Supplementary Note for:
Scribe: next generation library searching for DDA experiments
Brian C. Searle[1, 2, 3, 4 *] , Ariana E. Shannon[1, 2, 3], and Damien Beau Wilburn[1, 2, 3]

[1]Department of Biomedical Informatics, The Ohio State University Medical Center, Columbus, Ohio 43210, United States.
[2]Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States.
[3]Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio 43210, United States.
[4]Proteome Software Inc., Portland, OR

*Corresponding Address: Brian C. Searle, Ph.D., brian.searle@osumc.edu

Peptide detection with Scribe

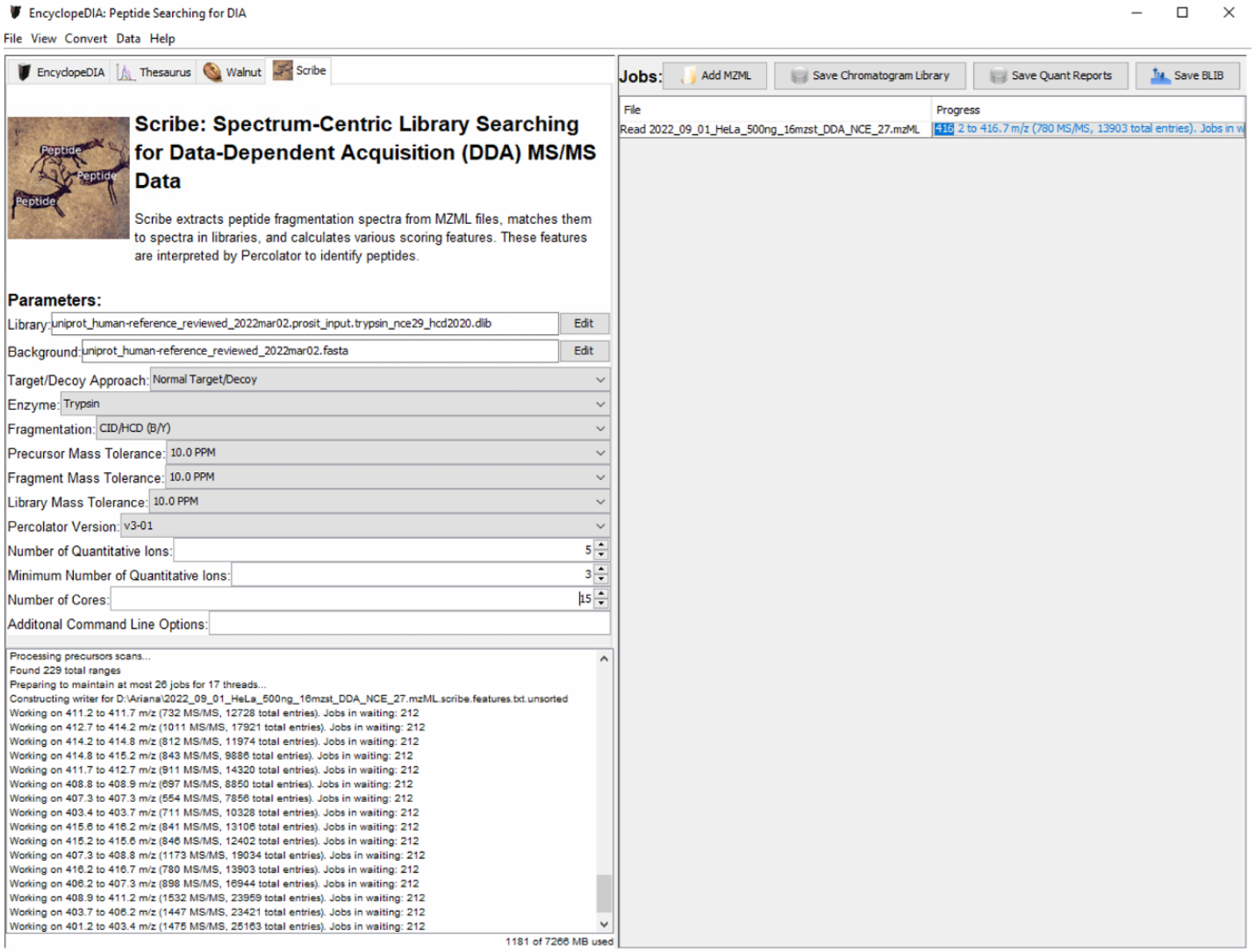
Tutorials based on Scribe version 2.11.2, last update on December 22, 2022
Written by Ariana Shannon (shannon.225@buckeyemail.osu.edu) and Brian Searle (brian.searle@osumc.edu)

Spectrum library searching is a powerful alternative to database searching for data dependent acquisition (DDA) experiments. Scribe is a new library search engine designed to leverage deep learning fragmentation prediction software such as Prosit. Scribe embeds Percolator for false discovery rate (FDR) correction and an interference-tolerant label-free quantification integrator to enable an end-to-end proteomics workflow.

TABLE OF CONTENTS

OVERVIEW

Scribe comes with a user-friendly GUI interface. The upper left pane contains search options, while the right pane contains a process queue. The bottom left contains a console that provides specific information about the process EncyclopeDIA is running.



This manual outlines how to use Scribe, what outputs are given by Scribe, and how to use the outputs to interpret data.

1. PREREQUISITES AND INSTALLING SCRIBE

Software requirements needed to process data using Scribe.

Scribe is a cross-platform Java application that has been tested for Windows, Macintosh, and Linux. Scribe requires 64-bit Java 1.8. While it is possible to use higher versions of Java, many versions are untested. Using untested versions of Java may result in unknown errors. In particular, Java versions 17 and 18 are known to cause stability issues in the current release.

A. If you don't already have Java, install Java 1.8 on your computer. if you are using Windows, you can download the Windows "x64 Installer" from: https://www.oracle.com/java/technologies/downloads/#java8. Other operating system options are available at this URL as well.

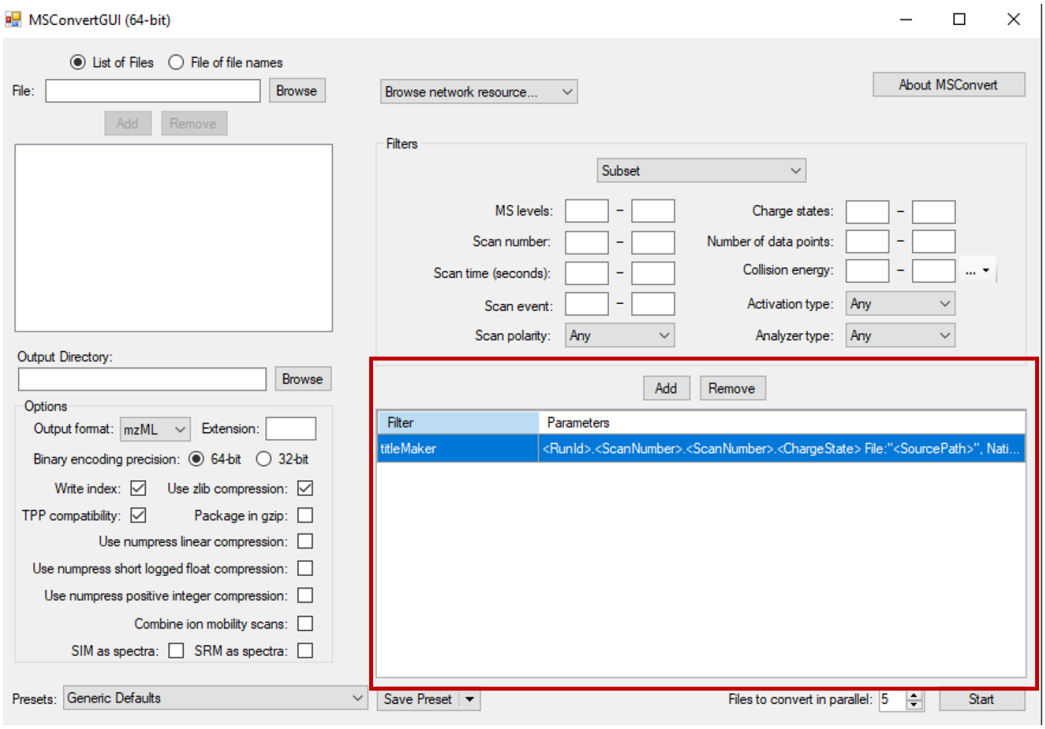
B. Scribe is folded into the EncyclopeDIA software package. After you have 64-bit Java 1.8, go to EncyclopeDIA's bitbucket page (https://bitbucket.org/searleb/encyclopedia/wiki/Home) and download the most recent stable version. Once downloaded, double-click on the EncyclopeDIA .JAR file to launch the GUI interface. If you are using a Macintosh, you may need to right-click on the EncyclopeDIA .JAR and select "Open" to execute it for the first time with the proper permissions. Click on the tab named Scribe at the top to search DDA data.

C. We recommend using Proteowizard to create mzML files from your RAW files. You can freely download Proteowizard from here: https://proteowizard.sourceforge.io/download.html.

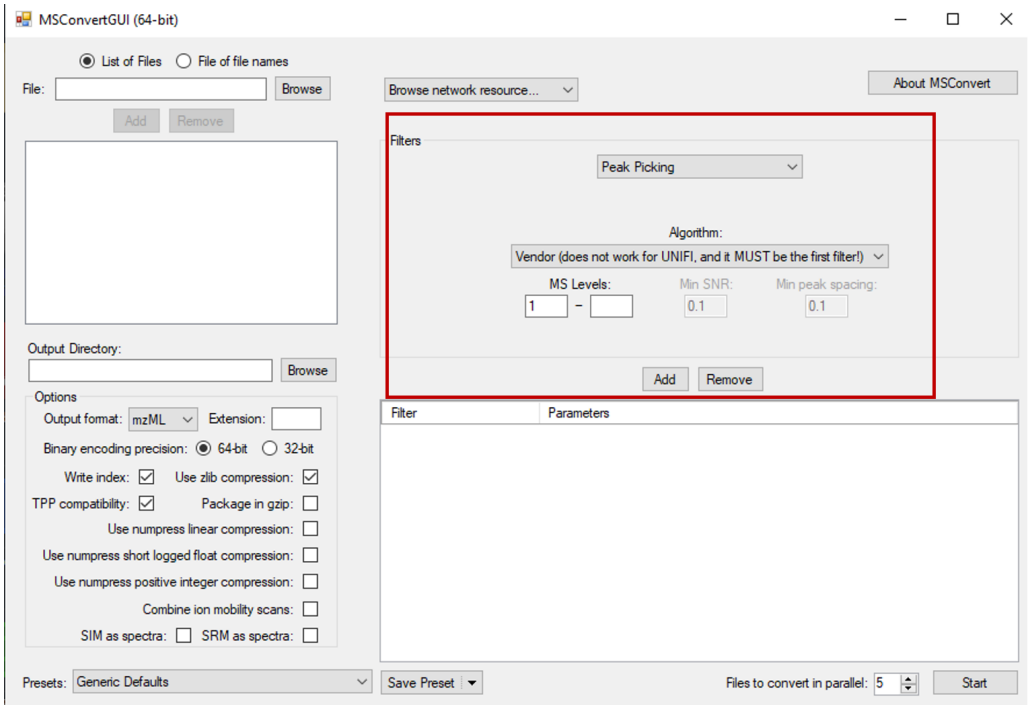
2. GENERATING MZMLS

How to use Proteowizard to generate vendor-neutral .mzML files from vendor-specific RAW files.

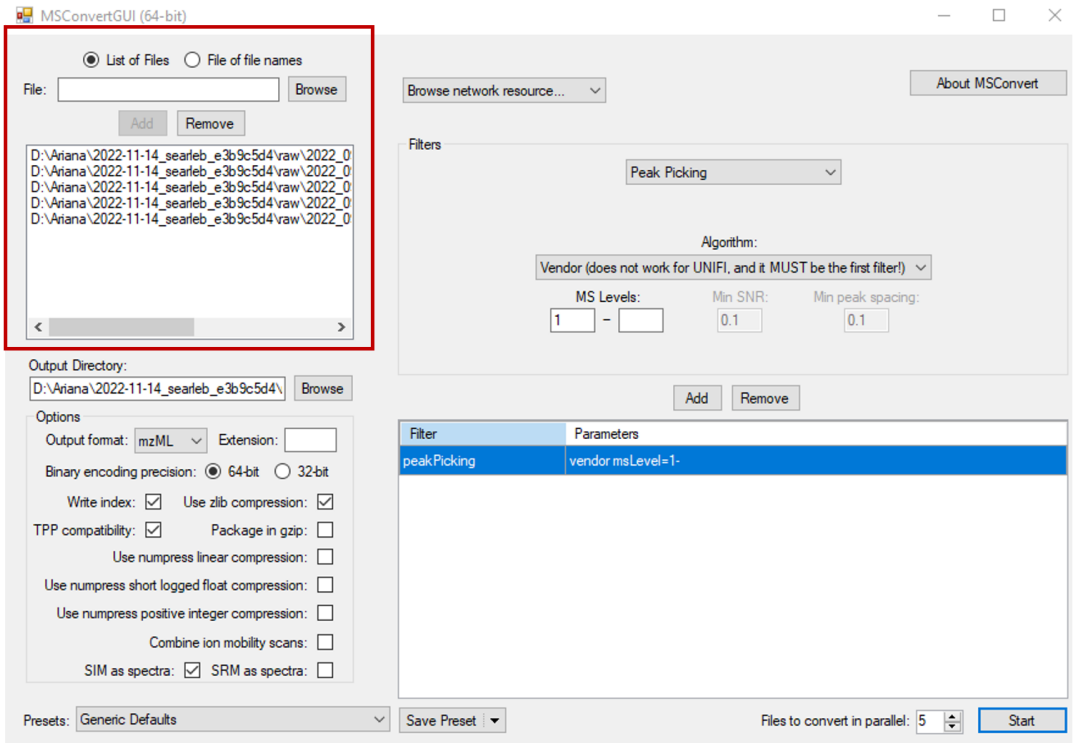
A. Before searching files in Scribe, RAW files must be converted to .mzML files using Proteowizard. To do so, open Proteowizard. Remove the “titleMakers” filter by selecting the filter in the parameters box, and clicking remove.



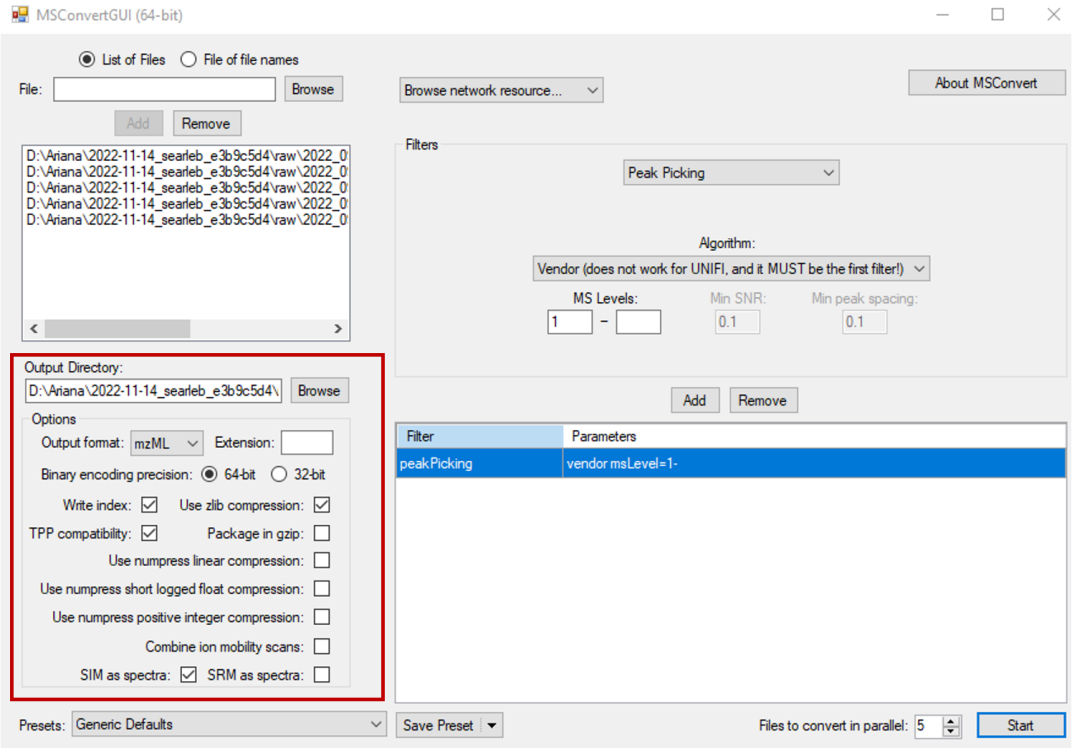
B. Add “Peak Picking” by selecting the peak picking option under filters, then click add. This should be the only filter in the box.



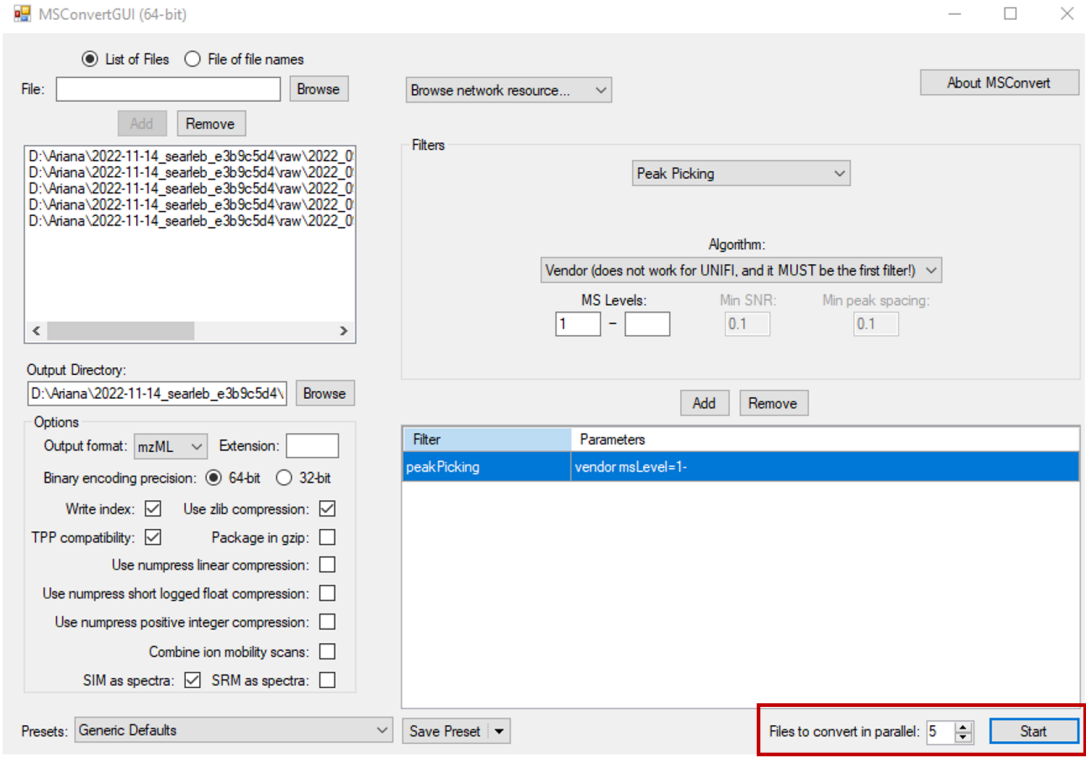
C. Select your files by clicking “browse”, and locate the desired DDA files in your directory. Click “add” on the left-hand side of the screen. The output directory will automatically populate where your files were selected from.

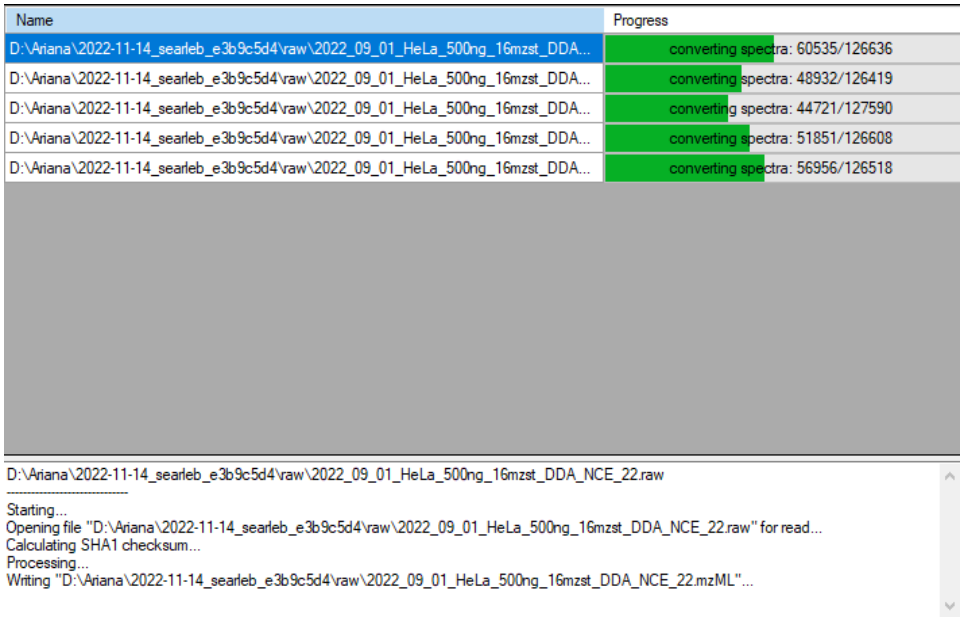


D. Under the options box, the output format should be mzML. You want “write index,” “use zlib compression”, “TPP compatibility, and “SIM as spectra” selected. Your settings should match the screenshot below.

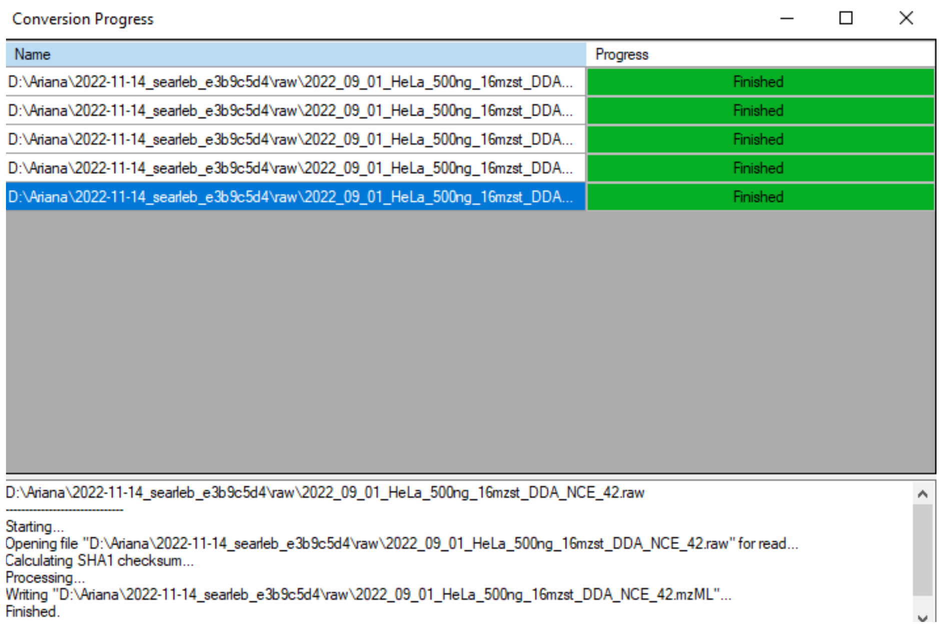


E. Click “Start” in the lower right hand corner to start the file conversion.



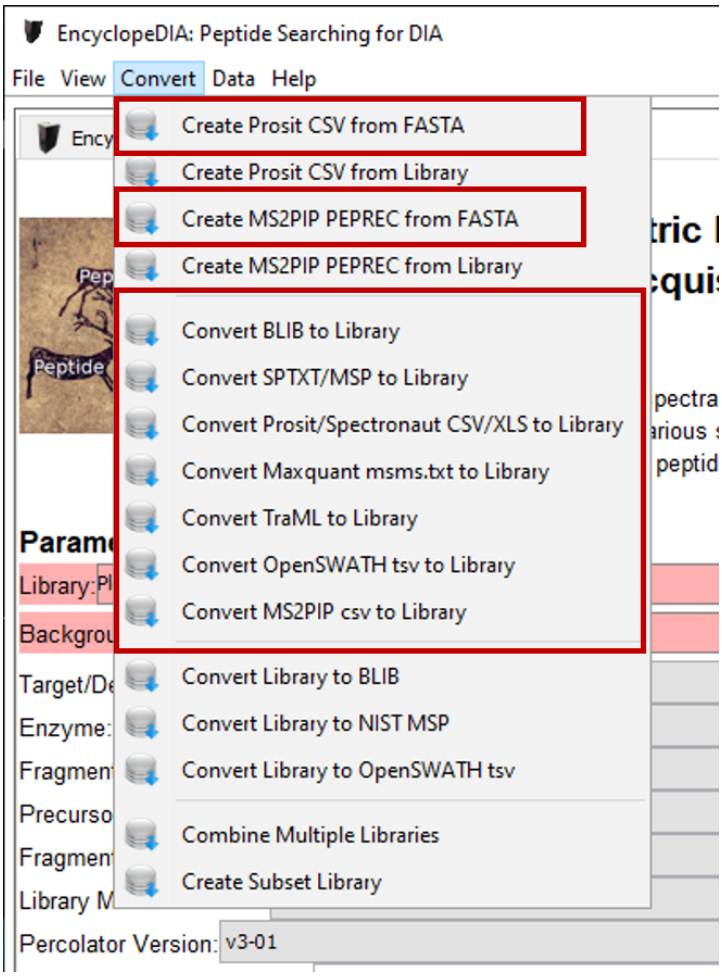


F. Once complete, the window will look like this. You can exit out of MSConvert once the conversion is done. Your files will be in the location previously specified.


Conversion Progress
Name
Progress
D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA...
Finished
D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA...
Finished
D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA...
Finished
D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA...
Finished
D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA...
Finished
D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA_NCE_42.raw
Starting...
Opening file "D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA_NCE_42.raw" for read...
Calculating SHA1 checksum...
Processing...
Writing "D:\Ariana\2022-11-14_searleb_e3b9c5d4\raw\2022_09_01_HeLa_500ng_16mzst_DDA_NCE_42.mzML"...
Finished.

