GlycoPeptide Spectrum Annotation Program (gpAnnotate)

Document version: October 5, 2019

Table of Contents

1. Overview

The Glycopeptide Spectrum Annotation program (**gpAnnotate**) is designed to annotate individual glycoproteomics MS/MS spectrum for several established MS fragmentation modes:

- **CID**: collision induced dissociation
- **HCD**: beam-type CID or higher-energy collision dissociation
- **ETD**: electron transfer dissociation
- **ETciD**: electron-transfer-CID, and
- **EThcD**: electron-transfer-HCD.

In addition to preset options that are fixed for the above modes, gpAnnotate also allows 'Custom' definition of fragmentation rules in order to accommodate additional user preferences. Thus, in principle, it can handle a wide variety of scenarios.

A key feature of the program is the use of DrawGlycan-SNFG (version 2, [1]) for the annotation of MS/MS spectrum. Many of the scoring algorithms incorporated into gpAnnotate are derived from the GlycoProteomics Analysis ToolBox (GlycoPAT, [2]). Additional updates to GlycoPAT for large scale glycoproteomics studies will be released shortly (Cheng *et al.*, manuscript in preparation). The most recent release of both DrawGlycan-SNFG and GlycoPAT are available at VirtualGlycome.org.

gpAnnotate includes 3 modules, illustrated below in the main GUI (graphical user interface). These enable: i. Preprocessing of experimental data; ii. Annotation of MS/MS spectrum, and iii. DrawGlycan-SNFG (version 2). Installation details are provided in Section 2. Technical details follow in Section 3 below, and GUI usage instructions in Section 4.

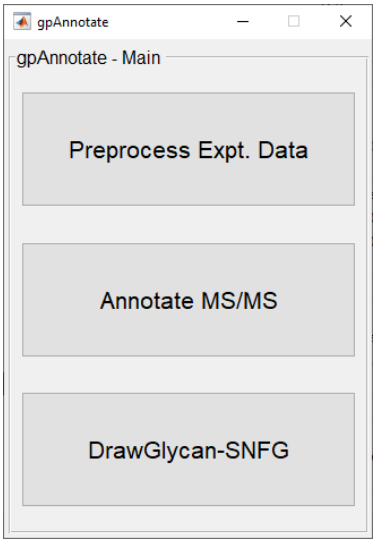


Fig. 1 Main GUI of gpAnnotate

2. Specifications and installation instructions

2.1 Compatibility

gpAnnotate in written using MATLAB R2018b. It is released in three different versions: Windows, MacOS and Linux. We have tested the program on these operating systems: Windows 10 ver. 1903, MacOS 10.12 and Ubuntu 18.04 LTS.

2.2 Installation

Download ***gpAnnotate*** from the program website and extract it to the desired location (for example c:/gpAnnotate).

- For Windows users: Run the program "Install_gp_Annotate_Win.exe" and follow the on-screen instructions.
- For MacOS users: Double click the package "Install_gp_Annotate_Mac.app" and follow the on-screen instructions.
- For Linux users (here we use Ubuntu as an example): In terminal, navigate to the location of the downloaded file "Install_gp_Annotate_Linux.install", right click this file then click "run".

  Super-user/Administration privilege may be necessary for installation. Also, you may have to temporarily disable anti-virus programs. Rest assured, we do not collect user data or install any spyware.

Proteowizard [3] is needed for data pre-processing, particularly for handling Thermo's proprietary .RAW files. To download and install ***Proteowizard*** go to http://proteowizard.sourceforge.net/. Get the most recent version. Follow default installation instructions, and note the folder where the file 'msconvert.exe' is stored, as this is needed by gpAnnotate.

3. *Technical details*

The following text describes the technical details of gpAnnotate.

1. Preprocess Expt. Data: This module extracts and processes selected data from MS experimental data files. Accepted input file types include: .RAW (Thermo-Fisher), .mgf (mascot generic format files), .mzML and .mzXML files (open-source). The output, which is stored in a .mat (MATLAB) format file, includes a data structure called 'msdata', which collates 11 fields for each MS-Spectrum:

i. sourceFileList: the source file name.
ii. scannum: scan number.
iii. precursormz: the monoisotopic m/z of the precursor ion. For MS^1 spectra, this value is -2.
iv. charge: charge state of precursor ion. 0 for MS^1 spectra.
v. spectra: Full MS^1 and MS^2 spectrum (i.e. Intensity vs. m/z data).
vi. mslvl: MS level of the scan.
vii. retime: retention time of scan.
viii. fragmode: fragmentation mode (CID, HCD etc.). Empty for MS^1 spectra.
ix. allprecursormz: This is a cell array with each element containing three values: a. m/z of selected ion, b. isolation window target and c. monoisotopic ion. These elements are empty for MS^1 spectra.
x. totIonCurrent: sum of all ion intensities in this spectrum.
xi. precursorScanNum: scan number of the spectrum where MS/MS precursor ion appeared. For MS^1 this value is -1.

