Journal of proteome research

pubs.acs.org/jpr

Article

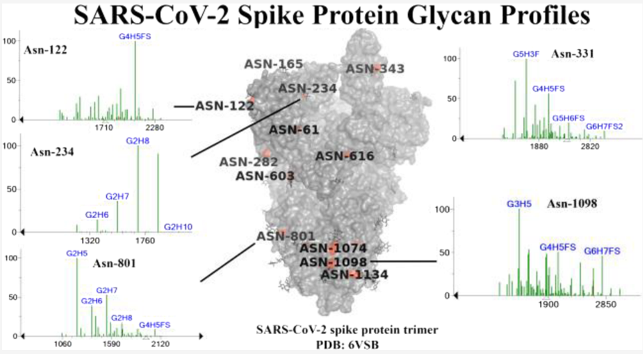
Determining Site-Specific Glycan Profiles of Recombinant SARS-CoV-2 Spike Proteins from Multiple Sources

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

Cite This: *J. Proteome Res.* 2023, 22, 3225−3241

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: Glycopeptide Abundance Distribution Spectra (GADS) were recently introduced as a means of representing, storing, and comparing glycan profiles of intact glycopeptides. Here, using that representation, an extensive analysis is made of multiple commercial sources of the recombinant SARS-CoV-2 spike protein, each containing 22 N-linked glycan sites (sequons). Multiple proteases are used along with variable energy fragmentation followed by ion trap confirmation. This enables a detailed examination of the reproducibility of the method across multiple types of variability. These results show that GADS are consistent between replicates and laboratories for sufficiently abundant glycopeptides. Derived GADS enable the examination and comparison of the glycan profiles between commercial sources of the spike protein. Multiple distinct glycopeptide distributions, generated by multiple proteases, confirm these profiles. Comparisons of GADS derived from 11 sources of recombinant spike protein reveal that sources for which protein expression methods were the same produced near-identical glycan profiles, thereby demonstrating the ability of this method to measure GADS of sufficient reliability to distinguish different glycoform distributions between commercial vendors and potentially to reliably determine and compare differences in glycosylation for any glycoprotein under different conditions of production. All mass spectrometry data files have been deposited in the MassIVE repository under the identifier MSV000091776.

KEYWORDS: *site-specific glycosylation, N-linked glycopeptides, mass spectral library, SARS-CoV-2*

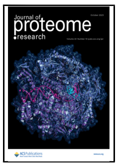
INTRODUCTION

Glycopeptide Abundance Distribution Spectra (GADS) have recently been described as an effective method for capturing glycoform distributions of intact glycopeptides for a given glycosylation site. These are expressed in a spectrum-like format that can be easily searched and compared using existing software for mass spectral library searching.[1] Such a format, in which all glycoforms for a given sequence are represented by the mass of the glycan composition on the *x*-axis and relative MS1 abundance on the *y*-axis, has enabled a straightforward method for comparing complex N-linked glycan distributions. As the quality of a GADS is dependent on the quality of the separation, abundance, fragmentation, and identification of a glycopeptide sequence, this work aims to evaluate the technical and interlaboratory reproducibility of the method and then apply it to examine glycosylation variability among multiple sources of the highly complex SARS-CoV-2 spike protein.

The SARS-CoV-2 spike protein is a viral surface glycoprotein that has been the target of vaccine development due to its role in host-cell receptor binding and cell entry.[2−4] It comprises three monomers each containing 22 N-linked glycosylation sites or sites in which a glycan is attached to the nitrogen of an asparagine side chain, which is followed by any amino acid (abbreviated as X), other than proline, and serine or threonine. This sequence motif, N-X-S/T, is termed a sequon.[5] Because of the role of the spike protein in vaccine development, one important question is the effect of different recombinant protein expression methods on the resulting glycan distribution as it may affect antigenicity.[6,7]

GADS libraries for 11 sources of recombinant spike protein were constructed from high-resolution LC-MS/MS data obtained from intact glycopeptides, where at least five different combinations of proteases were used to produce glycopeptides containing a single sequon. Glycopeptides containing multiple sequons were excluded. This requirement precluded the use of a nonspecific protease which, due to the low specificity, would produce multiple cleavage sites per sequon, thereby decreasing

Received: May 4, 2023
Published: August 30, 2023



https://doi.org/10.1021/acs.jproteome.3c00271

Table 1. Summary of Recombinant Spike Protein Sources Analyzed and the Information Provided by Each Vendor[a]

source	sequence	cells	furin cleavage site	proline substitutions	mutations	C-term tag
A	16−1213	HEK293	RAAA			T4, 10xHis
B	16−1213	HEK293	RAAA	F817, A892, A899, A942, K986, K87		T4, 10xHis
C	16−1213	HEK293	RAAA			His
D	1−1273	HEK293 Expi	GSAG	K986, V987		Rho 1D4
E	1−1273	HEK293 Expi	GSAG	K986, V987	del 69−70 and 144, N501Y, A570 D, D614G, P681H, T716I	Rho 1D4
F	15−1208	HEK293	GSAS	K986, K987		6xHis
G	1−1208	CHOExpress	GSAS	F817, A892, A899, A942, K986, V987	del 69−70 and 144−145, N501Y, A570D, D614G, P681H	8xHis
H	1−1208	CHOExpress	GSAS	K986, K987		8xHis
I	15−1208	HEK293	GSAS	K986, K987		His
J	16−1188	HEK293	RAAA			T4, His+Avi
K	16−1213	HEK293	RAAA			T4, 10xHis

[a] The sequence length is shown relative to that of the full length sequence (UniProt Accession P0DTC2)

the signal for a given sequon. Steps taken to acquire information-rich tandem mass spectra included a 490 min separation using a 75 cm C18 column, beam-type collision cell fragmentation (HCD) using stepped collision energies, and a contingent ion trap fragmentation triggered by the presence of an oxonium ion in the HCD spectrum. The long column and separation time aid in the separation of glycopeptides that differ in the number of sialic acid residues, as each additional sialic acid significantly increases the retention time. Furthermore, stepped HCD allows different collision energies to be combined to form a single tandem mass spectrum that contains fragment ions due to cleavages at glycosidic linkages, formed at lower collision energies, and the peptide backbone, which occurs at higher collision energies. The contingent ion trap spectrum generates a low energy spectrum that largely leaves the initially formed glycopeptide fragment ions intact. The resulting information-rich tandem mass spectra are required for the automated validation of glycopeptide spectral assignments.

After tentative glycopeptide identification, spectral peaks were annotated[8] and a variety of information was then used to confirm their identity. This information included relative retention time, where for a given peptide each sialyl group significantly delayed retention, a range of oxonium "marker" ions, and glycopeptide product ions, including the characteristic Y1 ion (a single sugar residue remaining on the peptide). Lastly, ion trap tandem mass spectra were used to verify the glycopeptide assignment of the corresponding HCD spectrum.

While a principal objective of this work is to report site-specific glycan variability among 11 commercial sources of the recombinant SARS-CoV-2 spike protein, an in-depth characterization of consistency for specific glycopeptides and sequons is presented to illustrate the level of confidence and variability associated with the present automated determination of site-specific glycan distributions. In addition to the well-known problems of accidental overlap of glycan and amino acid masses, other challenges include incorrect assignment of the monoisotope and incorrect computation of the area of the precursor isotope signals, which underscore the importance of using both MS1- and MS2-level information in glycopeptide identification. Details of the contribution of individual glycans to GADS provide users with a level of detail lost when a broad categorization is used to describe this distribution. After characterizing GADS variability, use of this method provided a detailed comparison of N-linked glycoforms distributions between multiple sources of recombinant protein that have been made available to be viewed, compared, and searched in a manner similar to that employed in mass spectral analysis.

■ MATERIALS AND METHODS

Materials

Sources of recombinant SARS-CoV-2 spike protein included Acro Biosystems (SPN-CH52H9), Creative Biolabs (VreP-Wyb141), Creative Biomart (Spike-208 V), Cube Biotech SARS CoV-2 spike protein (20782) and B.1.1.7 Mutation (28716), Elabscience (PKSR030489), ExcellGene D614G and Wuhan trimeric spike protein, ProSci (10−121), Reprokine (RKNCOVST), and speed BioSystems (YCP8621). Table 1 shows a summary of all 11 sources of the recombinant spike protein. Proteases rAspN (VA1160), Chymotrypsin (V1062), GluC (1651), LysC (V167A), and trypsin (V5111) were purchased from Promega, and both alpha lytic (wild type, A6362) and alpha lytic M190A mutants were purchased from Sigma.

Digestion of Recombinant Spike Protein

For each source of recombinant protein, 50 μg of protein was prepared by first performing three buffer exchanges using 50 mmol/L ammonium bicarbonate in a 10K Amicon molecular weight cutoff filter. For each buffer exchange, the filter was centrifuged at 14,000 rpm for 20 min. Next, the spike protein was brought to a final concentration of 50 mmol/L ammonium bicarbonate, 0.1% Rapigest-SF-Surfactant (Waters), and 20 mmol/L dithiothreitol. The solution was allowed to incubate at 60 °C for 60 min. Next, the solution was brought to a final concentration of 55 mmol/L iodoacetamide followed by incubation at 25 °C for 45 min in the dark. The alkylation step was then quenched by adding 20 mmol/L dithiothreitol in triplicate. A buffer exchange was repeated, and each sample was resuspended in 50 mmol/L ammonium bicarbonate.

A maximum of seven different proteases or combinations of proteases were used. All proteases, except for chymotrypsin, were incubated at 37 °C. Chymotrypsin digestions were incubated at 25 °C using the enzyme-to-substrate ratio and incubation times listed in Table 2. It should be noted that each alpha lytic wild type and M190A mutant digestion was performed with two additions of protease per digestion, with each addition allowed to incubate for 1 h. Each digestion was terminated by heating at 95 °C for 10 min. Samples were then

Table 2. Digestion Conditions Used for the Intact Glycopeptide Analysis of Recombinant Spike Proteins[a]

protease(s)	enzyme-to-substrate ratio	digestion time (h)	cleavage
chymotrypsin*	1:10	2	C-term: FWYL
alpha lytic wild type*	1:100	1	C-term: TASV
alpha lytic M190A mutant	1:100	1	C-term: MFL
AspN	1:20	18	N-term: DE
trypsin and LysC*	1:20	18	C-term: KR
trypsin and GluC*	1:20 (trypsin), 1:10 GluC	18	C-term: KR, N-term: DE
GluC and chymotrypsin	1:10	18	C-term: FWYLDE
chymotrypsin and trypsin	1:10 (chymotrypsin), 1:20 (trypsin)	18	C-term: FWYLKR

[a]Proteases shown with an asterisk were used for all 11 sources of recombinant spike protein.

desalted with a MonoSpin C18 Centrifugal Column (GL Sciences) and resuspended in 0.1% (v/v) formic acid in 10% (v/v) acetonitrile for LC-MS/MS analysis.

LC-MS/MS Analysis

Spike protein digests were analyzed on an UltiMate 3000 (Thermo Fisher Scientific) in-line with an Orbitrap Fusion Lumos (Thermo Fisher Scientific) mass spectrometer. Reversed-phase separation was performed on an Acclaim PepMap RSLC 75 μm × 75 cm column (Thermo Fisher Scientific) at a flow rate of 200 nL/min. The gradient consisted of (min: % Solvent B) 0:2, 320:32, 366:80, 383:98, 396:98, 409:2, and 490:2, where solvent B is 0.1% (v/v) formic acid in acetonitrile.

