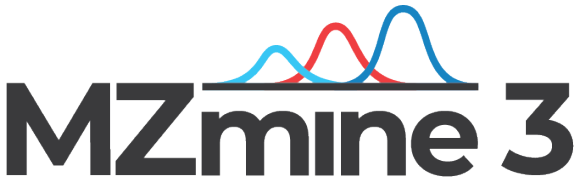
# MZmine 3 -GC-MS

Nicolas Barthes

# MZmine 3 - GC-MS

Nicolas Barthes

version 2.0




http://mzmine.github.io/

## 1 What will you learn?

This document contains the basics to compute the data obtained by gas chromatography coupled to mass spectrometry. We will learn how to convert the manufacturer files to open format files, create a reproducible pipeline to batch clean, integrate and align your chromatograms. At last, I'll give you some tips to annotate the molecules.
*I won't consider ESI, HRMS, MS2 spectra nor statistical analysis on the obtained data but* MZmine *can do it too!*

## 2 Prerequisites

You'll better understand this document if you first have some common notions about gas chromatography. For french people, I suggest you watch the FOQUAL Master youtube channel. Each year, students from Nice University are doing some short video tutorials on chemical analysis.
You'll find some help on:

- **sampling** : SBSE/HSSE, SPME, microwave liquid extraction...
- **injectors** : PTV, SSL
- **columns**

- **ionization** : chemical, electron
- **analyzers** : quad, ToF, FID
- **processing** : spectral databases

You'll find many more tutorials in english if you prefer.
Understanding how a TIC (Total Ion Current) chromatogram is built may help you to go through the deconvolution process easily.

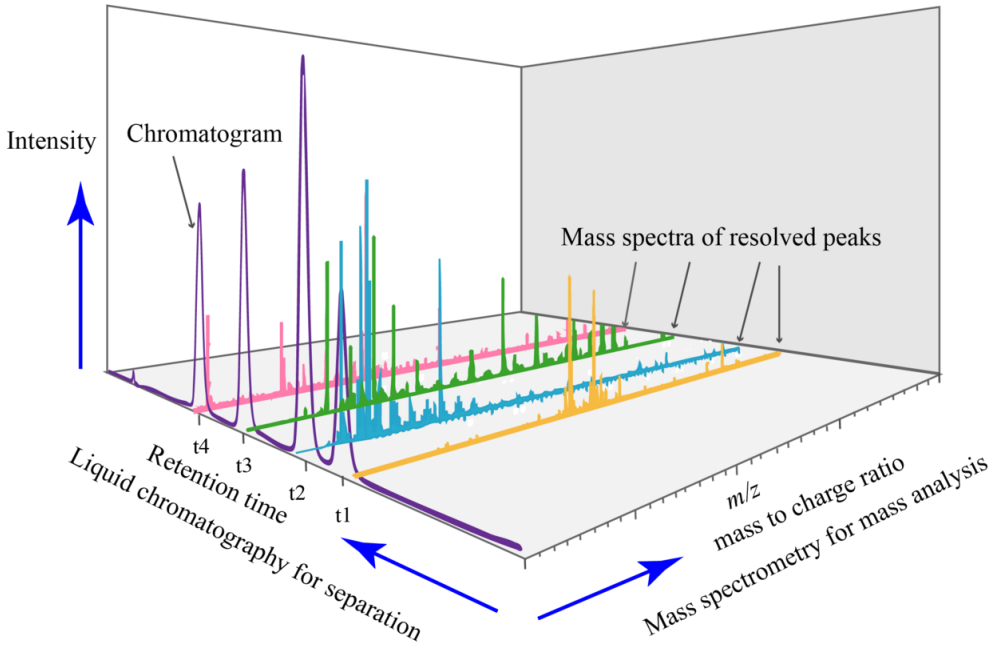



Figure 2.1: TIC 3D representation *from wikipedia*

Each point of the TIC chromatogram is a sum of all the intensities of the masses present in this particular scan mass spectra.

# 3 Formats

## 3.1 file format

Each manufacturer has its own file format which makes it uneasy for us to switch for one machine to another or compare chromatograms acquired on different machines.
Agilent use a .D (ChemStation) or .d (MassHunter) file format, Bruker a .d or .raw, Varian used a .ms or .sms file format while Perkin Elmer use a data.ms, Shimadzu a .qgd file format and Thermo a .raw - *which is incompatible with Bruker one...* Some manufacturer like LECO use a database to store data instead of files on the hard-drive.
For many of them, but not all, latest versions of their software allow to convert your analysis from the manufacturer file format to an open one.
The most used open file format for chromatography analysis are .mzML (previously .mzXML) and .cdf (or netcdf, or AIA(andi)). If your chromatography software can't export to one of those format, you can try with msconvert.exe from the **ProteoWizard** framework. A detailed

procedure - with alternatives - is given by the GNPS (Global Natural Products Social Molecular Networking) on their documentation page https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/.

### 3.2 open source softwares

If you want to process your chromatographic data without the manufacturer software, you'll have some choices: R packages as xcms or eRah - *on your computer or* via *a galaxy pipeline as the W4M* -, standalone software as OpenChrom or MZmine3 to quote the most famous.

