EncyclopeDIA Tutorial

Tutorials based on EncyclopeDIA version 4.6.0-SNAPSHOT
Last update on July, 8th 2024
Written by AE Shannon (shannon.225@osu.edu)

During this tutorial, you will learn how to

1. Convert Thermo .RAW files to a universal format (.mzML) using Proteowizard.
2. Search gas-phase fractionated samples against a prosit library to generate a chromatogram library.
3. Searching wide-window sample injections against a chromatogram library to detect peptides in a set of samples.

Additional material includes how to

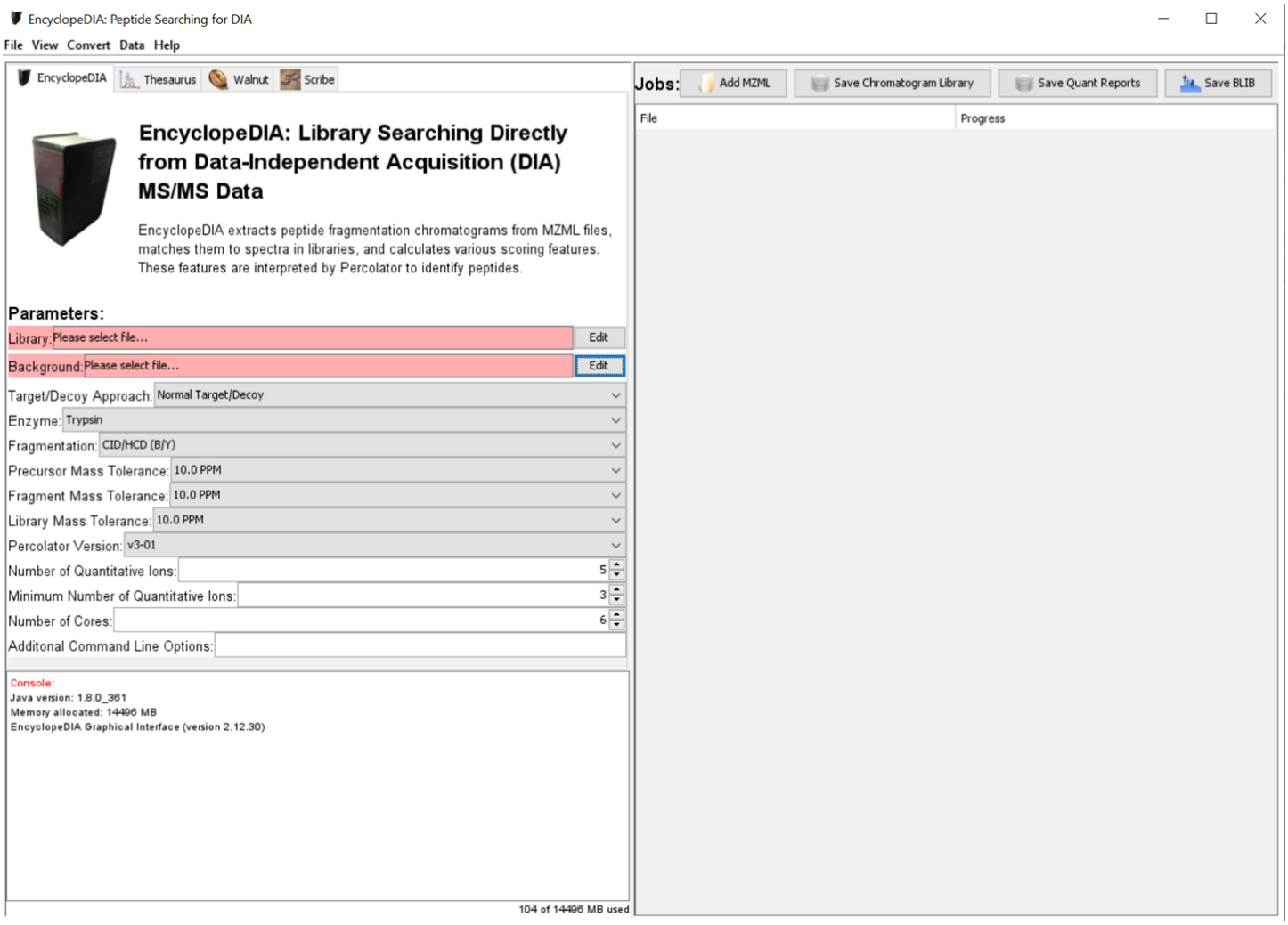
4. Make a prosit-generated, predicted spectral library using a FASTA through the prosit server.
5. View detections in EncyclopeDIA and other outputs
6. Upload detections to Skyline

For today's tutorial, we will focus on the bolded portions. We will do these together on the screen.

You can find the EncyclopeDIA processed libraries, FASTA file, and backup Skyline documents here.
Raw data can be acquired from the 2022 Wang paper https://doi.org/10.1038/s41597-022-01845-x.

OVERVIEW

EncyclopeDIA 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 processes EncyclopeDIA is running.



This tutorial outlines how to use EncyclopeDIA, what outputs are given, and how to use the outputs to visualize the data.

1. PREREQUISITES AND INSTALLING ENCYCLOPEDIA

Software requirements needed to process data using EncyclopeDIA.

EncyclopeDIA is a cross-platform Java application that has been tested for Windows, Macintosh, and Linux. EncyclopeDIA 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.

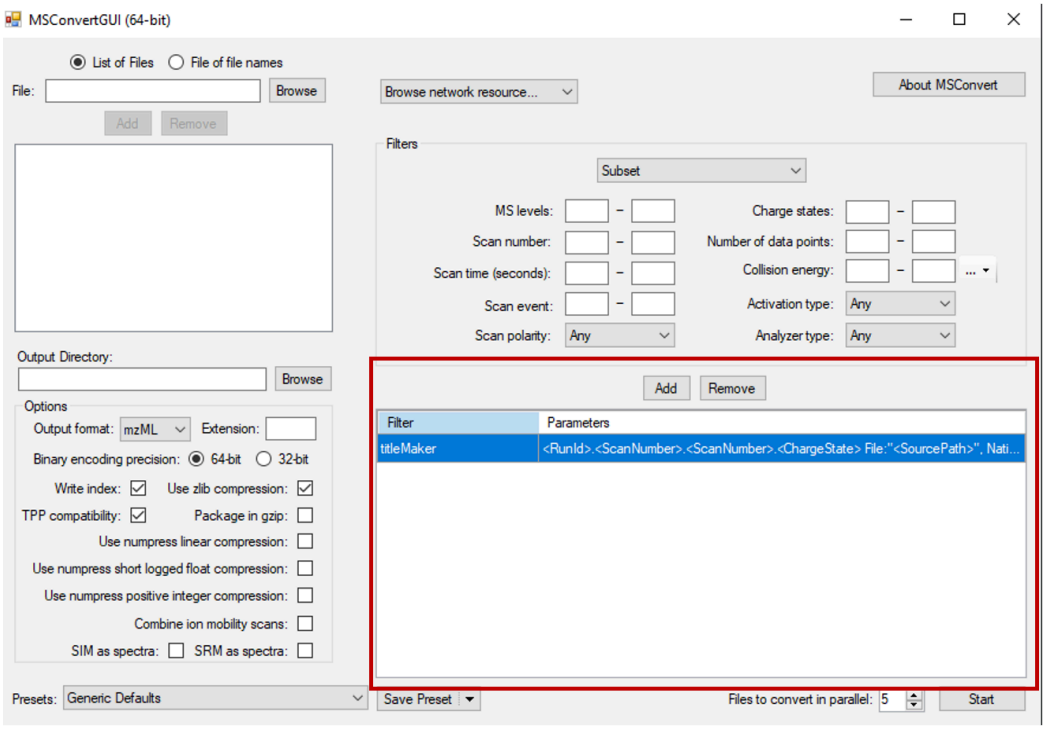
B. Scribe is folded into the EncyclopeDIA software package. After you have 64-bit Java 1.8, go to EncyclopeDIA's bitbucket page (https://bitbucket.org/searleb/encyclopedia/wiki/Home) and download the most recent stable version. Once downloaded, double-click on the EncyclopeDIA .JAR file to launch the 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.

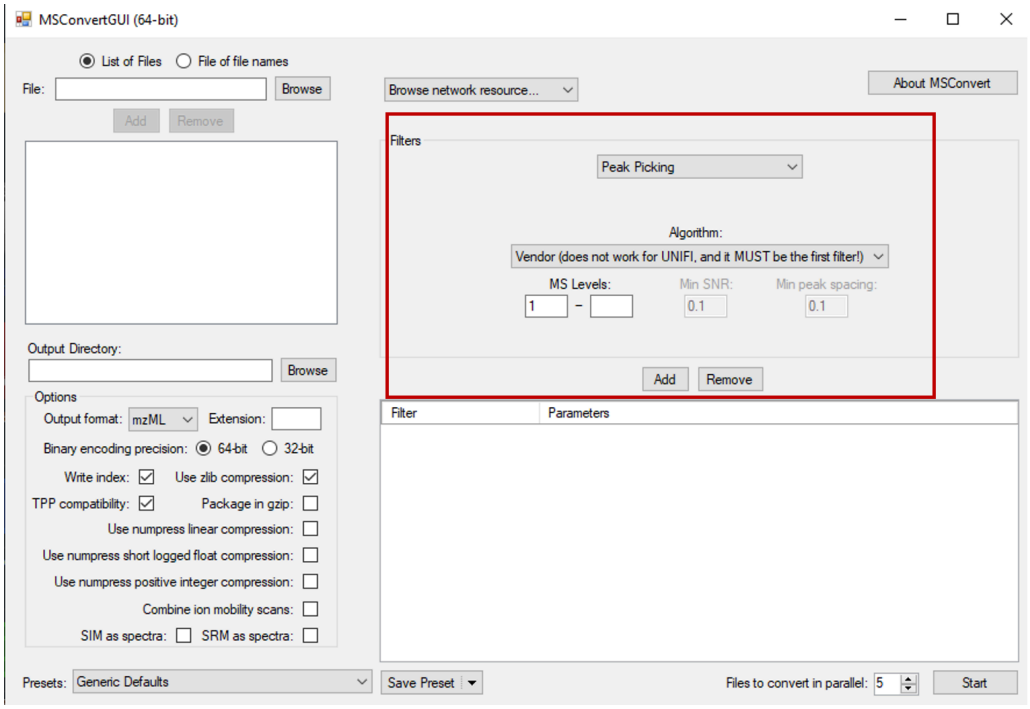
2. GENERATING MZMLS

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

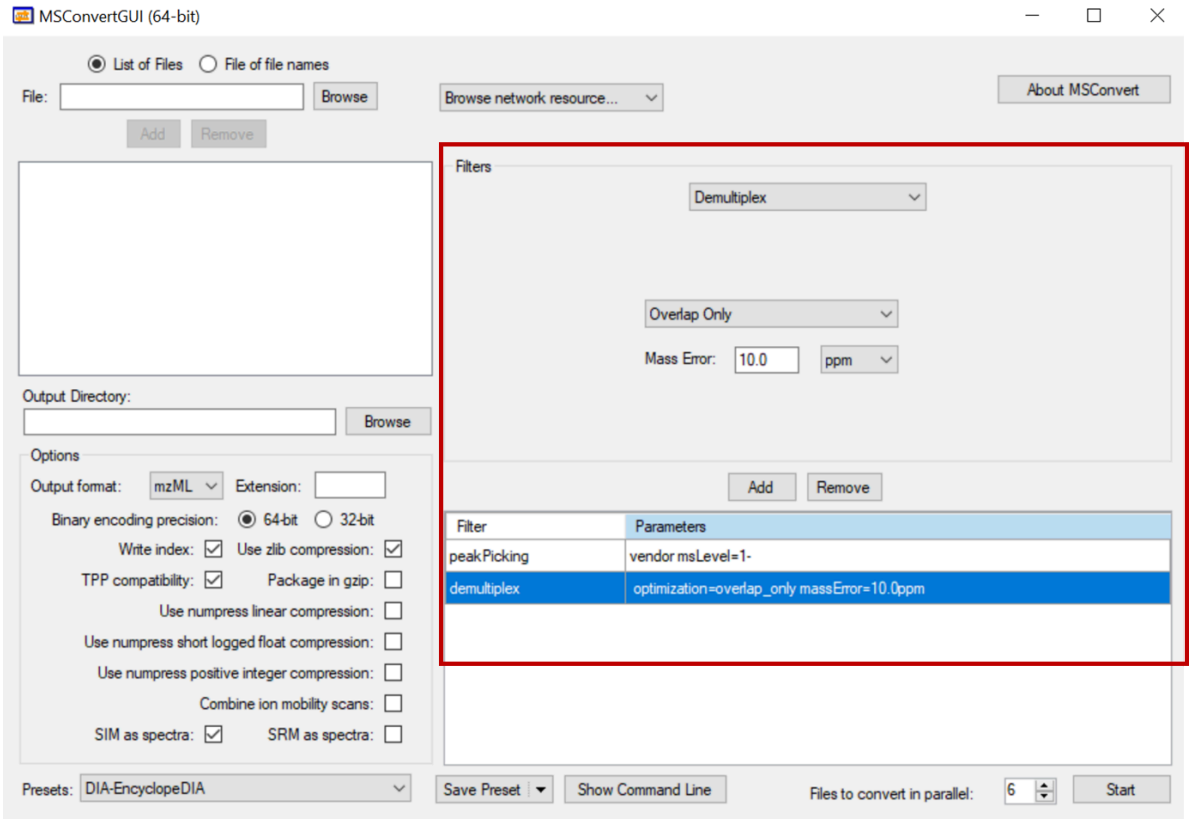
A. Before searching files in Scribe or EncyclopeDIA, RAW files must be converted to .mzML files using Proteowizard. To do so, open Proteowizard. Remove the “titleMakers” filter by selecting the filter in the parameters box, and clicking remove.



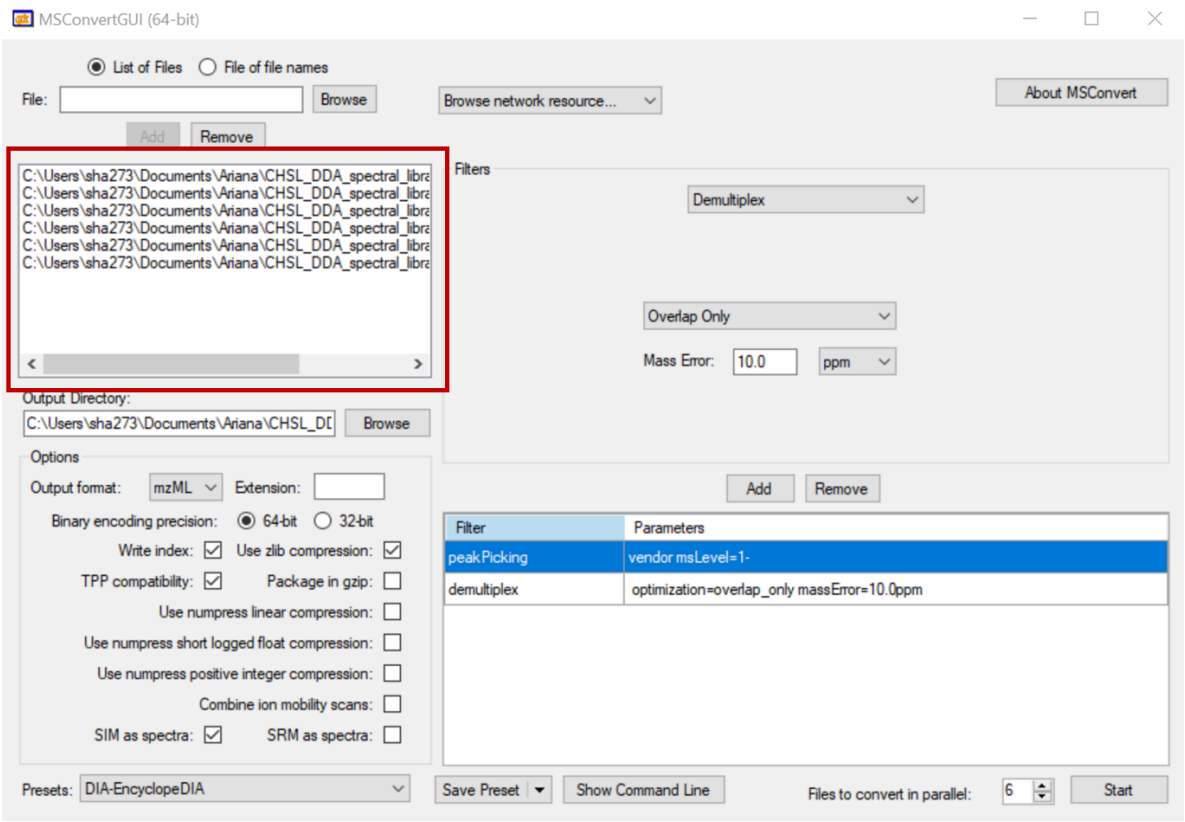
B. For converting DDA injections, or DIA injections that did not use an overlapping isolation windowing scheme, only add “Peak Picking” by selecting the peak picking option under filters, then click add. This should be the only filter in the box.