Data-dependent acquisition was performed in positive ion mode with an ion transfer tube temperature of 275 °C and a spray voltage of 1.80 kV. MS survey scans, or MS1 scans, were acquired with a scan range of m/z 380−2000, a maximum injection time of 50 ms, an AGC target of 400,000, and an RF lens value of 40%. Precursor ions were detected in the orbitrap with a resolution of 120,000. Precursor ions selected for fragmentation included those with charge states 2−8 with a minimum intensity of 50,000. A dynamic exclusion of 15 s was used with a tolerance of ±10 ppm. Tandem mass spectra were acquired with a cycle time of 5 s in between MS1 scans, a maximum injection time of 60 ms, an AGC target of 50,000, and a fixed starting m/z of 120. Beam-type collision cell spectra (HCD) using stepped collision energies of NCE 15, 25, and 35% were acquired and detected in the orbitrap with a resolution of 30,000. An ion trap scan was triggered if the GlcNAc oxonium ion at m/z 204.087 (15 ppm tolerance) was among the top 20 product ions in the HCD scan. The ion trap scan was acquired with a collision energy of 30%, activation Q of 0.25, and activation time of 10 ms. Product ions were detected in the orbitrap with a resolution of 30,000. All raw data files were made available for download on MassIVE (MSV000091776).

Data Analysis

Glycopeptide assignments were made using Byonic and Byologic software (version 3.10 Protein Metrics, Inc.). For each protease, with the exception of alpha lytic proteases, up to three missed cleavages were allowed using the cleavage rules listed in Table 2. Up to six missed cleavages were allowed for alpha lytic wild type and mutant proteases. Assignments were made using precursor and product ion tolerances of 5 and 20 ppm, respectively. Cysteines were required to contain a fixed carbamidomethyl while variable modifications included oxidation of methionine and N-terminal pyroglutamine. For glycopeptides, a library containing 445 N-glycans was used (Table S1).

GADS Library Construction

GADS were created using general methods described previously.[1,9] Briefly, this began with glycopeptide spectrum assignments from Byonic meeting a score threshold of 30 (see Figure S1 for Byonic score distribution). For a given glycopeptide sequence, each glycoform identified is represented by the mass of the glycan on the x-axis and its normalized relative abundance, derived from the XIC provided by Byonic output, on the y-axis. Peaks are labeled using the following compact abbreviation of the glycans components: G = HexNAc, H = Hexose, S = Sialic acid, F = Fucose, So = SO_3, and Po = HPO_3. The site of N-linked glycosylation is indicated by lower-case single-letter amino acid abbreviation (e.g., TQSLLIVNnATNVVIK/+2+3+4). Additional retention time, MS1- and MS2-level information, derived from annotated tandem mass spectra based on Byonic glycopeptide spectrum assignments, are used to annotate poor quality identifications.

Use of retention time to annotate poor or suspect glycoforms is based on the observation that, for a given glycopeptide sequence, the retention time is approximately the same for glycan compositions containing the same number of sialyl residues. Here, the median retention time is computed using identifications for a given glycopeptide sequence containing 0 up to 4 sialyl groups. Those identifications, for the same raw file, in which the retention time differs by more than 4 min are colored red and the peak is annotated with a "<" or ">" followed by the retention time deviation.

Information derived from annotated tandem mass spectra, made by MS_Piano,[8] were also used to reject glycopeptide identifications and denote an identification as suspect. Here, a narrower product ion tolerance of 10 ppm was used. GADS libraries for all 11 sources of recombinant spike protein have been made available for download. Spectra included in these libraries were required to have at least 100 good quality, identified tandem mass spectra (nSpec), and less than 33% of abundance attributed to peaks with retention time errors.

Collectively, the use of retention time and fragment ion annotation criteria for validation of mass spectra obtained using both ion trap and beam-type collisional dissociation have been found to be more conservative than a score threshold alone, corresponding to 1% FDR, for our single protein studies. Moreover, the automated methods described here eliminate the need for subjective or manual analysis of glycopeptide spectrum matches, as is commonly done for glycopeptide confirmation.

■ RESULTS AND DISCUSSION

This presentation of recombinant spike protein-derived GADS will be divided into three main sections: (1) evaluation of technical reproducibility, (2) evaluation of sequon reproducibility, and (3) comparison of glycoform distributions captured by the GADS for proteins from different sources. Here, evaluation of technical reproducibility extends previous work by examining highly complex glycan distributions for the same protein from different sources. It examines variation due to reversed phase separation and electrospray ionization of the same digest from multiple injections from the same vial.

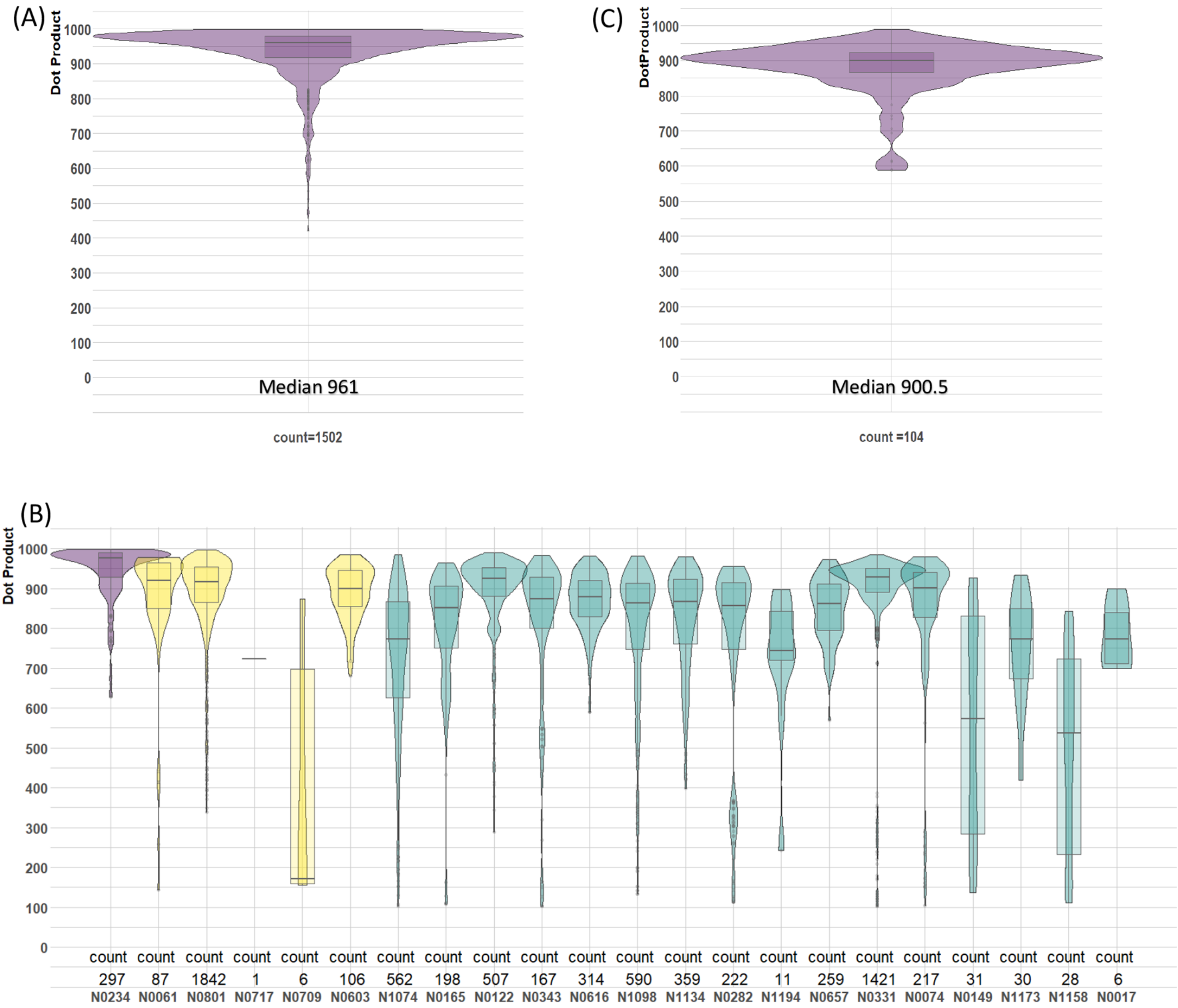


Figure 1. (A) Dot product distribution for inter-replicate GADS (same sequence/same charge state/same source) for all 11 sources of spike protein (1502 comparisons). (B) Dot product distribution for GADS corresponding to the same sequon with nSpec ≥100 (different sequence/same source). The sequons have been sorted in the order of descending total abundance of high mannose glycans. The colors are also used to highlight the oligomannose content of 80−100% (purple), 30−79% (yellow), and 0−19% (aquamarine). (C) Dot product distribution for NIST vs PNNL interlaboratory GADS for 2 sources of spike protein, D and E (104 comparisons).

Demonstration and discussion of sequon reproducibility examine the glycoform distribution identified in peptides produced by different proteases and, therefore, often differ in termini, length, elution time, and even charge state. This allows the glycan distribution on a single sequon to be identified across multiple distinct peptide sequences, adding to the confidence and providing a measure of variation in the distribution of glycoforms for a given sequon. Lastly, with measurement variabilities understood, those sequons that are well characterized using the methods described here are used to compare GADS across different recombinant spike protein sources for each of its 22 sequons. A key objective of this analysis is to confirm that these GADS measure characteristics of the proteins themselves and not artifacts of the measurement method.

These studies were conducted with the intention of generating as detailed and complete an analysis of the highly complex, recombinant form of the SARS-CoV-2 spike protein as currently possible. LC separation was done on a long column (75 cm − 490 min gradient) to ensure complete separation of sialylated glycans and maximal sampling time for abundant glycoforms. High loading (4 μg) was also employed to maximize the quality of glycopeptide fragmentation for each of the 11 recombinant spike protein sources. Due to space limitations, only selected illustrative GADS are shown here. All GADS are made available for download and can be perused and compared using NIST-provided mass spectral library software.

