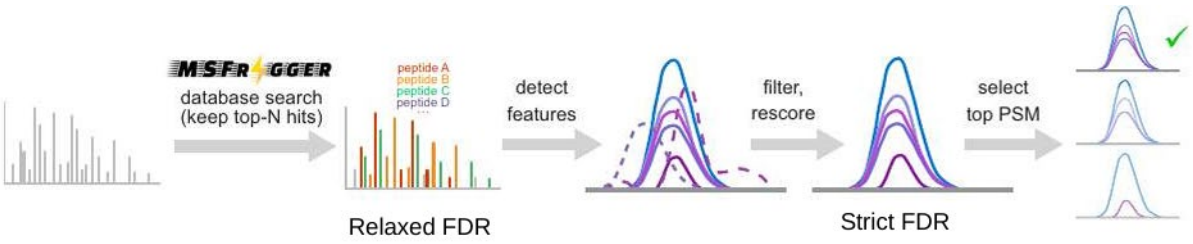
Skyline Webinar 26

Spectrum-centric analysis of DIA datasets

In this tutorial, we will conduct an untargeted spectrum-centric analysis of a data-independent acquisition (DIA) dataset utilizing the FragPipe computational tool collection. Our focus will be on a subset of samples from a larger study that compares the plasma proteomes of older and younger healthy individuals. Specifically, we prepared pooled plasma samples from five old (n=5) and five young (n=5) healthy participants. These pooled plasma proteomes were processed using the ENRICH-iST digestion kit from PreOmics, which enriches low-abundance proteins onto paramagnetic beads, effectively addressing the dynamic range challenge in plasma prior to analysis. The digested samples were subsequently analyzed with an Orbitrap Astral employing a DIA strategy, using 2-Da isolation windows across a mass range of 380 to 980 m/z. These analyses were performed with a 50-cm µPAC HPLC column and a chromatographic gradient of 30 SPD, approximately 38 minutes in duration.

We will use FragPipe for these analyses, which is an open-source Java Graphical User Interface (GUI) for a suite of computational tools enabling comprehensive analysis of mass spectrometry-based proteomics data. Fragpipe is powered by MSFragger, an ultrafast proteomic search engine suitable for both conventional and open (wide precursor mass tolerance) peptide identification. FragPipe includes the Philosopher toolkit for downstream statistical post-processing of MSFragger search results (PeptideProphet, iProphet, ProteinProphet), FDR filtering, label-free and label-based quantification, and multi-experiment summary report generation. The software is well documented (https://fragpipe.nesvilab.org/) and the original publication is Yu, F et al. (2023), Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nature Communications*, *14*(1), 4154.

In this tutorial, we will combine the MSFragger module with DIANN for direct analysis of data independent acquisition (DIA) data. We will first process the data with MSFragger to identify multiple peptides in chimeric spectra, then statistically validate the identification results with Percolator, and finally perform peptide quantification with DIA-NN. Once we get the identification and quantification results from FragPipe, we will load them in FragPipe-PDV to visualize the identifications, and we will perform some downstream analysis using FragPipe-Analyst. Finally, we will load the raw data in Skyline to see the extracted ion chromatograms for each of the identified peptides.



Yu, Fengchao, et al. "Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform." Nature Communications 14.1 (2023): 4154.

Parametrization of FragPipe graphical user interfase

In this first part of the tutorial, we will set up the graphical user interface of FragPipe and launch a spectrum-centric search combining MSFragger and DIA-NN. The result will be a collection of matrices with the quantification values at the precursor and protein levels, as well as a summary pdf file of the experiment.

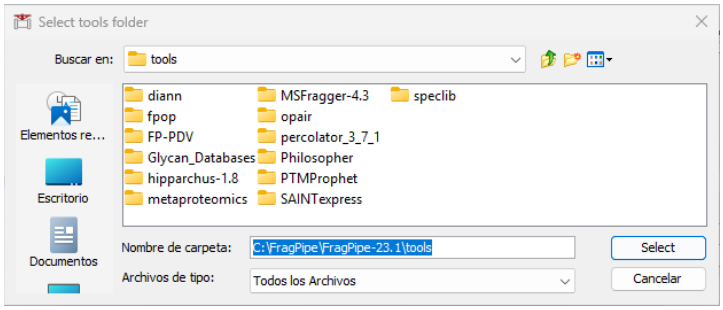
Note: This tutorial is based on FragPipe 23.1.

- Go to C:\FragPipe-23.1\bin\
- Click in the FragPipe-23.1.exe icon to open the graphical user interface.

Parametrization of the *Config* section

In this section we need to make sure that all the different tools that are required by FragPipe are installed in the system and provide FragPipe with the path to the corresponding executables.

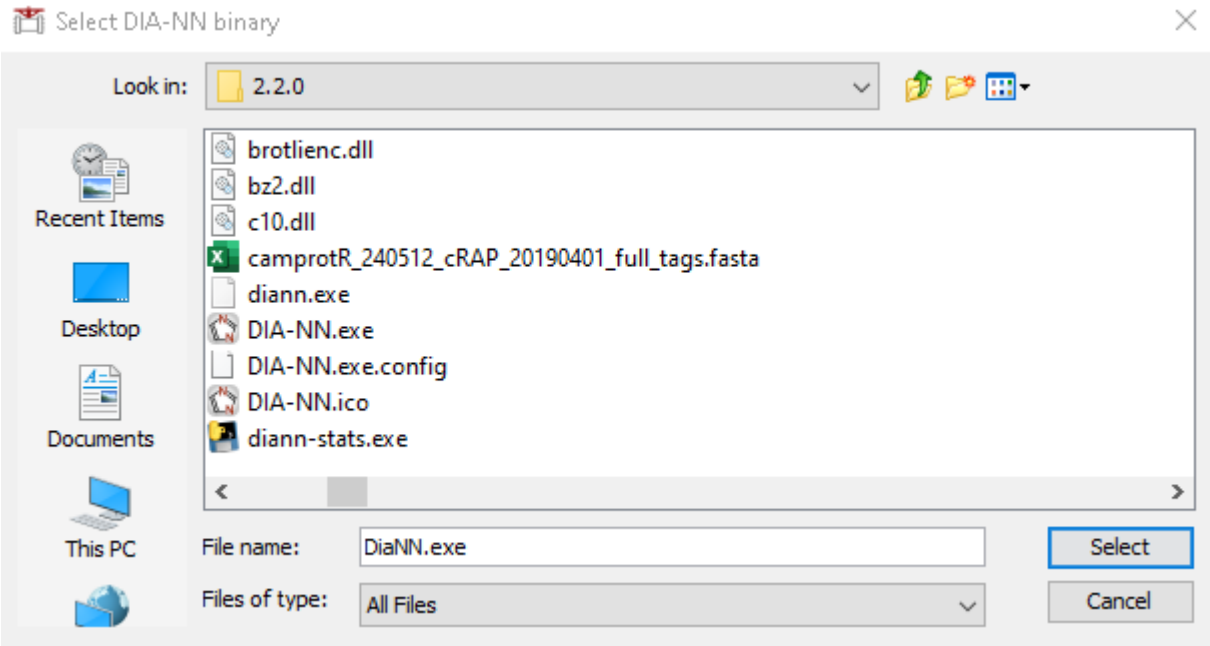
- Select the `Config` tab from the graphical user interface of FragPipe.