***2. Annotate MS/MS*:** The .mat file from the previous step is used by the *Annotate MS/MS* module (GUI shown here with additional information in program usage section). This is the core of gpAnnotate, and it has built-in facilities to handle CID, HCD, ETD, ETciD and EThcD fragmentation modes. Default parameter settings for each of these generic modes is specified in Table 1, along with detailed explanation of each of the parameters under footnotes. These settings are used either when the individual fragmentation modes are selected by the user, or when fragmentation mode is set to 'default' in which case the program uses the fragmentation mode specified in the "msdata" data structure.

The fragmentation rules in Table 1 are used to generate the theoretical MS/MS spectrum for the candidate glycopeptide. These are compared with the experimental spectrum for the specified MS data input file and scan number. Corresponding statistical/matching scores are generated.

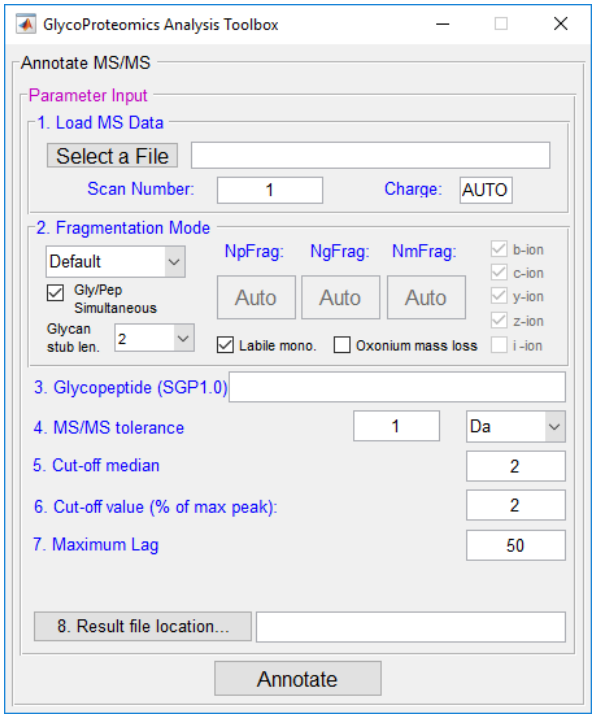


Fig. 2 Annotate MS/MS GUI interface

Table 1. Default fragmentation rules

Frag. mode	CID	HCD	ETD	ETciD	EThcD
i. NpFrag	0[a]	1	1	1	1
ii. NgFrag	2	2	0	1	1
iii. NmFrag	0	0	0	0	0
iv-v. Glycan/Peptide Simultaneous Fragmentation	No	Yes	No	Yes	Yes
(glycan stub length)	(-)	(2)	(-)	(2)	(2)
vi.-x. Peptide fragment type	b/y	b/y	c/z	c/z	b/c/y/z
xi. Labile monosaccharides	No	No	No	No	No
xii. Oxonium mass Loss	No	Yes	No	No	Yes

[a] In CID mode, if the glycan is too small (less than 4 monosaccharides in total), NpFrag will be changed to 1.

Np/Ng/NmFrag: The maximum number of cleavages allowed on peptide backbone (NpFrag), glycan (NgFrag) and non-glycan PTMs (NmFrag).

Glyco/peptide simultaneous fragmentation (glycan stub length): In some cases, it is possible that multiple fragmentation events may occur on a single glycopeptide with some of these occurring on the peptide backbone and others on the glycan chain. Such 'glycopeptide simultaneous fragmentation' is enabled using a checkbox (Y/N) in gpAnnotate (Fig. 2). "Glycan stub length" specifies the maximum number of glycosidic bonds between the peptide and the outermost reducing glycan. Stub length values from 0-7 are permitted. If stub length='unlimited', all glycan sizes are allowed. The default numerical value for standard fragmentation modes is presented in parenthesis. A stub length of 2 accommodates Y_0, Y_1 & Y_2 ions in HCD, ETciD and EThcD modes.

Peptide fragment type: Types of peptide fragments to include in the theoretical spectrum: b-/y-/c-/z-/i-, where i- is an "internal peptide fragment" generated due to two or more peptide backbone cleavages.

Labile monosaccharide: Neu5Ac and fucose may detach from glycans upon fragmentation, as they are labile. An option is available to include such losses in the theoretical MS/MS spectra, regardless of NgFrag settings.

Oxonium mass Loss: When collision energy is high, small B-ions tend to lose additional water and ammonia. This option adds common B-ions resulting from such losses of m/z: 145.0501 (for Hex containing glycans); 138.0555, 144.0661, 168.0661, 186.0766 (for HexNAc containing glycans); 274.0927 (for Neu5Ac containing glycans). If specific monosaccharides are absent in a glycan, corresponding oxonium ions are also omitted from the corresponding theoretical spectrum.

'Custom' fragmentation mode can be selected in gpAnnotate if either: i. the standard fragmentation modes are not sufficient for the experiment; or ii. the user wishes to change the default rules. In this case, the user may change the four numerical inputs (i.-iv.) and eight Y/N options (v.-xii) detailed in Table 1 using the Annotate MS/MS GUI (Fig. 2. see section 4 for procedural details).

