Peak List Generation

The output type of a mass spectrometer varies depending on the instrument vendor, and the first step of most workflows therefore consists of converting these raw (binary) files into a standard open format, often mzML[1]. An mzML file contains all the unprocessed spectra (MS1 and MS2) plus additional spectrum and instrument annotation.

But while mzML may be the standard format, the community often prefers the simpler mgf format for spectrum identification (see www.matrixscience.com/help/data_file_help.html#GEN). We will therefore convert our raw data into mgf, which only contains the MS/MS peak lists with some basic information about the precursors, for example (showing one spectrum):

```
BEGIN IONS
TITLE=File_A_Spectrum_1
RTINSECONDS=173.824
PEPMASS=1467.619628 1671.478512
CHARGE=3
438.253997      3469.398926
470.861908      1319.134888
        .
        .
1986.876587     2016.473755
END IONS
```

The general converter of choice is MSConvert, part of the ProteoWizard[2] package, as it supports multiple vendor formats. However, given that it requires Windows in order to to convert vendor-specific raw files, it is generally not an option for Linux and Mac. For Thermo Fisher raw files, the most common raw file format, several cross-platform converters have therefore been developed, including, but not limited to, RawTools[3] and ThermoRawFileParser[4].

Note: Below we will show how to use both ThermoRawFileParserGUI and ProteoWizard to generate mgf files. Both are good options. However, in order to get identical results as what is shown in the upcoming chapters, please use ThermoRawFileParserGUI when following the tutorial.

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

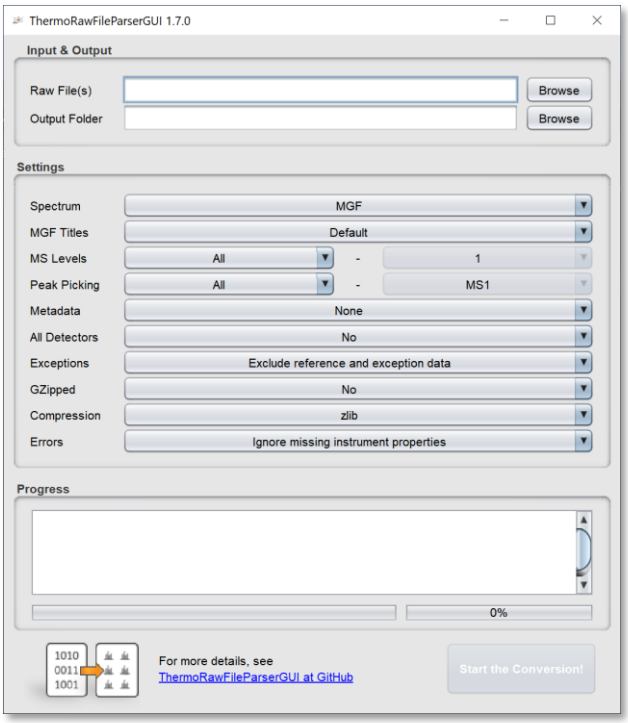
Raw Files Conversion - ThermoRawFileParserGUI

Raw file conversion will often be platform dependent as it requires vendor libraries, which for most raw file formats are only available for Windows. The exception is raw files from Thermo instruments, were several platform-independent options do exist. All based on the same ThermoFisher library.

For this tutorial we will rely on a user-friendly graphical user interface called ThermoRawFileParserGUI (https://github.com/compomics/ThermoRawFileParserGUI), basically a wrapper on top of the command lines from ThermoRawFileParser (https://github.com/compomics/ThermoRawFileParser).

In the resources folder, you will find an example file generated by a Q Exactive (Thermo Scientific, .raw file): qExactive01819.raw. (For experimental details see SupplementaryMaterial.pdf.)

Go to the Software folder and double-click the file ThermoRawFileParserGUI-X.Y.Z.jar (replace X.Y.Z with the current ThermoRawFileParserGUI version number):