The minimum number of good quality MS2 spectral identifications for a GADS, nSpec, was set to 100 for comparison of reproducibility in the following sections. This relatively high threshold was found to uniformly generate reproducible GADS, for well-characterized sequons, that include minor glycoforms for even the most complex sequons (Figure S2). Of course, other factors, such as the glycan heterogeneity of a given sequon, protease selectivity, and the ability to generate and separate

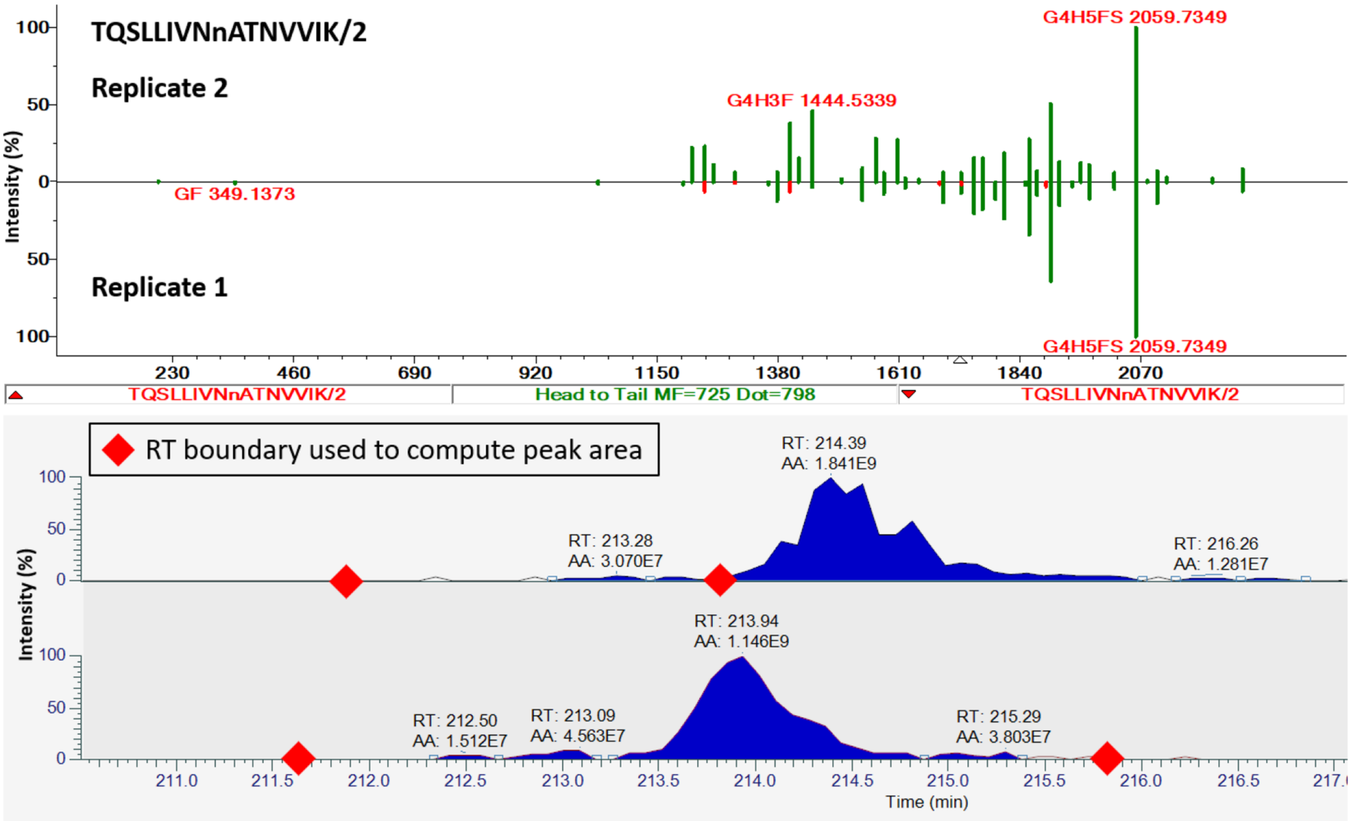


Figure 2. Example of irreproducible GADS between replicates due to an XIC issue in which the incorrect retention time boundaries were used to compute the XIC for replicate 1. In each XIC, red diamonds indicate the start and end retention times used for computing the XIC.

abundant glycopeptide ions, which may affect the number of glycoforms present and mass spectra acquired per glycoform contribute to the appropriate nSpec threshold for a given GADS. The modified cosine of the angle between GADS, the dot product, was used as a measure of similarity. This algorithm and extensions of it have long been used for measuring mass spectral similarity in library search applications.[10,11]

Reproducibility of GADS Across Technical Replicates

Here, an evaluation of the technical reproducibility of GADS is performed to determine the similarity expected for GADS derived from replicate measurements. These measure the degree of reproducibility of GADS possible using the present methods, where samples are different injections from the same vial. This arises from well-known variations in electrospray intensities, LC retention, MS/MS sampling, and identifications in nanospray ionization. Figure 1A shows the dot product distribution for inter-replicate GADS comparisons (same sequence and protein source) across all 11 sources of recombinant spike protein. The median dot product value of 961 indicates a very high degree of similarity between replicates. A dot product threshold of 850 was selected for GADS to be considered technically reproducible as extensive manual examination found that dot product values below this threshold may contain variation that could be ascribed to errors, such as assignment of the monoisotopic peak and glycopeptide sequence and not solely due to the variation in electrospray ionization and chromatographic separation (Supplemental Table S2 and Figure 2). In addition, the distribution of dot product values across all sequons in Figure 1B and Supporting Information Table S2B illustrates that those sequons with median dot product values less than 850 have broad dot product distributions consistent with poor reproducibility.

We found that 145 of all 1502 pairs of replicate GADS with nSpec ≥100 had dot product values less than 850 (Tables S3–S13). These were examined to find the origin of lower degrees of similarity, which were successfully identified for 138 of the 145 GADS (Table 3 and Supporting Information, Table S2).

Table 3. Principal Contributors to Variability of the Most Dissimilar Replicate GADS[a]

description	count	percent
poorly retained glycopeptide	61	44.2
incorrect monoisotopic assignment	32	23.1
incorrect ID/low abundance/low quality MS2	25	18.1
XIC uncertainty	18	13.0
probable ion suppression	1	0.72
possible in-source fragmentation	1	0.72

[a] Sources of variation were identified from manual inspection of all 145 inter-replicate GADS with dot products <850 (Supplemental Tables S3–S13), of which 138 were found attributable to the sources listed. Individual sequences for the GADS listed can be found in Table S2.

Sources included poorly retained chromatographic peaks, errors in extracted ion chromatogram (XIC) calculation (example shown in Figure 2), probable ion suppression, and errors in assignment of the glycan, peptide sequence, or monoisotopic peak. Using the established threshold, 90.2% of GADS with nSpec ≥ 100 were found to be technically reproducible, demonstrating that a high degree of similarity is expected for GADS derived from replicate measurements. The total number of GADS meeting the established threshold for each source of spike protein is given in Table 4.

Table 4. GADS Across 11 Sources (A−K) of Recombinant Spike Protein Meeting the Inter-Replicate Dot Product Threshold of 850[a]

spike source	total GADS (inter-replicate dot product ≥850)
A	152
B	143
C	103
D	79
E	73
F	168
G	76
H	112
I	108
J	175
K	168

[a] Here, four digests were performed for all 11 sources of recombinant spike protein: chymotrypsin, alpha-lytic wild type, trypsin and LysC, as well as Trypsin and GluC. All sources, sans sources D and E, were also digested using GluC and chymotrypsin as well as chymotrypsin and trypsin.

Consistency of Glycoform Abundance Distributions

Interlaboratory studies[12,13] have demonstrated that relative abundances of glycopeptide ions approximately correspond to their relative concentrations[14] and therefore should agree for different peptides containing the same sequon. The use of multiple proteases in this study, chosen in an effort to identify glycopeptides from all 22 sequons, enables the use of this idea to assess measurement variability by comparing GADS of different glycopeptide sequences containing the same sequon for a single source of recombinant spike protein.

The GADS used for this comparison combine multiple charge states into a single GADS. This is needed since the abundance of each glycoform for a given sequence can differ greatly for different charge states, with higher charge state ions preferring larger glycans.[1] Combining GADS for multiple charge states, by adding abundance values prior to normalization, into a single distribution therefore improves GADS similarity when comparing GADS for the same sequon (example shown in Figure S3).

Although glycopeptides were identified across all 22 sequons, not all sequons were found to have high-quality GADS from multiple distinct sequences. Comparing the nSpec corresponding to the most abundant peptide ion for each sequon (Figure 3A) highlights that some sequons, such as 122, 331, 801, and 1098, are highly abundant and well characterized. However, sequons 17, 709, 717, 1158, 1173, and 1194 have relatively low nSpec values due to their low abundance glycopeptides making them less confident and less reproducible. These differences in abundance, and therefore confidence, of each sequon are not unique to this study. In fact, the nSpec values identified here correlate with data published by Watanabe and co-workers, processed using the same methods described here, as shown in Figure 3B.[5] Interestingly, the two sequons that deviate from the correlation are those closest to the N- and C-terminus of the spike protein sequence, 17 and 1194. These deviations are the

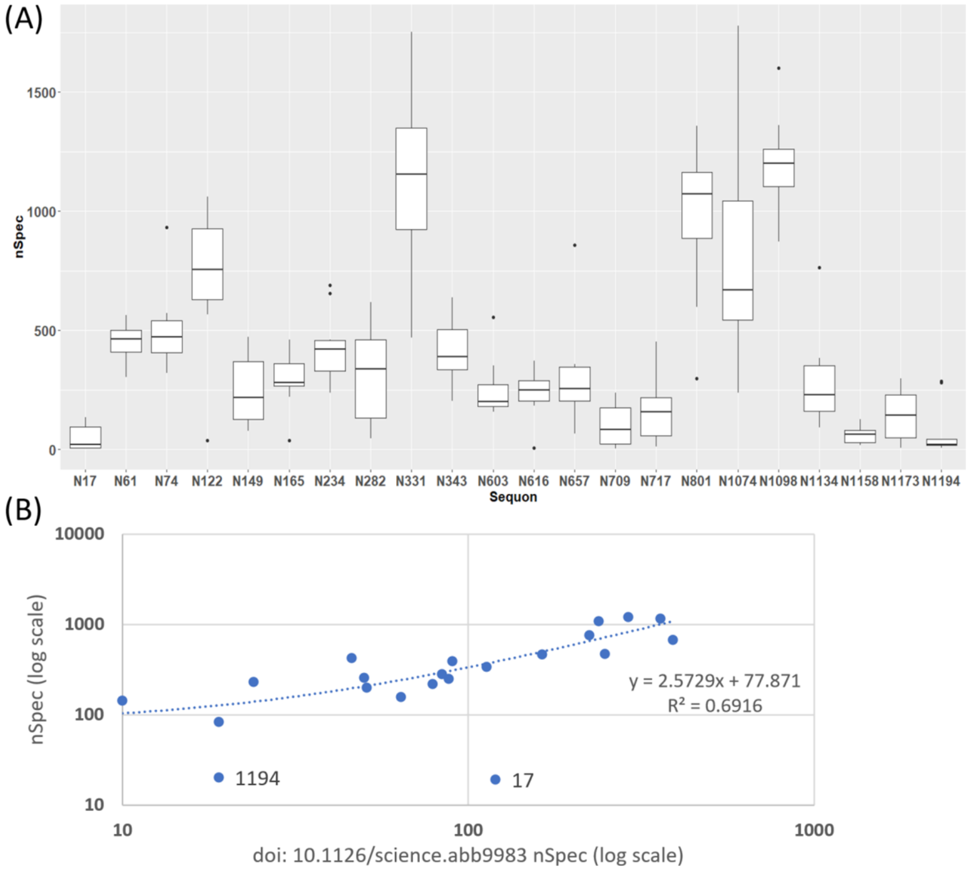


Figure 3. (A) Distribution of nSpec values per sequon for the most abundant peptide ion per source of the recombinant spike protein. These are listed in Table S15. (B) Correlation of nSpec values per sequon between the 11 sources of recombinant spike protein (*y*-axis) and those identified from raw mass spectral data published by Watanabe and co-workers.[5] Two outliers corresponding to sequons 17 and 1194 have been highlighted.