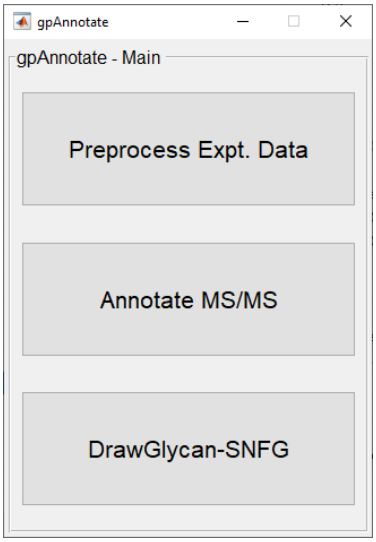
Additional fields are available in Annotate MS/MS to specify:

- The candidate glycopeptide sequence (in SGP 1.0 format). The theoretical spectrum is generated for this candidate and matched to the experimental scan. See ref. [2] and example .xlsx files for SGP1.0 specifications.
- MS/MS tolerance: This decides the maximum error allowed between theoretical and experimental spectrum peaks. For low-resolution quadrupole, ion trap and TOF analyzer, 1 Da is recommended. For orbitrap, 5-35 ppm may be appropriate. Higher value will match more peaks, but accuracy will be sacrificed.
- Denoising and cross-correlation parameters (do not change default values, unless you must). Cut-off median and cut-off value remove low intensity signal that may correspond to instrument noise [2]. Maximum Lag controls the number of iterations used during cross correlation calculations.
- Result File location (optional): Set path for .csv result file.

3. DrawGlycan-SNFG: This is a GUI version of DrawGlycan-SNFG (version 2). More details are provided in a separate DrawGlycan-SNFG manual (please see www.virtualglycome.org/drawglycan).

4. GUI usage instructions

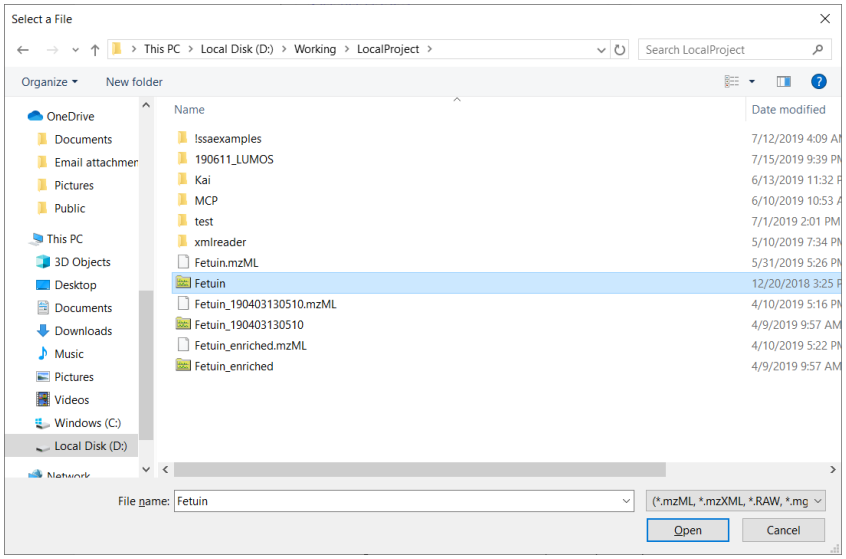
Double click the gpAnnotate.exe icon. This leads to the main ***gpAnnotate*** GUI, including the three modules described above.



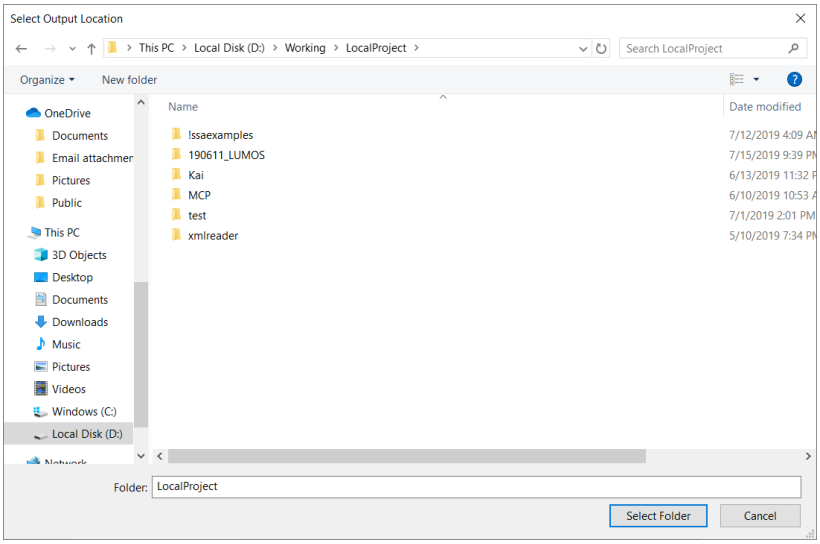
4.1 Preprocess Expt. Data

Choosing 'Process Expt. Data' will open the interface below:

Clicking 'Select Data File' will open the dialog box below. Four input file formats are supported: .raw, .mzML, .mzXML and .mgf.