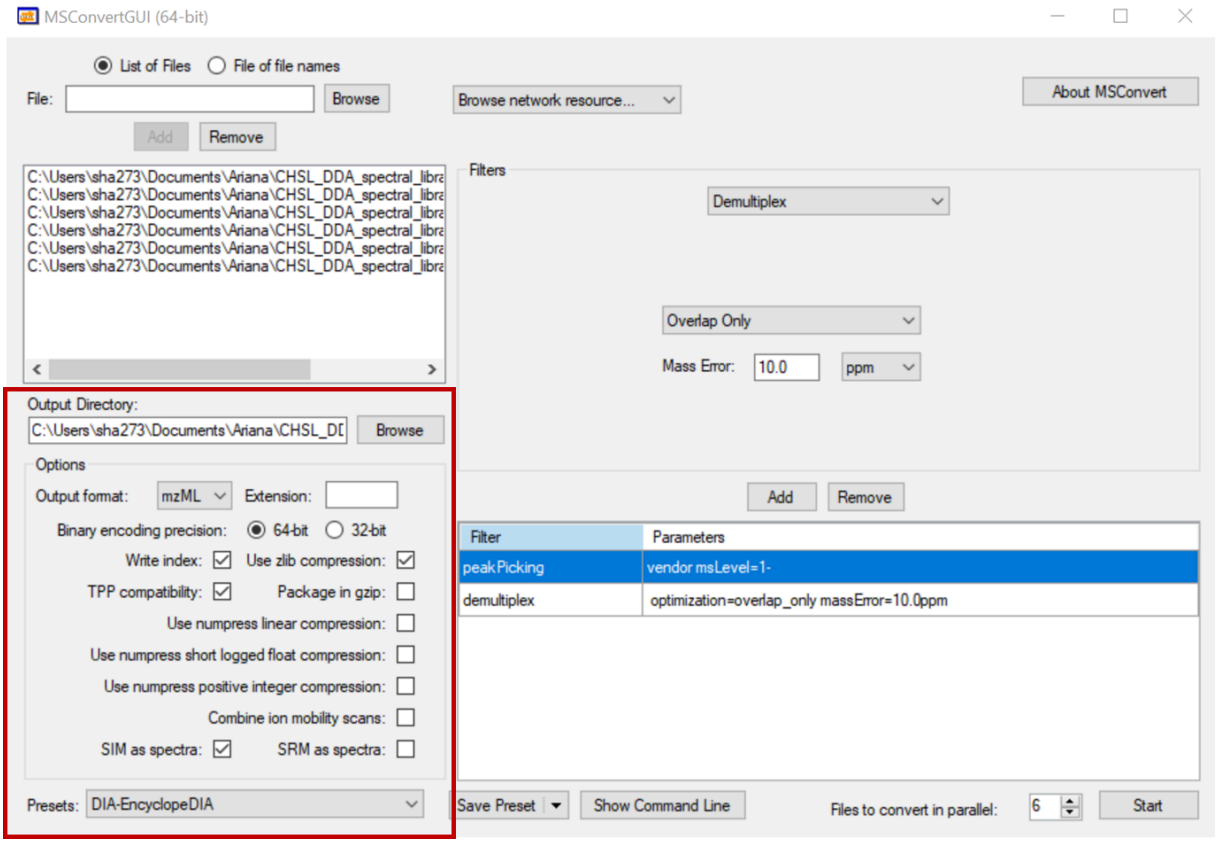
C. **ONLY FOR DIA INJECTIONS USING OVERLAPPING WINDOWS:** Click the drop down arrow and select demultiplex. Overlapping windows need to be demultiplexed after peak picking. Once you've found demultiplexing, you do not need to change anything else. Simply click "Add." Your window for overlapping DIA data should look at follows:



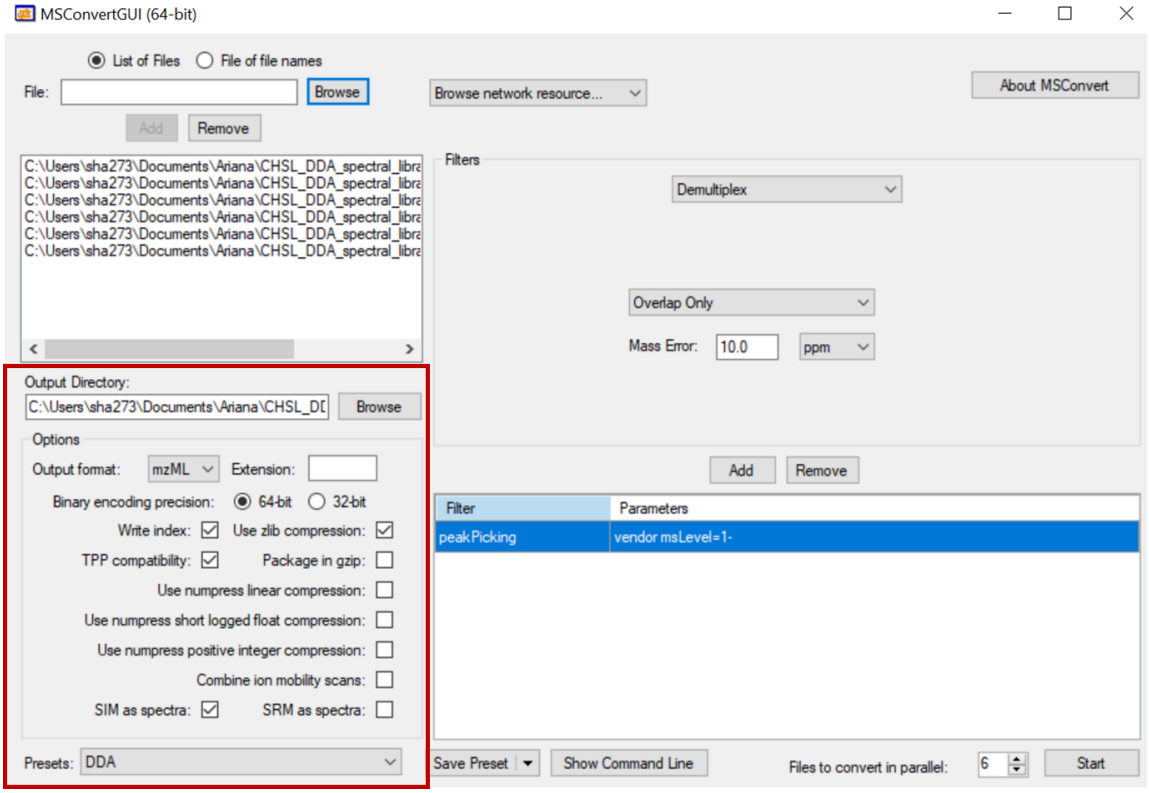
FOR DDA OR DIA WITH OVERLAPPING WINDOWS, your window will look like this:



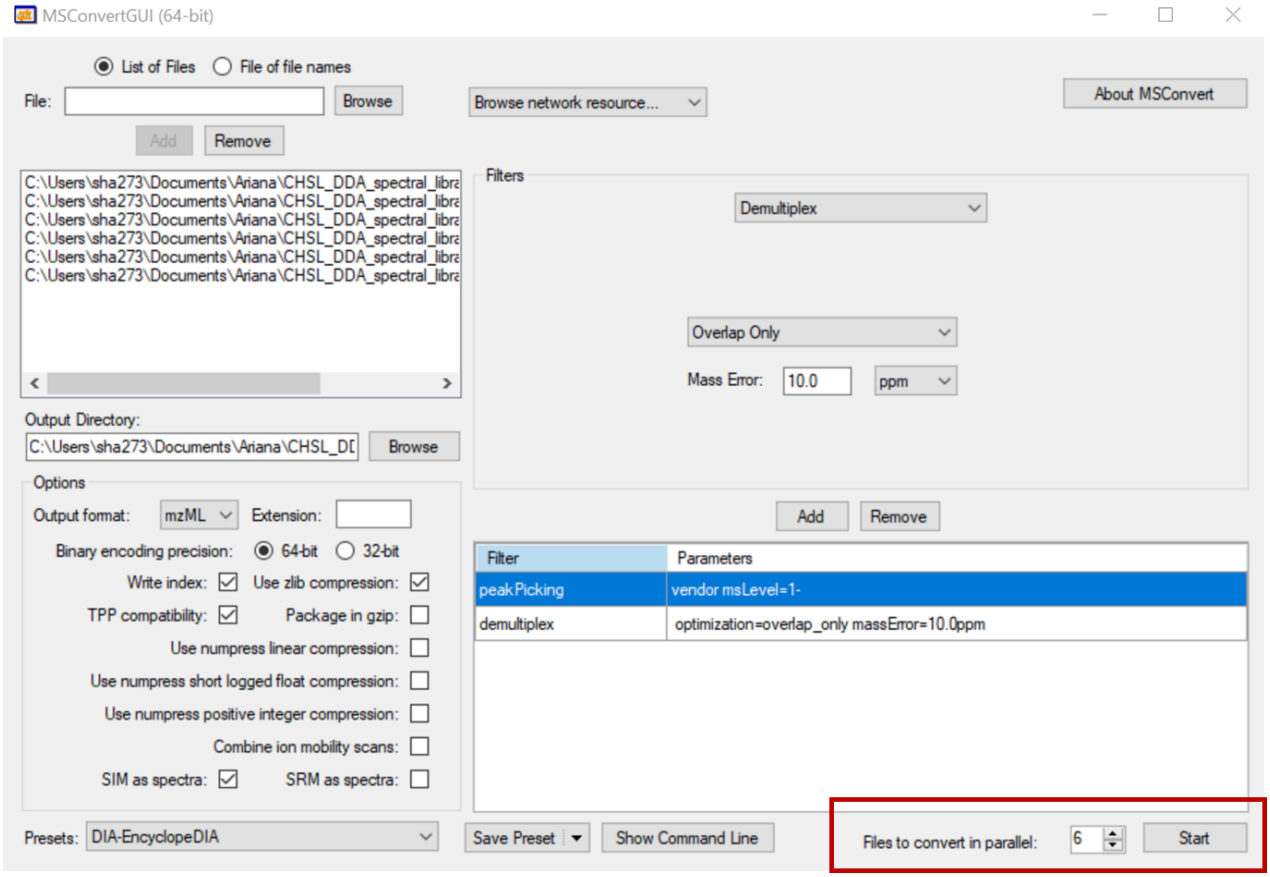
E. Select your files by clicking “browse”, and locate the desired DIA files in your directory. Click “add” on the left-hand side of the screen. The output directory will automatically populate where your files were selected from.

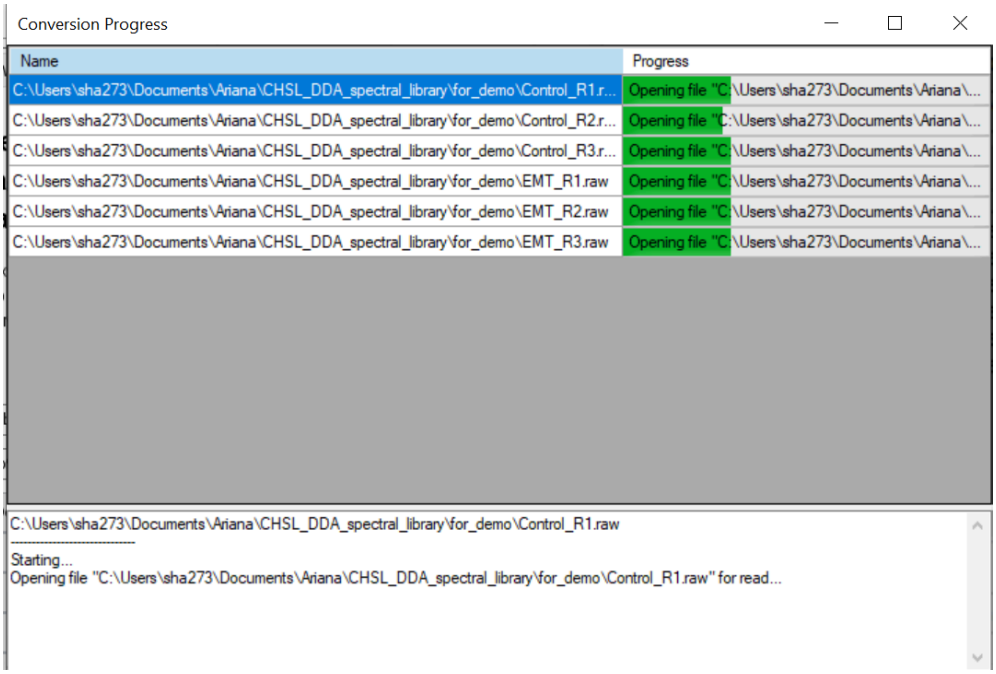


E. Under the options box, the output format should be mzML. You want “write index,” “use zlib compression”, “TPP compatibility, and “SIM as spectra” selected. Your settings should match the screenshot below.

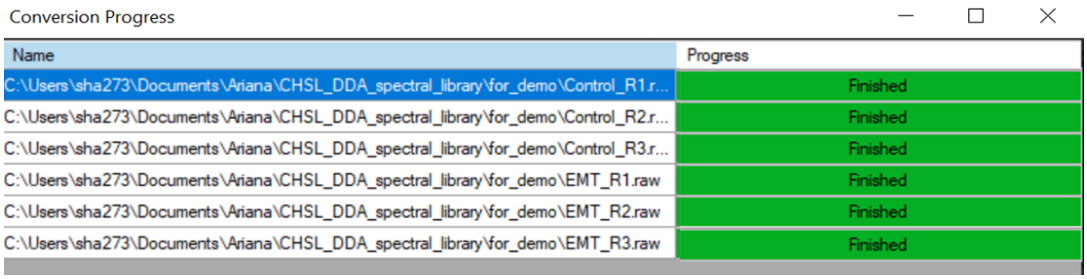


E. Click “Start” in the lower right hand corner to start the file conversion.



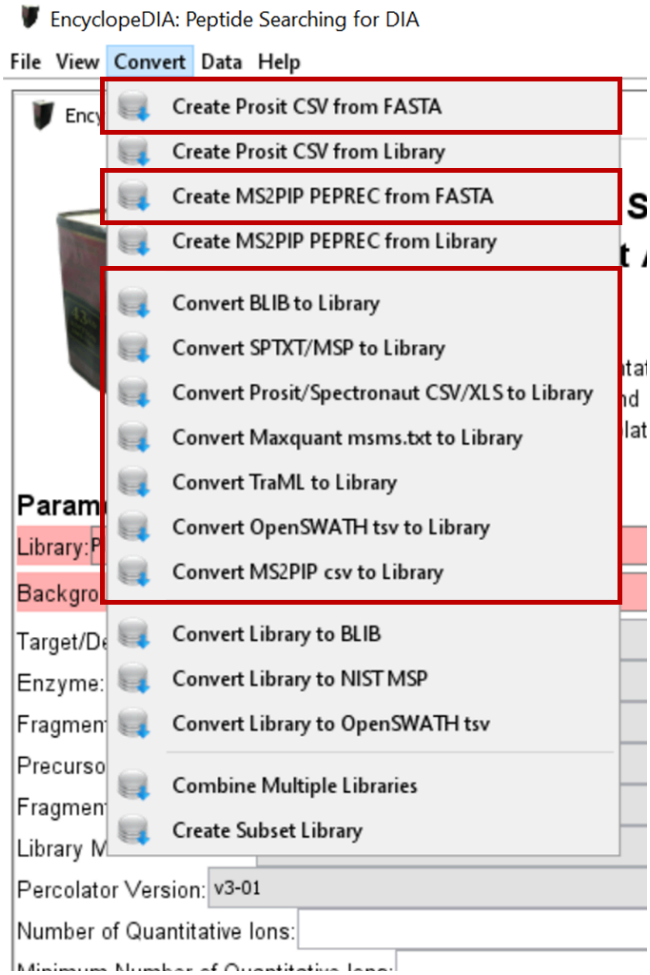


F. Once complete, the window will look like this. You can exit out of MSConvert once the conversion is done. Your files will be in the location previously specified.



