GlycanAnalyzer: Software for Automated Interpretation of *N*-Glycan Profiles After Exoglycosidase Digestions

Ian Walsh[1], Terry Nguyen-Khuong[1], Katherine Wongtrakul-Kish[1], Shi Jie Tay[1], Daniel Chew[1], Tasha José[2], Christopher H. Taron[2,†] and Pauline M. Rudd[1,†,*]

[1]Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), 20 Biopolis Way, #06-01 Centros, 138668, Singapore.[2]New England Biolabs, 240 County Road, Ipswich, MA, United States

[†]Christopher Taron and Pauline Rudd share senior authorship. *To whom correspondence should be addressed.

Table of Contents

1. Reading this supplementary 1

2. Releasing *N*-glycans from glycoproteins 2

3. Exoglycosidase overview 2

3.1 Exoglycosidases are highly specific 2

3.2 The retention time and mass matching problem 2

3.3 Exoglycosidase usage to discriminate similar *N*-glycans 3

4. GlycanAnalyzer web application 4

4.1 User input 5

4.1.1 The UPLC peaks and mass 5

4.1.2 Selecting a florescent label and permethylation 5

4.1.3 Where to focus the calculation 5

4.1.4 Constraining the search by selecting a predefined library 5

4.2 Software output 7

5. GlycanAnalyzer algorithm and scoring 10

5.1 Algorithm 10

6.2 Scoring 13

6. Additional results and tables. 16

7. Experimental protocol 17

8. Glycanalyzer support, limitations and future work 18

8.1 What does GlycanAnalyzer support? 18

8.2 Limitations and future work 18

References 18

Abbreviations

UPLC: Ultra-performance liquid chromatography

LC: Liquid chromatography

MS: Mass spectrometry

GU: Glucose Units

Acknowledgements

The authors would like to thank Dr. Sanjay Kumar (New England Biolabs, U.S.A) for his help in discussing and debugging the GlycanAnalyzer application, and Dr. Stephen Shi (New England Biolabs) for insightful discussions on the use of exoglycosidases. We would like to thank Dr Song Zhiwei (Bioprocessing Technology Institute) for kindly providing anti-Her2 antibody which was used throughout the manuscript to test the exoglycosidases.

1. Reading this supplementary

This supplementary serves as a general overview of the GlycanAnalyzer software. Within the supplementary, experimental results from anti-Her2 antibody using the 2-AB fluorescent label is presented. All the data and glycan annotations for the anti-Her2 antibody presented here are contained in an accompanying supplementary file titled IgG_complete_annotation.xls. The data was derived from a previous experiment (Chan, et al., 2016). It is important to note that GlycanAnalyzer can operate on released *N*-glycan LC/MS data from any glycoprotein where the *N*-glycans can be analysed unmodified or tagged with any fluorescent label.

2. Releasing *N*-glycans from glycoproteins

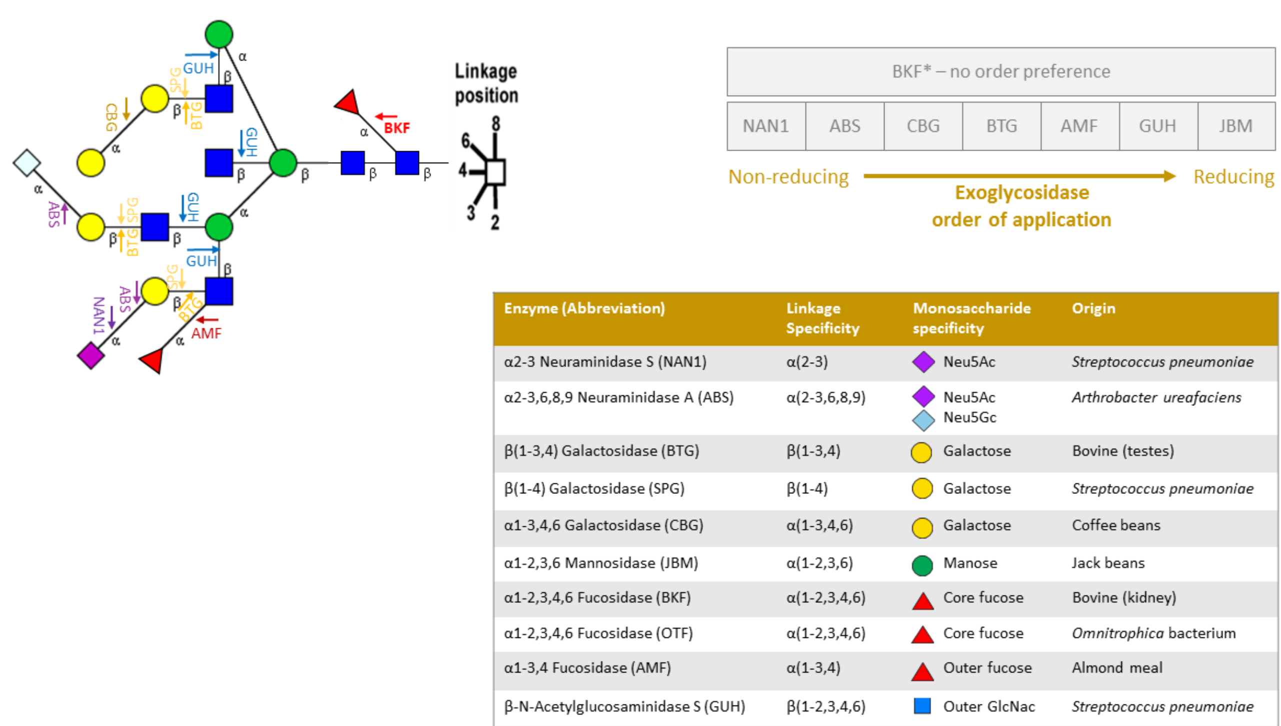
To fully characterize the structural complexity of a glycoprotein, it is essential to enzymatically release glycans from the protein, identify the glycan structure and measure their relative abundance. This is important, for example, in engineering biotherapeutics and identification of disease-related *N*-glycan alterations (glycan biomarkers). The experimental protocol for releasing *N*-glycans is beyond the scope of this article but further details can be found in (Mariño, et al., 2010). In high performance and ultra-performance liquid chromatography (H/UPLC) the relative abundance of glycans eluting at a peak can be measured by calculating the percentage peak area of the fluorescence signal. It is the goal of this article to present software for the automated identification of the eluting *N*-glycans in H/UPLC coupled to mass spectrometry (MS).

3. Exoglycosidase overview

Exoglycosidases are glycoside hydrolase enzymes which hydrolyze specific glycosidic bonds at terminal monosaccharides (non-reducing end) of an oligosaccharide. Consequently, monosaccharides in particular conformations are removed from the terminal side of the *N*-glycan in a step-wise manner. A series of exoglycosidases (also termed exoglycosidase array) applied to *N*-glycans can be used to improve the interpretation of glycan structural data obtained from H/UPLC–MS analyses. A number of enzymes have been fine tuned into the field's preferred exoglycosidases for use in glycan structure determination (Supplementary Figure 2).

3.1 Exoglycosidases are highly specific

Exoglycosidases are enzymes that release monosaccharides from the non-reducing end of *N*-glycans (Supplementary Figure 2). They not only have specificity toward monosaccharide units but also cleave specific anomeric (α/β) configurations and glycosidic linkages. Given that the H/UPLC and MS peaks change predictably in response to exoglycosidase application, the sequential application of an array of exoglycosidase will give the necessary information to infer exact structure of any *N*-glycan. Note that the nomenclature used in Supplementary Figure 2 is used throughout this manuscript as well as in GlycanAnalzyer's output.

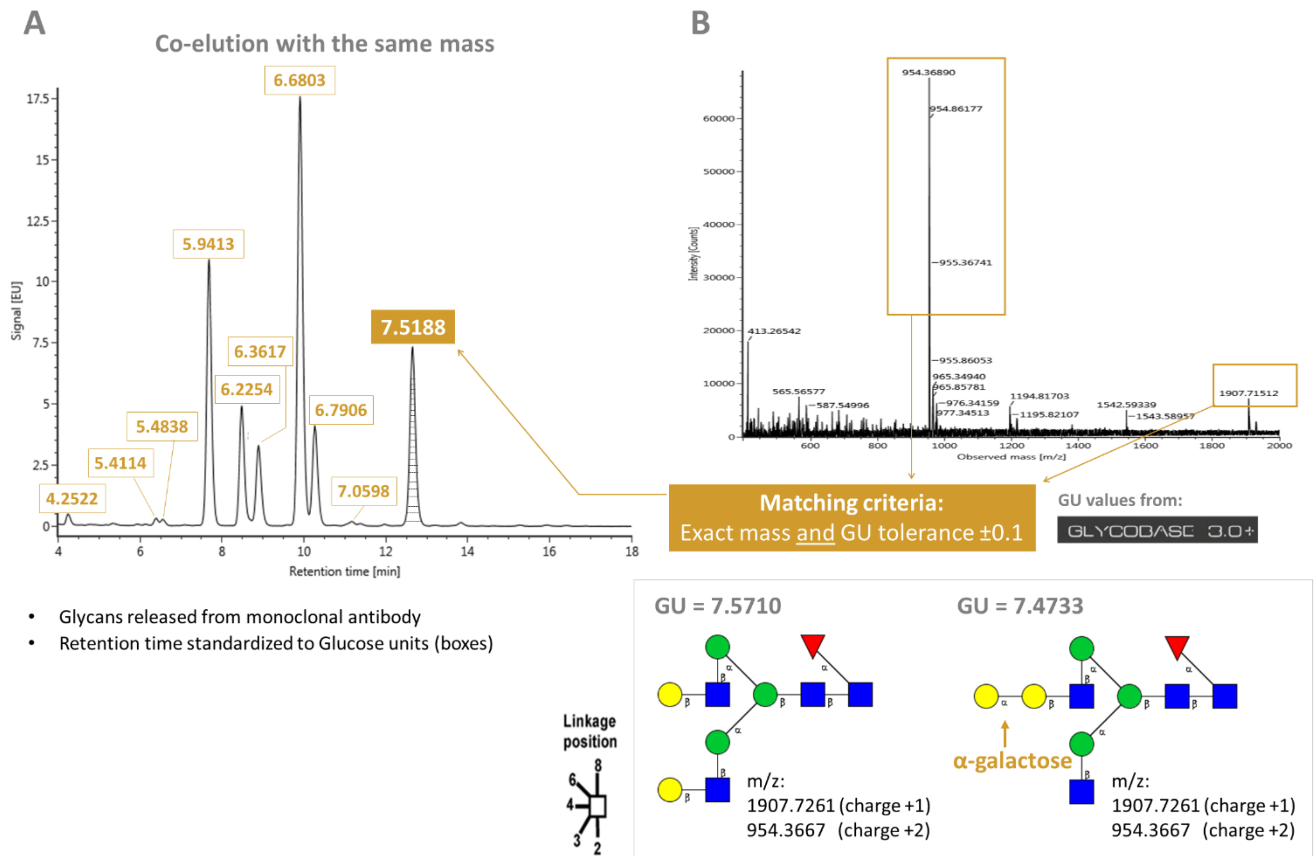


Enzyme (Abbreviation)	Linkage Specificity	Monosaccharide specificity	Origin
α2-3 Neuraminidase S (NAN1)	α(2-3)	Neu5Ac	*Streptococcus pneumoniae*
α2-3,6,8,9 Neuraminidase A (ABS)	α(2-3,6,8,9)	Neu5Ac Neu5Gc	*Arthrobacter ureafaciens*
β(1-3,4) Galactosidase (BTG)	β(1-3,4)	Galactose	Bovine (testes)
β(1-4) Galactosidase (SPG)	β(1-4)	Galactose	*Streptococcus pneumoniae*
α1-3,4,6 Galactosidase (CBG)	α(1-3,4,6)	Galactose	Coffee beans
α1-2,3,6 Mannosidase (JBM)	α(1-2,3,6)	Manose	Jack beans
α1-2,3,4,6 Fucosidase (BKF)	α(1-2,3,4,6)	Core fucose	Bovine (kidney)
α1-2,3,4,6 Fucosidase (OTF)	α(1-2,3,4,6)	Core fucose	*Omnitrophica* bacterium
α1-3,4 Fucosidase (AMF)	α(1-3,4)	Outer fucose	Almond meal
β-N-Acetylglucosaminidase S (GUH)	β(1-2,3,4,6)	Outer GlcNac	*Streptococcus pneumoniae*

Supplementary Figure 1. Linkage and monosaccharide specificity of the exoglycosidases supported in GlycanAnalyzer. Support for exoglycosidases currently in development will be implemented in future versions of GlycanAnalyzer. Note that OTF and BKF have the same specificity, however OTF was recently shown to efficiently hydrolyze in the presence of newer fluorescent dyes (Vainauskas, et al., 2018).