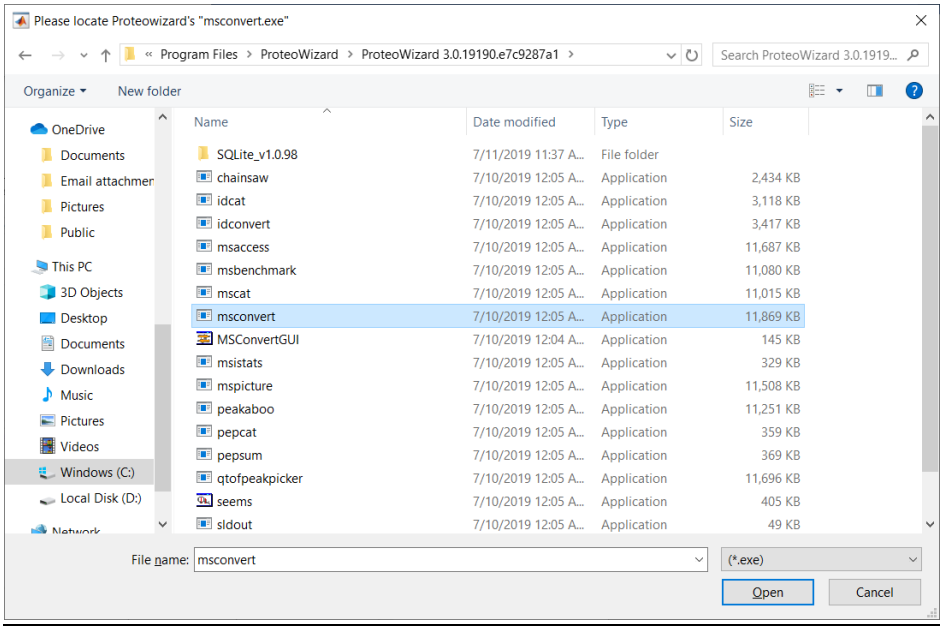


Click ‘Select Output .mat File Location’ to specify where to save the output data.

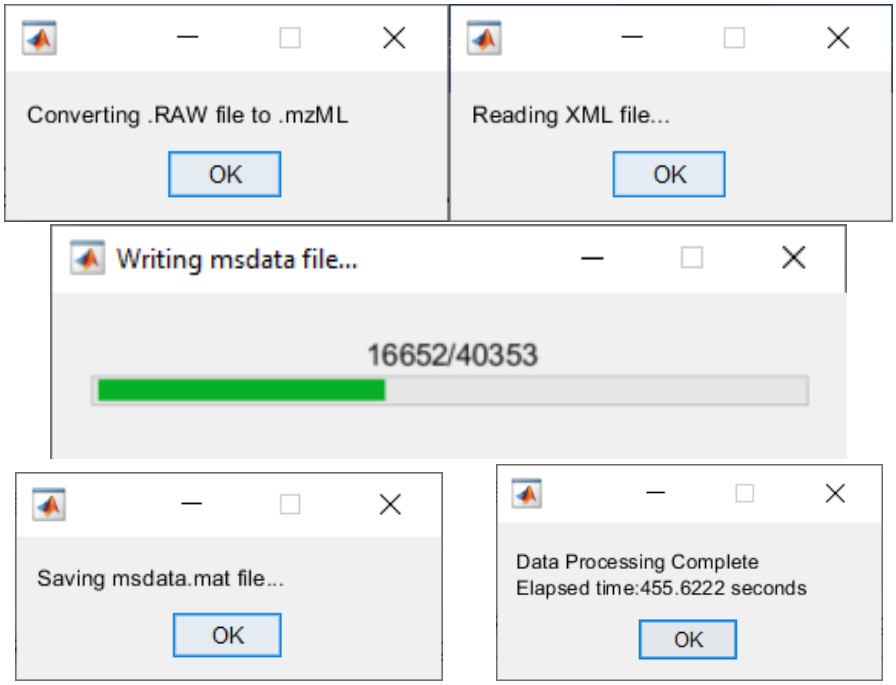


Press “RUN” to begin MS data extraction and processing.

If input data file has “.RAW” extension, a dialog box will appear requesting file path for the “msconvert.exe” conversion file in the Proteowizard folder. Browse to the location and click “Open” to continue.