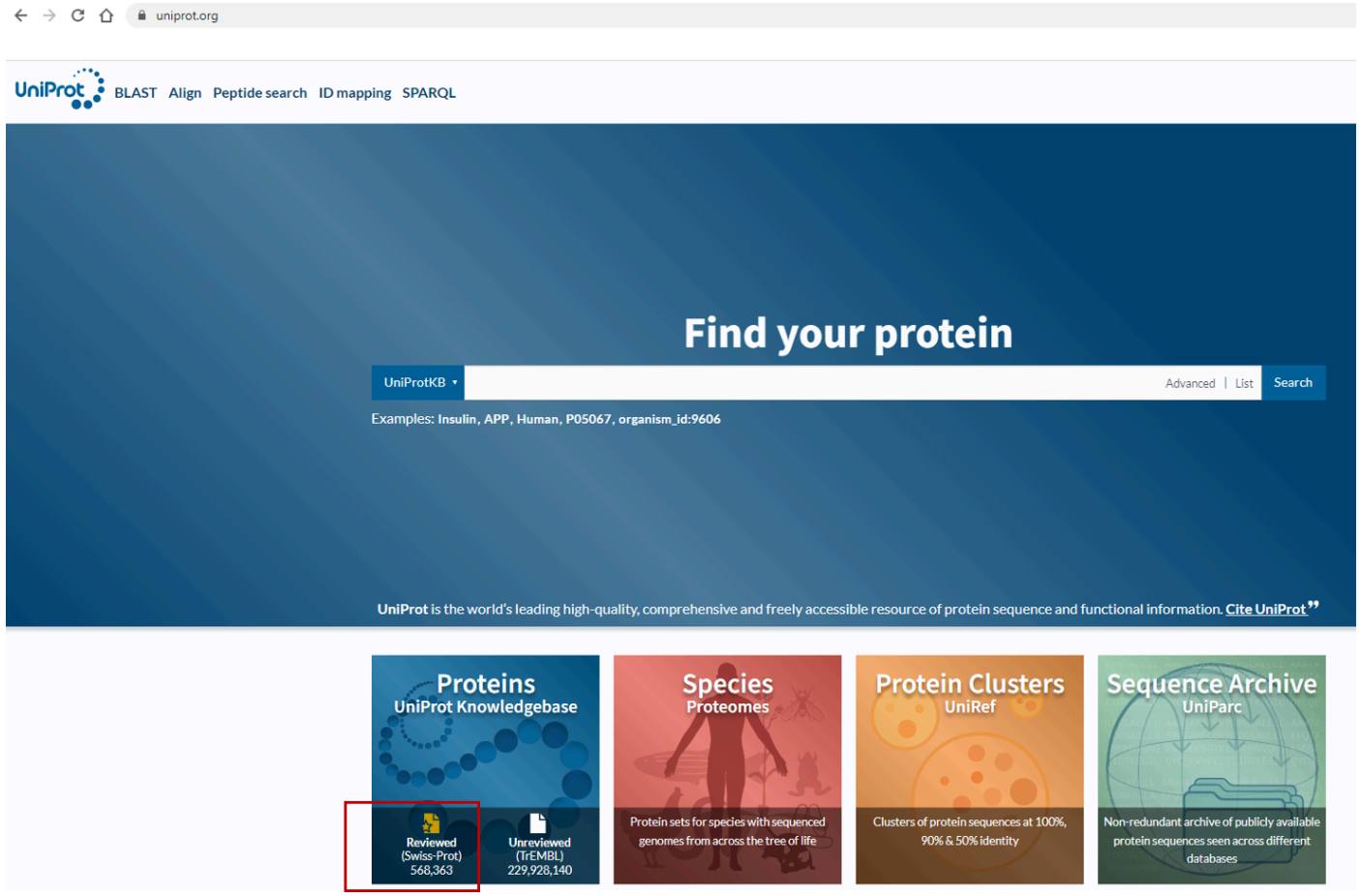
3. GENERATING LIBRARIES USING PROSIT AND ENCYCLOPEDIA

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

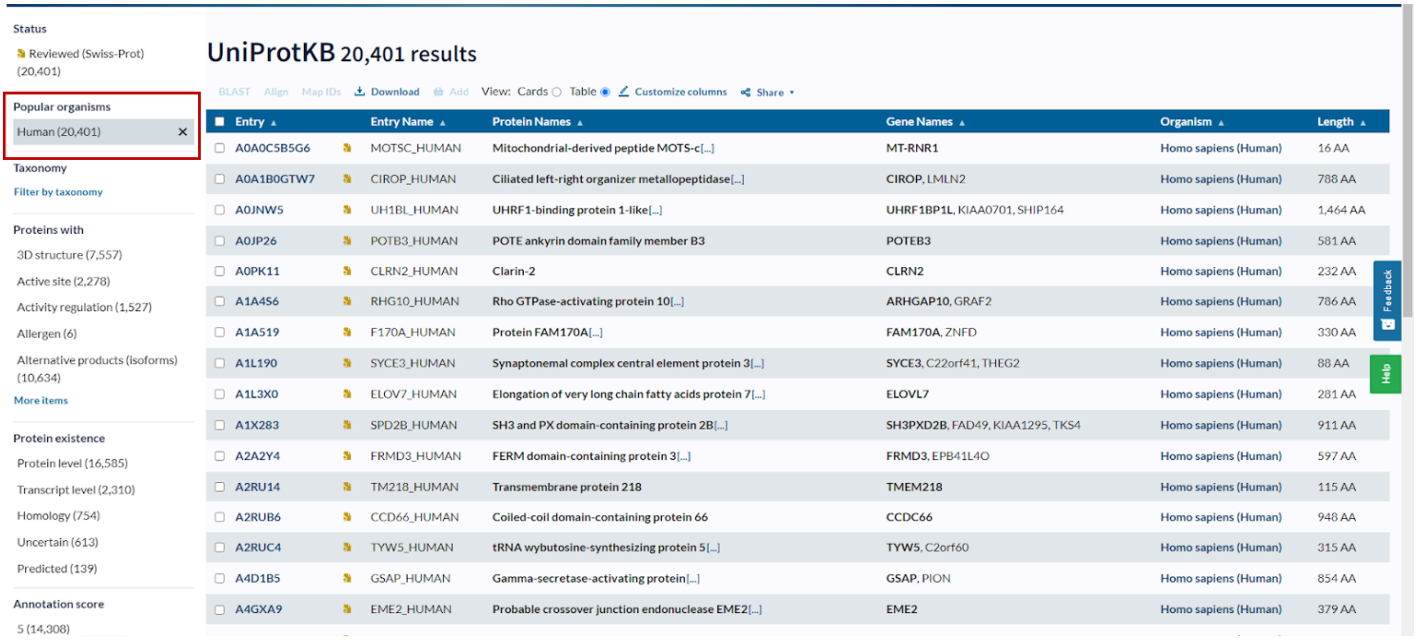


The primary library format EncyclopeDIA uses is .DLIB and .ELIB. Spectrum libraries from Skyline, NIST, TraML and other formats can be converted to .DLIB using the “Convert” menu. In particular, EncyclopeDIA can search fully predicted libraries generated by either Prosit or MS2PIP, and produce input 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. The files generated by EncyclopeDIA are .ELIBs, which are very similar to .DLIBs, but additionally contain retention time information.

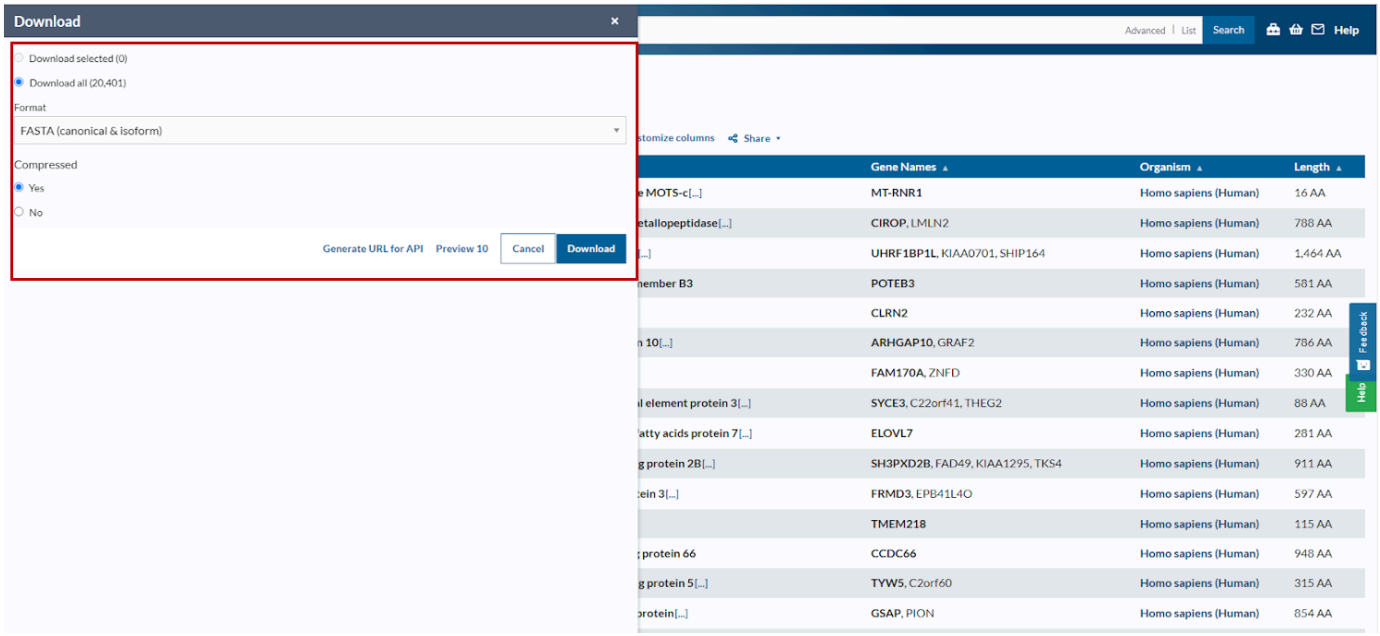
A. To generate a Prosit-predicted spectral library (.DLIB) for EncyclopeDIA, start by obtaining a FASTA file for your organism of interest. Given the computational expense of generating libraries, in most cases we recommend using only canonical protein sequences. For example, if you require a FASTA of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



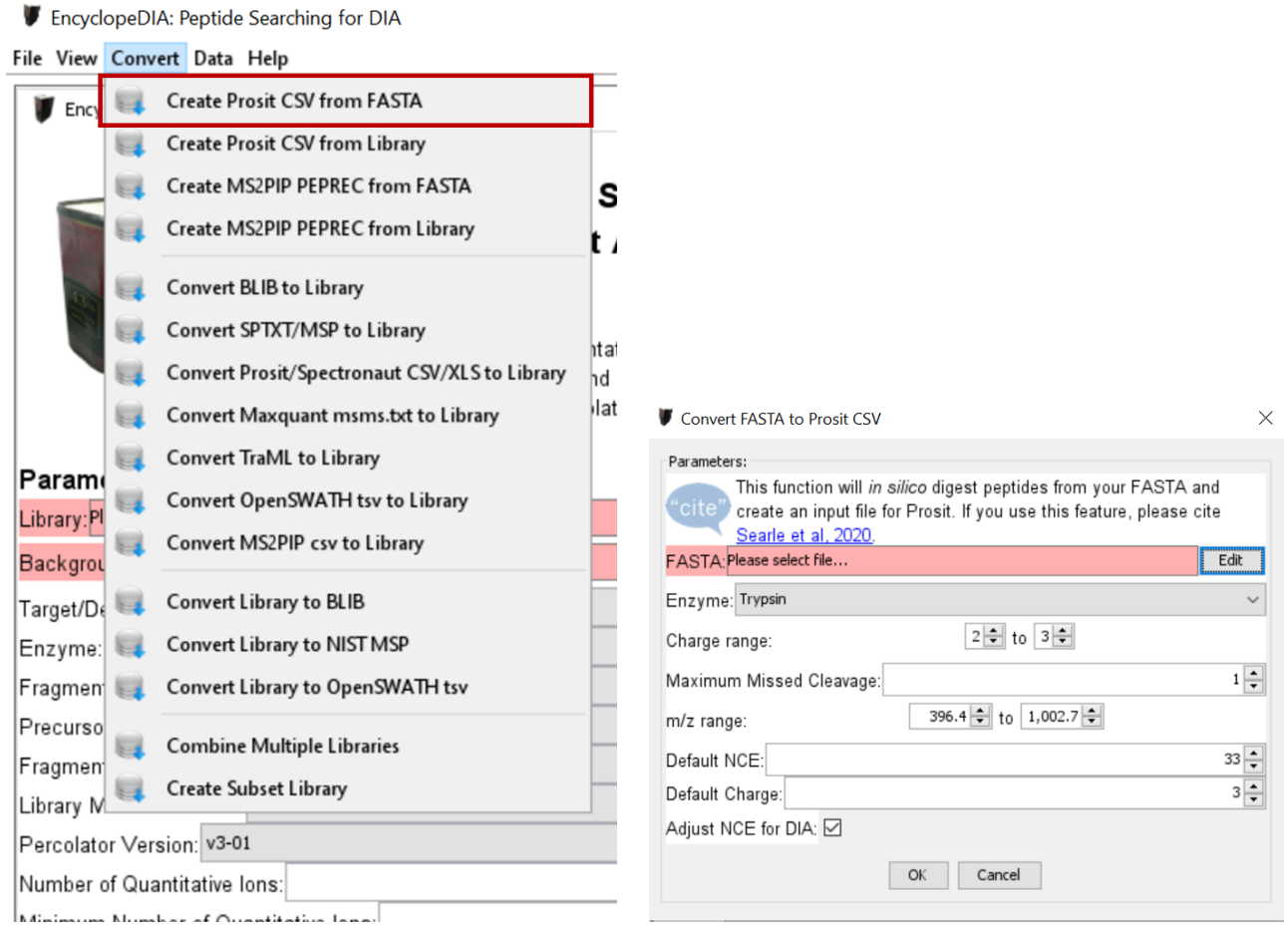
Specify “Human (20,401).”



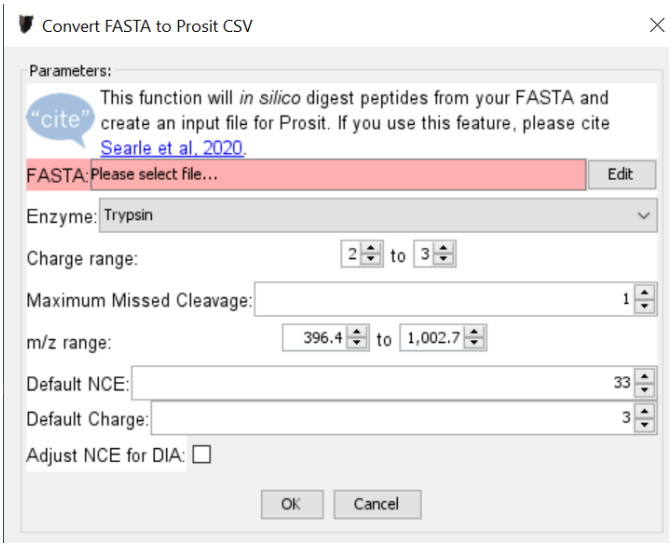
Select “Download.” If you download the compressed version, you will get a .gz and have to uncompress the file. If you download the uncompressed version, the file may take longer, but it will be in a FASTA format for immediate use.