3.2 The retention time and mass matching problem

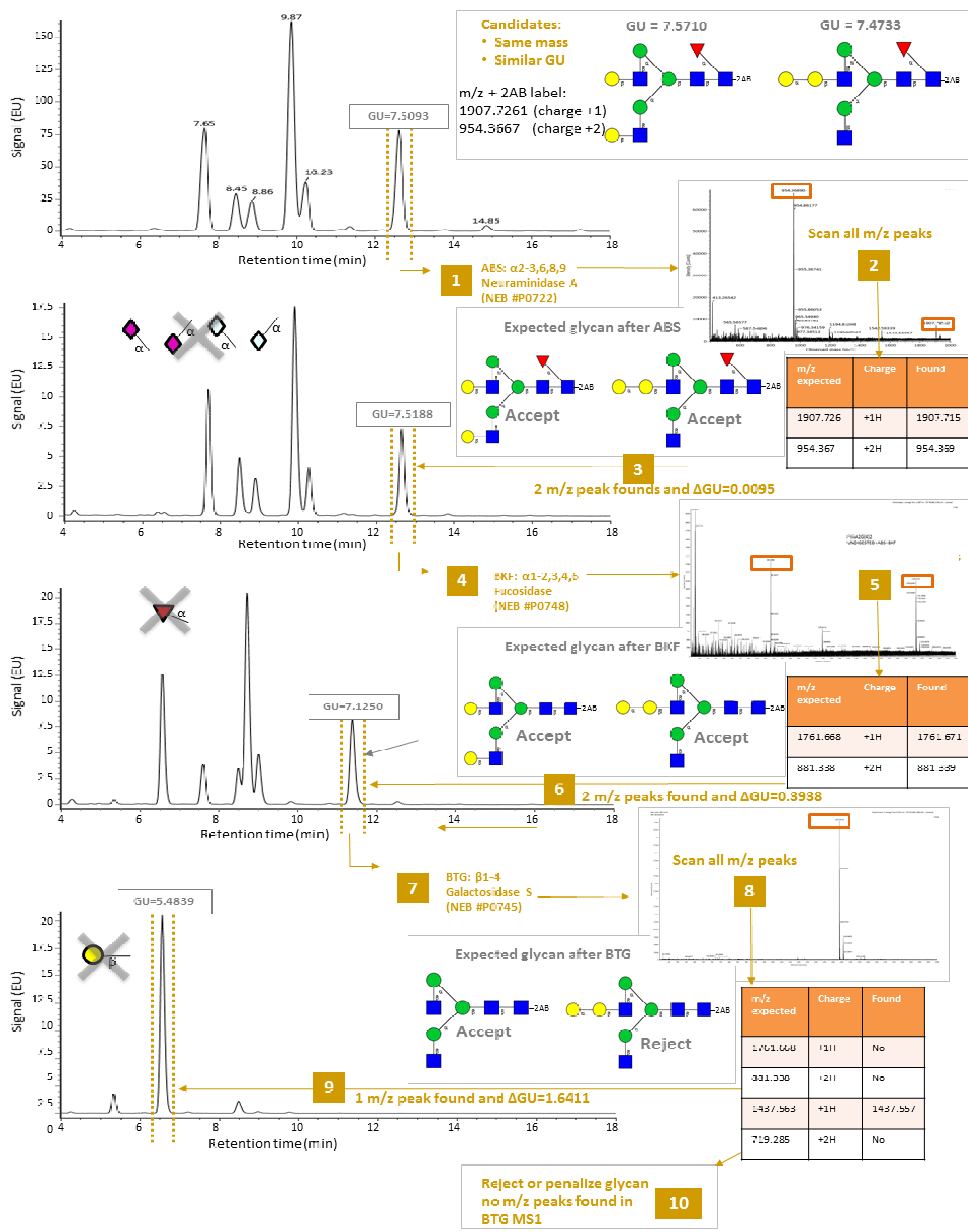
Building a database or library of *N*-glycans and matching them by observed mass and retention time is an effective way to automate assignment to H/UPLC-MS chromatograms. Additionally, retention times are often normalised to glucose units by matching peaks to maltooligosaccharide ladders. Glucose Units alleviate the problem of varying retention times with differing experimental conditions. Despite the usefulness of GU and mass matching, ambiguities often arise. For example, in Supplementary Figure 2 two N-glycans are possible when using GlycoStore (Zhao, et al., 2018) as our reference database. In fact, the structure with the α1-3 galactose epitope seems to be a better choice using GU similarity, but this was shown to be incorrect by manual analysis with exoglycosidases. This example highlights a need for exoglycosidase experiments, and GlycanAnalyzer can automate the annotation of these experiments.



Supplementary Figure 2. UPLC-MS data for glycans released from anti-Her2 antibody. GlycoStores (Zhao, et al., 2018) glucose unit search function was used to match similar *N*-glycans in A) and masses in (B), however, there is clear ambiguity when assigning glycan structure by mass and GU alone. For example, glycans 1 and 2 have identical masses and similar retention times and the software picks the incorrect isoform.

3.3 Exoglycosidase usage to discriminate similar *N*-glycans.

The ambiguity that arises by retention time and mass matching, highlighted in Supplementary Figure 2, is a common occurrence in any H/UPLC-MS analysis. This is especially true for isomers (i.e. glycans with the same mass but different conformation). The uncertainty can be overcome by using an exoglycosidase array defined using prior knowledge of the glycoprotein or sample under analysis. For example, the isomeric ambiguity in the example in Supplementary Figure 2 can be overcome using a simple exoglycosidase array (Supplementary Figure 3) which cleaves α(2-3, 6, 8, 9) sialic acid (ABS), α(1-2, 3, 4, 6) core fucose (BKF or OTF) and β(1-3,4) galactose (BTG). In other words, the isomers with identical mass and similar GU values can be distinguished using the information present in the exoglycosidase peak movements. It is exactly the interpretation shown in Supplementary Figure 3 that our GlycanAnalyzer software accomplishes.



Supplementary Figure 3. Monitoring UPLC-MS peak movements following sequential exoglycosidase digestions (ABS+BKF+BTG exoglycosidase array) enables the discrimination of isomers with identical mass and similar retention times. Data derived from anti-Her2 antibody experiment using the 2-AB fluorescent label. Steps 1-3: ABS is expected to have no impact on candidates confirmed by a small GU shift (ΔGU = 0.0095) and correct *m/z* found in charge +1H and +2H states. Steps 4-6: BKF is expected to cleave core fucose on both candidates; this is confirmed by a GU shift (ΔGU = 0.3868) and correct *m/z* found in charge +1H and +2H states. Steps 7, 8, 9: BTG is expected to cleave two β-galactose from one candidate; this is confirmed by a ΔGU shift of 1.6411. Additionally correct *m/z* is found in charge +1H and +2H states. Steps 7, 8, 10: BTG is expected to have no impact on the α-galactose candidate; the mass of this structure is not found in any peak therefore it should be penalized as a possible candidate. GU values are displayed above peaks. For higher resolution MS images see the supplementary data file called IgG_complete_annotation.xls.

4. GlycanAnalyzer web application

GlycanAnalyzer is a web application; meaning that all its functionality is executed through a web browser over the internet. The advantage of this approach is that all computations (some of them quite complex) are executed on our servers at New England Biolabs. Moreover,

maintenance of the software is carried out by our engineers at the server location and no installation is required on personal computers. The user interface is set up in a step-wise fashion to guide the user through the application; however, help pages are provided (using the '?' button) to further assist the user. It is important to note that the software can operate on data derived from any LC instrument using any separation column (BEH-Glycan column for HILIC-mode, BEH-C18 column, porous graphitized carbon etc.) as long as the retention times can be normalized to the glucose unit standard. Finally, tutorials using real datasets are provided to facilitate user training and support the program's use.

4.1 User input

We highly recommend that all software users first try the three tutorials that contain real experimental datasets. The tutorials will inform of the input process and the following text also gives a brief, general overview of it.

4.1.1 The UPLC peaks and mass

Variables such as GU, retention time, *m/z* etc are always supplied as columns separated in tabulated formatted (i.e. each column is separated by the tab buttn on a PC). Supplementary Figure 4a shows the first input scenario where users can directly upload a UPLC chromatogram and mass spectra. When uploading a UPLC chromatogram three variables must be supplied: (i) the retention time of the eluting peak, (ii) the Glucose Unit (GU) standardized retention time and (iii) the peak area (a measure of the relative abundance of the peak). The mass spectra can be provided for all possible retention times (represented as a 3D plot in Supplementary Figure 4a) through three variables: (i) all retention times, (ii) the *m/z* at each retention time and (iii) the intensity at each retention time and *m/z*. GlycanAnalyzer automatically extracts the mass for each peak from the provided information. Alternatively, the more experienced UPLC-MS user may prefer to extract the mass and charge themselves and input these four variables directly (see Supplementary Figure 4b and tutorial 2 online). Assignment without mass information is also possible (see 'NA' values in Supplementary Figure 4b and tutorial 3 online).

4.1.2 Selecting a florescent label and permethylation

N-glycans are commonly labelled with a fluorescent molecule which allows fluorescence-based quantitation after chromatographic separation. *N*-glycan labelling typically occurs via a reductive amination reaction (e.g. 2-AB, 2-AA and procainamide labelling) where the label is covalently attached to the reducing end terminus. A new fluorescent molecule from Waters Corporation, RapiFluor-MS™ (RFMS), however, performs rapid labelling of the *N*-glycan glycosylamine immediately following PNGase F treatment. The advantage of all these approaches is that labelling occurs in a stoichiometric manner where one label is attached to each *N*-glycan, allowing direct quantitation (Ruhaak, et al., 2010).

There are advantages and disadvantages when using different fluorescent labels. For example, 2-AB is a long-established label and the current gold standard in *N*-glycan profiling of biologics. The 2-AB labelled glycan database available on GlycoStore (Zhao, et al., 2018) is also the most extensive, allowing good coverage of most mammalian samples. However, in LC-MS workflows, 2-AB performs relatively poorly in MS detection because of its poor ionisation efficiency (Klapoetke, et al., 2010). In comparison, procainamide labelling has been shown to provide greater sensitivity in both fluorescence and MS detection compared to 2-AB, with up to 30 times higher MS signal intensity and up to three times higher fluorescence intensity for IgG *N*-glycans (Kozak, et al., 2015). RFMS provides the highest MS and fluorescence signal sensitivity of all currently available labels, with the added benefit of fast release and labelling times, but is also the most costly. Fluorescence signal intensity for RFMS-labelled N-glycans has been reported to be ten times greater than 2-AB, and 150 times greater in MS detection.

GlycanAnalyzer supports the use of three fluorescent labels: 2-aminobenzamide (2-AB), RapiFluor-MS™ (RFMS) and Procainamide. However, users can select the option label as 'free' if their glycan is unlabelled, or they can select the option 'Other' and enter the mass of the label they are using. GlycanAnalyzer does not currently support permethylation.

4.1.3 Where to focus the calculation

GlycanAnalyzer lets the user have complete control when annotating a chromatogram. To meet this goal, each peak can be selected one by one in the input, or the options "Assign all top hit" and "assign all top 5" can be selected. In the latter case, all peaks will be annotated with the best scored N-glycan or the top 5 scored N-glycans. This may take over 15-30 minutes depending on the complexity of the chromatogram (i.e. , number of peaks). Alternatively, processing peaks one-by-one allows users to concentrate on a peak of interest returning results quickly (1-2 minutes) and allowing users to reject poorly ranked glycans in a step-wise manner.

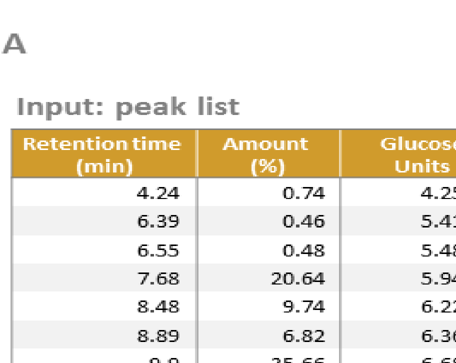
4.1.4 Constraining the search by selecting a predefined library

Our database GlycoStore, recently published (Zhao, et al., 2018) in collaboration with the authors of GlycoBase (Campbell, et al., 2008), contain experimental GU evidence for over 800 N-glycans from many different sources. Sources include human, mouse, serum, IgG, Erythropoietin and many different MAbs among others. For a particular N-glycan the GU stored in GlycoStore is the average GU across all sources. GlycanAnalyzer uses this average GU evidence for its H/UPLC analysis along with mass information (thus, aiding H/UPLC-MS annotation). Although there is slight variation of GU values across different sources we found that a standard deviation of less than 0.4 to be acceptable for our GlycanAnalyzer uses.

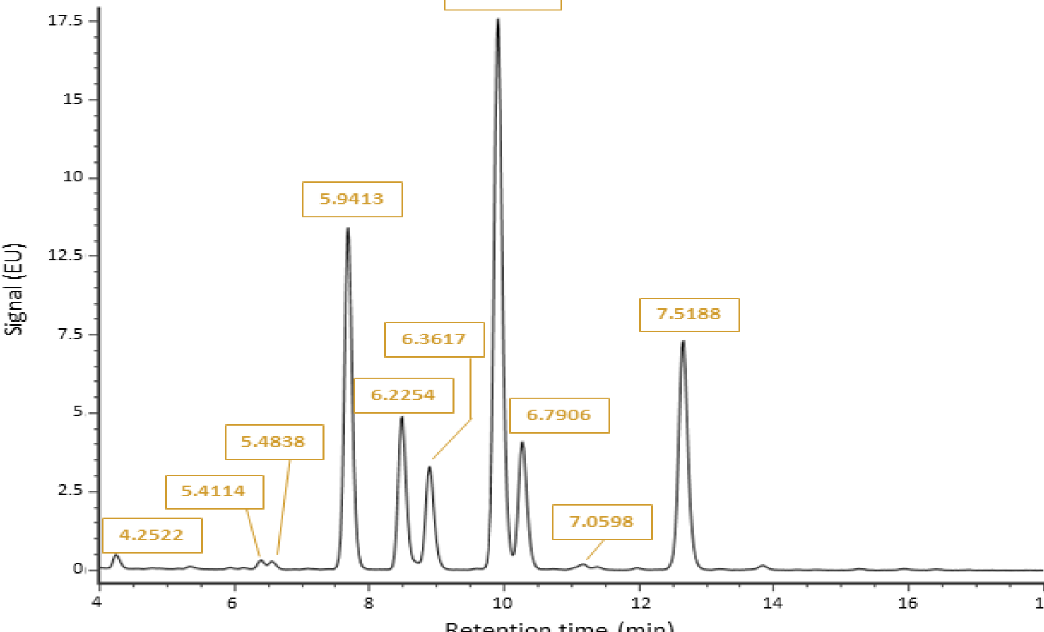
This originating source is a very valuable piece of information as it can constrain our search when the user knows what type of sample they are analysing. For example, if the user selects human IgG in the input page when analysing any monoclonal antibodies, the software only considers evidence from immunoglobulins (e.g. IgG, IgG1, IgA, etc.) in our database. It will not consider N-glycans with more than two galactose or sialic acid monosaccharides, thus greatly reducing the computational time and increasing the accuracy of the calculation.

