SOP of MS-PyCloud pipeline for LC-MS/MS data analysis

MS-PyCloud is a comprehensive software for mass spectrometry-based proteomics data analysis, encompassing peptide identification, protein inference, and quantitation of proteins, phosphosites, and glycan-specific glycopeptides. It processes raw Thermo LC-MS/MS data converted to mzML format, performing searches using GPQuest for glycan databases and MS-GF+ for protein databases, with results filtered based on PSM-level false discovery rates. Protein inference utilizes a bipartite graph analysis algorithm to group significant PSMs, assigning shared peptides to the most supported proteins. Quantitation supports various isobaric tags, including iTRAQ4, iTRAQ8, TMT10, TMT11, TMT16, and TMT18, and employs median normalization for accurate abundance, intensity, and Log2 ratios calculations. Enhancements like SQLite databases for peptide fragment indexing, Numba for JIT compilation, and a Streamlit GUI improve performance and usability. Integrated with AWS for scalable cloud computing, MS-PyCloud ensures efficient and high-availability peptide analysis.

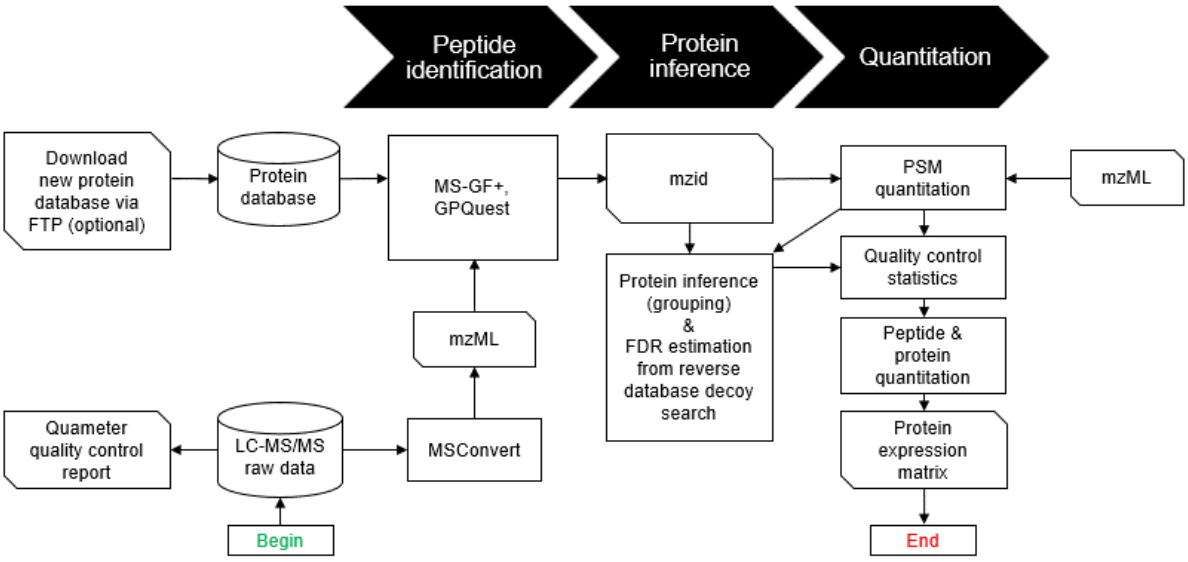

Figure 1. Schematic workflow of MS-PyCloud for LC-MS/MS data analysis.

MSPycloud tutorial:

1. ***Installing and setting up MSPycloud:***

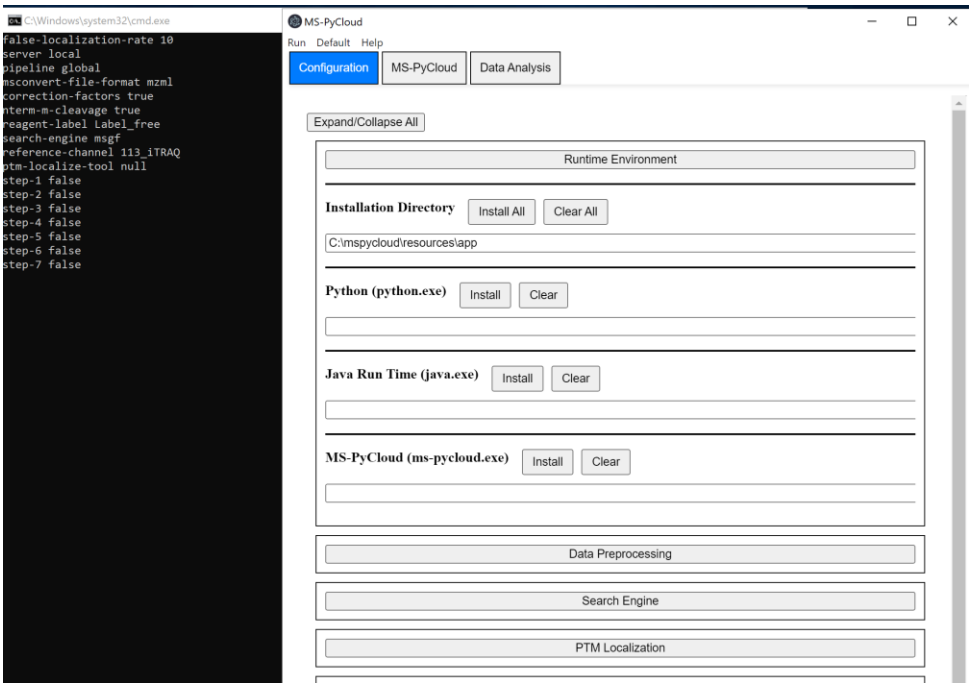

Figure 2. Graphical User Interface (GUI) of MS-PyCloud 'Configuration' tab.

- Download the zip folder on Github and extract the folder
- Launch the application either from 'ms-pycloud.exe' to open the interface or 'ms-pycloud.bat' to have the command prompt displaying console message.
- After launching, either install all (recommended) or install the needed configuration.

2. *Start the analysis:*

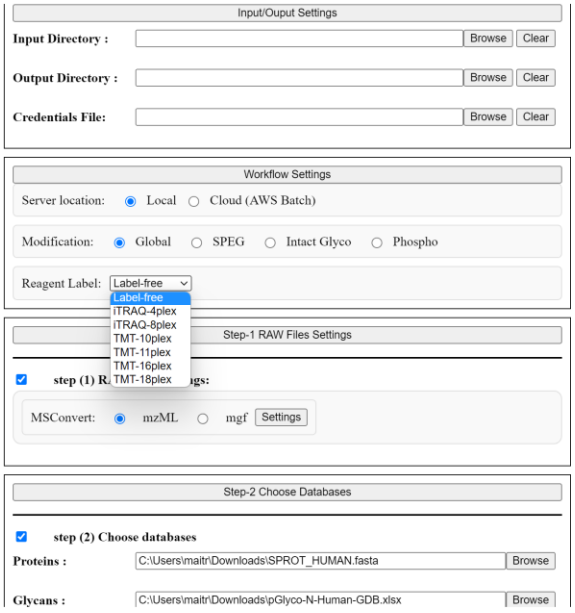

Figure 3. GUI of MS-Pycloud 'MS-Pycloud' tab

a. *Input/Output settings:*
   - Input Directory: Insert folder that contains the .raw data
   - Output Directory: Insert folder that collect the running results
   - Credential File: This file is needed to perform cloud run on AWS

b. *Workflow settings:*
   - Server location: *'Local'* is set as default. If choose *'Cloud (AWS Batch)'*, credentials file is needed.
   - Choose the analyzed modification and reagent label

c. *Step (1) RAW files settings:*
   - Use mzML for comprehensive analysis and detailed data.
   - Use MGF for faster, and simplified data processing, but less comprehensive.

d. *Step (2) Choose databases:*
   - Insert .fasta database for *'Proteins'* and .xlsx for *'Glycans'*.
   - The format of the .fasta (figure 4) and .xlsx (figure 5) files shown below:

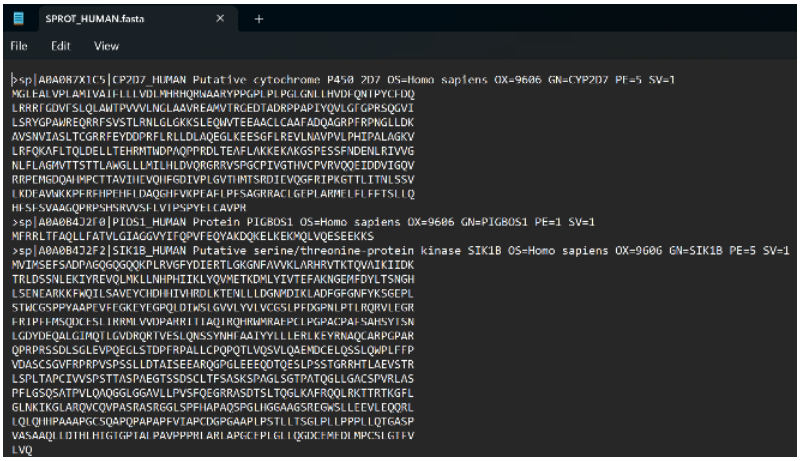

Figure 4. FASTA database format

	A	B	C	D	E
1	N	H	F	S	G
2	8	9	3	0	0
3	6	6	2	0	0
4	4	6	1	2	0
5	5	5	0	0	0
6	7	9	5	2	0
7	5	4	2	1	0
8	3	4	3	0	0
9	8	11	2	1	0
10	4	7	3	1	0
11	7	9	0	1	0
12	8	10	5	0	0
13	6	10	2	1	0
14	6	7	2	1	0
15	7	7	0	4	0
16	2	8	0	0	0