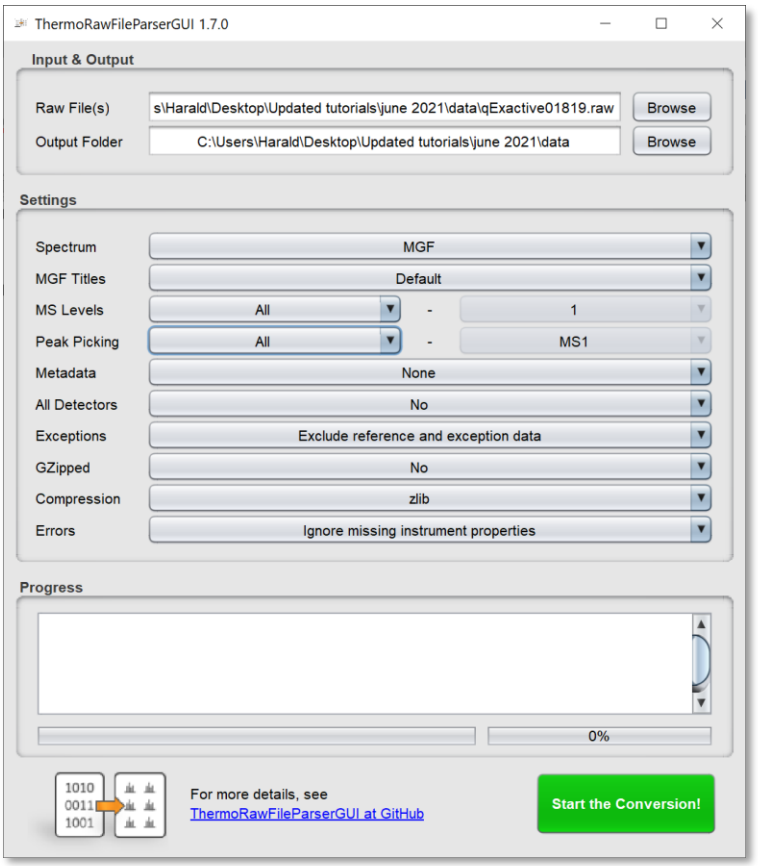
Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Using the “Browse“ button, select the qExactive01819.raw file mentioned above. Then select an output folder, and in the “Settings“ panel, make sure that mgf is selected as the Spectrum format and leave other settings as the defaults.

The data recorded by the Q Exactive is in so-called profile mode: spectra are a continuous line of data points. *The alternative to profile mode is centroid mode. How is this different from profile mode?* *[1.2a]*

In order to reduce the amount of data to interpret, we will apply a peak-picker, a program that transforms the bell-shaped profile mode peaks into single data points, i.e., a peak list.

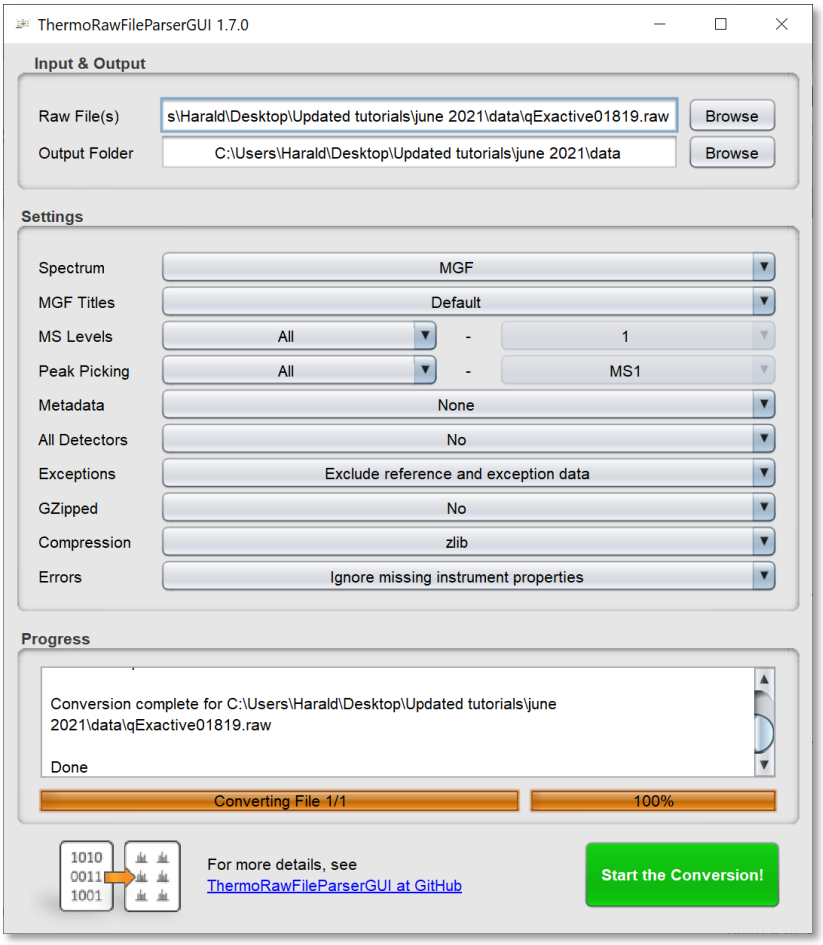
In the "Settings" panel make sure that “Peak Picking” is set to “All”. You should see the following:



How do you know whether or not you should apply a peak-picker? *[1.2b]*

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Upon clicking “Start the Conversion!“, the conversion will start and information about the progress will be displayed. When the conversion is complete, the file qExactive01819.mgf can be found in the folder specified.