- Go to the "MSFragger, IonQuant, diaTracer" section below and click "Browse". Navigate to C:\FragPipe-23.1\tools and click "Select".

Note: In case you don't have the MSFragger, IonQuant and diaTracer files in your current installation, you should click to "Download/Update" for automated installation of these three programs. These programs have specific academic licenses that need to be accepted in their first installation to be used (see below). Register, agree with the terms of use and enter the "Verification Code" that you will receive by email.

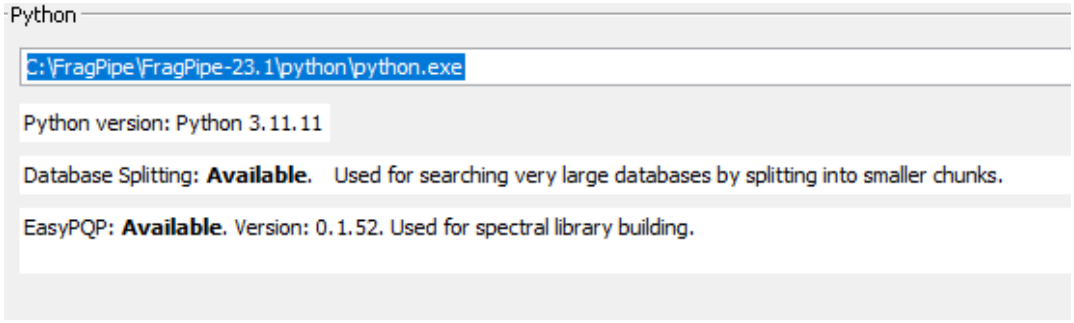
- Go to the DIA-NN section below and click "Browse". Navigate C:\DIA-NN\2.2.0\diann.exe and click "Select".



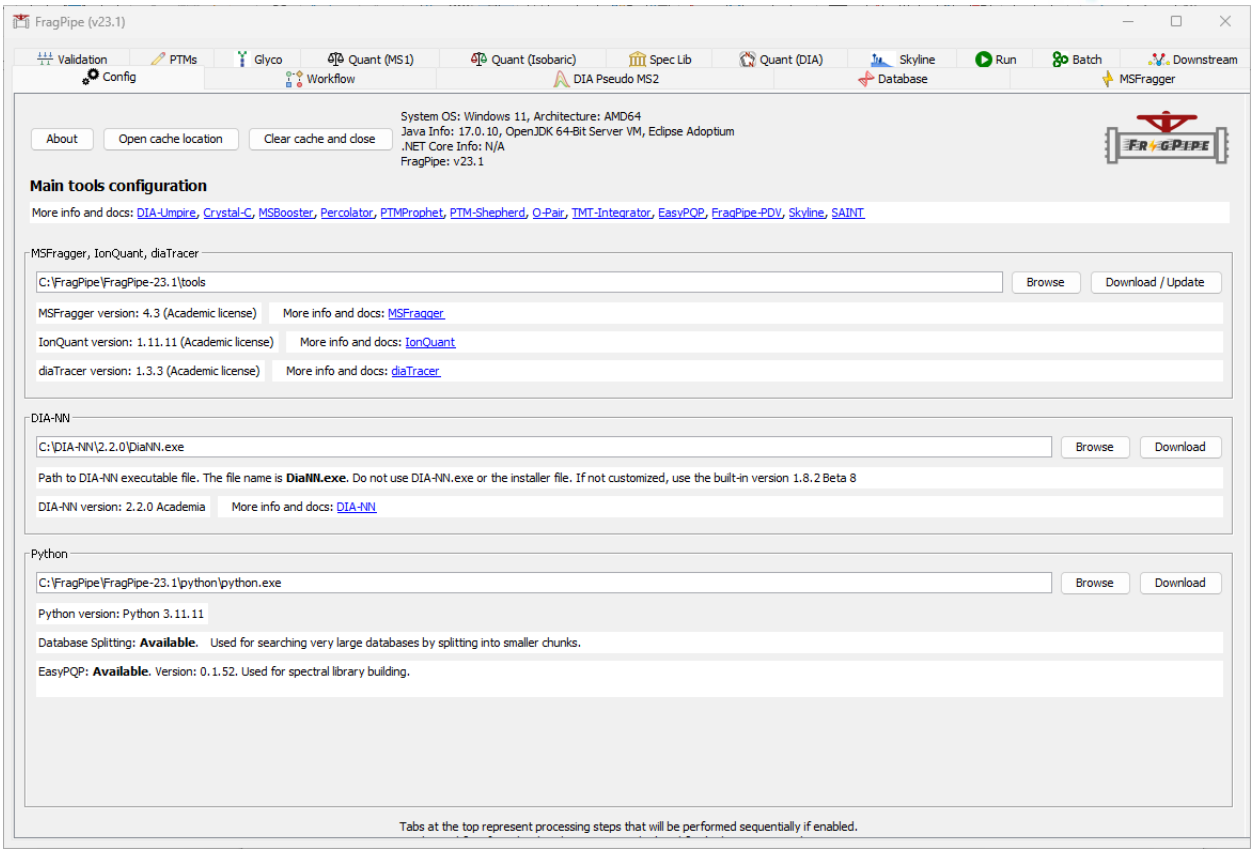
Note: If you don't have DIA-NN installed from previous tutorials, you can download it from the DIANN github repository https://github.com/vdemichev/DiaNN/releases or select the DIA-NN version 1.8.2 shipped together with FragPipe and that can be found at "C:\FragPipe\FragPipe-23.1\tools\diann".

- Go to the Python section and make sure that the correct path "C:\FragPipe\FragPipe-23.1\python\python.exe" is defined and points to Python version 3.11.11. Otherwise, click "Browse" and select "C:\FragPipe\FragPipe-23.1\python\python.exe".

- Check that the "Data Splitting" and "EasyPQP" modules are correctly installed and appear as "Available".



Your `Config` tab should look like this:



Parametrization of the *Workflow* section

FragPipe supports multiple proteomics workflows, which can be customized, saved, and shared with other users.

In the `Workflow` tab:

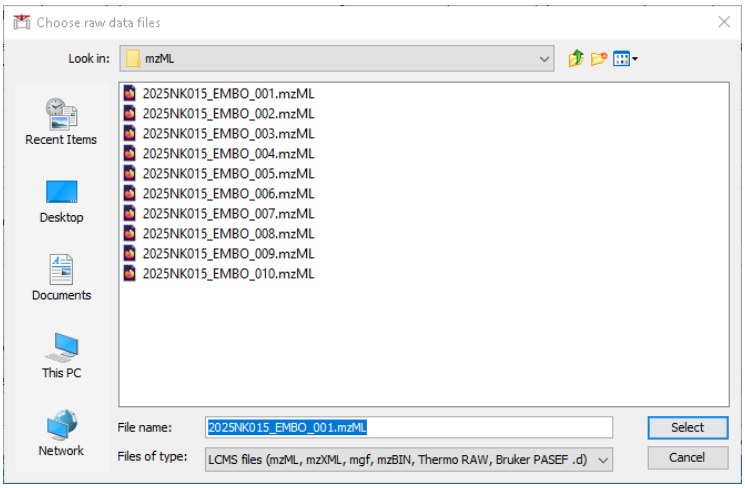
- Choose the workflow “DIA_SpecLib_Quant” workflows, which corresponds to the DIA spectral library generation and quantification using MSFragger and DIA-NN.
- In Global settings, set the amount of RAM memory to zero. A RAM setting of 0 will allow FragPipe to automatically detect available memory and allocate a safe amount.
- In the “Parallelism” you can select the number of logical cores to use. Set this to the number of cores that your computer has minus one.