B. Open EncylopeDIA 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. If you are generating a spectral library to search against DIA data, you should check “Adjust NCE for DIA.”

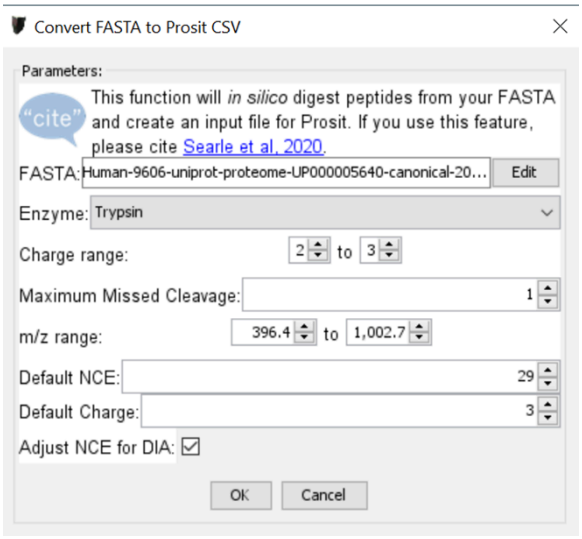


If you are building a prosit library to search against DDA data, you should leave the “Adjust NCE for DIA unchecked.



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 tutorial, you should load “Human-9606-uniprot-proteome-UP000005640-canonical-2022_04.fasta”

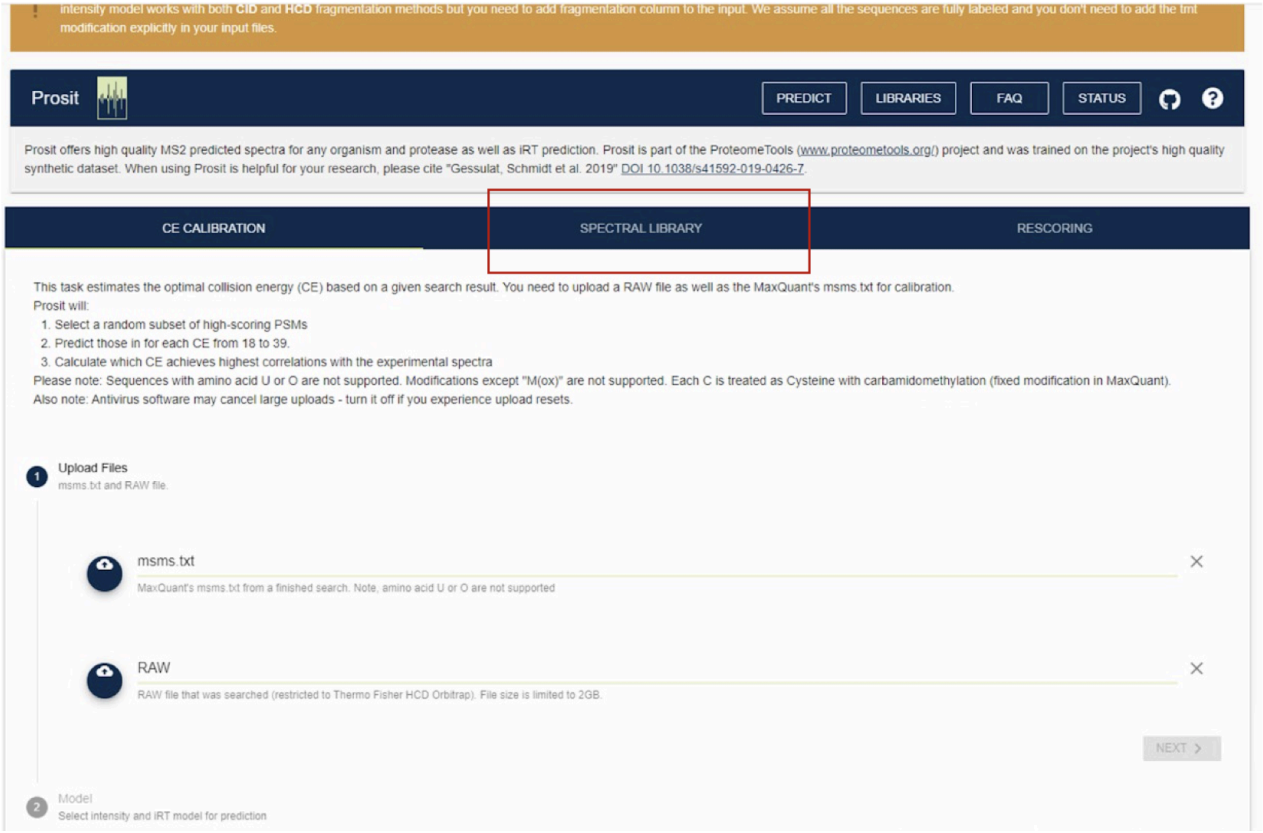
Settings for
“human_prosit_generated_abrf.dlib”



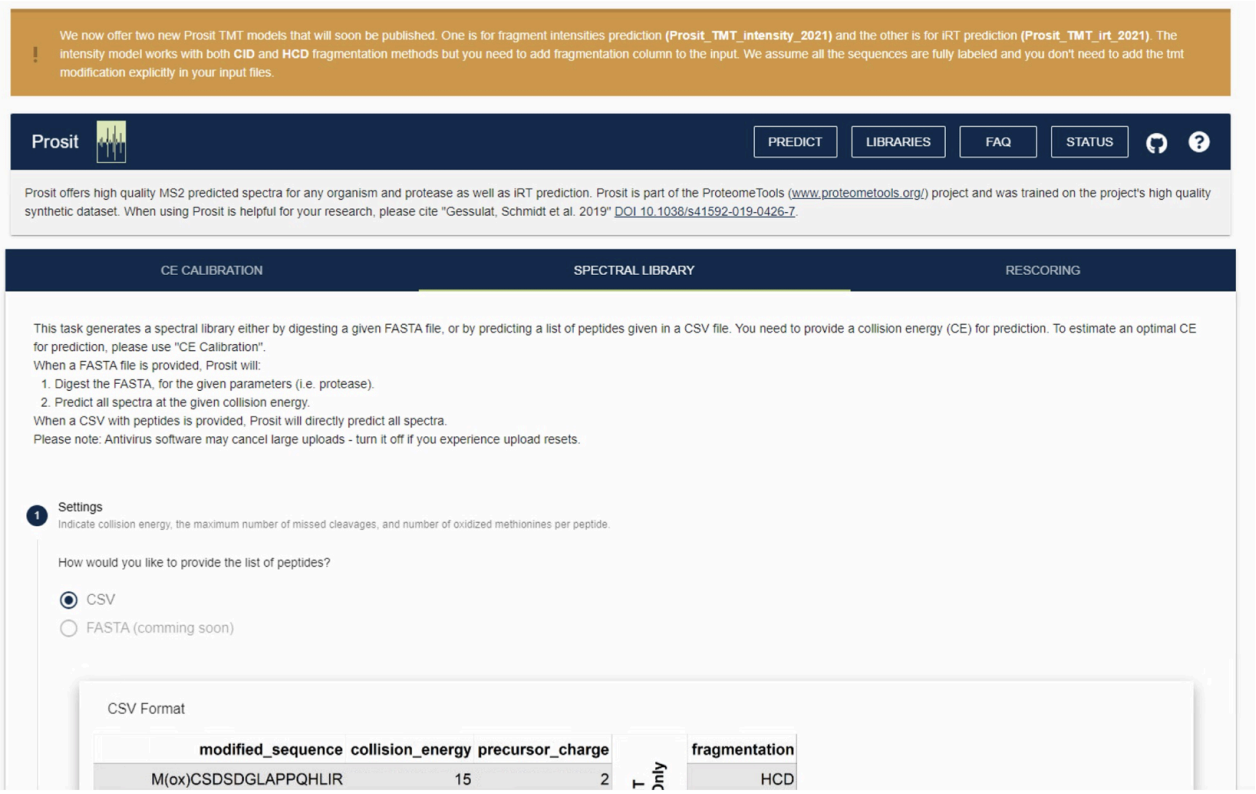
This will generate the file “Human-9606-uniprot-proteome-UP000005640-canonical-2022_04.fasta.fasta.trypsin.z3_nce29,” which is what you will use to put into Prosit.

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

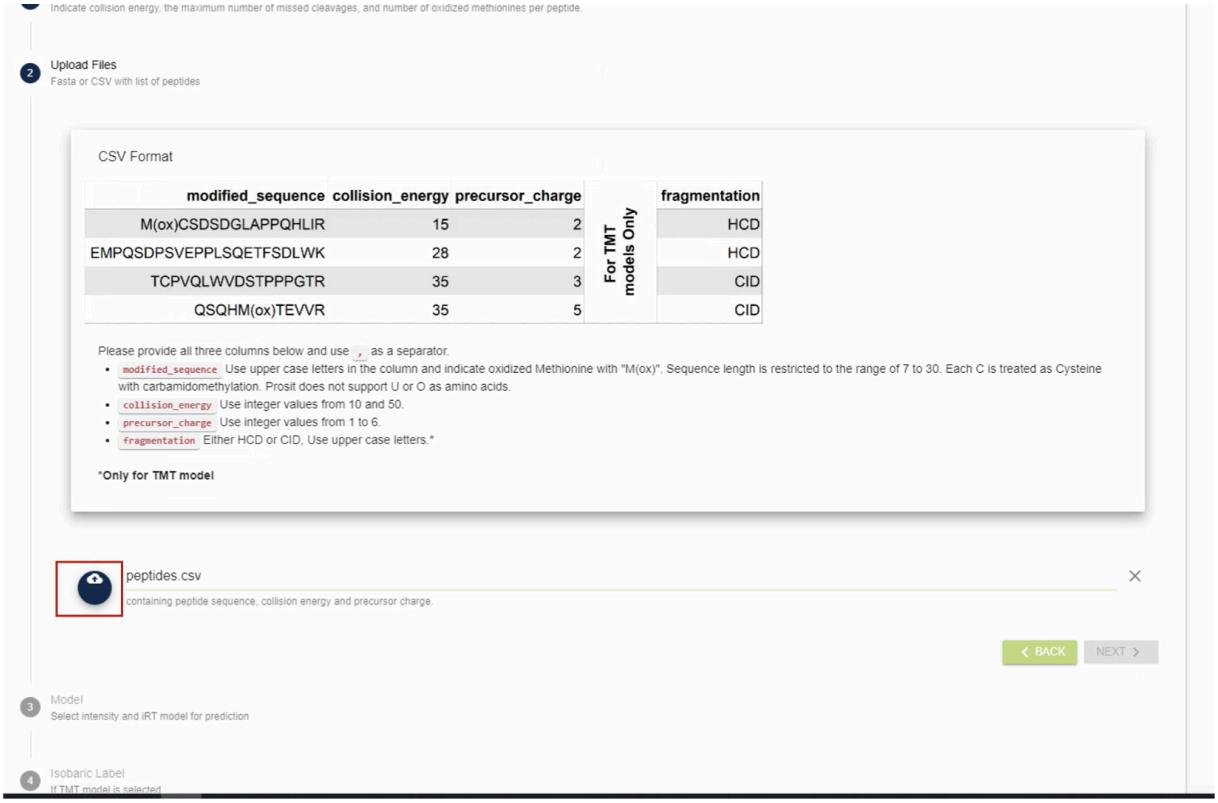
LIBRARY," and "RESCORING."



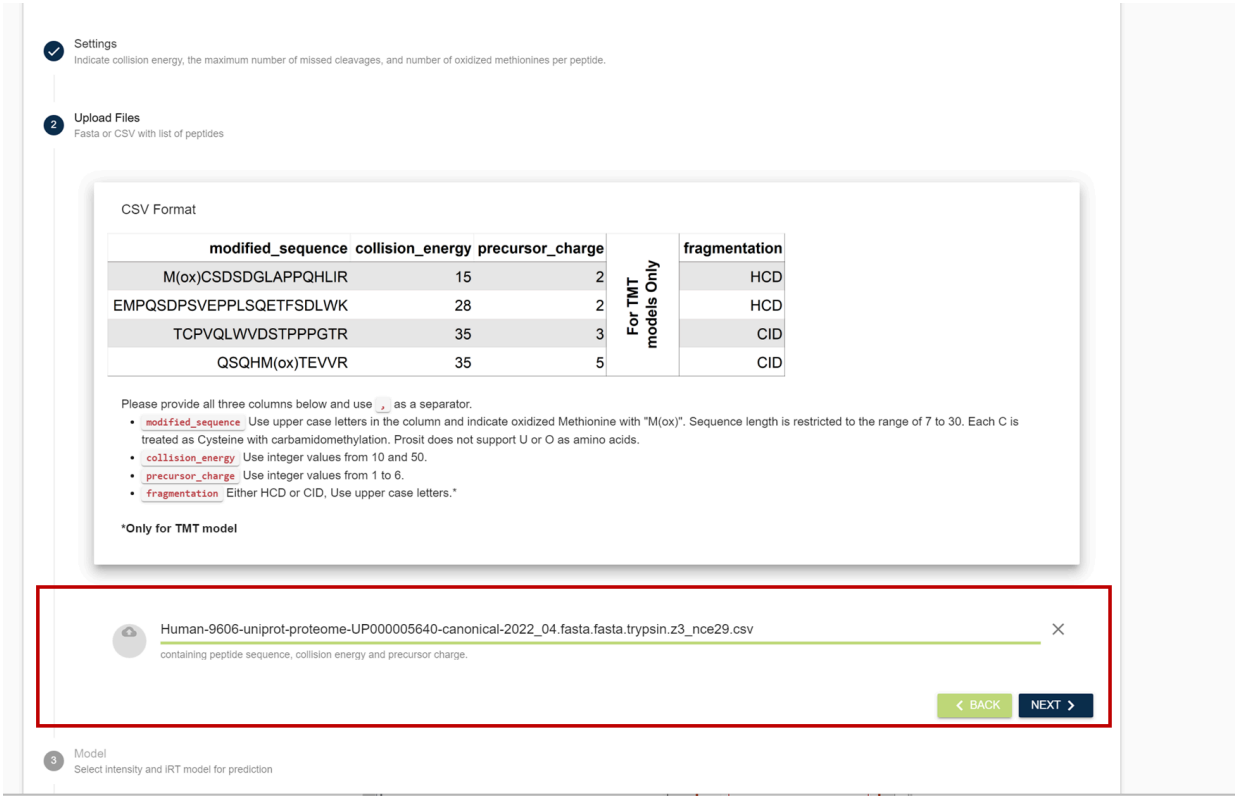
Go to the "SPECTRAL LIBRARY" tab.