Tip:

Many files will be generated as part of the spectrum interpretation workflow.

A good organization of the files will save you a lot of time!

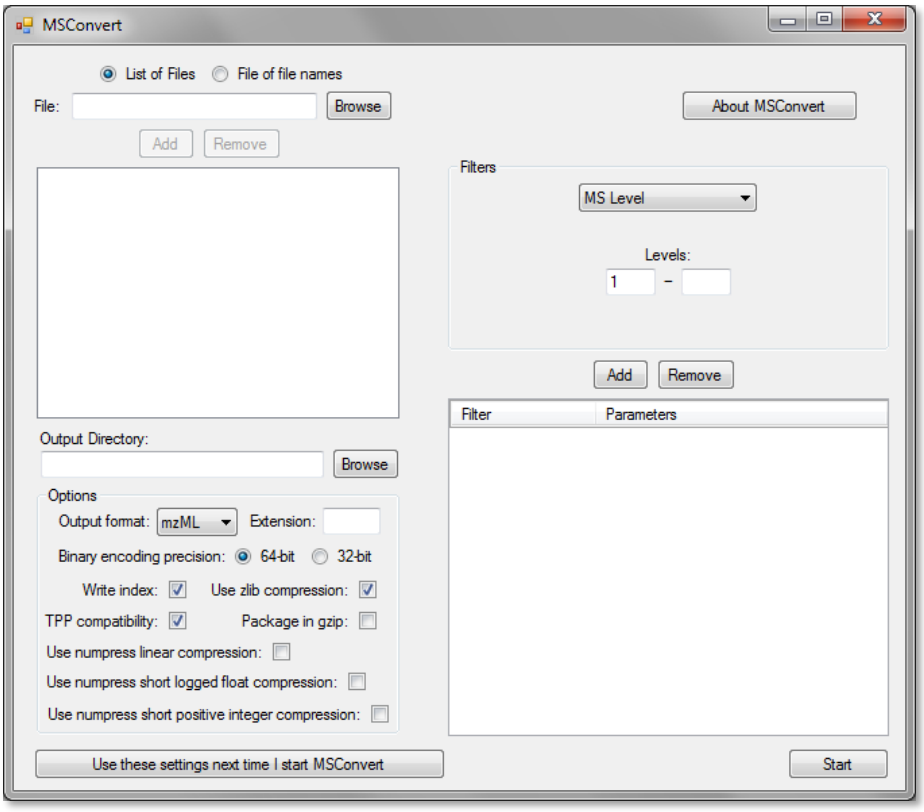
Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Raw Files Conversion - ProteoWizard

Note that this step can be platform dependent as it requires vendor libraries which, for most raw file formats, are only available for Windows. See http://proteowizard.sourceforge.net/formats.shtml for more information about the support for raw data formats for ProteoWizard.

In the resources folder, you will find an example file generated by a Q Exactive (Thermo Scientific, .raw file): qExactive01819.raw. (For experimental details see SupplementaryMaterial.pdf.)

Start MSConvertGUI.exe, delivered with the ProteoWizard package (see the software folder), you should see the MSConvert graphical user interface (GUI):



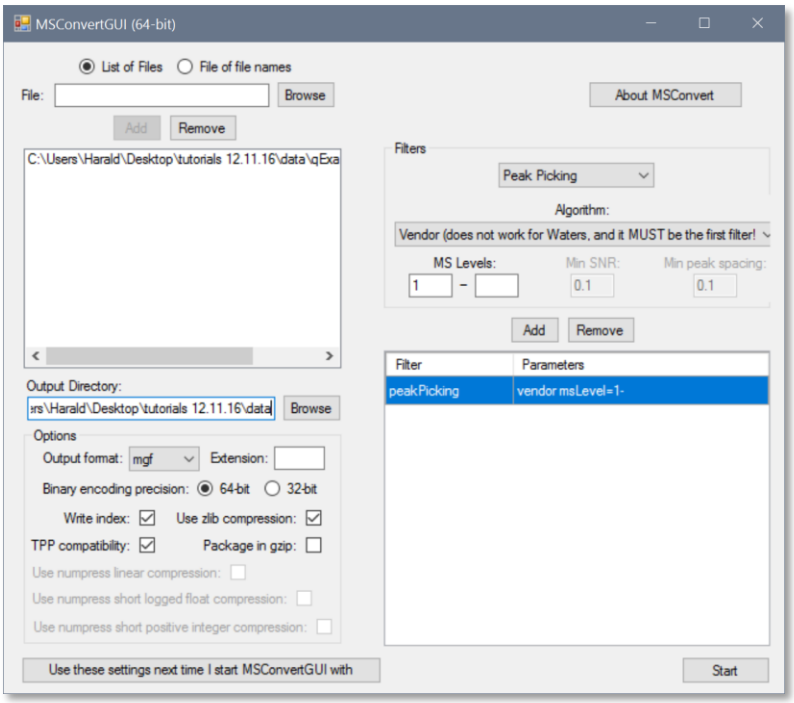
Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Using the “Browse“ button, select the qExactive01819.raw file and click on “Add“. (Note that MSConvert can process multiple files in parallel.) Select an output directory, and in the “Options“ panel, chose mgf as output format and leave other settings to default.