Note: You can check the number of logical cores by opening the "Task Manager" application in Windows and navigating to the "Performance" tab. Here, you will find the number of virtual cores available on your computer.

In “Input LC-MS Files” section we will load and annotated all the `.mzML` files that contain the raw mass spectrometry data acquired in the experiment.

- Check “Regular MS” as the MS data type. Note that the option ‘IM-MS’ is meant only for Bruker timsTOF PASEF data whereas ‘Regular MS’ is meant for all other data types (including FAIMS).

- Click “Add files” and navigate to “Webinar26\mzML” and select the 10 `.mzML` files. Click “Select”.



Note: The original `.raw` files were converted to the mzML open format and trimmed to keep only the first five minutes of each run with the purpose to optimize the file size and data analysis timings for this tutorial. This conversion was performed with the msconvert software that can be obtained from the Proteowizard web page (https://proteowizard.sourceforge.io/download.html).

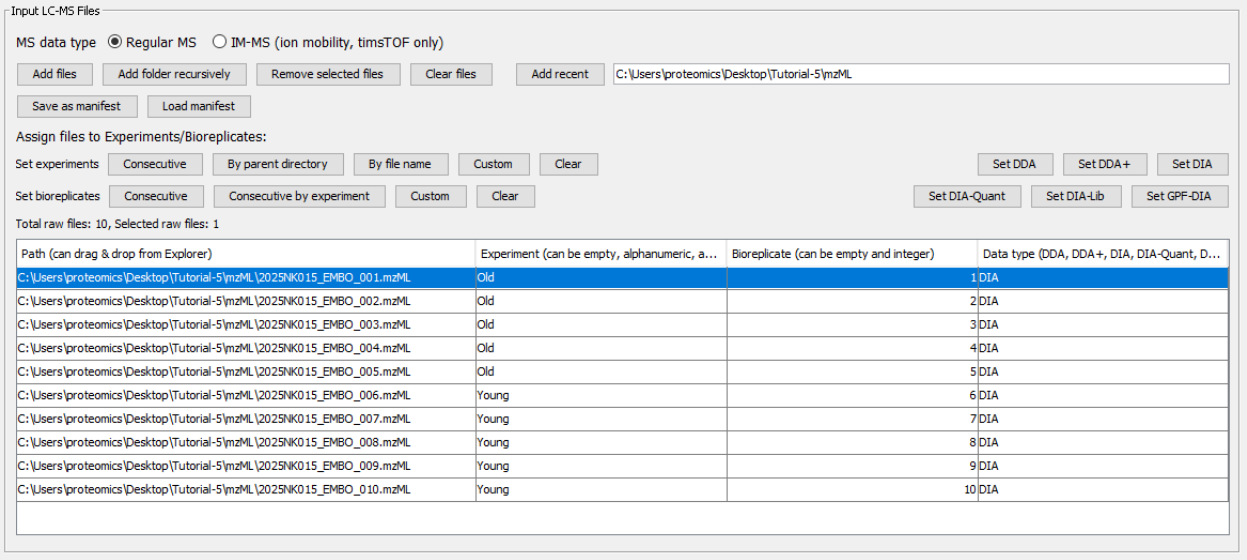
The command used was `msconvert.exe --32 --filter "scanTime [0,300]" --filter "peakPicking vendor msLevel=1-"`

Table 1: File names and conditions

File Name	Experiment	BioReplicate	Data Type
2025NK015_EMBO_001.raw	Old	1	DIA
2025NK015_EMBO_002.raw	Old	2	DIA
2025NK015_EMBO_003.raw	Old	3	DIA
2025NK015_EMBO_004.raw	Old	4	DIA
2025NK015_EMBO_005.raw	Old	5	DIA
2025NK015_EMBO_006.raw	Young	6	DIA
2025NK015_EMBO_007.raw	Young	7	DIA
2025NK015_EMBO_008.raw	Young	8	DIA
2025NK015_EMBO_009.raw	Young	9	DIA
2025NK015_EMBO_010.raw	Young	10	DIA

Now we need to annotate the “Experiment” and the “Bioreplicate” associated to each raw file according to the information provided in Table 1.

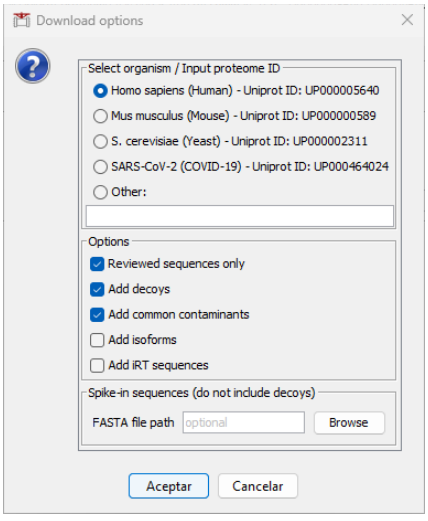
- First select all files in the table and click “Set DIA” to inform FragPipe that the raw files provided were acquired in DIA mode.
- Then, select all the specific raw files the “Control” key to select multiple rows simultaneously, and in the “Set experiments” section click “Custom” and type the condition “Young” or “Old” according to Table 1.
- For each file (row) select “Consecutive” in the “Set Bioreplicates” section.



Parametrization of the *Database* section

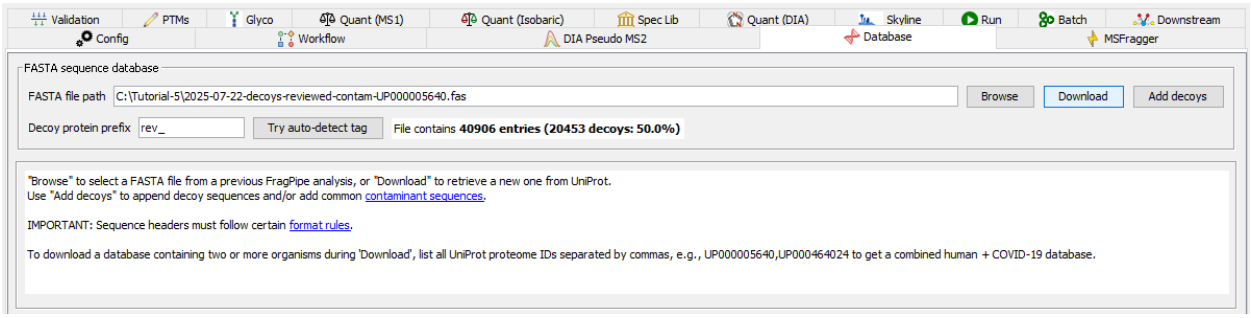
We will skip the `DIA Pseudo MS2` tab as it is not meant to be executed in the selected workflow, and move directly to the `Database` tab.

- Click “Download” to retrieve a fresh UniProt human database including reviewed sequences only, plus contaminants and decoys.