A

Input: peak list

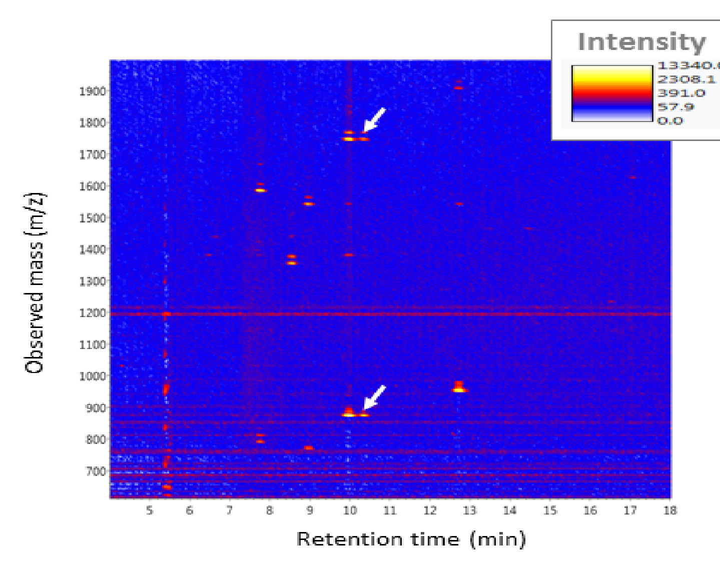


Retention time (min)	Amount (%)	Glucose Units
4.24	0.74	4.2522
6.39	0.46	5.4114
6.55	0.48	5.4838
7.68	20.64	5.9413
8.48	9.74	6.2254
8.89	6.82	6.3617
9.9	35.66	6.6803
10.27	8.63	6.7906
11.16	0.49	7.0598
12.65	16.34	7.5188

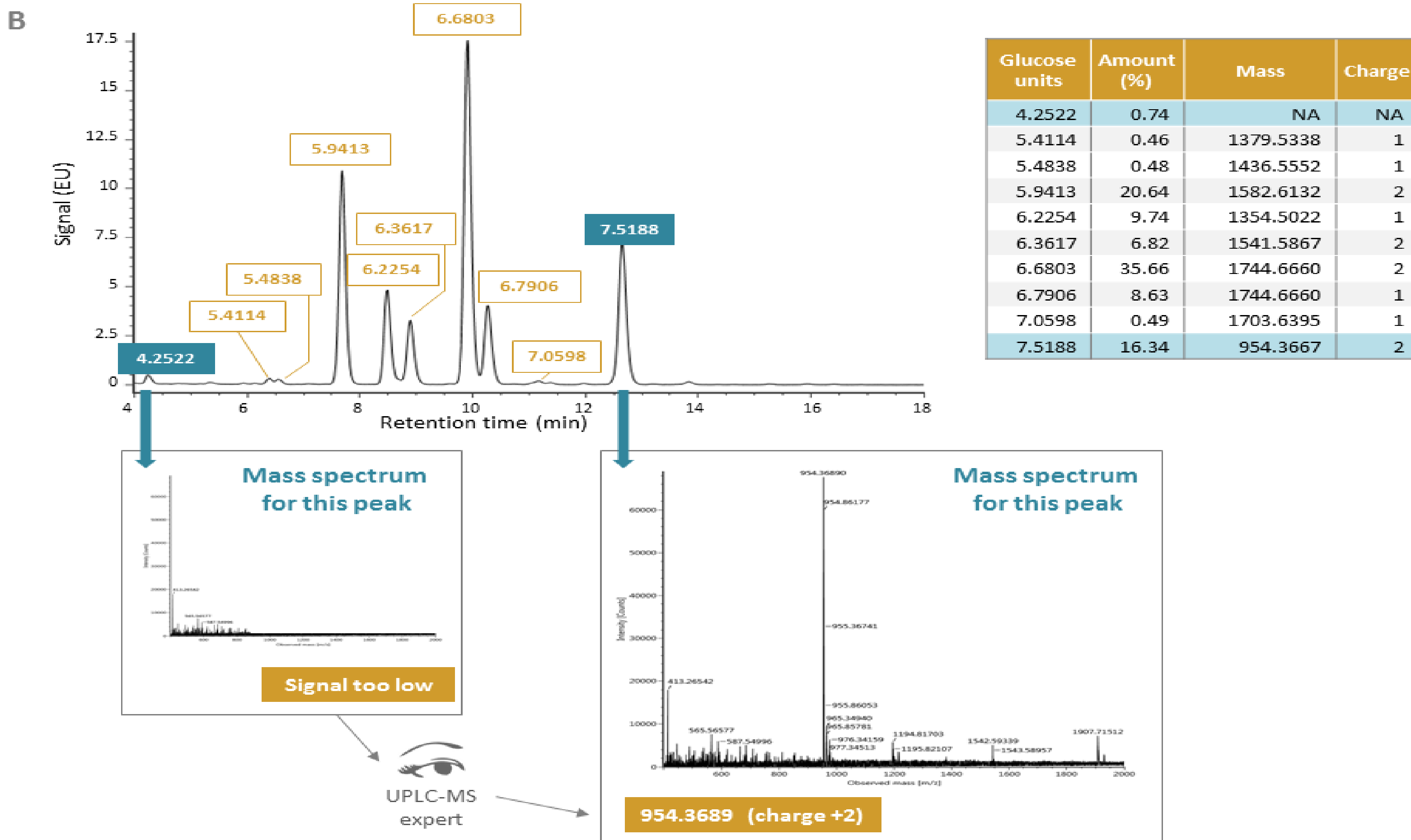


Input: 3D mass list

Retention time (min)	Observed mass (m/z)	Intensity (Counts)
...	...	...
...	...	...
10.33	1746.676	180934
10.33	1767.649	65209
10.33	1769.661	35679
10.33	1381.54	18326
10.33	885.3312	15456
10.33	892.3186	14856
10.33	895.319	13064
10.34	873.341	255720
10.34	1745.673	224206
10.34	873.8413	211368
10.34	874.343	105314
10.34	1747.677	100713
10.34	1768.656	55035
10.34	874.8418	48884
10.34	884.3282	32779
10.34	884.8331	30815
10.35	1748.673	41064
10.35	1380.542	34766
...	...	...
...	...	...



B



Glucose units	Amount (%)	Mass	Charge
4.2522	0.74	NA	NA
5.4114	0.46	1379.5338	1
5.4838	0.48	1436.5552	1
5.9413	20.64	1582.6132	2
6.2254	9.74	1354.5022	1
6.3617	6.82	1541.5867	2
6.6803	35.66	1744.6660	2
6.7906	8.63	1744.6660	1
7.0598	0.49	1703.6395	1
7.5188	16.34	954.3667	2

Supplementary Figure 4. GlycanAnalyzer uses the Glucose Unit (GU) mobility reference and all input is supplied through tab-separated text placed into text areas. (A) Tab separated input formats for UPLC peaks and the 3-dimensional mass information. Note that users must supply the 'Amount (%)' column which quantifies (relatively) the N-glycan(s) eluting at that peak. (B) The user can manually extract the mass and charge information. In this example no *m/z* information is supplied for UPLC peak at GU 4.2522 and only one *m/z* value (954.3689 in +2H charge state) is supplied for peak at GU 7.5188. However more than one *m/z* values can be supplied for any peak (e.g. 1907.715 in the +1H charge state is also found at the peak with GU 7.5188).

4.2 Software output

The example output shown in Supplementary Figure 5 shows the same ABS+BKF+BTG digestion highlighted in Supplementary Figure 3. In this example one peak was chosen to be analyzed and a ranked list of N-glycans was returned (shown in Supplementary Figure 5A). The software also allows us to visualize the peak movements after exoglycosidase digestion (Supplementary Figure 5B).

Single peak assignment page

GlycanAnalyzer's output has two features that help users to accept or reject the assignments. The first is the definition of a score used to rank the N-glycan assignments. The second is the incorporation of multiple sets of evidence for the assignment. The evidence include: (i) the GU similarity to known average GU values in GlycoStore and GlycoBase (Supplementary Figure 5, part 5), (ii) matching theoretical (expected) glycan masses to observed masses found in the MS spectra (Supplementary Figure 5, part 5 and 8) and, perhaps for the first time, (iii) tracing the sequential shifting of digested peaks (both mass and GU shifts). How both mass and GU shifts contribute to the score at each stage of the exoglycosidase array can be accessed in the "score calculation" link (Supplementary Figure 5, part 8). Assignments can be rejected by the user, for example the alpha-galactose structure can be rejected in Supplementary Figure 5, part 2 as it is ranked quite badly (as seen by its high score). Links to databases GlycoStore and GlyTouCan (Aoki-Kinoshita, et al., 2016) are also available to direct the user to further information for the assignment. Note that databases such as UniCarbKB and UniCarbDB were not chosen to be linked as GlyTouCan already has excellent linkage to them.

The output is graphical in nature with each N-glycan shown in established CFG/Oxford diagram form (Supplementary Figure 5, part 4). The shifting peaks can be visualized as a directed graph (Supplementary Figure 5, part 8; 'peak shift graph' link). Other graphical outputs include the highlighted shifting UPLC peak with corresponding mass spectra, both highlighted in red, with glycan signals annotated (Supplementary Figure 5B).

Summary page

The Summary page highlights all the peaks that were accepted by the user, gives distributions of monosaccharide types in currently accepted peaks and summarizes the assignments for each peak (Supplementary Figure 6). The summary page also allows users to reject assignments. For example, they may disagree with some of the assignments (i.e. no alpha galactose or mass does not match) in which case they can decide to reject them. All graphics and output on the summary page will be updated after removing the rejected glycan.

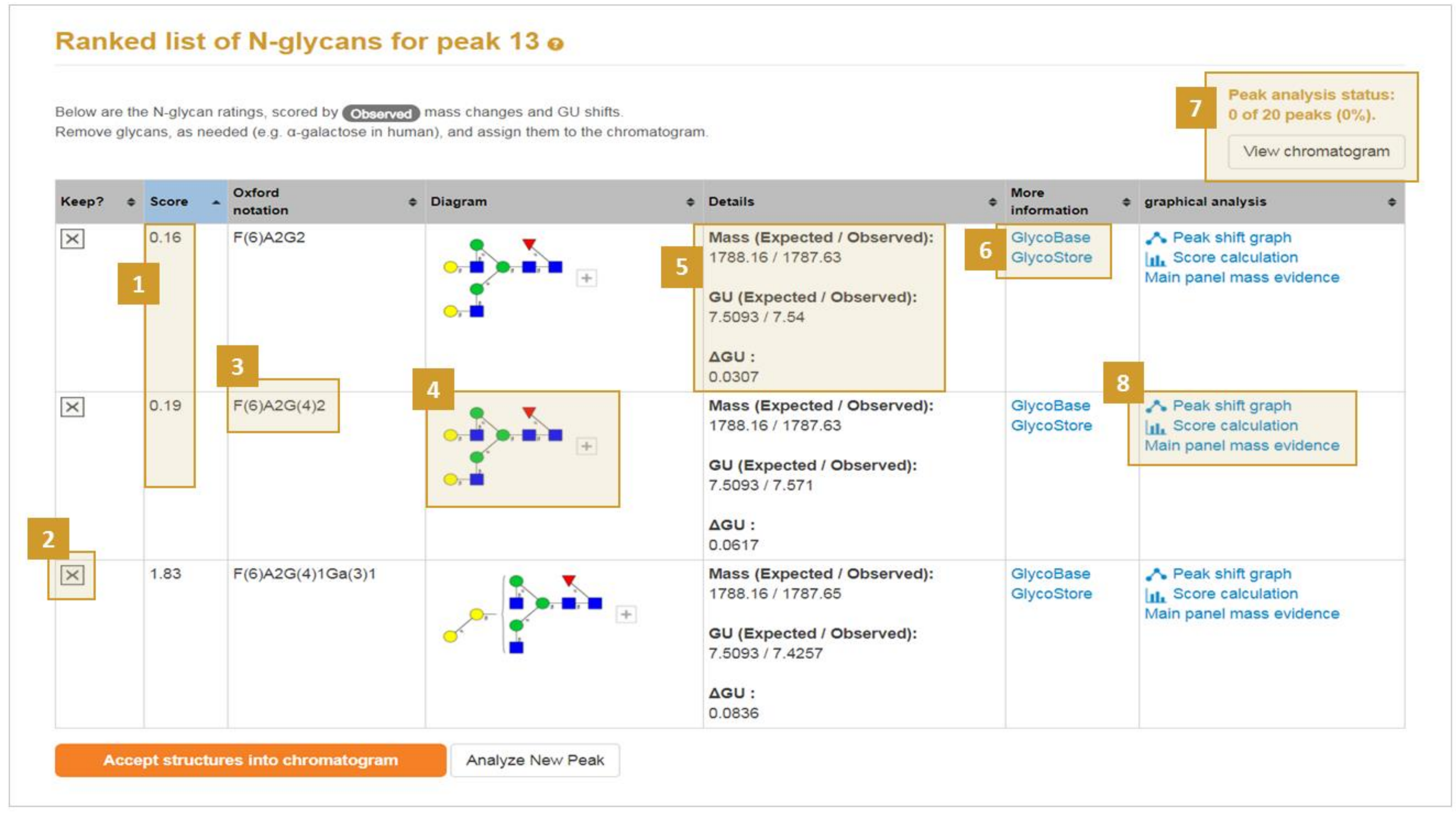
One of the most important sections on the summary page is the 'Evidence' column in the tables. Three possible evidences are given for the N-glycan assignment: (i) closeness of the experimental GU value to previous GU evidence in GlycoStore (termed GU similarity), (ii) if the mass was found in the peak of interest and (iii) if there is evidence of UPLC peak shifting after exoglycosidase digest.