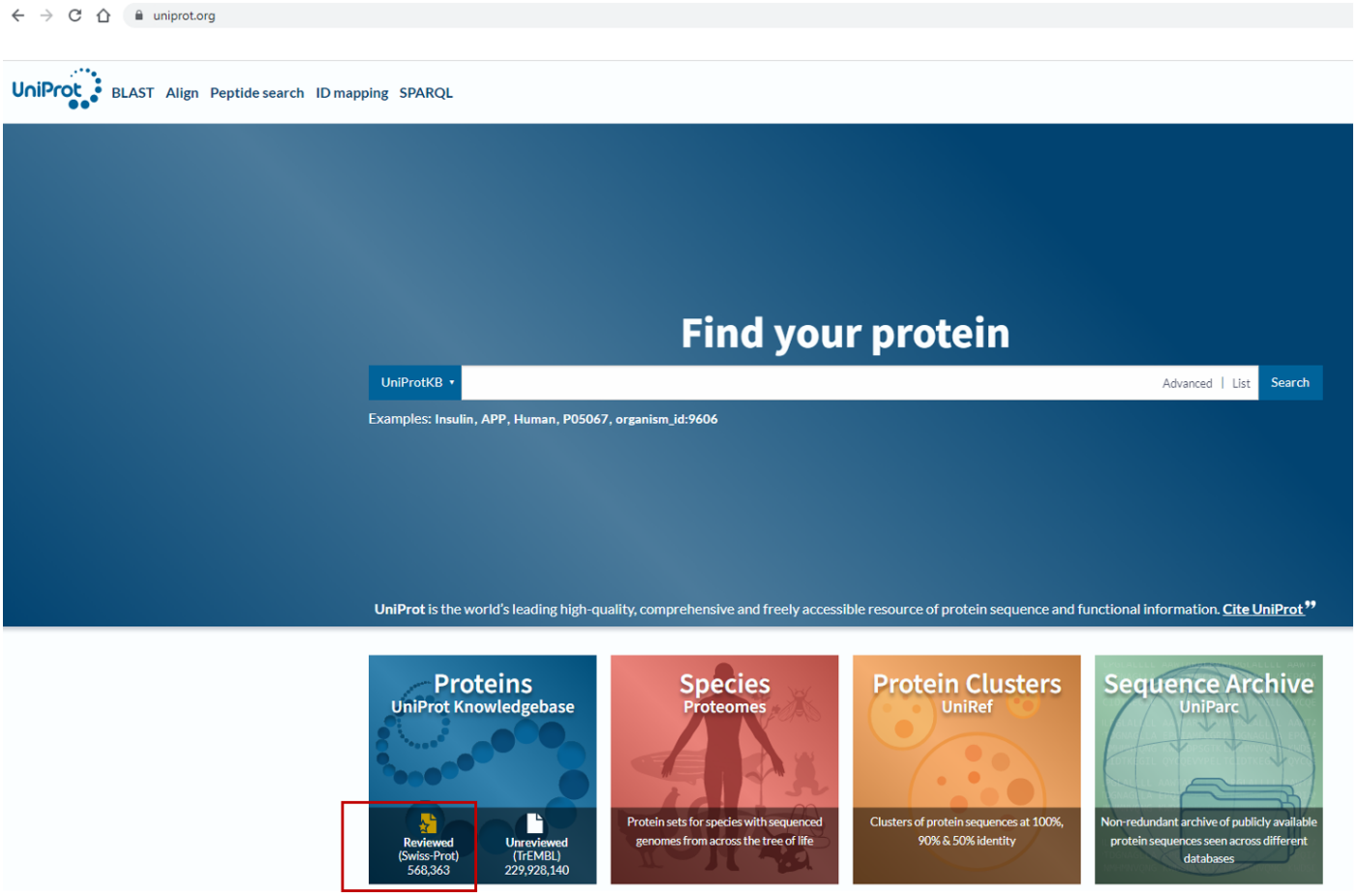
3. GENERATING LIBRARIES USING PROSIT AND SCRIBE

How to acquire a FASTA file and then use Prosit and Scribe to generate a spectral library file from it.

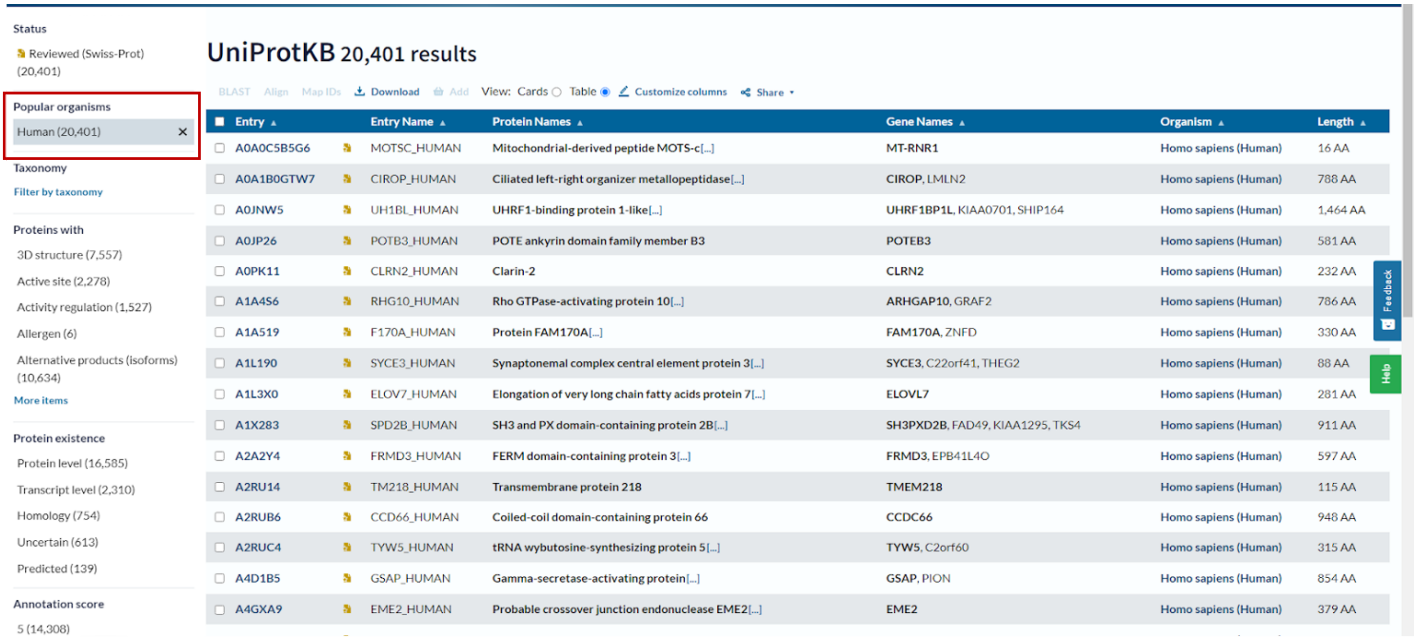


The primary library format Scribe uses is .DLIB. Spectrum libraries from Skyline, NIST, TraML and other formats can be converted to .DLIB using the “Convert” menu. In particular, Scribe can search fully predicted libraries generated by either Prosit or MS2PIP, and produce input files for both of those tools. The following tutorial is given on how to obtain a FASTA file and use Prosit to generate a predicted .DLIB.

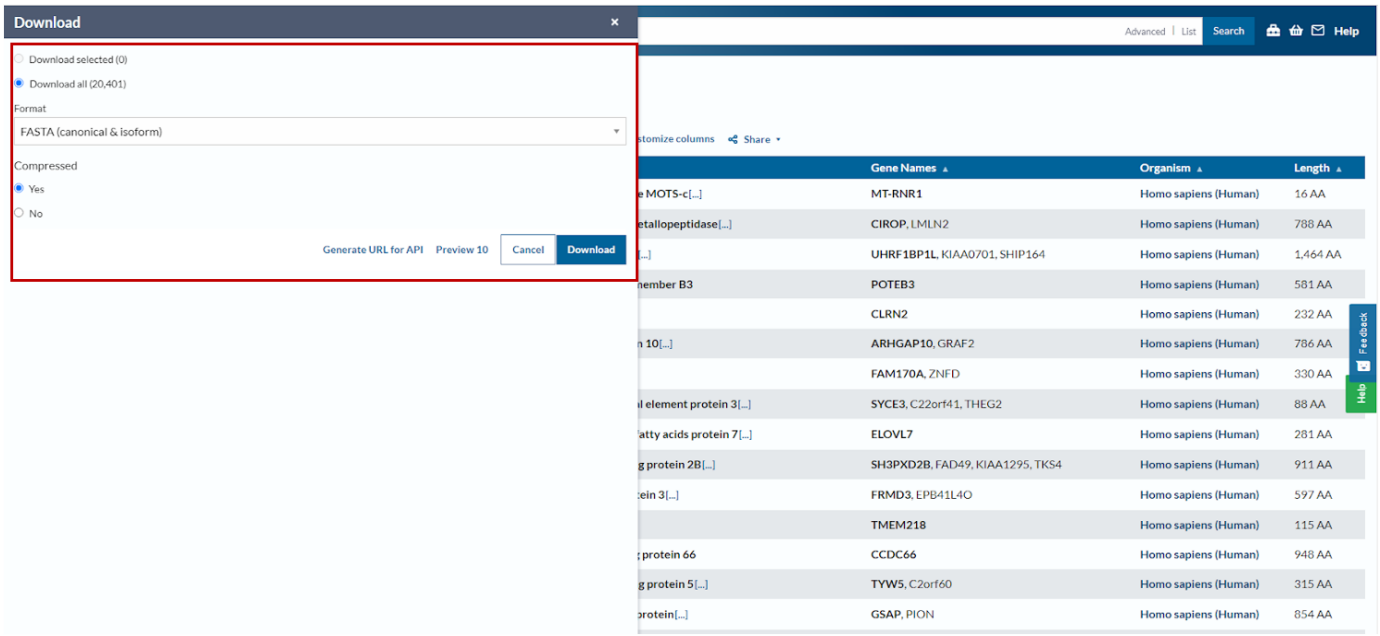
A. To generate a Prosit-predicted spectral library (.DLIB) for Scribe, start by obtaining a FASTA file for your organism of interest. Given the computational expense of generating libraries, in most cases we recommend using only canonical protein sequences. For example, if you require a FASTA of proteins sequences for Homo Sapiens, go to uniprot.org. Select “Reviewed.”



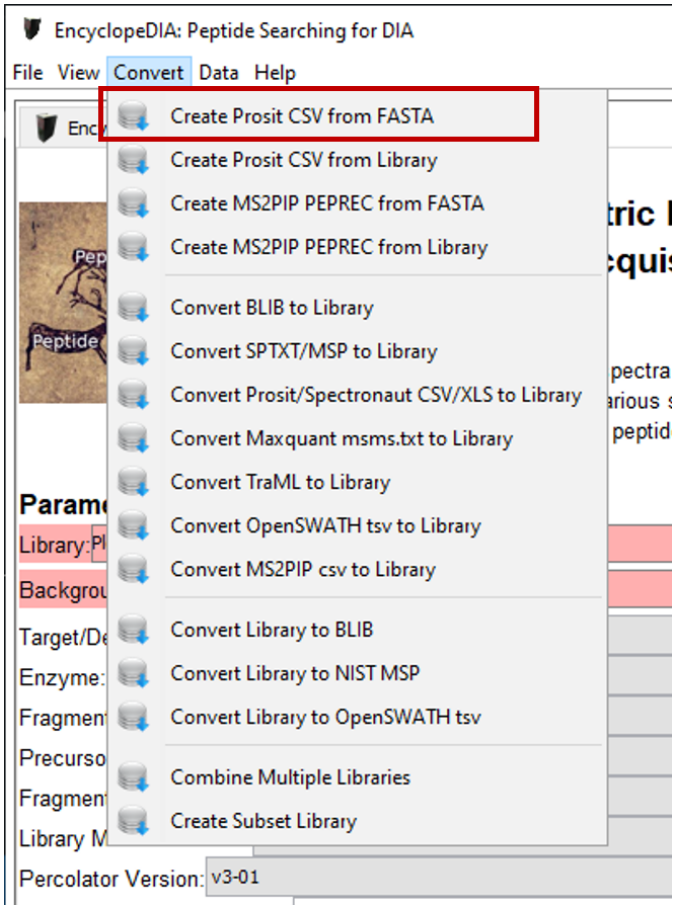
Specify “Human (20,401).”



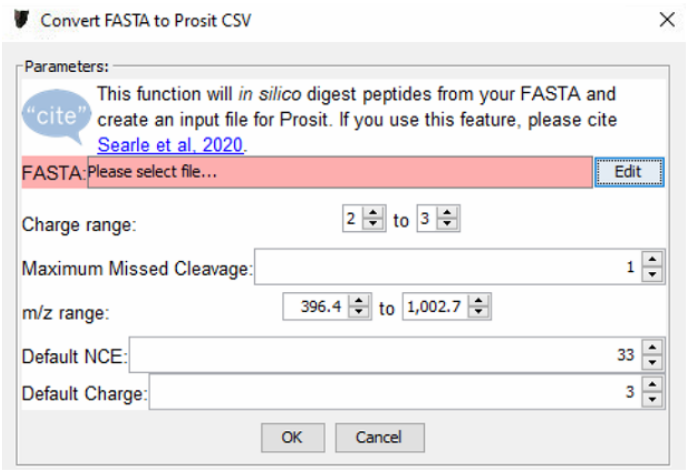
Select “Download.” If you download the compressed version, you will get a .gz and have to uncompress the file. If you download the uncompressed version, the file may take longer, but it will be in a FASTA format for immediate use.



B. Open Scribe to create a Prosit CSV from a FASTA file. Navigate to the “Convert.” Select “Create Prosit CSV from FASTA.” This will open a dialog window.

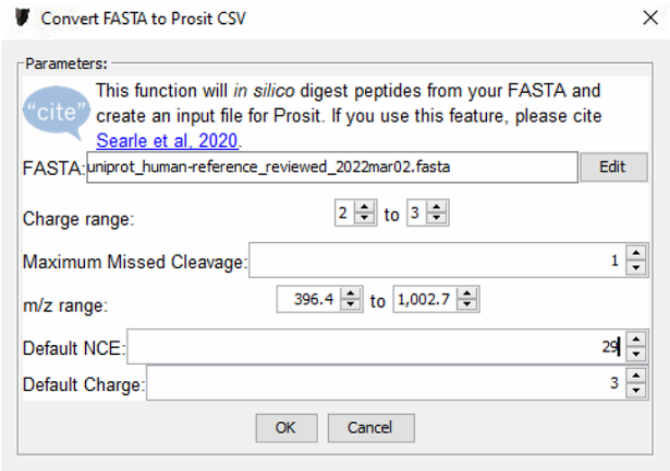


Default Settings

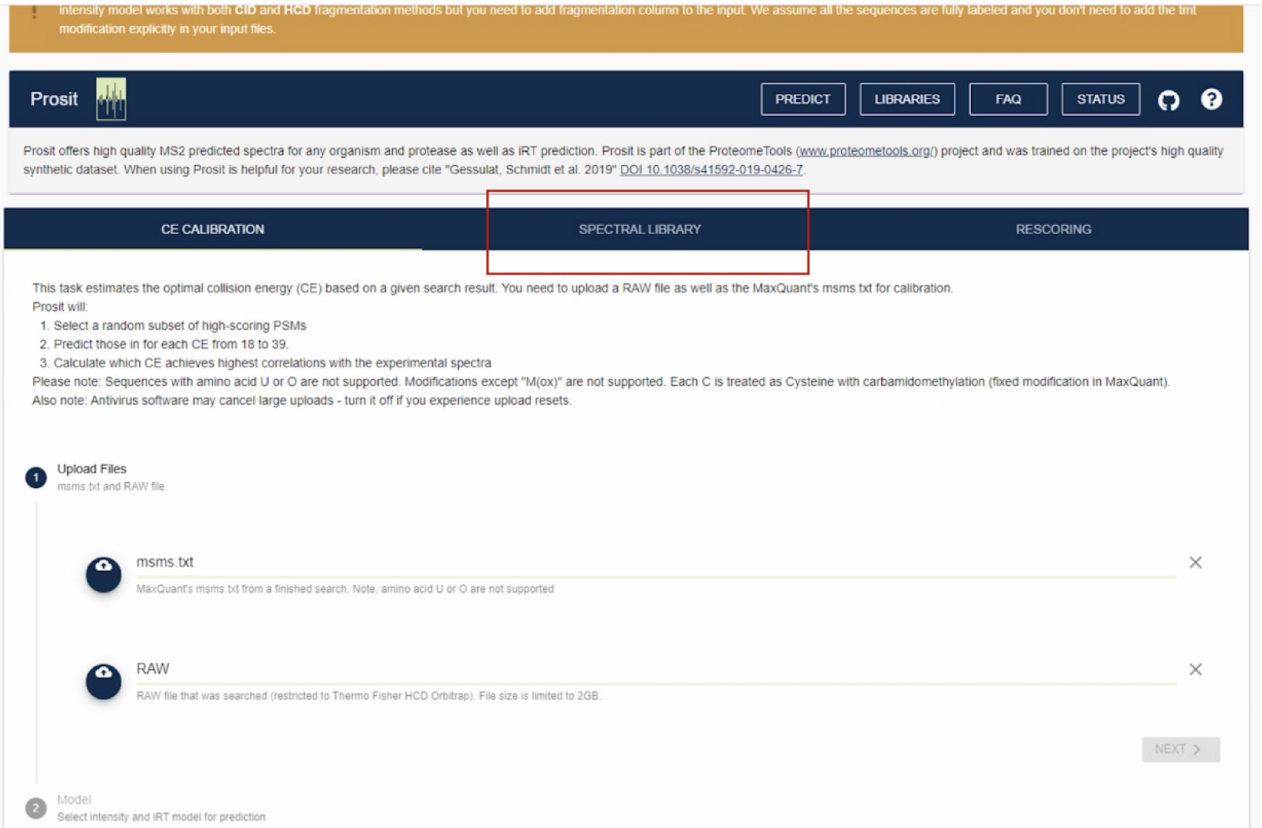


Upload the FASTA you have downloaded. The Prosit CSV will be generated in the same folder where the FASTA is held. If you use the example from the FASTA file provided in the Scribe manuscript, you should load “uniprot_human-reference_Reviewed_2022mar02.fasta.”

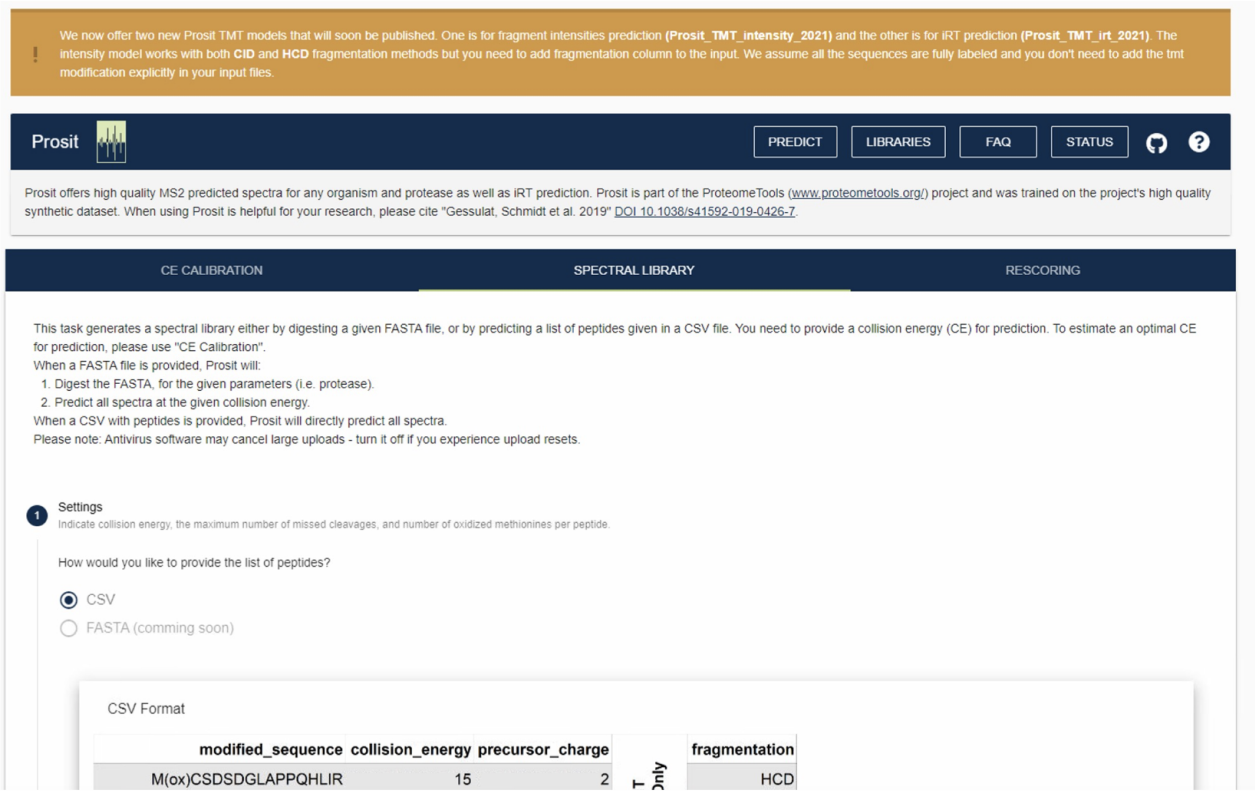
Settings for
“uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce29_hcd2020.dlib”



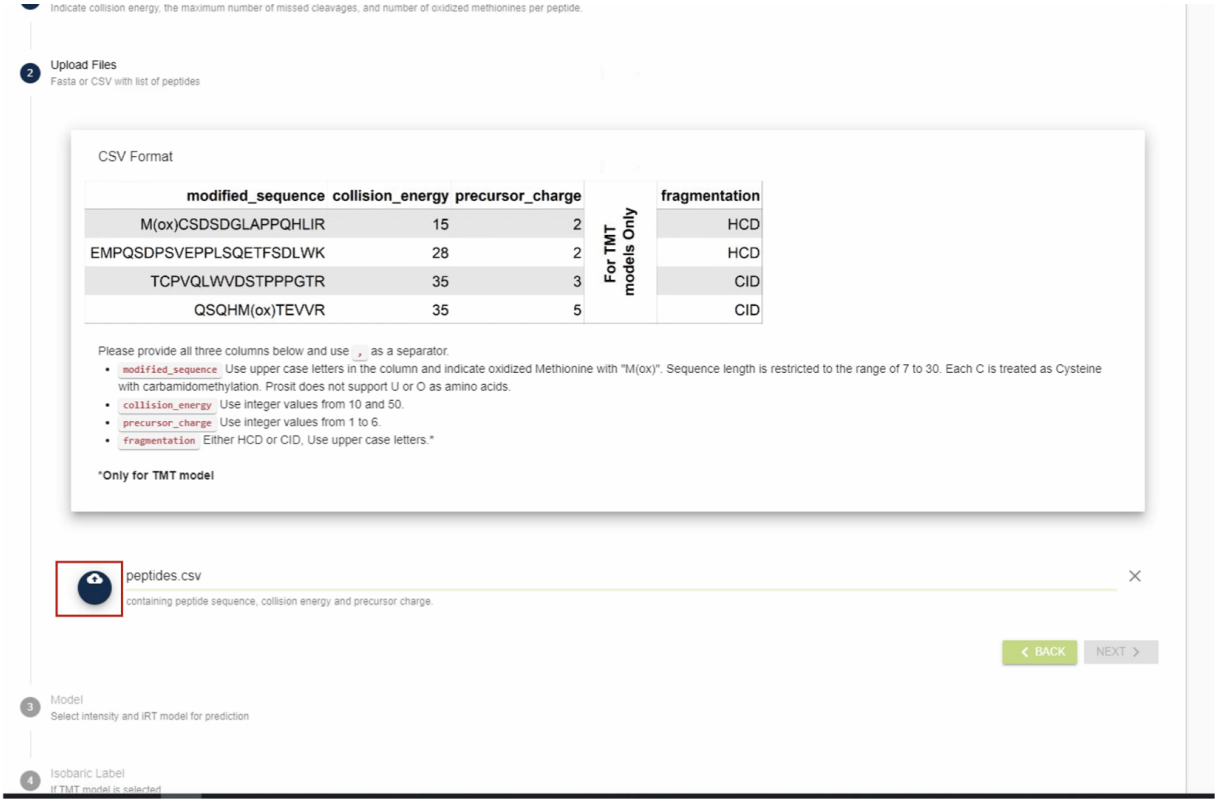
C. Use Prosit to perform an *in silico* digestion, and obtain an .MSP/NIST format text file. Navigate to the Prosit website, Prosit. When you get to the page, there are three tabs available; “CE CALIBRATION,” “SPECTRAL LIBRARY,” and “RESCORING.”



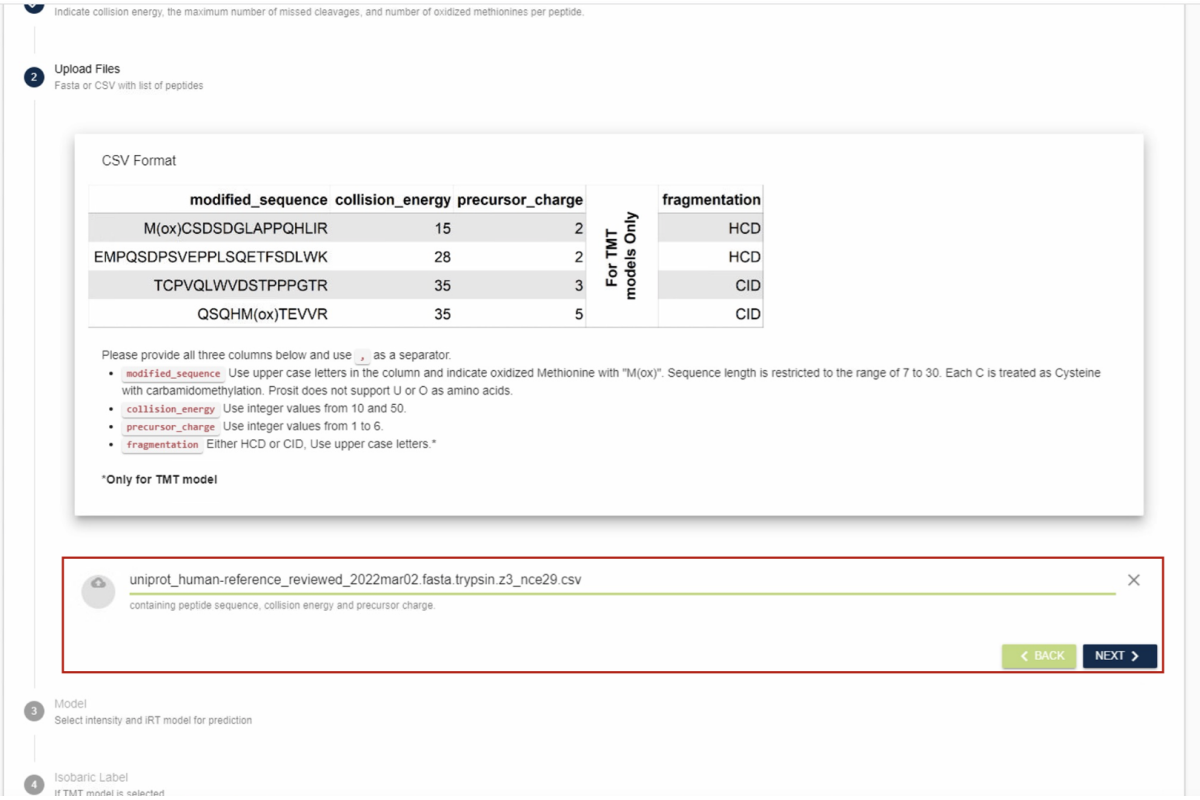
Go to the “SPECTRAL LIBRARY” tab.



