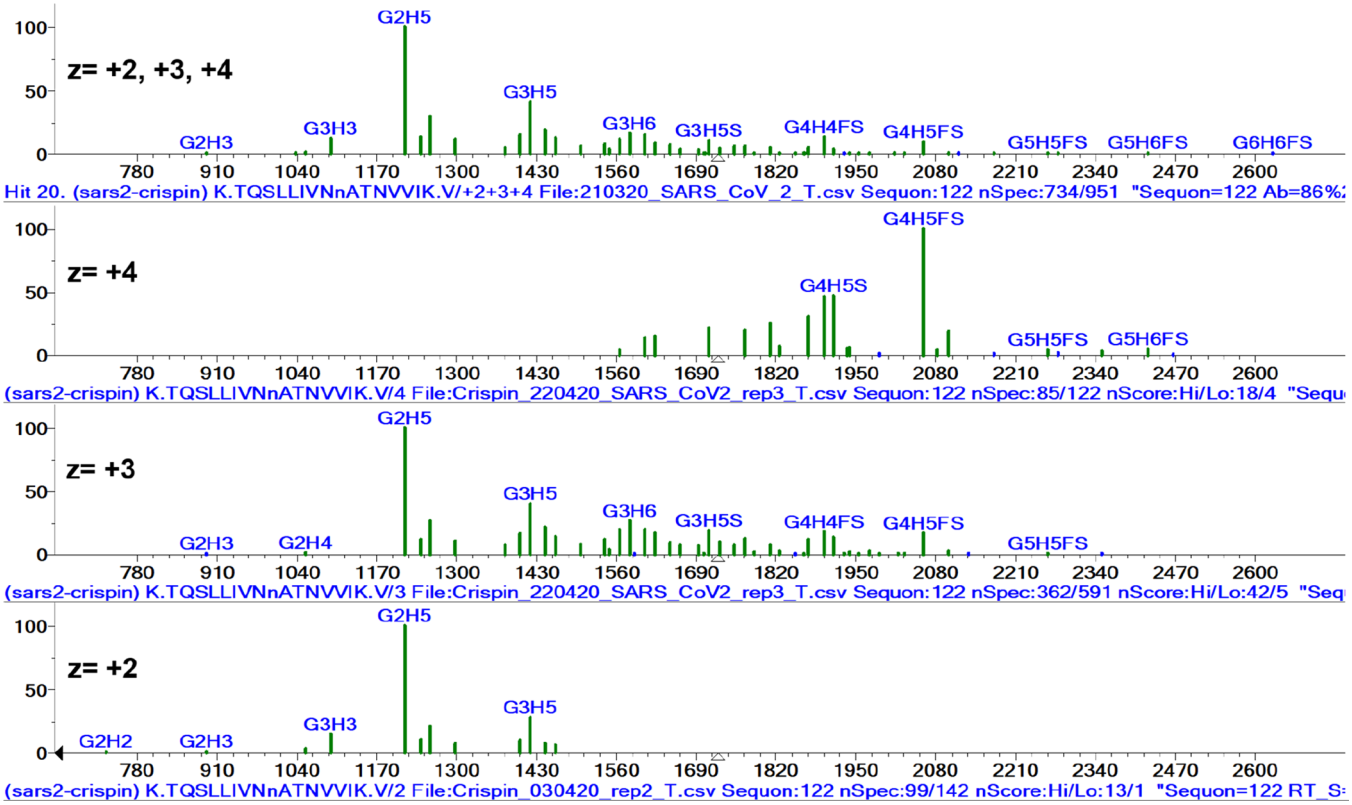


Figure 3. GADS can depend on the charge state (3A). For the tryptic peptide TQSLLIVNnATNVVIK (lower case n denotes asparagine at the point of glycan attachment), in a single digest, from bottom to top different charge states (z) are 2, 3, 4, and combined.