Table 5. Illustration of Those Sequons with GADS that were Found to be Reproducible, where the Numbers Represent the Number of Distinct, Reproducible (score > 850, nSpec ≥ 100) Glycopeptide Sequences between Replicates (Shown in Green) for All 11 Sources of Recombinant Spike Protein (A−K)

	17	61	74	122	149	165	234	282	331	343	603	616	657	709	717	801	1074	1098	1134	1158	1173	1194
A		2	2	7		1	3	1	9	3	2	6	5			9	1	6	4			
B		2	1	7	2	3	4	2	8	3	3	5	4	1	1	9	1	6	7			
C		2	8	6	1	2	3	1	7	3	1	1	1	1	1	6	4	6	4			
D	1	2	3	3		2	2	2	5		2	4	2	1		8	1	4	2		1	2
E		1	1	3		3	2	2	5	2	1	2	2			8	2	3	1		1	
F		2	1	3		2	3	1	9	3		4	2		1	6	2	4	3			1
G		4	2	6	1	3	4	2	12	3	2	2	2			14	2	6	3	1	1	2
H		2		3		1	3		5	1	1	1	1			5	1	3	3		1	
I		2		6	1		4	1	10	2		3	1		1	4	1	4	2	1	1	
J		2	1	3	1	3	4	2	11	3	1	2	1		1	8	3	6	2		2	
K		2	1	8	2	5	4	2	11	3	2	5	4			13	6	5	5			

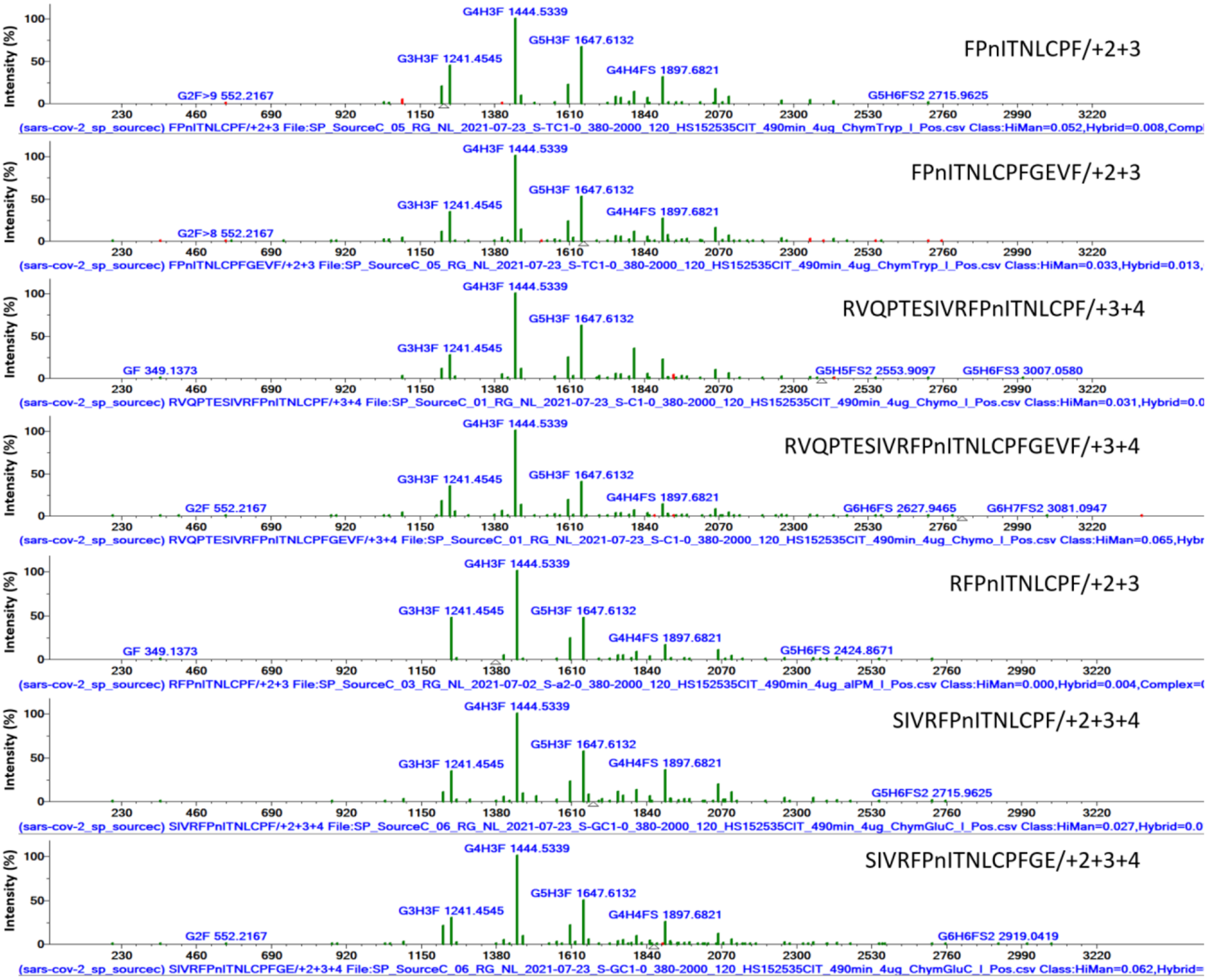


Figure 4. GADS derived from seven distinct glycopeptide sequences containing sequon 331 from the recombinant spike protein from Source C. Peaks shown in red have a retention time that differs by greater than four min from the median of glycopeptides with the same sequence and number of sialic acid residues. The open triangle along the *x*-axis corresponds to the monoisotopic mass of the nonglycosylated peptide sequence.

result of recombinant proteins of differing start and end positions relative to the full-length sequence (Table 1) which therefore produce different peptides. For sequons 709 and 717, in addition to peptides containing each sequon individually, peptides containing both sequons were also generated, which are not included in the GADS. Use of GADS highlights that

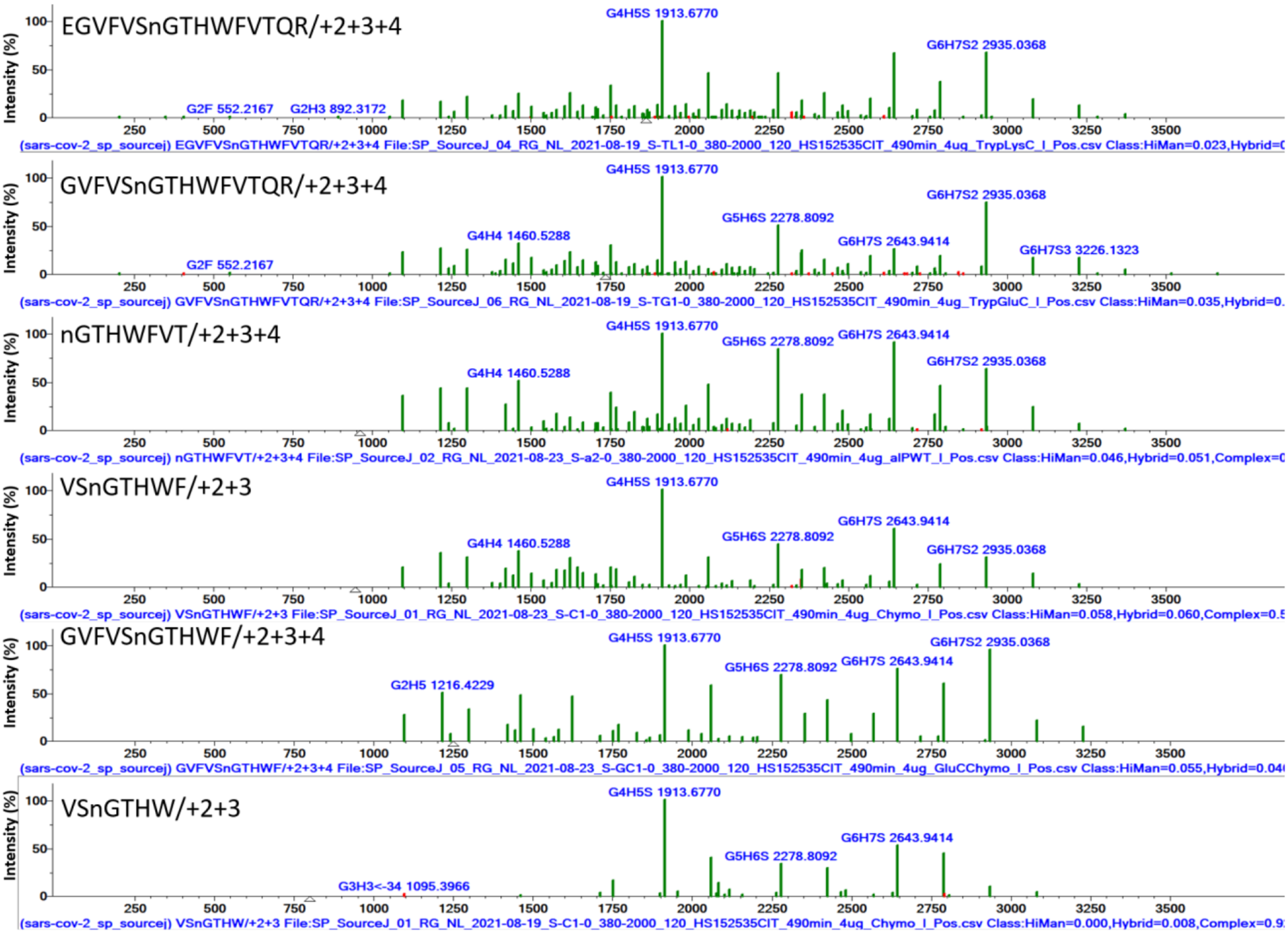


Figure 5. GADS derived from seven distinct glycopeptide sequences containing sequon 1098 from recombinant spike protein Source J. Peaks shown in red have a retention time that differs by greater than four min from the median of glycopeptides with the same sequence and number of sialic acid residues.

those challenging sequons with low nSpec values are also those with broad dot product distributions for different peptide sequences (Figure 1B) and, therefore, have poor reproducibility.

The total number of distinct glycopeptide sequences for which the GADS were found to be reproducible between technical replicates for each sequon is listed in Table 5 for each source of recombinant spike protein. In addition to being reproducible between technical replicates, the GADS listed meet additional chromatographic and fragmentation criteria outlined previously.[1] Altogether, these criteria mean that while glycopeptides were identified from all 22 sequons in each source of recombinant spike protein, 10 sequons were found to have high-quality GADS for all 11 sources of spike protein. Furthermore, 5 of these sequons (122, 234, 331, 801, and 1098) were characterized with at least 2 distinct glycopeptide sequences. The GADS derived from these five sequons is included in the GADS libraries available for download.

To examine GADS variability for the same sequon from different sequences (produced by different proteases) and thereby establish that they were largely independent of the method of measurement, a pairwise comparison by the dot product was made between each glycopeptide for the same sequon. Figure 1B shows that those sequons with at least 2 distinct glycopeptides per sequon for all 11 sources (sequons 122, 234, 331, 801, and 1098) each have median dot product values greater than 850. This figure highlights that GADS for peptides produced by different proteases and therefore contain different termini are indeed highly similar. Seven GADS from recombinant spike protein C for sequon 331, shown in Figure 4, further highlight the high degree of consistency. We note that a score of 850 has long meant a very high degree of similarity when matching mass spectra.