## 4 MZMINE 3

### 4.1 installation

MZmine installation is detailed on the website: http://mzmine.github.io/download.html. Since version 2.50, the Java RunTime Environment is now included, preventing version or operating system incompatibilities.
Don't forget to install R and its packages (needed for baseline correction and stats). Path to R must be set in the *Set preferences* module of the *Project* menu.
MZmine doesn't need "admin access" as it's only an archive file to be extracted and stored wherever you want.

### 4.2 Mzmine3 workflow

I will describe a general pipeline to process gas chromatography analysis. Up to you to refine the modules or their parameters to your experimental analysis.
To practice while reading, you can download an opensource dataset from D. Touboul publication [1].
First part of the process - import & cleaning of the data - takes place in the first tab (*MS data files*) of MZmine, using the *Raw data methods* menu whereas the last part (peak detection, integration, alignment, annotations) of the process takes place in the second tab (*Feature lists*), using the *Feature detection & list methods* menus.

CLEANING OF MASS VALUES For some reason, using a simple quadrupole, some manufacturer export more than one decimal mass value for each mass unit while a single quad isn't able of such precision. The **Scan by scan filtering** module with *Round resampling filter* can parse the data and round all mass values to the nearer unit, removing or summing the possible second value.

[1] Generation of a Molecular Network from Electron Ionization Mass Spectrometry Data by Combining MZmine2 and MetGem Software, N. Elie; C. Santerre; D. Touboul, *Anal. Chem*, **2019**, 91, 11489-11492

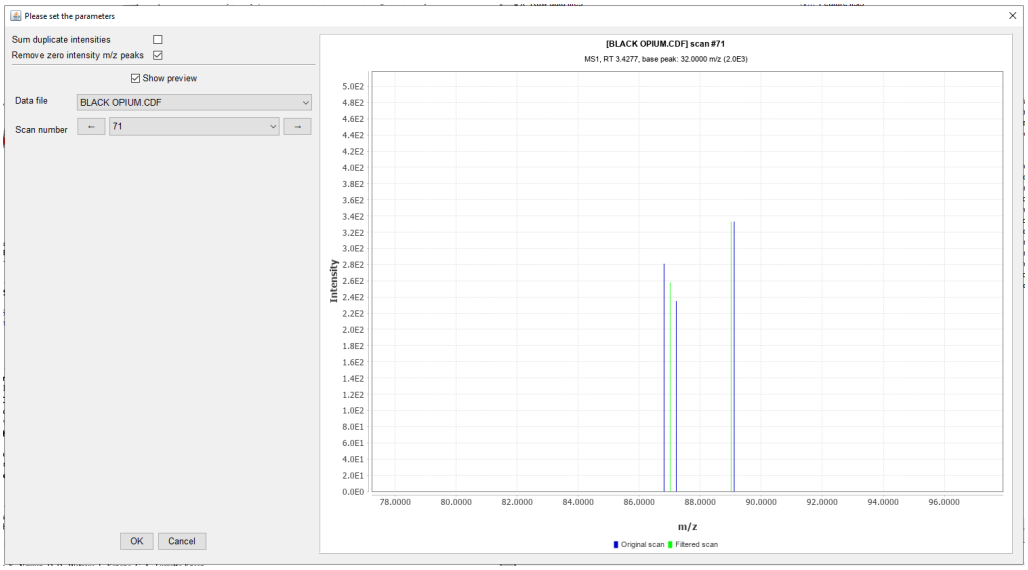



Figure 4.1: Scan by scan filtering module - *correction for ion m/z 87*

CROPPING The **Crop Filter** allows you to reduce the dataset in Retention Time range (*Scans* parameter) and/or m/z range. Working on a reduced dataset allows much faster processing.

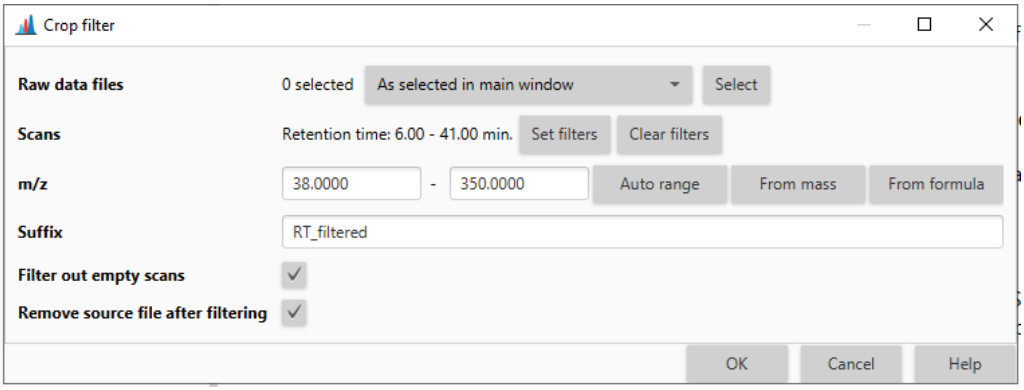



Figure 4.2: Crop module

BASELINE CORRECTION Depending on your column phase type and/or use, some drift may occur on the chromatogram baseline. It's often observed at the end of an analysis when you rise temperature to clean the column. If you can, just crop this part of the analysis but if you can't, the **Baseline correction Filter** can correct those effects. Be aware that the algorithm will remove some mass data to achieve the correction so use it with care.
This module use a connection to R and his *baseline* package through an R Engine (*Rserve* or *RCaller*). This part is often source of problems, in particular if you have more than one R version installed on your computer. As an alternative to handle the column bleed, you can also run separated mass detections in retention time intervals with different noise levels.