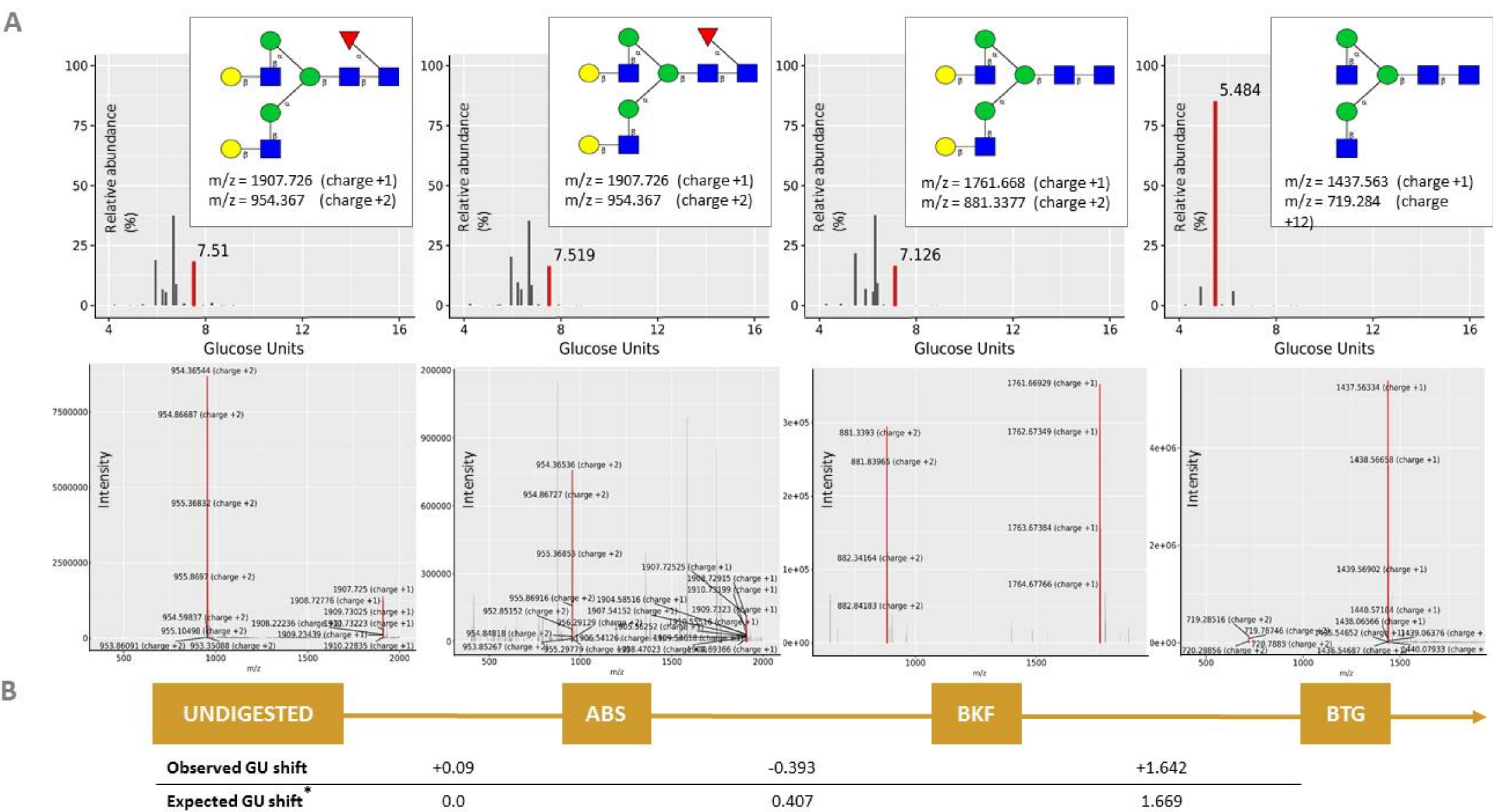
GlycanAnalyzer does not use threshold values to assign confidence levels to its annotations. Instead, three confidence level for each glycan annotation can be defined as follows:

(i) ***Weak***: There are no mass or peak shifts detected. Glycans are assigned using similarity to database GU values only.

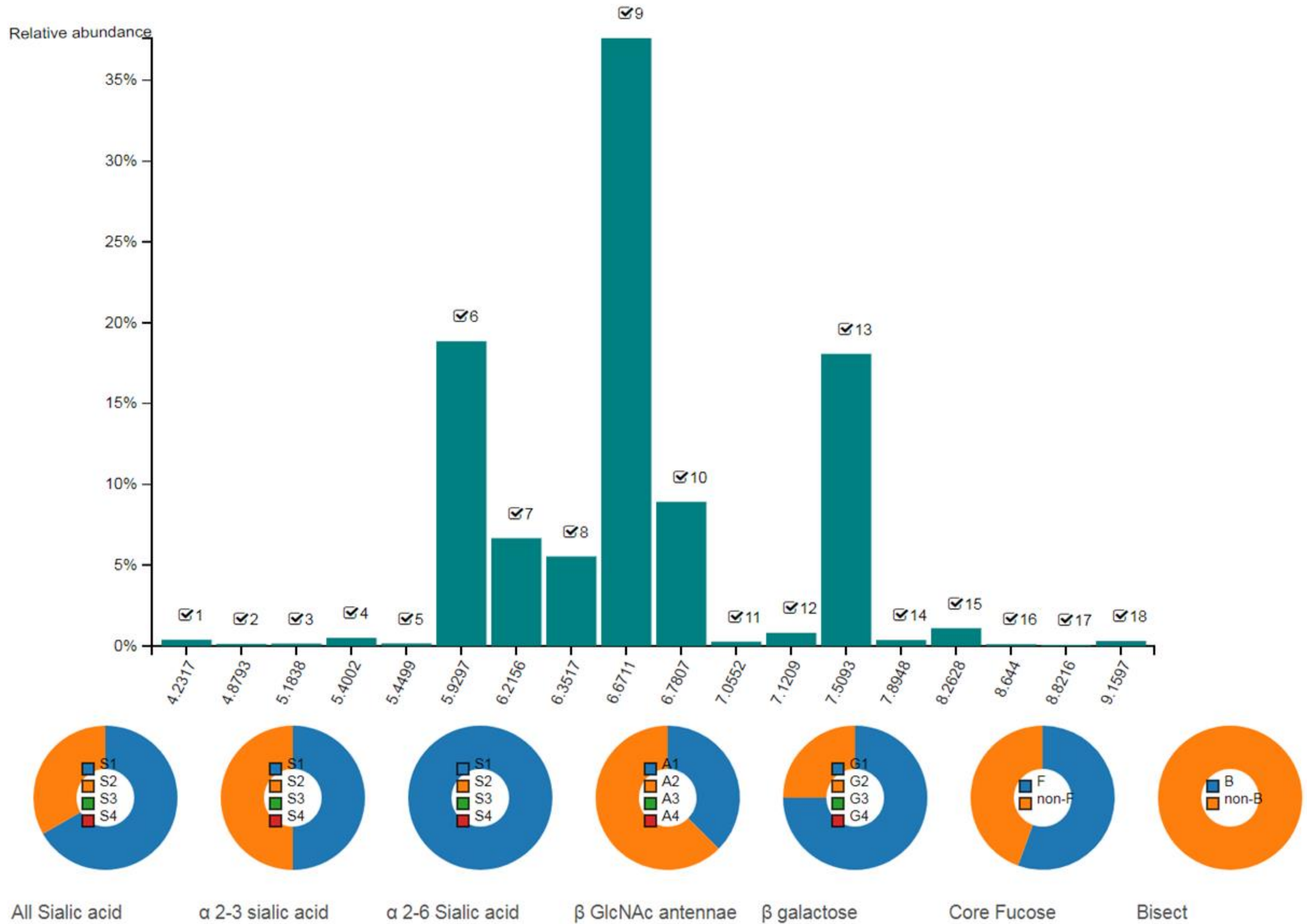
(ii) ***Medium***: There is a mass detected and a similar GU in the database for the glycan annotation. However, no peak shifts could be detected – this often happens for smaller peaks.

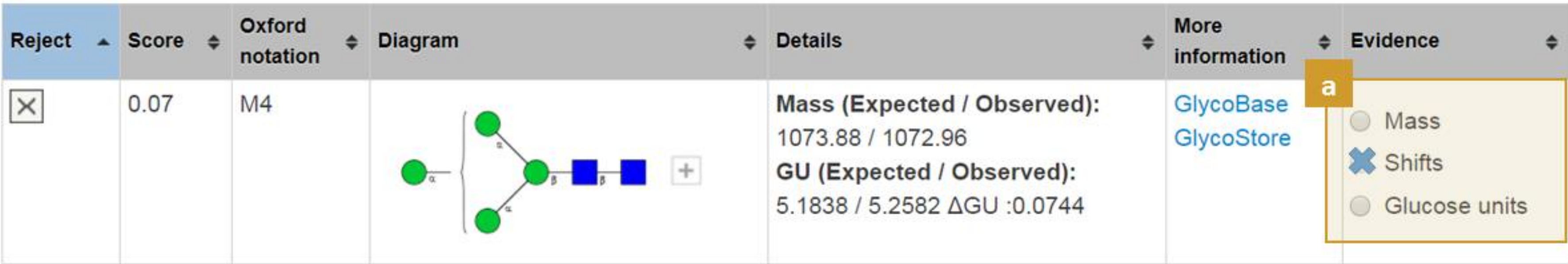
(iii) ***High***: There was a mass detected, a similar GU found in the database and the peaks movements could be traced from the Undigested to profile to the final exoglycosidase applied.

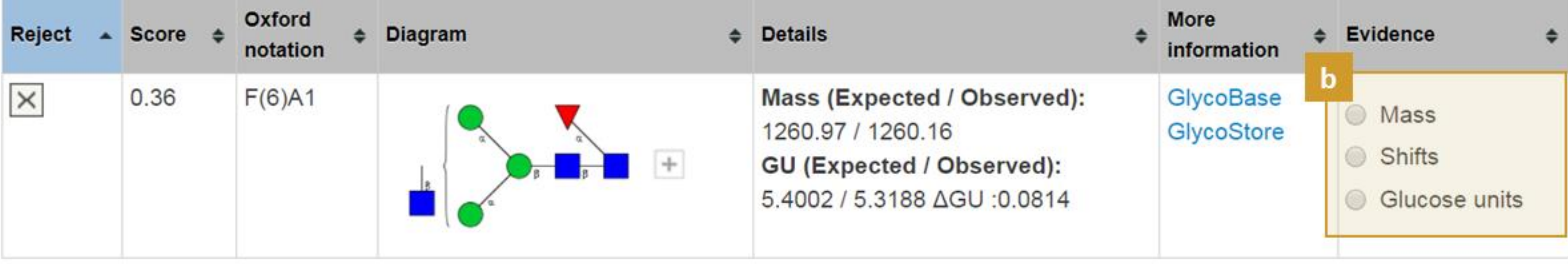




Supplementary Figure 5. Partial output in GlycanAnalyzer for peak 13 in our anti-Her2 antibody. The data for the antibody is available in the tutorials online. (A) 1: The returned N-glycans are ranked by a score (closer to zero the better). 2: The user has the option to reject any candidate N-glycan. 3&4: N-glycans are displayed in oxford notation and drawn in CFG/oxford notation - click the '+' symbol for higher resolution glycan image. 5: The details column gives some clues on the correctness of the assignment, for example how close the theoretical masses and GU values are to the isotopic average mass and the GU. 6: Links to other databases. 7: If the user chooses to 'accept structures into chromatogram' the number of completed peaks will increase. Clicking 'view chromatogram' gives a summary of currently completed peaks. 8: More visualization is available when these links are clicked. (B) Clicking 'Score calculation' (A part 8) reveals further graphics showing the peak movements and how the monosaccharide cleavages contribute to the score.







Supplementary Figure 6.The summary output page for the anti-Her2 antibody available in the tutorials (Supplementary Table 2). This view is presented when clicking 'view chromatogram' in Supplementary Figure 5. The bar chart shows each peak's relative abundance where a tick indicates user acceptance of the peak. Bars can be clicked to reveal peak shifting in the exoglycosidase array. The pie charts give the distribution of sialic acids (S1, S2, S3 and S4 means the % mono, bi, tri and tetra sialylated species respectively), GlcNAc antennae (A1, A2, A3 and A4 means the % mono, bi, tri and tetra GlcNAc antennae respectively), β-galactose (G1, G2, G3 and G4 means the % mono, bi, tri and tetra β- galactose species respectively), Core Fucose (F: % core fucose, non-F: % not core fucose) and Bisect (B: % core fucose, non-B: % not core fucose). Although the chromatogram has 18 peaks, for conciseness only peaks 3 and 4 are shown as an example. (a) Peak 3 has two pieces of evidence: an observed mass and similar GU to average evidence in public databases. Peak 4 has three pieces of evidence: mass, shifting peaks and similar GU to average evidence in public databases. N-glycan assignment for Peak 4 can be considered to have strong support while assignment for Peak 3 has medium support. The weakest level of supporting evidence is GU similarity alone. The Mass, Shifts and Glucose unit buttons can be clicked to visualize the evidence.