The broad distribution of dot product values observed in sequon 1098 underscores an important observation regarding the occurrence of adjacent cleavage sites for a given protease. The glycopeptide sequence associated with dot product values ranging from 620 to 737 is VSnGTHW, a chymotryptic peptide (Figure 5). Another glycopeptide sequence containing an additional phenylalanine, VSnGTHWF (Figure 5), was also produced by chymotrypsin. In this example, the peptide corresponding to the C-terminal cleavage of tryptophan is the minor peptide. The absence of low mass glycans in the GADS for VSnGTHW (Figure 5) is the result of a lack of sampling of the low-abundance analytes due, in part, to probable ion suppression caused by a highly abundant, unmodified coeluting peptide ion. This is one example illustrating how diluting the signal of a given sequon across two peptides can create challenges in acquiring high-quality tandem mass spectra and why having multiple sequences for a single sequon is needed for the confirmation of some glycan distributions. This is most

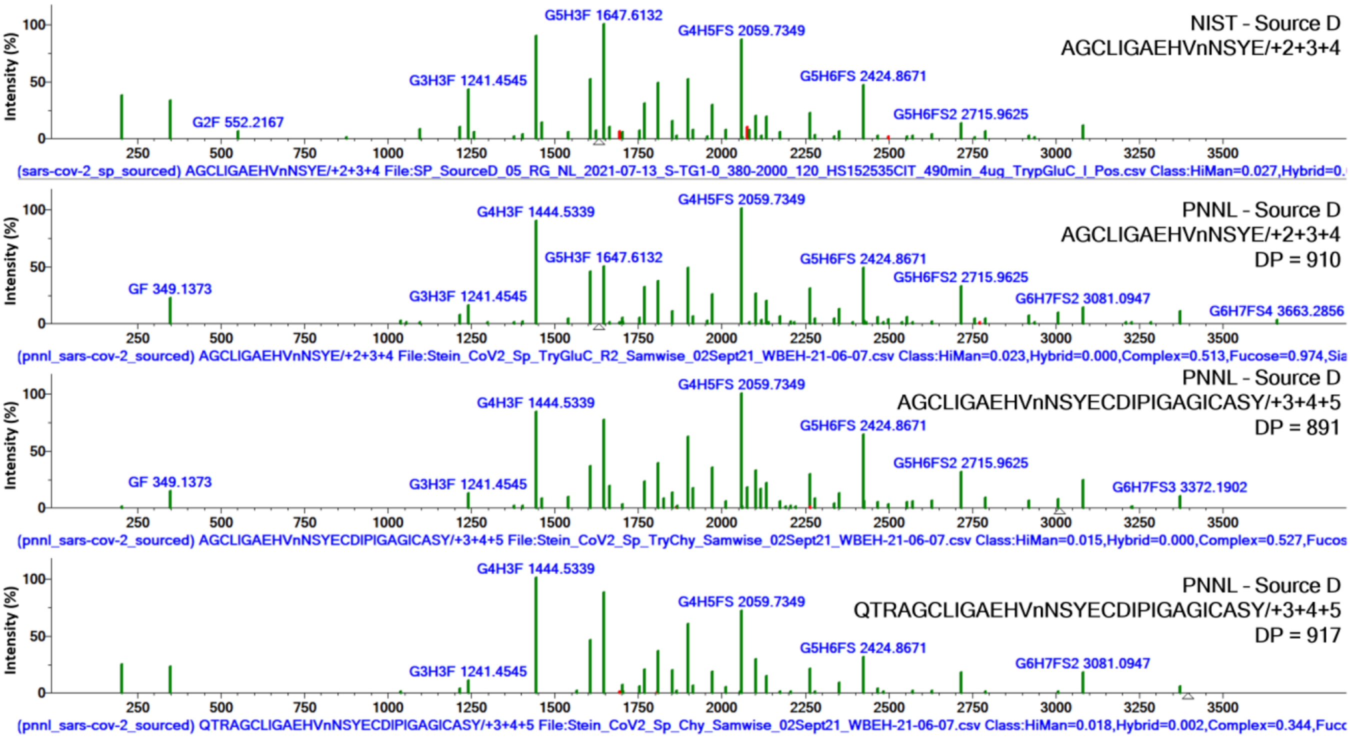


Figure 6. Comparison of GADS derived from sequon 657 for data acquired by PNNL and NIST for recombinant spike protein from Source D. Dot product (DP) values listed are relative to the GADS for AGCLIGAEHVnNSYE/+2+3+4 from data acquired at NIST.

notable for lower abundance glycoforms, as shown in Figure 5. This challenge may also occur for peptides containing missed cleavages and modifications, such as methionine oxidation. We note that such oxidation can "transform" a fucose to a hexose, potentially leading to an error in the assignment of peptide as well as glycan.

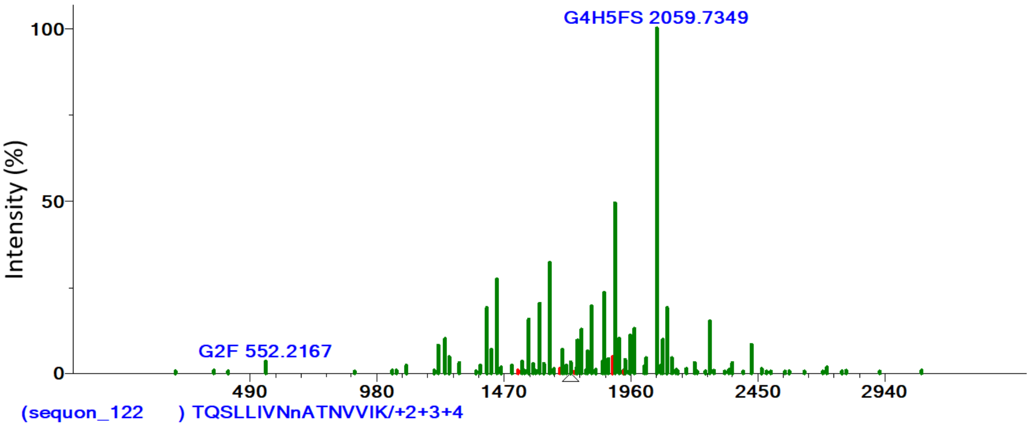
Reproducibility of GADS between Laboratories

Laboratories at NIST and Pacific Northwest National Laboratory (PNNL) were simultaneously sent spike proteins and analyzed using the same protocol to determine the degree of interlaboratory reproducibility for the GADS measurement. Both were sent two protein variants, the original spike (Source D) and a variant (Source E). A single protocol, described in the Supporting Information section, was used by both laboratories. Noteworthy differences between laboratories include the mass spectrometer used, column length, and gradient length. While these differences may cause variations in the amount of sampling of a given glycopeptide ion, we should expect the resulting GADS to be consistent.

When possible, the same glycopeptide sequence was used to compute the interlaboratory dot product values. However, due to the difference in the column length and gradient length, the nSpec for GADS acquired at PNNL were lower and an nSpec threshold of 100 could not be used. Instead, an inter-replicate dot product of 850 or greater was required. If the same glycopeptide sequence was not confidently identified, then the glycopeptide with the greatest nSpec value for that sequon was used for comparison (Supplemental Table S14). Indeed, the resulting interlaboratory dot product distribution has a median value of 900 (Figure 1C), indicating high similarity. The figure highlights that 22% of the dot product comparisons were below 850. A manual examination of MS1- and MS2-level information found that sources of lower dot product values were found for replicate runs in our own laboratory, including precursor ions that were not selected for fragmentation in the shorter gradient and XIC error (example shown in Figure S4). This is consistent with the sources of variation identified in the section on Reproducibility of GADS Across Technical Replicates. Taken together, these results show that the majority of GADS are reproducible between laboratories (example shown in Figure 6).

N-Linked Glycan Distributions for Different Sources of the SARS-CoV-2 Spike Protein. The distribution of glycoforms at each sequon, captured by the GADS, allows comparison across sources of recombinant spike protein, which vary not only in details of protein synthesis but also in the protein sequence length, cell expression system, furin cleavage site substitutions, and stabilizing proline substitutions. In view of this variability, we can only be descriptive in this analysis and choose illustrative sequons for a detailed examination. This also includes comparisons of results reported by Watanabe and co-workers.[5]

As previously reported,[5] the least processed sequons, defined as 80−100% high mannose-type glycans, included sequons 234 and 709. These sequons were also identified with 80−100% high mannose-type glycans in nine sources for sequon 234 and three sources for sequon 709. As discussed earlier, sequon 709 is poorly characterized due to low abundance, which explains the discrepancy. In addition, several sequons were reported with the largest degree of processing (0−29% high mannose): 17, 74, 149, 165, 282, 331, 343, 616, 657, 1098, 1134, 1158, 1173, and 1194.[5] Here, nine of the 14 sequons listed (74, 149, 282, 331, 343, 657, 1098, 1134, 1158) were also found to be highly processed across all 11 sources of recombinant spike protein. The sequons located near the C- and N-terminus (17, 1173, and 1194), which were not well characterized for all sources, were also found to be highly processed in all sources for which the sequons were well characterized. Lastly, sequons 165 and 616 were found to be highly processed in nine and ten sources of recombinant protein, respectively. Together, these results show



Source	Dot Product (DP)	High Mann DP	Hybrid DP	Complex DP	Sialyl DP
A	981	992	996	974	984
B	963	993	999	988	992
C	469	977	710	609	675
D	677	988	837	841	816
E	881	931	955	933	904
F	868	993	861	909	932
G	566	905	562	576	642
H	526	932	501	537	567
I	838	889	883	890	828
J	832	901	827	862	805

Figure 7. Comparison of GADS corresponding to sequon 122 across all 11 sources of the recombinant spike protein. Dot product values were computed relative to source K. Peaks shown in red have a retention time that differs by greater than 4 min from the median of glycopeptides with the same sequence and number of sialic acid residues.

that GADS reported here are in good agreement with previously reported characterizations of the spike protein.

As presented in the previous section, five of the twenty-two sequons in the spike protein were identified with multiple technically reproducible distinct glycopeptide distributions across all 11 sources. This allows confirmation of the identified glycoform distribution, and therefore, we use the same five sequons (122, 234, 331, 801, and 1098) to illustrate the nature of the site-specific variability found. All comparisons are made relative to recombinant spike protein source K, which generated the greatest total number of GADS (Table 4).

Sequon 122

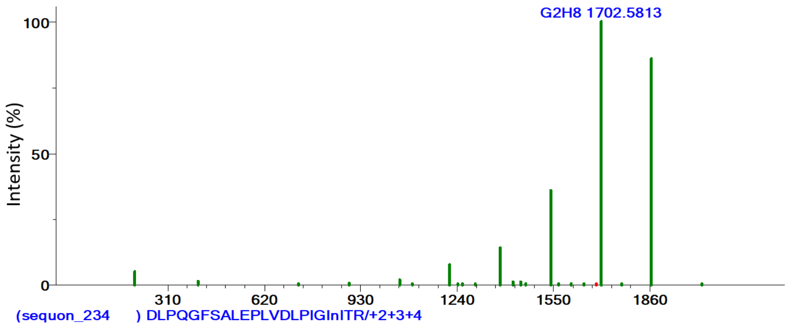
Consistent with previous reports,[5,6] sequon 122 was found to contain a mixture of high mannose-, hybrid-, and complex-type glycans (Figure 7, Supplementary Figure S5) with dot product values against other protein sources ranging from 469 to 981. Sources A and B were found to contain highly similar glycoform distributions (Dot product ≥963) to that of source K. These three sources, which differ only in proline substitutions (Table 1), were found to contain predominantly complex-type N-linked glycans with G4H5FS as the most abundant glycoform.