The data recorded by the Q Exactive is in so-called profile mode: spectra are a continuous line of data points. *The alternative to profile mode is centroid mode. How is this different from profile mode?* *[1.2a]*

In order to reduce the amount of data to interpret, we will apply a peak-picker, a program that transforms the bell-shaped profile mode peaks into single data points, i.e., a peak list.

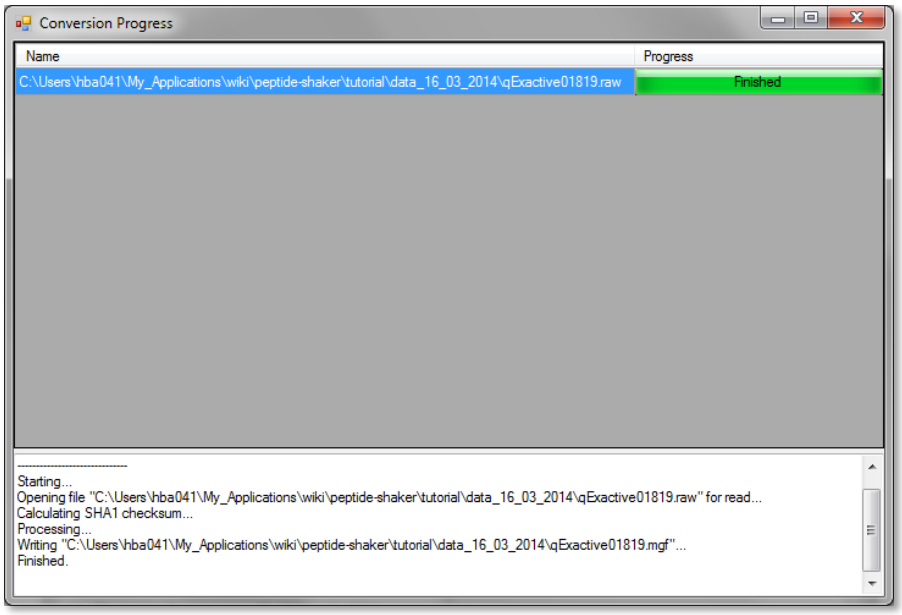
In the "Filters" panel select “Peak Picking”, choose “Vendor” as the algorithm, and click on “Add“. You should see the following:



How do you know whether or not you should apply a peak-picker? *[1.2b]*

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

When clicking “Start“, the following screen will appear and the file qExactive01819.mgf will be generated in the location specified.



Tip:

Many files will be generated as part of the spectrum interpretation workflow.

A good organization of the files will save you a lot of time!

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Advanced: Signal Processing

Depending on the mass spectrometer, the MS/MS spectra used for identification can require more advanced processing steps. Note that this step can be crucial as any imprecision made at this point will affect the rest of the workflow. One tool for processing spectra is OpenMS[5], which is a suite of tools – so called TOPP tools – dedicated to gel free proteomics.