Figure 5. Glycans database format

3. *Setting search parameters*

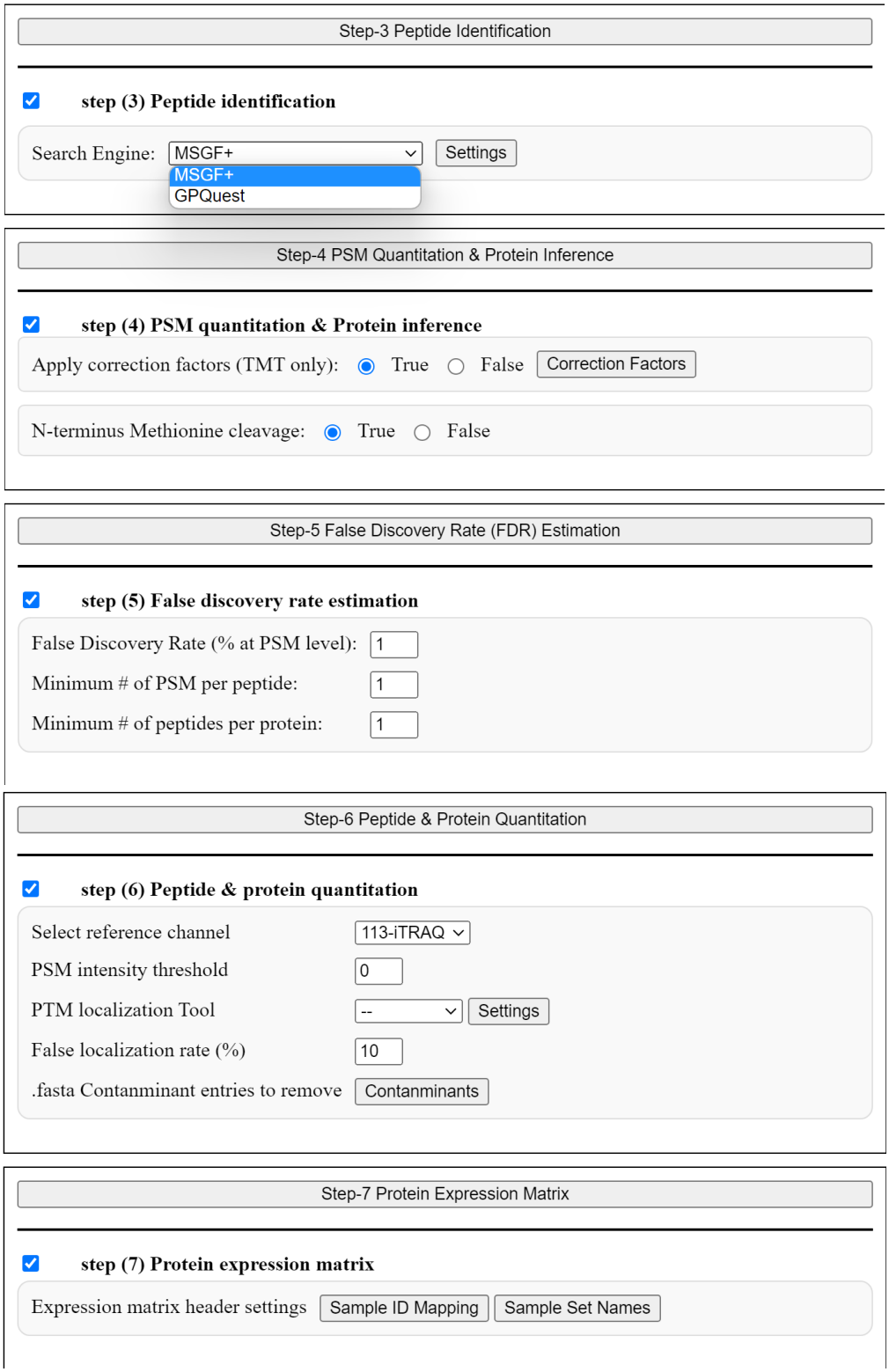

Figure 6. GUI of MS-Pycloud 'MS-Pycloud' tab (cont)

a. *Step (3) Peptide identification*
   - Search engine:
     - **MSGF+** works best for Global and Phospho search
     - **GPQuest** works best for Glyco search
     - The search engine will be automatically chosen based on the chosen 'Modification' in 'Workflow Settings'
     - Click *'Settings'* to view the search parameter that can also be found in the 'configs' folder in the output directory with the following names:
       - MSGF: MSGFPlus_Params.txt
       - GPQuest: gpquest3_params.yml