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

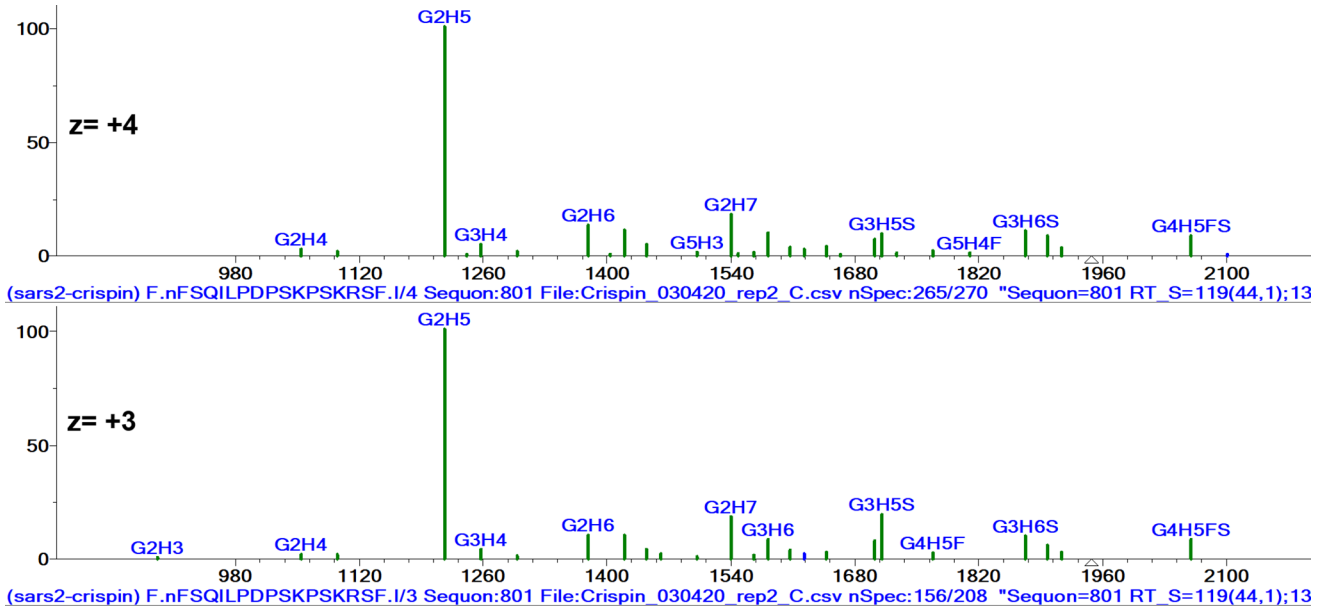
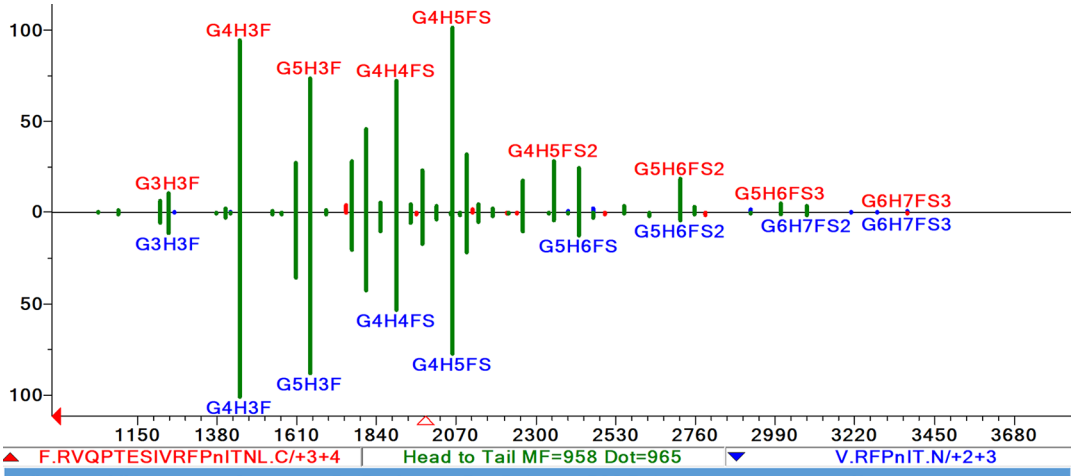


Figure 4. GADS of nFSQILPDPSKPSKRSF for two different charge states (z) from the same digest have similar glycan distributions.



#	Library	Score	DotProd	Rev-Dot	PSS-Dot	Name
1	sars2-crispin	999	999	999	999	F.RVQPTESIVRFPnITNL.C/+3+4
2	sars2-crispin	967	973	974	987	F.RVQPTESIVRFPnITNLCPF.G/+3+4+5
3	sars2-crispin	959	963	989	964	F.RVQPTESIVRFPnITNL.C/+3+4
4	**sars2-crispin**	**958**	**965**	**970**	**971**	**V.RFPnIT.N/+2+3**

Figure 5. GADS should be similar for different peptides containing a sequon. Peptides from two proteases contain the same sequon in the SARS-CoV-2 spike protein. The upper GADS (hit #1 in the hit list shown below the plot) was generated by chymotrypsin. The lower GADS corresponding to hit #4 (highlighted in the hit list) was generated by alpha-lytic WT. Note that hit #2 was also generated by chymotrypsin but with one missed cleavage. The data used are from the SARS-CoV-2 spike protein data set.[6,29] Note that scores shown here (max 999) are widely used to measure spectrum similarity for library searching and identification. The origin of differences in abundance is not known with certainty and will be the subject of future studies.

same digest. For example, in studies of various proteins, our laboratory results showed no systematic differences in GADS attributable to different denaturants, including those derived from urea in-solution and SDS in-gel digestion (Figure 2). A variety of other denaturants also showed no systematic effects on glycan distributions, including organics, guanidine, and even simple heating (Supplementary Section S1 Figure S2). Differences in GADS were primarily a result of differing abundances of the underlying glycopeptides. However, digestion at low pH, such as that required for pepsin digestion, was found to alter the glycosylation pattern significantly.

Significant changes in glycan distribution, especially involving the degree of sialylation, can result from differences in protein handling. An example is shown in Supplementary Section S1 Figure S3, where differences are observed for a protein after being stored in solution at 4 °C for 1 month.