While seven sources contained >30% sialylated glycans (A, B, E, F, H, J, and K), two sources (C and D) have a larger proportion of high mannose-type glycans (DP of 469 and 677 relative to source K, Figure 7, Supporting Information, Figure S5). However, dot product values used to compare the relative abundances of each individual class of glycans, such as high mannose, hybrid, complex and sialylated, highlight similar (DP ≥ 905) relative abundances of high mannose glycans across all sources. This underscores that while the degree of processing may differ between select sources, the relative abundance of specific glycoforms for a general class, such as high mannose in this example, may be similar.

Interestingly, the two recombinant spike protein sources produced in CHO Express cells, G and H (Table 1), also had low dot product values of 566 and 526, respectively, when compared to source K. While all three sources have a significant proportion of abundance belonging to complex glycans, they differ in the relative abundances of the complex glycoforms, as demonstrated with Complex dot product values of 576 and 537 for sources G and H, respectively.

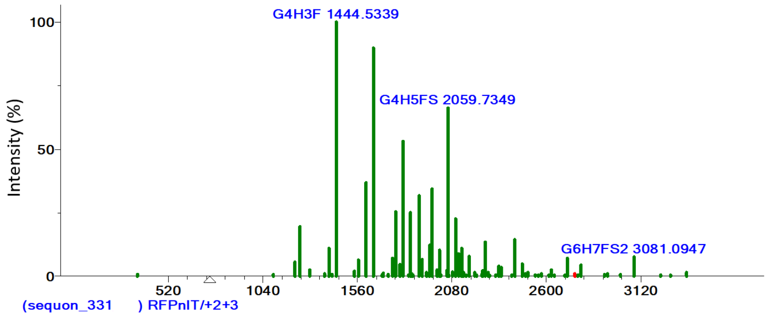
Sequon 234

While the most abundant glycoform for sequon 234 varies across the 11 sources of recombinant spike proteins, they are all high mannose-type N-linked glycans (Figure 8, Figure S6). This is consistent with an earlier report[6] and with the observation, based on molecular dynamics studies, that sequon 234 may be located in a trimer-associated mannose patch to aid in stabilizing



Source	Dot Product (DP)	High Mann DP	Hybrid DP	Complex DP
A	995	999	786	547
B	990	996	879	
C	872	884	941	
D	984	990	999	455
E	981	986	943	335
F	906	916	780	855
G	912	920	941	
H	863	892	790	661
I	351	505	710	594
J	554	679	724	568

Figure 8. Comparison of GADS corresponding to sequon 234 across all 11 sources of recombinant spike protein. Dot product values were computed relative to source K. Peaks shown in red have a retention time that differs by greater than 4 min from the median of glycopeptides with the same sequence and number of sialic acid residues.



Source	Dot Product (DP)	High Mann DP	Hybrid DP	Complex DP	Sialyl DP
A	965	981	994	953	949
B	979	999	977	996	994
C	837	854	948	812	834
D	885	969	899	883	907
E	933	881	927	947	932
F	954	963	965	953	983
G	905	751	83	863	935
H	799	773		831	892
I	876	751	933	862	913
J	868	874	917	829	900

Figure 9. Comparison of GADS corresponding to sequon 331 across all 11 sources of the recombinant spike protein. Dot product values were computed relative to source K.

the "up" conformation of the RBD.[15] However, a recent study of a spike protein prepared without proline substitutions or furin cleavage site modification[16] was found to contain predominantly processed glycans.

Seven of the 11 recombinant spike protein sources, A, B, D, E, F, H, K, share G2H8 as the most abundant glycoform (Supplemental Figure S6) and have highly similar GADS with dot product values ≥863. Two sources, C and G, have G2H9 as

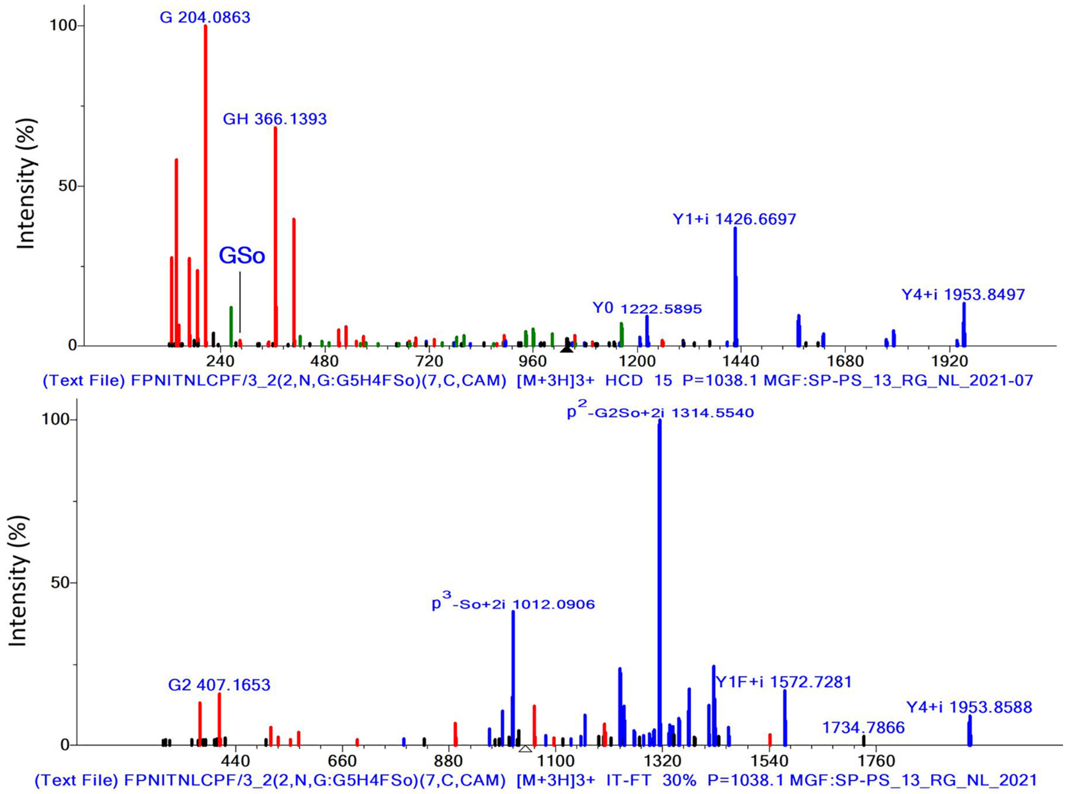
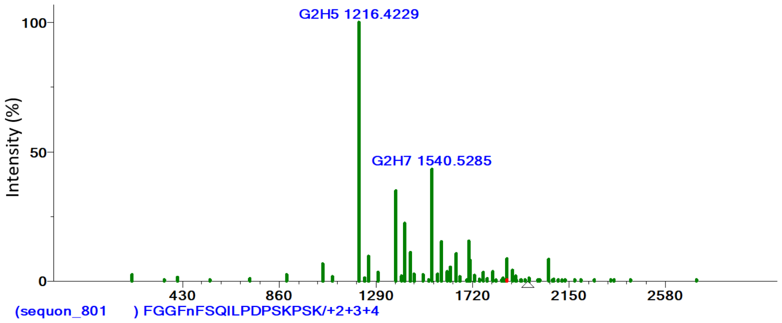
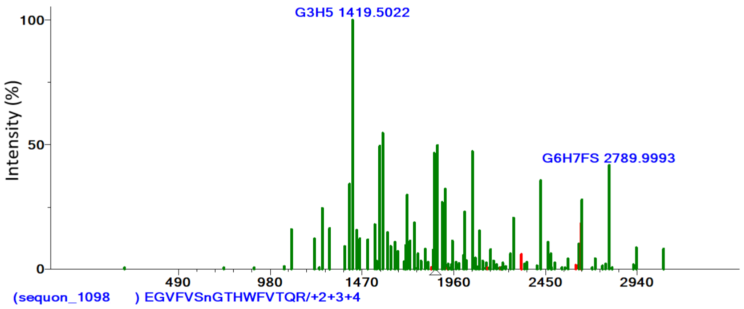


Figure 10. Annotated HCD (top) and IT (bottom) tandem mass spectra of FPNITNLCPF/3_2(2,N,G:G5H4FSo)(7,C,CAM) identified from a trypsin/chymotrypsin digest of recombinant spike protein source I. Colors are used to annotate peaks as unassigned (black), oxonium ions (red), peptide sequence or backbone ions (green) and glycopeptide ions (blue). The GSo oxonium ion at *m*/*z* 284 is highlighted in the HCD tandem mass spectrum.



Source	Dot Product (DP)	High Mann DP	Hybrid DP	Complex DP	Sialyl DP
A	995	999	999	987	988
B	994	998	999	995	997
C	800	939	905		789
D	951	962	993	899	967
E	929	968	969	875	923
F	874	995	945	888	810
G	915	977	944	861	915
H	685	972	928	888	757
I	731	921	970	923	801
J	734	972	951	909	790

Figure 11. Comparison of GADS corresponding to sequon 801 across all 11 sources of recombinant spike protein. Dot product values were computed relative to source K. Peaks shown in red have a retention time that differs by greater than 4 min from the median of glycopeptides with the same sequence and number of sialic acid residues.



Source	Dot Product (DP)	High Mann DP	Hybrid DP	Complex DP	Sialyl DP
A	994	999	999	990	996
B	995	998	999	993	995
C	801	982	929	691	785
D	886	928	994	922	924
E	920	925	988	909	901
F	869	911	994	901	880
G	816	973	920	828	844
H	714	892	774	853	771
I	741	746	932	868	795
J	767	860	846	808	762

Figure 12. Comparison of GADS corresponding to sequon 1098 across all 11 sources of recombinant spike protein. Dot product values were computed relative to source K. Peaks shown in red have a retention time that differs by greater than 4 min from the median of glycopeptides with the same sequence and number of sialic acid residues.

the most abundant glycoform, rather than G2H8, but nonetheless have highly similar GADS with dot product values of 872 and 912, respectively.

G2H5 is the most abundant glycoform present in recombinant spike protein sources I and J, resulting in a distinct distribution of high mannose glycans when compared to the other 9 sources (DP values of 351 and 554, respectively). Interestingly, these two recombinant proteins differ in the sequence length, furin cleavage site, and proline substitutions. This suggests that the sequence alone cannot account for the lack of processing at this sequon.

Sequon 331

Sequon 331 appears highly processed across the majority of recombinant spike protein sources, comprising predominantly complex- and hybrid-type glycans (Figure 9, Supplementary Figure S7), consistent with a previous report.[6] All 11 sources were found to be highly (>80%) fucosylated. Examination of the dot product values for each individual class of glycans highlights that fewer hybrid-type glycans were identified in recombinant spike protein sources G and H, both produced in CHOExpress cells (Table 1), resulting in a hybrid dot product value of 83 for source G. A hybrid dot product value could not be computed for source H as no hybrid-type glycans were identified.