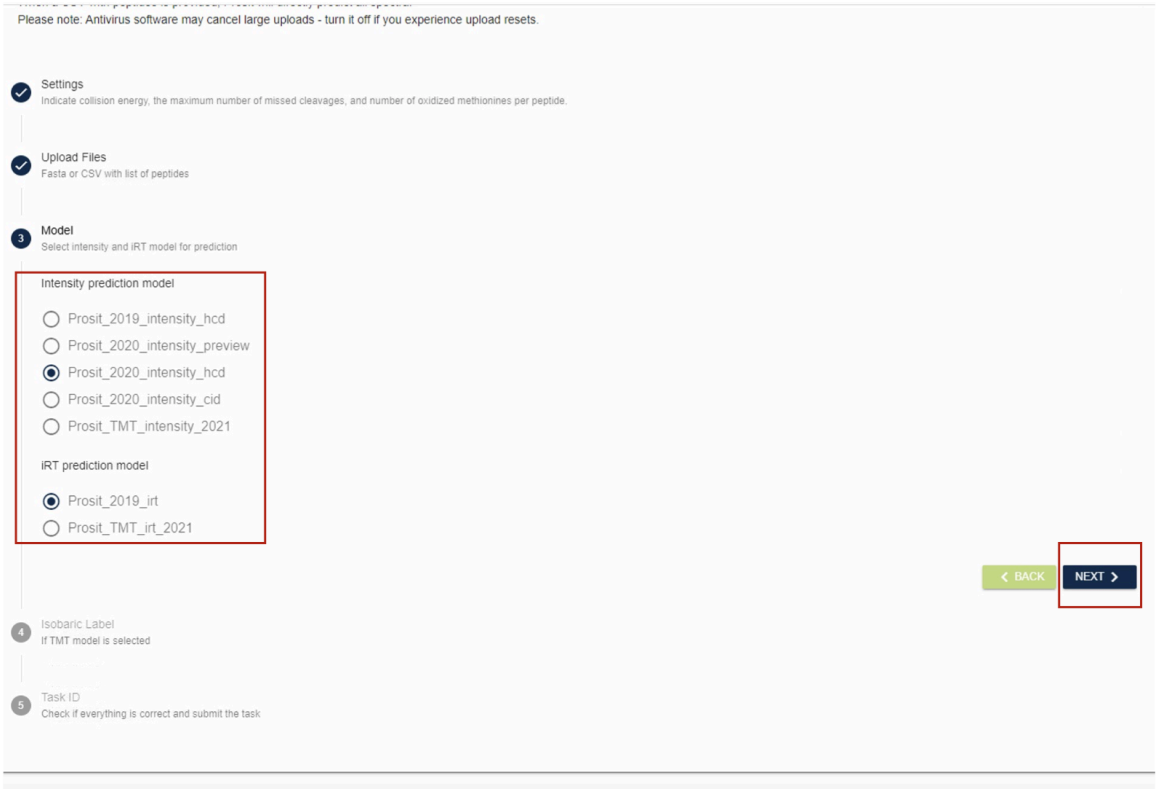
For the 1st step, CSV is already selected. You can click next to get to the 2nd step. Upload the Prosit CSV created in EncyclopeDIA.



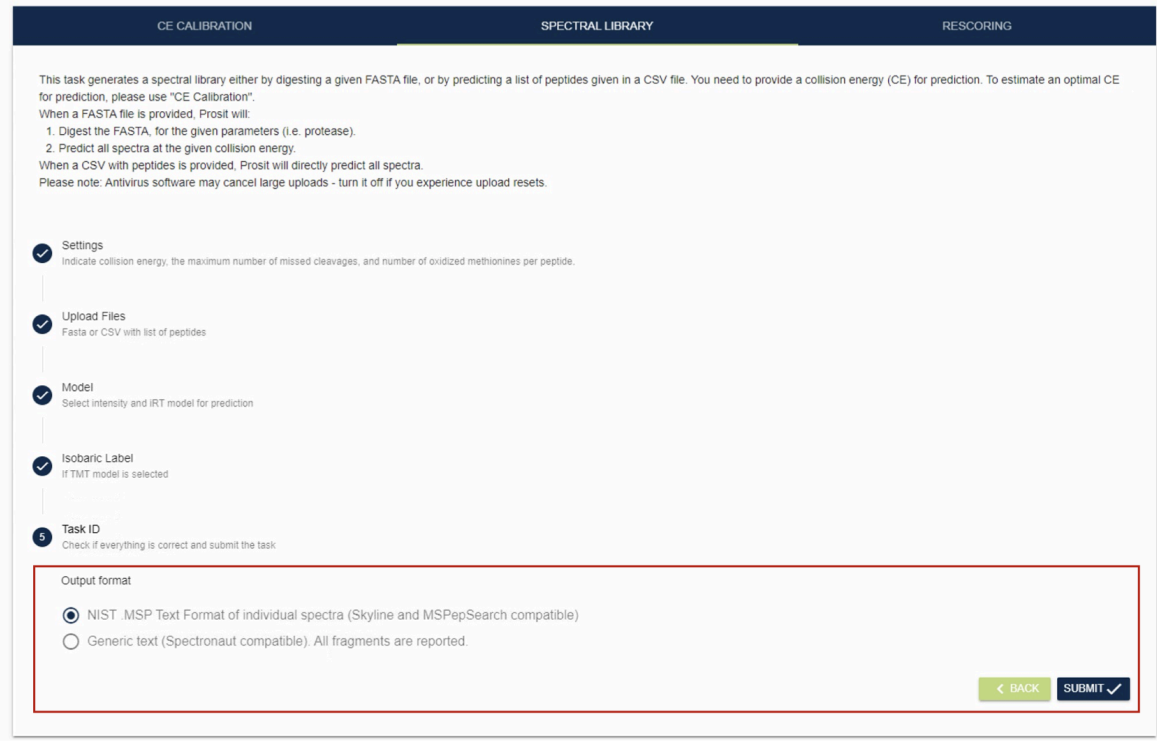
Once the CSV has once loaded, click next.



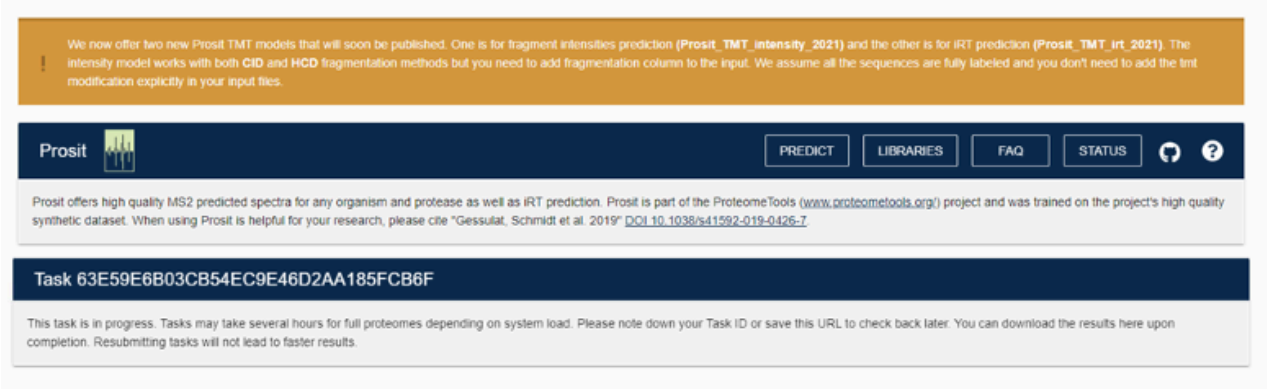
Select the desired model. For this example, we want to use the “Prosit_2020_intensity_hcd” for the Intensity prediction model, and the “Prosit_2019_irt” for the iRT prediction model. Click next.



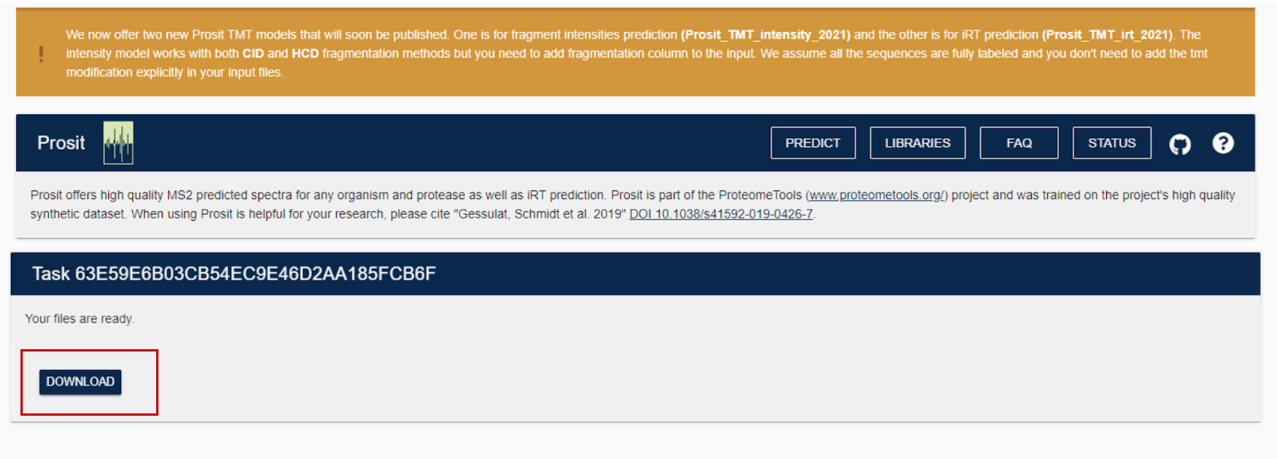
For Task ID Output format, select “NIST .MSP Text Format of individual spectra (Skyline and MSPepSerach compatible).



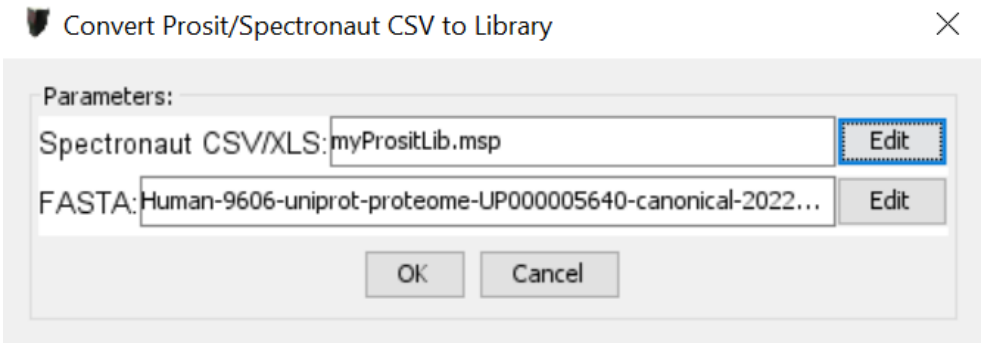
Submit the task. It is helpful to record the task number or bookmark the page shown after you submit the task to come back to once the job is complete.



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

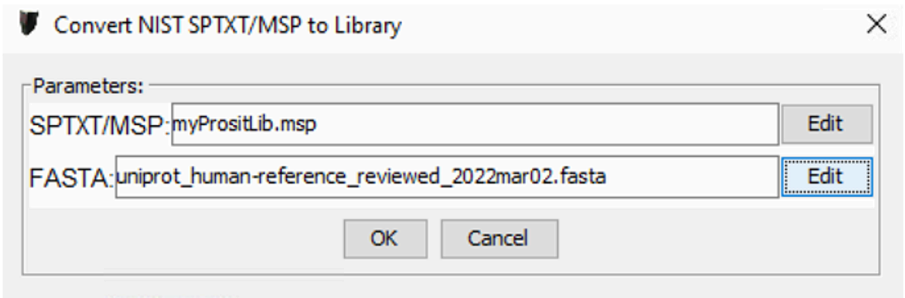


D. Convert the Prosit output (.MSP/NIST) to a Library (.DLIB) using EncyclopeDIA. At the top of the GUI, go to “Convert,” then navigate to “Convert SPTXT/MSP to Library.” This window will pop up.

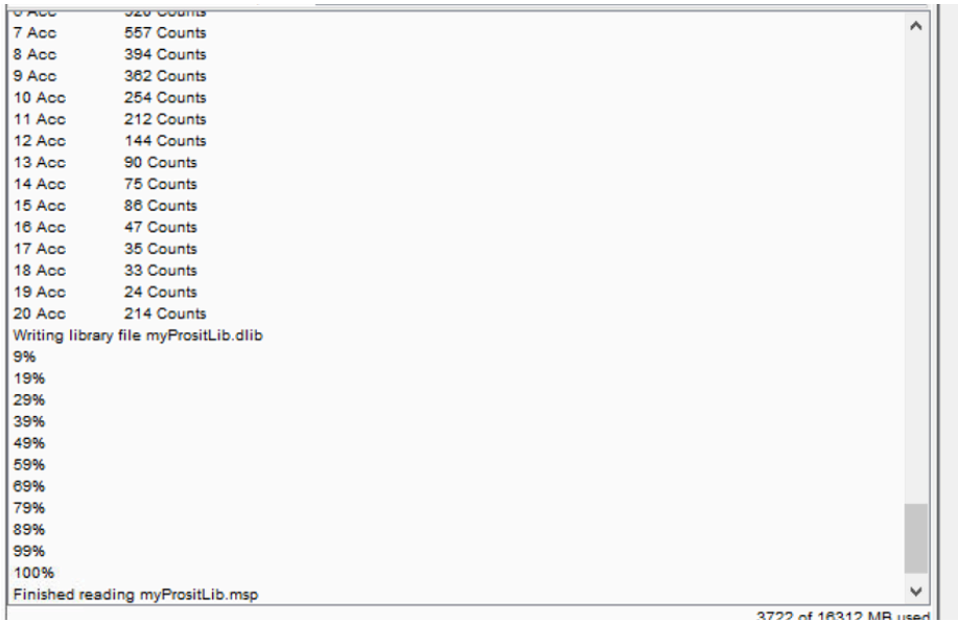


Upload the downloaded .MSP file from Prosit, and the FASTA file used to generate the .MSP.

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.

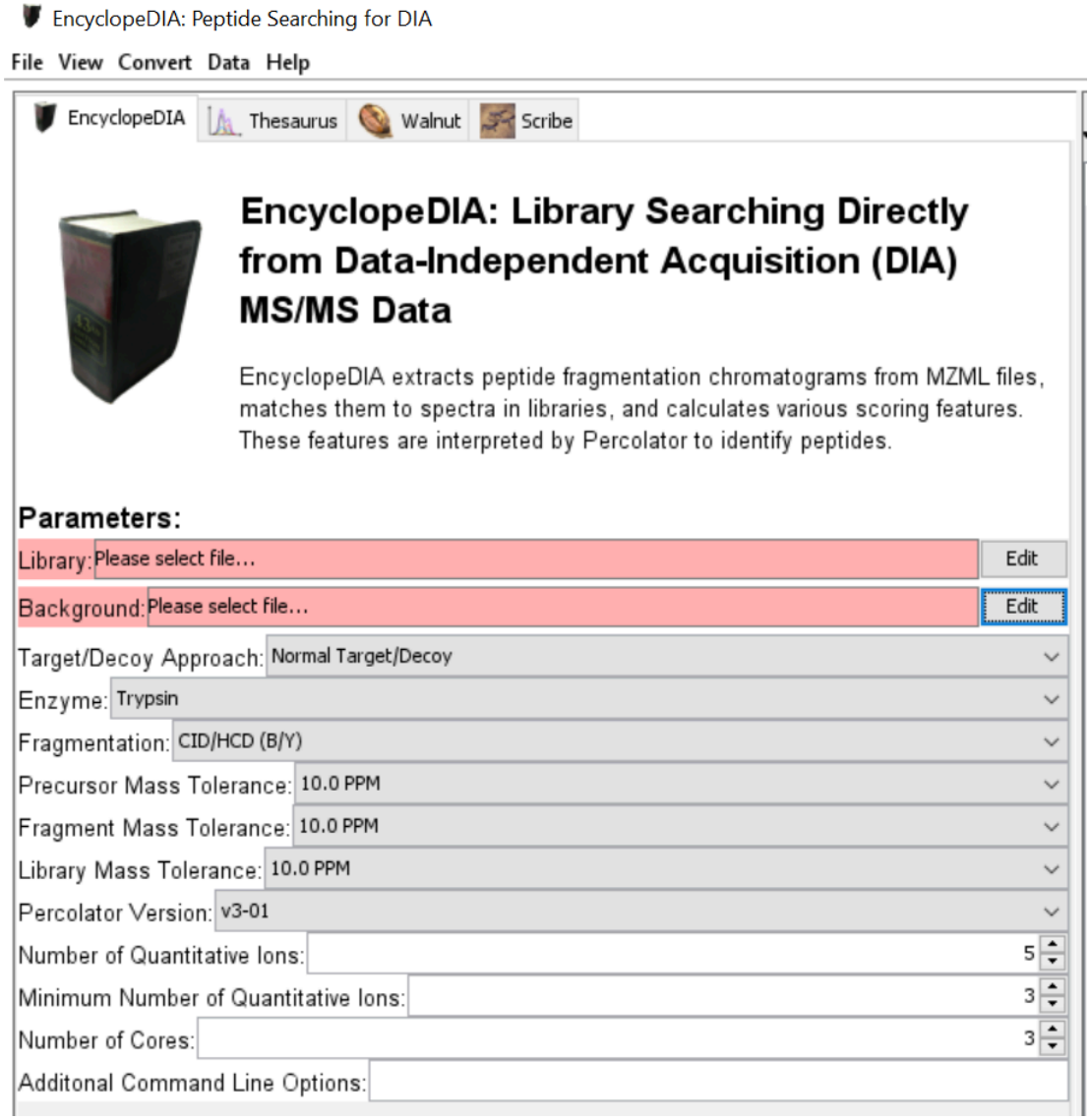


The resulting .DLIB should be analogous to human_prosit_generated_abrf.dlib.