Peptide Ion Variability

Charge state: Peptide sequences containing N-glycans are often identified at multiple charge states, which are generally combined for the final glycopeptide distribution. These different charge states can have quite different distributions of glycans, with higher charge states enriched in higher

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

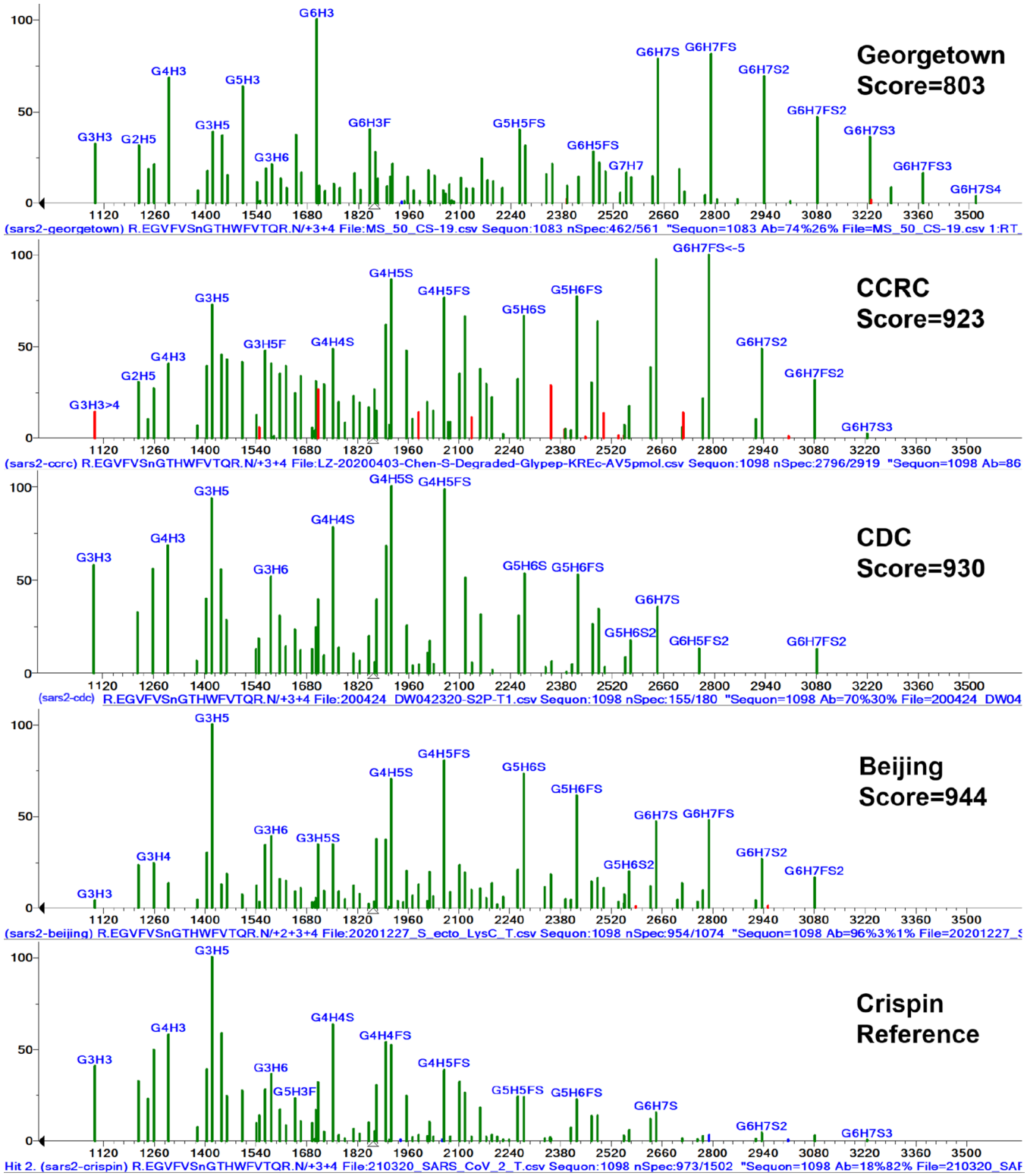


Figure 6. GADS of a SARS-CoV-2 spike tryptic peptide from four laboratories. From bottom to top, laboratories are Crispin (Reference data),[6] Beijing (PXD023346),[28] CDC,[12] CCRC,[18] and Georgetown.[13] Dot product scores comparing GADS from the bottom with each remaining four (in the above order) were 944, 930, 923, and 803. Sources of protein were different for each.

molecular weight glycans. This is illustrated in Figure 3 for cases where distributions differ significantly or little (Figure 4) between charge states.

Peptide Sequence. Assuming that GADS reflect the underlying distribution of glycans attached to a sequon, they should not depend on the specific peptide sequence containing that sequon. In fact, agreement of such GADS is generally observed and provides important confirmation of a correct determination. Figure 5 illustrates this point for GADS derived from peptides of very different lengths. A missed cleavage peptide containing the sequence of first GADS also matches