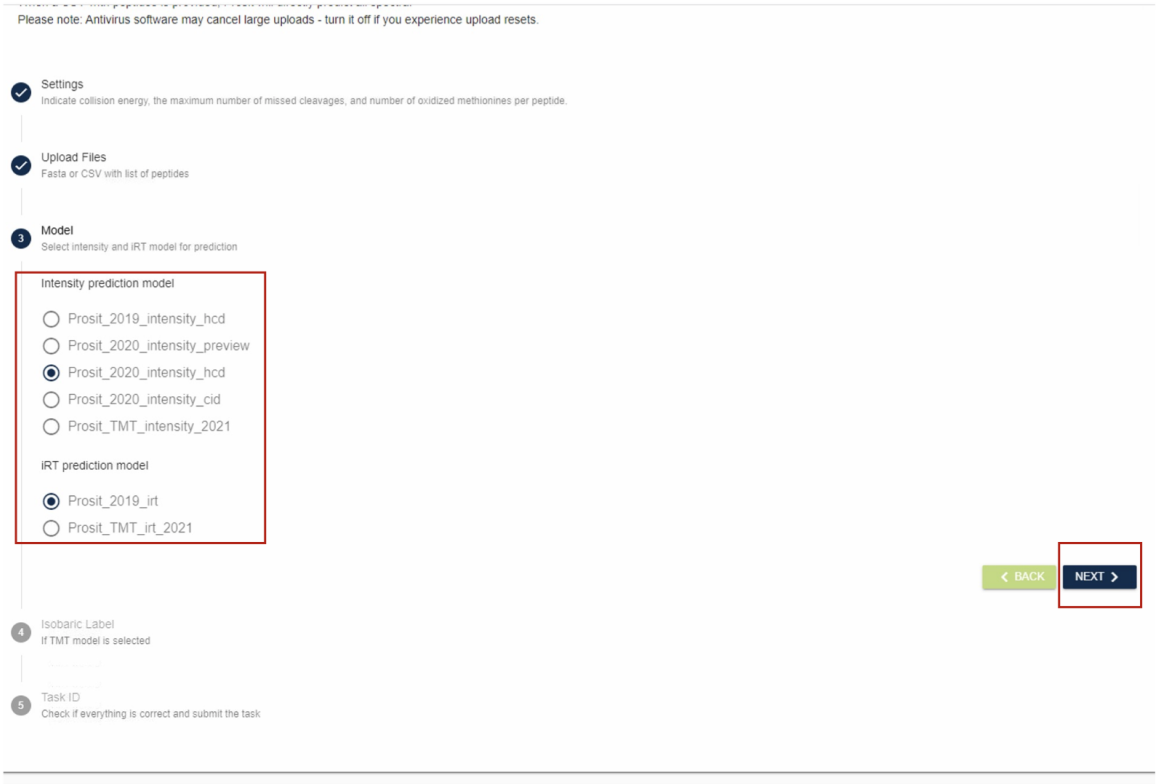
For the 1st step, CSV is already selected. You can click next to get to the 2nd step. Upload the Prosit CSV created in Scribe.



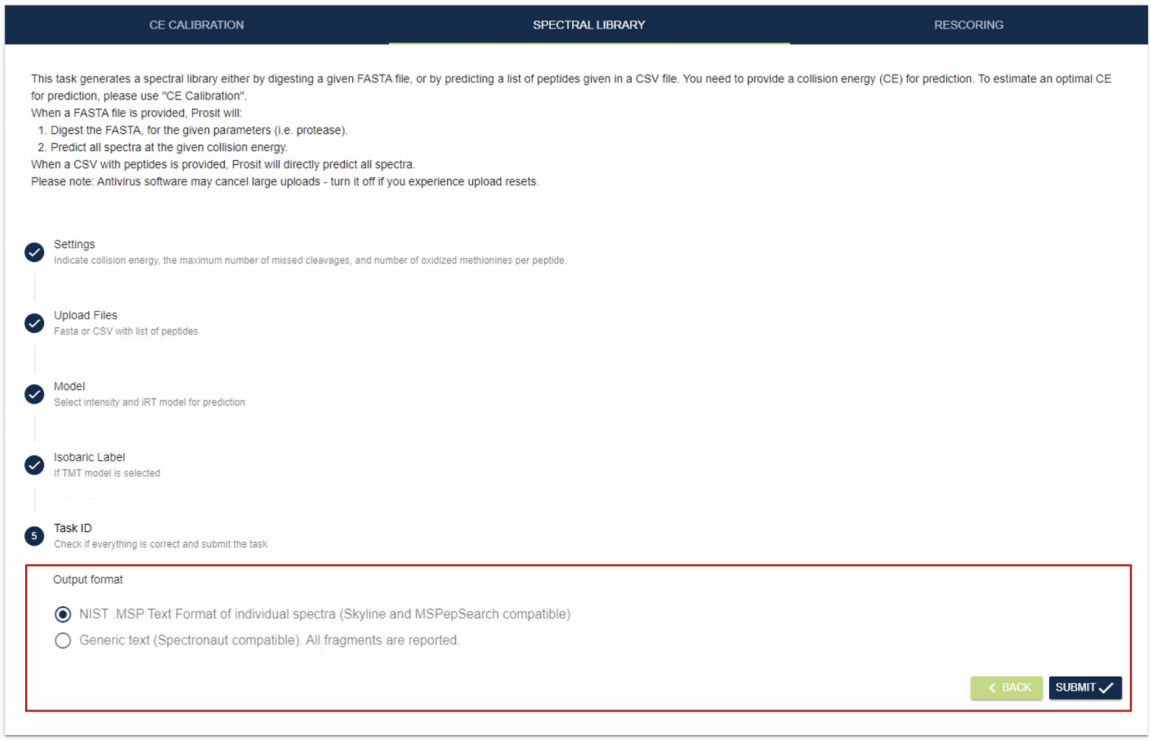
Once the CSV has once loaded, click next.



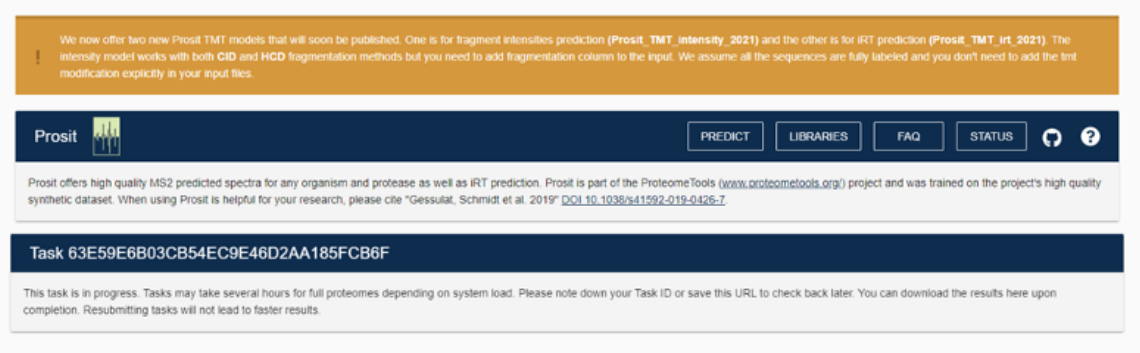
Select the desired model. For this example, we want to use the “Prosit_2020_intensity_hcd” for the Intensity prediction model, and the “Prosit_2019_irt” for the iRT prediction model. Click next.



For Task ID Output format, select “NIST .MSP Text Format of individual spectra (Skyline and MSPepSerach compatible).



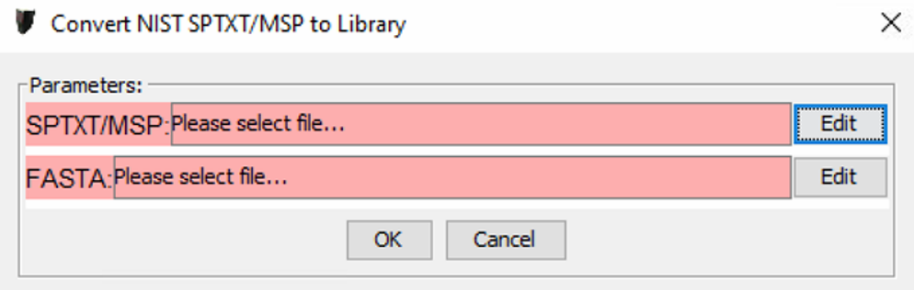
Submit the task. It is helpful to record the task number or bookmark the page shown after you submit the task to come back to once the job is complete.



Once the task is done, download the .MSP file.

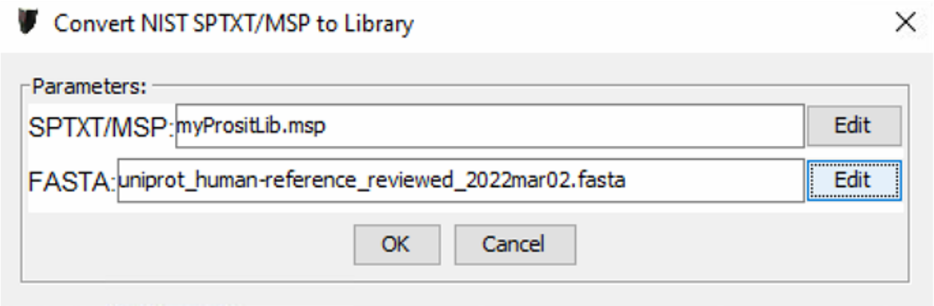
D. Convert the Prosit output (.MSP/NIST) to a Library (.DLIB) using Scribe. In Scribe, go to “Convert,” then navigate to “Convert SPTXT/MSP to Library.” This window will pop up.

Default

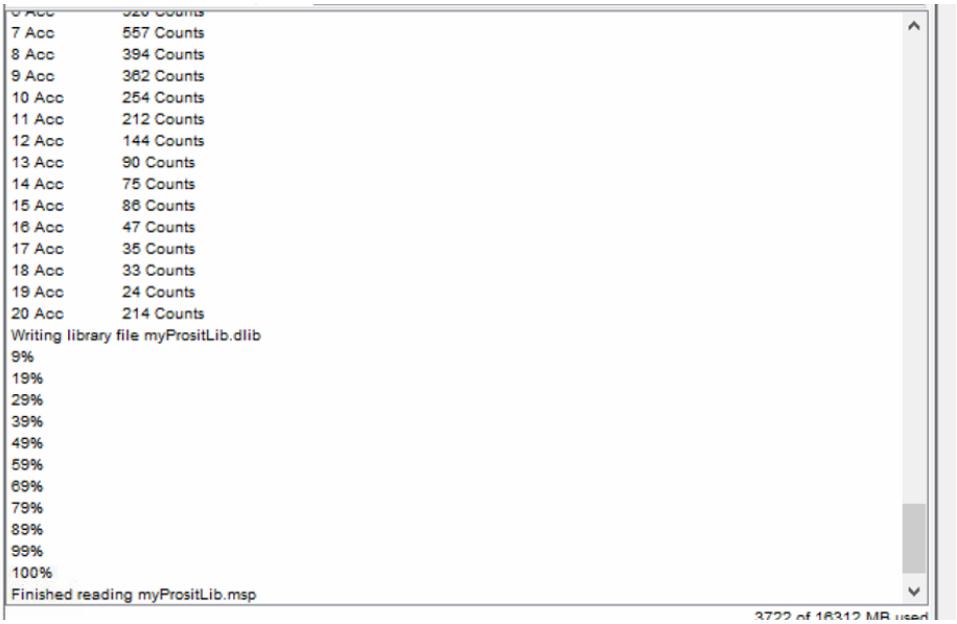


Upload the downloaded .MSP file from Prosit, and the FASTA file used to generate the .MSP.

After Uploading Files



Click okay. The dialog box will indicate the .MSP file is being converted to a .DLIB. Once it is complete, the dialog box will tell you.

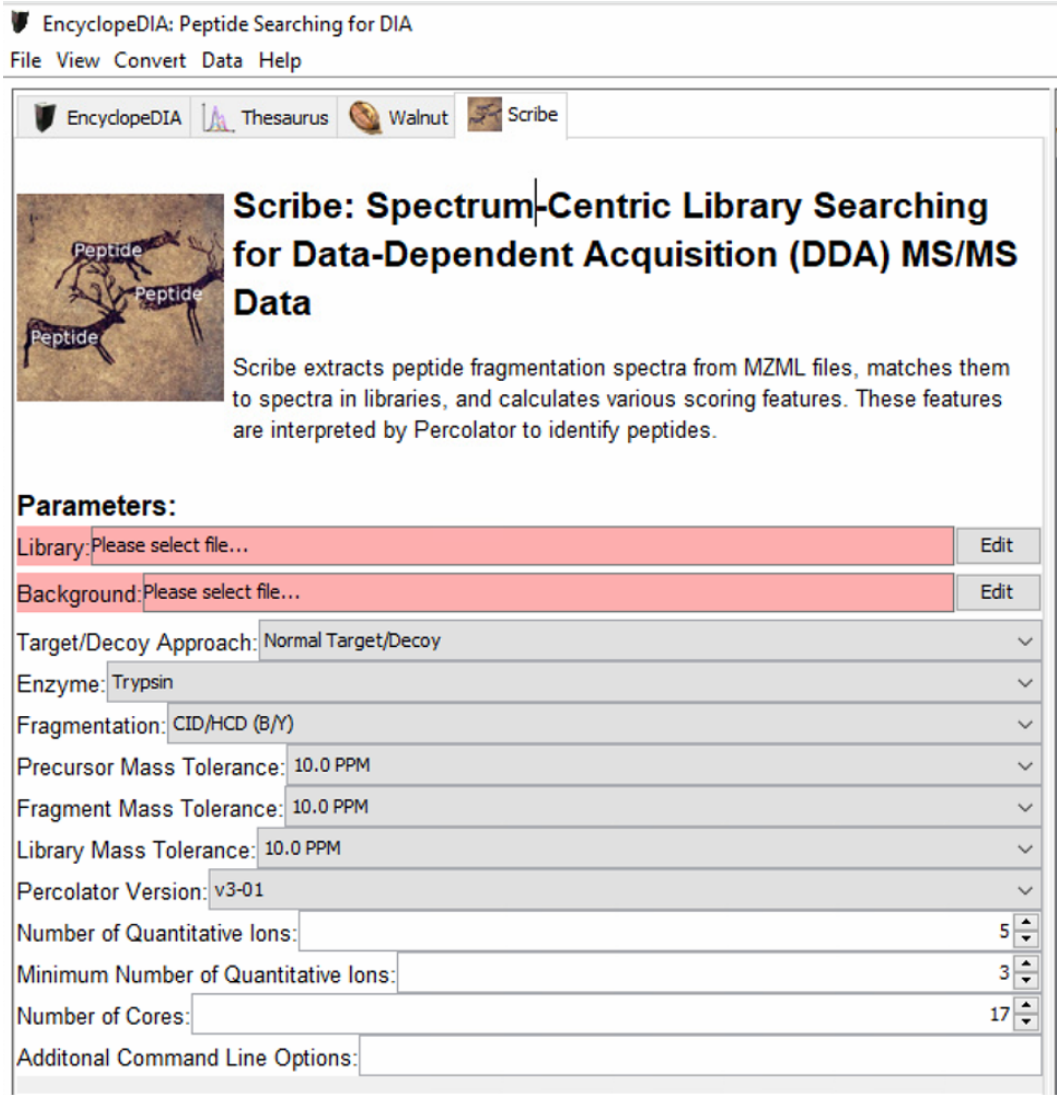


The resulting .DLIB should be analogous to uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce29_hcd2020.dlib.

4. SCRIBE SEARCH OPTIONS

Descriptions of the search options available in Scribe.

Here is a screenshot of the Scribe default search options:



A. Scribe has several options for searching files. Before you can start loading data, you need to specify both a .DLIB or .ELIB library to search as well as a background FASTA. These will be shaded red until they are properly specified. Libraries can be either in the .ELIB (chromatogram library) or .DLIB (spectrum library) format.

B. Scribe has several other search settings. As a general rule, we recommend using the default search parameters first. Other settings are defined below:

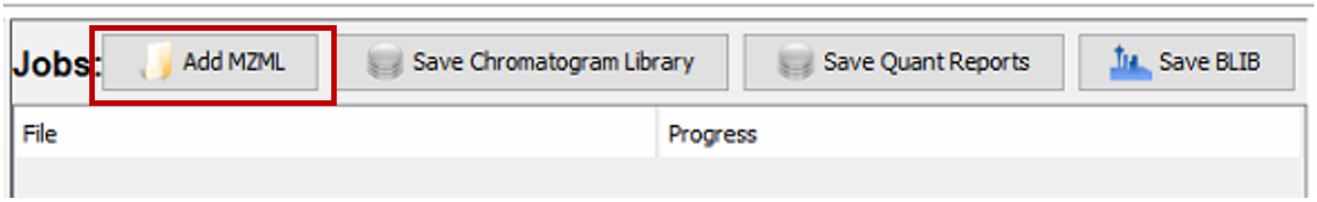
Target/Decoy Approach: In some circumstances, it may be necessary to add additional decoys to improve statistical analysis. However, as a general rule, this should be left at “Normal Target/Decoy”.

Enzyme: Several common digestion enzymes are supported.

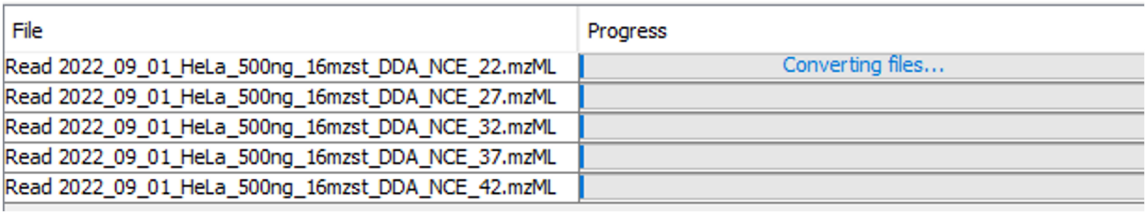
Fragmentation: In general, we recommend using CID/HCD (B/Y) fragmentation for most CID or HCD experiments. However, if your library is particularly large or messy you may get improved results with “HCD (y- only)”.
Precursor/Fragment/Library Mass Tolerance: Tolerances can be specified in PPM, AMU, or resolution.
Percolator Version: Percolator 3.1 is recommended for most experiments.
Number of Cores: This is the number of CPU cores you allow EncyclopeDIA to use. The maximum value you should set this to is one less than the number of cores your computer has. You need to leave at least one core for background processes.

C. Additional command line options can be specified in the command line options box. For example:

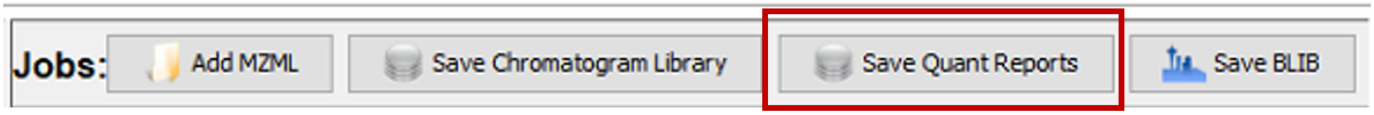
D. Use the “Add MZML” button to add RAW files for library searching. These will be automatically placed on the queue and executed in order using the current settings. If an MZML has been previously analyzed, Scribe will remember where it left off and try to not process it a second time.



Here is a screenshot of the Scribe search queue:



E. Search results can also be saved for downstream quantitative assessment using the “Save Quant Reports” button. RAW files between these experiments are expected to contain shared peptides so retention-time alignment is performed and match-between-runs quantification is calculated.

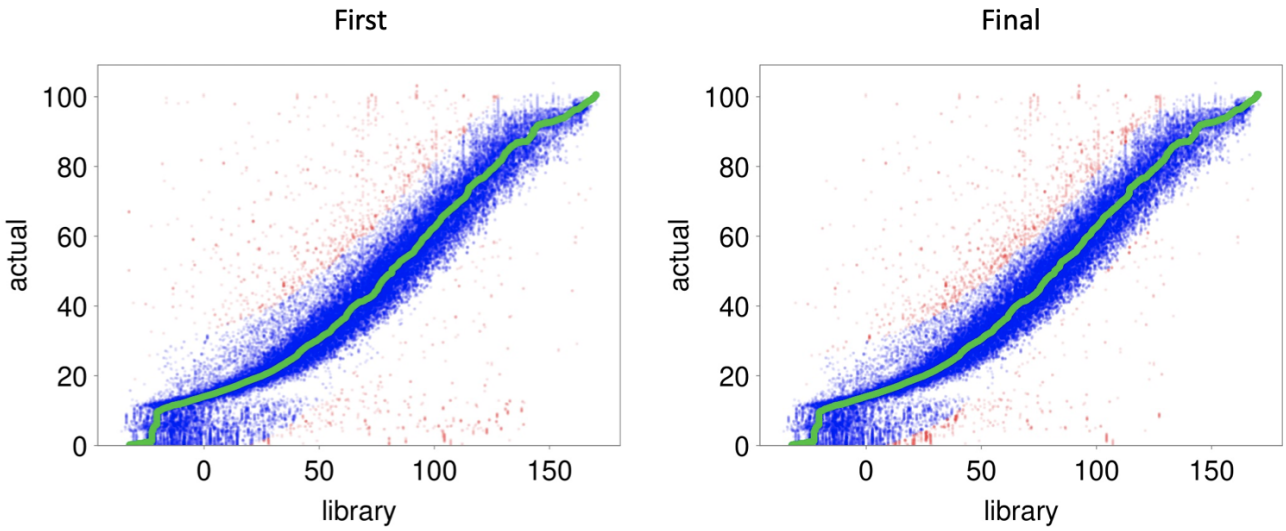


5. INTERPRETATION OF SCRIBE OUTPUTS

Descriptions of the output files given from Scribe.

Scribe outputs figures as PDFs displaying fits for retention time (RT), MS1, MS2, delta MS1, and delta MS2. These files can be used to partially assess Scribe's performance in matching spectra to peptides. Additional metrics can be found using the Features viewer, described in part 6 of this manual.

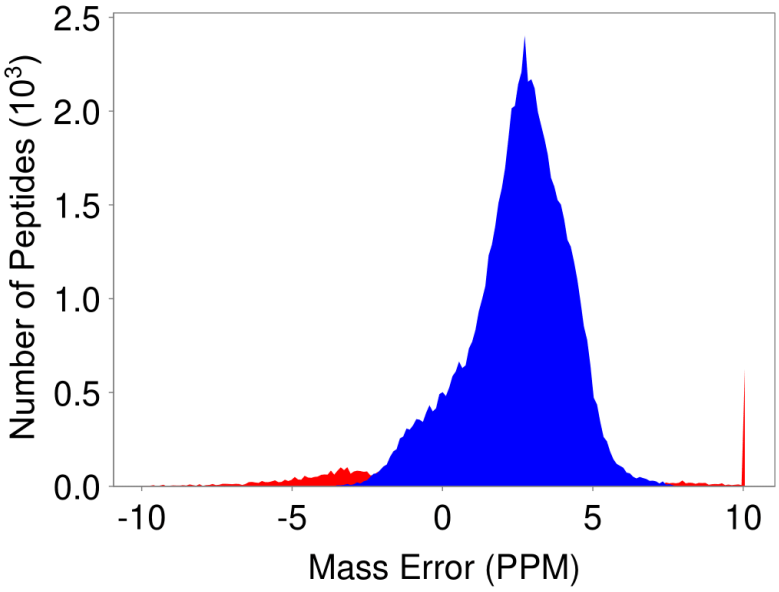
A. PDFs and corresponding text files



RT plots show the Prosit-predicted RT to the actual RT detected. The green line shows the trace of the kernel density estimation (KDE) RT fit. The blue dots indicate peptides that are matched within 3 standard deviations of the green fit. The red dots represent peptides that are rejected by the alignment, which are outside of 3 standard deviations of the median fit.

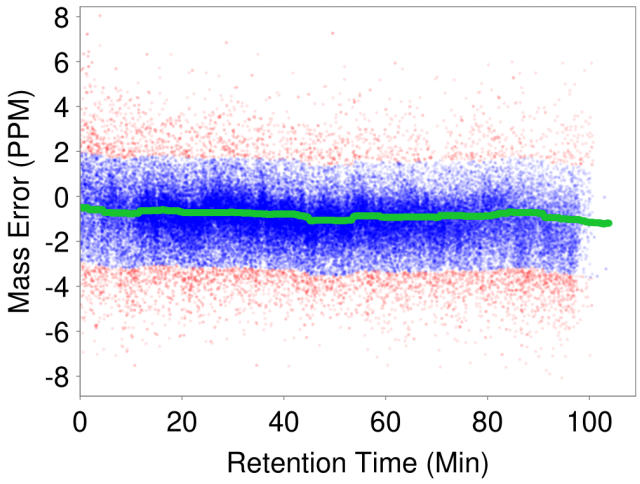
In the first fit, peptide FDRs are calculated using Percolator, then fed back into the RT algorithm for further refinement. The refined, or final retention time fit is represented in the second plot. Peptides that are sent through the second pass of the Percolator are a low fraction of the peptides that were fed into the first pass. As a result, the first and final traces only look subtly different. For the remaining plots, only the final fit will be shown.

Final Delta MS1 Fit



Similarly to the RT fit, the delta MS1 fit displays the mass error of peptide detections. The error is calculated by comparing the expected mass of a given precursor to the detected mass. The blue portion of the peak represents peptides found within 3 standard deviations of this mass error fit, while the red represents peptides outside of this range.

Final MS2 fit



Precursor m/z and fragment m/z have outputs that plot RT vs precursor mass error. We expect the acceptable error to be within the mass error that is selected during the peptide search. For this search, a mass tolerance of 10 PPM was selected for fragment ions, which gives the green line, or the median mass error fit. The blue dots represent peptides identified within 3 standard deviations of the median mass error, while the red dots are peptides that fall outside of this range. Although the MS1 and delta MS2 fit are not shown, both plots can be interpreted using the same principles described here.

D. Other files generated during sample searches and quantitation

Scribe.txt, decoy.txt, and features.txt

The ending in ".scribe.txt" contains 6 columns. The first column is the PMSid or sample name concatenated to intensity, peptide sequence, and charge. The next columns contain the scribe score, q-value, and posterior error probability. The Scribe score is a sum of squares error calculated from the actual intensity and the predicted intensity. This value reports the difference between the observed intensity and the predicted intensity. The last two columns contain peptides with protein matches. You can use this text file to asses the peptide spectral match for targets of interest

Files ending in .scribe.decoy.txt contain analogous scoring and information for the decoys generated for FDR estimation.

Features.txt files contain the complete 17 scores the Scribe algorithm uses to match peptide spectra to their target proteins using Prosit-predicted libraries.

Protein and Peptide.txt

After clicking "Save Quant Reports," a quantitative output is formatted in easy-to-use .txt files at both the protein and peptide levels. The protein and peptide reports use match-between-runs for quantification and therefore are only generated if you are searching more than one .mzML file. Quant report files are formatted as expression matrices that can be analyzed by hand in a program such as excel, or reformatted for use in a package using software like R or Python. Samples are in columns, while peptides or proteins are in rows. The number of peptides and the number of fragments are included as an additional column.

Diffacto inputs

Scribe also produces easy-to-outputs to use in Python with the differential expression tool Diffacto. This software, developed by Lukas Käll's research group, is capable of processing and normalizing data before performing a differential expression analysis using bayesian modeling to select only peptides that behave similarly to each other. Scribe writes the necessary .csv file for sample abundances, and the .lst file containing annotations. The .lst output contains sample annotations. The default lists all samples as separate sample groups. You will need to manually enter sample groupings before using Diffacto, if you wish for samples to be grouped by condition. To use Diffacto, consult its Github page.

Diffacto citation: Zhang, Bo et al. Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences Molecular & Cellular Proteomics, Volume 16, Issue 5, 936 - 948. https://doi.org/10.1074/mcp.O117.067728.

F. Library, .dia, and model files

When Scribe performs quantitation on individual files, or match-between-run quantitation of multiple files, an ELIB file is generated as a part of the output. ELIB files contain the same information present in DLIBs, with retention time information added. The ELIB output can be uploaded to Skyline or EncyclopeDIA to visualize peptides, or opened in a SQLite DB Browser to extract specific information using a query. Refer to part 6 for more information on how to visualize ELIBs and DLIBs.

Scribe writes .dia files as it is processing samples against predicted libraries. These files are read by EncyclopeDIA when visualizing .mzML files, and can also be read in Skyline.

MODEL files are necessary for visualizing data in EncyclopeDIA, and should not be deleted or moved.