There are many *Correction method* algorithm. The *RollingBall baseline corrector* is an easy one to understand. A ball is rolling on the baseline of your chromatogram and the path of this ball is the reflect of the correction applied. If the ball is small enough, it will go deeper into your peaks and the correction will be strong. If the ball is big, it will just roll through the chromatogram

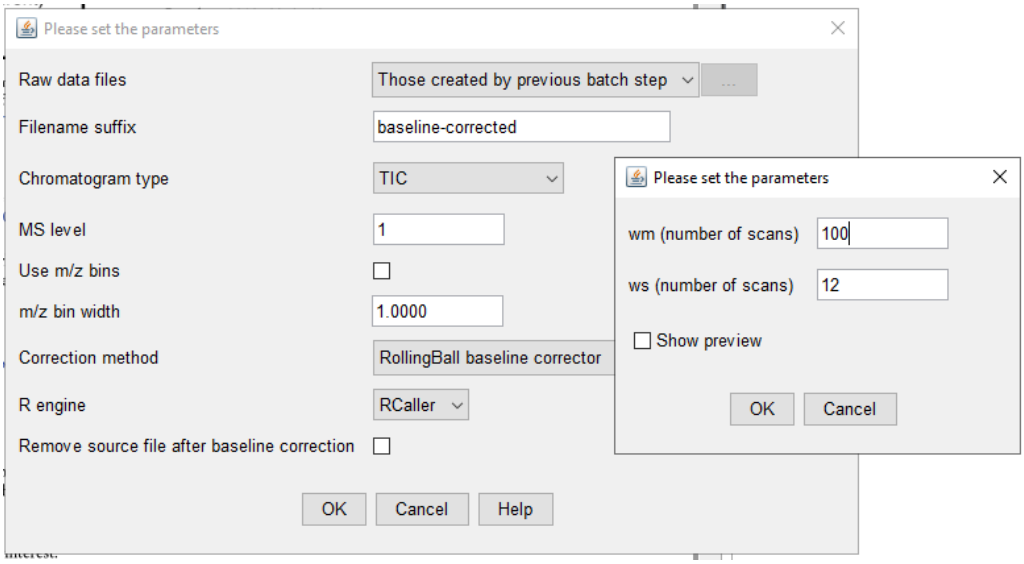



Figure 4.3: Baseline Correction module

without correcting anything. The size of the ball is set by the *wm* parameter while the *ws* parameter set the scan number used to average the new point created by the correction. Those parameters strongly depend on the acquisition parameters (scan frequency) but *wm = 100, ws = 12* is probably a good starting point. The *Show preview* option will help you to visually set up those parameters.

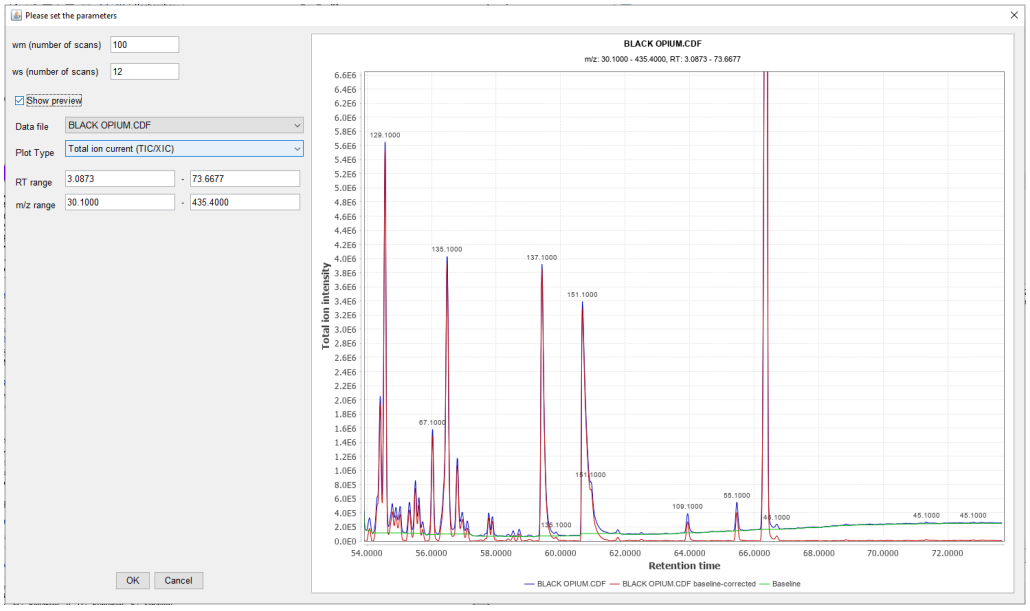



Figure 4.4: Baseline Correction wm/ws parameters

MASS DETECTION Most of the time, the mass data you'll have to deal with in gas chromatography will be in *centroid mode*. Just select this parameter in the **Mass detection** module, and set a *Noise level* threshold to delete background mass noise. Again, the *Show preview* will help you set this threshold level.
The peaks above the threshold (those with red dots on the right on Figure 4.5) will be used for