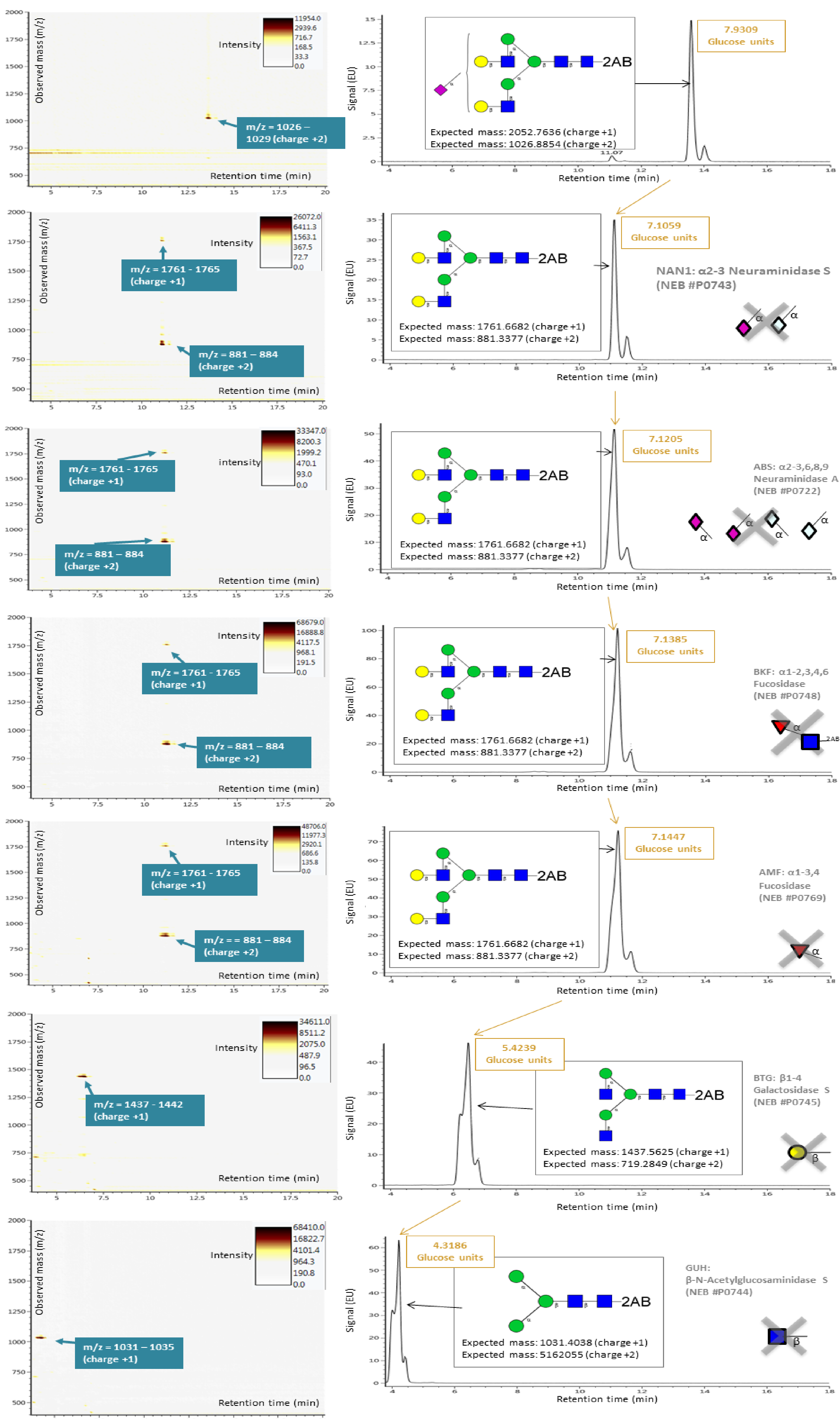
5. GlycanAnalyzer algorithm and scoring

5.1 Algorithm

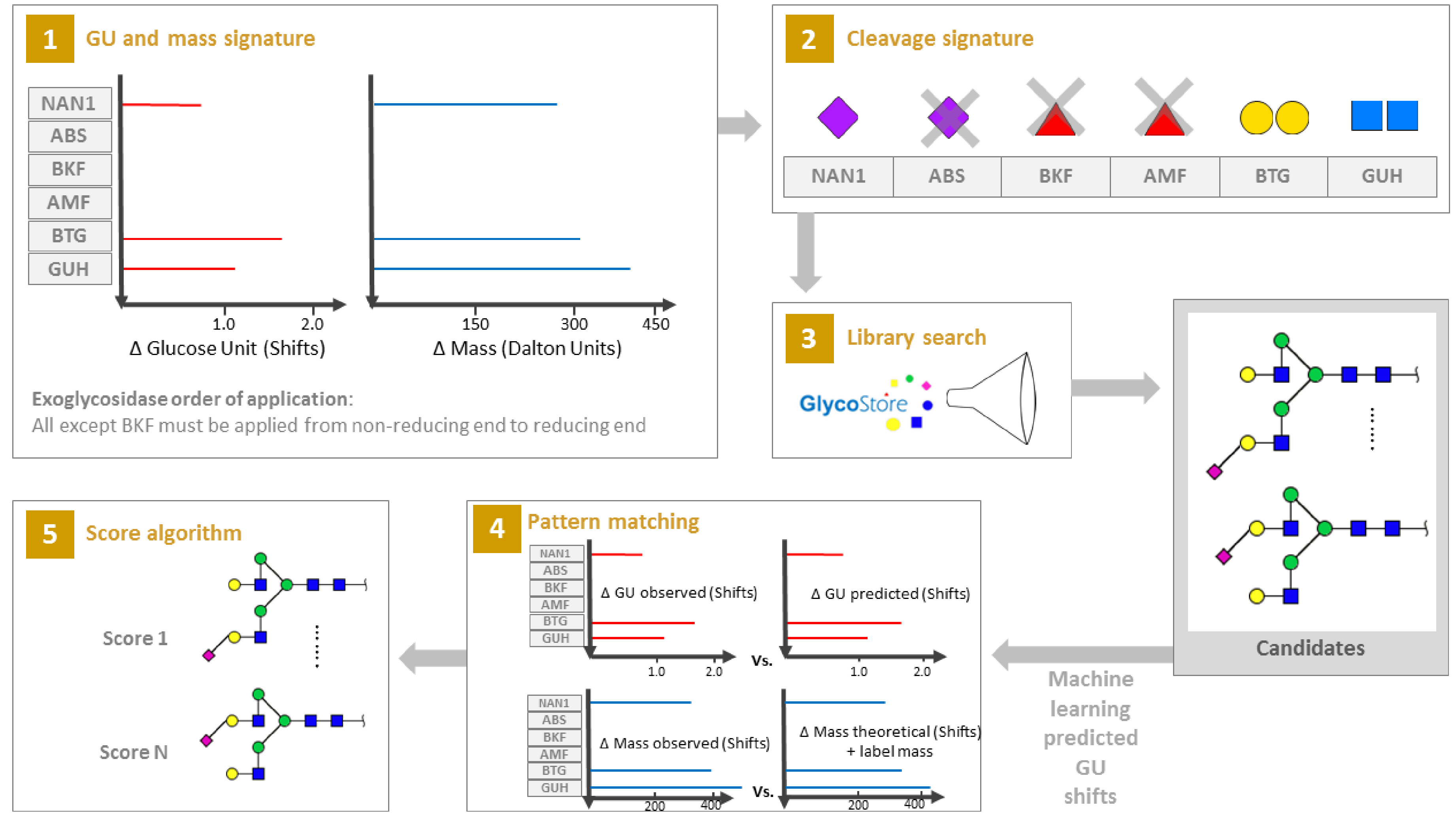
In essence, the algorithm monitors the peak movements in the H/UPLC and MS, looks for patterns of peak movements and assigns N-glycan(s) with a score dependent on the closeness of the pattern match. In order to describe the algorithm in simple terms and to test the software, four standards were purchased (Chemily, GlycoScience) and run on a HILIC-UPLC on an ACQUITY UPLC H-Class (Waters Corporation, MA, USA) with a fluorescence detector coupled online to a Xevo G2-S QTof mass spectrometer. Supplementary Figure 7 shows the mass spectra and UPLC peak shifts for one of the standards (the other three are shown in the Appendix). From the MS data, mass changes corresponding to monosaccharide losses are observed. In other words the *m/z* changes are precise, for example if we apply NAN1 and cleave one sialic acid then we know what mass to expect in the corresponding exoglycosidase (see Supplementary Figure 7 NAN1 profile). In the H/UPLC a difficulty lies in the fact that the GU shifts vary considerably depending on the glycan undergoing exoglycosidase application. Therefore GlycanAnalyzer implemented a predictive machine learning model for each exoglycosidase that predicts a theoretical GU shift. The five main stages of the algorithm are highlighted in Supplementary Figure 8 using the standard shown in Supplementary Figure 7 as an example. The five stages include (see Supplementary Figure 8 for reference):

1. Generation of all the possible GU and mass shifts for the peak of interest – two signatures are produced for the GU and mass separately.
2. Using the mass shifts to generate a cleavage signature
3. Using the cleavage signature to search libraries in GlycoStore. The more libraries we have the more intelligent the assignment will be.
4. Generation of a theoretical GU shift for the exoglycosidase array using predictive machine learning modules. Generation of predicted/theoretical mass shifts for the exoglycosidase array (note that mass of monosaccharides are known this is a direct calculation). Matching the theoretical GU/mass shifting pattern to the observed pattern generated from the UPLC-MS experiment.
5. Scoring the glycans using the difference between the predicted/theoretical GU/mass shifts and observed GU/mass shifts. The scoring is vital as many candidates can be returned by step 3 and also many possible peak movements are possible as can be seen in Supplementary Figure 9.

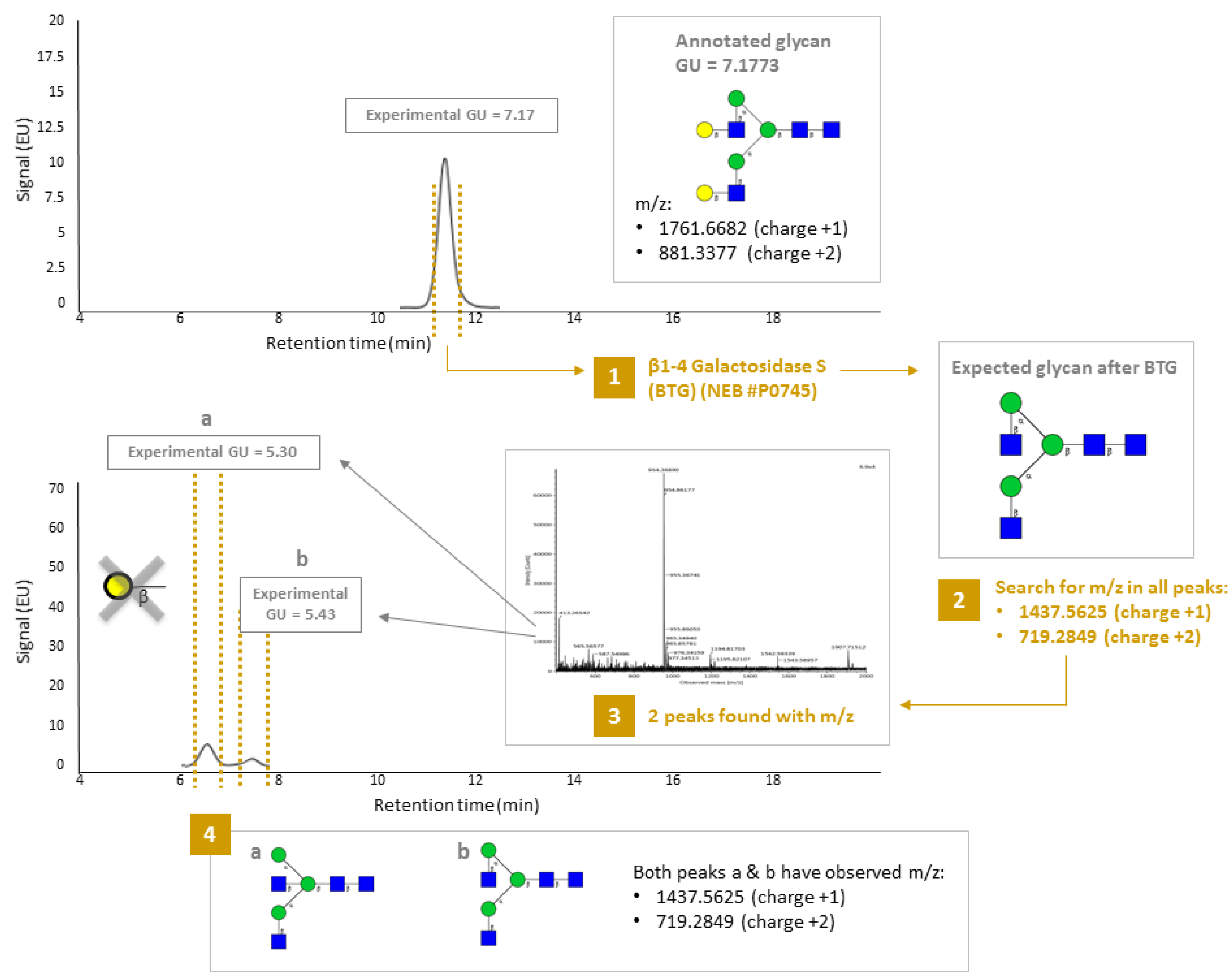
Note when mass is not supplied as an experimental value it is estimated as shown in Supplementary Figure 10.



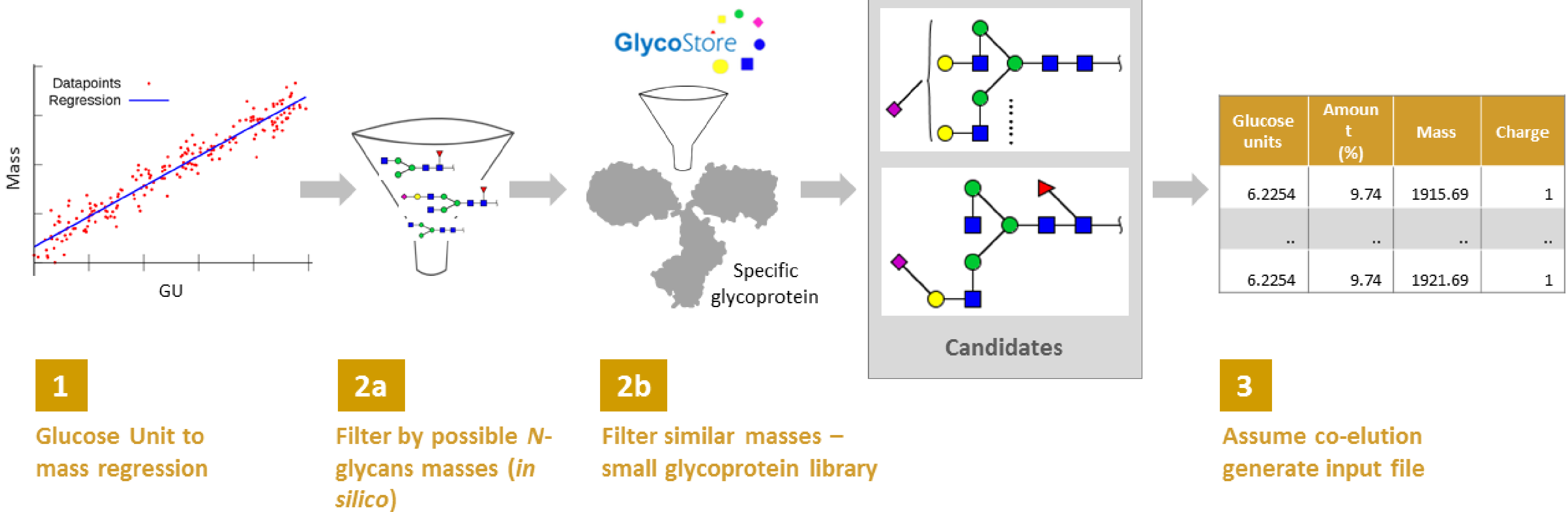
Supplementary Figure 7. The GU, RT and *m/z* movements for a synthetic N-glycan standard containing 1 α2-3 sialic acid, 2 β1-4 galactose, and 2 GlcNAc antennae. The glycan was digested sequentially by the exoglycosidases: ABS+NAN1+ABS+BKF+AMF+BTG+GUH (see supplementary Figure 2 for specificities). All GU and *m/z* shifts and were as expected for the cleaved monosaccharides.



Supplementary Figure 8. The five basic steps of the algorithm. The key information lies in the GU and mass shifts from the UPLC and MS peaks respectively. Another important component is the use of libraries contained in GlycoStore.