4. ENCYCLOPEDIA SEARCH OPTIONS

Descriptions of the search options available in EncyclopeDIA.

Here is a screenshot of the EncyclopeDIA default search options:



A. EncyclopeDIA 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. EncyclopeDIA 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.

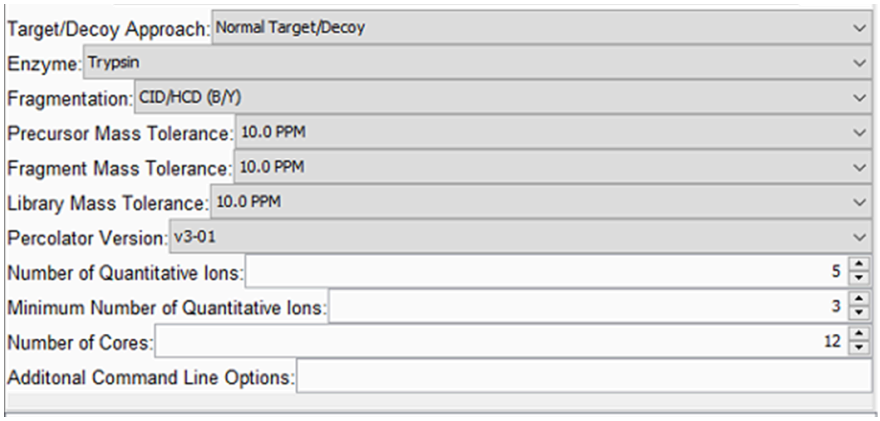
Fragmentation: In general, we recommend using CID/HCD (B/Y) fragmentation for most CID or HCD experiments. However, if your library is particularly large or messy you may get improved results with “HCD (y- only)”.
Precursor/Fragment/Library Mass Tolerance: Tolerances can be specified in PPM, AMU, or 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:

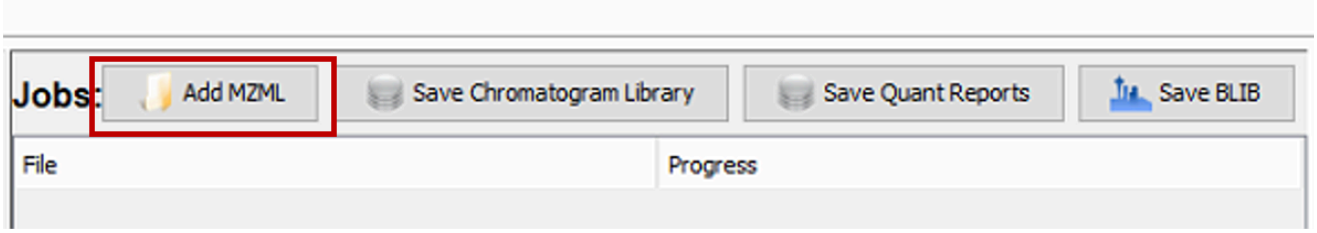
5. GENERATING A CHROMATOGRAM LIBRARY

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

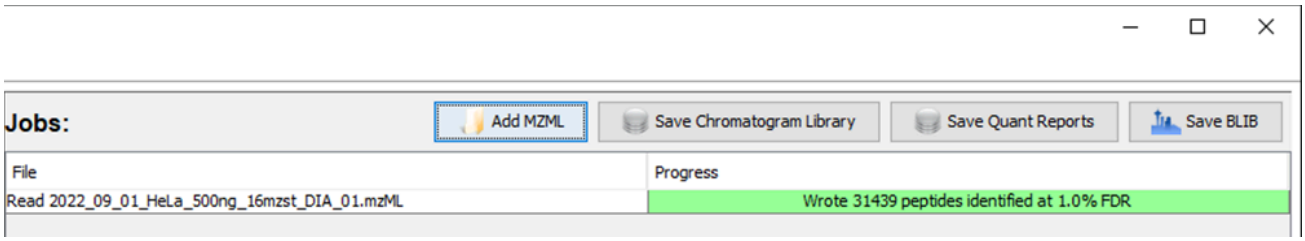
A. A chromatogram library, .DLIB, file is output from EncylopeDIA. With the latest version of EncyclopeDIA, you can use the acquired spectral library, or prosit-predicted library to search DIA injections. In this example, we are using 6 gas-phase fractionated injections of a pool of the 3 ABRF proteome mixtures. Open EncyclopeDIA, and upload “combined_abrf_6x_gpf.DLIB.” Specify “LakeTrout-Human-Cow-Contaminants-sPRG2022.fasta” as the background file. The settings should match the screenshot below.



Click “Add mzML,” then find the 6 DIA injections to run against the spectral library.



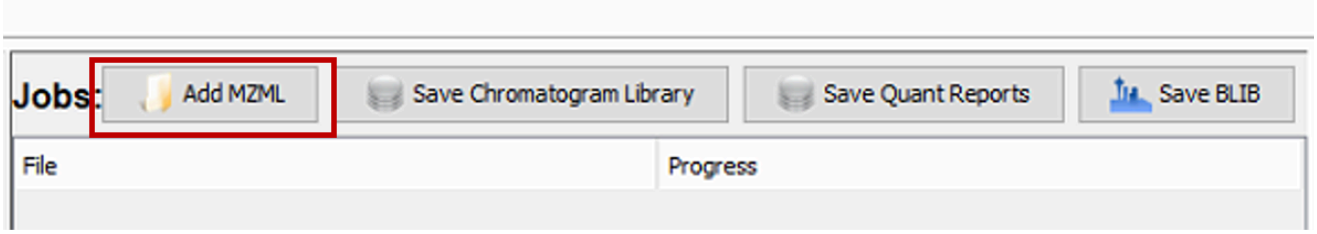
Once the search is complete, all 6 files will say the number of peptides detected. This is not exactly what your screen will look like, but representative



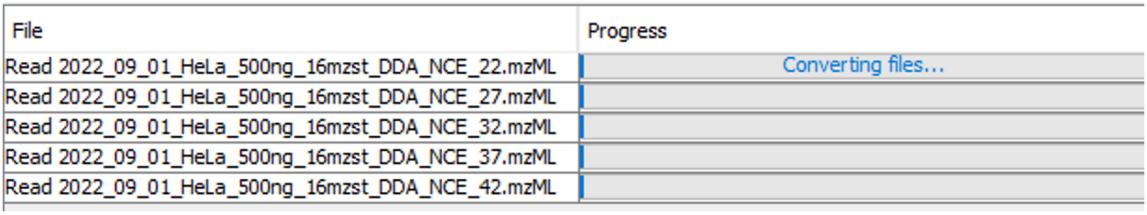
You can then visualize detected peptides as chromatograms using the “ELIB Browser.” The peptide below is an example of a peptide that was detected with little interference and multiple fragment ions.

6. SEARCHING SAMPLES AGAINST A CHROMATOGRAM LIBRARY

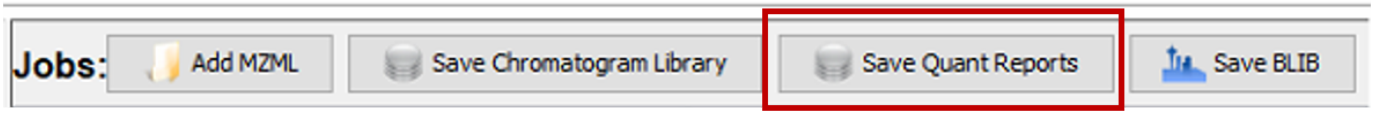
A. Use the “Add MZML” button to add RAW files for library searching. These will be automatically placed on the queue and executed in order using the current settings. If an MZML has been previously analyzed, the GUI will remember where it left off and try to not process it a second time.



Here is a screenshot of an EncyclopeDIA search queue. Again, these are representative, not what your screen will look like:

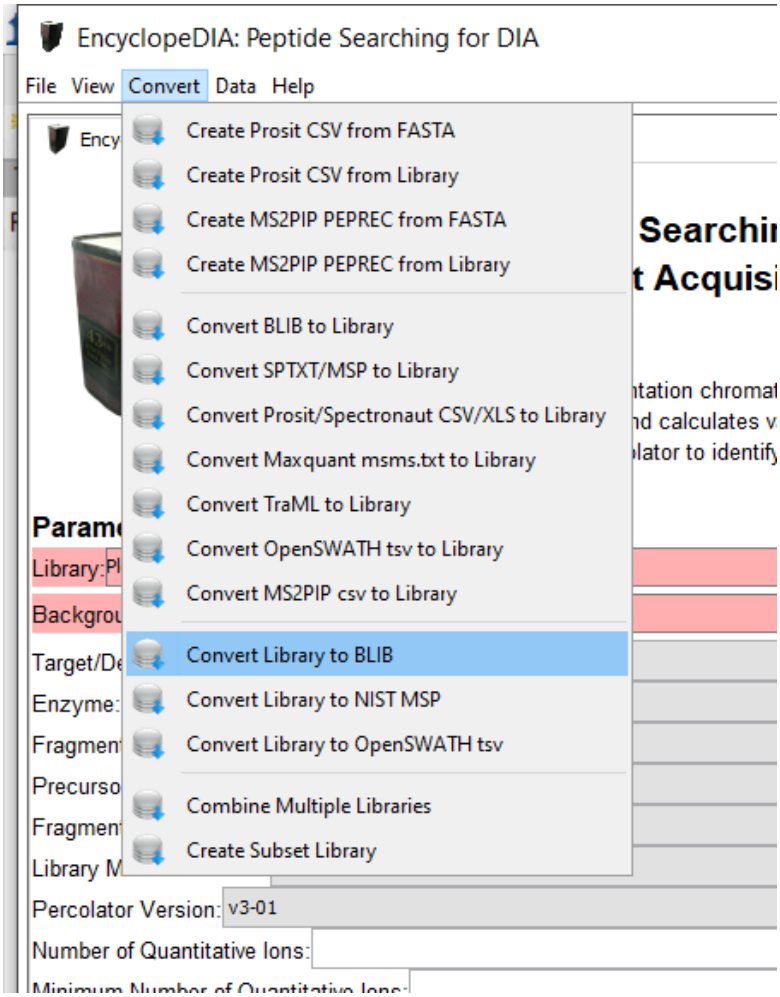


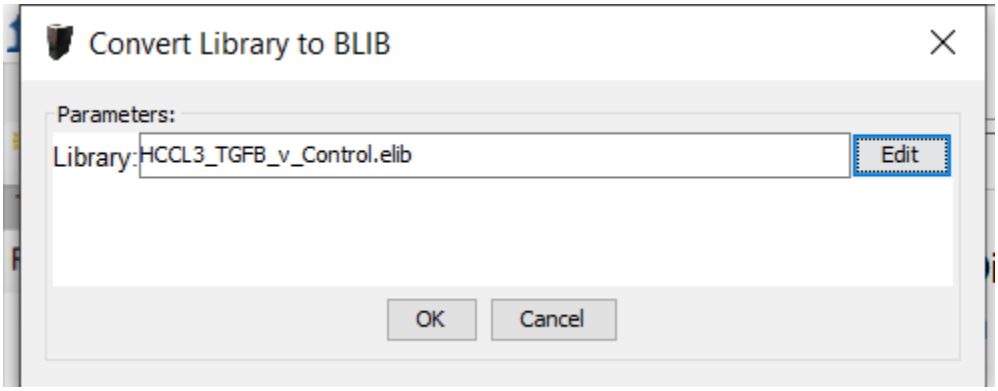
B. Search results can also be saved for downstream quantitative assessment using the “Save Quant Reports” button. RAW files between these experiments are expected to contain shared peptides so retention-time alignment is performed and match-between-runs quantification is calculated. This should generate a quant report (peptide.txt and protein.txt),



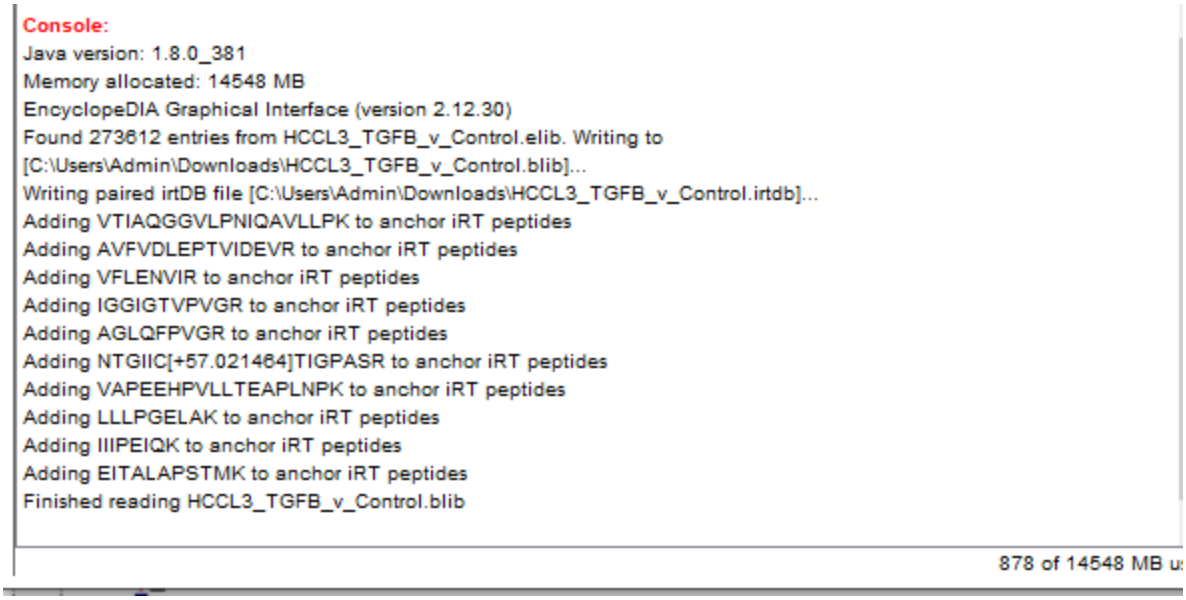
C. Additionally, save a BLIB version that we can put into Skyline. This should generate an indexed retention time database (.irtdb), and library (.blib) files.