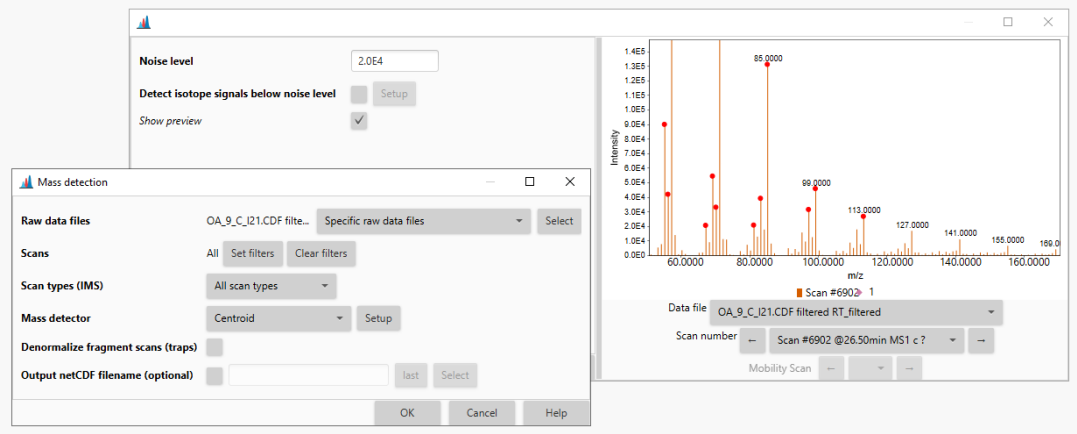



Figure 4.5: Mass detection parameters and preview

following steps. This allows faster processing as well as better accuracy for alignment or annotation.

CHROMATOGRAM BUILDER The **ADAP Chromatogram builder** module (*Feature detection* menu - *GC-MS* - *ADAP Chromatogram builder*) create a chromatogram for each m/z from the mass data detected in the previous step, plotting EICs (extracted-ion chromatograms). The ADAP Chromatogram builder module parameters are detailed in the ADAP module documentation (section 4.1) or *via* the *Help* button.
Working with the example dataset, using a minimum group size of 5 scans, a group intensity at 100 and a minimum group intensity at 300 will give satisfying results.
Working with centroid data, *m:z tolerance* should be setup around 0.5 m/z.

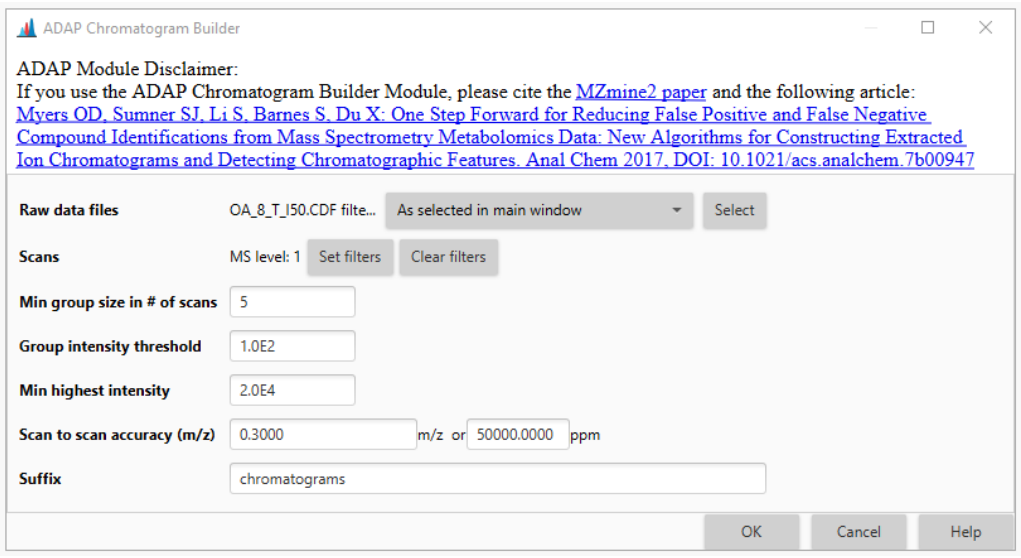



Figure 4.6: ADAP Chromatogram builder parameters

CHROMATOGRAM RESOLVING This step is generally known as "peak detection" or "peak picking". **ADAP resolver** module (*Feature detection* menu - *Chromatogram resolving* - *ADAP resolver*) is used. Again, the ADAP module documentation (section 4.2) or the *Help* button de-

tails every parameter. They are classical combination of *minimum height* to select only higher peaks and *threshold* that filters peak shape.