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

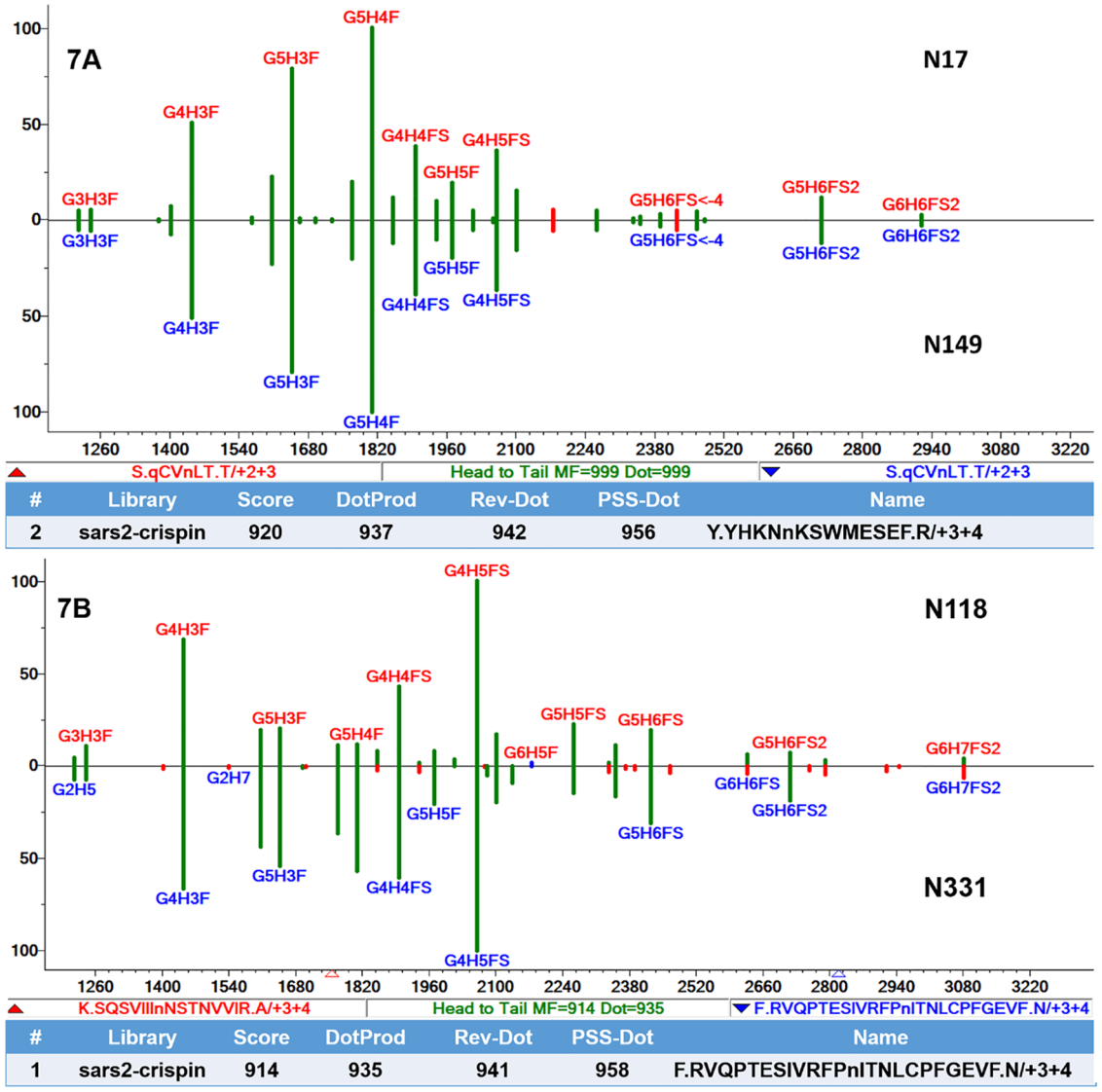


Figure 7. Comparison of GADS for two different sequons: (7A) within the SARS-CoV-2 spike protein[29] for sequons N17 to N149 with scores below comparing lower to upper highlighted at the bottom. (7B) Between two different spike proteins, SARS-CoV-1[5] (sequon N118) and SARS-CoV-2[6] (sequon N331).

with a high score. A perusal of the libraries provided with the GADS system strongly reinforces this finding.

Sample Variability

Biological Replicates. Differences in GADS between protein samples generated under identical conditions set the baseline for expected variations for differences in GADS for proteins generated under conditions where glycosylation might be expected to differ, including proteins from different sources and variants. Triplicate measurements obtained from the Crispin laboratory for the SARS-CoV-2 spike protein[6] have been analyzed to examine this variation. For 42 sequences of 22 sequons observed in multiple replicates, Table S1 reports median dot product scores of 957 and 939 between the first and second replicate and first and third, respectively. For the three replicates, median quality values for individual GADS are 27 or 28, and median numbers of identified glycopeptide spectra range from 225 to 246, with a maximum of 1540. Low values were generally associated with relatively small numbers of identified spectra and quality values and were often raised significantly in value by the new PSS, reflecting a difference in the degree of sampling between the replicates. For the third replicate, GADS for five glycopeptide sequences were not identified due presumably to differences in digestion efficiency for these peptides in that sample.

Interlaboratory Variations. For the multiple laboratories that have determined site-specific profiles for the SARS-CoV-2 spike and related proteins, GADS enable the direct comparison of results. Although each laboratory employs its own protein sources and methods and can have small differences in sequence, some GADS have a high degree of similarity. Two examples are shown in Figure 6, each of which involves two methods of glycopeptide identification, HCD (CID) and EThcD (combined ETD and CID fragmentation).