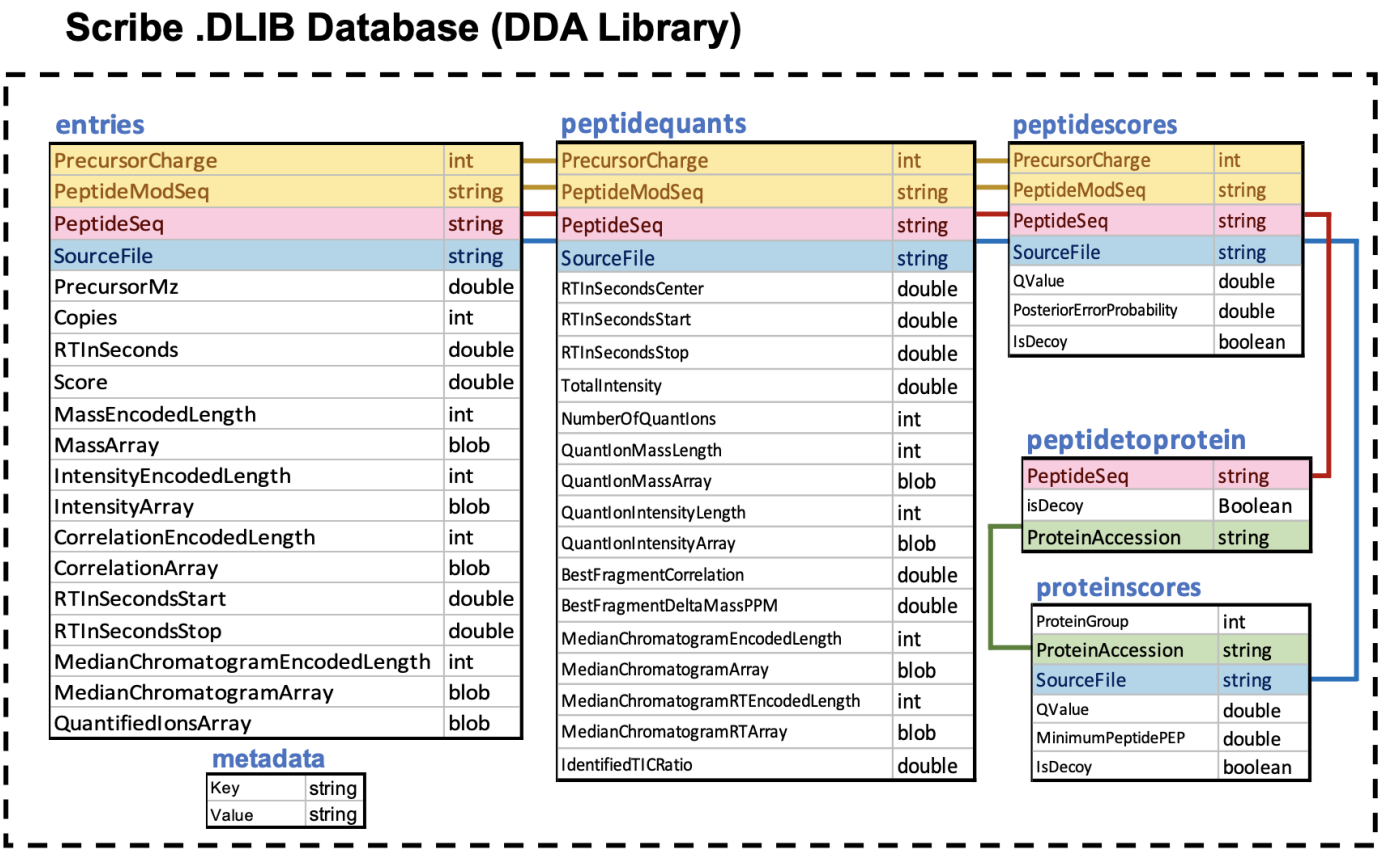
6. VISUALIZING DATA WITH BUILT-IN FEATURES

Three ways to access data processed by Scribe,

Using the outputs you obtain from Scribe or EncyclopeDIA, you can visualize the data using three methods; through an SQLite database browser, using Scribe, or looking through the data manually within excel. This is not an all-inclusive list of ways to analyze or visualize the data.

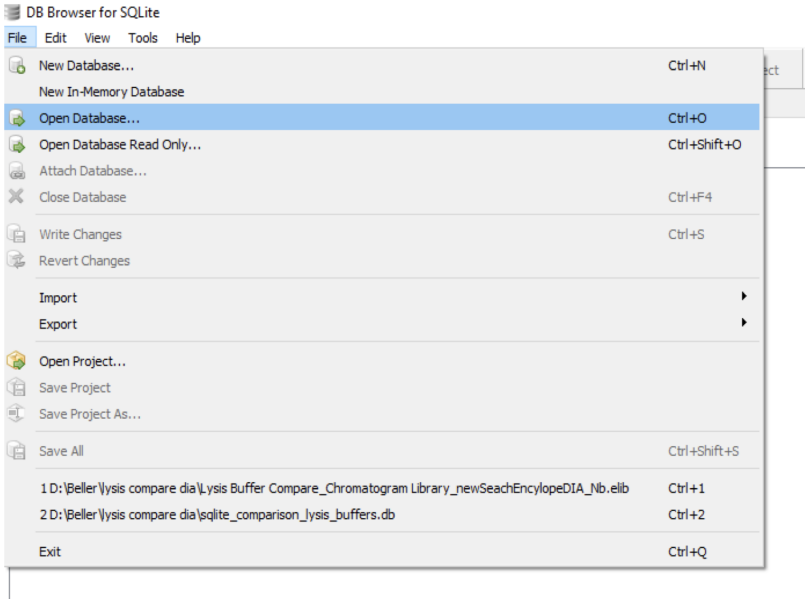
A. Accessing the data through SQLite

ELIB and DLIB files are formatted as SQLite databases, allowing users to access the data through an SQLite viewer, such as DB Browser for SQLite. Below is a map of how DLIB databases are structured. SQL databases contain tables, which are represented as boxes in the map, the names of these tables are “entries,” “metadata,” “peptidequants,” “peptidescores,” “peptidetoprotein,” and “proteinscores.”Each box contains the column names within the tables and the type of data each column contains. For example, “PeptideCharge” is formatted as “int,” indicating the peptide charges are integers in the data table.

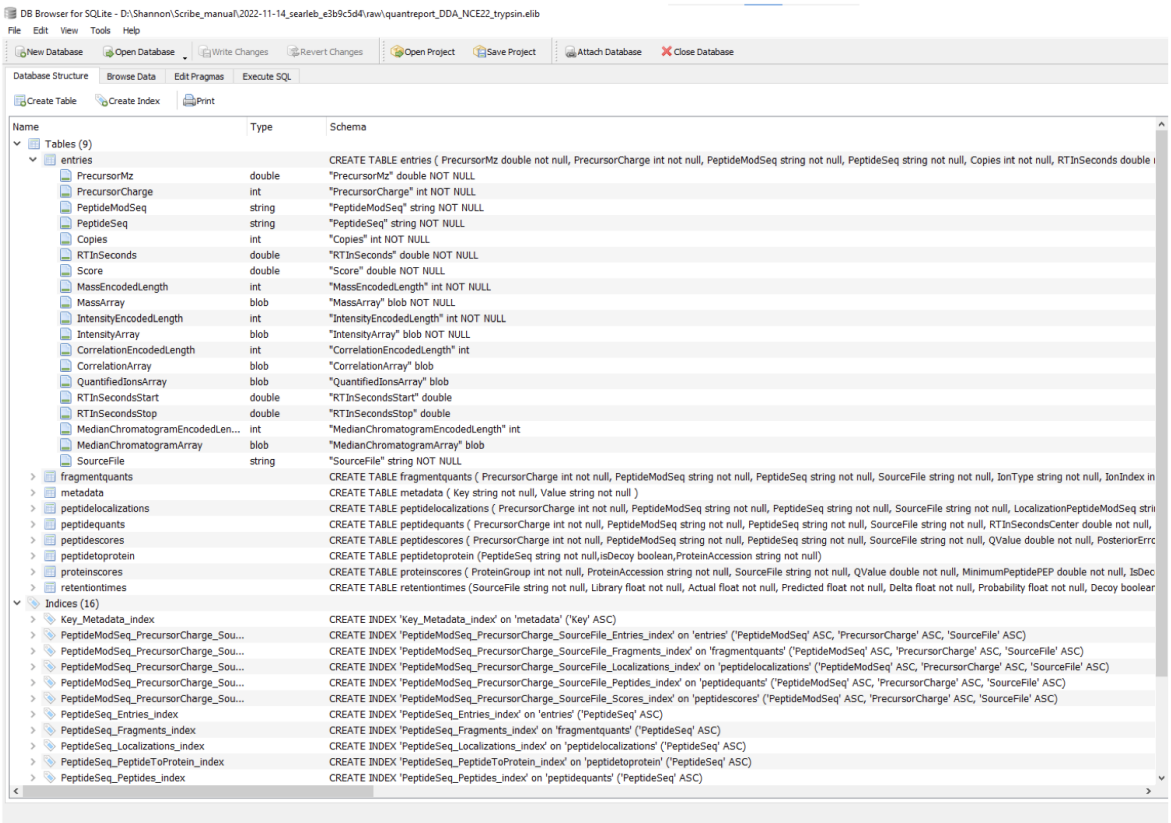


To open the browser, double-click the icon on the desktop, or find the program in the location it was downloaded.

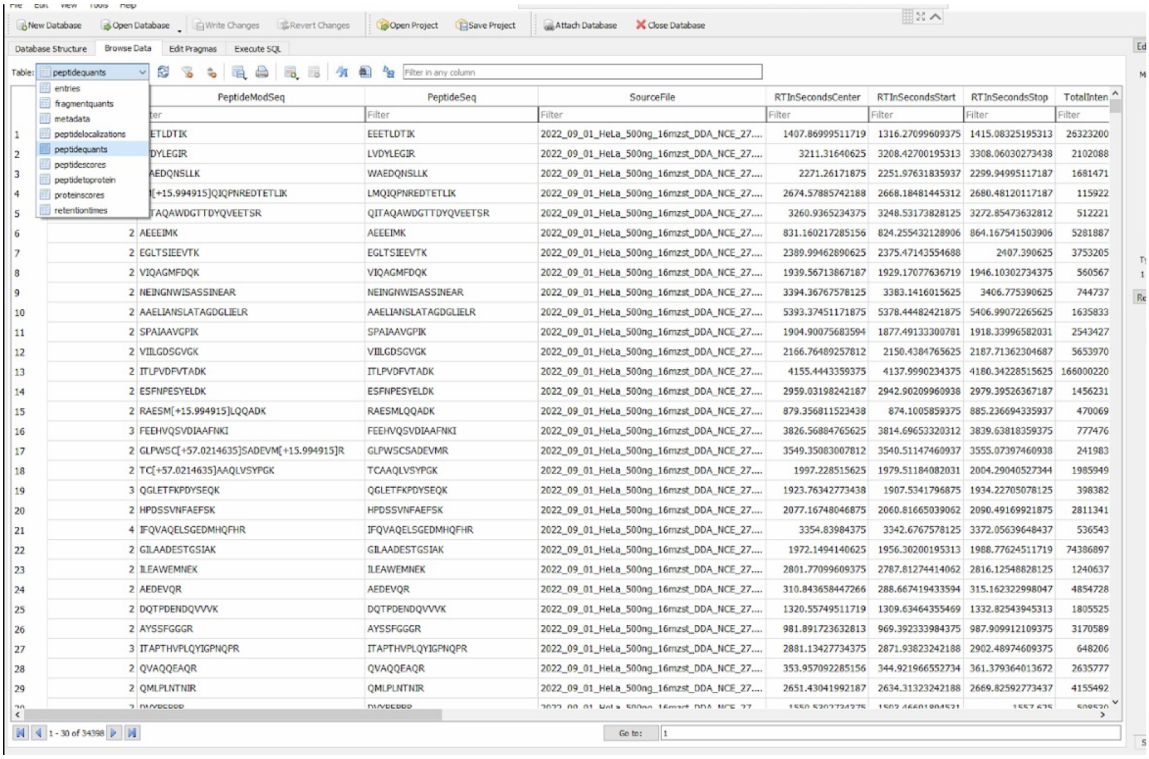
Once the database browser is open, go to the File > Open Database menu option, then find the ELIB in the location where it is stored.



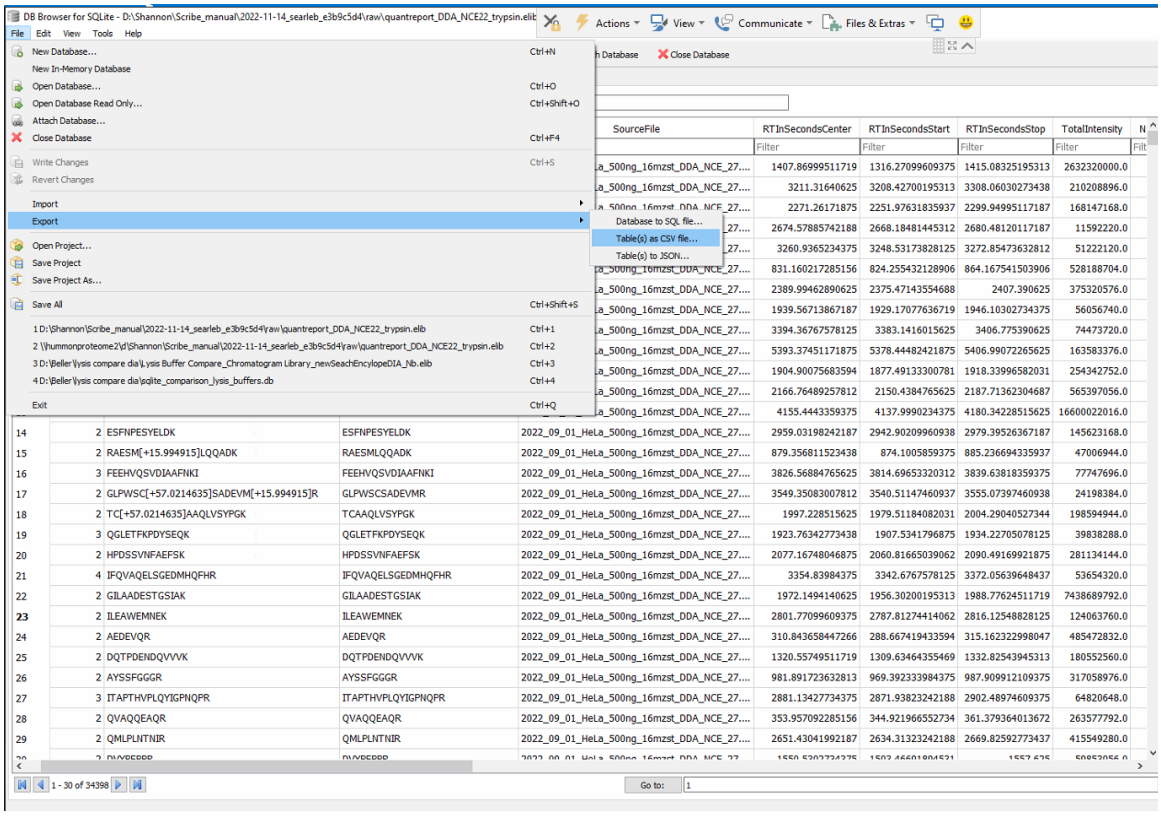
On the first tab, you can visualize the database structure. For example, the first file in the database is the “entries” table, as you can see below. This matches the information seen in the map above. You can see there are 9 tables in total. Below the tables, are indices, which are lookups for finding information within tables.



If you click on the next tab, “Browse Data,” you will be able to look through tables.



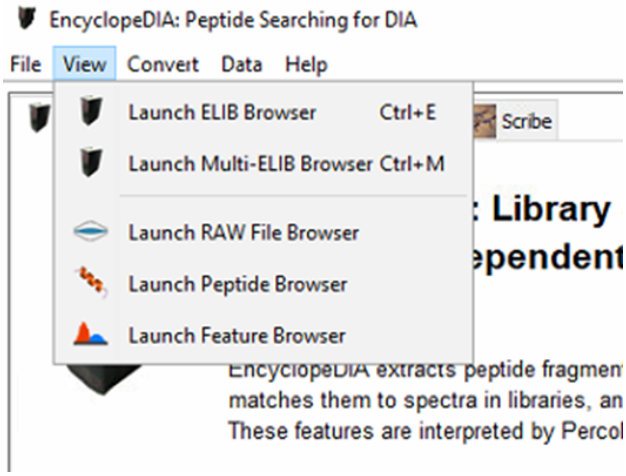
You can use filters to obtain specific data you are interested in, or use regular expressions for a more powerful query. To export a table you are interested in specifically, click “File > Export > Table(s) as CSV.”



For more information on using DB Browser for SQLite check out the SQLite Browser Wiki Github Page.

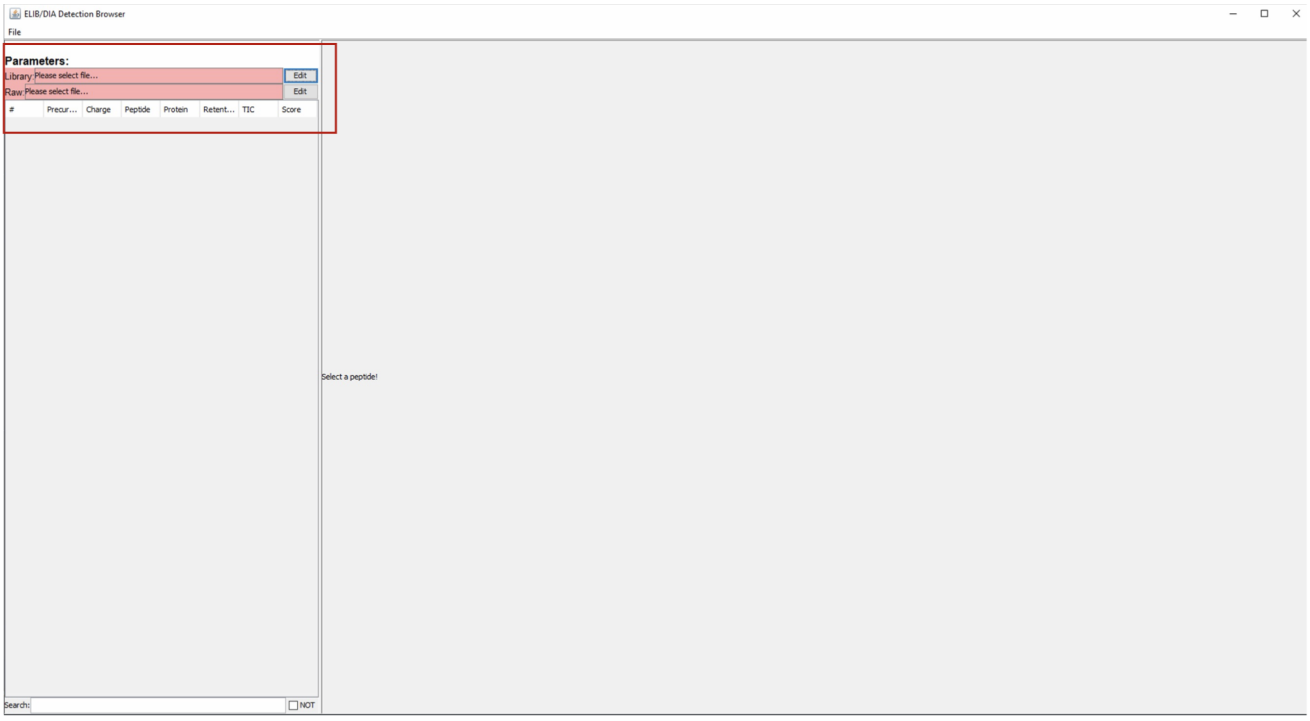
B. Looking at data through Scribe's built-in visualization tools.

EncyclopeDIA and Scribe have built-in tools to help you visualize raw data, in .mzML format, libraries, and quant reports. Use the View menu to access the following options for visualizing data.

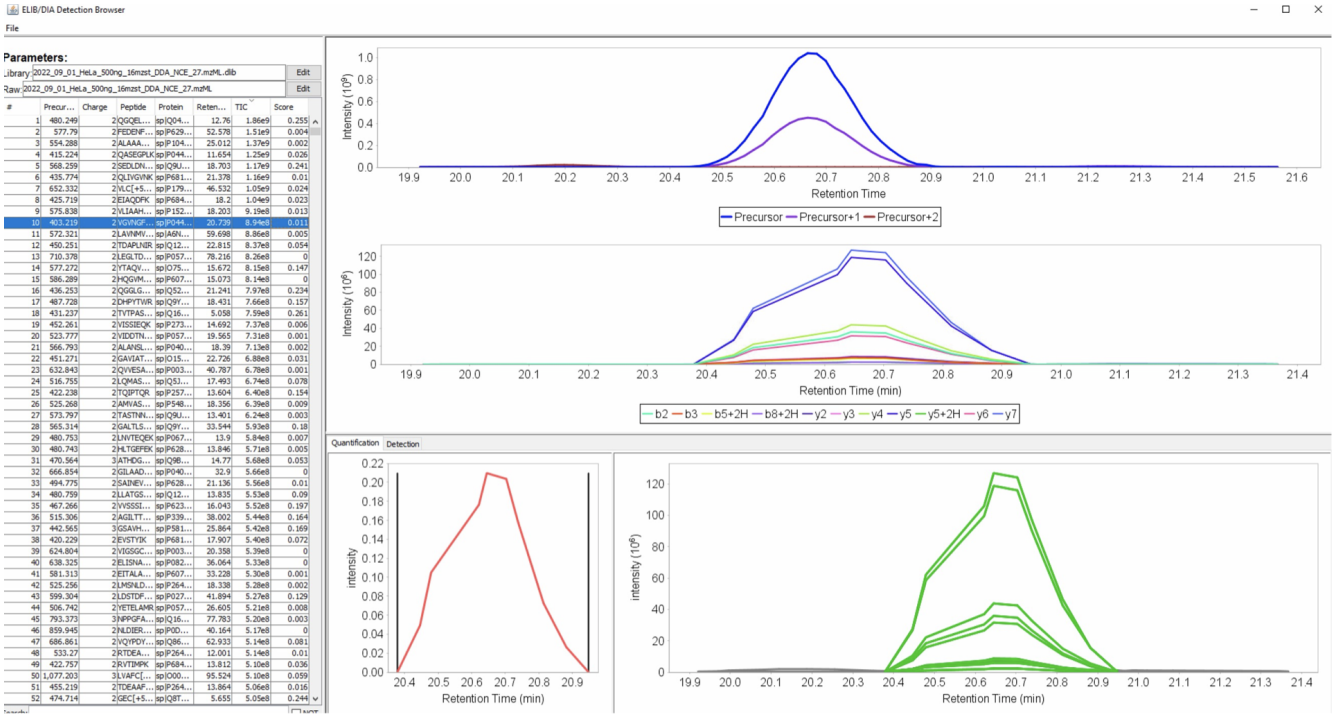


ELIB Browser

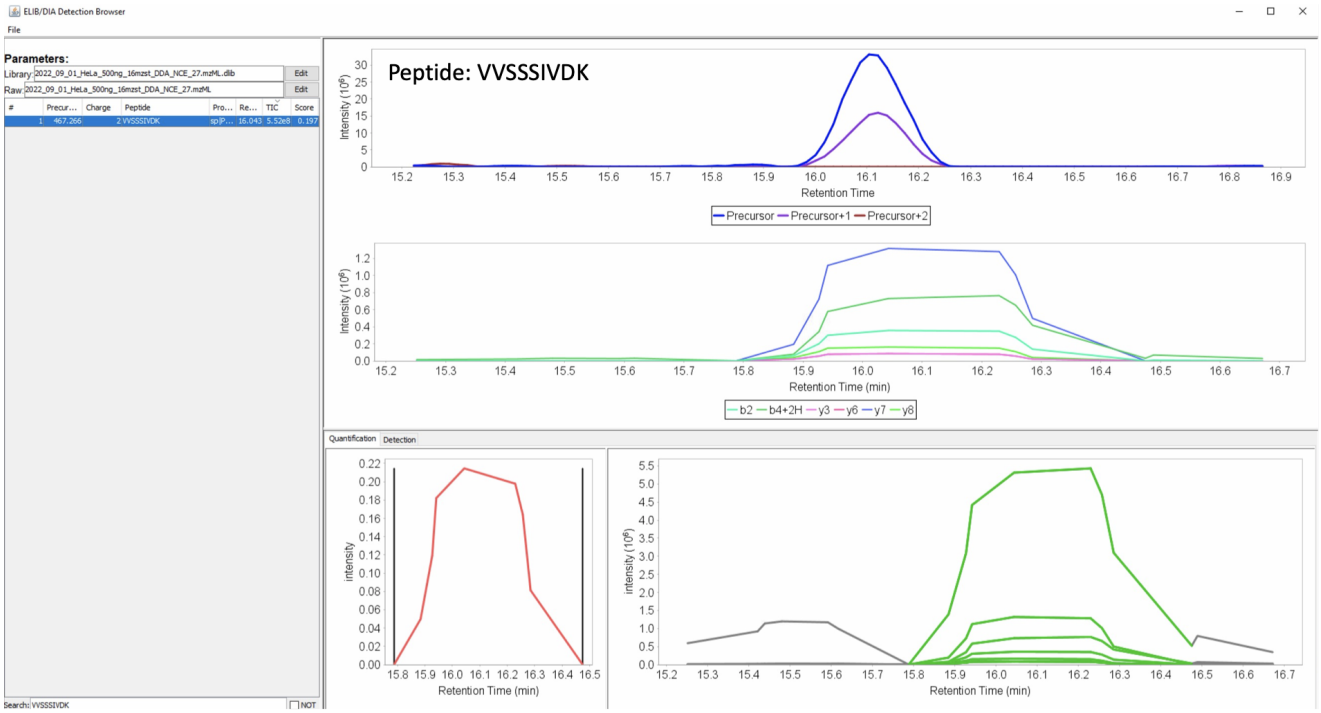
Click on "Launch ELIB Browser" to visualize a library relative to an individual sample. Click "edit" to select the library file you would like. Then do the same for the RAW file, in .mzML format. Note: EncyclopeDIA cannot read files in RAW form.



Once the files are loaded, peptides will populate. The screen will pull up this view.

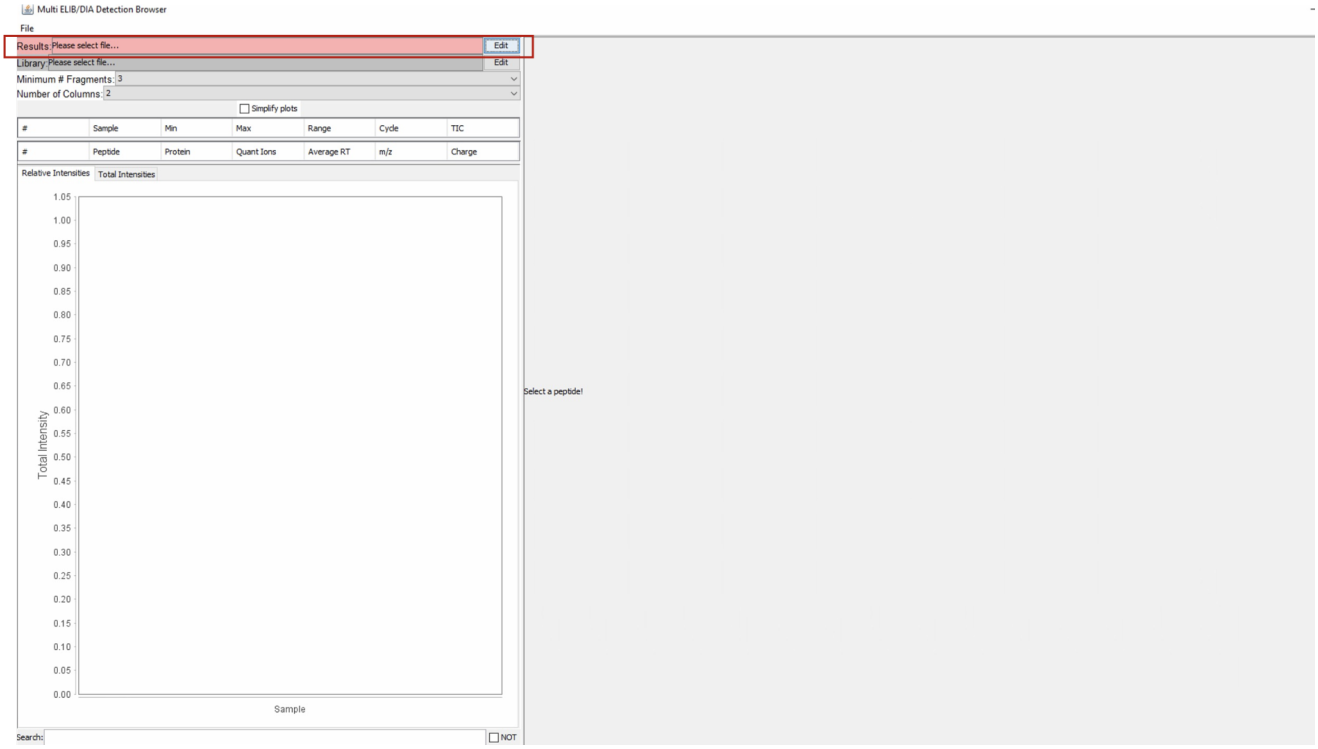


At the bottom, use the search bar to search for specific proteins or peptides. You can see in this example, the precursor spectra for the peptide “VVSSSIVDK” is symmetrical. For the MS2 spectra, the b2, b4+2H, y3, y6 y7 and y8 peptides were detected without interference. Visualization is presented as chromatograms, showing the relationship between ions in multiple spectral acquisitions of the same peptide.