- Save the fasta in the Webinar26 folder.

Note: Alternatively, one can also use the `fasta` file database provided in the "Webinar26\" folder with the name `2025-07-22-decoys-reviewed-contam-UP000005640.fas`

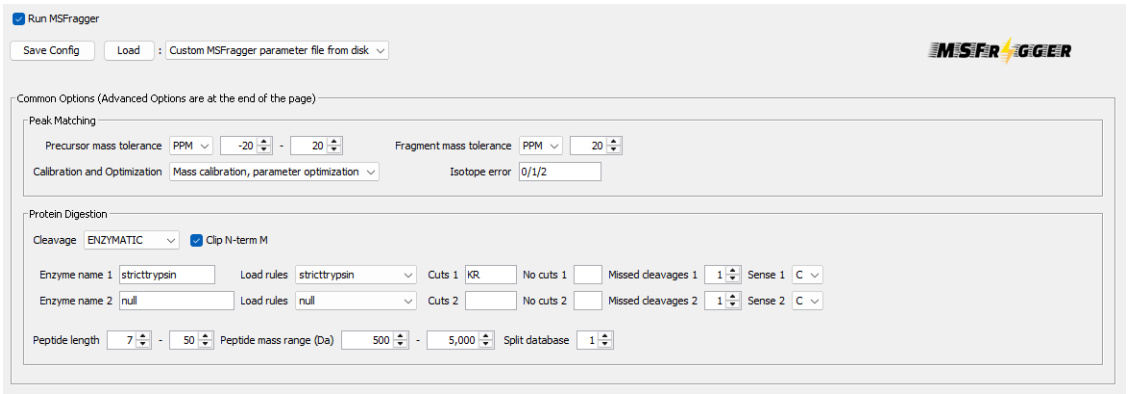


Parametrization of the *MSFragger* section

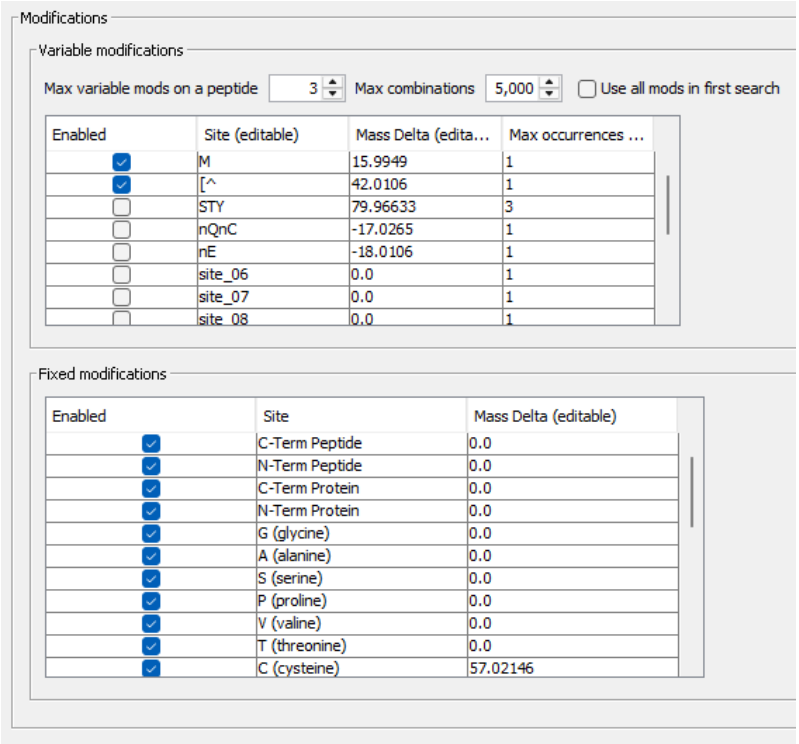
In the `MSFragger` tab you can check the search parameters that will be used to interpret the spectra acquired in our analysis. The parameters have already been filled with the default values associated to the workflow selected. Let's review them.

- In **Peak Matching** the precursor and fragment mass tolerances are specified. In this case, it is set to 20 ppm, which is a standard value in a so-called closed data search using data acquired in a high-resolution mass spectrometry analyzer. Options on whether there is the need for automatic mass calibration, parameter optimization and isotope peak selection correction are also present.
- In **Protein Digestion** we define that an enzymatic digestion was used to prepare the sample, which in this case was trypsin. The range of peptide length and mass range allowed are also specified in this section. All these parameters are important when interpreting the spectra *in silico* because they define the set of potential peptides that can be present in the sample.

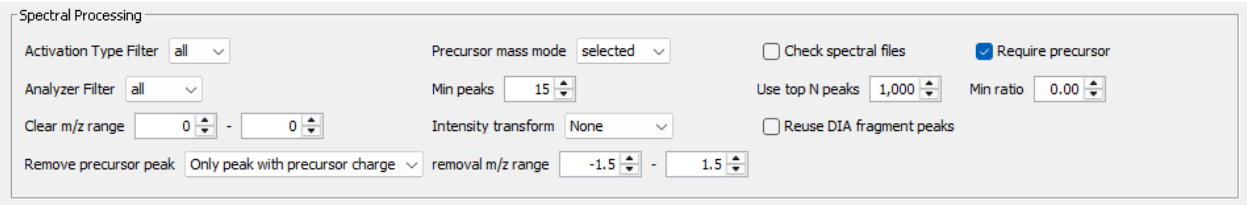
Note: *Calibration and Optimization* is set by default to "Mass Calibration, Parameter Optimization". This option will effectively perform multiple simplified MSFragger searches with different parameters to find the optimal settings for your data. In practice, it results in 5-10% improvement in the number of identified PSMs at the expense of increased analysis time. To save time, you can consider changing this option to "Mass Calibration" or even "None", especially if you already know your data (e.g. from previous searches of the same or similar files) and can adjust the corresponding MSFragger parameters (fragment tolerance, number of peaks used, intensity transformation) manually, if needed.

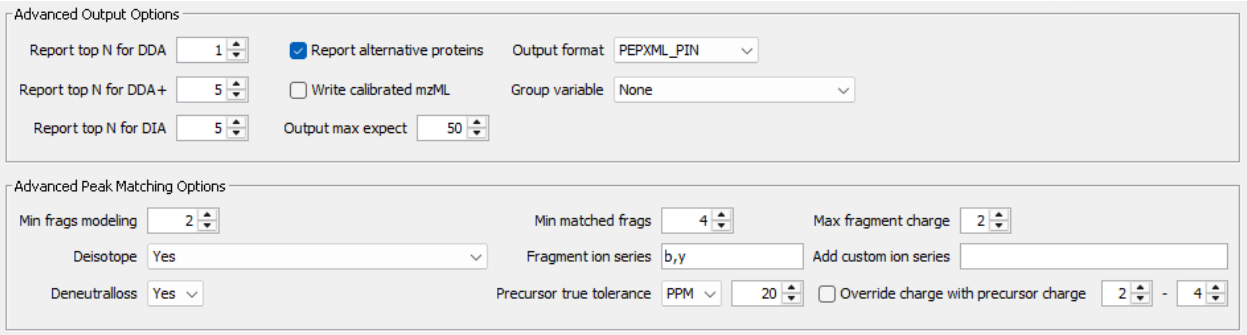


- In **Modifications** both variable and fixed modifications that can be found in the analysed peptides are specified, as well as the maximum number of allowed modifications per peptide and the maximum number of occurrences per single modification.