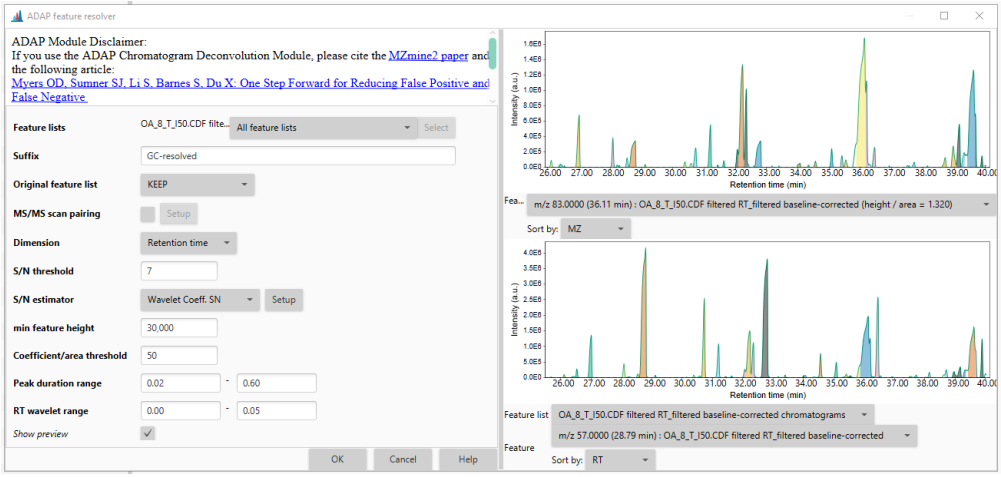



Figure 4.7: ADAP Chromatogram deconvolution parameters

SPECTRAL DECONVOLUTION **Spectral Deconvolution** (*Feature list methods* menu - *Spectral deconvolution(GC)*) detects analytes by combining synchronized mass peaks into clusters and using their intensities to construct fragmentation mass spectra. This step allows to discriminate coeluted compounds by assigning mass peaks to one or another.
Two sub-modules are available in MZmine. The *Hierarchical Clustering* module is faster with lots of parameters. The *Multivariate Curve Resolution* module is generally slower and has only a few parameters. It needs the outputs of the *ADAP resolver* and the ones from *ADAP Chromatogram builder* to process features clustering based on models.
Parameters are detailed in the ADAP module documentation (section 4.3) or *via* the *Help* button. *Deconvolution window width* is chosen based on the average width of peaks in the dataset. Generally, 0.2 min suit well to GC-MS analysis.
*Retention time tolerance* must be smaller than the window width and describe the longest drift mass peaks could have, belonging to the same analyte. *Minimum number of peaks* set the minimum mass values an analyte could have. Working with GC-MS, this number is generally greater than 2 or 3 as we are working with fragmentation spectra.

ALIGNMENT Last step to obtain an AUC (Area Under the Curve or Peak Area) matrix of analytes in your samples, the alignment uses the *ADAP aligner (GC)* or the *RANSAC aligner* if your dataset is clean enough. You'll generally be using the **ADAP aligner (GC)** for analysis sampled outside of a lab.
The *Minimum confidence* is the min percentage of samples in which the analyte must be detected. Then, you have to decide the *Retention time tolerance* you allow for an analyte in-between samples. For samples analysed in a row on the same chromatograph, this value should be around 0.1 or 0.2 minutes. If you did analyse your samples over multiple months or years, or if you want to compare samples across multiple chromatograph, then this value can be up to 1 or 2 minutes. The *Score threshold* set the minimum similarity score you want for analytes to be pooled together and *Score weight* is the balance between RT and m/z in the score calculation (1 is RT only; 0 is m/z only).