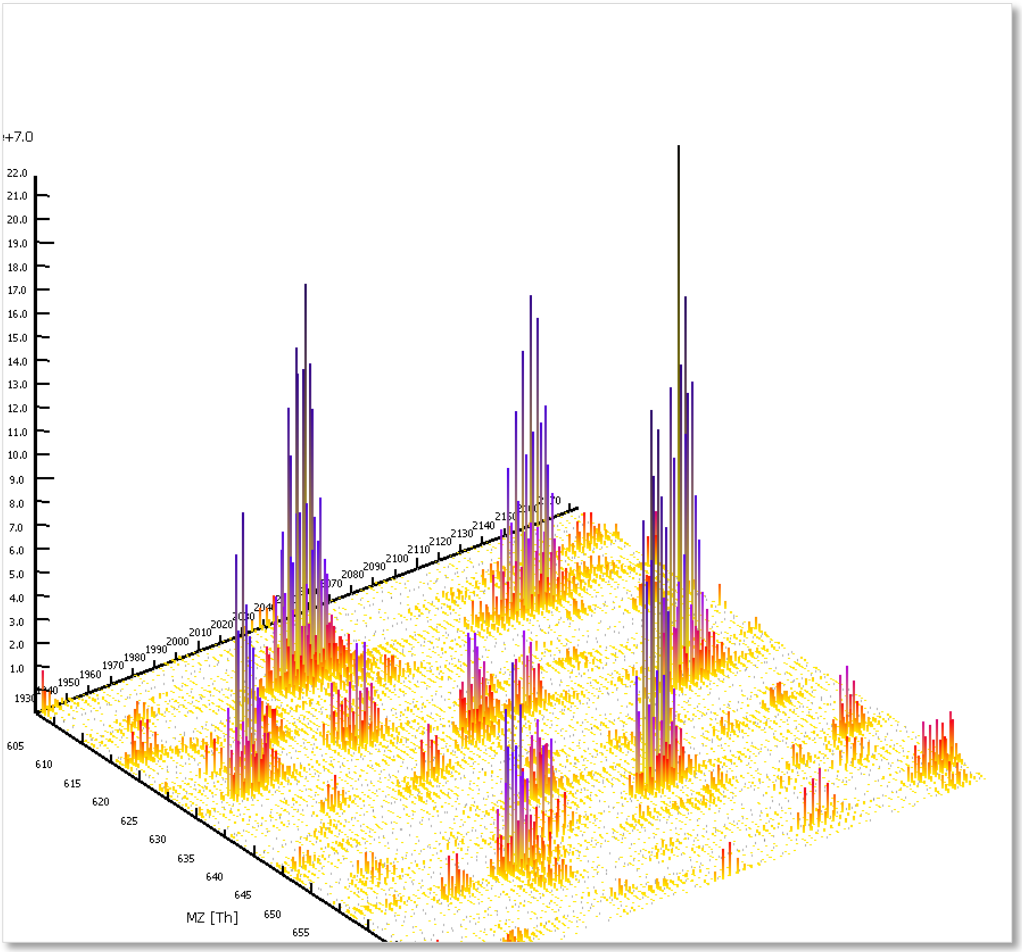
Our example dataset was acquired on a high-resolution mass spectrometer and does not require further processing.

Common processing steps for other datasets are:

- Baseline reduction: typically for TOF mass spectrometers, use this tool if the zero-intensity line of the spectrum is not stable or presents an offset.
- Noise filtering: for low resolution mass spectrometers, OpenMS provides a Savitzky-Golay[6] filter in order to reduce the noise.
- Peak picking: when the data is acquired in profile mode, every peak consists of several points which need to be summarized into one single peak before further processing. This step reduces the amount of data to be handled in the following. OpenMS provides two peak pickers: a wavelet based peak picker dedicated to low resolution mass spectrometers and a high resolution peak picker for high resolution mass spectrometers. OpenMS peak pickers are usually more efficient than the vendor's peak pickers,[7] they are thus advised for quantitative studies.[8] All these tools can be applied by the TOPPAS interface.

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Two graphical interfaces allow you to look at your data (TOPPview, see below) and to draw pipelines (TOPPAS).



Both tools include examples allowing you to familiarize with the software.

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

References

1. Martens, L. et al. mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **10**, R110 000133 (2011).
2. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534-2536 (2008).
3. Kovalchik, K.A. et al. RawTools: Rapid and Dynamic Interrogation of Orbitrap Data Files for Mass Spectrometer System Management. *J Proteome Res* **18**, 700-708 (2019).
4. Hulstaert, N. et al. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J Proteome Res* **19**, 537-542 (2020).
5. Bertsch, A., Gropl, C., Reinert, K. & Kohlbacher, O. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **696**, 353-367 (2011).
6. Savitzky, A. & Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* **36**, 1627-1639 (1964).
7. Lange, E., Gropl, C., Reinert, K., Kohlbacher, O. & Hildebrandt, A. High-accuracy peak picking of proteomics data using wavelet techniques. *Pac Symp Biocomput*, 243-254 (2006).
8. Vaudel, M., Sickmann, A. & Martens, L. Peptide and protein quantification: a map of the minefield. *Proteomics* **10**, 650-670 (2010).

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)