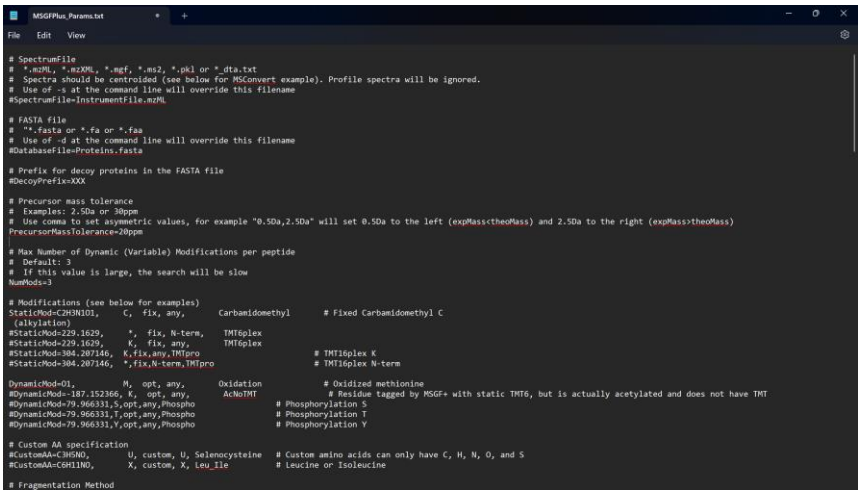

Figure 7. Parameter configuration file of MSGF+ (MSGFPlus_Params.txt)

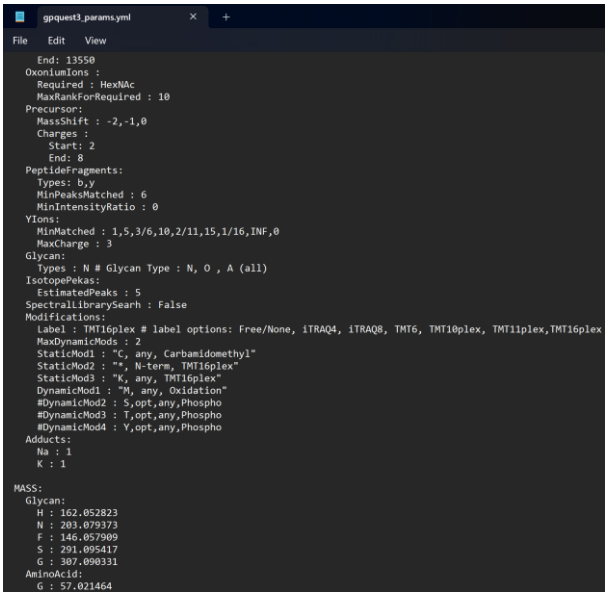

Figure 8. Parameter configuration file of GPQuest (gpquest3_params.yml)

- Notes:
    - MSGF+ can automatically modify the parameters ('# Modifications') to make it compatible with the chosen TMT-labeling option
    - GPQuest require manual change at 'Label', 'StaticMod2', and 'StaticMod3' to be compatible with the chosen TMT-labeling option

b. *Step (4) PSM quantitation and Protein Interference*
    - Apply correction factors:
        - Only applicable with TMT-labeling data
        - Must insert a compatible correction factor to process
        - Click 'Correction Factors' to view the correction factor, which can be found in 'configs' folder: correction_factors.tsv (figure 9).
        - This file currently offers the correction factor for TMT16plex, TMT11plex, and TMT10plex. Add extra correction factor (e.g. TMT18plex), if needed