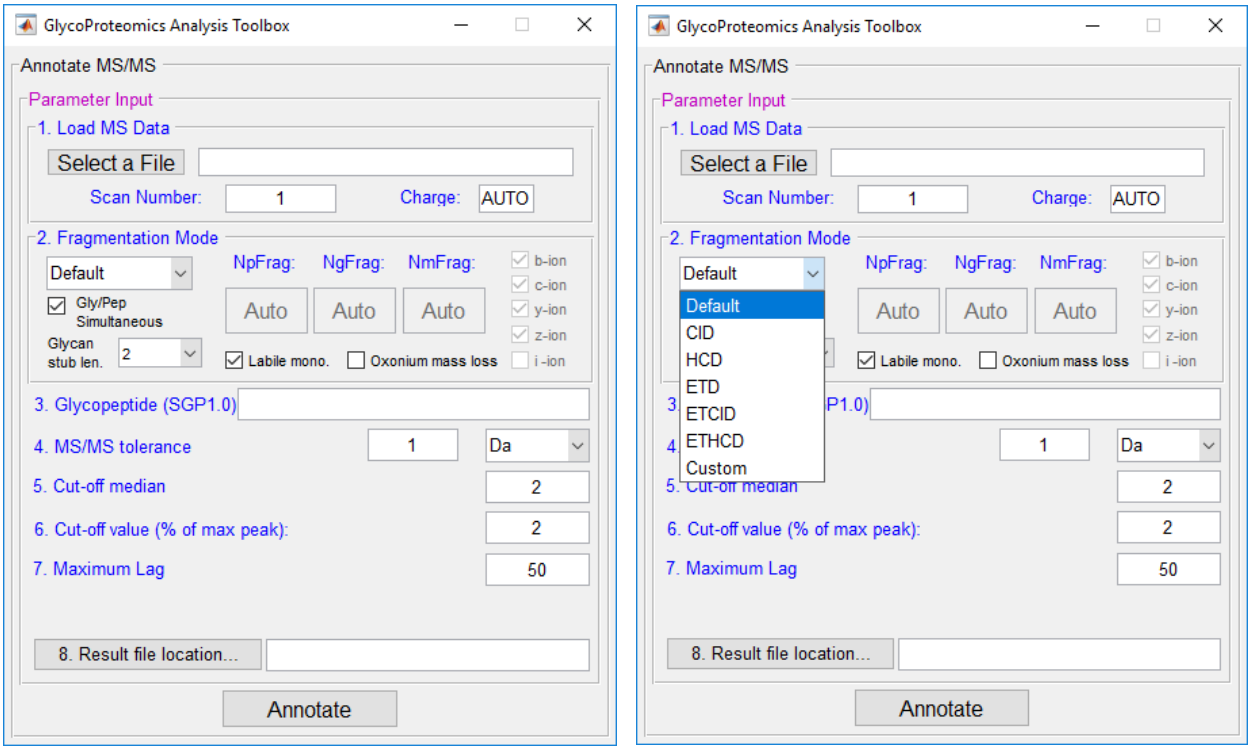
A series of dialog boxes and wait bars will appear to show progress.



The last dialog box marks the end of the pre-processing step.

4.2 Annotate MS/MS

Choosing 'Annotate MS/MS' in the main gpAnnotate GUI will open the interface (shown in two forms, with one displaying the pull-down menu):



To use this module:

1. a. *'Select a File'* in order to load MS data. This is the .mat file built in the last step.

b. Set '*scan number*' corresponding to the MS/MS spectrum that needs to be annotated.

c. Set '*charge*' state (optional) or leave as "AUTO" so that default charge value from scan header is used.

2. Set *'fragmentation mode*' from pulldown menu. Available options include: CID, HCD, ETD, ETCID, EThcD, 'Default' and 'Custom'. Here, 'Default' is the preferred fragmentation option. In this case, the program will use the fragmentation mode specified in the .mat file/msdata data structure. Fragmentation mode settings for all the above instances are specified in Table 1. Section 3 also described facilities available for 'Custom' fragmentation.

3. Specify the candidate Glycopeptide (SGP1.0 format) to score.

4. Set instrument MS/MS tolerance.

5.-7. Optional: Change default cut-off median, % of max peak and Maximum Lag if necessary.

8. Optional: Specify location to save result .csv file.

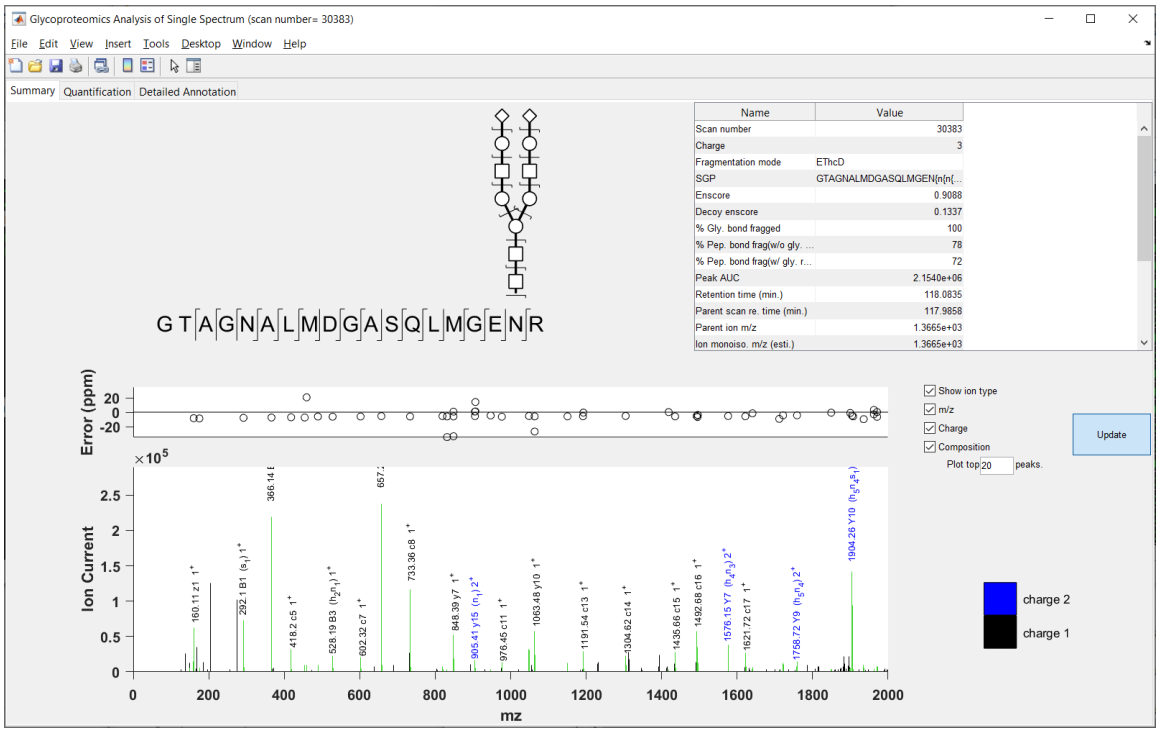
Click 'Annotate' to begin the analysis.