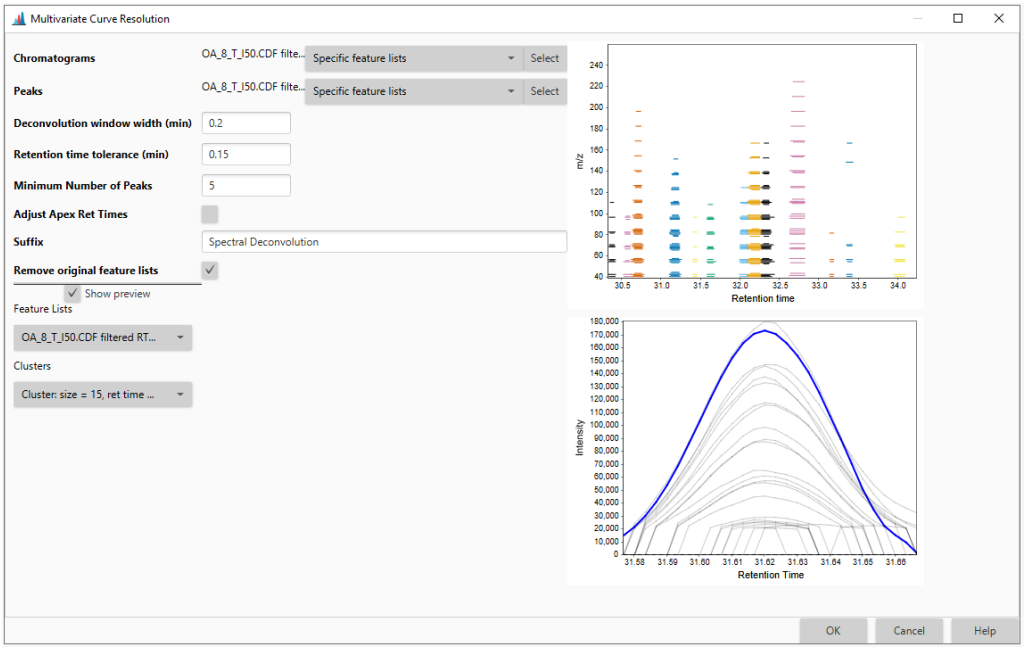



Figure 4.8: MCR Spectral deconvolution parameters

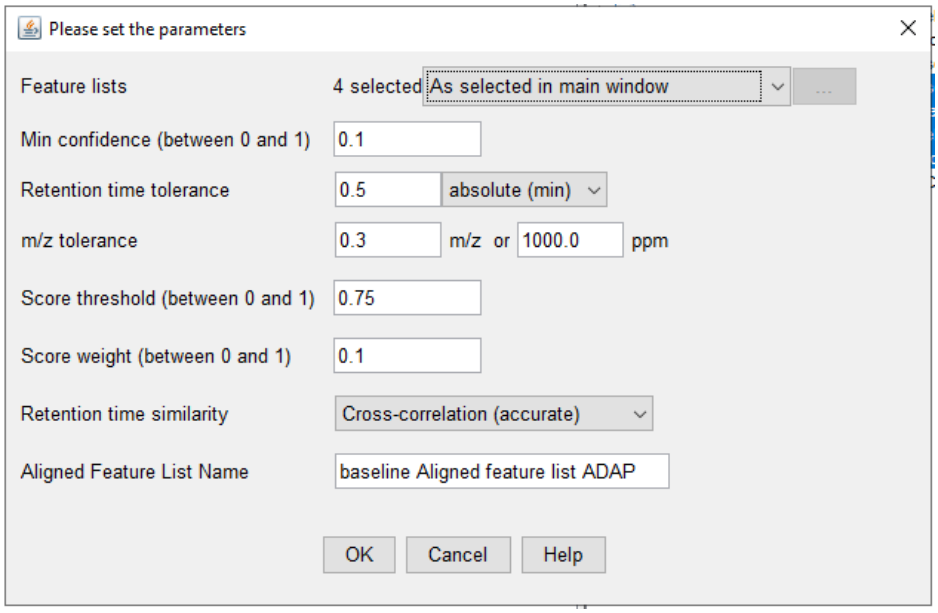



Figure 4.9: ADAP aligner parameters

BLANK SUBTRACTION Since v3.4, you don't need to filter out the compounds from the blanks or controls outside of MZmine. The module *Feature list blank subtraction* can do it for you. Given your aligned feature list, this module can check if identified features are present in raw blank/control chromatograms. You can specify the presence percentage of an 'has-to-be-removed' compound in the blanks/controls and a 'Fold change increase' that allows you to keep features more present in your samples than in the blanks/controls.
The module will export a new filtered feature list and if you ask for it, another feature list of removed compounds.

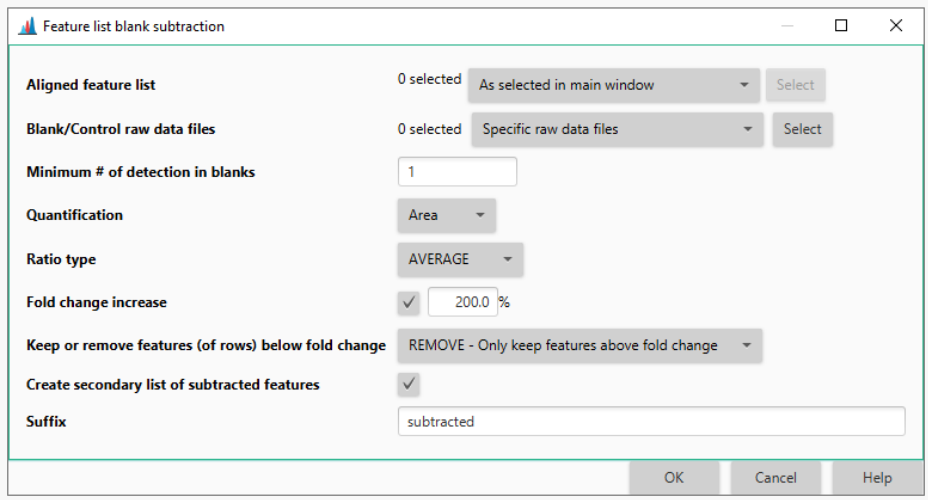



Figure 4.10: Feature list Blank Subtraction

GAP FILLING This module is optional. It checks if empty cells in the matrix means that the anayte is absent or if hasn't been deconvoluted due to the parameters you choose. **Peak finder** checks for "false negative" values.