Interestingly, sulfated glycoforms were identified in six of the 11 sources of recombinant spike protein and in multiple distinct glycopeptide sequences. Klein and co-workers also identified sulfated glycoforms at this sequon.[17] The sulfated glycoforms included G5H4FSo (sources A, B, F, I, J, and K) and G5H3FSo (sources F, I, K). Figure 10 includes an example of annotated HCD and IT tandem mass spectra for FPn(+G5H4FSo)-ITNLCPF (z = +3). Annotated tandem mass spectra for the sulfated glycoforms have also been included as a tandem mass spectral library available for download, for which all of the HCD tandem mass spectra were required to contain the GSo oxonium ion at m/z 284.

Sequon 801

Glycoforms at sequon 801 were found to contain minimal processing with seven sources of recombinant spike protein (A, B, D, E, F, G, K) containing G2H5 as the predominant glycoform (Figure 11, Supporting Information Figure S8). The significant proportion of high mannose-type glycans present in all sources is consistent with work performed by Watanabe and co-workers.[5] However, sources I and J both have G4H5FS as the predominant glycoform, resulting in lower dot product values of 731 and 734, respectively, when compared to source K. As noted for sequon 234, these two sources differ in their sequence length, furin cleavage site, and proline substitutions.

Sequon 1098

For sequon 1098, eight of the sources of recombinant spike protein (A, B, C, D, E, F, G, and K) contain a mixture of hybrid- and complex-type glycans (Figure 12, Supporting Information, Figure S9). Three of the sources (H, I, J) have a greater abundance of complex-type glycans with less than 5% of the abundance attributable to hybrid-type glycans, similar to the distribution reported by Watanabe and coauthors.[6] For these three sources, the recombinant proteins differ in the protein sequence length, cell expression system, furin cleavage site, and stabilizing proline substitutions (Table 1). While the abundance of high mannose-type glycans was low (<17%) across all 11 sources of recombinant spike protein, the amount of sialylation did vary, as reflected in the Sialyl DP values shown in Figure 11 which ranged from 762 to 996.

Comparison of Different Glycosites within Proteins from the Same Source. Comparison of GADS derived from different sequons within the same source of recombinant spike protein, which has also be referred to as glycan meta-

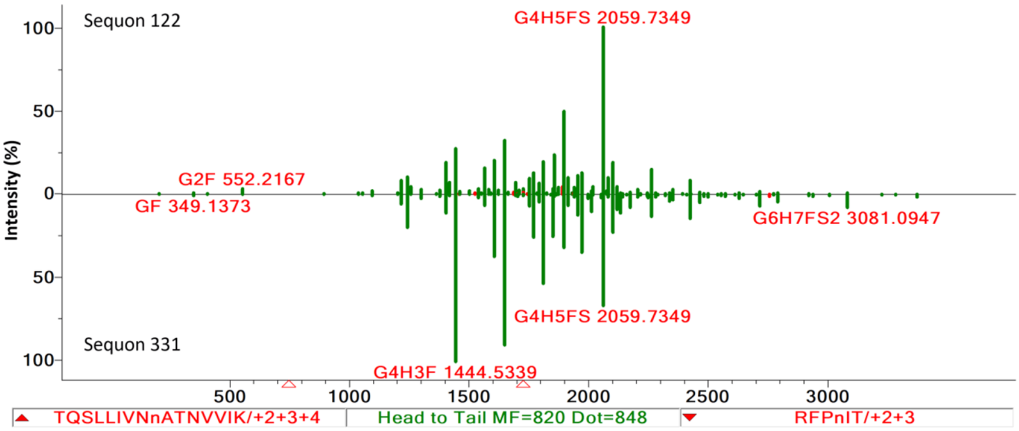


Figure 13. Head-to-tail comparison of GADS derived from sequons 122 (top) and 331 (bottom) for recombinant spike protein source K. Peaks shown in red have a retention time that differs by greater than four min from the median of glycopeptides with the same sequence and number of sialic acid residues.

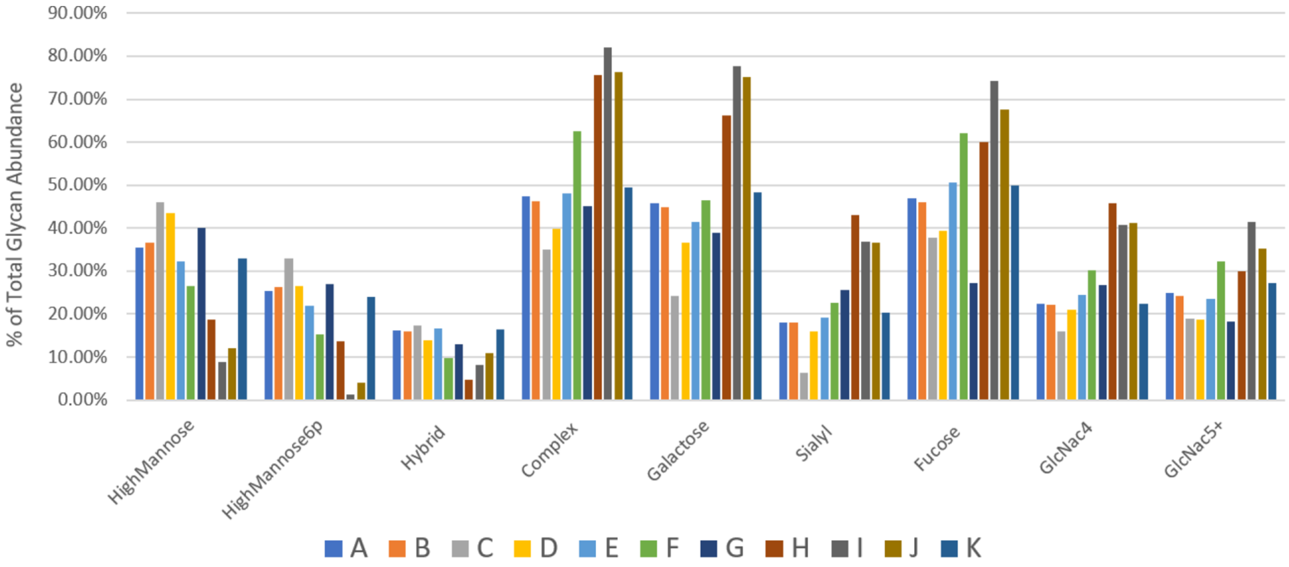


Figure 14. Comparison of glycan class abundance across all 11 sources of the recombinant spike protein. Here, the classes are defined as high mannose (G(2)H(5−9)), high mannose 6p (G(2)H(6-)), hybrid (G(3)H(5−8)), complex (G(4−6)H(5−7)), galactose-containing (non-high mannose with >3 hexose), sialylated, fucosylated, and complex-type glycans containing four GlcNAc and ≥5 GlcNAc.

heterogeneity,[18] can provide insight into the similarity in accessibility and degree of processing for different sites on the same protein. For each source of recombinant spike protein, dot product values were computed between GADS for sequons 122, 234, 331, 801, and 1098, as these distributions could be confirmed by multiple distinct glycopeptide sequences in all 11 sources.

The results, included in Supporting Information, Table S16, show that sequons 122 and 331 have similar glycoform distributions (dot product ≥800) in five (A, B, E, F, K) of the 11 sources of recombinant spike protein analyzed. Both sequons are located within the S1 domain, although sequon 122 is located within the N-terminal domain while 331 is located in the receptor binding domain. For these recombinant proteins, both sequons contained predominantly fucosylated glycans (>56%, see example in Figure 13). Interestingly, sequon 122 has a significantly higher abundance (>40%) of high mannose-type glycans compared to that of sequon 331 for sources C and D. This suggests that the accessibility of sequon 122 for sources C and D may differ from that of sources A, B, E, F, and K, as both sequons had a low (≤20%) abundance of high mannose-type glycans. Sources C and D differ in the sequence length, furin cleavage site, and proline substitutions (Table 1). Furthermore, sources A, B, E, F, and K also differ in the sequence length, furin cleavage site, and proline substitutions (Table 1). Taken together, these results highlight the variability and, therefore, the meta-heterogeneity across sequons within the same source of recombinant spike protein.

Comparison of N-Linked Glycosylation for Proteins from Different Sources

Comparing glycoform distributions at the sequon level has revealed similarities and differences across sources of recombinant spike protein that cannot be explained by sequence alone. As individual sequons may differ in their accessibility or shielding, an additional perspective may be gleaned from comparing the abundance of nine classes of glycans across sources of recombinant spike protein. These classes, presented in Figure 14, include high mannose, high mannose with 6 or more hexose residues, hybrid, complex, galactose-containing (nonhigh mannose containing >3 hexoses), sialylated, fucosy-

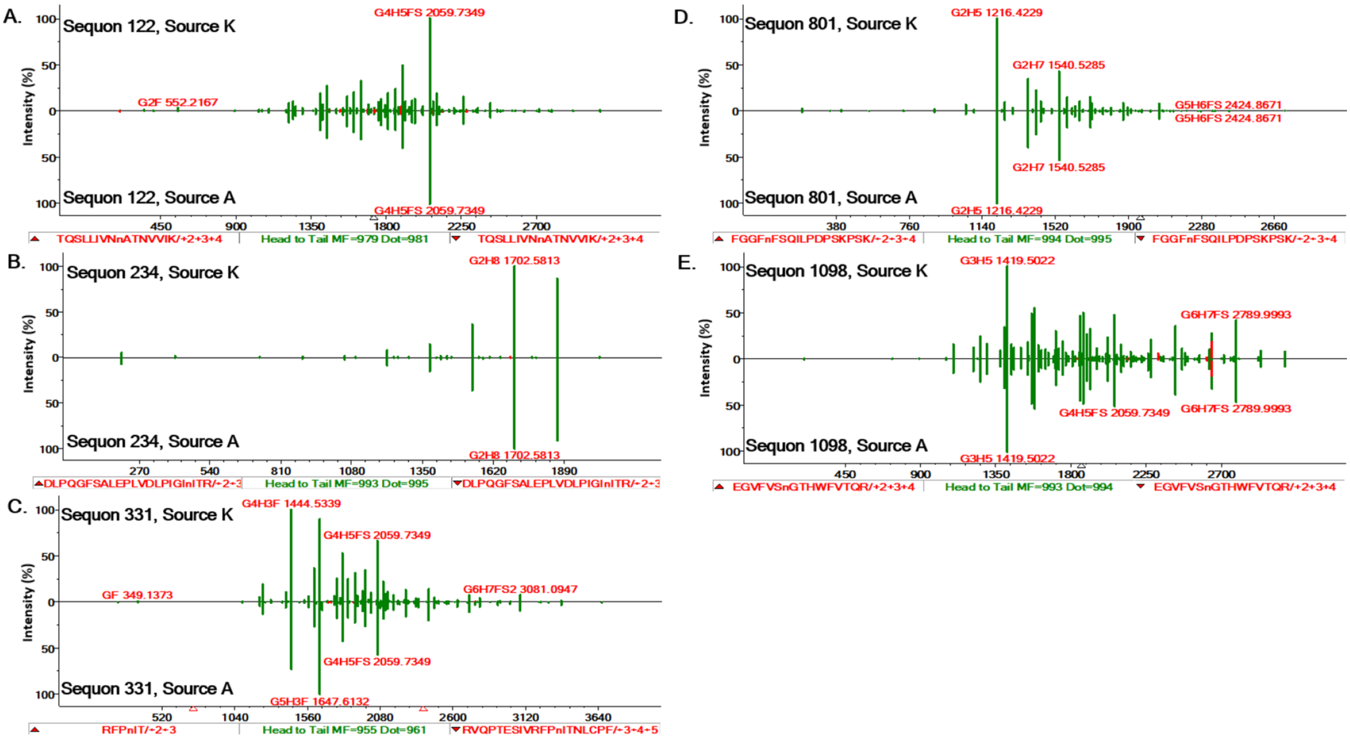


Figure 15. Head-to-tail comparison of source A (bottom) and K (top) for sequons (A) 122, (B) 234, (C) 331, (D) 801, and (E) 1098. Sources A and K have the same sequence length, cell expression system, furin cleavage site, stabilizing proline substitutions, and C-terminal tag (Table 1).

lated, and complex-type glycans containing 4 GlcNac or ≥5 GlcNAc.