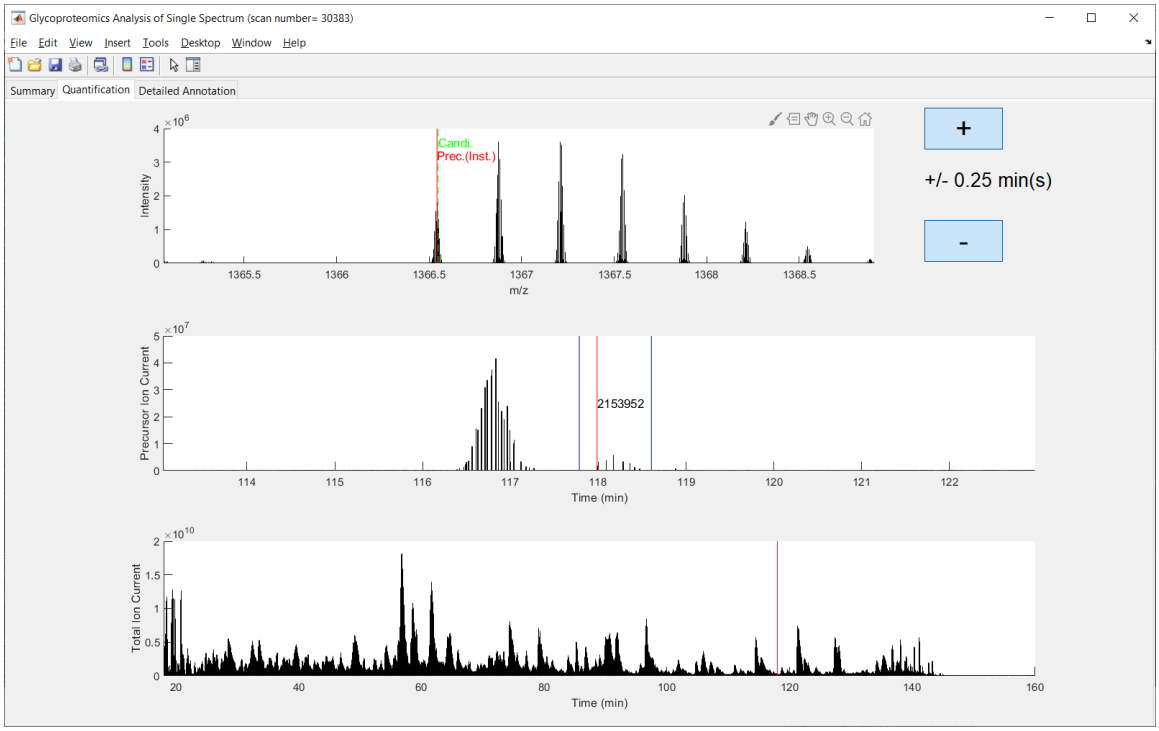
Notes: **i.** Loading the .mat file into memory can take time, during the first annotation. However, these data are stored in the memory, and used to generate additional annotations more rapidly for other glycopeptides.