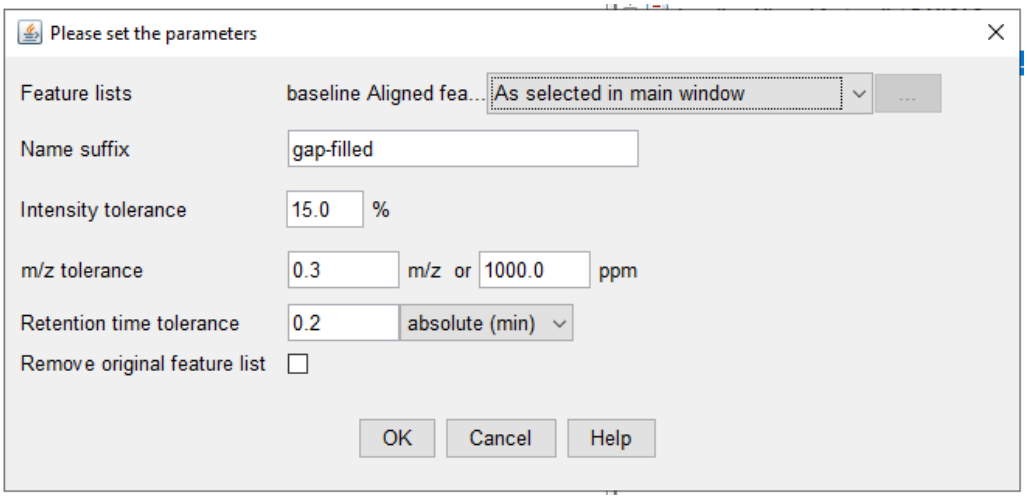



Figure 4.11: Gap Filler parameters

You simply have to set the *Intensity tolerance*, difference you allow concerning the peak shape and intensity, and the *Retention time tolerance*.

ANNOTATION Recent versions of MZmine can handle identification of fragmentation spectra obtained by EI MS although there is not as much options as for ESI MS*n* spectra.
If you are running MZmine in a Windows OS, you can use a direct connexion to the "NIST MS search" program using the *NIST MS search* module (*Annotation* menu - *Search spectra*). An easiest way to run identifications across different computers and OSs is to use the *Spectral library search* module. This one use local files as spectral databases: you can make your own database, selecting the compounds you are looking for (targeted search) or that might be present in your samples (untargeted search), depending on the biological model you are working on. Several spectra file formats are used by MZmine : MoNA json, NIST msp, GNPS json (internal library

submission format) or JCAMP-DX jdx.
If you don't own spectra libraries, you can download some at *Fiehn Lab* for example.
You can also access free spectra at the *NIST WebBook* or the Golm Metabolome Database[2].
Those databases can be imported to the third MZmine tab *via* the *Annotation* menu - *Import spectral libraries* or the *Raw data methods* menu - *Raw data import - MS data (advanced)*.

Last, MZmine offers you a powerful **Kovats Index extraction** module in the *Tools* menu. This video will show you how to use it. It's quite visual and easy. Note that you can combine 2 analyses to extract all the RI at once, useful if you have several complementary alcane mixes.

EXPORT The **Feature list methods** menu - *Export feature list* is feature rich! You can export the average mass spectra of the analytes for further processing: statistical analysis, annotations, online import (GNPS, MetaboAnalyst, SIRIUS), molecular network... A CSV export allows you to choose which data and metadata to export for statistical analysis.

## 4.3 Batch processing

All the MZmine modules can be compiled in a batch file describing the whole pipeline. You can use this batch file to share and publish your analytical workflow, or to run your process on a cluster if you have lots of samples and analytes.