The results shown in Figure 14 illustrate that sources A–G and K have roughly similar distributions across all 9 classes presented, with the exception of sialylated glycans in source C. As shown, sialylated glycans comprise 6.2% of the glycoforms in source C, while sources A–B, D–G, and K have values ranging from 17.2 to 25.4% abundance. Additionally, sources H, I, and J have a greater abundance of mature complex glycans (range 75.8–82.1%) relative to the remaining sources (range 34.9–62.5%), including galactose-, sialyl-, and fucose-containing glycans. While the GADS for sources I and J have been highly similar across sequons 122, 234, 331, 801, and 1098 (Figures 7–9), the glycoform distribution for source H differed for sequons 122, 234, and 801 (Supplemental Figures S5, S6, and S8). Using sequon 801 as an example, sources I and J share G4H5FS as the most abundant glycoform, while G4H3F is the most abundant glycoform for source H (Supplemental Figure S8). It has been noted that sources I and J differ in the sequence length, furin cleavage site, and proline substitutions; however, sources I and H differ in the sequence length and cell line used (Table 1). Source H was produced in CHOExpress cells while sources I and J were produced in HEK cells. Therefore, while the degree of processing appears similar across sources H, I, and J, differences in the relative abundance of individual glycoforms may be related to the CHO cell line used to produce Source H.

Overall, the results at the sequon- and protein level suggest that discrete differences in sequence or cell expression system alone cannot account for the differences observed in glycan processing and distribution. However, the results shown in Figure 14 show that two sources, A and K, with the same sequence length, cell expression system, furin cleavage site, stabilizing proline substitutions, and C-terminal tag have nearly identical glycan distributions for all five sequons described here (DP 961–995, Figure 15). Similarly, source B, which only differs from Source A and K in stabilizing Pro substitutions, is also consistently reproducible (DP 958–995, Supplemental Figure 10) across all five sequons, relative to source K. Collectively, the data presented highlight that recombinant proteins produced under the same protein expression conditions contain highly similar (dot product > 958) glycoform distributions that have been confirmed across multiple distinct glycopeptide sequences. Therefore, harmonized protein expression conditions were found to produce reproducible N-linked glycan profiles.

■ CONCLUSIONS

The in-depth characterization of glycoform variability across 11 sources of recombinant spike protein has established that GADS derived from replicate measurements and different laboratories are reproducible and that distributions from seemingly similar sources can be very different, most notably in their degree of sialylation. Sources of variability in measuring glycoform abundance were found to include poorly retained chromatographic peaks, errors in XIC calculation, ion suppression, and misassignment of monoisotopic peak or glycopeptide sequence. Together, the results show that after accounting for known sources of variation, GADS may be used to reliably compare the relative abundances of glycoforms identified at a given site.

In this work, the use of multiple proteases enabled the confirmation of glycoform distribution for a given site across multiple distinct glycopeptide sequences. This allowed a detailed comparison of glycoform distributions for five sequons (122, 234, 331, 801, and 1098) across 11 sources of recombinant spike protein. Importantly, the results presented illustrate that the recombinant proteins prepared under the same conditions yield reproducible glycoform distributions from different commercial vendors. This indicates that this method can be effective in determining differences in glycosylation of glycoproteins from different sources as both recombinant and native sources. The GADS described has been made available as a library that can be examined, compared, and searched in mass-spectral-like libraries using NIST MS Search software.

ASSOCIATED CONTENT

Supporting Information

The following files are available free of charge. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00271.

Supporting materials with tables listing glycan library used in the assignment of glycopeptide sequences (Table S1); inter-replicate GADS with dot product (DP) values less than 850 and sources of variability, when possible (Table S2); inter-replicate GADS comparisons for recombinant spike protein source A (Table S3) and source B (Table S4), C (Table S5), D (Table S6), E (Table S7), F (Table S8), G (Table S9), H (Table S10), I (Table S11), J (Table S12) and K (Table S13); interlaboratory GADS comparisons between NIST and PNNL (Table S14); nSpec values for the most abundant peptide ions per sequon for spike protein sources A−K (Table S15); and figures related to Byonic score distribution across relative abundance thresholds (bin = 10%) for GADS with a minimum nSpec of 100 (Figure S1); dot product distribution for GADS corresponding to the same sequon (Figure S2); head-to-tail comparison of GADS for the same peptide before and after charge states are combined (Figure S3); XIC error resulted in an incorrect abundance of G2H5 in the PNNL GADS for FGGFnFSQILPDPSKPSK/+2+3+4 (Figure S4); GADS comparison of sequon 122 (Figure S5), 234 (Figure S6), 331 (Figure S7), 801 (Figure S8), and 1098 (Figure S9) across all 11 sources of recombinant spike protein; and head-to-tail comparison of source B (bottom) and K (top) for sequons 122, 234, 331, 801, and 1098 (Figure S10) (PDF)

Table of contents with figures related to Byonic score distribution across relative abundance thresholds (bin = 10%) for GADS with a minimum nSpec of 100; dot product distribution for GADS corresponding to the same sequon; head-to-tail comparison of GADS for the same peptide before and after charge states are combined; (PDF)

Spike glycan profile details with sequon and peptide information (XLSX)

PNNL LC−MS/MS analysis (PDF)

AUTHOR INFORMATION

Corresponding Author

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; Email: meghan.burke@nist.gov

Authors

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

Concepcion 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

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

Sergey L. Sheetlin − *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*

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

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

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.3c00271

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare the following competing financial interest(s): Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

ACKNOWLEDGMENTS

The authors would like to thank Karl K. Weitz (PNNL) and Heather Olsen (PNNL) for LC-MS/MS analysis of two commercial sources of recombinant spike protein used in the interlaboratory analysis and Drs. Zheng Zhang (NIST), Yuxue Liang (NIST), and Oleg Toropov (NIST) for their suggestions and technical support.

ABBREVIATIONS

GADS=Glycopeptide Abundance Distribution Spectra; H=Hexose; S=Sialic acid; F=Fucose; So=SO_3; Po=HPO_3; nSpec=number of good quality tandem mass spectra identified; LC=liquid chromatography; MS/MS=tandem mass spectrum; XIC=extracted ion chromatogram; G(2)H(5−9)=High mannose-type glycans; G(3)H(5−8)=Hybrid-type glycans; G(4−6)H(5−7)=Complex-type glycans

REFERENCES

(1) 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* (9), 4475−4486.

(2) Letko, M.; Marzi, A.; Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **2020**, *5* (4), 562−569.

(3) Wrapp, D.; Wang, N.; Corbett Kizzmekia, S.; Goldsmith Jory, A.; Hsieh, C.-L.; Abiona, O.; Graham Barney, S.; McLellan Jason, S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367* (6483), 1260−1263. (acccessed Feb 02, 2022)

(4) Amanat, F.; Krammer, F. SARS-CoV-2 Vaccines: Status Report. *Immunity* **2020**, *52* (4), 583−589.

(5) Watanabe, Y.; Allen Joel, D.; Wrapp, D.; McLellan Jason, S.; Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **2020**, *369* (6501), 330−333. (accessed Feb 02, 2022)

(6) Allen, J. D.; Chawla, H.; Samsudin, F.; Zuzic, L.; Shivgan, A. T.; Watanabe, Y.; He, W.-T.; Callaghan, S.; Song, G.; Yong, P.; et al. Site-specific steric control of SARS-CoV-2 spike glycosylation. *bioRxiv* **2021**, No. 433764.

(7) Watanabe, Y.; Bowden, T. A.; Wilson, I. A.; Crispin, M. Exploitation of glycosylation in enveloped virus pathobiology. *Biochim. Biophys. Acta, Gen. Subj.* **2019**, *1863* (10), 1480−1497.

(8) 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* (9), 4603−4609.

(9) Remoroza, C. A.; Burke, M. C.; Yang, X.; Sheetlin, S.; Mirokhin, Y.; Markey, S. P.; Tchekhovskoi, D. V.; Stein, S. E. Mass Spectral Library Methods for Analysis of Site-Specific N-Glycosylation: Application to Human Milk Proteins. *J. Proteome Res.* **2022**, *21* (10), 2421−2434.

(10) Stein, S. E.; Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (9), 859−866.

(11) Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing similar spectra: From similarity index to spectral contrast angle. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (1), 85−88.

(12) Wada, Y.; Azadi, P.; Costello, C. E.; Dell, A.; Dwek, R. A.; Geyer, H.; Geyer, R.; Kakehi, K.; Karlsson, N. G.; Kato, K.; et al. Comparison of the methods for profiling glycoprotein glycans—HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study. *Glycobiology* **2007**, *17* (4), 411−422. (accessed Aug 24, 2022)

(13) Leymarie, N.; Griffin, P. J.; Jonscher, K.; Kolarich, D.; Orlando, R.; McComb, M.; Zaia, J.; Aguilan, J.; Alley, W. R.; Altmann, F.; et al. Interlaboratory Study on Differential Analysis of Protein Glycosylation by Mass Spectrometry: The ABRF Glycoprotein Research Multi-Institutional Study 2012. *Mol. Cell. Proteomics* **2013**, *12* (10), 2935−2951. (accessed Aug 24, 2022)

(14) Nilsson, J. Liquid chromatography-tandem mass spectrometry-based fragmentation analysis of glycopeptides. *Glycoconj. J.* **2016**, *33* (3), 261−272.

(15) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent. Sci.* **2020**, *6* (10), 1722−1734.

(16) Zhang, S.; Go, E. P.; Ding, H.; Anang, S.; Kappes, J. C.; Desaire, H.; Sodroski, J. G. Analysis of Glycosylation and disulfide bonding of wild-type SARS-CoV-2 spike glycoprotein. *J. Virol.* **2022**, *96* (3), No. e0162621, DOI: 10.1128/JVI.01626-21. (accessed Feb 02, 2022)

(17) Klein, J. A.; Zaia, J. Assignment of coronavirus spike protein site-specific glycosylation using GlycReSoft. *Biorxiv* **2020**, No. 125302.

(18) Caval, T.; Heck, A. J.; Reiding, K. R. Meta-heterogeneity: evaluating and describing the diversity in glycosylation between sites on the same glycoprotein. *Mol. Cell. Proteomics* **2021**, *20*, 100010.