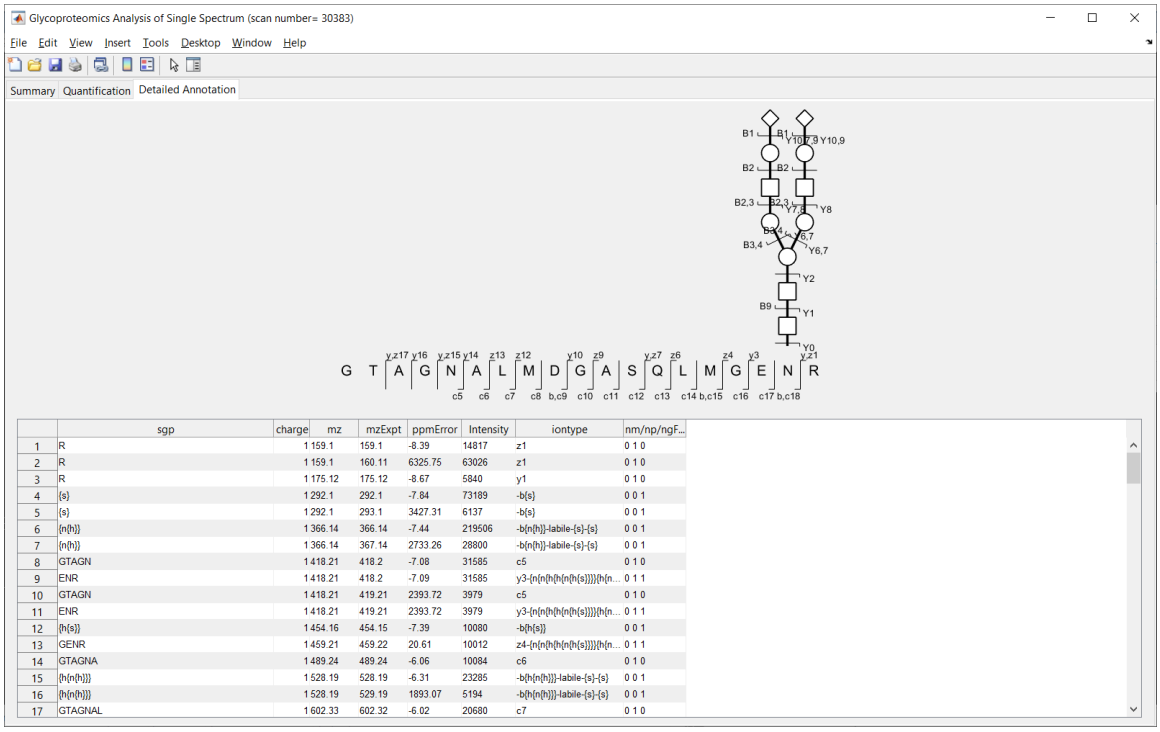
ii. Clicking the "Select a File" will erase all data and start the annotation afresh. Do not press unless you want to start from scratch.

Annotation results description

Results are presented in a 3 tabbed window shown below. Scan number in displayed in the header.







	sgp	charge	mz	mzExpt	ppmError	Intensity	iontype	nm/np/ngF...
1	R	1	159.1	159.1	-8.39	14817	z1	0 1 0
2	R	1	159.1	160.11	6325.75	63026	z1	0 1 0
3	R	1	175.12	175.12	-8.67	5840	y1	0 1 0
4	{s}	1	292.1	292.1	-7.84	73189	-b{s}	0 0 1
5	{s}	1	292.1	293.1	3427.31	6137	-b{s}	0 0 1
6	{n{h}}	1	366.14	366.14	-7.44	219506	-b{n{h}}-labile-{s}-{s}	0 0 1
7	{n{h}}	1	366.14	367.14	2733.26	28800	-b{n{h}}-labile-{s}-{s}	0 0 1
8	GTAGN	1	418.21	418.2	-7.08	31585	c5	0 1 0
9	ENR	1	418.21	418.2	-7.09	31585	y3-{n{n{h{n{h{s}}}}}{h{n...	0 1 1
10	GTAGN	1	418.21	419.21	2393.72	3979	c5	0 1 0
11	ENR	1	418.21	419.21	2393.72	3979	y3-{n{n{h{n{h{s}}}}}{h{n...	0 1 1
12	{h{s}}	1	454.16	454.15	-7.39	10080	-b{h{s}}	0 0 1
13	GENR	1	459.21	459.22	20.61	10012	z4-{n{n{h{n{h{s}}}}}{h{n...	0 1 1
14	GTAGNA	1	489.24	489.24	-6.06	10084	c6	0 1 0
15	{h{n{h}}}	1	528.19	528.19	-6.31	23285	-b{h{n{h}}}-labile-{s}-{s}	0 0 1
16	{h{n{h}}}	1	528.19	529.19	1893.07	5194	-b{h{n{h}}}-labile-{s}-{s}	0 0 1
17	GTAGNAL	1	602.33	602.32	-6.02	20680	c7	0 1 0

The following data are conveyed in the three tabs:

Tab 1 (Summary tab): This tab contains:

i. The glycopeptide structure drawn using DrawGlycan-SNFG (version 2) including fragmentation data indicating which bond fragments were observed in the experimental spectrum.

ii. A table showing summary statistics. The individual fields in this table are specified in Table 2 and they include results from the 1. pre-processing steps; 2. selected user specified inputs used by Annotate MS/MS; and 3. summary statistics for scoring that are derived from GlycoPAT (ref. [2], manuscript in preparation).

iii. Plot showing mass error between experimental and theoretical m/z for all matched peaks (middle)

iv. Annotated MS/MS spectrum (bottom) with matched peaks appearing in green, and unmatched peaks in red. Options (checkboxes) are available to annotate individuals peaks based on ion type, charge state, m/z values and/or glycan composition. Additional facilities are also available to limit the number of highest peaks that are annotated, based on user specified numerical value. The text used for such annotation are color coded based on ion charge state (specified in legends). An “Update” button is provided in order to apply changes.