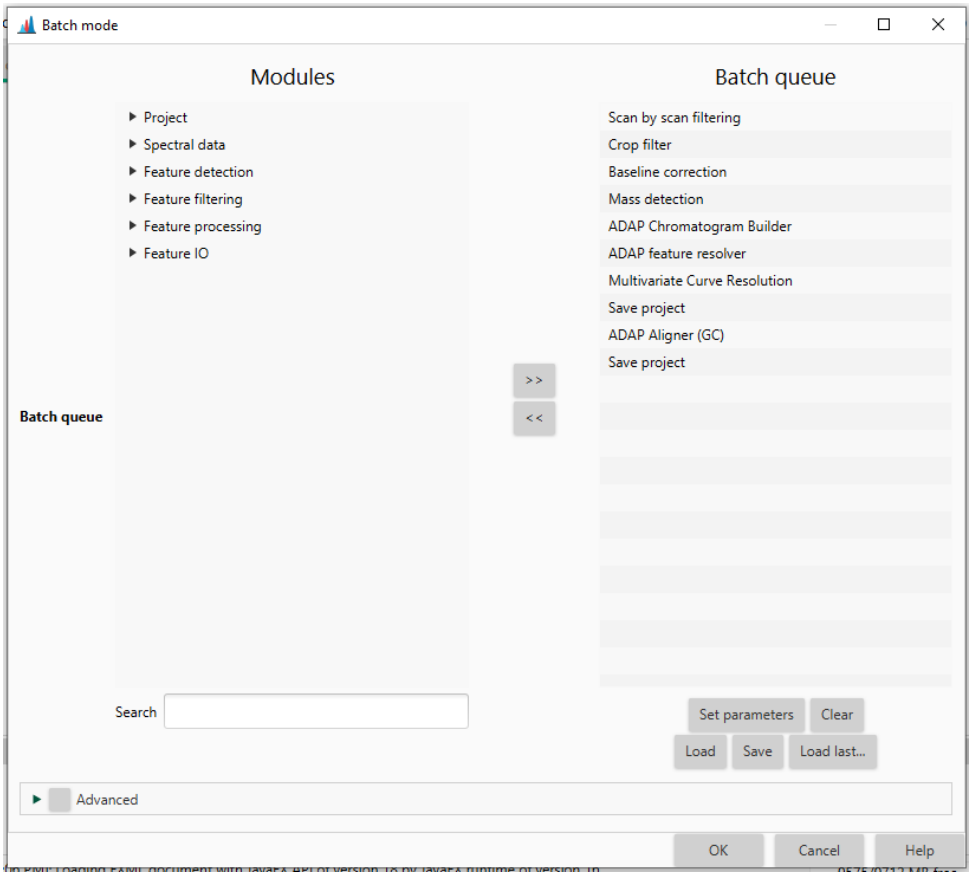



Figure 4.12: Batch window

Just *Add* > > the modules in the desired order and set them up with the *Set parameters* option,

[2]https://link.springer.com/chapter/10.1007/4735_2007_0229

the same way as seen above. You can also create a batch process from the right click on an analysis and *Show feature list summary*. This will automatically select and set up every modules used to the batch queue.

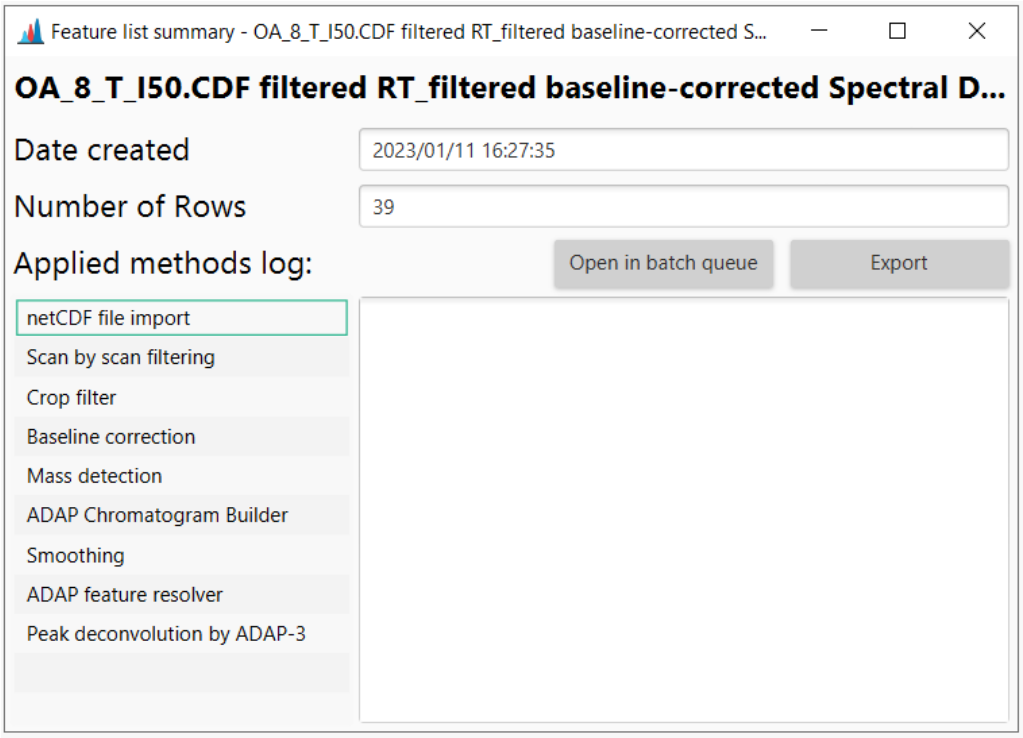



Figure 4.13: Batch creation from summary

### 4.4 Documentation

MZmine is upgraded frequently. You'll find new features every few months so stay tuned to the website or use the *Check for upgrade* in the *Help* menu.
New features are described on the MZmine Changelog page.
The main part of this tutorial is based on this publication by David Touboul and his colleagues.
The MZmine User Documentation page is now fully up to date and you'll find interesting links, especially to the ADAP Tutorial. The "International Summer School on Non-Targeted Metabolomics 2022" conference video playlist will introduce the new features of MZmine3. It is a great way to learn how to use this software.
You can also read the MetSoc forum dedicated to MZmine.

The GNPS provide great documentation working with MZmine. A major change last months is that many of the tools from the GNPS are now available for GC-MS data interpretation. You can check the GNPS documentation dedicated to GC-MS EI Data Analysis, this video and the corresponding publication.
Read this documentation carefully as it explains how to export MZmine results, import them on the GNPS online workflow, visualize its molecular network results in Cytoscape or do stat analysis with Metaboanalyst. *maybe next tutorial !*
At last, you can join an online community I manage where you can ask for help or answer questions. This app has Slack-like functions, but open-source and respectful of your privacy. Just create an account on any Matrix server and join us.