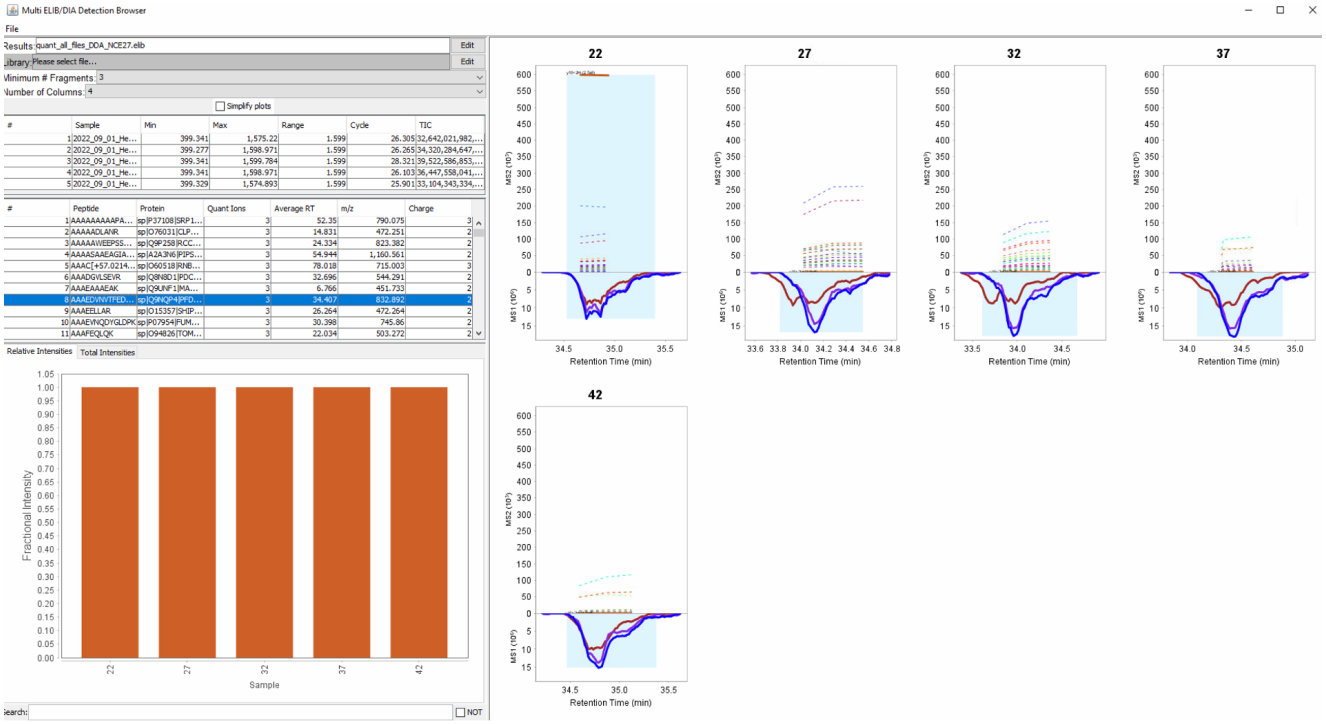


Multi-ELIB Browser

Click on “Launch Multi-ELIB Browser” to visualize data across all raw files in a quantitative result. Using this view, you can visualize peptides present in multiple samples.

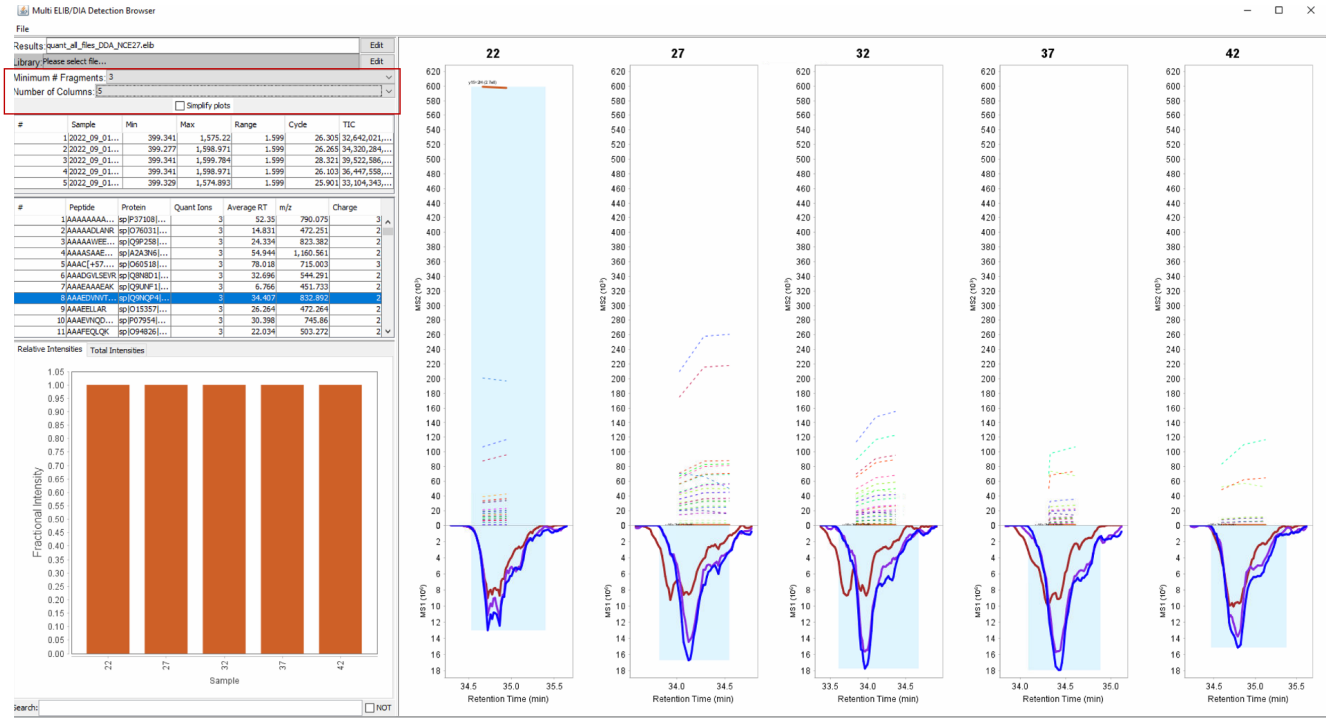


Click “edit” to upload the results file you are interested in viewing. You can insert the library file as well, although it is not necessary to visualize samples. Once uploaded, the view should look like this:

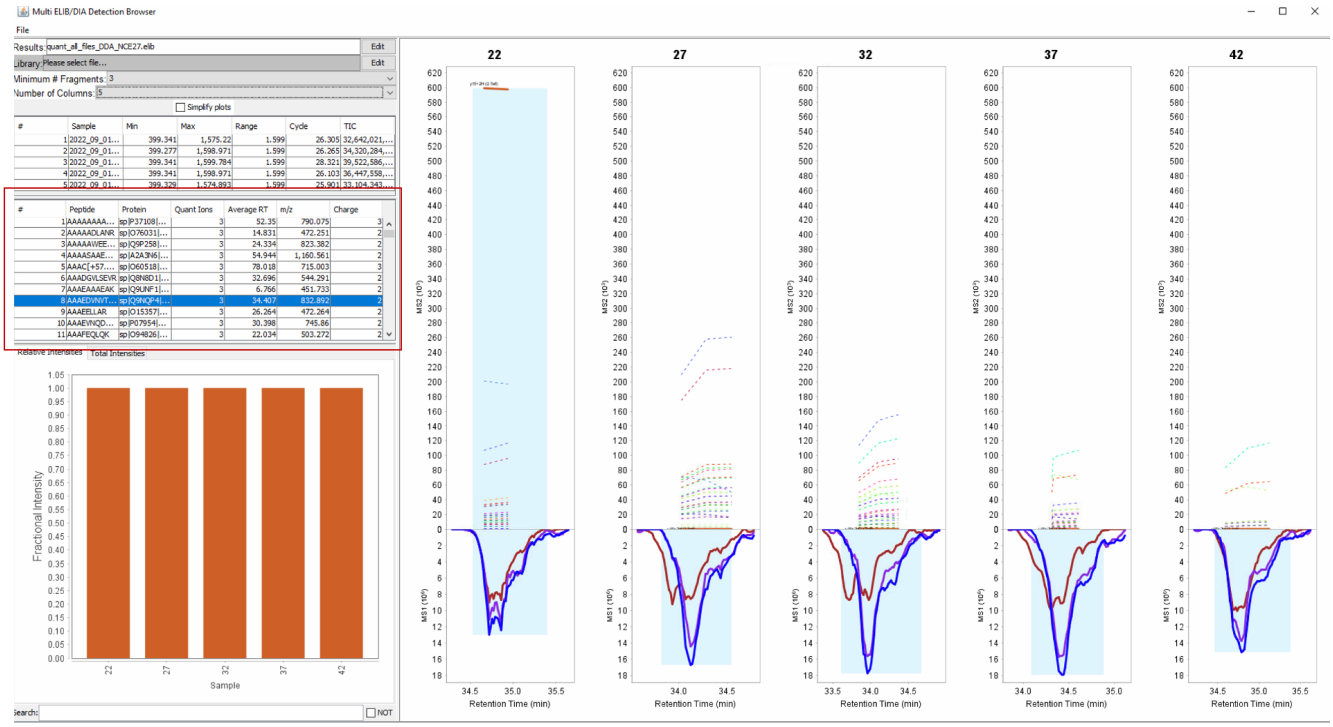


The minimum number of fragments per peptide can be set to limit the peptides that need to be examined. The number of columns can also be changed. The sample order will be in either numerical

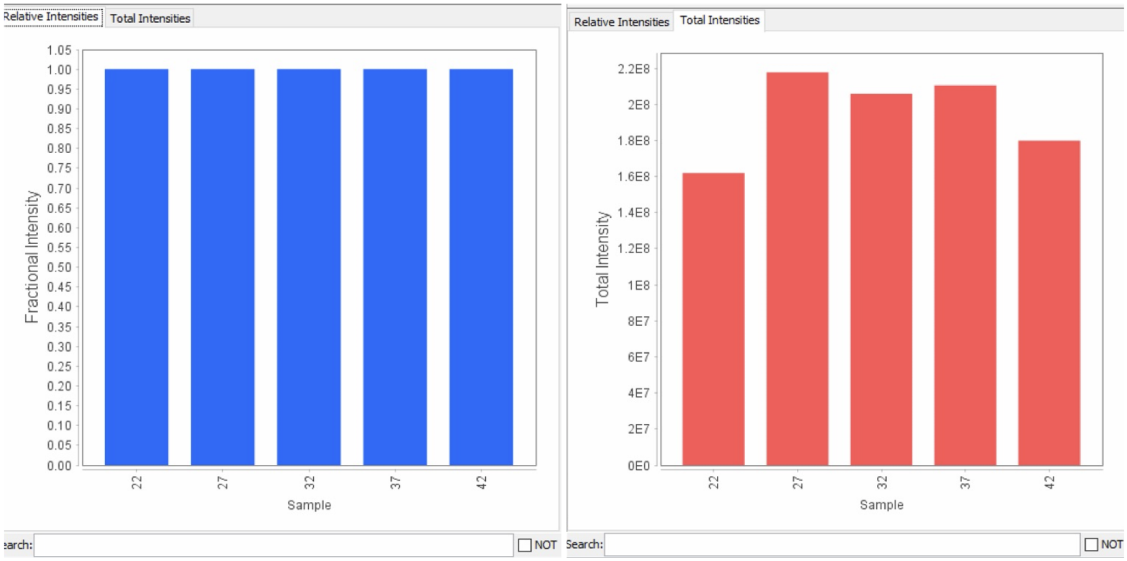
or alphabetical order. Visualization is presented as chromatograms that show the relationship between ions in multiple spectral acquisitions of the same peptide.



Similar to the ELIB browser, you can search for specific peptides, or click through the list of detections:

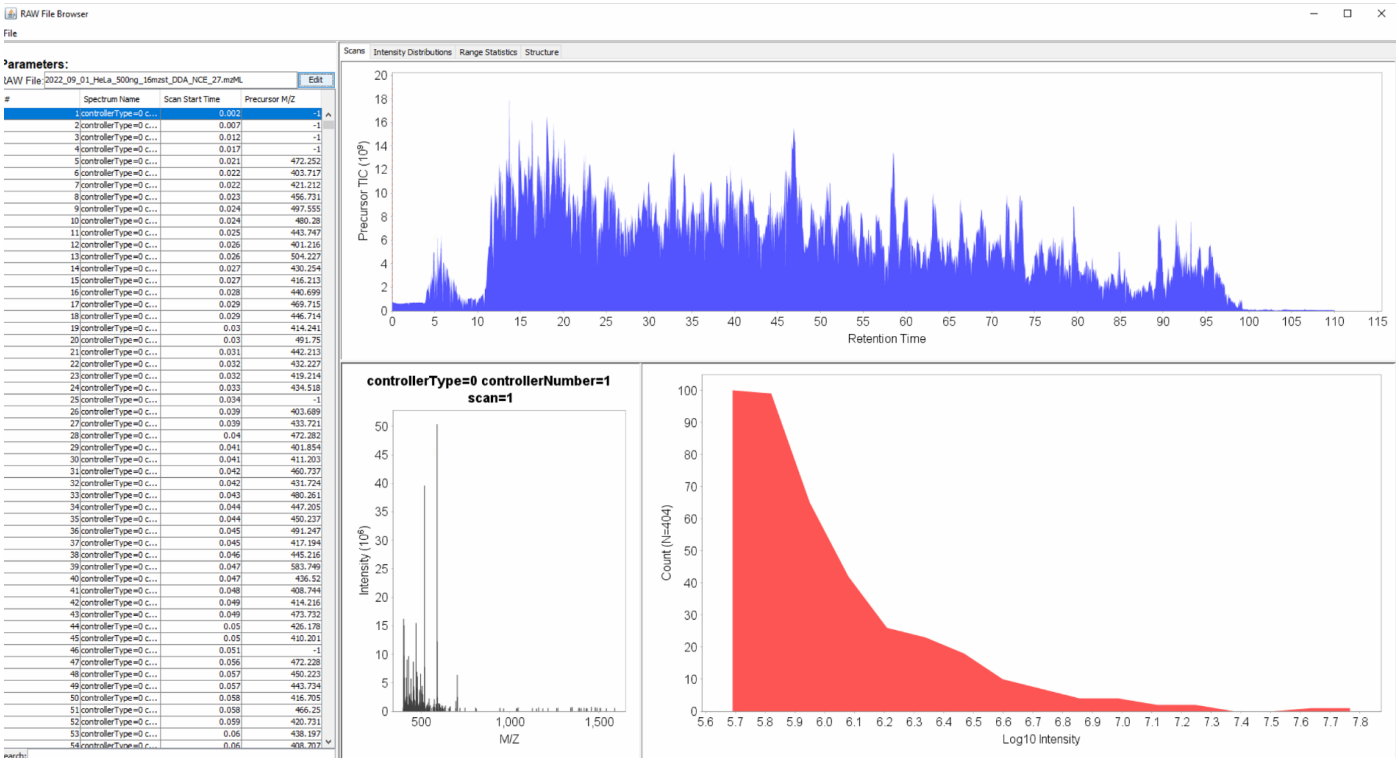


In the lower left panel, there are intensity bar graphs for “Relative Intensities” and “Total Intensities” across samples. Scribe uses precursors for quantitation, therefore the “Relative Intensities” bar graph will always consist of the intensity values for a single precursor across all samples. “Total Intensities” will allow you to compare quantitative intensities across samples. Below are both graphs for the peptide selected in the screenshot above.

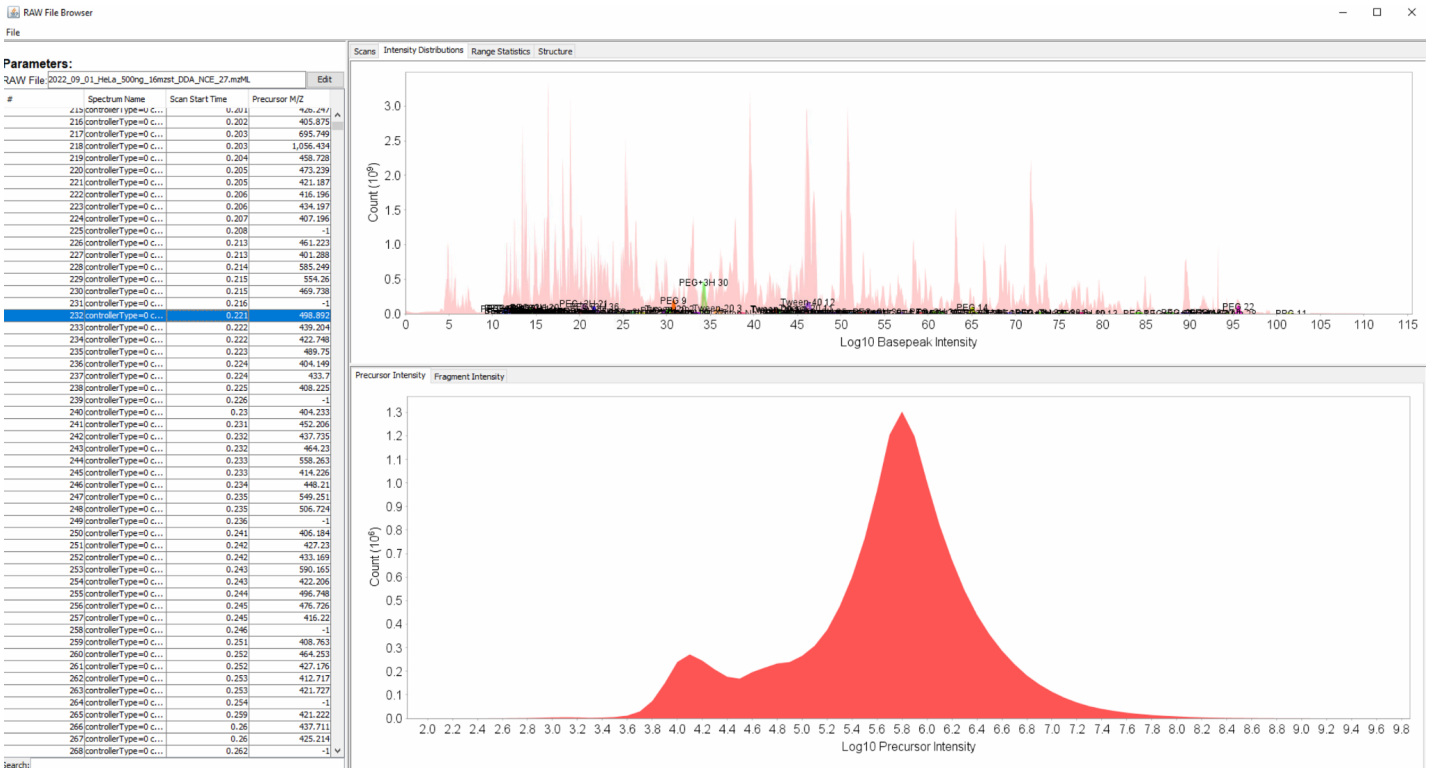


Raw File Browser

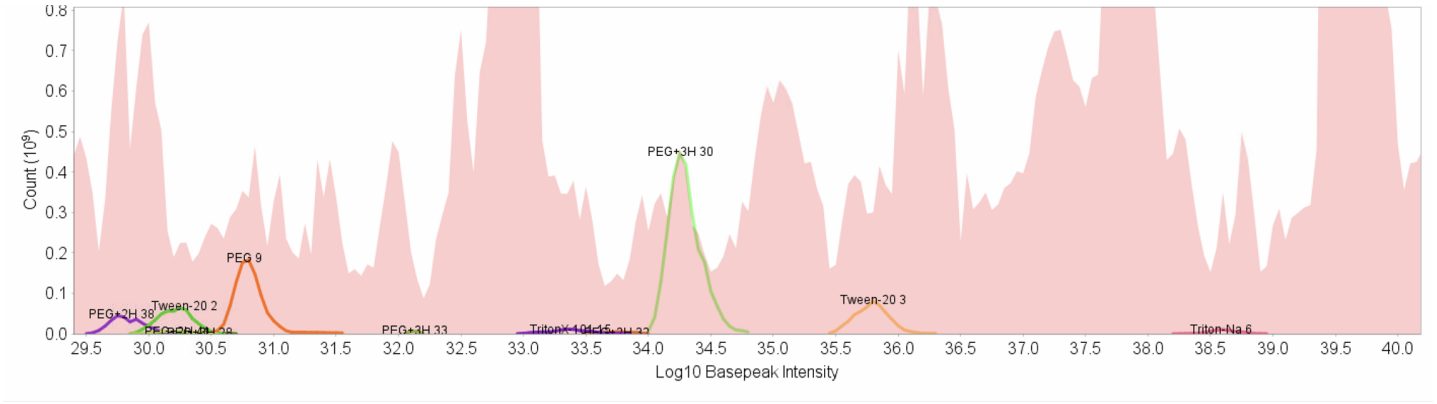
“Launch raw File Browser” will allow you to visualize samples, in mzML format. In this view, you can visualize unsearched data. The data is present in scans, allowing you to visualize the MS1 and MS2 spectra for specific scans throughout your injection



The next tab contains intensity distributions for precursor and fragment ions present in each scan in the bottom pane. In the top pain, you can see the base peak chromatogram trace.

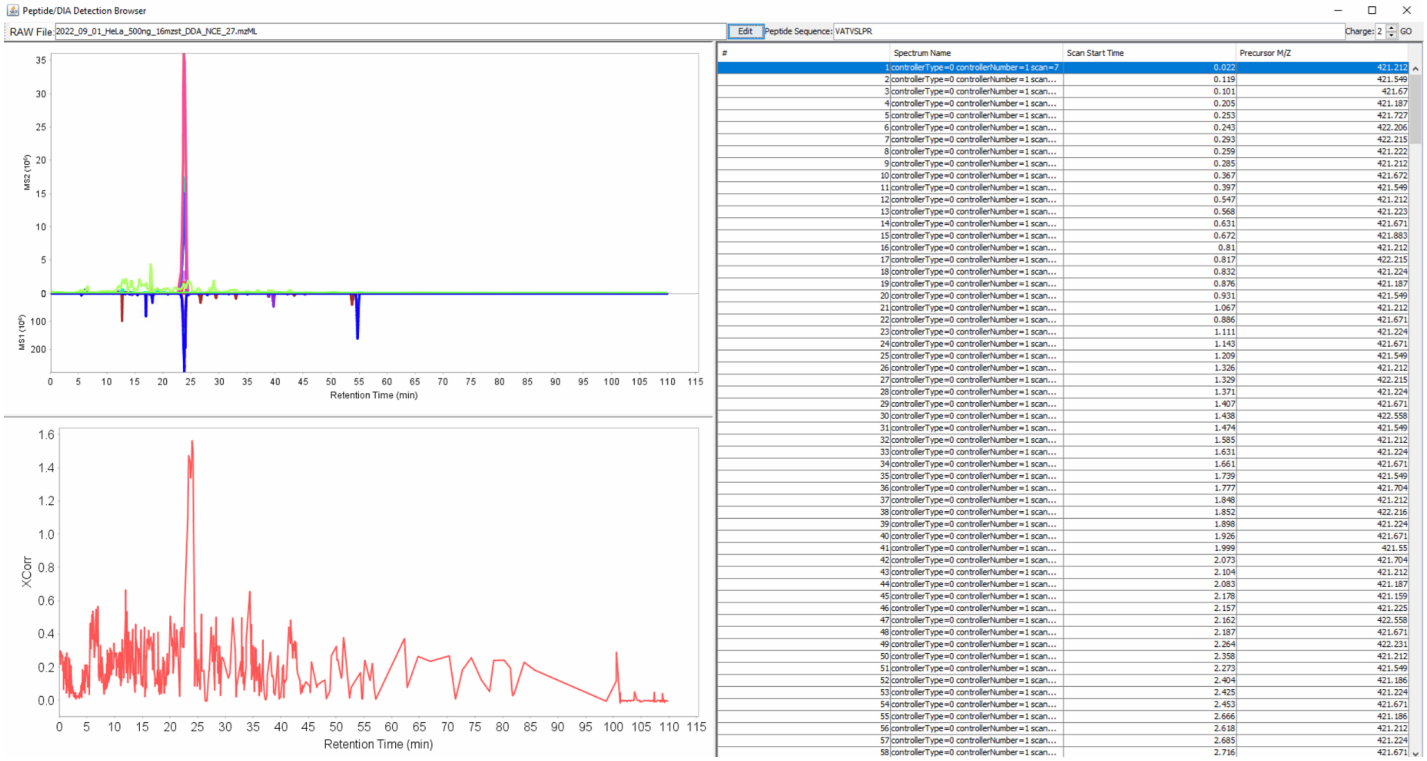


Contaminant peaks from common molecules, such as PEG, triton, tween and polysiloxane are present in the base peak trace. If you look in the screenshot below, you will see PEG +3H 30 base peak present in the chromatogram around 34.5 minutes.