TMT16plex

Mass tag	"""-2x 13C"""	"""-13C -15N"""	"""-13C"""	"""-15N"""	Monoisotopic	"""+15N"""	"""+13C"""	"""+15N +13C"""	"""+2x 13C"""
126	0	0	0	100	0.34	9.31	0.02	0.32	
127N	0	0	0	0.78	100	0	9.41	0	0.33
127C	0	0	0.93	0	100	0.35	8.63	0.01	0.27
128N	0	0	0.95	0.79	100	0	8.38	0	0.26
128C	0	0	1.47	0	100	0.34	6.91	0	0.15
129N	0	0	1.46	1.28	100	0	6.86	0	0.15
129C	0.51	0	2.74	0	100	0.36	6.15	0	0.11
130N	0.13	0	2.41	0.27	100	0	5.58	0	0.1
130C	0.04	0	3.1	0	100	0.42	4.82	0.02	0.06
131	0.04	0.04	3.09	1.36	100	0	4.75	0	0.06
131C	0.08	0	3.81	0	100	0.4	3.26	0.03	0.03
132N	0.04	0	2.84	0.79	100	0	3.51	0	0.02
132C	0.11	0	4.55	0	100	0.43	1.86	0	0
133N	0.36	0.01	3.64	0.82	100	0	1.94	0	0
133C	0.22	0	4.96	0	100	0.34	1.03	0	0
134N	0.3	0.03	5.49	0.62	100	0	1.14	0	0

TMT11plex

Mass tag	-2	-1	Monoisotopic	1	2
126	0	0	100	7.2	0.2
127N	0	0.4	100	7.3	0.2
127C	0	0.5	100	6.3	0
128N	0	0.7	100	5.7	0
128C	0	1.4	100	5.1	0
129N	0	2.5	100	5	0
129C	0	2.3	100	4.3	0
130N	0	2.7	100	3.9	0
130C	0.4	2.9	100	3.3	0
131	0	3.4	100	3.3	0
131C	0	2.6	100	2.9	0

Figure 9. Correction factor of TMT-labeling (correction_factors.tsv)

c. *Step (5) False discovery rate estimation (Default set at 1)*
    - False Discovery Rate: Threshold of PSM false discovery rate estimated using a concatenated target-decoy database search. Adjusted based on confidence requirement.

- Minimum of % PSM per peptide and peptides per protein: Increase if a more stringent peptide and protein filtering is needed.

d. Step (6) Peptide and protein quantitation

- Select reference channel: Choose channel that represents baseline condition
- PSM intensity threshold: Set an appropriate threshold to retain reliable PSMs (*default set a 0*)
- PTM localization tool:
  - Have the option to use LuciPHOr2 for phospho localization.
  - Leave blank if utilize other modifications search
- False localization rate: Adjust based on confidence requirements (*default set at 10*)
- .fasta contaminants entry to remove: Click 'Contaminants' to get access to 'contaminants.fasta'

e. Step (7) Protein expression matrix

- Sample ID Mapping:
  - Click to get access to the file 'header_expression_matrix.tsv' in the 'configs' folder (figure 10)
  - This file currently offers the header expression for TMT16plex, TMT11plex, TMT10plex, iTRAQ8plex, and iTRAQ4plex. Add extra header expressions (e.g. TMT18plex), if needed.
- Sample Set Names:
  - Click to get access to the file 'sample_filenames.txt' in the 'configs' folder (figure 11)
  - This file will record all of the raw files sample name. In the case of seeing a blank document or missing filenames, annually add all of the raw files sample names or re-insert the input and output directory.

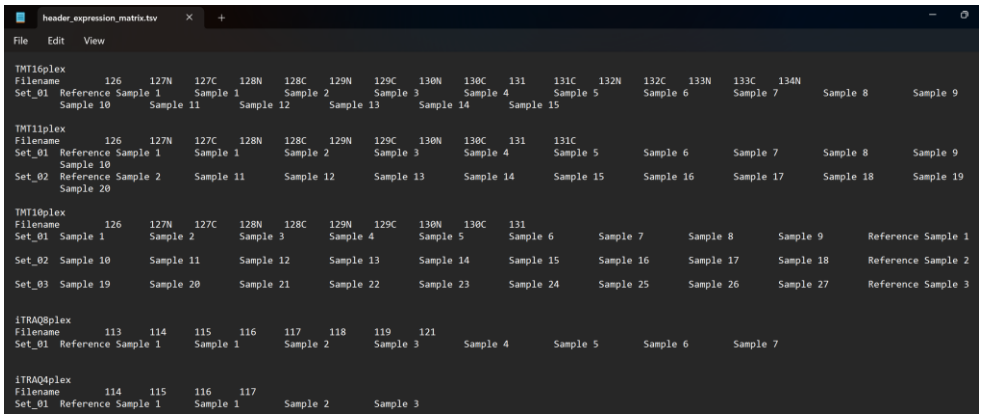

Figure 10. Header expression for TMT-labeling (header_expression_matrix.tsv)

Figure 11. Sample filenames of the raw files in input directory (sample_filenames.txt)