Intrasequon Similarity

Meta-Heterogeneity. Different sequons on a protein can have similar glycan distributions, a concept discussed as "meta-heterogeneity" in a recent review.[30] Since the same sample is used for such comparisons, accuracy can potentially be very high. The comparison is illustrated in Figure 7A for two SARS2 spike protein sequons (N17 and N149).[6]

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

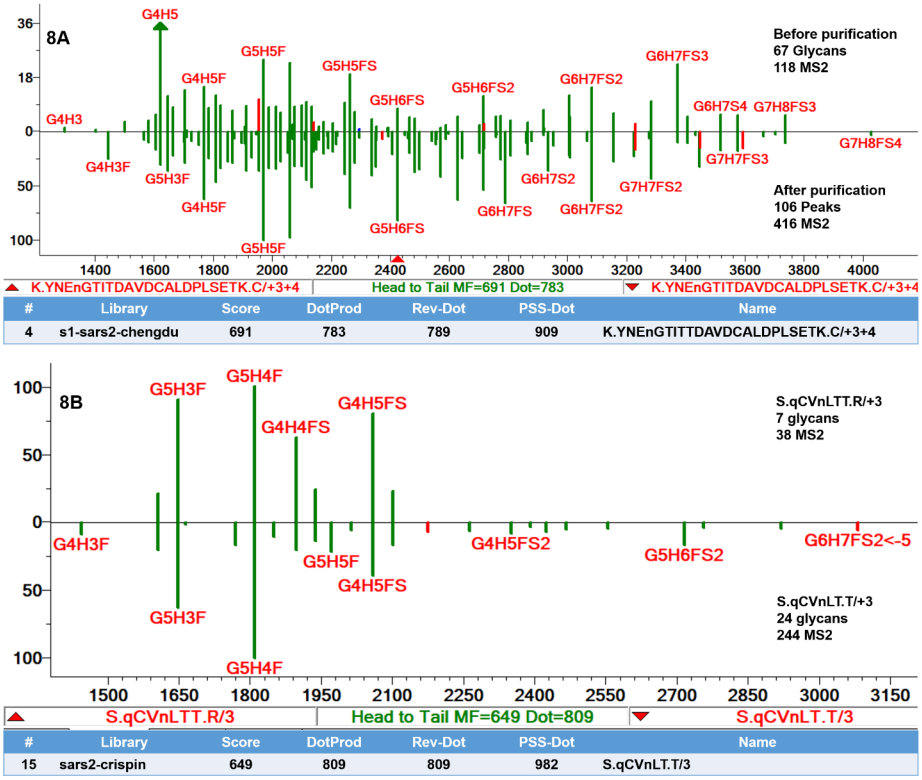


Figure 8. Comparing GADS with different depths of coverage (numbers of peaks). (A) Comparison of GADs before and after enrichment for a digest of a SARS-CoV-2 S1 protein.[27] Purification increases the concentration, leading to more identified glycopeptides. The PPS ignore glycans present in the higher concentration GADS but missing in the lower concentration GADS, thereby producing a high score. This indicates that almost all peaks in the low concentration GADS were also identifiable in the purified sample. (B) Comparison of GADS for two peptides containing the same sequon in the same SARS-CoV-2 digest[6] with one having a missed cleavage at a lower abundance, hence with fewer peaks than the peptide (lower GADS) and compared to DotProd (highlighted hit) with no missed cleavage. The high PSS shows that the two GADS are consistent with a single sequon.

Different Proteins. A variety of GADS libraries have been compiled from various MS data repositories,[10] enabling the facile comparison of GADS from different glycoproteins. Examples of matches between different sequons in the SARS-CoV-1[5] and SARS-CoV-2[6] spike proteins are shown in Figure 7B.

DISCUSSION

Influence of the Charge State and Sequence

As shown in Figure 3, glycopeptides at higher charge states tend to contain N-linked glycans of higher mass, presumably due to the increased charge stabilization provided by the higher mass glycans. Since relative charge state abundances can depend on electrospray conditions, comparisons of GADS are most reliably made with charge-combined spectra. Missing higher or lower charge state glycopeptides can result in missing higher or lower molecular mass glycans. It is important to note that a mass spectral abundance distribution does not directly reflect distributions of glycans on a sequon. This is due to expected differences in response factors for different glycans and charge states as well as possible nonlinearity of ion intensity concentrations.

Except for differences in interferences and relative depth of coverage, different sequences covering the same sequon are found to generally have highly similar GADS (Figure 5). Many such examples of different sequences containing the same sequon from various proteases or numbers of missed cleavages can be seen in the GADS libraries provided with the software.[29] The most significant differences originate from differences in depth of coverage (numbers of glycopeptide peaks).

GADS Properties and Scoring

Since GADS determinations depend on multiple MS^1 ion intensities and MS^2 identifications, they inherently differ in their behavior compared to individual MS^2 spectra in spectral libraries. Since a GADS peak is based on the identification of the underlying glycopeptide, a GADS contaminant is a false MS^2 identification, not a contaminating MS^2 peak common to mass spectra. Consequently, a principal source of GADS variation is the number of individual peaks that it contains, which depends primarily on the ion intensities of the underlying glycopeptides. This, in turn, depends on protein concentration, specific sequence detectability, and digestion efficiency. Therefore, to acquire a good quality GADS for a given sequon, it is necessary to employ a protease with sufficient specificity to generate a highly abundant peptide capable of being identified in an LC/MS experiment.