Supplementary Figure 9. In GlycanAnalyzer, computation of multiple peak shifts are possible. This hypothetical example shows a very simple case for the application of BTG exoglycosidase to a G2 structure. Since there are two possibilities for peak movement GlycanAnalyzer's machine learning module will predict which peak is the correct one based on the observed ΔGU peak shift. With the application of multiple enzymes on complex samples (i.e., a lot of peaks), multiple peak shifts can make assignment quite difficult. This example does not represent any data in the anti-Her2 antibody used throughout this supplement.



Supplementary Figure 10. A technique to estimate the mass using the GU values. The advantage of this approach is the reuse of the five step algorithm shown in Supplementary Figure 8 without changing the GlycanAnalyzer implementation. Note that 2b does not have to be selected and in this case all feasible N-glycans are used as a filter. However, the assignment accuracy is diminished when mass is estimated.

5.2 Scoring

The following is a mathematical description of the score used in GlycanAnalyzer. Supplementary Figure 11 shows the calculation for two example candidates. If the reader finds the mathematical terms difficult to understand, Supplementary Figure 11 can be used for clarity.

Candidate and score

Let **c** be an N-glycan candidate. One of the major goals of GlycanAnalyzer is to rank the candidate **c** using a score S(**c**). We define the best score to be 0 and the following are the major components of the score.

GU definitions

1. *GU similarity,* $\Delta GU^i()$: Let GU_o^i be the observed GU at peak *i* in an UPLC chromatogram and $GU_{db}(\boldsymbol{c})$ be the average GU for the candidate **c** found in the GlycoStore database. One factor in the candidate's score is the similarity between the observed GU and the average GU for the candidate: $\Delta GU^i(\boldsymbol{c}) = |GU_0^i - GU_{db}(\boldsymbol{c})|$. Note that GlycoStore GU values using different florescent labels do not vary much, hence an average of all fluorescent labels is used for $GU_{db}(\boldsymbol{c})$.
2. *Peak shift score,* $PSS^{ij}()$: A regression model, R(**c**, e, θ), was defined given a candidate, c, and exoglycosidase, e, by machine learning optimization of parameters, θ, using the following input and target:

   Input: the candidate **c** and the exoglycosidase applied e.

   Target: Experimental derived GU shifts in the databases GlycoStore and GlycoBase given a candidate **c** and exoglycosidase applied e.

   Optimization: find θ using gradient descent and neural network.

   The model, R(**c**, e, θ), once optimized can then predict GU shifts for candidate **c** given an applied exoglycosidase e. For example, R(, ABS, θ) = 1.96 ΔGU as it cleaves two sialic acids in the α2-6 linkage conformation. Using this regression model a peak shift can be scored.

   Let $GU_o^i(e_-)$ be the observed GU at peak *i* in the UPLC profile before digestion with exoglycosidase e and let $GU_o^j(e_+)$ be the observed GU at peak *j* in the UPLC profile after digestion with exoglycosidase e. Then $\Delta GU^{ij}(e) = |GU_o^j(e_-) - GU_o^i(e_+)|$ is the GU difference between peaks *i* and peak *j* after application of exoglycosidase e. A peak shift can then be scored given a candidate **c** as follows: $PSS^{ij}(\mathbf{c}, \mathrm{e}, \theta) = |\mathrm{R}(\mathbf{c}, \mathrm{e}, \theta) - \Delta GU^{ij}(e)|$.
3. *Total GU score,* $S_{GU}(\boldsymbol{c}) = \Delta GU^i(\boldsymbol{c}) + PSS^{ij}(\mathbf{c}, e_1, \theta) + \cdots + PSS^{ij}(\mathbf{c}, e_n, \theta)$ where the exoglycosidase array is $\underset{e}{\rightarrow} = [e_1, \ldots, e_n]$ (e.g. e = [Undigested, ABS, BKF, BTG, GUH] in our human IgG data described throughout this supplement)

Mass definitions

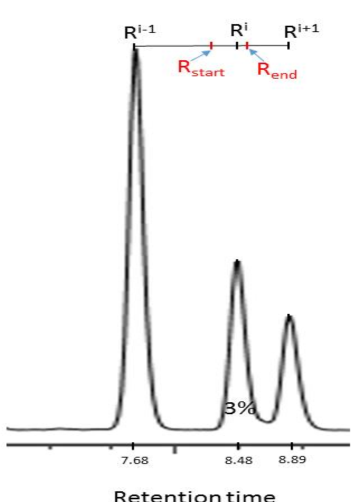
Let **c** be an N-glycan candidate

1. *Expected mass of unmodified candidate **c** in neutral state*, EM(**c**): This is the mass of the N-glycan in its neutral label free state. It is the sum of the individual monosaccharide masses in candidate c plus water loss from hydrolysis. For example, EM( ) is: 1 fucose + 4 GlcNAc +3 mannose + 2 galactose + 2 sialic acid + H_2O = 1x146.0579+4x203.0794+3x162.0528+2x162.0528 +2x291.0954 +18.01528 = 2368.846 Daltons.

2. In positive mode electrospray mass spectrometry the candidate glycan must be positively charged to be detected. Depending on the buffer used, glycan adducts containing protons (H), sodium (Na), lithium (Li), potassium (K) and combinations thereof may be detected in the mass spectrum. In this case, there can be a mixture of ions (+1H), (+1H+1Na), (+1H+1K), (+1K), (+1Na), (+1K+1Na) etc. The simplest expected mass to calculate is the addition of 1 proton: EM(**c**+1) = (EM(**c**) + 1.00730) / 1 (where 1.00730 is the mass of a proton). This is known to have a charge of +1. Addition of two protons results in an expected mass of EM(**c**+2) = (EM(**c**) + 1.00730 + 1.00730) / 2. In general, for p protons the charged mass is EM(**c**+p) = (EM(**c**) + p*1.00730) / p. For an adduct with n ions the mass becomes: EM(**c**, n)$^{1,..,n}$ = (EM(**c**) + mi_1+..+mi_n) / n where mi_x is the monoisotopic mass of ion x in the adduct. The monoisotopic masses supported by GlycanAnalyzer are K=38.96320, Na=22.98920, Li= 7.01550 and H=1.00730.

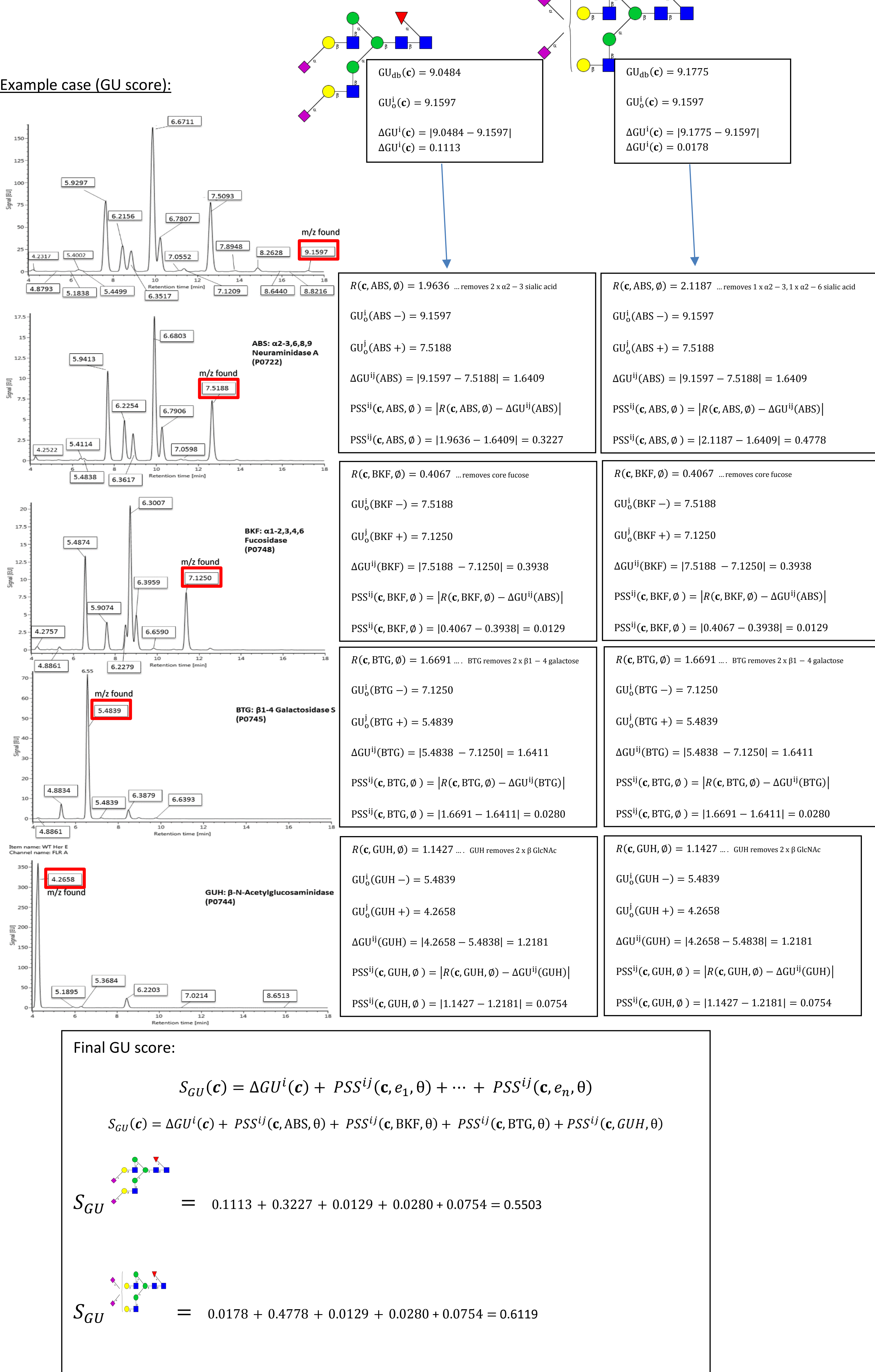
3. In GlycanAnalyzer, the user can select if the mass data was generated using a fluorescent label. GlycanAnalyzer supports labels 2-AB (mass=120.0687 daltons), RFMS (mass=311.1746 daltons), procainamide (mass=219.173 daltons) and the option to explicitly input any labels mass. When a label has been selected the estimated mass of an adduct with n ions becomes: EM(**c**, n, lm)$^{1,..,n}$ = (EM(**c**) + mi_1+..+mi_n) / n + lm where lm is the mass of the fluorescent label and mi_x is the monoisotopic mass of ion x (either K, Na, Li or H). Note lm=0 if user selects label free.

4. *Retention time range for peak i*, $[R_{start},..,R^i,..,R_{end}]$. Each peak in the UPLC chromatogram is defined by its GU value (GU_o^i above), its percentage area (relative abundance of the N-glycans in that peak), and its retention time (Supplementary Figure 4a). Let R^i denote the retention time of peak *i*. Since the eluting N-glycans will not always have their mass spectra exactly at the retention time R^i GlycanAnalyzer examines the observed masses for a range/window surrounding R^i. This range is denoted $[R_{start},..,R^i,..,R_{end}]$. If the peak has a relative abundance above 5% (i.e. a large peak), R_{start} is the retention time halfway between R^i and the previous peak's retention time R^{i-1} and R_{end} is the retention time halfway between R^i and the next peak's retention time R^{i+1}. If the peak is medium sized (1-5% relative abundance) then R_{start} is the retention time 1/3 of the way between R^i and R^{i-1} and R_{end} is the retention time 1/3 of the way between R^i and R^{i+1}. If the peak is small (<1% relative abundance) then R_{start} is the retention time 1/4 of the way between R^i and the previous peaks retention time R^{i-1} and R_{end} is the retention time 1/4 of the way between R^i and R^{i+1}. An example is shown in the diagram below for a medium peak:



5. Let $MS[e-][R_{start},..,R^i,..,R_{end}]$ be the mass spectrum between the retention time range $[R_{start},..,R^i,..,R_{end}]$ at peak i in the H/UPLC before application of exoglycosidase e. Let $MS[e+][R_{start},..,R^j,..,R_{end}]$ be the mass spectrum between the retention time range $[R_{start},..,R^j,..,R_{end}]$ at peak j in the H/UPLC after application of exoglycosidase e.
6. Using at most one of the K, Na, Li or H ions (+1 charge) define the set of possible +1 charged masses for candidate **c** to be C1M=[EM(c,1,lm)^{k}, EM(c,1,lm)^{Na}, EM(c,1,lm)^{Li} , EM(c,1,lm)^{h}] where lm is the mass of the label (lm=0 if label free). Using all combinations of at most two K, Na, Li and H ions define the set of possible +2 charged mass candidates to be C2M=[EM(c,2,lm)^{H,K}, EM(c,2,lm)^{H,Na}, EM(c,2,lm)^{H,Li}, EM(c,2,lm)^{K,Na}, EM(c,2,lm)^{H,Li}, EM(c,2,lm)^{Na,Li}, EM(c,2,lm)^{Na,Na}, EM(c,2,lm)^{Li,Li}, EM(c,2,lm)^{K,K}, EM(c,2,lm)^{H,H}] where lm is the mass of the label (lm=0 if label free).
7. Let $C1M[e-]$ be the +1 charged masses before application of exoglycosidase e. Let $C1M[e+]$ be the +1 charged masses after application of exoglycsidase e. $C1M[e+]$ can be easily calculated using the specificities in Figure 1 and known masses of monosaccharides. Similarly for C2M.
8. *Mass shift score applying exoglycosidase e*, MSS^{ij}(e): if any mass in $C1M[e-]$ or $C2M[e-]$ is found in $MS[e-][R_{start},..,R^i,..,R_{end}]$ then $SS^i = 0$ otherwise $SS^i = 1$. if any mass in $C1M[e+]$ or $C2M[e+]$ is found in $MS[e+][R_{start},..,R^j,..,R_{end}]$ then $SS^j = 0$ otherwise $SS^j = 1$. MSS^{ij}(e) = $SS^i + SS^j$.
9. *Total mass score*, $S_{mass}(\boldsymbol{c}) = MSS^{ij}(e_1) + \cdots + MSS^{ij}(e_n)$ where the exoglycosidase array is $\underset{e}{\rightarrow} = [e_1,..,e_n]$ (e.g. e = [Undigested, ABS, BKF, BTG, GUH] in our human IgG data described throughout this supplement)

The glycan total score output by GlycanAnalyzer is $S_{mass}(\boldsymbol{c}) + S_{GU}(\boldsymbol{c})$. Note if no mass information is supplied then $S_{mass}(\boldsymbol{c}) = 0$ by default.