4. *Data analysis*

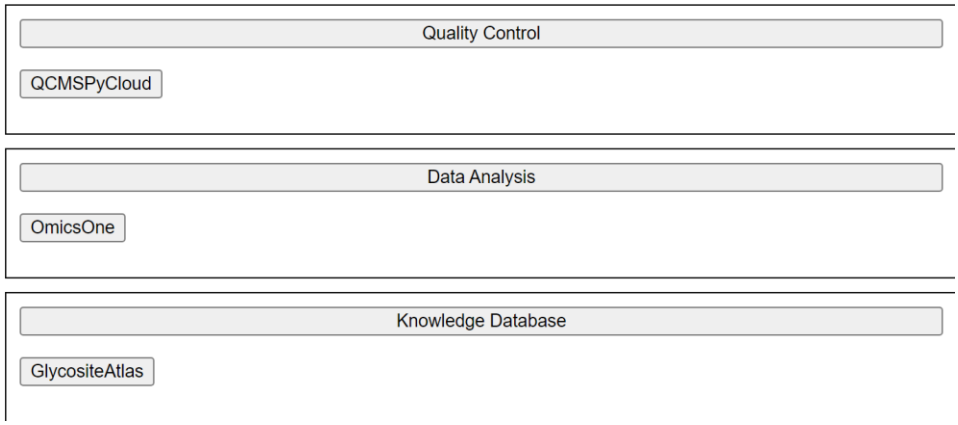

Figure 12. GUI of MS-Pycloud 'Data Analysis' tab

a. Quality control (QCMSPycloud)
   - Click 'QCMSPycloud' to get access to the application and view the data visualization for quality control
     - Access this way will have the application automatically extracted the data from the output directory
     - The user has the option to operate QCMSPycloud as a standalone application as they can input the data folder directory and choose the designated modification in the UI (figure 13)

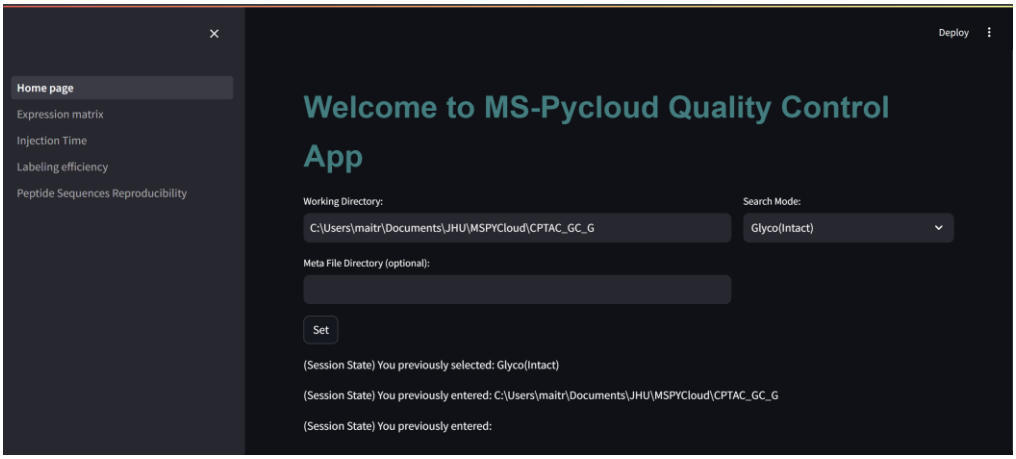

Figure 13. GUI of QCMSpyloud 'Home page' tab

- Four main analysis components:
  - Expression matrix
    - Have the following options for matrix types (Gene, Peptide, Protein, Site)
    - Have the following analysis type: Abundance, Intensities, and Log2Ratio
    - Have the following normalization options: Original, MD_norm, MD_norm_MAD_scaling
    - MSPycloud does not provide gene matrix for glyco search, and protein matrix for both glycol and phospho search.
    - For data visualization (figure 14):
      - A box plot of the sample with their analysis type
      - A bar chart of number of the matrix type counts in each sample and a table including
      - There is an option to download the png file of the box plot, bar chart, and the table