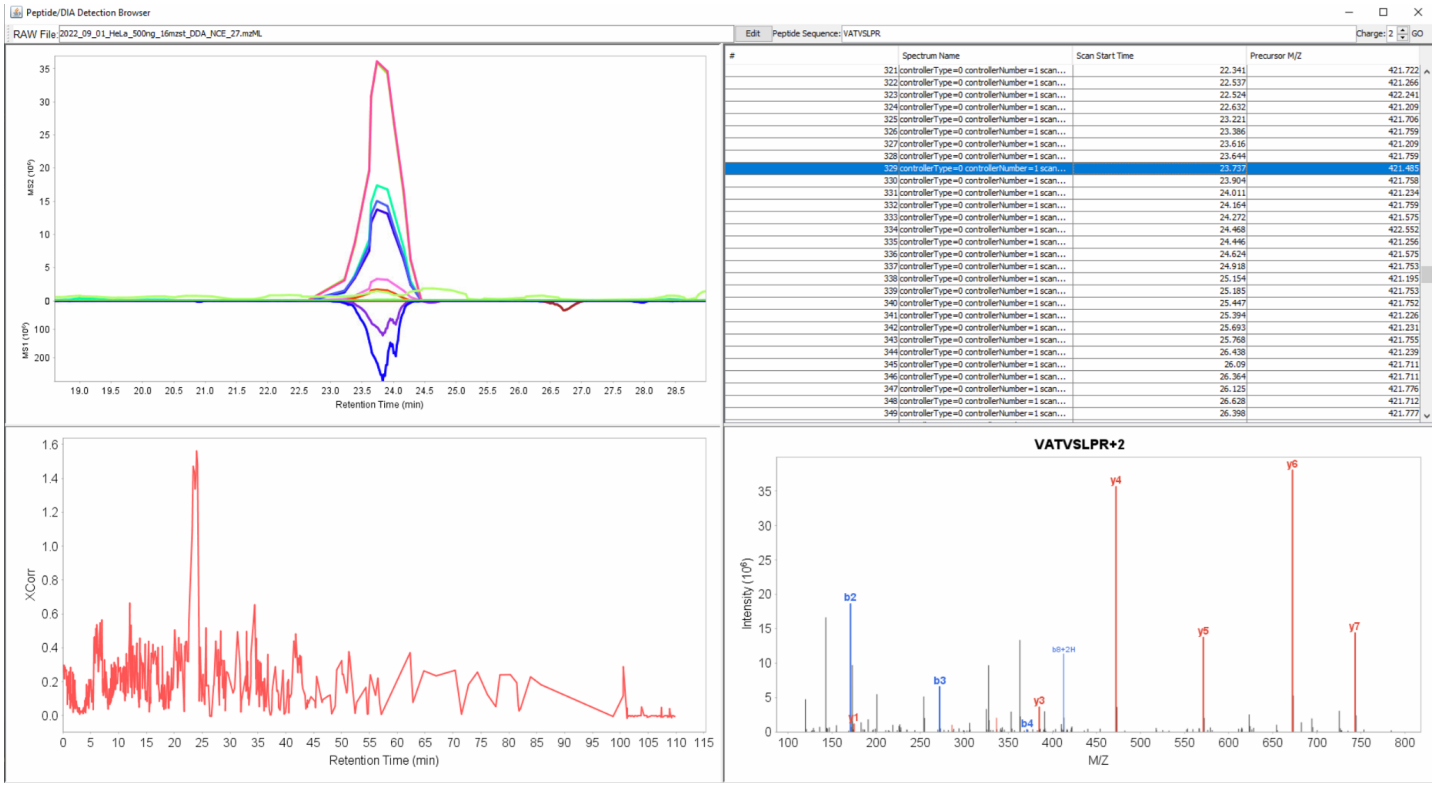


Peptide Browser

“Launch Peptide Browser” will allow you to search for a specific peptide in a sample file of interest. When you first open this browser, the peptide “VATVSLPR”, a common tryptic peptide, is selected. The top left panel displays the chromatogram trace for this peptide across the entire gradient. The bottom left panel shows the XCorr score across the gradient. You should see the highest score where the most intense peak is for the peptide. On the right side of the screen are the scans throughout the gradient. Visualization is presented as chromatograms that show the relationship between ions in multiple spectral acquisitions of the same peptide.

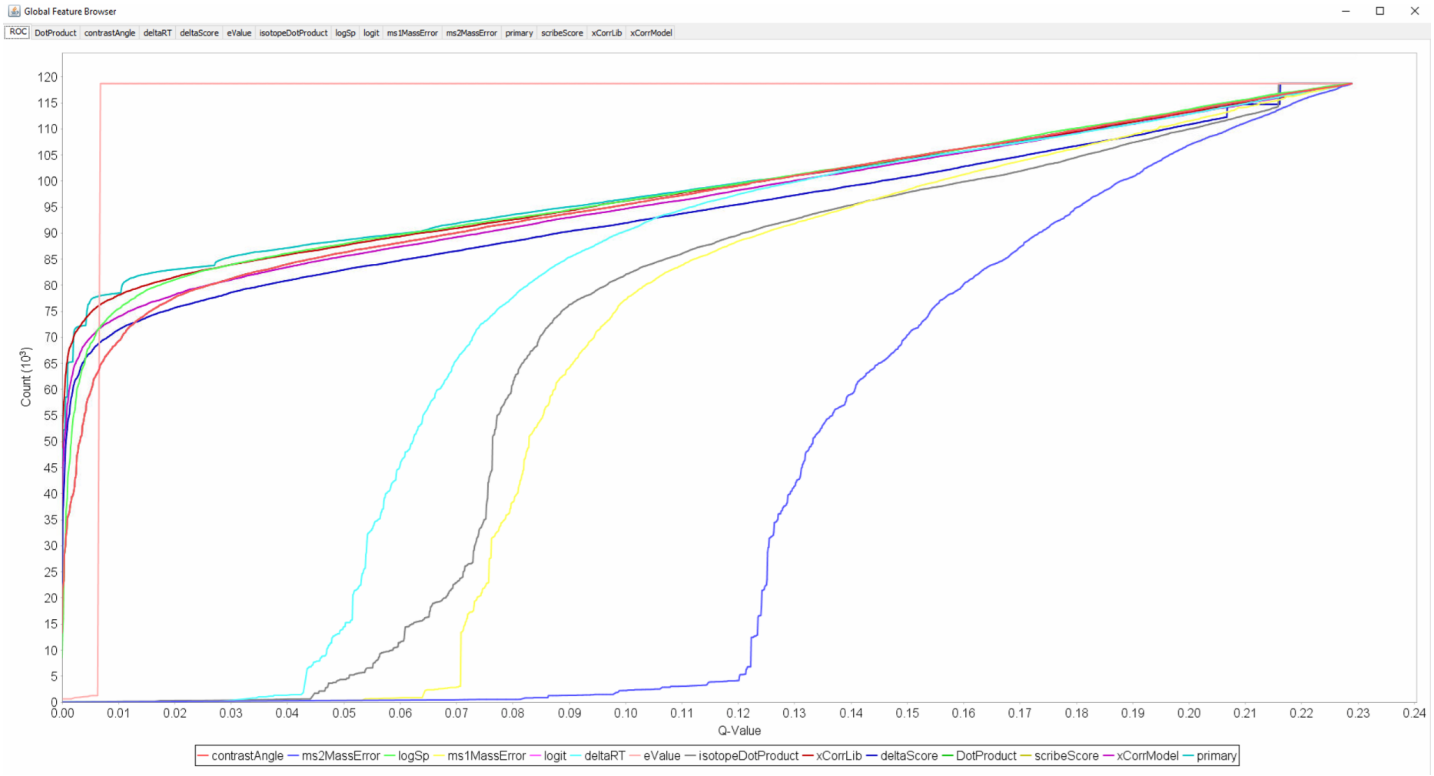


In this next view, the base peak has been zoomed in under the chromatogram panel in the top left. You can see this peak elutes around 23.8 min. A scan acquired at 23.737 min has been selected to pull up the MS2 spectra in the bottom-right corner.

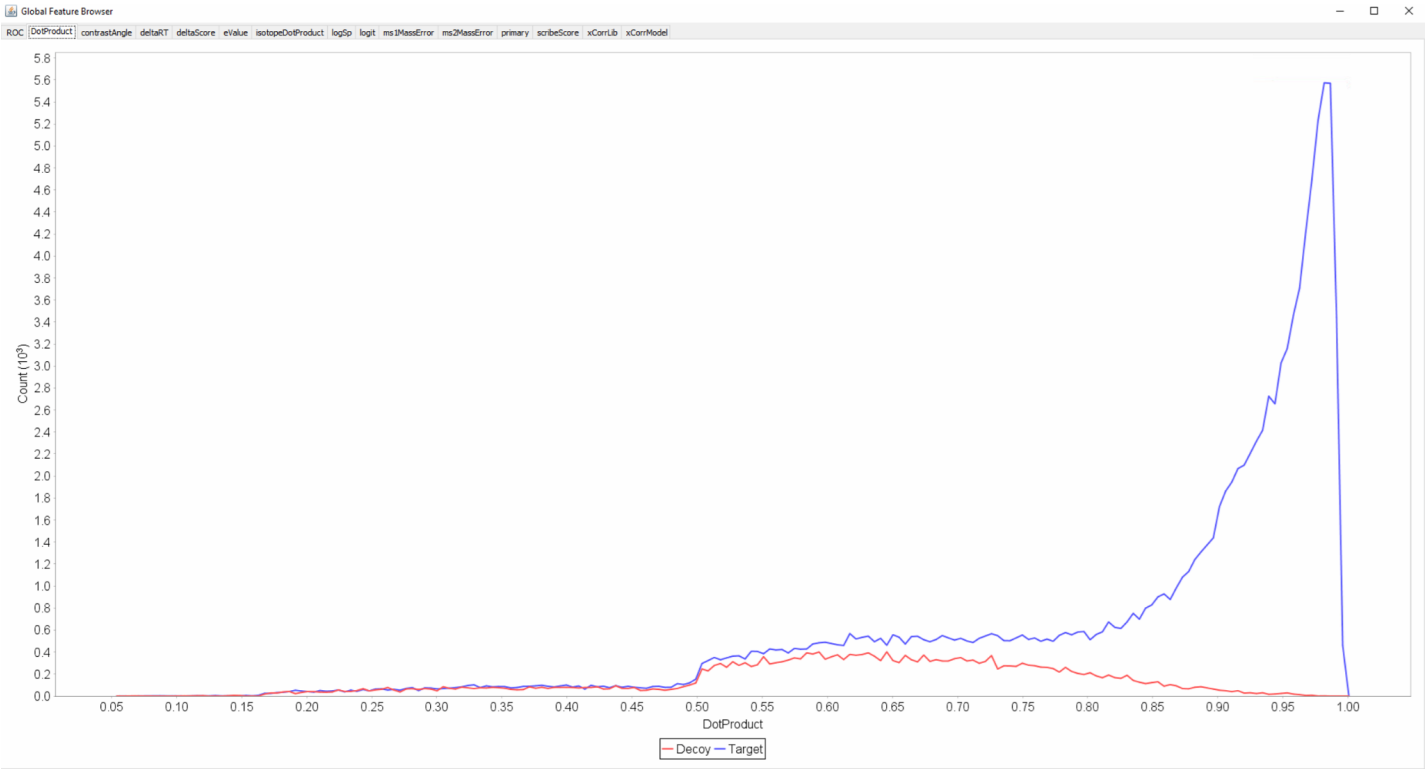


Feature Browser

“Launch feature browser” will allow you to visualize the features.txt files output from Scribe or EncyclopeDIA. The tabs of the features browser contain various metrics used in matching predicted spectra to peptides. Below is a screenshot of the ROC plot, which shows the performance of the 13 features available through this browser. On the x-axis is the Q-value, or the chances of a given feature resulting in a false positive rate (FPR). On the y-axis are counts, indicating the number of peptides detected at a given FPR. Features that are lower on the curve have worse performance than features higher on the curve. For example, “ms2MassError,” denoted by the dark blue line, performs the worse of the given characteristics and is only useful as a correction factor in this case.



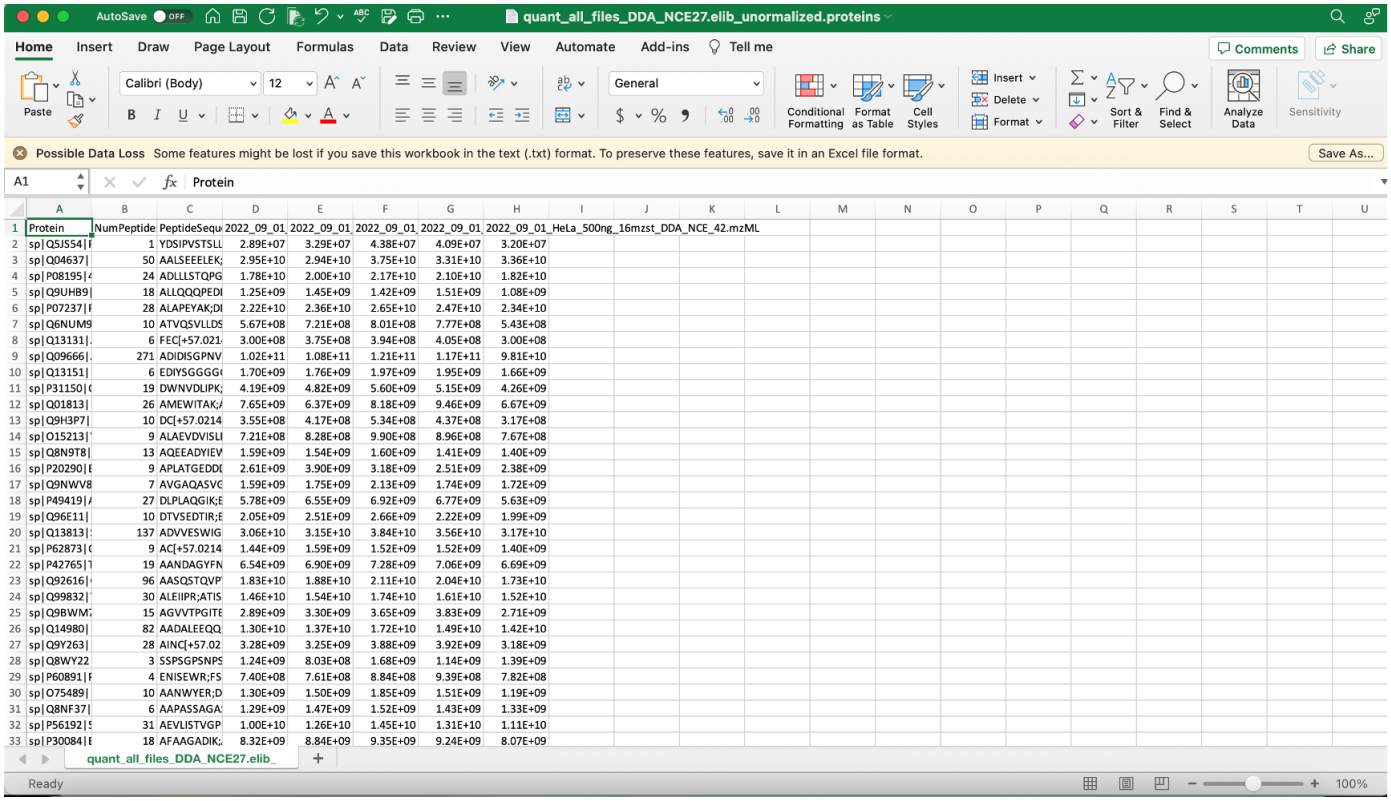
Each of the other tabs contains histograms of the individual scores for target and decoy peptides. For example, below is the dot product plot where a value closer to 1 indicates greater similarity between the actual and predicted spectra. In this plot, you should see the decoy trace (in red) reside lower than the target trace (in blue).



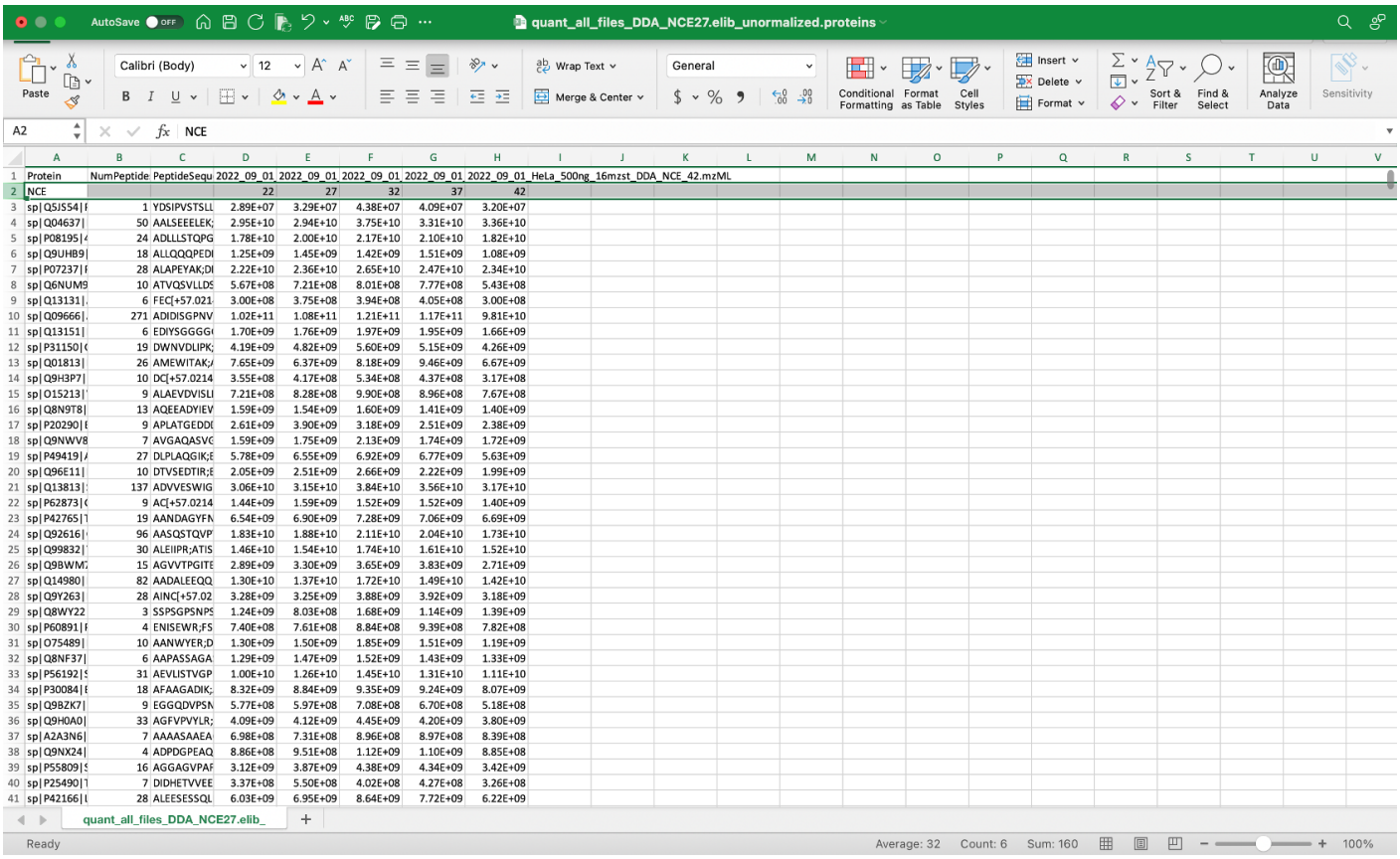
For the remaining features, you should also see the target trace (blue) relative to the decoy trace (red). This indicates the target spectra are performing better than the decoy spectra, which should occur when a peptide match is matched correctly.

C. Analyzing text files using Microsoft Excel

After saving a quant report, Scribe will perform a match-between-runs quantitation of all samples you obtained and then write a protein and peptide report. These are formatted as expression datasets of the proteins or peptides detected. This expression dataset allows you to use Open the text file in Excel by right-clicking on the file, open with excel. The following example shows how to process the protein reports for HeLa DDA injections at varying NCE processed against a library optimized against NCE = 27. Within this example, the file was “saved as” .XLS file to ensure that the equations are saved. Below is a screenshot of the file once it has been opened.

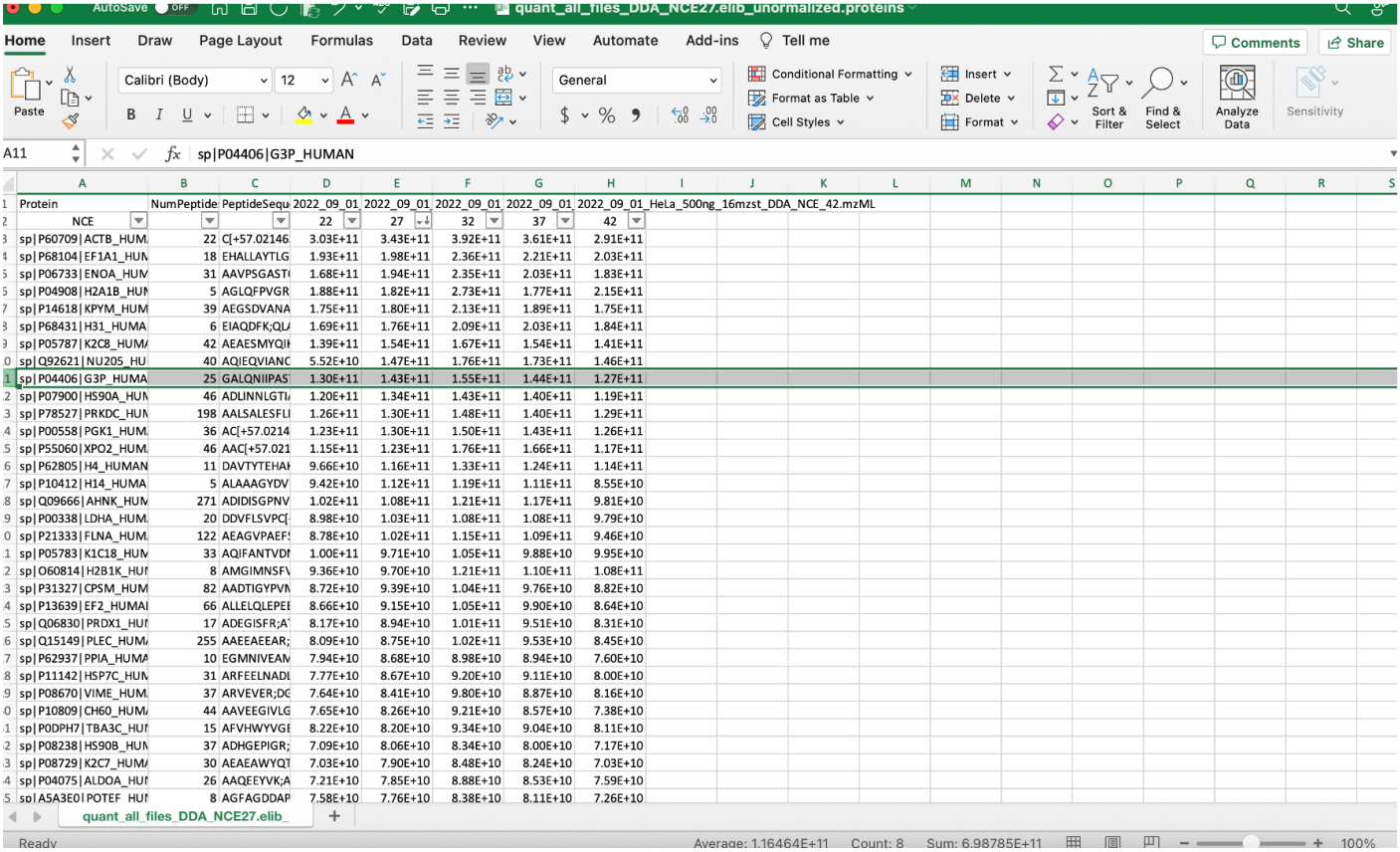


Add column annotations to describe the sample names/conditions for each injection. In this example, NCE was added as an annotation row.



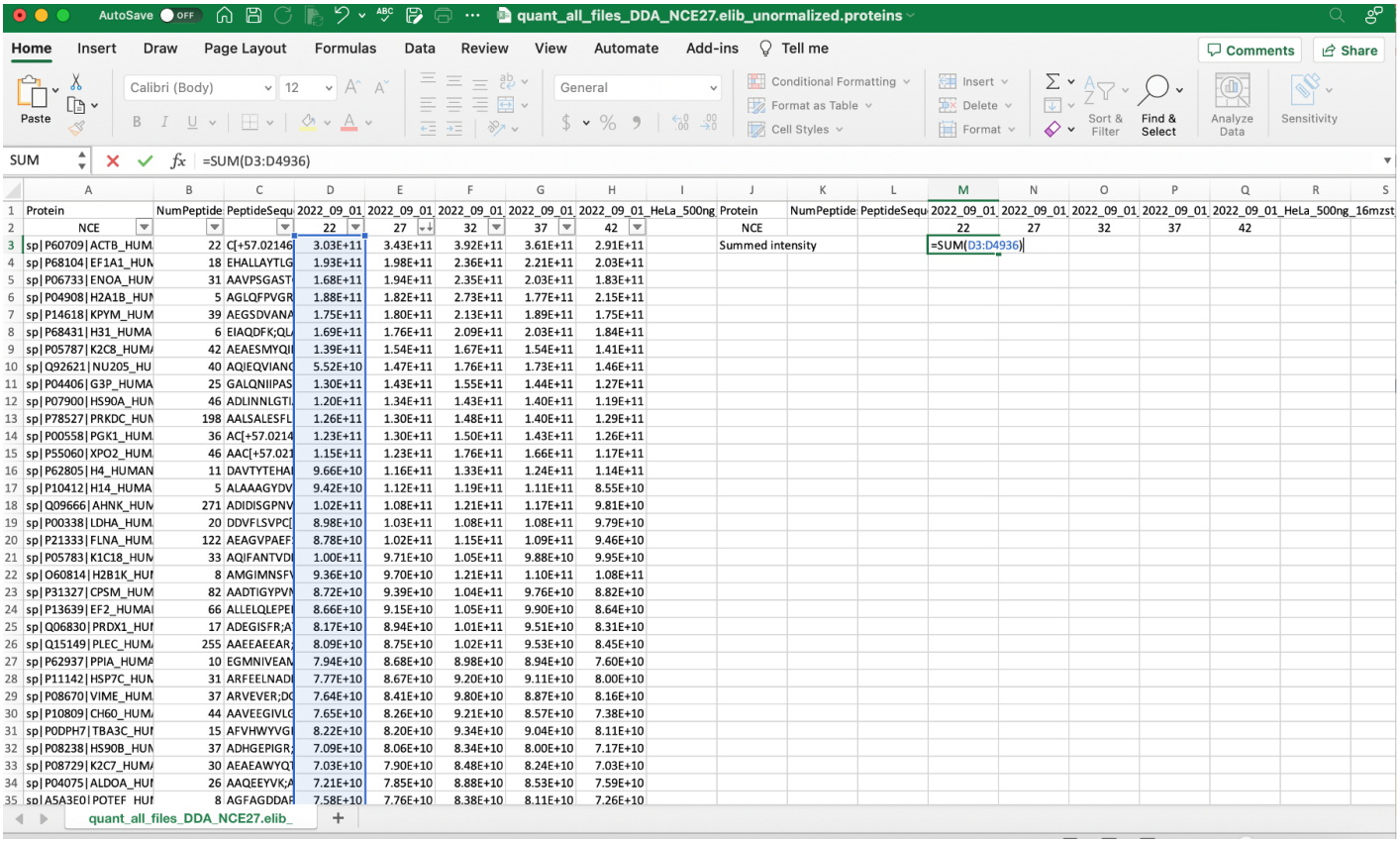
Protein	NumPeptide	PeptideSequ	2022_09_01	2022_09_01	2022_09_01	2022_09_01	2022_09_01_HeLa_500ng_16mzst_DDA_NCE_42.mzML
NCE			22	27	32	37	42
sp|Q5JS54|F	1	YDSIPVSTSLL	2.89E+07	3.29E+07	4.38E+07	4.09E+07	3.20E+07
sp|Q04637|	50	AALSEEELEK;	2.95E+10	2.94E+10	3.75E+10	3.31E+10	3.36E+10
sp|P08195|4	24	ADLLLSTQPG	1.78E+10	2.00E+10	2.17E+10	2.10E+10	1.82E+10
sp|Q9UHB9|	18	ALLQQQPEDI	1.25E+09	1.45E+09	1.42E+09	1.51E+09	1.08E+09
sp|P07237|F	28	ALAPEYAK;DI	2.22E+10	2.36E+10	2.65E+10	2.47E+10	2.34E+10
sp|Q6NUM9	10	ATVQSVLLDS	5.67E+08	7.21E+08	8.01E+08	7.77E+08	5.43E+08
sp|Q13131|.	6	FEC[+57.021	3.00E+08	3.75E+08	3.94E+08	4.05E+08	3.00E+08
sp|Q09666|.	271	ADIDISGPNV	1.02E+11	1.08E+11	1.21E+11	1.17E+11	9.81E+10
sp|Q13151|	6	EDIYSGGGGG	1.70E+09	1.76E+09	1.97E+09	1.95E+09	1.66E+09
sp|P31150|G	19	DWNVDLIPK;	4.19E+09	4.82E+09	5.60E+09	5.15E+09	4.26E+09
sp|Q01813|	26	AMEWITAK;A	7.65E+09	6.37E+09	8.18E+09	9.46E+09	6.67E+09
sp|Q9H3P7|	10	DC[+57.0214	3.55E+08	4.17E+08	5.34E+08	4.37E+08	3.17E+08
sp|O15213|'	9	ALAEVDVISLI	7.21E+08	8.28E+08	9.90E+08	8.96E+08	7.67E+08
sp|Q8N9T8|	13	AQEEADYIEV	1.59E+09	1.54E+09	1.60E+09	1.41E+09	1.40E+09
sp|P20290|E	9	APLATGEDDI	2.61E+09	3.90E+09	3.18E+09	2.51E+09	2.38E+09
sp|Q9NWV8	7	AVGAQASVC	1.59E+09	1.75E+09	2.13E+09	1.74E+09	1.72E+09
sp|P49419|A	27	DLPLAQGIK;E	5.78E+09	6.55E+09	6.92E+09	6.77E+09	5.63E+09
sp|Q96E11|	10	DTVSEDTIR;E	2.05E+09	2.51E+09	2.66E+09	2.22E+09	1.99E+09
sp|Q13813|S	137	ADVVESWIG	3.06E+10	3.15E+10	3.84E+10	3.56E+10	3.17E+10
sp|P62873|(	9	AC[+57.0214	1.44E+09	1.59E+09	1.52E+09	1.52E+09	1.40E+09
sp|P42765|T	19	AANDAGYFN	6.54E+09	6.90E+09	7.28E+09	7.06E+09	6.69E+09
sp|Q92616|(	96	AASQSTQVP'	1.83E+10	1.88E+10	2.11E+10	2.04E+10	1.73E+10
sp|Q99832|T	30	ALEIIPR;ATIS	1.46E+10	1.54E+10	1.74E+10	1.61E+10	1.52E+10
sp|Q9BWM7	15	AGVVTPGITE	2.89E+09	3.30E+09	3.65E+09	3.83E+09	2.71E+09
sp|Q14980|	82	AADALEEQQ	1.30E+10	1.37E+10	1.72E+10	1.49E+10	1.42E+10
sp|Q9Y263|	28	AINC[+57.02	3.28E+09	3.25E+09	3.88E+09	3.92E+09	3.18E+09
sp|Q8WY22	3	SSPSGPSNPS	1.24E+09	8.03E+08	1.68E+09	1.14E+09	1.39E+09
sp|P60891|F	4	ENISEWR;FS	7.40E+08	7.61E+08	8.84E+08	9.39E+08	7.82E+08
sp|O75489|	10	AANWYER;D	1.30E+09	1.50E+09	1.85E+09	1.51E+09	1.19E+09
sp|Q8NF37|	6	AAPASSAGA	1.29E+09	1.47E+09	1.52E+09	1.43E+09	1.33E+09
sp|P56192|S	31	AEVLISTVGP	1.00E+10	1.26E+10	1.45E+10	1.31E+10	1.11E+10
sp|P30084|E	18	AFAAGADIK;	8.32E+09	8.84E+09	9.35E+09	9.24E+09	8.07E+09
sp|Q9BZK7|	9	EGGQDVPSN	5.77E+08	5.97E+08	7.08E+08	6.70E+08	5.18E+08
sp|Q9H0A0|	33	AGFVPVYLR;	4.09E+09	4.12E+09	4.45E+09	4.20E+09	3.80E+09
sp|A2A3N6|	7	AAAASAAEA	6.98E+08	7.31E+08	8.96E+08	8.97E+08	8.39E+08
sp|Q9NX24|	4	ADPDGPEAQ	8.86E+08	9.51E+08	1.12E+09	1.10E+09	8.85E+08
sp|P55809|S	16	AGGAGVPAF	3.12E+09	3.87E+09	4.38E+09	4.34E+09	3.42E+09
sp|P25490|T	7	DIDHETVVEE	3.37E+08	5.50E+08	4.02E+08	4.27E+08	3.26E+08
sp|P42166|L	28	ALEESESSQL	6.03E+09	6.95E+09	8.64E+09	7.72E+09	6.22E+09