Supplementary Figure 11. An example calculation of the score for two glycan candidates. This is an actual calculation from the human IgG dataset used throughout this supplement.

6. Additional results and tables.

Software	H/UPLC input/ confirmation	Mass spectra input/confirmation	Accuracy~	Machine learning#	Parameters optimized on	Supports Co-elution^	Handles partial digestion*	Annotation Evidence
GlycoProfileAssigner	Yes	No	Medium	No	Human IgG only	No	No	GU only
GlycanAnalyzer	Yes	Yes	High	Yes	Over 6 monoclonal antibodies, human IgG, α1-antitrypsin	Yes	Yes	1. GU 2. Mass 3. GU & *m/z* peak shifts

Supplementary Table 1. GlycanAnalyzer is a clear improvement over GlycoProfileAssigner.

~ See Table 1 in main text for accuracy on Anti-Her2 antibody.

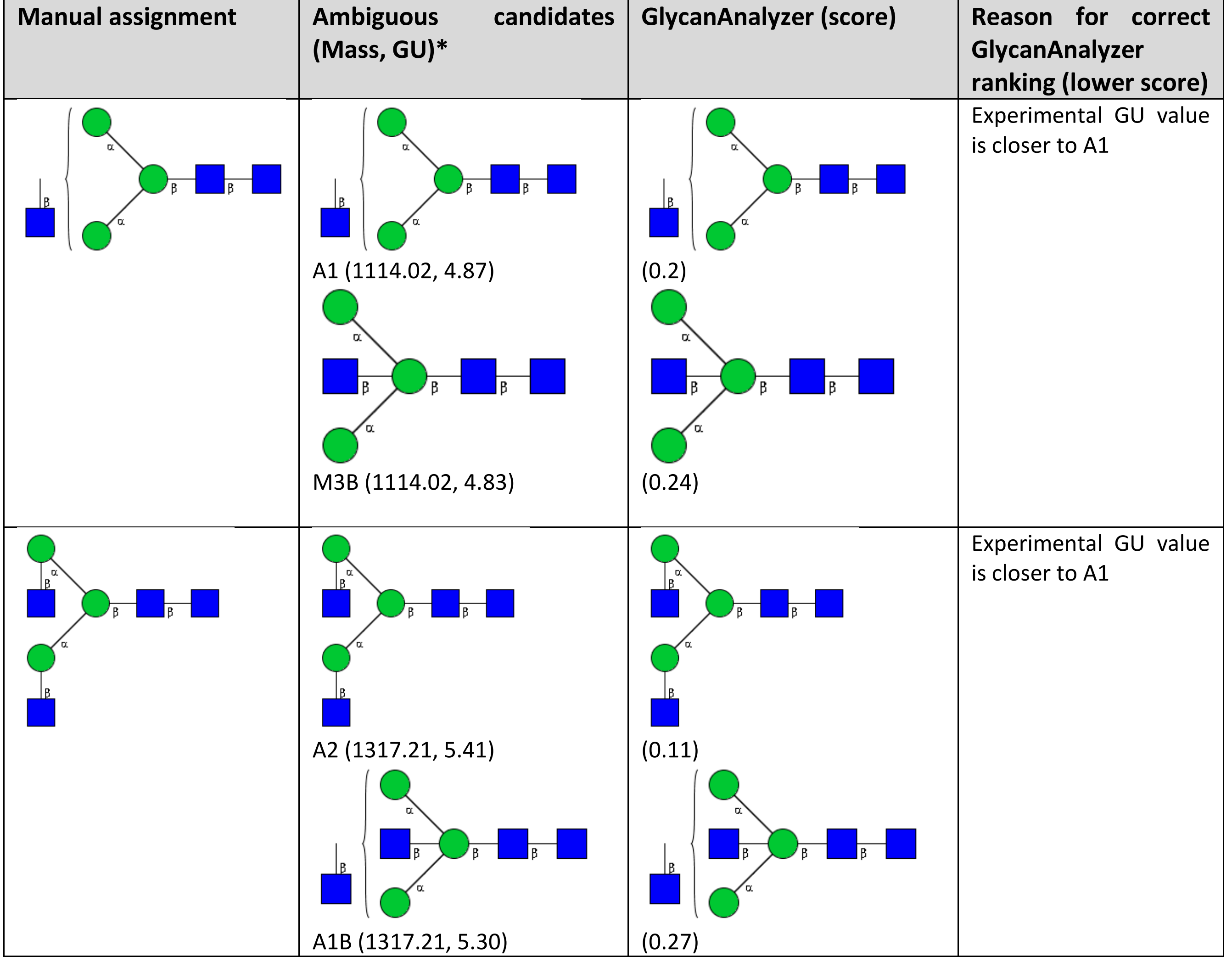
Theoretically as we generate more data the algorithm can be re-optimized to perform even better (i.e. learn).

^ Often more than one glycan elute in the same peak.

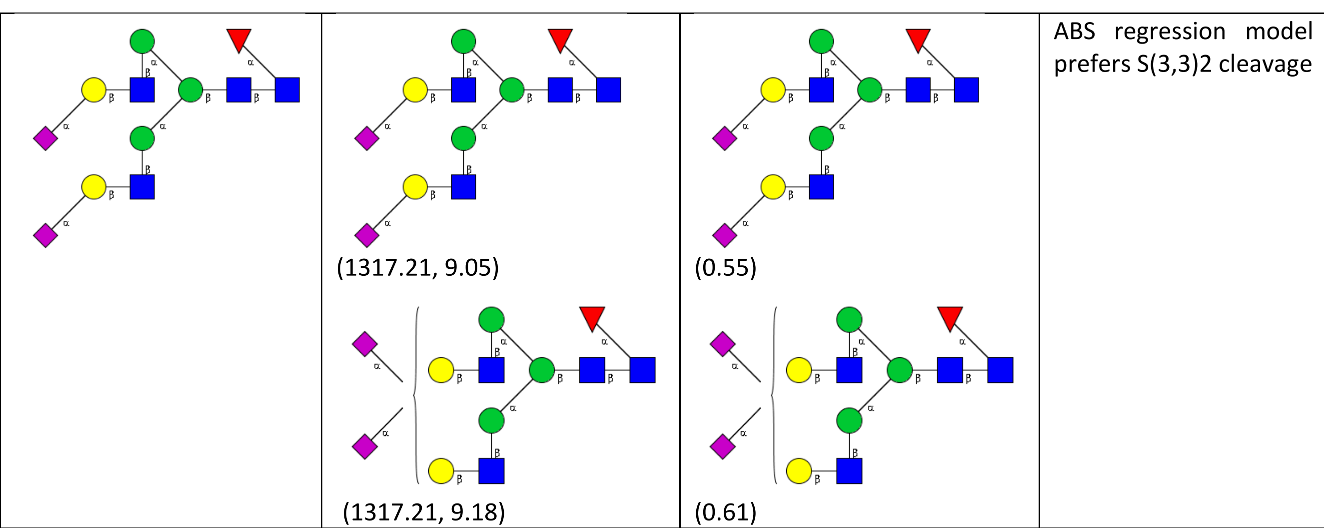
* Often due to certain experimental conditions an exoglycosidase digestion might not completely remove/digest a peak, this is often troublesome for calculations.

Tutorial	Goal
1	Guides the user through the input process and explains the output. The input process includes a peak list (for UPLC) and retention time, *m/z* and intensity (from the MS1).
2	Guides the user though the input process when they have determined the mass and charge states themselves. Shows how we can assign all peaks with one button click.
3	Guides the user though the input process when they have no mass information. The only information is the peaks (GU and area) from the UPLC. Accuracy will be lost in this situation.

Supplementary Table 2. A list of the tutorials available online with a brief description of what they achieve.



Manual assignment	Ambiguous candidates (Mass, GU)*	GlycanAnalyzer (score)	Reason for correct GlycanAnalyzer ranking (lower score)
	A1 (1114.02, 4.87) M3B (1114.02, 4.83)	(0.2) (0.24)	Experimental GU value is closer to A1
	A2 (1317.21, 5.41) A1B (1317.21, 5.30)	(0.11) (0.27)	Experimental GU value is closer to A1



Supplementary Table 3. Example cases where GlycanAnalyzer detected the correct glycan but Mass & GU matching alone was ambiguous. N-glycans released from monoclonal antibody. Note that software systems such as UNIFI do not rank but select. * Average GU and Mass extracted from GlycoStore.

Peak number*	Observed RT (min)	% Amount (%)	Observed Glucose Units	Observed *m/z*	Database Glucose Units	Theoretical *m/z*	GlycanAnalyzer assignment	Manual Assignment
6a	7.65	18.85	5.9297	1583.6194 (+1H) 792.3127 (+2H)	5.8534	1583.6205 (+1H) 792.3139 (+2H)		
6b	7.65	18.85	5.9297	1396.5378 (+1H)	5.7842	1396.5360 (+1H)		Not found
8a	8.86	5.54	6.3517	1599.6118 (+1H)	6.3572	1599.6154 (+1H)		Not found
8b	8.86	5.54	6.3517	1542.5938 (+1H) 771.8004 (+2H)	6.3578	1542.5939 (+1H) 771.8006 (+2H)		
12a	11.36	0.81	7.1209	1761.6770 (+1H)	7.1773	1761.6682(+1H)		Not found
12b	11.36	0.81	7.1209	1704.6439 (+1H) 852.8250 (+2H)	7.0347	1704.6467 (+1H) 852.8270 (+2H)		
14a	13.8	0.37	7.8948	1679.6153 (+1H)	7.8983	1679.6150 (+1H)		
14b	13.8	0.37	7.8948	1866.6932 (+1H) 933.8518 (+12H)	7.8436	1866.6995 (+1H) 933.8534 (+12H)		Not found

Supplementary Table 4. Example cases where GlycanAnalyzer detected co-eluting glycans in the anti-Her2 antibody. Our manual assignment failed to pick up this co-elution. See supplement file IgG_complete_annotation.xls for details. * in total there were 18 UPLC peaks integrated in the anti-Her2 antibody data for co-elution we number the glycans #a, #b etc..

7. Experimental protocol

Instruments: Chromatography was performed by HILIC-UPLC on an ACQUITY UPLC H-Class (Waters Corporation, MA, USA) with a fluorescence detector using an ACQUITY UPLC® BEH-Glycan column (1.7 µmn, 2.1 x 150 mm). The UPLC setup consist of a sample manager (at 10°C), a quaternary pump, a column oven (at 40°C) containing a Waters BEH Glycan column (2.1 mm inner diameter, 150 mm length, 1.7µm, 130 Å), and a fluorescence detector (Ex:330nm Em:420nm). The UPLC was coupled online to a Xevo G2-S QTof mass spectrometer (Waters Corporation, MA, USA). Data was acquired and processed using the UNIFI Scientific Information System (version 1.8.2).

Fluorescent labeling of four glycan standards: 0.5 nmol of each glycan standard (Chemily GlycoScience) was reconstituted in 25 uL of water, and was labelled with 20 µL of 2-AB labelling reagent (0.35M 2-AB, 1M Sodium cyanoborohydride dissolved in 7:3 DMSO: Acetic acid mixture) for 18 h at 37°C in a water bath. The 2-AB labeled samples were passed through a PD MiniTrap G-10 column (GE Healthcare) to remove excess dye, dried down in vacuum and subjected to an array of exoglycosidases in 20 µL of 50 mM sodium acetate buffer (pH 5.5) for 18 h at 37°C. The exoglycosidase array consisted of the following enzymes in order of their application: NAN1+ABS+BKF+AMF+BTG+GUH (New England Biolabs) where NAN1 is α2-3 Neuraminidase S, ABS is α2-3,6,8,9 Neuraminidase A, BKF is α1-2-4-6-fucosidase, BTG is β-1-3-4-galactosidase, AMF is α1-2 Fucosidase and GUH is β-n-acetylglucosaminidase-s (see Supplementary Figure 1 for specificities). Exoglycosidases were applied according to the manufacturer's protocol (New England Biolabs). The digested mixture was passed through a centrifugal filter cartridge with 10 kDa nominal molecular weight limit (NMWL) cut-off (Merck Millipore, Cork, Ireland) to separate the enzymes from the digested labelled glycans. The eluent containing digested labelled glycans was collected, dried in a vacuum, and reconstituted in 70% (v/v) acetonitrile in water for injection into a HILIC-UPLC-QTOF system.

Anti-Her2 antibody: A 100 µg aliquot of the anti-Her2 antibody was mixed with 500 U of PNGase F in the reaction buffer in a total volume of 100 µL and incubated at 37°C for 1 h. Such conditions will result in complete deglycosylation of IgG as indicated by SDS-PAGE-capillary gel electrophoresis (data not shown). The released glycans were then purified by HyperCarb porous graphitized carbon cartridge (Thermo Fisher Scientific, CA),and dried by CentriVap

(Labconco, Kansas City, MO) for ultra performance liquid chromatography combined with quadrupole time-of-flight (HILIC-UPLC-QTOF). *N*-glycans were labelled with 2-aminobenzamide (2-AB) according to a published protocol (Dell, et al., 1994). The excess 2-AB was removed by passing the labelling mixture through a MiniTrap G-10 desalting column (GEHealthcare) and the purified 2-AB-labeled glycans were then dried under vacuum. Before analysis, the dried samples were reconstituted in 250 µL of solvent consisting of 70% v/v acetonitrile in water and 10 µL injected for analysis as described above. The exoglycosidase array consisted of the following enzymes in order of their application: ABS+BKF+BTG+GUH (New England Biolabs) where ABS is α2-3,6,8,9 Neuraminidase A, BKF is α1-2-4-6-fucosidase, BTG is β-1-3-4-galactosidase and GUH is β-n-acetylglucosaminidase-s (see Supplementary Figure 1 for specificities). Exoglycosidases were applied according to manufacturer's protocol (New England Biolabs).

8. GlycanAnalyzer support, limitations and future work

8.1 What does GlycanAnalyzer support?

Currently GlycanAnalyzer can only annotate N-glycan LC-MS or LC only datasets where LC retention times have been normalised to glucose units. For glycan mass detection, the software is able to recognise different charge states (charge +1 and +2) in addition to various adducts (e.g. potassium (K), lithium (Li) and sodium (Na) adducts).