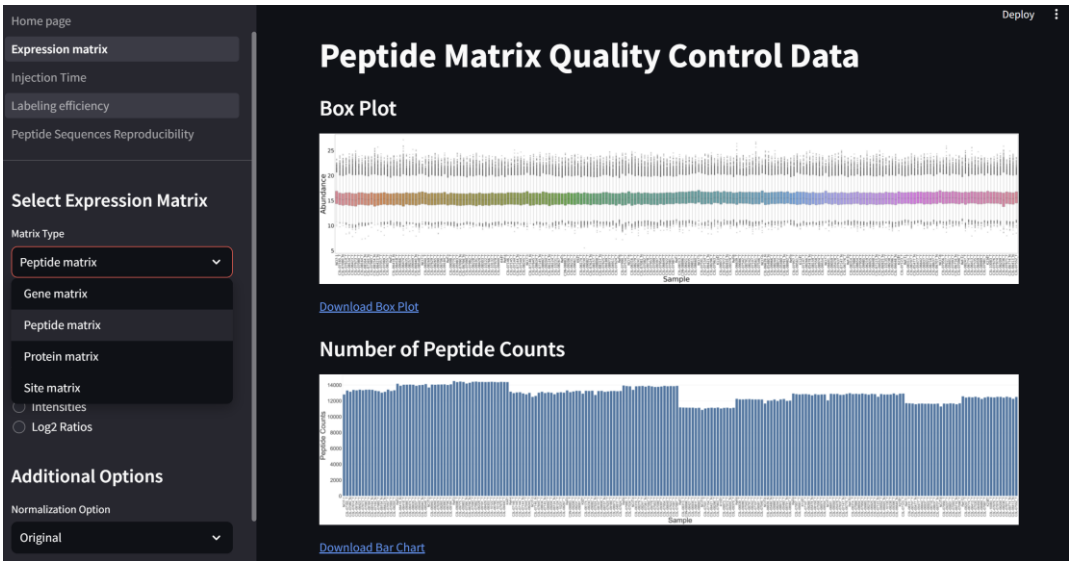

Figure 14. GUI of QCMSPycloud 'Expression matrix' tab

  - Injection time
    - Compose of histograms showing the frequency of the injection time (figure 15)
    - Each data file in the folder has their own histogram for injection time analysis and offer the download option

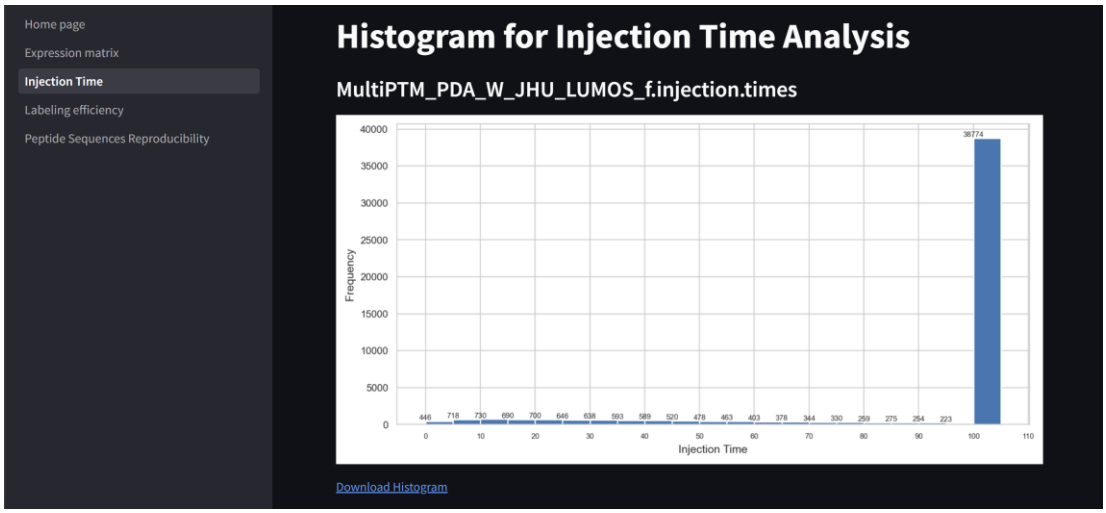

Figure 15. GUI of QCMSPycloud 'Injection time' tab

- Labeling efficiency
  - Compose of bar charts with the following components: Partially labeled peptides (%), Sites labeled (%), No TMT16plex modifications (%), Summed percent to verify unity (%), and [reference channel] null (%)
  - Each data file in the folder has their own bar chart for labeling efficiency analysis and offer the download option (figure 16)
  - Not supported for glyco search