To assist in matching GADS with varying numbers of peaks, a new scoring function, termed the "partial spectrum" search, has been added to current search methods. It uses the current dot product scoring but simply ignores nonmatching library peaks. In contrast to MS^2 spectra, where major fragment ions at a given energy are virtually never absent, a peak may be absent in a GADS if it was not sampled or did not produce a sufficiently high identification score, generally due to a low signal level or contaminant cofragmentation. This scoring

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

method is, in effect, the inverse of the commonly used "reverse" search, which ignores nonmatching query peaks, by assuming that they may originate from contaminants. This scoring method was described in 1978 as the "forward" score to distinguish it from the "reverse" score.[31] Two instances of the utility of the new search method are shown in Figure 8A and B. One is for the same sequence and sample but with one acquired prior to protein purification and the other afterward. Applying PSS increased the dot product score from 783 to 909. While conventional "dot product" scores were in the low 800s, indicative of only a moderate quality match, PSS were primarily in the high 900s, indicating a near perfect match. A more extreme example is shown for two sequences containing the same sequon but one with and another without a missed cleavage. The former generates a relatively low coverage GADS, leading to a conventional low dot product score of 809 compared to the latter. In contrast, a PSS of 982 indicates consistency with the higher coverage GADS. This score will also provide a reasonable score when matching single charge state GADS with those derived from multiple charge states.

Many examples of scores between GADS measured in replicate digests of the SARS2 spike protein[29] are shown in Supplementary Section S1 Table S1. Dot product values provide a simple measure of overall similarity. The "score" reduces these values for smaller numbers of peaks, as is done by the tandem spectrum score for matching spectra of metabolites. Note the significantly higher values for many reverse scores and PPS, which compensate for different numbers of glycans in the GADS being compared. Since the first replicate serves as the reference spectrum, when the replicate has significantly more peaks, then a higher PSS is generated. Conversely, when the replicate has fewer peaks, the reverse score is higher. In a large fraction of comparisons, one of these two numbers is over 950, indicating a generally very high degree of agreement for these biological replicate measurements.

Limitations and Extensions

As noted in the literature,[14,15] high scores for glycopeptide matching do not guarantee correct identification. This can occur for several reasons, including accidental matching of formulas of modifications and glycan compositions, wrong selection of the C^{12} precursor ion, and isobaric overlap of glycans with different chemical formulas. Random matching of different peptides is reduced using retention time constraints for GADS determinations. Higher mass glycopeptides are inherently prone to misidentification due to the increasing uncertainty with increasing total mass and the increasing difficulty in observing the monoisotopic precursor ion.[32] This problem, which is generally detected by human inspection, can be avoided mainly by confirmation of GADS from multiple peptides containing a sequon. An example is shown in Supplementary Section S1 Figure S4 for a high mass peptide where the loss of ammonia compensates for glycan differences within instrument accuracy (see the caption for details).

Other factors may lead to incomplete GADS for a given sequon. For example, sialylated N-glycan forms of peptides have insufficient solubility to be observed. If a higher charge state and/or late eluting glycopeptides are missed, one may also expect to miss higher mass sialylated N-glycans. In such cases, the partial spectrum scoring can help by ignoring the missed glycans.

Using mass alone to represent a glycan, of course, isomers cannot be represented. However, the present methods are extensible. If glycan isomers can be distinguished, they can be added to the annotation with fractions of each isomer given, even using multiple color schemes. Depending on needs, further scoring or peak representation schemes can be developed based on annotation (for example, high-mannose vs hybrid vs complex).

For reliable GADS comparison, both the query and library GADS should be built from the same N-glycan library. The present system uses a 173-member human N-glycan library.[29] Future work will examine the degree to which this list should be expanded.

These results demonstrate the key importance of signal strength of the glycopeptide ions released by digestion for creating a complete and reproducible GADS. GADS quality depends on several factors, including protein concentration and sample loading, chromatographic separation, instrument sensitivity, and use of the optimal proteases and digestion protocols for peptide release.

CONCLUSIONS

A method has been described for the effective representation and comparison of site-specific distributions of N-linked glycans derived from abundances of glycopeptides derived from glycoproteins. This mass spectrum-like format, called a GADS, represents each identified glycan attached to a given peptide sequence as a peak at the mass of the glycan and the abundance of the glycopeptide. This enables display, comparison, and library storage of glycan distributions associated with a "sequon" on a glycoprotein using available mass spectral library-search software.

By providing a complete and reproducible glycan distribution of a sequon, GADS can enable a facile comparison of site-specific distributions. This can provide a means of monitoring changes in protein glycosylation and documenting differences in two samples of the same protein. It can also establish effects on glycosylation from any alteration in sequence and find similar glycan distributions on different sequons on the same or different proteins. To the degree that distributions remain constant, these libraries can provide a community-wide means of storing reference site-specific glycan distributions of common interest. However, changes in MS1 signal intensities can lead to relative abundances that depend to some degree on specific experimental conditions.

GADS processing software[29] is derived from the widely used NIST-MS library-search software. This is made possible by the high degree of logical similarity of GADS representations with those of reference mass spectra. A new "partial spectrum" GADS score was created to provide a measure of similarity of an acquired GADS having fewer glycan peaks to reference GADS with more complete coverage. Differences in the depth of coverage are a principal factor leading to variations in GADS for the same sequon acquired in different experiments. Moreover, as the abundance of glycopeptides diminishes, not only are fewer glycopeptides identified, but their variability in intensity will tend to increase.

The software enables the facile comparison of any user-derived or library-selected GADS with all others. Included with this software is a set of annotated GADS libraries derived from publicly available data. GADS can be imported and converted to a searchable library format with a straightforward, flexible

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX

text format described in the GADS User's Manual provided in Supplementary Section S2.