In excel, sort values in increasing order. For example, in HeLa, a mammalian immortalized cancer cell line, the most abundant proteins are often metabolic enzymes (G3P) or filament proteins (ACTB).



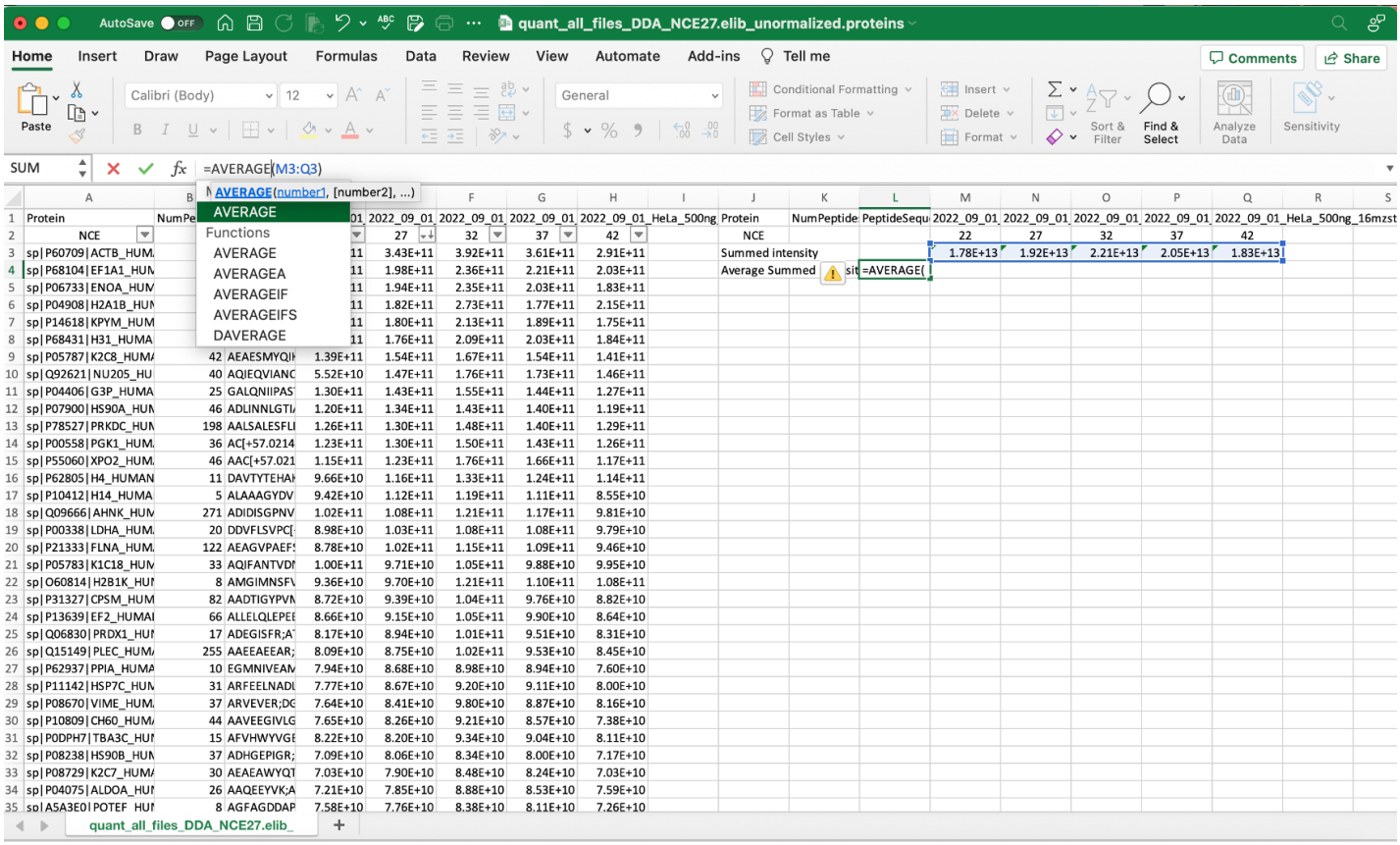
	A	B	C	D	E	F	G	H
1	Protein	NumPeptide	PeptideSequ	2022_09_01	2022_09_01	2022_09_01	2022_09_01	2022_09_01_HeLa_500ng_16mzst_DDA_NCE_42.mzML
2	NCE			22	27	32	37	42
3	sp|P60709|ACTB_HUM	22	C[+57.02146	3.03E+11	3.43E+11	3.92E+11	3.61E+11	2.91E+11
4	sp|P68104|EF1A1_HUN	18	EHALLAYTLG	1.93E+11	1.98E+11	2.36E+11	2.21E+11	2.03E+11
5	sp|P06733|ENOA_HUM	31	AAVPSGAST(	1.68E+11	1.94E+11	2.35E+11	2.03E+11	1.83E+11
6	sp|P04908|H2A1B_HUN	5	AGLQFPVGR	1.88E+11	1.82E+11	2.73E+11	1.77E+11	2.15E+11
7	sp|P14618|KPYM_HUM	39	AEGSDVANA	1.75E+11	1.80E+11	2.13E+11	1.89E+11	1.75E+11
8	sp|P68431|H31_HUMA	6	EIAQDFK;QL/	1.69E+11	1.76E+11	2.09E+11	2.03E+11	1.84E+11
9	sp|P05787|K2C8_HUM/	42	AEAESMYQIH	1.39E+11	1.54E+11	1.67E+11	1.54E+11	1.41E+11
10	sp|Q92621|NU205_HU	40	AQIEQVIANC	5.52E+10	1.47E+11	1.76E+11	1.73E+11	1.46E+11
11	sp|P04406|G3P_HUMA	25	GALQNIIPAS	1.30E+11	1.43E+11	1.55E+11	1.44E+11	1.27E+11
12	sp|P07900|HS90A_HUN	46	ADLINNLGTI/	1.20E+11	1.34E+11	1.43E+11	1.40E+11	1.19E+11
13	sp|P78527|PRKDC_HUN	198	AALSALESFLI	1.26E+11	1.30E+11	1.48E+11	1.40E+11	1.29E+11
14	sp|P00558|PGK1_HUM/	36	AC[+57.0214	1.23E+11	1.30E+11	1.50E+11	1.43E+11	1.26E+11
15	sp|P55060|XPO2_HUM/	46	AAC[+57.021	1.15E+11	1.23E+11	1.76E+11	1.66E+11	1.17E+11
16	sp|P62805|H4_HUMAN	11	DAVTYTEHAI	9.66E+10	1.16E+11	1.33E+11	1.24E+11	1.14E+11
17	sp|P10412|H14_HUMA	5	ALAAAGYDV	9.42E+10	1.12E+11	1.19E+11	1.11E+11	8.55E+10
18	sp|Q09666|AHNK_HUM	271	ADIDISGPNV	1.02E+11	1.08E+11	1.21E+11	1.17E+11	9.81E+10
19	sp|P00338|LDHA_HUM/	20	DDVFLSVPC[	8.98E+10	1.03E+11	1.08E+11	1.08E+11	9.79E+10
20	sp|P21333|FLNA_HUM/	122	AEAGVPAEF!	8.78E+10	1.02E+11	1.15E+11	1.09E+11	9.46E+10
21	sp|P05783|K1C18_HUN	33	AQIFANTVDI	1.00E+11	9.71E+10	1.05E+11	9.88E+10	9.95E+10
22	sp|O60814|H2B1K_HUI	8	AMGIMNSF\	9.36E+10	9.70E+10	1.21E+11	1.10E+11	1.08E+11
23	sp|P31327|CPSM_HUM/	82	AADTIGYPVN	8.72E+10	9.39E+10	1.04E+11	9.76E+10	8.82E+10
24	sp|P13639|EF2_HUMAI	66	ALLELQLEPE	8.66E+10	9.15E+10	1.05E+11	9.90E+10	8.64E+10
25	sp|Q06830|PRDX1_HUI	17	ADEGISFR;A	8.17E+10	8.94E+10	1.01E+11	9.51E+10	8.31E+10
26	sp|Q15149|PLEC_HUM/	255	AAEEAEEAR;	8.09E+10	8.75E+10	1.02E+11	9.53E+10	8.45E+10
27	sp|P62937|PPIA_HUMA	10	EGMNIVEAV	7.94E+10	8.68E+10	8.98E+10	8.94E+10	7.60E+10
28	sp|P11142|HSP7C_HUN	31	ARFEELNADI	7.77E+10	8.67E+10	9.20E+10	9.11E+10	8.00E+10
29	sp|P08670|VIME_HUM/	37	ARVEVER;DC	7.64E+10	8.41E+10	9.80E+10	8.87E+10	8.16E+10
30	sp|P10809|CH60_HUM/	44	AAVEEGIVLG	7.65E+10	8.26E+10	9.21E+10	8.57E+10	7.38E+10
31	sp|P0DPH7|TBA3C_HUI	15	AFVHWYVGE	8.22E+10	8.20E+10	9.34E+10	9.04E+10	8.11E+10
32	sp|P08238|HS90B_HUN	37	ADHGEPIGR;	7.09E+10	8.06E+10	8.34E+10	8.00E+10	7.17E+10
33	sp|P08729|K2C7_HUM/	30	AEAEAWYQT	7.03E+10	7.90E+10	8.48E+10	8.24E+10	7.03E+10
34	sp|P04075|ALDOA_HUI	26	AAQEEYVK;A	7.21E+10	7.85E+10	8.88E+10	8.53E+10	7.59E+10
35	sp|A5A3E0|POTEF_HUI	8	AGFAGDDAP	7.58E+10	7.76E+10	8.38E+10	8.11E+10	7.26E+10

Calculate the sum of all intensities in each sample. Divide each intensity by this sum.

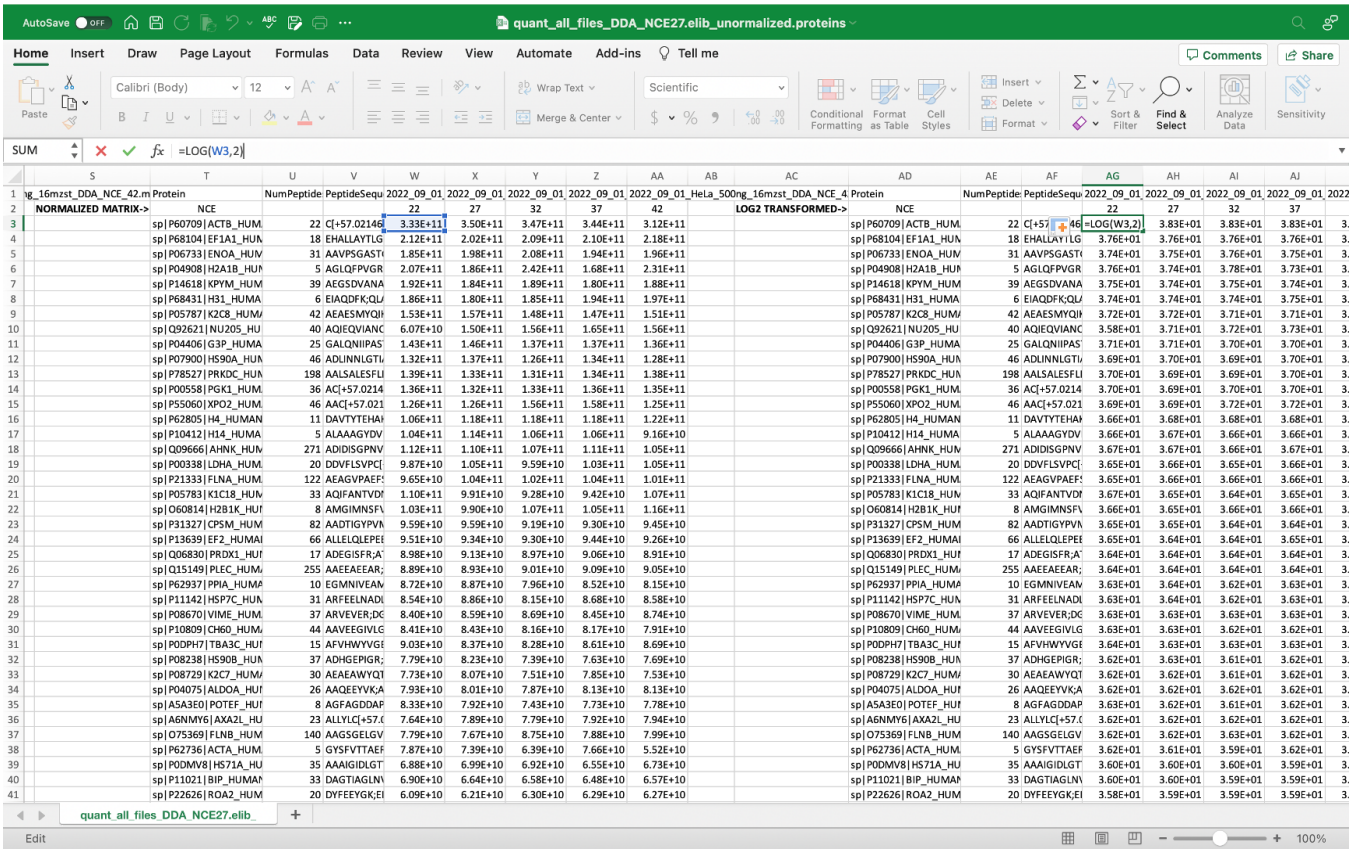


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Protein	NumPeptide	PeptideSequ	2022_09_01	2022_09_01	2022_09_01	2022_09_01	2022_09_01_HeLa_500ng		Protein	NumPeptide	PeptideSequ	2022_09_01	2022_09_01	2022_09_01	2022_09_01	2022_09_01_HeLa_500ng_16mzst		
2	NCE			22	27	32	37	42		NCE			22	27	32	37	42		
3	sp|P60709|ACTB_HUM	22	C[+57.02146	3.03E+11	3.43E+11	3.92E+11	3.61E+11	2.91E+11		Summed intensity			=SUM(D3:D4936)						
4	sp|P68104|EF1A1_HUN	18	EHALLAYTLG	1.93E+11	1.98E+11	2.36E+11	2.21E+11	2.03E+11											
5	sp|P06733|ENOA_HUM	31	AAVPSGAST	1.68E+11	1.94E+11	2.35E+11	2.03E+11	1.83E+11											
6	sp|P04908|H2A1B_HUN	5	AGLQFPVGR	1.88E+11	1.82E+11	2.73E+11	1.77E+11	2.15E+11											
7	sp|P14618|KPYM_HUM	39	AEGSDVANA	1.75E+11	1.80E+11	2.13E+11	1.89E+11	1.75E+11											
8	sp|P68431|H31_HUMA	6	EIAQDFK;QL	1.69E+11	1.76E+11	2.09E+11	2.03E+11	1.84E+11											
9	sp|P05787|K2C8_HUM	42	AEAESMYQIH	1.39E+11	1.54E+11	1.67E+11	1.54E+11	1.41E+11											
10	sp|Q92621|NU205_HU	40	AQIEQVIANC	5.52E+10	1.47E+11	1.76E+11	1.73E+11	1.46E+11											
11	sp|P04406|G3P_HUMA	25	GALQNIIPAS	1.30E+11	1.43E+11	1.55E+11	1.44E+11	1.27E+11											
12	sp|P07900|HS90A_HUN	46	ADLINNLGTI	1.20E+11	1.34E+11	1.43E+11	1.40E+11	1.19E+11											
13	sp|P78527|PRKDC_HUN	198	AALSALESFL	1.26E+11	1.30E+11	1.48E+11	1.40E+11	1.29E+11											
14	sp|P00558|PGK1_HUM	36	AC[+57.0214	1.23E+11	1.30E+11	1.50E+11	1.43E+11	1.26E+11											
15	sp|P55060|XPO2_HUM	46	AAC[+57.021	1.15E+11	1.23E+11	1.76E+11	1.66E+11	1.17E+11											
16	sp|P62805|H4_HUMAN	11	DAVTYTEHA	9.66E+10	1.16E+11	1.33E+11	1.24E+11	1.14E+11											
17	sp|P10412|H14_HUMA	5	ALAAAGYDV	9.42E+10	1.12E+11	1.19E+11	1.11E+11	8.55E+10											
18	sp|Q09666|AHNK_HUM	271	ADIDISGPNV	1.02E+11	1.08E+11	1.21E+11	1.17E+11	9.81E+10											
19	sp|P00338|LDHA_HUM	20	DDVFLSVPC[	8.98E+10	1.03E+11	1.08E+11	1.08E+11	9.79E+10											
20	sp|P21333|FLNA_HUM	122	AEAGVPAEF	8.78E+10	1.02E+11	1.15E+11	1.09E+11	9.46E+10											
21	sp|P05783|K1C18_HUN	33	AQIFANTVDI	1.00E+11	9.71E+10	1.05E+11	9.88E+10	9.95E+10											
22	sp|O60814|H2B1K_HUI	8	AMGIMNSF\	9.36E+10	9.70E+10	1.21E+11	1.10E+11	1.08E+11											
23	sp|P31327|CPSM_HUM	82	AADTIGYPVN	8.72E+10	9.39E+10	1.04E+11	9.76E+10	8.82E+10											
24	sp|P13639|EF2_HUMAI	66	ALLELQLEPE	8.66E+10	9.15E+10	1.05E+11	9.90E+10	8.64E+10											
25	sp|Q06830|PRDX1_HUI	17	ADEGISFR;A	8.17E+10	8.94E+10	1.01E+11	9.51E+10	8.31E+10											
26	sp|Q15149|PLEC_HUM	255	AAEEAEEAR;	8.09E+10	8.75E+10	1.02E+11	9.53E+10	8.45E+10											
27	sp|P62937|PPIA_HUMA	10	EGMNIVEAV	7.94E+10	8.68E+10	8.98E+10	8.94E+10	7.60E+10											
28	sp|P11142|HSP7C_HUN	31	ARFEELNADI	7.77E+10	8.67E+10	9.20E+10	9.11E+10	8.00E+10											
29	sp|P08670|VIME_HUM	37	ARVEVER;DC	7.64E+10	8.41E+10	9.80E+10	8.87E+10	8.16E+10											
30	sp|P10809|CH60_HUM	44	AAVEEGIVLG	7.65E+10	8.26E+10	9.21E+10	8.57E+10	7.38E+10											
31	sp|P0DPH7|TBA3C_HUI	15	AFVHWYVGI	8.22E+10	8.20E+10	9.34E+10	9.04E+10	8.11E+10											
32	sp|P08238|HS90B_HUN	37	ADHGEPIGR;	7.09E+10	8.06E+10	8.34E+10	8.00E+10	7.17E+10											
33	sp|P08729|K2C7_HUM	30	AEAEAWYQT	7.03E+10	7.90E+10	8.48E+10	8.24E+10	7.03E+10											
34	sp|P04075|ALDOA_HUI	26	AAQEEYVK;A	7.21E+10	7.85E+10	8.88E+10	8.53E+10	7.59E+10											
35	sp|A5A3E0|POTEF_HUI	8	AGFAGDDAF	7.58E+10	7.76E+10	8.38E+10	8.11E+10	7.26E+10											

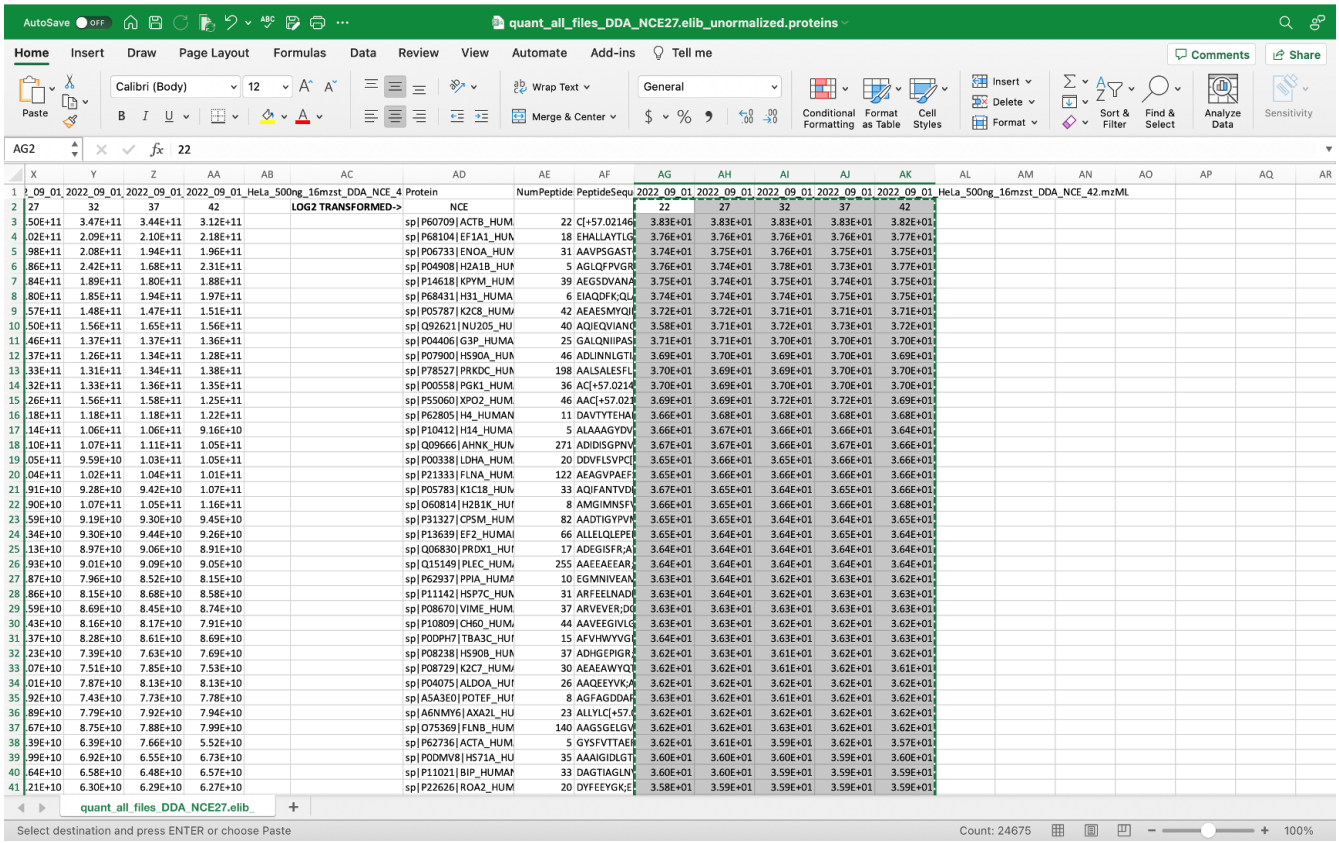
Calculate the average of the summed intensity for all samples (using the values calculated in step 2). Multiply each value by the average summed intensity. Repeat for all values in the data matrices.