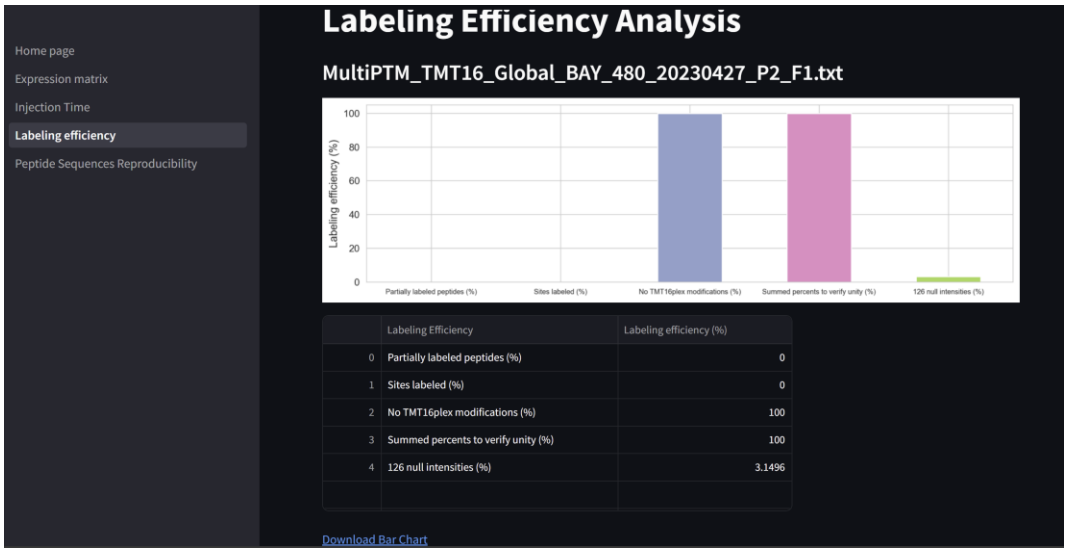

Figure 16. GUI of QCMSPycloud 'Labeling efficiency' tab

- Peptide Sequence Reproducibility
  - Display the reproducibility of peptides identification across different data files using the bar chart and a table
  - For Global and Phospho search:
    - Simpler method to extract data with more limited option (Figure 17)
  - For Glycol search:
    - Have the option to insert the meta.tsv in the Home page UI
    - Meta.tsv is only compatible for Glyco search
    - This option allows to look at the peptide sequence reproducibility of a variability of matrix files in step7 (Figure 18)

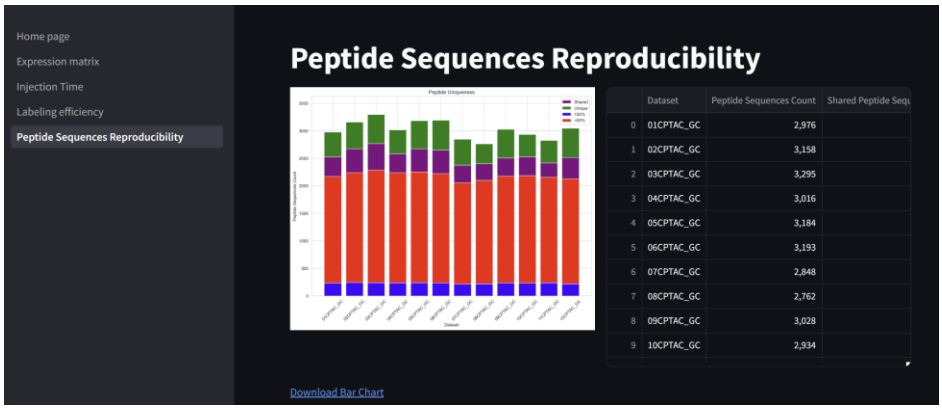

Figure 17. GUI of QCMSPycloud 'Peptide Sequences Reproducibility' tab (apply for all run)

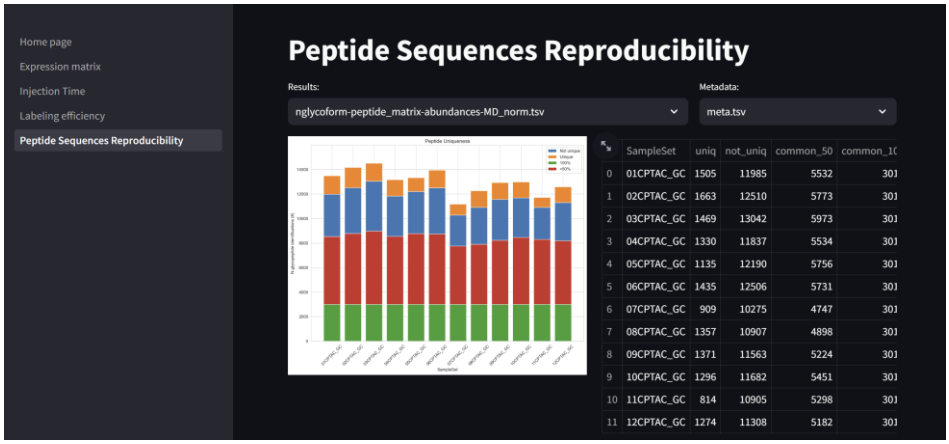

Figure 18. GUI of QCMSPycloud 'Peptide Sequences Reproducibility' tab (applicable for glyco search if meta.tsv available)

b. Data analysis (OmicsOne)
c. Knowledge Database (GlycositeAtlas)