Results show that good quality GADS derived from different peptides containing the same sequon on a given sample can be very similar, behaving in a manner similar to mass spectra in current spectral libraries. Moreover, results show that the same sequon for a given protein, in this case from SARS-CoV-2 proteins, generated in different laboratories using different methods for glycopeptide generation and identification (HCD and EThcD) can have very similar GADS. On the other hand, further work characterizing the behavior of GADS is needed to reliably separate analytical from actual glycosylation differences.

In addition to increasing the number and depth glycoproteins represented, future work will focus on improving GADS quality and glycan coverage and employing lower energy fragmentation methods to confirm glycan identity. Extensions to O-linked glycans are also planned.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00442.

(Supplementary Section S1) Supplemental protocol on sample preparation, digestion, and LC−MS/MS analysis; (Figures S1 to S5) GADS illustrations; and (Table S1) GADS scores for three replicate samples (PDF)

(Supplementary Section S2) User Guide for NIST-MS Glycopeptide Abundance Distribution Spectra (GADS) libraries, software, and documentation (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Stephen E. Stein − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, Unites States;* orcid.org/0000-0001-9384-3450; Email: steve.stein@nist.gov

Authors

Concepcion A. Remoroza − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, Unites States;* orcid.org/0000-0003-1540-1635

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

Yi Liu − *Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, Unites States*

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

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

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

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00442

Notes

The authors declare no competing financial interest.

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 intend that the materials, software, or instruments used are necessarily the best available for the purpose.

■ ACKNOWLEDGMENTS

We thank Drs. Oleg Toropov, Zheng Zhang, and Yuxue Liang for the technical support in the study.

■ REFERENCES

(1) Hua, S.; Nwosu, C. C.; Strum, J. S.; Seipert, R. R.; An, H. J.; Zivkovic, A. M.; German, J. B.; Lebrilla, C. B. Site-specific protein glycosylation analysis with glycan isomer differentiation. *Anal. Bioanal. Chem.* **2012**, *403*, 1291−1302.

(2) Ritchie, G.; Harvey, D. J.; Feldmann, F.; Stroeher, U.; Feldmann, H.; Royle, L.; Dwek, R. A.; Rudd, P. M. Identification of N-linked carbohydrates from severe acute respiratory syndrome (SARS) spike glycoprotein. *Virology* **2010**, *399*, 257−269.

(3) Zhu, Z.; Desaire, H. Carbohydrates on Proteins: Site-Specific Glycosylation Analysis by Mass Spectrometry. *Annu. Rev. Anal. Chem.* **2015**, *8*, 463−483.

(4) Cao, L.; Diedrich, J. K.; Kulp, D. W.; Pauthner, M.; He, L.; Park, S.-K. R.; Sok, D.; Su, C. Y.; Delahunty, C. M.; Menis, S.; Andrabi, R.; Guenaga, J.; Georgeson, E.; Kubitz, M.; Adachi, Y.; Burton, D. R.; Schief, W. R.; Yates Iii, J. R.; Paulson, J. C. Global site-specific N-glycosylation analysis of HIV envelope glycoprotein. *Nat. Commun.* **2017**, *8*, 14954.

(5) Watanabe, Y.; Berndsen, Z. T.; Raghwani, J.; Seabright, G. E.; Allen, J. D.; Pybus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B.; Crispin, M. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat. Commun.* **2020**, *11*, 2688.

(6) 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.

(7) Zhou, D.; Tian, X.; Qi, R.; Peng, C.; Zhang, W. Identification of 22 N-glycosites on spike glycoprotein of SARS-CoV-2 and accessible surface glycopeptide motifs: Implications for vaccination and antibody therapeutics. *Glycobiology* **2021**, *31*, 69−80.

(8) Shajahan, A.; Supekar, N. T.; Gleinich, A. S.; Azadi, P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology* **2020**, *30*, 981−988.

(9) NIST MS Search Software version 2.3 https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nistlibs.

(10) Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; Del-Toro, N.; Rurik, M.; Walzer, M. W.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcaíno, J. A. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* **2016**, *13*, 651−656.

(11) 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.

(12) Wang, D.; Baudys, J.; Bundy, J. L.; Solano, M.; Keppel, T.; Barr, J. R. 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*, 14730−14739.