Table 2: Summary statistics	
Experimental data & pre-processing results:	
Scan number	Scan number of the spectrum.
Ion monoiso. m/z (estimated)	The m/z of "monoisotopic ion". As described in .mzML standard.
Retention time (min.)	The retention time of scan
Parent ion m/z	The m/z of "selected ion". As described in .mzML standard (http://www.peptideatlas.org/tmp/mzML1.1.0.html#selectedIon).
Parent scan re. time (min.)	The retention time of the parent spectrum (The MS1 scan that contains the parent ion).
Annotate MS/MS settings:	
Charge	Charge of precursor ion
Fragmentation mode	Fragmentation mode used for analysis
SGP	Candidate glycopeptide sequence to score in SGP1.0 (SmallGlycoPeptide 1.0) format
NpFrag/NgFrag/NmFrag	Maximum number of cleavage events allowed on peptide /glycan/ non-glycan PTM
MS/MS scoring results:	
Candi. theo. mass	The mass of the candidate glycopeptide.
Enscore	Ensemble score of candidate
Decoy enscore	Ensemble score of a decoy (created by scrambling amino acid sequence and monosaccharide mass)
% Gly. bond fragged	% of theoretical glycosidic bond fragments that were matched/identified
% Pep. bond frag (w/o gly. residue)	% of peptide fragments (with glycan residue attached) that were matched
% Pep. bond frag (w/ gly. residue)	% of peptide fragments (without glycan residue attached) that were matched
% Theo. frag. ion match	The percentage of all theoretical fragments that were matched.
Peak lag	Offset between theoretical & experimental MS/MS spectrum where maximum Xcorr is achieved.
P-value	Poisson probability for false spectrum matching, calculate using candidate-decoy approach.
Top 10 peaks hit	How many of the 10 tallest peaks in the spectrum are matched.
Peak AUC	Label-free quantitation: Area under the curve (AUC) of candidate in elution curve

Tab 2 (Quantification tab): This tab shows:

i. *Top window*: The isotopic distribution from the experimental data collated in a specified time window, along with monoisotopic mass predicted by the instrument (in red), and monoisotopic mass of the user specified glycopeptide (in green). The +/- buttons on the right allow the assembly of this isotopic distribution using different MS^1 time windows. Here, besides the m/z of interest, additional peaks in

surrounding m/z are also accumulated and displayed. In the example presented above, the monoisotopic m/z of the instrument matches that specified by the user.

ii. *Middle window*: Ion current of the unfragmented candidate, extracted from the precursor MS^1 spectrum. This candidate ion current plot is shown for a ±5 min time interval, with mass tolerance of ±20 ppm. Area-under-the-curve ('Peak AUC', tabulated in Summary Tab) is calculated from this chromatogram for a continuous peak that is enclosed by the pair of blue lines. This AUC value is provided in this tab and also in the Summary tab. The red line indicates the precursor MS^1 spectrum location, which just precedes the specified MS/MS scan number.

iii. *Bottom window*: Total ion current-chromatogram for the entire experiment, including all species at all times. The location of the candidate spectrum us shown using a red line.

Tab 3 (Detailed annotation): The upper half shows the detailed DrawGlycan-SNFG sketch. Unlike Tab 1, all cleavages in this figure are fully annotated. The lower part shows a detailed table with information on each of the matched peaks. The fields in this table are listed in Table 3.

Table 3: Detailed peak matching table

Table fields	*Explanation*
sgp	SGP1.0 sequence of the matched glycopeptide fragment
charge	charge state of the matched peak
mz	Theoretical m/z of matched fragment
mzExpt	Experimental m/z of the matched peak.
ppmError or DaError	Difference between mz and mzExpt (unit depends on MS tolerance setting)
Intensity	Intensity of the matched peak.
Ion type	Classification of the fragment using the following scheme: B-ions: "-B" + *information on residue structure*. Y-ions: "-Y" "-" *residue lost 1* "-"*residue lost 2* "-" … b,c,y,z,i ions: b1, b2, c1, c2,…
NpFrag/NgFrag/NpFrag	Number of peptide/glycan/non-glycan PTM bond cleavages for that fragment

In addition to the three tabbed GUI window, detailed summary of results is also saved in a .csv file which is organized into two parts: A header section with summary parameters listed in Table 2 followed by a detailed table containing Table 3 data fields.

4.3 DrawGlycan-SNFG

Detailed usage instructions for DrawGlycan-SNFG (version 2) is provided in a separate manual (also please see virtualglycome.org/drawglycan).

5 Troubleshooting and help

Why does the first annotation take a lot of time?

Loading the .mat/data file into memory can take time, and this occurs during the first annotation. However, once these data are stored in the memory, additional annotations proceed more rapidly.

Why are there error or warning messages?

The program will display warning messages when user input contradicts experiment data. Error messages display when input is not acceptable. Please consider changing the inputs to avoid error messages, and ensure that the warnings are ok.

Should I disable virus protection during installation?

Yes. This may be necessary for some machines.

I have more questions or suggestions on how to improve the program?

Write to the authors: kaicheng@buffalo.edu or neel@buffalo.edu.

6 Bibliography

1. Cheng, K., Y. Zhou, and S. Neelamegham, *DrawGlycan-SNFG: a robust tool to render glycans and glycopeptides with fragmentation information.* Glycobiology, 2017. **27**(3): p. 200-205.
2. Liu, G., et al., *A Comprehensive, Open-source Platform for Mass Spectrometry-based Glycoproteomics Data Analysis.* Mol Cell Proteomics, 2017. **16**(11): p. 2032-2047.
3. Chambers, M.C., et al., *A cross-platform toolkit for mass spectrometry and proteomics.* Nat Biotechnol, 2012. **30**(10): p. 918-20.