D. (Alternative to C) Convert > Library to Blib and select the ELIB output from generating a quant report.





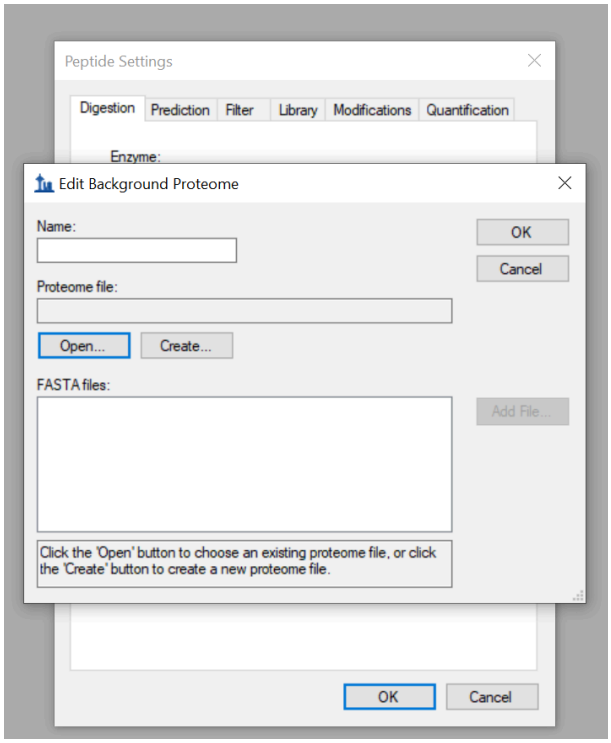
This will generate an .irtdb and .blib files.



The files you will obtain

7. IMPORTING DETECTIONS INTO SKYLINE

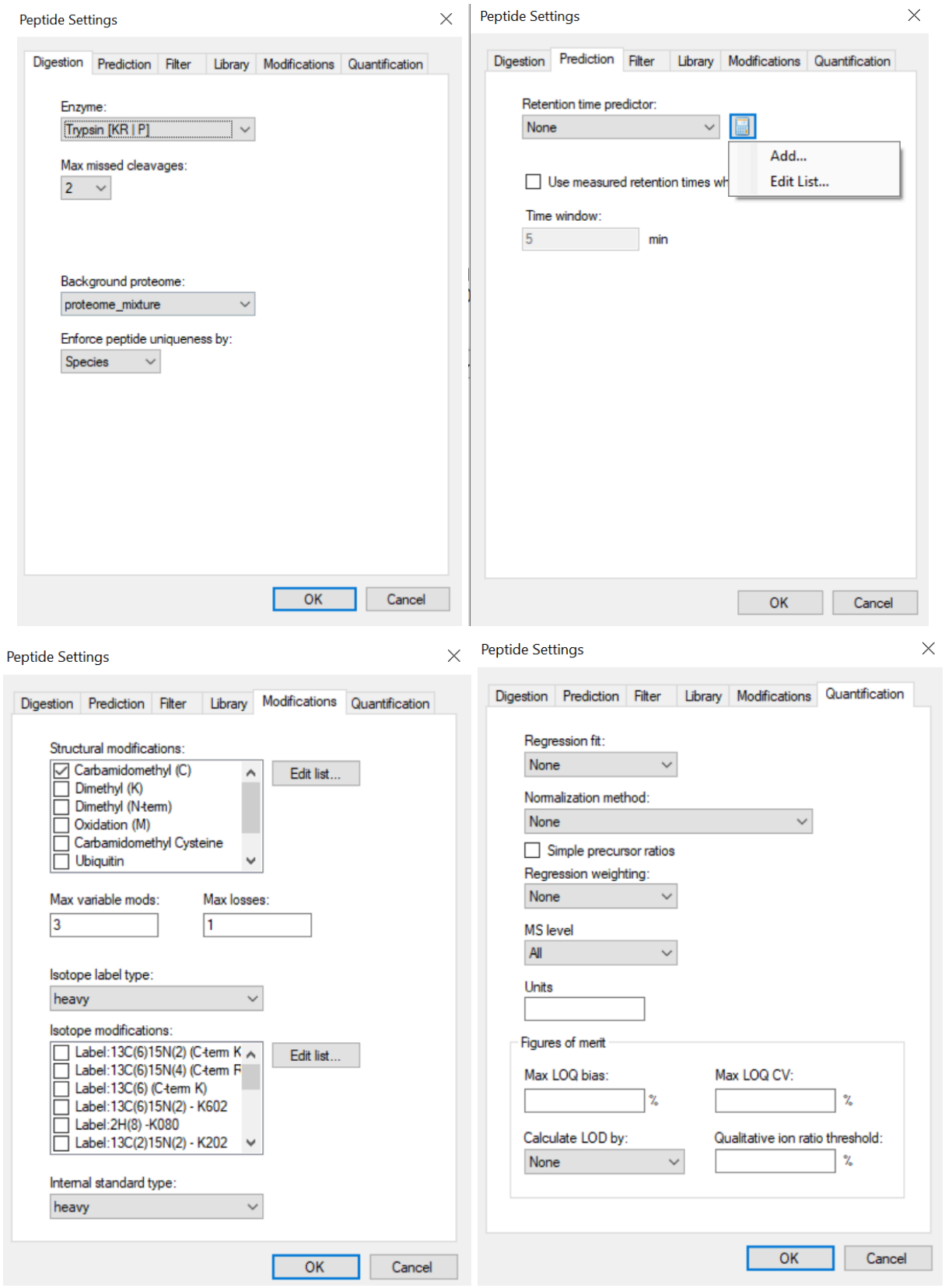
A. The first window that pops up will be the “Digestion tab.” We will add a background proteome. To do so, click the down arrow and select “add” to make a background proteome. You will see this window:



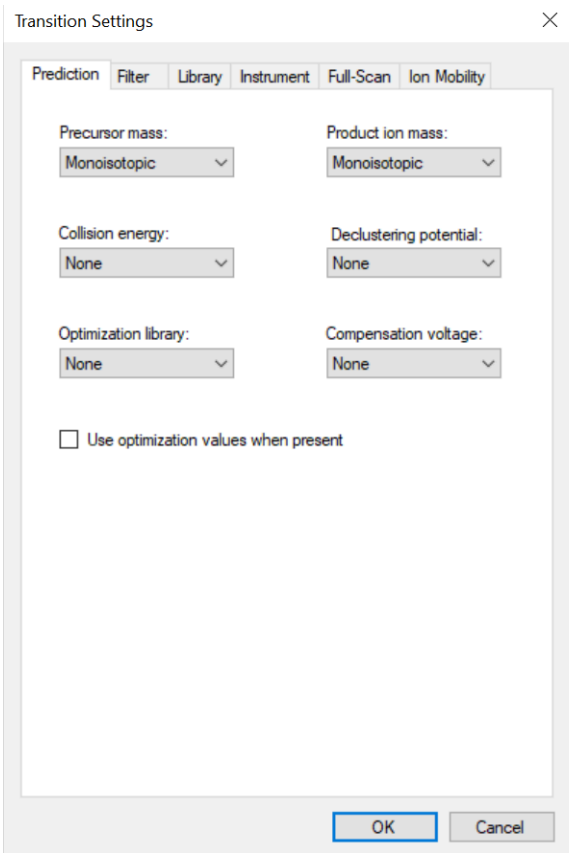
Click “Open…” and find the FASTA file named “Human-9606-uniprot-proteome-UP000005640-canonical-2022_04.fasta.” You will see proteins importing from the FASTA file.

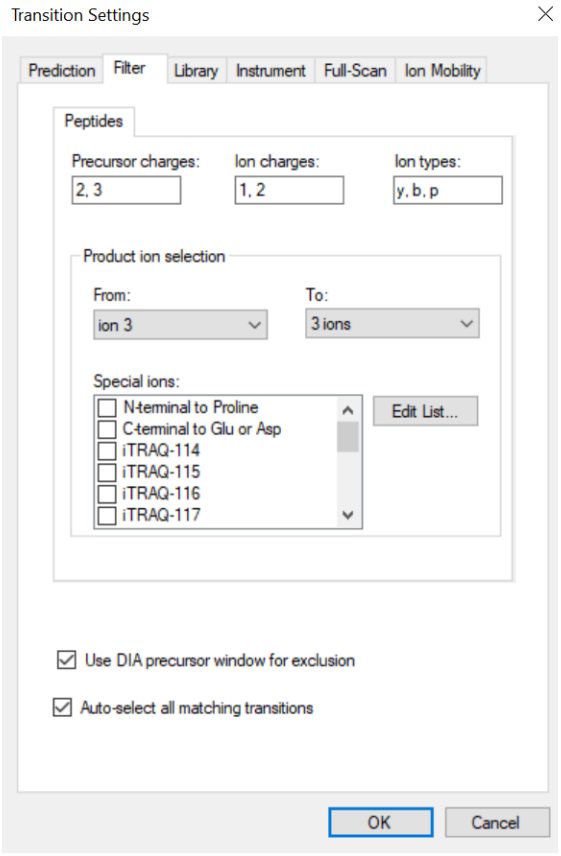
Once loaded, click “OK.” You will get a warning telling you that there are repeated peptide sequences. This is because the FASTA file was formed from separate FASTAs for each species, concatenated together. Proteins across species may share sequences. Click “OK” to accept the warning.

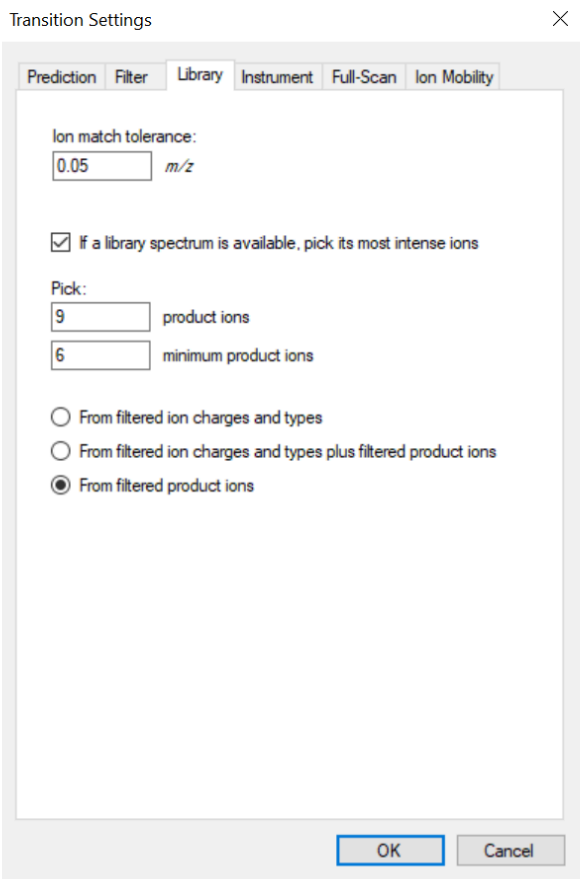
B. The rest of your Peptide settings will look as follows. Skip the Library tab for now, we will come back to that tab.

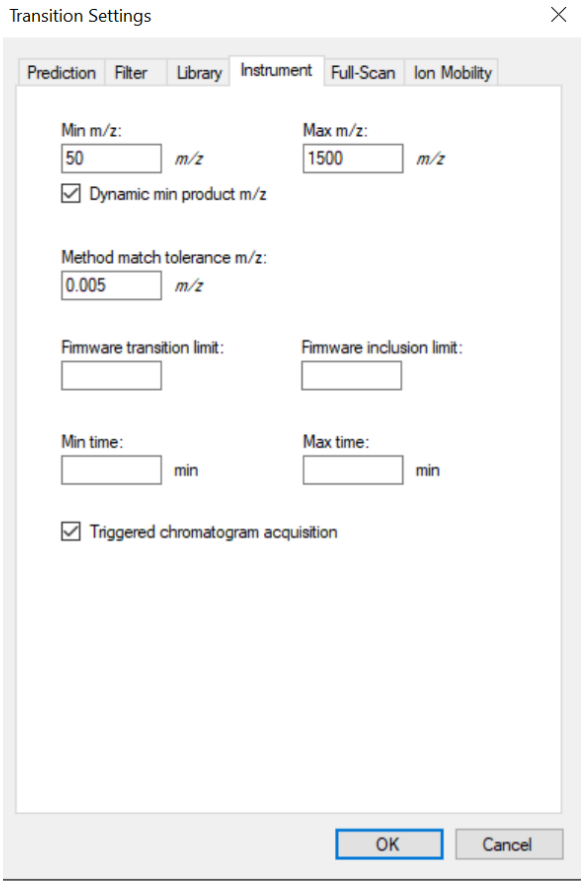


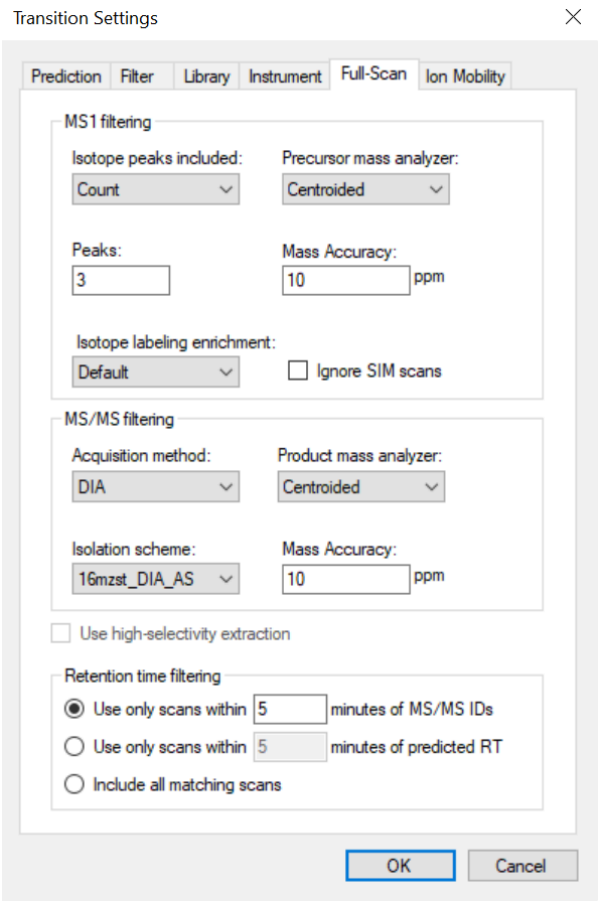
B. Set transition settings - set to as follows:


Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Ion Mobility
Precursor mass:
Monoisotopic
Product ion mass:
Monoisotopic
Collision energy:
None
Declustering potential:
None
Optimization library:
None
Compensation voltage:
None
Use optimization values when present
OK
Cancel

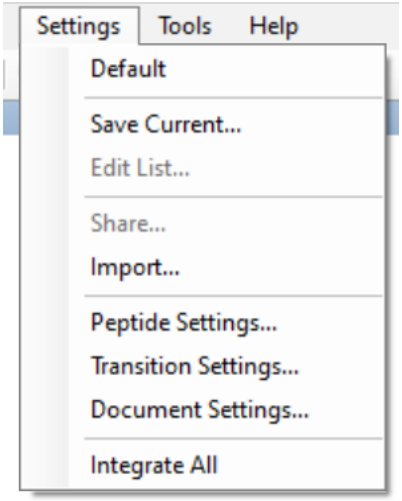

Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Ion Mobility
Peptides
Precursor charges:
2, 3
Ion charges:
1, 2
Ion types:
y, b, p
Product ion selection
From:
ion 3
To:
3 ions
Special ions:
N-terminal to Proline
C-terminal to Glu or Asp
iTRAQ-114
iTRAQ-115
iTRAQ-116
iTRAQ-117
Edit List...
Use DIA precursor window for exclusion
Auto-select all matching transitions
OK
Cancel


Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Ion Mobility
Ion match tolerance:
0.05
m/z
If a library spectrum is available, pick its most intense ions
Pick:
9
product ions
6
minimum product ions
From filtered ion charges and types
From filtered ion charges and types plus filtered product ions
From filtered product ions
OK
Cancel