8.2 Limitations and future work

O-glycan, glycosphingolipid head-group and other types of oligosaccharide analysis are equally as important as N-glycan analysis. We hope to extend GlycanAnalyzer to support automated glycan annotations of any oligosaccharide. LC-MS analysis is extremely powerful, nevertheless we recognise that the glycomics community also carries out analyses using MS only (i.e. without LC). We hope to release a version of GlycanAnalyzer that takes simple MS (e.g. MALDI) data as input alone. In addition, a major update to GlycanAnalyzer that can interpret tandem MS data after application of exoglycosidases will be implemented once the datasets are generated to optimize new parameters. The addition of tandem MS data will allow greater confidence and confirmation of glycan assignments.

References

Aoki-Kinoshita, K., *et al.* GlyTouCan 1.0–The international glycan structure repository. *Nucleic acids research* 2016;44(D1):D1237-D1242.

Campbell, M.P., *et al.* GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* 2008;24(9):1214-1216.

Chan, K.F., *et al.* Inactivation of GDP-fucose transporter gene (Slc35c1) in CHO cells by ZFNs, TALENs and CRISPR-Cas9 for production of fucose-free antibodies. *Biotechnology journal* 2016;11(3):399-414.

Dell, A., *et al.* [8] Mass spectrometry of carbohydrate-containing biopolymers. In, *Methods in enzymology*. Elsevier; 1994. p. 108-132.

Klapoetke, S., *et al.* The evaluation of a novel approach for the profiling and identification of N-linked glycan with a procainamide tag by HPLC with fluorescent and mass spectrometric detection. *Journal of pharmaceutical and biomedical analysis* 2010;53(3):315-324.

Kozak, R.P., *et al.* Comparison of procainamide and 2-aminobenzamide labeling for profiling and identification of glycans by liquid chromatography with fluorescence detection coupled to electrospray ionization–mass spectrometry. *Analytical biochemistry* 2015;486:38-40.

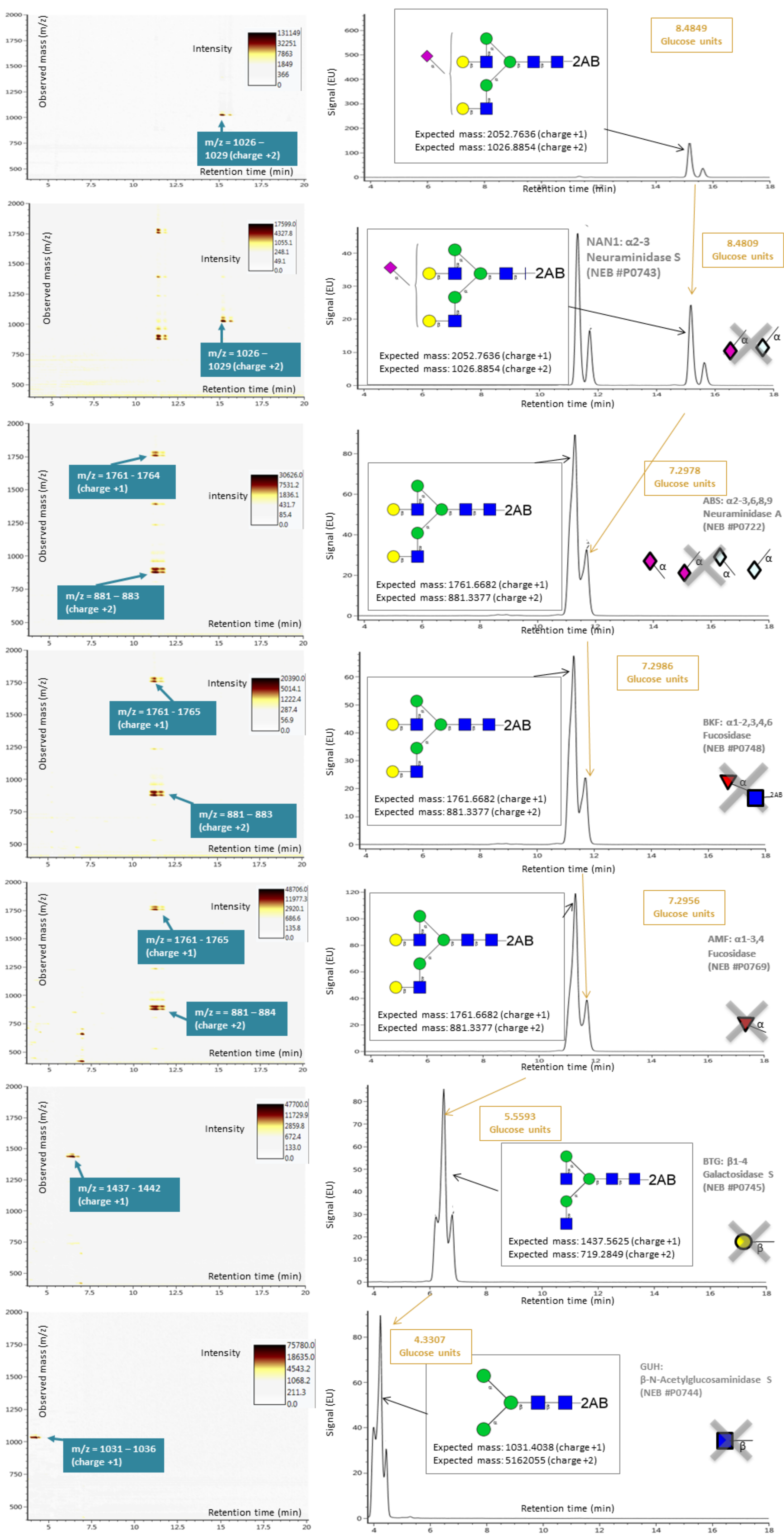
Mariño, K., *et al.* A systematic approach to protein glycosylation analysis: a path through the maze. *Nature chemical biology* 2010;6(10):713-723.

Ruhaak, L., *et al.* Glycan labeling strategies and their use in identification and quantification. *Analytical and bioanalytical chemistry* 2010;397(8):3457-3481.

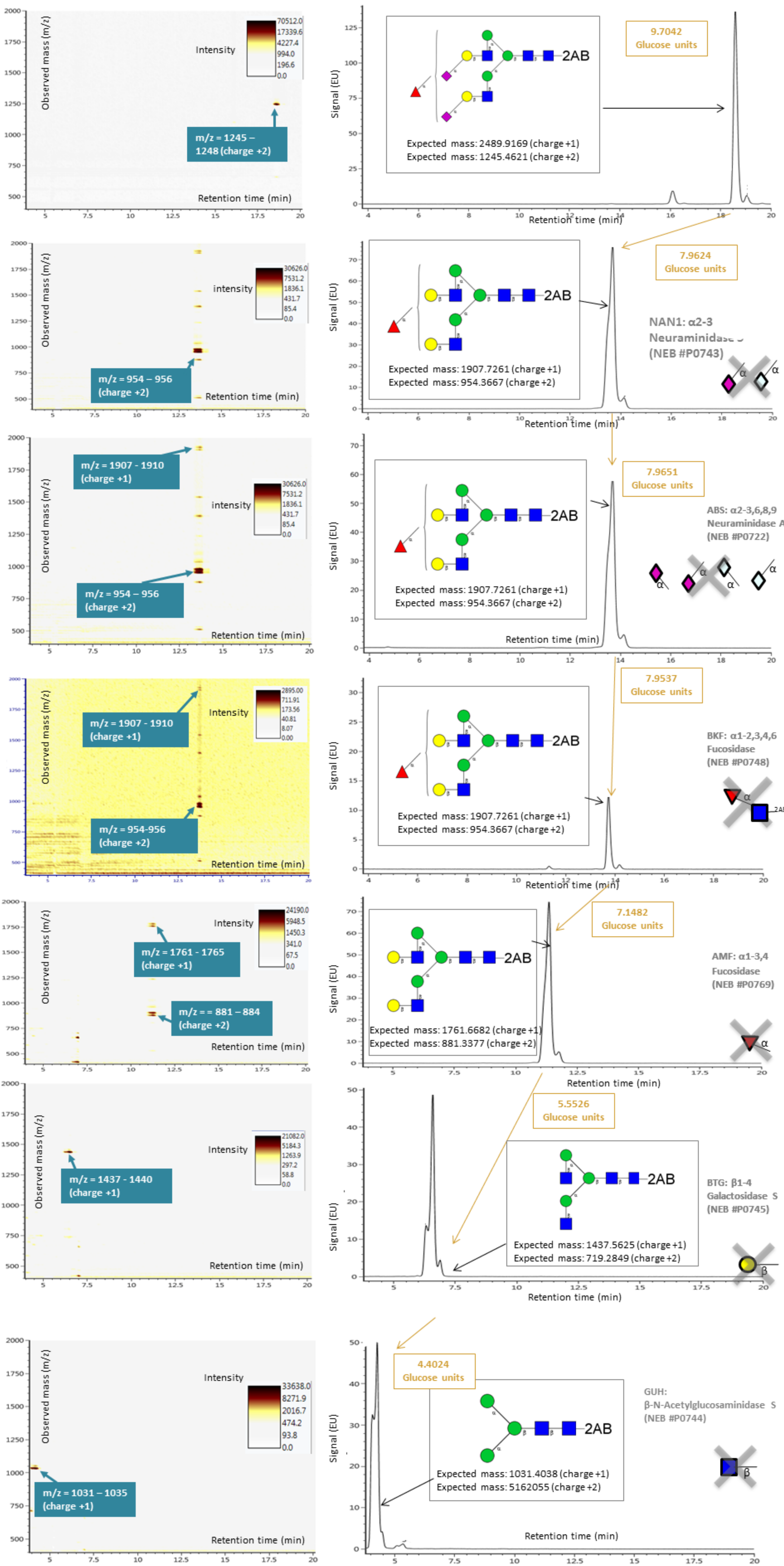
Vainauskas, S., *et al.* A novel broad specificity fucosidase capable of core α1-6 fucose release from N-glycans labeled with urea-linked fluorescent dyes. *Scientific Reports* 2018;8(1):9504.

Zhao, S., *et al.* GlycoStore: a database of retention properties for glycan analysis. *Bioinformatics* 2018;1:2.

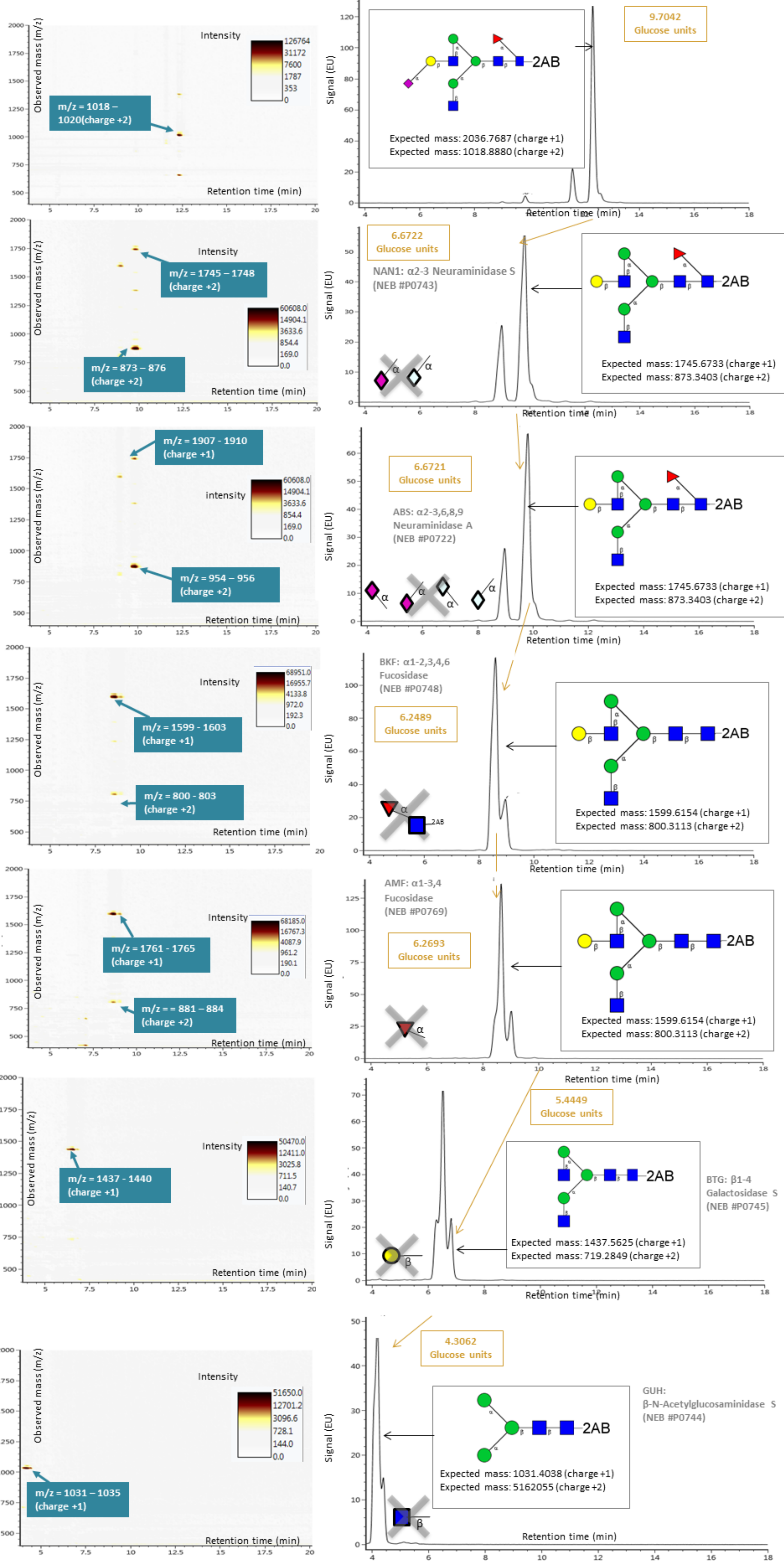
APPENDIX



Appendix Figure 1. The second standard used to successfully test the GlycanAnalyzer software.



Appendix Figure 2. The third standard used to test the GlycanAnalyzer software.



Appendix Figure 3. The fourth standard used to test the GlycanAnalyzer software.