(13) Sanda, M.; Morrison, L.; Goldman, R. N-and O-Glycosylation of the SARS-CoV-2 Spike Protein. *Anal. Chem.* **2021**, *93*, 2003−2009.
(14) Klein, J.; Zaia, J. Relative Retention Time Estimation Improves N-Glycopeptide Identifications by LC−MS/MS. *J. Proteome Res.* **2020**, *19*, 2113−2121.
(15) Wang, B.; Tsybovsky, Y.; Palczewski, K.; Chance, M. R. 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*, 729−741.
(16) Ang, E.; Neustaeter, H.; Spicer, V.; Perreault, H.; Krokhin, O. Retention Time Prediction for Glycopeptides in Reversed-Phase Chromatography for Glycoproteomic Applications. *Anal. Chem.* **2019**, *91*, 13360−13366.
(17) Lib2NIST Library Conversion Tool https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nist17#lib2nist_library_conversion_tool. 2020.
(18) Zhao, P.; Praissman, J. L.; Grant, O. C.; Cai, Y.; Xiao, T.; Rosenbalm, K. E.; Aoki, K.; Kellman, B. P.; Bridger, R.; Barouch, D. H.; Brindley, M. A.; Lewis, N. E.; Tiemeyer, M.; Chen, B.; Woods, R. J.; Wells, L. Virus-Receptor Interactions of Glycosylated SARS-CoV-2 Spike and Human ACE2 Receptor. *Cell Host Microbe* **2020**, *28*, 586−601.e6.
(19) Shajahan, A.; Archer-Hartmann, S.; Supekar, N. T.; Gleinich, A. S.; Heiss, C.; Azadi, P. Comprehensive characterization of N- and O-glycosylation of SARS-CoV-2 human receptor angiotensin converting enzyme 2. *Glycobiology* **2021**, *31*, 410−424.
(20) Snijder, J.; Ortego, M. S.; Weidle, C.; Stuart, A. B.; Gray, M. D.; McElrath, M. J.; Pancera, M.; Veesler, D.; McGuire, A. T. An Antibody Targeting the Fusion Machinery Neutralizes Dual-Tropic Infection and Defines a Site of Vulnerability on Epstein-Barr Virus. *Immunity* **2018**, *48*, 799−811.e9.
(21) Chang, D.; Hackett, W. E.; Zhong, L.; Wan, X.-F.; Zaia, J. Measuring Site-specific Glycosylation Similarity between Influenza a Virus Variants with Statistical Certainty. *Mol. Cell. Proteomics* **2020**, *19*, 1533−1545.
(22) Juno, J. A.; Tan, H.-X.; Lee, W. S.; Reynaldi, A.; Kelly, H. G.; Wragg, K.; Esterbauer, R.; Kent, H. E.; Batten, C. J.; Mordant, F. L.; Gherardin, N. A.; Pymm, P.; Dietrich, M. H.; Scott, N. E.; Tham, W.-H.; Godfrey, D. I.; Subbarao, K.; Davenport, M. P.; Kent, S. J.; Wheatley, A. K. Humoral and circulating follicular helper T cell responses in recovered patients with COVID-19. *Nat. Med.* **2020**, *26*, 1428−1434.
(23) Hurdiss, D. L.; Drulyte, I.; Lang, Y.; Shamorkina, T. M.; Pronker, M. F.; van Kuppeveld, F. J. M.; Snijder, J.; de Groot, R. J. Cryo-EM structure of coronavirus-HKU1 haemagglutinin esterase reveals architectural changes arising from prolonged circulation in humans. *Nat. Commun.* **2020**, *11*, 4646.
(24) Walls, A. C.; Xiong, X.; Park, Y.-J.; Tortorici, M. A.; Snijder, J.; Quispe, J.; Cameroni, E.; Gopal, R.; Dai, M.; Lanzavecchia, A.; Zambon, M.; Rey, F. A.; Corti, D.; Veesler, D. Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion. *Cell* **2019**, *176*, 1026−1039.e15.
(25) Walls, A. C.; Tortorici, M. A.; Frenz, B.; Snijder, J.; Li, W.; Rey, F. A.; DiMaio, F.; Bosch, B.-J.; Veesler, D. Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. *Nat. Struct. Mol. Biol.* **2016**, *23*, 899−905.
(26) Bouwman, K. M.; Tomris, I.; Turner, H. L.; van der Woude, R.; Shamorkina, T. M.; Bosman, G. P.; Rockx, B.; Herfst, S.; Snijder, J.; Haagmans, B. L.; Ward, A. B.; Boons, G.-J.; de Vries, R. P. Multimerization- and glycosylation-dependent receptor binding of SARS-CoV-2 spike proteins. *PLoS Pathog.* **2021**, *17*, e1009282−e1009282.
(27) Zhang, Y.; Zhao, W.; Mao, Y.; Chen, Y.; Wang, S.; Zhong, Y.; Su, T.; Gong, M.; Du, D.; Lu, X.; Cheng, J.; Yang, H. Site-specific N-glycosylation Characterization of Recombinant SARS-CoV-2 Spike Proteins. *Mol. Cell. Proteomics* **2021**, *20*, 100058.
(28) N- and O-glycosylation patterns of the SARS-CoV-2 spike protein reveal a new "O- follow-N" rule http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD023346. 2021.
(29) NIST Glycopeptide Search 1.0 Software and Libraries https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:nistmsgads. 2021.
(30) Caval, T.; Heck, A. J. R.; Reiding, K. R. Meta-heterogeneity: evaluating and describing the diversity in glycosylation between sites on the same glycoprotein. *Mol. Cell. Proteomics* **2021**, *20*, 100010.
(31) Sokolow, S. K. J.; Gustafson, P., *The Finnigan Library Search Program: Finnigan Application Report 2*, Finnigan Corp., New York. 1978.
(32) Park, G. W.; Kim, J. Y.; Hwang, H.; Lee, J. Y.; Ahn, Y. H.; Lee, H. K.; Ji, E. S.; Kim, K. H.; Jeong, H. K.; Yun, K. N.; Kim, Y.-S.; Ko, J.-H.; An, H. J.; Kim, J. H.; Paik, Y.-K.; Yoo, J. S. Integrated GlycoProteome Analyzer (I-GPA) for Automated Identification and Quantitation of Site-Specific N-Glycosylation. *Sci. Rep.* **2016**, *6*, 21175−21175.

https://doi.org/10.1021/acs.jproteome.1c00442
J. Proteome Res. XXXX, XXX, XXX−XXX