- Finally, the section of **Advanced Options**, which encompasses several key parameters, including *Mass Offsets*, which allow for the introduction of specific fixed delta masses to be considered in a restricted open search; *Glyco/Labile Mods*, which enable the inclusion of specific diagnostic fragment masses for identifying glycans and other labile modifications; and *Spectral Processing*, which determines the number of peaks considered for interpreting each spectrum. Additionally, this section includes *Open Search Options*, which are not activated in this DIA analysis, as well as *Advanced Output Options* that define the output format and related settings. Finally, the *Advanced Peak Matching Options* specify the number of top N peptides to be considered for identification per spectrum, detail the fragment ion series, establish the fragment charge range, and set the minimum number of matched fragments required during the search.



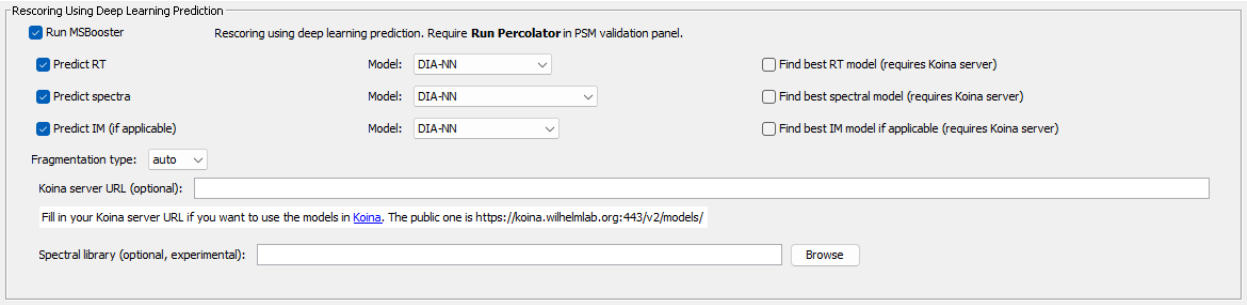


You can choose to save a customized parameter file to load for future use, or save the entire workflow (from either the 'Workflow' or the 'Run' tab).

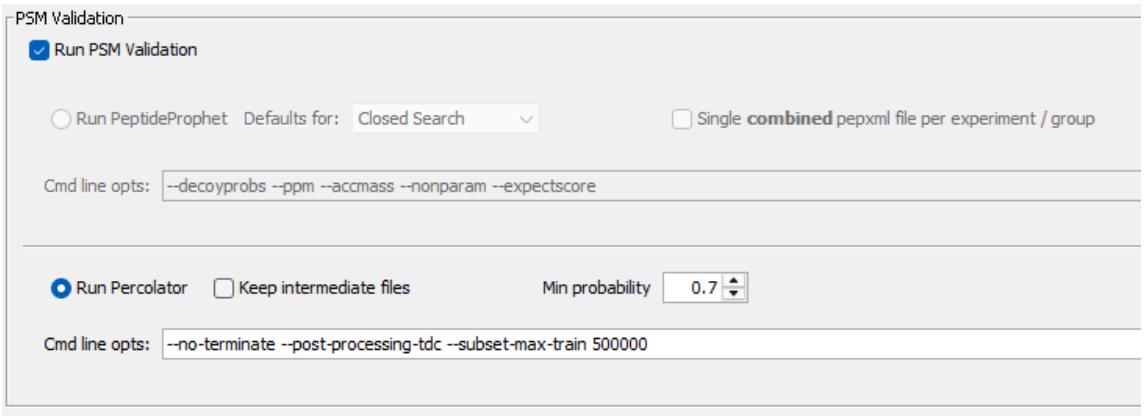
Parametrization of the *Validation* section

The `Validation` section will also be executed as part of the selected workflow. The search results obtained from MSFragger will be further analyzed by MSBooster, Percolator and ProteinProphet to get confident peptide identifications.

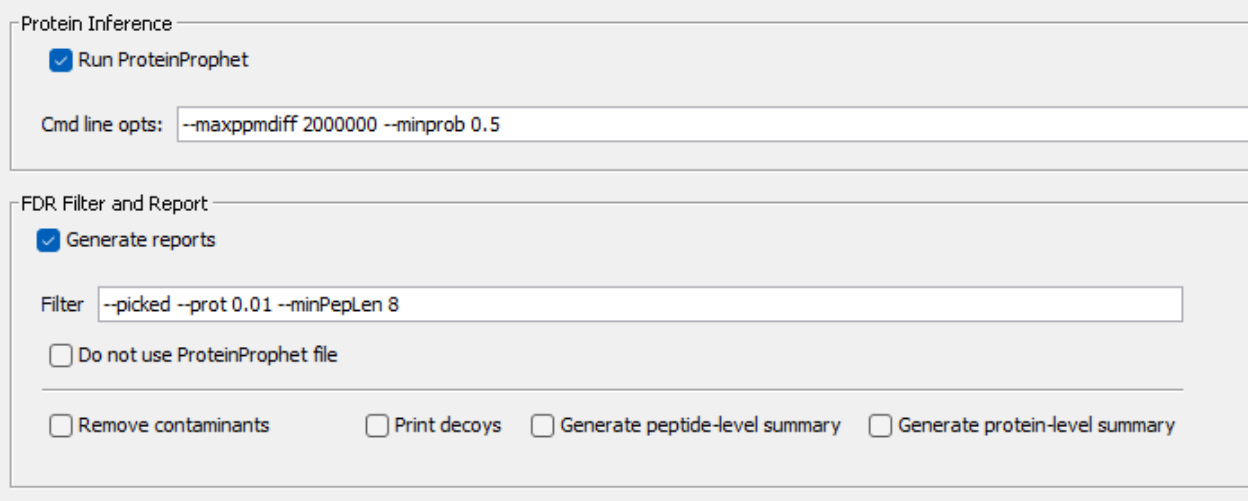
In this process, MSBooster will first use the DIA-NN deep learning model to predict additional features of the identified peptides including fragmentation spectra and retention time. Note that other models of peptide retention time and peptide fragmentation are available.



These features will be used to modify the initial identification scoring, and then Percolator will use them to improve its discrimination model to increase the number of confident identifications in the DIA dataset.