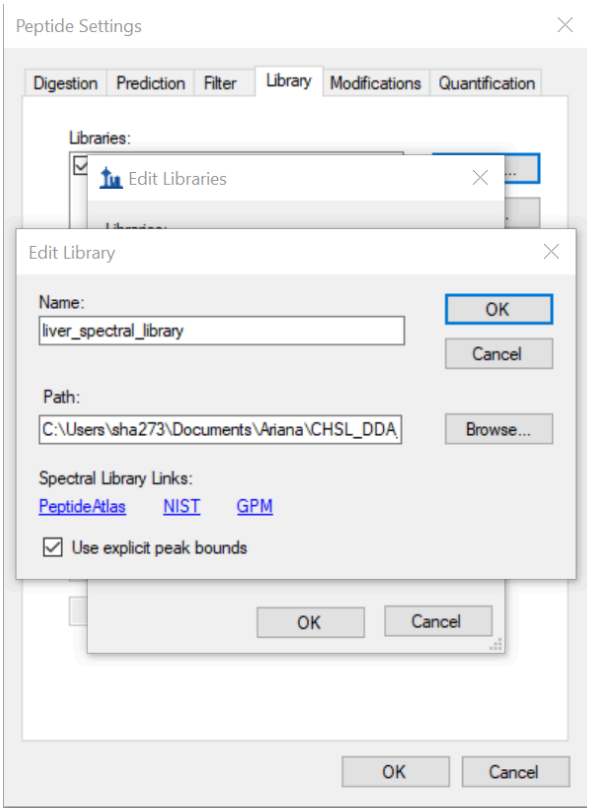

Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Ion Mobility
Min m/z:
50
m/z
Max m/z:
1500
m/z
Dynamic min product m/z
Method match tolerance m/z:
0.005
m/z
Firmware transition limit:
Firmware inclusion limit:
Min time:
min
Max time:
min
Triggered chromatogram acquisition
OK
Cancel



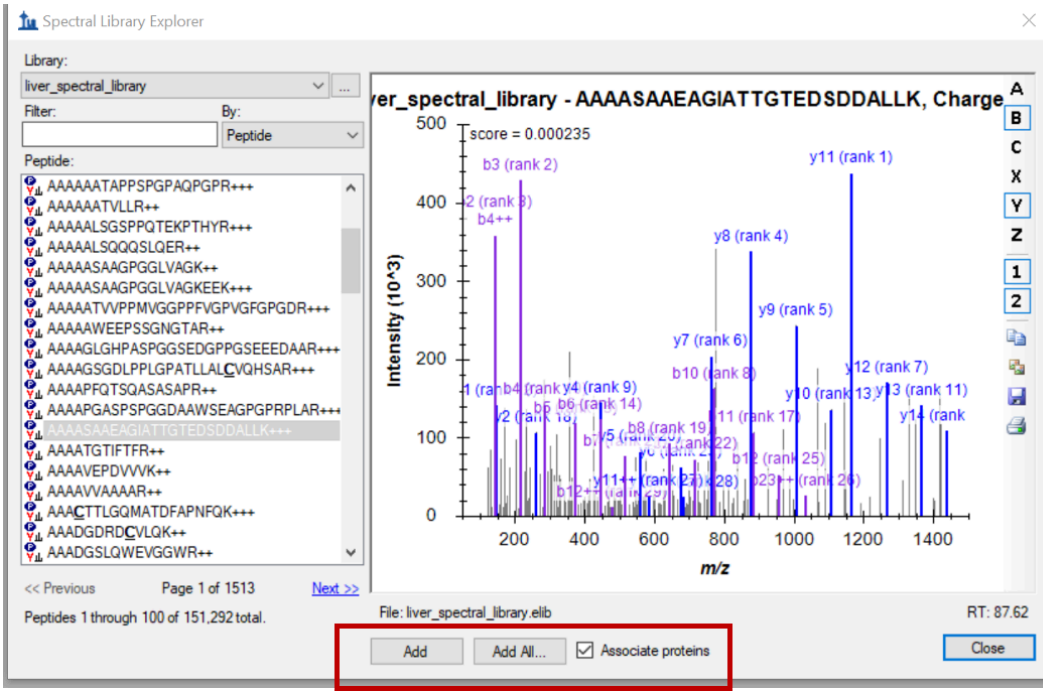
C. We will ignore the ion mobility window since we did not use any ion mobility with the data acquisition. Click OK to get back to your Skyline document. Next, go to Settings > Integrate all. This will check the integrate all and ensure that all peptides imported into the document will be integrated.



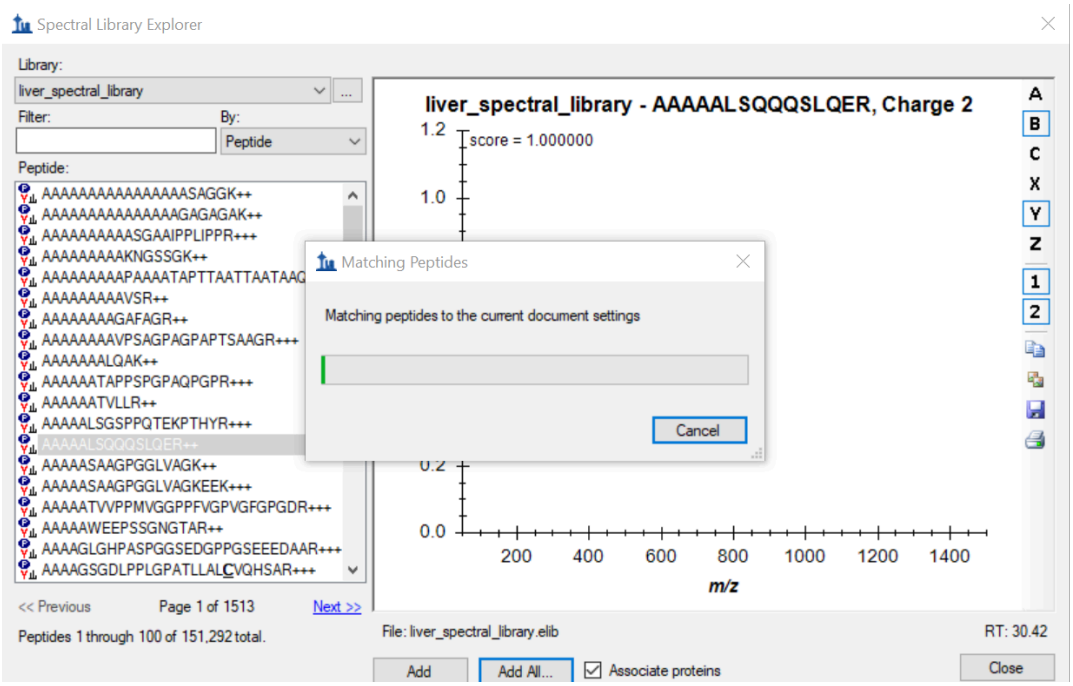
D. First, we will import the spectral library. Go to Peptide Settings > Library and click Edit List > Add. Then select Browse and find the “liver_spectral_library.blib” file. Name the library file “liver_spectral_library.”



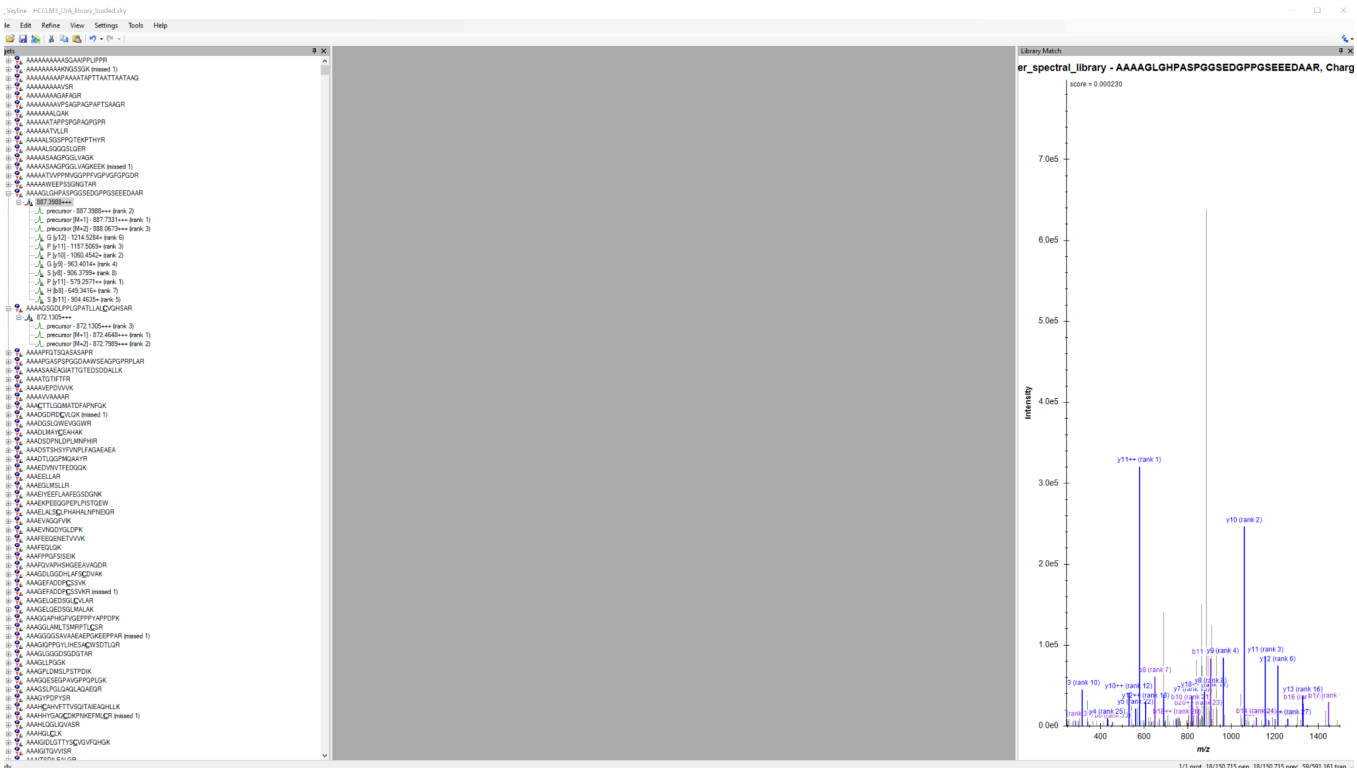
You will be taken back to the Library tab. Check the box next to the library you just imported, and select Explore. This will pop up the spectral library viewer



Check “Associate Proteins” and click the “Add All” button at the bottom of the Spectral Library Explorer window. This will add peptides from the spectral library to Skyline. The screen will look like this while is loading peptides according to the Transition and Peptide settings we have set.



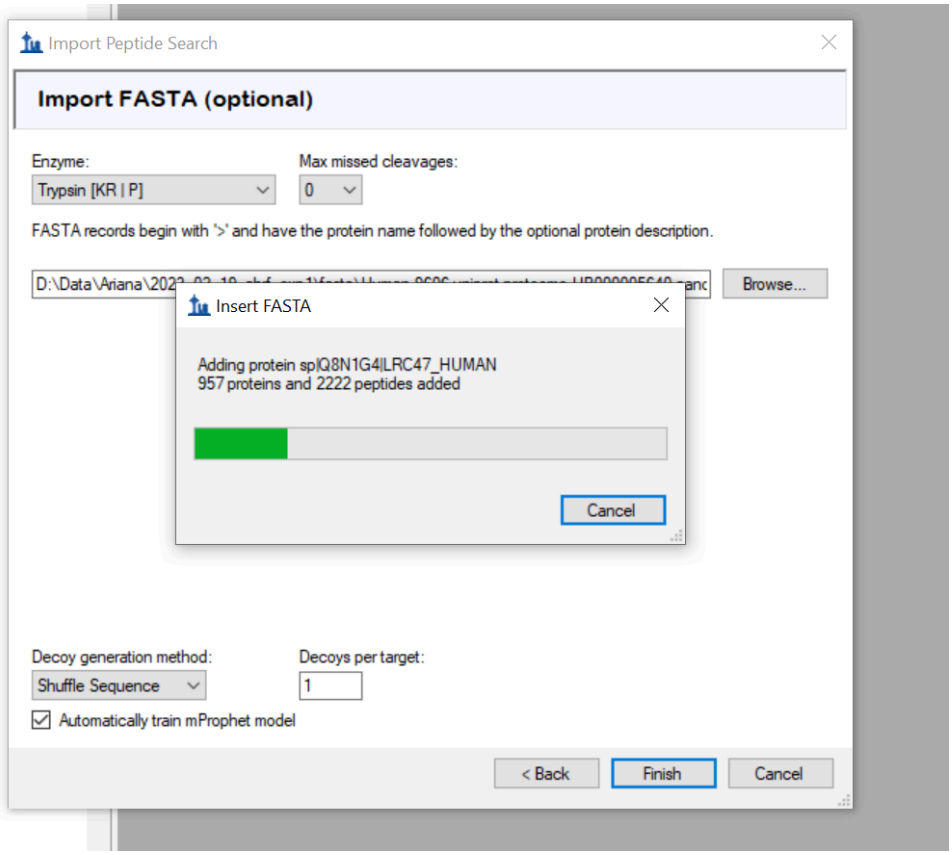
You will get a message that a portion of the peptides do not match the filter settings. We will add these for now, and then perform filtering downstream.



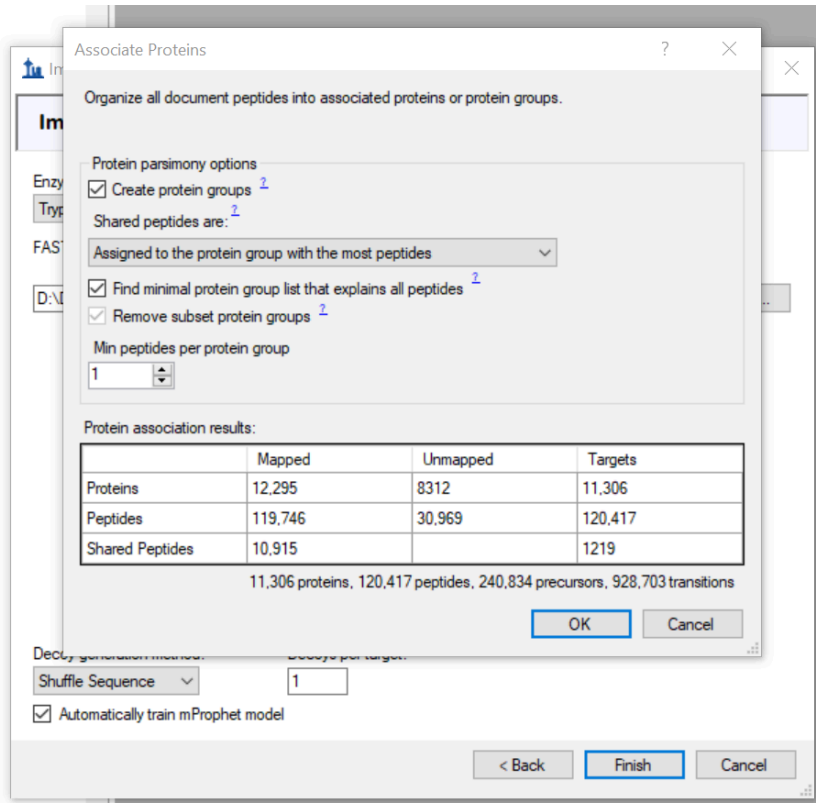
E. Go to File > Import Peptide Search. The first window that pops up will give you the option to generate a library or make a pre-made library. We will select a pre-made library and click on DIA. Find the .BLIB that we printed after the quant report that was generated from EncyclopeDIA.

F. Next, select the .mzML files used in the search of our individual samples.
    a. EMT_R1.raw
    b. EMT_R2.raw
    c. EMT_R3.raw
    d. Control_R1.raw
    e. Control_R2.raw
    f. Control_R3.raw

G. Your settings should be set to reflect the following for this search:
H. Find the Human FASTA we set as the background proteome, and select this for the FASTA. Ensure the enzyme is set to trypsin, the decoy generation method is Shuffle, and the number of Decoys is set to 1. Automatically train mProphet model should also be checked.



To associate proteins, we will check the box to “Create protein groups,” and click the drop-down menu to set Shared peptides to be “Assigned to the protein group with the most peptides,” and check “Find minimal protein list to explain all peptides”, which will also select the “Remove subsetted proteins” box. In general, we want to filter out duplicated peptides/peptides that match more than one proteins.



I. You will see your data importing, and once it is in Skyline, an mProphet module will pop up. Train the mProphet according to the decoys generated or second best peak.
J. Your loaded document should contain the full list of peptides.