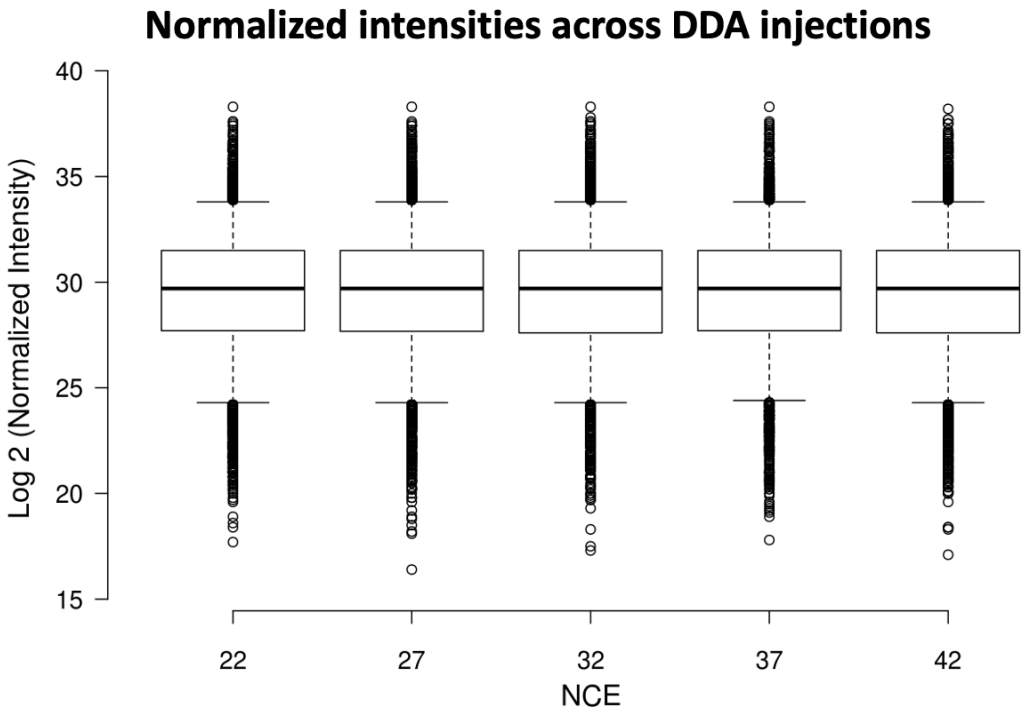


Log$_2$ transform the data matrix.



To visualize relative intensities across injections, you can copy and paste your matrix into a downstream tool, such as BoxplotR: http://shiny.chemgrid.org/boxplotr/




Normalized intensities across DDA injections
Log 2 (Normalized Intensity)
40
35
30
25
20
15
22
27
32
37
42
NCE

7. WALKTHROUGH WITH DEMO DATA

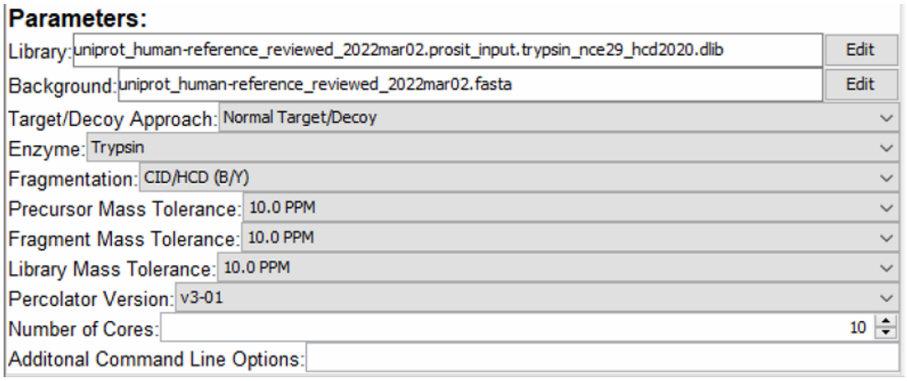
Tutorial for how to process data using Scribe and a Prosit generated library.

A. You can download data used in the Scribe manuscript from:
https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=99b781c9c0b84ac9bfc3f93734e9ddab

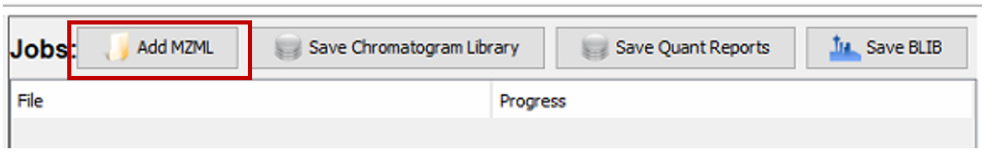
To start, we recommend downloading the following files for this walkthrough:

- 2022_09_01_HeLa_500ng_16mzst_DDA_NCE_27.mzML
- Uniprot_human-reference_reviewed_2022mar02.fasta
- uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce29_hcd2020.dlib

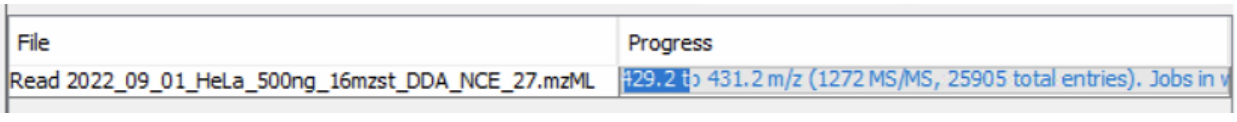
B. Specify [uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce29_hcd2020.dlib] as the library and [uniprot_human-reference_reviewed_2022mar02.fasta] as the background FASTA. The DLIB library was generated using the method explained in part 3 of the manual. We will use a “Normal Target/Decoy” approach, and select “Trypsin” as the enzyme. The fragmentation is set to CID/HCD (B/Y). Precursor mass tolerance, fragment mass tolerance, and library mass tolerance are all set at 10 PPM. The latest Percolator version should be selected, which at this time is v3-01. The settings should match the screenshot below.



C. Then queue up the [2022_09_01_HeLa_500ng_16mzst_DDA_NCE_27.mzML] mzML file by clicking the “Add MZML” button.”

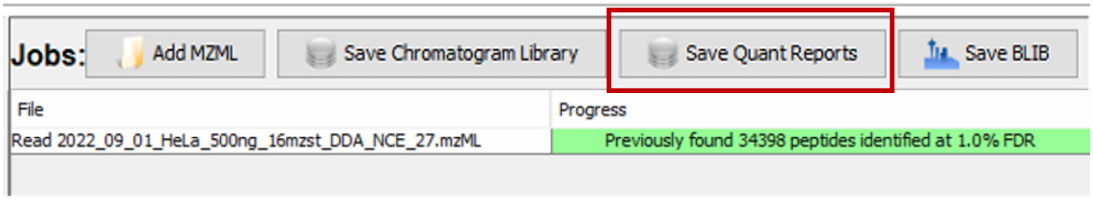


Processing should begin immediately. Detailed analysis information will be printed in the console portion of the screen, and a blue bar will move across as the job is processed:

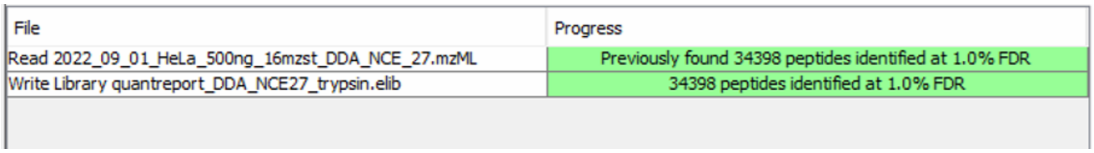


Scribe will complete several steps in order: conversion of the mzML to DIA format, library search scoring, Percolator filtering, retention time alignment, Refined Percolator filtering, and then quantification. The resulting search should detect and quantify around 34,398 unique peptides, matching the manuscript result from Figure 3. The specific number may change by a small amount depending on your operating system because of sort ordering in Percolator.

D. Once the file for this library has concluded running, the file bar will turn green. Although there is only one file, and match-between-runs quantitation will not be performed, we can click “Save Quant Reports.”



Once the quantitation has concluded, the screen will look the same as the screenshot below. At this point, you should exit from Scribe, and re-open a new window before running another search.

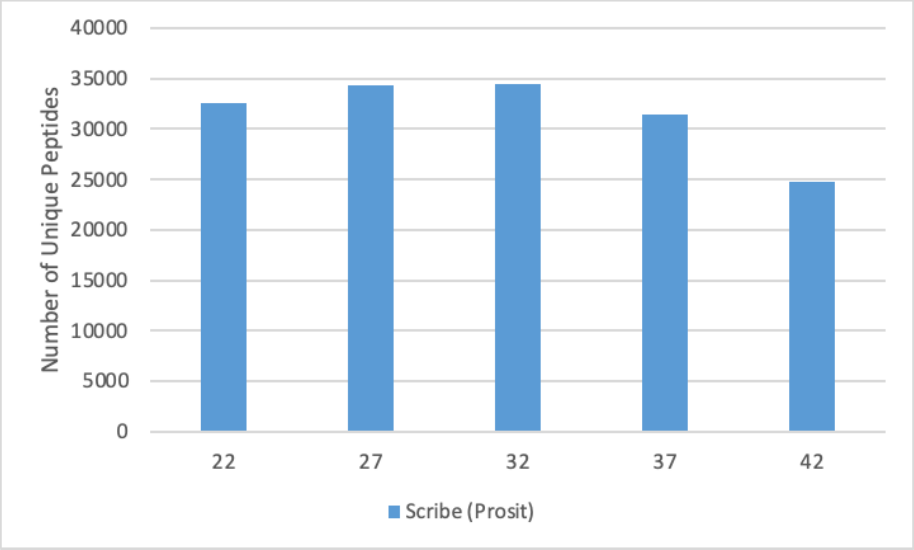


E. To process the other files found in the repository, consult the table below. DLIB Libraries are paired to HeLa mzMLs based on experimental NCE and Prosit NCE, where NCEs are tuned to the specific instrument. Here are the pairings for this experiment in Figure 3 and Figure 4:

Experimental NCE	Experimental File (mzML)	Library File (DLIB)
22	2022_09_01_HeLa_500ng_16mzst_DDA_NCE_22.mzML	uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce16_hcd2020.dlib
27	2022_09_01_HeLa_500ng_16mzst_DDA_NCE_27.mzML	uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce29_hcd2020.dlib
32	2022_09_01_HeLa_500ng_16mzst_DDA_NCE_32.mzML	uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce34_hcd2020.dlib
37	2022_09_01_HeLa_500ng_16mzst_DDA_NCE_37.mzML	uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce41_hcd2020.dlib
42	2022_09_01_HeLa_500ng_16mzst_DDA_NCE_42.mzML	uniprot_human-reference_reviewed_2022mar02.prosit_input.trypsin_nce50_hcd2020.dlib

If you run into error messages or Scribe shuts down, please check your version of Java. Ideally, you will be using Java version 1.8. We know of errors that were introduced with Java versions 17 and 18 and we recommend using Java 16 or earlier to ensure stability.

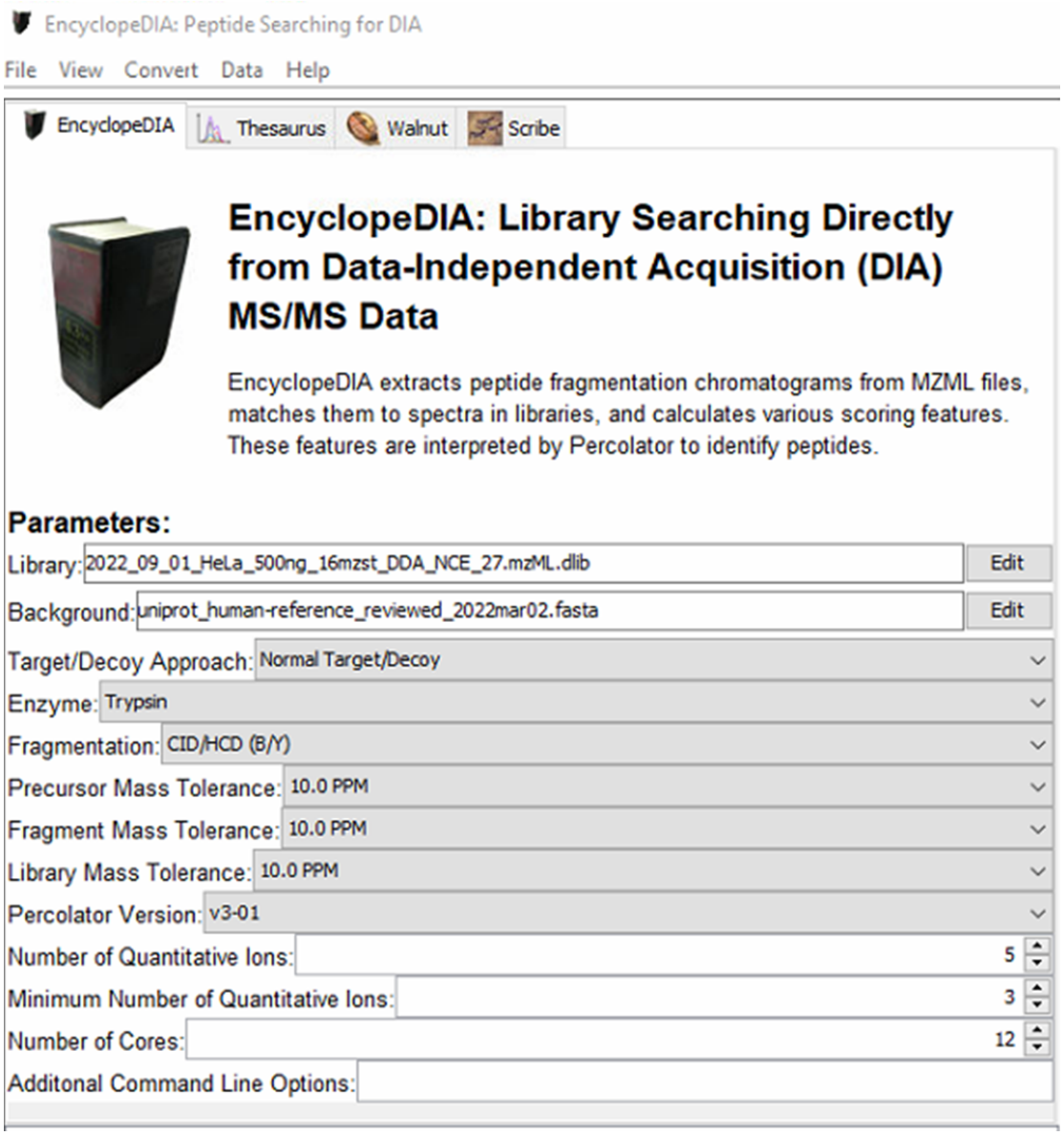
F. Plot the number of peptides identified in each run in any downstream software (e.g., R or Excel). The resulting graph should be similar to the figure below.


40000
35000
30000
25000
20000
15000
10000
5000
0
Number of Unique Peptides
22
27
32
37
42
Scribe (Prosit)

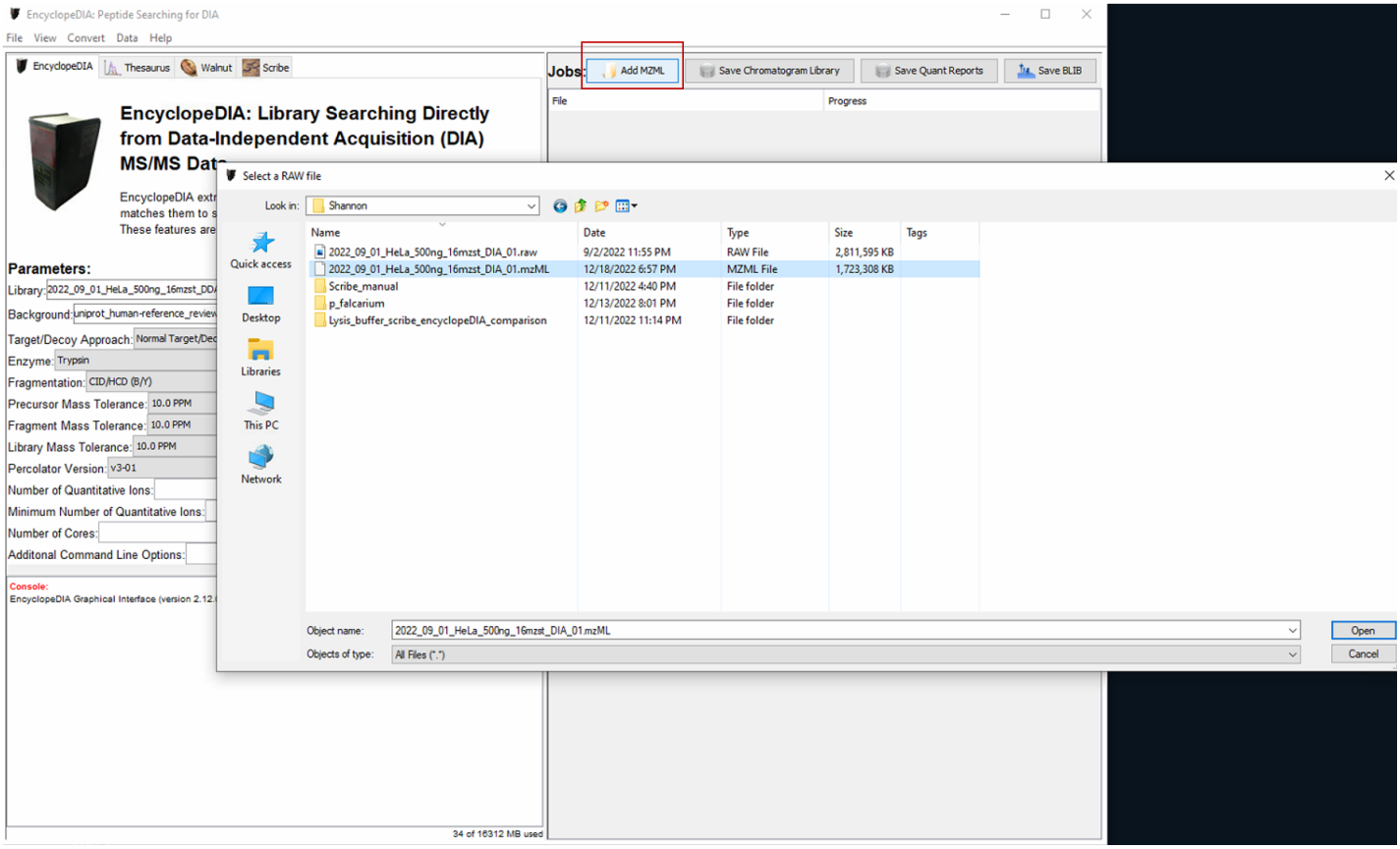
8. USING SCRIBE TO BUILD LIBRARIES FOR DIA SEARCHES

How to use spectral libraries made from DDA to search against DIA experiments

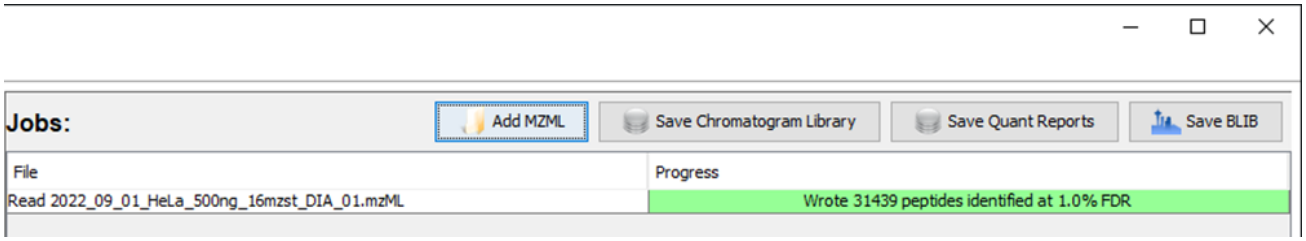
A spectral library, .DLIB, file is output from Scribe. With the latest version of EncyclopeDIA, you can use the acquired spectral library to search DIA injections. In this example, we are using the collision energy optimized library acquired from a HeLa DDA injection with NCE = 27. Open EncyclopeDIA, and upload "2022_09_01_HeLa_500ng_16mzst_DDA_NCE_27.DLIB." Specify "uniprot_human-reference_reviewed_2022mar02.fasta" as the background file. The settings should match the screenshot below.



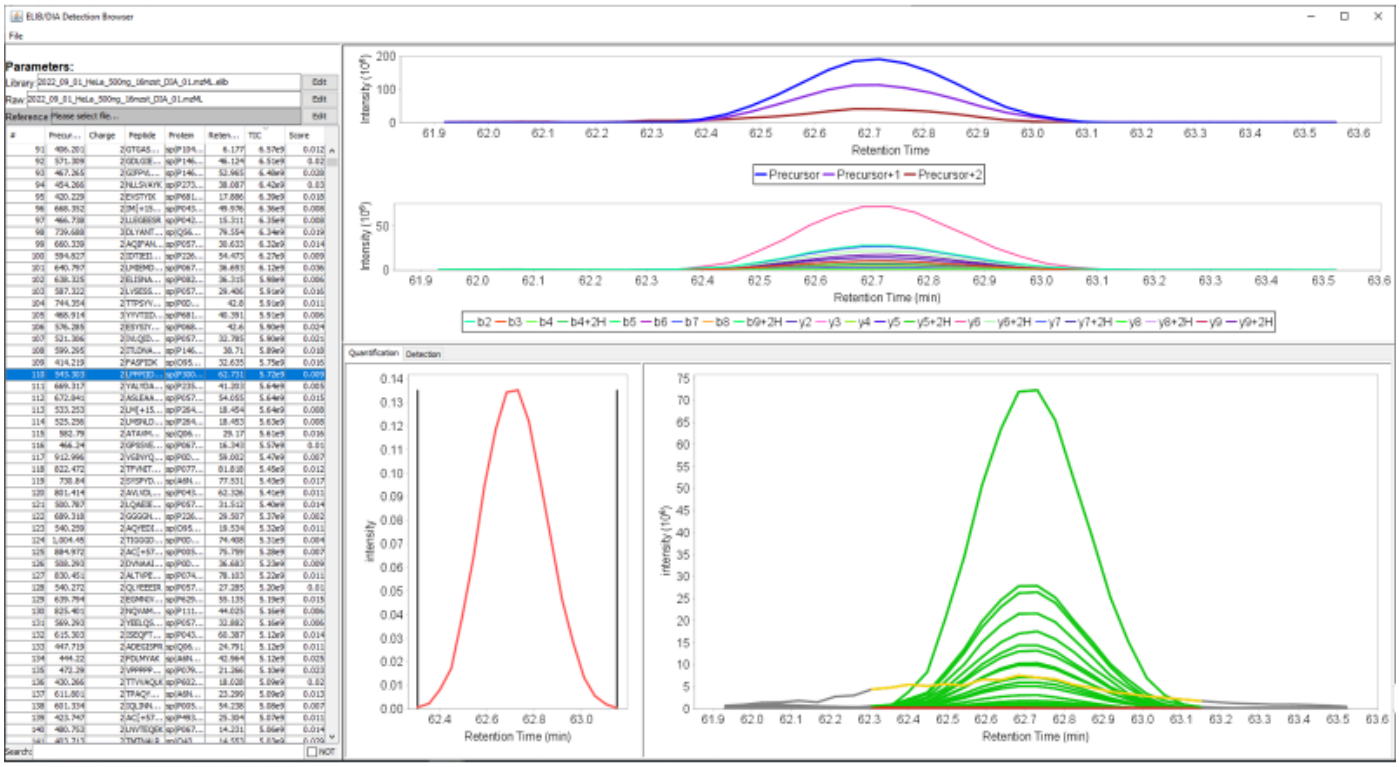
Click “Add mzML,” then find the DIA injection named “2022_09_01_HeLa_500ng_16mzst_DIA_01.mzML” to begin the search.



Once the search is complete, the detected number of peptides should be “31,349.”



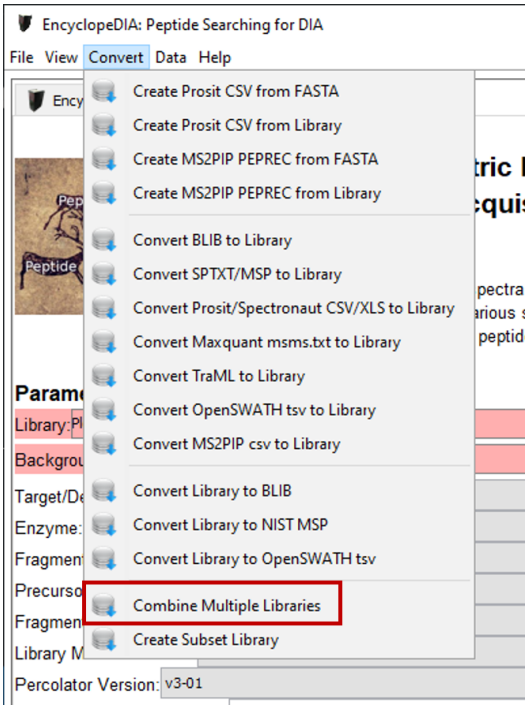
You can then visualize detected peptides as chromatograms using the “ELIB Browser.” The peptide below is an example of a peptide that was detected with little interference and multiple fragment ions.



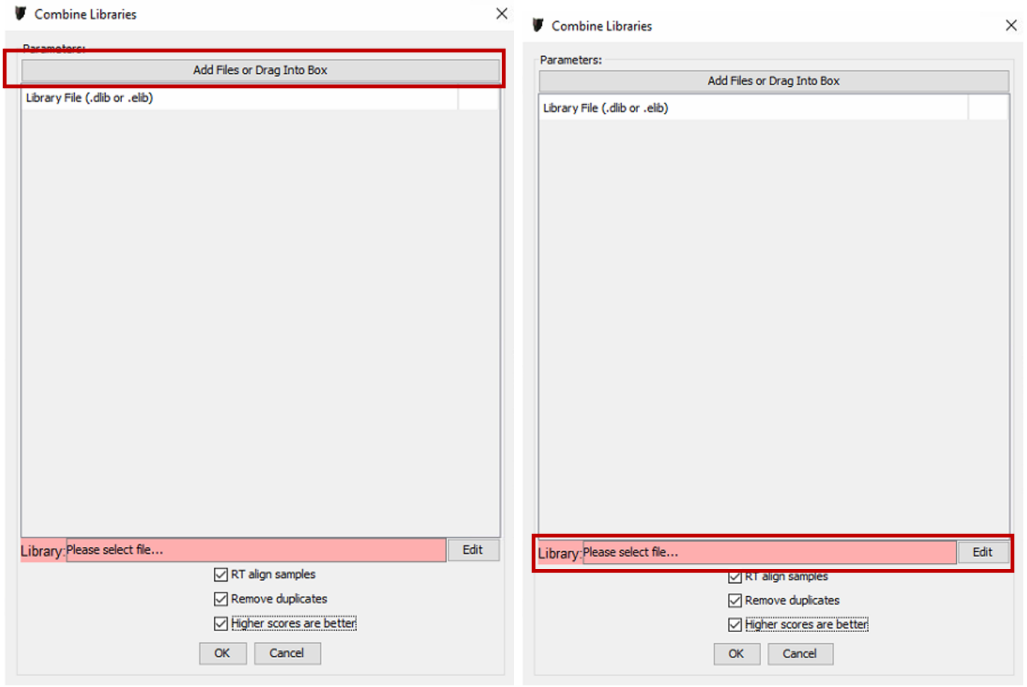
9. FAQ

A. Can multiple organisms be searched at once?

Yes, you can by generating a Prosit library for each organism using the method outlined in part 2, then using Encyclopedia/Scribe's "Combine Multiple Libraries" feature to merge both DLIBs. With this method, both organisms can be searched at once. The specified FASTA file contains both organisms. The "Combine Multiple Libraries" feature is outlined below:



Click on "Add Files or Drag Into Box" to add files to the queue. Then select "Library" at the bottom of the window to name the merged library. Click okay to start the process.



B. When are additional decoys necessary?

Scribe can search for additional decoys as entrapment peptides by adjusting the “Target/Decoy Approach”. When doing this, Scribe adds additional entrapment peptides that are generated by shuffling the sequence of target peptides. Decoy entrapment peptides (reverse shuffled) are also added. If the percentage of detected entrapment peptides is higher than the filtered FDR percentage, then this can help indicate potential errors that escape target/decoy analysis performed by Percolator.

For more information on entrapment as a tool for estimating confidence in proteomics datasets, we recommend the following paper:
Granholm, V., Fernández Navarro, J., Noble, W.S., Käll, L.: Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. J. Proteom. 80, 123–131 (2013)