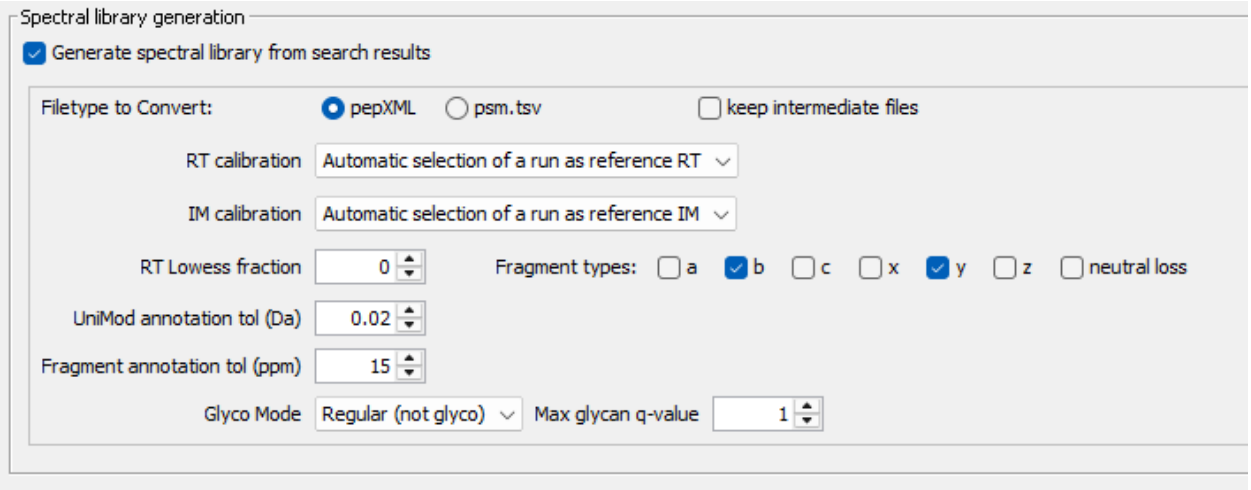


Finally, based on the identified peptides we will run the Protein Inference together with ProteinProphet to generate a confident list of protein groups identified in the sample at a maximum of 1% false discovery rate and a minimum peptide length of 8 amino acids.



Parametrization of the *Spec Lib* section

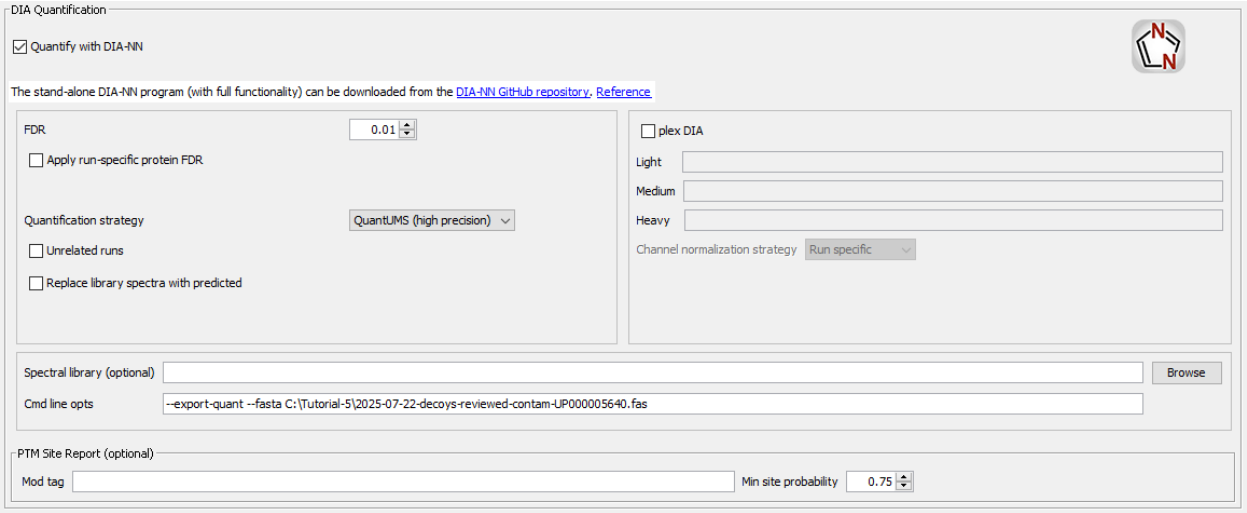
Next, we will jump directly to the `Spec Lib` tab as the other ones (`PTMs`, `Glyco`, `Quant (MS1)`, and `Quant (isobaric)`) are not relevant for the selected workflow and will not be executed. In the Spec Lib section we will generate a spectral library in pepXML format from the search results, containing b and y fragment ions, allowing for automatic selection of the runs to be used as reference for the retention time.



Parametrization of the *Quant (DIA)* section

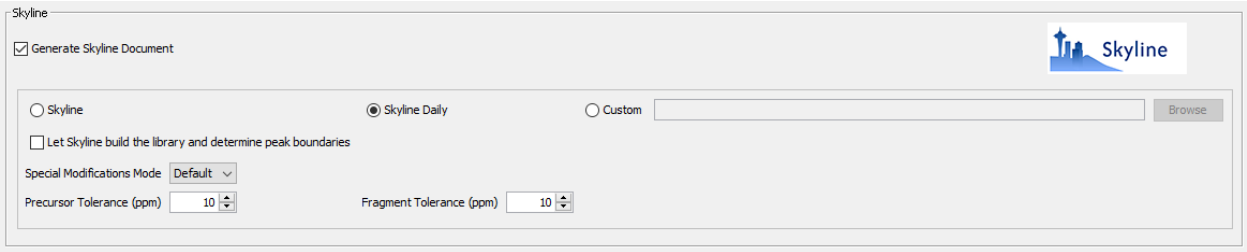
In the `Quant(DIA)` section, we will configure the quantification to be conducted by DIA-NN, setting a maximum false discovery rate of 1%. Additionally, we will include the command "--export-quant" and indicate the path to the fasta file with the command "--fasta C:\Webinar26\2025-07-22-decoys-reviewed-contam-UP000005640.fas" in the `Cmd line opts` section below. These commands instruct DIA-NN to export quantitative values at the fragment level and annotate the proteins names, which are essential steps in case the DIA-NN results are further analyzed with the MSstats statistical framework (https://msstats.org/).

Note: Msstats is an R package for statistical inference of proteomics data and can be accessed through the web at www.msstatsshiny.com. Kohler D, et al. MSstats Version 4.0: Statistical Analyses of Quantitative Mass Spectrometry-Based Proteomic Experiments with Chromatography-Based Quantification at Scale. J Proteome Res. 2023 May 5;22(5):1466-1482. doi: 10.1021/acs.jproteome.2c00834.



Parametrization of the *Skyline* section

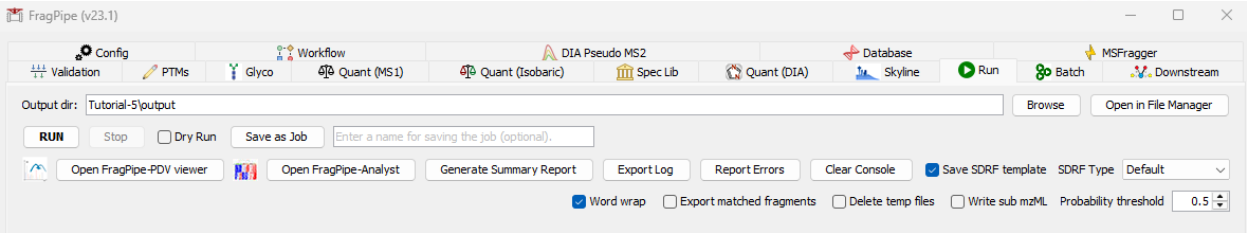
In this section we instruct FragPipe to generate a Skyline Document for the Skyline daily version.



Parametrization of the *Run* section

In this final section, we will indicate the output directory and run the analysis.

- Create a new folder in “Webinar26\” called `output` and set is as the output directory of the results.
- Click “RUN”. The analysis will now be launched, and it will take about 30 minutes to complete all the tasks. Once finished, do not close the FragPipe window. We will need it later to visualize the results.



Exploration of the FragPipe main results tables

In this second part of the tutorial, we will go through the main results tables generated by FragPipe and some of its intermediate files.

Inspection of the FragPipe main output

- Go to Webinar26\output and locate the “diann-quant-output” folder that has been generated by FragPipe.
- Inside the “diann-quant-output” folder locate the `report.pg_matrix.tsv` file and open it in Excel to inspect the protein-level output from the DIA-NN quantification module. You will see columns with the information such as protein group identifiers, gene names, and intensities from DIA runs calculated with the MaxLFQ algorithm embedded in DIA-NN.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Protein.Gi	Protein.N	Genes	First.Prote	N.Sequen	N.Proteot	C:\Users\	C:\Users\	C:\Users\	C:\Users\	C:\Users\	C:\Users\	C:\Users\	C:\Users\	C:\Users\	C:\Users\
2	A0A075B6H9		IGLV4-69		2	2	3.62E+07	1.04E+07	9.09E+06	4.15E+07	1.24E+07	4.56E+07	7.71E+06	5.51E+07	3.78E+07	1.12E+07
3	A0A075B6I1		IGLV4-60		2	2	1.82E+07	777622	2.76E+06	9.29E+06	5.44E+06	1.03E+07		8.85E+06	7.42E+06	
4	A0A075B6I7		IGLV5-48		1	1	881128	1.19E+06	457158		661236	1.83E+06	669138	962258	4.32E+06	
5	A0A075B6K5		IGLV3-9		1	0	3.40E+06	4.69E+06	292372	672406	590654	2.32E+06	253115	804634	2.01E+06	156050
6	A0A075B6S2		IGKV2D-29		1	0	1.02E+08	4.55E+07	1.04E+07	5.21E+07	8.72E+07	6.97E+07	1.02E+07	7.11E+07	1.12E+08	
7	A0A087WSX0		IGLV5-45		2	0	1.12E+07	1.54E+07	1.33E+07	3.40E+07	2.35E+07	3.30E+07	6.22E+06	2.48E+07	1.17E+08	1.89E+07
8	A0A0C4DH30		IGHV3-16		1	1	188564	715053	57310.5	569411		254652				249511
9	A0A0C4DH34		IGHV4-28		1	1	1.93E+08	3.04E+07	2.50E+07	1.87E+08	4.53E+07	1.14E+08	2.06E+07	8.85E+07	9.34E+07	3.70E+07
10	A0A0C4DH35		IGHV3-35		1	1	2.29E+07	5.15E+06	7.44E+06	1.41E+07	8.32E+06	3.67E+07	3.85E+06	2.02E+07	1.58E+07	5.78E+06
11	A0A3B3IRV3		MCTS2		1	0	2.14E+06	1.37E+06	1.83E+06	1.99E+06	1.49E+06	1.64E+06	1.63E+06		2.07E+06	1.64E+06
12	A0M8Q6		IGLC7		2	2	1.20E+08	2.31E+08	7.50E+07	1.41E+08	5.82E+07	1.84E+08	3.06E+07	3.16E+08	2.75E+08	3.89E+07
13	A1X283		SH3PXD2B		1	1					228449		72823.6			
14	A2PYH4		HFM1		1	1	1.09E+07	3.74E+06	1.64E+07	1.87E+07	3.51E+06	4.17E+07	2.23E+06	2.24E+07	1.14E+07	4.96E+06
15	A5D8V6		VPS37C		1	1	44015.2	36630.2	41823.7		27605.1	23567.6	52884.1	26211.2		21179.8
16	A6NDG6		PGP		1	1	170442	477806	572668		348631	814463	233671	684912	942986	
17	A6NIZ1		RAP1BL		2	0	4.94E+08	4.69E+08	2.87E+08	4.50E+08	3.75E+08	5.06E+08	3.70E+08	4.04E+08	4.02E+08	3.93E+08
18	B1AJZ9		FHAD1		2	2	433134		1.01E+06	5.22E+06			1.01E+06	3.49E+06	319954	1.19E+06
19	B2RPK0		HMGB1P1		2	0	1.49E+06	816174	2.12E+06	1.05E+06	1.29E+06	889203	885724	920782	1.06E+06	
20	B9A064		IGLL5		1	0				2.63E+09						
21	O00139		KIF2A		5	4	1.44E+07	1.05E+07	1.24E+07	1.14E+07	1.30E+07	1.17E+07	1.26E+07	1.25E+07	1.15E+07	1.27E+07

Inspection of intermediate FragPipe output files

If you are curious, you can explore FragPipe output files in Webinar26\output to get a better understanding of various FragPipe modules.

- Open the `psm.tsv` file with Excel and inspect the information that you have for each peptide-spectrum match. You will see the “SpectralSim” column, which indicates how well each PSM’s experimental fragment intensities match predicted intensities from the spectral prediction model; this is the ‘spectral entropy’ score, a value between 0 and 1, with 1 being a perfect match. You will also see the “RTScore”, which shows how much the experimental retention time of each PSM deviates from what is expected based on retention time predictions; the lower the value, the better.

- Open now the `.png` files inside *MSBooster*_plots both inside the *RT_calibration_curves* and *score_histograms* subfolders. In the *RT_calibration_curves* subfolder, each file corresponds to a different `mzML` file and shows the calibration curve fit between the predicted iRT scale and the experimental RT from this experiment's chromatography setup. In the score_histograms subfolder, there are the scoring distributions that use retention time and spectra entropy to differentiate between the targets and the decoy candidate peptides.
- Open the `log.txt` file with a text editor or your Notepad. In this file you will find all the commands that have been executed by the FragPipe workflow. Note the mass correction values printed at the mass calibration step of MSFragger. Inspect also the Percolator weights. The greater the magnitude of the weights, the more influence that variable has in Percolator rescoring.
- Open the `library.tsv` file with Excel. This is the library files built using EasyPQP from `PSM.tsv` and the `.mzML` files, and contains peptide ions passing 1% protein-level, peptide-level, and PSM-level FDR. For precursors identified from multiple PSMs (in the same or different runs), fragment ion intensities and retention time (after alignment to the reference run) from the best scoring PSM are used. This file is used as input to DIA-NN for extracting quantification.

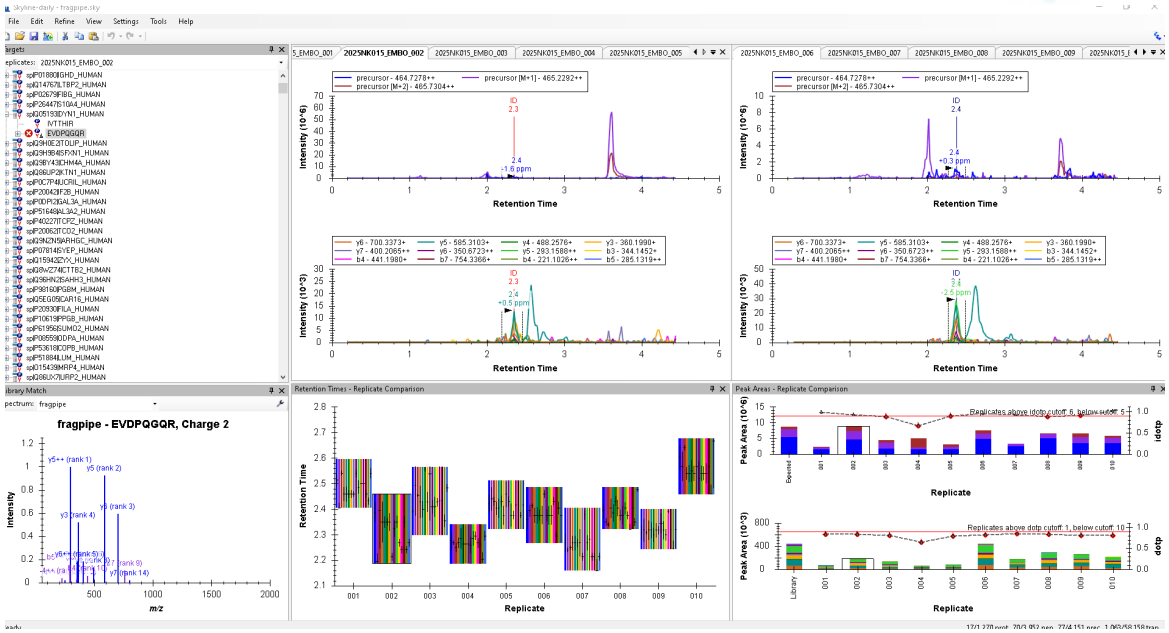
Visualization of raw data in Skyline

In this section we will use Skyline to visualize the results in terms of extracted chromatograms generated by FragPipe during the spectrum centric analysis of our dataset. Since its version 22.0 FragPipe can create a Skyline document with the results as defined in the Skyline tab during the configuration of FragPipe at the beginning of this tutorial.

Note This part of the tutorial is based in Skyline-daily 25.1.1.206

- Open Skyline.
- Select "Open file…" and browse to Webinar26\output\skyline_files and select the `fragpipe.sky` file.

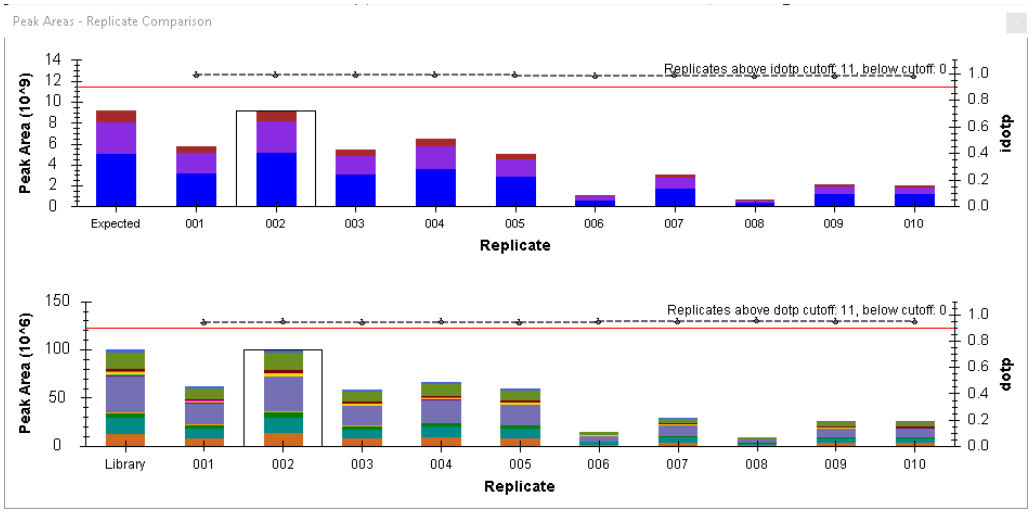
Skyline will now open a session with all the proteins and peptides identified by FragPipe. Re-arrange the panels to set your favourite window layout by simply dragging them around. See the screenshot below for an example layout.



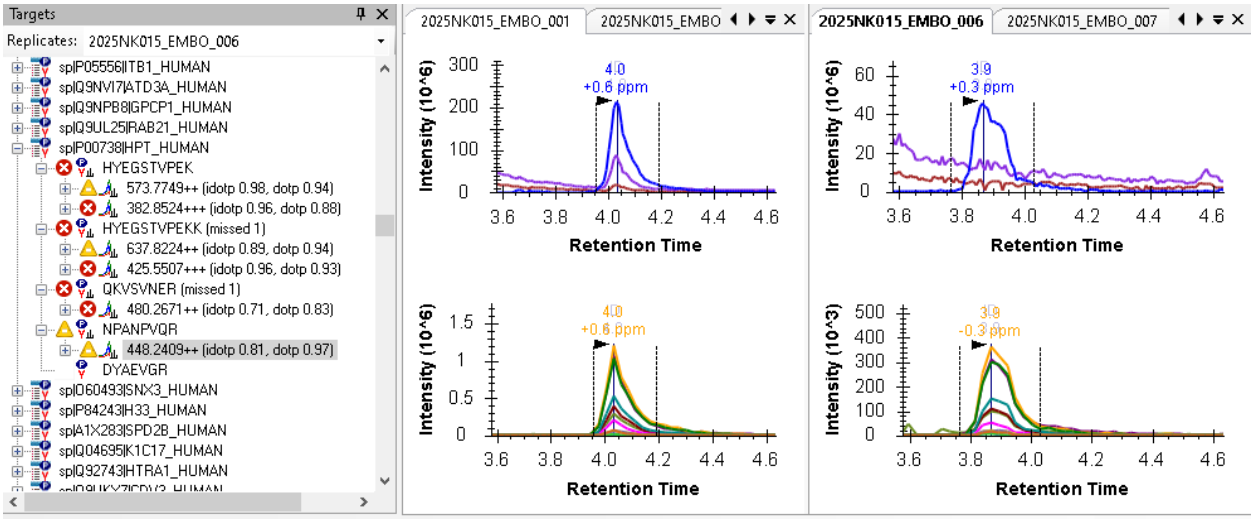
Finally, we will explore the extracted chromatogram in Skyline to inspect the intensity of the peptide HYEGSTVPEK (charge +2) from protein sp|P00738|HPT_HUMAN.

- Go to Edit → Find… and search for the peptide HYEGSTVPEK
- In the 'View' menu, select 'Peak Areas' and then 'Replicate Comparison' to visualize the intensities among all samples, to confirm the upregulation of this peptide in tumor samples. Make sure that the option "Normalized To" is set to "Default (None). You can right click in the Peak Areas plot to check it.

How does the data look like? In which samples do you have an associated identification? Do the signals in these samples look better? In which samples do the peptide have higher area values? Does this match with the statistical results obtained in FragPipe-Analyst?



- Inspect the different peptides that have been identified for this protein.
- Browse the different peptide precursors and inspect their signal in the different samples. How to they look like? How many points per peak are there?



- Select peptide AASGTQNNVLR and explore its signal in each sample. Do you see anything strange? What happens to this peptide in sample 2025NK015_EMBO_001, 003 and 010?

Finally, take your time to browse other proteins and precursors. How are their peaks? Why were they identified?

Credits

Proteomics Unit, Centre for Genomic Regulation, Dr Aiguader 88, 08003 Barcelona

Proteomics Unit, Universitat Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona

Creative Commons license